

# Specifications of tools to acquire valence patterns and morphosyntactic restrictions

Begoña Villada Moirón  
University of Groningen  
m.b.villada.moiron@rug.nl

Draft, January 31,2007

## Abstract

## 1 Introduction

Multiword expressions (MWEs) typically exhibit some idiosyncratic behavior at the morpho-syntactic level of analysis. The motivation to specify the morpho-syntactic restrictions of MWEs is two-fold: (i) to decide the lexical representation of the MWE in a lexical database (here, the IRME lexical database) and (ii) to specify what morpho-syntactic constraints need to apply to obtain a MWE interpretation in ambiguous expressions (i.e. expressions that could be interpreted literally or figuratively).

Regarding **morphology**, the component words may (or not) exhibit productive morphology. The morphological productivity (or lack of it) observed in the MWE use may differ from the productivity observed when the words are used in isolation (literal use). If a different behavior is observed, morphological restrictions need be specified in the lexical entry of the MWE.

Let us look at an example. *Schoen* is a countable noun. The noun may allow number variation (singular:*schoen*, plural:*schoenen*) and also diminutive *schoentje*. In its literal use, we find occurrences of singular (1), plural (2) and diminutive (3). In a MWE use, we typically find only one form, for example in (4), (5) and (6). Note, however, that morphological productivity is possible in some MWEs. Therefore, we need tools to establish how much productivity MWEs exhibit in large corpora.

- (1) Onder de [schoen] is een extra platte zool aangebracht om de stabiliteit te vergroten .
- (2) De conditie was goed , de [schoenen] ingelopen en drinken en druiven-suiker in de rugzak .

- (3) Zij, als enige, past het schoentje.
- (4) Je moet wel stevig in je [schoenen] staan.
- (5) Onze politici schuiven te gemakkelijk de schuld in de schoenen van de transportsector”, aldus de organisatie.
- (6) Alles wat fout is in deze wereld wordt in de schoenen van het Amerikaanse beleid geschoven (TwNc02, ad20011002\_13.xml)

At the morphological level, important restrictions to specify are:

- number variation: singular, plural
- diminutive
- compounding (prefix or other POS attached to the base noun)
- comparative and superlative use in adjectives
- tense inflection in verbs

At the **syntactic** level, the combination of words and/or phrases in the MWE could also exhibit a syntactic behavior that cannot be predicted from the behavior of the individual words in isolation. The following aspects need to be taken into account:

- valence pattern of the verb: list of the verb’s required syntactic dependents (phrasal categories)
- with regard to NPs whose head noun is restricted to a particular lexeme
  - determiner variation in NPs
  - prenominal modification: insertion of adjective
  - post-nominal modification (PPs, relative clause, clausal complement)
- presence/absence of adverbs between the MWE constituents
- passive
- topicalization
- wh-extraction
- clefting, etc.

Often, a phrase in a MWE shows more syntactic rigidity than the same phrase in a productive and literal expression use. Once again, we need to describe which syntactic ‘transformations’ a MWE may undergo while still keeping its meaning/interpretation as MWE.

Level		Possible values
Morphology	number	singular ( <b>sg</b> ), plural ( <b>pl</b> )
	diminutive	yes ( <b>dim</b> ), no ( <b>nodim</b> )
Syntax	determiner	possessive, (in)definite, quantifier
	prenominal adjective	none or lexeme
	postnominal modifier	PP head preposition
		relative pronoun
	adverbs	none, lexeme
	valence pattern	list of syntactic constituents
	passive	(yes,no), auxiliary verb

Table 1: Morpho-syntactic restrictions specified in the IRME lexical database

### 1.1 Focus of our work

In the IRME MWE database, each MWE specifies the morpho-syntactic restrictions shown in table 1 (if applicable):

### 1.2 Overview of the remainder of the paper

Section 2 describes our attempts to extract the morpho-syntactic restrictions from large syntactically annotated corpora. Section 3 briefly describes a first attempt to establish the valence pattern shown by the verb in the MWE. Section 4 outlines the next steps to follow in automatically acquiring valence patterns.

## 2 Acquisition of morphosyntactic restrictions

Before applying the automatic identification methods described in deliverable 2.4, potential candidate *ntuples* (tuples, triples) are extracted from a syntactically annotated corpus. Together with candidate triples, other information is extracted from the corpus, such information is used to code learning features when applying machine learning techniques for MWE identification. I re-use such information to establish the necessary morpho-syntactic restrictions that apply to a MWE.

### 2.1 Corpus

I used the CLEF corpus that contains approx. 80M words. The corpus was automatically annotated with the Alpino parser. The syntactically annotated data is encoded as XML trees. This facilitates the search and extraction of interesting linguistic contexts concerning any expression in the corpus.

## 2.2 Method

To facilitate the extraction of morpho-syntactic restrictions, Prolog is used to extract information from dependency structures. Such dependency structures can easily be accessed with the Alpino parser.<sup>1</sup> These structures have been successfully used in many applications, such as feature extraction for parsing disambiguation, extraction of terms, synonyms, improving question answering, etc. A dependency structure corresponding to example (5) is shown in figure 1.

Generally speaking, the same extraction procedure is applied to all MWES (candidates), with minor differences depending on the syntactic pattern in which we are interested.

- With `dtsearch`, collect all corpus sentences in which we find one of the following dependency patterns: (i) verb+NP (object1), (ii) verb+(NP, object1)+PP, (iii) adjective+noun, etc.
- Using Alpino, XML dependency structures are accessed as Prolog dependency triples. From these dependency triples, we extract the necessary information from each pattern instance observed in the corpus. From the example (6), the information collected about the VERB NP PP pattern instance (IETS IN DE SCHOENEN VAN SCHUIVEN) is shown in figure 2. Figure 3 shows rather similar information about another instance of the pattern IETS IN DE SCHOENEN VAN SCHUIVEN.
- After collecting all the described information for all instances of the desired syntactic pattern, we have frequency counts of each realization.

## 2.3 Results

For each piece of information (determiner, adjective, subject, etc.) we give the 5 most frequent realizations (if at least 5 were observed) and their relative frequency.<sup>2</sup> Figure 4 shows the summary of the information we collected regarding the VERB NP pattern *de aftocht blazen*.

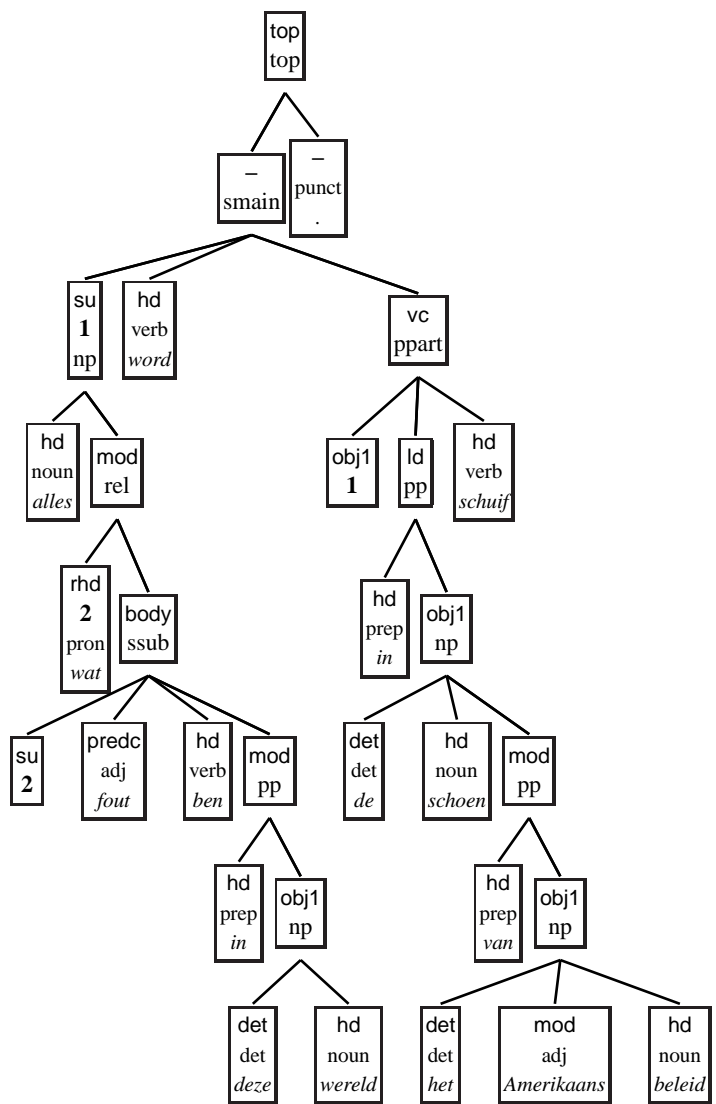
## 2.4 Limitations encountered

- Meaning ambiguity: A pattern instance (e.g. IN HAND HOUDEN) may have a literal use and one/more idiomatic uses. The extracted realizations and frequency counts could mix up morpho-syntactic information belonging to one or more uses of a pattern. This is actually a general

---

<sup>1</sup>Documentation about how to extract information from dependency structures is available at <http://odur.let.rug.nl/~vannoord/alp/Alpino/TreebankTools.html>.

<sup>2</sup>Instead of relative frequency, raw frequency will be given in the new data delivery.



```

in,schoen,ipr|      # preposition,noun,PP position
schuif,0|           # verb,non finite
np_ld_pp|          # verb frame
ld|                # dependency relation
na|                # subject
alles|             # accusative object
yes,word|          # passive, auxiliary verb
[det,de]|          # determiner POS, lexeme
[adj,none]|        # prenominal modifier
[prep,van]|        # postnominal modifier
pl,nodim,n|        # NOUN number, diminutive, non-pronoun
none|              # adverbs between constituents
TwNC-02/ad2001/ad20011002/ad20011002_13.xml #file identifier

```

Figure 2: Information extracted from the dependency structures in the expression IETS IN DE SCHOENEN VAN SCHUIVEN.

```

in,schoen,pr3|     # preposition,noun,PP position
schuif,0|          # verb,non finite
np_np_ld_pp|      # verb frame
ld|               # dependency relation
vrouw|           # subject
moord|            # accusative object
no,_|             # passive, auxiliary verb
[det,zijn]|       # determiner
[adj,none]|       # prenominal modifier
none|             # postnominal modifier
pl,nodim,n|       # NOUN number, diminutive, non-pronoun
none|             # adverbs between constituents
TwNC-02/ad2001/ad20011023/ad20011023_620.xml #file identifier

```

Figure 3: Information extracted from the dependency structures in the expression IETS IN DE SCHOENEN VAN SCHUIVEN.

```

aftocht blaas#aftocht
  freq      24
  file      clef/AD19940107/AD19940107-0060-321-4.xml
  sub       hij (0.083)
  frame     transitive (1.000)
  dr        obj1 (1.000) np_ld_pp ()
  pas       NO (na) YES (word,na)
  mor       sg (1.000)
  dim       nodim (1.000)
  pre-mod   NONE (0.792) eervol (0.042) elegant (0.042)
           smadelijk (0.042) NA (0.042)
  post-mod  uit (0.042)
  det       hun (0.042) de (0.833) een (0.125)
  adv       NONE (0.917) opnieuw (0.042) vervolgens (0.042)

```

Figure 4: Most frequent morpho-syntactic realizations observed in the pattern *de aftocht blazen* and their relative frequency in the `clef` corpus.

problem applying to automated corpus-based statistical methods. Further research is needed (not within this project) to develop a tool that disambiguates literal from figurative uses of MWEs.

- Annotation: MWEs described in the Alpino lexicon are almost always annotated as a fixed expression. Alpino annotates those instances showing no variation as a fixed SVP, instances allowing modification or variation as a flexible SVP and alternatively, as a syntactically regular expression. Consequently, the DR dependency relation between the verb and the required constituent will show at least two values: SVP and some other relation.
- Annotation: For expressions annotated as a fixed MWE, we ignore morphological information about the head noun (number, diminutive use, etc.). These affects all instances analyzed as SVP. Output information displays `na` (not available) value.
- Morpho-syntactic restrictions of unparseable sentences containing MWEs cannot be extracted with this method.
- Neither noun compounding nor adjective inflection (comparative, superlative) is recorded at the moment.
- Gathering the same information from the TwNC02 corpus, ca. 500M words, takes more time than I expected. The problem is that the files with all raw features have a big size which requires a lot of processing memory while keeping frequency counts, lexical realizations, etc. Other computing techniques have been tried, such as, storing

tables on temporary files, storing tables on disk using Perl Berkeley database files however, unsuccessfully.

### 3 Acquisition of valence pattern

At the moment the simplest technique is used. Thus, we extract the frame that has been assigned by the parser to the head verb in a MWE. After seeing all instances of a given MWE, the frame with the highest relative frequency is chosen. Sometimes, frames are erroneous, thus we need a better technique. Further investigation onto existing probabilistic techniques to extract subcategorization information is needed.

### 4 Future work plan

1. Decide a general format to represent morpho-syntactic restrictions applying to the different MWE syntactic patterns.
2. scaling up to using all TwNC02 data
3. deliver data to Nicole
  - Morpho-syntactic restrictions of (i) expressions common to VLIS or RBN and TwNC02 and (ii) expressions proposed as MWEs by automated identification methods.
  - 10 examples of each expression from existing corpora.
4. First, establish valence pattern of a MWE, then, extract morpho-syntactic restrictions when we know the valence pattern of the MWE. This should slightly improve the problem encountered with ambiguity of the literal/figurative meaning.
5. A more general method that works for all subtypes of MWE and all syntactic patterns.