Phonological similarity in multi-word units*

STEFAN TH. GRIES

Abstract

In this paper, I investigate the phonological similarity of different elements of the phonological pole of multi-word units. I discuss two case studies on slightly different levels of abstractness. The first case study investigates lexically fully-specified V-NP_{DirObj} idioms such as kick the bucket *and* lose one's cool*; the idioms investigated are taken from the Collins Cobuild Dictionary of Idioms (2002). The second case study investigates the lexically less specified* way-construction*, which is exemplified by* He fought his way through the crowd *(cf. Goldberg 1995: Ch. 9), on the basis of data from the British National Corpus 1.0.*

I show that both patterns exhibit a strong phonological within-pole relation, namely a strong preference for having their slots filled with phonologically similar elements, where phonological similarity is manifested in alliteration patterns. These preferences are statistically significant when compared to chance-level expectations derived both from the corpora and from the CELEX database (Baayen et al. 1995) and are explained on the basis of Langacker's concept of syntactic and phonological constituents as well as current exemplar-/usage-based approaches.

Keywords: *Semantic and phonological constituents, semantic and phonological poles, alliteration, corpus data, CELEX.*

[*] I thank audiences at CSDL 2006 and 2010 (in particular Luca Onnis) as well as the anonymous reviewers and the editors for their comments. The usual disclaimers apply. Email address for correspondence: <stgries@linguistics.ucsb.edu>.

Cognitive Linguistics 22–3 (2011), 491–510
DOI 10.1515/COGL.2011.019

0936–5907/11/0022–0491
© Walter de Gruyter

1. Introduction

A central notion within many frameworks that can be subsumed under the heading of Cognitive Linguistics is that of a unit. In Langacker's Cognitive Grammar, for example, a unit is defined as

> a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement [. . .] he has no need to reflect on how to put it together. (Langacker 1987: 57)

In Cognitive Grammar, the linguistic system is argued to consist of symbolic units, i.e., units that are conventionalized associations of a phonological pole (i.e., a phonological structure) and a semantic/conceptual pole (i.e., a semantic/conceptual structure). The notion of a symbolic unit is not restricted to morphemes or words, but also comprises more abstract grammatical patterns. More specifically, symbolic units as defined above can exhibit different degrees of complexity: they can range from morphemes or monomorphemic words to polymorphemic words, fully-fixed multi-word expressions, partially-filled multi-word expressions, up to lexically completely unspecified syntactic and/or argument structure constructions. The more often a speaker/hearer encounters a symbolic unit, the more entrenched this unit becomes in his linguistic system and the more automatically the unit is accessed. Thus, unit status correlates positively with a speaker/hearer not analyzing the internal structure of a unit.

The relations between the parts of a unit can be explored in two different ways. First, the relation between the phonological pole(s) and the semantic pole(s), which I refer to as *between-pole relation*, is usually not motivated such that the conceptual content of a unit is not predictable from the phonological unit with which it forms a symbolic unit; this is of course the arbitrariness of the sign as already discussed by de Saussure. However, between-pole relations and the notions that are relevant in their discussion—arbitrariness, iconicity, and motivation (cf. Van Langendonck 2007 for a recent overview)—are *not* my concern here (but cf. below, in particular n. 12). This paper is about what I will call *within-pole relations*. In Cognitive Linguistics, there has been a lot of work on relationships holding between the different parts of a unit's semantic pole, but there has been little work that specifically addresses phonological within-pole relations. Sometimes, however, such relations surface in surprisingly clear ways. In Gries (2006a), I studied the verb *to run* on the basis of corpus data and noted that most of the idiomatic expressions that *to run* participated in involved alliterations: *to run rampant*, *to run riot*, *to run roughshod*, *to run the risk*, *to run into rapture*. This unexpected phenomenon stimulated the exploration reported on in this paper. More specifically, I explore to what degree this is an isolated instance or whether this is actually a more widespread phenomenon in need of an explanation. In this paper, I therefore investigate this specific kind of phonological within-pole relations, namely the relation of phonological similarity of different elements of the phonological pole of symbolic units. I discuss two case studies on slightly different levels of abstractness. The first case study in Section 2 investigates lexically fully-specified V-NP$_{\mathrm{DirObj}}$ idioms such as *kick the bucket* and *lose one's cool*; the idioms investigated are from the *Collins Cobuild Dictionary of Idioms*. The second case study in Section 3 investigates the *way*-construction exemplified by *He fought his way through the crowd*, which is lexically only partially specified: the direct object must be *way*, but many different verbs can be inserted into the verb slot. This construction will be investigated on the basis of data from the British National Corpus 1.0. As I will show below, both case studies show a strong and statistically highly significant alliteration effect: verbs and DO head nouns in V-NP$_{\mathrm{DirObj}}$ idioms are much more often alliterative than expected by chance, and verbs in the *way*-construction are much more likely to begin with [w] than expected by chance; in both case studies, chance is computed in four different ways. In addition, the verb and the head nouns in alliterative V-NP$_{\mathrm{DirObj}}$ idioms exhibit markedly larger degrees of collocational attraction to each other than in non-alliterative idioms; similarly, the verb of the *way*-construction is much more strongly attracted to the *way*-construction if it begins with [w]. Section 4 will discuss motivations for, and implications of, these findings and draw conclusions.

2. Lexically-specified V-NP$_{\mathrm{DirObj}}$ idioms

2.1. *Data and methods*

The kind of idioms to be investigated in this section are lexically-specified V-NP$_{\mathrm{DirObj}}$ idioms. The 211 lexically-specified V-NP$_{\mathrm{DirObj}}$ idioms to be investigated here can be characterized as follows:

- they feature a full lexical verb;
- they feature an NP as a direct object of said verb;
- the V usually takes no further complements or adjuncts;
- the idiom is reasonably frequent, which is operationalized such that it occurs at least 105 times in the 211 m word corpus on which the Collins Cobuild Dictionary of Idioms is based.[1]

[1] This frequency threshold is based on the dictionary's ordinal frequency labeling of idioms: idioms with a frequency of once per 2 million words or higher were marked with three diamonds and included. This yields 211 idioms with a frequency of 105 or more. Thanks to Stefanie Wulff for making this list available to me.

Some representative examples are listed in (1).

(1) a. spill the beans
b. gain some ground
c. get the boot
d. lend a hand
e. bite the bullet

These idioms were explored according to the following four methodological steps:

(i) measuring the amount of alliteration effects for the idioms; and for comparison;
(ii) computing baseline amounts of alliteration that are based on the word-initial phonemes and their frequencies; these baseline computations can be and were done in three different ways;
(iii) computing a baseline amount of alliteration based on a control group of non-idiomatic V-NP$_{DirObj}$ sequences;
(iv) computing collocational statistics for the verbs and nouns in the idioms and control structures.

As for step (i), for each of these idioms, I noted the initial segment of the verb and the initial segment of the head noun of the NP$_{DirObj}$. In the case of *bite the bullet*, this means noting [b] and [b]; for *lose face*, it means [l] and [f]; etc. If the NP$_{DirObj}$ also involved additional content words as part of the idiom, the initial segments of these were also noted.[2] Thus, for *fight a losing battle*, the three pairs [f] [l], [f] [b], and [l] [b] were noted; similarly, for *keep a straight face*, I noted [k] [s], [k] [f], and [s] [f]. All pronunciations of words were straightforward to code and automatically extracted from the phonological data available in the CELEX database (cf. below) and I then counted the observed number of alliterations, i.e., the number of instances where one content word in the V-NP$_{DirObj}$ idiom begins with the same sound as another content word.

As for step (ii), it is clear that the observed percentage of alliterations must be compared to some kind of baseline to determine whether it is greater or less than chance would lead one to expect. However, there are at least three different ways in which this expected baseline frequency can be computed:

- without regard to any frequencies;
- with regard to type frequencies;
- with regard to token frequencies.

[2] The notion of *content word* is used here merely as a convenient traditional cover term for nouns, verbs, adjectives, and adverbs.

In what follows, each of these methods will be characterized briefly.[3] The logic of the first method is this: each word in the phonological part of the CELEX database (<EPW.CD>) begins with one out of 47 different phonemes; these run the whole gamut from highly frequent consonant phonemes to much less frequent diphthongs. Thus, there are 47 · 47 different possible combinations of word-initial segments of two words. Of these 47 · 47 different possible combinations of word-initial segments of two words, 47 will involve the same segment at the beginnings of both words; thus, the expected baseline percentage is 1 ÷ 47.

While this method is simple and straightforward, it also comes with one big problem: it does not take into consideration the frequencies with which each of the 47 phonemes occurs word-initially in differently frequent words. Thus, the high likelihood of two *s*'s at the beginning of words—because [s] is a highly frequent segment—is severely downplayed. This may therefore strongly decrease the expected baseline percentage and, thus, lead us to believe in an effect that a more careful operationalization would not identify. The second method therefore takes frequencies of *types* into account. There are 87,263 different word types in the CELEX database, each beginning with one of 47 different phonemes. Thus, one can use each phoneme's probability to occur *type*-initially in the computation of the expected baseline percentage such that

– the probability *p* that both words of a V-NP$_{\text{DirObj}}$ idiom begin with [s] is the squared percentage of word types starting with [s], or more formally: p_{type}([s. . .] . . . [s. . .]) = p_{type}[s. . .] · p_{type}[s. . .];
– p_{type}([s. . .] . . . [t. . .]) = p_{type}[s. . .] · p_{type}[t. . .];
– p_{type}([s. . .] . . . [ɪə. . .]) = p_{type}[s. . .] · p_{type}[ɪə. . .]; etc.

Of course, the first two examples are rather likely whereas the third one is not. In this method, the expected baseline percentage is therefore the sum of all probabilities of all pairs with identical segments.

The third method is very similar to the second, but differs in one crucial respect. In the previous method, the probabilities *p* were based on the frequencies of word types with initial phonemes in the CELEX database. This, however, disregards the frequencies with which these types occur. The third method, therefore, goes yet another step further and also includes the *token* frequencies of the relevant words. There are 18,580,121 word tokens in the CELEX database, again each beginning with one out of 47 different phonemes. Thus, one can use each phoneme's probability to occur *token*-initially in the computation of the expected baseline percentage such that

[3] All the data extraction, computations, and graphs were performed/created with R for Windows 2.11.1 patched (cf. R Development Core Team 2010).

- the probability p that both words of a V-NP$_{\mathrm{DirObj}}$ idiom begin with [s] is the squared percentage of word tokens starting with [s], or more formally: p_{token}([s. . .] . . . [s. . .]) = p_{token}[s. . .] · p_{token}[s. . .];
- p_{token}([s. . .] . . . [t. . .]) = p_{token}[s. . .] · p_{token}[t. . .];
- p_{token}([s. . .] . . . [ɪə. . .]) = p_{token}[s. . .] · p_{token}[ɪə. . .]; etc.

with the only difference to the above being that now the percentage is based on token—not type—frequencies. Again, the expected baseline percentage is therefore the sum of all probabilities of all pairs with identical segments.[4]

As for step (iii), I randomly sampled two transitive clauses from each of the 170 spoken data files whose names began with S1A or S2A in the fully parsed British Component of the International Corpus of English (ICE-GB) and counted alliterations in this control group of V-NP$_{\mathrm{DirObj}}$ structures as above for the idioms (i.e., including adjectival modifiers etc.).

Finally as for step (iv), I explored whether the idioms—both with and without alliterations—exhibited higher collocational attractions between the verb and the noun. To that end and as an approximation, I retrieved the frequency in the British National Corpus World edition of each verb and noun lemma from the idioms and the control verbs as well as the number of times they co-occurred in the same sentence. These frequencies were then used to compute two measures of collocational strength, Mutual Information (*MI*) and the *t*-score, since these are known to exhibit very different statistical behavior and, thus, cover different possible outcomes. These measures of collocational strength were then used as dependent variables, while the independent variables were V-NP$_{\mathrm{DirObj}}$ group (idiom vs. control) as well as Alliteration (yes vs. no).

2.2. *Results and interim conclusion*

The V-NP$_{\mathrm{DirObj}}$ idioms contained 35 alliterations out of 310 content word pairs (211 lexically specified idioms many of which had additional content words that were added to the overall number of content word pairs as explained above). Consequently, the observed percentage of alliterations is 35 ÷ 310 = 0.1129 and some random examples are listed in (2).

[4] There are actually analogous ways to compute baseline percentages which are not based on the (type or token) frequencies of word-initial segments, but which are based on the (type or token) frequencies of segments *anywhere in the word*. However, this would introduce strong biases into the computation. Since this study is concerned with alliteration effects which by definition occur word-initially, it is less than desirable to deal with co-occurrence frequencies of [s] / [z] and [ŋ], which are strongly inflated due to the role these phonemes play in plural or progressive-*ing* suffixes (especially the latter would be particularly problematic since [ŋ] does not even occur word-initially in English.

(2) a. bite the bullet
b. burn bridges
c. gain ground
d. make a mark
e. turn the tables

Alliterations with *s* (7) and *b* (6) were most common, but let us now look at the results of the three baseline computation approaches of step (ii), which involved the phonemes and their frequencies. As for the first method, in the CELEX database, the number of different phonemes is 47 so that, according to method 1, the expected baseline percentage is 1 ÷ 47 = 0.0213. As for the second method, I generated a symmetric 47 × 47 co-occurrence matrix that contained all 47 phonemes in the rows and in the columns and the probability of co-occurrence in the word-initial slots for each of the $47 \cdot 47 = 2{,}209$ possible combinations of two phonemes in the cells. For example, p([s. . .]) is 0.1186 so the cell for [s] / [s] contains p([s. . .] . . . [s. . .]), which amounts to approximately $0.1186 \cdot 0.1186 = 0.0141$ etc. Summing up the main diagonal, which contains the positions where the first and second phoneme are identical, results in the expected baseline percentage, which turns out to be 0.0595. This figure is higher than 1 ÷ 47 since it now includes the information that some phoneme repetitions are rather likely, given the high word-initial frequencies of phonemes such as [s], [t], etc. As for the third method, the logic is exactly the same, and the sum of the main diagonal of the 47 × 47 co-occurrence matrix yields an expected baseline percentage of 0.0473.

Step (iii), the analysis of control V-NP$_{\text{DirObj}}$ structures from the ICE-GB spoken data yielded altogether 32 alliterations out of 667 content word pairs, i.e., a proportion of 0.04798.

These results are represented in Figure 1 and can be summarized very straightforwardly.

The observed tendency for alliterations in the analyzed lexically-specified V-NP$_{\text{DirObj}}$ idioms is indicated by the solid horizontal line. It is between 1.9 and 5.3 times as strong as expected, and all differences between the observed baseline and the three baseline percentages (the three bars from the left) as well as the non-idiomatic V-NP$_{\text{DirObj}}$ structures are highly significant according to exact binomial tests (all p's < 0.001).[5]

[5] An exact binomial test is a test that can be used to determine the probability to get *n* or more white balls out of an urn when one draws *d* balls (with replacement) and the urn contains *x* white and *y* red balls. For example, if an urn contains $x = 10$ red and $y = 10$ white balls, the probability to draw a red ball (with replacement) is always 0.5. To now compute how likely it is that you get $n = 3$ or $n = 4$ red balls when you draw $d = 4$ balls (with replacement) from that urn, you can enter the following `sum(dbinom(`*n*, *d*, $x/(x+y) = p$(red ball))) into R: `sum(dbinom(3:4, 4, 0.5))`, which would return that this probability is 0.3125. Thus, the

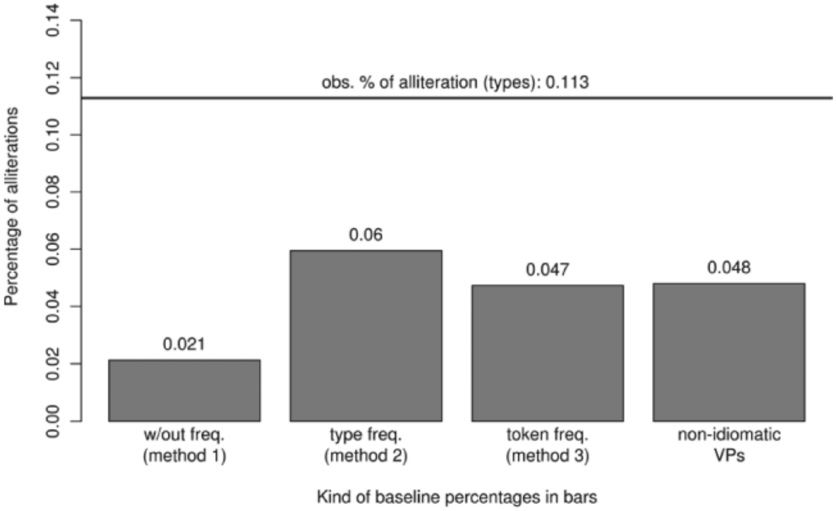

Figure 1. *Observed and expected percentages of alliterations in V-NP$_{DirObj}$ idioms*

Now, what about step (iv), the collocational attraction of verbs and nouns in the four groups? These data are somewhat difficult to evaluate with the usual statistical tools since they involve (i) very different sample sizes and (ii) very heterogeneous data. This is because (i) there are many more non-alliterating idioms than alliterating ones and many more control V-NP$_{\mathrm{DirObj}}$ structures than idioms, and (ii) collocation strengths can be extremely variable across all groups, which makes their variances so large as to not allow ANOVAs or similar procedures. (The ANOVAs I ran were in fact characterized by many violations of distributional assumptions.) The problem is exacerbated by the fact that the tested hypothesis only affects, and hence constrains, the collocation strengths of the lowest-frequency sets of items—alliterative idioms—whereas the collocation strengths of the other sets will run the whole gamut from very high to very low. Put differently, the data require that one test a small high-variance sample to a partially-overlapping even higher-variance sample. I am therefore using a descriptive approach and summarize the results with medians in Figure 2.

exact binomial test is the better (since exact) counterpart of a chi-square goodness-of-fit test applied to a binary variable. In this case, `chisq.test(c(3, 1), p=c(0.5, 0.5))` would return the very similar *p*-value of 0.3173; cf. Sheskin (2007) or Gries (2009) for details. The exact binomial tests were therefore computed as follows:

- method 1: `sum(dbinom(35:310, 310, 1/47))` = 2.011489e-15;
- method 2: `sum(dbinom(35:310, 310, 0.05950193))` = 0.0002375707;
- method 3: `sum(dbinom(35:310, 310, 0.04727673))` = 2.228605e-06.

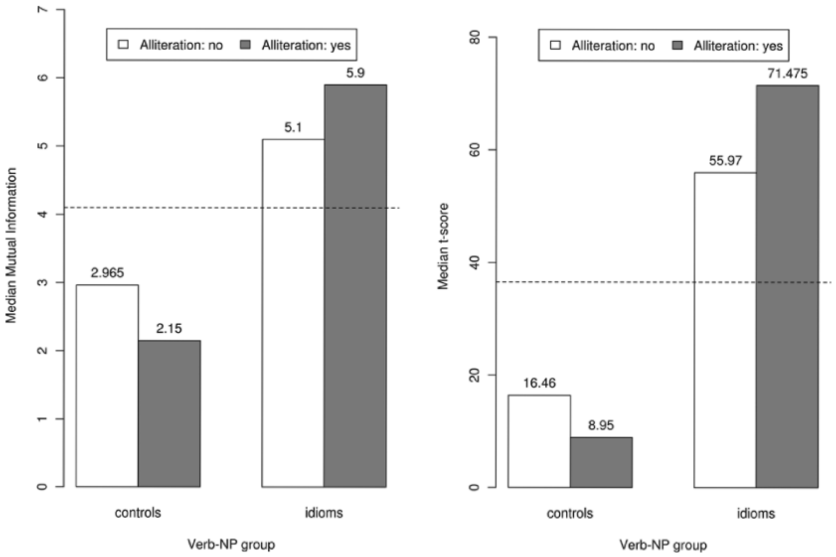

Figure 2. *Collocational attractions in V-NP$_{DirObj}$ idioms as a function of V-NP structure type and alliterations (left panel: MI, right panel: t)*

Both measures yield very similar results: Unsurprisingly, the idioms exhibit higher collocational strengths than the control V-NP$_{DirObj}$ structures. More interesting, however, is the suggestive interaction: Idioms exhibit a higher median collocational strength than controls, but in idioms the alliterative expressions have a higher collocational strength than the non-alliterative ones, whereas in controls this effect is reversed. In other words, the components of V-NP$_{DirObj}$ idioms exhibit the highest collocational attraction if they also alliterate.

3. The *way*-construction

3.1. *Data and methods*

The construction to be investigated in this section is the lexically partially specified *way* construction (cf., among others, Goldberg 1995: Ch. 9). The *way*-construction has the formal characteristics listed in (3a) and is exemplified in (3b) and (3c); cf. Stefanowitsch and Gries (2005) for comprehensive corpus data.

(3) a. SUBJ$_{theme}$ V$_{move}$ POSS *way* [$_{PP}$ P NP/S]$_{path}$
b. . . . as the British Task Force made its way across the Atlantic. (BNC: FNX)

c. . . . some expanse of water found its way into the picture. (BNC: C9W)

The semantics of the construction can be characterized as follows: the referent of the subject moves along, or creates, the path denoted by the PP, and the movement of the referent of the subject usually does not come easily in the sense that it involves laboriously circumventing or forcefully overcoming some obstacles on the way or creating the path in the first place. As with the V-NP$_{\text{DirObj}}$ idioms, several methodological steps are necessary:

(i) measuring the amount of alliteration effects for the *way*-constructions (once with verb types, once with verb tokens); and for comparison;
(ii) computing baseline amounts of alliteration that are based on the word-initial phonemes and their frequencies; these baseline computations were done in three different ways;
(iii) computing a baseline amount of alliteration based on a control group of transitive VPs with *way* as the direct object;
(iv) computing the collostructional attractions of the verbs to the *way*-construction.

As for step (i), the data set used in the present analysis is based on a concordance of any verb followed by any possessive pronoun (tag: DPS) followed by *way* in the British National Corpus 1.0 (British National Corpus Consortium 1994), which was subsequently cleaned manually; this procedure yielded 5,831 instances of the *way*-construction.[6] For each of these constructions, I noted the initial segment of the verb in the verb slot. In the case of *banged her way*, this means noting [b]; for *wound your way*, it means [w] etc. Again, all pronunciations of words were obvious and automatically extracted from the CELEX database. The observed percentage of alliterations was then computed in two ways, one for types and one for tokens. For types, I counted the number of verb types beginning with [w] and divided this number by the number of all verb types. For tokens, I counted the number of verb tokens beginning with [w] and divided this number by the number of all verb tokens.

As for step (ii) and as before, one needs expected baseline percentages against which we can compare the observed percentages; again as before, there are three different ways to arrive at such baseline percentages: without frequency information, with type-based frequency information, and with token-based frequency information. As for the first method, all verbs in the phonological part of the CELEX database start with one out of 39 different phonemes, which is why the expected baseline percentage will be 1 ÷ 39. For the same reasons as above, however, we will also want to compute expected baseline

[6] This data set is the one used in Stefanowitsch and Gries (2005).

percentages on the basis of the frequencies of the verbs in the CELEX database. Thus and as for the second method, there are 8,504 different verb *types* in the CELEX database. Of these 8,504 different verb types, x will begin with [w], so the expected baseline percentage will be $x \div 8{,}504$. As for the third method, there are 3,310,984 different verb *tokens* in the CELEX database. Of these, x will begin with [w], so the expected baseline percentage will be $x \div 3{,}310{,}984$.[7]

As for step (iii) and to be able to compare the *way*-constructions to cases where *way* is a direct object but not part of the *way*-construction, I also retrieved all instances of *way* used as a direct object in a transitive VP from the ICE-GB and annotated those that were not *way*-constructions for whether the first segment of the verb was a [w] or not.

Finally for step (iv), I computed a collexeme analysis for the *way*-construction using the data from Stefanowitsch and Gries (2005). That is, for each verb attested in the *way*-construction, I computed how much it 'likes' to occur in the *way*-construction given its overall frequency of occurrence. Again, to cover different measures of association, I used a bi-directional measure ($-\log_{10} p_{\text{Fisher-Yates exact}}$ as used by, say, Stefanowitsch and Gries 2003 or Gries et al. 2005) and a uni-directional measure, namely ΔP (cf. Ellis and Ferreira-Junior 2009). Then I computed the median degrees of attraction for alliterative and non-alliterative *way*-constructions.

3.2. *Results and interim conclusion*

The *way*-construction data from the BNC contained 32 alliteration types out of 492 types, i.e., an observed alliteration percentage of $32 \div 492 = 0.065$. For the tokens, I found 764 alliteration cases out of 5,831, i.e., an observed alliteration percentage of $764 \div 5{,}831 = 0.131$.

Let us again now look at the results of step (ii), the three methods to compute the baseline percentages. As for the first method, in the CELEX database, there are 39 different phonemes that verbs begin with; thus, the expected baseline percentage is $1 \div 39 = 0.0256$. As for the second method, $x = 226$, i.e., there are 226 verb types beginning with [w] in the CELEX database so the expected baseline percentage is $226 \div 8{,}504 = 0.0266$. As for the third method, $x = 167{,}254$, i.e., there are 167,254 verb tokens beginning with [w]

[7] The frequencies for this comparison were also based on CELEX as opposed to the BNC because (i) that makes sure that the source of the type and token frequencies used for computing the baselines is the same in both case studies and (ii) it is virtually impossible to use the BNC for this in the first place: contrary to the CELEX database the more than 900,000 word types of the BNC are not phonologically annotated so that no (semi-)automatic extraction of their pronunciation is possible.

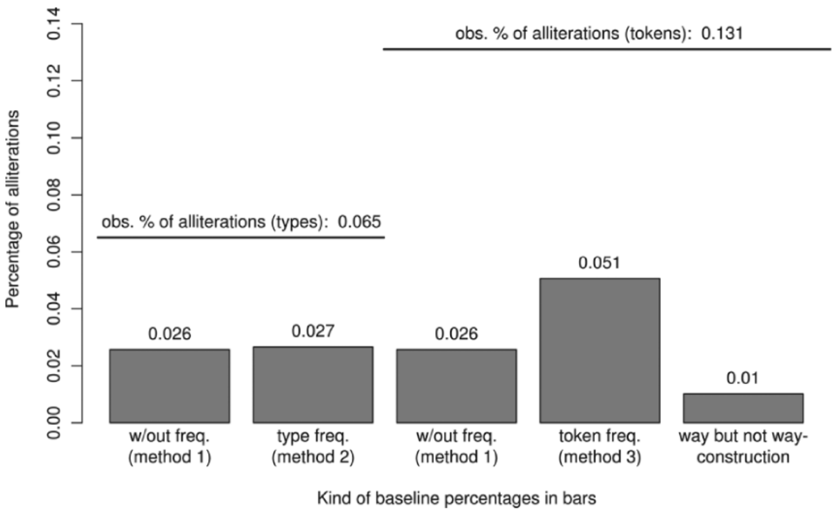

Figure 3. *Observed and expected percentages of alliterations in* way-*constructions*

in the CELEX database so the expected baseline percentage is 167,254 ÷ 3,310,984 = 0.0505.

As for step (iii), the search of *way* used as a direct object in a transitive VP yielded 99 instances that were not *way*-constructions, of which only one featured an alliteration; this baseline is therefore 1 ÷ 99 = 0.01.[8]

As in the case of V-NP$_{\text{DirObj}}$ idioms, these results can be summarized very straightforwardly; cf. Figure 3, the type alliteration data are represented on the left, the token data on the right. The observed tendency for alliterations in the partially lexically-specified *way*-construction for types and tokens are again indicated by the solid horizontal lines. For types, the percentage of alliterations is 2.54 times as high as the baseline percentage computed without regard to frequencies and 2.44 times as high as the baseline percentage computed with regard to type frequencies; both of these differences are highly significant according to exact binomial tests.[9] For tokens, the percentage of alliterations is 5.1 times as high as the baseline percentage computed without regard to frequencies, 2.59 times as high as the baseline percentage computed with regard to token frequencies, and 13 times as high as in the non-*way*-constructions;

[8] For these uses of *way*, no type counts were made because it is not clear how much in common these uses would have to have to constitute different types.

[9] The exact binomial tests for the type-based tests were computed as follows:

- method 1: `sum(dbinom(32:492, 492, 1/39))` = 2.323575e-06;
- method 2: `sum(dbinom(32:492, 492, 226/8504))` = 4.797799e-06.

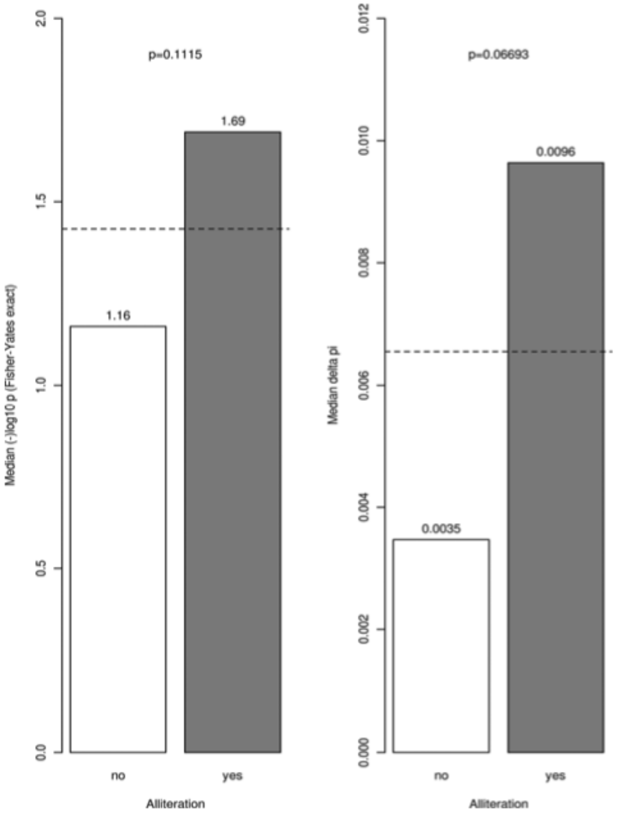

Figure 4. *Median association strengths for alliterative as well as non-alliterative* way-*constructions (left panel: –*$log_{10}\, p_{Fisher\text{-}Yates\ exact}$*; right panel:* ΔP*)*

all of these differences are highly significant according to exact binomial tests.[10]

With regard to step (iv), the median collostruction strengths are represented in Figure 4. Again, the data exhibit a large amount of heterogeneity and the values of interest are from the lowest-frequency set of items: alliterative *way*-constructions. As before, the results are suggestive. While they do not reach standard levels of significance, the median collostruction values are higher for

[10] The exact binomial tests for the token-based tests were computed as follows:

- method 1: `sum(dbinom(764:5831, 5831, 1/39))` = 8.516058e-292;
- method 3: `sum(dbinom(764:5831, 5831, 167254/3310984))` = 1.302444e-123;
- non-*way*-constructions: `sum(dbinom(764:5831, 5831, 1/99))` = 0.

the alliterative *way*-constructions: for ΔP, the alliterative *way*-constructions' median is nearly three times as high as for the non-alliterative ones. For the other measure, the difference is not as pronounced, but it may be interesting to point out that the median of the alliterative *way*-constructions is above the collostruction strength value that represents significant attraction (namely 1.301, as $-\log_{10}$ of 0.05 = 1.301) whereas the other one is below that.

4. Conclusions and outlook

For two kinds of symbolic units differing in schematicity, fully lexically-specified V-NP$_{\text{DirObj}}$ idioms and partially lexically-specified *way*-constructions, the results are unambiguous:

- there are strong alliteration effects;
- these differ significantly from baselines regardless of how expected and/or observed frequencies are computed;
- these differ significantly from non-conventionalized but otherwise parallel structures;
- these are weakly but suggestively correlated with measures of collocational/collostructional attraction, which they appear to reinforce.

These findings raise several questions. First, does this phenomenon serve some function? Second, if so, which function is that? Third, why is this effect observable in the form of alliterations? I do not have definitive answers to all these questions, but some speculations are possible, and the speculations regarding the questions are interrelated and compatible with an exemplar-/usage-based approach. I believe this phenomenon is not just an aberration or an accident and also not just due to subconscious priming effects. The lexically-specified idioms may at least in part be due to alliteration effects in the sense that the alliteration facilitated the lexicalization of some of the idioms. The *way*-construction, by contrast, is of course not due to lexicalized word-play—since the filler of the verb slot may vary—but the frequencies of some verbs may be boosted by the alliteration effect.

For both kinds of construction, one may speculate that, at some point in time, people created an expression, and because of the alliteration effect, the creation was both fun to produce and easy to memorize (if only unintentionally) and continued to be used until it became entrenched enough to be part of the language system, a process not unlike that undergone by, for example, new subtractive word-formations such as blends and complex clippings such as *chunnel*, *foolosopher*, etc. This account would fit in well with two different notions or strands of research in usage-based cognitive linguistics.

First, the present findings fit in with the growing recognition of the relevance of analogy and similarity for language learning and processing as well as the role of chunking (particularly for prefabs) in contemporary exemplar-/usage-based approaches. (I am following Gentner and Markman's [1997: 48] definitions.) For instance, it is well-known that (i) similarity of novel utterances to conventionalized utterances is correlated with the novel utterances' acceptability (Bybee 2010: 59), that (ii) a structure *S* at some point in time in a text/discourse primes a itself again at a later point more if the structures as well as how its slots are filled exhibit similarity (cf. Gries 2005, Szmrecsanyi 2005, Snider 2009), and that (iii) similarity on various levels (graphemic, segmental, phonemic, syllabic) facilitates the emergence and perseverance of newly-coined subtractive word-formation processes (cf. Gries 2004, 2006b).

However, while these types of similarity are at work over larger time periods—from priming effects within one text/discourse to the larger time periods time involving language acquisition and change—the present data add to our inventory of similarity effects an interesting much more *local*, within-VP phenomenon. But how can this be explained—what is the mechanism giving rise to this? I propose that this finding can be accounted for elegantly on the basis of the second notion, namely Langacker's (from my point of view under-appreciated) approach towards constituency. Cognitive Grammar does not view constituency as it is seen in, say, generative approaches to grammar. Rather, it distinguishes semantic/conceptual constituents and phonological constituents. Semantic/conceptual constituents are considered to be based on links that combine corresponding elements such that one element fulfills a valence requirement and elaborates an element of another element. One of the other semantic/conceptual groupings mentioned by Langacker (1997) is actually the kind of V-$\text{NP}_{\text{DirObj}}$ idiom investigated here; it is worth quoting a passage here at length:

> Another kind of conceptual group is the semantic pole of a complex lexical item such as *make headway* or *the cat is out of the bag*. It is well known that idioms are often phonologically discontinuous [. . .], hence not symbolized by a classical phonological constituent. This is unproblematic in cognitive grammar, which regards such symbolization as a minor and dispensable part of an idiom's characterization. An idiom resides primarily in a complex of semantic correspondences and symbolic links involving individual elements [. . .] (Langacker 1997: 15)[11]

With regard to phonological constituents, Langacker argues for the existence of several kinds of phonological groupings, of which temporal contiguity

[11] It is intriguing to notice that the two expressions Langacker refers to here also exhibit phonological similarity even though not in the form of alliterations: in *make headway*, both words feature a continuant followed by the same vowel ([**meɪ**k hed**weɪ**]), and in *the cat is out of the bag*, the two content words are both monosyllabic $\text{plosive}_{\text{bilabial}}$-æ-plosive ([ðə **kæt** ɪz aʊtə ðə **bæg**]).

is of course the most basic. Others include rhythmic cohesiveness, "stress, pitch level, and even *similarity in segmental content* (e.g., Spanish *la gata blanca* 'the white female cat')" (Langacker 1997: 22; my emphasis, STG). One example he discusses is that in.

(4) Any linguist is capable of making theoretical proposals, but only an *MIT* linguist is capable of making *interesting* theoretical proposals.

The semantic/conceptual constituent in question here is the focus indicated by the italicized words. Thus, the constituent is not held together by valence links, he argues, but by the abstract similarity of "degree of interest or informativeness." However, the focus is symbolized phonologically "by a phonological grouping based on unreduced stress."[12] This approach is in turn very much related to the facilitatory effects of similarity in exemplar-based approaches towards grammaticalization, onstructionalization, or conventionalization of prefabs as discussed by Bybee. For example, this is how Bybee (2010: 62–63) characterizes the workings of exemplar-based models: "entries sharing phonetic and semantic features are *highly connected* depending upon the degree of similarity" (my emphasis, STG).

Given all of the above, my hypothesis is that the recognition of the semantic constituent of the idioms/*way*-constructions is facilitated by the recognition of the phonological similarity based on the alliterations that the constructions studied here exhibit; recall the emphasized part of the Langacker quote regarding segmental similarity above. And in fact it is well-known by now that phonological information facilitates learning of higher-level sequences. Saffran et al. (1996) is perhaps the most-cited study to have shown that infants and young children can identify probabilistic tendencies that facilitate word segmentation in a stream of syllables. Even more pertinently, Onnis et al. (2005), for example, have shown that artificial language learners can identify words with non-

[12] Reviewers of this paper were convinced that the discussion of the within-pole phonological relation of alliteration should involve the notion of iconicity. While I disagree with that assessment, it is possible to see a relation between Langacker's phonological constituency on the one hand and diagrammatic iconicity on the other hand, where I follow Van Langendonck's (2007: 398) definition:

> a diagram is a systematic arrangement of signs that do not necessarily resemble their referents but whose mutual relations reflect the relations between their referents. More specifically, the constellation of the object and of its diagram is similar, but the individual referents and the individual signs themselves need not resemble each other.

From that perspective, one could argue that the observed alliteration effects constitute a case of diagrammatic iconicity: the *s*-*s*, *b*-*b*, etc. alliterations are not similar to their referents, but the relation of similarity that they exhibit reflects that together they make up a unit. However, Langacker does not appear to consider this connection necessary himself, and I concur.

adjacent syllable dependencies when the two non-adjacent syllables exhibit phonological similarity, where phonological similarity was operationalized on the basis of the first segment of the syllable, which fits nicely with the present case where content words in constructions exhibit phonological similarity, also in the form of alliterations.

Thus, the recognition of the phonological similarity of the elements studied here results in a higher degree of connectivity of the idioms' discontinuous constituents, which in turn facilitates and feeds back into (i) their perception as component parts of a greater whole and (ii) their undergoing the processes of chunking and subsequent constructionalization (cf. Bybee 2010: Section 3.2), but also their greater perseverance and internal collocational/collostructional coherence.

But if there is something to the above hypothesis, the question remains, why alliterations (rather than rhymes or, as in Langacker's example, other characteristics of words)? Currently, it is not clear which characteristic will be most likely to exhibit such similarity effects, but this is an empirical question and, thus, a problem shared by many researchers; for example Bybee (2010: 62) faces a similar explanatory problem for her treatment of *strung* verbs. The present study focused on word-initial alliteration effects, and Onnis et al.'s (2005) artificial language learning data exhibited a similar alliterative effect. And while Bybee (2010: 62) states that "the *final* consonants of the *strung* verbs are more important than the *initial* ones," (my emphasis, STG) six pages later she also does point out that some verbs that have been added to the class of *strung* verbs "also begin with a sibilant or sibilant cluster, increasing the phonetic similarity of the words as wholes." For yet another phenomenon that is connected to the present one theoretically (in terms of having recently been studied from a cognitive usage-based perspective) as well as empirically (in terms of exhibiting statistically significantly overrepresented word-initial phonological similarities), consider phonaesthemes (cf. Bergen 2004). Phonaesthemes, i.e., "frequently recurring sound-meaning pairings that are not clearly contrastive morphemes" (Bergen 2004: 290), are often observed from many examples of words sharing a particular sub-morphemic onset (such as *sl-*, *sn-*, or *gl-*, to name but the most widely-cited examples). Finally, previous studies have shown that *x* segments of the beginning of a word increase its chance of being recognized more than the same number of segments of its end (cf. Noteboom 1981).

In sum, the present study's observation that word beginnings are important and that the perception of phonological similarity may aid the identification of semantic/conceptual constituents/poles and their constructionalization is not as isolated or arbitrary an observation as it may seem. If (some part of) the function of these alliteration effects was to support the recognition of semantic/conceptual constituents by providing support for the recognition of a phonological constituent, then word beginnings would be a good place for this kind of support. This does of course not mean that word beginnings are always or mostly the most important determinant of language processing and change. For example, I fully accept Bybee's analysis that *strung* verbs are more revealingly analyzed on the basis of their rhymes. All of the above merely goes to show that different parts of words can be (more) relevant to different phenomena, and I have already pointed to cases where, say, the phonological similarity of words may be more distributed across the word (e.g., recall note 11 and the above allusion at blends). Thus, ultimately a broader and multiply granular view of similarity may be required for further study, and I will return to this now.

In this connection, there are several possible ways to follow up on the results of this study. The most obvious of these is to enlarge the database to see whether the same results will be obtained. This can mean increasing the numbers of V-NP$_{\mathrm{DirObj}}$ and related idioms and *way*-constructions. For example, a few examples I randomly overheard or noticed in writing are suggestive: *going great guns*, *give the devil his due*, *cut corners*, *pull the plug*, *do the trick*, and *gimme a break*. Similarly, Bybee's (2010: 60) mention of three prefabs includes *black and blue* and *bread and butter*. Another possible extension would of course be to investigate more and more different multi-word symbolic units, i.e., other constructions. On the one hand, just studying more constructions and/or idioms could be interesting. On the other hand, it could be worthwhile to find out how schematic, or lexically-filled, the constructions in focus have to be. This study looked at fully-filled and partially-filled structures but what about even more rigid constructions (e.g., proverbs) or, on the other side of the continuum, what about constructions with two schematic slots? Preliminary analysis of the *into*-causative (e.g., *he tricked her into buying that thing*) suggest an absence of alliteration effects; it seems that such similarity effects are not just a function of conventionalization but also of degree of schematicity. This would be a computationally and data-intensive task, but it could also open up interesting perspectives for our understanding of, say, the syntax-phonology interface, especially since it is already well-known that, for example, rhythmic alternation patterns influence both morphological and syntactic variation phenomena (cf. Schlüter 2003 and Gries 2007 respectively for examples).

More interestingly, one could extend and/or refine the notion of phonological similarity that is used. This can on the one hand mean looking for similarities in places other than word beginnings (although we have seen above why word beginnings may be a particularly salient point to begin with): in studies on blends and complex clippings, Gries (2004, 2006b) discusses a variety of other ways in which phonological similarity can be observed on the level of words and word combinations, and some or even many of these, such as string-edit distances, may be applicable here, too. For example, even if one restricts one's attention to word beginnings, the present analysis can be refined by widening the scope to (i) encompass not just initial segments, but complete onsets or even complete words and/or (ii) along the lines of note 11, include phonological similarity below the segmental level: *do the trick* and *gimme a break* do not involve segmental identity, but in *do the trick*, both the verb and the noun feature alveolar plosives, and in *gimme a break*, both the verb and the noun involve voiced plosives. In that regard: if features do play a role, which kinds of features are relevant: manner? place? voicing? These extensions and others can provide interesting insights regarding (i) the role that phonological constituency plays in multi-word units, (ii) the overall characterization of the forms in which phonological constituency can be manifested, as well as (iii) providing additional evidence for the multitude of similarity-, and thus categorization-based, processes in language that are at the heart of current theories involving exemplar-/usage-based models.

Received 10 August 2008
Revision received 24 September 2010

University of California,
Santa Barbara

References

Baayen, R Harald, Richard Piepenbrock, & Leon Gulikers (ed.). 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bergen, Benjamin J. 2004. On the psychological reality of phonaesthemes. *Language* 80 (2). 290–311.

British National Corpus Consortium. 1994. *The British National Corpus 1.0*.

Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Collins *Cobuild Dictionary of Idioms*. 2002. 2nd edition. London: Harper Collins.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220.

Gentner, Dedre & Arthur B. Markman. 1997. Structure mapping: a theoretical framework for analogy. *Cognitive Science* 7 (2). 155–170.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago, IL: The University of Chicago Press.

Gries, Stefan Th. 2004. Isn't that fantabulous? How similarity motivates intentional morphological blends in English. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture, and mind*, 415–428. Stanford, CA: CSLI.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34 (4). 365–399.

Gries, Stefan Th. 2006a. Corpus-based methods and cognitive semantics: The many meanings of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57–99. Berlin, New York: Mouton de Gruyter.

Gries, Stefan Th. 2006b. Cognitive determinants of subtractive word-formation processes: A corpus-based perspective. *Cognitive Linguistics* 17 (4). 535–558.

Gries, Stefan Th. 2007. New perspectives on old alternations. In Jonathan E. Cihlar, Amy L. Franklin, & David W. Kaiser (eds.), *Papers from the 39th regional meeting of the Chicago Linguistics Society: Vol. II. The Panels*, 274–292. Chicago, IL: Chicago Linguistics Society.

Gries, Stefan Th. 2009. *Statistics for linguistics with R: a practical introduction*. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th., Beate Hampe, & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16 (4). 635–676.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar Vol. I: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 1997. Constituency, dependency, and conceptual grouping. *Cognitive Linguistics* 8 (1). 1–32.

Noteboom, Sieb G. 1981. Lexical retrieval from fragments of spoken words: Beginnings vs. endings. *Journal of Phonetics* 9 (4). 407–224.

Onnis, Luca, Padraic Monaghan, Korin Richmond, & Nick Chater. 2005. Phonology impacts segmentation in online processing. *Journal of Memory and Language* 53 (2). 225–237.

R Development Core Team. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical. Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Saffran, Jenny R., Richard N. Aslin, & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274 (5294). 1928–1928.

Schlüter, Julia. 2003. Phonological determinants of grammatical variation in English: Chomsky's worst possible case. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 69–118. Berlin, New York: Mouton de Gruyter.

Sheskin, David J. 2007. *Handbook of parametric and nonparametric statistical procedures*. 4th edition. Boca Raton, FL: Chapman & Hall/CRC.

Snider, Neal. 2009. Similarity and structural priming. In Niels A. Taatgen & Hedderik van Rijn (eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, 815–820. Austin, TX: Cognitive Science Society.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8 (2). 209–243.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1 (1). 1–43.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: a corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1 (1). 113–150.

Van Langendonck, Willy. 2007. Iconicity. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, 394–418. Oxford & New York: Oxford University Press.