



Universiteit Utrecht

The STEVIN IRME Project

Jan Odijk

STEVIN Midterm Workshop
Rotterdam, June 27, 2008

Utrecht Institute of Linguistics OTS

U | STEVIN



IRME

- Identification and lexical Representation of **M**ultiword Expressions (MWEs)
- Participants:
 - Uil-OTS, Utrecht
 - Nicole Grégoire, Jan Odijk, André Schenk
 - Alpha-Informatica, Groningen
 - Begoña Villada Moirón, Gertjan van Noord, Gosse Bouma
 - Van Dale, Utrecht
 - Johan Zuidema
- Start: Jun 1, 2005; End: Aug 31, 2007
- <http://www.uil-ots.let.uu.nl/irme/>



MWEs

- Multiword Expressions (MWEs) are combinations of words that have (linguistic) properties that cannot be derived from the properties of the individual words or the way they are combined by some grammar / language model



MWEs

- Zo laat hij tegen het einde van de oorlog de beide NSB-directeuren **Rost van Tonningen** (president) en Robertson, die **het** na **Dolle Dinsdag op een lopen zetten**, niet **zonder meer de plaat poetsen**, maar (...).
- Om zijn gebrekkigheden **dragen** de Engelsen Fitz **op handen**.
- “**Sterker nog**, ze proberen werknemers een pakket verslechtingen **in de maag** te **splitsen** waar je **koud van wordt**.”
- De Feyenoorders **lopen de longen uit hun lijf** en spelen de beste wedstrijd van het seizoen.
- De operatie in maart **had scherpe kritiek tot gevolg**.

Translated MWES



Major Goals

- Research
 - Research into innovative methods for identification of MWEs and their properties in text corpora.
 - Research into an innovative manner of lexically representing MWEs for NLP based on the Equivalence Class Method (ECM)
- Resources
 - Set of MWEs identified in text corpora and their properties
 - Corpus-based MWE Database lexically represented in accordance with the ECM.



Identification

- Research into innovative methods for identification of MWEs and their properties in text corpora.
- Set of MWEs identified in text corpora and their properties



Identification

- Main Idea
 - Idiosyncratic behavior → likely MWEs
 - Strong lexical affinity
 - Limited morphological productivity
 - Limited syntactic flexibility
 - Non- or partial semantic compositionality



Identification

- Modelling linguistic properties
 - Strong lexical affinity
 - High co-occurrence frequency, log-likelihood
 - Limited morphological productivity
 - Entropy over noun/verb realizations
 - Non- or partial semantic compositionality
 - Predictable translations in parallel corpora
 - Substitution with semantically related nouns



Identification

- Supervised Methods
 - Decision Trees (C4.5)
 - Maximum Entropy Classifiers
- Test data based on Van Dale VLIS and RBN databases
- Corpora used
 - Twente News Corpus
 - CLEF Corpus
 - Both automatically parsed by Alpino



Identification

- Unsupervised Methods
 - Using Automatic Word-Alignment in Parallel Corpora
 - Translation does not result from combining translations of individual components → likely MWE
 - Assess translations of candidate expressions and measure translational entropy
 - Using Semantic Clustering and Selectional Preference
 - Noun does not allow substitution by semantically related noun → non-compositional, likely MWE
 - Noun clusters built using distributional similarity measures
 - Candidates with no selectional preference between head lexeme and noun replacement → non-compositional, likely MWE



Identification

- Resource created
 - Patterns extracted:
 - NP V, (NP) PP V, NP NP V, A N, N PP
 - Properties:
 - Frame,
 - frequency and realization of subject, complement, determiners, adjectival modifiers, postmodifiers
 - Morphological variation



Identification: Conclusions

- State of the art results
- Methods are very useful for updating and enlarging lexica (as also confirmed by Van Dale).
- But several improvements are still possible
 - Improved recall and precision
 - Actually identified is a tuple or triple of words
 - Which is often not the full MWE
 - Which sometimes are part of multiple different MWEs
 - Heb#hand (have#hand)→
 - de vrije hand hebben ‘to have the free hand’ = ‘to be unrestrained’
 - een gelukkige hand hebben ‘to have a lucky hand’ = ‘be lucky’
 - ergens de hand in hebben ‘have the hand in something’ = ‘control something’
 - de handen vol hebben aan iets ‘to have the hands full to something’ = ‘be occupied with something’



Identification: Conclusions

- improvements are still possible (cont.)
 - automatic methods cannot distinguish literal from idiomatic uses of polysemous expressions.
 - Tested methods issue a binary classification; ranking method may be more suitable to model the continuum of multiword expression-hood.
 - Further improvements are possible by combining the above methods.



Lexical Representation

- Research into an innovative manner of lexically representing MWEs for NLP based on the Equivalence Class Method (ECM)
- Corpus-based MWE Database lexically represented in accordance with the ECM.



Equivalence Class Method

- Classification of MWEs in equivalence classes on the basis of their syntactic structures
- Procedure to incorporate MWEs thus represented into NLP systems



Equivalence Class Method

- ECM originally only for idioms
- Extended to other types of MWEs, esp. Support Verb Constructions
- Parameterised variant of ECM further developed and elaborated for Dutch



Equivalence Class Method

- ECM tested successfully by incorporating MWEs into two completely different NLP systems for Dutch
 - Alpino
 - Robust syntactic parser for Dutch
 - implemented
 - Rosetta
 - MT System NL-EN-ES
 - Specified in detail



MWE Database

- Lexical database containing 5000 MWEs created in accordance with the parameterised ECM
- MWEs selected from MWEs identified in large text corpora (frequency)
 - TWNC02 (500M tokens)
 - CLEF corpus (80M tokens)
- MWE properties supported by occurrences in these corpora



MWE Database- Extra

- Syntactic structure for each equivalence class
- based on CGN syntactic structures (adapted for MWEs) – *de facto* standard for Dutch
- → allows one to use other methods than ECM for NLP systems that can deal with CGN(-like) structures
 - E.g. Alpino
 - But not Rosetta
- Carried out experiments with such a different method for Alpino
- Carried out small experiment on Alpino with lexicon extended with MWEs derived from MWE database:
 - `concept accuracy' for sentences containing MWEs improves
 - `concept accuracy' for sentences not containing MWEs does not deteriorate



MWE Database

- Lexical database and documentation available
- Validated by CST, Copenhagen
- Small corrections made
- Revised version available via HLT-Central
- Extra: Web-based GUI to the database



Dissemination

- 17 presentations on IRME
- 6 publications in conference and workshop proceedings
- 1 journal submission
- Co-organized 3 workshops on MWEs
- <http://www.uil-ots.let.uu.nl/irme/>
- data via HLT-Central



Summary

- Research Goals achieved
- Resource Goals achieved
- Some extra's
- Results available to everyone via articles and reports and the data via HLT-Central
- Still a lot of research questions remain open, and resources can be extended and enhanced
- Part of the research continued in a PhD project
- <http://www.uil-ots.let.uu.nl/irme/>



Universiteit Utrecht



Utrecht Institute of Linguistics OTS

ŋ | STEVIN



Translation

- **Rost van Tonningen** (name)
- **Het op een lopen zetten** 'put it on a run' = 'run away'
- **Dolle Dinsdag** 'Crazy Tuesday' (name)
- (niet) **zonder meer** '(not) without more' = (not) simply
- **de plaat poetsen** 'polish the plate' = 'bolt'
- (iemand) **op handen dragen** = 'carry someone on hands' = 'admire someone'
- **Sterker nog** 'stronger yet' 'even less likely'
- (iemand) (iets) **in de maag splitsen** 'split something into someone's stomach' = 'put someone up with something'
- (ergens) **koud van worden** 'get cold from something' = 'get frightened from something'
- **de longen uit zijn lijf lopen** 'walk the lungs out of one's body' = 'run very much/very fast'
- **tot gevolg hebben** 'have till consequence' = 'have as a consequence'
- **scherpe kritiek** 'sharp criticism' = 'strong criticism'
- **Back to MWE Examples**



Identification

(NP) PP V pattern					
	Method	Accuracy	Precision	Recall	F-score
	Baseline	77,49	–	–	–
	Supervised				
	Decision Trees	83,4	0,66	0,53	0,59
	Maximum Entropy	81,79	0,62	0,39	0,47
	Unsupervised				
	Translation-based	93,2	–	–	–
	Substitution-based	82,84	0,55	0,47	0,55