Introduction to the Special Issue on Multiword Expressions: having a crack at a hard nut

Aline Villavicencio

Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester CO4 3SQ, United Kingdom

Francis Bond

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0237, Japan

Anna Korhonen

University of Cambridge Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 OFD, United Kingdom

Diana McCarthy

Department of Informatics
University of Sussex
Falmer,
Brighton BN1 9QH, United Kingdom

Abstract

Multiword expressions are an integral part of language. Their heterogeneous characteristics have proved a challenge to both linguistic and computational analysis. Their importance to language technology has long been recognised. In this special issue we include ten papers which propose a variety of approaches for finding and handling these expressions, both for building general purpose lexical resources and in the context of specific applications. In this introduction we give a brief summary of what multiword expressions are, the challenges that they pose and some open areas of research. We then highlight the contributions that the ten papers make to these areas.

Preprint submitted to Elsevier Science *9 March 2005*

Key words:
Multiword Expressions, Collocations, Idioms, Noun-Noun Compounds, Verb-Particle Constructions

1 Introduction

A multiword expression (MWE) is an expression for which the syntactic or semantic properties of the whole expression cannot be derived from its parts. This definition covers a large number of related but distinct phenomena, such as phrasal verbs (e.g. *add up*), nominal compounds (e.g. *telephone box*), institutionalised phrases (e.g. *salt and pepper*), and many others. They are used frequently in everyday language, usually to express precisely ideas and concepts that cannot be compressed into a single word. They are syntactically and/or semantically idiosyncratic in nature but can have a great deal of flexibility and variation in their form, with complex interrelations that can be found between their components. For instance, some MWEs are fixed in the sense that they do not present internal variation, such as *by and large* and *ad hoc*, whilst others are much more flexible and allow different degrees of internal variability and modification, such as *touch a nerve* (*touch/find a raw nerve*) and *spill the beans* (*spill/spilt the/several/some of/all the beans*).

MWEs are a major part of language. In English, Jackendoff (1997a, 156) estimates that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. This is borne out by most on-line lexical resources where almost half of the entries are multiword expressions. For example, in WordNet 1.7 (Fellbaum, 1998b), 41% of the entries are multiword.

MWEs are challenging for both linguistic and computational work due to their heterogeneous characteristics which pose problems for successful (computational) linguistic treatment (Sag et al., 2002). However, the importance of MWEs and their impact in linguistics and natural language processing (NLP) has long been recognised. In linguistics, for example, they have been often used to validate the properties of grammatical theories (e.g. should a syntactic theory include transformational operations or not? (Nunberg et al., 1994)). In NLP applications such as machine translation, recognition of MWEs is necessary for systems to preserve the meaning and produce appropriate translations and avoid the generation of unnatural or nonsensical sentences in the target language.

Email addresses: avill@essex.ac.uk (Aline Villavicencio), bond@cslab.kecl.ntt.co.jp (Francis Bond), Anna.Korhonen@cl.cam.ac.uk (Anna Korhonen), dianam@sussex.ac.uk (Diana McCarthy).

Despite increasing interest in idiomaticity within linguistic research, there is still a gap between the needs of NLP and the descriptive tradition of linguistics. Most real-world applications tend to ignore MWEs or address them simply by listing. The problem is that an encoding that treats them as invariant strings (a *words with spaces* approach), will not adequately describe any such expression, except the simplest fixed cases such as *ad hoc* (Sag et al., 2002; Calzolari et al., 2002). Successful applications need to be able to identify MWEs, given their potential for variation, and interpret or use them in a meaningful way.

The papers in this special issue include articles with a variety of proposals and approaches for handling MWEs both when constructing general purpose lexical resources, and in the context of NLP tasks and applications. We start this introduction with a section on the definition of the term "Multiword Expression" (section 2). We describe the various types of MWEs in section 3, using a classification from (Sag et al., 2002). Section 4 identifies some open areas of research for the computational treatment of MWEs and section 5 introduces the articles in this special issues with respect to these areas.

2 What are MWEs?

The term "Multiword Expression" has been defined slightly differently by different researchers [1] . Calzolari et al. (2002) gives a general definition as "a sequence of words that acts as a single unit at some level of linguistic analysis", which in addition must exhibit (some of) the following characteristics to a smaller or greater extent:

(1) reduced syntactic and semantic transparency;
(2) reduced or lack of compositionality;
(3) more or less frozen or fixed status;
(4) possible violation of some otherwise general syntactic patterns or rules;
(5) a high degree of lexicalisation (depending on pragmatic factors);
(6) a high degree of conventionality.

Calzolari et al. (2002) consider MWEs to include fixed or semi-fixed phrases, compounds, support verbs, idioms, phrasal verbs, collocations, etc.

Alegria et al. (2004) define MWEs as referring to "both semantically compositional and non-compositional combinations, and both syntactically regular and idiosyncratic phrases" including idioms, proper names, compounds, lexical and grammatical collocations, institutionalised phrases, date and number expressions. Frequency

[1] Other terms used to refer to MWEs include "multiword expressions", "multiword units" (Dias et al., 2004) and "fixed expressions and idioms" Moon (1998).

factors can also be considered in the definition of MWEs, as for Pereira et al. (2004) "sequences of words that co-occur more often than expected by chance".

Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". The focus of this definition is on the mismatch between the interpretation of a MWE as a whole and the standard meaning of the individual words that compose the expression. Within MWEs, they include fixed and semi-fixed expressions, idioms, compound nominals, proper names, verb-particle constructions, institutionalised phrases, and light verbs. This is the definition we adopt in this special issue.

3 A Classification of MWEs

As the term MWEs is used to refer to such a heterogeneous class of phenomena, it is difficult to attribute a set of characteristics that are intrinsic to the class as a whole. In this section, we give an overview of different types of MWE, concentrating on those that are relevant for this special issue, adopting the classification and terminology used by Sag et al. (2002). The first division is into two broad classes: *institutionalised* and *lexicalised phrases*.

3.1 Institutionalised Phrases

Institutionalised Phrases refer to MWEs that are syntactically and semantically compositional but whose co-occurrence is conventionalised so that variations that might be expected, given the compositionality of the phrase, do not occur. For instance the MWE *strong tea* is formed by retaining the senses of these two words and compositionally combining them, however alternative forms (anti-collocations (Pearce, 2001)) like *powerful tea* and *potent tea* are found with extremely low or zero frequency in comparison with the dominant accepted form. There are also phrases which are institutionalised in terms of their order, that is one hears *fish and chips* much more frequently than *chips and fish* (in the UK at least).

3.2 Lexicalised Phrases

Lexicalised Phrases on the other hand correspond to those expressions that are syntactically and/or semantically idiosyncratic to some extent (e.g. *jump on the bandwagon*), or that include words that do not occur in isolation (e.g. *ad hoc*). These can be further divided according to how lexically flexible they are into **syntactically flexible**, **semi-fixed** and **fixed expressions**.

3.2.1 Syntactically Flexible Expressions

These MWEs exhibit a large range of morphological and syntactic variation. They include decomposable idioms and verb-particle constructions.

Whilst idioms are characterised by their opaque semantics and lack of compositionality and syntactic variations (e.g. *kick the bucket*), they are in fact a rather heterogeneous group of MWE and some are more flexible (e.g. *spill the beans*) and can undergo different types of syntactic variation and modification (e.g. *The beans were spilled in the latest edition of the report*). The type of syntactic variation that these idioms allow is highly unpredictable (Riehemann, 2001). However, the variation seems to be linked to their decomposability (Nunberg et al., 1994) in the sense that many idioms seem to be compositional if we consider that some of their component words have non-standard meanings. Then, using compositional processes, the meaning of a **semantically decomposable idiom** can be derived from the meanings of its elements. One example is *spill the beans*, where if *spill* is paraphrased as *reveal* and *beans* as *secrets*, the idiom can be interpreted as *reveal secrets*. On the other hand, an idiom like *to kick the bucket*, meaning *to die*, which only allows morphological inflection, is non-decomposable according to this approach.

Verb-particle constructions (VPCs) are formed by the combination of a verb and one or more particles (e.g. *eat up, make up, fall down, fire away, come up with...*) (Bolinger, 1971; Fraser, 1976; Jackendoff, 1997b; Dehé, 2002), where the particle may come immediately after the verb (e.g. *tear up the letter*), or they might be separated by an NP (e.g. *dance the night away*). Some VPCs require the particle to be adjacent to the verb (e.g. *come up with an idea* vs. **come with an idea up*) while others accept both forms (e.g. *clean up the place* vs. *clean the place up*). In addition, they may allow some adverbs to intervene between the verb and the particle (e.g. *hand right in*). VPCs can occur in many different subcategorisation frames, ranging from intransitive VPCs (e.g. *shut up*) to VPCs with sentential complements (e.g. *find out that he went to the office*), and their semantics can range from opaque cases (e.g. *cock up* "ruin") to more compositional ones (e.g. *carry up*) with variable productivity (e.g. *clear/vacuum/clean/polish up*).

3.2.2 Semi-Fixed Expressions

Some MWEs are much more rigid in terms of word order and compositionality, but still allow a certain degree of lexical variation, such as the choice of a determiner and the possibility of undergoing morphological inflection. This is enough to create problems for a simple encoding that merely lists them using a words with space approach, since all possible variations of these MWEs would have to be included. This class includes compound nominals and non-decomposable idioms.

Apart from allowing morphological inflection, **compound nominals** are syntactically inflexible MWEs. Thus, a compound such as *coffee machine* can be inflected

for number (*coffee machines*), but it cannot be e.g. internally modified (**coffee powerful machine*). For some compounds the head is the leftmost element inflecting accordingly (e.g. *attorney(s) general*, while for others it is the rightmost element (e.g. *coffee machine(s)*).

Non-decomposable idioms are semantically opaque, and are very rigid in terms of word order. Their semantics cannot be straightforwardly inferred from the meanings of the component words. For example, the idiom *kick the bucket* meaning *to die* cannot be passivised (**the bucket was kicked by him*), topicalised (**that bucket, he kicked*), or internally modified (**he kicked many buckets*), but it can vary as to the morphological inflection of the verb *kick*: *kick/kicks/kicked the bucket*. Other idioms like *wet oneself* also allow variation in the reflexive form (e.g. *wet herself/himself*) (Sag et al., 2002).

3.2.3 Fixed Expressions

This term is used to describe fully lexicalised invariant expressions like *ad hoc*, *in addition* and *as well as* that do not follow grammatical and compositional processes. Thus, such expressions do not allow morphosyntactic variation or internal modification (e.g. *in addition/*in some addition/* in additions*). These MWEs can be more straightforwardly accommodated in current language technology: for example, in terms of encoding, they can be handled even using a simple word-with-spaces approach.

4 Current Research on MWEs

There has been a growing awareness in the NLP community of the problems that MWEs pose and the need for their robust handling. This interest is very much related to the relevance of MWE for applications such as information retrieval, machine translation, question-answering, word sense disambiguation (WSD), and text summarization. Several projects have focused on MWEs, including the XMELLT project Cross-lingual Multiword Expression Lexicons for Language Technology (`http://www.cs.vassar.edu/ ide/XMELLT.html`), the project Collocations in the German Language of the 20th Century (`http://www.bbaw.de/forschung/kollokationen/index.html`) and the MWE project (`http://mwe.stanford.edu/`). In addition, there have been three recent workshops on the topic, two at ACL (Bond et al., 2003; Tanaka et al., 2004) and one at LREC Dias et al. (2004).

Research has included the following topics:

- linguistic treatments for MWEs

- the construction of lexical resources, grammars and ontologies
- the definition of standards for lexical resources
- the identification and extraction of MWEs and the acquisition of syntactic and semantic information from written and spoken data
- the usage of MWEs in computational applications
- the evaluation of MWE related tasks

MWEs have been investigated from a theoretical perspective, defining appropriate linguistic descriptions for these expressions (Bolinger, 1971; Fraser, 1976; Nunberg et al., 1994; Jackendoff, 1997a; Moon, 1998; Riehemann, 2001; Dehé, 2002). Research on MWE has been done on many languages, including Basque (Alegria et al., 2004), Italian (Calzolari et al., 2002), German (Trawiński, 2003; Neumann et al., 2004), Japanese (Uchiyama and Ishizaki, 2003; Tanaka and Baldwin, 2003; Baldwin and Tanaka, 2004; Fujita et al., 2004), Portuguese (Dias, 2003; Villavicencio et al., 2004a; Baptista et al., 2004), Russian (Sharoff, 2004) and Turkish (Oflazer et al., 2004). Efforts have also been made to investigate MWEs from a multilingual perspective (Calzolari et al., 2002; Villavicencio et al., 2004a), since languages differ as to what subtypes of MWEs they have, and how they are realised in other languages. For example, the English idiom *in the red* has an idiomatic equivalent in Portuguese, *no vermelho* (*in+the red*) that is almost an exact translation and has very similar characteristics (Villavicencio et al., 2004a). However, while some MWEs have (multiple) equivalents in a second language, others have none, and this information is also of importance (see Tanaka and Baldwin (2003) for a discussion of English and Japanese compound nouns in the context of a machine translation task).

There has also been an immense amount of work on multiword expressions in the field of terminology (e.g. the Computerm workshops and many domain specific workshops). The emphasis is generally practical, with special emphasis on standardisation of terminological resources and constructing and maintaining domain specific dictionaries and thesauri. Most work has been done on nominal expressions, as they are both common and informative. For information extraction and data mining, mapping terminological variants to a canonical form, and/or clustering terms is an important part of identifying relevant knowledge. There is also considerable work, particularly in the medical domain, on normative use — using NLP techniques to identify canonical terms and then encouraging their use.

Different strategies for encoding MWEs in lexical resources have been employed with varying degrees of success, depending on the type of MWE. One case is the Alvey Tools Lexicon (Carroll and Grover, 1989), which has a good coverage of phrasal verbs, providing extensive information about their syntactic aspects (variation in word order, subcategorisation, etc), but which does not distinguish compositional from non-compositional entries or specify entries that can be productively formed. WordNet, on the other hand, covers a large number of MWEs (Fellbaum, 1998a), but does not provide information about their variability or type. Neither of

these resources covers idioms.

The challenge in designing adequate lexical resources for MWEs is to ensure that the variability shown by the different types of MWE can be captured. Such a move is called for by Calzolari et al. (2002) and Copestake et al. (2002). Calzolari et al. (2002) discuss these problems in the context of establishing standards for MWE description for multilingual lexical resources. They focus on MWEs that are productive and present regularities which are applicable to other classes of words with similar properties. Villavicencio et al. (2004b) propose an architecture for a lexical encoding of MWEs which allows for a unified treatment of different kinds of MWE (e.g. nominal compounds, verb-particle constructions and idioms).

From a more applied view, research has been done on designing resources, algorithms and systems for dealing with MWEs, sometimes in the context of applications, such as Question Answering and Machine Translation (Tanaka and Baldwin, 2003; Baldwin and Tanaka, 2004; Miyazaki et al., 1993). The identification of MWEs and the extraction of multiword expressions and collocations of certain types, such as noun-noun compounds, institutionalised expressions and verb particle constructions have received a good deal of the attention to date. Here, the flexibility of MWEs is a major factor that needs to be taken into account. The elements of a MWE may not occur adjacently to one another, but may be separated due to passivisation, lexical insertion, variation in word order, etc. For example, in the VPC *look up*, the verb and the particle may be separated by an arbitrarily long NP, as in *They will have to look the words that they cannot spell up in the dictionary*, [2] where they are separated by 6 words.

Any NLP application which involves some level of syntactic and/or semantic processing requires inventories of MWE for appropriate interpretation. In addition, ambiguity can be reduced when MWEs are identified correctly. A statistical parser can reduce ambiguity by using inventories of MWE to identify phrases which may be functioning as a MWE rather than analysing the sentence from the words as individual units. For example, in *The army blew up the bridge*, knowledge that there is a strong likelihood of the verb-particle construction *blow up* can help reduce the likelihood of analyses with *up* as a preposition. In WSD, identification of MWEs is important so that time is not wasted on resolving ambiguity for component words. Thus, in the *blow up* example, it is necessary to identify the verb-particle construction before trying to determine the sense of *blow*: there are 22 senses of the verb *blow* in WordNet 2.0, but only 7 senses of the *blow up* combination.

Investigation has also been done on the evaluation of the success of these tasks (e.g. what is the precision and recall of a given method for extraction of MWEs from corpora). For evaluation researchers have tended to rely on available man-made lexical resources or to use manual annotation of either the input data or the automatically extracted lists. As well as a binary decisions on whether something is, or is not a

[2] From `www.lessonplanspage.com/~LAWritingContractionsInAPoem35.htm`

given MWE there has also been some work on obtaining compositionality judgements from humans. There is considerable scope for further proposals of standard evaluation metrics, test and training data and for task-based evaluation, since evaluation is difficult and the problems are exacerbated because of the great variety of MWEs.

5 Introduction to the Articles in this Special Issue

There was a lot of interest in this special issue, reflecting the interest in MWEs at this time. We had 34 submissions in total, and after an extensive review process [3] we were able to select 10 papers from the initial submissions.

The task of acquisition of inventories of MWEs is essential for any NLP system requiring some level of semantic processing. The extraction, and subsequent evaluation of acquired inventories is extremely important since it is not feasible to develop such inventories manually. However, issues such as where exactly in the sentence to look for the elements of an MWE, and how far apart they are going to be found present a tough challenge for methods for the automatic identification/extraction of MWEs. Several approaches have already been proposed, both from a more symbolic perspective and from a more statistically driven view, with varying degrees of success for different types of MWE. This taks is the broad topic of the first five articles of our special issue.

Comparing and Combining A Semantic Tagger and A Statistical Tool for MWE *Extraction*
Scott Songlin Piao, Paul Rayson, Dawn Archer and Tony McEnery

Scott Songlin Piao, Paul Rayson, Dawn Archer and Tony McEnery approach the task of extracting MWEs with the combination of symbolic and statistical algorithms. They then employ a semantic analysis system which automatically annotates (English) corpora with semantic category information. In this way, through the use of a semantically classified MWE template database, the system not only extracts MWEs, but also assigns semantic information to these MWEs. The authors then explore how the addition of a statistical algorithm to the symbolic semantic analysis system based on collocational information, can improve the results obtained. As expected, these different approaches prove to have different strengths: while the symbolic system is unable to identify some domain-specific MWEs, the statistical tool has problems with low-frequency MWEs. Therefore, their combination significantly improves MWE coverage.

Deep Lexical Acquisition of Verb Particle Constructions

[3] We are extremely grateful to our reviewers for the effort that went into this process.

Timothy Baldwin

The extraction of a specific type of MWE is addressed by Timothy Baldwin, who concentrates on English verb-particle constructions. Baldwin investigates the difficult task of extracting VPCs with valence information from raw text, exploring a range of techniques. He proposes four basic methods: using the output of (a) a POS tagger, (b) a chunker, (c) a chunk grammar and (d) a dependency parser. The use of a combined classifier is found to consolidate the strengths of the component methods.

The Availability of Verb-Particle Constructions in Lexical Resources: How Much is Enough?
Aline Villavicencio

Because not all verbs can combine with (all) particles it is important to know which verbs form constructions with specific particles. Aline Villavicencio explores how lexical resources, corpora and the World Wide Web can be used to determine whether an arbitrary combination of verb and particle is valid. In particular, given that there are verbs that can productively combine with particles (e.g. *ring/call/phone up*) and that semantic properties of verbs influence their ability to combine with particles, she investigates how a classification of verbs (such as that proposed by Levin (1993)) can be used as an indicator of acceptability of combinations of verbs and particles and how the Web can be used to validate them. In this way, the evidence gathered from the Web for each VPC can help differentiate the genuine VPCs among those automatically generated.

Multiword Expressions in Spoken Language: an exploratory study on pronunciation variation
Diana Binnenpoorte, Catia Cucchiarini, Lou Boves and Helmer Strik Radboud

Unlike other works in this special issue, Diana Binnenpoorte, Catia Cucchiarini, Lou Boves and Helmer Strik Radboud investigate MWEs in spoken rather than written corpora. They use a large corpus of spoken Dutch containing orthographic transcriptions of spontaneous speech to derive an inventory of frequently found N-grams. They then look at pronunciation characteristics of MWEs in a subset of this data from the perspective of automatic speech recognition (ASR) and automatic phonetic transcription (APT). When looking at the canonical form of pronunciation for individual words, they find that there is a significant difference in the pronunciation of these words in the context of the N-grams when compared to any other context. Based on this finding Binnenpoorte et al. suggest that words in the N-grams should be treated as lexical entries with their own specific pronunciation variants in lexicons used in ASR and APT.

Using Small Random Samples for the Manual Evaluation of Statistical Association Measures
Stefan Evert and Brigitte Krenn

The issue of evaluation is an important aspect of research. However, due to the heterogeneity of MWEs, it is also a difficult problem to address, since what works for a given task and class of MWEs may not work well for another. On this subject Stefan Evert and Brigitte Krenn look into the empirical evaluation of statistical association measures for the extraction of lexical collocations from corpora. Their research indicates that the results of an evaluation experiment cannot easily be generalised to a different setting. From that perspective, when carrying out experiments it is important to ensure that the experiments are done under conditions that are as similar as possible to the intended use of the measures. Evert and Krenn also discuss how the amount of manual annotation work can be reduced by using an evaluation strategy based on random samples. This enables one to perform a greater number of evaluation experiments under specific conditions.

As well as automatically acquired lists and gold standards for evaluation, other information on MWE is also required. Semantic information is particularly needed, due to the variation in compositionality exhibited by MWEs. Some MWEs can be interpreted in a more compositional way (e.g. *coffee machine, carry up*), while others have a more idiomatic interpretation and their meaning cannot be straightforwardly inferred from their components (e.g. *kick the bucket, make up*). Even when the interpretation is compositional, there may be a variation as to precisely how much each component of a MWE contributes to the semantics of the MWE. It is not always evident even for humans whether an MWE is compositional or not. The next three articles in this special issue involve specific topics of semantic interpretation.

Learning about the Meaning of Verb Particle Constructions from Corpora
Colin Bannard

For the case of verb-particle constructions Colin Bannard looks at a distributional approach to determine when and to what extent the components of a VPC contribute their simplex meanings to the interpretation of the VPC. The basic idea is to look at the lexical contexts in which a VPC occurs and to compare it with the lexical contexts in which each component occurs. Bannard's hypothesis is that if similar lexical contexts are found for a VPC and a given component, then that component is contributing its simplex meaning to the VPC. The results he reports indicate that a correlation can indeed be found between contextual similarity and compositionality judgements.

On the Semantics of Noun Compounds
Roxana Girju, Dan Moldovan, Marta Tatu and Daniel Antohe

Roxana Girju, Dan Moldovan, Marta Tatu, Daniel Antohe focus on the semantics of noun compound MWEs. Their work involves a corpus analysis of the semantics of two and three noun compounds such as *((coffee maker) industry)*. Annotation of corpus is used to construct a mapping between two sets of semantic classification categories: a list of 8 prepositional paraphrases (e.g. *make of coffee*) and a

more finer-grained set of 35 semantic relations (e.g. MAKE/PRODUCE). The corpus data is then used to supervise models using WordNet-based WSD and lexical specialisation for semantic classification of two and three noun compounds and the bracketing of three noun compounds. The results improve on previous models which use less semantic information.

Disambiguating Japanese Compound Verbs
Kiyoko Uchiyama, Timothy Baldwin and Shun Ishizaki

The problem of WSD has been attentively investigated, and considerable advances have been made to the area. In terms of MWEs, however, there is still much work to be done. If in some cases there are advantages to working with MWEs (the different elements of an MWE can help to reduce the number of senses being considered), in others, they can bring further complexities to the task and the problem of polysemy can be as acute or worse than for simplex words. Kiyoko Uchiyama, Timothy Baldwin and Shun Ishizaki focus on the problem of sense disambiguation of Japanese compound verbs, using both statistical and symbolic methods. They first explore the use of a statistical sense discrimination method based on verb combinatoric information, and combine it with a first-sense statistical sense disambiguation method. Then from a symbolic perspective they also investigate a manual rule-based method which uses information about argument structure and verb semantics. When comparing the two methods, they find that the rule-based method with its detailed semantic information performs better than the statistical method. However, the latter performs surprisingly well and without any access to manually added syntactic or lexical semantic information to help the process.

The last two articles deal with MWEs in lexical ontologies. The construction of a lexical ontology is a demanding task which involves complex organisational decisions in order to maintain consistency throughout the ontology. This task is further complicated when dealing with MWEs, because even for a specific kind of MWE, there may be a number of alternative strategies that can be employed based on different criteria (e.g. the compositionality of the MWE and the degree of binding between its elements).

Creative Discovery in the Lexical "Validation Gap"
Jer Hayes, Nuno Seco and Tony Veale

Jer Hayes, Nuno Seco and Tony Veale look at the organisation of lexical ontologies concentrating on issues involved in the inclusion of compound terms, like completeness and consistency. They propose that for literal compounds the use of creative exploration can help to find terms that need to be added to an ontology to systematically balance it. In addition, they argue that through this process of creative exploration, it is possible to discover new concepts. They use shared modifier head relations to identify new candidate terms and use both information within the lexical resource (WordNet) and the Web (as Aline Villavicencio does) for valida-

tion.

A Symbolic Approach to Automatic MultiWord Term Structuring
Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan and Fabio Rinaldi

The relation between different words and how they can be structured in an ontology is an important question that has received considerable attention. Many terms are only used within specific domains, and their meaning can be opaque to outsiders. Further, different subdomains often use different terms, and mapping between them is important for exchanging information. Research on this area can benefit several information-oriented tasks like Text Mining and Question Answering. In terms of MWEs, Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan and Fabio Rinaldi propose a three-level structuring of MWEs based on lexical inclusion, WordNet similarity and a clustering approach. They discuss how the knowledge structures of a particular domain can be organised using a method for term clustering by automatic data analysis. To evaluate the approach proposed, a comparison is carried out between the mapping of the domain topics from a corpus of genomics and the hand-built GENIA ontology. The impact of this research for NLP applications is addressed by a discussion of how these results can be used in a Question-Answering system.

6 Conclusions

The importance of MWEs for language technology is increasingly being recognised. Whilst much needed research on extraction of MWEs is being undertaken, more effort is required to support this work. Acquisition of the phonological, morpho-syntactic and semantic properties of MWEs is essential so that lexical resources can be successfully deployed in language technology applications. Expressive representations and rigorous evaluation methodologies are required alongside proposals for acquisition. Additionally, as well as techniques for extracting MWE types for lexical resources, methods for identification of token instances in text and speech are necessary since not all candidates will be genuine MWEs. Whilst there are a great many possibilities for further research, we hope that this collection of papers describing some of these issues will lay the basis for future advances on the area.

References

Alegria, I. n., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., Urizar, R., 2004. Representation and treatment of multiword expressions in Basque. In: Tanaka et al. (2004), pp. 48–55.

Baldwin, T., Tanaka, T., 2004. Translation by machine of complex nominals: Getting it right. In: Tanaka et al. (2004), pp. 24–31.

Baptista, J., Correia, A., Fernandes, G., 2004. Frozen sentences of Portuguese: Formal descriptions for NLP. In: Tanaka et al. (2004), pp. 72–79.

Bolinger, D., 1971. The phrasal verb in English. Harvard University Press, Harvard, USA.

Bond, F., Korhonen, A., McCarthy, D., Villavicencio, A. (Eds.), 2003. ACL-03 Workshop on Multiword Expressions: Analysis, Acquistion and Treatment. ACL, Sapporo, Japan.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A., 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.

Carroll, J., Grover, C., 1989. The derivation of a large computational lexicon of English from LDOCE. In: Boguraev, B., Briscoe, E. (Eds.), Computational Lexicography for Natural Language Processing. Longman.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., Flickinger, D., 2002. Multiword expressions: Linguistic precision and reusability. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, Canary Islands, pp. 1941–1947.

Dehé, N., 2002. Particle verbs in English: syntax, information structure and intonation. John Benjamins, Amsterdam/Philadelphia.

Dias, G., 2003. Multiword unit hybrid extraction. In: Bond et al. (2003), pp. 41–48.

Dias, G. H., Lopes, J. G. P., Vintar, S. (Eds.), 2004. LREC-2004 Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications. 4th International Conference On Languages Resources and Evaluation, Lisbon, Portugal.

Fellbaum, C., 1998a. Towards a representation of idioms in WordNet. In: Proceedings of the workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL 1998). Montreal, pp. 52–57.

Fellbaum, C. (Ed.), 1998b. WordNet: An Electronic Lexical Database. MIT Press.

Fraser, B., 1976. The Verb-Particle Combination in English. Academic Press, New York, USA.

Fujita, A., Furihata, K., Inui, K., Matsumoto, Y., Takeuchi, K., 2004. Paraphrasing of Japanese light-verb constructions based on lexical conceptual structure. In: Tanaka et al. (2004), pp. 9–16.

Jackendoff, R., 1997a. The Architecture of the Language Faculty. MIT Press, Cambridge, MA.

Jackendoff, R., 1997b. Twistin' the night away. Language 73, 534–59.

Levin, B., 1993. English verb classes and alternations - a preliminary investigation. The University of Chicago Press.

Miyazaki, M., Ikehara, S., Yokoo, A., 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).

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

Neumann, G., Fellbaum, C., Geyken, A., Herold, A., Huemmer, C., Koerner, F., Kramer, U., Krell, K., Sokirko, A., Stantcheva, D., Stathi, K., 2004. A corpus-based lexical resource of German idioms. In: Saint Dizier, P., Zock, M. (Eds.), Proceedings of the COLING Workshop on Electronic Lexicons. Geneva, Switzerland.

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

Oflazer, K., Çetinoğlu, O., Say, B., 2004. Integrating morphology with multi-word expression processing in Turkish. In: Tanaka et al. (2004), pp. 64–71.

Pearce, D., 2001. Synonymy in collocation extraction. In: Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. CMU, pp. 41–46.

Pereira, R., Crocker, P., Dias, G., 2004. A parallel multikey quicksort algorithm for mining multiword units. In: Dias et al. (2004), pp. 17–24.

Riehemann, S., 2001. A constructional approach to idioms and word formation. Ph.D. thesis, Stanford University.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP. In: Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002). Mexico City, Mexico, pp. 1–15.

Sharoff, S., 2004. What is at stake: a case study of Russian expressions starting with a preposition. In: Tanaka et al. (2004), pp. 17–23.

Tanaka, T., Baldwin, T., 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In: Bond et al. (2003), pp. 17–24.

Tanaka, T., Villavicencio, A., Bond, F., Korhonen, A. (Eds.), 2004. Second ACL Workshop on Multiword Expressions: Integrating Processing. ACL, Barcelona, Spain.

Trawiński, B., 2003. Licensing complex prepositions via lexical constraints. In: Bond et al. (2003), pp. 97–104.

Uchiyama, K., Ishizaki, S., 2003. A disambiguation method for Japanese compound verbs. In: Bond et al. (2003), pp. 81–88.

Villavicencio, A., Baldwin, T., Waldron, B., 2004a. 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., Copestake, A., Waldron, B., Lambeau, F., 2004b. Lexical encoding of MWEs. In: Tanaka et al. (2004), pp. 80–87.