No more collostructions, please! (At least not the traditional type …)

Stefan Th. Gries
UC Santa Barbara & JLU Giessen
https://www.stgries.info

Corpora are distributional data

- Corpus linguistics is a distributional discipline: everything is about
  - the frequency of occurrence of something
  - the frequency of co-occurrence of something with something else
  - both of the above somewhere
- thus, we often report
  - frequencies
  - association measures (AMs)
  - dispersion measures (DMs, way too rarely)
- what are those statistics used for, or grounded in?
- what motivates the use of such statistics?

The distributional hypothesis

- Central to much of corpus linguistics: different versions of the distributional hypothesis
  - "You shall know a word by the company it keeps" (Firth (1957:11)
  - "[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution." (Harris 1970:585f.)
- see also Goldberg's (1995:50) semantic coherence principle, according to which constructions attract lexical items that are compatible with the semantic specifications of particular slots

In the context of Construction Grammar …

- … **what** is typically done?
- the two by far most frequent applications are
  - simple collexeme analysis
    - the association of 'words' to a slot in a construction
    - probably most often, the association of a verb to a slot in an argument structure construction
  - distinctive collexeme analysis
    - the association of 'words' to one of several functionally similar constructions
    - probably most often, the association of a verb to a slot in two alternating constructions
- note: there's a reason for the quotes around *words*

In the context of Construction Grammar …

- … **how** is this typically done?
- on the basis of 2×2-tables
- for simple collexeme analysis

	construction	other	Sum
$word_x$	A	b	A+B
other	C	d	c+d
Sum	A+C	b+d	A+B+C+D

- for distinctive collexeme analysis

	$construction_1$	$construction_2$	Sum
$word_x$	A	B	a+b
other	C	d	c+d
Sum	A+C	B+D	a+b+c+d

- quantify the association of rows & columns to each other with typically, $p_{\text{Fisher-Yates exact test}}$ or G^2 (signed)
- sort $word_{1-n}$ by the AM

Very wide applicability …

- These methods have been applied in a very wide variety of contexts
  - general corpus-linguistic studies
  - psycholinguistic applications
  - second/foreign language acquisition/learning
  - diachronic studies of language change
  - experimental studies w/ collostr. results as predictors
  - experimental studies w/ collostr. results as controls
  - registers
  - languages (eg English, German, Dutch, Italian, Standard Arabic, Mandarin Chinese, …)
- and these methods have been hugely successful
- but …

Very wide applicability … but they are >20 years old!

- … and (corpus) linguistics has become so much more quantitative, and quantitatively better during that time!
- thus, what are all the ways in which traditional collexeme analyses are often less than ideal?
  - 1: they use, i.e. quantify things with, one measure only that, typically, combines frequency and association
  - 2: they use, i.e. quantify things with, one measure only that, typically, combines both directions of association
  - 3: they do not include dispersion
  - 4: one variable only: this is essentially monofactorial
  - 5: are actually not really following the maybe most central tenet of Construction Grammar and 'disregard' aspects of form AND meaning for all constructions involved
  - 6: they do not provide any indications of uncertainty or variability (or 'significance')
- thus, what might – should! – a reviewer say?

Conflating frequency and association

- In one and the same corpus, which of the following verb forms is more attracted to the ditransitive?

	ditransitive	other	Sum
assured	10	2	12
other	1810	137085	138895
Sum	1820	137087	138907

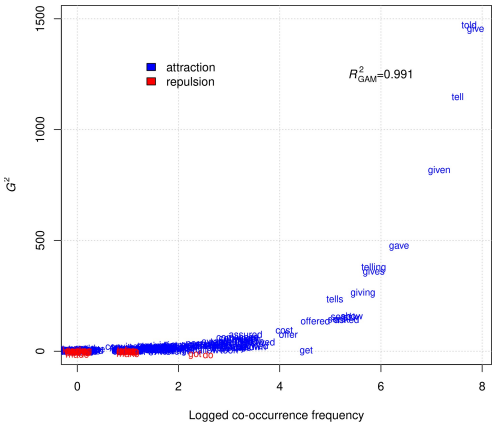
	ditransitive	other	Sum
offer	18	55	73
other	1802	137032	138834
Sum	1820	137087	138907

- In which of these two corpora is *give* more strongly attracted to the ditransitive?

	ditransitive	other	Sum
give	234	209	443
other	1586	136878	138464
Sum	1820	137087	138907

	ditransitive	other	Sum
give	2340	2090	4430
other	15860	1368780	1384640
Sum	18200	1370870	1389070

Conflating frequency and association



Conflating directions of attraction

- An example from collocations: would you want to consider all of these collocations 'the same'?
  - *of course*, *for instance*, *for example*, *old fashioned*, *status quo*, *pot pourri*, *coup d'etat*, *grand prix*, *de facto*
  - *instead of*, *according to*, *owing to*, *apart from*, *volte face*, *habeas corpus*, *ipso facto*
- look at *in vain* in a spoken corpus:
  - observed freq: 3 vs. expected freq: ≈0.1092748
  - G^2≈15.313 & *PMI*≈4.779
- look at *pursuant to* in a spoken corpus:
  - observed freq: 2 vs. expected freq: ≈0.044897976
  - G^2≈15.186 & *PMI*≈5.477
- but if we look at directional measures like the *KLD*:
  - *in→vain*: 0.0000715 vs. *in←vain*: 1.38
  - *pursuant→to*: 5.466 vs. *pursuant←to*: 0.0000348

KLD – what is that?

FREQ	ditrans	other	Sum
assured	10	2	12
other	1810	137085	138895
Sum	1820	137087	138907

ROWPERC	ditrans	other	Sum
assured	0.83333	0.16667	1
other			
Sum	0.01310	0.98690	1

posterior div by prior	63.6021	0.1689
log of that	5.9910	-2.5659
log x posterior	4.9925	-0.4277
Sum	4.56485	

assured → ditrans

- The KL-divergence quantifies the divergence of
  - a prior distribution, eg the freqs of constructions
  - a posterior distribution, eg the freq of a cx w/ some w
- *KLD* is related to G^2, but
  - can be computed directionally
  - is less correlated w/ frequency, which is good: it provides new information rather than replicates old info
- the *KLD* can be
  - normalized (odds-to-probabilities transf. from log reg)
  - signed (so >0 = attraction and <0 = repulsion)
  - stretched (so attraction covers all of [0, 1])
  - cleaned with regard to frequency (see below)

KLD – what is that?

FREQ	ditrans	other	Sum
assured	10	2	12
other	1810	137085	138895
Sum	1820	137087	138907

COLPERC	ditrans	other	Sum
assured	0.00549		0.00009
other	0.99451		0.99991
Sum			

posterior div by prior	63.6021	0.9946
log of that	5.9910	-0.0078
log x posterior	0.0329	-0.0078
Sum	0.02514	

assured ← ditrans

- The KL-divergence quantifies the divergence of
  - a prior distribution, eg the freqs of constructions
  - a posterior distribution, eg the freq of a cx w/ some w
- *KLD* is related to G^2, but
  - can be computed directionally
  - is less correlated w/ frequency, which is good: it provides new information rather than replicates old info
- the *KLD* can be
  - normalized (odds-to-probabilities transf. from log reg)
  - signed (so >0 = attraction and <0 = repulsion)
  - stretched (so attraction covers all of [0, 1])
  - cleaned with regard to frequency (see below)

Neglecting/ignoring dispersion

- Stefanowitsch & Gries (2003): verbs' attraction to the imperative in the ICE-GB: *let*, *see*, *look*, *listen*, *worry*, verb$_1$, *remember*, *check*, verb$_2$, *try*, *hang on*, …
- makes perfect sense, but what are verb$_1$ and verb$_2$?
  - verb$_1$: *fold*:
    - 6th strongest with 16 occurrences in the imperative
    - *fold*->imperative: *KLD*=2.095 & *fold*<-imperative: *KLD*=0.031
    - -log$_{10}$ p_{FYE} of 21.03
  - verb$_2$: *process*:
    - 9th strongest with 15 occurrences in the imperative
    - *proc*->imperative: *KLD*=3.301 & *proc*<-imperative: *KLD*=0.073
    - -log$_{10}$ p_{FYE} of 16.68
- in how many files of the ICE-GB are these found? $^{1}/_{500}$
  - KLD_{hofreq} for *fold*=0.9; KLD_{hofreq} for *process*=0.91
- Gries (2024) on ditransitives: *tell* is
  - about as frequent in the ditransitive as *send*
  - more strongly attracted to the ditransitive than *send*
  - less well dispersed with the ditransitive than *send*

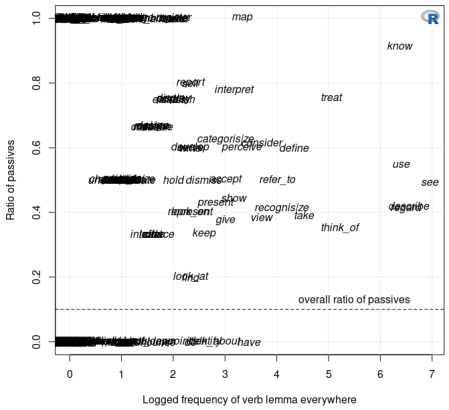
Collostructions are monofactorial, given that …

- … they are computed from 2×2 tables cross-tabulating a construction as a function of a lexical predictor
- … they can be re-written as a logistic regression: glmer(CONSTRUCTION ~ 1 + (1|WORD))
- but it is statistically obvious that a constructional choice will not be due only to lexical items
- but it is linguistically/theoretically obvious that constructions involve more information than that:
  - "On a usage-based model then, constructions are simply conventionalized chunks of linguistic knowledge... From this, it follows that the storage and organization of grammatical knowledge is dependent on, and can change according to, patterns of use" (Patten 2014)

Collostructions are monofactorial

- But it is linguistically/theoretically obvious that constructions involve more information than that:
  - Fillmore (1988:36f.) distinguishes between internal and external properties of constructions:
    - internal: formal, semantic, and symbolic structures
    - external: recurring contextual patterns
    - thus, structural knowledge of constructions is not the only kind of knowledge in speakers. Knowledge of contexts of use is also bona fide constructional knowledge.
    - In a usage-based perspective, external properties range from co-textual ones (including colligational & collostructional) to situational, social, and intertextual ones
- early example: Gries, Hampe, & Schönefeld (2005): verbs are attracted to the *as*-predicative, but that construction is in turn strongly attracted to the passive construction

Collostructions are monofactorial



Many collostructional studies actually undermine their CxG background

- They do usually *not* quantify the attraction of a 'more grammatical construction' and a lexical construction ("it's constructions all the way down")
- they quantify the attraction of a 'more grammatical construction' with letter sequences! – why?
- because they treat a sequence of letters as a construction when a sequence of letters is only the form side of a construction but not also its function
- let me ask you this:
  - in a distinctive collexeme analysis of the dative alternation – ditransitives (*He gave him the book*) vs. prepositional datives (*He gave the book to him*) – *play* is attracted to the prepositional dative – how/why?!
  - in a distinctive collexeme analysis of the verb-particle alternation – particles before DO (*He picked up the book*) vs. particle after DO (*He picked the book up*) –
    - *carry out* prefers particle before DO
    - *pick up* and *put down* have no strong preference – how/why?

How certain, robust, variable are the results?

- There is a notable overall increase of statistical sophistication in linguistics
- more significance testing, more effect sizes, more confidence intervals, more Bayesian methods, ...
- but collostructional applications are lagging behind
  - some measure 'are' (essentially) *p*-values, ok, but
    - they are usually not corrected for post-hoc testing (Gries 2005, *CLLT*)
    - they arguably run into problems with the random-sampling assumption of null hypothesis significance testing
  - some measures are effect sizes, but
  - we do not get (equivalents of) confidence intervals, which obfuscates how much trust we can place in the ranked lists we usually interpret

If you really submit a traditional collexeme analysis in 2026, …

- … you should get maximally a *revise and resubmit* ;-)
- reviewers should say
  - tupleize
    - measure frequency
    - measure association without frequency
    - measure association directionally
    - measure dispersion without frequency
  - add other variables
  - take constructions and their senses more seriously
  - add uncertainty assessments
- let's look at what this can add to your work

"The authors should tupleize to keep dimensions of information separate"

- "They should not compute
  - one measure that conflates a lot of frequency and association (p_{FYE}, G^2, chi-squared, *t*, *z*, …)
  - one measure that is 'only' (similar to) an effect size (*PMI*, (log) *OR*, …)
- "they should determine/compute
  - frequencies of (co-)occurrence
  - one measure of association for each direction, but make that measure
    - one that distinguishes attraction from repulsion
    - one works the same in each direction
    - one that is not/little affected by frequency
      - in one case study, the most widely-used AM – G^2 – is R^2_{GAM}=0.9465 correlated with co-occurrence frequency
      - in another (see above), G^2 is R^2_{GAM}=0.991 correlated with co-occurrence frequency
- "the authors might use Gries's (2024) suggestion and
  - compute the KL-divergence in both directions
  - use the version that 'controls' for frequency

Why one needs to control for frequency

LOW	C=*ditrans*	C=other	Sum
V=*assured*	0.1	12.1	12.2
V=other	1820.1	137075.1	138895.2
Sum	1820.2	137087.2	138907.4

-0.6769

OBS	C=*ditrans*	C=other	Sum
V=*assured*	10	2	12
V=other	1810	137085	138895
Sum	1820	137087	138907

5.9910

UPP	C=*ditrans*	C=other	Sum
V=*assured*	12	0	12
V=other	1808	137087	138895
Sum	1820	137087	138907

6.2540

LOW	C=*ditrans*	C=other	Sum
V=*offer*	0.1	73.1	73.2
V=other	1820.1	137014.1	138834.2
Sum	1820.2	137087.2	138907.4

-3.2618

OBS	C=*ditrans*	C=other	Sum
V=*offer*	18	55	73
V=other	1802	137032	138834
Sum	1820	137087	138907

4.2341

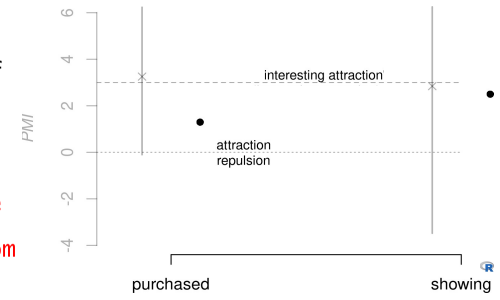
UPP	C=*ditrans*	C=other	Sum
V=*offer*	73	0	73
V=other	1747	137087	138834
Sum	1820	137087	138907

6.2540

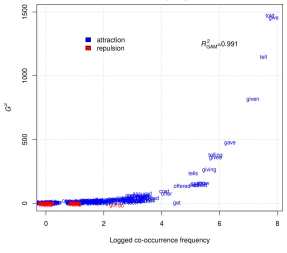
- *PMI* theoretically falls into the interval [-∞,∞]
- but practically, ie given marginal totals
  - *PMI* for *assured* & ditrans falls into [-0.68, 6.25]
  - *PMI* for *offer* & ditrans falls into [-3.26, 6.25]
- seeing this, can we really compare
  - *assured*'s PMI of 5.99 to
  - *offer*'s PMI of 4.23?
- look what can actually happen …

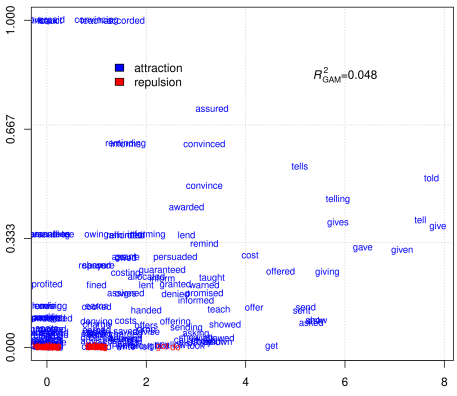
Why one needs to control for frequency

- *PMI* to the ditrans:
  - *purchased*: 3.254
  - *showing*: 2.845
- but here are the possible ranges of *PMI*-values
  - *purchased*: [-0.103, 6.236]
  - *showing*: [-3.481, 6.252]
- how can we compare the two *PMI*-values when they come from very different value ranges?



What happens if one 'takes freq out' …

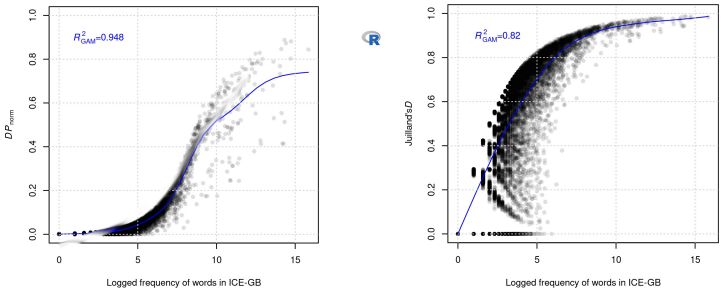




"The authors should tupleize to keep dimensions of information separate"

- "They should check for all co-occurrences the degree to which they are widely dispersed
- "however, here, they must be even more aware of the high correlation of nearly all DMs with frequency
- "this is esp. true of *range*, but also other DMs
- "the authors might use Gries's (2024) suggestion and
  - compute the KL-divergence
    - of the distribution of co-occurrences (posterior) from the file sizes (prior)
    - that controls for frequency
- "that is the measure that
  - can reflect dispersion differences between lower-frequency words
  - has been shown to boost predictive modeling of reaction times best (Gries 2022, *JSLS*)

"The authors should tupleize to keep dimensions of information separate"



"The authors should include other variables that co-vary with collexemes"

- "For example, Olguín Martínez & Gries (2024) explore 3 counterfactual constructions
  - *if not for* X, [main clause …]
  - *if it wasn't for* X, [main clause …]
  - *if it weren't for* X, [main clause …]
- "a traditional analysis would involve this, but …
- "they did a hierarchical configural frequency analysis
- "given a set of *f* annotated features, the HCFA generates all possible sub-tables of the *f* features
  - if *f* is 3 with features A, B, and C, then it generates
    - the levels of A (not
    - the levels of B that
    - the levels of C interesting)
    - the levels of A:B (A cross-tabulated w/ B)
    - the levels of A:C (A cross-tabulated w/ C)
    - the levels of B:C (B cross-tabulated w/ C)
    - the levels of A:B:C (A cross-tabulated w/ B and C)

"The authors should include other variables that co-vary with collexemes"

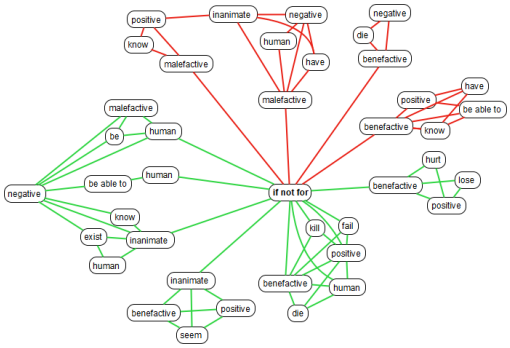
	A	B
1	CONSTRUCTION	LEMMAmainclause
2	if it weren't for	*be*
3	if it wasn't for	*know*
4	if not for	*be*
5	if it wasn't for	*have*
6	if it weren't for	*know*
7	if not for	*go*
8	if not for	*know*
9	if it weren't for	*be*

	A	B	C	D	E
1	CONSTRUCTION	LEMMAmainclause	CONSTRUCTION TYPE	ANIMACY	POLARITY main clause
2	if it weren't for	*be*	malefactive	inanimate	positive
3	if it wasn't for	*know*	malefactive	inanimate	positive
4	if not for	*be*	malefactive	human	negative
5	if it wasn't for	*have*	benefactive	human	negative
6	if it weren't for	*know*	benefactive	human	negative
7	if not for	*go*	benefactive	human	positive
8	if not for	*know*	benefactive	human	negative
9	if it weren't for	*be*	benefactive	human	positive
10	if it wasn't for	*win*	benefactive	human	positive
11	if not for	*have*	malefactive	inanimate	positive
12	if it wasn't for	*know*	benefactive	inanimate	negative
13	if not for	*collapse*	benefactive	inanimate	positive
14	if it wasn't for	*be*	malefactive	human	positive

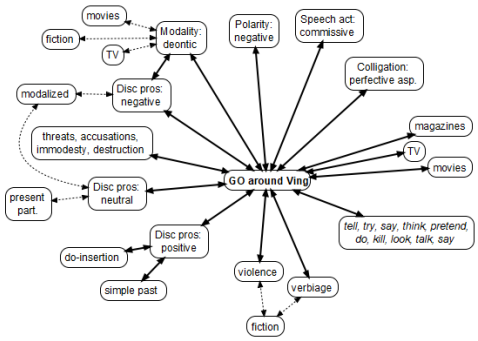
"The authors should include other variables that co-vary with collexemes"

- "Apart from the the observed co-occurrence frequencies, the HCFA determines
  - the expected co-occurrence frequencies (under H_0)
  - contributions to X^2 and to G^2
  - the (adjusted) *p*-value of an exact binomial test
  - an r/R^2-like effect size measure
- "the main output then is
  - a list of preferred co-occurrences (types), i.e. combinations of variable levels that attract each other
  - a list of dispreferred co-occurrences (antitypes), i.e. combinations of variable levels that repel each other
- "these can be explored for interesting yet parsimonious combinations of features
- "Jensen & Gries (2025) did something similar using the exploratory algorithm of association rules for
  - 2 constructions: *go (a)round* Ving & *go (a)round and* V
  - 8 additional variables: LEMMA, COLLIGATION, MODALITY, POLARITY, SEMPROSODY, DISCPROSODY, SPEECHACT, & REGISTER

"The authors should include other variables that co-vary with collexemes"



"The authors should include other variables that co-vary with collexemes"



"The authors should consider more than just letter sequences"

- Above, I said most studies treat a sequence of letters as a construction …
- … when a sequence of letters is only the form side of a construction but not also its function
- then, I asked why is/are
  - *play* attracted to the prepositional dative
  - *carry out* attracted to verb-particle-DO (VPO)?
  - *pick up* and *put down* not attracted to either construction?
- you see what this presupposes?
- to answer some of these, we need to answer:
  - what is the construction [*carry out*]?
  - what is the construction [*pick up*]?
  - what is the construction [*put down*]?

Constructions, not letter sequences: [verb particle] + 'meaning'!

- [*carry out*] is the formal pole – but the functional?
  - 'move outside' (eg *a box out of an apartment*)
  - 'execute/perform' (eg *an order*)
- [*pick up*] is the formal pole – but the functional?
  - 'register/notice/learn' (eg *I don't pick things up quite as quickly as he does*)
  - 'take' (eg *He picked up a pencil*)
  - 'meet/hit on' (eg *He picked a girl up on the train*)
  - 'address' (eg *Let's pick up the point about membership fees again*) …
- [*put down*] is the formal pole – but the functional?
  - 'kill' (eg *put down the terminally ill dog*)
  - 'place' (eg *put down the cup*)
  - 'write/register' (eg *put me down for some pudding*)
  - 'denigrate' (eg *must you put him down like that? He's only just started practising this!*)

Collostructions require a different annotation

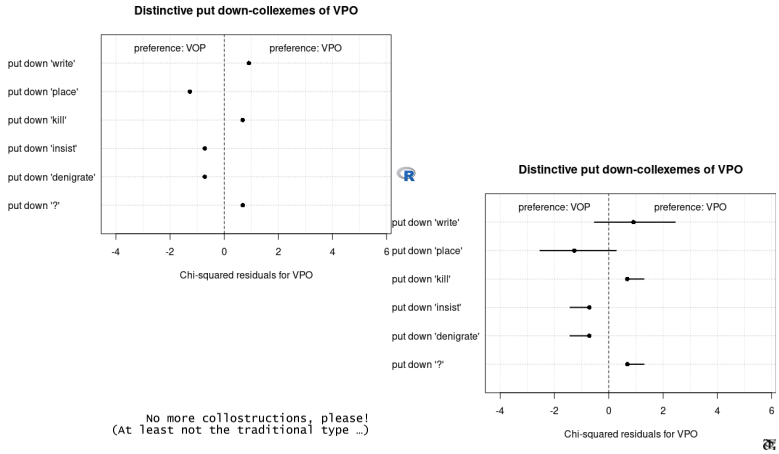
- "Here's what everyone is doing collostructions on
- "here's what we **should be** doing collostructions on
- "what are the results?
  - for the 2 constructions of the letter sequence *carry out*
    - [*carry out*] 'execute': 49 VPO : 0 VOP
    - [*carry out*] 'move out': 1 VPO : 1 VOP
  - for the 6 constructions of the letter sequence *put down*
- "it is self-contradictory to
  - self-identify as Construction Grammarians
  - do collostructions
  - take seriously only the constructionhood – a pairing of form and function – of one of the constructions in one's study, the 'grammatical'/argument structure construction
    - the ditransitive or prepositional dative
    - verb-particle-DO or verb-DO-particle
  - but reduce the other one to the form side of 'what letters are being used' and disregard their meaning!

	A	B	C	D	E	F	G	H
1	PRECEDING	VERB	DO	PART	SUBSEQUENT	CX	LEMMA-LETTER	
2	old of a book and you just don't want to	put	it	down		VOP	put down	
3	I	put	it	down	on my C V though	VOP	put down	
4	God you really know how to	put	someone	down	don't you	VOP	put down	
5	I	put	my foot	down	at being a medic so this is a compromi	VOP	put down	
6	You	put	them	down	there	VOP	put down	
7	said well fuck off then <laughter> and	put	the phone	down	on him you know	VOP	put down	
8	Now on my uh my manifesto do I	put	Greek and La	down	cos when I'm when I'm listing my seco	VPO	put down	
9	Well I putting		it	down		VOP	put down	
10	To move your hand and then I	put	it	down		VOP	put down	
11	and then you	put	it	down	by your side cos you've only half used	VOP	put down	
12	Uhm now I	put	a lie	down		VPO	put down	
13	I'm an assistant manager and they	put	me	down		VOP	put down	

	A	B	C	D	E	F	G	H
1	PRECEDING	VERB	DO	PART	SUBSEQUENT	CX	LEMMA-LETTER	LEMMA-SENSE
2	old of a book and you just don't want to	put	it	down		VOP	put down	put down: move downwards
3	I	put	it	down	on my C V though	VOP	put down	put down: write
4	God you really know how to	put	someone	down	don't you	VOP	put down	put down: denigrate
5	I	put	my foot	down	at being a medic so this is a compromi	VOP	put down	put down: idiom
6	You	put	them	down	there	VOP	put down	put down: move downwards
7	said well fuck off then <laughter> and	put	the phone	down	on him you know	VOP	put down	put down: move downwards
8	Now on my uh my manifesto do I	put	Greek and La	down	cos when I'm when I'm listing my seco	VPO	put down	put down: write
9	Well I putting		it	down		VOP	put down	put down: move downwards
10	To move your hand and then I	put	it	down		VOP	put down	put down: move downwards
11	and then you	put	it	down	by your side cos you've only half used	VOP	put down	put down: move downwards
12	Uhm now I	put	a lie	down		VPO	put down	put down: write
13	I'm an assistant manager and they	put	me	down		VOP	put down	put down: write
14	So if you	put	both those	down	the whole point is which one is right	VOP	put down	put down: write
15	So you	put	drivel	down	and just present it so that it's all pagin	VPO	put down	put down: write
16	I	put	ds a month or	down		VPO	put down	put down: write
17	All you can really do then is just is just	put	e names as the	down	on the form with the new address	VPO	put down	put down: write
18		put	s but it's not b	down	uhm	VPO	put down	put down: write
19	And just	put	the cost prices	down		VPO	put down	put down: write

	VPO	VOP	Sum
put down$_{c11}$	a	b	a+b
other	c	d	c+d
Sum	a+c	b+d	a+b+c+d

Collostructions require a different annotation



Some collostructional studies did much better

- Wiechmann (2008) found that collostructional results that distinguished senses of verbs (whether they take a nominal or a sentential complement) outperform collostructional results that did not distinguish senses when it comes to predicting reading latencies
- Gilquin (2013) showed different senses of verbs have v different preferences for causative constructions
- Colleman & Bernolet (2012) showed how polysemy effects can partially explain the lack of (better) agreement between corpus & exp. data on the dative alternation
- Bernolet & Colleman (2016) illustrated how verb senses can have widely different alternation biases (in a study of the Dutch dative alternation) and how, in an experiment, priming was affected by an interaction of the strength of priming and sense-specific biases of target verbs

"The authors should indicate ranges of plausible values"

- "However, this is tricky
- "they cannot use parametric confidence intervals that come with many ready-made statistics because
  - corpus data often violate the general distributional-shape assumptions of parametric confidence intervals
  - the bag-of-words model isn't an appropriate one for corpus data
- "they cannot use bootstrapping of individual uses / data points because corpus data involve repeated measurements
- "they need to use bootstrapping based on texts/files

Quantifiying uncertainty – why and how?

- "For the right kind of bootstrapping, they need the input in a format different from what we're used to (which was already used in all examples above)
- "instead of

Word/lemma	FreqInDitr	FreqElsewhere
accord	3	3
ask	92	428
give	566	603
…	…	…

or

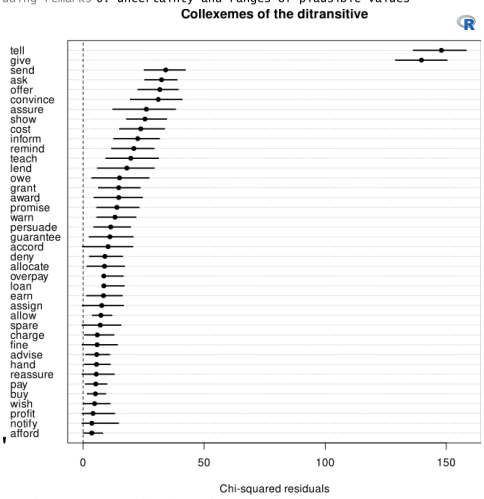
Word/lemma	FreqInDitr	FreqCorpus
accord	3	6
ask	92	520
give	566	1169
…	…	…

FILE	LEMMA	DITR
s1a-001	*tell*	TRUE
s1a-001	*be*	FALSE
…	…	…
s1a-002	*graduate*	FALSE
s1a-002	*give*	TRUE
…	…	…

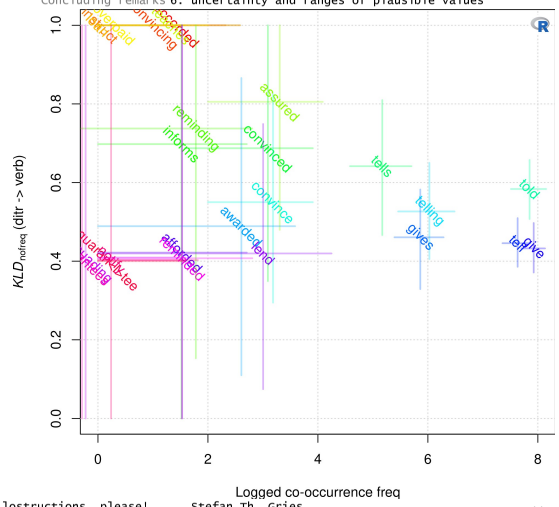
- "they need something like this, which states for every verb form/lemma
  - whether it is used ditransitively or not
  - which file it occurred in (so we can use that for our legitimate bootstrapping!)
- "Then they can do the following 500 times
  - sample from all *n* files *n* w/ replacement
  - compute the desired AM for each lemma in the sample
  - collect & then plot the observed value in the corpus & the 2.5% and 97.5% quantiles

Results for a collexeme a.

- Interestingly,
  - *tell* is a bit ahead of *give*, but really not significantly different
  - even among the top 30 verbs (w/ very fitting semantics), we find a few verbs that are not significantly attracted (e.g., *accord* or *assign*)
  - in fact, of all 88 verbs in the ditransitive, 34 aren't 'significant'

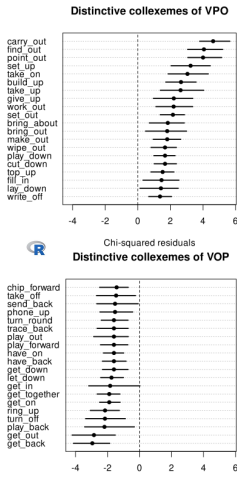


Results for a collexeme analysis



Results for a dist. collexeme analysis

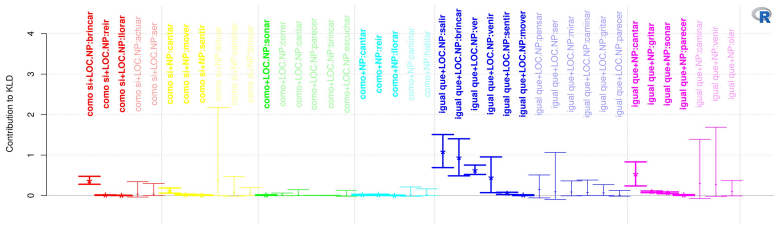
- Interestingly, we find a variety of significant phrasal verbs for each construction, but also fairly high variability – i.e. little significance – among them
- 160 verb lemmas have no significant preference, 540 do (this is not surprising, given how a dist. coll. anal. focuses only on the two constructions of interest)



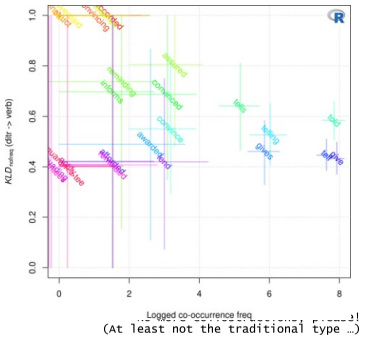
"As an editor, …"

- "… I think the reviewers identified a variety of areas in which the empirical analysis in your ms can be improved
- "I don't want to force you to do all these things.
- "I think the reviewers' suggestions could be ranked in terms of simplicity/'involvedness' as follows:
  - distinguish frequency and association and different directions of association
    - because you already have freq and just need different AMs
  - compute dispersion
    - because you already know where what occurred
  - compute ranges of plausible values
    - because it's just one loop on top of dispersion
  - annotate senses
    - because semantic annotation is hard and requires another thorough look at the data
  - annotate more variables
    - because annotation is hard and requires even more looks

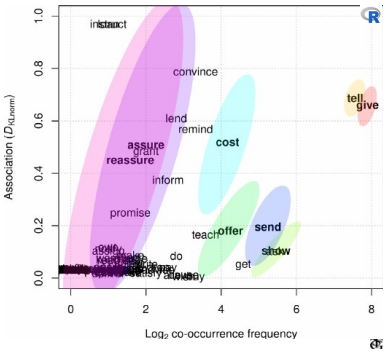
"As an editor, …"



"As an editor, …"



"As an editor, …"



"As an editor, …"

- "However, it is also necessary to consider the importance of the suggestions in terms of how much they are required and improve the analysis
- "For that, the ranking is probably a bit different:
  - annotate senses
    - because otherwise you're not exploring co-occurrence of constructions
  - distinguish frequency and association and different directions of association
    - because otherwise you're not 'measuring' what you say you're 'measuring'
  - annotate other variables
    - because of course there's more to the uses of these constructions than just the words/lemmas in one slot
  - compute dispersion
    - because you need to at least make sure you're not overinterpreting outlier results
  - compute ranges of plausible values
    - because it's just one loop on top of dispersion and will help streamline your interpretation

Take-home message

- Collostructions is a general insightful approach, …
- but linguistics as a discipline has evolved a lot when it comes to the quantitative/statistical analysis of (corpus) data
- it's not 2003/2004 anymore
- just like variationist sociolinguistics moved on from Varbrul (at some point, finally, thankfully …), …
- … collostructional studies, and quantitative approaches to CxG in general, need to move on from 'basic collostructions'
- collostructions need to be 'updated' and I hope this talk has shown you why and especially how
- so, go and preempt this reviewer! ;-)

Thank you!

https://www.stgries.info