Unifying everything: Integrating quantitative effects into formal models of grammar

Matías Guzmán Naranjo
Universität Leipzig

I. Introduction

Quantitative effects can be divided into two general groups: syntagmatic effects like collocations and grammaticalization, and paradigmatic effects like family size, entropy, alternations and collostructions. By now we have more than ample evidence that quantitative effects play a role in language processing, acquisition and language change. It is also clear that core grammar is, at least for the most part, independent from quantitative effects, eg. the frequency of a construction says nothing about its grammatical properties. What is still sorely lacking is an interface between grammar and usage that allows the usage module to access grammar in a systematic way.

Meanwhile, a problem for non-formal approaches to quantitative effects (like most construction grammar based corpus studies) is that they have no way of imposing constraints on which quantitative effects are possible, and which aspects of the grammar they can refer to. There is no known way to disallow extremely non-local and unmotivated associations to take place. With the mainstream approach it is not possible to distinguish explicitly between nonsensical correlations and real correlations in a corpus. There have been some attempts to address this problem, like collosctructional analysis, but these are always done informally, and there are no clear constraints on what is and what is not possible.

I present a model that solves both problems. On the one hand it provides an interface for grammar and quantitative effects to interact, and on the other hand it formally constraints how quantitative effects can take place. This model is based on Sign-Based Construction Grammar [7], [8], [9], but it should also work for related frameworks like HPSG, and possibly Fluid Construction Grammar.

II. The solution

The model I propose makes use of two key aspects of the SBCG architecture: (1) the ARG-ST feature, and (2) the type hierarchy in the signature. The ARG-ST feature constraints syntagmatic qunatitative effects, and the type hierarchy provides almost all the requirements for deriving paradigmatic quantitative effects.

The key insight is that most of the relevant, syntactically motivated syntagmatic effects can be derived if we allow verbs and adjectives to list their preference for complements and selected heads (denoted as a set of sign-weight pairs in curly brackets). We can achieve this in the form of a set of weights of the relative attraction strength for each possible complement or head. Since the ARG-ST feature specifies types, and the type hierarchy is complete, the ARG-ST feature has access to each and every possible maximal feature structure it can license as a complement or head. The weights can be calculated using most measurements of association strength (mutual information, exact Fisher T-test, bayesian contingency tables, Δ p, etc.).

A. Collocations

Collocations are the easiest to handle in this model, so we start with them. A simple example would be that of the possible collocates of *brush* like *teeth* and *hair*. Conceptually, we do not want to claim that the attraction is just between the word *brush* and the word *teeth*, because we would like to include phrases like *brush your teeth* or *brushed my teeth*. This means that the attraction needs to be encoded at the lexeme level previous to the application of the inflectional rule. Additionally, the determiner should be (partially) invisible to the attraction, since in this case we would like to treat all different options (*my, your, their* etc.) equally. In traditional collocational analysis the issue of inflection is usually ignored (though it could be overcome by using the lemma instead), and the issue of linear distance is "solved" by using spans of more than one word. The problem with this solution of longer spans from the node is that there is no way of determining how long a span should or should not be. Additionally, some spans have the problem that they can include many words that make no theoretical sense. The solution to this problem is simple: we only consider as collocates other words selected for by the node. For the particulate collocates mentioned above a possible analysis is given in (1)[1] as follows:

(1)
$$\begin{bmatrix} \textit{str-v-lxm} \\ \text{SYN} \begin{bmatrix} \text{CAT} \begin{bmatrix} \textit{verb} \\ \text{LID} \left[\text{l-brush}\right] \end{bmatrix} \end{bmatrix} \\ \text{ARG-ST} \begin{bmatrix} \text{NP}, \\ \left\langle \text{NP} \begin{Bmatrix} \dots \\ \text{NP[L-ID } \textit{l-teeth}\text{]}, \ \Delta_p a \\ \text{NP[L-ID } \textit{l-hair}\text{]}, \ \Delta_p b \\ \dots \end{Bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

where $\Delta_p a$ and $\Delta_p b$ are the respective attraction strengths for each *teeth* and *hair*, calculated with any desired measure

[1]The LID feature is a unique identifier of every lexical entry in the lexicon.

(here Δ_p makes reference to the directional measure proposed by Gries [4] but almost anything else would also be possible). More fixed cases of collocations can also be accounted for in a similar fashion, but allowing less freedom to the dependents.

Just from this simple approach to collocations two clear advantages emerge. The first one is that we do not need to explain unreasonably distant collocates, at, say, span 20 from a given node. This model only allows for locally selected collocates to be considered as having a real effect. Nevertheless, non-locally realized complements (displaced or extracted) will be counted and accounted for. The second advantage is that it also allows us to dismiss linearly close but syntactically superfluous collocates, like attraction between *and* and *the* as a side effect of the actual [*and* NP] phrase.

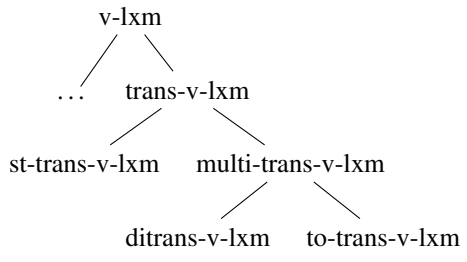
B. Collostructions

Collostructions were initially treated as essentially the same phenomenon as collocations [10], but from this perspective we have to make a distinction between two kinds of collostructions. On the one hand we have fixed pattern constructions as the classical *X is waiting to happen*, which can be modeled in exactly the same way as a collocation because X is selected for lexically (lexical treatments of similar patterns are given in Sag [9]). On the other hand, more abstract constructions like argument structure constructions require a different treatment. Because SBCG does not have argument structure constructions, we need a lexical solution [6].

The example given by Stefanowitsch and Gries [10] of the dative argument structure construction can illuminate this point. In their analysis Stefanowitsch and Gries[10] measure what they claim is the attraction between different verbs and the [NP V NP NP] pattern. However, in this case, since argument structure patterns are lexical properties of the verb encoded in the ARG-ST feature, we cannot claim that these patterns attract verbs in the same way that verbs attract complements. I will claim that SBCG provides a better structure to model these effects.

The type hierarchy of SBCG has the right form to encode all quantitative effects associated with abstract collostructions. In SBCG the type hierarchy captures generalizations common to classes of signs that share some feature or set of features. A simplified example of the type hierarchy for verbs adapted and slightly modified from [9] is given in (2)[2].

(2)



[2]v-lxm=verb lexeme, trans-v-lxm=transitive verb lexeme, st-trans-v-lxm=strictly transitive verb lexeme, multi-trans-v-lxm=multi-transitive verb lexeme, ditrans-v-lxm=ditransitive verb lexeme, to-trans-v-lxm=*to*-transitive verb lexeme.

The relevant sign for the verb *give* is given in (3).

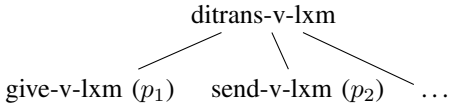
(3)
$$\begin{bmatrix} \textit{multi-trans-v-lxm} \\ \text{SYN|CAT|LID} \left\langle \textit{l-give} \right\rangle \\ \text{SEM} \left[\text{FRAMES} \left\langle \left[\textit{giving-fr} \right] \right\rangle \right] \end{bmatrix}$$

And the sign for all ditransitive verbs is given in (4).

(4) *ditrans-v-lxm* $\Rightarrow$
$$\begin{bmatrix} \text{ARG-ST} \left\langle \text{NP}_x, \text{NP}_z, \text{NP}_y \right\rangle \\ \text{SEM} \left[\text{FRAMES} \left\langle \begin{bmatrix} \textit{giving-fr} & \\ \text{AGENT} & \text{x} \\ \text{THEME} & \text{y} \\ \text{RECIPIENT} & \text{z} \end{bmatrix} \right\rangle \right] \end{bmatrix}$$

But since the type hierarchy is fully specified, there is a complete mapping from the *ditrans-v-lxm* type to each ditransitive verb:

(5)



Which means that the structure works as a directed graph, and it can specify weights or probabilities for each verb (expressed in parenthesis in (5)). These weighs or probabilities would then constitute what in collostructional analysis is seen as the attraction between the argument structure construction and the verb. Notice that any quantitative effect is only defined with respect to a given corpus, and corpora are finite, which means we do not have infinitely many verbs, and thus no infinitesimal probabilities. This is however only a practical issue, and we could assign "left-over" probabilities to unseen verbs.

The difference in treatment is empirically justified. There is a very fundamental difference in the way valence patterns interact with verbs and the way lexical items interact with other lexical items or even larger signs. Sign collocations are ope ended and grammatically driven (except for fixed expressions). There is, at least in principle, no end to the NPs that can be selected for by *brush*. On the other hand, verbs are restricted to relatively few valence patterns (be it through ambiguity or lexical rules for valance augmentation, etc. [5], [6], see the case study), while any valance pattern can, in principle, have and unbounded number of different members. This difference is clearly captured by the present proposal, but missed by the classical approach to collostructions.

Interestingly, the related question of collocates between a valence pattern and the possible valents would receive a more classical treatment. If we want to investigate, for example, the kinds of subjects of ditransitive verbs, we could measure the attraction of all corpus instances of ditransitive verbs to their subjects. This would give us the attraction of the sign in (4) to its possible subjects. The same principle would apply to any other abstract grammatical pattern we may want to investigate.

C. Alternations

Alternation modelling is in the current proposal a consequence of the two previous effects. The classical alternation example is the dative alternation between ditransitive (*I gave Martha the book*) and *to*-ditransitive (*I gave the book to Martha*) patterns. To model this alternation we need the additional sign in (6) which contains the template for *to*-ditransitive verbs.

(6) *to-ditrans-v-lxm* $\Rightarrow$

$$\left[\begin{array}{l}\text{ARG-ST}\left\langle \text{NP}_x,\ \text{NP}_y,\ \text{PP[to]}_s \right\rangle \\ \text{SEM}\left[\text{FRAMES}\left\langle \begin{bmatrix} \textit{giving-fr} & \\ \text{AGENT} & \text{x} \\ \text{THEME} & \text{y} \\ \text{PATH} & \text{s} \end{bmatrix} \right\rangle\right]\end{array}\right]$$

But because we have added this additional sign as a sister type of the *ditrans-v-lxm* and sub type of the *multi-trans-v-lxm*, we end up with verbs that are ambiguous between *ditrans-v-lxm* and *to-ditrans-v-lxm*, that is, verbs in the dative alternation. But now, because verbs like *give* can be of either type (notice that (3) is of type *multi-trans-v-lxm*, which means it can be instantiated by either of the sub-types of this type), their full probability is shared between both types. The fact the probabilites of v_1 to v_n in *ditrans-v-lxm* and *to-ditrans-v-lxm* do not add up to 1 in each one produces the lexical effects of some verbs preferring one construction over the other.

Other types of factors can be handle as collocational. The factors found to be relevant for predicting the dative alternation in the data set of Bresnan et al. [2] were of four kinds. The first were verb related factors: verb and semantic class of the verb. The second one were those related to the recipient: Length of the recipient, animacy of the recipient, definiteness of the recipient, accessability of the recipient and pronominality of the recipient. Length of the theme, animacy of the theme, definiteness of the theme, accessability of the theme and pronominality of the theme. And finally, those related to the modality of the corpus, and the speaker.

The variables animacy, definiteness and pronominality are all marked by different kinds of features, and are thus easy to integrate the same way we did with collocations:

(7)

$$\left[\begin{array}{c}\text{NP}_x,\ \text{NP}_z, \\ \text{ARG-STR}\left\langle \text{NP}_y \left\{ \begin{array}{l} \dots \\ \left[\text{CAT } \textit{pron}\right],\ \Delta_p a \\ \left[\text{FRAMES}\left\langle\left[\textit{anim}\right]\right\rangle\right],\ \Delta_p b \\ \left[\text{MARKING def}\right],\ \Delta_p c \\ \dots \end{array} \right\} \right\rangle\end{array}\right]$$

In (7) we model pronominality through the CAT feature which specifies the category of the sign (which can be a pronoun or not), we model animacy with an animacy frame shared by animate entities, and definiteness through the MARKING feature which indicates whether a noun has a given "marking" of a particular determiner. Because each element in the set has its own attraction strength, and each element is independent from each other, the weights can be added up (with the appropriate method depending of which regression model is used for estimating the weights). Length and accessibility are factors harder to capture but could be done in a similar fashion. Accessibility could be modeled with some CONTEXT feature which provides contextual information, and length through the FORM feature, which lists all morphemes of the sign.

Finally, speaker and speech modality variation are model as corpus variation. Since the weights are defined for a given corpus, each speaker and each modality constitutes an independent corpus, with an independent set of weights.

III. A case study: German verbal collostructions

In this section I present a case study to illustrate how collostructions can be captured by the present model. Since in this proposal argument structure collostructions are a lexical phenomenon, we can extract them easily and automatically from a parsed corpus. I am using the part A of the Hamburg Dependency Treebank [3], which contains around two million words. I extracted all valency patterns from the corpus in the form of an ordered list of pairs, where the first element of a pair is a simplified POS tag, and the second element is the grammatical function of the dependent. After minimal processing[3] I found a total of 163 valency patterns in the corpus. However, most valency patterns have an extremely low frequency, and the top 10 patterns account for 90% of observed verbs (Figure 1).

We can also check how verbs are distributed across valency patterns. Figure 2 presents the number of verbs that appear with a given number of valency patterns. We can see that the majority of verbs only appear with 2, 3 or 4 different patterns, while a tiny minority can be found with more than 20 (overall mean= 4.9, median=4). This confirms the observation above that most verbs seem to be confined to very few valency patterns.

This approach presents additional possibilities over the traditional approach. Because we have access to **all** verbal collostructions, we can calculate not only the interactions of one construction with its lexical items, but the interactions between and across constructions. This includes measurements like constructional entropy (Figure I) and verbal entropy (Figure II). The constructional entropy can be calculated by the distribution of the verbs within a construction. Constructions with many different verb types will have a much higher entropy than constructions with few types, or constructions where a single type has a much higher frequencies than the other types in the construction. Similarly, verbal entropy measures how disperse the verb is across constructions. Verbs that appear with very few constructions, or that are extremely

[3]I simplified pronouns, nonwords and foreing words to nouns. I also normalized particle verbs so that the particle is counted as part of the verb and not as a dependent.

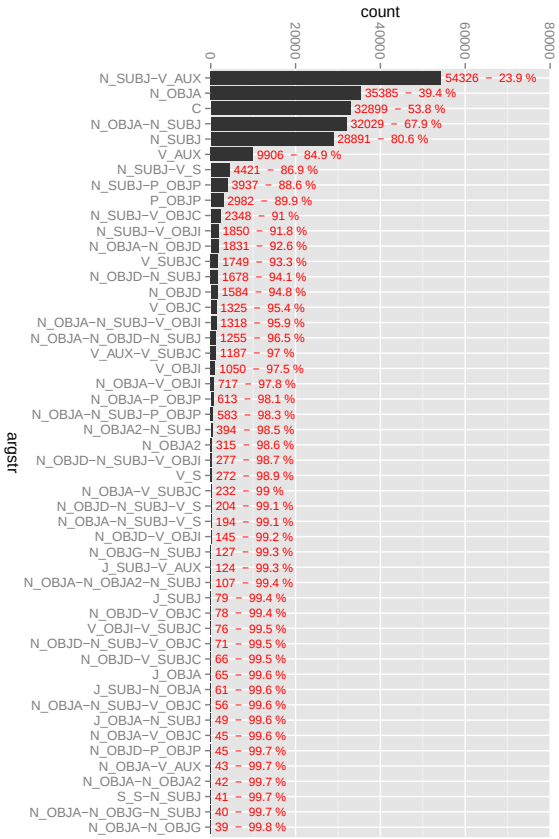


Fig. 1. Main 70 valency patterns in the corpus. Token frequencies are in red, followed by their cumulative percentage. POS: N=noun (also pronouns), P=preposition, A=adverb, J=adjective, V=verb. Syntactic functions: SUBJ=subject, OBJA=accusative object, OBJD= dative object, OBJP=prepositionl object, OBJI=infinitive verb as object, S=sentence, AUX=infinitive verb in verb chain, SUBJC=subject clause, OBJC=object clause.

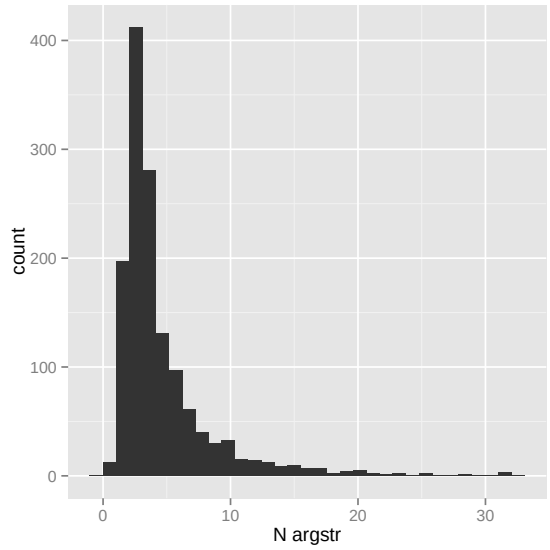


Fig. 2. Distribution of verbs across valency patterns.

frequent with only one construction will have a low entropy, while verbs that are evenly distributed across many different constructions will have high entropy.

pattern	N	types	entropy
C	32899	2758	6.4983
N_OBJA	35385	2417	6.4771
N_OBJA-N_SUBJ	32029	1817	5.7090
N_OBJD	1584	281	4.8892
V_OBJC	1325	231	4.6346
N_OBJA-N_OBJD	1831	245	4.6283
N_OBJA-N_OBJD-N_SUBJ	1255	186	4.3195
N_SUBJ-V_OBJC	2348	205	4.3126
J_OBJA	65	59	4.0464
P_OBJP	2982	138	3.9878
...			

TABLE I
HIGHEST ENTROPY VALENCY PATTERNS

verb	constructions	tokens	entropy
erzählen	16	29	2.655638
helfen	16	171	2.462284
fragen	19	132	2.452410
bitten	14	84	2.368767
überlegen	15	45	2.359551
empfehlen	18	143	2.305882
erinnern	16	56	2.292603
zwingen	14	143	2.218775
lehren	10	21	2.202521
versichern	17	108	2.178640
...			

TABLE II
HIGHEST ENTROPY VERBS

We can measure attraction to a given construction, in this case the dative construction defined by the valency pattern: N_OBJA-N_OBJD-N_SUBJ. Within this approach a p-value could be in principle calculated like in the traditional method, but it makes relatively little sense within the formal model because p-values do not follow organically from the signature distribution. There are however many alternatives. Some possibilities are:

$$WP_1(v) = \frac{P(v|c)}{H(v)} \quad (1)$$

$$WP_2(v) = P(v|c) \quad (2)$$

where:

$P(v|c)$ = probability of the verb in the construction ($\frac{N(v|c)}{N(c)}$: number of occurrences of the verb in the construction divided by the number of occurrences of the construction)

$H(v)$ = entropy of the verb: $-\sum P(x_i) log_2 P(x_i)$ for $i \in C(onstructions)$. That is, the dispersion of the verb across constructions.

WP_1(v) measures the probability of a given verb appearing in a particular construction, divided by the entropy (dispersion) of the verb. WP_2(v) simply measures the probability of a given verb within a particular construction. Notice that with these elements further attraction strength measures are possible, and the decision of picking any particular one should be based on empirical work.

We can test how WP_1 and WP_2 compare to p-value ranking. We see in Figure 3 a scatter plot of the log transformed p-values of a Fisher's Exact Test against PW_1, and in Figure 4 the corresponding plot for PW_2.

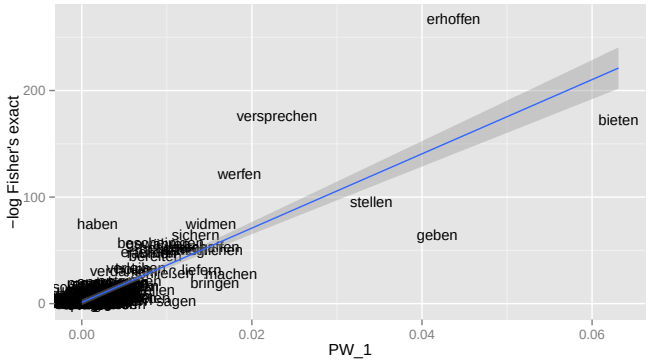


Fig. 3. PW_1 vs Fisher's exact

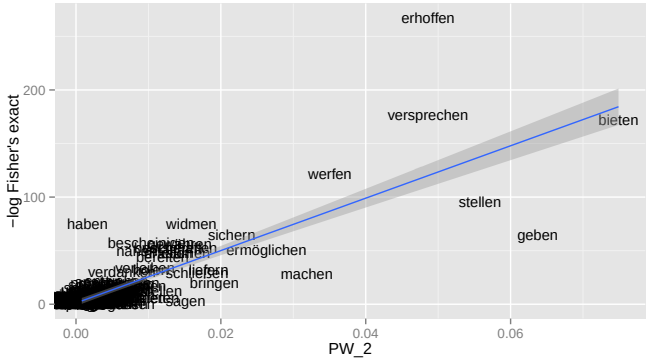


Fig. 4. PW_2 vs Fisher's exact

We can see that both attraction measurements provide rankings that are actually very similar to that provided by the p-values of the Fisher's Exact test. Even more, if we compare the rankings produced by the Fisher Test with the one produced by the Chi-Square[4] (Figure 5) test we see that the differences are more or less the same as with WP_1 and WP_2 with Fisher's Exact Test. There is no single one ranking for collostructions, and no right ranking we can compare against. Both WP_1 and WP_2 calculate different things, but both are perfectly valid for their individual purposes.

Having access to the whole collostructional space allows us to perform analysis that are not possible with the traditional method. We can, for example, try to test whether collostructional analysis should be seen as lexical or as phrasal. I claim that a prediction of the lexical view of collostructional analysis, is that lexically related valency patterns (by the inheritance hierarchy or lexical rules) will show high correlation in the

[4]For the cases where the p-value of the Chi-Square test was zero, I assigned a ranking value equal to the maximum ranking value found in the data set for the non-zero p-values.

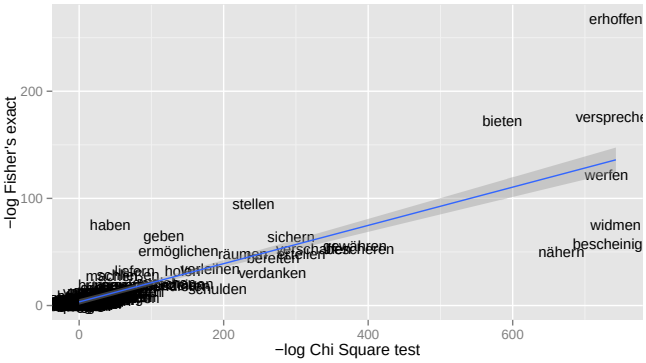


Fig. 5. Chi-Square vs Fisher's exact

proportions of verbs that appear with them. We can see a correlation matrix

Fig. 6. Correlations between ten most frequent valency patterns. The upper triangle shows the correlation values, and the lower triangle shows the scatter plots.

We can clearly see in Figure 6 that the valency patterns with highest correlations are those that are related by just 'subject deletion'. This pattern is easily explained in the lexical approach because a lexical rule can easily relate [A_SUBJ-B_y] to [B_Y]. A phrasal approach would require something like a transformation to link to valency patterns this way. An alternative could be, in the particular case of null subjects sentences, to claim that these instantiate two phrasal constructions: one "regular" argument structure construction (eg. transitive) construction and a null instantiation construction. But such an account would predict that the null instantiation construction should also attract verbs of its own and thus correlate across argument structure constructions, but this is something we clearly do not find in the data.

Finally, this same correlation matrix can be used to induced clusters of "verb types" from the data. Figure 7 presents a simple cluster analysis with the most frequent valency patterns. We see once more that the clusters are mostly related by argument deletion, and in some cases argument substitution, and that some very clear groups emerge. In group (3) we have

mostly transitives (with an accusative object); group (1) has mostly intransitives and modals taking an infinitive (V_AUX); group (8) has verbs with double accusatives; group (6) verbs with prepositional objects, etc.

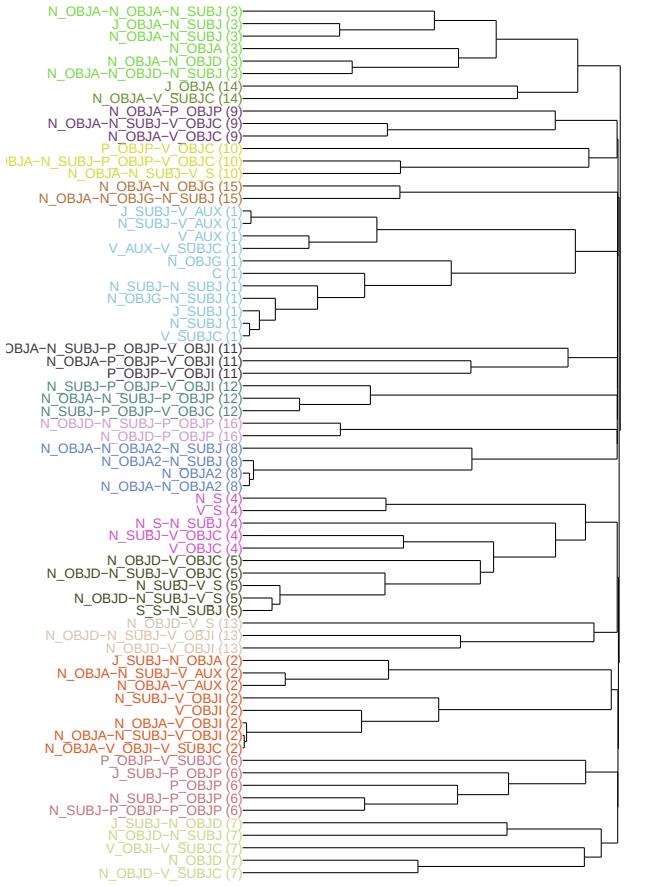


Fig. 7. Cluster analysis of the 70 most frequent valency patterns, based on correlation distance. Numbers mark group.

Without something very similar to the type hierarchy in SBCG/HPSG it is not clear how phrasal approaches could explain these data.

IV. Conclusion

The key innovation in this approach is that it allows us to model collocations, collostructions and alternation distribution with a single, unified mechanism. SBCG is a fundamentally lexicalist theory, and formalizing quantitative effects within this framework means also making all these effects lexical properties. Although originally collostructions and collocations where understood as related phenomena because construction grammar treats the difference between grammar and lexicon as a gradient, with the current proposal it is made explicit in which way they are identical and in which ways they are different. Additionally, because the model is designed as an interface, it allows some grammar and usage to retain some modularity while closely interacting (eg. most descriptions of grammatical patterns do not need to make reference to the frequency of the patterns).

One advantage of this approach for formal linguistics is that we can argue on the basis of quantitative evidence for changes to some aspects of the formal model. A concrete example would be the direction of some selectional features. It is possible that we should find that multiple cases of collocational attraction are in reverse direction, which would argue for reverse selectional features where complements can select for their heads in some contexts, as it has been argued for before in cases like periphrasis [1]. Conversely, theoretical findings like the fact lexical approaches to argument structure are superior to phrasal ones [6] can inform how we think about and how we model quantitative effects.

Formalizing quantitative effects is in the interest of both formal linguists and quantitative linguists. For formal linguists this model offers a way of meeting most of the challenges put forward by usage-based approaches, and a solid proposal for what the interface between grammar and usage is. For quantitative linguists it offers a way of formalizing their findings, and of imposing clear constraints on their models.

References

[1] Olivier Bonami. Periphrasis as collocation. *Morphology*, 25(1):63–110, 2015.

[2] Joan Bresnan, Anna Cueni, Tatiana Nikitina, R Harald Baayen, et al. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.

[3] Kilian Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. Because size does matter: The hamburg dependency treebank. In *Proceedings of the Language Resources and Evaluation Conference 2014 / European Language Resources Association (ELRA)*. Universität Hamburg, 2014.

[4] Stefan Th Gries. 50-something years of work on collocations: what is or should be next. *International Journal of Corpus Linguistics*, 18(1):137–166, 2013.

[5] Stefan Müller. Phrasal or Lexical Constructions? *Language*, 82(4):850–883, 2006.

[6] Stefan Müller and Stephen Wechsler. Lexical approaches to argument structure. *Theoretical Linguistics*, 40:1–76, 2014.

[7] Ivan A. Sag. English filler-gap constructions. *Language*, 86:486–545, 2010.

[8] Ivan Sag and Thomas Wasow. Performance-Compatible Competence Grammar. In K. Börjars R. D. Borsley, editor, *Non-Transformational p Syntax: Formal and Explicit Models of Grammar*, pages 359–377. Wiley, 2011.

[9] Ivan Sag. Sign-Based Construction Grammar: An Informal Synopsis. In Ivan A. Sag Hans C. Boas, editor, *Sign-Based Construction Grammar*, pages 69–202. University of Chicago Press, 2012.

[10] Anatol Stefanowitsch and Stefan Th Gries. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243, 2003.