Stefan Th. Gries

The role of quantitative methods in cognitive linguistics

Corpus and experimental data on (relative) frequency and contingency of words and constructions

Abstract: One relatively frequently used corpus-based method in cognitive/usage-based linguistics is collexeme analysis, the study of the function of a construction on the basis of the words that are strongly attracted to particular slots of that construction. This approach has recently been discussed critically particularly in Schmid and Küchenhoff (2013). This paper argues that many of the points of critique raised in that paper are invalid and that its exploratory analysis – monofactorial rank correlations – are far from sufficient to identify and tease apart (i) the many interrelated ways in which the association of a word and a construction can be measures and (ii) how these operationalizations are correlated with experimental data. Two more appropriate statistical approaches – mixed-effects modeling with model selection and multimodel inferencing – are applied to the data to showcase not only what kinds of issues analysts face in the study of association measures, but also how these methods can contribute to more sophisticated analyses.

1 Introduction

1.1 General introduction

One of the greatest paradigmatic changes in theoretical linguistics over the past few decades has been the joint way in which (i) cognitive, or usage/exemplar-based, linguistics has developed into a full-fledged attractive theoretical approach to language competing with generative linguistics and (ii) how this development brought about, and was in turn reinforced by a similarly profound change in linguistic methodology, namely the more and more widespread adoption of quantitative methods in theoretical linguistics. Dirk's work and impact

Stefan Th. Gries: University of California, Santa Barbara

on both the theoretical and the methodological side of this field has been profound and it is with great honor that I accepted an invitation to participate in this volume celebrating Dirk's 60th birthday; the present paper will hopefully be a good way to congratulate him by discussing and involving things central to Dirk's research – usage-based linguistics (specifically, the association of verbs and constructions), the use of non-introspective methods (specifically, corpus and experimental data), and, I hope, careful statistical analysis using both methods that bring together hypothesis-testing and exploratory work.

More precisely, the focus of the present paper is a re-analysis of previous experimental results on one method of the family of collostructional analysis, viz. collexeme analysis (CA). CA is essentially a very basic application of the corpus-linguistic notion of (lexical) association measures as applied to the co-occurrence of words to the co-occurrence of words (often verbs) and constructions (often sentence-level/argument structure constructions). As outlined in the first publication on CA, Stefanowitsch and Gries (2003), CA is typically done as follows:

i. retrieve all instances of a construction *cx* in question (such as the ditransitive) from a corpus;
ii. compute an association measure (AM) for every word type *v* that occurs in the relevant slot of construction *cx* (these are referred to as *collexemes*) (such as *give*, *send*, *tell*, ...). Such AMs are usually computed on the basis of a 2×2 co-occurrence table that cross-tabulates token (non-)occurrences of *cx* against every single co-occurring element/type *v* as schematically represented in Table 1; thus, for instance, *a* is the number of times v_1 occurs in *cx*, etc.

Tab. 1: Schematic frequency table of verb v_1 and *cx* and their co-occurrence

	***cx* is present**	***cx* is absent**	**Totals**
v_1 is present	a	b	a+b
v_1 is absent	c	d	c+d
Totals	a+c	b+d	a+b+c+d=N

iii. rank all types $v_{1\text{-}n}$ by the value of the AM;
iv. explore the top *n* (often 10-50) co-occurring types for functional patterns.

Crucially and as stated by Stefanowitsch and Gries (2003: 217) or Gries (2012: 480), pretty much any AM can be used to compute what has been called *collexeme strength*, but most published studies have chosen the negative log of the *p*-value of the Fisher-Yates exact test (for collexemes that occur more often than expected with the construction), henceforth *FYE*; this is because (based on Stefanowitsch and Gries 2003: 239, n. 6):

– since FYE involves an exact test, it does not come with distributional assumptions and can handle small frequencies well (see also Evert 2009);
– since it is based on a significance test, its results incorporate both observed frequencies and effect size.

CA has recently been discussed critically in two publications, Bybee (2010) and Schmid and Küchenhoff (2013, henceforth S&K). The shortcomings of Bybee (2010) were addressed comprehensively in Gries (2012) and will not be repeated here; the many problems of Schmid and Küchenhoff (2013) are discussed in Gries (to appear) and will be recapitulated here only to the extent that is required for the present analysis.

First and as Bybee before them, S&K criticize the choice of FYE as an AM because it is not a significance measure and not intuitively easy to understand. Second, they problematize the computation of AMs based on tables like Table 1 as discussed above on the assumption that defining the frequency to insert in cell *d* – i.e. the frequency of constructions that are not *cx* and that do not involve *v* – is difficult/treacherous. Third, they argue that FYE is a bidirectional AM, i.e. an AM that cannot distinguish the attraction of *v* to *cx* from the attraction of *cx* to *v*, which they claim is desirable. Finally, they criticize a study attempting to validate CA – Gries, Hampe, and Schönefeld (2005), henceforth GHS – for how in that study the effects of frequency and FYE were compared. In that study, GHS used corpus data to determine verbs of high/low frequency and high/low collexeme strength in the *as*-predicative construction (e.g. *She is regarded as an astute observer or He sees himself as a total fraud*). Then, they asked subjects to complete sentence fragments ending in such verbs, and S&K criticize the way in which the numeric variables of frequency and FYE were dichotomized in GHS's study of how much particular verbs lead to *as*-predicative completions.

Many of their points of critique are problematic on several levels, however (see Gries to appear for comprehensive discussion). Their first criticism misses the point because it is like criticizing the whole paradigm of reaction time studies in psycholinguistics because they often use linear models for the statistical

analysis of their data – even if the choice of FYE were as problematic as they claim, which has been shown it is not (see Gries 2012, to appear), that does not invalidate the idea of exploring constructions on the basis of which words are attracted to their (main) slots. In addition, their argumentation ignores the fact that FYE is merely used to then rank all collexemes in step iii. and ranks of types are certainly intuitively straightforward. Also, they ignore the fact that, unlike an effect size, FYE can distinguish identical effects that result from high- or low-frequency co-occurrence.

Their second criticism ignores Gries (2012), which has shown on the basis of statistical simulations that the impact of different frequencies of *d* (and thus, different corpus sizes) on the overall ranking of word/verb types (recall step iii. from above) is negligible.

Their third point of critique, the bidirectionality of FYE, is a more useful observation and leads to two related suggestions of theirs: First, they discuss directional alternatives to FYE, namely two conditional probabilities: $p(v|cx)$ (i.e., $^{a}/_{a+c}$), which they call *attraction*, and $p(cx|v)$ (i.e., $^{a}/_{a+b}$), which they call *reliance*. Somewhat confusingly, they also discuss another alternative to FYE, namely the odds ratio (i.e., $^{a/b}/_{c/d}$). This is confusing because (i) the odds ratio requires filling cell *d* in the cross-tabulation (just like FYE), (ii) is bidirectional (like FYE), and (iii) contributes very little that is not already covered by their proposed measure reliance: In both the *as*-predicative data to be discussed below as well as their own N-*that* construction data, the Spearman rank correlations between the odds ratio and reliance exceed >0.99! The only major theoretical difference between FYE and the odds ratio is that the latter is an effect size, which a priori is neither good nor bad. A final issue to be mentioned here is that they do not discuss in this regard is that attraction and reliance per se do not reveal whether a word is attracted to a construction or repelled by it – for that, the measures $\Delta P_{\text{construction} \rightarrow \text{word}} = (^{a}/_{a+c})\text{-}(^{b}/_{b+d})$ and $\Delta P_{\text{word} \rightarrow \text{construction}} = (^{a}/_{a+b})\text{-}(^{c}/_{c+d})$ (cf. Ellis 2007; Gries 2013), which have been outputted by the *R* script most people have been using to do CAs, are more useful (because they reflect attraction/repulsion with positive/negative signs).

As for the final point of critique, S&K are right in pointing out that the dichotomization GHS employed is sub-optimal: While the cut-off points to dichotomize frequency and FYE into *low* and *high* were chosen in a bottom-up fashion, they lose too much information compared to what now, 10 years later, is more profitably explored using numeric predictors in a more appropriate statistical analysis. That being freely admitted, unfortunately, the kind of analysis that S&K then report on themselves is even more problematic: At a time where the state-of-the-art in cognitive/usage-based linguistics has evolved to multivariate

exploratory methods and/or multifactorial regression (see Levshina, Geeraerts, and Speelman 2013 for an excellent recent example, whose combination of exploration and hypothesis-testing is mirrored in this article), they merely report a variety of monofactorial Spearman rank-order correlations between all the different measures for the corpus-based and experimental data of GHS as well as their N-*that* construction data[1], which is problematic given that we know that this, like every other linguistic phenomenon, is not mono- but multifactorial. Nonetheless, it is revealing to see that even in their own re-analysis of the by-verb data of GHS, it is the supposedly inferior measure of FYE that is most strongly correlated with GHS's experimental results.

1.2 The present paper

In this paper, I will explore ways in which CA in general and the potential confluence of GHS's CA results and their experimental completion results in particular can be explored in more detail. While both space and the size of the data set do not allow for a fully comprehensive re-analysis of the data, the focus of this brief exploration here is to showcase what the current state-of-the-art in cognitive/usage-based linguistics might allow one to begin to do to shed more light on the doubtlessly complex interplay between corpus-based frequency and contingency and speakers' experimental reactions. Two approaches will be presented. Section 2 discusses a regression-based approach to GHS's data, in which (i) an exploratory principal components analysis (PCA) will be applied to a variety of different AMs for the verbs in GHS's experiment (to address the problem that AMs are highly correlated with each other), followed by (ii) a generalized linear multilevel modeling (GLMM) approach that controls for subject-specific variation as well as verb-specific variation as well as experimental-stimulus variation nested into the verbs.

Section 3 takes a slightly different approach with regard to the policy of model selection and how collinearity might be addressed: On the basis of the above-mentioned principal components, I use multimodel inferencing (cf. Burnham and Anderson 2002; Kuperman and Bresnan 2012), a regression approach that generates a variety of models and weighs their regressions' coefficients proportionally to the degree to which each model deviates from the best model's performance/*AIC*.

---

1 Admittedly, they presumably did not have access to the whole set of experimental data of GHS, however, they also didn't conduct an experiment for their own data.

The input data analyzed in both ways are 512 sentence completions by native speakers of English to sentence fragments ending in verbs that are differently frequent and differently strongly attracted to the *as*-predicative. For each of the completions, the following data are available, which were used in the following two sections:

- SUBJECT: the speaker who provided the sentence completion;
- ASPRED: a binary variable coding whether the subject used an *as*-predicative, *no* versus *yes*;
- VERB: the verb used in the experimental stimulus;
- ITEM: the experimental stimulus;
- VOICE: the voice of the stimulus: *active* versus *passive*;
- COLLSTR: FYE as defined above;
- ATTRACTION: the attraction value of the verb in the stimulus as defined by S&K;
- DPC2W: $\Delta P_{\text{construction} \rightarrow \text{word}}$, i.e. essentially a normalized ATTRACTION value as defined above;
- KLDATTRACTION: the Kullback-Leibler divergence of how the distribution of the verb in and outside of the construction differs from the distribution of everything else in and outside of the construction (cf. Baayen 2011 for discussion);
- RELIANCE: the reliance value of the verb in the stimulus as defined by S&K;
- DPW2C: $\Delta P_{\text{word} \rightarrow \text{construction}}$, i.e. essentially a normalized RELIANCE value as defined above;
- KLDRELIANCE: the Kullback-Leibler divergence of how the distribution of the construction with and without the verb differs from the distribution of everything else with and without the verb (cf. again Baayen 2011);
- ORLOG: the odds ratio as computed above and logged to the base of 2.

As mentioned above, the analyses below can only be first steps towards future research, but they do indicate the complexities usage-based linguistics will need to deal with if it wants to stay true to its promise of taking usage and its effect one representation, processing, and use seriously.

2 Approach 1: PCA and GLMM

2.1 The principal components analysis

In a first step, the data were subjected to two PCAs. One was done on the four columns containing AMs that are related to $p(v|cx)$, i.e. CollStr, Attraction, DPc2w, and KLDAttraction, the other on the columns with AMs that are related to $p(cx|v)$, i.e. ORlog, Reliance, DPw2c, and KLDReliance. Both PCAs indicated that the four variables were extremely highly correlated and that each could be well summarized by their first principal component. In the case of the $p(v|cx)$, that first principal component accounted for 94% of the variance of the four variables; in the case of the $p(cx|v)$, the first principal component accounted for 96.3% of the variance of the four variables. In each case, I then computed principal component scores that summarized the original four predictors that had been entered into the analysis. These were very highly correlated with the four predictors that they reflected and little with the other four; the two principal components, PCcx|v and PCv|cx, were still somewhat, but just about not significantly correlated with each other: $r_{\text{over verb types}}=0.357$, $p>0.06$. These result show that, on the whole, the used AMs capture two dimensions of the association of words to constructions and the other way round – a bidirectional exploration of association is therefore useful, see Gries (2013) and of course S&K – but also that these two dimensions are still related to each other – in other words, there may well be yet one "deeper" dimension that underlies even these two principal components. (In fact, a follow-up exploratory PCA on just PCcx|v and Pv|cx suggests just that because it returns one such "deeper" component, which accounts for more than 69% of the variance of these two, indicating that the last word on how many dimensions AMs need to cover has not yet been spoken[2].) These factor scores were then added to the original data set and used in the regression-based approach discussed in the next section.

2 Also, Gries (to appear) performed one PCA on all eight variables and found that the first principal component of that analysis accounts for 66% of the variance of all eight measures. This strategy is not pursued here because that component is uninterpretable: all eight AMs load highly on it.

2.2 The generalized linear multilevel model

In order to determine how the two factors co-determine the subjects' fragment completions, a series of generalized linear multilevel models was fit. The maximal model involved ASPRED as a dependent variable, the fixed effects of Voice as well as the two principal components PCCX|V and PCV|CX and all their interactions. In addition, I added varying intercepts for each experimental subjects (1|SUBJECT) as well as varying intercepts for all experimental stimuli, which in turn were nested into varying intercepts for all verbs (1|VERB/ITEM). In a first series of tests, it became obvious that the varying intercepts for the subjects were not required and thus omitted whereas the varying intercepts for verbs and stimuli were required; after the PCA, neither collinearity nor overdispersion were a problem. A subsequent model selection process resulted in the deletion of several interaction as well as the main effect of VOICE: The minimal adequate model contained only the two principal components and their significant interactions, as shown in the results in Table 2.

Tab. 2: Results of the GLMM on ASPRED

	coef	*se*	*z*	*p*
Intercept	−1.3583	0.3398	−3.997	<0.001
Pv|cx	−0.5628	0.1915	−2.939	0.0033
Pcx|v	−0.6376	0.1812	−3.518	<0.001
Pcv|cx : Pv|cx	−0.1734	0.0575	−3.017	0.0026

This model was significantly better than an intercept-only model (Chi-squared =20.228, df=3, p<0.001) and came with moderately high correlations: $R^2_{marginal}$=0.3, $R^2_{conditional}$=0.6 (computed as suggested by Nakagawa and Schielzeth 2013); also, the classification accuracies with and without random effects were 83.2% (C=0.91) and 75.8% (C=0.77) respectively; both these results point to the fact that the subjects' completions were affected to quite some degree by the specific verbs used. The main result, the effect of the interaction of the two principal components on the predicted probability of *as*-predicatives by the subjects (fixed effects only) is represented in Figure 1: PCV|CX is shown on the x-axis, PCCX|V is shown on the y-axis, and the plotted numbers represent the predicted probability of an *as*-predicative (higher/larger numbers meaning higher probabilities); in addition, all experimental verbs are plotted at their PCA-scores coordinates in a size reflecting their proportion of *as*-predicatives.

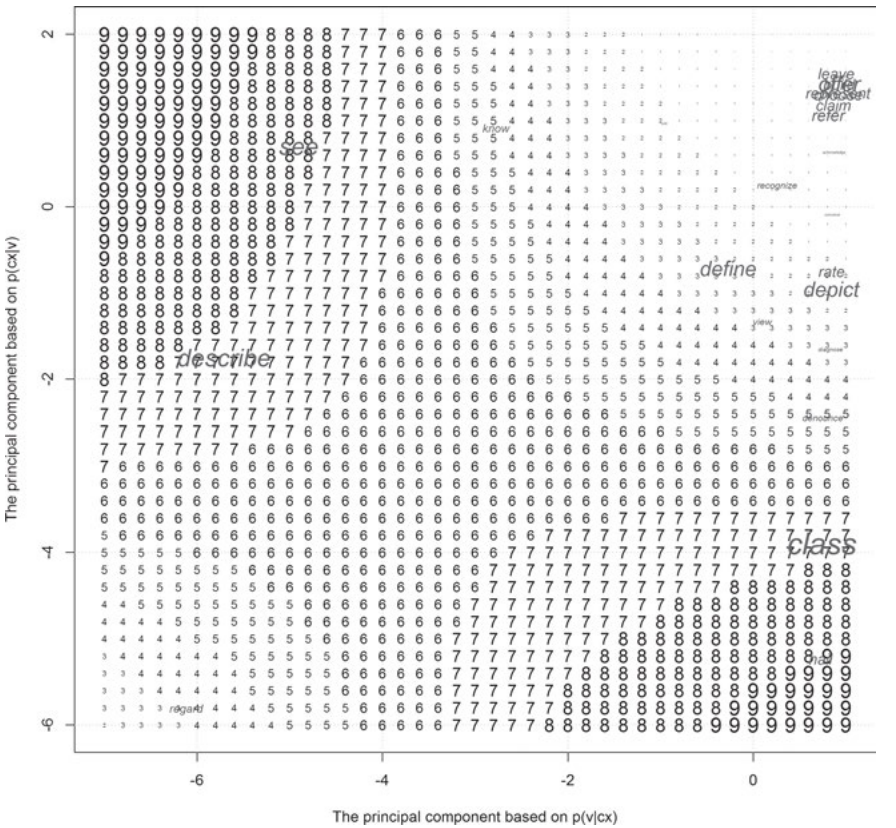

Fig. 1: The interaction of PCcx|v : Pv|cx in the minimal adequate model

What does the visualization reflect with regard to the roles of the two perspectives on association? Before we can begin to answer that question, two things need to be pointed out. First, the graph needs to be interpreted with some caution since the two principal components are from different PCAs so they are *not* orthogonal, even if that is what the traditional 90º angle between the *x*- and the *y*-axis suggests! Second, the orientation of the two axes is what might seem counterintuitive, because, on both the *x*- and the *y*-axis, highly negative values mean that the verb "likes to occur" in the construction or that the construction "likes to occur" with the verb, and values around 0 or positive values reflect an absence of such a preference; this is why *regard* is located in the lower left corner of the plot: *regard* occurs with the *as*-predicative so frequently that both perspectives reflect that fact.

With these things in mind, the interaction indicates that each principal component seems to have an *as*-predicative boosting effect when the other component is at its weakest:

- in the top left corner, *as*-predicatives are strongly predicted to be produced, which is where PV|CX has its strong effect and PCX|V has its weaker effect; this is characteristic of verbs like *see* and *describe*, which, e.g., have high COLLSTR values and low $\Delta P_{\text{construction} \rightarrow \text{word}}$ values but lead to *as*-predicative completions >80% of the time (compared to an overall baseline of 29.3%);
- in the bottom right corner, *as*-predicatives are also strongly predicted to be produced, which is where PV|CX has its weaker effect and PCX|V has its stronger effect; this is characteristic of verbs like *class* and *hail*, which, e.g., have low COLLSTR values and high $\Delta P_{\text{construction} \rightarrow \text{word}}$ values (both >0.59) but lead to *as*-predicative completions 100% and 50% of the time respectively;
- in the bottom left corner, where both principal components would lead one to expect very high numbers of *as*-predicatives, we only find the verb *regard*, which is an interesting case: Its overall proportion of *as*-predicatives (37.5%) is only slightly above average, but that is largely due to the fact that 75% of the responses to the active experimental item were not *as*-predicatives. Thus, while that experimental item's effect on the overall regression results is probably not too damaging (because it was "dealt with" by the multilevel structure of the model), this individual verb's result are a bit unexpected;
- in the top right corner, we see many different verbs that do not have high scores on either principal component and thus do not lead to *as*-predicative completions much, and in fact the average proportion of *as*-predicatives for all verbs with positive principal component scores is 14.2%, i.e. not even half the overall baseline.

In sum, the two principal components capture what are two somewhat different but nonetheless related distributional dimensions. Probably in part because of unexpected results for the verb *regard*, however, the interaction of these two dimensions reveals that each of these dimensions is strongest in co-determining sentence completions when the other dimensions does not have a strong effect itself.

3 Approach 2: Multimodel inferencing (MuMIn)

One of the trickiest aspect of the current debate surrounding AMS for both lexical collocation and word-construction associations (colligation/collostruction) is that the many different measures that have been proposed (see Pecina 2009 for an overview of >80 measures) are so highly correlated that a simple regression-based approach will run into huge problems of collinearity, i.e. the fact that sizes and even signs of regression coefficients – the very measures intended to reflect the importance of predictors – will vary erratically. The above approach was a traditional method to deal with collinearity: use a PCA to capture what is shared among collinear predictors and proceed with a regression-based approach on the basis of the PCA scores. In this section, I am using a different, more recent approach: multimodel inferencing. This approach begins by fitting one maximal model (maximal in terms of both its fixed- and random-effects structure), of which then all possible sub-models are fit, i.e. all subsets of the predictors of the maximal model. For each of these models, coefficients and *AICc*-values are computed and stored. Once that process is done, the best model (in terms of *AICc*) identified and the coefficient values of all regressions are averaged such that each model's contribution to these averages are weighted by how much the model deviates from the best model. Because of this averaging of the standard errors, collinearity is less of an issue than it would be if only one regression was run on the raw data[3].

As mentioned above, this particular application is based on the same two principal components the previous section, Pv|cx and Pcx|v. The first/maximal model that was fit had ASpred as the binary dependent variable and involved the two principal components and Voice as well as all their interactions as fixed effects and, as before, (1|Subject) and (1|Verb/Item) as random effects; in addition, all submodels of this maximal model were fit with an eye to determine (i) which model provides the best fit for the data (measured in terms of *AICc*) and (ii) which predictors are most important in predicting the sentence completions.

In this particular case, the results are very compatible with those of the model selection procedure in the previous section. The best model contains an intercept, the two principal components, and their interaction (*AICc*=453.8).

3 The degree to which multimodel inferencing helps is determined in part by the amount of collinearity in the data. In this particular case, the above-mentioned correlation between the two principal components is of a size that multimodel inferencing is supposed to be able to handle well (see Freckleton 2011).

More specifically even, of all 19 possible submodels, only five have *AICc*-values less than 4 higher than the optimal model and all these models contain these three predictors[4]. For the shrinkage-corrected coefficients and variable importance measures of all predictors in these five models, see Table 3.

Tab. 3: Results of the MuMIn approach on ASPRED (full model-averaged coefficients)

Predictors	*coef*	*se*	adj. *se*	*z*	*p*	importance
PCX|V	−2.8	0.81	0.81	3.45	<0.001	1
PV|CX	−2.43	0.85	0.85	2.87	0.004	1
PCX|V : PV|CX	−2.55	0.85	0.85	3	0.003	1
$\text{VOICE}_{\text{active} \rightarrow \text{passive}}$	−0.2	0.37	0.37	0.54	0.59	0.49
PV|CX : $\text{VOICE}_{\text{active} \rightarrow \text{passive}}$	−0.03	0.19	0.19	0.14	0.89	0.11
PCX|V : $\text{VOICE}_{\text{active} \rightarrow \text{passive}}$	0.001	0.18	0.18	0.01	1	0.1

While these overall results are very similar to the ones from Section 2 above, they are nonetheless important to arrive at: First, the MuMIn regression results are less likely to be affected by all the risks of model selection processes (most importantly, a high confirmation bias) and are more robust (since they are obtained from multiple different statistical models). Second, the fact that multiple models are studied makes it possible to compute an overall variable importance score ranging from 0 to 1 to determine how important each predictor is. In this case, the two principal components and their interactions all score the maximal value; if this computation is done on the basis of all 19 models regardless of their quality, then the value for PCX|V remains at 1, and the values for PV|CX and PCX|V : PV|CX change minimally to 0.97 and 0.93 respectively.

In sum, the results of the MuMIn approach are conceptually very similar to those of the model selection procedure and point again to the fact that both perspectives on AMs have something to offer although future work is needed to determine to what information exactly it is that the two separately derived principal components share (see the above-mentioned correlation between the two).

4 The value of 4 is a difference threshold mentioned by Burnham & Anderson (2002: 70) and indicates that a model that has an *AIC*-difference of >4 is "considerably less" likely to be the best model.

4 Concluding remarks

Given the size of both the currently available experimental data on word-construction associations as well as limitations of space, this paper cannot be, but only hope to stimulate, a full-fledged discussion on what different association measures exactly reflect/operationalize and how that is related to subjects' behavior in different experimental tasks. More specifically, I hope to have shown two kinds of things: First, with regard to recent critiques of CA, I hope to have shown that

- the critique of CA by S&K is problematic in a variety of theoretical aspects, some of which were mentioned above and more of which are discussed in Gries (to appear);
- the suggestion made by S&K to take the potential bidirectionality of association into consideration is potentially useful (both principal components return significant results but are correlated with each other) and compatible with existing claims in that regard for lexical and colligational/collostructional co-occurrence (Ellis 2007; Gries 2013);
- the way in which S&K study word-construction associations is not useful: instead of recognizing the complex multifactoriality of the phenomenon in question, their exploration is restricted to mere monofactorial rank correlations, which actually return FYE as the strongest predictor.

Second, I hope to have given a first impression of the actual complexity of the phenomenon and how the current methodological state-of-the-art in cognitive/usage-based linguistics can begin to address it. Specifically,

- instead of monofactorial correlations, we need to use more advanced regression-based methods that can handle the *multivariate nature* of the issue while at the same time avoiding, or at least checking to, potential pitfalls of model selection procedures;
- at the same time, we need to be able to address in some way the obvious fact that AMs from both the $P_{V|CX}$ and $P_{CX|V}$ perspectives exhibit *intercorrelations* with each other;
- we need to be able to handle the ways in which corpus and experimental data violate the *independence-of-datapoints* assumptions. Much existing work uses mixed-effects modeling to handle crossed random effects such as speakers and lexical items, but we also need to take nested random effects into consideration as when verbs are tested with multiple different experimental stimuli or when multiple data points come from the same file and thus sub-register and thus register (see Gries 2015);

– we need to be able to add *more predictors* into the mix. For instance, Gries (2012, to appear) discusses the role that verbs' constructional entropies may play. In order to explore this possibility, I used the data of Roland, Dick, and Elman (2007) to compute for each verb used in the sentence-completion experiment the difference between the entropy of all construction frequencies with and without the transitive+PP uses (like the *as*-predicative), which (i) in a GLMM turned out to interact marginally significantly ($p<0.1$) with each principal component and (ii) in a MuMIn scored an importance value of 0.72 even in the tiny data set that is left once all verbs not attested in Roland, Dick, and Elman (2007) are left out.

Again, while I cannot provide hard-and-fast solutions here, I hope it has become obvious what to consider in future research and how – given the complexities involved, methodological simplification is certainly not the answer, which I am certain is a statement that Dirk would subscribe to whole-heartedly. Congratulations, Dirk, and many happy returns!

References

Baayen, R. Harald. 2011. Corpus linguistics and naïve discriminative learning. *Brazilian Journal of Applied Linguistics* 11(2). 295–328.

Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. New York: Springer.

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

Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 29/2), 1212–1248. Berlin & New York: Mouton De Gruyter.

Freckleton, Robert P. 2011. Dealing with collinearity in behavioural and ecological data: Model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology* 65(1). 91–101.

Gries, Stefan Th. 2007. Coll.analysis 3.2a. A script for *R* to compute perform collostructional analyses. http://tinyurl.com/collostructions (accessed 3 April 2015).

Gries, Stefan Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language* 36(3). 477–510.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics* 18(1). 137–165.

Gries, Stefan Th. 2014. Coll.analysis 3.5. A script for *R* to compute perform collostructional analyses (major update to handle larger corpora/frequencies). http://tinyurl.com/collostructions (accessed 3 April 2015).

Gries, Stefan Th. 2015. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.

Gries, Stefan Th. To appear. More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff. *Cognitive Linguistics* 26(3).

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

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: CSLI.

Kuperman, Victor & Joan Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66(4). 588–611.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013. Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics* 51(4). 825–854.

Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142.

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1–2). 137–158.

Roland, Douglas, Frederick Dick & Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: a corpus analysis. *Journal of Memory and Language* 57(3). 348–379.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.

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