Modeling the Frequency of Phrasal Verbs with Search Engines

Grażyna Chamielec¹  Dawid Weiss²

¹ SuperMemo (consultant)
² Institute of Computing Science, Poznan University of Technology, Poland
Phrasal verbs

[... ] a simple verb combined with an adverb or a preposition, or sometimes both, to make a new verb with a meaning that is different from that of the simple verb [... ]
Phrasal verbs

[…] a simple verb combined with an adverb or a preposition, or sometimes both, to make a new verb with a meaning that is different from that of the simple verb […]

BREAK [UP/DOWN/…]
Phrasal verbs

[... ] a simple verb combined with an adverb or a preposition, or sometimes both, to make a new verb with a meaning that is different from that of the simple verb [... ]

BREAK [UP/DOWN/... ]

Non native speakers → must memorize
Motivation

SuperMemo’s training course
Motivation

SuperMemo’s training course

- Which PVs should be memorized first?
- Which PVs are rare enough to be omitted?
Measure PVs occurrence frequency on a corpus
Methodology

Measure PVs occurrence frequency on a corpus

Extract frequent and infrequent PV groups
QUANTITATIVE OR QUALITATIVE?
QUANTITATIVE OR QUALITATIVE?

corpus size, extraction accuracy, effort required, approximation quality, subjective bias, ...
How To Ask Out a Girl | eHow.com
How to ask out a girl is a question many will pose, but there usually lies a different question underneath. Everyone that asks "how to ask a girl out", ...
www.ehow.com/how_2088131_ask-out-girl.html - 62k - Cached - Similar pages - Note this

How to Ask Out - Asking Someone Out - Tips for Asking Out - Ask ...
Fear not, as we bring you the answer to the question of how to ask out. ... Always remember, it is better to ask out and get a no, than not to have asked ...
festivals.iloveindia.com/valentines-day/how-to-ask-out/index.html - 21k - Cached - Similar pages - Note this

Ask out - English Phrasal Verb - UsingEnglish.com
Ask out - To invite someone for a date. Example: He wanted to ASK her OUT but was too shy. Notes: - Separable [obligatory] - International English ...
www.usingenglish.com/reference/phrasal-verbs/ask+out.html - 14k - Cached - Similar pages - Note this
• Document count, occurrence count.
• Issues of bias and trust.
• Possible errors.
Sources of errors:

- Document count is a rough estimate.
- False hits: I’m in Wisła right now.
- Multiple-meaning PVs.
- Separable PVs: I will ask that girl out.

Remedies:

- Use independent search providers.
- Multiple samples, cross-validatation.
- Wildcard queries: ask * out.
EXPERIMENT
Data set:

- Manually assembled list of 991 PVa.
- Separable and inseparable PVs.
- Word form variations.
- Total: 10633 entries (‘queries’)
Data set:
- Manually assembled list of 991 PVa.
- Separable and inseparable PVs.
- Word form variations.
- Total: 10633 entries (‘queries’)

Search engines:
- Google, Yahoo, AllTheWeb, Live, Gigablast
Data set:
- Manually assembled list of 991 PVa.
- Separable and inseparable PVs.
- Word form variations.
- Total: 10633 entries (‘queries’)

Search engines:
- Google, Yahoo, AllTheWeb, Live, Gigablast

Data collection process:
- APIs, scraping.
- Terms of license.
- Approx. 10-samples for each PV form.
Results

- peculiarities
- PV ordering
GOOGLE HAS IT ALL
GOOGLE HAS IT ALL

Well, yes. Maybe. We can’t say.
Which SE has the largest index? Which SE has the largest doc. count variation? Are search engines consistent between each other in reported document counts?
Index size and distribution differences

- AllTheWeb
- Yahoo!

- exponential distribution
- near-exact match

all phrasal verbs (forms), by Yahoo's order
Index size and distribution differences

![Graph showing distribution of phrasal verbs (forms) by Yahoo!'s order.}]
Index size and distribution differences

all phrasal verbs (forms), by Yahoo's order

Gigablast

Yahoo!
Index size and distribution differences

- still high correlation
- different than others
Document count estimation stability (min-max)
PV rankings

Several ways of aggregating PV forms and samples:

- min, max, (truncated) average for all samples,
- with or without forms with wildcards,
- ordering by min, max, (truncated) average,
- PV ordering by min, max, (truncated) average of positions of its forms.

Most combinations had little influence on the final rankings.
Several ways of aggregating PV forms and samples:

- min, max, (truncated) average for all samples,
- with or without forms with wildcards,
- ordering by min, max, (truncated) average,
- PV ordering by min, max, (truncated) average of positions of its forms.

Most combinations had little influence on the final rankings.
XY plots of PV rankings, sort by $\text{avg}^2$, zero wildcards
XY plots of PV rankings, sort by $\text{avg}_2$, zero wildcards
XY plots of PV rankings, sort by $\text{avg2}$, zero wildcards
## Top and bottom of the ranking

<table>
<thead>
<tr>
<th></th>
<th>Top ↓</th>
<th></th>
<th>Bottom ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td></td>
<td>Google</td>
<td>Live</td>
</tr>
<tr>
<td>sign up</td>
<td>sign up</td>
<td>sign up</td>
<td></td>
</tr>
<tr>
<td>look for</td>
<td>look for</td>
<td>look for</td>
<td></td>
</tr>
<tr>
<td>check out</td>
<td>be in</td>
<td>check out</td>
<td>be in</td>
</tr>
<tr>
<td>be in</td>
<td>check out</td>
<td>check out</td>
<td>find out</td>
</tr>
<tr>
<td>look at</td>
<td>go back</td>
<td>find out</td>
<td></td>
</tr>
<tr>
<td>find out</td>
<td>look at</td>
<td>look at</td>
<td></td>
</tr>
<tr>
<td>arise from</td>
<td>find out</td>
<td>set up</td>
<td></td>
</tr>
<tr>
<td>come to</td>
<td>be after</td>
<td>come to</td>
<td></td>
</tr>
<tr>
<td>set up</td>
<td>look in</td>
<td>get to</td>
<td></td>
</tr>
<tr>
<td>go back</td>
<td>start off</td>
<td>work on</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Yahoo</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>slug out</td>
<td>sob out</td>
</tr>
<tr>
<td>winkle out</td>
<td>slog out</td>
</tr>
<tr>
<td>fur up</td>
<td>swirl down</td>
</tr>
<tr>
<td>satire up</td>
<td>nestle up</td>
</tr>
<tr>
<td>skirt round</td>
<td>fur up</td>
</tr>
<tr>
<td>sponge down</td>
<td>rein back</td>
</tr>
<tr>
<td>tail away</td>
<td>skirt round</td>
</tr>
<tr>
<td>be bombed out</td>
<td>sponge down</td>
</tr>
<tr>
<td>slog out</td>
<td>scorch along</td>
</tr>
<tr>
<td>hiss off</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** e-commerce, **be in**, overall impression.
Observations and conclusions:

- PV rankings biased with the Web and e-commerce.
- Sensible for our task.
- Interesting search engine differences.

Possible extensions:

- Validate against a more accurate method.
- HPSG + map-reduce.