Lexical Encoding of MWEs

Aline Villavicencio
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ, UK
and
Computer Laboratory, University of Cambridge
avill@essex.ac.uk

Ann Copestake, Benjamin Waldron, Fabre Lambeau
Computer Laboratory
University of Cambridge
William Gates Building, JJ Thomson Avenue
Cambridge, CB3 0FD, UK
{aac10,bmw20,faml2}@cl.cam.ac.uk

Abstract

Multiword Expressions present a challenge for language technology, given their flexible nature. Each type of multiword expression has its own characteristics, and providing a uniform lexical encoding for them is a difficult task to undertake. Nonetheless, in this paper we present an architecture for the lexical encoding of these expressions in a database, that takes into account their flexibility. This encoding extends in a straightforward manner the one required for simplex (single) words, and maximises the information contained for them in the description of multiwords.

1 Introduction

Multiword Expressions (MWEs) can be defined as *idiosyncratic interpretations that cross word boundaries (or spaces)* (from Sag et al. (2002). They comprise a wide-range of distinct but related phenomena like idioms, phrasal verbs, noun-noun compounds and many others, that due to their flexible nature, are considered to be a challenge for many areas of current language technology. Even though some MWEs are fixed, and do not present internal variation, such as *ad hoc*, others are much more flexible and allow different degrees of internal variability and modification, as, for instance, *touch a nerve* (*touch/find a nerve*) and *spill beans* (*spill several/musical/mountains of beans*). In terms of semantics, some MWEs are opaque and their semantics cannot be straightforwardly inferred from the meanings of the component words (e.g. *to kick the bucket* as *to die*). In other cases the meaning is more transparent and can be inferred from the words in the MWE (e.g. *eat up*, where the particle *up* adds a completive sense to *eat*).

Given the flexibility and variation in form of MWEs and the complex interrelations that may be found between their components, an encoding that treats them as invariant strings (a *words with spaces* approach), will not be adequate to fully describe any such expression appropriately with the exception of the simplest fixed cases such as *ad hoc* ((Sag et al., 2002), (Calzolari et al., 2002)). Different strategies for encoding MWEs have been employed by different lexical resources 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 it does not distinguish compositional from non-compositional entries neither does it specify entries that can be productively formed. WordNet, on the other hand, covers a large number of MWEs (Fellbaum, 1998), but does not provide information about their variability. Neither of these resources covers idioms. The challenge in designing adequate lexical resources for MWEs, is to ensure that the variability and the extra dimensions required 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 while attempting to establish the standards for MWE description in the context of multilingual lexical resources. Their focus is on MWEs that are productive and that present regularities that can be generalised and applied to other classes of words that have similar properties. Copestake et al. (2002) present an initial schema for MWE description and we build on these ideas here, by proposing an architecture for a lexical encoding of MWEs, which allows for a unified treatment of different kinds of MWE.

In what follows, we start by laying out the minimal encoding needed for simplex (single) words. Then, we analyse two different types of MWE (idioms and verb-particle constructions), and discuss their requirements for a lexical encoding. Given these requirements, we present a possible encoding for MWEs, that uniformly captures different types of expressions. This database encoding minimises the amount of information that

needs to be specified for MWE entries, by maximising the information that can be obtained from simplex words, while requiring only minimal modification to the encoding used for simplex words. We finish with some discussion and conclusions.

2 Simplex Entries

Simplex entries, in this context, refer to simple standalone words that are defined independently of others, and form the bulk of most lexical resources. For these entries, it is necessary to define at least their orthography, and syntactic and semantic characteristics, but more information can also be specified, such as particular dialect, register, and so on, and table 1 shows one such encoding. In this minimal encoding a lexical entry has an identifier (to uniquely distinguish between the different entries defining different combinations of parts-of-speech and senses for a given word), the word's orthography, grammatical (syntactic and semantic) type and predicate name.[1] In the case of this example, the identifier is **like_tv_1**, which is an entry for the verb *like*, with type **trans-verb** for transitive verbs, and predicate name **like_v_rel**. A type like **trans-verb** embodies the constraints defined for a given construction (in this case transitive verbs), in a particular grammar, and these vary from grammar to grammar. Thus, these words can be expanded into full feature structures during processing according to the constraints defined in a specific grammar.

Table 1: LINGO ERG lexical database encoding

identifier	orthography	type	predicate
like_tv_1	like	trans-verb	like_v_rel

This table shows a minimal encoding for simplex words, but it can serve as basis for a more complete one. That is the case of the LinGO ERG (Copestake and Flickinger, 2000) lexicon, which adopts for its database version, a compatible but more complex encoding which is successfully used to describe simplex words (Copestake et al., 2004). In the next sections, we investigate what would be necessary for extending this encoding for successfully capturing MWEs.

3 Idioms

Idioms constitute a complex case of MWEs, allowing a great deal of variation. Some idioms are very flexible and can be passivised, topicalised, internally modified, and/or have optional elements (e.g. *spill beans* in *those beans were spilt*, *users spilt password beans* and *judges spill their musical beans*), while others are more inflexible and only accept morphological inflection (e.g. *kick/kicks/kicked the bucket*).

In order to verify empirically the possible space of variation that idioms allow, we analysed a sample of some of the most frequent idioms in English. This sample was used for determining the requirements that an encoding needs in order to provide the means of adequately capturing idioms.

The Collins Cobuild Dictionary of Idioms lists approximately 4,400 idioms in English, and 750 of them are marked as the most frequent listed.[2] From these, 100 idioms were randomly selected and analysed as described by Villavicencio and Copestake (2002).

A great part of the idioms in this sample seems to form natural classes that follow similar patterns (e.g. the class of verb-object idioms, where an idiom consists of a specific verb that takes a specific object such as *rock boat* and *spill beans*). The remaining idioms, on the other hand, cannot so easily be grouped together, forming a large tail of classes often containing only one or two idioms (e.g. *thumbs up* and *quote, unquote*).

Most of the idioms in this sample present a large degree of variability, especially in terms of their syntax, also allowing variable elements (*throw* SOMEONE *to the lions*), and optional ones (*in a (tight) corner*). The type of variation that these MWEs allow 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 an idiom can be derived from the meanings of its elements. Thus, in these idioms, referred to as **semantically decomposable**

[1] The identifier and semantic relation names follow the standard adopted by the LinGO ERG (Copestake and Flickinger, 2000), while the grammatical type names are also compatible with it.

[2] These idioms have at least one occurrence in every 2 million words of the corpus employed to build this dictionary.

idioms, a meaning can be assigned to individual words (even if some of them are non-standard meanings) from where the meaning of the idiom can be compositionally constructed. 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*, according to this approach is non decomposable.

When semantic decomposability is used as basis for the classification, the majority of the idioms in this sample is classified as decomposable, and a few cases as non-decomposable. The decomposable cases correspond to the flexible idioms, and the non-decomposable to the fixed ones, providing a clear cut division for their treatment. For the non-decomposable idioms, a treatment of idioms as *words with space* can be adopted similar to that of simplex words, where in a single entry the orthography of the component words is specified, along with the syntactic and semantic type of the idiom, and a corresponding predicate name. In addition, for the cases that allow morphological inflection, it is also important to define which of the elements of the MWE can be inflected. In this case, an idiom like *kick the bucket*, is given the type of a normal intransitive verb, except that it is composed of more than one word, and only the verb can be inflected (e.g. *kick/kicked/kicks the bucket*,...). Consequently, an encoding for non-decomposable idioms needs to allow the definition of several orthographic elements for an entry, as well as the specification of the entry's orthographic element that allows inflection.

In order to capture the flexibility of decomposable idioms, a treatment using normal compositional processes can be employed as discussed by Copestake (1994). In this approach, each idiomatic component of an idiom could be defined as a separate entry similar to that of a simplex word, except that it would also be possible to specify a paraphrase for its meaning. In the case of *spill beans*, it would mean defining an entry for the idiomatic *spill*, which can be paraphrased as *reveal* and another for the idiomatic *beans* paraphrased as *secrets*. Moreover, as an idiomatic entry for a word may share many of the properties of (one of) the word's non-idiomatic entries (sometimes differing from the latter only in terms of their semantics), it is important to define also for each idiomatic element a corresponding non-idiomatic one, from which many aspects will be inherited by default. For example, in an idiom such as *spill beans*, the idiomatic entry for *spill* shares with the non-idiomatic entry the morphology (*spilled or spilt*) and the syntax (as a transitive verb), and so does the idiomatic *beans* with the non-idiomatic one. In addition, as there is a variability in the status of the words that form MWEs, with some words having a more literal interpretation and others a more idiomatic one, only the idiomatic words need to have separate entries defined. For example in the case of the idiom *pull the plug*, *pull* can be interpreted as contributing one of its non-idiomatic senses (that of *removing*), while *plug* has an idiomatic interpretation (that can be understood as meaning *support*). Thus, only an idiomatic entry (like that for *plug*) needs to be defined, while the contribution of a non-idiomatic entry (like that for *pull*) to the idiom comes from the standard entry for that word.

Having idiomatic and non-idiomatic entries available for use in idioms is just the first step in being able to capture this type of MWE. For a precise encoding of idioms, it is also necessary to define a very specific context of use for the idiomatic entries, to avoid the possibility of overgeneration. Thus, the verb *spill* has its idiomatic meaning of *reveal* only in the context of *spilt the beans* but not otherwise (e.g. in *spill the water*). The definition of these idiomatic contexts is important to ensure that idiomatic entries are used only in the context of the idiom, and that outside the idiom these entries are disallowed. Conversely, it is important to be able to define for each idiom, all the elements that need to be present for the idiomatic interpretation to be available. An idiom is only going to be understood as such if all of its obligatory components are present. In addition, it is necessary to ensure that the appropriate relationship among the components of an idiom is found, for the idiomatic meaning to be available, in order to avoid the case of false positives, where all the elements of an idiom are found, but not with the relevant interrelations. Thus, a sentence like *He threw the cat among the pigeons* has a possible idiomatic interpretation available, but this interpretation is not available in a sentence like *He held the cat and she threw the bread among the pigeons*, even though it has all the obligatory elements for the idiom (*throw, cat, among, pigeons*), because *cat* did not occur as a semantic argument (the agent) of *throw*. Many idioms also present some slight variation in their components, accepting any one of a restricted set of words, as for example *on home ground* and *on home turf*. Each of these possibilities corresponds to the same idiom realised in a slightly different way, but which nonetheless has the same meaning. Some idioms have also optional elements (such as *in a corner* and *in a tight corner*), and for these it is necessary to indicate which are the optional and which are the obligatory elements.

Idioms also present variation in the number of (obligatory) components they have, with some as short as

two words (e.g. *pull strings*) to others as long as 10 words (e.g. *six of one and half a dozen of the other*) or more, but with no lower and upper bound, or standard size. Consequently, an adequate treatment of idioms cannot assume that idioms will have a specific pre-defined size, but instead it needs to be able to deal with this variability.

4 Verb Particle Constructions

Verb Particle Constructions (VPCs) are combinations of verbs and prepositional or adverbial particles, such as *break down* in *The old truck broke down*. In syntactic terms, VPCs can be used in several different subcategorisation frames (e.g. *eat up* as intransitive or transitive VPC). In semantic terms VPCs can range from idiosyncratic or semi-idiosyncratic combinations, such as *get along* meaning *to be in friendly terms*, where the meaning of the combination cannot be straightforwardly inferred from the meaning of the verb and the particle, (in e.g. *He got along well with his colleagues*), to more regular ones, such as *tear up* (in e.g. *In a rage she tore up the letter Jack gave her*). The latter is a case where the particle compositionally adds a specific meaning to the construction and follows a productive pattern (e.g. as in *tear up*, *cut up* and *split up*, where the verbs are semantically related and *up* adds a sense of completion to the action of these verbs).

In terms of inflectional morphology, the verb-particle verb follows the same pattern as the simplex verb (e.g. *split up* and *split*). Other characteristics, like register and dialect are also shared between the verb in a VPC and the simplex verb. If the VPC and corresponding simplex verb are defined as independent unrelated entries, these generalisations about what is common between them would be lost. One option to avoid this problem is to define the VPC entry in a lexical encoding in terms of the corresponding simplex verb entry.

As discussed earlier for many VPCs the particle compositionally adds to the meaning of the verb to form the meaning of the VPC, and this provides one more reason for keeping the link between the VPC entry (e.g. *wander up*) and the simplex verb entry (e.g. *wander*), which share the semantics of the verb. Moreover, some of the compositional VPCs seem to follow productive patterns (e.g. the resultative combinations *walk/jump/run up/down/out/in/away/around/...* from joining these verbs and the directional/locative particles *up*, *down*, *out*, *in*, *away*, *around*, ...). This is discussed in Fraser (1976), who notes that the semantic properties of verbs seem to affect their possibility of combination with particles. For productive VPCs, one possibility is then to use the entries of verbs already listed in a lexical resource to productively generate VPC entries by combining them with particles according to their semantic classes, as discussed by Villavicencio (2003). However, there are also cases of semi-productivity, since the possibilities of combinations are not fully predictable from a particular verb and particle (e.g. *phone/ring/call/*telephone up*). Thus, although some classes of VPCs can be productively generated from verb entries, to avoid overgeneration we adopt an approach where the remaining VPCs need to be explicitly licensed by the specification of the appropriate VPC entry.

To sum up, for VPC entries an appropriate encoding needs to maintain the link between a VPC and the corresponding simplex form, from where the VPC inherits many of its characteristics, including inflectional morphology and for compositional cases, the semantics of the verb. On the other hand, for a non-compositional entry, like *get along*, it is necessary to specify the resulting semantics. In this case, the semantics defined in the VPC entry overrides that inherited by default from its components.

5 A Possible Encoding for MWEs

Taking the encoding of simplex entries as basis for an MWE encoding, we now discuss the necessary extensions to the former, to be able to provide the means of capturing the extra dimensions required by the latter. While taking these requirements into account, it is also desirable to define a very general architecture, in which simplex and MWE entries can be defined quite similarly, and in which different types of MWE can be captured in a uniform encoding.

In the proposed encoding, simplex entries are still defined in terms of orthography, grammatical type and semantic predicate, in the Simplex table (table 2). The same encoding can be used for fixed MWEs, which are treated as words with space, except that it also allows for the definition of the element in the MWE that can be inflected. This is the case of *kick the bucket*, which is defined as an intransitive construction whose first orthographic element (*kick*) is marked as allowing inflection, and from where variations such as *kicks the bucket* can be derived, table 2.

Table 2: Simplex Table: Extended Encoding for Simplex Entries

identifier	orthography	type	predicate	inflectional position
find_tv_1	find	trans-verb	find_tv_rel	
look_tv_1	look	trans-verb	look_tv_rel	
mention_tv_1	mention	trans-verb	mention_tv_rel	
pull_tv_1	pull	trans-verb	pull_tv_rel	
reveal_tv_1	reveal	trans-verb	reveal_tv_rel	
spill_tv_1	spill	trans-verb	spill_tv_rel	
touch_tv_1	touch	trans-verb	touch_tv_rel	
wander_tv_1	wander	trans-verb	wander_tv_rel	
up_prt_1	up	particle	up_prt_rel	
bean_n_1	bean	noun	bean_n_rel	
nerve_n_1	nerve	noun	nerve_n_rel	
secret_n_1	secret	noun	secret_n_rel	
unmentionable_n_1	unmentionable	noun	unmentionable_n_rel	
kick-the-bucket_iv_1	kick, the, bucket	intrans-verb	kick-the-bucket_iv_rel	1
walk_tv_1	walk	intrans-verb	walk_iv_rel	

Table 3: MWE Table:Encoding for Idiomatic Entries

identifier	base form	type	predicate	paraphrase
i_find_tv_1	find_tv_1	idiomatic-trans-verb	i_find_tv_rel	mention_tv_rel
i_spill_tv_1	spill_tv_1	idiomatic-trans-verb	i_spill_tv_rel	reveal_tv_rel
i_touch_tv_1	touch_tv_1	idiomatic-trans-verb	i_touch_tv_rel	mention_tv_rel
i_bean_n_1	bean_n_1	idiomatic_noun	i_bean_n_rel	secret_n_rel
i_nerve_n_1	nerve_n_1	idiomatic_noun	i_nerve_n_rel	unmentionable_n_rel

The encoding of flexible MWEs, on the other hand, is done in 3 stages. In the first one, the idiomatic components of an MWE are defined in a similar way to simplex words, in terms of an identifier, grammatical type and semantic predicate, in the MWE table (table 3). In addition, they also make reference to a non-idiomatic simplex entry (**base form** in table 3) from where they inherit by default many of their characteristics, including orthography. This is done by means of the non-idiomatic entry's identifier. In the case of e.g. the idiomatic *spill* (**i_spill_tv_1**), the corresponding non-idiomatic entry is the transitive *spill* defined in the simplex table, and whose identifier is **spill_tv_1**. Moreover, when appropriate, a non-idiomatic paraphrase for the idiomatic element can also be defined. This is achieved by specifying, in **paraphrase** the equivalent non-idiomatic element's semantic predicate. The idiomatic *spill*, for example, is assigned as corresponding paraphrase the non-idiomatic *reveal* (**reveal_tv_rel**) defined in the simplex table. This can be used to generate a non-idiomatic paraphrase for the whole MWE (e.g. *reveal secrets* as paraphrase of *spill beans*, as defined in table 3).

However, in order to be able to encode precisely an MWE, in the second stage its context is specified, where all the elements that make that MWE are listed. This ensures that only when all the core elements defined for an MWE are present, is that the MWE is recognised as such (e.g. *spill* and *beans* for the MWE *spill beans*), preventing the case of false positives (e.g. *spill the milk*) from being treated as an instance of this MWE. Likewise, this prevents idiomatic entries from being used outside the context of the MWE (e.g. the idiomatic *spill* being interpreted as *reveal* in *spill some water*). This is done in the table known as MWE Components, table 4. In this table each entry is defined in terms of an identifier for the MWE (e.g. **i_spill_beans_1**), and identifiers for each of the MWE components (e.g. **i_spill_tv_1** and **i_bean_n_1**), that provide the link to the lexical specification of these components either in the simplex table (table 2), or in

the MWE table (table 3). In order to allow MWEs with any number of elements to be uniformly defined, (from shorter ones like *spill beans*, rows 1 to 2 in table 4, to longer ones like *pull the curtain down on*) we propose an encoding where each element of the MWE is specified as a separate contextual entry (row). Thus, what links all the components of an MWE together, specified each as an entry, is that they have the same MWE identifier (e.g. **i_spill_beans_1**). Moreover, to account for MWEs with optional elements, like *in a corner* and *in a tight corner* where *tight* is optional, each of the elements of the MWE needs to be marked as obligatory or optional in this table.

For some MWEs, such as VPCs, one of the components may be contributing a very specific meaning in the context of that particular MWE, and often the meaning is more specific than the one defined in the corresponding base form entry for the component, from when the meaning is obtained by default. Thus, for non-compositional VPCs, such as *look up*, the particles can be assumed to have a vacuous semantic contribution, and the semantics of these VPCs are contributed solely by the verbs. For *look up*, the verbal component, **look_tv_1**, defines the meaning of the VPC as **look-up_tv_rel** while *up* is assigned a vacuous relation (**up-vacuous_prt_rel**). Similarly, *up* in a VPC such as *wander up* has either a directional or locational/aspectual interpretation, which in both cases can be regarded as qualifying the event of wandering and can be compositionally added to the meaning of the verb to generate the meaning of the combination. For these cases, it is important to allow the semantics of the component in question to be further refined in its entry for that MWE (e.g. *up* with semantics **up-end-pt_prt_rel** in table 4). The approach taken means that the commonality in the directional interpretation between *wander up* and *walk up*, where the semantics of the particle is shared, is captured by means of the specific semantic type defined for the particle, which means that generalizations can be made in an inference component or in semantic transfer for Machine Translation. Similarly, by defining a VPC from the base form of the corresponding verb, it is possible to capture the fact that the semantics of verb is shared between the verb *wander* and the VPC *wander up*.

Table 4: MWE Components

Phrase	Component	Predicate	Slot	Optional
i_spill_beans_1	i_spill_tv_1		PRED1	no
i_spill_beans_1	i_bean_n_1		PRED2	no
i_find_nerve_1	i_find_tv_1		PRED1	no
i_find_nerve_1	i_touch_tv_1		PRED1	no
i_find_nerve_1	i_nerve_n_1		PRED2	no
walk_up_1	walk_iv_1		PRED1	no
walk_up_1	up_prt_1	up-end-pt_prt_rel	PRED2	no
wander_up_1	wander_tv_1		PRED1	no
wander_up_1	up_prt_1	up-end-pt_prt_rel	PRED2	no
look_up_1	look_tv_1	look-up_tv_rel	PRED1	no
look_up_1	up_prt_1	up_vacous_prt_rel	PRED2	no

Finally, in order to specify the appropriate relationships between the elements of the MWE, a set of labels is used (PRED1, PRED2,...), which refer to the position of the element in the logical form for the MWE. This can be seen in the MWE Type table (table 5). The basic idea behind the use of these labels, defined in the column **slot**, is that they can be employed as place holders in the semantic predicate associated with that particular MWE. The precise correspondences between these place holders and the predicates are specified in meta-types defined for each different class of MWE. Thus the particular meta-type verb-object-idiom is for idioms with two obligatory elements, where PRED1 corresponds to pred1(X,Y) and PRED2 to pred2(Y), and PRED1 (corresponding to the verb) is a predicate whose second semantic argument (Y) is coindexed with the second predicate (the object). When this meta-type is instantiated with the entries for an MWE like *spill beans* (**i_spill_beans_1**) the slots are instantiated as **i_spill_rel(X,Y)**, and **i_bean_rel(Y)**.[3]

These meta-types act as interface between the database and a specific grammar system. As mentioned before MWEs can be grouped together in classes according to the patterns they follow (in terms of syn-

[3]For reasons of clarity, in this paper we are using a simplified but equivalent notation for the meta-type description.

Table 5: MWE Type Table

mwe	meta-type
i_find_nerve_1	verb-object-idiom
i_spill_beans_1	verb-object-idiom
walk_up_1	verb-particle-np
wander_up_1	verb-particle-np
look_up_1	verb-particle-np

tactic and semantic characteristics). Therefore, for each particular class of MWE, a specific meta-type is defined, which contains the precise interrelation between the components of the MWE. This means that for a particular grammar, for each meta-type there must be a (grammar-dependent) type that maps the semantic relations between the elements of the MWE into the appropriate grammar dependent features. Thus, in the third stage, it is necessary to specify the meta-types for the MWEs encoded.

In order to test the generality of the meta-types defined, a further sample of 25 idioms was randomly selected, and an attempt was made to classify them according to the meta-types defined. The majority of these idioms could be successfully described by the available types, with only a few for which further meta-types needed to be defined.

The same mechanisms are also used for defining MWEs which have an element that can be realised in different ways, but as one of a restricted set of words like *touch a nerve* and *find a nerve* which are instances of the same MWE. For these cases, it is necessary to define each of the possible variants and the position in the idiom in which they occur. This is done in table 4, where *find* and *touch*, the variants of the idiom *find/touch a nerve* are defined as occurring in a particular slot, PRED1 (and *nerve* as PRED2): **i_touch_rel(X,Y) i_nerve_rel(Y)** and **i_find_rel(X,Y) i_nerve_rel(Y)**. By using the same identifier (**i_find_nerve_1**) and slot (PRED1) in both cases, *find* and *touch* are specified as two possible distinct realizations of the slot for that same idiom.

6 Discussion

Multiword Expressions present a challenge for language technology, given their flexible nature. In this paper we described a possible architecture for the lexical encoding of these expressions. Even though different types of MWEs have their own characteristics, this proposal provides a uniform lexical encoding for defining them. This architecture takes into account the flexibility of MWEs extending in a straightforward manner the one required for simplex words, and maximises the information contained for them in the description of MWEs while minimising the amount of information that needs to be defined in the description of these expressions.

This encoding provides a clear way to capture both fixed (and semi-fixed) MWEs and flexible ones. The former are treated in the same manner as simplex words, but with the possibility of specifying the inflectional element of the MWE. For flexible MWEs, on the other hand, the encoding is done in three stages. The first one is the definition of the idiomatic elements, in the MWE table, the second the definition of an MWE's components, in the MWE Components table, and the third is the specification of a class (or meta-type) for the MWE, in the MWE Type table. Different types of MWEs can be straightforwardly described using this encoding, as discussed in terms of idioms and VPCs.

A database employing this encoding can be integrated with a particular grammar, providing the grammar system with a useful repertoire of MWEs. This is the case of the MWE grammar (Villavicencio, 2003) and of the wide-coverage LinGO ERG (Flickinger, 2004), both implemented on the framework of HPSG and successfully integrated with this database. This encoding is also used as basis of the architecture for a multilingual database of MWEs defined by Villavicencio et al. (2004), which has the added complexity of having to record the correspondences and differences in MWEs in different languages: different word orders, different lexical and syntactic constructions, etc. In terms of usage, this encoding means that the search facilities provided by the database can help the user investigate MWEs with particular properties. This in turn can be used to aid the addition of new MWEs to the database by analogy with existing MWEs with similar characteristics.

7 Acknowledgements

This research was supported in part by the NTT/Stanford Research Collaboration, research project on multiword expressions and by the Noun Phrase Agreement and Coordination AHRB Project MRG-AN10939/APN17606. This document was generated partly in the context of the DeepThought project, funded under the Thematic Programme User-friendly Information Society of the 5th Framework Programme of the European Community (Contract No IST-2001-37836).

References

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio 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.

John Carroll and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev and E. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag, and Dan Flickinger. 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.

Ann Copestake, Fabre Lambeau, Benjamin Waldron, Francis Bond, Dan Flickinger, and Stephan Oepen. 2004. A lexicon module for a grammar development environment. In *To appear in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Ann Copestake. 1994. Representing idioms. Paper presented at the HPSG Conference.

Christiane Fellbaum. 1998. 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.

Dan Flickinger. 2004. Personal Communication.

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

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language*, 70:491–538.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Aline Villavicencio and Ann Copestake. 2002. Aspectual on the nature of idioms. LinGO Working Paper No. 2002-04.

Aline Villavicencio, Timothy Baldwin, and Benjamin Waldron. 2004. A multilingual database of idioms. In *To appear in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Aline Villavicencio. 2003. Verb-particle constructions and lexical resources. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64, Sapporo, Japan.