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
  $\langle|\text{ncmod}|_\_|\text{point}:16\_\text{VV0}|\text{out}:17\_\text{RP}|\rangle$
  $\langle|\text{dobj}|_\_|\text{point}:16\_\text{VV0}|\text{hotel}:19\_\text{NN1}|\rangle$

- vs Prepositional verb e.g. refer to the map
  $\langle|\text{iobj}|\text{to}:12\_\text{II}|\text{refer}:11\_\text{VV0}|\text{map}:14\_\text{NN1}|\rangle$
Parser Evaluation

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

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASP</td>
<td>87.6</td>
<td>49.4</td>
</tr>
<tr>
<td>RASP with ANLT list</td>
<td>92.6</td>
<td>64.2</td>
</tr>
<tr>
<td>MINIPAR</td>
<td>78.9</td>
<td>44.1</td>
</tr>
</tbody>
</table>
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:

– \textit{climb+down}: \textit{clamber+up} .248 \textit{slither+down} .206 \\
\textit{creep+down} .183 ...

– \textit{climb}: \textit{walk} 0.152 \textit{jump} .148 \textit{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 $\leq 0.000001$
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 $X = 30, 50, 100, \text{ and } 500$
- **overlapS** of first $X$ neighbours, where $X = 30, 50, 100, \text{ and } 500$, 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
- climb 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 phrases as simplex:
clamber
slither
creep
scramble
skip
glance
stumble

step
wander
walk
skip
swing
leap
rush
disappear
fly

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)
  – \( \chi^2 \) (chi-squared)

• Man-Made resources
  – WordNet
  – ANLT lists (phrasal and prepositional verbs)
## Results

<table>
<thead>
<tr>
<th>Overlap</th>
<th>$r_s$</th>
<th>$Z$ score</th>
<th>$p$ under $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 30$</td>
<td>0.166</td>
<td>1.74</td>
<td>0.04</td>
</tr>
<tr>
<td>$X = 50$</td>
<td>0.136</td>
<td>1.43</td>
<td>0.08</td>
</tr>
<tr>
<td>$X = 100$</td>
<td>0.037</td>
<td>0.39</td>
<td>0.35</td>
</tr>
<tr>
<td>$X = 500$</td>
<td>-0.032</td>
<td>-0.38</td>
<td>0.35</td>
</tr>
<tr>
<td>OverlapS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X = 30$</td>
<td>0.306</td>
<td>3.21</td>
<td>&lt;0.00007</td>
</tr>
<tr>
<td>$X = 50$</td>
<td>0.303</td>
<td>3.18</td>
<td>&lt;0.00007</td>
</tr>
<tr>
<td>$X = 100$</td>
<td>0.263</td>
<td>2.76</td>
<td>0.0030</td>
</tr>
<tr>
<td>$X = 500$</td>
<td>0.167</td>
<td>1.75</td>
<td>0.040</td>
</tr>
</tbody>
</table>
## Results continued…

<table>
<thead>
<tr>
<th>X=500</th>
<th>statistic</th>
<th>Z score</th>
<th>p under H₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>sameparticle</td>
<td>$r_s=0.414$</td>
<td>4.34</td>
<td>&lt; 0.000003</td>
</tr>
<tr>
<td>sameparticle-simplex</td>
<td>$r_s=0.49$</td>
<td>5.17</td>
<td>&lt;0.000003</td>
</tr>
<tr>
<td>simplexasneighbour</td>
<td>MW</td>
<td>0.950</td>
<td>0.171</td>
</tr>
<tr>
<td>simplexrank</td>
<td>$r_s=-0.115$</td>
<td>-1.21</td>
<td>0.113</td>
</tr>
<tr>
<td>simplexscore</td>
<td>$r_s=0.052$</td>
<td>0.54</td>
<td>0.295</td>
</tr>
</tbody>
</table>
## Correlations of GS with man-made resources and statistics

<table>
<thead>
<tr>
<th></th>
<th>statistic</th>
<th>Z score</th>
<th>P under $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLR</td>
<td>$r_s = -0.168$</td>
<td>-1.76</td>
<td>0.0392</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>$r_s = -0.213$</td>
<td>-2.22</td>
<td>0.0139</td>
</tr>
<tr>
<td>MI</td>
<td>$r_s = -0.248$</td>
<td>-2.60</td>
<td>0.0047</td>
</tr>
<tr>
<td>Phrasal freq</td>
<td>$r_s = -0.096$</td>
<td>-1.01</td>
<td>0.156</td>
</tr>
<tr>
<td>Simplex freq</td>
<td>$r_s = 0.092$</td>
<td>0.96</td>
<td>0.169</td>
</tr>
<tr>
<td>WordNet</td>
<td>MW</td>
<td>2.39</td>
<td>0.0084</td>
</tr>
<tr>
<td>ANLT phrasals</td>
<td>MW</td>
<td>3.03</td>
<td>0.0012</td>
</tr>
<tr>
<td>ANLT prep ns</td>
<td>MW</td>
<td>0.430</td>
<td>0.334</td>
</tr>
</tbody>
</table>
Correlation of measures with man-made resources

<table>
<thead>
<tr>
<th></th>
<th>In WordNet</th>
<th>In ANLT phrasals</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>-2.61</td>
<td>-4.53</td>
</tr>
<tr>
<td>sameparticle-simplex</td>
<td>3.71</td>
<td>4.59</td>
</tr>
</tbody>
</table>
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