

IRME

Identification and Representation of Multiword Expressions

Gosse Bouma¹ Nicole Grégoire² Gertjan van Noord¹
Jan Odijk² Begoña Villada Moirón¹ Johan Zuidema³

¹Alfa Informatica
University of Groningen

²Uil-OTS
University of Utrecht

³Van Dale Lexicografie BV

- IRME
- Automatic identification of multiword expressions
- Lexical representation of multiword expressions

Outline

- **IRME**
- Automatic identification of multiword expressions
- Lexical representation of multiword expressions

Project aims

- DEVELOP methods and tools for handling MULTIWORD EXPRESSIONS (MWES)
 - **automatic identification** in a large corpus
 - **lexical representation** in a lexical database
- PRODUCE corpus-based lexical database of Dutch MWES (5,000 entries)
- EVALUATE database externally; its usability in two independent NLP-systems for Dutch

Outline

- IRME
- **Automatic identification of multiword expressions**
- Lexical representation of multiword expressions

Automatic identification of multiword expressions

- What qualifies as multiword expression (MWE)?
- How to capture distinction between MWEs and productive combinations?
- Can one method perform well on all subtypes of MWEs?
- How to establish subcategorization frame, variability and modification?

What are multiword expressions?

- Dat compliment **steek** ik natuurlijk graag **in** mijn **zak**
(NH2003/05/31)
- Rosenmöller (GroenLinks) voorspelde dat de vakbonden weer "cao-machines" zullen worden, die elkaar met looneisen **de loef afsteken** en de goede doelen **uit het oog verliezen** (AD1994/02/09)
- Hadden deze Raspoetins **het** Tsjetsjeense **varkentje** afgelopen weken geruisloos **gewassen**, dan zouden ze het door het Westen bejubelde democratiseringsproces **onder leiding** van Jeltsin ook stilletjes **ten grave** hebben kunnen **dragen** (NH1994/12/28)

What qualifies as multiword expression?

Expressions whose linguistic behavior is not predictable from the linguistic behavior of its component words when they occur in isolation.

Idiosyncratic behavior manifest as . . .

- 1 *de laatste hand *op/aan iets *plaatsen/leggen*
- 2 . . . die ook Douglas tijdlang *aan het *lijn/lijntje houdt*.
- 3 Directeur Bontrop steekt niet onder stoelen of banken dat hij op deze manier ook een **wetenschappelijke** *oog in het zeil* kan *houden*.
- 4 Maar schaatsen *zit ons in het bloed*, zegt Karlstad.

Identification seen as classification

Given a large corpus, automatically annotated with the Alpino parser,

- 1 extract instances of syntactic pattern (VERB PP, VERB NP)
- 2 quantify linguistic properties as features
- 3 apply binary classifier to label candidates
- 4 evaluate and select discriminating features

Extraction of VERB PP patterns

component	feature	value
VERB	lexeme	zit
	tense	1 (finite)
	frame	ld_pp
SUBJ	lexeme	muziek
OBJ1		
PP	lexemes	in,bloed
	NP_det	haar
	NP_mod	
	NP_num	sg
	NP_pronoun	0
	location	fo3
	dependency relation	ld

Quantifying linguistic properties

Property	Feature	Description
Lexical		
Lexical affinity	$c(V, P, N)$ $sal(V, PN)$	count instances statistical dependence between words
Local context		
Head dependence	$H(P, N)$	(P,N) distribution entropy
Dependency relation	mod, pc	assigned by the parser
Morpho-syntactic		
PP position in verb cluster modifiability passive	$f((V, P, N), pos='ipr')$	how often (P,N) argument immediately precedes Verb entropy on determiners, adj. how often passivized

Quantifying linguistic properties. An example

- 1 *iets in het bloed zitten*
- 2 *iets uit het oog verliezen*
- 3 *iets krijgen in het jaar ...*

Property	Feature	(1)	(2)	(3)
Lexical affinity	$c(V, P, N)$	89	118	211
	$sal(V, PN)$	31.09	70.18	2.54
Head dependence	$H(P, N)$	3.069	0.93	5.87
Dependency relation	mod, pc	svp	svp	mod
PP position in verb cluster	$f((V, P, N), pos='ipr')$	96.15	1	0.51
modifiability		4.69	0	7.622
passive		0	0.09	0.004

Classification results

- Decision trees classifier (WEKA J48) applied on
 - 8,501 candidates (verb ObjectNoun prep noun, verb prep noun)
- Gold standards: Van Dale data and RBN
- 10-fold cross validation
- Baseline (salience) 80%
- Classification accuracy 90%

Results (2)

- Overall accuracy is pretty good.
- Precision and recall of actual MWEs needs improvement (F-Measure, 0.478 (y) vs. 0.937 (n))
- Annotation errors, among these
 - + 100 expressions classified as MWE missing in gold standards or not considered a MWE
- Approximation of idiosyncratic behavior at different linguistic levels improves performance.
- Modifiability and passive, not sufficiently discriminating
- Method identifies MWEs not found in gold standards

Not found in gold standards

np ten grave dragen
aan de macht komen
iem. in de war brengen
np te dood brengen
in opstand komen
in premiere gaan
tot leven komen
np te leen krijgen
bij kas zitten
in opstand komen
np in gijzeling nemen
in het midden laten
naar werk gaan
ter beschikking komen
np naar buiten brengen

Future work

- Capturing linguistic properties
 - semantic compositionality using (i) selectional restrictions (van de Cruys' method) and (ii) translations (Villada Moirón and Tiedemann, 2006)
- Other syntactic patterns: VERB NP, VERB ADJ, VERB PART
- Subcategorization frame, variability and modification

Outline

- IRME
- Automatic identification of multiword expressions
- **Lexical representation of multiword expressions**

Goal

Making available a large number of lexical entries for MWEs for the use in various NLP systems.

Goal

Making available a large number of lexical entries for MWEs for the use in various NLP systems.

Criterion

High degree of theory- and implementation-independence

Focus

Verbal idioms

MWEs headed by a verb (non-finite in the canonical form) with a fixed (or very limited) item selection and which meaning cannot be obtained compositionally from the meaning of its parts when used in isolation.

Equivalence Classes (EC)

Group idioms according to their structure.

EC example

Expressions headed by a verb taking a direct object NP that consists of a determiner and a singular noun:

- de plaat poetsen
- de boot missen
- de kar trekken

Why ECs?

Converting the standard representation into a system specific representation:

manually

de plaat poetsen



automatically

de boot missen

de kar trekken

.

.

Parameterized ECs

Group idioms that are for a large part identical and only differ e.g. in the number of the noun it requires:
de plaat-SG poetsen vs. de benen-PL nemen

Parameters

- parameter = <parameter category (PC),parameter value (PV)>
- PC refers to the aspect we parameterize.
- PV refers to the value a PC takes.
- We distinguish 7 PCs and 23 PVs.
- E.g.: <num,SG>, <num,PL>, <afirm,SUP>

First results

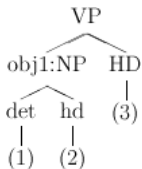
- The introduction of parameters decreases the number of equivalence classes needed with almost 90% with respect to the numbers of equivalence classes needed in the original ECM.
- A total of 15 parameterized equivalence classes are needed to cover 95% (or 1,109) of the three- and four-word idioms.

Idiom description

- 1 PATTERN: EC1
- 2 EXPRESSION: de benen nemen
- 3 ICL (Idiom Component List): de been[het][pl] nemen
- 4 OPTIONAL: NULL
- 5 EXAMPLE: hij heeft de benen genomen
- 6 MEANING: ervandoor gaan
- 7 CONJUGATION: hebben
- 8 VERB TYPE: transitive
- 9 POLARITY: none

Idiom pattern description

- 1 PATTERN NAME: EC1
- 2 PATTERN: [.VP [.obj1:NP [.det (1)] [.hd (2)]] [.HD (3)]]



- 3 DESCRIPTION: Expressions headed by a verb, taking a fixed direct object consisting of a determiner and a noun.

Initial version of the database

1,000 VP idioms taken from the *Referentiebestand Nederlands*

Future work

- Extend the method to other types of MWEs.
- Test the method in two Dutch parsers: Alpino and Rosetta MT.