

Author: Tyler Schnoebelen
University/Affiliation: Stanford University, Department of Linguistics
Email address: tylers@stanford.edu

Measuring compositionality in phrasal verbs

Introduction

This paper demonstrates how to measure the compositionality of phrasal verbs using corpus frequencies from the BNC. This allows us to distinguish semantically transparent phrasal verbs (*they lifted up their hats*) from opaque ones (*they summed up their feelings*). Working by analogy to paradigmatic approaches to morphology (Moscoso del Prado Martín et al 2004), I use information theoretic terms to reveal and express a complicated web of relationships between verbs and particles. In so doing, I am able to predict two different sets of data—one semantic, the other syntactic.

Experiment one: Semantics and parsability

Hay (2002) looked at the ordering of English affixes in terms of their “parsability”—that is, a word like *government* is unlikely to be parsed as *govern+ment* since *government* is more frequent than *govern*. On the other hand, *discern* is more common *discernment*, so the affixed word is likely to be parsed. Similarly, I show that we can determine how distinct the parts of a phrasal verb are by counting the relative frequencies of the verbs participating in phrasal verbs. The prediction, which is borne out, is that literal phrasal verbs will be more obviously made up of parts than opaque ones.

I extract 789 different phrasal verbs from the BNC (3,190 tokens), as well as all tokens of the simplex verbs. Bannard (2002) gives the entailment characteristics of 180 phrasal verbs, and I analyze the parsability of the 124 that are either fully entailed (*lift up* entails both lifting and something going up) or fully unentailed (there is no literal *summing* or *up* in *summing up feelings*). Thinking about phrasal verbs in terms of entailment is a key notion in Bannard (2002); it’s also used to good effect in Lohse et al (2004). In a paradigmatic approach like Hay and Baayen (2005), the idea is that forms can get support from other words in the lexicon occupying similar positions. These results suggest that *up* simply isn’t as present in *sum up* as it is in *lift up*.

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 by comparing the frequencies of the simplex verb with the verb in its phrasal verb combinations. As long as there are more examples of the simplex verb, the phrasal verb will be parsed. For the verbs that are parsed, I add up how many different “types” there are—this means adding up the number of different particles that they take. For *kick off* we see that its verb combines with not just *off* but *through*, *around*, *up*, and *in*. Thus its “number of types parsed” is 5.

To determine the “average type-parsing ratio” I simply divide the number of parsed types by the total number of types for the verb. There are 18 examples of *wind*; 11 of them with *up*, three of them with *down*, four of them without any particle at all. That means that *wind* has a type-parsing ratio of $1/2=0.5$ since *wind down* is parsed but

wind up is not. The bottom two rows in the table are built similarly, only using tokens instead of types.

Table 1. Phrasal verbs behave similar to Hay (2002)’s investigation of affixes.

	Opaque/fully unentailed	Transparent/fully entailed	Significance of difference (by Wilcoxon test)
Avg number of types parsed	2.49	5.52	p=3.76e-06
Avg type-parsing ratio	0.704	0.957	p=0.00336
Avg number of tokens parsed	18.1	33.5	p=0.00156
Avg token-parsing ratio	0.680	0.960	p=0.00336

For each row, it is the transparent column that has the higher value—just as in Hay (2002), where it’s the more decomposable/parsable level 2 affixes (*-less*, *-ness*) that score higher than the level 1 affixes (*-ity*, *-ic*). Hay’s prediction is that highly parsable affixes “will contain predictable meaning, and will be easily parsed out. Such affixes can pile up at the ends of words, and should display many syntax-like properties” (Hay 2002: 535). Here, in the realm of phrasal verbs, we recall Gries (2002)’s finding that literal items like *lift up* take more advantage of the “actually syntactic” property of flexible alternation between NP objects and particles.

Experiment two: Semantics and information theory

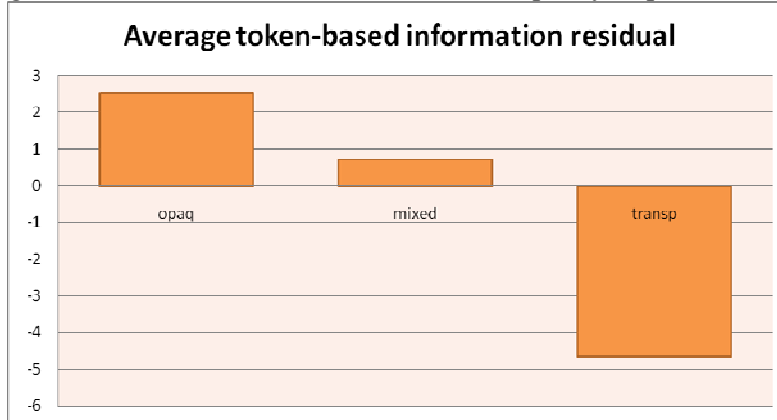
Moscoso del Prado Martín et al (2004) use information theory to develop measures for the amount of information contained in a particular word and the amount carried by the different morphological paradigms it’s a part of—in other words, how does a word get composed of meaning? How much does each part and paradigm contribute? Specifically, they calculate the “information residual” based on the overall amount of information ($-\log_2(\text{frequency of phrase/size of the corpus})$) minus the support from its various paradigms, which is measured by a verbal entropy score and a particle entropy score. These numbers are calculated twice—once using token counts and once using type counts. In the type-based calculations, the verb entropy is determined by the number of particles that a verb combines with; the particle entropy is likewise determined by counting how many verbs a particular particle combines with.

Entropy is the number of bits that are necessary to express an outcome—the greater the number of outcomes (and the greater the variety in those outcomes), the greater the entropy (Cover and Joy 2006). Here, there are more outcomes possible for exactly the phrasal verbs that have the largest number of paradigm members: the literal phrasal verbs. Literal phrasal verbs are the most flexible, productive, and intelligible.

There are correspondingly fewer outcomes possible for opaque phrasal verbs, which are more restricted in their meaning and syntax and which are less capable of being parsed into separate pieces. Because the “amount of information” is relatively constant across all phrasal verbs—and because entropy values are subtracted from it—the smaller the entropy values, the larger the information residual. Again, that’s the amount of meaning not explained by the parts.

Using 6,793 phrasal verbs, consisting of 2,318 verbs and 48 particles from Baldwin and Villavicencio (2002), I create informational residual scores for all of the phrasal verbs that Bannard (2002) describes. I find that the token-based “information residual” scores for fully untailed phrasal verbs are reliably higher than that of fully entailed phrasal verbs ($p=2.49e-06$). The same thing happens in type-based analyses: the informational residual scores for untailed phrasal verbs are higher than entailed ($p=0.01530$).

Figure 1. Information residual describes the opacity of phrasal verbs.



Experiment 3: Syntax and information theory

Turning to syntactic realization, I create a generalized linear mixed-effects model with the actual data Gries (2002) used in describing factors that matter for predicting the particle placement of transitive phrasal verbs (*V NP Prt* or *V Prt NP*). Where Gries uses 15 fixed effects, my model has only seven fixed effects and one random effect (the verb itself). Despite the fact that I have simplified the model, I still achieve slightly higher classification accuracy.

Having experimented with no fewer than 26 different variables (including simple log frequency measurements), my final model is comprised of the (i) length of the direct object (DO) in syllables, (ii) the number of times the DO’s referent is mentioned in prior discourse, (iii) whether there is a directional adverbial following the DO/particle, (iv) the type of DO (pronominal, lexical, etc.), (v) whether the DO has a definite/indefinite/absent determiner, (vi) the token-based information residual, and (vii) Gries’ hand-coded measurement of idiomaticity (idiomatic/metaphorical/literal).

All the factors in the final model are significant and G2 tests demonstrate that removing any of the factors creates a weaker model, while adding any others fails to improve it. This model achieves 87.22% classification accuracy.

Conclusion

While opaque phrasal verbs share the characteristic of “opacity” with idioms, it seems difficult to actually relegate them into the idiom-corner of the lexicon—they don’t alternate quite as much or as easily as transparent phrasal verbs, but they still alternate. They also fail other heuristics for idioms (for example, they can passivize). It may be difficult to capture this in grammatical rules unless individual lexical items are marked

and there are different (but very similar) rules that are sensitive to what they find in each lexical entry. Yet even if this approach is tenable, it may not capture the observation that phrasal verbs and their pieces are connected to each other through patterns of usage.

In the first experiment, I used corpus frequencies to demonstrate a difference between semantically opaque and semantically transparent phrasal verbs. The difference lies in the fact that opaque phrasal verbs don't combine with as many particles and the fact that their frequencies, relative to other instances of their simplex verbs, make them more likely to be treated as a single entity.

The next two experiments found the same patterns as the first, but measured them in terms of information theory. While experiment one established that the relationships between particular verbs and particular particles mattered, experiments two and three went further and modeled the relationship between all verbs and particles. By positioning each individual verb and particle in the context of how other verbs and particles were behaving, I showed even stronger results for estimating the entailment characteristics (experiment two) and I was even able to improve models of the "syntactic" phenomena of particle alternation (experiment three).

These corpus experiments establish that analogies to morphology are apt and that it is possible to bring frequencies into syntax and semantics in a meaningful way. Information theoretic terms give us a rich and elegant model for investigating patterns that emerge from actual language use. Such measurements ultimately lead us to ask rather indelicate questions: can generative approaches be adequate if they don't take usage into account? Is compositionality really a categorical phenomenon?

References

- Baldwin, T. and A. Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.
- Bannard, C. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. *LinGO Working Paper No. 2002-06*.
- Bolinger, D. 1971. *The phrasal verb in English*. Cambridge: Harvard University Press.
- Cover, T. M. and J. A. Thomas. 2006. *Elements of information theory, 1st Edition*. New York: Wiley-Interscience.
- Gries, S. T. 2002. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum International Publishing Group Ltd.
- Hay, J. 2002. From Speech Perception to Morphology: Affix-ordering revisited. *Language* 78 (3): 527-555.
- Hay, J. and Baayen, R. H. 2002. Parsing and Productivity. In *Yearbook of Morphology 2001*, G.E. Booij and J. V. Marle (eds.), 203-235. Kluwer Academic Publishers, Dordrecht.
- Lohse, B., J. Hawkins, and T. Wasow. 2004. Processing Domains in English Verb-Particle Constructions. *Language* 80 (2): 238-261.
- Moscoso del Prado Martín, F., A. Kostić, and R. H. Baayen. 2004. Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition* 94 (1): 1-18.
- Nunberg, G., I. Sag, and T. Wasow. 1994. Idioms. *Language* 70 (3): 491-538.