

Quantifying and qualifying lexicalized and idiomatic expressions

Begoña Villada Moirón
Alfa Informatica
University of Groningen
The Netherlands
M.B.Villada.Moiron@rug.nl

Nicole Grégoire
Uil-OTS
University of Utrecht
The Netherlands
Nicole.Gregoire@let.uu.nl

Collocations and Idioms 1. Joensuu (Finland), May 20th 2006

Towards a classification of fixed expressions

- Compile a dictionary of Dutch fixed expressions [Moon, 1998] that require lexical mention.
- What do we want to include exactly?
 - ★ “Combinations of two or more words whose linguistic properties cannot be inferred from the linguistic properties of its component words”.
- Phrasal lexical items [Everaert and Kuiper, 1996] or multiword expressions [Sag et al., 2001].

Towards a classification of fixed expressions (continued)

How?

- Apply corpus-based method that captures the linguistic properties of fixed expressions.
- Approximate such linguistic properties with quantitative techniques.
- Classify expressions using quantitative measurements.

Focus of this research:

- Do quantitative measurements of the linguistic properties mirror a theoretical classification of fixed expressions?

Overview

- A selection of descriptive properties.
- Quantifying linguistic properties. Two example measurements.
- Classification based on quantitative measurements vs. theoretical classification.
- Discussion and conclusion.

A selection of descriptive properties

- **Lexical level**
 - ★ strong lexical affinity between component words
 - ★ strong dependence between selecting verb and selected argument
- **Morphology**
 - ★ non-productivity (number morpheme, diminutive)
- **Syntax**
 - ★ restrictions on determiners, quantifiers and (adjectival) modification
 - ★ strong preference for a syntactic position (specific to Dutch)
 - ★ Ignored: passivization, topicalization, extraction
- **Semantics**
 - ★ partially or non-compositional meaning
 - ★ consistency in translation into another language

Lexical affinity between component words

- Combinations that form a FIXED EXPRESSION show stronger lexical affinity between component words than regular word combinations.
- The more fixed the combination the stronger the lexical affinity.

(from Twente Nieuws Corpus, newspaper text)

```
**fixed wilden minister Kok kennelijk een loer DRAAIEN.  
        Alleen de aluminiumpoot DRAAIT nog met verlies.  
        Thyssen Draht DRAAIT echter minder goed.  
We werken met verschillende disc-jockeys, die steeds weer andere  
        platen DRAAIEN.  
        Alles DRAAIT in Japan om de poen.  
**fixed die etter hem zo'n vernederende loer DRAAIDE.  
        's Avonds DRAAIT 'Natural born  
killers' in kino Bosna.
```

Dependence between selecting verb and selected XP

A PP selected by a large number of verbs is likely to be an optional adjunct. A PP selected by one or a few verbs is probably a required argument in fixed expression.

PP	# selecting verbs	verbs	fixed expression
aan pols	3	houden,komen,geven	yes
bij paaltje	2	komen,terugschrikken	yes
in cassatie	2	gaan,maken	yes
aan beurt	3	komen,zijn,inspringen	yes
in jaar	823	zijn,brengen,doen,gaan,nemen	no
aan kant	249	zijn,houden,staan	no
in brief	145	schrijven,stellen	no
van ministerie	43	krijgen,zijn,vragen,weten	no

English: *take into account* vs. *be in the year*.

Preference for specific syntactic contexts

PP complements closer to the verb in **verb final context** than adverbial PPs [Broekhuis, 2004].

hij laat zich te gauw [van de wijs] BRENGEN door hobbels en bobbels in
Maar als puntje [bij paaltje] KOMT, is de Nederlander toch
een koopman en zal hij Duits praten.

Elias zou (in een brief) aan de hoofdredactie hebben GESCHREVEN
Over Toni Morrison werd (in deze krant) al uitgebreid GESCHREVEN.
Minister Ter Beek zegt vandaag in een interview
(in deze krant) blij te ZIJN dat...

Arguments in fixed expressions **tend to immediately precede** the verb group in v-final contexts (Jack Hoeksema (p.c.)).

Morphology and syntax . . . productive?

- (1) *iets in zijn bezit hebben/krijgen*
sth in his property have/get
'to get to own; to own'

	DET	N-MORPH	ADJ/MOD
in bezit hebben	het	sg	van-PP
in bezit hebben		sg	
in bezit hebben	hun	sg	
in bezit hebben	mijn	sg	
in bezit hebben	haar	sg	
in bezit krijgen		sg	
in bezit hebben	zijn	sg	voor-PP

Morphology and syntax . . . productive?

- (2) *iemand aan het lijntje houden*
s.o. on the little-line hold
'keep s.o. dangling'

DET N-MORPH ADJ/MOD

aan lijntje houden	het	sg	-
aan lijntje houden	een	sg	opportunistisch
aan lijntje houden	het	sg	-
aan lijntje houden	het	sg	-
aan lijntje houden	het	sg	-

Transparent or opaque meaning?

Assume:

- We can approximate the underlying meaning of an expression by its translation into a foreign language.
- Parallel corpora provide us with translations of fixed expressions.

Villada Moiron and Tiedemann [2006] showed that the meaning of an expression is

- **transparent** if it receives a word-by-word literal translation. Consistent translation throughout corpus.
- **semi-transparent** if one part receives a literal translation and the rest of the expression doesn't.
- **opaque** if it receives different translations. Unpredictable translation.

Transparent or opaque meaning?

Dutch	Literal translation	Found translations
-----	-----	-----
feliciteer met werk (transparent)	congratulate with work	'congratulate on work' 'congratulate for work' 'congratulate'
-----	-----	-----
tot akkoord komen (semi-transparent)	to agreement come	'reach an agreement' 'come to agreement' 'secure an agreement'
-----	-----	-----
naar voren brengen (opaque)	to forward bring	'present' 'indicate' 'highlight' 'raise'

Quantifying linguistic properties

- Lexical affinity measured as the log-likelihood score.

word combination	translation	log-likelihood
(van,wijs,brengen)	mislead s.o.	806.5314
(aan,pols,houden)	keep sth. under control	564.5474
(bij,paaltje,komen)	conclude	228.0177
(aan,kant,staan)	stay on the side	153.4448
(in,jaar,zijn)	be in the year	4.5073

Quantifying linguistic properties

- Preference for specific syntactic context measured as the relative frequency of seeing the PP immediately preceding the verb group.

word combination	translation	score
(van,wijs,brengen)	mislead s.o.	1.00
(bij,paaltje,komen)	conclude	1.00
(aan,pols,houden)	keep under control	0.92
(aan,kant,staan)	stay on the side	0.73
(in,jaar,zijn)	be in the year	0.16

What's next?

- We selected a bunch of descriptive properties of fixed expressions.
- Using a corpus-based approach, we measured to which extent a descriptive property is observed in a list of expressions.
- We want to establish whether these quantitative measurements split apart subtypes of lexicalized phenomena.

Empirical vs. theoretical classification

Scores	lexical affinity	head dependence	preference PP_pos	productivity	translation consistency
low	light verb constructions	idiomatic		idiomatic	idiomatic
medium		lexicalized		lexicalized	
		light verb constructions	light verb constructions	light verb constructions	light verb constructions
high	idiomatic		idiomatic		lexicalized

Lexical affinity measure

Expression	freq	lexical affinity	Class
van mening zijn	1317	4896	lexicalized
op agenda staan	438	2662	LVC/lex.
iets naar voren brengen	292	2089	idiom.
tot doel hebben	281	1823	lexicalized
tot akkoord komen	203	1366	LVC/lex.
iets in bezit hebben	363	1137	LVC
van wijs brengen	124	1084	idiom.
ter sprake brengen	158	893	LVC/lex.
uit zijn dak gaan	71	482	idiom.
iem. aan lijntje houden	57	473	idiom.
puntje bij paaltje komen	58	372	idiom.
aan het werk zijn	450	340	LVC

Head dependence measure

Expression	freq	head dependence	Class
aan het werk zijn	450	2.999	LVC
iets in bezit hebben	363	1.804	LVC
op agenda staan	438	1.655	LVC/lex.
tot doel hebben	281	1.048	lexicalized
van mening zijn	1317	1.043	lexicalized
uit zijn dak gaan	71	0.994	idiom.
iets naar voren brengen	292	0.680	idiom.
tot akkoord komen	203	0.555	LVC/lex.
ter sprake brengen	158	0.508	LVC/lex.
van wijs brengen	124	0.386	idiom.
iem. aan lijntje houden	57	0.257	idiom.
puntje bij paaltje komen	58	0.086	idiom.

Other measures

Expression	PP_pos	productivity	translation	Class
sta op agenda	0.94	2.845	1.4117	LVC
iets in bezit hebben	1.00	2.255		LVC
aan het werk zijn	1.00	1.834		LVC
tot akkoord komen	0.91	2.750	1.5404	LVC/lex.
ter sprake brengen	1.00	0.00	3.0795	LVC/lex.
tot doel hebben	1.00	1.762	2.4526	lexicalized
van mening zijn	1.00	1.310	2.2465	lexicalized
uit zijn dak gaan	1.00	2.080		idiom.
van wijs brengen	1.00	0.047		idiom.
iets naar voren brengen	1.00	0.000	3.8576	idiom.
puntje bij paaltje komen	1.00	0.581		idiom.
iem. aan lijntje houden	1.00	0.441		idiom.

Some conclusions

- Some empirical properties distinguish LVCs from idiomatic expressions
 - ★ head dependence
 - ★ consistency across translations
 - ★ morphological productivity and syntax flexibility
- Empirical properties that are non-discriminating
 - ★ lexical affinity measured with log-likelihood statistic
 - ★ preference for a specific syntactic context (in subordinate clauses)
- Quantitative measurements do not mirror categorical distinctions between the 'subtypes' of phenomena. Measurements rank expressions in a continuum from productive to fossilized combinations.

Corpus-based approach

- Corpus:
 - ★ newspaper text (Dutch)
 - ★ 78 million words; 4 million sentences
 - ★ parsed with Alpino, a Dutch syntactic analyzer. Available at <http://www.let.rug.nl/~vannoord/alp/Alpino>.
 - ★ each sentence has a full syntactic analysis (no error correction)
- Text processing tools (Perl scripts) used to
 - ★ find all VERB PREPOSITIONAL PHRASE occurrences
 - ★ PP is a syntactic dependent of the VERB
 - ★ collect necessary frequency counts
- Statistics used: absolute frequency, log-likelihood **Dunning [1993]**, entropy and relative frequency.

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.
- Hans Broekhuis. Het voorzetselvoorwerp. *Nederlandse Taalkunde*, 9(2):97—131, 2004.
- Miriam Butt. *The structure of complex predicates in Urdu*. PhD thesis, Stanford University, 1995.
- Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61—74, 1993.
- Gregor Erbach. Head-driven lexical representation of idioms in HPSG. In Martin Everaert, Erik-Jan van der Linden, Andre Schenk, and Rob Schreuder, editors, *Proceedings of Idioms. International conference on Idioms.*, volume 1, pages 11–24. Tilburg. The Netherlands, 1992.
- Martin Everaert and Koenraad Kuiper. Theory and data in idiom research. In L. McNair, K. Singer, L. M. Dobrin, and M. Aucoin, editors, *Papers from the Parasession on Theory and Data in Linguistics*, volume CLS 32, pages 43–58. Chicago Linguistics Society, 1996.
- Paola Merlo and Matthias Leybold. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Procs of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 121–128, Toulouse. France, 2001.

Rosamund Moon. *Fixed expressions and Idioms in English. A corpus-based approach*. Clarendon Press, Oxford, 1998.

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.

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.