On the study of co-occurrence in corpus data: against frequency and towards a more multi-dimensional view

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Corpus data in cognitive and psycholinguistics

- One can distinguish two ways in which corpus data can inform linguistic research
  - to constitute/define/operationalize predictors, or independent variables (esp. in studies that might not be corpus-based otherwise)
    - e.g., when frequency data from corpora are used as statistical predictors or controls in lexical decision task studies or in self-paced reading time studies
    - e.g., when frequency and/or association data from corpora are used to construct stimuli (e.g., in multi-word unit reading time studies)
  - to constitute predictors and responses in corpus-based studies
    - e.g., in studies on phonetic reduction (based on phonetically/phonologically-annotated corpora)
    - e.g., in studies of syntactic production (based on morphosyntactically-annotated corpora)
- the former is less controversial, the latter more so

Against using corpus data for both predictors & responses

- Corpus data are
  - (too) noisy
  - (too) heterogeneous
  - (too) unbalanced/skewed (in particular Zipfian)
- in addition, corpus data often pose additional challenges
  - predictors may be (highly) collinear
  - non-independence of data points / autocorrelation
    - e.g., due to speakers providing more than one data point
    - e.g., due to words 'providing' more than one data point
    - e.g., due to priming effects (from many interrelated levels)
- that has led some to conclude corpus data should
  - only be used for exploratory studies
  - not be used for hypothesis-testing studies
  - "Corpora have proved useful as a means of hypothesis generation, but unequivocal demonstrations of syntactic priming effects can only come from controlled experiments" (Branigan et al. 1995:492)

In favor of using corpus data for both predictors & responses

- Yes, much of the above is true, but it is possible to address many of these challenges, plus corpus data also offer some advantages
  - statistical control: many of the challenges can be addressed statistically
    - collinearity can be diagnosed well and addressed
    - unbalancedness, dependencies, and autocorrelation can be addressed using advanced (regression) modeling techniques
  - ecological validity (see Jaeger's 2010 discussion)
    - experiments often introduce stimuli in balanced designs, which are not compatible with the relevant expressions' frequencies in the 'real world'
    - subjects learn distributional facts and, in particular, can do so even after only short periods of input
      - see phonological learning in Saffran & colleagues' work
      - subjects became more accepting of unconventional uses of constructions during just 8 trials (Doğruöz/STG 2014)

Corpus data always involve frequencies, but we can and should do better

- Consider, for example, how frequency has assumed a central role in cognitive/usage-based linguistics
- it has been assumed for a long time now that token frequency (co-)determines degree of entrenchment …
- … which (co-)determines speed/ease of lexical access, retrieval, production, …:

- "continuous scale of entrenchment in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence." (RWL 1987:59)
- "[t]his seems highly convincing, not least in view of the considerable body of evidence from psycholinguistic experiments suggesting that frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse […]. As speed of access in, and retrieval from, the mental lexicon is the closest behavioural correlate to routinization, this indeed supports the idea that frequency and entrenchment co-vary." (Schmid 2010a:115f.)

Kinds of frequencies

- We can distinguish different kinds of frequencies
  - conceptual frequency (see Hoffmann 2004)
  - type frequency
  - token frequency
- following Schmid, token frequency then can be divided into
  - absolute frequency (→ cotext-free entrenchment)
    - counts of x in a corpus (maybe normalized)
  - relative frequency (→ cotextual entrenchment)
    - counts of x with/close to y in a corpus
- for example,
  - abs. freq. of *x* : $a+b$
  - rel. freq of *x* given *y*: $a/_{a+c}$

	y	others	Σ
x	*a*	*b*	*a+b*
others	*c*	*d*	*c+d*
Σ	*a+c*	*b+d*	*a+b+c+d=N*

- it seems as if most of usage-based linguistics focuses on the latter
  - *p*(unit *x* | unit *y*) (the probability of *x* given *y*)
  - *p*(function *x* | unit *y*)

(Schmid discusses reasons to also consider co-text freq)

Seems reasonable/useful, no?

- But then Schmid also says

  It seems to me that many researchers, inclusing [sic!] myself, have had a great deal too much confidence in the potential of quantitative methods for the study of aspects of the linguistic and cognitive system. All quantitative methods that I am aware of ultimately boil down to counting the frequencies of tokens and types of linguistic phenomena. What I have tried to show here, however, is that so far we have understood neither the nature of frequency itself nor its relation to entrenchment, let alone come up with a convincing way of capturing either one of them or the relation between them in quantitative terms. (Schmid 2010a:125)

- I believe this assessment is
  - too pessimistic
  - in part due to a bit of a lack of understanding of the methods that paper discusses
  - sub-optimal handling of some issues that come up when dealing with frequency (I am guilt of that myself)
- on the other hand, …

On the other hand, maybe token frequency is less important than usually thought

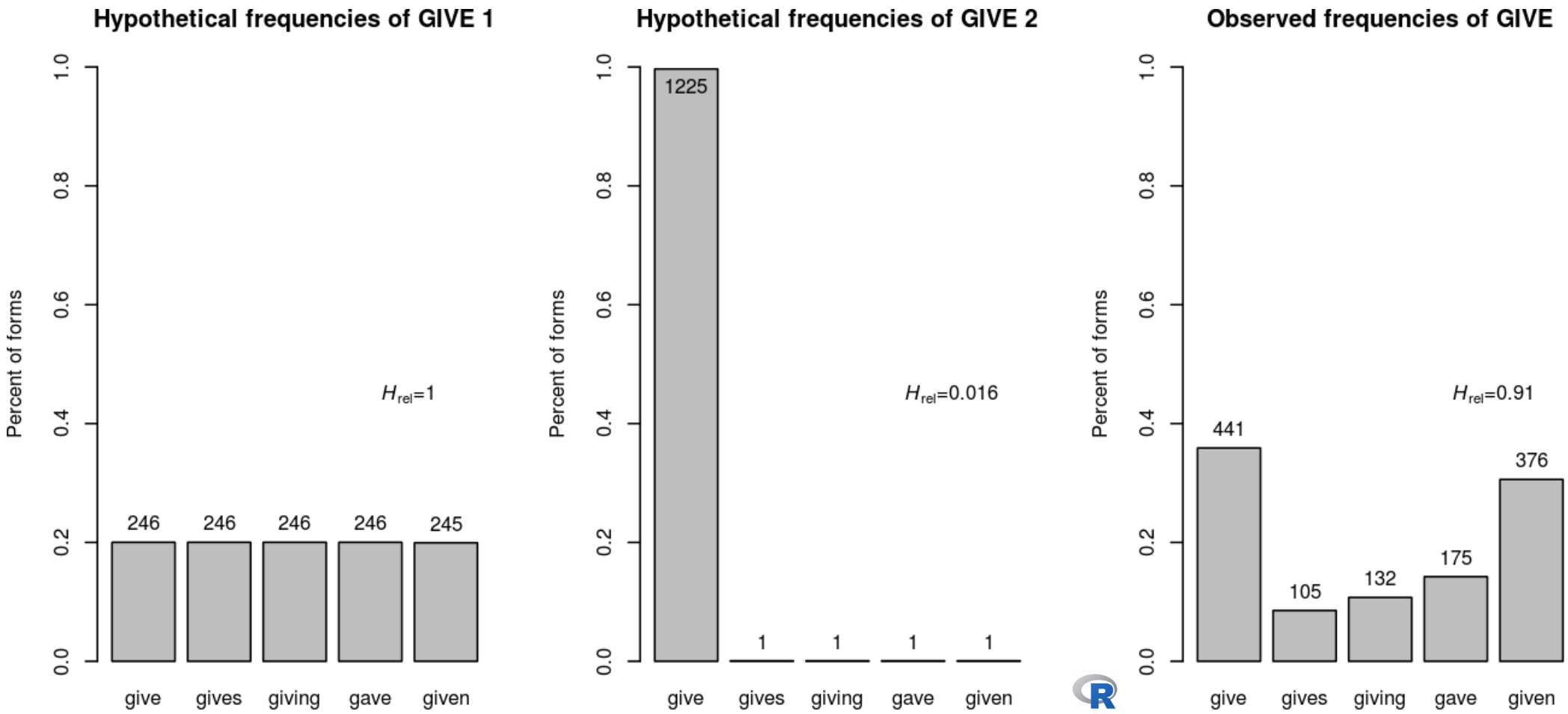
- Ellis (2011) has emphasized that
  - "it [is] contingency, not temporal pairing, that generated conditioned responding" in classical conditioning"
  - "human learning is […] perfectly calibrated with normative stat. measures of contingency like r, χ^2 & ΔP"
  - "[l]anguage learning can thus be viewed 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"
- Shillcock & McDonald's (2001) *CD* outperforms absolute freq. as predictor of lexical decision time
- Baayen (2010:436)
  - the word frequency effect in the sense of pure repeated exposure accounts for only a small proportion of the variance in lexical decision
  - local syntactic & morphological co-occ. probs are what makes word frequency a powerful predictor for lexical decision latencies

What about type frequency though?

- Similar comments may apply to type frequency
- units can have identical token and token frequencies yet still differ in their entropy $H = -\sum_{i=1}^{n} p(x) \cdot \log_2 p(x) \quad (w/\log_2 0 = 0)$
  - for instance, Goldberg, Casenhiser, & Sethuraman (2004) find that subjects learn novel words with identical type & token frequencies better in the low-*H* condition
  - for instance, Linzen & Jaeger (2015) find that the *H* reduction of potential parse completions is correlated with reading times of sentences involving the DO/SC alternation
  - for instance, Lester & Moscoso del Prado (2017) find that *H*s of syntactic distributions affect response times of Ns in isolation and the ordering in coordinate NPs
  - Lester et al. (2017) find that words occurring in similar distributions of syntactic constructions prime each other: Ns' representations appear to be connected to syntactic structures ≈ how often they occur in them

What about type frequency though?

$$H = -\sum_{i=1}^{n} p(x) \cdot \log_2 p(x) \quad (w/\log_2 0 = 0)$$



On contingency …

- The following are considerations that are relevant to choosing a measure of contingency/association
  - symmetry: is the AM supposed to be symmetric or not?
    - nearly all AMs are: p_{FYE}, LLR, X^2, MI, t, z, odds ratio, …
    - some are not: $p(y|x)$, ΔP, $D_{KL}(P||Q)$ …
  - metric type: +effect – freq. vs +effect +frequency
    - the former: odds ratio, the asymmetric ones from above, …
    - the latter: p_{FYE}, LLR, X^2, …
  - frequency information: token vs token+type frequency
    - the former: all but one
    - the latter: lexical gravity G
  - dispersion: is dispersion information included? so far, virtually never …
- probably best settings in an ideal world:
  - symmetry: no
  - metric type: +effect
  - frequency: token+type
  - dispersion: included (see below)

On context(s) …

- Most corpus-based work explores context by studying/annotating concordances and then compute freqs/percs
- while this allows to see minute details, it may also
  - lead to easy overestimates of the role of frequency
  - obscure bigger distributional trends
- McDonald & Shillcock (2001) suggest that what seem like frequency effects may in fact be epiphenomenal
- they propose the measure Contextual Distinctiveness (CD), D_{KL}(P||Q) (D_{KL}(to posterior||from prior))
  for words, e.g., that is the (asymmetric!) divergence
  - from the probability distributions of words in a corpus
  - to the probability distributions of words in some context/with some function

Words	a	b	c	d	e	Sum
Freqs of words in corpus	10	15	5	10	15	55
Freqs of words around *state*	8	1	2	1	3	15
% of words in corpus	0.1818	0.2727	0.0909	0.1818	0.2727	1
% of words around *state*	0.5333	0.0667	0.1333	0.0667	0.2	1

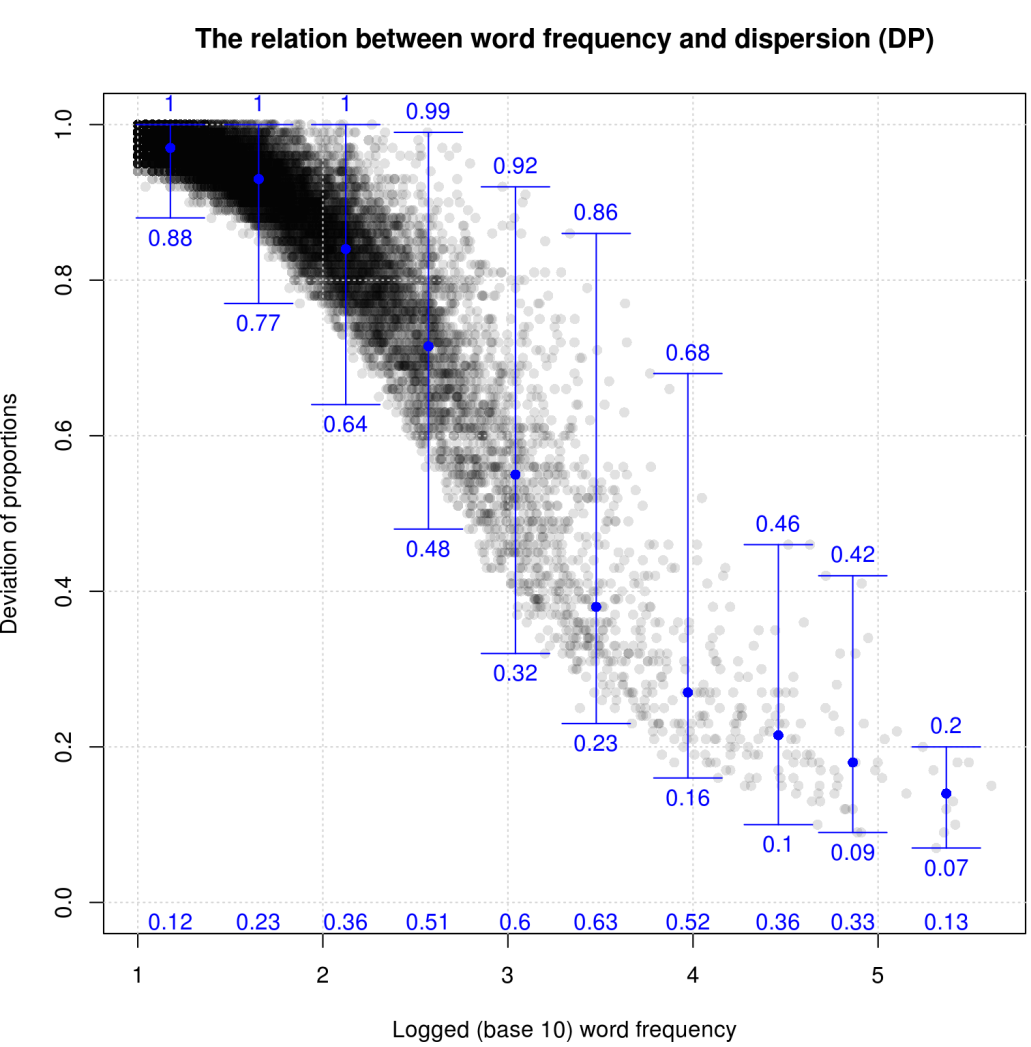
On context(s) …

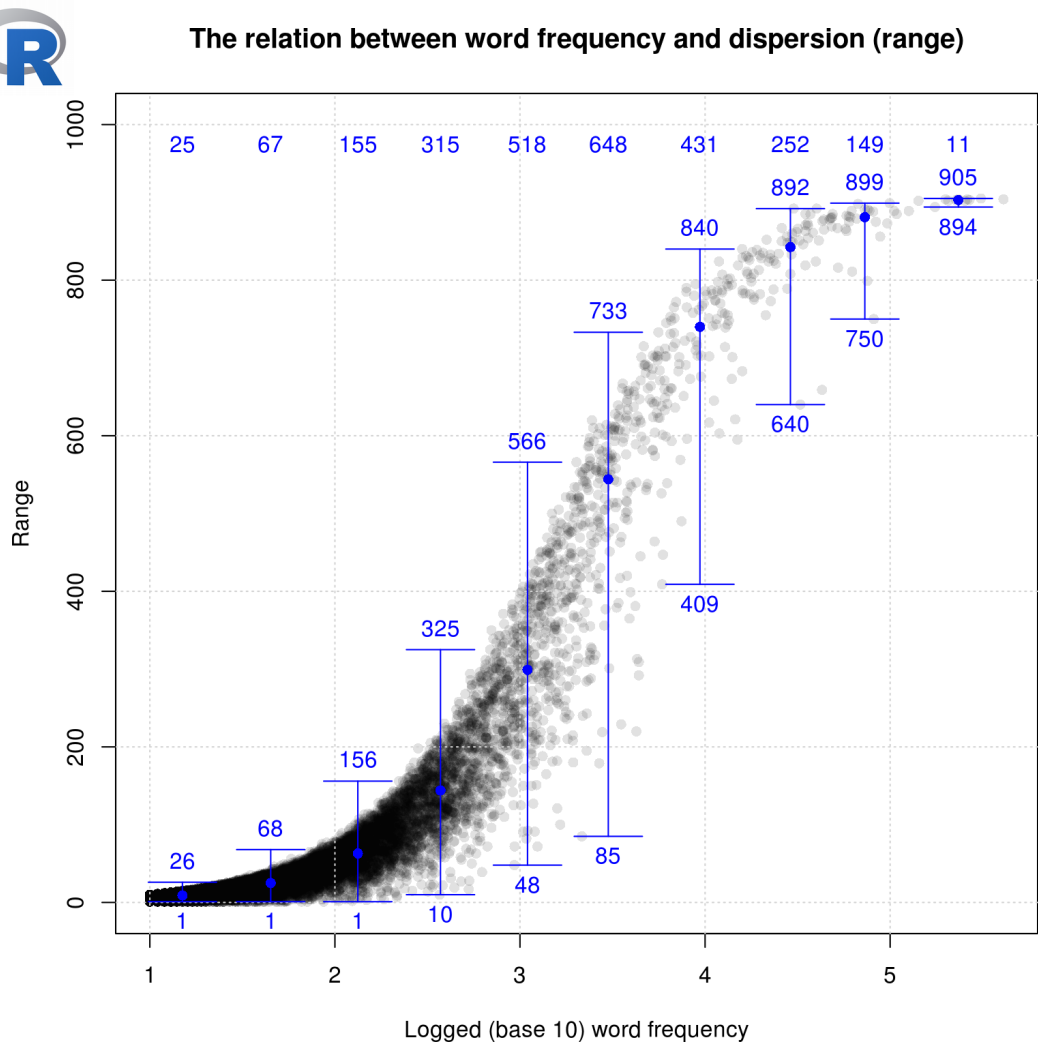
- Most corpus-based work explores context by studying/annotating concordances and then compute freqs/percs
- while this allows to see minute details, it may also
  - lead to easy overestimates of the role of frequency
  - obscure bigger distributional trends
- McDonald & Shillcock (2001) suggest that what seem like frequency effects may in fact be epiphenomenal
- they propose the measure Contextual Distinctiveness (CD), D_{KL}(P||Q) (D_{KL}(to posterior||from prior)) for words, e.g., that is the (asymmetric!) divergence
  - from the probability distributions of words in a corpus
  - to the probability distributions of words in some context/with some function
- crucially, this measure is correlated with freq altho abs freq does not enter into its computation
- it accounts for variance in RTs when word freq & length are controlled for whereas freq does not when length and CD are controlled for

On recency/dispersion

- Let us return to Schmid (2010): "frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention"
- recency (*Contextual Diversity* (??) in Adelman, Brown, & Quesada 2006) is
  - important because of how it relates to acquisition, learning, and forgetting: "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" (ABQ)
  - hardly ever utilized outside of the context of priming
  - either a better or at least an additional measure of how likely it is something is encountered by a speaker, e.g.
    - the words *enormous* and *staining* are equally freq in the Brown corpus but as unequally dispersed as possible
    - BNCspoken: *council* = same freq bin as *nothing*, *try*, *whether*
- in ABQ, range is a better and more unique predictor of lex dec times (see also STG 2010)

On recency/dispersion





There is more to do …

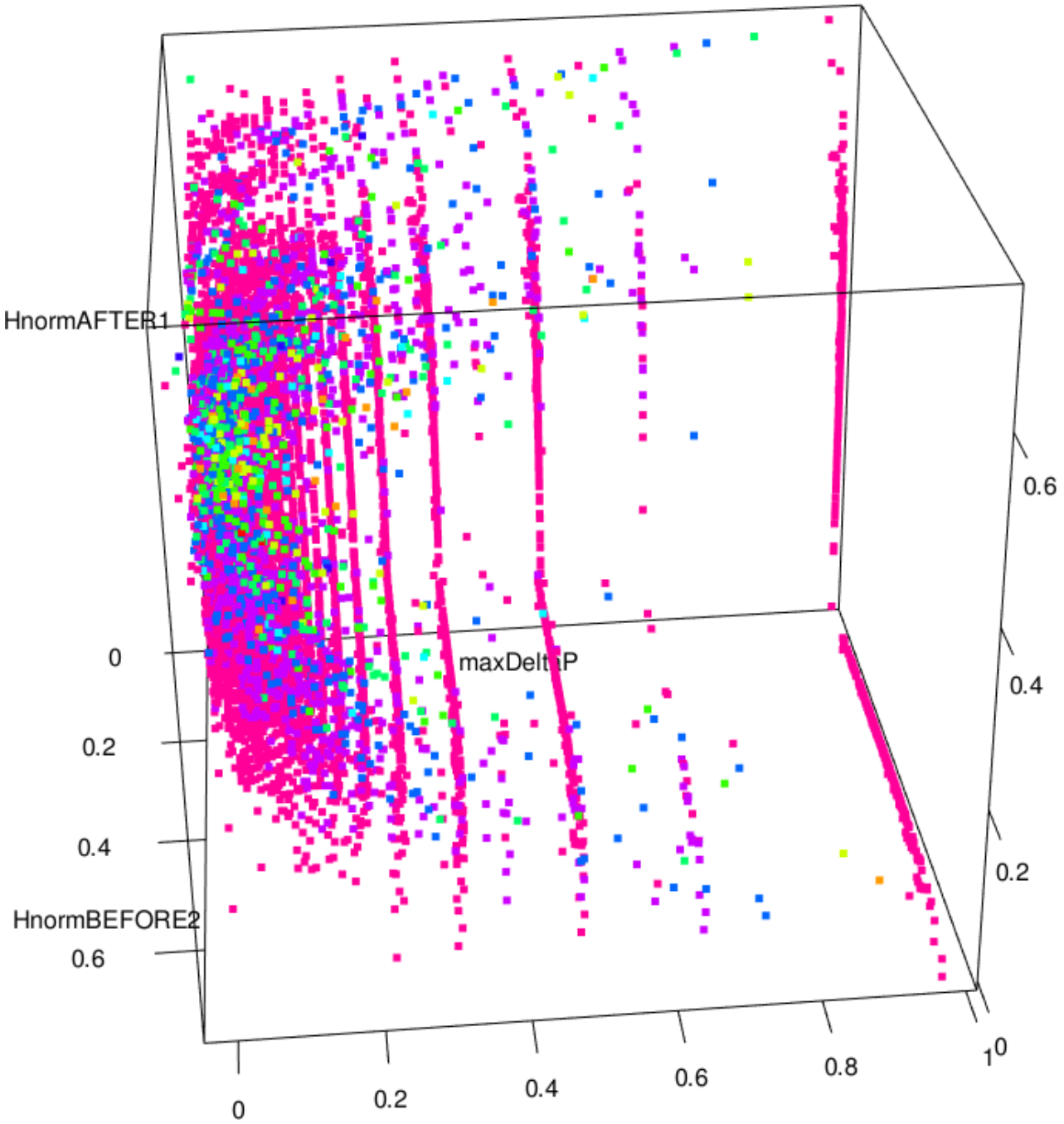
- 1st, there are more cognitive mechanisms that require corpus-linguistic operationalizations
- 2nd, corpus linguistics must look more to what other disciplines do with corpus data, e.g., cognitive and, more importantly, psycholinguistics
- 3rd, linguistics must stop underutilizing corpus data – raw freqs & probs aren't all there is
  - let's not fall into the trap and say 'everything's due to frequency'
  - this is not just number-crunching but also theoretically important
    - if frequency is a causal factor (as a repetition counter), then maybe entrenchment / resting-activation-level accounts are correct
    - if it isn't and it is contingency, entropy, etc., then we probably need different psycholinguistic models
- 4th, let's drop the pessimism before we have explored more and better options

One example of what I have in mind: the identification of multi-word units

- Input
  - a corpus with every file being its own (list) element
  - a number of iterations (e.g., 10000)
  - threshold values for critical variables
- procedure for each iteration
  - generate all *n*-grams in the corpus
  - compute their (normalized?) frequencies and store
  - compute ΔPs for each *n*-gram and store (positive only)
    - left-to-right and right-to-left; maybe conflate (max? sum?)
  - compute *DP* for each *n*-gram and store
  - compute normalized entropies for each *n*-gram
    - probability distribution of units after unit 1
    - probability distribution of units before unit 2
  - pick *n*-gram to merge into new unit
    - tuple? Euclidean distance? weighted Euclidean distance?
  - merge and iterate …

For predictors: frequency, ok, 'what else you got?'

For predictors and responses: things we can do

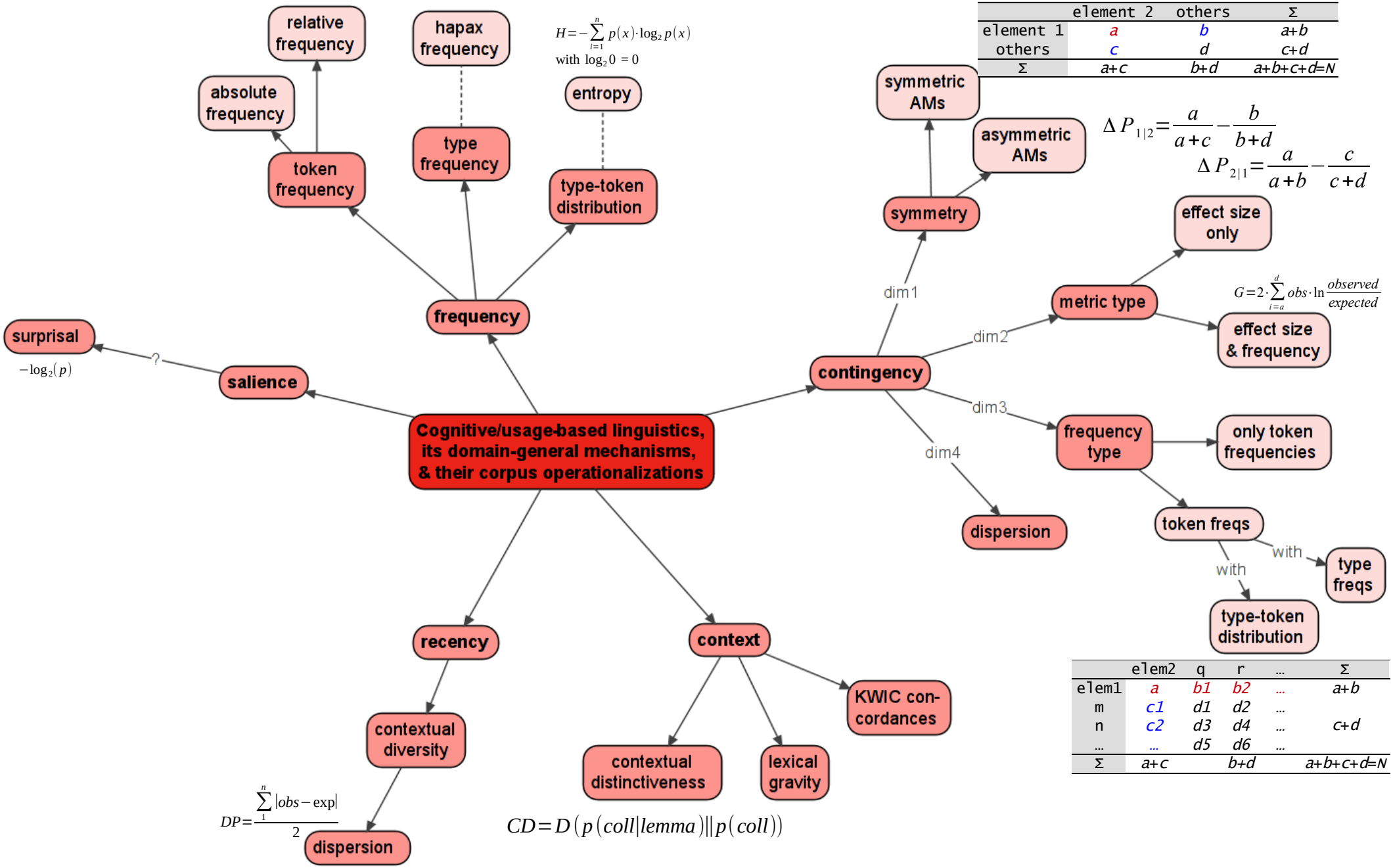


Overall conclusion

- Yes, corpus data are only frequencies
  - only frequencies (in a sense)
  - noisy/messy
  - skewed and Zipfian-distributed
  - full of collinearity, sequential data/autocorrelation …
- but
  - more creative ways of looking at frequencies and their distributions allows us to go way beyond frequency-as-entrenchment claims
  - corpus data are also ecologically more valid in many respects
- that means, we
  - do need very careful statistical control/modeling tho …
  - a kind of multidimensional thinking that most of corpus and cognitive linguistics hasn't adopted yet

Thank you!

http://tinyurl.com/stgries



	element 2	others	Σ
element 1	a	b	a+b
others	c	d	c+d
Σ	a+c	b+d	a+b+c+d=N

	elem2	q	r	…	Σ
elem1	a	b1	b2	…	a+b
m	c1	d1	d2	…	
n	c2	d3	d4	…	c+d
…	…	d5	d6	…	
Σ	a+c		b+d		a+b+c+d=N

On co-occurrence in corpus data: against frequency and towards a more multi-dimensional view

Stefan Th. Gries
University of California, Santa Barbara