STEVIN-IRME project

MWE Lexicon for Dutch
Report on the extraction and the selection of MWEs

Nicole Grégoire and Begoña Villada Moirón

October 27, 2007

Contents

1 Introduction

The *Lexicon of Dutch MWEs* is one of the results of the project *Identification and Representation of Multiword Expressions* (IRME). The IRME project has been carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments.[1]

The *Lexicon of Dutch MWEs* contains lexical descriptions of 5,000 multiword expressions (MWEs), which meets the criterion of being highly theory- and implementation-independent. The main purpose of the lexicon is for it to be used in various Dutch NLP systems.

This document describes the extraction and selection of the MWEs included in the lexicon. The identification of candidate expressions has been done by Begoña Villada Moirón working at the University of Groningen. The selection of true MWEs and their representation in the lexicon has been done by Nicole Grégoire under supervision of Jan Odijk, both affiliated with the University of Utrecht.

The document starts with discussing the data extraction in section 2. This is followed by describing the selection procedure of MWEs and their representation in the lexicon in section 3. The appendix A gives an overview of the data records format that formed the input for the MWE selection.

[1] http://taalunieversum.org/stevin

2 Data extraction

Source material The candidate expressions[2] for the *Lexicon of Dutch MWEs* are extracted from the Dutch CLEF corpus, a collection of newspaper articles from 1994–1995, taken from the Dutch daily newspapers Algemeen Dagblad and NRC Handelsblad. The corpus contains 80 million words and 4 million sentences. 90.8% of the sentences have been annotated automatically with the Alpino parser.[3]

Evidence of morpho-syntactic information is extracted from the Twente Nieuws Corpus (TwNC) (Ordelman, 2002). The TwNC comprises 500 million words of newspaper text and television news reports. The corpus has also been syntactically annotated with the Alpino parser (in Fall 2006).

Automated MWE identification The automated extraction of MWEs requires predefined patterns. In this case five syntactic patterns, shown in (1), are chosen. The choice for these patterns was made after an exhaustive study of a random selection of MWEs taken from the Van Dale Lexical Information System (VLIS) database.

(1)
1. direct object - verb (OB1_V)
2. (variable noun phrase) - prepositional phrase - verb ((NP)_PP_V)
3. indirect object - direct object - verb (OBJ2_OBJ1_V)
4. adjective - noun (A_N)
5. noun - prepositional phrase (N_PP)
6. preposition - noun - preposition (P_N_P)

The tuples extracted from the corpus form the input for the identification models. After experimenting with various machine learning techniques, we applied a decision tree classifier since this method worked best for us. The classifier learns a notion of MWE-hood on the basis of training data that consists of a collection of tuples and a number of features that encode linguistic information. Such features measure the lexical affinity between the component words, the syntactic flexibility, the strength of the dependence between the words, passivizability, etc. In addition, each tuple in the training data specifies its class, whether an MWE or non-MWE. To annotate the training data, two existing lexical databases have been used, VLIS and the RBN (*Referentiebestand Nederlands* 'Reference Database of The Dutch Language') (Martin and Maks, 2005).

The identification model also makes use of an absolute frequency threshold, i.e. tuples that occur infrequently are not taken into account since they would introduce noise and degrade the performance of the classifier. Empirically, we established a desirable threshold per syntactic pattern, see table 1. The chosen threshold was the one yielding the best performance of the classifier.

The decision tree classifier proposes a class (MWE|noMWE) for each input tuple. Although the class is accompanied by a probability that suggests how confident the classifier is in assigning a given class to a tuple, no use has been made of this probability. The identification provides a list of candidate expressions, i.e. tuples that are assigned the class

[2]For convenience we speak of *candidate expressions*, in practice, the expressions extracted from the corpus are actual lemma pairs, triples or quadruples, i.e. combinations of two, three or four words, depending on the pattern of the extracted data, that may form an MWE or may be part of an MWE.

[3]http://www.let.rug.nl/vannoord/alp/Alpino/

pattern	threshold used
OB1_V	f>=10
(NP)_PP_V	f>=10
OBJ2_OBJ1_V	f>=10
A_N	f>=50
N_PP	f>=30
P_N_P	f>=50

Table 1: Used absolute frequency threshold for each pattern.

MWE, yielding a total of 9,451 expressions, see table 2. No manual filtering or correction has been applied to this list at this stage.

pattern	# of candidate expressions
OB1_V	3,894
(NP)_PP_V	2,405
OBJ2_OBJ1_V	202
A_N	1,001
N_PP	1342
P_N_P	607
Total	**9,451**

Table 2: Distribution of candidate expressions over the extracted patterns.

Collect morpho-syntactic information Next, morpho-syntactic information about the candidate expressions is collected from the bigger corpus. The decision of using the CLEF corpus for the extraction of candidate expressions and the TwNC corpus for the extraction of morphosyntactic information was purely pragmatic. A syntactic pattern may be very productive showing in a huge number of tuples. For each of these tuples, many features have to be collected from the annotated corpus, which means that we could easily end up with gigabytes of data for further processing. To keep the data down to a manageable size, data extraction (for identification) of a very productive pattern was done from the CLEF corpus, and data extraction of a less productive pattern was done from the TwNC corpus.

MWEs allow morpho-syntactic variation, e.g. verbs may show different forms depending on tense, person, number, voice; nouns may allow number alternation, etc. To facilitate both the automated extraction of MWEs from the corpus and the collection of morpho-syntactic information, we adopted the following representation of an expression: all related forms of a content word (noun, verb and adjective) are represented with root forms, whereas functional words are represented with their surface form. Inside noun phrases, determiners are ignored. Examples of the representation of candidate expressions are shown in (2)-(4).

(2) *de benen nemen* (lit. 'to take the legs', id. 'to escape'): neem#been

(3) *dikke maatjes* (lit. 'fat friends', id. 'good friends'): dik#maat

(4) *de kat de bel aanbinden* (lit. 'to fasten the bell to the cat', id. 'to bell the cat'): bind_aan#bel#kat

For each candidate expression a set of properties was extracted:

1. the subcategorization frame assigned by the Alpino parser
2. the absolute frequency
3. the size of the used corpus
4. subject information
5. number information of the nouns (values: *sg* for singular and *pl* for plural)
6. diminutive information (values: *dim* for diminutive and *nodim* for no diminutive)
7. determiner information
8. pre-modifier information
9. post-modifier information

The tuples and their properties are represented in predefined formats, see appendix A.

3 Data processing

The candidate expressions and their properties form the input for the data processing. The processing of the data is a manual procedure that includes two stages:

1. Selecting true MWEs, i.e. analysing the tuples, their properties and examples and determining true MWEs according to predefined criteria, and
2. representing the selected MWE and its properties in the *Lexicon of Dutch MWEs*.

MWE selection The MWEs are selected according to the definition given in (5).

(5) A multiword expression is 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 further specified as:

- Lexical properties: one component within the expression is selected by another component. The lexical item selection is fixed or very limited, i.e. a selected component cannot be substituted without changing the meaning of the whole expression. Two Dutch examples are:

  (6) zware/*sterke shag
  heavy/*strong tobacco
  'heavy tobacco'

  (7) een fout maken/begaan/*doen
  a mistake make/commit/*do
  'make a mistake'

- Morphological properties, e.g. *e*-inflection on the noun: *ten gevolg**e** van* ('because of').
- Syntactic properties, e.g. the lack of a determiner preceding a singular count noun, which is prohibited in standard Dutch grammar: *in opdracht van* ('by order of').
- Semantic properties: the meaning of the expression cannot be deduced from the meaning of the individual components, e.g.:

  (8) uit de boot vallen
  out the boat fall
  'to be eliminated'

  (9) met de handen in het haar zitten
  with the hands in the hair sit
  'to be at a loss what to do'

The morphological, syntactic and semantic properties of an analysed expression often lead to a clear non-discussable decision of whether the expression is a true MWE. Deciding whether a combination is a true MWE solely on the basis of its lexical properties is not always as clear-cut, especially not for direct object - verb combinations, since in many cases not all properties of the individual components are known.

An example of a clear MWE is *een gesprek voeren* ('have a conversation'): although one meaning of *voeren* is "being actively occupied with", and although one can be actively occupied with a conversation, the combination is unpredictable since *gesprek* cannot be substituted by its synonym *praatje* ('chat'), i.e. *een praatje voeren* ('have a chat') is out. For this reason, *een gesprek voeren* is a true MWE and thus entered in the lexicon.

An illustration of a not so clear-cut example is the expression *een getuigenis afleggen* ('to give a testimony'), the extracted data contain five other nouns that occur with *afleggen*, three of which requiring the same meaning of *afleggen* as required by the noun *getuigenis*: *verklaring* ('statement/testimony'), *eed* ('oath'), and *bekentenis* ('confession'). The question is whether the lexical selection of the noun is predictable according to its semantic properties. In this case we are not sure, since we do not know which semantic proporties a noun that selects the verb *afleggen* requires. Although the expression seems semantically regular, we have no clear evidence since we have no acces to a full list of synonyms of one of the components. Concretely this means that in this case all four expression are included in the lexicon.

One test that can be used in cases where it is hard to tell whether a candidate can be classified as a true MWE, is to translate the expression to another language. If the translation of one component varies based on the component it co-occurs with, then the combination is likely to be selected for the lexicon. This can be illustrated with the verb *afleggen*: *afleggen* translates into *make/give* when it combines with *verklaring*, to *give* when it combines with *getuigenis*, to *take* when it combines with *eed*, and to *make* when it combines with *bekentenis*. Since the English translation of *afleggen* depends on the noun it combines with, it needs to be included in the lexicon. Although this is a valid criterion, it is not decisive.

One incompletion of the extracted data is that a single data record can contain a lemma tuple that is part of multiple MWEs. An example of such a data record is:

```
heb#hand
frame transitive_ndev 1280,np_ld_pp 181,aci_simple 22,np_aan_het 14,
freq 1497
corpus 500M words
hd heb
subject hij 149,die 96,ik 70,ze 67,je 46,zij 28,we 27,politie 20,god 12,wie 12,
compl1 hand
hd1 hand
hdcomp1
dep1 obj1 1497,
mor1 sg 908,pl 589,
dim1 nodim 1497,
det1 de 696,een 235,NO 208,geen 90,zijn 74,hun 62,haar 33,twee 15,onze 15,mijn 13,
premod1 NO 875,gelukkig 123,vrij 118,schoon 75,vuil 46,groot 17,goed 17,vast 13,
postmod1 NO 1186,in 115,van 99,op 24,bij 14,vol 12,voor 11,aan 9,om 5,die 5,
328.xml|Mijn vader had het druk - hij had zijn handen vol om een boterham te verdienen .
```

```
234.xml|De Kustwacht in het gebied is sinds begin 1996 operationeel en heeft de handen
        vol aan drugssmokkelaars op de route Colombia-Europa .
469.xml|Hij is ' een pianist die vier handen leek te hebben , zoveel noten speelde hij .
958.xml|Het Iraakse regime heeft de hand gehad in de dood van meerdere vooraanstaande
        sjiitische leiders .
452.xml|Ook daar had God de hand in , meent T.Q. In ieder geval gaf het hem een nieuw
        doel in z'n leven : T.Q. besloot zanger te worden .
796.xml|De Amerikaanse meisjes hadden hun handen op de gebogen knieen .
```

The example sentences in this record contain at least four different expressions, each containing the extracted tuple *hand* and *hebben*:

(10) a. de vrije hand hebben
the free hand have
'have a free hand'

b. een gelukkige hand hebben
a lucky hand have
'be lucky'

c. de hand hebben in iets
the hand have in sth.
'have a hand in sth.'

d. de handen vol hebben aan iets
the hands full have on sth.
'have one's hands full with sth.'

Knowing that the extracted pattern was direct object - verb, we know that *hebben* is a verb and also the head of the expression and that *hand* is a noun and the head of the direct object. Since *hand hebben* is not an expression, the first step is to take into account the determiners: the determiner *de* ('the') is the most frequent determiner, but *de hand hebben* is not an MWE, in fact no combination of a determiner and *hand hebben* forms an expression. This means, that knowing that the extracted data is of the form direct object - verb, and knowing that the combination determiner+*hand*+*hebben* is not an MWE, we may reject this data record and move on to the next one. However, from experiences with the data we know that we must also take into account the premodifier values (*premod1*), postmodifier values (*postmod1*), and morphology and diminutive information (*mor1* and *dim1*). Since no explicit search has been done for e.g. a direct object that contains an adjective and a noun, these combinations have been created and checked manually using both language knowledge and in some cases a dictionary.

To illustrate, looking at the *premod1* values, we see that *gelukkig* ('lucky') and *vrij* ('free') are the most occurring premodifiers of *hand* (when combined with *hebben*). Given our knowledge of the Dutch language, we know that this information yields the expressions *de vrije hand hebben* and *een gelukkige hand hebben*. Moreover, from language knowledge and the presence of example sentences we can add the expressions *de hand hebben in iets* and *de handen vol hebben aan iets* to the *Lexicon of Dutch MWEs*.

To conclude, MWEs for the lexicon are selected from lists of candidate expressions, their properties and example sentences according to the definition given in (5). One data record may contain a lemma tuple that is part of multiple MWEs, the lexicon entries of which are created using both language knowledge and in some cases a dictionary.

MWE representation Various aspects played a role in the representation as it is in the *Lexicon of Dutch MWEs*. The main requirement of the standard encoding is, however, 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) (Odijk, 2004). 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 consisting of MWEs with the same pattern, it requires some manual work to convert one instance of each EC into a system specific representation, but all other members of the same EC can be done fully automatically.

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 and the Rosetta MT system (Rosetta, 1994).

The selected MWEs are represented in the *Lexicon of Dutch MWEs* in a uniform format, which contains the following description fields:

1. ID
2. PATTERN_NAME1
3. PATTERN_NAME2
4. PATTERN_NAME3
5. EXPRESSION
6. CL
7. LISTA
8. LISTB
9. SUBJECT
10. OBJECT
11. RPRON
12. MODIFIER
13. EXAMPLE1
14. EXAMPLE2
15. EXAMPLE3
16. CONJUGATION
17. POLARITY
18. COMMENTS

The data record(s) corresponding to an MWE description are stored in a file with the same name as the entry's ID. As one can see, the MWE description fields contain a PATTERN-field, the value of which refers to an MWE pattern description. The majority of the MWEs (over 4,500) have been assigned a pattern value, yielding 2 or more MWEs having the same pattern, i.e. belonging to the same EC. A small number of MWEs in the lexicon have a unique pattern, and although these expressions must be analyzed properly, creating a new EC for each of the expressions does not contribute the main requirement of the lexicon, see Grégoire (2007b) for more information on this point.

The encoding guidelines of the description fields, as well as the description fields of the MWE patterns, are documented in the *MWE lexicon for Dutch: Encoding protocol* (Grégoire, 2007a).

References

Grégoire, N. (2007a), Design and implementation of a lexicon of Dutch multiword expressions, *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, Association for Computational Linguistics, Prague, Czech Republic, pp. 17–24.

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

Martin, W. and Maks, I. (2005), Referentie bestand nederlands documentatie, *Technical report*, INL.

Odijk, J. (2004), 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.

A Format of the data records

This appendix gives an overview of the format of the various data records. Besides a description of the attributes, a concrete example is given for each extracted pattern. The examples have been adapted for display reasons.

A.1 General format

1. Data records are separated by a *blank line*.
2. Attributes within a data record are separated by a *new line*.
3. An attribute and its first value are separated by a *tab*.
4. Multiple values of one attribute are separated by a *comma*.
5. Values and absolute frequencies are separated by a *space*.
6. NO must be used for no premodifier, no postmodifier and empty determiner values: de 4,NO 3,een 2.
7. Represent a maximum of six corpus examples for each candidate expression.

A.2 OBJ1_V

root forms of the candidate expression separated by #

frame: subcategorization frame assigned by the Alpino parser

freq: absolute frequency of the tuple

corpus: corpus size

hd: head of the candidate expression

subject: subject information (with a maximum of 10 values)

compl1: complement

hd1: head of *compl1*

dep1: dependency label of *compl1* (value: *obj1*)

mor1: number information of *hd1* (values: *sg pl*)

dim1: diminutive information of *hd1* (values: *dim nodim*)

det1: determiner information of *hd1*

premod1: premodifier information of *hd1* (with a maximum of 10 values)

postmod1: postmodifier information of *hd1* (with a maximum of 10 values)

up to six examples sentences

```
leg_af#verklaring
frame    ninv(transitive,part_transitive(af)) 805,ninv(np_ld_pp,part_np_ld_pp(af)) 64,
freq     883
corpus   500M words
hd       leg_af
subject  hij 121,die 67,na 58,ze 35,verdachte 31,getuige 27,Van der G. 21,ik 19,
compl1   verklaring
hd1      verklaring
dep1     obj1 883,
mor1     sg 515,pl 368,
dim1     nodim 883,
det1     een 363,NO 275,geen 99,de 35,zijn 27,hun 13,deze 12,zulk 9,die 8,veel 7,
premod1  NO 483,belast 96,tegenstrijdig 50,vals 39,ontlast 18,beken 11,kort 9,
postmod1 NO 773,over 58,voor 13,in 13,onder 9,bij 4,van 3,tegen 2,daarover 2,die 1,
parool19990105_1036.xml|Zijn tegenstanders zeggen dat de premier een tribunaal vreest,
omdat zijn vroegere commandanten dan belastende verklaringen kunnen afleggen.
parool19990114_1064.xml|Nadat Burns die verklaring had afgelegd,
parool19990121_1072.xml|Velthuis heeft voor Willing belastende verklaringen afgelegd.
parool19990122_735.xml|Van de brandweer zal onder anderen Carel Boer een verklaring
afleggen en van de belangengroep Klankbord Bos .
parool19990123_1743.xml|Er waren zo veel verklaringen afgelegd.
parool19990126_837.xml|Daar moest ik verklaringen over afleggen;
```

A.3 (NP)_PP_V

root forms of the candidate expression separated by #. The variable noun phrase is either *nul*, i.e. there is no variable NP, or *np*, i.e. there is a variable NP.

frame: subcategorization frame assigned by the Alpino parser

freq: absolute frequency of the tuple

corpus: corpus size

hd: head of the candidate expression

subject: subject information (with a maximum of 10 values)

comp1: first complement. This takes either the value NO if there is no variable NP, or a list of the 10 most occurring values.

dep1: dependency label of *comp1* (value: *obj1*)

comp2: second complement

dep2: dependency label of *comp2*

hd2: head of *comp2*

hdcomp2: head of the complement of *comp2*

mor2: number information of *hdcomp2* (values: *sg pl*)

dim2: diminutive information of *hdcomp2* (values: *dim nodim*)

det2: determiner information of *hdcomp2*

premod2: premodifier information of *hdcomp2* (with a maximum of 10 values)

postmod2: postmodifier information of *hdcomp2* (with a maximum of 10 values)

up to six examples sentences

```
sta#nul#onder#druk
frame    nonp_copula 4833,ld_pp 819,so_nonp_copula 140,intransitive 127,
freq     5987
corpus   500M words
hd       sta
subject  die 232,hij 140,koers 134,relatie 117,resultaat 117,prijs 109,marge 108,
comp1    NO 5987,
dep1     obj1
comp2    onder druk
dep2     predc 4946,ld 613,mod 428,
hd2      onder
hdcomp2  druk
mor2     sg 5987,
dim2     nodim 5987,
det2     NO 5915,een 26,de 17,welk 6,zo'n 6,geen 4,enig 3,veel 2,die 2,meer 2,
premod2  NO 4457,groot 695,zwaar 447,neem_toe 88,enorm 51,sterk 36,hoog 29,
postmod2 NO 4987,van 525,door 243,na 40,vanwege 29,als 29,in 20,wegens 19,
```

A.4 OBJ2_OBJ1_V

root forms of the candidate expression separated by #.

frame: subcategorization frame assigned by the Alpino parser

freq: absolute frequency of the tuple

corpus: corpus size

hd: head of the candidate expression

subject: subject information (with a maximum of 10 values)

comp1: first complement

det1: determiner information of *comp1*

premod1: premodifier information of *comp1* (with a maximum of 10 values)

postmod1: postmodifier information of *comp1* (with a maximum of 10 values)

mor1: number information of *comp1* (values: *sg pl*)

dim1: diminutive information of *comp1* (values: *dim nodim*)

dep1: dependency label of *compl1* (value: *obj1*)

hd2: head of the second complement

mor2: number information of *hd2* (values: *sg pl*)

dim2: diminutive information of *hd2* (values: *dim nodim*)

det2: determiner information of *hd2*

premod2: premodifier information of *hd2* (with a maximum of 10 values)

postmod2: postmodifier information of *hd2* (with a maximum of 10 values)

dep2: dependency label of *hd2* (value: *obj2*)

```
bind_aan#bel#kat
frame    ninv(np_np,part_np_np(aan)) 121,
freq     121
corpus   500M words
hd       bind_aan
subject  die 53,iemand 12,hij 10,initiatief_groep 7,fractievoorzitter 3,beide 3,wij 3,
comp1    bel 121,
det1     de 121,
premod1  NO 121,
postmod1 NO 121,
mor1     sg 121,
dim1     nodim 121,
dep1     dr1(obj1) 121,
hd2      kat
mor2     sg 121,
dim2     nodim 121,
det2     de 121,
premod2  NO 121,
postmod2 NO 121,
dep2     dr2(obj2) 121,
```

A.5 A_N

root forms of the candidate expression separated by #

freq: absolute frequency of the tuple

corpus: corpus size

hd: head of the candidate expression

hdmod: head of the modifier

dep: dependency information of the whole candidate expression

mor1: number information of *hd* (values: *sg pl*)

dim1: diminutive information of *hd* (values: *dim nodim*)

det1: determiner information of the whole candidate expression

premod1: premodifier information of the whole candidate expression (with a maximum of 10 values). The first value is the adjective that forms the candidate expression.

postmod1: postmodifier information of the whole candidate expression (with a maximum of 10 values)

```
open#dag
freq     161
corpus   80M words
hd       dag
hdmod    open
dep      mod 89,obj1 33,su 22,pc 10,ld 4,predc 2,obj2 1,
mor1     sg 130,pl 31,
dim1     nodim 161,
det1     de 94,een 34,NO 19,deze 3,haar 3,die 2,hun 1,of 1,zulk 1,tien 1,
premod1  open 151,jaarlijks 4,eerste 3,landelijk 1,jaar 1,jaarlijke 1,
postmod1 NO 94,van 38,in 7,voor 6,op 6,die 2,bij 2,naast 1,11 1,Oude 1,
```

A.6 N_PP

root forms of the candidate expression separated by #.

freq: absolute frequency of the combination noun1#prep followed by the absolute frequency of the whole candidate expression

corpus: corpus size

hd: head of the candidate expression

det1: determiner information of *hd*

premod1: premodifier information of *hd* (with a maximum of 10 values)

postmod1: postmodifier information of *hd* (with a maximum of 10 values)

mor1: number information of *hd* (values: *sg pl*)

dim1: diminutive information of *hd* (values: *dim nodim*)

compl: complement followed by

1. The difference between the relative frequency of the noun with the highest relative frequency and the average relative frequency of each noun in the set of all nouns, and
2. the label VAR if this difference is smaller than 0.8, assuming that the noun with the highest frequency is a variable direct object of the preposition, and the label FIXED if this difference is bigger than 0.8, assuming that the noun with the highest frequency is a fixed direct object of the preposition.

hd1: head of the complement followed by

hdcomp: head of the complement of *hd1*

det2: determiner information of *hdcomp*

premod2: premodifier information of *hdcomp* (with a maximum of 10 values)

postmod2: postmodifier information of *hdcomp* (with a maximum of 10 values)

mor2: number information of *hdcomp* (values: *sg pl*)

dim2: diminutive information of *hdcomp* (values: *dim nodim*)

```
raad#van#bestuur
freq     5745 2583
corpus   ca.160M words
hd       raad
det1     de 5112,een 384,NO 99,zijn 42,deze 24,Fokker 18,geen 15,soort 9,hun 6,zo'n 6,
premod1  NO 5040,nieuw 69,Europees 66,wijs 27,heel 27,ook 27,eigen 18,voltallig 15,
postmod1 van 5739,die 3,council 3,
mor1     sg 5568,pl 177,
dim1     nodim 5745,
compl    van bestuur
hd1      van
hdcomp   bestuur 0.45 (VAR)
det2     NO 2577,het 6,
premod2  NO 2583,
postmod2 NO 1536,van 987,die 6,waarvan 6,Publieke 6,waaronder 6,met 3,over 3,uit 3,
mor2     sg 2583,
dim2     nodim 2583,
```

A.7 P_N_P

root forms of the candidate expression separated by #

freq: absolute frequency of the tuple

corpus: corpus size

hd: head of the candidate expression

compl: complement

det: determiner information of the noun

premod: premodifier information of the noun

postmod: postmodifier information of the noun

mor: number information of the noun

dim: diminutive information of the noun

```
in#plaats#van
freq    10242
corpus  ca.400M words
hd      in
compl   plaats van
det     NO 9811,de 404,een 7,zijn 7,hun 2,die 2,het 2,deze 1,600 1,elk 1,
premod  NO 10196,eerste 33,tweede 3,ander 2,meest 1,smerig 1,divers 1,belangrijk 1,
postmod NO 9835,van 398,om 2,maar 2,voor 1,Jan 1,Pordenone 1,in 1,op 1,
mor     na 9835,sg 400,pl 7,
dim     nodim 10242,
```