On recency and dispersion

Stefan Th. Gries
UC Santa Barbara & JLU Giessen
http://tinyurl.com/stgries

A very brief recap

- I mentioned earlier this week that (token) freqs alone are maybe not as important as much work in psycholinguistics & cognitive/usage-based linguistics has assumed – we saw
  - doubts from Schmid regarding the nature between freq & entrenchment & this quote "frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse."
  - results from Baayen and others that seem to indicate that frequency-as-repetition is not that important
  - quotes from Ellis et al. pointing towards other factors

> Practice promotes proficiency (eg, Anderson, 2009; Bartlett, [1932] 1967; Ebbinghaus, 1885). Learning, memory and perception are all affected by frequency, recency, and context of usage: The more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed (Ellis, Römer, & O'Donnell 2016:45f.)

Today, we will talk about the 2nd crucial mechanism in this quote: recency

- Recency can be seen as being manifested corpus-linguistically in two ways
  - short-term: priming/autocorrelation
  - long-term: dispersion
    - across speakers (ie often files) (recall Dąbrowska 2016)
    - across registers/genres/other corpus parts
- recency is hardly ever utilized outside of the context of priming in both cogn & corpus linguistics
- this is unfortunate, because we know that
  - priming/autocorrelation has a lot of predictive power
  - aggregated freqs disregarding dispersion mean little
- let's unpack those things …

each instance redefines the system,
however infinitesimally,
maintaining its present state or
shifting its probabilities
in one direction or the other

(Halliday 1991/2005:67)

Stefan Th. Gries
UC Santa Barbara & JLU Giessen

Recency as priming: what's that?

- The first manifestation of recency is priming, ie the fact that an occurrence of x increases the probability of x recurring beyond its (frequency-based) baseline
  - if you've just described a transitive scenario w/ a passive sentence, you're more likely to describe the next transitive scenario also w/ a passive than if you'd just described a transitive scenario w/ an active sentence
  - if you've just described a transfer scenario w/ a prepositional dative, you're more likely to describe the next transfer scenario also w/ a prepositional dative than if you'd just described a transitive scenario w/ a ditransitive
- words can prime themselves like that, too, and they can prime semantically related words, etc
- ie there's different kinds of priming: syntactic, lexical, semantic, phonological, non-linguistic, …

Recency as priming in corpora

- How would you even recognize it in corpus data?

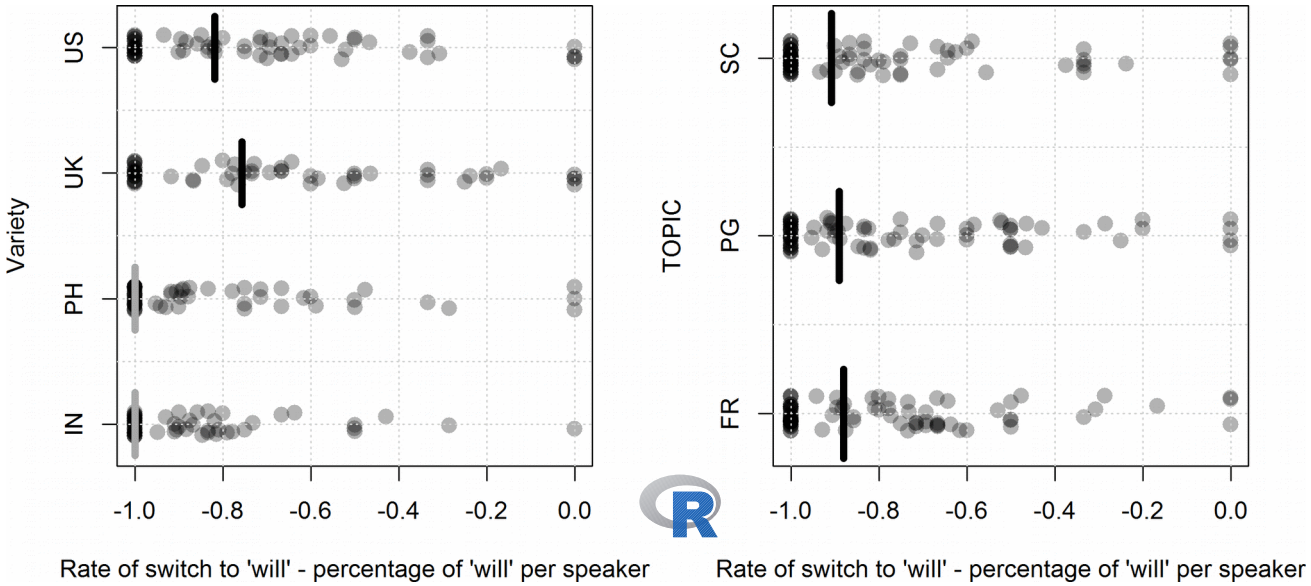
	CORPUS	FILE	LINE	PRECEDING	MATCH	MATCHLEMMA	SUBSEQUENT	COMPLEMENTIZER	LengthMatrixSubj	ComplementSubjLength
2	ICE-GB	S1A-001	#12:1:B	I	think	think	the m <,> the main perception	absent	1	79
3	ICE-GB	S1A-001	#12:1:B	n <,> uhm <,> unbalanced <,	was	be	I think the m <,> the main per	present	79	65
4	ICE-GB	S1A-001	#1251B	Uh I	was	shocked	<,> I mean I wasn't shocked	present	1	12
5	ICE-GB	S1A-001	#127:1:B	powerful and moving uhm <,,	is	be	Uhm so one of the things that	absent	48	5
6	ICE-GB	S1A-001	#13:1:B	m something that I I saw a lot	was	be	<,,> that when people were <,	present	28	6
7	ICE-GB	S1A-001	#2:1:B	I	think	think	the main things that I saw as	absent	1	85
8	ICE-GB	S1A-001	#28:1:B	I	think	think	that the <,,> what I get out of	present	1	38
9	ICE-GB	S1A-001	#29:1:B	obody is left out of this group	is	be	Uhm <,> the difference <,> I	present	75	18
10	ICE-GB	S1A-001	#29:1:B	Uhm <,> the difference <,> I	think	think	the main difference that I feel	absent	1	75
11	ICE-GB	S1A-001	#31:1:B	e work that I was involved in	was	be	Uhm <,> and I think one of th	absent	55	1
12	ICE-GB	S1A-001	#34:1:B	lop <,> uhm physical skills <,	was	be	One was that I was being give	present	3	1
13	ICE-GB	S1A-001	#35:1:B	m all sorts of other people <,	was	be	The other was that this was <	present	9	4
14	ICE-GB	S1A-001	#38:1:B	e w we 're working with now	is	be	that those <,> movement skill	present	77	21
15	ICE-GB	S1A-001	#48:1:B	We	decided	decide	that we would work together	present	2	2
16	ICE-GB	S1A-001	#53:1:B	I	think	think	that would be <,> that 's going	absent	1	4
17	ICE-GB	S1A-001	#54:1:B	And and I	think	think	the question can <,> is <,> is	absent	1	12
18	ICE-GB	S1A-001	#55:1:B	I	think	think	that the problems of working	present	1	32
19	ICE-GB	S1A-001	#64:1:B	And I	think	think	that 's very much <,> the patt	absent	1	4
20	ICE-GB	S1A-001	#66:1:B	egin to dance with each other	think	think	Uhm <,> so I think there is th	absent	1	5
21	ICE-GB	S1A-001	#71:1:B	Some peo I	think	think	some people come initially to	absent	1	11
22	ICE-GB	S1A-001	#74:1:B	se the requirement from them	will be	be	that they dance <,> that prima	present	33	4
23	ICE-GB	S1A-001	#74:1:B	But I	think	think	they very those people are ver	absent	1	4
24	ICE-GB	S1A-001	#98:1:B	nce a week uhm which which	means	mean	that the the pressures on us ar	present	5	20
25	ICE-GB	S1A-002	#10:1:C	rested particularly because he	was saying	say	that <,> it was a lot of a lot	present	2	2

- in some parts of (cogn) ling that's well known
  - eg, lg acq research carefully controls for priming
- elsewhere, not so much, which is tricky …

Ok then, but why do we care?

- Recency as priming
  - priming is a threat to common regression modeling
    - as a form of autocorrelation, it amounts to a violation of the independence-of-data points assumption
    - it has high predictive power that, if overlooked, can make other predictors seem stronger than they are, ie it makes studies anticonservative
  - example: *will* vs. *going to* vs. *shall* in the Q&A corpus
    - switch % to *will* – % of *will* per speaker
      - many speakers don't alternate, but many also exhibit strong priming effects (esp. in the non-native varieties), variation of priming strengths is less high across topics than across varieties
    - predictive power: 80.9% correct just by
      - choosing a speaker's last construction
      - choosing *will* as the first future

Ok then, but why do we care?



When we plan on including the role of autocorrelation we should …

- … consider (cumulative) priming/learning effects – not just in corpora, also within experiments!
  - Scheepers (2003) explores long-term priming within an experiment by splitting the data into an early and a late half, but finds no significant effect w/ that
  - Jaeger/Snider (2008): cumulativity, "the number of primes of each structure previously encountered or produced […] (excluding the most recent prime)"
    - study voice & *that*-relativizer omissions in corpus data
    - find significant effects of cumulativity
  - STG/Wulff (2009): *to* vs. *ing* complementation in L2 Eng
    - use a sentence-completion experiment w/ German learners
    - find a suggestive tendency for within-subject-accumulative priming
  - Doğruöz/STG (2012): satiation (Francom 2009)
    - find that speakers of Turkish become more accepting of unconventional syntactic expressions over 8 stimuli
  - STG (to app): verb-specific learning effects in dat.alt.
- thus, we could add a predictor CUMPRIM …

When we plan on including the role of autocorrelation, we should …

- … consider Szmrecsanyi's two kinds of persistence (Szmrecsanyi's 2005, 2006 word for 'priming')
  - α-persistence – what we called recency-as-priming
    - previous exposure to **the same variable**: the use of variant X will facilitate/make more likely a subsequent use of X
      - e.g., analytic comparatives prime analytic comparatives
      - e.g., *going-to* futures prime *going-to* futures
  - β-persistence
    - previous exposure to **a related/similar variable**: the use of variant X will facilitate/make more likely the use of a similar/related variant of Y
      - e.g., uses of *more* outside of analytic comparatives prime analytic comparatives
      - e.g., uses of *go* as a motion verb prime *going-to* futures
- thus, adding predictors such as LASTCHOICE and LASTCHOICEWGHT is good, but not even enough since they do not consider β-persistence – we might need LASTSIMILAR

When we plan on including the role of autocorrelation we should …

$$d_{ij} = \frac{\sum_{k=1}^{p} w_k \delta(ij;k) d(ij;k)}{\sum_{k=1}^{p} w_k \delta(ij;k)}$$

- … consider a role of prime-target similarity going beyond β-persistence
  - there is a well-documented lexical-identity boost (Pickering/Branigan 1998, STG 2005, Szmrecsanyi 2005)
  - there is an effect of global prime-target similarity: Snider (2009) finds that "When the prime construction is PO, the PO construction is 10.6 times more likely in the target for every one-unit decrease in [GlobalSim]"
  - there is an effect of verb-sense-identity boost for dat.alt. (Bernolet/Colleman/Hartsuiker 2014)
- … consider the role of surprisal for priming (later)
- … thus, we could add predictors LEXID and PRIM2TARGSIM and SURPRISAL – all of this requires very careful planning, annotation, & evaluation

Recency as dispersion: what's that?

- The second manifestation of recency is dispersion, ie the fact that occurrences of x are usually not evenly distributed across the parts of a corpus
- this affects
  - freqs of occurrence: *HIV*, *keeper*, & *lively* are equally freq in the BNC (16 pmw) but differ
    - re range: 62, 97, and 97, of 100 equally-sized corpus parts
    - re Juilland's *D*: 0.56, 0.87, 0.92 respectively
  - freqs of co-occurrence: verbs most attracted to the imperative in the ICE-GB: *see*, *let*, *look*, *fold*, *worry*, *listen*, *take*, *remember*, 5 more, *process* (15), but *fold* & *process* in imperatives occur in only $^{1}/_{500}$ files (*D*=0)
  - everything: dispersion affects every single kind of frequency you can get from a corpus

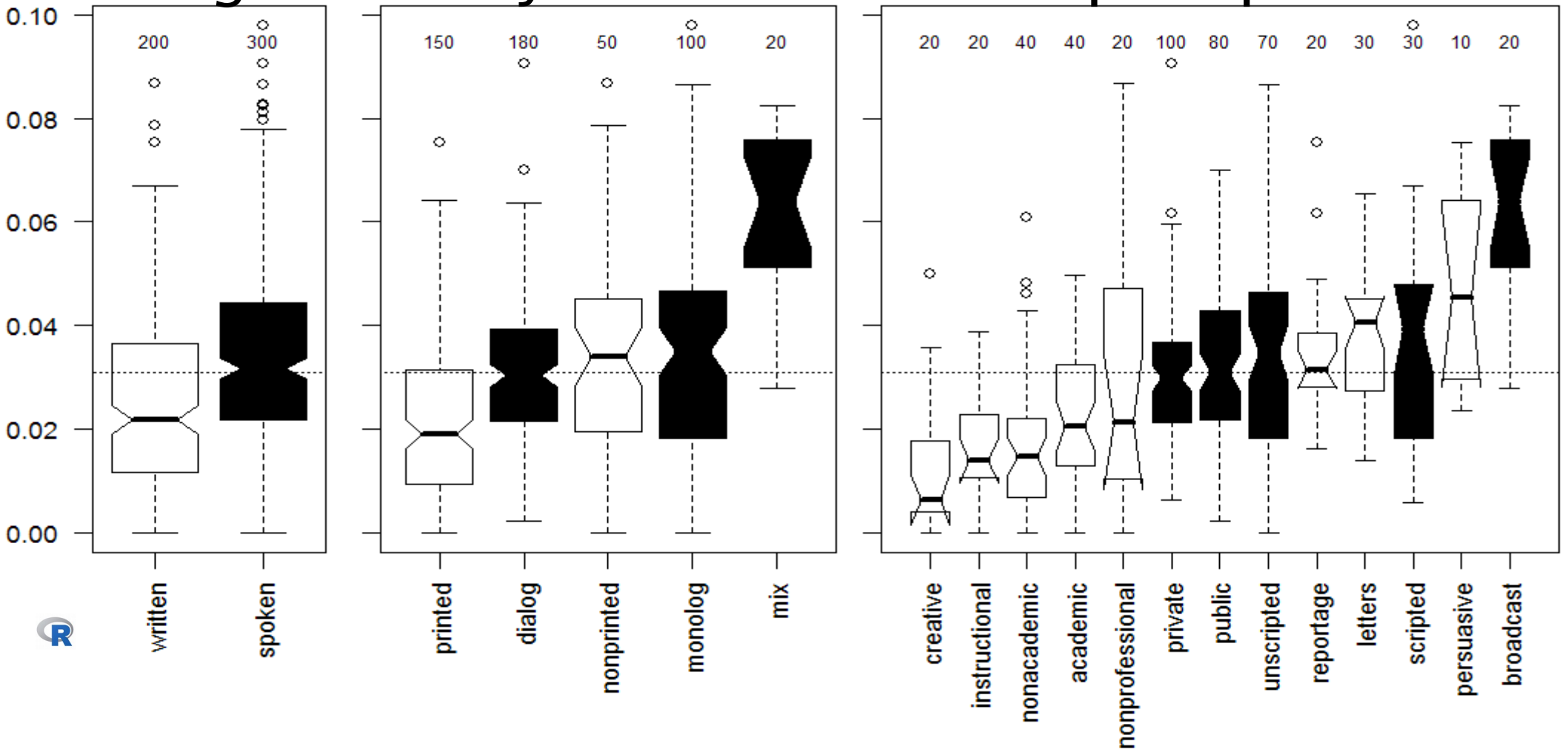
Recency as dispersion: what's that computed on (ideally)?

- Note: most (of the too few) applications of dispersion measures are based on dividing the corpus into parts that are linguistically irrelevant
- but corpora usually come with a linguistically meaningful substructure, which provides levels of resolution over which to compute dispersion
  - files (if those correspond to speakers/texts/otherwise meaningful sampling units)
  - registers, subregisters, genres, modes, …
- if you do not consider dispersion, any statement about 'what's in a corpus' is
  - a generalization over parts of a corpus that may be valid, but also …
  - a generalization over parts of a corpus that
    - hopes that the H0 of equal distributions is right
    - may be terribly wrong or oversimplified if said H0 is wrong
- what, you don't believe me?

How frequencies of present perfects change when you look at corpus parts …

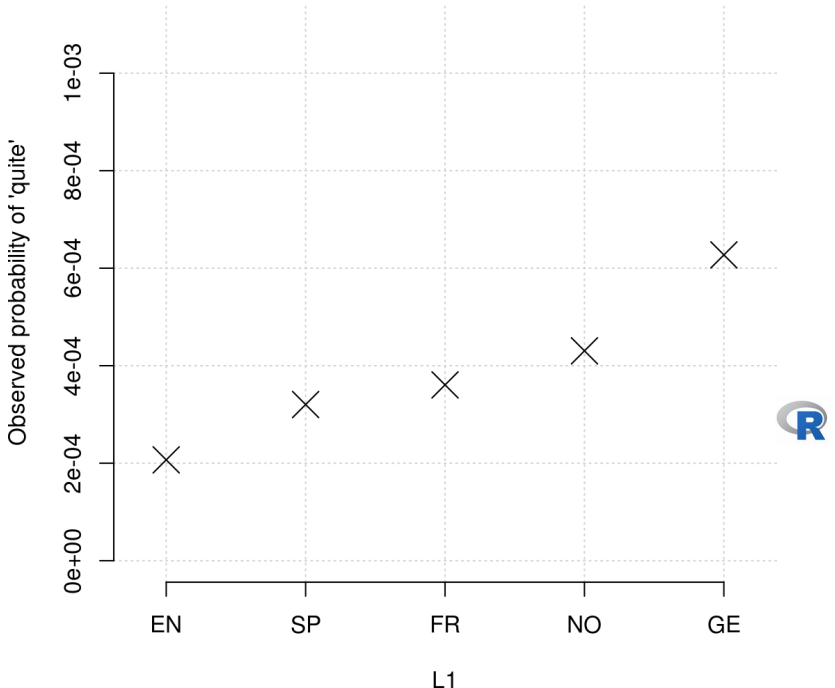
Mode	Register	Sub-register	File
spoken	dialog	private	f1, f2, f3, …
		public	f10, f11, …
	monolog	scripted	f20, f21, …
		unscripted	f30, f31, …
	mix	broadcast	f40, f41, …
written	print	academic	f50, f51, …
		creative	…
		instructional	…
		non-academic	…
		persuasive	…
		reportage	…
	non-printed	letters	…
		non-professional	…

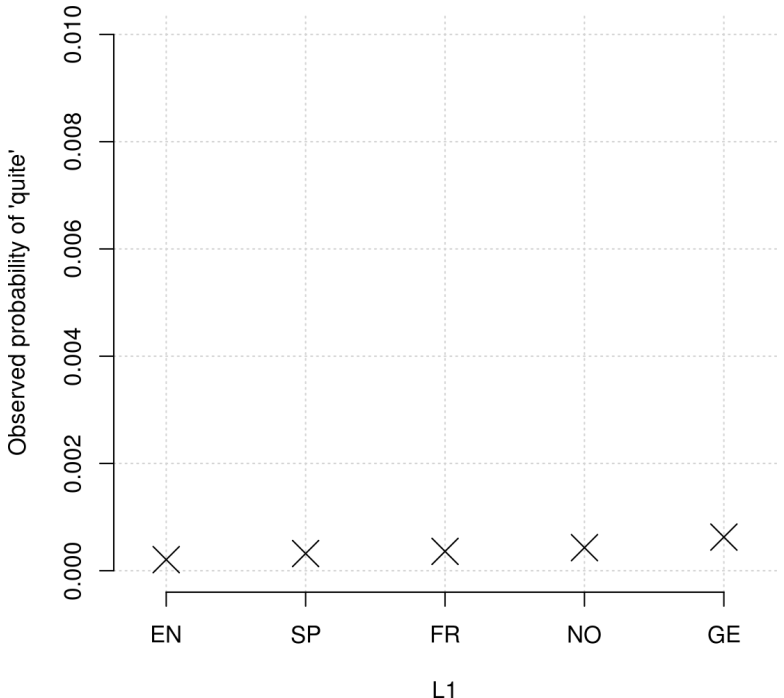
How frequencies of present perfects change when you look at corpus parts …



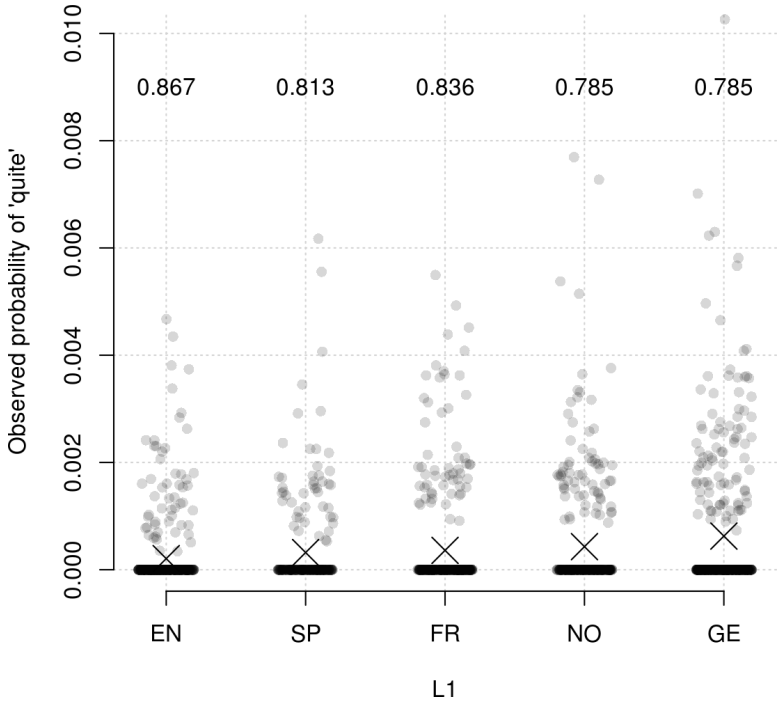
How frequencies of *quite* change when you look at speakers in corpus parts …

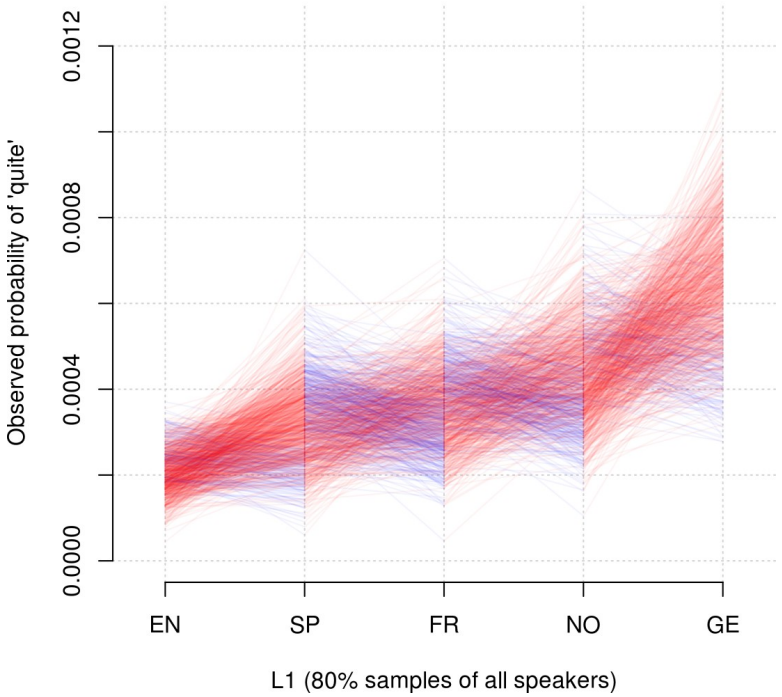
- Here are some %s of the word *quite* in
  - native speaker data (EN)
  - non-native speaker data (SP, FR, NO, GE)
- here's the speaker variability …
- most don't even use quite
- here's how sampling-dependent these results are
- many similar results: STG (2006), Callies (2013), Gablasova, Brezina, & McEnery (2017), …





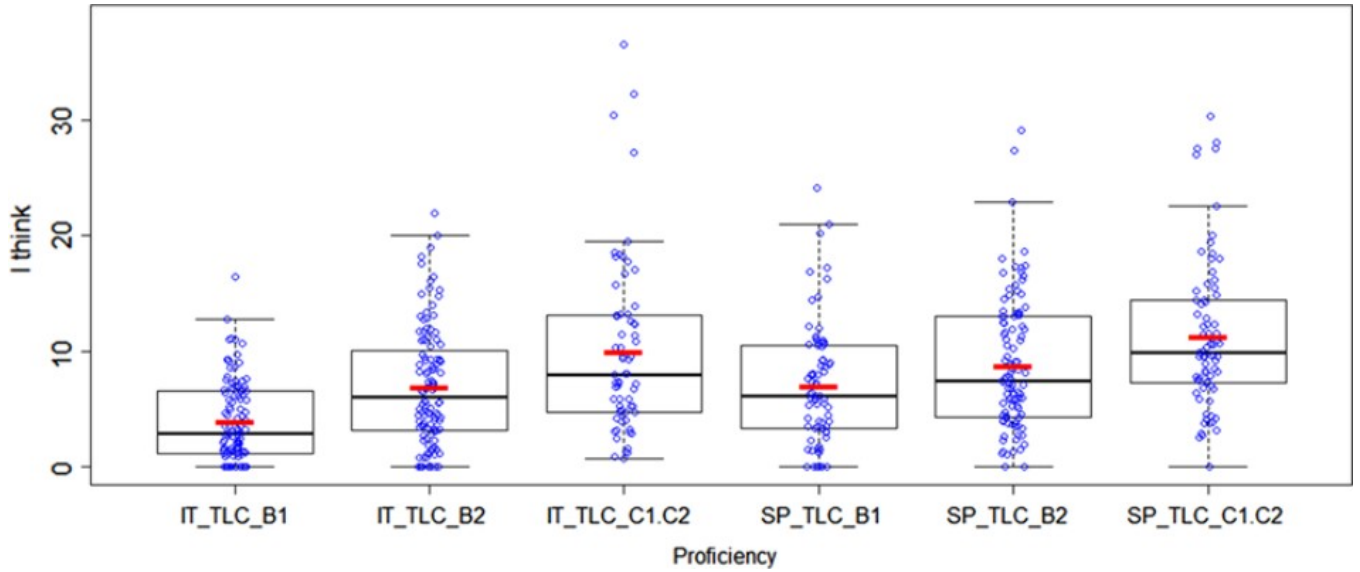
How frequencies of *quite* change when you look at speakers in corpus parts …

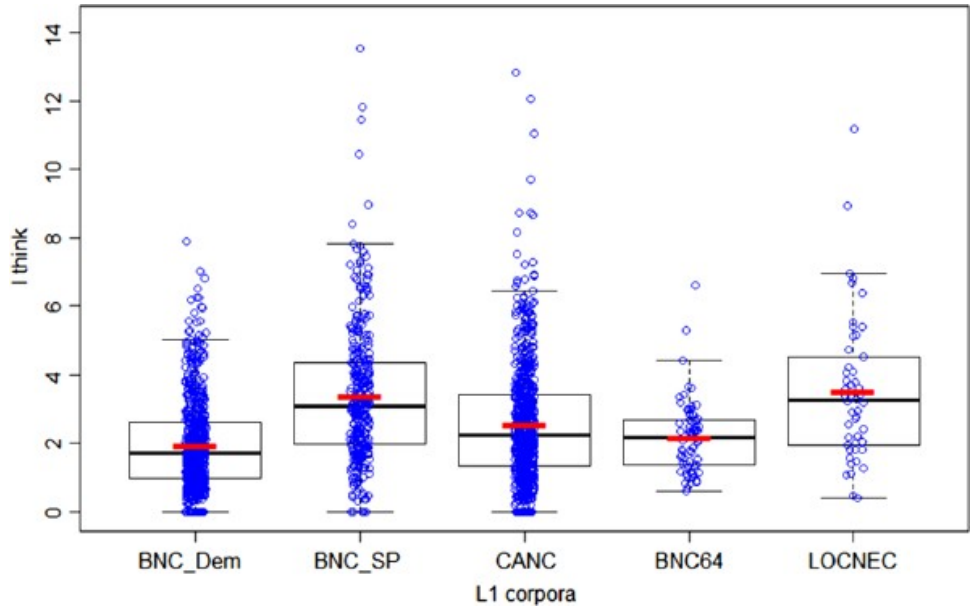




Studies relying on aggregate frequencies are ~~potentially~~ likely useless

- What do these studies show?
  - Gablasova, Brezina, & McEnery (2017) on *I think* in the Trinity Lanc Corpus of learner language
  - Callies (2013) documents similar variability for first-*person* pronouns in MICUSP and CALE
  - and of course this isn't just the case in learner language: Gablasova, Brezina, & McEnery (2017) again
  - in fact, without any success whatsoever, STG (2006) promoted a whole research agenda on corpus homogeneity using permutation/resampling methods
    - no, his focus wasn't highlighting differences between speakers
    - but yes, every corpus result is affected to some degree by corpus homogeneity
- no study that wants to be usage-based can afford to not at least explore different speakers' usages!


I think
0
10
20
30
IT_TLC_B1
IT_TLC_B2
IT_TLC_C1.C2
SP_TLC_B1
SP_TLC_B2
SP_TLC_C1.C2
Proficiency


I think
0
2
4
6
8
10
12
14
BNC_Dem
BNC_SP
CANC
BNC64
LOCNEC
L1 corpora

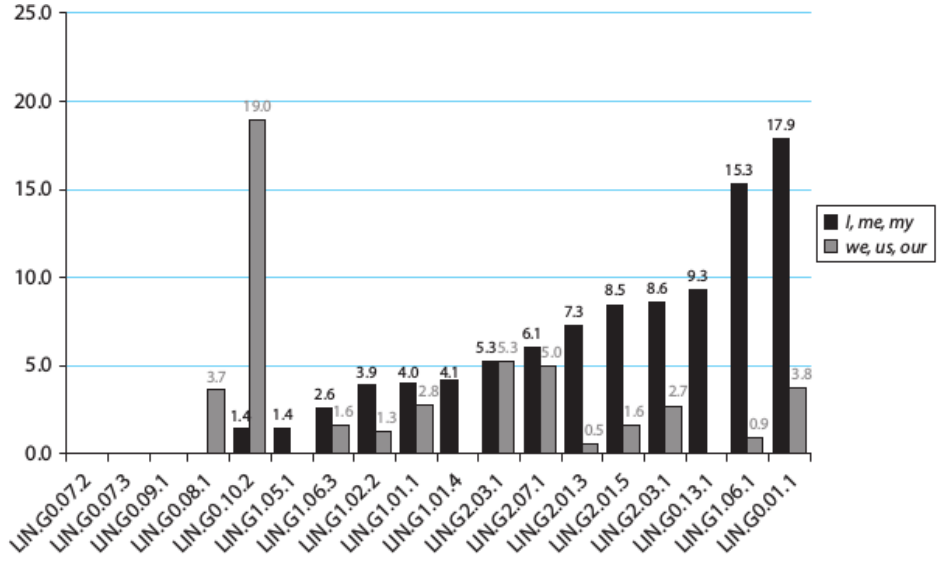
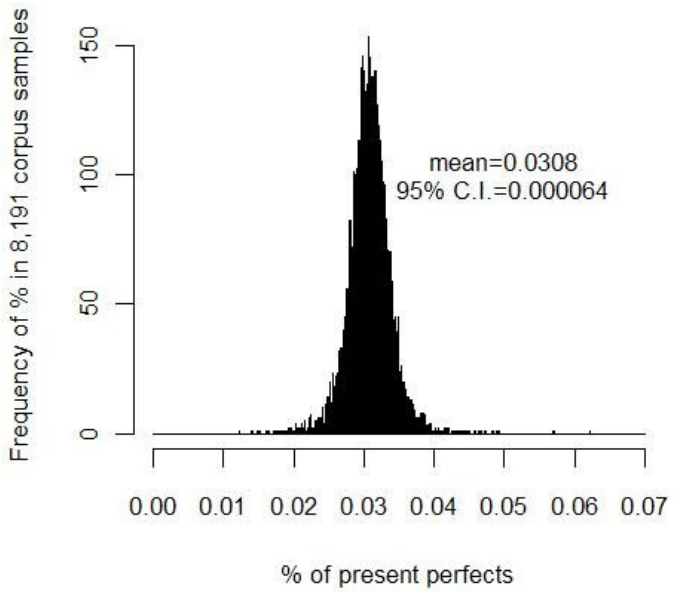


Figure 4. Use of first person pronouns per 1,000 words by individual writers in MICUSP



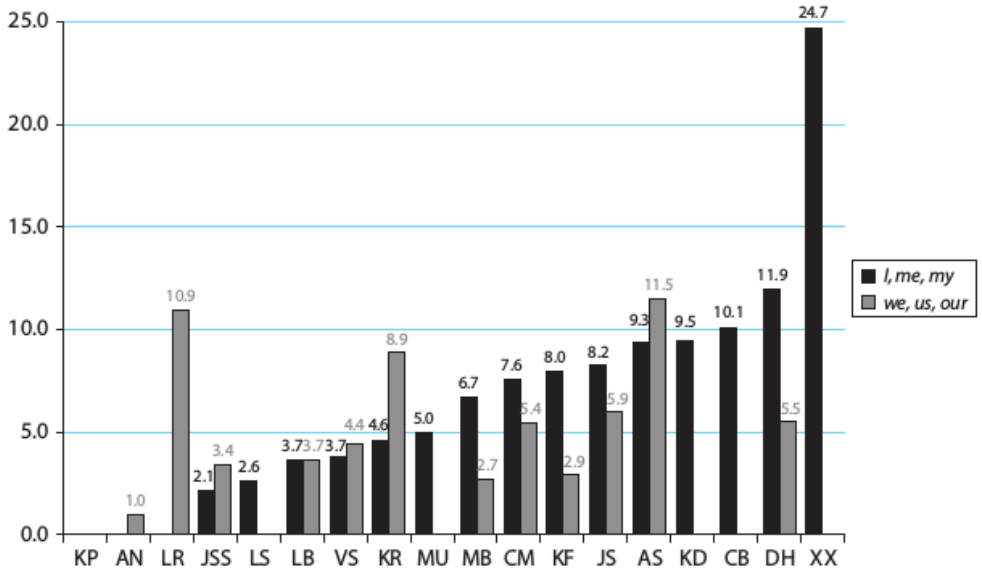


Figure 5. Use of first person pronouns per 1,000 words by individual writers in CALE

How frequencies can be misleading …

- Imagine you're looking for verbs/adjectives from some frequency range in the Brown corpus (35-40 pm)
  - because you need stimuli for a psycholinguistic experiment or a vocabulary test
  - because you need words for a vocabulary list …
- so you find these two: *enormous* & *staining* (*n*=37)
- but you probably didn't at all reach your goal (of finding words that are identified equally fast/accurately, that learners are equally likely to know, …)
  - *enormous*: 1 each in 35 corpus parts and 2 in 1
  - *staining*: 37 in 1 corpus part
- "Language users are more likely to experience constructions that are widely or evenly distributed in time or place. When they do so, contextual dispersion indicates that a construction is broadly conventionalized, temporal dispersion shares out recency effects." – this supersedes frequency!

What measure of dispersion to use …

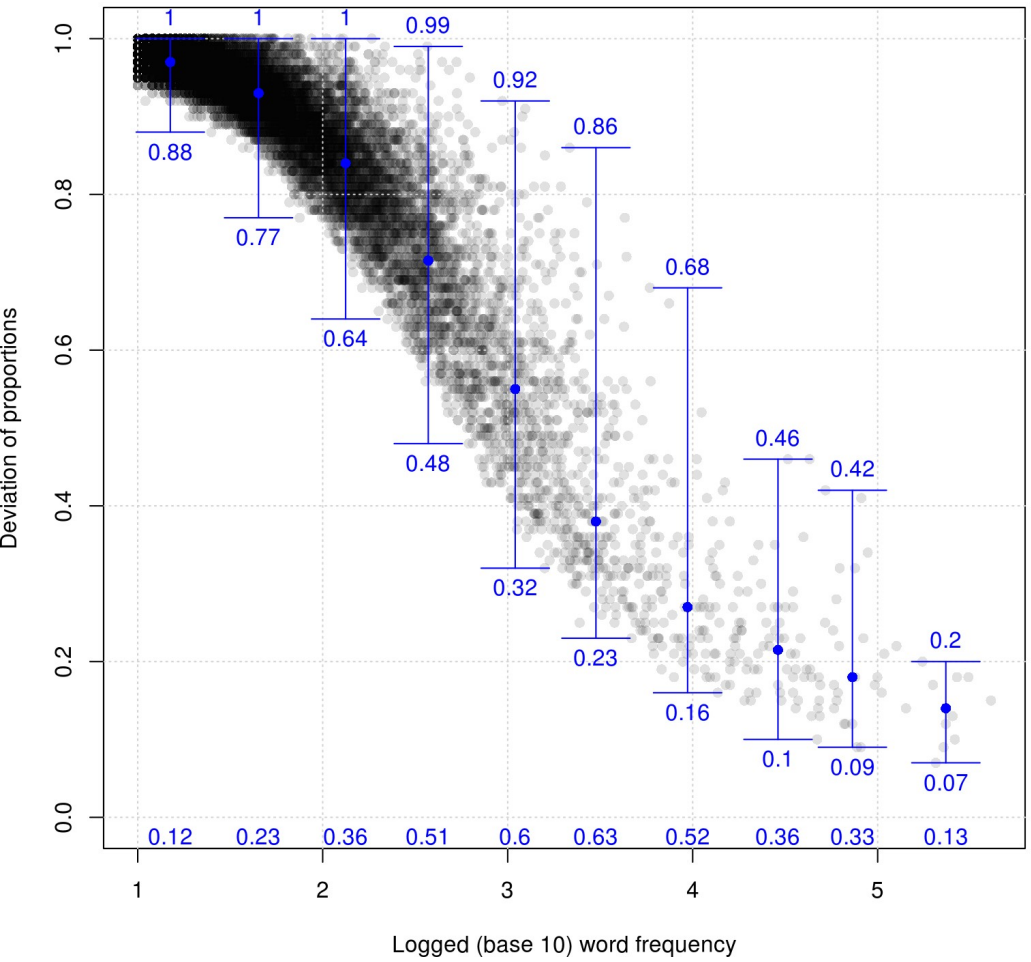
- Many measures of dispersion have been proposed …
  - range: the number of corpus parts containing x
  - *sd/vc* of the frequencies of x in the corpus parts
  - Juilland's *D* $1-\frac{sd_{population}(p)}{mean(p)}\times\frac{1}{\sqrt{(n-1)}}$ Carroll's $D_2=\frac{-\sum_{i=1}^{n}(\frac{p_i}{\sum p}\times\log_2\frac{p_i}{\sum p})}{\log_2 n}$
  - Rosengren's *S*= $(\sum_{i=1}^{n}\sqrt{s_i\cdot v_i})^2\times\frac{1}{f}$ $(\text{with } min\,S=1/n)$ *DP*= $0.5\times\sum_{i=1}^{n}\left|\frac{v_i}{f}-s_i\right|$
- Deviation of Proportions *DP* (Gries 2008), ie sum(abs(OBS-EXP))/2
  - stays within its defined comparable range
  - distinguished distributions other measures can't
  - doesn't overly penalize 0s
  - has been shown to be better than the standard of Juilland's *D* (Biber et al. 2016, Burch et al. 2017)
- how does *DP* behave when applied to pseudo-randomly sampled words from the BNC sampler?

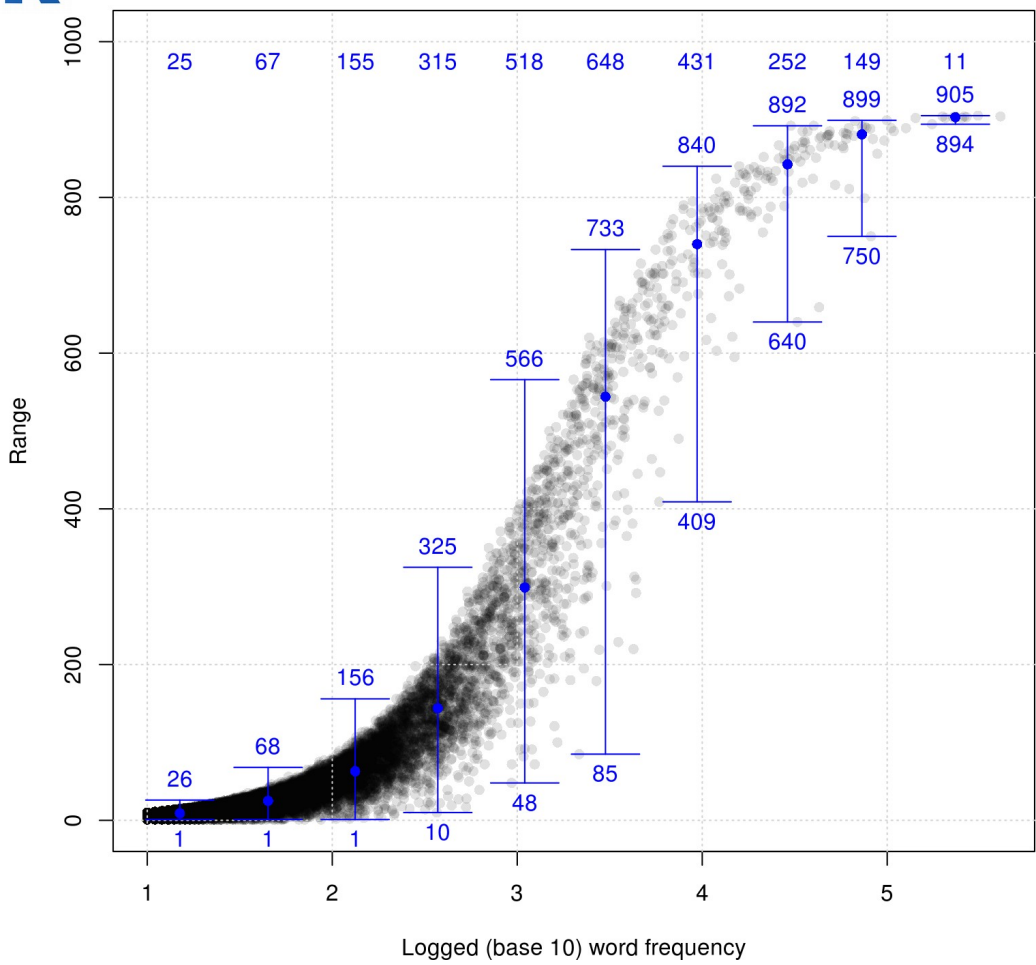
Example number	Exp (sizes of parts)	Obs (distribution)	abs diff	sum of abs diff	divide by 2 *DP*
	0.33	0.33	0		
1	0.33	0.33	0	0	0
	0.33	0.33	0		
	0.33	1	0.67		
2	0.33	0	0.33	1.33	0.665
	0.33	0	0.33		
	0.01	0.98	0.97		
3	0.01	0.01	0	1.94	0.97
	0.98	0.01	0.97		
	0.01	0	0.01		
4	0.01	0	0.01	0.04	0.02
	0.98	1	0.02		
	0.45	1	0.55		
5	0.35	0	0.35	1.1	0.55
	0.2	0	0.2		
	0.45	0	0.45		
6	0.35	1	0.65	1.3	0.65
	0.2	0	0.2		
	0.45	0	0.45		
7	0.35	0	0.35	1.6	0.8
	0.2	1	0.8		

What measure of dispersion to use …

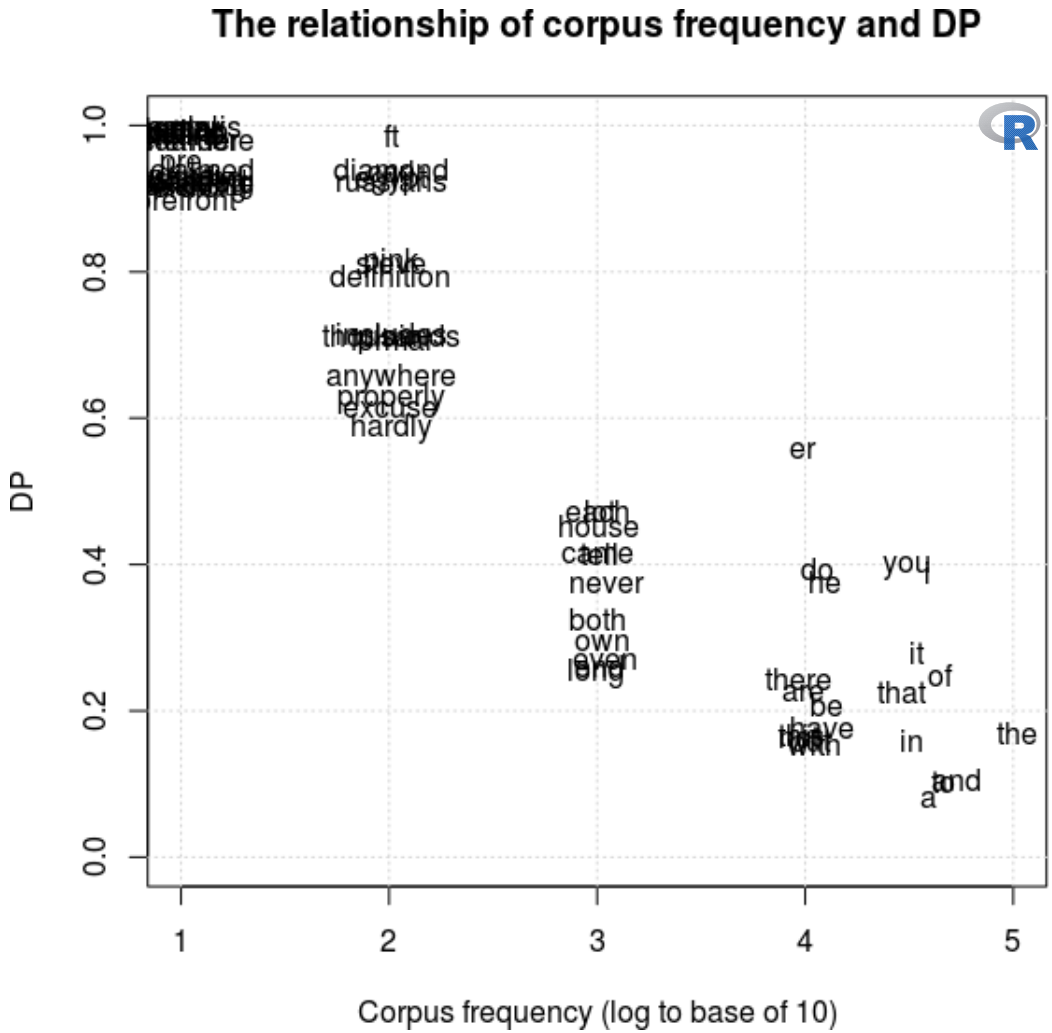
Minimal *DP*s			Intermediate *DP*s			Maximal *DP*s		
Word	*DP*	Freq	Word	*DP*	Freq	Word	*DP*	Freq
a	0.08	39122	*definition*	0.8	102	*macari*	1	10
to	0.1	46188	*includes*	0.72	102	*mamluks*	1	10
and	0.11	53224	*thousands*	0.71	102	*lemar*	1	10
with	0.16	11138	*plain*	0.71	102	*sem*	0.99	10
but	0.16	10569	*formal*	0.71	102	*hathor*	0.99	10
in	0.16	32201	*anywhere*	0.65	102	*tatars*	0.99	10
not	0.17	9211	*properly*	0.63	102	*scallop*	0.99	10
this	0.17	9652	*excuse*	0.61	102	*malins*	0.99	10
the	0.17	104253	*hardly*	0.59	102	*ft*	0.99	102
have	0.18	11929	*er*	0.56	9721	*defender*	0.98	10
be	0.21	12735	*each*	0.47	1007	*scudamore*	0.98	10
are	0.22	9771	*lot*	0.47	1032	*pre*	0.95	10
that	0.23	29283	*house*	0.45	1024	*diamond*	0.94	102
there	0.24	9243	*tell*	0.41	1023	*carl*	0.94	102

What that measure of dispersion does & how it relates to frequency …





What that measure of dispersion does & how it relates to frequency …



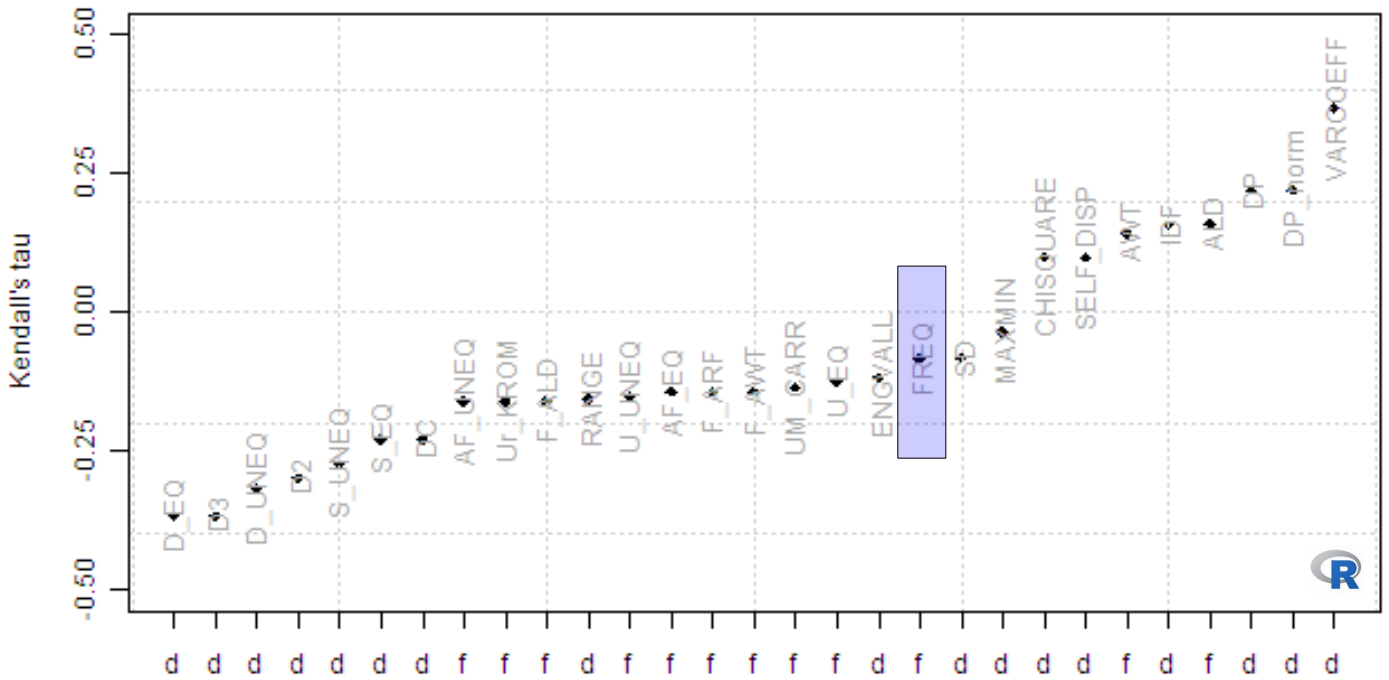
What theoretical motivation do we have to use dispersion?

- "Given a certain number of exposures to a stimulus, or a certain amount of training, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition." (Ambridge et al. 2006:175)
- learning is related to separations of exposures in time & context (Glenberg 1976, 1979)
- 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 et al. 2006:814)
- Schooler & Anderson (1997) also demonstrated that there is a power (i.e., log-log linear) function relating probability of a word occurring in the headline in the NYT on day n to how long it has been since the word previously occurred in that context. The human forgetting curve (Ebbinghaus, 1885) is rational in that it follows this trend. (Ellis, Römer, & O'Donnell 2016:37f.)

What empirical motivation do we have to use dispersion?

- You mean apart from all of the above?!
- Ellis & Simpson-Vlach (2005) & Ellis et al. (2007) show that range) has significant predictive power above & beyond raw frequency
- Adelman, Brown, & Quesada's (2006) show that range is a better & more unique predictor of RTs
- Gries (2010) shows that some dispersion measures correlate more highly than raw frequencies with
  - response time latencies from Balota & Spieler (1998)
  - lexical decision task times from Baayen (2008)
- Baayen's (2010) comprehensive analysis mentioned earlier finds that dispersion is the second strongest of 19 predictors of lex dec times
  - yes, in that study frequency is the strongest, but
  - frequency is 91% explainable from everything else, &
  - repetition frequency does little else

What empirical motivation do we have to use dispersion?



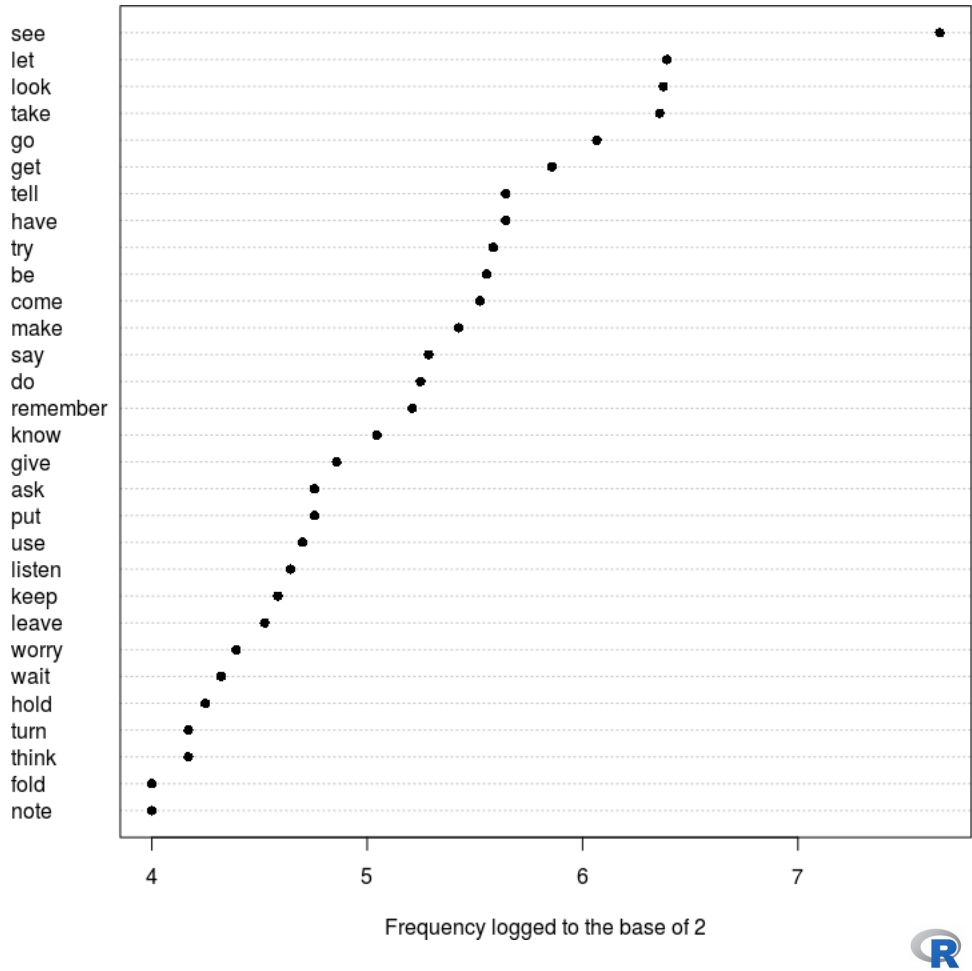
What empirical motivation do we have to use dispersion?

- An extended example: Balota & Spieler's RT data for 2820 words measured for both older & younger subjects
  - I took 6 corpora
    - BNC Baby, BNC Sampler, BNC, BNC spoken, Brown, ICE-GB
  - computed DP_{norm} of each word type across files
  - correlated each word's RT w/ frequency & DP_{norm} to determine which predicts RTs best
- results
  - frequency is never the best predictor (despite its ubiquity, strong support for Baayen, Adelman et al.)

Deviance expl. by GAM	FREQ		DP_{norm}	
	young	old	young	old
BNC Baby	4.96	7.06	8.48	14.9
BNC Sampler	5.22	6.44	9.07	13
BNC	5.06	7.57	9.26	17.3
BNC spoken	4.26	5.88	8.64	14.3
Brown	4.78	6.77	7.85	13.2
ICE-GB	3.79	4.78	6.1	9.3

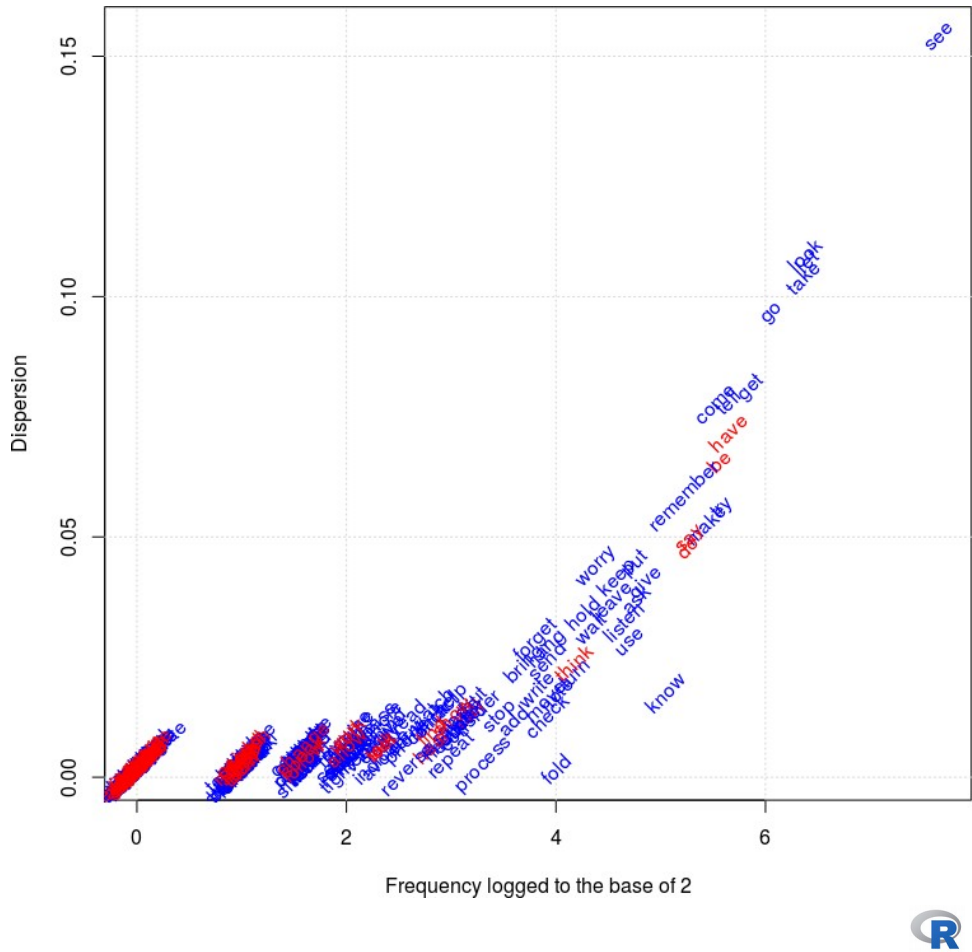
Note: I am not saying 'use dispersion *instead of* frequency' – use both!

- Let's look at a constructional slot: verbs in the English imperative
- step 1: frequency
  - ok-ish results
    - *see*, *let*, *look*, *take*, *go*, *get*, *tell*, …



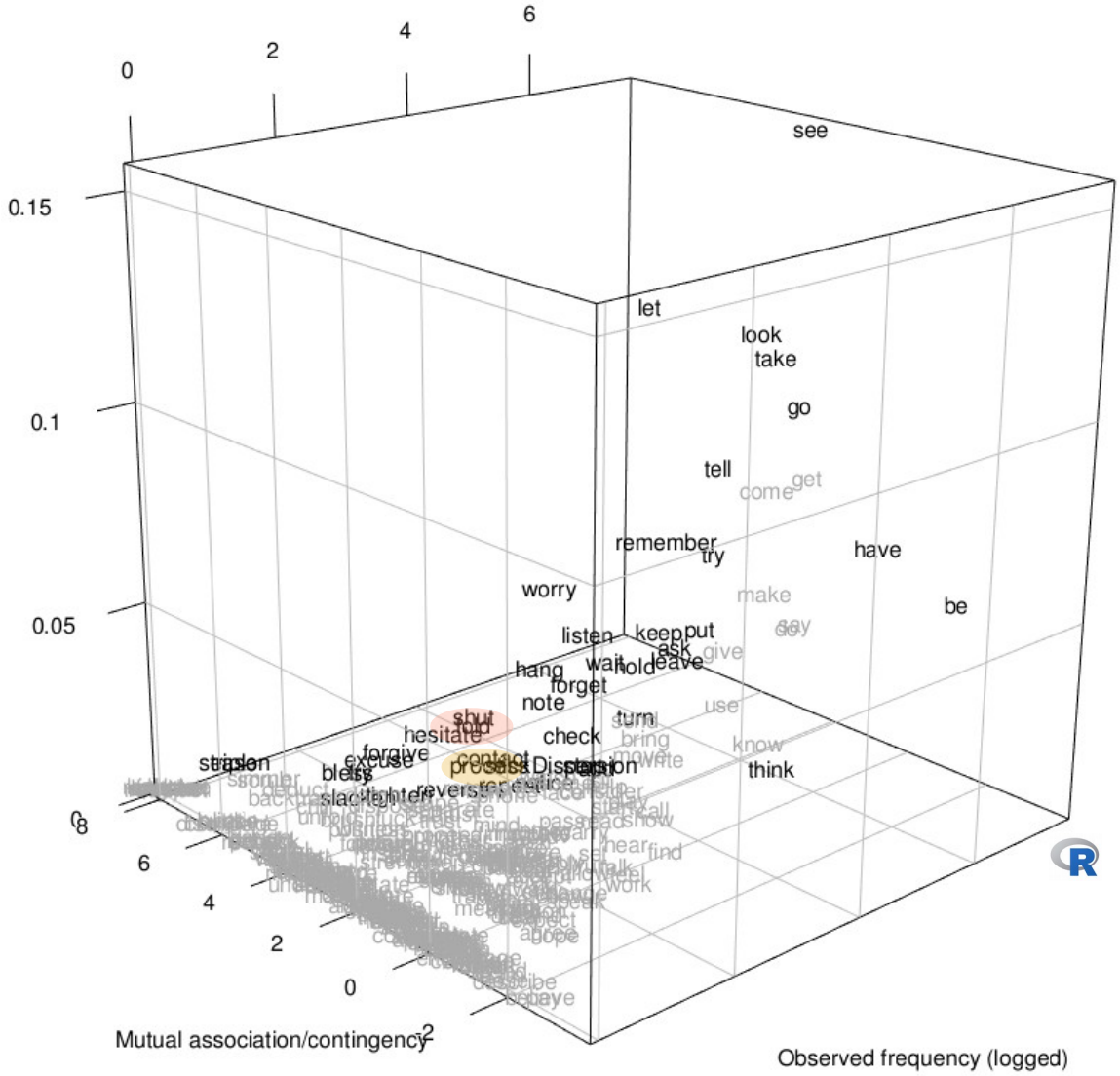
Note: I am not saying 'use dispersion *instead of* frequency' – use both!

- Let's look at a constructional slot: verbs in the English imperative
- step 1: frequency
  - ok-ish results
    - *see*, *let*, *look*, *take*, *go*, *get*, *tell*, …
- step 2: frequency & dispersion (*DP*)
  - *fold* = *process* < *know* < *use* = underdispersed
  - correlation between frequency & dispersion obvious again



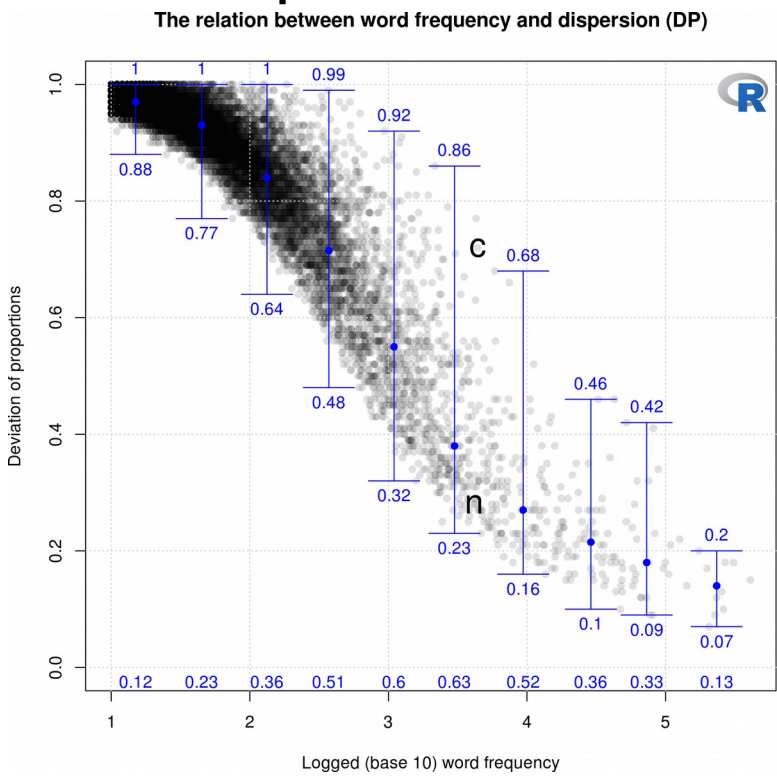
Note: I am not saying 'use dispersion *instead of* frequency' – use both!

- Step 3: frequency & mutual contingency & dispersion
  - *see* has high frequency & dispersion, but lower contingency than *let*, *worry*, *listen*, *shut* …
  - *fold* and *process* get completely downgraded for lack of dispersion



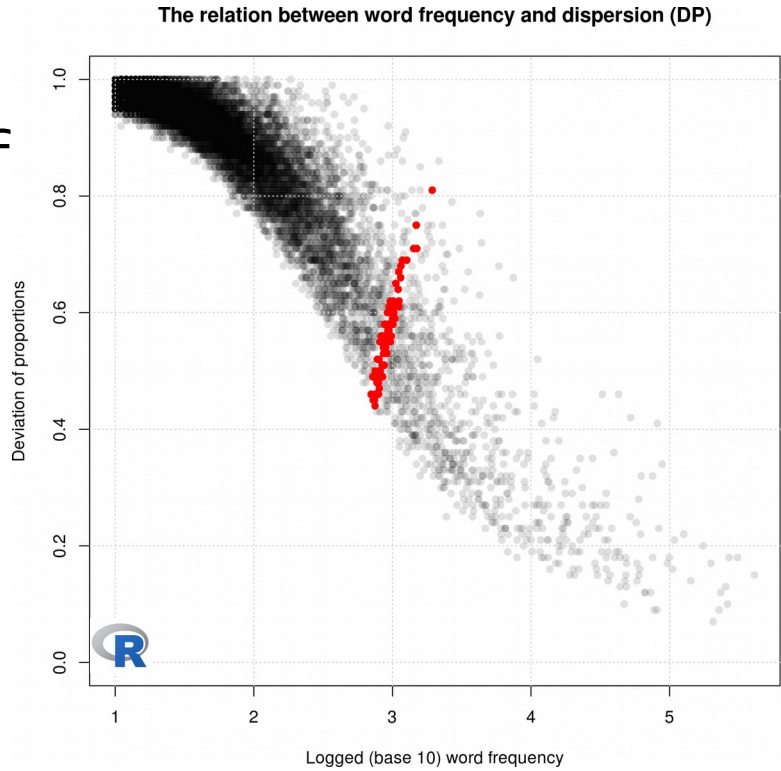
Use both, but separately!

- Freq & disp are correlated (R^2=0.83 BNCsp), but
- in the middle range of frequencies, words can have very similar frequencies but unequal dispersions
  - *staining* vs. *enormous* (in
  - *church* vs. *place* Brown)
  - in the 6th freq bin of BNCspkn
    - *council*: *freq*=4386, *DP*=0.72, range=292 *out* of 905
    - *nothing*: freq=4159, *DP*=0.28, range=652 out of 905
    - *try*: freq=4199, *DP*=0.28, range=664 out of 905
    - *whether*: freq=4490, *DP*=0.32, range=671 out of 905
    - the correlation between frequency & dispersion is low: R^2=0.08



Use both, but separately!

- Finally, the graphs also show that forcing frequency and dispersion into a single value – an adjusted frequency of the kind often used in lexicography – is a bad idea because of the information loss
  - theoretically, an adjusted freq. of 35 could result from
    - freq=350 & Juilland's *D*=0.1
    - freq= 35 & Juilland's *D*=1
  - yes, that's a hypothetical, but
    - adj.freq. for *pull* in BNCspkn≈375
    - adj.freq. for *chairman* ≈368
    - *pull*: freq= 750, *DP*=0.5
    - *chairman*: freq=1939, *DP*=0.81
  - in the plot on the right, all the red dots represent words with 365 ≤ adj. freq ≤ 434, but with 701 ≤ freq ≤ 1939



Conclusions

- All the things that frequency was supposed to affect
  - learning & acquisition, memory, processing/cognition, …
- are correlated w/ freq, but recency overrides freq
  - in the short term, as priming
  - in the longer term, as dispersion
- all corpus stats are at risk from such recency effects – the fact that occurrence or co-occurrence for any and all phenomena might not be evenly distributed across parts of a corpus: aggregate freqs are mostly useless for anything cognitive
- priming is highly predictive, cumulative, & moderated by distance, similarity
- dispersion explains more than freq-as-rep and should be computed over meaningful corpus parts
- but: keep dimensions of information separate
- with all that, freq effects we arrive at will be more accurate/reliable

Thank you!

http://tinyurl.com/stgries