

# Identifying idiomatic expressions using automatic word-alignment

Begoña Villada Moirón and Jörg Tiedemann  
University of Groningen, The Netherlands

'MWES in a multilingual context' EACL Workshop

Trento, April 3, 2006

## Defining the problem

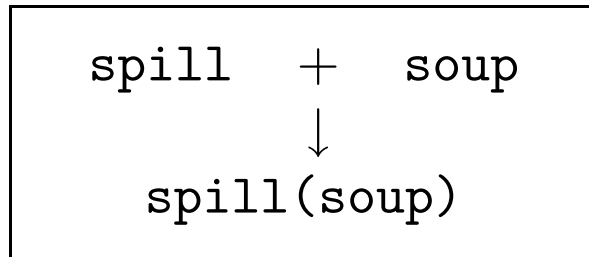
- We investigate automated methods to rank a list of candidate expressions in terms of their idiomaticity.
- We explore whether automatic word alignment can be useful to identify idiomatic expressions.
- Compile a lexicon of idiomatic multiword expressions to be used in NLP systems.

## Idiomatic expressions . . .

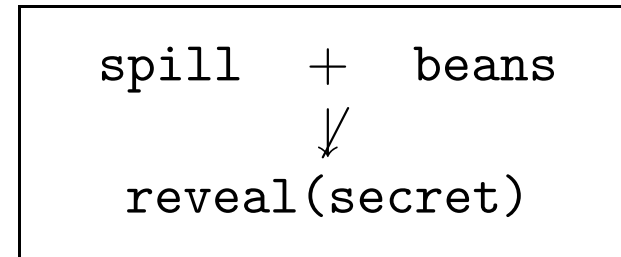
- constitute a subset of multiword expressions [Sag et al., 2001].
- show idiosyncratic behavior at various levels of analysis
  - ★ rigidity in syntax and morphology
  - ★ strong lexical affinity
  - ★ meaning is conventionalized
- have a non-compositional meaning.

## How can we capture non-compositional meaning?

### compositional

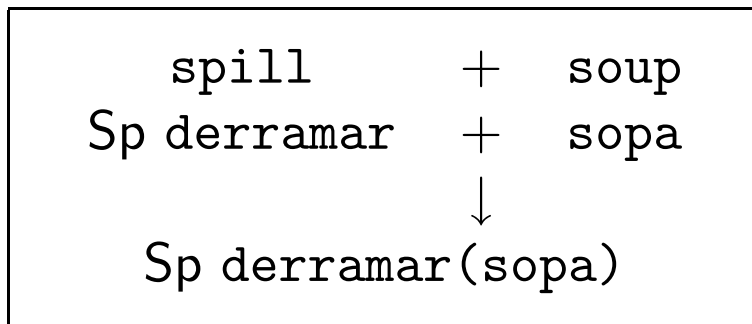


### non-compositional

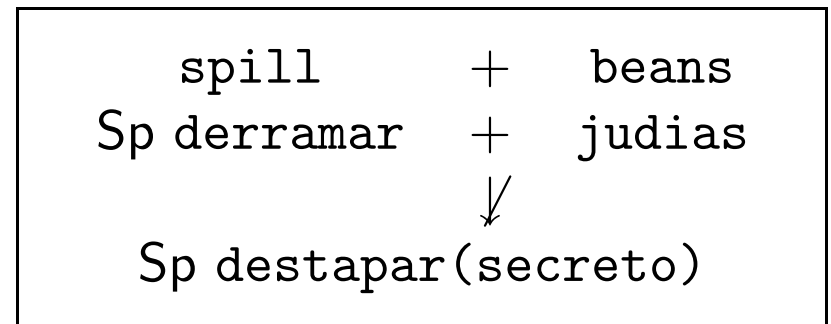


- Approximate meaning by looking up translation in a foreign language.

### literal



### idiomatic



## Our approach

- To what extent can parallel corpora help us to find out the type of meaning an expression has?

### Assumptions:

- words in their literal use are translated rather consistently
- **literal meaning** if translation results from combining each word's translation in isolation; **non-compositional meaning** otherwise.
- automatic word-aligner has more difficulties with non-compositional expressions

## Overview of our approach

1. Extract candidate MWEs from a source language in a parallel corpus.
  - Dutch
2. For each candidate, collect translation alignments in a target language.
  - English
  - Spanish
  - German
3. Score initial candidates and rank them in terms of idiomaticity.

## 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.

$$H(T_s|s) = - \sum_{t \in T_s} P(t|s) \log P(t|s) \quad (1)$$

- $H(S)$  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  $T_s$

$$pda(S) = \frac{\# \text{ of alignments} = \text{default alignments}}{\# \text{ of alignments}} \quad (2)$$

- 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), <b>to</b> (1) <b>requirements</b> (3), comply (1) requirements (2), <b>met</b> (1), <b>satisfy</b> (1)	<b>to</b> , on, in, for demand, <b>requirements</b> , call, demands meet, fulfil, <b>met</b> , <b>satisfy</b>
aan kaak stellen	NO LINKS (3), <b>to</b> (1) criticised (1), <b>condemn</b> (1), challenged (1), NO LINKS (1) criticised (1), challenged (1), <b>condemn</b> (1), unacceptable (1)	<b>to</b> , 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	<b>0.864</b>
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	<b>0.911</b>
pda	<b>0.891</b>	<b>0.894</b>	0.894
baseline	0.755	0.755	0.755

## Improvements. Lemmatization and no prepositions

Setting	NL-EN	NL-ES	NL-DE
using entropy scores			
<b>with prepositions</b>			
wordforms	0.864	0.892	0.907
lemmas	0.873	–	0.906
<b>without prepositions</b>			
wordforms	0.906	0.923	<b>0.932</b>
lemmas	0.910	–	0.931
using pda scores			
<b>with prepositions</b>			
wordforms	0.891	0.894	0.894
lemmas	0.888	–	0.903
<b>without prepositions</b>			
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'
	73.33	1.8687	*	bedank voor feit 'thank for fact'
---	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.

**Thanks for your attention!**

# References

- 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.
- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61—74, 1993.
- 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 Exploitation'*, 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.
- 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.
- Ivan Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03, 2001.

## 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