STEVIN-IRME project

MWE Lexicon for Dutch
Encoding protocol

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

October 27, 2007

Contents

1 Introduction

This document is meant as a guideline for the encoding of MultiWord Expressions (MWEs) in the *Lexicon for Dutch MWEs*. The instructions should help lexicographers to add new data to the lexicon and NLP builders should use them in combination with the conversion procedure manual to incorporate the lexicon into a system specific representation (Grégoire, 2007a).

The description of an MWE consists of a list of properties specific for a certain MWE and a pattern name that refers to the description of an MWE pattern. Although multiword expressions and their behavior in standard Dutch were extensively studied before finalizing the representation format as described in this document, some properties and special behavior cannot be captured in the current representation. Known limitations of the representation are explicitly stated in the relevant sections.

The document starts with an overview of the ingredients of an MWE pattern description in section 2. This is followed by an overview of the ingredients of an MWE description in section 3. The appendices A and B show examples of actual MWE pattern descriptions and MWE descriptions respectively.

2 MWE pattern description

The current section elaborates on the ingredients that are part of the MWE pattern description. Each unique pattern is described with the following fields:

1. ID
2. PATTERN_NAME
3. POS
4. PATTERN
5. MAPPING
6. MWE
7. EXAMPLE
8. DESCRIPTION
9. COMMENTS

All fields, except for the comment field, are required. The coding instructions for each field are described in the subsections below. Examples of actual encodings are given in appendix A. A complete overview of the pattern descriptions including a graphical representation of the dependency tree is given in the document *MWE Lexicon for Dutch: Overview of MWE pattern descriptions* (Grégoire, 2007b).

2.1 id

Automatically assigned unique identifier.

2.2 pattern_name

The PATTERN_NAME is an identifier that uniquely identifies the pattern description. Each pattern name is automatically generated and starts with *EC* followed by a unique number.

2.3 pattern

In order to form equivalence classes, each expression should be assigned an MWE pattern. The notation used to describe the patterns is based on a formalization of dependency trees, in particular CGN (*Corpus Gesproken Nederlands* 'Corpus of Spoken Dutch') dependency trees (Hoekstra et al., 2003). CGN dependency structures are based on traditional syntactic analysis described in the Algemene Nederlandse Spraakkunst (Haeseryn et al., 1997) and are aimed to be as theory neutral as possible.

An example of a CGN dependency structure is given in Figure 1. Category labels (*c*-labels: in the cylinder-boxes) are assigned to each mothernode. Dependency labels (*d*-labels: in the square boxes) denote the relation of a certain constituent with respect to another constituent dominated by the same mother node (Hoekstra et al., 2003).

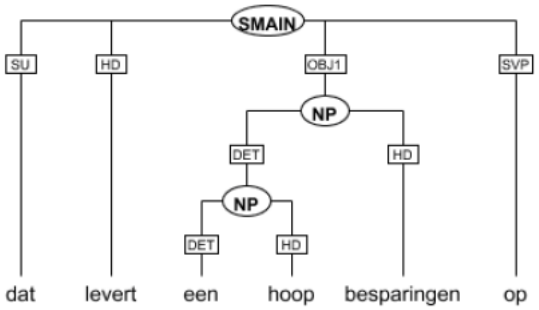


Figure 1: An example of a CGN dependency tree.

Notation LaTeX-code is used as a basis for the notation of the patterns.[1] The *d*-label (in lower case) and *c*-label (in upper case) are separated by a colon (:), e.g. *obj1:NP*. For leaf nodes, the part-of-speech is represented instead of the *c*-label. Leaf nodes are followed by an index that refers to the MWE component as represented in the CL-field (see section 3.1.3), e.g. (1) refers to the first component of the CL, (2) to the second, etc.[2]

Representation rules The following rules apply to the representation of the patterns:

1. The available *c*-labels are shown in Table 1.[3]

c-label	Description
AP	adjectival phrase
AdvP	adverbial phrase
CP	complementizer phrase
INF	infinitive
NP	noun phrase
OTI	(*om*)-*te*-infinitive phrase[4]
PP	prepositional phrase
SSUB	subordinate clause
TI	*te*-infinitive phrase
VP	verb phrase

Table 1: An overview of *c*-labels used in the MWE patterns.

2. The available *PoS*-labels used on leaf nodes are shown in Table 2.

   The N, N1, N2 and A, A1 labels are created to cover the modifiability of the noun and adjective respectively.[5] The examples (1)–(5) illustrate the use of these labels.

[1] The LaTeX-notation is used for two reasons: (1) the notation is short, i.e. it fits one line opposed to e.g. XML code, and (2) an image file can be created using a LaTeX compile command, which makes it possible to show a graphical dependency tree. The package that is used is *qtree*. For more information see http://www.latex-project.org/.

[2] This implies that the number of indices in the PATTERN is the same as the number of elements in the CL-field.

[3] The sets of labels used in the notation can be extended, but must be properly documented accordingly.

[4] In an OTI phrase *om* is optional which means that it can either be included or left out. In a TI phrase *om* is prohibited.

[5] Modifiability of the adjective includes variation of the form, e.g. comparative and superlative.

c-label	Description
A	not modifiable adjective
A1	modifiable adjective
Adv	adverb
C	complementizer
D	determiner
N	not freely modifiable noun
N1	modifiable noun
N2	limitedly modifiable noun
P	preposition
PRON	pronoun
REFLV	reflexive verb
V	verb

Table 2: An overview of *PoS*-labels used on leaf nodes used in the MWE patterns.

(1) de broek dragen ('wear the pants')
[.VP [.obj1:NP [.det:D (1)] [.hd:**N** (2)]] [.hd:V (3)]]

(2) zijn debuut maken ('make one's debut')
[.VP [.obj1:NP [.det:D (1)] [.hd:**N1** (2)]] [.hd:V (3)]]

(3) zijn best doen ('do one's best')
[.VP [.obj1:NP [.det:D (1)] [.hd:**N2** (2)]] [.hd:V (3)]]

(4) de grote stad ('the big city')
[.NP [.det:D (1)] [.mod:**A** (2)] [.hd:N (3)]]

(5) grote dag ('big day')
[.NP [.mod:**A1** (1)] [.hd:N (2)]]

It should be noted that often it is not clear whether a noun is limitedly modifiable or free modifiable, and whether the limitedly modifiability of the noun is the result of the combination of the noun with the other components of the expression or that it is a property of the noun itself. The determination of whether the noun is limitedly modifiable or free modifiable is merely based on corpus information, which may not be exhaustive and may lead to an incorrect pattern allocation.

3. The available *d*-labels are shown in Table 3.

4. Fixed expressions[6] that are difficult to assign an internal structure should be represented with the label *fixed* followed by round brackets with indices referring to the components:

(6) ten onder gaan ('go down')
[.VP [.ld:PP fixed(1 2)] [.hd:V (list)]]

[6]The term *fixed expressions* refers to expressions that always occur in the same word order and that do not allow variation in lexical item choice. Furthermore, they cannot undergo morphosyntactic variation and are contiguous, e.g. no other elements can intervene between the words that are part of the fixed expression.

d-label	Description
BODY	body of the complementizer phrase
CMP	head of the complementizer phrase
DET	determiner
HD	head
LD	locative/directional complement
ME	measure complement
MOD	modifier
OBCOMP	comparison complement
OBJ1	direct object
OBJ2	indirect object
PC	PP-argument
PREDC	predicative complement
PREDM	predicative adjunct
SE	obligatory reflexive
SU	subject
VC	verbal complement

Table 3: An overview of *d*-labels used in the MWE patterns.

5. If the determiner position is free, then no determiner position should be specified in the pattern. From this it follows that if a determiner position is specified, the determiner variation is fixed and specified in the EXPRESSION-field, cf. (7)-(8).

   (7) academisch ziekenhuis ('university hospital')
   [.NP [.mod:A (2)] [.hd:N (3)]]

   (8) de grote stad ('the big city')
   [.NP [.det:D (1)] [.mod:A (2)] [.hd:N (3)]]

6. A **verb list**[7] should be represented as: [.hd:V (list)]

   (9) blunder maken ('make a mistake')
   [.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)]]

7. An **adjective list** should be represented as: [.mod:A1 (list)]

   (10) zwaar accent ('strong accent')
   [.NP [.mod:A1 (list)] [.hd:N1 (1)]]

8. A **reflexive verb** should be represented as: [.hd:REFLV ()], where the index refers to the position of the verb, and no index is stated for the reflexive *zich*.

   (11) zich wassen ('have a wash')
   [.VP [.hd:REFLV (2)]]

9. **Obligatory variable direct objects** → [.obj1:NP (var)]

[7] See section 3.1.4

(12) iemand aan de tand voelen (‘grill s.o.’)
[.VP [.obj1:NP (var)] [.ld:PP [.hd:P (1)] [.obj1:NP [.det:D (2)] [.hd:N (3)]]] [.hd:V (4)]]

10. **Obligatory variable indirect objects** → [.obj2:NP (var)]

(13) iemand zijn woord geven (‘give s.o. one’s word’)
[.VP [.obj2:NP (var)] [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]

11. **Obligatory indirect object consisting of the preposition *aan* and a variable complement** → [.obj2:PP [.hd:P (aan)] [.obj1:NP (var)]]

(14) het woord geven aan iemand (‘give the floor over to s.o.’)
[.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)] [.obj2:PP [.hd:P (aan)] [.obj1:NP (var)]]]

12. **PPs with a variable NP complement** → [.pc:PP [.hd (1)] [.obj1:NP (var)]]

(15) in de rij staan voor iets (‘stand in line for sth.’)
[.VP [.ld:PP [.hd:P (1)] [.obj1:NP [.det:D (2)] [.hd:N (3)]]] [.hd:V (list)] [.pc:PP [.hd:P (4)] [.obj1:NP (var)]]]

13. **Obligatory variable locative/directional complements** → [.ld (var)]

(16) ergens zijn dagen slijten (‘end one’s days somewhere’)
[.VP [.ld (var)] [.obj1:NP [.det:D (1)] [.hd:N1 (2)]] [.hd:V (3)]]

14. **Obligatory variable measure complements** → [.me (var)]

(17) elkaar ME ontlopen (‘differ ME from each other’)
[.VP [.obj1:NP [.hd:PRON (1)]] [.me (var)] [.hd:V (3)]]

15. **Obligatory variable modifiers headed by an adverb** → [.mod:AdvP (var)]

(18) MOD uit de verf komen (‘MOD live up to its promise’)
[.VP [.ld:PP [.mod:AdvP (var)] [.hd:P (1)] [.obj1:NP [.det:D (2)] [.hd:N (3)]]] [.hd:V (4)]]

16. **Obligatory variable modifiers headed by an adjective** → [.mod:AP (var)]

(19) het MOD voorbeeld geven (‘set a MOD example’)
[.VP [.obj1:NP [.det:D (1)] [.mod:AP (var)] [.hd:N1 (2)]] [.hd:V (3)]]

17. **Verbal complements consisting of a variable (*om*)-*te*-infinitive phrase** → [.vc:OTI (var)]

(20) ambitie hebben om te (‘have ambition to’)
[.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)] [.vc:OTI (var)]]

18. **Verbal complements consisting of a variable *te*-infinitive clause** → [.vc:TI (var)]

(21) het gevoel hebben te (‘have the feeling to’)
[.VP [.obj1:NP [.det:D (1)] [.hd:N2 (2)]] [.hd:V (3)] [.vc:TI (var)]]

19. **Verbal complements consisting of a variable subordinate clause** → [.vc:SSUB (var)]

    (22) het gevoel hebben dat ('have the feeling that')
    [.VP [.obj1:NP [.det:D (1)] [.hd:N2 (2)]] [.hd:V (3)] [.vc:SSUB (var)]]

20. To avoid the creation of multiple patterns for one type of MWEs, the following order of the constituents should be used in the lexicon:

    se - ld(var) - obj2:NP - obj1 - me(var) - mod:PP - ld:PP - predc - pc:PP(fixed) - **verb** - predm:PP - obj2:PP - pc:PP(var) - vc

2.4 pos

The POS-field is mainly used for maintenance reasons, i.e. with the help of the part-of-speech of an expression it is possible to limit the number of candidate pattern descriptions for that expression. The encoding of this field consists of a part-of-speech tag for each leaf node in the PATTERN-field, thus including verb lists, obligatory NP-arguments, locative complements, etc. The tags should be separated by a space. Table 4 gives an overview of the part-of-speech tags.[8]

PoS-tag	Description	PoS-tag	Description
a	adjective	advx	variable AdvP
adv	adverb	ax	variable AP
c	complementizer	ldx	variable locative/directional complement
d	determiner	npx	variable NP
n	noun	reflv	reflexive verb
p	adposition	sx	variable clause
pron	pronoun	x	fixed component
v	verb		

Table 4: An overview of part-of-speech tags.

2.5 mapping

The numbers in the MAPPING-field indicate the relation between the position of a component in the Component List (CL) and its position in the EXAMPLE-field, both part of the MWE descriptions,[9] i.e. the relation between non-inflected forms and full forms. The numbers should be separated by a space. For example, the CL of *de broek dragen* is 'de broek[sg] dragen', EXAMPLE*n* is 'zij heeft de broek gedragen' and MAPPING is '3 4 5': the first component of CL refers to the third component of EXAMPLE*n*, the second component of CL refers to the fourth component of EXAMPLE*n*, and the third component of CL refers to the fifth component of EXAMPLE*n*.

[8] The lexicon contains one class (EC70) with a mixture of expressions having a unique pattern (Grégoire, 2007b). Since the POS-field cannot be empty, the POS of this class is set to *mixture*. However, this tag should not be used for other classes.

[9] See section 3.1.3 and 3.1.5.

2.6 mwe

This field should contain an example expression representing the current pattern taken from the EXPRESSION-field in the MWE descriptions.

2.7 example

This field should contain an example sentence of MWE taken from EXAMPLE*n* in the MWE descriptions.

2.8 description

The DESCRIPTION-field should be used to describe the uniqueness of the pattern in plain text.

- *de broek dragen*: Expressions headed by a verb, taking a direct object constisting of a fixed determiner and an unmodifiable noun.
- *iemand aan banden leggen* ('restrain s.o.'): Expressions headed by a verb, taking (1) a variable direct object, and (2) a locative/directional complement headed by a fixed preposition taking a complement consisting of a fixed determiner and a limitedly modifiable noun.

3 MWE description

The description of an MWE consist of two parts:

1. A basic MWE description, which contains the following fields:
   (a) PATTERN_NAME
   (b) EXPRESSION
   (c) CL (Component List)
   (d) LISTA
   (e) LISTB
   (f) EXAMPLE
2. An additional MWE description, which contains the following fields:
   (a) SUBJECT
   (b) OBJECT
   (c) MODIFIER
   (d) RPRON
   (e) CONJUGATION
   (f) POLARITY

Section 3.1 details the encoding guidelines for the basic description fields, whereas the encoding instructions of the additional description fields are outlined in section 3.2.

It must be noted that the main focus is on representing those properties that are needed for a successful implementation of the MWE lexicon in any specific NLP system. This means that the priority is on properly describing the fields that are part of the *basic MWE description*, and although the additional description fields also form an important part of the description it cannot be guaranteed that these fields are completely filled or free from errors. Any comments regarding the MWE description should be entered in the optional COMMENT-field.

3.1 Basic MWE description

The coding instructions for each field that is part of the basic MWE description are described in the subsections below. All fields are required. Examples of actual encodings are given in appendix B.

3.1.1 pattern_name

The PATTERN_NAME is used to assign an MWE pattern description to the expression, see section 2.3. Up to three patterns can be specified for each MWE. An example of an entry with multiple patterns represented is *college geven* ('lecture'): the assignment of PATTERN_NAME1 yields the MWE *college geven*, and the assignment of PATTERN_NAME2 yields the MWE *college geven aan iemand* ('lecture s.o.').

It is allowed to mix patterns with and without a LIST-index provided that the number of indices in the PATTERN-field is equal to the number of elements in the CL-field. See section 3.1.4.

3.1.2 expression

The EXPRESSION-field should contain the obligatory fixed components of an MWE in the full form.

Representation rules The following rules apply to the representation of this field:

1. The components should be separated by a space.
2. The order of the components should match the order of the pattern (see section 2.3).
3. **Obligatory variable arguments** should not be represented in this field, but in the PATTERN-field.
4. **Possessive NPs** should be represented as *iemands*. This notation implies that the possessive NP can be realized as either an NP+*s*, or a possessive pronoun, or *de/het* NP *van* NP ('the NP of NP'). Different realizations should be explicitly marked in the COMMENT-field of an expression. An illustration is given in (23), where the possessive NP can only be realized as an NP+*s*, or a possessive pronoun, but not as *de/het* NP *van* NP ('the NP of NP').

   (23) het naar iemands zin maken
   it to s.o.'s pleasure make
   'please s.o.'

5. **Reflexives** should be represented with the canonical forms *zich* and *zichzelf*. Since these reflexives are subject bound, the form will change depending on the subject.
6. **Bound possessive pronouns** should be represented as *zijn*. Whether the pronoun is (indirect) object bound or subject bound is indicated in the CL-field.[10]
7. For pragmatic reasons, generally the main **determiners** are used in the representation: *de* ('the'), *het* ('the'), *een* ('a'), *zijn* ('one's'), *zijneigen* ('one's own'), and in an exceptional case *alle* ('all'). This approach is of course not desirable, but an attempt to group determiners in such a way that it covers all possible determiner combinations failed. Although the majority of the expressions containing an empty determiner or *een* can also occur with *geen* ('no'), one should be aware of over-generalizations.
8. **Variation of the determiner** should be represented in this field using a slash (/) to separate the alternations:

   (24) de/zijn hand ophouden
   the/his hand hold-out
   'hold out one's hand for a tip'

9. **Polarity items** should not be represented in this field, except for items that are regarded as part of the idiom:

[10] see section 3.1.3.

(25) Ik krijg *geen* gehoor.
I get no ear.
'There's no reply.'

10. There are special rules for the encoding of **modifiers**:

    (a) Expressions with a fixed modifier: the modifier should be represented in the EXPRESSION-field.

    (b) Expressions with a limited selection of known modifiers (maximum of three): for each modifier a new entry is created.

    (c) Expressions with a limited selection of modifiers, but the boundaries are unknown or more than three modifiers are known: the modifier component in the EXPRESSION-field should be represented by the most frequent modifier, the other known modifiers should be listed in the MODIFIER-field.[11]

    (d) Expressions with an unknown number of modifiers: no modifier component should be represented in the EXPRESSION-field, but the pattern should contain a variable modifier node and a list of the most frequent modifiers should be represented in the MODIFIER-field.

3.1.3 Component List (CL)

The Component List (CL) should contain the same components as the EXPRESSION-field. The difference is that the components in the CL should be in the canonical (or non-inflected) form, instead of in the full form. The following rules apply to this field:

1. The components should be separated by a space.

2. **Parameters** are used to specify the full form features of each component.[12] Table 5 gives an overview of the parameter categories and corresponding parameter values distinguished in the MWE lexicon. Since each parameter value is unique, i.e. belongs to one parameter category, we only represent the parameter value of each parameter. The parameter values are realized between square brackets directly on the right of the item they parameterize. Default values (see table 5) are not represented. Because of their uniqueness, there is no restriction on the order of values.

3. **Fixed components** are represented in their full form and are not assigned a parameter.

4. **Preposition-components** always precede their complement, even if they must be realized as a postposition.

5. **Determiners** should be represented with the form they take in the MWE (as in the EXPRESSION-field), e.g. the determiner is represented as *de* if the MWE component is a plural definite noun, irrespective of the gender of the noun.

[11] see section 3.2.4.

[12] The term *parameter* can be defined as an occurrence of the pair <parameter category,parameter value>, where *parameter category* refers to the aspect we parameterize, and *parameter value* to the value a parameter category takes (Grégoire, 2006).

PoS	PC	PC description	PV	PV description	Default
d	*dbin*	binding type	sb	subject bound	√
			dob	direct object bound	
			iob	indirect object bound	
n	*ngen*	the gender of the noun	de	definite article for masculine and feminine nouns	√
			het	definite article for neutral nouns	
n	*ncount*	the countability of the noun[13]	count	count noun	√
			mass	mass noun	
n	*nnum*	the number of the noun	infl	inflectable (when count noun)	√
			sg	singular	
			pl	plural	
n	*nfrm*	the form of the noun	pos	positive	√
			dim	diminutive	
a	*afrm*	the form of the adjective	norm	normal	√
			noe	never *-e*inflection	
			optepl	no *-e* inflection when noun is singular, optional when noun is plural.	
			noesg	no *-e* inflection when noun is singular	
			opte	*-e* inflection is optional	
			comp	comparative	
			sup	superlative	
v	*vfrm*	the form of the verb	fin	finite	√
			inf	infinitive	
			part	particle verb	
			presp	present participle	
			passp	passive participle	
p	*ppos*	the way the preposition must be realized	prep	preposition	√
			post	postposition	

Table 5: Overview of parameters, with descriptions of the parameter category (PC) and the parameter value (PV).

6. **Determinerless NPs** should be represented with the label **EMP** in the determiner position.
7. **Possessive NPs** should be represented with the label **PNP** in the determiner position.
8. NPs that can solely take **indefinite determiners** should be represented with the label **INDEF** in the determiner position.
9. **Particle verbs** should be represented with a _ between the particle and the verb.
10. If a noun does not have a singular form, then the plural form should be represented without a parameter.

[13] Ideally, a distinction between count, uncount and mass should be made.

11. **[*:canonical form*]** should be used in special cases where the full form of a component in an MWE is the colloquial form of a word, i.e. the form that is generally not used in formal speech or writing. In these cases the full MWE form should be listed in the CL followed by [*:canonical form*], where *canonical form* is replaced with the non-inflected form of the component. Examples of colloquial word forms in Dutch are *rooie* en *dooie* in (26). These components are represented in the CL as *rooie[:rood]* and *dooie[:dood]*. It must be noted that only the colloquial form can be used when the expression is used idiomatically. Substituting the normal forms *rode* and *dode* for *rooie* and *dooie*, respectively, changes the meaning of the expression in (26).

    (26) a. over de *rooie* gaan
    over the red-one go
    'flip one's lid'

    b. op zijn *dooie* akkertje
    on his dead little-land
    'dawdling'

12. **[NIL]** should be used for so-called frozen words, i.e. words that do not have a meaning when used in isolation and do not occur outside the expression. An example is *loer* in the expression *iemand een loer draaien* (lit. 's.o. a LOER turn', id. 'trick s.o.').

3.1.4 lista and listb

The LISTA-field and LISTB-field should be used to store components that can be substituted for the LIST-index in the pattern, yielding one or more expressions. This means that these fields are required for expressions with a LIST-index in the pattern. Both fields should be assigned the value *n.a.* if the expression has no LIST-index in the pattern. If only one of the fields is filled, then the other field should be represented with a hyphen (-).

The use of these fields is restricted to three types of expressions:

1. Combinations of a verb that seems to have very little semantic content and a prepositional phrase, a noun phrase or an adjectival phrase. 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.

2. Combinations of a noun and a verb that may be a regular combination, but since the exact properties of the individual components are unknown, the combination is treated as an MWE.[14]

3. Combinations of an adjective that seems to have very little semantic content and a noun that is used in its literal sense. Both components are subject to standard grammar rules, and also occur in grammatical constructions other than in the same NP, e.g. the noun as a subject and the adjective as a predicative complement.

The lexical selection of the verb and the adjective is highly restricted, but not always limited to one. The alternation of the verb or the adjective should be specified in the LIST-fields.

[14] More information on the MWE selection procedure can be found in the document *MWE Lexicon for Dutch: Report on the extraction and the selection of MWEs* (Grégoire and Villada Moirón, 2007).

Representation rules The following rules apply to the representation of these fields:

1. 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, especially with PPs. Two sets of verbs are predefined:

   (a) *blijken* ('appear') *blijven* ('remain') *gaan* ('go') *komen* ('come') *lijken* ('appear') *raken* ('get') *vallen* ('fall')[15] *worden* ('become') *zijn* ('be')

   (b) *brengen* ('bring') *doen* ('do') *geven* ('give') *hebben* ('have') *houden* ('keep') *krijgen* ('get') *maken* ('make') *zetten* ('put')

   A complement co-occurs either with verbs from set 1 or with verbs from set 2. 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 self-constructed data, it is deleted from the LISTA-field. The LISTB-field contains lexemes, either verbs or adjectives, that are not in the predefined set but do co-occur with the component(s) in the EXPRESSION-field. The information in the LISTB-field is merely based on corpus data and therefore may not be exhaustive.

2. An expression that is described as a non VP MWE may also combine with a verb. These expressions are assigned multiple patterns and the corresponding verbs are represented in the LIST-fields. It should be noted that non VP MWEs that include a noun, which alone forms an expression with a list verb, have not been represented with list verbs. An example is the NP MWE *verwoede poging* ('frantic attempt'). Since *poging* is described in the lexicon as a modifiable noun combining with the verbs *doen* ('do'), *ondernemen* ('undertake') and *wagen* ('dare'), it is assumed that the NP *verwoede poging* also combines with these verbs and this is not explicitly described in the entry of *verwoede poging*.

3. It is assumed that nouns, which are described in the lexicon as being modifiable and combining with a list verb, are taking a modifier or a complement when used in the definite form.[16] Whether the noun in these combinations can take a subordinate clause or an *om-te*-infinitive clause as a complement is explicitly represented in the description of the expression.

4. To make sure that example sentences are available for each individual expression, the lexemes in the LISTB-field must be followed by its passive participle form between round brackets, only if the lexeme is not used in the example sentence. Since the lexemes in LISTB are either verbs or adjectives, it should be noted that this rule only applies to lexemes that are verbs.

5. **Particle verbs** in the LISTB-field should be represented as in de CL-field, i.e. with a _ between the particle and the verb and with the [part] parameter value.

[15] The literal meaning of *vallen* is 'fall', but it has a variety of different meanings in MWEs of this type, including 'become', 'is experienced as', etc.

[16] Besides taking a modifier or a complement, it is also possible that the intonation implies that the noun contains given information.

3.1.5 example

The EXAMPLE-field contains an example sentence with the expression. An example sentence should be specified for each chosen pattern, which means that up to three example sentences can be specified. The only requirement of this field is that the structure should be identical for each expression with the same PATTERN_NAME. In general, this field complies with the following guidelines for MWEs that contain a verb (either a fixed verb or verb list):

1. The subject of the sentence is *hij*.
2. *iets* should be used for non-human subjects.
3. Variable components should be represented as *iets*, *iemand* or *ergens*.
4. Negative or positive components of polarity MWEs should be left out, if they would cause an extra lexeme in the EXAMPLE-field.
5. The present perfect tense is preferred to avoid particle verbs occuring as two words.

3.2 Additional MWE description

The coding instructions for each field that is part of the additional MWE description are described in the subsections below. Except for the MODIFIER-field, all fields are exclusive to VP MWEs. The default value for the MODIFIER-field is a hyphen (-). The default value for the others fields is a hyphen (-) when the expression is a VP MWE and *n.a.* in all other cases. Examples of actual encodings are given in appendix B.

3.2.1 subject

This field should be used to cover subject restrictions and can contain both a list of heads of possible subjects extracted from annotated corpora and predefined labels:

- [**het**] expletive *het* subject
- [**fem**] female subject
- [**male**] male subject
- [**sg**] singular subject
- [**pl**] plural (or mass) subject
- [**pl/sgmet**] the subject should be either plural (or mass), see (27), or else combined with a comitative-*met* clause, see (28).

  (27) zij zitten op één lijn
  they sit on one line
  'they are on the same wavelength'

  (28) hij zit op één lijn met haar
  he sits on one line with her
  'he and she are on the same wavelength'

- [**pl/sgals**] the subject should be either plural (or mass), see (29), or else combined with a comparative-*als* clause, see (30).

  (29) zij zitten op dezelfde golflengte
  they sit on the-same wavelength
  'they are on the same wavelength'

  (30) hij zit op dezelfde golflengte als zij
  he sits on the-same wavelength as she
  'he and she are on the same wavelength'

- [**anim**] animate subject
- [**non-anim**] non-animate subject
- [**het**] the subject must be the expletive *het* ('the').

- [**hetssub**] besides a normal realization of the subject, the subject can also be a clause starting with a complementizer and either occupying the first position of the sentence, see (31), or occupying any other position combined with the anticipatory subject *het* in the subject position, see (32).

  (31) dat hij komt speelt geen rol
  that he comes plays no role

  (32) het speelt geen rol dat hij komt
  it plays no role that he comes

- [**nohetssub**] besides a normal realization of the subject, the subject can also be a clause starting with a complementizer occupying any position of the sentence other than the first, without the anticipatory subject *het* in the subject position.

  (33) daarbij speelde een belangrijke rol, dat hij niet komt
  with-that played an important role, that he not comes

- [**hetvp**] besides a normal realization of the subject, the subject can also be an infinitive clause occupying any position of the sentence other than the first combined with the anticipatory subject *het* in the subject position.

  (34) het heeft geen zin eerder weg te gaan
  it has no sense sooner away to go
  'it makes no sense to go sooner'

notation Predefined labels are between square brackets and should precede subject examples. The subject examples are either heads of possible subjects taken from corpus data followed by a space and the absolute frequency that denotes the number of occurrences of the head in combination with the tuple specified in the data record, see Grégoire and Villada Moirón (2007), or heads of possible subjects not taken from corpus data and therefore not followed by a frequency. Pairs of a subject and its optional frequency are separated by a comma. An example is given in (35).

(35) [no-anim][hetssub] gerucht 820,verhaal 522,die 369,naam 90,bericht 71,dat 59,speculatie 57,theorie 48,grap 30,anekdote 17,

3.2.2 object

This field should be used to cover object restrictions and can contain both a list of possible objects extracted from annotated corpora and predefined labels:

- [**pl**] plural (or mass) object
- [**anim**] animate object
- [**non-anim**] non-animate object

- **[hetssub]** besides a normal realization of the object, the object can also be a clause starting with a complementizer and either occupying the first position of the sentence, see (36), or occupying any other position combined with the anticipatory object *het* in the object position, see (37).

  (36) dat zij wegging nam hij mij kwalijk
  that she left took he me badly
  'he blamed me that she left'

  (37) hij nam het mij kwalijk dat zij wegging
  he took it me badly that she left
  'he blamed me that she left'

- **[nohetssub]** besides a normal realization of the object, the object can also be a clause starting with a complementizer occupying any position of the sentence other than the first, without the anticipatory object *het* in the object position.

  (38) daarbij nam hij mij kwalijk dat zij wegging
  with-that took he me badly that she left
  'furthermore he blamed me that she left'

- **[hetvp]** besides a normal realization of the object, the object can also be an infinitive clause combined with the anticipatory object *het* in the object position.

  (39) hij heeft het in zijn hoofd gehaald weg te gaan
  he has it in his head get away to go
  'he got the idea to leave'

- **[nohetvp]** besides a normal realization of the object, the object can also be an infinitive clause without the anticipatory object *het* in the object position.

  (40) daarbij maakte hij duidelijk weg te willen
  with-that made he clear away to want
  'furthermore he made it clear that he wants to leave'

notation Predefined labels are between square brackets and should precede object examples. The object examples are either heads of possible objects taken from corpus data followed by a space and the absolute frequency that denotes the number of occurrences of the head in combination with the tuple specified in the data record, see Grégoire and Villada Moirón (2007), or heads of possible objects not taken from corpus data and therefore not followed by a frequency. Pairs of an object and its optional frequency are separated by a comma. An example is given in (41).

(41) [no-anim] schuld 105,moord 19,verantwoordelijkheid 11,dood 6,deel 4,brand 4,nederlaag 3,alles 3,opvatting 3,ding 3,

3.2.3 rpron

This field should be used to encode pronominalized PP realizations, and can contain the following predefined labels:

- **[ssub]** the complement of the pronominalized PP must be realized as a clause starting with a complementizer.

  (42) hij had er een hekel aan dat zij kwam
  he had there a hackle to that she came
  'he hated it that she came'

- **[vp]** the complement of the pronominalized PP must be realized as an infinitive clause.

  (43) hij heeft er een hekel aan weg te moeten
  he has there a hackle to away to have
  'he hates to leave'

3.2.4 modifier

This field should be used to list both obligatory variable modifiers and modifiers that modify a (limitedly) modifiable noun. In the current encoding this field is mainly filled with modifiers coming from extracted data. These modifiers are in the non-inflected form, followed by a space and the absolute frequency denoting the number of occurrences of the modifier in combination with the tuple specified in the data record. *NO* stands for no modifier. When multiple data records are included in a single MWE description, then the modifier examples of the most frequent tuple are taken. Modifiers that should not be in the list are deleted, and in some cases more modifiers are added without a frequency. An example is given in (44).

(44) NO 142,groot 64,eerste 13,nieuw 7,duur 7,spectaculair 6,goed 6,klein 5,gericht 5,

There are no actual encodings with multiple modifier constituents, but if this occurs, the following rules should be followed:

1. Adjectives modifying the first noun of the expression should be preceded by {1}.
2. Adjectives modifying the second noun of the expression should be preceded by {2}.
3. All other modifiers should precede premodifiers.

3.2.5 conjugation

This field is used to specify whether the head of the expression conjugates with *zijn* ('to be') represented as *z*, or *hebben* ('to have') represented as *h*, or both represented as *b*.

3.2.6 polarity

This field takes a hyphen (-) by default and takes the value *NPI* if an expression can only occur in negative environments, and *PPI* if an expression cannot occur in positive environments.

References

Grégoire, N. (2006), Elaborating the parameterized equivalence class method for Dutch, *in* N. Calzolari (ed.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, ELRA, Genoa, Italy, pp. 1894–99.

Grégoire, N. (2007a), MWE lexicon for Dutch: Conversion procedure manual, *Technical report*, STEVIN IRME.

Grégoire, N. (2007b), MWE lexicon for Dutch: Overview of MWE pattern descriptions, *Technical report*, STEVIN IRME.

Grégoire, N. and Villada Moirón, B. (2007), MWE lexicon for Dutch: Report on the extraction and the selection of MWEs, *Technical report*, STEVIN IRME.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, M. (1997), *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff and Wolters Plantyn, Groningen en Deurne.

Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I. and van der Wouden, T. (2003), CGN syntactische annotatie.

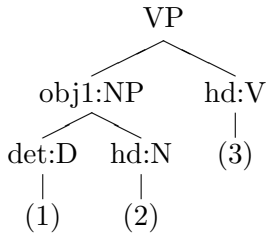
A Examples of MWE pattern descriptions

This appendix shows some examples of MWE pattern descriptions in the following format:

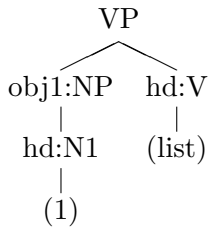
- PATTERN_NAME – POS
- PATTERN
- MAPPING – EXAMPLE_MWE – EXAMPLE_SENTENCE
- DESCRIPTION
- COMMENT
- graphical representation of the PATTERN

A complete overview of the pattern descriptions including a graphical representation of the dependency tree is given in the document *MWE Lexicon for Dutch: Overview of MWE pattern descriptions* (Grégoire, 2007b).

- ec1 – d n v
- [.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]
- 3 4 5 – de stormbal hijsen – hij heeft de stormbal gehesen
- Expressions headed by a verb, taking a direct object constisting of a fixed determiner and an unmodifiable noun.

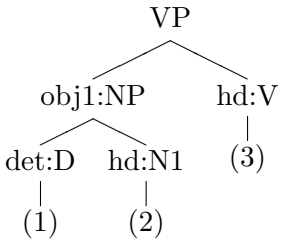


- ec2 – n v
- [.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)]]
- 4 – blunder – hij heeft een blunder gemaakt
- Expressions headed by a verb, taking a direct object consisting of a modifiable noun (list).

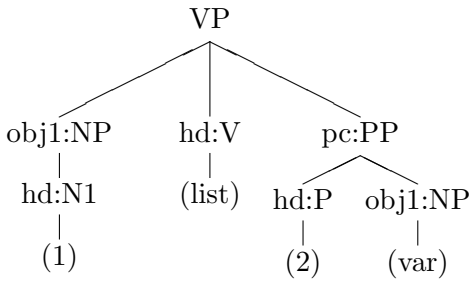


- ec5 – n v
- tba5
- 3 – actie – hij heeft actie gevoerd
- Expressions headed by a verb, taking a direct object consisting of a modifiable noun (list).
- The nouns in this class behave both like count nouns (as they can both be plural and singular) and as mass nouns (as they can take an empty determiner and indefinites such as "veel' and 'weinig'). [.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)]]

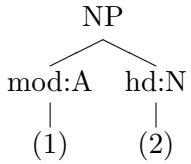
- ec7 – d n v
- [.VP [.obj1:NP [.det:D (1)] [.hd:N1 (2)]] [.hd:V (3)]]
- 3 4 5 – zijn debuut maken – hij heeft zijn debuut gemaakt
- Expressions headed by a verb, taking a direct object consisting of a fixed determiner and a modifiable noun.



- ec9 – n v p npx
- [.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)] [.pc:PP [.hd:P (2)] [.obj1:NP (var)]]]
- 4 6 – bod op – hij heeft een bod gedaan op iets
- Expressions headed by a verb, taking (1) a direct object consisting of a modifiable noun, and (2) a PP-argument consisting of fixed preposition and a variable complement (list).



- ec101 – a n
- [.NP [.mod:A (1)] [.hd:N (2)]]
- 1 2 – achtergestelde lening – achtergestelde lening
- Expressions headed by a fixed noun, taking a modifier consisting of an unmodifiable adjective.



B Examples of MWE descriptions

This appendix shows some examples of MWE descriptions in the following format:

- EXPRESSION – CL
  - PATTERN1 – EXAMPLE1
  - LISTA – LISTB
  - CONJUGATION – POLARITY
  - SUBJECT
  - OBJECT
  - RPRON
  - MODIFIER
- zijn kansen waarnemen – zijn kans[pl] waar_nemen[part]
  - ec1 – hij heeft zijn kansen waargenomen
  - n.a. – n.a.
  - h – -
  - -
  - -
  - -
  - -
- een oog dichtdoen – een oog[het][sg] dicht_doen[part]
  - ec1 – hij heeft geen oog dichtgedaan
  - n.a. – n.a.
  - h – NPI
  - -
  - -
  - -
  - -
- concert – concert[het]
  - ec2 – hij heeft een concert gegeven
  - - – geven
  - h – -
  - -
  - -
  - -
  - NO 1057,eenmalig 73,eerste 42,laat 34,extra 32,gratis 18,uitverkocht 13,tweede 12,groot 8,slechts 6,

- oorlog – oorlog
  - ec5 – hij heeft oorlog gevoerd
  - - – voeren
  - h – -
  - -
  - -
  - -
  - NO 1028,heilig 45,eigen 24,koud 15,bloedig 15,vuil 13,echt 10,jarenlang 9,psychologisch 7,klein 6,
- een reputatie opbouwen – een reputatie[sg] op_bouwen[part]
  - ec7 – hij heeft een reputatie opgebouwd
  - n.a. – n.a.
  - h – -
  - -
  - -
  - -
  - NO 139,goed 26,groot 19,stevig 14,uitstekend 11,internationaal 7,zeker 7,enorm 6,behoorlijk 6,bedenkelijk 6,
- zin in – zin[mass] in
  - ec9 – hij heeft geen zin gehad in iets
  - hebben houden krijgen – -
  - h – -
  - -
  - -
  - [vp]
  - NO 8581,goed 17,ontzettend 8,zeker 6,slechts 6,enorm 5,alleen 5,gewoon 4,juist 4,erg 3,
- respect voor – respect[het][mass] voor
  - ec9 – hij heeft veel respect gehad voor iemand
  - hebben houden krijgen – -
  - h – -
  - -
  - -
  - [ssub]
  - NO 1509,groot 154,diep 57,enorm 15,echt 9,min 7,grenzeloos 6,meest 4,zeker 3,slechts 3,

- de helpen[presp] hand[sg] bieden – n.a.
  - ec52 – hij heeft iemand de helpende hand geboden
  - n.a. – -
  - - –
  - -
  - -
  - -
  - hij heeft de helpende hand geboden
- achtergestelde lening – achter_stellen[part][passp] lening
  - ec101 – achtergestelde lening
  - n.a. – -
  - –
  - n.a.
  - n.a.
  - n.a.
  - -