

# Report on the lexical representation of subclasses of MWEs

Nicole Grégoire  
Uil-OTS, Utrecht University  
Nicole.Gregoire@let.uu.nl

March 9, 2007

## 1 Introduction

This report focuses on various subtypes of MultiWord Expressions (MWEs) and the way they are dealt with in the *MWE lexicon for Dutch*. The lexicon is developed as part of the STEVIN IRME project,<sup>1</sup> which aims at creating an electronic resource of 5,000 Dutch expressions that meets the criterion of being highly theory- and implementation-independent, and which can be used in various Dutch NLP systems.

The description of an MWE consists of a list of properties including a pattern name that refers to the description of an MWE pattern. This document solely addresses the description fields that are relevant for the discussion on classes of MWEs. For an elaborate overview of the encoding guidelines of an MWE description and an MWE pattern description, I refer to the *Encoding Protocol* (Grégoire, 2007).

Prior to the overview of subclasses of MWEs in our lexicon, given in section 3, we discuss subclasses and their representation described in related work in section 2. The report ends with a conclusion in section 4.

## 2 Related research: Classes and representations

In recent years, the NLP community has increasingly become aware of the problems that multiword expressions pose. A considerable amount of research has been conducted in this area. Most progress has been made especially in the field of multiword extraction. Moreover, interesting papers have been written on the representation of MWEs, most of them focusing on one class of MWEs. In section 2.2 I summarize related work on the representation of MWEs, but first I briefly discuss some subtypes of MWEs that are relevant for this report.

### 2.1 Classes of MWEs

The area of multiword expressions includes many different subtypes, varying from fixed expressions to syntactically more flexible expressions. Sag et al. (2001) wrote a paper on subclasses of MWEs, in which they make a distinction between *lexicalized phrases* and *institutionalized phrases*. Lexicalized phrases are subdivided into fixed, semi-fixed and flexible expressions. The most important reason for this subdivision is the variation in the degree

---

<sup>1</sup><http://taalunieversum.org//stevin/> and <http://www-uilots.let.uu.nl/irme/>

of syntactic flexibility of MWEs. Roughly they claim that syntactic flexibility is related to semantic decomposability. Semantically non-decomposable idioms are idioms the meaning of which cannot be distributed over its parts and are therefore not subject to syntactic variability. Sag et al. (2001) state that “the only types of lexical variation observable in non-decomposable idioms are inflection (*kicked the bucket*) and variation in reflexive form (*wet oneself*).” Examples of non-decomposable idioms are the oft-cited *kick the bucket*, *shoot the breeze* and *trip the light fantastic*. On the contrary, semantically decomposable idioms, such as *let the cat out of the bag* and *spill the beans*, tend to be syntactically flexible to some degree. Mapping the boundaries of flexibility, however, is not always easy and no one can predict exactly which types of syntactic variation a given idiom can undergo.

One subtype of flexible expressions discussed in Sag et al. (2001) are *Support Verb Constructions* (or *Light Verb Constructions*). SVCs are combinations of a verb that seems to have very little semantic content (hence the names *support verb/function verb/light verb*) and a prepositional phrase, a noun phrase or adjectival phrase. An SVC is often paraphrasable by means of a single verb or adjective. Since the complement of the verb is used in its normal sense, the constructions are subject to standard grammar rules, which include passivization, internal modification, etc. Examples of English SVCs are *give/\*make a demo*, *make/\*do a mistake*.

## 2.2 Representation

Most work on the representation of MWEs focuses on a single subtype. Related research described in this section concern verbal idioms, collocations, verb-particle constructions and light verb constructions. It is notable that three out of four papers discussed include syntactic transformation and/or semantic decomposability in their description of an MWE. Deciding on which syntactic operations a certain expression can undergo, requires a large amount of (manual) work and a large-scaled corpus investigation to avoid analyses entirely based on speaker-specific intuitions. Since the resources described contain no more than 1,000 high-frequent expressions, an approach that includes syntax and semantics is feasible, although it does require a considerable amount of corpus data to extract sufficient examples of the various syntactic transformations.

Below, I give a summary of four papers describing research on the representation of subclasses of MWEs, starting with work done on German verbal idioms in 1998 and concluding with research also on German verbal idioms finished in 2006.

**Dormeyer and Fischer (1998)** report on work on a computational dictionary for German verbal idioms, called *Phraseo-Lex*. *Phraseo-Lex* was designed both as a computational dictionary for humans and as a source to generate lexicons for NLP systems. The *Phraseo-Lex* dictionary contains syntactic, semantic and pragmatic information in the description of the German verbal idioms. The paper solely describes the representation features that are relevant for the representation of idioms in NLP systems.

Besides standard information about the idiom, such as the components it contains and their syntactic category, *Phraseo-Lex* has the option to mark for each idiom whether a given set of syntactic transformations is possible, impossible, or undecidable. This set includes, inter alia, passivization, relativization, negation and quantification.

Regarding the semantic features, it is decided for each idiom part whether it has a meaning of its own or not, i.e. they distinguish between meaningful and meaningless idiom parts, yielding three classes of idioms:

1. Compositional idioms, of which the idiom parts are all meaningful.
2. Non-compositional idioms, of which the idiom parts are all meaningless.
3. Partially compositional idioms, which consists of both meaningful and meaningless parts.

The paraphrase of the idiom is supposed to reflect the idiom’s semantic type, which means that for compositional idioms, the paraphrase should have the same syntactic structure as the idiom. Apart from the mapping between idiom parts and paraphrase parts, a semantic role is assigned to each internal or external valency. Internal valencies that do not carry independent meaning are marked as having no role. Last point to mention is that in *Phraseo-Lex* the modifiable parts of an idiom are explicitly listed.

**Krenn (2000)** worked on a database of lexical collocations. She uses the term collocation for “word combinations that are lexically determined and constitute particular syntactic dependencies such as verb-object, verb-subject, adjective-noun relations, etc.” The database contains more than 1,000 German PP-verb collocations (PNV-combinations).

In her approach, the representation of a collocation consists of a *competence base* and an *example base*. In the competence base, collocation instances (CI) and CI-analyses are stored. CIs are representations of the major lexical elements of a collocation, in which nouns are represented as full forms, and verbs as infinitives. The CI-analysis lists the values of eight attributes for each collocation, among which causativity, Aktionart, the syntactic arguments required by the collocation, modification of the noun, modification of the whole expression. The example base contains example sentences extracted from corpora for each CI.

**Villavicencio et al. (2004)** propose a possible architecture for the lexical encoding of MWEs. They analyse two different types of expressions, viz. idioms and verb-particle constructions. The central idea behind the design of the encodings is to minimize the amount of information that needs to be specified for MWE entries by maximising the information that can be inherited from simplex verbs. For each simplex verb, its orthography and its syntactic and semantic type are stored in the database. Villavicencio et al. (2004) subdivide the class of idioms into non-decomposable idioms, which they classify as fixed MWEs, and decomposable idioms, grouped as flexible MWEs. Fixed MWEs are treated as words with spaces and encoded as simplex entries. Elements that can inflect, e.g. *kick* in *kick the bucket*, are marked as such.

The encoding of flexible MWEs is dealt with in three stages. First the idiomatic components of an MWE are defined in the same way as simplex verbs. In addition, each component is linked to a non-idiomatic simplex entry from which they obtain by default many of the characteristics. Furthermore a non-idiomatic paraphrase for the idiomatic element is defined. In the second stage, all the components that make the MWE are listed. This is done to ensure that the idiomatic reading of the individual components is only used when it occurs in the presence of the other components. Meta-types, in the sense of predefined semantic relations, are specified in the last stage. Examples of meta-types specified are *verb-object-idiom* and *verb-particle-np*. The majority of the MWEs in their database could be described by the meta-types defined, yielding a classified list of MWE entries.

**Fellbaum et al. (2006)** discuss the motivation as well as the design and development of a large lexical resource focusing on German verb phrase idioms and light verbs. Lexical annotation of the idioms is based on corpus-based investigation. Both the annotations and the corpus data are accessible. The properties of an idiom are recorded in eight so-called data sheets. The following components are part of the idiom description:

1. The citation form of the idiom.
2. Some corpus examples.
3. A paraphrase of the idiom.
4. Information about usage and alternations.
5. A dependency structure, including core components, obligatory components, optional components, optional elements, idiom-external components, and idiom-external optional components. Each element is assigned a lexical realization, which may consist of more than one lexeme.
6. Morphosyntactic properties of each element.
7. Lexical and phrasal variation from the citation form.
8. The syntactic transformations found in the example corpus. The following transformations are provided: pronominalization, passivization, question formation, relativization, affirmation, conversion (change in the syntactic category of a core component), autonomization (occurrence of only one of the core components), focusing, zeugma, and anaphorization.
9. Semantic properties.
10. Paradigmatic relations with other idioms.

A total of 1,000 multiword units have been investigated in this project.

### **3 Classes of MWEs and their representation in the MWE lexicon for Dutch**

In our research multiword expressions are defined as a combination of words that has linguistic properties not predictable from the individual components or the normal way they are combined. The linguistic properties can be of any type (Odiijk, 2004a):

- Lexical: *de plaat poetsen, zware/\*sterke shag*
- Orthographic: *Yahoo!*
- Phonological: *over de rooie/\*rode* (gaan/zijn/raken)
- Morphological: *ten gevolge van*
- Syntactic: *in opdracht van, bijvoeglijk(\*e) naamwoord*
- Semantic: *de plaat poetsen, een bok schieten*

- Pragmatic: *dames en heren*

We are creating a resource in which we want to cover as many different types of MWEs and as many properties of MWEs as possible. The possibilities of what to include in a standard representation of MWEs are numerous, but time is limited. Various aspects played a role in the representation as it is in our lexicon. First of all, the main requirement of the standard encoding is that it can be converted into any system specific representation with a minimal amount of manual work. The method adopted to achieve this goal is the Equivalence Class Method (ECM) (Odiijk, 2003, 2004b). The idea behind the ECM is that MWEs that have the same pattern require the same treatment in an NLP system. MWEs with the same pattern form so-called Equivalence Classes (ECs). Having the ECs, it requires some manual work to convert one instance of an EC into a system specific representation, but all other members of the same EC can be done in a fully automatic manner. The target system’s grammar must include a way to handle MWEs and lexical entries must be available.

The creation of MWE descriptions is a very time-consuming task and of course we aim at an error-free result. Accordingly, we decided to describe the minimal ingredients of an MWE that are needed for successful incorporation in any Dutch NLP system. For the development of the representation two Dutch parsers are consulted, viz. the Alpino parser<sup>2</sup> and the Rosetta MT system (Rosetta, 1994).

Rosetta is the result of seven years of research on machine translation started in 1985 at the Philips Research Laboratories in Eindhoven. This system is meant to translate between English, Dutch and Spanish and has been developed according to an approach called compositional translation. The type of grammar used in Rosetta is called *M-grammar*, a computationally feasible variant of *Montague Grammar*.

Alpino is a dependency parser for Dutch, developed in the context of the NWO PIONIER Project *Algorithms for Linguistic Processing*. Alpino is based on the Head-Driven Phrase Structure Grammar (HPSG).

Another requirement of the lexicon structure is that the information needed for the representation must be extractable from corpora. Unfortunately, occurrences of MWEs in corpora are very rare, and not totally reliable due to corruption of the data and absence of a distinction between literal and idiomatic interpretations. In our research we use data extracted from the Twente Nieuws Corpus (TwNC) (Ordelman, 2002) as empirical material. The corpus comprises a 500 million words of newspaper text and television news reports. In addition to the corpus data we make use of our linguistic knowledge and intuitions.

### 3.1 Classes of MWEs

Despite the fact that we aspire to give a full account of Dutch multiword expressions, time restrictions and lack of adequate data force us to focus on those properties that are sufficient for a successful conversion of the standard representation into any Dutch NLP system. Since the treatment of MWEs in NLP systems currently does not involve sophisticated syntactic and semantic analyses, the following properties of MWEs are not part of our standard representation:

1. **Meaning** No meaning is assigned to the MWEs in our lexicon. This means that nothing is and can be said about the semantic decomposability of the expressions.

---

<sup>2</sup>See <http://odur.let.rug.nl/~vannoord/alp>.

2. **Syntactic flexibility** Expressions may or may not undergo certain syntactic operations, such as topicalization, passivization, relativization etc. Although including this information definitely enriches the lexicon, time constraints prevent us from providing idiosyncratic syntactic flexibility information for individual MWEs.<sup>3</sup>

Although we do not explicitly specify the syntactic operations that a certain expression can undergo, we do make a distinction between fixed, semi-flexible and flexible expressions, based on generalisations about the morpho-syntactic flexibility of MWEs. The characteristics of these three classes and their representation are discussed in the subsections below.

### 3.1.1 Fixed MWEs

**Characteristics** Fixed MWEs always occur in the same word order and there is no variation in lexical item choice. Fixed MWEs cannot undergo morphosyntactic variation and are contiguous, e.g. no other elements can intervene between the words that are part of the fixed MWE.

Examples of fixed MWEs are: *ad hoc*, *ter plaatse* ‘on the spot’, *van hoger hand* ‘from higher authority’.

**Representation** Fixed expressions can be represented in two ways depending on its internal structure:

1. For fixed expressions that are difficult to assign an internal structure, we introduced the dependency label *fixed* (Grégoire, 2007). The pattern for expressions such as *ad hoc* and *ter plaatste* is: [. fixed(1 2) ]

Currently, all fixed expressions with the same number of components are listed with the same pattern, disregarding the category of the expression. When the fixed expressions are entered in the lexicon, they will be subdivided according to their category. As a result, the patterns will change to: [.:X fixed(1 2) ], where X is replaced with a category label.

2. Fixed expressions with an analyzable internal structure are represented according to the normal pattern rules:
  - EXPRESSION de volle buit
  - CL de vol buit[sg]
  - PATTERN [.NP [.det:D (1) ] [.mod:A (2) ] [.hd:N (3) ]]

### 3.1.2 Semi-flexible MWEs

**Characteristics** The following characteristics are applicable to the class of semi-flexible MWEs in our lexicon:

1. The lexical item selection of the elements of the expression is fixed or very limited.

---

<sup>3</sup>Many linguists share the claim that the syntactic behaviour of idioms can be predicted, at least in part, on the basis of their meanings (Nunberg et al., 1994). As we leave out both meaning and syntactic behaviour in our representation of MWEs, we cannot elaborate on their possible relation.

2. The expression can only be modified as a whole,<sup>4</sup> i.e. the components cannot be modified individually.
3. The head of the expression can inflect, unless explicitly marked otherwise with a parameter.<sup>5</sup>

Examples of Dutch semi-flexible MWEs are: *de plaat poetsen*, *witte wijn* ‘white wine’, *bijvoeglijk naamwoord* ‘adjective’.

**Representation** Semi-flexible MWEs are represented according to the pattern rules described in section 2.2 of the *Encoding Protocol*:

1.
  - EXPRESSION *de plaat poetsen*
  - CL *de plaat[sg] poetsen*
  - PATTERN [.VP [.obj1:NP [.det:D (1) ] [.hd:N (2) ]] [.hd:N (3) ]]
2.
  - EXPRESSION *zoete broodjes bakken*
  - CL EMP *zoet broodje[pl] bakken*
  - PATTERN [.VP [.obj1:NP [.det:D (1) ] [.mod:A (2) ] [.hd:N (3) ]] [.hd:N (4) ]]

To make a distinction between (1) an NP of which all elements are fixed, and (2) an NP of which some elements are lexically fixed, but which is still subject to standard grammar rules, we introduce a new syntactic category *N1*.<sup>6</sup>

1.
  - EXPRESSION *bijvoeglijk naamwoord*
  - CL *bijvoeglijk[noesg] naamwoord[hct]*
  - PATTERN [.N1 [.mod:A (1) ] [.hd:N (2) ]]
2.
  - EXPRESSION *ongewenst bezoek*
  - CL *ongewenst[noe] bezoek[hct][sg]*
  - PATTERN [.N1 [.mod:A (1) ] [.hd:N (2) ]]
3.
  - EXPRESSION *witte bonen*
  - CL *wit boon[pl]*
  - PATTERN [.N1 [.mod:A (1) ] [.hd:N (2) ]]

In the examples above, the lexical realization of adjective and noun is fixed, but since the syntactic category is *N1*, the expression can be modified as a whole and specify a determiner: *de correcte bijvoeglijk naamwoorden*.

---

<sup>4</sup>We abstract away from the reason why some external modifiers, such a *proverbial* in *he kicked the proverbial bucket*, may intrude in these semi-flexible expressions.

<sup>5</sup>It must be noted that adjectives are subject to standard Dutch grammar rules, which means that they inflect according to the gender of the noun and the form of the determiner. The parameter value *[noe]* is used when an adjective cannot occur with *-e* inflection. For an elaborate description of the encoding of the CL-field and parameters, see section 3.3 of the *Encoding Protocol*.

<sup>6</sup>Note that the noun in the expressions in 2 and 3 cannot inflect due to the parameter value.

### 3.1.3 Flexible MWEs

**Characteristics** Flexible MWEs can be characterized as follows:

1. They are not subject to a fixed word order and they allow intrusion by other words or phrases.
2. Individual components can inflect and be modified: *een blunder begaan* – *blunders begaan*, *scherpe kritiek* – *heel scherpe inhoudelijke kritiek*

The main characteristic of flexible MWEs is the fact that, contrary to semi-fixed MWEs, the individual components within flexible MWEs can inflect and be modified. This contrast accounts for differences between *de plaat poetsen* vs. *een bok schieten* and *blunder maken/begaan*. Although both *een bok schieten* and *blunder maken/begaan* are flexible MWEs, there is a difference between the two expressions. According to the classification proposed by Sag et al. (2001), *een bok schieten* is a decomposable idiom and *een blunder maken* is a support verb construction. In our resource, we subdivide the class of flexible MWEs into:

1. Expressions of which one part is fixed and the other part is a list of one or more co-occurring lexemes. Dutch examples are: *bar/bitter koud*, *scherpe/stevige kritiek*, *blunder maken/begaan*.
2. Expressions of which the lexical realization of each component consists of exactly one lexeme.

The latter subtype solely comprises **verbal** flexible expressions. The difference between the two subtypes is made visible in the MWE pattern and the MWE description.

**Representation** Flexible expressions are encoded using the syntactic category N1:

- EXPRESSION bok schieten
- CL bok schieten
- PATTERN [.VP [.obj1:N1 [.hd:N (1) ]] [.hd:V (2) ]]

Expressions of which one part is fixed and the other part is a list of one or more co-occurring lexemes are represented with a so-called LIST-index in the pattern. The theory is that the fixed part of the expression is in its literal sense. The combination of the literal part with other lexemes is not predicable from the meaning of the combining lexeme. Since we do not include the meaning of the MWEs or its parts in our representation, we can list every single component with which the fixed part can combine in the same MWE entry. For this list of components we created a LIST-A-field and LIST-B-field in the MWE description. Each component in the LIST-A-field and LIST-B-field can be substituted for the LIST-index in the pattern, yielding one or more different expressions.

The reason for using two LIST-fields is to separate predefined list values from special list values. The predefined list values are high frequent verbs that are known to occur often as so-called light verbs. Two sets of verbs are predefined:

1. blijken blijven gaan komen lijken raken schijnen vallen worden zijn

2. brengen doen hebben houden krijgen maken zetten

Since it is not possible that a complement co-occurs with verbs from both set 1 and set 2, keeping the same pattern, either one set is selected for the LIST-A-field. Each verb from the chosen set is checked against the occurrences found in the corpus data. If a verb does not occur in the corpus data and also not in constructed data, it is deleted from the LIST-A-field.

The LIST-B-field contains lexemes that are not in the predefined set but do co-occur with the component(s) in the EXPRESSION-field. The information in the LIST-B-field is merely based on corpus data and therefore may not be exhaustive.

Examples of flexible MWEs that are assigned a LIST-index and LIST-A-field/LIST-B-field are given in the following table:

EXPRESSION	CL	LIST-A	LIST-B	PATTERN
blunder	blunder	maken	begaan	[.VP [.obj1:N1 [.hd:N (1) ]] [.hd:V (list) ]]
aan de gang	aan de gang	blijken blijven lijken raken zijn		[.VP [.predc:PP [.hd:P (1) ] [.obj1:NP [.det:D (2) ] [.hd:N (1) ]]] [.hd:V (list) ]]
koud	koud		bar bitter	[.AP [.mod:A (list) ] [.hd:A (1) ]]
kritiek	kritiek		scherp stevig	[.N1 [.mod:A (list) ] [.hd:N (1) ]]

## 4 Conclusion

In this report, we elaborated on the lexical representation of Dutch multiword expressions. We distinguish between fixed, semi-flexible and flexible expressions. Although we agree that a complete representation of MWEs also includes information about the semantic decomposability and syntactic variability of each expression, we omit these details in our representation. The main reason for this is the lack of sufficient data to draw any conclusions regarding these aspects. The differences between the three subclasses distinguished in our lexicon are mainly based on the degree of morpho-syntactic variability of the expressions, i.e. whether parts of the expression can inflect and be modified.

Besides giving an overview of the subclasses distinguished in our research, we discussed a subdivision of MWEs as proposed in Sag et al. (2001) and summarized several papers on the representation of MWEs that are related to our research.

## References

- Dormeyer, R. and Fischer, I. (1998), Building lexicons out of a database for idioms, *in* A. Rubio, N. Gallardo, R. Castro and A. Tejada (eds), *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 833 – 838.
- Fellbaum, C., Geyken, A., Herold, A., Koerner, F. and Neumann, G. (2006), Corpus-Based Studies of German Idioms and Light Verbs, *International Journal of Lexicography* **19**(4), 349–361.

- Grégoire, N. (2007), MWE lexicon for dutch: Encoding protocol, *Technical report*, STEVIN IRME.
- Krenn, B. (2000), CDB - a database of lexical collocations, *2nd International Conference on Language Resources & Evaluation (LREC '00)*, May 31 - June 2, ELRA, Athens, Greece.
- Nunberg, G., Sag, I. and Wasow, T. (1994), Idioms, *Language* **70**, 491–538.
- Odiijk, J. (2003), Towards a standard for multi-word expressions. ISLE Project Report.
- Odiijk, J. (2004a), Multiword expressions in NLP, Course presentation, LOT Summerschool, Utrecht.
- Odiijk, J. (2004b), A proposed standard for the lexical representation of idioms, *EURALEX 2004 Proceedings*, Université de Bretagne Sud, pp. 153–164.
- Ordelman, R. (2002), Twente nieuws corpus (TwNC).
- Rosetta, M. T. (1994), *Compositional Translation*, Kluwer Academic Publishers, Dordrecht.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2001), Multiword expressions: A pain in the neck for NLP, LinGO Working Paper, (2001-03).
- Villavicencio, A., Copestake, A., Waldron, B. and Lambeau, F. (2004), The lexical encoding of MWEs, in T. Tanaka, A. Villavicencio, F. Bond and A. Korhonen (eds), *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.