Stefan Th. Gries (Santa Barbara)

Operationalizations of domain-general mechanisms cognitive linguists often rely on: a perspective from quantitative corpus linguistics

Abstract: In this paper, I discuss ways to corpus-linguistically operationalize and explore four domain-general cognitive mechanisms that have figured prominently in cognitive-linguistic studies of the lexicon/constructicon.

1 Introduction

Over the last few decades, linguistics has changed considerably, both with regard to theory and methodology. As for theory, after a long time during which "theoretical linguistics" was fairly synonymous with "generative linguistics", more and different theories have emerged; of interest for this paper is the 'family of theories' variously referred to as *cognitive linguistics*, (*cognitive*) *construction grammar*, and/ or *usage-based linguistics*, which hold that much of linguistic acquisition, representation, processing, use, and change can be explained with regard to actual usage of language. In these approaches, 'the stimulus' is considered to be less impoverished than is often assumed, and frequency matters (cf. Bybee 2010). Also, language is considered less modular than often assumed: Much of language learning, processing, etc. is argued to be explainable by domain-general cognitive processes/mechanisms (or should be until this approach breaks down and needs domain-specific mechanisms); examples of such mechanisms include *frequency*, *contingency*, *context*, and *recency* (cf. Ellis 2011; Gries/Ellis 2015 for overviews).

As for methods, after a long time during which much of "theoretical linguistics" was dominated by judgment data, more studies are now using a wider range of empirical data, including experimental, observational, and simulation data. Observational data in the form of corpus data have seen a particularly strong increase, as have, more slowly, statistical methods.

These two developments can be seen as somewhat related because corpus data are distributional in nature and, thus, often require statistical methods to make sense of. However, when cognitive linguists study the mental lexicon or, since cognitive linguists/construction grammarians typically eschew a qualitative distinction between lexis and syntax, the constructicon, using corpora, then

corpus and psycholinguists cannot help but feel that corpora are considerably underutilized. More bluntly, while there is a lot of talk about multidimensional exemplar space and cognitive mechanisms, cognitive-linguistic studies using corpus data often do not go beyond observed frequencies and what those mean for entrenchment and processing. In this paper, I will discuss the role of the above four domain-general mechanisms and how they might be studied using corpus-linguistic tools in ways that help cognitive linguists do more justice to what is captured in the well-known and much-endorsed 'cognitive commitment', "to make one's account of language accord with what is generally known about the mind and brain from disciplines other than linguistics" (Lakoff 1991, p. 54).

2 Frequency

In this section, I will discuss an example of how sometimes raw or relative *frequency* is really all that a cognitive/usage-based approach to language requires (based on case study 1 in Gries 2011). I am concerned with the question of whether the frequency of a certain phenomenon, within-unit alliteration, is higher than one might expect. For this, we need to first define a central concept in Langacker's Cognitive usage-based Grammar, the unit:

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

In Cognitive/Construction Grammar, units can exhibit different degrees of complexity; from low-complexity/abstractness cases (morphemes, monomorphemic words) via polymorphemic words, fully-fixed multi-word expressions, and partially filled multi-word expressions to high-complexity/abstractness syntactic/argument structure constructions. In Cognitive Grammar, symbolic units are conventionalized associations of a phonological pole (the form aspect of a unit/construction) and a conceptual pole (the meaning aspect of a unit/construction), whose relationships can be looked at in various ways: For instance, relations between the different conceptual poles of a unit are instances of polysemy or homonymy and have been widely studied in cognitive work. For instance, relations between the phonological and the conceptual pole of a unit are addressed in work on the Sapir-Whorf hypothesis, sound symbolism, phonaesthemes, onomatopoeia, and the arbitrariness-vs.-motivation of the sign in general. However, relations within the phonological pole of a unit have received much less attention

(although see Boers/Lindstromberg 2005; Lindstromberg/Boers 2008a, b; Gries 2011; Boers/Lindstromberg/Eyckmans 2012).

While annotating data for a conceptual-pole study, I noticed that many of the idioms *run* participated in involved alliterations: *run rampant*, *run riot*, *run roughshod*, *run the risk*, which raised the question of whether this pattern was systematic in the sense of 'non-random'. To any quantitatively-minded reader, this means one will need to compare some observed frequency of alliteration against an expected/random one – the questions remaining are where to get such frequencies and what kinds of frequencies to use for this.

To study this effect, I retrieved V-NP idioms from the "Collins Cobuild Dictionary of Idioms" (2002); the V had to be a full lexical verb with no additional complements/adjuncts and the idioms had to occur ≥1/2m words in the corpus on which the dictionary was based; examples include *spill the beans*, *gain some ground*, *get the boot*, etc. For the observed frequency of alliteration, I noted the initial segments of the verb, the head of the NP_{DO}, and any other content word (e.g., *fight a losing battle*) and computed an observed relative frequency of alliterations, 11.3%.

For the expected frequency – the baseline to compare 11.3% to – things are more complex because, even though frequency appears to be a straightforward notion, that simplicity is deceptive: the question arises as to how to compute it or what to put into the denominator of the fraction that computes the expected frequency. In that study, to insure against statistical artifacts, four different questions were asked whose answers determined the denominator of the relative frequencies:

- how many *different phonemes* are there that English words begin with (47 in the CELEX database, Baayen/Piepenbrock/Gulikers 1995) and what is, thus, the chance that two content words from the same idiom begin with the same phoneme?
- in how many *word types* do different phonemes occur word-initially and what is, thus, the chance that two content words from the same idiom begin with the same word token?
- in how many *word tokens* do different phonemes occur word-initially and what is, thus, the chance that two content words from the same idiom begin with the same one?
- in how many of a random sample of transitive clauses from a corpus (the British Component of the International Corpus of English) do two content words begin with the same phoneme?

Here, all results led to the same conclusion: The baselines of the above strategies amounted to 2.1%, 6%, 4.7%, and 4.8%, which are all significantly lower than the

observed alliteration percentage in the idioms: idiom status seems correlated with phonological similarity, at least when operationalized as alliteration.

There are various implications of this findings and various follow-up questions. As for the former, Gries (2011) explored how this finding can be related to idiom lexicalization from a Cognitive Grammar perspective (using Langacker's 1997 definition of phonological and conceptual constituents), and much of the work by Boers, Lindstromberg and colleagues is concerned with mnemonic effects of alliteration (plus assonance and rhyme) on the learnability of phrasal chunks, etc. As for the latter, it would be useful to explore what the exact locus of similarity is – just the beginning sound, the onset, the rhyme, etc., but crucially, studying any of these still requires the analyst to decide on ways to compute the right baseline. This is not as trivial a question as it may seem: First, not all studies have computed baselines in the first place (e.g. Boers/Lindstromberg 2009) and, second, not all studies compute even the observed frequencies correctly, which can impact the results.

As for the latter point, Schlüter (2005) reports frequencies of present perfects for several different corpora, but does so by dividing the observed numbers of present perfects by the numbers of words of the corpus, not the numbers of verbs. Thus, any differences found may be due to different frequencies of present perfects, but also just due to different frequencies of verbs! Thus, even though the notion of frequency appears straightforward, it need not be: trivially, how observed and expected frequencies are computed affects the results.

That being said and in spite of the huge role that frequency plays in cognitive-linguistic publications, it is also necessary to 'put it in its place' a bit: Not only is it probably always necessary to augment it with other information (such as contingency or entropy), but a consensus seems to emerge that frequency as a mere repetition-counter is much less a proxy of entrenchment than is often assumed and requires more fine-grained data on context and recency.

3 Contingency

Given the important role of frequency in cognitive-linguistic publications, why exactly would it needed be put in its place and what would contingency add to the picture? As in particular Ellis has argued in a variety of insightful overview papers (see esp. Ellis 2011), frequency is not the only determinant of learning, but the three other central notions at the core of this paper – contingency, recency, and context – are all relevant, too (as are others), to understand how the human processor deals with our probabilistic uncertain world. For contingency in particu-

lar, Ellis (ibid., p. 7) discusses (referencing Peterson/Beach 1967) how "human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like r, χ^2 and ΔP" and how "it [is] contingency, not temporal pairing, that generated conditioned responding" in classical conditioning; he concludes that "[l]anguage learning can thus be viewed as a statistical process in that it requires the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations" (Ellis 2011, p. 12).

One area in corpus linguistics in which contingency has played a huge role is that of co-occurrence of lexical items (collocations) and of lexical items with syntactic constructions (colligations/collostructions) and much debate has been concerned with finding measures of contingency, so-called association measures (AMs), whose application to data yields results that appear reasonable/useful and/or are correlated with other (often experimental) data; two examples of such studies include Mollin (2009) and Michelbacher/Evert/Schütze (2011). As different AMs are being debated, the perhaps most important question is actually a somewhat more general one: Should an AM be a measure that reflects both frequency/sample size and effect size (such as (logged) $p_{\text{Fisher-Yates exact}}$, the log-likelihood statistic G^2, χ^2, t, ...) or a measure that reflects only effect size (odds ratio, Cramer's V / φ, ΔP, ...).

Currently, G^2 is probably the most widely-used measure, which is largely due to the reasons that (i) it is the best approximation to $p_{\text{Fisher-Yates exact}}$, (ii) Dunning (1993) did much to popularize it, and (iii) several software packages provide it even for corpus linguists with no statistical and/or programming knowledge. In this section, I will discuss a recent proposal to use G^2 not just for collocate rankings of x collocates to one node word, but for the bottom-up identification of multi-word units (MWUs)/constructions, or n-grams, where n is not set by the user, but 'decided' by an algorithm.

This approach (Wahl/Gries under review a, b) is called MERGE (for Multi-word Expressions from the Recursive Grouping of Elements). Similar to some other work, it embodies a recursive 2-gram approach, but unlike other work, our algorithm is designed to extract all MWUs in a corpus and not just those that contain a particular node word. MERGE begins by extracting all 2-gram tokens in a corpus, which may include contiguous 2-grams, as well as bigrams with gaps (at present, we have only experimented with 1-word gaps). The tokens for each 2-gram type are counted, as are the tokens for each individual word type as well as the corpus size. Next, these values are used to calculate G^2-scores. The highest-scoring 2-gram is selected as 'the winner' and merged into a unit and all instances of that 2-gram are replaced by instances of the new, merged unit. That means that all frequency information – 1-gram and 2-gram statistics and the corpus size –

must be updated as new candidate n-grams are created through the co-occurrence of individual 1-grams with tokens of the newly-merged 2-gram. After all these adjustments have been made, new 2-gram strengths can be calculated and the cycle iteratively repeats from the point at which a winning 2-gram is chosen above; this continues until the lexical association strength of the winning 2-gram reaches some minimum cutoff threshold or a user-defined number of iterations has been processed, after which the output of the algorithm is a corpus, parsed in terms of MWUs, and a list of 1- to n-grams of different sizes, with and without gaps.

As discussed above, MERGE uses an AM that conflates frequency/sample size and contingency, a decision that, in spite of the widespread use of G^2, is not uncontroversial – MERGE needs to be demonstrated to work in and of itself, but maybe also in comparison with competing approaches such as O'Donnell's (2011) Adjusted Frequency List (AFL), which uses only frequency and not also contingency. Specifically, the AFL works by first identifying all n-grams up to some size threshold in a corpus. Then, only n-grams exceeding some frequency threshold – 3 in O'Donnell (2011) – are retained along with their frequency and, for each n-gram, starting with those of threshold length and descending by order of length, the two component n-minus-1-grams are derived. Finally, the number of tokens in the frequency list of each n-minus-1-gram is decremented by the number of n-grams in which it is a component and no smaller n-grams can be 'taken out of' larger n-grams anymore. In the next two sub-sections, I discuss results of two case studies (from Wahl/Gries under review a, b) that aim at testing MERGE's performance against the AFL; for two additional case studies, see those two papers.

3.1 Validation 1: MERGE vs. AFL in ratings

The input data for the algorithm comprised two corpora: The Santa Barbara Corpus of Spoken American English (SBC; Du Bois et al. 2000-2005) and the spoken component of the Canadian subcorpus of the International Corpus of English (ICE-Canada Spoken; Newman/Columbus 2010). SBC includes about 250K words, while ICE-Canada Spoken includes about 450K. First, the formatting of both corpora was standardized: All tags and transcription characters not part of the lexical representation of the words were removed, including markers of overlap in talk, laughter, breathing, incomprehensible syllables, pauses, and other non-lexical vocalizations, among other features.

Next, both MERGE and the AFL were run and the top 1000-ranked items from the output of each were selected for further consideration. In the case of MERGE, this involved simply running the algorithm for 1000 iterations. In the case of the AFL, the minimum frequency threshold was set to five and the 1000 items with

highest frequencies were selected. We then decided to focus on the MWEs that the two algorithms did not agree on rather than the MWEs that they had in common. Two groups of items were created: The first group was those items found in the AFL output but not in MERGE's; the second group was those items found in MERGE's output but not in the AFL's. This allowed a highly tractable examination of how the respective performances of the two algorithms contrasted, as stimulus items fell into one of two categories. From each of the two groups of disjunctive outputs, 180 items were then randomly sampled and an even distribution of sampling from across the range of items was achieved by partitioning the two rank-ordered item groups into 10 bins and randomly sampling 18 items from each bin. We then created four different questionnaires of 90 items each by combining 45 randomly-chosen MERGE items with 45 randomly-chosen AFL items and randomized their presentation orderings per subject. Study participants were asked to rate sequences based on whether, in their opinion, they represented a *complete unit of vocabulary*. The hope was that participants' understanding of the notion of *vocabulary* would be compatible with the notion of a lexicon, since these U.S. students would have grown up learning vocabulary lists in spelling classes, etc. 20 participants were recruited from an introductory linguistics course at the University of California, Santa Barbara and each participant was placed in a quiet room by themselves and given as much time as they needed to complete the survey. Only data from native speakers of English was used in the final analysis.

Table 1: Results for the fixed-effects part of the regression model (REML)

Predictor	*coef*	*se*	*df*	*t*	$p_{one\text{-}tailed}$
Intercept	3.93	0.27	19.7	14.6	$<10^{-11}$
ORIGIN: *AFL → MERGE*	0.59	0.25	22.8	2.31	0.0151

The data were analyzed with a linear mixed-effects model. The dependent variable was RATING, i.e., the numerical rating provided by subjects for the MWUs; the independent variable was the binary variable ORIGIN, which specified where the rated MWU was from – AFL vs. MERGE; the random-effects structure was maximal. The linear mixed-effects model we fitted resulted in a significant fit (LR chi-squared = 5, *df*=1, *p*=0.0254, from a ML-comparison to a model without fixed effects) but only a weak correlation: $R^2_{marginal}$=0.02, $R^2_{conditional}$=0.37; see Table 1 for the corresponding results. As is obvious from the above statistics, the overall effect is weak – although the product-moment correlation between the observed ratings and the one predicted by our model is *r*=0.68 – and the random-effects

structure explains more of the variance than the fixed effects. Nevertheless, the significant main effect of ORIGIN, while not strong and variable across subjects/ MWUs, provides support for the hypothesized usefulness of the MERGE algorithm: The randomly-sampled MWUs from MERGE score higher average formulaicity judgments than those from the AFL. Given the small effect size, the evidence is probably not conclusive but, here, the measure that combines, in fact conflates, frequency and contingency outperforms the one that only uses frequency.

3.2 Validation 2: MERGE vs. AFL in finding the BNC's multi-word units

In the second small case study to be reported here, we test which of MERGE and AFL is better at finding the ≈400 MWUs that corpus compilers annotated in the spoken component of the British National Corpus. In other words, while in the previous study we assessed the performance via naïve intuitions, here we are testing performance via specialist knowledge. Specifically, we applied both MERGE and AFL to the complete spoken component of the BNC to determine how well both methods can identify 388 expressions that were tagged as multi-word units (using <mw></mw>). We took the top 10,000 items from either approach and used one-tailed exact binomial tests to compare the proportions of BNC MWUs that either approach would identify; given the previous results and our belief that contingency information is vital (even if only conflated with frequency rather than kept separately), we expected MERGE to find a higher percentage of MWUs than the AFL.

Then, we checked how many of the 388 formulaic sequences from the BNC spoken were identified by the top 10,000 MERGE and AFL items: MERGE found 112 of the 388 MWUs whereas the AFL found only 93. According to a one-tailed binomial test, MERGE finds a significantly higher number than the AFL ($p_{\text{one-tailed}}$ =0.0152); conversely, according to a second one-tailed binomial test, the AFL performs significantly worse than MERGE ($p_{\text{one-tailed}}$=0.0178).

Again, the effect is not large, but in the predicted direction and significant, lending another bit of support for the view that contingency is useful. While undoubtedly more study is necessary, these results are instructive insofar as they support previous work on collostructions – the occurrence of verbs in syntactically-defined slots of constructions – and their experimental validation, which has also provided results in favor of AMs that reflect frequency *and* contingency over AMs that only reflect frequency (see Gries/Hampe/Schönefeld 2005, 2010; Gries 2015; among others). That being said, the main point of what follows is that any study of co-occurrence cannot be reliably, replicably, sufficient if it only includes frequency and contingency.

4 Context

The previous two sections discussed ways in which two central notions in cognitive linguistics – frequency and contingency – can be operationalized and instructive. However useful these two notions are, they nevertheless only scratch the surface of how cognitive linguists would need to characterize the multidimensional exemplar space they postulate constitutes linguistic (and other) knowledge. While both corpus-linguistic and psycholinguistic work has discussed useful extensions, some important findings have not yet led to an adjustment/extension of the main cognitive-linguistic toolbox. The first of these is concerned with a quantitative, more big-picture view of the context of linguistic expressions under consideration.

Much corpus-/cognitive-linguistic work explores the context of a linguistic item under consideration on the basis of concordance lines. While that allows an analyst to discover potentially every detail of each usage event, it also makes it harder to recognize higher-level patterns of usage and what they reveal about mental representation as well as how that may force us to rethink certain notions. For example, for many decades now frequency of use (as operationalized by (logged) corpus frequencies) has been seen as a major determinant of speed of access (as operationalized by, e.g., reaction times in experiments). However, for some time now, it has been suggested that frequency as a mere repetition counter may not be as useful as often assumed but that frequency is epiphenomenal due to its high correlation with other, truly relevant, measures. One of these is McDonald/ Shillcock's (2001) contextual distinctiveness (*CD*). Measuring the *CD* of some lemma *l* involves

- retrieving all instances of *l* within its context;
- computing the relative frequencies of a set of *n* collocates within a context window around *l* (e.g., ±5 words); this is the so-called *posterior distribution*, essentially the list of conditional probabilities $p(\text{collocate}|l)$;
- computing the relative frequencies of those *n* collocates in the corpus in general; this is the so-called *prior distribution*, essentially the list of probabilities $p(\text{collocate})$;
- compute the relative entropy/Kullback-Leibler divergence from the prior to the posterior distribution as in (1).

(1)
$$CD = \sum_{i=1}^{n} p(coll_i|lemma) \cdot log_2 \frac{p(coll_i|lemma)}{p(coll_i)}$$

This measure is correlated with observed raw frequency, but its computation does not involve it – *CD* is blind to the sparseness of the posterior distribution (if necessary or appropriate, smoothed relative frequencies can be used). In addition,

CD incorporates prior knowledge (in the form of the probabilities of collocates in the corpus at large) and, most importantly, in McDonald/Shillcock (2001, exp. 1), *CD* accounts for variance in reaction times even when word frequency and length are statistically controlled for, whereas frequency did not when word length and *CD* were statistically controlled: "[w]ords that appear in relatively constrained (or distinctive) linguistic contexts have high *CD*-scores and tend to attract longer lexical decision latencies" (ibid., p. 312).

Interestingly, this result is corroborated in a very comprehensive study of Baayen (2010), who found that

> the word frequency effect in the sense of pure repeated exposure accounts for only a small proportion of the variance in lexical decision, and that local syntactic and morphological co-occurrence probabilities are what makes word frequency a powerful predictor for lexical decision latencies. (Baayen 2010, p. 436)

This is of crucial theoretical interest to cognitive linguistics, which has emphasized the role of frequency as a repetition counter and, thus, as one of the most important driving forces of entrenchment (e.g., Langacker 1987, p. 59, 100; Schmid 2010, pp. 118 f.); this in turn has influenced the discussion of (interactive) activation in the cognitive-linguistic literature (e.g., Langacker 1991, p. 45). However, findings such as the above are very compatible with many recent findings of the relevance of other information-theoretic notions such as surprisal in corpus-based psycholinguistics. For example, Jaeger (2011) studies subject-extracted relative clauses (SRC) and finds that using the full version is highly significantly correlated with the surprisal of seeing an SRC given the noun as well as the surprisal of seeing an SRC given the participle. Similarly, Linzen/Jaeger (2015) find that the entropy reduction of potential parse completions is correlated with reading times of sentences involving the DO/SC alternation. In other words, in more and more studies, there is evidence for the importance of predictors involving planning (to produce) and predicting (to comprehend) upcoming constituents in ways that go beyond frequency/entrenchment *per se* and that point to the need to re-conceptualize frequency as maybe just a crude proxy towards contextual diversity (e.g., within Anderson's 1990 theory of rational analysis as discussed in detail by Ellis 2011).

5 Recency

The last domain-general cognitive notion to be discussed here is recency. Recency is obviously strongly related to acquisition/learning as well as forgetting because "the extent to which the number of repeated exposures to a particular item affects that item's later retrieval depends on the separation of the exposures in time and

context" (Adelman/Brown/Quesada 2006, p. 814). The corpus-linguistic equivalent to this "separation of the exposures in time and context" is dispersion, the degree to which occurrences of something are spread out evenly in, say, a corpus (even dispersion) or are concentrated in maybe just a small part of it (uneven/clumpy dispersion). Dispersion is as important as it is underutilized in corpus-linguistic but especially cognitive-linguistic studies. The reason it is so important is that it can provide a more reliable view of how regularly some linguistic item appears in language, or how likely it is that some speaker would encounter that item, than a frequency estimate, a claim that is supported in Figure 1.

In both panels, logged word frequency (from the spoken BNC for all words occurring 10+ times) is on the *x*-axis. In the first/upper panel, a dispersion measure called *DP* (Deviation of Proportions) is on the *y*-axis: $DP{\approx}1$ means a word is distributed extremely clumpily whereas $DP{\approx}0$ means a word is distributed extremely evenly; correspondingly, the dots to the very right are all function words. In the second/lower panel, the *y*-axis is range, i.e. the number of files (out of all 905) in which a word is attested. In both panels, word frequency has been binned into 10 equally wide bins and, for each bin, the interval represents the minimum and maximum of the *DP*/range values, with the corresponding ranges at the bottom (left panel) and the top (right panel). It is plain to see that there is a huge variability in dispersion in particular in the middle range of frequencies where many 'normal content words' are. For instance, a relatively 'specialized' word like *council* is in the same (6th) frequency bin (*freq*=4386, *DP*=0.72, range=292 out of 905) as intuitively more 'common/widespread' words like *nothing*, *try*, and *whether* (*freqs*=4159, 4199, 4490; *DPs*=0.28, 0.28, 0.32; ranges=652, 664, 671 out of 905). Also, even just in the 6th frequency band, the extreme range values values that are observed are $^{85}/_{905}$=9.4% vs. $^{733}/_{905}$=81% of the corpus files, i.e. huge differences between words that in some less careful study might well be considered 'similar in frequency'.

The above study by Adelman/Brown/Quesada (2006) is concerned with dispersion, although they refer to it as *contextual diversity* – an unfortunate misnomer, given that the use of a word in different corpus files does by no means imply that the actual contexts of the word are different: No matter in how many different files *hermetically* is used, it will probably nearly always be followed by *sealed*. That mislabeling aside, they do show that dispersion is a better and more unique predictor of word naming and lexical decision times and they, too, draw an explicit connection to Anderson's rational analysis of memory and conclude "number of contexts has an effect because the more contexts an item has occurred in, the more likely that item is to be needed in any new context" (ibid., p. 822), a finding that chimes in well with the kind of information-theoretic results mentioned above and some others to be briefly mentioned below in the conclusion.
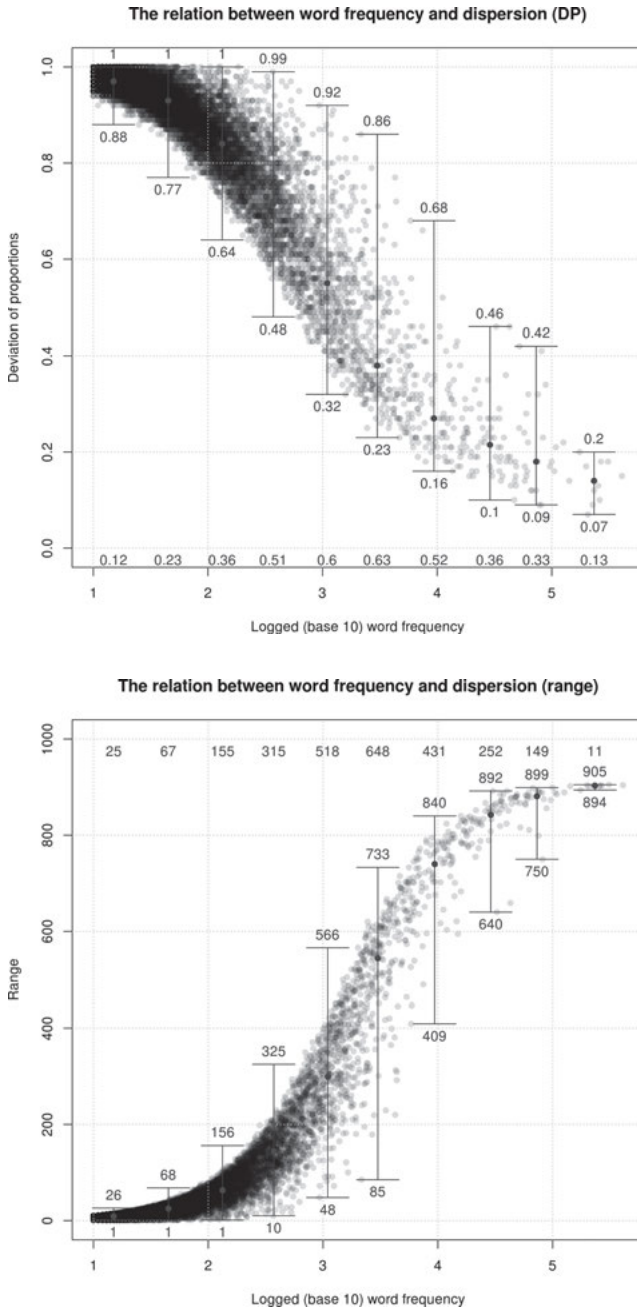
The relation between word frequency and dispersion (DP)

The relation between word frequency and dispersion (range)

Fig. 1: The relation between frequency and dispersion in the spoken part of the BNC

6 Conclusion

What follows from all this? Starting out from the observation how cognitive-linguistic or usage-based linguistics has developed into a widely-used alternative theoretical framework to the generative framework that has been predominant for decades, I discussed a variety of domain-general cognitive mechanisms that cognitive linguistics often relies on to explain acquisition, representation, processing, use, and change: frequency, contingency, context, and recency. However, while the rise of cognitive-linguistic work has coincided with a greater use of corpus data in linguistics, it is my impression that cognitive/usage-based linguistics under-utilize what corpus data have to offer and what they have already contributed in neighboring fields such as corpus linguistics proper and psycholinguistics. Maybe a bit polemically, it seems as if (large? parts of) cognitive linguistics are falling into a similar trap as the field did in the 1990s when polysemy and network analyses were proposed left and right that, as convincingly shown especially in Sandra/Rice's (1995) landmark article, were simplistic and/or vague when it came to explicating the status of network notations, the role of different kinds of evidence etc. etc. How so? Well, from a quantitative corpus linguist's point of view, now there still seems to be too much simplifying 'frequency is proportional to entrenchment' or 'frequency/entrenchment is proportional to resting levels of activation/connection weights' and similar talk around – how many publications in cognitive linguistics talk about frequency and entrenchment and their role for acquisition, processing etc. compared to how many relate their data to findings like those discussed ever so briefly above – like how contingency adds to frequency (as long argued by Ellis), how context is more uniquely important than frequency-as-repetition (as long argued in some psycholinguistic circles), how dispersion may be more reliable than frequency in particular for anything having to do with learning?

Maybe it is about time that we let go of the relatively simple ways in which we have been using 'frequency' as a cure-all and realize that frequency is just one and, at least in some case, apparently quite epiphenomenal notion and that more refined and in particular information-theoretic measures have more to offer in terms of explanation and convergence with neighboring fields (recall Lakoff's cognitive commitment). Just two examples of findings that an overly strong reliance on 'frequency → entrenchment → processing speed' could never have uncovered: First, Lester/Moscoso del Prado Martín (2016) find that nouns that project a diverse array of structures are produced faster and conclude

> words are finely articulated syntactic entities whose history of use partially determines how efficiently they are processed [...] Perhaps words and syntactic structures are much more tightly linked than is typically acknowledged. (ibid., p. 2589)

More concisely, entropies of syntactic distributions affect response times of nouns in isolation and the ordering in coordinate NPs. Second and more interestingly even, Lester/Feldman/Moscoso del Prado Martín (2017) find that words occurring in similar distributions of syntactic constructions prime each other: Nouns' representations appear to be connected to syntactic structures in proportion to how often they occur in them in a way modular accounts do not predict. And what exactly is the connection between notions I haven't talked about in this paper – e.g., salience – and information-theoretic measures such as surprisal?

These kinds of findings and these kinds of questions should trigger something in the field, both for their content – empirical priming results appearing to support the position of abandoning the lexis-syntax boundary that cognitive linguistics has adopted, how great is that? and there are corpus-based measures for cognitively relevant notions such as salience? – but also for how they are and would be arrived at, by rigorous and comprehensive analysis of corpus data that goes beyond what is found in much cognitive-linguistic corpus work. I hope this paper can provide a small impetus in this direction ...

References

Adelman, James S./Brown, Gordon D.A./Quesada, Jose F. (2006): Contextual diversity, not word frequency, determines word-naming and lexical decision times. In: Psychological Science 19, 9, pp. 814–823.

Anderson, John R. (1990): The adaptive character of thought. Hillsdale, NJ.

Baayen, R. Harald (2010): Demythologizing the word frequency effect. A discriminative learning perspective. In: The Mental Lexicon 5, 3, pp. 436–461.

Baayen, R. Harald/Piepenbrock, Richard/Gulikers, Leon (1995): The CELEX lexical database. Philadelphia.

Boers, Frank/Lindstromberg, Seth (2005): Finding ways to make phrase-learning feasible: The mnemonic effect of alliteration. In: System 33, 2, pp. 225–238.

Boers, Frank/Lindstromberg, Seth (2009): Optimizing a lexical approach to instructed second language acquisition. Basingstoke.

Boers, Frank/Lindstromberg, Seth/Eyckmans, June (2012): Are alliterative word combinations comparatively easy to remember for adult learners? In: RELC Journal 43, 1, pp. 127–135.

Bybee, Joan (2010): Language, usage, and cognition. Cambridge/New York/Melbourne.

Collins Cobuild (2002): Collins Cobuild Dictionary of Idioms. For use with the Collins Cobuild Dictionary of Idioms. 2nd ed. London.

Du Bois, John W. et al. (2000–2005): Santa Barbara corpus of spoken American English. Part 1–4. Philadelphia.

Dunning, Ted (1993): Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics 19, 1, pp. 61–74.

Ellis, Nick C. (2011): Language acquisition as rational contingency learning. In: Applied Linguistics 27, 1, pp. 1–24.

Gries, Stefan Th. (2011): Phonological similarity in multi-word symbolic units. In: Cognitive Linguistics 22, 3, pp. 491–510.

Gries, Stefan Th. (2015): More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). In: Cognitive Linguistics 26, 3, pp. 505–536.

Gries, Stefan Th./Ellis, Nick C. (2015): Statistical measures for usage-based linguistics. In: Language Learning 65, 1, pp. 1–28.

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

Gries, Stefan Th./Hampe, Beate/Schönefeld, Doris (2010): Converging evidence II: More on the association of verbs and constructions. In: Rice, Sally/Newman, John (eds.): Empirical and experimental methods in cognitive/functional research. (= Conceptual Structure, Discourse and Language 9). Stanford, pp. 59–72.

Jaeger, T. Florian (2011): Corpus-based research on language production: Information density and reducible subject relatives. In: Bender, Emily M./Arnold, Jennifer (eds.): Language from a cognitive perspective: grammar, usage, and processing. Stanford, pp. 161–197.

Lakoff, George (1991): Cognitive versus generative linguistics. How commitments influence results. In: Language and Communication 11, 1/2, pp. 53–62.

Langacker, Ronald W. (1987): Foundations of Cognitive Grammar. Vol. 1: Theoretical prerequisites. Stanford.

Langacker, Ronald W. (1991): Foundations of Cognitive Grammar. Vol. 2: Descriptive applications. Stanford.

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

Lester, Nicholas A./Moscoso del Prado Martín, Fermín (2016): Syntactic flexibility in the noun: Evidence from picture naming. In: Papafragou, Anna et al. (eds.): Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci 2016), Austin, TX, pp. 2585–2590.

Lester, Nicholas A./Feldman, Laurie B./Moscoso del Prado Martín, Fermín (2017): You can take a noun out of syntax...: Syntactic similarity effects in lexical priming. In: Gunzelmann, Glenn et al. (eds.): Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017). Austin, TX, pp. 2537–2542.

Lindstromberg, Seth/Boers, Frank (2008a): The mnemonic effect of noticing alliteration in lexical chunks. In: Applied Linguistics 29, 2, pp. 200–222.

Lindstromberg, Seth/Boers, Frank (2008b): Phonemic repetition and the learning of lexical chunks: The power of assonance. In: System 36, 3, pp. 423–436.

Linzen, Tal/Jaeger, T. Florian (2015): Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. In: Cognitive Science 40, 6, pp. 1382–1411.

McDonald, Scott A./Shillcock, Richard C. (2001): Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. In: Language and Speech 44, 3, pp. 295–323.

Michelbacher, Lukas/Evert, Stefan/Schütze, Hinrich (2011): Asymmetry in corpus-derived and human word associations. In: Corpus Linguistics and Linguistic Theory 7, 2, pp. 245–276.

Mollin, Sandra (2009): Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. In: Corpus Linguistics and Linguistic Theory 5, 2, pp. 175–200.

Newman, John/Columbus, Georgie (2010): The International Corpus of English – Canada. Edmonton.

O'Donnell, Matthew Brook (2011): The adjusted frequency list: A method to produce cluster-sensitive frequency lists. In: ICAME Journal 35, pp. 135–169.

Peterson, Cameron R./Beach, Lee Roy (1967): Man as an intuitive statistician. In: Psychological Bulletin 68, 1, pp. 29–46.

Sandra, Dominiek/Rice, Sally (1995): Network analyses of prepositional meaning: Mirroring whose mind—the linguist's or the language user's? In: Cognitive Linguistics 6, 1, pp. 89–130.

Schlüter, Norbert (2005): How reliable are the results? Comparing corpus-based studies of the present perfect. Paper presented at the workshop 'The scope and limits of corpus linguistics – empiricism in the description and analysis of English'. Berlin.

Schmid, Hans-Jörg (2010): Entrenchment, salience, and basic levels. In: Geeraerts, Dirk/Cuyckens, Hubert (eds.): The Oxford handbook of Cognitive Linguistics. Oxford, pp. 118–138.

Wahl, Alexander/Gries, Stefan Th. (under review a): Multi-word expressions: A novel computational approach to their bottom-up statistical extraction.

Wahl, Alexander/Gries, Stefan Th. (under review b): Computational extraction of formulaic sequences from corpora: two case studies of a new extraction algorithm.