Detecting a Continuum of Compositionality in Phrasal Verbs

Diana McCarthy & Bill Keller & John Carroll

University of Sussex

This research was supported by:
the RASP project (EPSRC)
and
the MEANING project (EU 5th framework)

Overview

- Phrasal verbs
- Motivation for detecting compositionality
- Related research
- Using an automatically acquired thesaurus
- Evaluation
- Results
- Comparison
  - With some statistics used for multiword extraction
  - With entries in man-made resources
- Conclusions, problems and future directions

Phrasals: syntax and semantics

- Syntax, e.g. particle movement, adverbial placement
- Some productive combinations better handled in grammar (Villavicencio and Copestake, 2002)
- Want different treatment depending on compositionality e.g. *fly up, eat up, step down, blow up, cock up*
- use neighbours from thesaurus to indicate degree of compositionality
- Compare to compositionality judgements – on an ordinal scale
- Cut-off points to be determined by application

Motivation

- Selectional Preference Acquisition (*eat* & *eat up* vs *blow* & *blow up)*
- Word Sense Disambiguation –importance of identification depends on degree of compositionality, and granularity of sense distinctions
- Multiword Acquisition – relate phrasal sense to senses of simplex verb, how related are they?

RASP Parser Output

- Phrasal Verb e.g point out the hotel

(|ncmod| _ |point:16_VV0| |out:17_RP|)
(|dobj| _ |point:16_VV0| |hotel:19_NN1|)

- vs Prepositional verb e.g. refer to the map

(|iobj| |to:12_II| |refer:11_VV0||map:14_NN1|)

Parser Evaluation

- For verb and particle constructions identified as such in the WSJ
- Use phrasal lists (such as in ANLT) to improve parser performance

	Precision	Recall
RASP	87.6	49.4
RASP with ANLT list	92.6	64.2
MINIPAR	78.9	44.1

Related Research

- Extraction
  - Blaheta and Johnson (2001) Phrasality and `good collocation' correlated with opaqueness
  - Baldwin and Villavicencio (2002)

- Compositionality
  - Lin (1999) thesaurus filtered with Log-likelihood ratio, used to obtain substitutes, test significance of difference in mutual information of substitute MW to original.
  - Schone and Jurafsky (2001) LSA for multiword induction
  - Bannard et al. thesaurus and LSA, evaluation for verb and particle contribution
  - Baldwin et al. LSA compared with WordNet based scores

Acquiring the Thesaurus

- Thesaurus acquired from RASP parses of the written portion of the BNC data
- Phrasal verbs (*blow up*) and their simplex counterpart (*blow*) listed with all subjects and direct objects
- Thesaurus obtained following Lin (1998)
- Output: top 500 nearest neighbours listed (with similarity score)

Using the Thesaurus:

  - *climb+down*: *clamber+up* .248 *slither+down* .206 *creep+down* .183 …
  - *climb*: *walk* 0.152 *jump* .148 *go+up* .147…
- Position and similarity score of simplex verb within phrasal neighbours
- Overlap of neighbours of simplex with neighbours of phrasal
- How often the same particle occurs in neighbours
- Evaluation – no cut off, see correlation between measures and ranks from human judges

Evaluation

- 100 phrasal verbs selected randomly from 3 partitions of the frequency spectrum, + 16 verbs selected manually
- 3 judges : native English speakers
- List of 116 verbs, score between 0 and 10 (fully compositional)
- Removed any verbs with *don't know* category (5 such verbs)
- Scores treated as ranks, look at correlation of ranks
- Average ranks used as a gold-standard

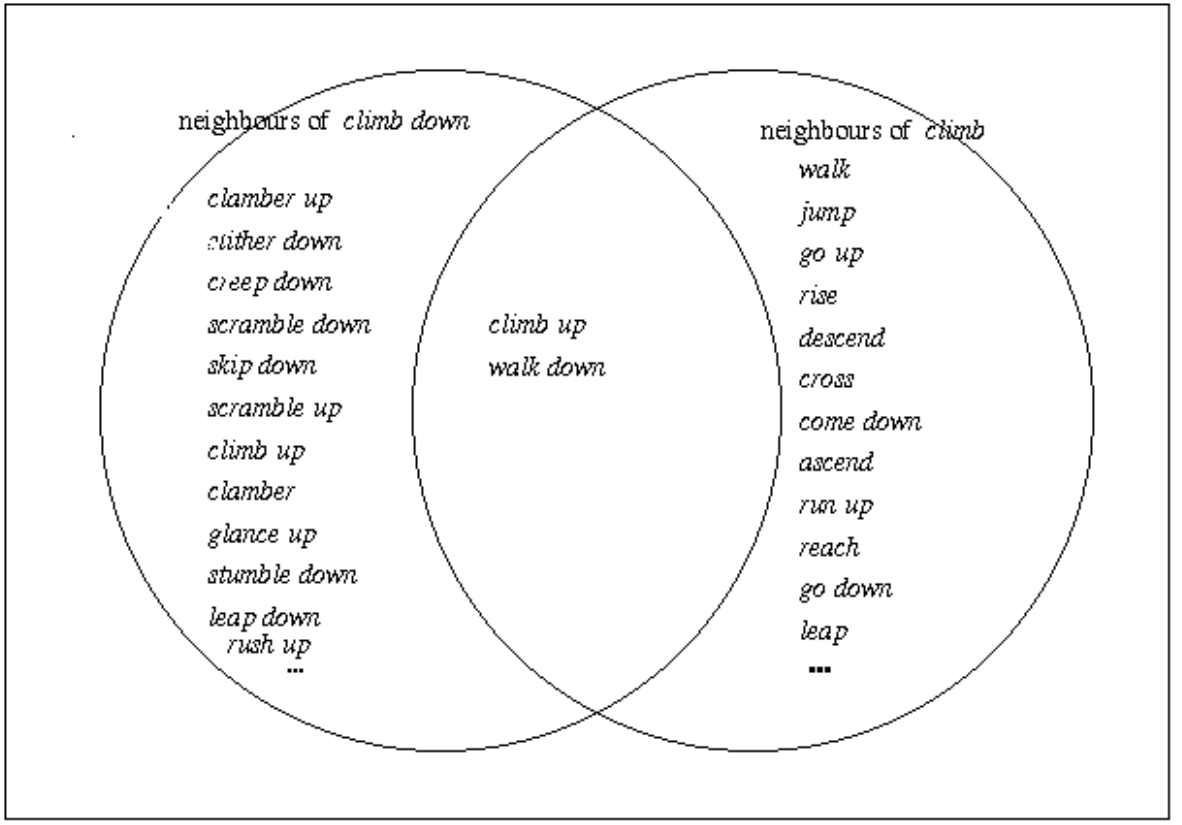
Inter-Rater Agreement

- Kendall Coefficient of Concordance (Siegel and Castellan, 1988)
- useful for 3 or more judges giving ordinal judgements
- linear relationship to the average Spearman Rank-order Correlation Coefficient taken over all possible pairs of rankings
- highly significant $W = 0.594$, $\chi^2 = 196.30$
- probability of this value by chance <= 0.000001

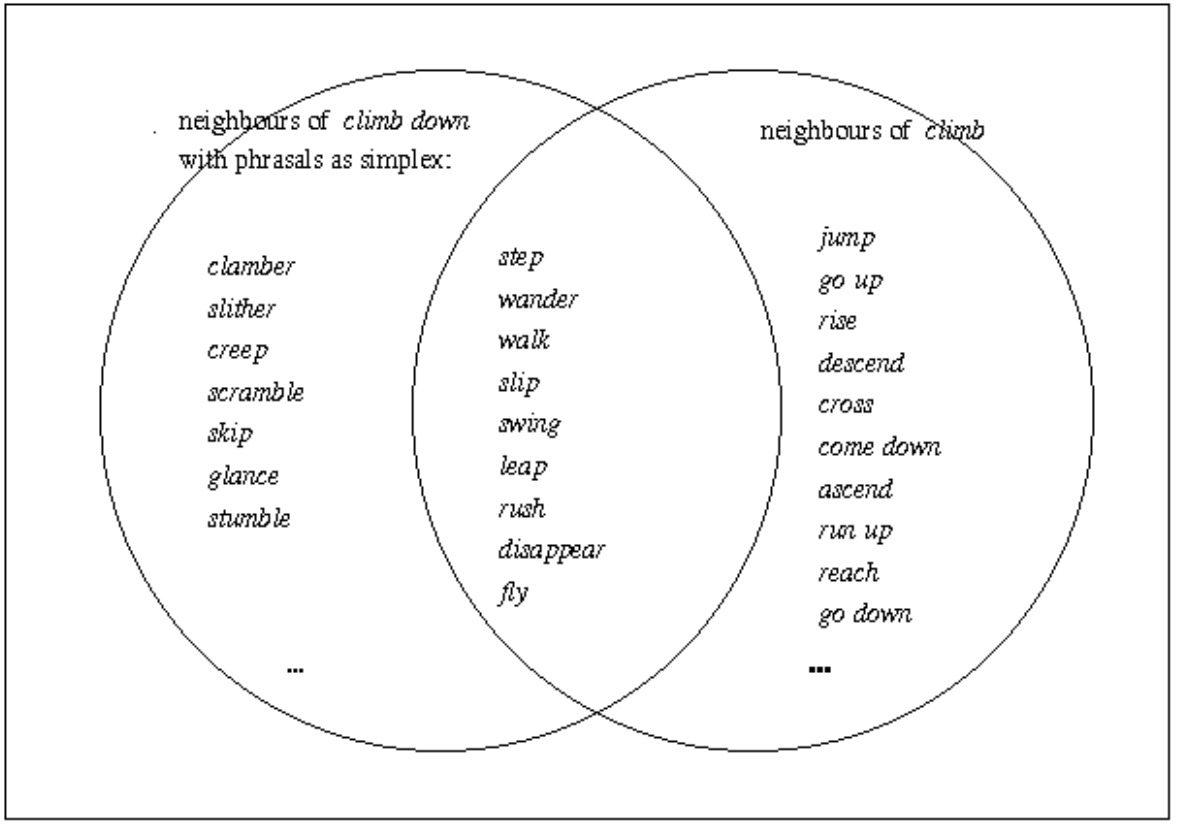
Measures

- **simplexasneighbour** *X* = 500
- **rankofsimplex** X=500
- **scoreofsimplex** The similarity score of the simplex in top *X* = 500 neighbours
- **overlap** of first *X* neighbours, where *X*= 30, 50, 100, and 500
- **overlapS** of first *X* neighbours, where *X*= 30, 50, 100, and 500, with simplex form of neighbours in phrasal neighbours
- **sameparticle** number of neighbours with same particle as phrasal *X*=500
- **sameparticle-simplex** as above minus the number of simplex neighbours with the same particle *X* = 500

Overlap


neighbours of *climb down*
clamber up
clither down
creep down
scramble down
skip down
scramble up
climb up
clamber
glance up
stumble down
leap down
rush up
...
climb up
walk down
neighbours of *climb*
walk
jump
go up
rise
descend
cross
come down
ascend
run up
reach
go down
leap
...

OverlapS


neighbours of *climb down* with phrasals as simplex:
clamber
slither
creep
scramble
skip
glance
stumble
...
step
wander
walk
slip
swing
leap
rush
disappear
fly
neighbours of *climb*
jump
go up
rise
descend
cross
come down
ascend
run up
reach
go down
...

For Comparison

- Statistics
  - Log-likelihood ratio test (Dunning, 1993)
  - Mutual Information (point-wise) Church and Hanks (1990)
  - χ^2 (chi-squared)
- Man-Made resources
  - WordNet
  - ANLT lists (phrasal and prepositional verbs)

Results

Overlap	r_s	Z score	p under H_0
X = 30	0.166	1.74	0.04
X = 50	0.136	1.43	0.08
X = 100	0.037	0.39	0.35
X = 500	-0.032	-0.38	0.35
OverlapS			
X = 30	0.306	3.21	<0.0007
X = 50	0.303	3.18	<0.0007
X = 100	0.263	2.76	0.0030
X = 500	0.167	1.75	0.040

Results continued…

X=500	statistic	Z score	p under H_0
sameparticle	$r_{s=}0.414$	4.34	< 0.00003
sameparticle-simplex	$r_{s=}0.49$	5.17	<0.00003
simplexasneighbour	MW	0.950	0.171
simplexrank	$r_{s=}-0.115$	-1.21	0.113
simplexscore	$r_{s=}0.052$	0.54	0.295

Correlations of GS with man-made resources and statistics

	statistic	Z score	P under H_0
LLR	r_s= -0.168	-1.76	0.0392
χ^2	r_s =-0.213	-2.22	0.0139
MI	r_s =-0.248	-2.60	0.0047
Phrasal freq	r_s =-0.096	-1.01	0.156
Simplex freq	r_s =0.092	0.96	0.169
WordNet	MW	2.39	0.0084
ANLT phrasals	MW	3.03	0.0012
ANLT prep^{n}s	MW	0.430	0.334

Correlation of measures with man-made resources

	In WordNet	In ANLT phrasals
MI	-2.61	-4.53
sameparticle-simplex	3.71	4.59

Conclusions, Problems and Future Directions

- Thesaurus measures worked better than statistics, especially looking for neighbours having the same particle
- Straight overlap of neighbours – not as good as hoped,
- Overlap taking particles into account helps.
- May help to use similarity scores or ranks of neighbours.
- Polysemy is a problem for both methods and evaluation.
- Continuum of compositionality useful for exploring relationship – still need cut-offs for application