

Capturing idiosyncratic linguistic behavior for automatic Multiword Expression identification

Begoña Villada Moirón
University of Groningen, The Netherlands

Saarbrücken, November 23, 2006

The issue investigated . . .

- Automated method to extract multiword expressions (MWES) from large corpora.
- Ideally, a technique applicable to all subtypes of MWES.
- Lists of MWES together with syntactic frame, modifiability information and frequency.

Talk outline

- What are Multiword Expressions (MWES)?
- The landscape of MWES
- What linguistic characteristics?
- Motivation
 - ★ In general, for NLP
 - ★ Specific to the IRME project
- Recent approaches to MWE identification
- Identification of Dutch MWES
- Discussion
- Conclusions and future work

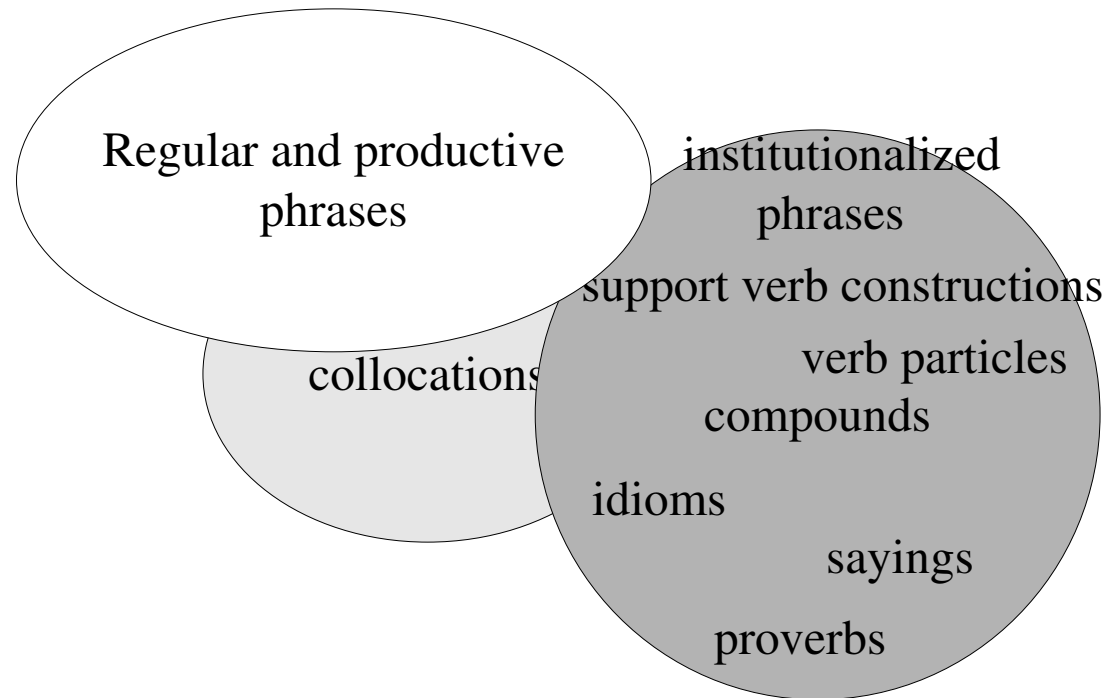
What are Multiword Expressions?

“expressions whose linguistic behavior is not predictable from the linguistic behavior of its component words when they occur in isolation [Calzolari et al., 2002]”.

“idiosyncratic combinations that cross word boundaries (or spaces) [Sag et al., 2002]”

Sag et al. [2002] focus on the mismatch between the interpretation of the MWE as a whole and the standard meaning of the individual words that compose the expression.

The landscape of Multiword Expressions



What linguistic characteristics define MWEs?

The idiosyncratic behavior of MWEs characterized by “a lack of compositionality manifest at different levels of analysis, namely, lexical, morphological, syntactic, semantic, pragmatic and statistical [Baldwin, 2006]”.

1. A strong lexical affinity between component words.
2. Restricted morpho-syntactic flexibility.
3. Partially or non-fully compositional meaning.
4. Statistically, an idiosyncratic behavior.

Idiosyncratic linguistic behavior

Linguistic properties	MWES	Productive expression
Lexeme level		
ordinary lexemes	✓	✓
<i>nonce</i> words		?
strong lexical affinity	✓	✓ (houden, van)
Morphology level		
singular/plural morpheme		✓
diminutive	✓	✓
archaic forms		
tense inflection (verb)	✓	✓
prefix/compounding		✓
Semantics (meaning)		
(fully) transparent meaning		✓
opaque meaning (idiomatic)	✓	✓
Syntax (structure)		
modification,	LIMITED NP ₁	✓
quantification, determiners		✓
syntactic versatility (topicalization, passivization, etc.)	?	✓
open slots (words/phrases)	NP ₁ ADJ NP ₁	✓ NP

Motivation. NLP

NLP applications that require some sort of semantic interpretation encounter difficulties with multiword expressions

- Machine translation
- Text summarization and paraphrasing
- (Deep) parsing
- (Multilingual) information retrieval

Motivation. IRME project

Identification and Representation of Multiword Expressions (IRME)

- Create a lexical database of Dutch Multiword Expressions (5,000 entries)
- Investigate automated identification methods that identify MWEs in large corpora
- Each MWE entry specifies lexical components, syntactic frame, morpho-syntactic productivity and lexical variation.
- Establish a theory-neutral lexical representation of MWEs
- Task-based evaluation: test database usability in wide-coverage parsing.

MWE **identification methods**

Dictionaries

Translations

Substitution

Statistical

Distributional

Classification taken from [Baldwin \[2006\]](#) and [McCarthy \[2006\]](#)

Dictionary-based methods

Hashimoto et al. [2006] idiom identification and literal-figurative sense disambiguation. A syntactic pattern in a text is checked against an idiom dictionary. If idiom description constraints satisfied, tag figurative sense, else literal sense.

Piao et al. [2006] method to measure compositionality (and to identify non-compositional expressions). Measures semantic distance between a MWE expression and its component words.

Translation-based methods

Melamed [1997] measures semantic entropy among translations of a word. A word with high semantic entropy is potentially very ambiguous and/or highly context-dependent.

Villada Moiron and Tiedemann [2006] measure translational entropy and overlap between actual translation and word-to-word literal translation in a parallel corpus. Expressions with high translational entropy and low overlap are likely to be non-compositional.

Substitution methods

Only compositional expressions allow synonym replacement [Pearce, 2001, 2002]. Using WordNet and large corpus, measure in what degree an expression allows synonym replacement.

Compare pointwise mutual information (PMI) value of an expression e_1 with average PMI value of alternative realizations e_2 to e_n . Bigger divergence found in non-compositional expressions. Requires automatically generated thesaurus [Lin, 1999]. Properties aimed at: collocativity [Venkatapathy and Joshi, 2005] and lexical fixedness [Fazly and Stevenson, 2006].

Statistical methods

Wermter and Hahn [2004] Quantitative measure of collocativity based on the frequency of the most characteristic realization of an expression (modifiability) and the likelihood of seeing the expression in a corpus.

Baldwin [2005] compares various techniques: (i) purely statistical (frequency, PMI, log-likelihood, Dice, χ^2), (ii) purely linguistic (stranded preposition freq, distance conditioned frequency, VP distance ratio), (iii) Skew divergence and (iv) a combined system. Combined system improved performance. F-scores 0.41 (BNC).

Fazly et al. [2005] model distinction between literal and figurative usages of polysemous verbs. Compositionality measured as (i) association between light verb and complement added up to (ii) the difference between the association (LV,N) with positive syntactic patterns and their association with negative syntactic patterns.

Venkatapathy and Joshi [2005] model compositionality. Collocation-based features and context-based features fed into a Support Vector Machine algorithm. Reasonable correlation established between an SVM ranking function and human judgements (Pearson's rank cor. = 0.448)

Fazly and Stevenson [2006] use lexical fixedness and syntactic flexibility as partial indicators of semantic analysability and hence idiomaticity (VP). Combining both metrics, accuracy goes up to 80%.

Tan et al. [2006] view *light verb construction* identification as a supervised classification problem. Use random forest classifier. Features: association measures (PMI, salience), metrics from earlier work [Fazly et al., 2005], lexical features (light verb (Y/N), DET, ADJ), etc. Best performing system F-measure 0.576.

Distributional methods

Using selectional preferences, [McCarthy et al. \[2003\]](#) builds on Lin's (1998) work on constructing a thesaurus automatically. Positive correlation between `sameparticle` and `simplexasneighbor` and, human compositionality judgements. `sameparticle` - number of neighbors of the phrasal showing same particle. The more neighbors, the more likely that particle keeps meaning. `simplexasneighbor` checks whether simplex verb occurs as neighbor of phrase; if yes, phrase is probably compositional.

Using Latent Semantic Analysis (LSA). Compare vector representing a MWE candidate to vector corresponding to individual verb component [Baldwin et al. \[2003\]](#), [Katz and Giesbrecht \[2006\]](#)

Use LSA to compute (i) dissimilarity between a MWE candidate and its VERB and (ii) similarity between a MWE candidate and verbal form of object N [Venkatapathy and Joshi \[2005\]](#),

Talk outline

- What are Multiword Expressions (MWES)?
- The landscape of MWES
- What linguistic characteristics?
- Motivation
- Recent approaches to MWE identification
- **Identification of Dutch MWES**
- Discussion
- Conclusions and future work

Identification of Dutch MWES

Assume the following are discriminating features:

- A strong lexical affinity between component words.
- Restricted morpho-syntactic flexibility.
- Partially or non-fully compositional meaning.
- Statistically, an idiosyncratic behavior.

Furthermore, we know that,

- Not all idiosyncrasies are observed in MWES.
- Not all MWES exhibit the same idiosyncrasies.
- MWES may undergo productive morpho-syntactic processes (like productive word combinations).

Identification = classification task

- A successful **identification model** has to check the candidate expressions against a list of properties (or idiosyncrasies) manifest at different linguistic levels,

CANDIDATES	properties				MWE yes/no
	a_1	a_2	a_3	a_4	
exp ₁			✓	✓	yes
exp ₂		✓	✓		yes
exp ₃				✓	yes
exp ₄	✓				no
exp ₅	✓				no
...					
exp _n		✓	✓	✓	yes

- ... afterwards, the model can better decide whether a candidate is a MWE or not.

Procedure

Given a collection of candidates,

1. Select characteristics that may split MWEs from productive expressions.
2. Measure the strength of each **characteristic** in candidate expressions.
3. Use quantitative measurements as **attributes** in classification task.
4. Candidate expressions showing MWE characteristics get YES label.

Checking linguistic characteristics at different levels

Lexical lexical affinity and local context

Morpho-syntax morphosyntactic flexibility

Semantics semantic compositionality

Measuring lexical affinity

Co-occurrence frequency

Saliency [Kilgarriff and Tugwell, 2001] and/or **log-likelihood** [Dunning, 1993] scores

Support verb: boolean (yes/no)

- Verbs that can function as main but also as light verbs are *brengen* 'bring', *doen* 'do', *gaan* 'go', *geven* 'give', *hebben* 'have', *komen* 'come', *krijgen* 'get', *maken* 'make', *nemen* 'take' and *stellen* 'state' [Hollebrandse, 1993]. I added *houden* 'hold'.

Assessing local context

Head dependence two techniques proposed by Merlo and Leybold [2001]

- number of verbs that select a given PP: *hd_d_int* and,
- entropy of the distribution among the verbs that select for a given PP ((*P*, *N*) tuple)

$$H(P, N) = - \sum_j \frac{f((P, N), V_j)}{f(P, N)} \log \frac{f((P, N), V_j)}{f(P, N)} \quad (1)$$

Assessing local context (2)

Dependency relation and frequency

Dependency relation assigned by the Alpino parser to every dependent of a verb, e.g. direct object (obj1), prepositional complement (pc), separable particle or fixed phrase (svp), modifier (mod), etc.

Measuring morpho-syntactic flexibility in VERB (NP) PP

Preference for a specific syntactic context Is the PP typically found in the same preverbal position in the verbal cluster? Verb-final contexts (subordinate clauses).

Modifiability does an expression show many variants? are such variants very frequent or not? Measure the entropy observed among the realizations of the determiner, adjective and noun inflection slots in a PP argument.

Passivization how often a candidate expression has been observed in a passive construction.

Pronominalization NPs in non-compositional expressions fail to pronominalize. Is NP inside argument PP a pronoun? (yes/no)

Semantic compositionality

Measures make use of selectional preferences (Resnik 1993, 1996)

Semantic uniqueness which is approximately a ratio between the selectional preference of a *verb prep* for a *noun* and the selectional preference of a *verb prep* for the cluster in which that noun falls.

Selectional association of a NOUN for a VERB PREP.

Data for classification

houd_aan_afspraak,	251,0.88,	19,0.825,	1428.3465,?
heb_in_handen,	207,1.00,	7,1.215,	576.6651,?
kom_tot_resultaat,	59,0.98,	5,0.942,	113.5017,?
sta_ter_discussie,	85,0.98,	2,0.445,	276.8422,?
ga_op_dag,	67,0.48,	328,4.610,	9.0425,?
sta_in_krant,	67,0.71,	124,3.676,	159.0070,?
krijg_de_tijd,	368,0.97,	3,1.059,	1237.7703,?
krijg_bij_er,	472,0.14,	82,2.633,	520.3565,?
heb_van_doen,	390,1.00,	1,0.000,	2993.8551,?
kom_naar_land,	63,0.97,	76,3.034,	5.3596,?
laat_met_rust,	72,1.00,	9,0.920,	788.4035,?

Data for classification (2)

Syntactic pattern	Types	Tokens
V PP	4,969	
V NP PP	3,519	
total	8,488	1,140,800
V NP	10,211	2,053,286

Reference data

Vlis database More than 61,150 idioms and collocations collected by Van Dale lexicographers.

RBN *Referentie Bestand Nederlands* (reference lexical database). Includes singleton lexical entries but also MWES. Currently, a list of 3,805 MWES used.

Data annotation

All candidate expressions were annotated *automatically* using the two lexical resources VLIS and RBN.

Syntactic pattern	Freq	Types	MWES	non-MWES
V (NP) PP	≥ 10	8,488	1,910 (22.5%)	6,578 (77.49%)
V NP	≥ 10	10,211	2,771 (27.13%)	7,440 (72.86%)
	≥ 50	1,769	917 (51.83%)	852 (48.16%)

Evaluation methodology

Correct MWES: classified as MWE and present in reference data.

Correct non-MWES: classified as non-MWES and missing in reference data.

True positives (TPs): number of *correct* MWES.

True negatives (TNs): number of *correct* non-MWES.

$$\text{Accuracy} = \frac{|TPs + TNs|}{|\text{candidates}|}$$

$$\text{Precision}_{mwe} = \frac{|TPs|}{|TPs + FPs|}$$

$$\text{Recall}_{mwe} = \frac{|TPs|}{|\text{MWES in data}|}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Precision, Recall and F-measure given per class.

Baseline: accuracy of a naive classifier that chooses the most frequent class.

Experiments

Individual classifiers (C1,C2,C3) test effect of features 'assessing' one linguistic property (at a time).

Superior classifiers (C0_1,C0_2) test effect of combining features that simultaneously assess several linguistic properties.

Classifier	Linguistic properties	
	V (NP) PP	V NP
C1	lexical affinity	lexical affinity
C2	local context	local context
C3	syntactic flexibility	syntactic flexibility
C0_1	all properties	all properties
C0_2	all properties + semantic compositionality	

Results VERB (NP) PP

Classif	Dataset	Accuracy	class 'y'			class 'n'		
			P	R	F	P	R	F
Baseline	All	77.49	0	0	0	1.0	1.0	1.0
C1	Training (10fcv)	81.46	0.62	0.43	0.51	0.85	0.92	0.88
	Test	81.00	0.62	0.43	0.5	0.84	0.92	0.88
	All (10fcv)	81.22	0.64	0.36	0.46	0.83	0.94	0.88
C2	Training (10fcv)	80.55	0.6	0.37	0.46	0.83	0.93	0.88
	Test	80.93	0.61	0.44	0.51	0.84	0.91	0.88
	All (10fcv)	81.52	0.62	0.43	0.51	0.85	0.92	0.88
C3	Training (10fcv)	80.02	0.62	0.27	0.37	0.81	0.95	0.88
	Test	81.29	0.7	0.31	0.43	0.82	0.96	0.88
	All (10fcv)	80.53	0.66	0.26	0.38	0.81	0.96	0.88
C0_1	Training (10fcv)	82.52	0.64	0.5	0.57	0.86	0.91	0.89
	Test	82.07	0.62	0.47	0.54	0.86	0.91	0.89
	All (10fcv)	82.99	0.67	0.483	0.561	0.86	0.93	0.89
C0_2	Training (10fcv)	82.71	0.65	0.5	0.56	0.86	0.92	0.89
	Test	82.75	0.64	0.49	0.56	0.86	0.92	0.89
	All (10fcv)	83.4	0.66	0.53	0.59	0.87	0.92	0.89

Results VERB NP (freq $\geq 10, 50$)

Classif	Dataset	Accuracy	class 'y'			class 'n'		
			P	R	F	P	R	F
C1	All (10fcv)	76.81	0.66	0.29	0.4	0.78	0.94	0.85
C2	All (10fcv)	73.52	0.54	0.13	0.22	0.74	0.95	0.84
C3	All (10fcv)	74.05	0.6	0.13	0.21	0.74	0.96	0.84
C0	All (10fcv)	77.06	0.61	0.41	0.49	0.8	0.9	0.85
Baseline	All	72.86	0	0	0	1	1	1

Classif	Dataset	Accuracy	class 'y'			class 'n'		
			P	R	F	P	R	F
C1	All (10fcv)	63.03	0.66	0.58	0.62	0.6	0.68	0.63
C2	All (10fcv)	65.34	0.62	0.8	0.7	0.7	0.48	0.57
C3	All (10fcv)	64.33	0.65	0.67	0.66	0.63	0.61	0.62
C0	All (10fcv)	65.29	0.65	0.68	0.67	0.64	0.61	0.63
Baseline	All	51.83	1	1	1	0	0	0

Which linguistic properties are most useful?

Applying feature selection, feature importance measured with InfoGain.

$$\begin{aligned} \text{dep_rel}(lo) &\prec \text{rel_freq_pn_vfi}(sy) \prec \text{salience}(la) \prec \text{hd_ent}(lo) \\ &\prec \text{freq}(la) \prec \text{hd_int}(lo) \prec \text{modif}(sy) \prec \text{s2}(s) \prec \text{s1}(s) \\ &\prec \text{dep_rel_freq}(lo) \prec \text{passive}(sy) \prec \text{supp}(la) \prec \text{pron}(sy) \end{aligned} \quad (2)$$

- Head dependence (entropy), salience, modifiability and semantic scores have a small effect on precision but a positive effect on recall.
- Frequency has a positive effect on precision but a negative effect on recall.

Correctly classified MWES

- MWES from all frequency ranges observed.
- Low head dependence scores.
- Argument PP high relative frequency of immediately preceding verb in verb-final context (≥ 0.81).
- Modifiability varies a lot, more idiomatic expressions no variation (*iets uit het oog verliezen* 'lose sight of'), almost compositional expressions showing substantial variation (*tot conclusie komen* 'reach a conclusion')
- Semantic scores high values.

Misclassified MWEs

False negatives

- 36% annotation errors
- 64% classifier errors: (i) compositional MWEs (*naar de stembus gaan* go to vote), metaphors and idioms (*in iem. vaarwater zitten* 'work against s.o.');
- (ii) triple is ambiguous (HEB IN HAND representing one literal use and 2 idiomatic)

False positives

- 23% annotation errors but true positives
- 14% annotation errors matching a MWE with another syntactic pattern
- 60% classifier errors

False positives: not all are errors

- Metaphorical expressions: *op losse schroeven zetten* ''
- Transparent (institutionalized?) expressions: *iets in het nieuws komen* 'come up in the news'
- Predicative PPs with a verb: *onder druk komen* 'come under pressure', *in zwang zijn* 'be on fashion'
- Grammatical collocations with open NP slot: *van je houden* 'love you', *tegen mij zeggen* 'tell you'
- Directional and locative PPs with a verb: *naar school brengen* 'bring to school', *iets op de agenda zetten* 'place sth in the agenda'

Summarizing our findings

- High proportion of annotation errors likely to have negative effect on performance.

Captured:

- More idiomatic expressions, MWES allowing no/little modifiability.
- Active metaphors and institutionalized phrases.

Not captured:

- Predicative, locative or directional PPs. Similar behavior that actual PP arguments in V (NP) PP MWES. Hard to classify.
- MWES that allow modifiability (DET alternation, poss.).
- Grammatical collocations, constructions (*het gaat om 'it concerns'*)

Conclusions

MWE envisaged as a binary classification task. Best performing classifier reaches 83.4% accuracy. Performance is expected to be higher once annotation errors are corrected.

Among false positives, there are metaphors, institutionalized phrases and some idioms. Human judges need to check this.

Classifier identifies MWEs from all frequency ranges.

22.5% of V (NP) PP and 27% of V NP patterns are MWEs. Surely, MWEs are non-negligible phenomena.

Precision (0.66), recall (0.53) and f-measure (0.59) fair well if compared to identification of MWEs in other languages. Baldwin [2005] reports f-score 0.41 in prepositional verbs; Tan et al. [2006] f-score 0.576 in light verb constructions.

Future work

- Larger extraction corpus: Twente Nieuws Corpus, ca. 600M word.
- Correct annotation errors.
- Include translational entropy and overlap between actual translation and word-for-word translation [[Villada Moiron and Tiedemann, 2006](#)] as semantic features, too.
- View identification task as a ranking task, not as classification task.
- Apply MaxEntropy model on same data and features.

The End

Thanks a lot for your attention !

References

- T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. An Empirical Model of Multiword Expressions Decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, 2003.
- Tim Baldwin. Compositionality and Multiword Expressions: Six of One, Half a Dozen of the Other? Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July 2006.
- Timothy Baldwin. Looking for prepositional verbs in corpus data. In *Proc. of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications*, Colchester, UK, 2005.
- N. Calzolari, C.J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli. Towards best practice for Multiword Expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, 2002.
- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61—74, 1993.
- A. Fazly and S. Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy, 2006.
- Afsaneh Fazly, Ryan North, and Suzanne Stevenson. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In T. Baldwin, A. Korhonen, and A. Villavicencio, editors, *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 38–47, Ann Arbor, Michigan, 2005.

C. Hashimoto, S. Sato, and T. Utsuro. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 353–360, Sydney, Australia, 2006.

Bart Hollebrandse. Dutch light verb constructions. Master's thesis, Tilburg University, the Netherlands, 1993.

G. Katz and E. Giesbrecht. Automatic identification of non-compositional multi-word expressions using Latend Semantic Analysis. In *Proc. of the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, 2006.

Adam Kilgarriff and David Tugwell. Word sketch: Extraction & display of significant collocations for lexicography. In *Proceedings of the 39th ACL & 10th EACL -workshop 'Collocation: Computational Extraction, Analysis and Explotation'*, pages 32–38, Toulouse, 2001.

Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft, available from <http://people.csail.mit.edu/koehn/publications/europarl/>, 2003.

Dekang Lin. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324. University of Maryland, 1999.

Diana McCarthy. Automatic methods for detecting compositionality. Invited talk given at the Collocations and Idioms 2006 Workshop, November 2006. Berlin, Germany.

Diana McCarthy, Bill Keller, and John Carroll. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, 2003.

I. Dan Melamed. Measuring semantic entropy. In *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46, Washington, 1997.

Paola Merlo and Matthias Leybold. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Proc of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 121–128, Toulouse. France, 2001.

Franz Josef Och. GIZA++: Training of statistical translation models. Available from <http://www.isi.edu/~och/GIZA++.html>, 2003.

Darren Pearce. Synonymy in collocation extraction. In *WordNet and Other lexical resources: applications, extensions & customizations (NAACL 2001)*, pages 41–46, Pittsburgh, 2001. Carnegie Mellon University.

Darren Pearce. A comparative evaluation of collocation extraction techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002.

S. Piao, P. Rayson, O. Mudraya, A. Wilson, and R. Garside. Measuring mwe compositionality using semantic annotation. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 2–11, Sydney, Australia, 2006. Association for Computational Linguistics.

Ivan Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword Expressions: a pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City, Mexico, 2002.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*", pages 49–56, Trento, Italy, 2006.

S. Venkatapathy and A. Joshi. Measuring the relative compositionality of verb-noun collocations by integrating features. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 899–906, Vancouver, 2005.

Begoña Villada Moiron and Jörg Tiedemann. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*", pages 33–40, Trento, Italy, 2006.

Joachim Wermter and Udo Hahn. Collocation extraction based on modifiability statistics. In *Proceedings of Coling 2004*, Geneva, Switzerland, 2004. COLING.

Data and Resources

- Europarl corpus
 - ★ tokenized and aligned at sentence level [Koehn, 2003]
 - ★ Dutch part ca. 29 million tokens, 1.2 million sentences.
- Automatic word alignment
 - ★ using GIZA++ [Och, 2003]
 - ★ alignments produced for both translation directions (source to target and target to source)
 - ★ combination of both directional alignments (refined alignment)
- LINK LEXICA: For each pair of aligned corpora, for each word in source language (NL), collect all its alignments in target language (EN,ES,DE). Frequency of observing (source,target) alignment.
- List of candidate MWES.

1. Extraction of candidate MWES

- VERB PP patterns extracted from Dutch Europarl section; fully parsed with Alpino.
- using log-likelihood [Dunning, 1993], salience [Kilgarriff and Tugwell, 2001] and head dependence [Merlo and Leybold, 2001, Baldwin, 2005]
- among 191,000 types, we select 200 potential MWES to test our method.
- list of 200 potential MWES classified into idiomatic and literal expressions (precision = 0.64, uap = 0.75)

2. Collecting translation alignments

- For each expression candidate, we collect all translation alignments of its component words in the context of the triple.
- *aan eisen voldoen* 'satisfy the requirements' and *iets aan de kaak stellen* 'denounce'

Source	Translation alignments in English				T_s
	instance 1	instance 2	instance 3	instance 4	
aan	with	met	to	with	T_{aan}
eisen	requirements	requirements	requirements	comply	T_{eisen}
voldoen	requirements	met	satisfy	requirements	$T_{voldoen}$
	instance 1	instance 2	instance 3	instance 4	
aan	NO LINKS	NO LINKS	to	NO LINKS	T_{van}
kaak	criticised	challenged	condemn	NO LINKS	T_{mening}
stellen	criticised	challenged	condemn	unacceptable	T_{zijn}

Approach

1. Extract candidate MWES from a source language.
2. Collect translation alignments in a target language.
3. **Score (initial) candidate MWES and rank in terms of idiomaticity.**
 - Translational entropy
 - Proportion of default alignments

Translational entropy

- Idiomatic expressions more difficult to align than literal expressions.
- Expect a larger variety of translation alignments for words in idiomatic expressions.
- Measure unpredictability of an event.

$$H(T_s|s) = - \sum_{t \in T_s} P(t|s) \log P(t|s) \quad (3)$$

- $H(S)$ is the average of $H(\text{preposition})$, $H(\text{noun})$ and $H(\text{verb})$

Proportion of default alignments

- **Link lexica** provide us with
 - ★ default alignments D_s : alignments of a word s in whole corpus (4 most frequent alignments)
- alignments of a word s in the context of a triple T_s

$$pda(S) = \frac{\# \text{ of alignments} = \text{default alignments}}{\# \text{ of alignments}} \quad (4)$$

- Large proportion of default alignments suggests literal meaning; low proportion suggests idiomatic meaning.

Computing scores. An example

Source	Translation alignments	Default alignments
	T_{word}	D_{word}
aan eisen voldoen	with (2), met (1), to (1) requirements (3), comply (1) requirements (2), met (1), satisfy (1)	to , on, in, for demand, requirements , call, demands meet, fulfil, met , satisfy
aan kaak stellen	NO LINKS (3), to (1) criticised (1), condemn (1), challenged (1), NO LINKS (1) criticised (1), challenged (1), condemn (1), unacceptable (1)	to , on, in, for the, condemn condemned, .. am, I, would, propose

$$H(\text{eisen}) = -\left(\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4}\right) = 0.562$$

$$pda(\text{eisen}) = \frac{3}{4} = 0.75$$

$$H(\text{kaak}) = -\left(4 * \frac{1}{4} \log \frac{1}{4}\right) = 1.386$$

$$pda(\text{kaak}) = \frac{1}{4} = 0.25$$

Results. Alignment types and scoring metrics.

Word alignment helps. Source to target best performance.

Alignment	uap
src2trg	0.864
trg2src	0.785
refined	0.765
baseline	0.755

Entropy or pda.

Score	NL-EN	NL-ES	NL-DE
entropy			
- without NO_LINKS	0.864	0.892	0.907
- NO_LINKS=many	0.858	0.890	0.883
- NO_LINKS=one	0.859	0.890	0.911
pda	0.891	0.894	0.894
baseline	0.755	0.755	0.755

Improvements. Lemmatization and no prepositions

Setting	NL-EN	NL-ES	NL-DE
using entropy scores			
with prepositions			
wordforms	0.864	0.892	0.907
lemmas	0.873	–	0.906
without prepositions			
wordforms	0.906	0.923	0.932
lemmas	0.910	–	0.931
using pda scores			
with prepositions			
wordforms	0.891	0.894	0.894
lemmas	0.888	–	0.903
without prepositions			
wordforms	0.897	0.917	0.905
lemmas	0.900	–	0.910
baseline	0.755	0.755	0.755

rank	pda	entropy	MWE	triple
1	9.80	8.3585	ok	breng tot stand 'create'
2	9.24	8.0923	ok	breng naar voren 'bring up'
3	16.40	7.8741	ok	kom in aanmerking 'qualify'
4	15.33	7.8426	ok	kom tot stand 'come about'
5	8.70	7.4973	ok	stel aan orde 'bring under discussion'
6	5.65	7.4661	ok	ga te werk 'act'
7	17.46	7.4057	ok	kom aan bod 'get a chance'
8	9.38	7.1762	ok	ga van start 'proceed'
9	14.15	7.1009	ok	stel aan kaak 'expose'
10	18.75	7.0321	ok	breng op gang 'get going'
17	10.25	6.4893	ok	neem onder loep 'scrutinize'
18	7.83	6.4666	ok	breng aan licht 'reveal'
19	5.99	6.4049	ok	roep in leven 'set up'
20	15.89	6.3729	ok	neem in aanmerking 'consider'
102	23.56	4.6865	ok	kom te weten 'find out'
103	15.38	4.6713	ok	neem in ontvangst 'receive'
104	31.57	4.6556	*	ga om waar 'go about where'
105	35.95	4.6380	*	houd met daar 'keep with there'
106	34.86	4.6215	*	ga om zaak 'go about issue'
107	28.33	4.5846	ok	kom tot overeenstemming 'come to terms'
180	70.53	2.7395	*	voldoe aan criterium 'satisfy criterion'
181	52.33	2.7351	*	beschik over informatie 'dispose of information'
182	74.71	2.6896	*	stem voor amendement 'vote for amending'
183	76.56	2.5883	*	neem_deel aan stemming 'participate in voting'
187	80.39	2.0992	*	stem tegen amendement 'vote against amending'
188	78.04	2.0924	*	onthoud van stemming 'withhold one's vote'
189	77.63	1.9997	*	feliciteer met werk 'congratulate with work'
190	82.21	1.9020	*	stem voor verslag 'vote for report'
191	77.78	1.9016	*	schep van werkgelegenheid 'set up of employment'
193	73.33	1.8687	*	bedank voor feit 'thank for fact'
198	85.56	1.1779	*	dank voor antwoord 'thank for reply'
199	90.55	1.0398	*	ontvang overeenkomstig artikel 'receive similar article'
200	87.88	1.0258	*	recht van vrouw 'right of woman'

Discussion

Top scores assigned to idiomatic or metaphorical expressions. Lower scores assigned to literal expressions.

- syntactic construction specific to source language: *(het) gaat om* 'the issue/question is'
 - ★ translations include multiple paraphrases
- particle verbs: *verzoek aangeven* 'comply with request'
- verb PP part of a MWE: *rekening houden met* 'consider'
- support verb constructions ranked lower than idiomatic expressions (*van mening zijn* 'believe')
- similar expressions in source and target language (*te ver gaan* 'go too far, be unreasonable')

Conclusions

- Word alignment in parallel corpora provides evidence of the type of meaning of expressions.
- Translational entropy measures predictability of the translation of an expression. In Dutch to German gives 75.5% to 93.2% improvement.
- Proportion of default alignments measures consistency between translation of individual words in the context of the expression and when in isolation. In Dutch to Spanish, pda gives 91.7%.
- Better results obtained for Dutch to German and Dutch to Spanish.
- Ranking mirrors scale from non-compositional to compositional. Metaphorical expressions and support verb constructions located in the middle.

Word alignment types

src2trg		trg2src	
source	target	target	source
gesteld	appreciate	NO_LINK	stellen
prijs	appreciate	much appreciate indeed	prijs
op	appreciate	NO_LINK	op
gesteld	be	keenly appreciate	stellen
prijs	delighted	fact	prijs
op	NO_LINK	NO_LINK	op