A lesson from associative learning

Guillaume Desagulier

A lesson from associative learning: asymmetry and productivity in multiple-slot constructions.

Guillaume Desagulier[1]

[1]MoDyCo — Université Paris 8, CNRS, Université Paris Ouest Nanterre La Défense

July 12, 2018

Abstract

Non-redundant taxonomic models of construction grammar posit that only fully productive patterns qualify as constructions because they license an infinity of expressions. Redundant models claim that, despite subregularities and exceptions, partially productive patterns also count as constructions, providing the overall meanings of such patterns are not the strict sums of their parts. Because productivity is a major bone of contention between redundant and non-redundant construction grammar taxonomies, I examine the productivity of *A as NP* which, according to Kay (2013), is not a "construction" but merely a "pattern of coining" due to its limited type productivity. Expanding on Gries (2013), this paper explores how a combination of symmetric and asymmetric association measures can contribute to the study of the "Productivity Complex" described in Zeldes (2012). Although the productivity of *A as NP* is admittedly limited at its most schematic level, some partially-filled subschemas such as *white/black as NP* or *A as hell/death* are arguably productive.

Keywords: asymmetry, association measures, ΔP (delta P), collostructional analysis, construction grammar, adjectives, intensification, productivity

1 Introduction

In Paul Kay's view of Berkeley Construction Grammar, only general and productive schemas such as *let alone* (Fillmore, Kay, & O'Connor 1988), *what's X doing Y* (Kay & Fillmore 1999), or *all*-clefts (Kay 2013) qualify as constructions. Patterns such as *A as NP* do not contain what is necessary and sufficient to interpret and generate other linguistic expressions. They are therefore at the periphery of grammar (Kay 2013):

(1) stiff as a board

(2) cool as a cucumber

(3) flat as a pancake

Admittedly, knowing the meaning of *cool* and the meaning of *cucumber* fails to predict the meaning of their combination in (2). Furthermore, this schema cannot be extended freely to similar A-NP pairs (e.g. $^{??}$*nervous as a zucchini*). Kay concludes that *A as NP* is not a "construction" but a "pattern of coining".

By comparison, Cognitive Construction Grammar (Goldberg 1995, 2006, 2009) is redundant: any pattern that is sufficiently entrenched and whose overall meaning is not the sum of the meaning of its parts counts as a construction. Based on these premises, even *A as NP*, counts as a construction.

Since the constructional status of *A as NP* differs depending on the theoretical perspective that one adopts, my aim is to test the boundary between what counts as a construction and what does

not. To do this, I assess the productivity of *A as NP* in the 100-million-word British National Corpus (XML Edition).

In both Construction Grammar approaches, the productivity of a complex expression is generally indexed on type frequency (Croft & Cruse 2004, 309). As regards *A as NP*, this is problematic because the productivity of the pattern is more complex than it seems. Even though semantic constraints block some A-NP pairings (e.g. $^{??}$*safe as buildings*) and make *A as NP* less productive than alledgedly full-fledged "constructions" such as *let alone*, *what's X doing Y*, or *all*-clefts, the pattern generates subschemas whose productivity is non-negligible, e.g. *white as NP* (*white as snow/a sheet/a whiteboard/etc.*), *pale as NP* (*pale as a ghost/death/parchment/etc.*), *A as hell* (*mad/hot/angry/etc. as hell*), or *(as) A as a cat* (*(as) jumpy/smug/tense/etc. as a cat*).

The productivity of *A as NP* is therefore more complex than Kay suggests, and the difference between a "construction" and a "pattern of coining" needs deeper empirical foundations. Because *A as NP* has two lexically unspecified slots, a complete assessment of its productivity should not be restrained to the schematic level. It should also involve a complementary inspection of productivity at subschematic levels, where the construction is partially or fully specified. These levels concern: (a) the first slot, (b) the second slot, and (c) their combinations.

Hapax-based productivity measures have been implemented in morphology (Baayen 1989; Baayen & Lieber 1991; Baayen 1992, 1993) and applied with success in the syntactic domain, especially in the context of multiple-slot constructions (Zeldes 2012). Such measures imply that productive linguistic units produce a large number of rare tokens. I will use these measures to show to what extent *A as NP* is productive and deserves a constructional status according to BCG (a position that I adopt for methodological purposes, without theoretical endorsement).

Zeldes (2012) also demonstrates that lexical choice is often mutually constrained in multiple-slot constructions. More precisely, observing one slot influences the expected realization of the other slot. This is in keeping with a key principle from associative learning: speakers tend to infer a linguistic outcome from cues available in their linguistic environment in an asymmetric fashion (Ellis 2006; Gries 2013). To develop this point further, I will apply a directional association measure, ΔP (Allan 1980), to the study of *A as NP*. Given a specific instance of *A as NP*, if ΔP reveals that the adjective is a better cue of the NP than vice versa, it is because the NP co-occurs with more adjectives. Conversely, if ΔP reveals that the NP is a better cue of the adjective than vice versa, it is because the adjective co-occurs with more NPs. ΔP alone cannot be used as a measure of productivity because it is based on token frequencies, not type frequencies. For a subschema indexed on an adjective to be considered productive, it must attract a significant number of distinct NP types. Likewise, for a subschema indexed on a NP to be considered productive, it must attract a significant number of distinct A types. However, once combined with the abovementioned productivity measures, ΔP can contribute to a better understanding of the productivity of multiple-slot constructions. Indeed, if it is confirmed that the choice of a slot influences the realization of the other slot, ΔP can specify the direction of this association.

In parallel, I will also examine attraction patterns between A and NP using a symmetric measure, namely log-likelihood ratio (Dunning 1993). My assumption will be that the most strongly associated A-NP pairs are also the most conventional, the most autonomous, and the least productive. Again, this measure will be evaluated against hapax-based productivity measures and ΔP.

This paper is organized as follows. Section 2 reviews competing accounts of what counts as a construction and suggests ways in which opposing views can be reconciled thanks to an empirical assessment of productivity in construction grammar. Section 3 compares the assets of symmetric and asymmetric association measures and discusses to what extent these measures have a bearing in the productivity assessment of a two-slot construction. Section 4 combines hapax-based measures and association measures to assess the productivity of *A as NP* and maps the results with principal component analysis. Section 5 discusses the results and argues that by augmenting productivity measures with a combination of symmetric and asymmetric association measures, one can estimate the productivity of a pattern in more details than if one were to rely only on traditional productivity

measures.

2 Theoretical issues

2.1 What counts as a construction: plain as day or clear as mud?

Construction grammar is a generic term for a family of approaches that focus on the grammatical status of syntax-semantics pairings: Berkeley Construction Grammar (Fillmore 1997; Fillmore et al. 1988; Kay & Fillmore 1999), Cognitive Construction Grammar (Goldberg 1995, 2003, 2006, 2009), Cognitive Grammar (Langacker 1986, 1987, 2008, 2009), Radical Construction Grammar (Croft 2001), Fluid Construction Grammar (Steels 2011, 2012), and Embodied Construction Grammar (Bergen & Chang 2005). All construction grammar approaches offer a non-derivational representation of grammatical knowledge and posit that grammar consists of a large inventory of constructions, varying in size and complexity, and ranging from morphemes to fully abstract phrasal patterns. Yet, not all theories agree as to how grammatical information is stored in the construction taxonomy, among other criteria. According to Croft & Cruse (2004, 257–290), there exist four taxonomic models. In the full-entry model, information is stored redundantly at various levels of the taxonomy. In the usage-based model, grammatical knowledge is acquired inductively, speakers generalizing over recurring experiences of use. In the normal-inheritance model, constructions with related forms and meanings are part of the same network. In the complete inheritance model, grammatical knowledge is stored only once at the most superordinate level of the taxonomy. The first three models are not mutually exclusive, each targeting a specific problem (e.g. the full-entry model resolves multiple inheritance conflicts). The last model is the odd one out because it leaves no room for redundancy.

Berkeley Construction Grammar (henceforth BCG) is a complete-inheritance model. It is non-redundant and it maintains a sharp distinction between "grammar" and the "meta-grammar".[1] "Grammar" is understood as "the strictly linguistic information required to produce and understand all possible utterances of a language, and no more" (Kay 2013, 30), whereas the "meta-grammar" consists of "patterns [that] are neither necessary nor sufficient to produce or interpret any set of expressions of the language (...)" (Kay 2013, 33). Given the above, only patterns that are general and fully productive qualify as constructions. Such is the case of *all*-cleft constructions:

(4) All I want is to stand and look at you, dear boy! (BNC-FPU)

Along with other cleft structures such as *wh*-clefts, *all*-clefts inherit their syntax from an abstract construction. What makes an *all*-cleft unique, and justifies its place as a construction in the grammar, is the scalar function, which Kay names the "'below-expectation' reading". This reading is distinct from the default universal quantification reading illustrated in (5).

(5) All that I command is yours now. (BNC-FR0)

All-clefts such as (4) are fully productive because they are only constrained at the lexical level "with respect to the filler constituent of the subject phrase" (Kay 2013, 37), and because they offer a way to rule in the good cases and rule out the bad cases without listing them.

In contrast, patterns that are partially or imperfectly productive are relegated to the meta-grammar. Such is the case of *A as NP*, as exemplified in (1)–(3) above, and (6) below:

(6) a. dark as night
b. black as pitch
c. thin as a rail
d. cold as hell

Despite the pairing of an identifiable syntax (*A as NP*) and a specific reading ("very A"), Kay considers *A as NP* "non-constructional" and "non-productive" because (a) knowing the pairing is not enough to license and interpret existing tokens (especially when there is no obvious semantic link between the

adjective and the NP, as in *easy as pie*), and (b) speakers cannot use the pattern freely to coin new expressions. Kay concedes that the pattern has "many members" but each addition appears via "analogical creation" (Kay 2013, 38). Two idiosyncrasies further block *A as NP* from qualifying as a construction (Kay 2013, 39). First, some expressions are motivated by a literal association between the adjective and the NP (*tall as a tree*, *white as snow*),[2] whereas others hinge on figurative associations between A and NP, including possible puns (*safe as houses*)[3], and yet others are the sign that the NP has grammaticalized to intensifying functions (*jealous as hell*, *sure as death*). Second, some expressions are compatible with a *than*-comparative (*flat as a pancake* > *flatter than a pancake*) whereas others are not (*happy as a lark* > ??*happier than a lark*). Such idiosyncrasies provide BCG with further evidence that expressions of the *A as NP* kind are not generated by a rule (i.e. a "construction") but must be learned individually.

A pattern of coining may grow into a construction diachronically, thus augmenting the existing grammar via the creation of a new linguistic type. But in BCG, diachronic productivity is of little help because it fails to explain how new utterance tokens are created in accordance with the existing grammar (Kay 2013, 44). In other words, BCG's sole focus is on synchronic productivity.

Because it is a "severe view of grammar" (Kay 2013, 33), BCG treats as non-constructional patterns that would count as full-fledged constructions in normal-entry models of construction grammar such as Cognitive Construction Grammar (henceforth CCG). CCG considers that grammatical knowledge is acquired inductively. Therefore, the inventory of constructions is internally redundant and allows for subregularities and exceptions (Goldberg 1995, 73–74). Consequently, a CCG account of grammar contradicts each of the tenets of BCG outlined in the previous paragraphs. First, according to (a) above, all expressions based on a construction are interpretable regardless of previous exposure. Because *A as NP* is not a construction in BCG, the meaning of *easy as pie* and *easy as duck soup* is supposedly inaccessible to "a young, foreign, or sheltered speaker of English" even if the speaker knows the meanings of *easy*, *pie*, *duck* and *soup* (Kay 2013, 38). However, putting aside the (hopefully) rare case of a sheltered language learner, we could reasonably posit that when there is no semantic motivation from the NP, such motivation can arise from the adjective. Indeed, we may reasonably expect that through exposure to tokens of *A as NP*, a learner will infer that expressions such as *easy as spit*, *easy as wink*, or *easy as apple pie* mean "very easy" in contexts where no spit, wink, or apple pie is readily available.

Second, according to (b) above, only a construction allows the speaker to coin new expressions *freely*. Accordingly, Kay distinguishes cases of full creativity (a construction can generate an infinite number of expressions) from "analogical creations" (typical of patterns of coining and more limited in terms of the number of generated expressions). Admittedly, *A as NP* cannot be applied freely to build new expressions on the model of some highly conventional expressions. Even though its NP has nothing to do with vegetables, the idiomatic expression in (7c) is a far more likely antonym of (7a) than (7b):

(7) a. cool as a cucumber
    b. ??nervous as a tomato/potato/zucchini/etc.
    c. as nervous as a long-tailed cat in a room full of rocking chairs

Yet, there is evidence that, contrary to what Kay (2013, 38) claims, not all novel *A as NP* expressions "die aborning", and not all of them are the product of "self-conscious, literary usage". Some patterns, such as *easy as pie*, are sufficiently established to serve as templates for the creation of syntactically- and semantically-related expressions:

(8) a. What I'm sayin', Benedict can take you out **as easy as cake**. (*Last Action Hero*)
    b. Buying Wireless: **Easy As Candy**? (connection.ebscohost.com/)
    c. Her father groans and leans for his glass. "**Simple as pie**."
    (*The Significance of Birds*, Lynn Hanson)

(9) a. **Easy as a wink**, so I'm told. (www.wattpad.com)
    b. (...) you'll find managing your parameters **as easy as a blink**. (www.soundonsound.com)

From a BCG perspective, full productivity is blocked when there is no straight semantic correspondence between the adjective and the NP. Nevertheless, *A as NP* is flexible enough to accommodate antonymic pairings:

(10) I'm **as heavy as a feather**. Hallelujah. (from the song *Hello, Dear Wind* by Page France)

(11) Ladies' golf shafts are about **as stiff as rubber bands**. (forums.quattroworld.com)

More generally, even when some expressions are lexically filled, speakers are not necessarily blocked from creating new expressions when the adjective and the NP are semantically compatible. For example, since a hot-air balloon is big and can fly high, it is a good candidate for the intensification of *big* and *high*, including when these adjectives are used figuratively:

(12) a. (. . .) he had a head **as big as a hot air balloon**.
(*The Adventures of Jack Lime*, James Leck)

b. My denial bubble was about **as big as a hot air balloon**.
(*Close to Mummy's Heart*, Mary Cromwell-Hillenburg)

(13) a. Taking agritourism **as high as a hot-air balloon**. (smallfarms.cornell.edu)

b. I was just **high as a hot-air balloon**. I didn't hear a word he said.
(*Paradise Park*, Allegra Goodman)

Whether some uses are considered as self-conscious or literary depends largely on the corpus that one adopts, and even nonce expressions are the sign that not yet attested expressions should not be ruled out as ungrammatical in the absence of empirical evidence.[4] Examples (10) and (11) above, and example (14) below make it clear that *A as NP* patterns are available for the creation of new tokens, or for the modification of existing tokens, even when the A-NP pairing appears to be fixed:

(14) "**Quicker Than a Wink**:" Photographs by Harold Edgerton. (www.amherst.edu)

Excluding these examples from a BCG inventory of constructions is consistent with the tenets of a non-redundant view of grammar, but somewhat at odds with usage. I will argue below that it takes more than a cursory corpus search to assess the productivity of a pattern and decide whether it has the status of a construction, regardless of the theoretical premisses. At this point, productivity cannot be discussed further without a clear understanding of how the concept is used and how it can be operationalized in construction grammar approaches.

2.2 Productivity in construction grammar

According to Bybee (1985, 2001, 2010), the account of morphological productivity outlined above also applies to morphosyntactic constructions in an exemplar framework.[5] As an exemplar of a construction reaches a certain token-frequency threshold, it is likely to be stored as a unit in the minds of speakers. As high-frequency tokens and formulaic expressions lose in terms of analyzability, compositionality, and productivity, they gain in terms of autonomy. In other words, autonomous units do not reinforce a schema if the schema can be generalized.

2.2.1 Theoretical approaches

Turning now to what Kay considers full-fledged constructions, *What's X doing Y* (*WXDY*) and *all*-clefts have a high type frequency because they generate a potentially infinite number of instances:

(15) a. What is this fly doing in my soup?

b. What is my best friend doing under the bed?

c. What is it doing raining?

d. . . .

(16) a. All I want is get out of here.

b. All she does is sing.

c. All he reads is pulp fiction.

d. ...

Although productive with respect to type, *WXDY* and *all*-clefts generate instances that are not themselves so productive given the lexical openness and contextual specificity of their respective schemas. In other words, tokens modelled on *WXDY* or *all*-clefts rarely occur more than once, and this is precisely what keeps the mental representation of the constructions active in the minds of speakers. Such schemas are therefore productive at a the highest level of schematicity. In contrast, even though *A as NP* is somewhat restrained as to the number of A-NP pairs that it sanctions, among the many instances that are built on this schema, some have a relatively high token frequency (in the BNC, such is the case of *good as gold*, *cold as ice*, *large as life*, *bold as brass*, *safe as houses*, etc.). Some pairs are fixed (*cool as a cucumber*), but many others are not. Among the latter, some instances are modelled on productive subschemas, e.g. *A as hell/death*, *white/black/old as NP*, etc. As will appear below, deciding on the productivity of *A as NP* requires inspecting more than the highest level of schematicity.

Three issues may be raised with respect to productivity in Kay's view of BCG. Firstly, as pointed out in Barðdal (2008), BCG does not provide a unified definition of productivity, especially with regards to whether it is graded. Kay and Fillmore (1999, 31) seem to recognize several levels of productivity, ranging from partially productive (idiomatic, lexically filled) to fully productive (schematic) patterns:

> (...) a construction-based approach appears to provide promise of accounting both for the relatively idiomatic and for the abstract and more fully productive aspects of a language.

However, in more recent works, Kay and Fillmore do not express similar views anymore. On the one hand, Kay (2013) understands productivity in an all-or-none fashion: patterns that are completely and fully schematic are productive whereas other patterns are not. On the other hand, Fillmore (2002) recognizes that "productivity is a notion of degree". Fillmore does distinguish productivity (i.e. applying a general rule without listing exceptions) from coining (exploiting unproductive patterns in the creation of new expressions). However, he does not relegate coining to the periphery of the "grammar proper" on account of partial or lacking productivity:

> There is a view of grammar according to which the grammar proper will identify only the productive processes. Since the ability to create new words, using non-productive processes, is clearly a linguistic ability, it is my opinion that a grammar of a language needs to identify constructions that exist for "coining" purposes as well.

Fillmore admits that constructions are assumed to exist at all levels of schematicity (i.e. productive and less productive), a tenet shared by CCG. Elaborating on the terminology proposed by Barðdal (2008, 37), we can summarize the difference between Kay and Fillmore as follows: the former posits a regularity vs. extensibility *distinction* (productivity being indexed on regularity alone) whereas the latter posits a regularity vs. extensibility *continuum* (productivity being indexed on both). There being no general consensus on positing a radical view such as Kay's, one can choose either to disregard it on the grounds that it is a matter of theoretical preference (Barðdal 2008, 36–37) or to test it empirically. In the sections that follow, I choose the second option.

Secondly, the correlation between schematicity and productivity assumed by Kay is problematic. Bybee (2010, 94–95) observes that, in the construction *V someone A* (e.g. *she drove me crazy*)[6], the adjective slot is very productive because it is quite open. In comparison, the verb slot is more schematic yet less productive, because open to a limited number of types, such as *drive*, *send*, or *make*. Just because a pattern is schematic does not mean that it is productive. Furthermore, as Croft and Clausner note, there are various levels of schematicity: "The instanciation of one schema may function as an intermediate schema for other more specific instantiations" (Croft & Clausner 1997, 251). Just because a pattern is not productive at the highest schematic level does not mean that it is unproductive at subschematic levels.

Finally, if we accept the view that morphosyntactic productivity corresponds to the likelihood that a construction will apply to a new item, we may still question that the productivity of a construction should be entirely determined by its type frequency. Other factors have been recognized to play a role, such as semantic similarity between tokens (Croft & Clausner 1997) or semantic coherence (Barðdal 2008). Moreover, cognitive-functional approaches to frequency effects on productivity posit a clear-cut distinction between high and low frequencies (whether type or token) (Langacker 1987; Bybee 1985, 2001; Croft & Clausner 1997). Once we try to operationalize this theoretical threshold, we realize that the boundary between what is frequent and infrequent in a corpus is relative. Bybee (2010) illustrates her claims on frequency effects with several case studies (e.g. the English auxiliary sequence, the difference between *no* and *not* negation, etc.), but most of them are broad enough to return contrasted frequency scores. We should not be blinded to the fact that many patterns occur at intermediate frequency levels, where the difference between what is highly frequent and what is not is intuitively fuzzy.

2.2.2 Quantitative approaches

Productivity has received sustained attention in morphology. Broadly defined, morphological productivity is the property of an affix or a word-formation rule to generate new formations in a systematic way (Plag 2003). The more an affix is available for the creation of new words, the more productive it is. Less general processes generate new words in a local, *ad-hoc* fashion. For example, speakers may use existing affixes to derive new words analogically, e.g. *beigeness* and *beigity* (Bauer 2001). At a qualitative level, given the existence of rare but actual coinages, it is difficult to draw the line between productive and less productive processes.

In response, quantitative approaches have developed measures to capture the productivity of morphological processes in corpora. The first measure is $V(C, N)$, i.e. the type count V of the members of a morphological category C in a corpus of N tokens. It captures what Baayen (2009) calls "realized productivity". One problem with indexing a measure on type is that it does not discriminate between established forms and new forms. Indeed, types are not distributed uniformly in a corpus, and the larger the corpus, the harder it is to find innovations. To keep track of vocabulary development across a given corpus, one can plot the number of types against the number of tokens at multiple intervals. One obtains a vocabulary growth curve (henceforth VGC) (Baayen 1993). At first, the curve is expected to rise steeply as most tokens define a new type. Then, as more text is scanned, more and more tokens are included in already defined types, and the curve flattens out gradually. VGCs are a handy way of visualizing productivity: the steeper the curve, the more productive the linguistic unit/process under study.

A second set of measures, based on Baayen (1989) and further described in Baayen and Lieber (1991), and Baayen (1993) *inter alia*, draws on the idea that the number of hapax legomena of a given morphological category correlates with the number of neologisms in that category, which in turn correlates with the productivity of the rule at work. Baayen estimates "expanding productivity" (also known as $\mathscr{P}^*$, the "hapax-conditioned degree of morphological productivity") as the number of hapax legomena in the category divided by the number of hapax legomena in the corpus:

$$\mathscr{P}^* = \frac{V(1, C, N)}{V(1, N)} \tag{i}$$

$\mathscr{P}^*$ does not account for whether a morphological process, although productive, is saturated, unlike $\mathscr{P}$, which measures "potential productivity". $\mathscr{P}$ is the ratio of the number of hapax legomena with a given affix and the sum of all tokens that contain the affix:

$$\mathscr{P} = \frac{V(1, C, N)}{N(C)} \tag{ii}$$

$\mathscr{P}$ corresponds to the probability of encountering new types. The larger the number of hapax legomena, the larger $\mathscr{P}$ and the productivity of the affix. Conversely, the larger the sum of all tokens, the smaller

$\mathscr{P}$ and the productivity of the affix. One mathematical property of $\mathscr{P}$ is that it is the slope of the tangent to a VGC at its endpoint. It indicates the rate at which vocabulary increases at the end of a vocabulary growth curve. Like VGCs, $\mathscr{P}$ values must be compared at equal sample sizes. Baayen (1993) explores further the correlation between productivity and frequency and observes that $\mathscr{P}$ captures only the probability of possible types to the detriment of observed types. He advocates for a measure of global productivity (P^*) which evaluates $\mathscr{P}$ in terms of type frequency (i.e. actual use):

$$P^* = \frac{V}{\mathscr{P}} \tag{iii}$$

Underlying Baayen's hapax-based measures is the following principle: productivity is a factor of both a large number of low-frequency words and a low number of high-frequency words. This combination contributes to parsing effects which maintain the affix activated in the minds of speakers.[7] Zipf (1949) showed that even the largest collection of text in a given language does not contain instances of all types in that language. Therefore, type count and distribution in a corpus cannot reliably estimate type count and distribution in a language. To circumvent this issue and predict the vocabulary size in a larger corpus or a whole language based on a much smaller corpus, Baayen (2001) uses LNRE models (the acronym stands for "large number of rare events"). First, one converts the data into a frequency spectrum (i.e. a table of frequencies of frequencies). Next, one fits a model to the spectrum object, choosing from a range of LNRE models available in the `zipfR` package (Evert & Baroni 2006, 2007) for R (R Core Team 2014), namely the Zipf-Mandelbrot model (ZM), the finite Zipf-Mandelbrot model (fZM), or the Generalized Inverse Gauss-Poisson model (GIGP). Once a satisfactory LNRE model has been fitted to the frequency spectrum, one obtains expected values for the vocabulary size and the spectrum elements within the range of the corpus/sample size (interpolation) and beyond (extrapolation). One can then plot the results on an enhanced VGC. LNRE models make it possible to compute S, the limit of V when N approaches infinity (Baayen 2001, Section 2.4):

$$S = \lim_{N \to \infty} E[V(N)] \tag{iv}$$

S is recognized to be an estimate of the total possible number of types for a given process.[8]

Zeldes (2012) shows that Baayen's family of hapax-based measures and LNRE models apply equally well to morphosyntactic constructions, providing one is willing to adapt the definitions of type and type count to the structural configuration of the construction under study. In the context of a construction with an invariable part and one slot to be filled, productivity is measured at the level of that single open slot.[9] In the context of a two-slot construction,[10] Zeldes (2012, 127) presents two options to compute hapax-based measures, plot growth curves, and fit LNRE models. The first option consists in concatenating the two slots to produce a single fused type – e.g. *good_gold* – which is then tested for productivity. If a given exact concatenation has not been observed before in the corpus, it is treated as a hapax legomenon. If it is observed again, it leaves the hapax list. However this first option makes little sense in the context of *A as NP*, where we can reasonably expect the A slot and the NP slot to be interdependent both semantically (A is intensified with reference to NP) and structurally (both A and NP are required by the form of the construction). The second option consists in treating as new types innovations in at least one slot. In this case, "novel combinations of lexemes already observed in their respective slots are considered repetitions" (Zeldes 2012, 128). Both options will be compared in Section 4.1.

The next sections will address the question of how we can determine the multilayered productivity of *A as NP* at various levels of schematicity. Qualitatively speaking, I will assume that productivity is a gradient that involves autonomy and analyzability, in the wake of Bybee (1985, 2001, 2010), Croft and Clausner (1997), and Barðdal (2008). Quantitatively speaking, I will follow Zeldes (2012, 135), who substantiates the idea that syntactic productivity is multidimensional: different measures will sometimes yield different productivity rankings, even at equal sample sizes. To capture what Zeldes calls the "Productivity Complex", I will examine frequency data not only in the light of the traditional productivity measures outlined above, but also in the light of symmetric and asymmetric

association measures. I am aware that the question of whether association measures can even begin speak to the issue of productivity may raise serious doubts among linguists. Yet it is my intuition that valuable insights can be also drawn from such measures regarding the inner workings of productivity in multiple-slot constructions. I justify this intuition below.

3 Methods

3.1 Symmetric association measures

Because *A as NP* has two lexical slots, an obvious starting point would be to measure the collocativity between A and NP, choosing from a vast inventory of popular association measures many of which are reviewed in Church, Gale, Hanks, and Hindle (1991), Evert (2005, 2009), and Pecina (2010).[11] Most of these measures compare the observed occurrences of two forms in a contingency table with their expected frequencies to determine which collocations are statistically significant. Hypothesis tests such as t-score, z-score, Dunning's log-likelihood ratio (Dunning 1993) or Pearson's χ^2 select a statistical test so that its distribution converges to a well-known distribution when the null hypothesis is true. These tests are powerful and easy to compute, but they suppose that the data is similar to one of these known distributions. Since co-occurrence data does not always satisfy these distributional assumptions, exact tests such as the binomial, Poisson, and Fisher's exact tests may be more appropriate. Exact tests compute the p-value of all possible outcomes that are similar to or greater than the observed frequencies of a contingency table.

Neither hypothesis tests nor exact tests are without problems. The χ^2 test presupposes that the linguistic phenomenon under scrutiny is distributed randomly across a corpus, but it is a well-know fact that "language is never, ever, ever, random" (Kilgarriff 2005). Fisher's exact test (Yates 1984) is considered more appropriate than parametric, asymptotic tests in the identification of dependent word pairs because it is not affected by the skewed and sparse nature of data samples typical of natural language corpora (Pedersen 1996). In contrast to the above tests, pointwise mutual information (Church & Hanks 1990) does not depend on distributional assumptions and can be computed at minimal computational cost. A product of information theory, pointwise MI determines to what extent the occurrences of a word w_1 influence the occurrences of another word w_2. Because it is intuitively and methodologically simple, British lexicographers have made pointwise MI a standard association measure. Yet, pointwise MI is sensitive to data sparseness and invalid for low-frequency word pairs (Evert 2009; Kilgarriff 2001; Manning & Schütze 1999). Consequently, pointwise MI favors low-frequency collocates (Church et al. 1991, 133). All things considered, no association measure is fully satisfactory, and the choice of one over another depends on what aspects of collocativity one wishes to reveal (Evert 2005; Wiechmann 2008). For all the above, one should should always interpret ranked lists with caution.

In the Construction Grammar framework, Stefanowitsch and Gries have developed a family of methods known as collostructional analysis. Collexeme analysis measures the mutual attraction of lexemes and constructions (Stefanowitsch & Gries 2003), distinctive collexeme analysis contrasts alternating constructions in their respective collocational preferences (Gries & Stefanowitsch 2004b). To measure the association between the two lexical slots of *A as NP*, we can use a third method known as covarying-collexeme analysis (Gries & Stefanowitsch 2004a; Stefanowitsch & Gries 2005), using Dunning's log-likelihood ratio (G^2) as an association metric. To evaluate if such a method has any bearing in productivity assessment, let us anticipate briefly Section 4. We might be tempted to use the metric to determine those A-NP pairs which are highly conventionalized (due to verbatim recurrence), and therefore unproductive in Baayen's sense. Such would be the case for instance of the top elements ranked by collostructional strength in Table 1. That *good* and *gold* are the most strongly associated covarying collexemes in this construction should come as no surprise for distributional reasons. The method determines for each potential A occurring in the first slot which potential NPs in the second slot co-occur with it more often than expected. Table 2 shows that *good* and *gold* co-occur almost exclusively with each other in the sample (*gold* is used only once with another adjective). The differ-

ence between observed and expected frequencies is therefore large, which reflects on the collostruction strength.

Table 1: Top covarying collexemes in *A as NP* in the BNC

rank	A	NP	coll.strength (G^2)
1	*good*	*gold*	288.8138
2	*quick*	*flash*	189.2885
3	*right*	*rain*	175.9832
4	*large*	*life*	164.5468
5	*safe*	*houses*	148.3236
6	*sure*	*hell*	141.9763
...	...	...	...
25	*queer*	*folk*	56.9492

Table 2: Distribution table for a covarying collexeme analysis involving *good* and *gold*

	gold	other NPs
good	29	0
other As	1	3638

By comparison, Table 3 indicates that *queer* and *folk* co-occur exclusively with each other, with no exception. This does not reflect so much on the collostruction strength because of the relatively low type count and the smaller difference between observed and expected frequencies.

Table 3: Distribution table for a covarying collexeme analysis involving *queer* and *folk*

	folk	other NPs
queer	4	0
other As	0	3638

Finally, Table 4 shows that while *sure* and *hell* co-occur with each other significantly more than expected, which explains their relatively high ranking as a pair, each lexeme also co-occurs with other lexemes. This is especially true of *hell*, which appears with 44 adjective types in 98 tokens of *(as) A as hell*. *Sure* appears with only 11 NP types in 50 tokens of *sure as NP*.

Table 4: Distribution table for a covarying collexeme analysis involving *sure* and *hell*

	hell	other NPs
sure	33	17
other As	65	3638

Good as gold seems to have reached a very high level of conventionalization,[12] autonomy, and lack of analyzability. Therefore, we could tentatively use collostruction strength as an inverse measure of productivity: the higher the association strength, the higher the level of autonomy, the lesser the productivity. However, this should be done with caution as ranked lists based on most association measures have caveats (Schmid & Küchenhoff 2013). Being part of an informal saying (*there's nowt as/so queer as folk*), *as/so queer as folk* is hardly less conventionalized and autonomous than *good as gold*, despite what its lower collostruction strength suggests. As a lexically-filled schema, *sure as hell* is fairly conventional, autonomous, and hardly analyzable. Yet, we also learn a lot about the distributional behavior of the partially-filled schemas *A as hell* and *sure as NP* from inspecting the details of Table 4.

3.2 An asymmetric association measure: ΔP

In a recent paper, Gries (2013) identifies yet another issue with traditional association measures. Such measures commonly assume a bidirectional dependency between two word pairs. Therefore, given the pair $word_1$-$word_2$, $word_1$ attracts $word_2$ as much as $word_2$ attracts $word_1$. Yet, given *ipso*, the probability of obtaining *facto* is high, but given *facto*, the probability of obtaining *ipso* is not as high because of other words compete for the same slot (e.g. *de* or *post*). In other words, the collocation between *ipso* and *facto* is directional because one word is a better cue of the other than vice versa.

The idea that collocations are directional is inspired by studies on language acquisition (in particular Ellis 2006, Ellis & Ferreira-Junior 2009) which draw from a rejection of classical conditioning in the field of associative learning. In classical conditioning, learning occurs when a conditioned stimulus (CS, or "cue") is temporally paired with an unconditioned stimulus (US, or "outcome"). Typically, in Pavlov's experiments (Pavlov 1927), a CS (e.g. ringing a bell) was presented to a dog just before the US (delivering food). After several identical experiments, the dog emitted the conditional reflex (CR) of salivating upon being presented with the CS alone. Pavlov concluded that the US reinforced the CR of salivating to the CS. According to classical conditioning, it is the temporal pairing of the US and the CS that associates the two and therefore strengthens the CR. However, Rescorla (1968) showed that rats did not emit a CR when trials involving the US alone were added to trials where the temporal pairing between US and CS was preserved. He concluded that it is in fact contingency, not temporal pairing, that generates the conditioned response. According to Ellis (2006) and Baayen (2011), Rescorla's revision of Pavlovian conditioning is at the root of contemporary inferential methods in animal and human learning. In line with Bayesian reasoning, the Rescorla-Wagner equations state that learners predict an outcome from cues available in their environment if such cues have a value in terms of outcome prediction, information gain, and statistical association (Wagner & Rescorla 1972). Drawing a parallel with optimal word processors, whose algorithms take into account recency of prior usage, frequency of prior usage, and context of usage to perform automatic completion, the association between a cue and an outcome is directional. When an event comprises a cue and an outcome, there are four possible configurations, summarized in Table 5 (where a, b, c, and d are frequencies):

Table 5: Co-occurrence table involving a cue (C) and an outcome (O)

	O	¬O
C	a	b
¬C	c	d

To measure the asymmetric dependency between C and O, Ellis relies on ΔP, a one-way-dependency statistic developed by Allan (1980):

$$\begin{aligned} \Delta P &= p(O|C) - p(O|\neg C) \\ &= \frac{a}{a+b} - \frac{c}{c+d} \end{aligned} \tag{v}$$

The closer ΔP is to 1, the more C increases the likelihood of O. Conversely, the closer ΔP is to -1, the more C decreases the likelihood of O. If $\Delta P = 0$, there is no covariation between C and O.

Gries (2013) transposes the above to the study of bigrams. Given a bigram involving $word_1$ and $word_2$, four configurations are also possible (Table 6).

Table 6: Co-occurrence table involving a cue ($word_1$) and an outcome (here, $word_2$)

	$word_2$: present	$word_2$: absent
$word_1$: present	a	b
$word_1$: absent	c	d

This time, two ΔP values must be computed, depending on whether the cue is $word_1$ and the outcome

is word$_2$ or whether the cue is word$_2$ or and the outcome is word$_1$:

$$\begin{aligned}\Delta P_{(word_2|word_1)} &= p(word_2|word_1) - p(word_2|\neg word_1) \\ &= \frac{a}{a+b} - \frac{c}{c+d}\end{aligned} \tag{vi}$$

$$\begin{aligned}\Delta P_{(word_1|word_2)} &= p(word_1|word_2) - p(word_1|\neg word_2) \\ &= \frac{a}{a+c} - \frac{b}{b+d}\end{aligned} \tag{vii}$$

If $\Delta P_{(word_2|word_1)} - \Delta P_{(word_1|word_2)}$ is positive, then word$_1$ is a better predictor of word$_2$ than vice versa. Conversely, if $\Delta P_{(word_2|word_1)} - \Delta P_{(word_1|word_2)}$ is negative, then word$_2$ is a better predictor of word$_1$ than vice versa. If $\Delta P_{(word_2|word_1)} - \Delta P_{(word_1|word_2)}$ is null, then no word is a good predictor of the other. Table 2 and formulas (vi) and (vii) can be easily transposed to the study of *A as NP*. Arbitrarily, the adjective is the cue and the NP is the outcome in Table 7 :

Table 7: Co-occurrence table involving a cue (here, A) and an outcome (here, NP)

	NP: present	NP: absent
A: present	a	b
A: absent	c	d

$$\begin{aligned}\Delta P_{(NP|A)} &= p(NP|A) - p(NP|\neg A) \\ &= \frac{a}{a+b} - \frac{c}{c+d}\end{aligned} \tag{viii}$$

$$\begin{aligned}\Delta P_{(A|NP)} &= p(A|NP) - p(A|\neg NP) \\ &= \frac{a}{a+c} - \frac{b}{b+d}\end{aligned} \tag{ix}$$

Similarly, if $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ is positive, then the adjective is a better predictor of the NP than vice versa. Conversely, if $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ is negative, then the NP is a better predictor of the adjective than vice versa. If $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ is null, then neither the adjective nor the NP is a good predictor. For the sake of illustration, let us now apply Table 7 and formulas (viii) and (ix) to one concrete example of *A as NP*: *mad as a March hare*. We obtain Table 8 and formulas (x) and (xi):

Table 8: Co-occurrence table for *mad* and *March hare* in *mad as a March hare* in the BNC

	March hare: present	*March hare*: absent
mad: present	2	7
mad: absent	0	98363774

$$\begin{aligned}\Delta P_{(March\ hare|mad)} &= p(March\ hare|mad) - p(March\ hare|\neg mad) \\ &= \frac{2}{2+7} - \frac{0}{0+98363774} \\ &\approx 0.22\end{aligned} \tag{x}$$

$$\begin{aligned}\Delta P_{(mad|March\ hare)} &= p(mad|March\ hare) - p(mad|\neg March\ hare) \\ &= \frac{2}{2+0} - \frac{7}{7+98363774} \\ &\approx 1\end{aligned} \tag{xi}$$

$$\Delta P_{(March\ hare|mad)} - \Delta P_{(mad|March\ hare)} \approx -0.78 \tag{xii}$$

Unsurprisingly, *March hare* is a much better predictor of *mad* than vice versa. This is because *mad* co-occurs with other NPs such as *hatter* (another reference to *Alice in Wonderland*) or *hell*. On the other hand, *March hare* co-occurs exclusively with *mad* in the BNC. In other words, because *mad* is far more productive than *March hare* in *A as NP*, the former is much less of a good predictor than the other in the corpus.

The above suggests that we could use the asymmetric dependency between the adjective and the NP to locate the subschematic productivity of each type of *A as NP*. More precisely, instead of deciding whether *A as NP* is productive as a the most schematic level, we could use ΔP to spot productive subschemas. Thus, given an adjective A_i, the subschema built on A_i would be productive if it attracted several types of NPs. Given a noun phrase NP_j, the subschema built on NP_j would be productive if it attracted several types of adjectives. Yet, ΔP computes asymmetries from token frequencies and yields results that are valid at the token level. To avoid this pitfall, one must show how token frequencies may reflect tendencies observable at the type level. I address this issue in Section 4.

3.3 Data

The data comes from the British National Corpus (XML Edition), which consists of 98363783 words of spoken and written British English divided among over 4049 texts (Burnard 2000). Its size and its relatively balanced sampling scheme contribute to the reliability and validity of observations. 8 text types are represented in either written or spoken English: "academic writing", "non-academic writing", "published fiction", "news and journalism", "other published writing", "unpublished writing", "conversation" and "other spoken". Only instances of *A as NP* where the adjective is intensified were kept. Typically, such instances display non-literal comparisons. Examples involving a literal comparison and no intensification were discarded.

The idea that literal comparisons can be filtered out may seem intriguing to anyone used to handling natural language corpus data. For example, *he had a head as big as a hot air balloon* is clearly figurative in (12a). Yet, one can easily find a literal equivalent. In (17), the monster described by the child is literally as tall as a 50-floor building. Its head is therefore literally as big as a hot air ballon:

(17) My monster has a huge head, **as big as a hot air balloon**. The monster is huge, as big as a 50 story building. (*My Monster* by Sean, http://www.woodlandroad-p.schoolwebsites.com.au)

Sometimes, writers deliberately blur the lines between literal and figurative meanings, as in (18):

(18) Is mastering food idiom a piece of cake, **as easy as a pancake**, or a roll with butter? (*The Guardian*, 17 October 2012)

Once retrieved, the examples were therefore filtered manually with special attention paid to the context. Because the immediate context was not always decisive, it was sometimes necessary to resort to the extended context. By way of illustration, (19) is an excerpt from an ad for a company selling Afghan and Persian carpets. The text being meant for a British audience, it is very likely that the Northern Line refers to a London underground line. Because delays have made this line infamous, the author considers "a delay on the Northern Line" a paragon of commonness:

(19) Over the years, Liberty have maintained their reputation for Afghan and Persian carpets despite tough competition. On the Afghan border with Soviet Uzbekistan lies Mazar-e-Sharif. Not a city for the lily-livered or faint-hearted. Here a Kalashnikov in the back is **as common as a delay on the Northern Line**. (BNC-CFS)

Insofar as *as common as a delay on the Northern Line* can be paraphrased as "very common", the NP is assumed to intensify the adjective and this token of *A as NP* was kept. The remaining difficulty concerns the delimitation of the construction tokens. For the purpose of this study, NP determiners and modifiers were ignored except in cases where some degree of collocation between the NP and its modifier(s) was assumed to exist:

(20) The answers were **clear as ash in a smouldering grate**. (BNC-K5M)

(21) You were **as mad as a March hare**. (BNC-ACK)

In cases like these, the NP, its determiner, and its modifier(s) were amalgamated into one complex NP.

Taking the above difficulties into account, 1819 tokens of *A as NP* were extracted. Admittedly, these tokens attest a small vocabulary size for a study on productivity.[13] This could be a problem insofar as most construction samples are located in the "LNRE zone" (Baayen 2001, 55). Given the large number of rare constructions, asymptotic properties will not emerge easily (or even not at all) if the distribution is based on too small a sample size $N(C)$. A solution suggested by Zeldes (2012) is to obtain samples from very large corpora such as those compiled from the web by Baroni, Bernardini, Ferraresi, and Zanchetta (2009). Admittedly, the BNC is much smaller than, say, the 2.25-billion-token ukWaC corpus of English (Ferraresi 2007). However, to extract a large volume of error-free matches so as to minimize manual inspection, one must formulate a high-accuracy corpus query that will filter away unwanted data. As regards *A as NP*, this is easier said than done as both the BNC and ukWaC have a high proportion of patterns which search engines mistake for intensifying components (ex. *happy as a child*, *available as an alternation/an option/a remedy*, *likely as a result*, etc.) despite careful effort to run a highly accurate query. One could of course implement a heuristic such as a stoplist of unwanted patterns. One could also turn to some form of supervised learning, but then one would still have to tease apart those matching patterns that potentially have an intensifying function, like (22a), and those that clearly do not, like (22b):

(22) a. (...) Haig, who was an unemotional man with a character **as hard as granite** (...). (ukWaC)

b. But this basalt is **as hard as granite** (...). (ukWaC)

To the best of my knowledge, available techniques are far from reaching the accuracy of manual inspection. Therefore, I believe the BNC is a viable alternative for the task at hand, as long as the statistical models presented in Section 2.2.2 show a satisfactory goodness of fit.

4 Results

4.1 Type and vocabulary growth

The number of types depends on how we parse the construction. If we consider exact matches, which include variation as to the choice of the comparative structure (*as... as* vs. *as*), the NP determiner (indefinite, definite, and zero), number (*snug as a bug* vs. *snug as bugs*), an optional NP postmodifier (*sure as eggs* vs. *sure as eggs is/are/was eggs*), and the two lexical slots (A and NP), $V = 1316$. If we concatenate the two lexical slots (A-NP), the type count is 1206. There are 402 A types, and 876 NP types.

We might expect exact matches to be more productive than A-NP concatenations because the former have more loci of innovation than the latter. This is confirmed in the left part of Table 9.

Table 9: Left part: productivity measures for exact matches, concatenated slots, A, and NP; right part: corresponding LNRE models and goodness-of-fit (multivariate chi-squared test)

	$N(C)$	V	$V1$	$\mathscr{P}$	S	LNRE model	χ^2	df	p
exact matches	1819	1316	1116	0.613523914	1638648	Generalized Inverse Gauss-Poisson	0.9384721	4	0.9189776
A-NP	1819	1206	1011	0.55579989	415579.1	finite Zipf-Mandelbrot	8.447248	5	0.1332488
A slot	1819	402	217	0.119296317	65711.95	finite Zipf-Mandelbrot	7.038904	6	0.3172644
NP slot	1819	877	665	0.365585487	1.59E+14	finite Zipf-Mandelbrot	8.349324	6	0.2136115

Exact matches are more productive than concatenated slots in terms of V, $V1$, and $\mathscr{P}$. As expected, the elements of structural variation present in exact matches have a positive impact on the productivity of the construction. Once deprived of those elements, *A as NP* displays a higher degree of conventionalization due to the restrictions in the range of its A-NP combinations in the sample. As expected, again, A and NP slots are less productive when they are considered individually than when

they are combined. This is due to the verbatim recurrence of most adjectives and, to a lower extent, most NPs.

The high S scores in Table 9 mean that there are many types yet to be discovered if one went on browsing more data of the BNC kind. S is especially high for the NP slot, which is hardly surprising given the structural variation inherent to NPs (i.e. determiners + number agreement + the potential presence of postmodifiers). In theory, combining these elements yields a high number of yet unseen possible types.

The right part of Table 9 shows which LNRE model provides the most adequate fit for each distribution. The goodness of fit of the LNRE model to the empirical distribution of each configuration is evaluated with a multivariate chi-squared test (Baayen 2001, Section 3.3.). According to Baayen (2008, 233), "[f]or a good fit, the χ^2-value should be low, and the corresponding p-value large and preferably well above 0.05". The models are globally satisfactory and can therefore provide reliable interpolated and extrapolated expected values for the vocabulary size and the spectrum elements. These values are plotted on VGCs in Figure 1.

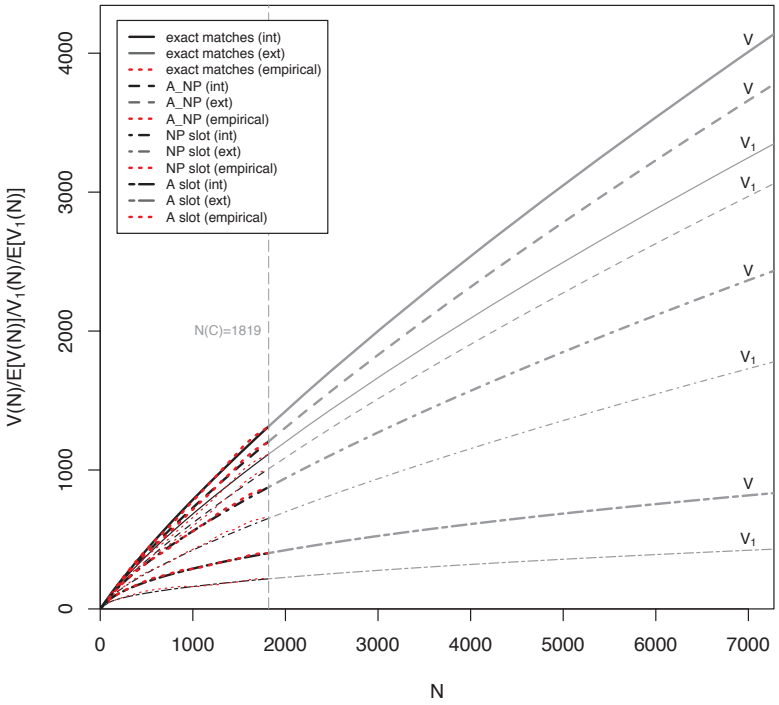

Figure 1: Vocabulary growth curves for exact *A as NP* matches, concatenated slots (A_NP), the A slot, and the NP slot with LNRE interpolations (int) and extrapolations (ext) to four times the sample size

Due to the limited size of $N(C)$, all configurations are located in the early LNRE zone, where V and $V1$ are all increasing, and asymptotic limits are far outside the boundaries of the plot (see S scores in Table 9). However, inspecting the VGCs reveals that the differences observed in Table 9 are likely to hold whatever the sample size (except for the NP slot judging from its large S score).

As seen at the bottom of section 2.2.2, a productivity measurement of the two slots is all the more relevant as we have good reasons to assume that the A slot and the NP slot are interdependent. Following Zeldes (2012, 130), we can show this interdependence by computing the probability of a *A as NP* hapax legomenon. For N tokens, this probability should be 1 minus the product of the complementaries to $\mathscr{P}$ in each slot:

$$P(HL_{A\,slot} \cup HL_{NP\,slot}) \approx 1 - (1 - \mathscr{P}_{A\,slot})(1 - \mathscr{P}_{NP\,slot}) \qquad \text{(xiii)}$$

The prediction in equation (xiii) is not good for *A as NP*, where $1-(1-\mathscr{P}_{A\,slot})(1-\mathscr{P}_{NP\,slot}) = 0.44$

(vs. $\mathscr{P}_{A-NP} = 0.56$). We observe a 20.6% deviation from the expected probability of hapax legomena. This means that lexical choice is mutually constrained in the two slots of the construction. It also means that if we see a given slot (A or NP), we have a reasonably strong expectation about the other slot. What is left to explore is the direction of this expectation. This is what we now turn to.

4.2 Symmetry & asymmetry

To measure the asymmetric dependencies at work in *A as NP*, $\Delta P_{NP|A}$ and $\Delta P_{A|NP}$ were computed for each A-NP type found in the BNC. Figure 2 displays the sorted pairwise differences $\Delta P_{NP|A} - \Delta P_{A|NP}$ for each *A as NP* subschema based on A-NP types. It reveals that most of them are asymmetric. Among the 1206 A-NP pairs in the sample, 686 display a large difference ($-0.5 \geq \Delta P \geq 0.5$) and 153 are characterized by an absence of covariation ($\Delta P = 0$). About 80% of those that show asymmetry are observed when the difference $\Delta P_{NP|A} - \Delta P_{A|NP}$ is negative, i.e. when the presence of the NP increases the likelihood of the adjective.

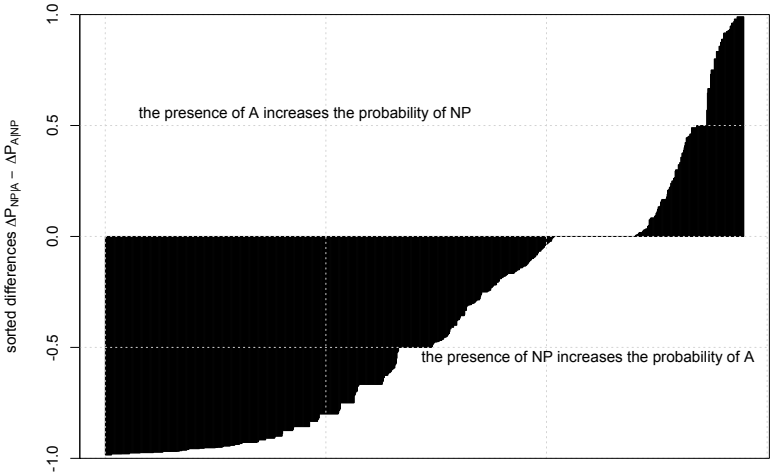

Figure 2: Sorted pairwise differences between ΔP values for 1206 A-NP pairs in the BNC

To appreciate the detail of those asymmetries, let us follow Gries (2013) and observe the distribution of A-NP pairs with regard to both the sorted pairwise differences of ΔP values and a bidirectional association measure (Dunning's log-likelihood ratio score). Initially, Gries's idea is to show how symmetric measures often return high association scores regardless of collocation asymmetries. At first sight, the shape of the cloud of data points in Figure 3 confirms that the degree of symmetric association and the lack of covariation between A and NP are not correlated. A line of data points stretches vertically where $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ is close to 0. Admittedly, the most strongly associated A-NP pairs at the top of this line are among the least asymmetric tokens: *good_gold*, *quick_flash*, *right_rain*, etc. Yet, such is also the case of some the least strongly associated A-NP pairs: *black_hell*, *strong_death*, *soft_silk*, etc. More interestingly, a denser line of data points follows a concave curve whose extremities converge to a value of $\log_{10} G^2$ that approximates 1 where $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ has extreme values (i.e. close to -1 and 1).[14] Table 10 zooms in on the hardly legible margins of Figure 3 and compares them to the top and bottom margins (i.e. collostruction strength).

Table 10a displays the A-NP pairs with the highest collostruction strengths, whereas Table 10b displays the A-NP pairs with the weakest association score. As seen above, all pairs with extreme G^2 values (whether high or low) are characterized by little or no covariation. All the most strongly associated A-NP pairs are highly conventional and used as popular idioms. For each A-NP pair in Table 10c, the NP is a much better predictor of the adjective than vice versa. Conversely, Table 10c displays pairs for which the adjective is a much better predictor of the NP than vice versa. We find 4 instances of *hell* in Table 10b, where the collostruction strength is the weakest. Interestingly, the

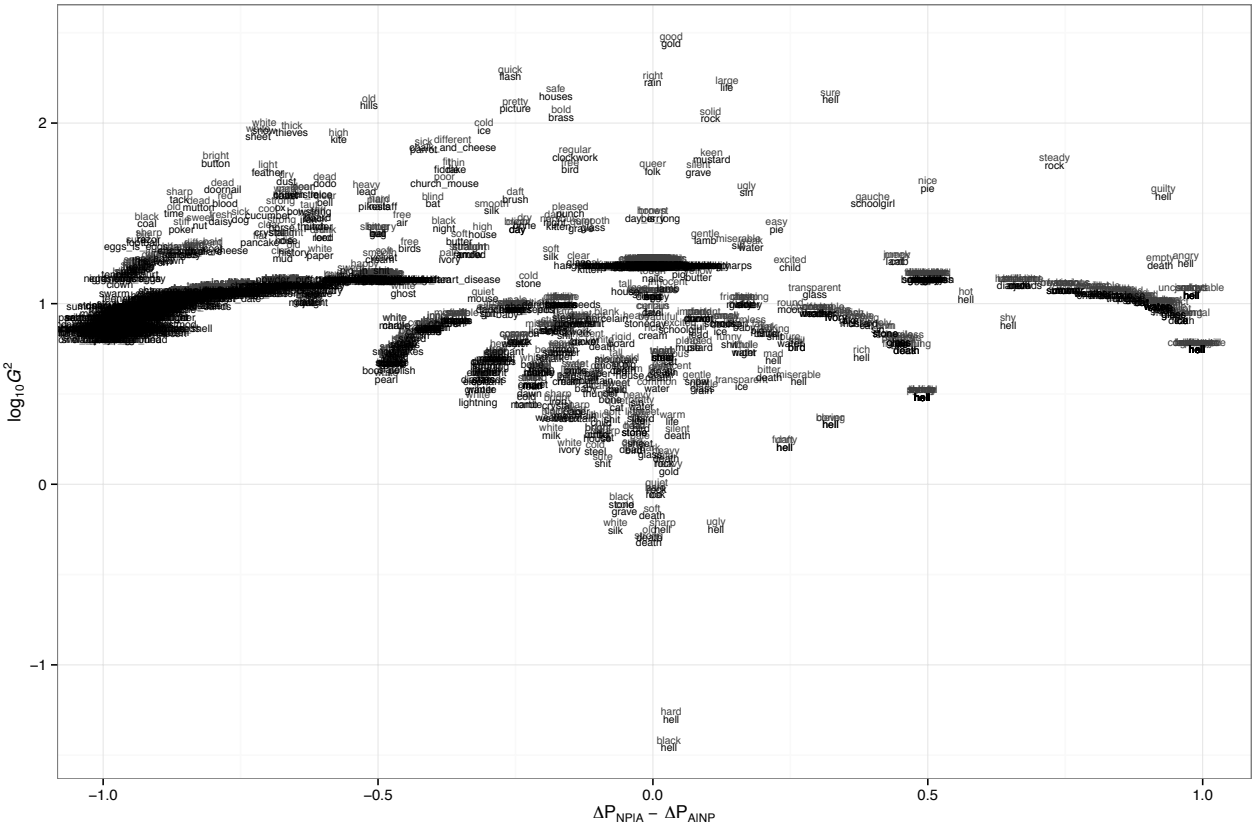

Figure 3: Distribution of A-NP pairs according to ΔP differences (x-axis) and log-likelihood ratio (y-axis, logged)

Table 10: A comparison of the top, bottom, left and right margins of Figure 3

a. Top 10 covarying collexemes (top margin)				**c. Top 10 negative ΔP values (left margin)**			
A	NP	G^2	$\Delta P_{NP\|A} - \Delta P_{A\|NP}$	A	NP	G^2	$\Delta P_{NP\|A} - \Delta P_{A\|NP}$
good	*gold*	288.8138	0.033333323	*white*	*ash*	6.6782	-0.984614734
quick	*flash*	189.2885	-0.259259188	*white*	*candlewax*	6.6782	-0.984614734
right	*rain*	175.9832	0	*white*	*desert*	6.6782	-0.984614734
large	*life*	164.5468	0.134920604	*white*	*fire*	6.6782	-0.984614734
safe	*houses*	148.3236	-0.176470558	*white*	*flour*	6.6782	-0.984614734
sure	*hell*	141.9763	0.323264818	*white*	*jellyfish*	6.6782	-0.984614734
old	*hills*	130.8729	-0.51612887	*white*	*office paper*	6.6782	-0.984614734
pretty	*picture*	126.4332	-0.249999959	*white*	*quicklime*	6.6782	-0.984614734
bold	*brass*	113.2006	-0.166666646	*white*	*snow* + postmod.	6.6782	-0.984614734
solid	*rock*	110.9541	0.104895064	*white*	*towel* + postmod.	6.6782	-0.984614734

b. Bottom 10 covarying collexemes (bottom margin)				**d. Top 10 positive ΔP values (right margin)**			
A	NP	G^2	$\Delta P_{NP\|A} - \Delta P_{A\|NP}$	A	NP	G^2	$\Delta P_{NP\|A} - \Delta P_{A\|NP}$
black	*hell*	0.0366	0.029387267	*awkward*	*hell*	5.8518	0.989794932
hard	*hell*	0.0527	0.033273417	*bleary*	*hell*	5.8518	0.989794932
strong	*death*	0.5004	-0.007211477	*conservative*	*hell*	5.8518	0.989794932
old	*death*	0.5367	-0.006203423	*depressed*	*hell*	5.8518	0.989794932
white	*silk*	0.585	-0.067948179	*difficult*	*hell*	5.8518	0.989794932
sharp	*hell*	0.5895	0.017573066	*frustrated*	*hell*	5.8518	0.989794932
ugly	*hell*	0.5927	0.114795003	*gloomy*	*hell*	5.8518	0.989794932
soft	*death*	0.7054	-0.001424491	*lonesome*	*hell*	5.8518	0.989794932
cold	*grave*	0.7417	-0.051020052	*nosy*	*hell*	5.8518	0.989794932
black	*stone*	0.8146	-0.056922701	*nutty*	*hell*	5.8518	0.989794932

same NP is found in all the A-NP combinations in Table 10d, where the positive ΔP values reach their highest. This suggests that there is local correlation in the data between a low collostruction score and a high ΔP score. This tendency is confirmed if we compare Tables 10b and 10c, where the negative ΔP values reach their lowest. All the adjectives in Table 10b also appear in pairs with extreme negative

ΔP values (i.e. $\Delta P \leq -0.9$). If we were to extend Table 10c to the top 300 values, we would find only 37 adjective types, all of which denoting basic properties such as colors and shades (*white*, *black*, *red*, *pale*, *clear*), texture and constitution (*big*, *stiff*, *hard*, *heavy*, *solid*, *strong*, *smooth*, *soft*, *sharp*, *thick*, *dry*), age (*old*), speed (*quick*), mental/physical states (*happy*, *sick*, *dead*), temperature (*hot*, *cold*), and epistemic judgment (*sure*). The semantic similarity between groups of tokens being a productivity criterion according to Croft and Clausner (1997), we could hypothesize that productive adjectives and NPs are found where asymmetry is attested.

To interpret the above results at the type level, the mean scores of $\Delta P_{(NP|A)} - \Delta P_{(A|NP)}$ and G^2 were calculated for each A and each NP. Collostruction strength and ΔP can now be compared to two productivity measures described in Section 2.2.2, namely P^*, and $\mathscr{P}$, notwithstanding caveats regarding the difficulties inherent to comparing the latter values at different sample sizes.

4.3 ΔP, collostruction strength, and hapax-based productivity measures

As originally designed in morphology, $\mathscr{P}$ is calculated by dividing the number of hapax legomena with a given affix by the total number of tokens of that affix in the corpus (see Section 2.2.2). Although $\mathscr{P}$ works well in the productivity assessment of a single affix, difficulties emerge when comparing the productivity scores of several affixes with different sample sizes. As pointed out by Gaeta and Ricca (2006), Baayen and Lieber (1991) obtain counterintuitive results when they compare deverbal suffixes with different token frequencies on the basis of the size of the whole corpus. According to Gaeta and Ricca, computing $\mathscr{P}$ values using the whole corpus size as an independent reference point amounts to comparing these values at the endpoints of as many curves as there are different suffixes, each endpoint corresponding to a distinct value of N. Given that $\mathscr{P}$ is a decreasing function of N, this can lead to overestimating $\mathscr{P}$ for the less frequent items, especially when token frequencies differ greatly in the sample. For this reason, Gaeta and Ricca (2006) insist on computing $\mathscr{P}$ for equal values of N.[15] This is no minor constraint as, at the current state of the art, hapax-based productivity measures best apply to macro-scale studies based on preferably very large corpora and case studies (such as affixes) with token frequencies that are *a priori* high enough to return a critical mass of hapax legomena.

Using texts from continuous issues of the Italian daily newspaper *La Stampa* over a period of 36 months, Gaeta and Ricca (2006, Section 2) adopt a variable-corpus approach to evaluate the productivity of Italian affixes. The corpus is split into subcorpora. For each subcorpus, a token-frequency spectrum of word forms is created, on the basis of which type, token, and hapax legomena are calculated for each affix. The procedure advocated by Gaeta and Ricca (2006) allows for a direct comparison of affix productivity scores because it is assumed that the distribution of a given affix in any subcorpus is proportional to its distribution in the whole corpus. For this to work, however, the subcorpora must be homogeneous with respect to their textual typology. Although the variable-corpus approach is feasible and relevant in the case of a study of affixes in the *La Stampa* corpus (36 months can easily be divided into 36 similar subcorpora, each subcorpus displaying a critical mass of tokens), respecting the distribution of text types found in the BNC in the study of *A as NP* is far more challenging and beyond the scope of this article.

In the context of *A as NP*, one canonical procedure could involve breaking down $N(C)$ into as many sample sizes as there are slot types, i.e. 402 adjective slot types + 876 NP slot types = 1278 sample sizes. In order to compare $\mathscr{P}$ values over the whole data set, one would need to set N to the smallest common sample size, which would make little sense given the presence of many rare types in the overall sample. Many of these rare types consist of a single hapax legomenon, in which case $\mathscr{P}$=1 for whichever slot is considered the variable. Such an extreme score may be poorly indicative of productivity (these rare types may just be archaisms or nonce innovations). As inspired by Zeldes (2012, Table 15), one could instead compare the potential productivity of only the most frequent A slots and NP slots, which would guarantee a maximal common sample size. However, that would be at the cost of losing potentially interesting items from the viewpoint of other measures, notably ΔP and G^2.

In the remainder of the paper, I adopt a workaround that deviates from the canonical procedure

outlined in Gaeta and Ricca (2006) and Zeldes (2012). The above caveats notwithstanding, I consider that slot types are extracted from the same sample of *A as NP* slot types and therefore $N(C)$=1278. This has the advantages of avoiding extreme productivity scores and paving the way for a complex scale of productivity. It may only be the second- or third-best workaround (and it is certainly not the only workaround available), but at least it is a procedure whose main shortcoming we are aware of, thanks to Gaeta and Ricca (2006), namely overestimating the productivity of low-frequency items when doing fixed-corpus calculations. The ensuing $\mathscr{P}$ and P^* measurements must therefore be considered with caution even though, fortunately, no such overestimations were found after cross-checking with the original frequency data. If the productivity of *A as NP* were assessed with hapax-based measures only, the smallest-common-sample-size issue might be insuperable. Yet, in this paper, $\mathscr{P}$ and P^* are but two among several factors in the multidimensional productivity assessment of *A as NP*. In Section 4.4, it is the role of Principal Component Analysis to weight hapax-based measure against other metrics and include illustrative frequency variables to help identify potentially counterintuitive data points.

Given formula (iii), P^* can be plotted with V on the y-axis and $\mathscr{P}$ on the x-axis, as in Figure 4. Each point corresponds to either an adjective type or a NP type. We should expect the most globally productive lexemes to appear in the upper-right part of the plot. In this respect, the NP *hell* stands out in the sense that *A as hell* is far from having exhausted its combinations with other A types. Next we find a cluster of adjectives that belong to highly asymmetric pairs. These adjectives denote the following properties: colors and shades (*white*, *black*, *bright*, *clear*, *pale*), texture and constitution (*big*, *sharp*, *thick*, *strong*, *stiff*), and temperature (*cold*).

To see how ΔP performs comparatively, the absolute value of $\Delta P_{NP|A} - \Delta P_{A|NP}$ is plotted instead of V in Figure 5. The same items as in Figure 4 stand out as indices of the most productive subschemas in the upper-right corner of the plot. The most potentially productive lexemes are also among the most asymmetric. However, despite being part of highly asymmetric pairs, the data points in the upper-left corner of Figure 5 are poorly productive. These items do not readily form nonce A-NP combinations.

Finally, to see if collostruction strength provides a reasonable estimation of the lack of productivity (see Section 4.2), the mean G^2 score for each A and NP is plotted against $\mathscr{P}$ in Figure 6. We observe an inverse correlation for extreme values. Indeed, the most productive items according to $\mathscr{P}$ have among the lowest G^2 scores: *hell*, *big*, *sharp*, *bright*, *cold*, *black*, *strong*, *pale*, *thick*, *white*, and *stiff*. Conversely, the items with the highest collostruction strengths are among the least potentially productive: *gold*, *bold*, *right*, *rain*, *queer*, and *folk*. Two A-NP pairs match highly conventional constructions: *right-rain* (as in *right as rain*) and *queer-folk* (as in *queer as folk*). Because these constructions combine a high degree of (exclusive) association and a low productivity score, they rank among the most autonomous units in our sample. The added value of collostruction strength is important for extreme values as well: items that occur in the most strongly associated pairs are far less productive than the other items.

Although insightful, two-dimensional plots such as the above make it difficult to weight the relative contributions of the horizontal and vertical dimensions. As Baayen (1992, 124) concedes, one cannot rely on plots such as Figure 4 to derive meaningful P^* rankings when both extent of use (V) and degree of productivity ($\mathscr{P}$) are different for non-competing items. For example, *gold* and *sure* have large P^* scores ($P^*_{gold} \approx 54570$ and $P^*_{sure} \approx 22738$), but such scores are not reflected in the plot. Criticisms can also be made against the other plots. In Figure 5, even though ΔP_diff and the probability of encountering new types are correlated locally, we are left to wonder how to account for the data points in the top-left corner. In Figure 6, even though an inverse correlation is observed for data points with extreme values, those in the bottom-left corner of the plot are not accounted for.

4.4 Summary

Ideally, we would want to summarize the contributions of all the above measures to the "Productivity Complex" of *A as NP* in the BNC. To this aim, the mean ΔP and collostructional strength scores for each adjective and NP appearing in *A as NP* were collected together with V, $V1$, $\mathscr{P}$, and P^*. The figures were summarized in a data frame which was submitted to principal component analysis, henceforth PCA (Baayen 2008; Husson, Lê, & Pagès 2011).[16] PCA is a multifactorial method whose purpose

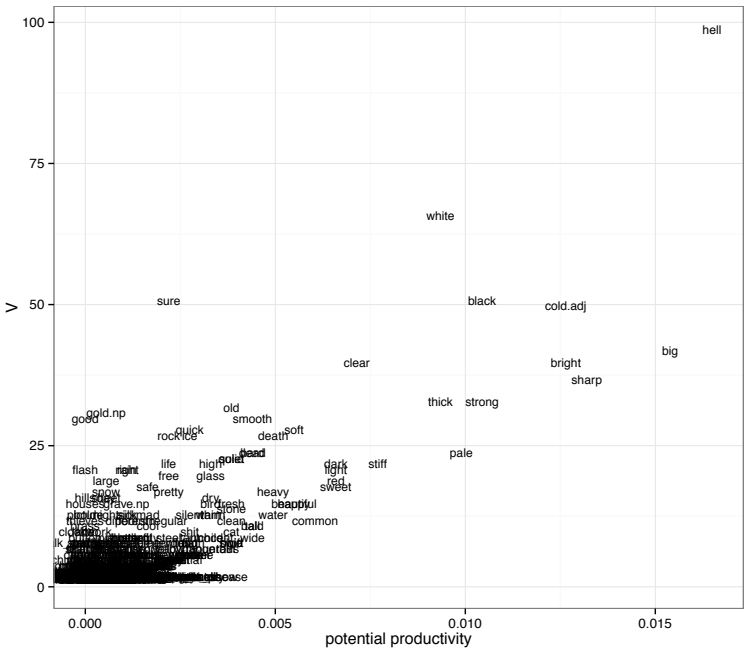

Figure 4: Global productivity for As and NPs in *A as NP* in the BNC

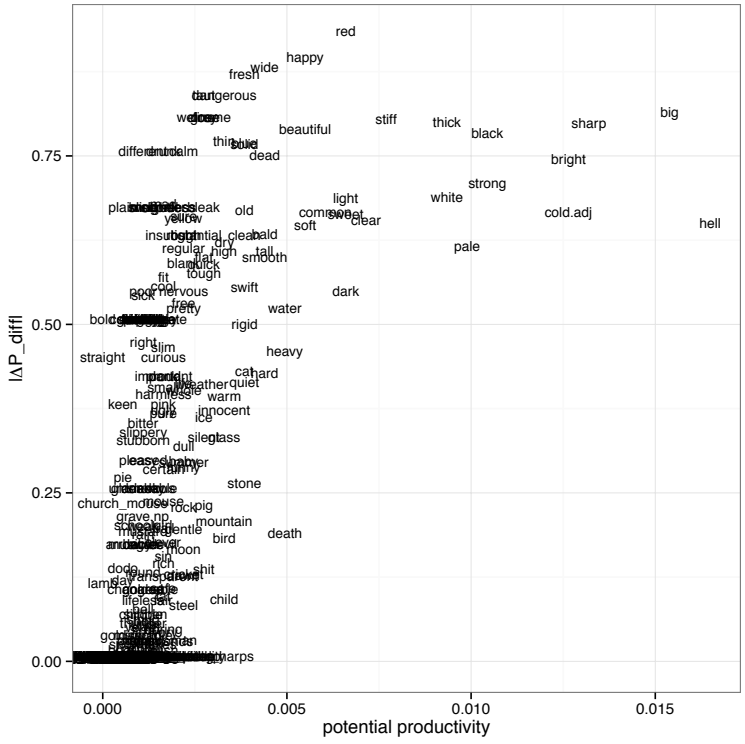

Figure 5: Distribution of A-NP pairs according to $\mathscr{P}$ (x-axis) and the mean absolute value of ΔP differences (y-axis)

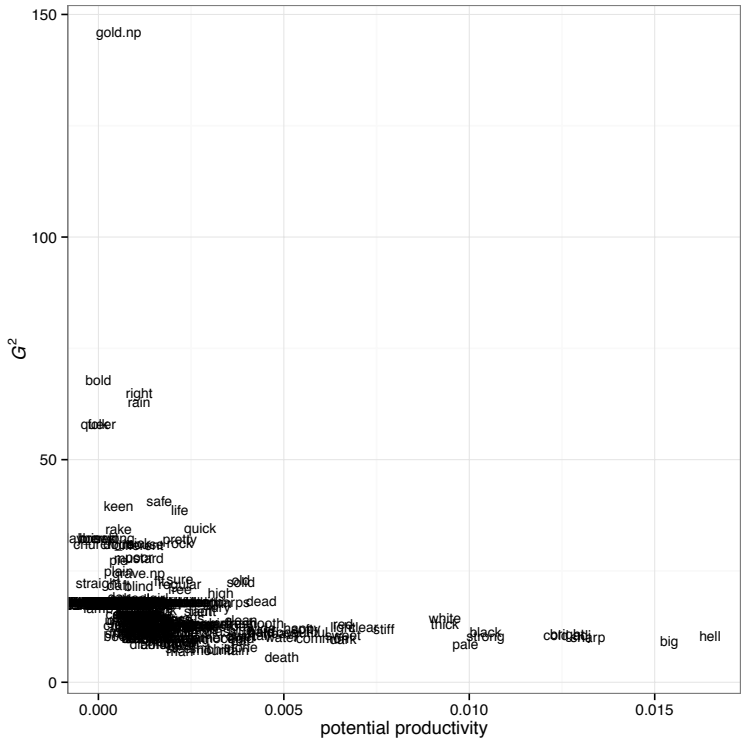

Figure 6: Distribution of A-NP pairs according to $\mathscr{P}$ (x-axis) and mean collostruction strength (y-axis)

is to explore multidimensional data tables where rows are considered as individuals and columns as variables. The individuals consist of all adjective and NP types of *A as NP* tokens. Each of the 1278 individuals (402 adjective types and 876 NP types) was examined in light of four active variables: the mean collostructional strength, the mean difference $\Delta P_{NP|A} - \Delta P_{A|NP}$ (henceforth ΔP_diff), $\mathscr{P}$, and P^*. Three supplementary quantitative variables were also projected to verify that no counterintuitive results were obtained with respect to the non-canonical computation of hapax-based measures (see Section 4.3): V, $V1$, and construction frequency (henceforth const_freq).[17] The variables were centered and standardized.

Inspection of Figure 7 reveals that the first component (i.e. the horizontal axis), which accounts for one third of the variance, is largely correlated with collostruction strength ($r \approx 0.79$) and P^* ($r \approx 0.75$). A positive correlation exists with ΔP_diff but it is much smaller than that of the other variables ($r \approx 0.42$). This is because ΔP_diff contributes more to the third component (see below). The second component (i.e. the vertical axis) accounts for 15.39% of the variance. It is mostly correlated with $\mathscr{P}$ and $V1$ ($r_{\mathscr{P}} = r_{V1} \approx 0.95$),[18] V ($r \approx 0.8$), and to a much lesser extent P^* ($r \approx 0.33$). In other words, the further up along the vertical axis, the more potentially productive the individuals.

Figure 8 charts the two-dimensional space spanned by the first two principal components. So as not to clutter the graph, only the individuals that contribute the most to each dimension were projected. Other individuals appear as dots without a label. Three clusters stand out: individuals that are productive according to $\mathscr{P}$ cluster along the vertical axis in the upper part of the plane, globally productive elements cluster in the upper-right corner of the plane along the horizontal axis, and individuals that occur in A-NP pairs with a high collostruction strength cluster along the same axis, but in the lower-right corner.

As seen in Figure 6, the most productive individuals according to $\mathscr{P}$ belong to weakly associated pairs. Some marginal tendencies observed in Figure 3 and Table 10 of Section 4.2 are confirmed. The subschemas indexed on productive adjectives denote basic properties such as:

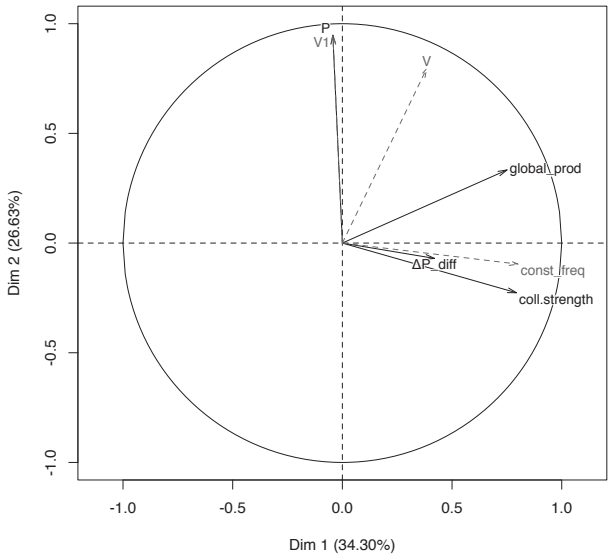

Figure 7: PCA – 4 active variables (plain arrows) and 3 supplementary variables (dashed arrows) in dimensions 1 & 2

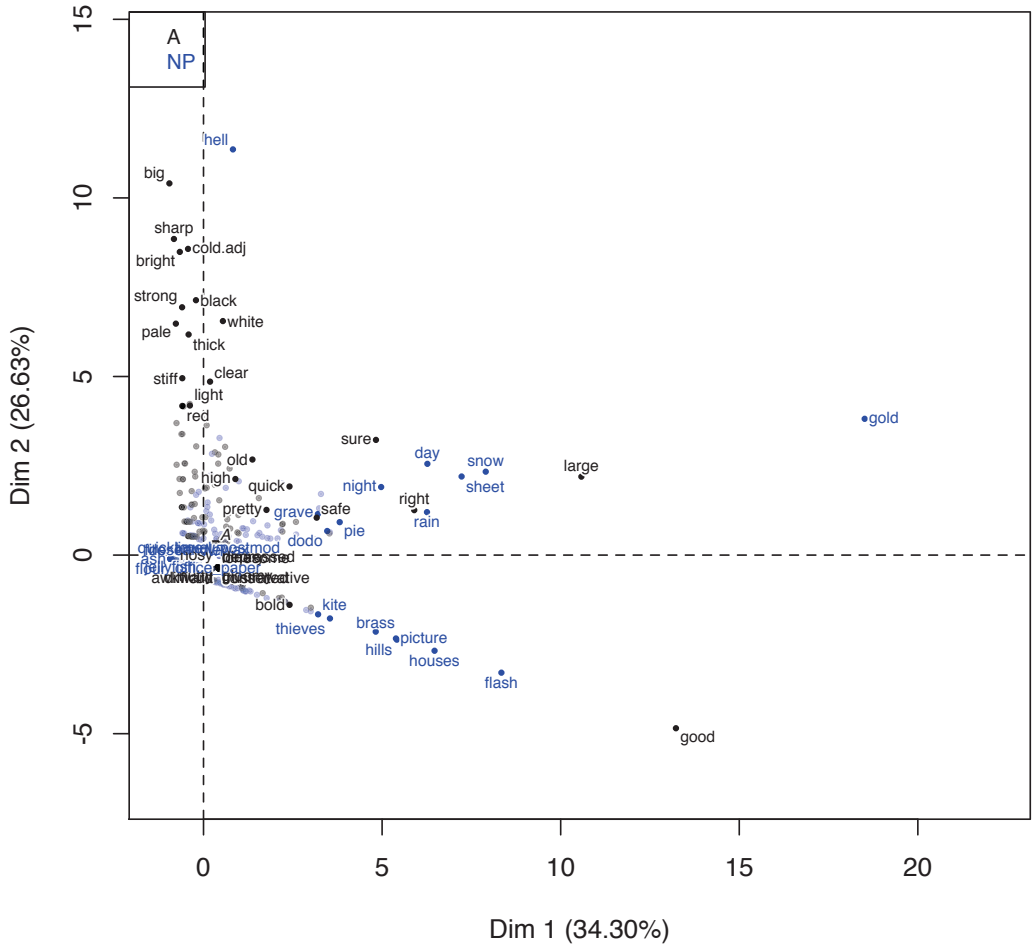

Figure 8: PCA – plane representation of individuals in dimensions 1 & 2

colors and shades	*black as*	*anthracite*, *coal*, *midnight*, *thunder*, etc.
	white as	*marble*, *paper*, etc.
	red as	*blood*, *a tomato*, *holly*, etc.
	clear as	*crystal*, *mud*,[19] *(a) spring*, etc.
	bright as	*stars*, *a button*, *day*, *copper*, etc.
	pale as	*ice*, *a nun*, *cream*, *ivory*, etc.
texture and constitution	*big as*	*a giant*, *a bus*, *an elephant*, *a house*, etc.
	sharp as	*a knife*, *a razor*, *a blade*, *daggers*, etc.
	strong as	*a horse*, *a tree*, *a viking*, *a bull*, etc.
	thick as	*thieves*, *mud*, *a book*, etc.
	stiff as	*a board*, *a rod*, *a fence post*, etc.
	light as	*a feather*, *a harebell*, *thistledown*, etc.
temperature	*cold as*	*ice*, *slate*, *debt*, *stone* etc.

On the one hand, a rather straightforward semantic correspondence can be found between the adjective and the NP even when the subschema is used figuratively: granite is hard, crystal is clear, ice is cold, thieves are thick in the sense that they are close, etc. On the other hand, regarding the most productive subschema *A as hell*, the NP seems to have lost its literal meaning to the benefit of an exclusively intensifying function: *angry as hell*, *guilty as hell*, *hot as hell*, *jealous as hell*, *rich as hell*, *rusty as hell*, *shy as hell*, *uncomfortable as hell*, etc. Yet, while conventional, *A as hell* is flexible enough to accommodate new adjectives.

Conversely, individuals that belong to highly associated pairs are among the least potentially productive. The positive correlation between P^* and G^2 along the horizontal axis suggests that the individuals characterized by these variables are used more than once. For G^2, this is evidenced by its proximity to the construction frequency variable ($r_{coll.strength} \approx r_{const\_freq} \approx 0.8$). For P^*, this is evidenced by its position in the upper part of the plane under the influence of V (see formula (iii)).

Individuals with extreme values for the first component are mostly nouns. Most nouns with the highest P^* values denote paragons whose semantic relation with the adjective can be easily accessed or reconstructed despite its conventional nature: day is bright or clear; night is black, dark, or dull; snow and sheets are white; a grave is easily conceived as cold, quiet, or silent, and a dodo as dead. Most nouns with the highest G^2 values denote paragons whose semantic relation with the adjective is more remote. Because these lexemes belong to highly conventionalized expressions, it takes more than common sense to guess the semantic relation that holds between the adjective and the noun in some expressions, e.g. *bold as brass*, *safe as houses*, and *old as the hills*. Often, it is the phonetic affinity between two lexemes that conditions their pairing, as in *good as gold* (both lexemes share /g/), *thick as thieves* (/θ/), *bold as brass* (/b/), *high as a kite* (/ai/), and *pretty as a picture* (/p/ + /t/).

The difference between P^* (upper part) and G^2 (lower part) can be explained as follows: globally productive individuals are more likely to be used in new formations than individuals belonging to strongly associated pairs. Despite being used often as an expression, *right as rain* is therefore not as fixed as Figure 6 suggests because *right* and *rain* are each likely to be used in other A-NP combinations. With respect to *good as gold*, only *gold* is more likely to be used in other combinations with other adjectives.[20] As far as *bold as brass* is concerned, neither *bold* nor *brass* are likely to be used in other A-NP pairs. In other words, as we move down from the upper-left to the bottom-right part of the plot, productivity declines and conventionalization and autonomy increase.

Figure 8 further confirms the relevance of a two-slot productivity assessment. The individual behavior of lexemes does not necessarily match the behavior of the same lexemes once they are paired. Two dimensions of productivity may underlie the same subschema. In *cold as a grave* and *black as night*, the adjective is potentially productive whereas the NP is globally productive. In *thick as thieves*, *thick* is potentially productive but *thieves* is not. In *sure as hell*, *sure* is globally productive whereas *hell* is potentially productive. In *old as the hills*, *pretty as a picture*, and *high as a kite*, the adjective is globally productive, whereas the NP is not.

Figure 9 shows that the contribution of ΔP_diff is manifest in the third component ($r \approx 0.9$)

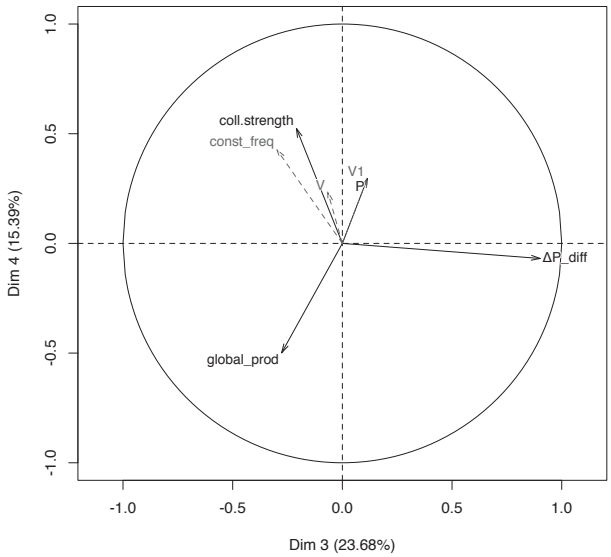

Figure 9: PCA – 4 active variables (plain arrows) and 3 supplementary variables (dashed arrows) in dimensions 3 & 4

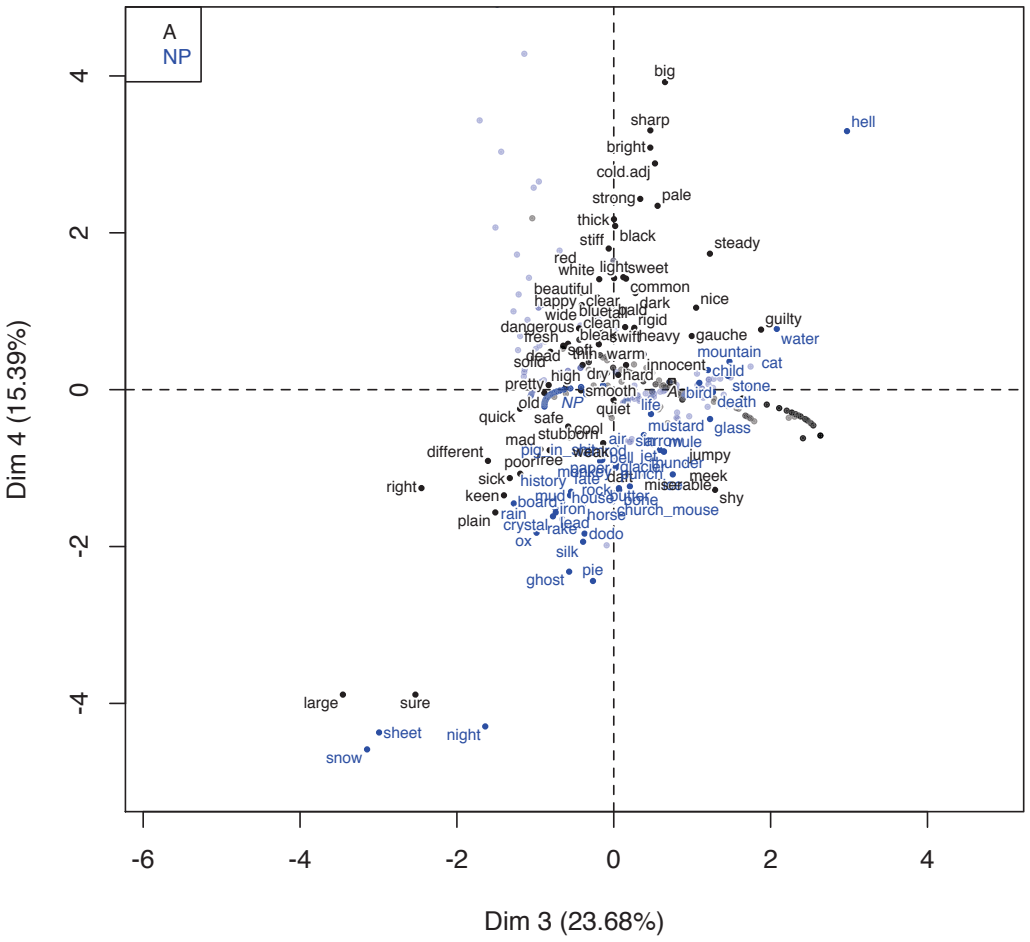

Figure 10: PCA – plane representation of individuals in dimensions 3 & 4

along with $\mathscr{P}$ and $V1$ ($r_{\mathscr{P}} \approx r_{V1} \approx 0.11$). This component is negatively correlated with coll.strength ($r \approx -0.21$), global productivity ($r \approx -0.28$), and construction frequency ($r \approx -0.3$). In the fourth and last component, coll.strength displays the maximum positive correlation ($r \approx 0.52$), followed by const_freq ($r \approx 0.43$), $\mathscr{P}$, $V1$ ($r_{\mathscr{P}} \approx r_{V1} \approx 0.3$), and V ($r \approx 0.23$). P^* displays the maximum negative correlation ($r \approx -0.49$), followed by ΔP_diff ($r \approx -0.07$). Individuals stretching towards the upper-right part of the plot belong to potentially productive asymmetric pairs. Individuals stretching towards the lower-right part of the plot belong to globally productive asymmetric pairs.

Figure 10 charts the two-dimensional space spanned by the third and fourth principal components, zooming in on the selected individuals with the highest $\mathscr{P}$ and P^* values. If the individuals belonging to asymmetric pairs were characterized by a high ΔP score only, they would stretch far along the right end of the horizontal axis. Such is not the case. The fact many data points stretch from the lower-right corner of the plot to the upper-right and lower left-corners suggests that most asymmetric subschemas are also potentially or globally productive.

Given the above, even if asymmetry alone is not sufficient to measure productivity, it nevertheless helps spot loci of productivity for a multiple-slot construction at subschematic levels. The precise nature of productivity (or the lack thereof) depends on how these subschemas are characterized by type frequency, the number of hapax legomena, or their ratio.

5 Discussion

It might be objected that ΔP applies too conviently to our case study because of the presence of two lexical slots which nicely fit in a co-occurrence table such as Table 6. Yet, this method applies equally well to constructions that are patterned differently. Gries (2013, 154–155) suggests one way in which ΔP can be extended to the study of multi-word units with more than two lexical items. Through an iterative process, collocations characterized by high ΔP values can be amalgamated into larger multi-word patterns until all amalgamation possibilities have been exhausted. As regards constructions with only one lexical slot (typically, such constructions are amenable to collexeme analysis),[21] one straightforward option – already implemented in Coll.analysis 3.2a (Gries 2007) – is to consider, for instance, the construction schema as outcome and the collexeme as cue and modify Table 6 accordingly. We obtain Table 11.

Table 11: Co-occurrence table involving a cue (construction) and an outcome (collexeme)

	collexeme: present	collexeme: absent
construction: present	a	b
construction: absent	c	d

Again, two ΔP values should be computed, depending on whether the cue is the construction schema and the outcome is the collexeme or vice versa:

$$\begin{aligned} \Delta P_{(collexeme|construction)} &= p(collexeme|construction) - p(collexeme|\neg construction) \\ &= \frac{a}{a+b} - \frac{c}{c+d} \end{aligned} \tag{xiv}$$

$$\begin{aligned} \Delta P_{(construction|collexeme)} &= p(construction|collexeme) - p(construction|\neg collexeme) \\ &= \frac{a}{a+c} - \frac{c}{b+d} \end{aligned} \tag{xv}$$

If $\Delta P_{(collexeme|construction)} - \Delta P_{(construction|collexeme)}$ is positive, then the construction schema is a better predictor of the collexeme than vice versa. Conversely, if the difference is negative, then the collexeme is a better predictor of the construction schema than vice versa. If the difference is negative, there is no covariation between the construction schema and the collexeme.

It could also be objected that some constructions are not equally productive despite identical ΔP scores.[22] Such would be the case of the *NPN* construction, where both NPs are identical, as

exemplified by *construction after construction* or *door to door* (Jackendoff 2008). Jackendoff (2008, 13) observes that "[i]n general, *N by N* seems more productive than *N to N*." Vocabulary measures could easily confirm this intuitive productivity assessment. If one computed attractions based on ΔP for NPN, as one would for any two-slot construction where the two slots are distinct, one would inevitably find no asymmetry. Yet, because the NPs are identical, it would make little sense to look for asymmetry between the two slots. One should instead adopt the methodology outlined in the above paragraph regarding one-slot constructions. More challenging would be the case of *the Xer the Yer*, which has lower productivity for *X* and *Y* than *the Xer NP VP, the Yer NP VP* (Zeldes 2012), although both adjectives can be novel in both constructions. Here, a combination of symmetric and asymmetric association measures would certainly not replace the productivity measures used in the previous section, but they would certainly be helpful to (a) compare the degrees of conventionalization and autonomy of the two forms of comparative correlatives both at schematic and subschematic levels, and (b) determine the direction of expectation between the two mutually constrained slots in each construction.

Gries (2013) provides two validations of ΔP, the first involving bigrams with a high mutual association, the second based on a study of randomly chosen bigrams. In line with the first validation, future research will have to show that our results differ significantly from those that one would obtain with syntactically similar but functionally unrelated patterns, such as those where *as* is not an adverb but a preposition, and where no intensification is expressed, as in (23), (24), and (25):

(23) I was never **happy as a child**. (BNC-FPK)

(24) Have your policy or claim number **ready as a reference**. (BNC-HB5)

(25) It is not **admissible as evidence**, (...). (BNC-JJV)

In line with the second validation, future research will have to show that my results differ significantly from those that one would obtain with random A-NP pairs, regardless of how they pattern.

Pending validation, I believe my findings have important implications for the study of constructions. If productivity is key in determining whether an expression is a construction, then we obtain a more detailed picture when we examine different levels of schematicity than by considering exclusively the most schematic level. We gain a lot by examining frequencies in a corpus rather than intuitively. In the context of multiword expressions, these frequencies make sense when inspected in light of both productivity and association measures.

The methodology presented here could be extended to the study of what BCG considers as fully productive patterns such as *all*-clefts, *what's X doing Y*, or *let alone*. It may well appear that syntactic, semantic, and contextual constraints influence the joint distribution of each of these constructions' constituents, thus putting their full productivity into perspective.

6 Conclusion

Some linguists might argue that whether construction grammar should be "severe" or less so is, after all, a matter of theoretical preference. Judging from the inherent bond established between productivity and constructional status in BCG, I believe otherwise. Firstly, arbitrary sets of bigrams may well be very productive according to any of the above measures, but this does not mean that they constitute a construction (Baayen 2001, 221). Secondly, even the most schematic construction will not generate an infinity of instances. Thirdly, intermediate productivity levels must be recognized. *A as NP* may not be fully productive at the schematic level, but it generates subschemas whose productivity is doubtless.

For a construction not to be fully schematic does not necessarily involve that it should not be granted constructional status or that it is unproductive. In fact, construction grammar approaches assume even fully specified schemas as constructions. For this reason, the productivity assessment carried out in this paper was not primarily meant to determine what is a construction and what is not (although I showed that *A as NP* is more productive than Kay suggests), but to recognize the existence of productive subschemas underlying *A as NP* and compare them.

All in all, the evidence from this study suggests that (a) there is more to the productivity of a multiple-slot construction than high-schematic-level type frequency, (b) the borderline between "constructions" and "patterns of coining" cannot be set introspectively, and (c) multiple-slot schemas, as statistically significant sequences of constituents, are amenable to principles of associative learning, especially the fact that some constituents are cues for the outcome of others.

This paper does not aim to replace existing measures of (constructional) productivity. The methodology I have proposed is but another way of exploring the multitiered productivity of complex, multiple-slot expressions. I believe that analyses based on traditional productivity measures can only gain from a combination with symmetric and asymmetric methods. Quantitative assessments of this kind may arguably find their place in the inventory of existing productivity measures in the capture of what Zeldes (2012) calls "the multidimensional productivity complex". More generally, any measure that reflects human language processing is an undeniable asset in corpus linguistics.

Notes

[1]The idea that grammar has a center and a periphery echoes one of the main tenets of generative linguistics, which BCG is close to alongside other sign-based theories such as HPSG.

[2]For some expressions, even though the correspondence between A and NP is literal, their combination in *A as NP* becomes figurative. Such is the case of *cold as ice*. Ice is cold, but *cold as ice* has little to do with low temperatures: e.g. *His mind went cold as ice.* (BNC-ARK)

[3]Although the origin of *safe as houses* is unclear, folk etymology may relate it to hiding places known as "safe houses".

[4]Kay (2013, 37) concedes: "corpora can never present direct evidence of ungrammaticality".

[5]The main difference with morphological constructions is that morphosyntactic constructions are all schematic to some degree, with the exception of fully substantive idioms.

[6]After Boas (2003).

[7]This idea finds empirical support in Hay (2001) and Hay and Baayen (2002, 2003). It also finds theoretical support in usage-based accounts of mental storage, frequency, and productivity.

[8]See Aronoff (1976, 76), Baayen and Lieber (1991, 817), and Zeldes (2012, Section 3.6) for discussion.

[9]E.g. *wegen* constructions in German (Zeldes 2012, Section 4.3)

[10]E.g. comparative correlatives (Zeldes 2012, Section 4.5).

[11]See also Stefan Evert's comprehensive inventory of existing association measures: http://www.collocations.de/AM/ (last accessed November 21, 2012).

[12]This is visible at the usage level: *good as gold* is used exclusively with a non-compositional, figurative meaning in the BNC.

[13]As pointed out by one anonymous reviewer.

[14]The log-transformation of G^2 is meant to stretch the plot vertically to make the top and bottom margins more legible. Without this log-transformation, all the data points whose G^2 score is below 50 would cluster at the bottom of the y axis all along the x axis.

[15]According to Gaeta and Ricca (2006, footnote 4), Baayen (1989, 117) disregards this possibility on the grounds that "what is being studied by comparing affixes for identical sample size is [...] pragmatic usefulness, a concept that, to our mind, is a component of the pretheoretical notion of morphological productivity".

[16]I used the package `FactoMineR` for R (Husson, Josse, Pagès, & Lê 2009)

[17]As opposed to active elements, supplementary elements, do not contribute to the construction of the principal components. The sole function of supplementary elements is to illustrate the principal components (Husson et al. 2011, 20).

[18]The near perfect correlation between $\mathscr{P}$ and $V1$ should come as no surprise since $\mathscr{P}$ is largely based on $V1$.

[19]*Clear as mud* is one of the very few examples in the data where *A as NP* means "not A", along with *as compatible as salt and strawberries* and *as compatible as oil and water*.

[20]In fact, we know from Table 2 that *gold* occurs only once with another adjective. In cases like this one, P^* should be used with caution.

[21]See Stefanowitsch and Gries (2003).

[22]I wish to thank one of the anonymous reviewers for pointing this out.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Aronoff, M. (1976). *Word formation in Generative Grammar*. Cambridge: Cambridge University Press.

Baayen, R. H. (1989). *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Amsterdam: Centrum Wiskunde en Informatica.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991* (pp. 109–149). Dordrecht & London: Kluwer.

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1992* (pp. 181–208). Dordrecht & London: Kluwer.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. an international handbook* (pp. 899–919). Berlin: Mouton de Gruyter.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *11*, 295–328.

Baayen, R. H. & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, *29*, 801–843.

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic*. Amsterdam: John Benjamins.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Bauer, L. (2001). *Morphological productivity*. Cambridge: Cambridge University Press.

Bergen, B. K. & Chang, N. (2005). Embodied Construction Grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammars: Cognitive grounding and theoretical extensions* (pp. 147–190). Amsterdam & Philadelphia: John Benjamins.

Boas, H. (2003). *A constructional approach to resultatives*. Stanford: CSLI Publications.

Burnard, L. (2000). Reference guide for the British National Corpus (World Edition). Web Page. Retrieved August 14, 2015, from http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Church, K., Gale, W. A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale: Lawrence Erlbaum.

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Croft, W. (2001). *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, W. & Clausner, T. C. (1997). Productivity and schematicity in metaphors. *Cognitive Science*, *21*(3), 247–282.

Croft, W. & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge; New York: Cambridge University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Ellis, N. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 187–220.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations* (PhD dissertation, Universität Stuttgart). Retrieved August 14, 2015, from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 1212–1248). Berlin & New York: Mouton de Gruyter.

Evert, S. & Baroni, M. (2006). *The* zipfR *library: Words and other rare events in R*. Presentation at useR! 2006: The Second R User Conference, Vienna, Austria.

Evert, S. & Baroni, M. (2007). *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th annual meeting of the association for computational linguistics on interactive posters and demonstration sessions* (pp. 29–32). (R package version 0.6-6 of 2012-04-03). Prague, Czech Republic.

Ferraresi, A. (2007). *Building a very large corpus of english obtained by web crawling: UkWaC* (Master's thesis, University of Bologna).

Fillmore, C. (1997). *Construction grammar lecture notes*. Retrieved August 14, 2015, from http://www.icsi.berkeley.edu/~kay/bcg/lec02.html

Fillmore, C. (2002). *"Idiomaticity"*. Retrieved August 14, 2015, from http://www1.icsi.berkeley.edu/~kay/bcg/lec02.html

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, *64*(3), 501–538.

Gaeta, L. & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, *44*(1), 57–89.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Cognitive theory of language and culture. Chicago: University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford & New York: Oxford University Press.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93–127.

Gries, S. T. (2007). Coll.analysis 3.2. a program for r for windows 2.x. Comp. software. Retrieved August 14, 2015, from http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, *18*(1), 137–166.

Gries, S. T. & Stefanowitsch, A. (2004a). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–236). Stanford: CSLI.

Gries, S. T. & Stefanowitsch, A. (2004b). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(4), 1041–1070.

Hay, J. & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 2001* (pp. 203–235). Dordrecht & London: Kluwer.

Hay, J. & Baayen, R. H. (2003). Phonotactics, parsing, and productivity. *Rivista di Linguistica*, *15*(1), 99–130.

Husson, F., Josse, J., Pagès, J., & Lê, S. (2009). FactoMineR, an R package dedicated for multivariate analysis. Retrieved August 14, 2015, from http://factominer.free.fr/index.html

Husson, F., Lê, S., & Pagès, J. (2011). *Exploratory multivariate analysis by example using R*. London: Chapman and Hall – CRC.

Jackendoff, R. (2008). Construction after construction and its theoretical challenges. *Language*, *84*(1), 8–28.

Kay, P. (2013). The limits of (Construction) Grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 32–48). Oxford: Oxford University Press.

Kay, P. & Fillmore, C. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language*, *75*, 1–33.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.
Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276.
Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*(1), 1–40.
Langacker, R. W. (1987). *Foundations of cognitive grammar.* Stanford: Stanford University Press.
Langacker, R. W. (2008). *Cognitive grammar: A basic introduction.* Oxford: Oxford University Press.
Langacker, R. W. (2009). *Investigations in cognitive grammar.* Berlin & New York: Mouton de Gruyter.
Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge: MIT Press.
Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex.* London: Oxford University Press: Humphrey Milford.
Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*(1), 137–158.
Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the south-central SAS users group conference* (pp. 188–200). Texas: SAS Users Group.
Plag, I. (2003). *Word-formation in English.* Cambridge: Cambridge University Press.
R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved August 14, 2015, from http://www.R-project.org/
Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*, 1–5.
Schmid, H.-J. & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, *24*(3), 531–577.
Steels, L. (2011). *Design patterns in Fluid Construction Grammar.* Amsterdam: John Benjamins.
Steels, L. (2012). *Computational issues in Fluid Construction Grammar.* Lecture Notes in Computer Science. Berlin: Springer.
Stefanowitsch, A. & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.
Stefanowitsch, A. & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, *1*(1), 1–46.
Wagner, A. R. & Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). New York: Appleton-Century-Crofts.
Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, *4*(2), 253–290.
Yates, F. (1984). Tests of significance for 2 x 2 contigency tables. *Journal of the Royal Statistical Society. Series A (General)*, *147*(3), 426–463.
Zeldes, A. (2012). *Productivity in argument selection: From morphology to syntax.* Berlin & New York: Mouton de Gruyter.
Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge: Addison-Wesley.