

Measuring compositionality in phrasal verbs

Tyler Schnoebelen, Stanford University

Abstract

Is compositionality the kind of thing that can be measured? Working by analogy to paradigmatic approaches to morphology, I investigate whether the patterns of token and type frequencies for verbs, particles, and whole phrasal verbs can be used to measure compositionality. The analogy proves to be apt and through three corpus experiments, I show that some phrasal verbs are less decomposable than others. Experiment one establishes that opaque phrasal verbs (like *give in*, whose components are unentailed) have different distributions than fully entailed phrasal verbs (e.g., *lift up*, which entails the simplex meaning of both *lift* and *up*). Experiment two uses information theoretic terms and an even larger corpus to confirm that opaque and transparent phrasal verbs are indeed different. It also shows that information theoretic measurements are able to predict the entailment characteristics of different phrasal verbs. Experiment three adds information theoretic measures to Gries (2002)'s model of particle alternation (V Prt NP vs. V NP Prt). These additions improve the classificatory adequacy of Gries' model while reducing the overall number of factors. 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.

Contents

1. Overview	3
What's ahead, section by section	4
2. Basics of phrasal verbs and idioms	5
3. Measuring compositionality	8
4. Analogies to morphology	10
5. Experiment one: opaque and transparent phrasal verbs are different	12
Discussion	16
6. Experiment two: Information theoretic terms predict entailment	17
Expanding to all verbs and all particles	21
Discussion	22
7. Experiment 3: Entropy measurements improve Gries (2002)'s model	24
Expanding to all verbs and particles	28

Discussion.....	30
8. General discussion	31
Review of findings.....	31
9. Next steps.....	34
Appendix A: Bannard (2002) entailment coding.....	35
Appendix B: Gries' definitions for IDIOM	36
Appendix C: Gries (2002) factors not in my model.....	37
Appendix D: Building a model with Gries (2002).....	38
Appendix E: What else is an idiom besides non-compositional?	46
Bibliography	49

1. Overview

This paper questions some of the old assumptions about what compositionality really is and offers new ways of measuring and conceptualizing it as a gradient phenomenon. I take 6,793 phrasal verbs, consisting of 2,318 verbs and 48 particles and look at how they pattern with each other and how they behave outside phrasal verb combinations, too. In so doing, I am able to predict two different sets of data—one semantic (entailment characteristics), the other syntactic (particle placement for transitive phrasal verbs).

Phrasal verbs offer a good testing ground for theories of compositionality because of their basic properties: (i) they are highly frequent in and of themselves, (ii) their component pieces are also highly frequent, (iii) in assessing compositionality there are three easy pieces: the phrase, the verb, and the particle,¹ (iv) for transitive phrasal verbs, the ability of the direct object to occur before or after the particle gives us a nice approximation of “tightness”.²

To assess compositionality, I use the notion of entailment. 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*). 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*).

In the first experiment, I demonstrate a difference in distributions between semantically opaque and semantically transparent phrasal verbs ($p=3.756e-06$). The difference lies in the fact that opaque phrasal verbs don’t combine with as many particles and that—relative to other instances of their simplex verbs—their frequencies make them more likely to be treated as a single, unparsed entity. So while transparent phrasal verbs seem classically composed, the whole of an opaque phrasal verb is more salient and meaningful than either of its parts.

The next two experiments find the same patterns as the first, but measure them in terms of information theory. While experiment one established that the relationships between particular verbs and particular particles mattered, experiments two and three go further and model the interconnected relationships between all verbs and particles. The information theory measurements I use are significantly related to the compositionality of these phrasal verbs ($p=2.49e-06$), and they are able to accurately predict the compositionality of 95 out of the 124 (76.61% accuracy).

Information theory has been important in engineering fields since Shannon and Weaver (1949), and computational linguists have been making good use of it for quite some time in speech recognition, machine translation, information retrieval, and even language analysis. More recently, morphologists like

¹ In fact, it would follow from my theory that we should also look at the NP object for transitive phrasal verbs. I have not done so here in order to keep my scope reasonable.

² This idea of the tightness of association between the verb and the particle is something that Kennedy (1920) noted over eighty years ago. He suggests, for example, that *put out* is a lot more tightly bounded than *blot out*, which is in turn tighter than *brush out*. By a tight bind, he meant that the connection between certain verbs and particles was stronger than others, but that such relationships existed along a continuum. The loosest among them “shades off imperceptibly into the great mass of adverbial modifications such as *fly away*” (Kennedy 1920: 9).

Moscoso del Prado Martín, Kostic, and Baayen (2004) have deployed it to explain response latencies in visual decision tasks. Though it can take a bit of effort to get one's head around it, information theory is mathematically quite elegant. In the sections that follow, I build towards information theory gradually, defining terms carefully and giving plenty of examples. Since "entropy" will come up a fair amount, let me give a first definition here (but more will come later).

Entropy is often defined as the number of bits that are necessary to express an outcome (see, for example, Cover and Joy 1991: 5). A die with its six possibilities therefore has greater entropy than a coin, which has only two. That means you'll be more uncertain about what will happen with any particular roll of a die than you will be about any flip of a coin. Entropy is formulated so that things that are more probable get shorter descriptions than things that are less probable. This is as we would want it—let's say we came up with a code because we send each other a lot of text messages but get charged per character. We'd want to make our most frequent (most probable) messages the shortest. It's really entropy that does the work in experiment two for predicting compositionality. In experiment three, entropy values also help predict whether a transitive phrasal verb will appear in the joined construction (V Prt NP) or the split construction (V NP Prt).

To examine this syntactic phenomenon, I turn to Gries (2002)'s multifactorial account of particle placement in phrasal verbs. By adding entropy 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 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.

What's ahead, section by section

I start out by outlining some of the basics of phrasal verbs. Since opaque phrasal verbs might just be idioms, I also survey the relationship between idiomaticity and compositionality. In section 3, I move on to thinking about how compositionality can be measured.³

Next, section 4 looks to morphology for analogies, since affixed words, like phrasal verbs, seem to be made up of multiple parts. In particular, I use Hay (2002)'s approach to parsability and Moscoso del Prado Martín et al (2004)'s information theoretic measurements of paradigm sizes to run three different analyses against the corpora (sections 5, 6 and 7). In section 8 I review the challenges that a traditional (generative) theory of linguistics will have based on the findings of the experiments. In section 9, I conclude with directions and predictions for future work.

³ My work here takes for granted that the use of corpus data and modern statistical modeling are important and valid resources for linguists—and this does seem to be an increasingly safe assumption. For answers to common criticisms about usage data, I recommend Bresnan et al (2007). For a thorough introduction to the statistical methods used in this paper and elsewhere, see Baayen (in press).

2. Basics of phrasal verbs and idioms

“I adopt the term phrasal verb purely for convenience” says Bolinger as he kicks off his survey of the phenomenon in English (1971: 3). I also find it a convenient and attractive term to use, though the use of “verb-particle construction” in Gries (2002) and elsewhere makes it feel a bit like a throw-back. In any case, these are the expressions that we native English speakers use to torture non-native speakers: why lock a verbal meaning in one word when you can spread it out across several? Among the most interesting consequences of using a verb and a particle (instead of a verb alone) is what happens to an NP object if the phrasal verb is transitive: the particle and the NP can alternate in their positions.

- (1) He picked up a book.
- (2) He picked a book up.

Are (1) and (2) really equally likely? Are there any conditioning factors or any patterns? Bolinger (1971) was interested in the possibilities, as were researchers before and since—though few seem either as nuanced or as discursive. To my mind, however, it is Gries (2002) that provides the most complete and satisfying analysis. By reviewing the literature on phrasal verbs, he was able to find 25 operationalizable hypotheses about what matters.⁴ Some of these had been studied empirically, but Gries combined them all in a multifactorial model that could show how the different effects contributed to the particle placement and how they interacted with one another. There is much to draw one in to Gries (2002), but I was struck by a particular factor that made it in among the top five: idiomaticity.

Determining the idiomaticity of a phrasal verb is not completely straight-forward, and unlike most of Gries’ factors it required explicit human judgment and rather coarse categorization: taking his 403-sentence sample of the British National Corpus, he coded each as having an idiomatic, metaphoric, or literal use of the phrasal verb. In the resulting models, Gries found that idiomatic phrasal verbs tended to occur in joined constructions (V Prt NP). Researchers like Bolinger (1971: 121) and Fraser (1974) had suggested this direction, but Gries (2002) was actually able to demonstrate that it mattered even when more important factors like heavy NP shift were taken into consideration.

The basic analysis, and one that the work in this paper hinges upon, is that the verb and the particle are more of a single unit in the case of idiomatic phrasal verbs.⁵ Transparent phrasal verbs, on the other hand, get their meaning from a combination of the verb’s meaning and the particle’s meaning and are therefore more loosely federated. Since the function is essentially additive and the pieces are distinct, you can put the NP before or after the particle without any penalty. In the extreme examples, you could construct a perfectly fine sentence with just the verb and the NP and then at the last millisecond add the particle to shift the meaning slightly.

- (3) She stretched the cord (out) along the table.
- (4) The factory spewed chemicals (out) that polluted the lake.

⁴ I treat each of his factors as a hypothesis, though Gries himself places all of the significant ones into an overall “Processing Hypothesis” in which “the choice of word order will serve to facilitate processing” (Gries 2002: 48). The joined construction will be preferred when the direct objects require a lot of effort, while the split construction is preferred if the direct objects don’t require such effort.

⁵ I will actually call these “opaque” phrasal verbs.

But things that are in some way idiomatic have a meaning that is more than the sum of their parts and thus these parts are less distinct, less separable. To summon up an opaque phrasal verb is to summon up the verb and the particle simultaneously since you need both to get the opaque meaning.

- (5) She's bringing up the children.
- (6) She's bringing the children up.
- (7) She's bringing the children.

You can say any of these, but you wouldn't get (6) by thinking (7) and then just tacking the particle on the end. What we probably want to say, then, is that opaque phrasal verbs are more frequently realized in the joined construction because of how they are stored in the lexicon. Almost all transitive phrasal verbs display some flexibility, but opaque phrasal verbs have a tighter relationship between the verb and the particle. Yet we don't want to lose sight of the fact that opaque phrasal verbs do actually alternate. Because of this, and because the same verbs and particles participate in both types,⁶ it isn't right to declare that opaque phrasal verbs are truly idiomatic. By looking at frequency data and examining the patterns, I will show that it is possible to measure how tight a relationship there is between a verb and a particle and how that corresponds to varying degrees of transparency/opacity.

Nearly everyone seems to accept the idea that the meaning of idioms isn't made up of their parts. Consequently, theorists believe that they have to be stored differently than other expressions that involve multiple words and use normal compositional semantics. Typically, idioms are imagined to be stored in the same way as single-word expressions like *book*, *mauve*, or *cavort* (for example, Fraser 1976: 104).

Fraser's definition of "idiom" has been adopted by many, including Gries (2002).⁷

I shall regard an idiom as a constituent or series of constituents for which the semantic interpretation is not a compositional function of the formatives of which it is composed. (Fraser 1976: 103)

As I will show, compositionality is a matter of degree, at least for phrasal verbs—and I presume, other phenomena as well. The generative tradition is, as is its wont, categorical about idioms and composition. There are plenty of people who are happy to argue whether a complex expression is or is not compositional—Fraser is part of a tradition alongside Katz and Postal (1963), Weinreich (1969; 1972), Chafe (1970), and many others.

There is no semantic information associated with the individual parts of the idiom but only a single set of markers associated with the entire idiom (Fraser 1970: 27).

Nunberg et al (1994) say that such misconceptions muddle the literature on the syntax of idioms: idioms *do* have some semantic compositionality (Nunberg et al 1994: 491). The authors divide English idioms

⁶ Consider opaque *give up* and transparent *give out*. Or opaque *trail off*, which shares a particle with transparent *lift off*. Most difficultly, *bring up* can be transparent or opaque, depending whether you mean *laundry* or *children*.

⁷ I adopt it as well, though in its weakest possible sense: you can't predict the meaning of an idiom just by knowing the meaning of the parts when they are used in isolation. This is inextricably tied not just to the concept of compositionality, then, but also to that of conventionality. See "Appendix E: What else is an idiom besides non-compositional?" for a rough outline about this and other factors that may define idiomaticity.

into two basic classes—idiomatic phrases (*kick the bucket, saw logs*), which really can't be decomposed and idiomatically combining expressions (*take advantage, pull strings*), where a language user can assign some sort of meaning to the parts. In this they are making a claim that constructions like the passive require some sort of meaning for the parts. It follows that you can't get (8) because idiomatic phrases can't be decomposed into meaningful parts. You can get (9), since it is an idiomatically combining expression and its parts do carry some meaning.

(8) *The bucket was really kicked by James Dean that night in his racecar.

(9) Those strings were pulled by people higher up the food chain than I am.

Note that even the most idiomatic of phrasal verbs fails to be an idiomatic phrase in these terms. Not only can they appear in passives and the like, but it is always possible to assign components of meaning to the verb and the particle.⁸

Wood (1986) sees compositionality as gradient, “shading in degrees from utterly opaque to the fully predictable” (Wood 1986: 2). However, she still wants to carve out a special place for idioms where only the subset of truly non-decomposable expressions is to be considered idiomatic. My approach takes essentially the same observations, but draws no boundaries. I want to say that opaque phrasal verbs get some support from their components—just not very much. Importantly, the measurements I use predict that compositionality has degrees (the verb and particle of a phrasal verb) can be closer to their core meanings or farther away. The relationship can be so attenuated, in fact, that most of the meaning comes from the combination and not from the pieces themselves. Those are the cases we call most opaque.

⁸ McCarthy et al (2003) rated 116 phrasal verbs for their transparency/opacity. This was only done by three linguists so the results are suspect in scientific terms, but consider the seven phrasal verbs that had the lowest average score for transparency: *cock up, rack up, whip off, write off, space out, clam up, stave off*. In each case, once an English speaker has deduced the meaning from contextual clues, the verb and the particle can be assigned interpretations. This makes even more sense if we consider that opaque phrasal verbs have antecedents in transparent ones, which have blossomed through a process of figuration. In many cases, the words that combine idiomatically can also combine transparently (to repeat the example from a few footnotes ago, you can *bring up the children* or *the laundry*). In Nunberg et al's terms, “the meanings of idiom chunks”—verbs and particles here—“are not their literal meanings. Rather, idiomatic meanings are generally derived from literal meanings in conventionalized, but not entirely arbitrary, ways” (1994: 503). Opaque phrasal verbs aren't idiosyncratic anomalies; they are tied to other uses of the verbs and other uses of the particles in patterns that are complex but which can be modeled with information theory, as I will show in upcoming sections.

3. Measuring compositionality

In the last section, I suggested that something like (10) would be an example of an opaque phrasal verb because knowing the central meanings of *bring* and *up* gets you no closer to understanding what *bring up* means.

- (10) I brought up my children to respect their elders.
- (11) I lifted up the marriage certificate so everyone could see it.

By contrast, (11) involves a very traditional use of both *lift* and *up*.⁹

In making these claims, I have appealed to your intuitions about English, but let me try to be a little more specific. Lohse et al (2004) use entailments as a measure of compositionality and this seems like a relatively useful place to begin. Thus, we would reformulate the preceding paragraph to say that neither *bring* nor *up* are entailed in *bring up* but that both *lift* and *up* are entailed in *lift up*.

Lohse et al (2004) are interested in an idea of “minimizing lexical dependency domains”, which really means that if things depend on each other, you want to keep them close together. When neither the verb nor the particle is entailed, they are said to depend on each other, whereas when both are entailed they are said to be independent. Hawkins proposed this idea to Lohse and Wasow and they took it up, though they never actually motivated it (Wasow p.c.). This is still about intuition, as most semantic notions are, but it is a little plainer to operationalize and a little more precise than a gut reaction about what is compositional and what isn't.¹⁰

Bannard (2002) uses entailment in a similar way:

It is still not trivial for a human judge to decide whether any given verb or particle is contributing simplex meaning. The criteria we chose to use in deciding this are essentially whether the semantics of the simplex verb or particle can be used to helpfully decompose the construction. This comes down to a question of entailment, and whether we can say that the sentence involving the [phrasal verb] entails certain statements involving the simplex verb or particle. If we can say that the statements involving the verb or the particle are entailed, then we can say that the item has standard semantics. (Bannard 2002: 8)

In both cases, entailment describes the relationship between the parts and the whole, which is a description of compositionality, as Bannard says. Let's look at the various combinations of entailment characteristics.

Imagine that you have verb_x and particle_y; generally verb_x has meaning_x and particle_y has meaning_y, these are what various authors call the “simplex” meanings and which can be thought of as the meaning that the word has “most of the time”. Now imagine the phrasal verb that can be formed with verb_x and particle_y.

⁹ This observation is also at the heart of Gries' classification system. See “Appendix B: Gries' definitions for IDIOM” for a complete description of his classification scheme.

¹⁰ This definition of an entailment test also resembles the “Is a”-condition that Allan (1978) uses in looking at compounds. *Lifting up* is a type of lifting, but *bringing up* is not a type of bringing.

We will say that the combination of these gives us meaning_z. Compositionality essentially marks the relationships between how much of x and y go into giving us z .

- $z=x+y$, that is, the phrase is really just meaning_x + meaning_y, as in *lift up*.
- $z=x+y+n$, that is, our phrase definitely contains meaning_x + meaning_y, but it also contains something extra.¹¹ It is conceivable that our previous expression ($z=x+y$) never really exists because there is *some* sort of additional meaning whenever you habitually pair words together. The n value could be infinitesimal or it could be large, making the combination of x and y more or less compositional.
- $z=(x/y)+n$, that is, the meaning of only one word (x or y) is part of the meaning of the whole. In principle, every word contributes *something* to meaning, so this really just expresses the case when either the verb or the particle seems to be contributing almost nothing.
- $z=n$, that is, the meaning doesn't seem to have anything to do with the components. Again, what's really happening is that meaning_x + meaning_y are contributing nothing to the meaning, so we're just simplifying our core expression: $z=x+y+n$.

It becomes rather obvious that the real calculation is $z=x+y+n$, but that the size of the contributions can range from zero on up for each variable. Entailment, as I will use it, will be rather ham-fisted about this. Lohse et al (2004) acknowledge the same fact:

Our entailment tests for verbs and particles force a binary coding, but one has to assume that verb and particle dependency is ultimately a graded concept, for which a binary opposition can only be an approximation. (Lohse et al 2004: 255)

Lohse et al (2004)'s idea of verb and particle dependency suggest that the relationship of $z=x+y+n$ itself may not capture a more complicated fact. To use entailment tests is always to assume that words have some sort of core meaning, but we also know that words often have multiple meanings and that their meanings shift as they are used alongside other words. A useful metaphor here is a network: the different senses of a word are connected to each other and they are connected to other words they have co-occurred with, too. At this point, however, it will suffice to say that entailment tests—despite their limitations—are a rather good approximation of “very compositional” and “very noncompositional”.

If we accept that words do have core meanings and see compositionality as representing how close a use is to the core, then to say that neither the verb nor the particle are entailed is to reduce $z=0+0+n$, where n comprises the meaning. To say that both the verb and the particle are entailed is to give proportionally less room for n to contribute to the meaning. We cannot control for the size of n , but if we think of phrasal verbs as having approximately the same amount of meaning (z), then n will always be bigger when a phrasal verb is opaque (fully unentailed, non-decomposable) than when it is transparent (fully entailed, compositional): For example, given a constant z , then $n_1 > n_2$ if $z=0+0+n_1$ and $z=10+10+n_2$.¹²

¹¹ The immediate example I have in my mind is “bacon and eggs”, which in a diner menu really does mean “bacon” and “eggs”, but also means that both are fried (at least most of the time).

¹² In the middle reaches, where either the verb or the particle is entailed but not both, we can't control for n at all. Thus in this paper I restrict myself to the extremes of fully entailed and fully unentailed phrasal verbs.

What we are left with, interestingly, is the idea that decomposable expressions have little extra meaning beyond their components, but that non-decomposable expressions have essentially nothing *but* extra meaning. In looking at compositionality we are most concerned with n and what happens when it is big and when it is small. This concept is obviously semantic in nature, yet it has consequences for syntax, too, as we'll see later.

But how do we distinguish meaning_x , meaning_y , and meaning_n so that we can have a measure of compositionality? Information theory gives us tools to assess these; generalized linear mixed-effect models give us the means of determining the validity. In the next sections, I build towards information theory by first laying the groundwork for paradigmatic approaches to linguistics. Information theory is coherent with linguistics to the extent that we see experiences with language as creating a vast network of relationships whose patterns can be measured and compared. Such relationships are built out of frequency data, which especially makes sense if we take the reasonable position that frequent uses of a word are more likely to be central to its semantic than infrequent ones. The information theoretic measurements I will employ are novel, but they are consistent with other lines of research that have shown that usage factors like frequency do affect compositionality (Bybee and Scheibman 1999, Haiman 1994, Hay 2001).

4. Analogies to morphology

Composition is essential to the field of morphology, which investigates how words are related to one another and how those relationships create meaning. In the traditional morpheme-based approach, it's easy to see how this works: you posit that morphemes are the smallest pieces in a language that have meaning and then create rules for combining them. *Dogs* is decomposable because we can break it into *dog+s*, but *dog* itself cannot be broken down into any other semantically meaningful pieces.

In phrasal verbs, we seem to see a range of compositionality, from the opaque *give up* to the transparent *move up*. What is striking is that *up* is the same in both expressions and that there are transparent phrasal verbs with *give* (*give out*). There is something strange going on; it isn't the parts qua parts that make the difference, but the relationship between them.

Morpheme-based approaches are not the only view of morphology and indeed, I think that the paradigmatic approaches developed by authors like Hay, Baayen, Ackerman, Blevins, and Moscoso del Prado Martín offer a richer analogy for our phrasal verbs precisely because they are more focused on the patterns of relationships between whole words.

Morpheme-based approaches use deterministic rules to put meaningful parts together and when a form is irregular, they simply say that the form is stored in memory and falls outside the rules (thus we generally add *-ed* to verbs to get the past tense, but for *bring* we just memorize the fact that the past tense is *brought*).

Alternatively, paradigmatic approaches are based on complete words, not morphemes. Though Stump's paradigmatic morphology is not probabilistic in nature, the theories are pretty amenable to probabilities:¹³

¹³ The paradigmatic view is bolstered by evidence from psycholinguistic experiments in which people are consistent in describing "more and less affixed" words and in which they have different reading time responses based on where a word falls on that scale (for example, Hay 2001).

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)¹⁴

A paradigmatic approach allows both the storage of full words, the storage of relationships between them, and the “parts of complex wholes”, too. Thus, not all words contain affixes the same way—*government* is less complex than *discernment*, for example. That is, even though both of these words include the suffix *-ment*, people see the parts of *discernment* (*discern+ment*), but they don’t see the parts of *government* (*govern+ment*). English speakers rate the complexity of these words differently and in speech production, the simpler form has fewer phonetic cues that there’s a juncture right before the affix (Hay 2000; 2001). Instead of basing complexity on the number of morphemes that linguists can count, complexity is defined as being about what is perceptible to language users—words with multiple morphemes can be simple.

But how do we decide if language users treat words as simple or complex? Frequency itself is too crude, says Hay (2002: 530), and others follow suit. Their models of the paradigms are built using the following types of data:

- Relative frequencies of bases and affixed forms
- Token-based families of forms
- Type-based families of forms

The “family size” of *x* is based on the number of things that *x* appears in. For phrasal verbs, the family size of a verb is the number of particles it combines with. It is possible, as in the next section, to limit a family to the things that are likely to be seen as parts of it—for example, since *government* is more frequent than *govern*, we may not want to put *government* into the *-ment* family. “Tokens” are individual occurrences and “types” are the abstract aggregations of these. Thus if all you have in your corpus is 10 instances of *discernment* and 50 instances of *indictment*, you have 60 tokens but only two types.

You can build a number of relationships between these numbers, but one of the simplest and least redundant methods involves measures from information theory. The goal is to have a measure of how much information is contained by a word and how much is carried by the paradigms it is a part of. These relationships seem to have psychological reality since, for example, “The amount of information carried by an inflected noun form is inversely proportional to its processing latency” in psycholinguistic tests like word-recognition (Moscoso del Prado Martín et al 2004: 4).

Before getting into these more complicated measurements, let’s start with a more straight-forward analogy between word occurrence patterns and parsability—the degree to which an expression seems composed of bits of meaning (*dog+s*, *discern+ment*, *move+up*) rather than having its bits recede into the whole (*dog*, *government*, *give up*).

¹⁴ Thus *up* simply isn’t as present in *bring up* as it is in *lift up*. What isn’t “as present” can’t really be entailed.

5. Experiment one: opaque and transparent phrasal verbs are different

Hay (2002) uses a variety of measures to investigate affix ordering. She believes that looking at some of the type and token relationships I've described can help explain the old observation that some affixes, like *-ic* and *-ity* seem to be less capable of affixing than other affixes like *-ness* and *-less*. These are traditionally called "level 1" and "level 2" affixes, and it is generally the case that you can't affix a level 1 affix outside of a level 2 affix. The idea is that decomposability is related to such ordering and that level 1 affixes tend to be semantically opaque.¹⁵

Hay (2002) offers six measures that show how the difference in parsing rates corresponds to the difference in productivity. To create the Table 1, Hay 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". Nevertheless, Hay does keep track of the total number of different words that the affix is part of (regardless of whether they are likely to be parsed), and is able to get a "type-parsing ratio" by dividing the "number of types parsed" by the total number of affixed words. This can be done for both types (each word counts once) and tokens (so that more frequent words are weighted higher than less frequent ones).¹⁶

Consider an affix that shows up in a corpus 1,000 times with 100 different bases. That means you have 100 affixed word-types and 1,000 affixed tokens. If all of the affixed forms are more common than base forms, then none of them are parsed for tokens or types. The 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.

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

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

Table 1: Hay (2002) summary of affix ordering

Do phrasal verbs have the same relationships between frequency and compositionality as affixes? If so, then we may have a way of determining how distinguishable the verb and the particle are from the phrasal

¹⁵ See, for example, Siegel (1979), Aronoff (1976), Kiparsky (1982), and/or Giegerich (1999) for a summary.

¹⁶ Note that I will not be reporting hapaxes, which are single-word occurrences. These are also part of the productivity value given by Baayen and Lieber (1991), so I won't be reporting that, either. The Bannard (2002) data that's being tested tends to have common phrasal verbs, so we don't really have evidence for these measures. Future work should look at what the "one off" phrasal verbs reveal. There are 162 verbs in my corpus that occur only one time and with only one particle. They include such things as *battle on*, *chip away*, *kill off*, *rig up*, and *yell out*.

verb as a whole. I extract 789 different phrasal verbs from the BNC (3,190 tokens), as well as all tokens of their base verbs. What I am going to build towards is a table just like Hay's Table 1 above (2002: 535), but I expect to have "opaque/fully unentailed" listed in place of the "Level 1 affixes" column and "transparent/fully entailed" for the "Level 2 affixes" column. That is, transparent phrasal verbs will be more obviously made up of parts than opaque ones: when you are entailed, you are separable; when you aren't entailed you melt into the whole phrasal verb.

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. For example, there are four examples of *shrug* in the portion of the corpus I extract, but three of them are actually *shrug off*. That means that *shrug off* will not participate in either the "average number of types parsed" or the "average number of tokens parsed" since it is not parsed: the combination is more frequent than the base.¹⁷ There are 29 examples of *kick*, on the other hand, but only 7 of these are in *kick off*. Therefore, under Hay's formulation, *kick off* gets parsed.

For each of the phrasal verbs that is parsed, I add up how many different "types" there are—this means adding up the number of different particles that they take. *Shrug off* has been ruled out, so we don't calculate anything for it. 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. Bannard (2002) gives the entailment characteristics of 180 phrasal verbs, and I perform a Hay-like analysis on the 124 that are either fully entailed (transparent) or fully unentailed (opaque).¹⁸

To determine the "average type-parsing ratio" I simply divide the number of parsed types by the total number. For *shrug*, the result is $0/1=0$; for *kick* it is $5/5=1$. 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. Finally, I average all of the transparent parsing ratios and the opaque ones to see if there's a meaningful difference.

Token-parsing and token-parsing ratios operate about the same way, except with tokens instead of types. Thus the token-parsing count for *wind* is three (the three from *wind down*—since the others weren't parsed). The token-parsing ratio is $3/14=0.2143$.¹⁹

¹⁷ Actually, *shrug off* still does participate—in the form of the denominators for the numbers below. The parsed forms are always divided by all of the forms (parsed and unparsed). So the data from *shrug off* is not lost; the fact that it isn't parsed is reflected in the fact that its token/type counts will enlarge the denominator but not the numerator. For each form that isn't parsed, then, the parsing ratios go down.

¹⁸ For details about Bannard (2002) and his coding, please see "Appendix A".

¹⁹ The denominator is 14 and not 18 because we only care about the tokens that are participating in some phrasal verb construction.

Here, then, is the table that matches Hay (2002):

	Opaque/fully unentailed	Transparent/fully entailed	Significance of difference (by Wilcoxon test)
Avg number of types parsed	2.49	5.52	p=3.756e-06
Avg type-parsing ratio	0.704	0.957	p=0.003360
Avg number of tokens parsed	18.10	33.48	p=0.001561
Avg token-parsing ratio	0.68	0.96	p=0.003363

Table 2: Transparent and opaque phrasal verbs behave differently, along the lines of Hay (2002)’s investigation of affixes.

For each row, it is the transparent column that has the higher value—just as in Hay (2002)’s table, where it’s the more decomposable/parsable level 2 affixes that score higher. The generalization that emerged for Hay was that generalizations about affix ordering can be reduced to perception: affixes that are easily parsed shouldn’t occur inside affixes that aren’t easy to parse. Building on Hay and Baayen (2002) as well, Hay’s prediction is that opaque affixes (Level 1) will be less productive, while 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 *lift up* take more advantage of the “actually syntactic” property of flexible alternation between NP objects and particles.

Note that the parsing calculations fit with the general pattern that verbs in opaque phrasal verbs take fewer particles than the transparent ones. The averages are: opaque=1.49 particles, transparent=2.03. The difference between these is significant (p=0.0204).

Of course it remains possible that the differences in Table 2 aren’t statistically significant—the ratios between the numbers in Hay (2002)’s columns is greater than the ratios in my table. The last column in Table 2 discloses that each measurement is significant, but I will walk through how these were calculated.

The first question is whether there are really two different groups. Is it reasonable to explain parsing numbers and ratios in terms of our notions of complete entailment and unentailment? Figure 1 and Figure 2 show that phrasal verbs do break into two groups with different distributions, regardless of whether you use token- or type-based measures. These figures plot the “density” of the distributions, measuring the relative probability of “getting a value close to x”.²⁰

²⁰ See Dalgaard (2002) for more about density and other statistical functions.

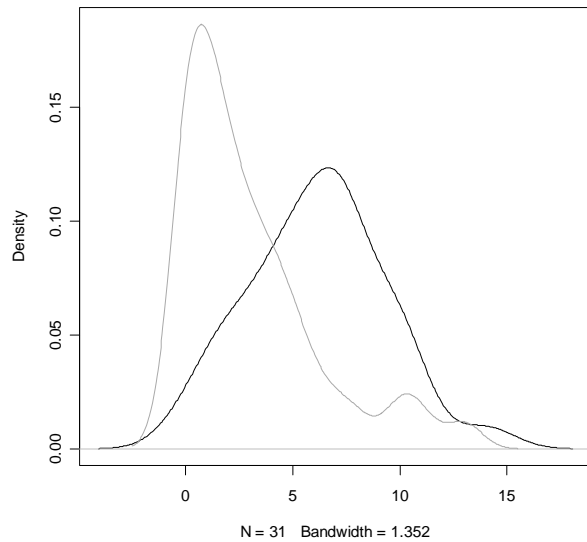


Figure 1: The “Type parsed” density plots for transparent (black) and opaque (grey) phrasal verbs; they have significantly different distributions ($p=3.756e-06$).

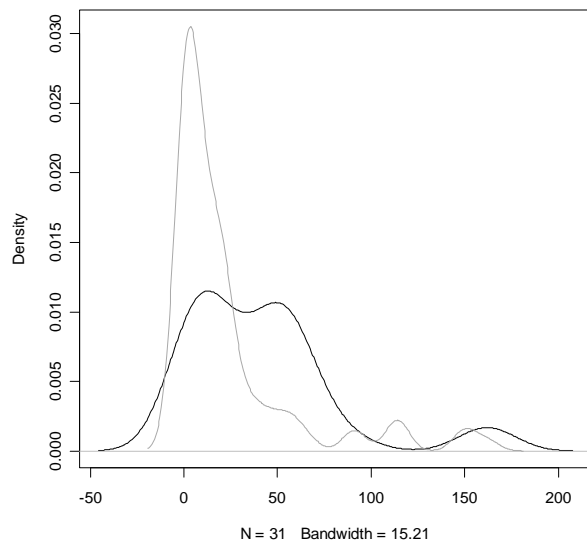


Figure 2: The “Token parsed” density plots for transparent (black) and opaque (grey) phrasal verbs; they also have significantly different distributions ($p=0.001561$).

As is often the case with language data, the numbers in the data are not normally distributed (you can guess that from Figure 2 especially).²¹ Most differences in means are assessed using t-tests, but because

²¹ A Shapiro-Wilk normality test shows that the chance of these measures occurring in normal distribution is highly unlikely: Types parsed ($W=0.8989$, $p=2.529e-07$), Type-parsing ratio ($W=0.5228$, $p<2.2e-16$), Tokens parsed ($W=0.726$, $p=2.031e-13$), Token-parsing ratio ($W=0.5265$, $p<2.2e-16$). Kolmogorov-Smirnov tests for normal

the data is skewed away from a normal distribution, this isn't appropriate. Instead, I use a Wilcoxon test. This test will be more conservative in its p-values, that is, it will give us larger p-values than a t-test would. Despite the conservatism, in each case the means of the transparent phrasal verbs are significantly different than those of the opaque phrasal verbs: Types parsed ($p=3.756e-06$), Type-parsing ratio ($p=0.003360$), Tokens parsed ($p=0.001561$), Token-parsing ratio ($p=0.003363$).

Some readers will recall that 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. While that seems like a reasonable rule-of-thumb for him, it might make us question the averages calculated here since we know that some phrasal verbs can be used opaquely and transparently (like *bring up*). To test whether this is a problem, I recalculated Table 2 while treating 20% of each of the opaque phrasal verbs as transparent and 20% of each of the transparent phrasal verbs as opaque. In other words, even though many of Bannard's phrasal verbs had entailment characteristics consistently higher than 80%, I chose to be conservative for all of them. Having done this, only the "average number of tokens parsed" becomes insignificant ($p=0.07455$). By the three other measures, transparent phrasal verbs are still significantly different from opaque phrasal verbs—Types parsed ($p=3.343e-05$), Type-parsing ratio ($p=0.0003102$), and Token-parsing ratio ($p=0.003490$). Both our division of the data into two groups and our use of Hay-like parsing measurements are supported statistically even when we account for problems of homonymy.

Discussion

Let me first acknowledge, as Hay has not, the possibility that there is some vast conspiracy of Free Masons or Bilderbergs that may have surreptitiously fixed the data to give nice results not only for her affixes but for my phrasal verbs, too. Looking at different phenomena from different corpora, we both found that corpus frequencies gave us a meaningful way of characterizing connections between components that carry meaning. Though I use the term compositionality, I am essentially showing what she calls "parsability".

The first generalizations about affix ordering used two buckets, Level 1 and Level 2, to explain which affixes could appear inside of which other affixes. Fabb (1988) showed that this approach generates many more combinations than actually happen. For that reason, he suggested that affix ordering is about selectional restrictions. Hay (2002) argues that this approach misses important generalizations about the sensitivity affixes have to the internal structure of their potential bases (its parsability).

Hay (2002) leads us away from the basic tenets of generative morphology where there are structures and rules to compose words. In using parsability to account for affix ordering, we need to store more information than traditional theories have called for—parsability is driven by perceptability, which is a property that emerges from experiences with language. If the lexicon only stores bases and affixes (*govern*, *discern*, *-ment*), it won't be easy to account for the difference in parsability between *government* and *discernment*. If we put all the parts into the lexicon (*govern*, *discern*, *-ment*, *government*, *discernment*), we will still have to keep information about relative frequency if we want to model Hay's generalizations and account for how different people, at different times will find different words more or less complex.

distribution confirm that these aren't normally distributed. (Tokens parsed: $D = 0.2264$, $p\text{-value} = 1.374e-05$; Types parsed: $D=0.1467$, $p=0.01355$).

Like affixes, we can say that there are different degrees of perceptibility in phrasal verbs and that it's the transparent phrasal verbs that are more obviously composed of pieces. Being parsable allows the verb and the particle to each carry meaning in a way that isn't possible for the opaque items. There is a little bit of a chicken-and-the-egg problem, but those phrasal verbs that have no entailment are not composed of distinct pieces. Without distinct pieces, it is the whole that carries meaning, not the parts.

This section has looked at relatively small families made up of individual verbs that take different particles. The next section expands to a much wider notion of family, allowing phrasal verbs to be situated in several paradigms simultaneously.

6. Experiment two: Information theoretic terms predict entailment

I ended section 3 on measuring compositionality with the question, "But how do we distinguish meaning_x, meaning_y, and meaning_n so that we can have a measure of compositionality?" This section may provide the answer. It builds off of the notion that we should be able to separate out the components of meaning (along the lines of our earlier equation, $z=x+y+n$) and measure their contributions. The prediction, which is borne out, is that opaque phrasal verbs get less meaning from their parts and more meaning from "something else"—the combination itself.

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? They apply their measure of "information residual" to a variety of data and get significantly better results than with previous measures. I have adopted "informational residual" as a measure of compositionality and also found significant results.

In information theory, it is possible to relate surface frequency and amounts of information. (12) gives the minimum amount of information necessary to encode something optimally in binary—for example, a word if all words in the lexicon were listed with their frequencies. The amount of information tells us how hard it is to recognize a word all on its own. Because of the way this is formulated, the more frequent a word is, the less information it is said to have. A word that is 90% of all tokens would have an information load of 0.152 while a word that is only 1% of tokens would have an information load of 6.64.²²

$$(12) \quad I(x) = -\log_2(\text{frequency of } x / \text{size of the corpus})$$

Since x may fit in to a variety of derivational or inflectional paradigms, the idea is that you will subtract the joint entropy of the various paradigms from the $I(x)$ value and get the amount of residual information that the paradigms can't account for, this is the n of "extra meaning" that I spoke about in section 3. For phrasal verbs, the joint entropy is what I called $x+y$ and is made up of a verbal entropy score and a particle entropy score.

We're back to the word "entropy". In the overview section, I gave a first definition—entropy is the number of bits that are necessary to express an outcome: the greater the number of outcomes, the greater

²² In Moscoso del Prado Martín et al (2004), this is crucially related to lexical decision tasks—the greater the amount of information contained by a word, the more costly it will be to recognize.

the entropy.²³ Here, there are more outcomes possible for exactly the phrasal verbs that have the largest number of paradigm members and get the most support from those paradigms: the transparent phrasal verbs. These transparent phrasal verbs are the most flexible, productive, and intelligible. But 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.

Lower entropy values mean less uncertainty. In my earlier terms, an opaque phrasal verb like *rack up* (joint entropy= 3.416) is a tighter unit than something transparent like *move up* (joint entropy= 10.08). Entropy captures this tightness as a measure of predictability: *up* is more predictable after *rack* than it is after *move*. If a usurious phone company was charging us per character for our text messages, we'd want our code to use fewer characters with *rack up* than with *move up*.

Our task in this section is to compare how much information phrasal verbs get from their parts compared to the whole. Again, to do this, we're going to take the amount of information ($I(x)$) for each phrasal verb and subtract the entropy values for its parts (the verbs and the particles). We expect that opaque phrasal verbs will get more of their meaning from the whole than the parts, so that for a given amount of information, opaque phrasal verbs will have lower entropy values (less support from their paradigms) and higher information residual scores.

This gets rather complicated, so I'll try to break it down. In the case of *drum up*, there are 65 tokens in the corpus of 310,941 phrasal verbs.²⁴ Thus if we restrict ourselves to this corner of the world, the amount of information is $I(x) = -\log_2(65/310,941) = 12.22$.

Next, I'll pretend that our phrasal verbs are comprised of two different paradigms, the verb paradigm and the particle paradigm. To calculate the entropy of any paradigm, we use the following equation. $F(x)$ is the frequency of x ; $F(P)$ is the frequency of the whole paradigm.

$$H(\mathcal{P}) = - \sum_{x \in \mathcal{P}} p(x|\mathcal{P}) \log_2 p(x|\mathcal{P}) \approx - \sum_{x \in \mathcal{P}} \frac{F(x)}{F(\mathcal{P})} \log_2 \frac{F(x)}{F(\mathcal{P})},$$

(13)

Drum up participates in two different sets of patterns, so we'll have to calculate two entropies. The verbal entropy describes the tendency of all phrasal verbs that have *drum* in them. There are a total of 66 *drum*'s in the phrasal verb corpus, 65 of them are with *drum up* and one is with *drum out*.

²³ In fact, entropy is also sensitive to how even the probabilities are for the outcomes. So if you had two different 32-sided dice and one of them was truly fair, its entropy would be 5 (the equation for entropy will be given in a few paragraphs). If your other die had 32 sides but one of them was weighted heavily, it would have a smaller entropy value, relatively speaking. These die have the same number of outcomes but they don't have the same distributions of probabilities. If the second die turns up "7" 50% of the time, even if the rest of the sides are equal amongst themselves, the entropy for the die will be 4.7. Literal phrasal verbs have less variance in the values for their token counts and the number of different particles the verbs combine with.

²⁴ Here, I restrict myself to examples of phrasal verbs because that's the paradigm in question. I use a corpus that Baldwin and Villavicencio (2002) developed out of the BNC. Their title gives you a hint about why I have chosen to use their data here rather than my own: "Extracting the Unextractable: A Case Study on Verb-particles". My own search for phrasal verbs relied upon the parsed part of the BNC, which is rather smaller. When calculating accurate measures for paradigms, the more data, the better. The Baldwin and Villavicencio (2002) data was used with minimal editing: strays like *that out* were removed and some variations like *git out* and *get out* were collapsed. All told, my edits affected only 812 tokens total, less than 0.3% of their data.

- Drum out: $-(1/66)*\log_2(1/66)= 0.0002324$
- Drum up: $-(65/66)*\log_2(65/66)= 0.008382$

Now we sum these together to get a *drum* entropy of 0.008614. We do the same thing with *up*, though there are actually 1,110 different verbs that go with it. After all of the *up* verbs are added together, we get a particle paradigm entropy value of 7.089. To calculate the information residual, we subtract these values from the amount of information. In this case, the information residual for *drum up* is $12.22-(0.008614+7.089)=5.1222$.

In fact, things are still a little more complicated. This process for calculating information residual works for nested paradigms. In Moscoso del Prado Martín et al (2004), the example is *thinkers*, which is nested inside *thinker*, which in turn is nested inside *think*. The calculation for a compound like *think-tank*, however, can't be calculated by simply adding the *think* paradigm and the *tank* paradigm together: "the members are mutually exclusive except for [think-tank] itself" (Moscoso del Prado Martín et al 2004: 10).

Similarly, our phrasal verbs have two direct, distinct ancestors. We therefore need to consider the union of the verb and particle paradigms as a single random variable. The exact steps are:

- For each phrasal verb, look up the total number of occurrences for both the verb and the particle. Add these together. *Drum*: $66+ up: 79,236=79,302$.
- Subtract the total number of occurrences of the phrasal verb from this to get the actual union of the two paradigms. $79,302-drum up:65=79,237$.
- Use the equation in (13) to calculate two separate scores, one for the verb and one for the particle: *drum*: $-(66/79,237)*\log_2(66/79,237)=0.008521$; *up*: $-(79,236/79,237)*\log_2(79,535/79,237)=1.821e-05$.
- 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*, so 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.
- Now we subtract this from the $I(drum up)$ score (which hasn't changed) and get an information residual of $12.22-3.432=8.788$.
- We can use the same steps to look at the impact of types.²⁵

Having calculated these measures across the complete corpus, I pull out the values for the Bannard test data to see if the opaque and transparent phrasal verbs are actually any different from one another in these terms. It turns out they are.

The token-based information residual scores for opaque phrasal verbs are reliably higher than that of transparent phrasal verbs ($p=2.49e-06$).²⁶ This is because both the verb entropy and particle entropies are

²⁵ In this case, all phrasal verbs have the same $I(x)$ value of $-\log_2(1/6,793)=12.73$, since each individual phrasal verb is a single type.

²⁶ Throughout this section, I report p-values based on generalized linear models built using transparency/opacity as a response to be predicted by the information theoretic measurement in question. Using the R system for statistical computation and graphics, the models take the basic form of, for example, "token.glm<-glm(ir\$Transp ~ ir\$Resid_token, family="binomial")".

bigger for transparent phrasal verbs and since these are subtracted from a relatively stable $I(x)$ amount, the bigger the entropies, the smaller the information residual.²⁷

- *Bog down* (opaque) scores 3.225 in token-based information residual
- *Walk in* (transparent) scores -5.366 in token-based information residual

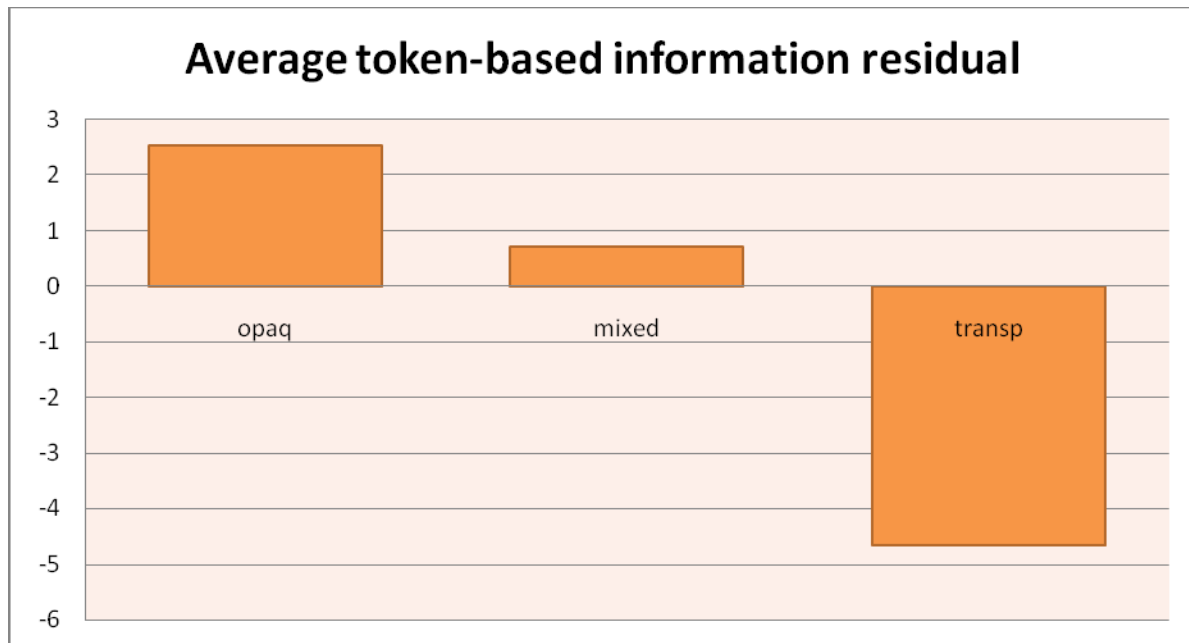


Figure 3: Opaque phrasal verbs have the highest information residual, transparent ones have the smallest. Phrasal verbs that are partly entailed (either the verb or the particle, but not both) are right in the middle.

The same thing happens in type-based analyses: the informational residual scores for opaque phrasal verbs are higher than transparents ($p=0.01530$) and it's also the case that the type entropies for both the verbs and the particles are higher for transparents than opaques.²⁸

- *Touch off* (opaque) scores 0.8804 in type-based information residual
- *Bring out* (transparent) scores -0.8517 in type-based information residual

Earlier, I gave the equation $z=x+y+n$, where z was the amount of meaning I held constant. A statistical test confirms that z (the amount of meaning otherwise known as $I(x)$) doesn't change significantly based on whether or not a phrasal verb is opaque ($p=0.0625$). This is what we expect and even hope. If this weren't the case, I'd have to say that one type of phrasal verb encoded more information than the other—that it in some way “meant more”. It would be nice if the significance value here wasn't lurking close to

²⁷ Despite the fact that we're considering the union of particles and verbs, we can sort out the patterns. The verb token entropy for transparents is significantly higher than that of opaques ($p=3.07e-06$), and the particle token entropy is higher, too ($p=3.81e-05$).

²⁸ The verb type entropy is significantly higher for transparents than opaques ($p=0.000434$), but the particle type entropy is not ($p=0.334$). This isn't terribly surprising since we expect most of the semantics to come from the verbs and we know that there's a lot more variability in verb types than in particle types.

0.05; the fact that it is further encourages follow-up studies that will test a greater number of phrasal verbs.

Expanding to all verbs and all particles

In these equations notice that I have limited myself to the tokens of verbs and particles that occur in phrasal verbs. It is also possible to expand to include all verb and particle tokens regardless of whether they are in phrasal verbs. This will not change any of the type-based measurements, but it will alter the results for the token-based measurements. The first part of that is particularly easy—extracting all of the verb lexemes from the BNC will have us consider 14,723,316 verb tokens (310,941 of which are participating in phrasal verbs).

It's a little harder to count particles since these are notoriously difficult for parsers to get right (see, for example, Toutanova and Manning 2000). The BNC is tagged with “adverbial particles”, which the Reference Guide for the British National Corpus describes as “preposition-type words that don't have complements”.²⁹ Thus both of the *out*'s in the following sentences are tagged as adverbial particles, even though only (15) is a phrasal verb.

(14) Come out here.

(15) I can't hold out any longer.

If we consider only the things that are tagged in the corpus as “adverbial particles”, we'll end up including both (14) and (15). That may be alright, depending on what we think the “particle paradigm” is. Unfortunately, if we only look at the BNC's adverbial particles, we will end up with a lot of holes—of the 48 particles that Baldwin and Villavicencio (2002) found occurring in phrasal verbs, only 15 particles are marked as adverbial particles in the BNC.

Moreover, the BNC only has 13 uses of *across* marked as adverbial particles, but the Baldwin and Villavicencio (2002) data found 605 uses of *across* in phrasal verbs. This also happens with *by*, which the BNC has 371 tokens of as an adverbial particle, but which Baldwin and Villavicencio (2002) have 621 examples of in phrasal verbs. Either the BNC parsing or the Baldwin and Villavicencio (2002) parsing could be wrong here, but finding the source of the problem is well beyond the scope of this paper. I will presume that (i) the Baldwin and Villavicencio (2002) tagging is more accurate than that in the BNC, and (ii) restricting the particle paradigm to actual particles (not including, say, prepositional homonyms) is a reasonable way to proceed.³⁰

²⁹ See <http://www.natcorp.ox.ac.uk/XMLedition/URG/posguide.html#m2prep>. Note that the BNC also tags words that are ambiguous between adverbial particle and preposition. There are 346,095 of these (there are only 656,784 words that are unambiguously tagged as adverbial particles. Comparing these two numbers illustrates some of the difficulty of parsing particles).

³⁰ Further research needs to be done here to define what should count as part of the particle paradigm. What Hay (2002) and experiment one have shown us is that related words (like homonyms) may not exert equal, consistent influence on each other. Ideally, we might want to include all homonyms but weight them by something like part of speech so that similar uses of a word influence each other more. In such a calculation, “part of speech” itself would be a rather coarse grouping since we'd expect even within a category, some items would be more alike than others.

The table below describes significance scores for information residuals for a variety of particle definitions. Based on the assumptions just described, it would appear that using all of the BNC verb lexemes but then only the Baldwin and Villavicencio (2002) particles makes the most sense (the last column in the table). In this case (as with almost any definition of ‘particle’), information residual is still a significant predictor of entailment ($p=0.00195$). So are the particle entropy ($p=0.004724$) and verb entropy ($p=0.00114$) that make it up.

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

Table 3: Information theoretic measurements predict entailment characteristics for Bannard (2002). Significance codes: $p<0.001$: ‘*’ ; $0.001<p<0.01$: ‘**’; $0.01<p<0.05$: ‘*’; $p<0.1$: ‘.’**

Discussion

The standard interpretation is that the higher the entropy of a paradigm, the more information and facilitation the words receive from the paradigm. In morphology, this corresponds to shorter response latencies in psycholinguistic tests (Moscoso del Prado Martín et al 2004). In our case, the high entropies are associated with the transparent phrasal verbs. This makes sense: if you’re a transparent phrasal verb, you get your meaning from the combination of the parts and these parts have a greater ability to be parsed and then combine.

Transparent phrasal verbs have verbs that combine with more particles and particles that combine with more verbs. This means that transparent phrasal verbs have more possible outcomes than opaque ones and therefore higher entropy values.³¹ The basic equation is still “information residual=amount of information

³¹ The greater the number of members in a paradigm, the greater the entropy will tend to be. Generally, if you increase the number of members, you decrease the probability of each occurring. That, in turn, means you have to increase the number of bits required to represent them. A die has greater entropy than a coin; a literal phrasal verb has greater entropy than an opaque one.

- joint entropy of the verb and particle”— a refinement and reordering of the section 3 equation “ $z=x+y+n$ ” into $n=z-(x+y)$. Because both types of phrasal verbs have similar frequencies, they have similar amounts of information ($I(x)$), it follows that higher entropy values mean lower information residual values.

So far I have shown that transparent and opaque phrasal verbs are significantly different from one another whether using token-based measurements ($p=2.49e-06$) or type-based ones ($p=0.01530$). This supports the findings from experiment one where I used non-information theoretic measures and found two different distributions.

Another way to think about the success of these measures is to look at the classification accuracy. To do this, I created a generalized linear model to predict what sort of phrasal verb is most likely given a particular information residual score as the input. If the model assigns something close to 0 to a particular phrasal verb because of its information residual, then it’s predicting an opaque phrasal verb. If it assigns something close to 1, the model is saying that the phrasal verb is transparent. If we take these predictions and say that everything above a 0.5 is a prediction of a transparent phrasal verb and everything below 0.5 is a prediction of an opaque phrasal verb, then we can compare the model’s predictions against Bannard (2002)’s classification. Doing this results in 76.61% accuracy (95 out of 124 phrasal verbs are correctly classified). That’s a pretty good result just by the look of it, but in truth we need to compare it to a baseline. What if we had just guessed “opaque” for all of them? In that case, we would’ve gotten 70.16% accuracy. Our measure represents a 9.195% relative improvement over baseline, which is nothing to sneeze at, but still isn’t the kind of result one plasters above the drinking fountains.

Nevertheless, these results are remarkable: they have emerged despite a flotilla of factors that should have drowned them out. Many of these factors come from the fact that paradigm sizes are sensitive to accurate frequency values; we have a large corpus but it is notoriously difficult to extract phrasal verbs and even the BNC with its 100 million words is a very feeble stand-in for the experiences English speakers actually have with language.

Moscoso del Prado Martín et al (2004) propose information residual in order to account for lexical decision response times more parsimoniously. Heretofore, results were attributed to surface frequency, base frequency, inflectional ratio, cumulative root frequency, and morphological family size. Information residual is “a measure of the cost of recognizing a word, considering the decreases and increases in uncertainty contributed by the morphological paradigms to which the word belongs” (Moscoso del Prado Martín et al 2004: 15). Practically, by dropping out all the excess terms, they reduce a lot of the collinearity that poses problems for regression models using multiple frequency counts. Theoretically, they like the fact that response latencies arise as differences in statistical distributions within paradigms.

It is exciting that phrasal verbs should be describable in similar terms. It once again suggests that the morphological analogy is not misplaced, and it points to the probabilistic explanations that are available to us. Most importantly, we start to see how we might be able to measure how words are related to one another and what impact this has on notions of compositionality—is it really an all-or-nothing game? Concepts and methodologies that have helped researchers understand word formation can be imported into areas of syntax, where we are concerned with the formation of even more complex ideas.

7. Experiment 3: Entropy measurements improve Gries (2002)'s model

In this section, I build upon Gries' work in predicting the placement of the direct object in transitive phrasal verbs. Specifically, I create a generalized linear mixed-effects model using the actual data Gries (2002) uses.³² Where Gries uses 15 fixed effects, my model has only seven fixed effects and one random effect.³³ Despite the fact that I have simplified the model, I still achieve slightly higher classification accuracy.³⁴ If you would like to see a play-by-play of the creation of a model, please see "Appendix D: Building a model with Gries (2002)".

As Gries (2002: 14) points out, nearly everyone understands that the length of the direct object is tremendously influential, pushing you towards the joined construction as the NP gets longer. So let's begin with a simple generalized linear model with only one factor—the length, in syllables, of the direct object.³⁵ Since this model predicts the occurrence of the split construction, the coefficient estimate of "-2.0119" means that as syllable length increases you are less and less likely to use the split construction (V NP Prt).

```
Call:
glm(formula = CONSTRUCTION ~ log(LENG_SYLL), family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2348 -0.7277  0.4145  0.7859  2.0665

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.4112    0.2675   9.014  <2e-16 ***
log(LENG_SYLL) -2.0119    0.2028  -9.922  <2e-16 ***

Null deviance: 552.83  on 398  degrees of freedom
Residual deviance: 385.02  on 397  degrees of freedom
AIC: 389.02
```

³² Actually, 4 of the sentences had phrasal verbs that weren't in Baldwin and Villavicencio (2002), so my test corpus is made up of 399 sentences, not 403.

³³ Gries (2002) is rather conspicuously missing some sort of measurement of frequency as a factor. These are added in my models by dint of adding information theoretic measurements, which are largely defined by the frequency patterns in the data.

³⁴ Gries uses a general linear model without any random effects (in particular he uses a GLM sub-model called "discriminate analysis"). I have found that mixed models perform better, whether one adds in Verb as a random effect or Phrase. Gries (2002: 112) reports 86.4% classification accuracy—I have not been able to come up with a fixed effects or mixed effects model that gets that high just by using his variables. When he has tested his sample with bootstrapping, the results seem to average around 83.9%, varying based on whether the test sample is drawn from oral or written sentences or a mixture of the two.

³⁵ We can also measure this in words instead of syllables. Both are significant, but since they overlap so much, it is best to choose one. The syllable measurement performs better, so I opt to go with that.

This model assigns the correct construction to 75.19% of the data (if we just guessed “split” for every sentence we would only have been correct for 51.38% of them).³⁶

But how strong are these predictions? 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. A typical rule of thumb suggests that anything about 0.8 may mean the model has some real predictive capacity (see, for example, Baayen (in press)). Dxy 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. In other words, a good model has a high value for C and for Dxy. The basic length-only model has a C value of 0.8000 and a Dxy value of 0.6000. Length has something to offer us, but isn’t the whole story.

Using Gries’ results as a guide, I build up the model variable-by-variable. Suspecting that there may be random effects, I try Phrase, Verb, and Particle.³⁷ Either Phrase or Verb would be a reasonable random effect to include, but I choose Verb because it performs best in various iterations of the model. Note that even a mixed model that simply has syllable length and verb as a random effect gets higher classification accuracy: 78.45% compared to 75.19% in the NP-length-only model. The C and Dxy scores for this simple mixed-effects model are better, too: C=0.8666 and Dxy=0.7333.

Having experimented with no fewer than 26 different variables, here are the ones that comprise my final model:³⁸

- 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.

³⁶ To calculate this model’s accuracy, we do the same thing we did in the previous section with information residual accuracy. Here, if you give a model the length of the direct object for a bunch of data, it will assign scores to each one, ranging from 0 to 1. These correspond to the model’s classification prediction (will the construction be split or joined?). If there were a phenomenon we could model with perfect accuracy, the model would assign 0’s to one category and 1’s to another category. With a perfect model and perfectly categorical data, there wouldn’t be any predictions of 0.2343 or 0.5 or 0.9999. Moreover, all of the things that were assigned 0’s would really be in Category A and all of the things assigned 1’s would really be in Category B.

³⁷ Random effects are used in a model to indicate that the model is using a particular factor that is non-exhaustive. The typical examples are subjects and items, since you want the model to be generalizable to other people and other stimuli. In this case, we want to be clear that there are other phrases and verbs that could be part of the model but which aren’t in the data the model is being built from. Our model doesn’t include all possible particles, either, though that is a much smaller set. In fact, making Particle a random effect explains such an infinitesimal amount of the variance that it can safely be discarded.

³⁸ For a list of other factors considered and dismissed, see “Appendix C: Gries (2002) factors not in my model”.

- **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.

Here are examples of average values for four of these factors:

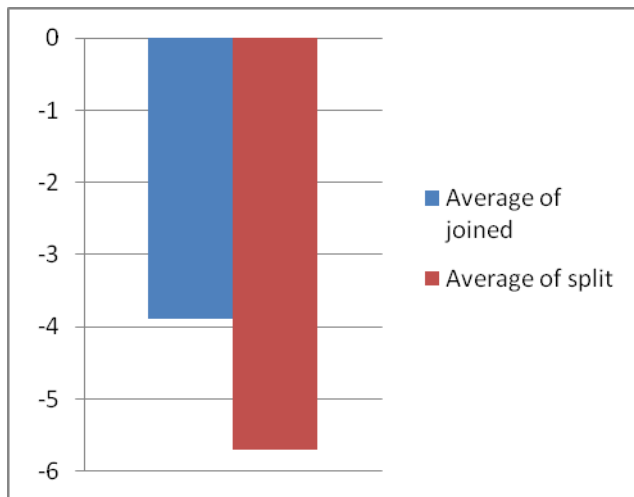


Figure 4: Info_resid; The greater the information residual, the more likely to be in the joined construction

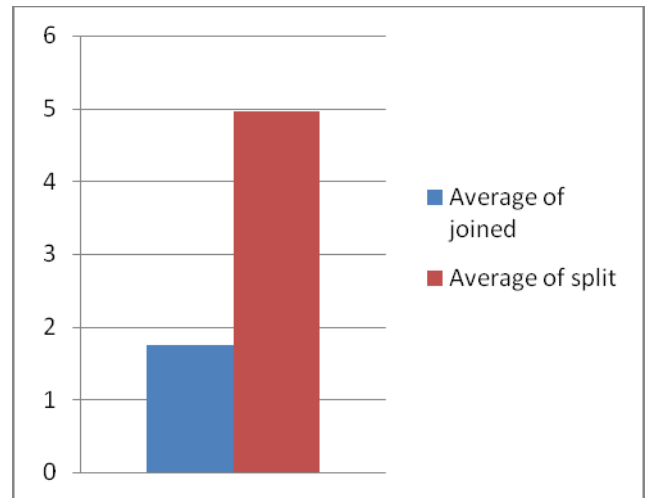


Figure 6: CohPC; the more often the direct object has been mentioned, the more likely it is that the split construction will be used

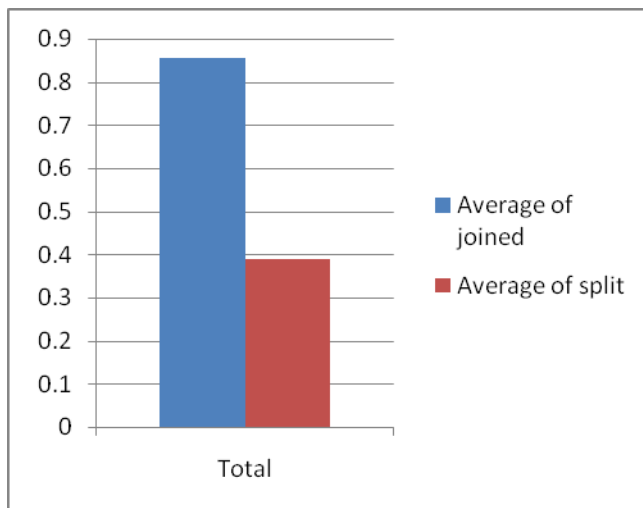


Figure 5: Log(LENG_SYLL); the longer the direct object is, the more likely it is to be joined

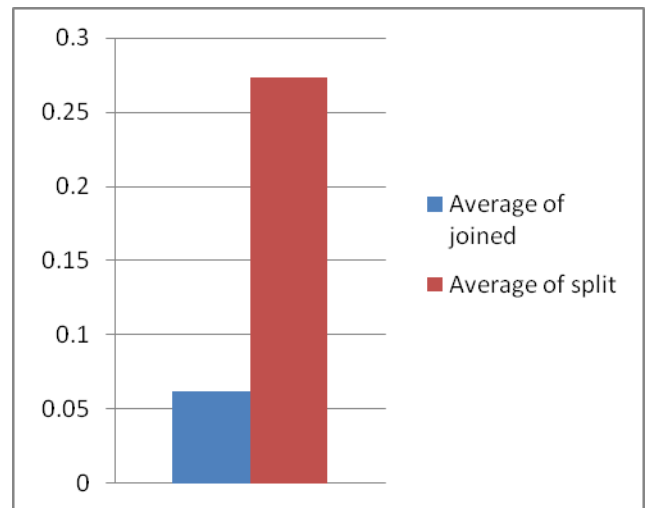


Figure 7: DIR_PP; if there is a following adverbial phrase, the split construction is more likely

All the factors in the final model are significant and G^2 tests demonstrate that removing any of the factors creates a weaker model. This model achieves 87.22% classification accuracy. It has a much higher score for C and D_{xy} — $C=0.9465$, $D_{xy}=0.8930$.

```

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:
Groups Name Variance Std.Dev.
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 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

There are several things that are immediately noticeable. The `Resid_token` coefficient (-0.09311) goes in the same direction as `LENG_SYLL`—as the direct object gets longer, you are less likely to get a split construction, as the information residual increases, you are also less likely to get a split construction. Remember that it is with opaque phrasal verbs that we get high information residuals. The `Resid_token` coefficient may look petite despite the fact that they achieve significance. Part of this has to do with the fact that they are doubtlessly overlapping with the random effect of “Verb”, which also contains part of the notion of interrelationships between individual verbs occurring in different phrasal verbs. I also expect that there is overlap with the `IDIOM` measurement. Ideally, I was hoping to *replace* the `IDIOM` measurement, so it is with some consternation that I leave it in. But the model’s likelihood is higher with it than without it. Its presence, in general, suggests to me that there is more to idiomaticity than just compositionality. See “Appendix E: What else is an idiom besides non-compositional?” for a quick survey about what else may matter.

Expanding to all verbs and particles

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 G^2 tests we see that removing any of these factors results in a weaker model.

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 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

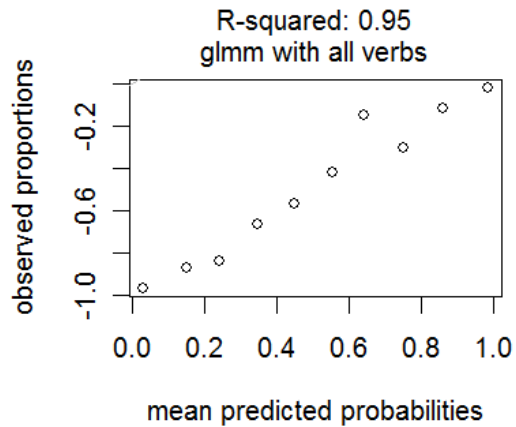


Figure 8: The goodness of fit for using all verbs (placing expected probabilities in ten bins and comparing each bin for its success against observed data).

So what happened to Resid_token? You’ll remember that residual information is made up of verb entropy and particle entropy. The basic problem here is that particle entropy is not significant ($p=0.4755$ when it is added to the model above) yet it dominates Resid_token. In fact, the average Prt_entropy is 1,966 times larger than the average Verb_entropy—it’s a complete wipe-out when they are combined. This turns on the definition of particle I have used (just those particles attested in the Baldwin and Villavicencio 2002 data).

Taking these results as-is means that the syntactic effect of alternation is caused by numerous factors, including the shape of the verbal paradigm, but it isn’t significantly influenced by the shape of the particle paradigm. This makes some sense if we consider that the verbs come first in speech and probably do dominate the meaning. It is also possible that the definition of “particle paradigm” needs to change—perhaps to include prepositional uses of the particle lexemes.

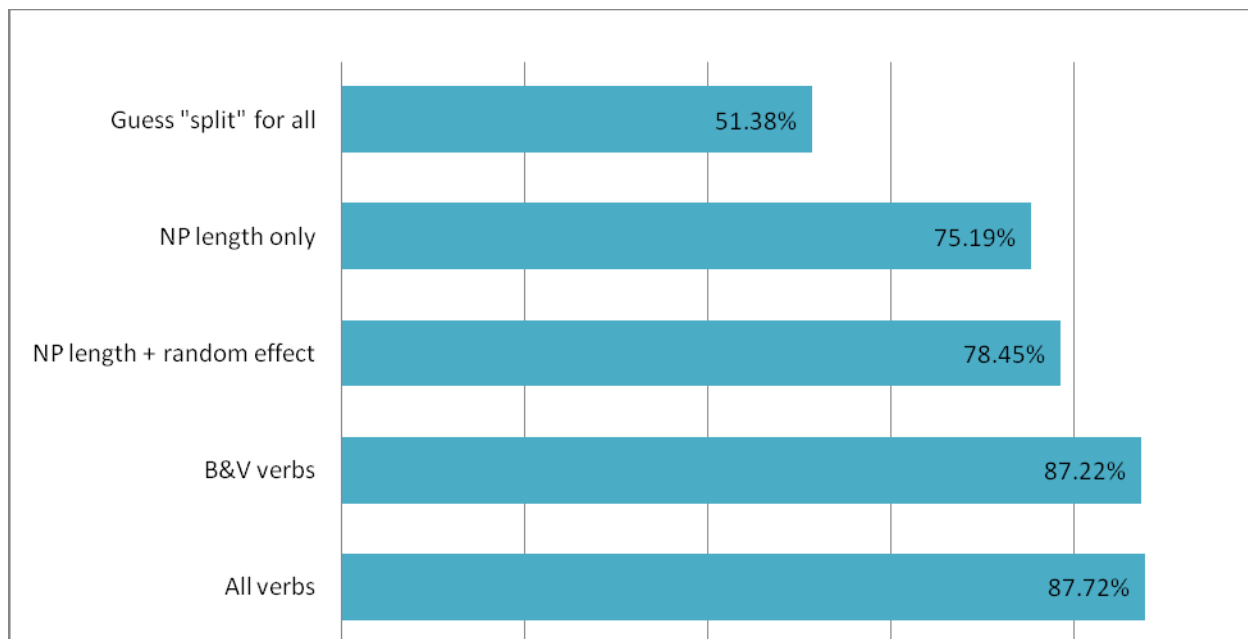


Figure 9: A comparison of various models' accuracy.

The all-verbs model is 87.72% accurate, with $C=0.9443$ and $D_{xy}=0.8887$. The accuracy is a tiny bit better, while C and D_{xy} are a little worse. Nonetheless, the all-verbs model may be preferable on theoretical grounds. We may well believe that all uses of the verb influence patterns of phrasal verbs, not just the instances of phrasal verbs themselves. All uses of the particles may also matter—if only we can identify the right way to define those and get accurate counts.

Discussion

In earlier experiments, I showed how to distinguish the compositionality of phrasal verbs by looking at the frequency of the whole and its parts—not just relative to each other (experiment one) but relative to all other phrasal verbs and all of their parts (experiment two). In this third corpus experiment, I showed that such measurements also help predict syntactic phenomena like particle alternation. And by including entropy values or information residuals, I am able to offer a significantly simpler and more accurate model.

The information theoretic measurements went in the direction one would predict: the higher the entropy, the more likely the phrasal verb is to occur in the split construction (V NP Prt). It's the transparent phrasal verbs that have higher entropy 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). This is indeed what happens.³⁹

Part of what this project boils down to is capturing the intangible concept of information. If this strikes you as strange, you aren't alone. Even the authors of the most cited textbook admit they were drawn to

³⁹ The information residual works the opposite way of entropy—the higher an information residual score, the more likely it is to appear in the joined construction—these are the more opaque phrasal verbs.

information theory because of this apparent impossibility (Cover and Joy 1991: vii). Nevertheless, it turns out that the field gives us real tools for expressing information.

But information theory isn't so far afield from the way most linguists treat language as a symbolic system. It is also typical in linguistics to worry about acquisition and storage. Information theory takes these fundamentals and adds to them a more controversial piece: frequencies. Once you acknowledge that frequencies belong in the domain of linguistics, it behooves you to use shorter descriptions for the more probable and the more basic terms. This means you give longer descriptions to the less probable/more complex terms.

8. General discussion

Throughout this paper I have shown that phrasal verbs can be more and less decomposable. 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 verb frequencies, relative to other instances of the verb, make them more likely to be treated as a single entity. In the terms of Hay and Baayen (2005), the whole is more salient than its parts.

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 compositionality (experiment two) and was even able to improve models of the "syntactic" phenomena of particle alternation (experiment three). Taken together, these experiments demonstrate the usefulness of modeling morphology, semantics, and syntax in the same (probabilistic) terms, rather than as three distinct (rule-based) components that interact.

Traditional theories of grammar were designed to formally describe structures for stable phenomena. Such theories deny that frequency and usage are a fundamental part about language; as such they are at odds with the findings reported in this paper (or if not "at odds", perhaps just "not interested"), which suggest that strength of the relationships between words influences the structures that they occur in.⁴⁰

Review of findings

Let's summarize the particular areas that a traditional, generative, categorical view of grammar will struggle with by breaking down the claims and evidence into four basic statements.

1. Transparent phrasal verbs and opaque phrasal verbs are different from one another.

⁴⁰ In fairness, some generativists have allowed for the importance of gradience (consider Jackendoff (2002), for example), though they haven't gone as far as others who think of language as fundamentally dynamic and think of linguistic structure as emerging during language use (for example, Hopper 1987, Bybee and Hopper 2001). Discovering and exploiting similarities in the data is a major research area for a number of different usage-based theories, including exemplar-based models and connectionist ones. See, for example, Bybee and McClelland (2005).

I defined transparent phrasal verbs as fully entailing the simplex meanings of their verbs and particles, while I defined opaque phrasal verbs as entailing neither the simplex verb or particle. This seemed to be an easy way to operationalize the difference in the fact that some phrasal verbs seemed to be easy to understand (*lift up*), while others seemed to require special learning (*give in*). These groups turned out to be distinct by several measures—first, I looked the relative frequencies between the simplex verbs and the particles they combined with. In experiment one, transparent and opaque phrasal verbs really did seem to come from two different distributions. In experiment two, the two groups were also found to be reliably different from each other, even after summing up the weighted entropies for all verbs that shared a particle in common and all particles that shared a verb in common.

A traditional grammar can handle distinct groups, no problem. If I had found that phrasal verbs really divided into literal phrasal verbs and idiomatic ones, then it would be straightforward to let the literal ones combine with normal grammatical processes whereby you scoop up entries from the lexicon and put them together using well-established rules. For idioms, well, those would just be listed in the lexicon.

The real problem comes in the second observation.

2. Opaque phrasal verbs can't be relegated to the lexicon.

As the previous section described, if there were only transparent or only opaque phrasal verbs, we wouldn't have a problem. But the truth is that we have both.⁴¹ We could make distinctions between these groups, handling transparent ones normally and calling the opaque ones idioms, but opaque phrasal verbs don't really seem to behave like idioms—what sort of syntactic reindeer games are they prohibited playing that transparents can? They passivize, they alternate between joined and split constructions. In short, while they share with idioms the characteristic of opacity, opaque phrasal verbs are not idioms themselves.

What, then, is the nature of opacity? This paper has gone about showing that opaque phrasal verbs do get some of their meaning from the verbs and particles they are comprised of, just not as much as the transparent ones.

(16) *Transparent*: She took away her daughter's iPod and grounded her for a week.

(17) *Opaque*: He played down his illness because he didn't want anyone thinking he wasn't up to the job.

If you know English, the first example is presumably trivial to understand even if you've never heard "took away" before (so long as you do know "took" and "away" separately). But with (16), you can't use the simplex meaning of "play" or "down"; you will have to learn the meaning from context and experience. The parts of the opaque phrasal verbs don't help us out a lot, but they do help us some, and to differing degrees depending on the particular phrasal verb.

(18) *Opaque*: They wiped out the invading army.

In this case, the meaning may not be entailed, but it's close: you require metaphorical extensions of *wipe* and *out*. Opaque phrasal verbs, as we saw in experiments one and two, are less decomposable, but that

⁴¹ In fact, we also have mixed examples that entail either the verb or the particle but not both. These pattern right in between the opaque/transparent patterns.

isn't to say they haven't any pieces. What our information theory measurements have given us is (i) a way of showing that the whole of *wipe out* is more salient than its parts, and (ii) that it is reasonable to connect all the different types of *wipes* and all the different types of *out* in assessing how the combination will be interpreted. So to *wipe out* we connect *knock out* and its unentailed *out*, *hammer out* and its entailed *out*, *wipe up* and its entailed *wipe*. These examples share in common making something vanish—an invading army, consciousness, a dent, spilt milk. Our information theoretic measures have helped to link these together and to weight them by the likelihood that they have actually contributed something to the understanding of one another.

3. Transparent phrasal verbs are more decomposable than opaque ones.

The basic finding is from experiment one, where I showed that phrasal verbs are more or less parsable—decomposable—and that it's the opaque phrasal verbs that pattern like the most opaque affixes in morphology (*-ity*, *-ic*).

The standard idealization of grammar uses generative syntax and compositional semantics to get meaning. It can handle transparent phrasal verbs quite easily because it posits a lexicon with words and meaning, grammatical rules that combine them, and basic semantic interpretation principles such that “speakers do not in general need to know in advance the meanings of complex structures...rather, the meaning of such larger structures simply follow from the knowledge of forms and rules that speakers have to know independently” (Fillmore et al 1998: 502, summarizing typical concepts of grammar).

Recall that Hay (2002) showed that some affixed forms are simple just because they are more frequent than the base forms that linguists think of as composing them. When we talk about compositionality and parsability, we really mean that the use of language shapes what is perceptible, and that this determines what are 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.

4. The lexicon is interconnected and syntax is sensitive to that.

In experiment two, I used information theoretic terms to interrelate 2,318 verbs and 48 particles (which form 6,793 phrasal verbs altogether). I took 124 of these that Bannard (2002) had already coded for entailment characteristics and found that the measures significantly predicted compositionality. These measures were sensitive not only to relative frequencies, but also to frequencies within “paradigms”, that is, the measures don't simply count occurrences of *drum up*, but also add in all the other phrasal verbs that share either a verb in common (*drum out*) or a particle (*move up*).

There are two ways of thinking of phrasal verbs in the generative tradition: either the verb and particle are listed together or they are listed separately (in which case you have compositional processes to combine them). Either way, we need to build in the connections between the different pieces and wholes—because it is the strength of those connections that determines when we see pieces and when we see wholes. Information theoretic terms can help us understand how the strength and patterns of paths translate to compositionality and particle placement.

Experiment two concluded we could use information theoretic measures for the semantic phenomenon of entailment, experiment three concluded that we could also use these measures for the syntactic phenomenon of particle placement (on a totally separate set of data) The best model for predicting Gries

(2002)'s corpus of particle alternations involves seven fixed effects, including information residual, which we know describes entailment characteristics (i.e., compositionality).

What I found in all three experiments is that transparent and opaque phrasal verbs have different distributions and behaviors. That can't be captured by grammatical rules unless you mark individual lexical items and have different (but very similar) rules that are sensitive to what they find in the lexical entry. But if lexical items are simply marked in their entries without having any connections tying them together, then we've missed important generalizations about "what behaves like what", *and* how that came about and is maintained.

9. Next steps

The findings in this paper will be strengthened (or refuted) by four additional lines of research.

First, experiments one and two make predictions about the entailment characteristics for 6,793 phrasal verbs, a sample of which should be tested. The most effective way to do this is a psycholinguistic experiment in which subjects are asked to rate the entailment characteristics of phrasal verbs.⁴² The phrasal verbs used as stimuli should be chosen to represent a range of entropy (or information residual) values. The prediction is that phrasal verbs with higher entropy/lower information residual values will be seen as more entailed.

Second, experiment three and Gries (2002) could obviously be applied to an even larger corpus of natural language sentences with particle alternation. The prediction is that the factors that were significant in experiment three would still be significant in a larger corpus.

Third, reading time experiments should establish that opaque phrasal verbs are read faster than transparent phrasal verbs. Encouragingly, Kuperman (2007) has found results that phrasal verbs in Dutch are faster to read when they are more predictable. Because entropy values encode probabilities, they measure predictability (and its inverse, surprisal). Something with a low entropy value (like the opaque *bog down*) has low surprisal, which is to say that it has high predictability.

Fourth, a phonetic analysis of phrasal verbs may show that opaque phrasal verbs are phonetically reduced in speech and/or are more likely to participate in an ongoing sound change. As with reading times, this follows from the idea that opaque phrasal verbs are more predictable. More specifically, predictable words do not need to carry as much semantic information as unpredictable ones, so they should be more amenable to changes and reductions.

There is one additional direction that will be fruitful, but for which it is harder to make predictions. An analysis of historical corpora should give us more insight about the origins of opaque phrasal verbs and what has happened to them over time. Were all of the first phrasal verbs transparent? Which phrasal verbs became more opaque over time and which seem to have started off opaque? The framework this paper has used has allowed for opacity and transparency to change over time but will be greatly enriched by additional research that shows how those changes have actually preceded.

⁴² One method to do this is to ask people "How closely related is *send* to *send in*?" as Kuperman (2007) did. Alternatively, we might ask "When you *send in* something does it get *sent*?"

Appendix A: Bannard (2002) entailment coding

Bannard's gold-standard data comprises 180 phrasal verbs. He initially identified 843 phrasal verbs from the Wall Street Journal. From this, he selected the 263 that occur more than four times. He rejected 25 of these because they didn't seem to be phrasal verbs. He rejected another 28 because they were prohibitively polysemous, and another 30 because they were too noisy (this is rather underspecified in his description). This left him 180 phrasal verb types with 2034 tokens.

The most important decision in annotation, Bannard says, was to annotate by type instead of token.

This raises the significant problem of polysemy. As is the case with any lexical item, VPCs can have more than one sense. When I attempt to describe all instances of a particular VPC in a corpus with one judgement, I am oversimplifying enormously, since the likelihood is that it will encompass items which have a different sense and for which the judgement simply isn't true. (Bannard 2002: 12)

This is indeed a major simplification. To get around this, Bannard makes sure that he is recording a dominate sense that has a minimum ratio of 4:1. This leaves him with a problem that he can tackle in the time allowed, but it is potentially problematic for me to adopt his scores for the "unseen" data in the BNC corpus I'm looking at. To the extent that I find null results, the problem may well be that the dominate senses in Bannard's data are different than those in the BNC. It is less likely, however, that I will find false patterns using Bannard's entailment judgments.

Having said that, I will move on to describing Bannard's coding. I quote at length from (Bannard 2002: 9):

1. If the VPC sentence is *Tom put the picture up*, does this entail that the picture has been put somewhere by Tom, and that as a consequence the picture is up? The answer seems to be yes, and we can classify the VPC here as fully compositional.
2. If the VPC sentence is *Richard finished up his paper*, then can we say that the paper is finished by Richard, or that as a result the paper is up? The answer to the first question is yes, but to the second is no. We can therefore say that the verb has standard semantics here, but the particle does not.
3. If the VPC sentence is *Philip gunned down the intruder*, can we say that the intruder has been gunned by Philip or that as a result the intruder is down? The answer to the former question would seem to be no, but to the latter question the answer is yes. We can therefore say that the particle but not the verb has standard semantics.
4. If the sentence is *Richard and Bethany made out*, then can we say that the two individuals involved made, or that they were out. The answer to both questions seems to be no, and we can therefore say that the verb is completely noncompositional.

Appendix B: Gries' definitions for IDIOM

Gries (2002) says that phrasal verbs are difficult to categorize if you only have “literal” (*Fred carried up the box*) or “idiomatic” (*I will turn over a new leaf*) boxes. For that reason, he comes up with a middle category of “metaphorical”, which is essentially all of the examples in between that isn't one of the other two.

Literal: Totally predictable from meaning of the parts: *You can stick the pin in.*

Metaphorical: Not fully predictable from the meaning of its parts because of, say, violations of selectional restrictions that could be accounted for with reference to simple metaphorical or metonymic mappings...or, more importantly, preference violations: *I put down comments.*

Idiomatic: The sentence isn't predictable on the parts and maximally two simple mappings. “In cases where this classification procedure seemed only slightly problematic, the degree of idiomaticity was checked using the Longman Dictionary of Phrasal Verbs”: *Divers should take out decompression insurance.*

He doesn't get an effect of metaphor in his results—they split evenly across the split and joined constructions that he's analyzing—but he does get a strong effect for idiomaticity, as mentioned above. One of the chief goals of this paper is to see whether I can come up with a graded measure of idiomaticity that has better predictivity. Just as important, this measure should connect to an overall theory of idioms. Gries' three buckets start us out on the road by telling us there's something significant to look at, but they are rather too coarse to tell us much about the phenomenon itself.

Appendix C: Gries (2002) factors not in my model

Here are factors also considered but dismissed. Adding these factors to the model only increases the AIC vales (Akaike Information Criterion), which we want to keep small and decreases the log likelihood values that we want to keep large. Pay special attention to COMPLEXITY, ActPC, TOPM, LM, and CONCRETE, all of which had higher factor loadings than IDIOM in Gries' model.

- ActPC, the number of clauses before in which there was mention of the direct object's referent.
- COMPLEXITY, Gries treats this as an ordinal scale of 0 (simple, like a pronoun), 1 (intermediate, like an NP with an adjectival modifier), and 2 (complex, as in an NP with an embedded clause). I treat this as a factor.
- TOPM, the number of times that the direct object has been mentioned (properly, the *referent* of the direct object).
- LASTMENTION, 0 if the direct object has not been mentioned before, otherwise 1.
- ANIMACY, is the direct object animate?
- ClusSC, the number of clauses to the next mention of the direct object's referent.
- CONCRETE, is the direct object something concrete?
- MEDIUM, is the example from the written or oral portion of the BNC?
- NM, for "next mention": 0 if the direct object isn't mentioned again in the discourse, 1 if it is.
- OM, for "overall mentions": 0 if the direct object is only mentioned in the phrasal verb in question, 1 if it is mentioned before or after the sentence in question.
- PREP.PA, is the particle identical to the preposition of the following directional adverbial?
- TOSM, how many times is the referent of the direct object mentioned subsequently?

Here are factors that Gries (2002) doesn't include, but which I did try (they didn't make it into the model because they were insignificant and/or because they didn't improve the model by AIC and G^2 tests).

- Mutual information, a score based on how much expectation there is of seeing the particle if you've seen the verb (and vice versa). This was calculated using two different corpora. Neither measurement was significant.
- Odds ratio, a score based on how often the verb and particle would be expected to occur together based on their overall frequencies. This was calculated using two different corpora. Neither measurement was significant.
- Resid_type, the same as Resid_token except that it is calculated based on the type counts for the phrasal verb, not the token counts.
- Verb_type_entropy, like Verb_entropy, but calculated using types.
- Prt_type_entropy, like Prt_entropy, but calculated by looking at the various phrasal verbs that the particles combine (using types instead of tokens).
- Phrase_freq, the frequency of the phrasal verb in the corpus didn't change how it was realized syntactically, nor did this change if we took the log of that frequency.
- Verb_freq, the raw frequency of the verb wasn't significant, nor was its log.

Appendix D: Building a model with Gries (2002)

This appendix shows how I arrived at a model for the data in Gries (2002). It uses commands from the R statistical program. Those begin, simply enough, with reading in a file with all the data and attaching it so that the data can be called without specifying the “gries” object each time.

```
> gries<-read.csv("Gries_for_R.csv", header=T)
> attach(gries)
```

I'll begin with loading the statistical packages we'll need.

```
> library(Design)
> library(lme4)
```

Since we'll begin using the Design package's `lrm()` function, we'll also need to build a data distribution, a couple steps Manning agrees are “magical incantations” (p.c.).

```
> gries.dd<-datadist(gries)
> options(datadist="gries.dd")
```

Now I build a model that is based solely on the length of the NP in words. The p-values reported in both the `anova()` and in the general model description suggest that we have something. Because we are predicting the split construction (V NP Prt), the coefficient of word length is negative (-0.847)—the longer the NP, the less likely the particle will follow it.

```
> griesW.lrm<-lrm(CONSTRUCTION ~ LENGTHW, x=T, y=T)
> anova(griesW.lrm)
```

Wald Statistics		Response: CONSTRUCTION	
Factor	Chi-Square	d.f.	P
LENGTHW	60.97	1	<.0001
TOTAL	60.97	1	<.0001

```
> griesW.lrm
```

Logistic Regression Model

`lrm(formula = CONSTRUCTION ~ LENGTHW)`

Frequencies of Responses

0 1

194 205

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
399	6e-10	125.58	1	0	0.794	0.588

Gamma	Tau-a	R2	Brier
0.722	0.295	0.36	0.179

	Coef	S.E.	Wald Z	P
Intercept	2.093	0.2579	8.12	0
LENGTHW	-0.847	0.1085	-7.81	0

We can also compare this to the NP length in syllables.

```
> griesS.lrm<-lrm(CONSTRUCTION ~ LENG_SYLL, x=T, y=T)
> anova(griesS.lrm)
```

```

Wald Statistics      Response: CONSTRUCTION

Factor      Chi-Square      d.f.      P
LENG_SYLL  74.05              1      <.0001
TOTAL      74.05              1      <.0001
> griesS.lrm
Logistic Regression Model
lrm(formula = CONSTRUCTION ~ LENG_SYLL)
Frequencies of Responses
 0  1
194 205

  Obs  Max Deriv      Model L.R.      d.f.      P      C      Dxy
 399  7e-09          158.98          1      0      0.841  0.683

Gamma      Tau-a      R2      Brier
0.744      0.342      0.438      0.165

      Coef      S.E.      Wald Z      P
Intercept      2.118      0.23950      8.84      0
LENG_SYLL      -0.532      0.06182      -8.61      0

```

Again we have small p-values, but look at the difference in some of the other statistics. “Model L.R.” expresses the model likelihood—the difference between the null deviance and the residual deviance. A higher number here is better (favoring the syllable measurement). R2 says how accurate the predictions of the model are and again the syllable measurement is better. 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. As mentioned earlier, the rule of thumb is that anything about 0.8 may mean the model has some real predictive capacity. Dxy stands for “Somers’ D_{xy} ”, which is a rank correlation between predicted probabilities and observed responses. In this case it is 0.683 that indicates randomness, while 1.0 indicates perfect prediction. In all of these measures, measuring length in syllables seems to out-perform measuring length in words.

Though the syllable length only ranges from 1 to 31, it seems reasonable to transform the measurement on a logarithmic scale. Doing so also increases the R2 value, which is nice.

```

> grieslogS.lrm<-lrm(CONSTRUCTION ~ (log(LENG_SYLL)), x=T, y=T)
> anova(grieslogS.lrm)
      Wald Statistics      Response: CONSTRUCTION
Factor  Chi-Square d.f. P
LENG_SYLL  98.44  1 <.0001
TOTAL    98.44  1 <.0001
> grieslogS.lrm
Logistic Regression Model
lrm(formula = CONSTRUCTION ~ (log(LENG_SYLL)))
Frequencies of Responses
 0  1
194 205

```

```

Obs Max Deriv Model L.R. d.f. P C Dxy
399 2e-09 167.8 1 0 0.841 0.683

Gamma Tau-a R2 Brier
0.744 0.342 0.458 0.161

```

```

Coef S.E. Wald Z P
Intercept 2.411 0.2675 9.01 0
LENG_SYLL -2.012 0.2028 -9.92 0

```

I continue trying out logarithmically transformed and untransformed versions in future models, as well as trying both LENGTHW and LENG_SYLL. In the interest of expedience (if I can be said to be interested in that), I don't include those in the notes to follow.⁴³

Let's go the opposite direction and include all of Gries' factors to see what that model looks like.

```

> griesall.lrm<-lrm(CONSTRUCTION ~ ANIMACY + COMPLEXITY + CONCRETE + DIR_PP + DET + IDIOM +
LASTMENTION + NM + MEDIUM + PREP.PA + log(LENG_SYLL) + TYPE + ActPC + ClusSC + CohSC + CohPC + TOSM +
TOPM + log(OM), x=T, y=T)
> griesall.lrm
Logistic Regression Model
lrm(formula = CONSTRUCTION ~ ANIMACY + COMPLEXITY + CONCRETE +
DIR_PP + DET + IDIOM + LASTMENTION + NM + MEDIUM + PREP.PA +
log(LENG_SYLL) + TYPE + ActPC + ClusSC + CohSC + CohPC + TOSM +
TOPM + log(OM), x = T, y = T)
Frequencies of Responses
0 1
194 205

Obs Max Deriv Model L.R. d.f. P C Dxy
399 0.009 311.54 23 0 0.942 0.885

Gamma Tau-a R2 Brier
0.885 0.443 0.723 0.094

Coef S.E. Wald Z P
Intercept -0.176297 0.8195 -0.22 0.8297
ANIMACY 0.090586 0.5951 0.15 0.879
COMPLEXITY -1.468123 0.4722 -3.11 0.0019
CONCRETE -0.116212 0.4937 -0.24 0.8139

```

⁴³ As we'd expect, complexity and length are highly correlated (i.e., they have a correlation coefficient greater than 0.5).

```

> cor(COMPLEXITY, log(LENG_SYLL))
[1] 0.7355457
> cor(COMPLEXITY, log(LENGTHW))
[1] 0.7711503
> cor(log(LENG_SYLL), log(LENGTHW))
[1] 0.8894477

```

DIR_PP	1.994678	0.4928	4.05	0.0001
DET=undef	-0.697337	0.4935	-1.41	0.1576
DET=none	-0.65534	0.4386	-1.49	0.1351
IDIOM=lit	2.119144	0.6966	3.04	0.0023
IDIOM=met	1.787632	0.624	2.86	0.0042
LASTMENTION	-0.853653	1.101	-0.78	0.4381
NM	1.054298	1.1201	0.94	0.3466
MEDIUM=wri	-0.795742	0.3588	-2.22	0.0266
PREP.PA	-13.24515	270.992	-0.05	0.961
LENG_SYLL	-0.95919	0.3943	-2.43	0.015
TYPE=pron	9.951034	27.0799	0.37	0.7133
TYPE=propN	1.367965	0.8478	1.61	0.1066
TYPE=spron	3.146759	1.1764	2.67	0.0075
ActPC	0.289542	0.1272	2.28	0.0228
ClusSC	0.005306	0.1207	0.04	0.9649
CohSC	-0.136899	0.1448	-0.95	0.3445
CohPC	0.364565	0.1274	2.86	0.0042
TOSM	0.955941	0.4484	2.13	0.033
TOPM	0.082567	0.4323	0.19	0.8485
OM	-3.200566	1.2319	-2.6	0.0094

I've highlighted in red and bolded items that seem significant at the $p < .01$ level (the others that are in green and bold-italics are significant at $0.05 < p < 0.01$).

This is a pretty otiose model and we know that when models have extraneous factors they get in the way of truthfully seeing what's really happening. Take, for example, LENG_SYLL, which is now only significant at the $p < 0.05$ standard. Yet Gries (2002) and our own intuitions confirm that this is probably the most important factor. The explanation is that other variables here are stealing some of its thunder.

We could rebuild the model with only the significant factors. In truth, I proceeded somewhat differently. Taking LENG_SYLL as being the one "must have" in my model, I combined it with each of the significant factors. In doing so, I was able to measure what model with only two variables had the best model characteristics (Model LR, C, Dxy, R2). CohPC turns out to be the best of these. I move to considering LENG_SYLL and CohPC as my basic terms and continue trying out the others one-by-one. This process demonstrates that IDIOM is the next best one to add, then TYPE, then DIR_PP, then DET. At this point, the model looks like this:

```
> gries.lrm21 <- lrm(CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + TYPE + DIR_PP + DET, x=T, y=T)
> gries.lrm21
Logistic Regression Model
lrm(formula = CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + TYPE +
    DIR_PP + DET, x = T, y = T)
Frequencies of Responses
  0  1
194 205
```

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
399	0.005	278.1	10	0	0.924	0.848

Gamma	Tau-a	R2	Brier
0.849	0.425	0.669	0.11

Coef	S.E.	Wald Z	P
Intercept	-0.3758	0.64694	-0.58 0.5613
LENG_SYLL	-1.5958	0.28663	-5.57 0.0000
CohPC	0.1904	0.05931	3.21 0.0013
IDIOM=lit	2.1627	0.56367	3.84 0.0001
IDIOM=met	1.6161	0.55360	2.92 0.0035
TYPE=pron	9.8544	27.96729	0.35 0.7246
TYPE=propN	1.1238	0.77042	1.46 0.1446
TYPE=spron	1.9364	0.78874	2.46 0.0141
DIR_PP	1.8058	0.43348	4.17 0.0000
DET=indef	-0.8805	0.44814	-1.96 0.0494
DET=none	-0.9234	0.39052	-2.36 0.0181

The next one to be added will be COMPLEXITY, though adding it directly results in the coefficient of LENG_SYLL dropping from -1.5958 to -0.8365 (the p-value goes from $p < 0.0000$ to $p = 0.0202$). Complexity, meanwhile, takes up a coefficient of -1.3684 ($p = 0.0015$). Other research has shown that complexity and length are not particularly different in their effects on syntactic data. Here I prefer to leave off COMPLEXITY and keep the LENG_SYLL effect strong. This is defensible because it keeps the model simpler and at this point, I take gries.lrm21 to be the most parsimonious way of accounting for the Gries factors.

We can also use bootstrapping validation to work backwards from complex-to-simpler models.⁴⁴

```
> griesall.lrm2<-lrm(CONSTRUCTION ~ ANIMACY + CONCRETE + DIR_PP + DET + IDIOM + LASTMENTION + NM +
MEDIUM + PREP.PA + log(LENG_SYLL) + TYPE + ActPC + ClusSC + CohSC + CohPC + TOSM + TOPM + log(OM), x=T,
y=T)
> validate(griesall.lrm2, bw=T, B=200)
```

Deleted	Chi-Sq	d.f.	P	Residual	d.f.	P	AIC
ClusSC	0.00	1	0.9849	0.00	1	0.9849	-2.00
PREP.PA	0.00	1	0.9617	0.00	2	0.9987	-4.00
CONCRETE	0.09	1	0.7661	0.09	3	0.9929	-5.91
ANIMACY	0.10	1	0.7543	0.19	4	0.9958	-7.81
TOPM	0.15	1	0.6974	0.34	5	0.9968	-9.66
LASTMENTION	0.27	1	0.6019	0.61	6	0.9962	-11.39
NM	2.63	1	0.1049	3.24	7	0.8618	-10.76
CohSC	2.47	1	0.1163	5.71	8	0.6799	-10.29
DET	5.28	2	0.0715	10.98	10	0.3588	-9.02

⁴⁴ If Resid_token is added to this model, it is preserved as a factor in the final model, too. When LENG_SYLL, LENGTHW and COMPLEXITY compete against each other most models opt for COMPLEXITY. If COMPLEXITY is kept, the resulting bootstrapping validation still suggests eliminating all of the factors but DIR_PP, IDIOM, and COMPLEXITY.

```

MEDIUM  4.05  1  0.0441 15.04  11  0.1808 -6.96
TOSM    3.20  1  0.0736 18.24  12  0.1087 -5.76
ActPC   2.47  1  0.1160 20.71  13  0.0789 -5.29
OM      1.51  1  0.2187 22.22  14  0.0742 -5.78
TYPE    6.77  3  0.0794 29.00  17  0.0346 -5.00
CohPC   6.20  1  0.0128 35.19  18  0.0089 -0.81

```

Approximate Estimates after Deleting Factors

```

      Coef S.E. Wald Z      P
Intercept -0.4161 0.6152 -0.6763 4.988e-01
DIR_PP    1.5154 0.4466  3.3931 6.911e-04
IDIOM=lit 2.1600 0.5803  3.7225 1.973e-04
IDIOM=met 1.4345 0.5784  2.4802 1.313e-02
LENG_SYLL -1.2653 0.2790 -4.5344 5.776e-06

```

Factors in Final Model

```
[1] DIR_PP IDIOM LENG_SYLL
```

Here's what that minimal model looks like.

```
> gries.lrm40<-lrm(CONSTRUCTION ~ DIR_PP + IDIOM + log(LENG_SYLL), x=T, y=T)
> gries.lrm40
```

Logistic Regression Model

```
lrm(formula = CONSTRUCTION ~ DIR_PP + IDIOM + log(LENG_SYLL), x = T,
     y = T)
```

Frequencies of Responses

```
0  1
```

```
194 205
```

```

      Obs Max Deriv Model L.R.  d.f.  P      C  Dxy  Gamma
399    2e-12  226.82    4    0  0.89  0.779  0.796

```

```
Tau-a    R2    Brier
```

```
0.39    0.578  0.133
```

```
      Coef S.E. Wald Z P
```

```

Intercept 0.730 0.4059  1.80 0.0721
DIR_PP    1.825 0.4184  4.36 0.0000
IDIOM=lit 2.224 0.4317  5.15 0.0000
IDIOM=met 1.383 0.4307  3.21 0.0013
LENG_SYLL -2.095 0.2333 -8.98 0.0000

```

If you compare this bare-bones model (`gries.lrm40`) to `gries.lrm21`, which adds `CohPC` and `TYPE`, you'll see that the extra two factors do better in terms of `C`, `Dxy`, and `R2`. I adopt `gries.lrm21` for that reason.

The question now becomes—can this model be improved by adding information residual? In both the case of token-based and type-based information residual, the basic model stats improve (`Model LR`, `C`, `Dxy`, and `R2`). However, a model with `Resid_token` is significant, while a model with `Resid_type` is not. (Since the rest of the models stay about the same, I extract only the lines having to do with the residuals themselves, again these are from two separate models.)

```
Resid_token -0.1061 0.03201 -3.32 0.0009 (from model gries.lrm41)
```

Resid_type -0.03644 0.10473 -0.35 0.7279 (from model gries.lrm42)

At this point it is worth considering random-effects. Generally, we think of these as the “non-repeatables” in data—the subjects and items in a psycholinguistic experiment, for example. Here, I would consider the phrase and verb as such since we know that English has no set cap on simplex verbs or phrasal verbs. (I do consider particles as random effects, too, but they don’t seem to account for any real variance. Particles are a much more closed group, too, and most of them are represented in our data and would be in any future studies, too.)

To build this model, I switch to the lme4 package and a command called “lmer”.⁴⁵

```
> gries.glmm70<-lmer(CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + TYPE + DIR_PP + DET + Resid_token +
(1|Verb), family="binomial")
> gries.glmm70
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:
Groups Name Variance Std.Dev.
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 **
```

In this model, TYPE is barely significant, but a model without TYPE is worse in terms of log likelihood and AIC, so we want to keep it.

⁴⁵ I have also built out a model including both phrase and verb as random effects, but it seems questionable to include both when there’s such overlap in them. Having the verb-only-as-random-effect model results in higher log likelihood and lower AIC than having phrase-only, despite the fact that the phrase accounts for more variance.

```

> gries.glmm71<-lmer(CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + DIR_PP + DET + Resid_token + (1|Verb),
family="binomial")
> anova(gries.glmm70, gries.glmm71)
      Df   AIC   BIC logLik Chisq Chi Df Pr(>Chisq)
gries.glmm71 10 298.00 337.88 -139.00                # no TYPE
gries.glmm70 13 278.04 329.90 -126.02 25.952    3 9.763e-06 ***

```

At this point, let's compare having a random-effect (gries.glmm70) vs. not having one (gries.lrm70). The stats reported for the models aren't quite the same, so we have to ask for the information we want. For example, this is how we get Somers' D_{xy} for the mixed-effects model.

```

> somers2(binomial())$linkinv(fitted(gries.glmm70)), CONSTRUCTION)
C           Dxy           n           Missing
0.9464798    0.8929595    399.0000000    0.0000000

```

The values here *are* higher than the fixed-effects-only model (C=0.931 and Dxy=0.862).

We should also verify that the addition of Resid_token improves the log likelihood of the model. Here's a model without Resid_token compared against one that has it. The model with the Resid_token has the higher log likelihood and the lower AIC value. Gries.glmm70 continues to be our best model.⁴⁶

```

> gries.glmm72<-lmer(CONSTRUCTION ~ log(LENG_SYLL) + CohPC + IDIOM + DIR_PP + DET + TYPE + (1|Verb),
family="binomial")
> anova(gries.glmm71, gries.glmm72)
      Df   AIC   BIC logLik Chisq Chi Df Pr(>Chisq)
gries.glmm72 12 282.64 330.50 -129.32                # no Resid_token and it's worse
gries.glmm70 13 278.04 329.90 -126.02 6.5926    1 0.01024 *

```

The final model, which is given in the body of the paper, is gries.glmm70, which has 87.22% accuracy.⁴⁷

⁴⁶ In some of the earlier LRM models, we also saw some significance for COMPLEXITY, MEDIUM, ActPC, TOSM, and OM. Due to the correlation of COMPLEXITY and LENG_SYLL, we won't put that back in the model. If we build a model with MEDIUM, ActPC, TOSM, and OM, everything will have significance, but correlation test show that there's a fair amount of overlap, especially between CohPC, ActPC, TOSM, and OM (all pairings have correlation coefficients greater than 0.68). It's more parsimonious to just keep CohPC. What about a model that takes gries.glmm70 and adds MEDIUM? It doesn't quite achieve 0.05 significance (p=0.05447) and log likelihood tests show that it isn't any better than gries.glmm70 (p=0.05623)—it's close, though, so it's definitely worth considering it in future research.

⁴⁷ To do this by hand, you can create a data frame with the predictions. The data frame gives a fitted prediction for each phrasal verb (when it's greater than 0, the model predicts the split construction, when it is less than 0, it predicts the joined construction).

```

> gries.fit<-data.frame(fit=fitted(gries.glmm71), Verb, Prt, CONSTRUCTION_spelledout)

```

Appendix E: What else is an idiom besides non-compositional?

Everyone seems to agree that compositionality is a major part of idiomaticity, but there are disagreements about what else counts.⁴⁸ Here are some of the main additions to the definition of “idiom”. As my discussion of each suggests, I am not really happy with any of them. Nonetheless, something beyond compositionality seems to be required by the fact that IDIOM remains important in modeling the data from Gries (2002) even after compositionality is included.

- Non-productivity
 - Wood (1986) restricts idioms to expressions that are completely non-decomposable and which are completely incapable of being productive. Here, non-productivity allows in *eke out* (because *eke* is so restricted), *cock a snook*, and *cook x’s goose*.⁴⁹ Given the oddness of these, we probably do want to count them as idioms.
 - But by this logic, Wood says that *take offense* and *catch hell* are not idioms because they have variants like *take exception/umbrage* and *catch whatfor/it*.
 - I admire that Wood is so clear about the degrees of productivity (so that *eke out* is less productive than *send forth*, which is less productive than *nail up* or *go away*), but I think that productivity is basically just an aspect of compositionality.
 - Consider the end of the spectrum where something is truly non-decomposable—the classic example is *kick the bucket*. You cannot really expect people to know what you mean if you say *jump the bucket* or *kick the bowl*. To the extent that these are interpretable, you are relying on the fact that even at the non-decomposable end of the spectrum, individual words are still discernable. The more decomposable you are, the more productive you are because the collocated words make fewer demands on each other. At the far end, they may merely make a demand that you can’t *drink magenta* or *bricks*, but you can *drink tea/coffee/water/blood/something cold/anything in the fridge/the sights/etc*.
 - Under Wood’s plan we drastically reduce the number of phrases we call idioms. Consider that Moon (1998: 120) shows that 40% of the fixed expressions and idioms in English have at least one lexical or grammatical variant, and as many as 14% have two or more. Limiting the scope of what one will consider in a paper is one thing, but to define away such an enormous part of formulaic speech diminishes our theories and just introduces the necessity of more machinery to deal with the middle degrees. Why disconnect the really set phrases from the somewhat set phrases? To appreciate the patterns of formulaic language, it is better to preserve the continuum than to build fences.
- Familiarity

⁴⁸ Researchers tend to describe idiomaticity as involving at least several of these components, yet even those who create classification schemes have rather flat continua. Barkema (1996: 133-152) is an exception.

⁴⁹ It also lets in *runcible spoon*, which appears as a nonsense word in Edward Lear’s “The Owl and the Pussycat.” I don’t think that Wood was aware that Lear also used *runcible* with *hat*, *cat*, *goose*, and *wall*. I think these uses themselves would make *runcible* productive. Indeed, this is the problem with nonsense words: they are inherently productive, aren’t they? And not only for the coiner—in his song “Heather”, Paul McCartney sings, “I will dance to a runcible tune.”

- Familiarity must be a function of frequency, but it is not purely word counts. For example, in Moon (1998)'s 18 million word corpus she found no occurrences of *kick the bucket*, *out of practice*, or *by hook or by crook*. This points to a problem of using large corpora to estimate human experiences with language, but it also suggests that words and phrases may be tied to specific contexts and resonate more in those areas than in others. Familiarity is different than frequency, then, because words may have different relationships with each other as well as different strengths for those relationships.
- I would've expected that my measurements could have captured this, though something more directly related to frequency may be more appropriate than building up descriptions of paradigms.
- Ambiguity
 - For Weinreich (1969: 44), ambiguity is an essential characteristic, but this is an odd thing to propose since it would allow in *cook x's goose* (you might actually be roasting a bird), but wouldn't allow in *eke out* since it is so restricted that it's hard to imagine any ambiguity. Nor would it meaningfully include *cock a snook*, unless you are to claim that someone could think you were treating a fish like a pistol.
 - When it comes to polysemy, much is made of the fact that red hair isn't the same as red velvet, nor is white coffee the same as white wine. Polysemy does play a role—especially in the historical development of idioms—but this doesn't mean it needs to be definitional. Compositionality does capture polysemy since any words that compose a phrase will likely have multiple meanings and you will have to deal with which ones are selected. The question is how you select the meanings for the words that best match the meaning of the phrase.
 - Finally, ambiguity is generally not as hard for people as linguists in their armchairs make it out to be. In any real world language sample, even a brief look at the neighborhood of the supposedly ambiguous phrase will suffice to disambiguate it (Tschichold 2004: 168). Meanings are rarely confused because they occur in such different circumstances.
- Conventionality, institutionality, lexicality
 - This idea is called various things by various researchers, but it is basically the concept that idioms are made up of pieces that are highly predictable given one other. (Fernando and Flavell 1981: 44, Makkai 1972: 160-161, Fawley and Syder 1983: 211.)
 - By building out an idea of compositionality in terms of frequency, I had expected to capture these notions. There are certainly other information theoretic terms this brings to mind, for example “mutual predictability” and “mutual expectancy”.
- Frozenness, fixedness, fossilization
 - The idea is that you can't substitute synonyms and you can only rarely rearrange the words. (See Skandera 2004: 26-27 for more.) It is also related to the idea of non-productivity, which tends to be more about word substitution while this is about syntactic modifications.
 - This is a rather weak formulation, but speaks to the “transformational deficiency” that people saw in idioms. Yet we know from corpus investigations that most idioms can be switched—intensifiers can be added, they can appear in the passive.

- But Bush's father, George H.W., was then a U.S. congressman from Houston, and strings were pulled.⁵⁰
- Close tabs were kept on them, including forage and browse studies within the pens, but the plan soon went awry.⁵¹

See also Nunberg et al (1994: 492-493) for their “more-or-less orthogonal properties”, which include conventionality, inflexibility, figuration, proverbiality, informality, and affect”. Several of these have to do with the circumstances and consequences surrounding a speaker who uses an idiomatic expression. For them, idioms “occupy a region in a multidimensional lexical space, characterized by a number of distinct properties: semantic, syntactic, poetical, discursive, and rhetorical” (1994: 492). Having failed in my own attempts to reduce idiomaticity to purely compositionality, I also come to the conclusion that idiomaticity is more like a syndrome of multiple properties, not a synonym for one.

⁵⁰ <http://www.straightdope.com/columns/030411.html>

⁵¹ <http://www.nps.gov/archive/thro/adhi/adhi9.htm>

Bibliography

- Allan, M. 1978. Morphological Investigations. Ph.D. dissertation, University of Connecticut, Storrs.
- Allopenna, P., J. Magnuson, and M. Tanenhaus. 1998. [Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models](#). *Journal of Memory and Language*, 38, 419-439.
- Aronoff, M. 2007. In the beginning was the word. *Language* 83: 803-830.
- Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Baayen, R. H. In press. [Analyzing Linguistic Data. A Practical Introduction to Statistics Using R](#). Cambridge: Cambridge University Press.
- Baayen, R. H. and R. Lieber 1991. Productivity and English Derivation: A Corpus Based Study. *Linguistics*, 29:801-843.
- 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*.
- Barkema, H. 1996. Idiomaticity and terminology: A multi-dimensional descriptive model. *Studia Linguistica*, 50(2):125-60.
- Bod, R. 2006. Exemplar-based syntax: How to get productivity from examples. *Linguistic Review* 23: 291-320.
- Bod, R. and D. Cochran. 2007. Introduction to Exemplar-Based Models of Language Acquisition and Use. *ESSLI Summer Workshop*.
- Bolinger, D. 1971. *The phrasal verb in English*. Cambridge: Harvard University Press.
- Breheny, R. 2003. On the dynamic turn in the study of meaning and interpretation. In J. Peregrin (ed.), *Meaning: The Dynamic Turn*. Amsterdam: Elsevier (Crispi Series).
- Bresnan, J. A. Cueni, T. Nikitina, and H. Baayen. 2007. [Predicting the Dative Alternation](#). In G. Boume, I. Kraemer, and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science: 69-94.
- Burnard, L. (ed.). 2007. *Reference Guide for the British National Corpus (XML Edition)*. Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. <http://www.natcorp.ox.ac.uk/XMLedition/URG/index.html>
- Bybee, J. and J. Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of 'don't' in English. *Linguistics* 37: 575-596.
- Chafe, W. 1970. *Meaning and the Structure of Language*. Chicago: University of Chicago Press.

- Cover, T. and J. Thomas. 2006. Elements of information theory. New York: Wiley-Interscience.
- Dalgaard, P. 200. Introductory Statistics with R. New York: Springer.
- de Hoop, H, P. Hendriks, and R. Blutner. 2004. [A new hypothesis on compositionality](#). In P. P. Slezak (ed.), Proceedings of the Joint International Conference on Cognitive Science. Sydney: ICCS/ASCS 2004.
- Evans, G. 1982. The Varieties of Reference. Oxford: Clarendon Press.
- Fabb, N. 1988. English suffixation is constrained only by selectional restrictions. *Natural Language and Linguistic Theory* 6: 527-539.
- Fodor, J. A. and Z. W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28: 3-71.
- Fraser, B. 1976. The Verb-Particle Combination in English. The Hague: Mouton.
- Fraser, B. 1974. The phrasal verb in English. By Dwight Bolinger. *Language* 50: 568-75.
- Frege, G. 1884. Grundlagen der Arithmetik. Eine logisch-mathematische Untersuchung über den Begriff der Zahl. W. Koebner, Breslau.
- Gahl, S. and S. Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation change. *Language* 80: 748-775.
- Giegerich, H. 1999. Lexical strata in English: Morphological causes, phonological effects. Cambridge: Cambridge University Press.
- Gibson, E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68.1: 1-76.
- 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. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 36(6): 1041-1070.
- Hay, J. 2000. Causes and consequences of word structure. Chicago: Northwestern University dissertation.
- Hay, J. and R. H. Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Science* 9: 342-348.
- Hay, J. and R. H. Baayen. 2002. Parsing and Productivity. In Booij, G. E. and J. v. Marle (eds.), *Yearbook of Morphology 2001*, Kluwer Academic Publishers, Dordrecht. 203-235.
- Hay, J. and J. Bresnan. 2006. "Spoken syntax: The phonetics of 'giving a hand' in New Zealand English." *The Linguistic Review* 26: 321-349.
- Haiman, J. 1994. Ritualization and the development of language. In William Pagliuca (ed.), *Perspectives on Grammaticalization*. Amsterdam: John Benjamins. 3-28.

- Hopper, P. 1998. Emergent Grammar. In M. Tomasello (ed.), *The New Psychology of Language*. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers. 155-175.
- Johnson, K. 2004. On the systematicity of language and thought. *Journal of Philosophy* 101: 111–139.
- Katz, J. and P. Postal. 1963. Semantic interpretation of idioms and sentences containing them. *Quarterly Progress Report 70* (MIT Research Laboratory of Electronics): 275-282
- Kennedy, A. G. 1920. The Modern English verb-adverb combination. *Language and Literature*, 1(1): 1-51.
- Kiparsky, P. 1982. Lexical morphology and phonology. In I. Yang (ed.), *Linguistics in the morning calm*. Seoul: Hanshin Publishing. 1-91.
- Kuperman, V. 2007. Visual processing of particle verbs in Dutch: An experimental approach. *Stanford Psychology of Language Tea*, 21 February 2007.
- Landauer, T. 1986. How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science* 10: 477–493.
- Lohse, B., J. Hawkins, and T. Wasow. 2004. "[Processing Domains in English Verb-Particle Constructions](#)". *Language* 80(2): 238-261.
- McCarthy, D., B. Keller and J. Carroll. 2003. [Detecting a Continuum of Compositionality in Phrasal Verbs](#), In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.
- Merrill, J. 1993. *The Changing Light at Sandover*. New York: Alfred A. Knopf.
- Montague, R. 1970. *Universal Grammar*. Reprinted in R. Thomason (ed.), *Formal Philosophy*. New Haven : Yale University Press, 1974: 222–246.
- Moon, R. 1998. *Fixed Expressions and Idioms in English: a Corpus-based Approach*. Oxford: Oxford University Press.
- Moscoso del Prado Martín, F., Kostić, A., and Baayen, R.H. 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.
- Sag, I. and T. Wasow. To appear. Performance-Compatible Competence Grammar. In R. Borsley and K. Börjars, (eds.) *Non-Transformational Syntax*.
- Siegel, D. 1979. *Topics in English morphology*. New York: Garland.
- Shannon, C. and W. Weaver. 1949. *The Mathematical Theory of Communication*. Univ of Illinois Press.
- Spenader, J. and R. Blutner. 2007. [Compositionality and systematicity](#). In G. Bouma, I. Krämer, and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*. Amsterdam: KNAW publications.

Szabó, Z. 2007. 'Compositionality' in Stanford Encyclopedia of Philosophy.
<http://plato.stanford.edu/entries/compositionality/>.

Toutanova, K. and C. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000).

van Gelder, T. 1990. Compositionality: A connectionist variation on a classical theme. *Cognitive Science* 14: 355–384.

Weinreich, U. 1972. *Explorations in Semantic Theory*. The Hague: Mouton.

Weinreich, U. 1969. Problems in the analysis of idioms. In J. Puhvel (ed.), *Substance and Structure of Language*, Berkeley, California: University of California Press. 23-81.

Wood, M. 1986. A definition of idiom. Master's thesis, University of Manchester (1981). Reproduced by the Indiana University Linguistics Club.