On frequency in corpora 1: frequencies vs. association measures

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

Collostructional analysis: its goals, computation, results, and interpretation

- Collostructional analysis (CA) applies association measures (AMs) to lexico-syntactic co-occurrence
  - 1: retrieve all instances of a construction *c*
  - 2: compute AMs for all collexemes *w* in a slot of *c*
  - 3: rank collexemes *w* by their association score
  - 4: explore the top *t* collexemes' functional patterns

	w	¬ *w*	totals
c	a	b	a+b
¬ *c*	c	d	c+d
totals	a+b	b+d	a+b+c+d

- any association measure can be used (S&G 2003:217)
  - by far most frequent approach: $(-)\log_{10} p_{\text{Fisher-Yates exact}}$
    - no distributional assumptions
    - can handle the low-frequency data following from Zipf
    - as a *p*-value, p_{FYE} is correlated with both effect size and sample size, weighing an observed effect higher, when it is found in a larger sample: '$p_{FYE}\ {}^{14}/_{35} > p_{FYE}\ {}^{8}/_{20}$'

The logic of collostructions: (distinctive) collexeme analysis :-|

- Collostructions are but one approach towards co-occurrence data, which is based on a statistical test of bi-directional association in 2×2 tables
  - collexeme analysis quantifies the degree of attraction/repulsion of one construction (a word) to another
  - distinctive collexeme analysis quantifies the degree of attraction/repulsion of one construction (a word) to one of *x* functionally similar constructions
- crucially, the approach
  - serves to rank-order collexemes: the top collexemes reflect the construction's prototypical sense(s)
  - normalizes frequency of occurrence in a construction
  - typically uses $-\log_{10} p_{\text{Fisher-Yates exact}}$ – because it reflects association, significance, and frequency (cf. below)
  - can distinguish attracted and repelled collexemes
  - CA has been applied to data from English, German, Dutch, Swedish, … synchronic and diachronic data, syntactic alternations, priming, second language acquisition, …

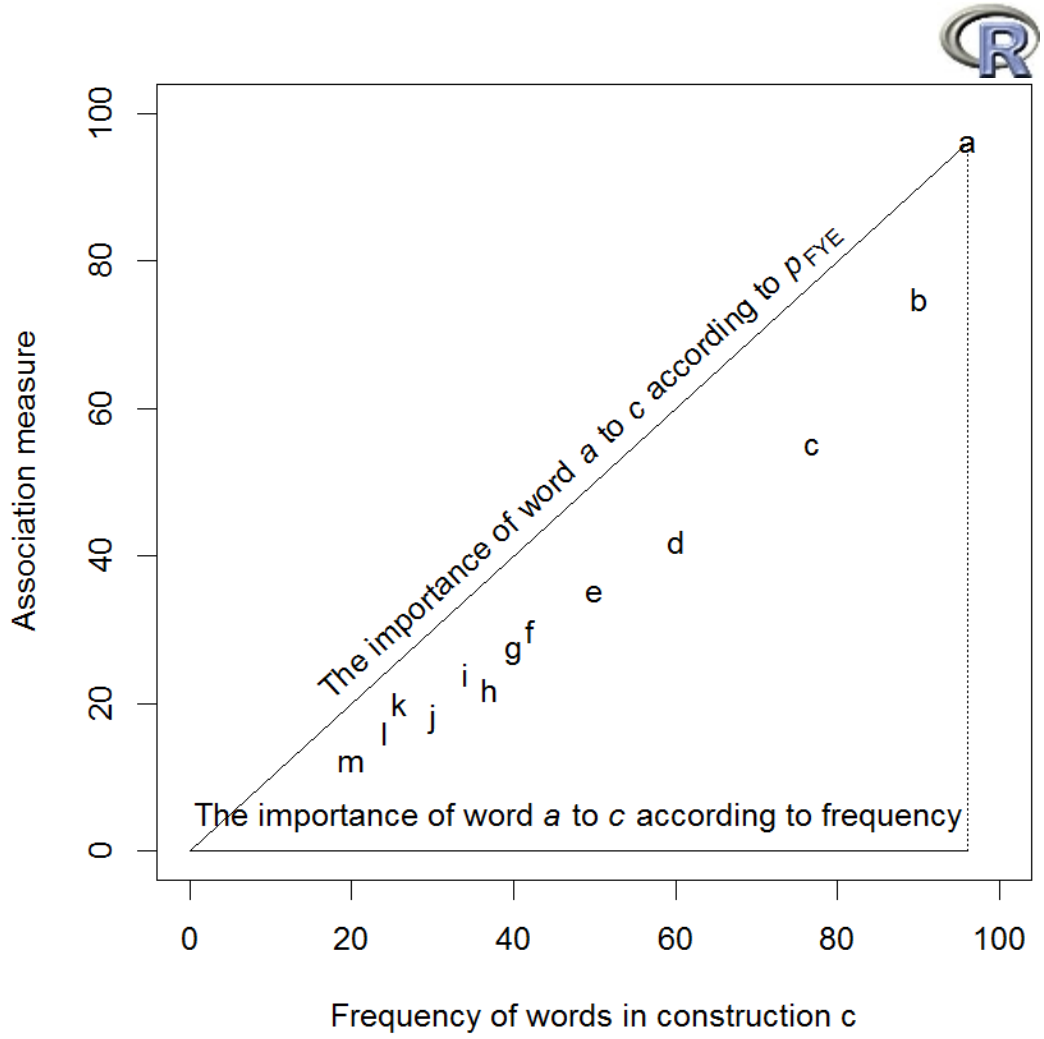
"Problems of collostructional analysis" (Bybee 2010: Section 5.12)

- Bybee criticizes the collostructional approach for what she considers "problems"
  - (1) the fact that a significance test is used because "lexemes do not occur in corpora by pure chance" and "the factors that make a lexeme high frequency in a corpus [may be] the factors that make it a central and defining member of the category" (p. 97) And, how is *d* (¬*w*, ¬*c*) calculated? (p. 98)
  - (2) no "cognitive mechanism" is proposed that "corresponds to their analysis" (p. 100)
  - (3) "Since no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" (p. 98) … "since it works only with numbers and not with meaning." (p. 100)
  - (4) collostructions do not distinguish low-frequency semantically related and semantically unrelated collexemes in the data of Bybee & Eddington (2006)

On (1) and what association measures do

- (1) misses the point completely
  - no corpus linguist would ever say words occur by chance – if they did, not only computing AMs but also counting frequencies would be pointless …
  - the use of a significance test is merely one statistical heuristic:
    - S&G (2003:217) explicitly said *any AM* could be used
    - measures not based on *p*-values have in fact been used
  - what this approach – any AM – does is
    - downgrade words that are highly promiscuous
    - upgrade words that are highly faithful to the construction under investigation
    - thus
      - it makes *regard* – not *see*, *describe*, or *know* – most representative of the *as*-predicative (V NP_{DO} *as* complement)
      - it recognizes that equal frequencies of occurrence may mask preferences (*consider* is equally frequent in *to*- and *ing*-complement constructions, but *to*- is much more freq)

On (1) and what association measures do

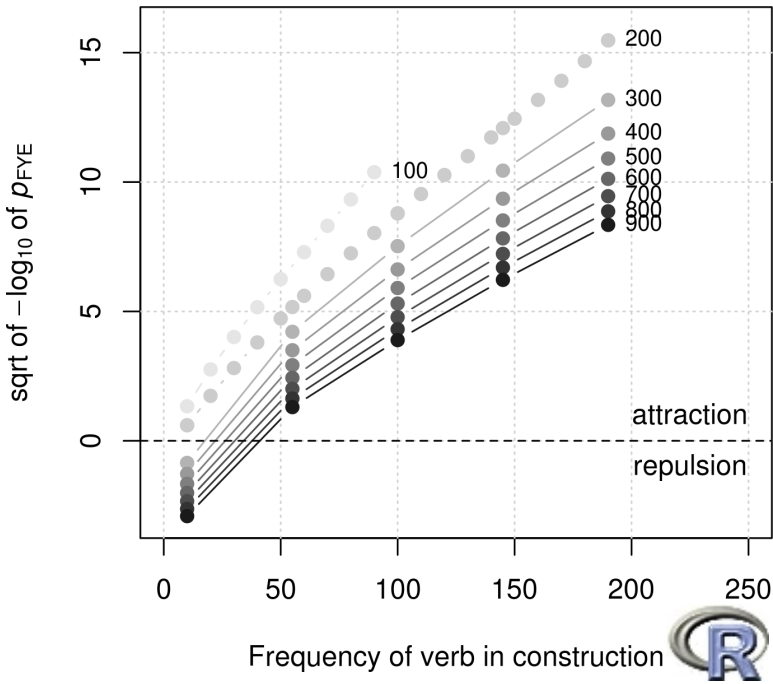


The most wide-spread measure of collostruction strength

The Fisher-Yates exact test (FYE) is

- statistically more appropriate than many other measures which make unwarranted distributional assumptions
- a significance test but unlike many other measures
  - it actually incorporates frequencies to make the measure more compatible with frequency/entrenchment-based accounts (cf. Stefanowitsch & Gries 2003)
  - with increasing frequency of co-occurrence, the collexeme strength goes up more (both is what Bybee would want!)

- statistics need not be simple to provide accurate representations of distributional properties of language (Stefanowitsch 2012)
  - Wiechmann's (2008) comparison ranks FYE second (after *MinSem*, which is theoretically problematic)



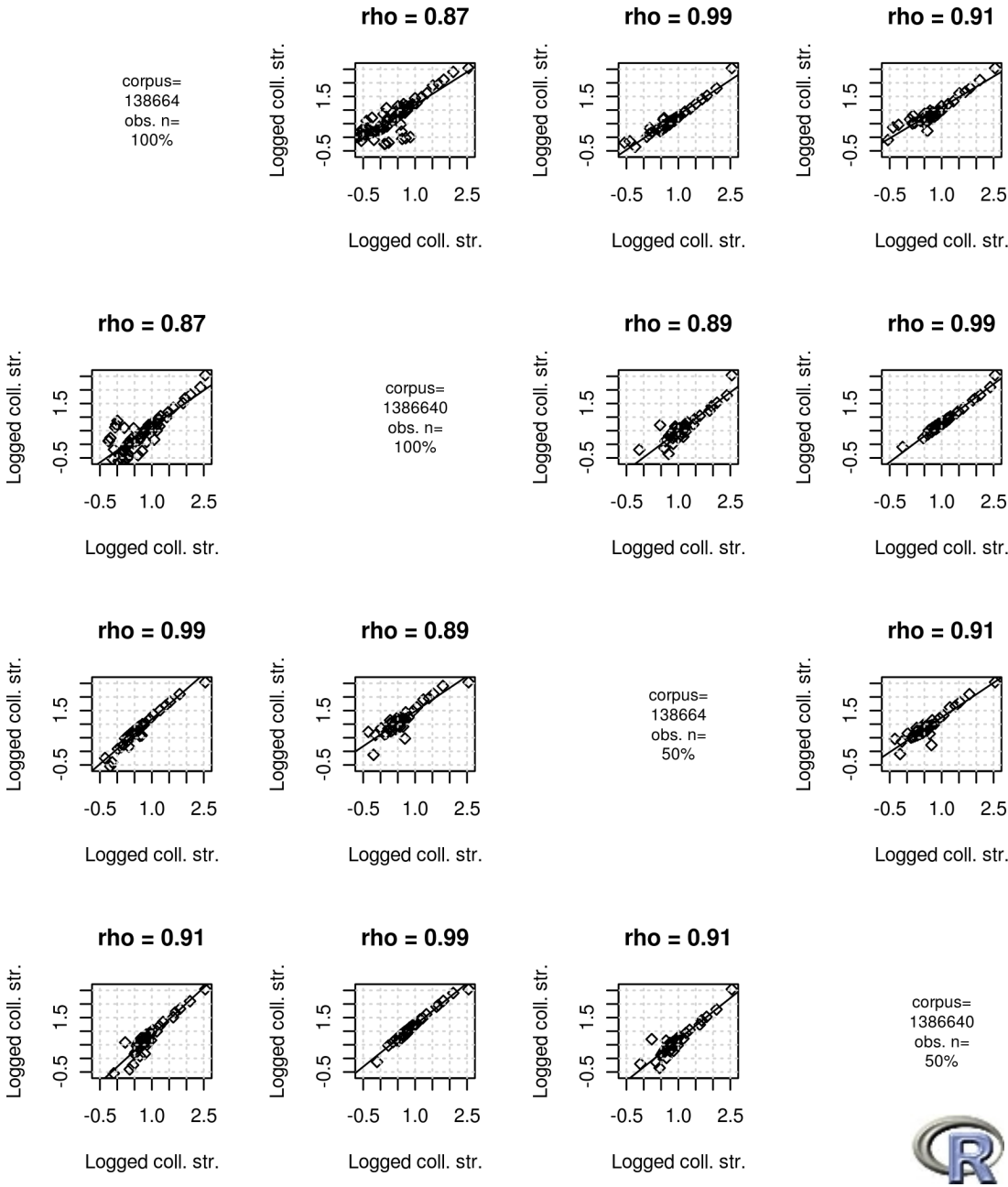
Which ranking of ditransitive verbs do you prefer?

Verbs	Fisher-Yates exact	Verbs	Log odds	Verbs	Mutual Information
give	308.25	*give*	2.20	*accord*	6.07
tell	126.80	*accord*	2.13	*award*	5.87
send	67.14	*award*	2.02	*give*	5.73
offer	48.48	*allocate*	1.73	*allocate*	5.26
show	32.65	*profit*	1.65	*profit*	5.07
cost	21.95	*owe*	1.63	*owe*	5.01
teach	15.36	*lend*	1.59	*lend*	4.92
award	10.86	*offer*	1.58	*offer*	4.86
allow	9.95	*cost*	1.52	*cost*	4.72
lend	8.55	*send*	1.51	*grant*	4.69
deny	8.35	*grant*	1.50	*send*	4.63

On (1) and what association measures do

- (1) misses the point completely
  - association measures actually have a variety of characteristics that should appeal to cognitive linguists
  - association measures typically result in frequency-and-their-effects relations that are non-linear, just as frequency effects (Tryk 1968), learning (Anderson 1982), forgetting curves, priming decay (Gries 2005, Szmrec-sanyi 2005), etc.
  - association measures reflect
    - that "[r]aw frequency of occurrence is less important than the contingency between cue and interpretation", and
    - the notion that "Contingency, and its associated aspects of predictive value, information gain, and statistical association, have been at the core of learning theory ever since." (Ellis & Ferreira-Junior 2009)
  - *d* is computed as it is for all AMs: on the basis of a reasonable approximation for a corpus; simulations show that CA rankings are very robust

On (1) and what AMs do



So, what if we pit frequency against association strength?

- Gries, Hampe, & Schönefeld (2005): *as*-predicative
  - *regard* (80) is returned as the central verb (*see* (111) and *know* (79) are more/as frequent but more general than the top-ranked *regard* and *describe*)
  - in a sentence completion task, p_{FYE} outperforms conditional prob. and freq. as predictors of completions
- Ellis & Simpson-Vlach (2009): *MI* outperfoms frequency as a predictor of formulaicity ratings
- Gries, Hampe, & Schönefeld (2010): in a self-paced reading times experiment, p_{FYE} (p=0.065) outperforms freq (p=0.293) with an effect size 3 times as high
- Colleman & Bernolet (2012) find seemingly erratic verb-specific preferences in the Dutch dative alternation that fall into place when explored with association strength

$-\log_{10} p_{FYE}$ yields good results, but can we do better than that?

frequency of *(!)x* and *(!)y* in some corpus		2 *y*	*!y*	Totals
	1: *x*	a	b	a+b
	1: *!x*	c	d	c+d

- Nearly all AMs are bidirectional – but (associative) learning is not
- thus, a uni-directional AM may be more useful

$$\Delta P_{2|1} = p(word_2|word_1 = present) - p(word_2|word_1 = absent) = \frac{a}{a+b} - \frac{c}{c+d}$$

$$\Delta P_{1|2} = p(word_1|word_2 = present) - p(word_1|word_2 = absent) = \frac{a}{a+c} - \frac{b}{b+d}$$

- consider *of course* in the spoken part of the BNC
  - MI=5.41, t=476.97, G^2=36,693.85, $p_{FYE}<10^{-320}$
  - ΔP(*course*|*of*)=0.032 and ΔP(*of*|*course*)=0.697
- discrepancies like this are rather common
- a reanalysis of Gries, Hampe, & Schönefeld (2005) with ΔP (construction|verb) shows that ΔP is a significant predictor of subjects' completions

On (2), cognitive mechanisms that may be involved / have been invoked

- In addition to the above comments, collostructional studies by Stefanowitsch and myself as well as others and other work by myself at least have embraced particular cognitive mechanisms
  - Stefanowitsch & Gries (2003:237)
    - discuss implications of collostructions to psycholinguistic studies of language acquisition (quoting Goldberg 1999)
    - discuss the relation of collostruction strength to raw token frequency in that connection
    - draw an explicit connection to the notion of cue validity, which is a well-established notion in the Competition Model, Goldberg's work, Ellis's work, …
  - Stefanowitsch & Gries (2003:239, n. 6) discuss the relation of collostruction strength to entrenchment in a way that is in fact sympathetic to Bybee (cf. also Stefanowitsch 2008 for more discussion)
  - Wiechmann (2008:257) discusses association strengths (incl. collostructions) and their relation to cue validity and cue strength (cf. also MacWhinney p.c.)

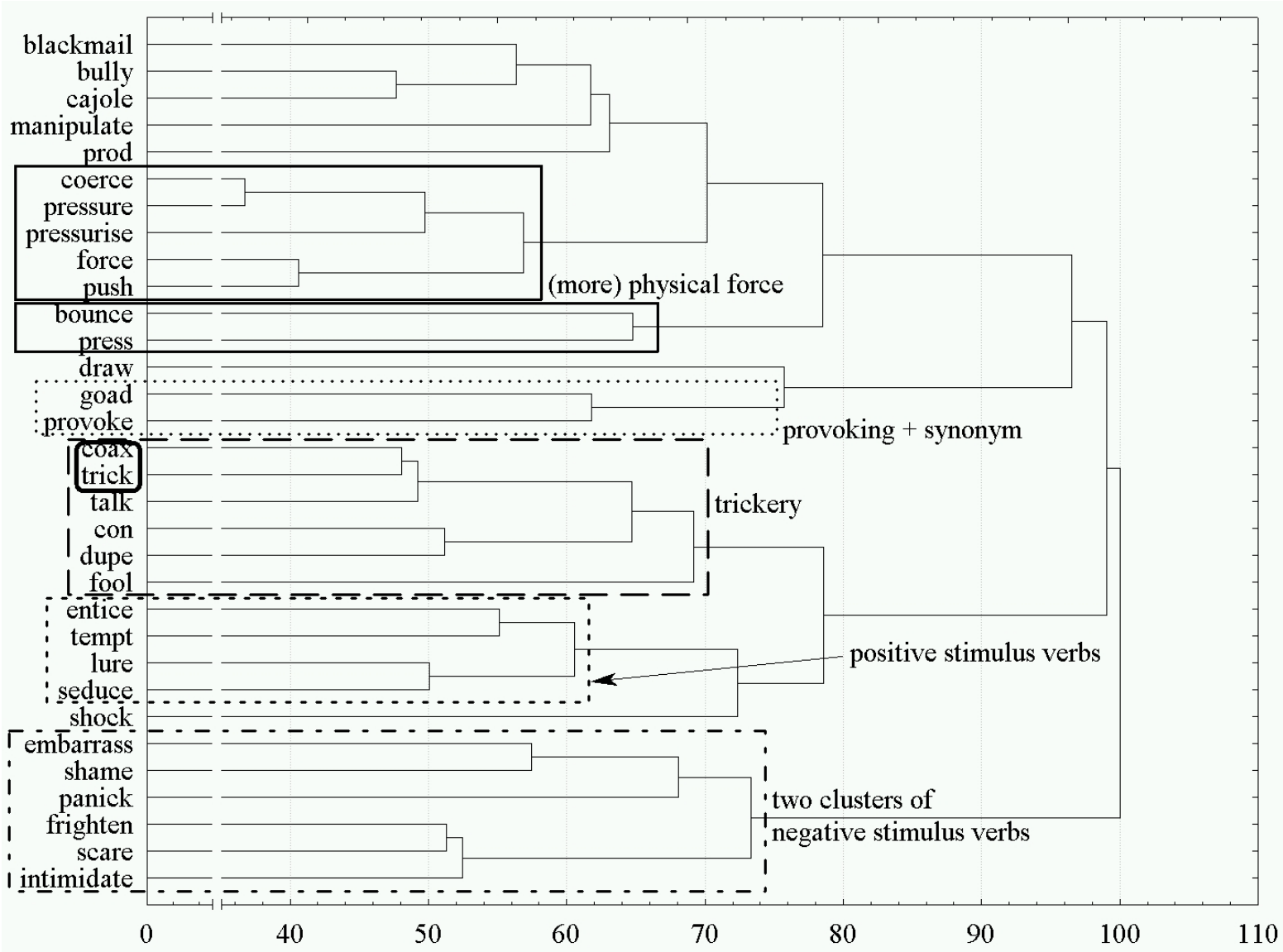
On (3) and on semantic analysis in general

- In general, the notion that semantics cannot emerge from analysis X if no semantics is entered into it is flawed (and I know that Joan knows that)
  - there is a whole body of work in computational (psycho)linguistics, or distributional linguistics, where purely number-based distributional analyses reveal functionally highly coherent clusters
    - Reddington, Chater, & Finch (1998)
      - input: co-occurrence frequencies of 150 bigrams before 1000 target words
      - output: cluster analysis that returns parts of speech
    - Mintz, Newport, & Bever (2002): similar analysis but different definition of context (boundaries at function words)
    - a lot of work in information retrieval, document summarization, latent semantic analysis, etc.

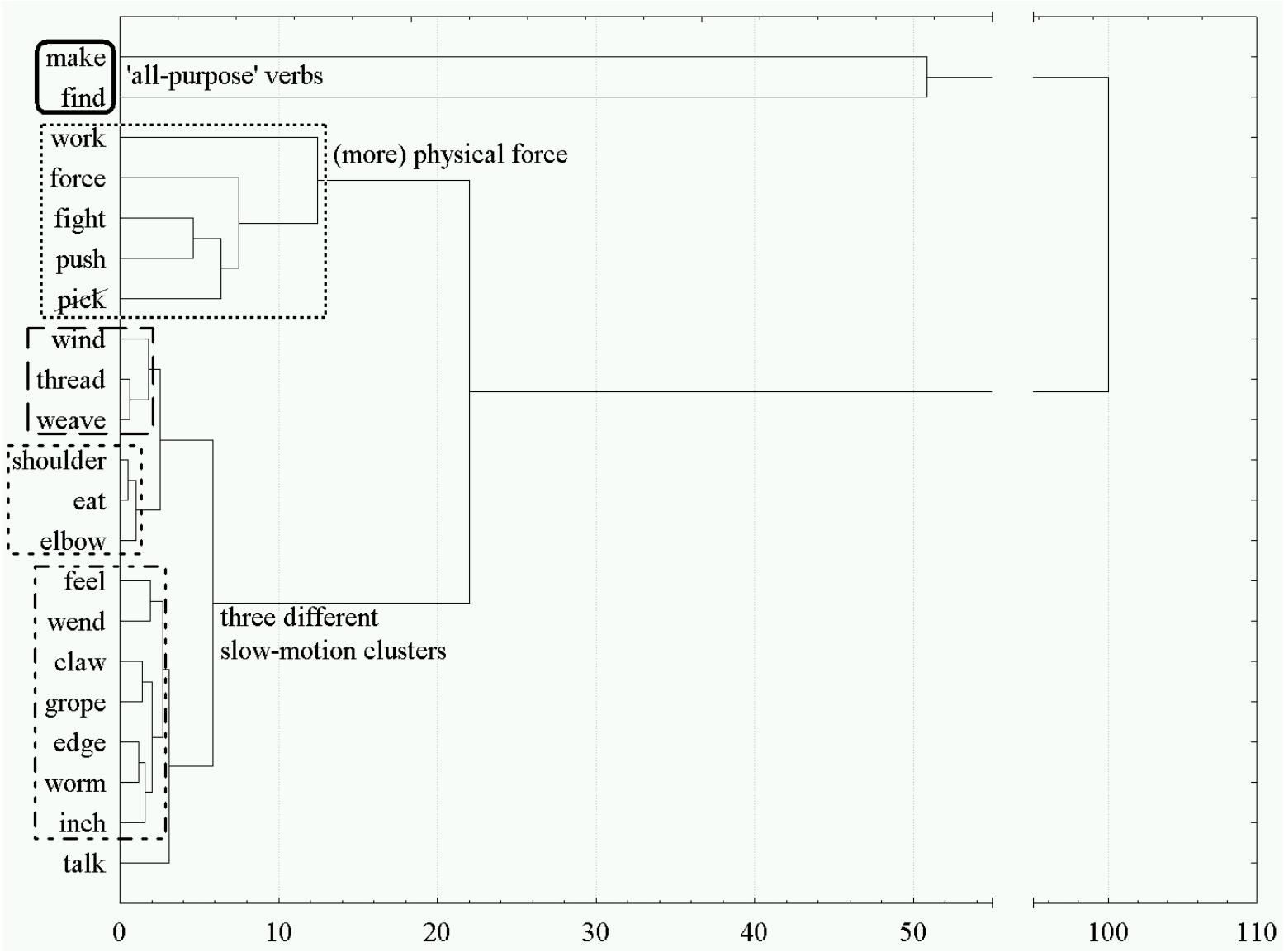
On (3) and semantic analysis in collostructional analysis

- In particular, in collostructional analysis semantic analysis follows the statistics, but it is false to assume that no semantic analysis can emerge from it
  - there are many studies now on very many different constructions and 'alternations' that have shown this, e.g. Stefanowitsch & Gries's (2003) on the ditransitive
    - the verb returned as prototypical/central is *give*
    - the next verbs instantiate the senses of Goldberg's (1992, 1995) polysemy analysis of the ditransitive: satisfaction conditions (*offer*, *owe*, *promise*), enablement (*allow*), cause non-transfer (*deny*), cause future transfer (*grant*), intention to receive (*earn*), communication as transfer (*tell*, *teach*), perceiving as receiving (*show*), … (cf. above)
  - Gries & Stefanowitsch's (2010) cluster
    - verbs in the *into*-causative based on the *ing*-verbs
    - verbs in the *way*-construction based on the prepositions

V_{action} $NP_{DirOb\ Patient}$ *into Ving*



V POSS *way* [$_{PP}$ PREP NP]



On (4) and (the perceived lack of) discriminative power

- Bybee: low-frequency collexemes do not distinguish between semantically related and semantically unrelated items (referring to Bybee & Eddington 2006)
- how does she 'show' that?

Step	CA	Bybee's 'CA'
1	retrieve all instances of *c*	-
2	compute AMs for all collexemes in a slot of *c*	compute AMs for 24 semantically-analyzed adjectives w/ freqs of ≤1
3	rank-order collexemes by the AMs	-
4	explore top *t* collexemes functionally	-

- by ignoring all previously published results
- by ignoring how collexeme analysis is actually done
  - semantic analysis is done after the stats – not before
  - all collexemes are included – not just those with $n \leq 1$
- note, she focuses only on whether collostructions can predict her positive judgments – what about negative judgments? (Gries & Wulff 2009)

On (4) and (the perceived lack of) discriminative power

- What do collostructions output when applied as intended – inspecting words ranked highest based on a measure combining frequency and contingency?
  - S & G (2003) and G & S (2004)
    - ditransitives, dative alternation, *as*-predicatives: as above
    - *into*-causatives: as above, and more comprehensive results than Hunston & Francis's (1999) frequency-based results
  - Ellis & Ferreira-Junior (2009): freq. of learner uptake is predicted by frequency, ΔP, $-\log_{10}(p_{FYE})$ – in fact, $-\log_{10}(p_{FYE})$ outperforms freq in $^{2}/_{3}$ constructions
  - Colleman & Bernolet (2012): verb-specific preferences (Dutch dative alternation) that appear erratic from a raw frequency perspective turn out to be systematic once distinctive collexemes are used
  - Gregory et al. (1999): MI is correlated significantly with 3 pronunciation effects (frequency with 2)

Some additional difficulties I have, and where that leads us

- Bybee criticizes collostructions for ignoring low-frequency collexemes, which may reflect productivity
  - but, again, that's not what collostructions are supposed to do: if one wants a productivity measure, one can count the number of hapaxes directly from the clx output (but note that only clx can rank low-freq words highly)
- Bybee argues that it is not pertinent how often a lexeme does *not* occur in a construction, but
  - we have seen the literature on learning says otherwise
  - CA just uses one more type of information than Bybee herself

(9)

SUBJECT [DRIVE] { *me* / *you* / *him* / *her* / *the producer* } { *mad* / *crazy* / *up the wall* }

CA just normalizes the frequency of *crazy*/*mad*/etc. against their overall frequencies

Some additional difficulties I have, and where that leads us

- Her example (9) simplifies but is instructive!
- ultimately, the situation is more complicated than Bybee AND collostructions assume: we really need
  - the type frequencies and token frequencies in all slots
  - the dispersion of the tokens
  - the distribution/entropy of the token frequencies
  - the frequencies and association strengths of elements to the slot and other slots
  - all of that sense-specific (cf. Roland & Jurafsky 2002; Colleman & Bernolet 2012) and in a probabilistic network of constructions (cf. Roland, Dick, & Elman 2007)

To conclude, this is what we agree on and what our agenda looks like …

- All the above must not distract from the fact that I strongly agree with most of what Bybee (and of course many others) say(s): of course,
  - usage-/exemplar-based approaches are on the right track
  - domain-general mechanisms such as analogy/similarity, chunking, and frequencies of exposure/processing are key
- but the devil lies in the detail (duh!) and we must be very careful to
  - not throw too many babies out with the bathwater and prematurely choose, or abandon, particular kinds of data and methods
  - not explore the true nature/structure of our data especially since … it doesn't get much messier / more chaotic than linguistic data :-|

Thank you!

http://tinyurl.com/stgries