Measuring compositionality in phrasal verbs

TYLER SCHNOEBELEN
STANFORD UNIVERSITY
DEPARTMENT OF LINGUISTICS
Phrasal verbs

- **Phrasal verbs** go by many names.
Phrasal verbs

- Phrasal verbs go by many names.
- But in short, they are the expressions that we native English speakers use to torture non-native speakers.
Phrasal verbs go by many names.
But in short, they are the expressions that we native English speakers use to torture non-native speakers.

- *Turn on*
- *Turn off*
- *Turn down*
- *Turn out*
- *Turn up*
- *Turn over*
- *Turn back*
- *Turn away*
- *Turn in*
Some are easier to learn than others

<table>
<thead>
<tr>
<th>Transparent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lift up</td>
</tr>
<tr>
<td>Spread out</td>
</tr>
<tr>
<td>Take away</td>
</tr>
<tr>
<td>Bring in</td>
</tr>
<tr>
<td>Pull back</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opaque</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give in</td>
</tr>
<tr>
<td>Chalk up</td>
</tr>
<tr>
<td>Wipe out</td>
</tr>
<tr>
<td>Touch off</td>
</tr>
<tr>
<td>Step down</td>
</tr>
</tbody>
</table>
Why phrasal verbs?

- They are highly frequent in and of themselves
- Their component pieces are also highly frequent
- For transitive phrasal verbs, the ability of the direct object to occur before or after the particle gives us a nice approximation of “tightness” between the verb and the particle.
Compositionality is the idea that the meaning of a complex expression comes from (i) its structure and (ii) the meanings of its constituents (see, for example, Fodor and Pylyshyn 1988).
Key definitions

- **Compositionality** is the idea that the meaning of a complex expression comes from (i) its structure and (ii) the meanings of its constituents (see, for example, Fodor and Pylyshyn 1988).

- To assess compositionality, I use the notion of entailment.
  - See also Lohse et al (2004) and Bannard (2002)

- Fully entailed phrasal verbs have standard semantics: they involve “literal” or “transparent” uses of the simplex meanings of their parts. When something is lifted up, it is both raised and it moves upward.

- **Opaque** phrasal verbs are those whose meaning you can’t predict just by knowing the meaning of their parts when they are used in isolation (*give in*).
The basic idea is a phrasal verb has some meaning $z$, which is made up of some amount of meaning from the verb ($x$), some from the particle ($y$), and something extra that just happens by combining them ($n$).

- $z = x + y + n$
- $z$ itself should be equivalent for transparent and opaque phrasal verbs
- But transparent phrasal verbs should get most of their meaning from ($x+y$)
- Opaque phrasal verbs should get most of their meaning from $n$
Opaque phrasal verbs have different distributions than transparent phrasal verbs.
- Experiments one and two

Information theoretic measurements are able to predict the entailment characteristics of different phrasal verbs.
- Experiment two

Adding information theoretic measures to Gries (2002)’s model of particle alternation (V Prt NP vs. V NP Prt) improves the classificatory adequacy of Gries’ model while reducing the overall number of factors.
- Experiment three

Taken together, the results suggest that the lexicon is not simply a list of entries but a highly interconnected network with patterns we can detect and describe.
Corpus experiment one
Background

- I work by analogy to Hay (2002), which shows that type and token relationships between parts and wholes can account for affix ordering better than absolute rules.
  - Hay (2002) is working on the old observation is that some affixes, like –ic and –ity seem to be less capable of affixing than other affixes like –ness and –less.
  - The idea is that decomposability is related to such ordering and that level 1 affixes (-ic, -ity) tend to be semantically opaque.
Hay (2002) takes all of the affixed words and looks at how many of them are likely to be parsed into base+affix.

- If the affixed word is more frequent than the base (government>govern), then it doesn’t count in the “average number of types parsed”.
- **Parsability** is central to Hay’s analysis—some things are seen as being composed of parts, others aren’t (she also looks at phonotactics and boundaries between morphemes).
Consider an affix that shows up in a corpus 1,000 times with 100 different bases.

That means you have 100 affixed word-\textit{types} and 1,000 affixed \textit{tokens}.

If all of the affixed forms are more common than base forms, then none of them are parsed for tokens or types.

The \textit{parsing ratios} will be $0/100=0$ (for types) and $0/1,000=0$ (for tokens). An affix like this one would be almost invisible to a speaker and the words that contain it would be likely to drift semantically.

At the other end, affixes that are highly separable will be parsed more often and have higher parsing ratios, too. These affixes will be much more productive when it comes to coining new words and the meaning of the affixed words will be much more predictable.
Hay (2002) on level ordering

<table>
<thead>
<tr>
<th></th>
<th>LEVEL 1 AFFIXES</th>
<th>LEVEL 2 AFFIXES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of types parsed</td>
<td>34.64</td>
<td>143.81</td>
</tr>
<tr>
<td>Average type-parsing ratio</td>
<td>0.3</td>
<td>0.61</td>
</tr>
<tr>
<td>Average number of tokens parsed</td>
<td>1139.21</td>
<td>3711.44</td>
</tr>
<tr>
<td>Average token-parsing ratio</td>
<td>0.12</td>
<td>0.34</td>
</tr>
<tr>
<td>Average number of hapaxes (V1)</td>
<td>22.79</td>
<td>77.31</td>
</tr>
<tr>
<td>Average productivity (P)</td>
<td>0.002</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table 1. Averaged figures for affixes typically classed as level 1 and level 2.
Methodology

- Bannard (2002) gives descriptions of verb and particle entailment for 180 phrasal verbs. 124 of these are either fully entailed (*lift up*) or fully unentailed (*give in*).
Bannard (2002) gives descriptions of verb and particle entailment for 180 phrasal verbs. 124 of these are either fully entailed (lift up) or fully unentailed (give in).

I extract 789 different phrasal verbs from the BNC (3,190 tokens), as well as all tokens of the base verbs. I perform an analysis of opaque and transparent phrasal verbs that is much like Hay (2002)’s approach to level 1 and level 2 affixes.
Methodology

- Bannard (2002) gives descriptions of verb and particle entailment for 180 phrasal verbs. 124 of these are either fully entailed (*lift up*) or fully unentailed (*give in*).
- I extract 789 different phrasal verbs from the BNC (3,190 tokens), as well as all tokens of the base verbs. I perform an analysis of opaque and transparent phrasal verbs that is much like Hay (2002)’s approach to level 1 and level 2 affixes.
- For each phrasal verb in the BNC, I calculate whether or not it was likely to be parsed as a single unit or broken into a verb and a particle.
  - Most of the examples of *shrug* are actually *shrug off*, so *shrug off* isn’t parsed.
  - There are lots of examples of *kick* that aren’t in a phrasal verb, so something like *kick off* is parsed.
The same trends as in Hay (2002)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Level 1 Affixes</th>
<th>Level 2 Affixes</th>
<th>Significance of difference (by Wilcoxon test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of types parsed</td>
<td>34.64</td>
<td>143.81</td>
<td></td>
</tr>
<tr>
<td>Average type-parsing ratio</td>
<td>0.3</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Average number of tokens parsed</td>
<td>1139.21</td>
<td>3711.44</td>
<td></td>
</tr>
<tr>
<td>Average token-parsing ratio</td>
<td>0.12</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Average number of hapaxes (Vi)</td>
<td>22.79</td>
<td>77.31</td>
<td></td>
</tr>
<tr>
<td>Average productivity (P)</td>
<td>0.002</td>
<td>0.030</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Averaged figures for affixes typically classed as level 1 and level 2.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Opaque/fully unentailed</th>
<th>Transparent/fully entailed</th>
<th>Significance of difference (by Wilcoxon test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg number of types parsed</td>
<td>2.49</td>
<td>5.52</td>
<td>p=3.756e-06</td>
</tr>
<tr>
<td>Avg type-parsing ratio</td>
<td>0.704</td>
<td>0.957</td>
<td>p=0.003360</td>
</tr>
<tr>
<td>Avg number of tokens parsed</td>
<td>18.10</td>
<td>33.48</td>
<td>p=0.001561</td>
</tr>
<tr>
<td>Avg token-parsing ratio</td>
<td>0.68</td>
<td>0.96</td>
<td>p=0.003363</td>
</tr>
</tbody>
</table>
Different distributions

Figure 1: The “Type parsed” density plots for transparent (black) and opaque (grey) phrasal verbs; they have significantly different distributions (p=3.756e-06).

Figure 2: The “Token parsed” density plots for transparent (black) and opaque (grey) phrasal verbs; they also have significantly different distributions (p=0.001561).
In the first experiment, I demonstrate a difference in distributions between semantically opaque and semantically transparent phrasal verbs ($p=3.756e-06$).

Following Hay (2002), we may say that perceptibility is a property that emerges from experiences with language: words can be seen as more and less complex (though this will vary by person and time).

- In phrasal verbs, too, we can say that there are different degrees of perceptibility.
- It’s the transparent phrasal verbs that are more obviously composed of pieces.
- Opaque phrasal verbs aren’t composed of such distinct pieces. The whole carries more meaning than the parts.
Corpus experiment two
Towards information theory

- While experiment one established that the relationships between particular verbs and particular particles mattered, experiment two goes further and model the relationship between all verbs and particles.
Recall the early intuition

- The basic idea is a phrasal verb has some meaning $z$, which is made up of some amount of meaning from the verb ($x$), some from the particle ($y$), and something extra that just happens by combining them ($n$).
  - $z=x+y+n$
  - $z$ should be equivalent for transparent and opaque phrasal verbs, but that transparent phrasal verbs get most of their meaning from ($x+y$) while opaque phrasal verbs should get most of their meaning from $n$

- Moscoso del Prado Martín et al (2004) provide a more rigorous way to measure these.
Information theoretic terms

- $z$, the meaning of the whole phrasal verb, is called “the amount of information”. It is the minimum amount of info necessary to encode something in binary.
  - $I(\text{phrasal verb}) = -\log_2(\text{frequency of phrasal verb/size of the corpus})$

- $x$ and $y$ are calculated as a joint entropy. They tell us how much help a phrasal verb gets from using a particular verb and a particular particle.

- $n$ is the “information residual”. To get it, just subtract the joint entropy of $x$ and $y$ from $z$. 

$$H(\mathcal{P}) = -\sum_{x\in\mathcal{P}} p(x|\mathcal{P}) \log_2 p(x|\mathcal{P}) = -\sum_{x\in\mathcal{P}} \frac{F(x)}{F(\mathcal{P})} \log_2 \frac{F(x)}{F(\mathcal{P})},$$
The amount of information theory being used across talks today is striking.
Notions like “entropy” are a bit complicated but really allow us quite elegant ways of expressing patterns.
And...
“You should call it entropy for two reasons: *first*, the function is already in use in thermodynamics under the same name...
“You should call it entropy for two reasons: **first**, the function is already in use in thermodynamics under the same name...

**Second**, and more importantly, most people don’t know what entropy really is, and if you use the word *entropy* in an argument you will win every time.”

– John von Newman to Claude Shannon
By this point in the day, “entropy” has been defined a couple different ways. Here’s how I’d sum the concept up:

- A die has six sides. That means there is a 1/6 chance of any particular event happening.
- A coin has two possibilities, each has ½ a chance.
- You’re going to be more uncertain about what’s going to happen when you roll a die than when you flip a coin.
- Entropy is the number of bits necessary to express an outcome. It takes more bits—more information—to express what happens with a die than a coin.

Entropy is formulated so that things that are more probable get shorter descriptions than things that are less probable.
Information residual

- Each phrasal verb is associated with an amount of information.
  - These are all roughly the same (there’s no significant difference between types of phrasal verbs).
- After you subtract away the verb entropy and the particle entropy, you’re left with the “information residual”.
  - That is, how much information is left over after you consider the verb and particle.
Methodology

- I use 6,793 phrasal verbs, consisting of 2,318 verbs and 48 particles that Baldwin and Villavicencio (2002) pulled out of the BNC (310,941 tokens).
- I calculate “amount of information”, “verb paradigm entropy”, “particle paradigm entropy”, and “information residual” for all of them.
- I test the results against Bannard (2002)’s gold standard of entailment characteristics.
Calculating information residual

1. Calculate the “amount of information” for the whole phrasal verb:
   - $I(\text{phrasal verb}) = -\log_2(\text{frequency of phrasal verb/size of the corpus})$
   - For drum up: $-\log_2(65/310,941) = 12.22$

2. Now start calculating the joint entropy. For each phrasal verb, look up the total number of occurrences for both the verb and the particle. Add these together.
   - There are 65 drum up’s and 1 drum out = 66 drums in phrasal verbs. There are a total of 79,236 tokens for “up”.

3. Subtract the total number of occurrences of the phrasal verb from this to get the actual union of the two paradigms.
Calculating information residual

4. Use the equation below to calculate two separate scores, one for the verb and one for the particle:
   - drum: \(-(66/79,237) \times \log_2(66/79,237)\) = 0.008521;

5. Do this for all 6,793 phrasal verbs. Now add every value of drum and up together.
   - We already calculated a score for drum in drum up, to that we add the score for drum in drum out (0.01812).
   - We also add all the 1,110 scores for up to get a joint entropy value of 3.432.

6. Now we subtract this joint entropy from the amount of information, \(I(drum \ up)\) and get the information residual.
I find a significant relationship between information residual and the compositionality of phrasal verbs ($p=2.49e-06$).

The measurements accurately predict the compositionality of 95 out of the 124 (76.61% accuracy).
Phrasal verbs differ by info residual

![Diagram showing average token-based information residual for different categories.](image)
Opaque phrasal verbs have higher info residuals

High information residual
- High predictability
- Presumably faster word recognition
- Presumably more phonetic reduction in speech

Low entropy values
- Low surprisal
- Low uncertainty
- Not much information from the “paradigms”
- Restricted in syntax and semantics
- Not productive
- Need to be learned
- Harder to parse into two pieces
Corpus experiment three
Corpus experiment three

- To examine syntactic phenomenon, I turn to Gries (2002)’s multifactorial account of particle placement in phrasal verbs.
- By adding information residual measurements to Gries’ analysis, I develop a more accurate model with fewer factors. The overall predictions for whether the particle comes before or after the direct object are 87.22% accurate, which is quite good.
These info residual values went in the direction one would predict: the lower the informational residual, the more likely the phrasal verb is to occur in the split construction (V NP Prt).

- It’s the transparent phrasal verbs that have lower information residual values, of course, and which receive so much support from the verb and particle paradigms that they can split—they are only loosely attached to each other after all.

- Similarly, we expect the “tightness” of opaque phrasal verbs—the fact that their meaning comes from the whole more than the parts—will lead them to prefer the joined construction (V Prt NP).
Methodology

- Gries (2002) pulls 403 sentences with phrasal verbs from the BNC and creates a generalized linear model (“discriminate analysis”) to determine which of 25 factors matter in determining particle placement.
- He sent me his data (thanks!), so I was able to add some new factors (entropy values, information residuals, log frequency, etc).
- I build a mixed-effects model up, factor by factor, testing the models with $G^2$ tests and bootstrapping to drop factors that weren’t worth keeping.
Factors in the final model

- **LENG_SYLL**: the length in syllables of the direct object. Longer direct objects are more likely to occur in joined constructions.
- **CohPC**: counts the number of times the direct object’s referent is mentioned in prior discourse, including superordinate and subordinate terms (*flowers/tulips*) (Gries 2002: 74). The more often the direct object is mentioned, the more likely it is to appear in the split construction.
- **DIR_PP**: describes whether there is a directional adverbial following the direct object or the particle. For example, “I would urge the panel to send out their proposed leaflet to the ministers in various areas” (Gries 2002: 75). The presence of a directional adverbial makes it more likely to get the split construction.
- **Resid_token**: if you calculate the amount of information in the phrasal verb and subtract the amount of support it receives from the verb and particle paradigms, this value is what left—the meaning of the phrasal verb that doesn’t come from its parts.
- **IDIOM**: Gries assigned a value to each example depending on whether it was a literal use of the phrasal verb, a metaphoric one, or an idiomatic one. The first two make it more likely that the construction will be split.
- **TYPE**: a nominal variable that describes the direct object: pronominal, semi-pronominal (“something else”), lexical, or proper name.
- **DET**: does the direct object have a definite determiner, an indefinite determiner, or no determiner.
- (Also, “Verb” as a random effect; note that I could also use information residual scores or entropy scores based on types.)
Final model for Gries (2002) data

- Generalized linear mixed model fit using Laplace
- Formula: CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + TYPE + DIR_PP + DET + Resid_token + (1 | Verb)
- Family: binomial(logit link)
- AIC  BIC logLik deviance
  278.0 329.9 -126.0  252.0
- Random effects:
- Verb  (Intercept) 0.43923  0.66275
- number of obs: 399, groups: Verb, 48
- Estimated scale (compare to 1)  0.9157284
- Fixed effects:

|                     | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | -0.66456 | 0.71390    | -0.931  | 0.351915 |
| log(LENG_SYLL)      | -1.65989 | 0.31986    | -5.189  | 2.11e-07 *** |
| CohPC               | 0.22400  | 0.06423    | 3.487   | 0.000488 *** |
| IDIOMlit            | 1.85312  | 0.61044    | 3.036   | 0.002400 ** |
| IDIOMmet            | 1.12563  | 0.61988    | 1.816   | 0.069387 . |
| TYPEpron            | 17.91050 | 1506.53681 | 0.012   | 0.990515 |
| TYPEpropN           | 1.37901  | 0.84562    | 1.631   | 0.102940 |
| TYPEspron           | 1.76149  | 0.89210    | 1.975   | 0.048320 * |
| DIR_PP              | 1.96163  | 0.47158    | 4.160   | 3.19e-05 *** |
| DETindef            | -1.26500 | 0.49900    | -2.535  | 0.011243 * |
| DETnone             | -1.03569 | 0.43006    | -2.408  | 0.016030 * |
| Resid_token         | -0.09311 | 0.03574    | -2.605  | 0.009185 ** |
Figure 4: Info_resid; The greater the information residual, the more likely to be in the joined construction

Figure 5: Log(LENG_SYLL); the longer the direct object is, the more likely it is to be joined

Figure 6: CohPC; the more often the direct object has been mentioned, the more likely it is that the split construction will be used

Figure 7: DIR_PP; if there is a following adverbial phrase, the split construction is more likely
When we fit the model’s predictions for the data against the actual data, we get 87.22% accuracy.

The final model also has a higher score for C and D_{xy} than other models: C=0.9465, D_{xy}=0.8930.

- C looks at an “index of concordance” between the predicted probability and the observed responses: a value of 0.5 would mean that the prediction might as well be random; 1.0 would be perfect prediction. The rule of thumb is that anything about 0.8 may mean the model has some real predictive capacity.

- D_{xy} stands for “Somers’ D_{xy},” which is a rank correlation between predicted probabilities and observed responses. In this case it is 0.0 that indicates randomness, while 1.0 indicates perfect prediction.
Conclusions
Recap

- Opaque phrasal verbs and transparent phrasal verbs have different distributions
  - Experiments one and two
- Information residual can predict entailment characteristics
  - Experiment two
- Information residual improves Gries (2002)’s model of particle alternation
  - Experiment three
1. Transparent phrasal verbs and opaque phrasal verbs are different from one another.
   - Practically any theory can handle this observation. The trouble comes in the next claim.
2. Opaque phrasal verbs can’t be relegated to the lexicon.
   - Opaque phrasal verbs don’t really behave like idioms. What syntactic reindeer games can’t they play? They still alternate, passivize, etc.
   - They share in common the characteristic of opacity with idioms, but they aren’t idioms themselves.
   - Opaque phrasal verbs get some meaning from their parts, just not as much as transparent ones do.
3. The use of language shapes the perception of parts and wholes.
   - Categories emerge from the patterns that the language user encounters; since they aren’t static, we expect some forms will, for example, become less compositional over time.
The interconnected lexicon

4. The lexicon is interconnected and syntax is sensitive to that.
   - Grammatical rules aren’t enough, you’d have to have something in each lexical entry.
4. The lexicon is interconnected and syntax is sensitive to that.
   - Grammatical rules aren’t enough, you’d have to have something in each lexical entry.
   - But even if lexical items are simply marked in their entries without have any connections tying them together, then we miss important generalizations.
The interconnected lexicon

4. The lexicon is interconnected and syntax is sensitive to that.
   - Grammatical rules aren’t enough, you’d have to have something in each lexical entry.
   - But even if lexical items are simply marked in their entries without having any connections tying them together, then we miss important generalizations.
   - Once we add in frequency and usage as fundamental parts of language, we are able to generalize about dynamic phenomena that traditional grammars can’t.
The lexicon is interconnected and syntax is sensitive to that.

- Grammatical rules aren’t enough, you’d have to have something in each lexical entry.
- But even if lexical items are simply marked in their entries without having any connections tying them together, then we miss important generalizations.
- Once we add in frequency and usage as fundamental parts of language, we are able to generalize about dynamic phenomena that traditional grammars can’t.
- We gain the ability to describe “what behaves like what”, and how that came about and is maintained.
Special thanks

- Arto Anttila
- Adrian Brasoveanu
- Joan Bresnan
- Penny Eckert
- Gus Elliott
- Stefan Gries
- Victor Kuperman

- Beth Levin
- Chris Manning
- John Rickford
- Tom Wasow

- And you, for listening
Questions?

- tylers@stanford.edu
- http://www.stanford.edu/~tylers
Appendix slides

(JUST IN CASE)
What about: *make out*, *bring up*, etc?

- Bannard only classified phrasal verbs as entailed or unentailed if there were at least four times as many examples of them being used one way than the other.
- To test if this was a problem, I recalculated the table matching Hay (2002), while treating 20% of the opaque phrasal verbs as transparents and 20% of each of the transparents as opaque.
  - Even though many of Bannard’s phrasal verbs had entailment characteristics higher than 80%, I chose to be conservative for all of them.
  - Everything stays significant except average number of tokens parsed (p=0.07455).
    - Types parsed (p=3.343e-05), Type-parsing ratio (p=0.0003102), and Token-parsing ratio (p=0.003490).
“Hear”
“Write out”
“Kiss”
Semantic heterogeneity in the paradigm does cause problems. Derivational measures like entropy still need more refinement for this.
Note that Ruhl (1976: 464) has fifty uses of “take off”
Flexibility

- A thrown-out piece of paper
  - Perhaps “throw out the garbage” is a morphological construct. (Toivonen 2002: 207 looking at the fact that verbs and particles don’t have to be head-final in English)
- Out! Throw the garbage out!
  - vs. ?Out! Throw out the garbage
- Overcome and come over
## All verbs

<table>
<thead>
<tr>
<th></th>
<th>All matching tokens, regardless of POS</th>
<th>BNC adverbial particle tokens + BNC prepositional tokens</th>
<th>All verb and adverbial particle lexemes in the BNC</th>
<th>BNC adverbial particles OR Baldwin and Villavicencio (2002) particles, whichever is larger</th>
<th>Baldwin and Villavicencio (2002) particle tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verb entropy</strong></td>
<td>2.17e-05 ***</td>
<td>2.34e-05 ***</td>
<td>p=0.00114 **</td>
<td>3.34e-05 ***</td>
<td>0.00195 **</td>
</tr>
<tr>
<td><strong>Particle entropy</strong></td>
<td>0.07009</td>
<td>0.3141</td>
<td>p=0.004724 **</td>
<td>0.0121 *</td>
<td>0.004724 **</td>
</tr>
<tr>
<td><strong>Information residual</strong></td>
<td>0.0109 *</td>
<td>0.06631</td>
<td>p=0.00195 **</td>
<td>0.0058 **</td>
<td>0.00114 **</td>
</tr>
</tbody>
</table>
All verbs

- As with the Bannard entailment data, we can also expand our verb counts to include all tokens of the verbs in question (not just those appearing in phrasal verbs). Starting from scratch, we still end up with all of the same factors from Gries, but we drop Resid_token in favor of Verb_entropy. This is the minimal model in which all factors are significant. Using G² tests we see that removing any of these factors results in a weaker model. The all-verbs model is 87.72% accurate, with C=0.9443 and Dxy=0.8887. The accuracy is a tiny bit better, while C and Dxy are a little worse.

Generalized linear mixed model fit using Laplace
- Formula: CONSTRUCTION ~ factor(DIR_PP) + DET + log(LENG_SYLL) + TYPE + CohPC + IDIOM + Verb_entropy + (1 | Verb)
- Family: binomial(logit link)
- AIC   BIC logLik deviance
  280.4 332.3 -127.2    254.4
- Random effects:
  Groups Name        Variance Std.Dev.
  Verb   (Intercept) 0.45902  0.67751
  number of obs: 399, groups: Verb, 48
  Estimated scale (compare to 1 )  0.8703255
- Fixed effects:
  Estimate Std. Error z value Pr(>|z|)
  (Intercept)       -1.74323    0.87699  -1.988 0.046840 *
  factor(DIR_PP)1    1.94069    0.46933   4.135 3.55e-05 ***
  DETindef -1.38953    0.49946  -2.782 0.005402 **
  DETnone -1.03017    0.42349  -2.433 0.014992 *
  log(LENG_SYLL)       -1.56241    0.31248  -5.000 5.73e-07 ***
  TYPEpron 18.36774 1597.43904   0.011 0.990826
  TYPEpropN 1.36916    0.81831   1.673 0.094295 .
  TYPESpron 1.85872    0.86751   2.143 0.032146 *
  CohPC 0.22642    0.06362   3.559 0.000373 ***
  IDIOMlit 2.09823    0.61616   3.405 0.000661 ***
  IDIOMmet 1.35849    0.62179   2.185 0.028903 *
  Verb_entropy 2.16729    1.00625   2.154 0.031254 *
Other things to talk about

- **Representations, cognition, exemplar-based theories**
  - And the network metaphor
  - What constrains them?
  - How do they encode social information?
  - How do they handle the phonetic/phonology interface?

- **Compositionality**
  - Gradience and its consequences
  - Productivity
  - Systematicity
# Toy entropies

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>A ent</th>
<th>B ent</th>
<th>C ent</th>
<th>D ent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>25</td>
<td>1</td>
<td>20</td>
<td>-0.46439</td>
<td>-0.5</td>
<td>-0.11288</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>25</td>
<td>30</td>
<td>20</td>
<td>-0.46439</td>
<td>-0.5</td>
<td>-0.44218</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>25</td>
<td>30</td>
<td>20</td>
<td>-0.46439</td>
<td>-0.5</td>
<td>-0.44218</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>20</td>
<td>20</td>
<td>-0.46439</td>
<td></td>
<td>-0.18575</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>13</td>
<td>20</td>
<td>20</td>
<td>-0.46439</td>
<td></td>
<td>-0.50529</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4</td>
<td>20</td>
<td>20</td>
<td>-0.46439</td>
<td></td>
<td>-0.29151</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>50</strong></td>
<td><strong>50</strong></td>
<td><strong>50</strong></td>
<td><strong>100</strong></td>
<td><strong>2.321928</strong></td>
<td><strong>1</strong></td>
<td><strong>1.537608</strong></td>
</tr>
</tbody>
</table>

**Formula:**

\[
H(\mathcal{D}) = - \sum_{x \in \mathcal{D}} p(x|\mathcal{D}) \log_2 p(x|\mathcal{D}) = - \sum_{x \in \mathcal{D}} \frac{F(x)}{F(\mathcal{D})} \log_2 \frac{F(x)}{F(\mathcal{D})},
\]
Transparent phrasal verbs: high entropy values

High entropy
- High surprisal
- High uncertainty
- Lots of information from the paradigms
- Longer descriptions
- Productive
- Intelligible
- Parsable

Low predictability
- Presumably slower to read
- Presumably less phonetic reduction

\[ H(\mathcal{P}) = - \sum_{x \in \mathcal{P}} p(x|\mathcal{P}) \log_2 p(x|\mathcal{P}) = - \sum_{x \in \mathcal{P}} \frac{F(x)}{F(\mathcal{P})} \log_2 \frac{F(x)}{F(\mathcal{P})}, \]
As Lüdeling and de Jong (2002) point out, large family sizes—shorter reaction times; you can limit family sizes to ‘relatedness’.

- “If particle verbs are unanalyzable nits at the semantic level, we expect to find a different effect of family size for transparent versus opaque particle verbs. Transparent particle verbs would inherit the total number of family members which are semantically consistent with the opaque meaning...On the other hand, if particle verbs in the mental lexicon are represented with a composite phrase-like structure, we expect activation to flow to all family members related to the base verb, as the composite nature of both kinds of particle verbs would allow the base verb, together with its family members to become activated in the central mental lexicon (2002: 326-327)

- They’d like to make the claim that phrasal verbs are listed in the lexicon as composite entries.
  - No reliable differences in reaction times between opaque and transparent phrasal verbs in reaction times. Family size didn’t predict transparent phrasal verbs, but full family did for opaque verbs, contra others’ work.
Compositionality and systematicity
The real goal is how a human being can understand sentences they’ve never heard before.

Usually this requires a static notion of meaning, but you can use dynamic logic (as in programming logic: x+1). Basically, you just use a more abstract notion of meaning.

- Similarly, you can use Situation Semantics to convey information about the world/state of mind as a relation.

- *Pet fish* and “holism”, the meaning of a lexical expression (at least in part) by the meanings of other expressions in the language.
  - Robbins (2001: 333) uses the term “superstrong holism” for the idea that there are words which do not have meaning independent of meaningful phrases containing them.
Frege and compositionality

- Early Frege (1884: xxii) writes this:
  - “One should ask for the meaning of a word only in the context of a sentence, and not in isolation.”
  - This notion of contextuality isn’t compatible with compositionality, since compositionality needs words in isolation so that meanings of compounds can be built out of them.

- Later Frege (1923) writes this:
  - “With a few syllables [language] can express an incalculable number of thoughts...This would be impossible, were we not able to distinguish parts in the thoughts corresponding to the parts of a sentence, so that the structure of the sentence serves as the image of the structure of the thoughts.”
  - This sounds more like compositionality.

- See also Janssen (to appear), his conclusion is that Frege didn’t come up with compositionality and that it wasn’t til later writing that you can even find support for it.
Compositionality gets you two things: productivity and systematicity.

On productivity: *The hippo loves the tortoise*
- Since competent speakers can understand a complex expression \( e \) they never encountered before, it must be that they (perhaps tacitly) know something on the basis of which they can figure out, without any additional information, what \( e \) means. If this is so, something they already know must determine what \( e \) means. And this knowledge cannot plausibly be anything but knowledge of the structure of \( e \) and knowledge of the meanings of the simple constituents of \( e \). (Szabo 2007)

On systematicity: if you understand *callow roustabout* and *mercurial polyglot*, then you understand *mercurial roustabout* and *callow polyglot*.
- Within an hour
- Without a watch
- Within a watch
- Without an hour
- Halfway closed
- Firmly believed
- Halfway believed
- Firmly closed
- Most people sleep at night.
- Most people drink at night.
To really understand systematicity in cognition, you need to get similarity right so that representations of similar things can be treated as similar (see also Van Gelder 1990: 379).
More on systematicity

- What we want to preserve about systematicity is the idea of generality from Evans (1982: 100):
  - The thought that John is happy has something in common with the thought that Harry is happy, and the thought that John is happy has something in common with the thought that John is sad.
- The “something in common” is accounted for by positing lengthy exemplar representations and threads between exemplars that are similar in one, two, or many ways.
  - Though there will be exemplars of full utterances and odd subcomponents of them, so long as we have a decomposition function, the majority of exemplars will be individual words, connected and defined by exemplars that offer fuller context. This allows many threads to follow John and Harry to happiness.
Cognitive models
Cognitive models

- **Transparent**: She took away her daughter’s iPod and grounded her for a week.
- **Opaque**: He played down his illness because he didn’t want anyone thinking he wasn’t up to the job.
- **Opaque**: They wiped out the invading army.
  - Knock out, hammer out, wipe up
- **Chaque mot a son histoire (chaque phrase aussi)**
Top phrasal verbs—a network
Conclusion

- What constrains exemplar theories or our network metaphor or our information theoretic terms?
- If everything is connected, then there always is a path from a to z.
- But we can constrain such paths—the paths are not equally short, nor equally well-trodden.
- We should assume, given finite mental resources and the finite time of language processing, that there are indeed constraints.
- Since language experiences are neurally encoded, their representations and activations are subject to the same memory effects that the rest of the cognitive system is.
- Frequency and recency are expected to play important roles.
  - The techniques employed in this paper do not capture recency, but measures like entropy are able to chart the effects of frequency, describing what combinations have more information, more possibilities, and more uncertainty. These will be things like literal phrasal verbs, whose parts have greater entropy values than their opaque cousins’.
- If the composition of meaning is about the paths activated in the speaker and in the hearer, we expect to find further evidence of gradience than that provided in this paper.
- With gradient compositionality, we predict that miscommunication will be aggravated as the distance between two speakers’ life experiences grows. But the majority of communication between fluent speakers will be relatively easy—they can fill in each other’s blanks because their own brains have similar paths that are similarly lit up.
- Deciding how similarity is defined is an important next step for researchers working on usage-based theories.
- What is the balance to strike between formulaic language and the creative?
Exemplar-based models

- Exemplar theory doesn’t have rules or grammar, just decomposition and recomposition functions and exemplars for all the utterances ever heard or produced (though these may decay and go away eventually).
  - Exemplars aren’t the same as the actual language tokens, though. They are categorized, classified, analyzed versions.
  - By storing all of the parts of each utterance—and all of the subparts—Bod (2006) is able to give an account for how exemplar theories deal with productivity.
  - The composition and decomposition functions aren’t really sensitive to anything except previous representations. They could produce gobbledygook, but are constrained by probabilities that reward structures that have occurred more frequently. Since they are probabilistic, they can handle the dependencies that rule-based accounts have difficulty with since the rules need to be independent (see also Bod 2006; 292).

- Our information theoretic measurements can help trace the strength of paths through the network, but there still needs to be a better notion of “similarity”.
  - When a new word is encountered, it is classified according to its similarity to the exemplars already stored. Its similarity to any single stored exemplar can be computed as its 'distance' from the exemplar in the parameter space. The most probable labelling given the labelling of the exemplars in the neighborhood is computed on the basis of stored exemplars" (Bod and Cochran 2007: 2).
For Pierrehumbert (2001), modeling frequency effects is an important thing to do and exemplars do it well.

Usage-based frameworks can handle the idea that there are no cases in which analogous phonemes in two different languages show the exact same phonetic target and patterns of variation.

Usage-based frameworks deal with this by saying the mental representations of phonological targets and patterns are gradually built from experience with speech.

No real separation between lexicon and grammar.

“In an exemplar model, each category is represented in memory by a large cloud of remembered tokens of that category. These memories are organized in a cognitive map, so that memories of highly similar instances are close to each other and memories of dissimilar instances are far apart.” (3)

Basically, you map between points in a phonetic parameter space and the labels of the categorization system. Labels are levels of representation or functional links to other levels of representations.

- You can be in more than one categorization scheme!
Hintzman/Goldinger use a perception model that encodes “more episodic information” in memory for rare events than frequent ones.

That means you can remember better that a word was spoken in a particular voice if that word is rare than if it’s common.

This has to do with entrenchment effects, which exemplar models aren’t always good with. Often, contra entrenchment effects, models predict that with frequency you get more production noise and therefore more variance over time. But think of kids taking up the cello—their tune has less variance with practice, not more.

One thing we might want to think about is how people adapt speech patterns to their audience even without overt pressures and rewards (communicative success and social attunement are probably rewards—how are they assessed?)
Spreading

- You sum up exemplar strengths because an exemplar spreads activation to “labels” that are similar.
- Production isn’t just a single target exemplar (selected at random). Instead, “a target location in the exemplar cloud is selected at random, and the exemplars in the neighborhood of this location all contribute to the production plan, to a degree which reflects their activation level” (Pierrehumbert 2001: 11).
Connectionism

- Van Gelder (1990: 363)
  - “Connectionists are finding various ways to implement nonconcatenative schemes making use of the massive and distinctive processing capacities of neural networks.
  - “Since representations in a nonconcatenative scheme do not contain tokens of their constituents, they are without syntactic structure...This gives rise to an important general point: A nontrivial representation of a complex structured item need not itself have an internal compositional structure.”
- If you don’t have to first extract “basic constituents” you can capitalize on the inherent and systematic structural similarities among the nonconcatenative representations you’re looking at.
- Note that Fodor and Pylyshyn (1988: 12) are clear that they think the Language of Thought uses a combinatorial syntax and semantics.
Connectionism II

- Connectionist representations are sort of vectors (patterns of activity over processing units or the connections between such units).
- Smolensky (1987: 2) defines the problem as “finding a mapping from a set of structured objects (e.g., trees) to a vector space”.
- He wants to preserve the various constituency relations among representations, so that a complex representation can still get you the representation of the parts, as necessary.
  - Cup of coffee=contains (for cup), contained by (for coffee)
Examples
Highest and lowest information residuals

- **Highest (all opaque)**
  - rack up
  - firm up
  - beef up
  - drum up
  - shape up
- **Lowest (all transparent)**
  - give away
  - go away
  - take away
  - walk away
  - walk around
- **Middle range (some of each)**
  - go through
  - run up
  - give out
  - go out
  - hold out
Predicted to be opaque, Bannard says transparent

- force out
- blurt out
- spread out
- go out
- push up
- scare off
- send out
- stand up
- give out
- call in
- push through
- pay down
- bring out
- go up
- move up
- buy back
- sit down
- bring in
- walk out
- move out
- pay back
- walk in
Predicted to be transparent, Bannard says opaque

- set aside
- set off
- go ahead
- turn down
- pass over
- take off
Highest and lowest probabilities

- **Opaque**
  - come back 0.83305389
  - give away 0.86125375
  - take away 0.89245151
  - walk away 0.9211611
  - walk around 0.92535493

- **Transparent**
  - rack up 0.0387375
  - chalk up 0.03889891
  - bottom out 0.04604793
  - snap up 0.04801753
  - scoop up 0.04888556

- **Right in the middle**
  - write down 0.47163784 o o
  - walk in 0.48451959 t o
  - put in 0.4863609 o o
  - lay off 0.49129119 o o
  - send back 0.5329444 t t
Call in

- My model predicts that it’s opaque
- Not listed in Bannard (2002)
- But McCarthy et al (2003) gives three linguists judgments on its compositionality (0=opaque, 10=transparent). Here are their scores:
  - 8
  - 8
  - 2
Other phenomena like phrasal verbs

- These also have varying degrees of “tightness”:
  - To fall ill
  - To turn turtle
  - To take to
Verb classes (Levin 1993)

- Only enough data to really analyze verbs of
  - Existence
  - Motion
  - Change of state
  - Removing
  - Sending/receiving

- None of the verbs in these classes were significantly more likely to be in split or joined construction, nor opaque/transparent.
“Morphological structure emerges from the statistical regularities that characterize the forms and meanings of words. In this view, morphological structure is inherently graded... The degree to which *ed* is ‘present’ in *walked* depends on the amount of analogical support from other words in the lexicon occupying similar positions in the inflectional paradigm (e.g. *thanked*, *warmed").”

(Hay and Baayen 2005: 342-343)
Earlier notions and intuitions
Intuitions about cognition

- “A language would be a difficult thing to handle if its speakers had the burden imposed on them of remembering every little item separately.” (Jesperson 1924/1966: 177)
- “Our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs.” (Bolinger 1976: 1)
- Meaning is indeterminate, clarified by context (see also Mitchell 1971: 37 which is skeptical that one measurement can capture meaning).
- Against this is the desire to make linguistics scientific and to define scientific as relying on stability and provability.
Saussure (1916)

“Synthesizing the elements of [a] syntagm into a new unit...[such that] when a compound concept is expressed by a succession of very common significant units, the mind gives up analysis—it takes a short cut—and applies the concept to whole cluster of signs, which then becomes a simple unit.”
Collocations

- “A collocation is defined as a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components.” (Choueka 1988)
- “Thus it is that the different uses, in which we have come to be associated with a word or a phrase, associate themselves with each other.” (Paul 1890/1970: 6)
- “One of the meanings of night is its collocability with dark, and of dark, of course, its collocation with night.” (Firth 1957)
Idioms

“An idiom is a construction whose words occur elsewhere but never with the same meaning as in this construction. This definition allows the possibility that the words may contribute to, yet not fully account for, the meaning of the construction.” (Ruhl 1977: 459)

See also Ruwet (1991) who argues that the syntactic frozenness of idiomatic expressions isn’t just simple or arbitrary marking but comes out of the meaning that the speakers assign (or don’t) to the individual parts.
Historical change

- Do they have natural completion states?
- Do they become stable?
- Can models predict that?
- “The less productive a pattern is, the more likely it is that if a new form does not get coined by the pattern it will have idiomatic value.” (Hockett 1958: 308)
- Traditional view: highly opaque idioms are more transparent from a diachronic point of view.
The goal of linguistics is to develop grammars
  ○ And meta-grammars! (Theories of grammars)
These should be formal accounts of the structure of human languages.
The problem: how do meaning and semantics get put into the structures?
  ○ (Recall Chomsky 1980’s distinction between problems we might be able to solve and mysteries that we may not. As Ruwet 1991: xxi points out: “mysteries interfere with problems—or, irony of ironies...the key to the problems lies in the mysteries!”)
The possibilities

- Is what you see of the world through this window so beautiful that you absolutely do not want to look through another window—that you even try to prevent others from doing? (Nietzsche, *Human All Too Human*, vol. 2, 359)