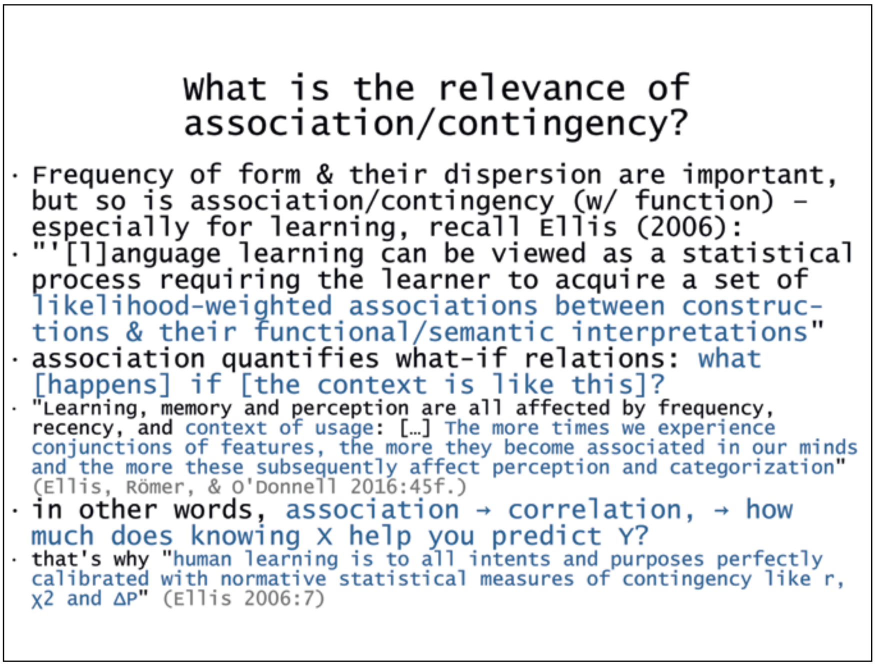
LECTURE 6

On Association

To recap, basically, I started out the talks here by talking about how frequency on its own might not be the best way to proceed. And in the last theoretical talk, I talked about one additional dimension that I think needs to be considered, which was that of recency, in particular recency, first, in the guise of priming and, second, in the guise of dispersion. So, today I want to talk about the third kind of variable or the third type of dimension that is useful for us when it comes to looking at how things behave in corpora.



What is the relevance of association/contingency?

- Frequency of form & their dispersion are important, but so is association/contingency (w/ function) – especially for learning, recall Ellis (2006):
- "'[l]anguage learning can be viewed as a statistical process requiring the learner to acquire a set of likelihood-weighted associations between constructions & their functional/semantic interpretations"
- association quantifies what-if relations: what [happens] if [the context is like this]?
- "Learning, memory and perception are all affected by frequency, recency, and context of usage: […] The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization" (Ellis, Römer, & O'Donnell 2016:45f.)
- in other words, association → correlation, → how much does knowing X help you predict Y?
- that's why "human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like r, χ2 and ΔP" (Ellis 2006:7)

FIGURE 1

All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website www.figshare.com. They may be accessed via this QR code and the following dynamic link: https://doi.org/10.6084/m9.figshare.9611465

| DOI:10.1163/9789004410343_007

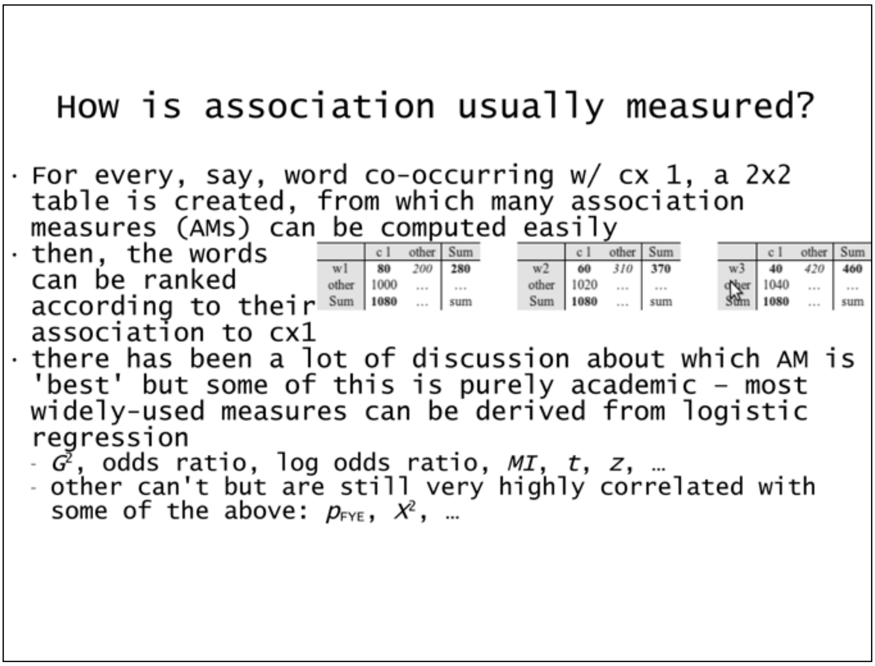
I want to start with sort of bringing back to you one of those several quotes that have been written up by Ellis (2006) and that are very insightful, and how they pinpoint a variety of things that need to come together in a good type of cognitive analysis. For instance, he said, “frequency of form and their dispersion are important, but so is association or contingency with function, and that’s especially true for learning.” Remember this quote, where he said “language learning can be viewed as a statistical process requiring the learner to acquire a set of a likelihood-weighted associations”—the topic of today’s talk—“between constructions and their functional or semantic interpretations”.

The interesting thing or the reason why association is so important is basically that it allows us to quantify *what-if* relations: What happens with some linguistic form, what happens with some linguistic function, if there is a certain context looking like this or like that, or something else? So, pretty much nothing in language happens without any context. We will always be interested in figuring out how the context of something affects either its form (the realization of it in sound or in writing), or its function, its meaning, its pragmatic intention or things like that.

The main other quote then into which we will launch from here again is this one, again you’ve seen it before every time, because it builds up very nicely to what I want to talk about. Again, the quote was that “learning, memory and perception are all affected by frequency” and that was the first theoretical talk; “recency”, that was the second, and now, third “context of usage: [...] The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization.” (Ellis, Römer, & O’Donnell 2016:45f.) So basically, we’re trying to build up, cover, all the aspects that Ellis et al. are discussing in this quote.

So, association basically is concerned with correlation again. Correlation is defined here as “how much does knowing one thing help you predict what something else will be doing? How much does knowing certain linguistic realization of something help you predict its functional impact? How much does the information structure of something help you predict a syntactic realization?” Or something like that. All of these things are what we want to look at.

Again, Ellis (2007:7) already put it very nicely by saying that “human learning is to all intents and purposes perfectly calibrated with these normative statistical measures of contingency [i.e., association like r, χ^2 and ΔP]” and then he actually lists a bunch of correlation coefficients. That essentially is what we want to look at in this talk today: So to what degree does knowing one thing help us predict some other thing where both of these things can be formal or functional realizations of various types of constructions at various levels of granularity or resolution?



How is association usually measured?

- For every, say, word co-occurring w/ cx 1, a 2x2 table is created, from which many association measures (AMs) can be computed easily
- then, the words can be ranked according to their association to cx1

	c 1	other	Sum
w1	**80**	*200*	**280**
other	1000	…	…
Sum	**1080**	…	sum

	c 1	other	Sum
w2	**60**	*310*	**370**
other	1020	…	…
Sum	**1080**	…	sum

	c 1	other	Sum
w3	**40**	*420*	**460**
other	1040	…	…
Sum	**1080**	…	sum

- there has been a lot of discussion about which AM is 'best' but some of this is purely academic – most widely-used measures can be derived from logistic regression
  - G^2, odds ratio, log odds ratio, *MI*, *t*, *z*, …
  - other can't but are still very highly correlated with some of the above: p_{FYE}, X^2, …

FIGURE 2

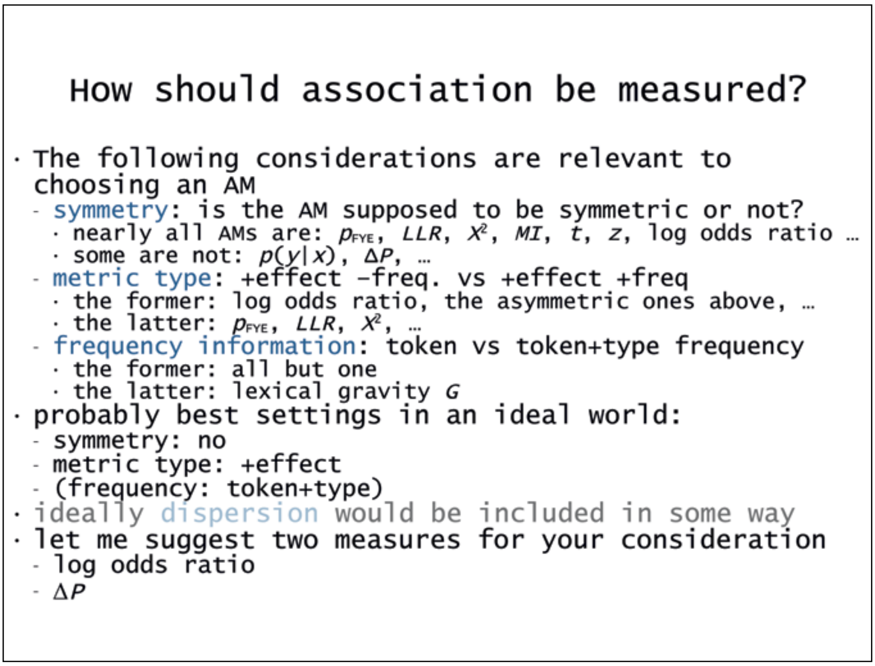
Now how is this usually measured? Typically, it proceeds in a way that you've actually seen one time before very briefly in the frequency talk. The example I want to use is that of verb-construction co-occurrence, because so much of my own work has been concerned with things like this. The idea is that for every word, let's say, that occurs with a certain construction, like construction$_1$ for lack of a better term right now, you draw up a 2×2 table like this [pointing to the first table in Figure 2] from which you can compute association measures. This kind of table here would say basically that word$_1$ occurs in your corpus 280 times; the construction that you're looking at occurs 1,080 times; and 80 of these two [refers to word$_1$ and the construction], they co-occur together. The other uses of word$_1$: the remaining 200 are with other constructions, not with that one that we're currently looking at. The idea is that you do that for multiple words, ideally for all the words that occur in that construction. So this is the same table, but for word$_2$ in construction$_1$. The construction frequency is the same, and it's still 1,080. But now there's another word that you're interested in, which occurs 370 times in the corpus, but only 60 times with this construction. And here's word$_3$ occurring 460 times in the corpus, 40 times in the construction, and it's the same construction, so again, 1,080, 1,080, 1,080. Then, when you have an association measure for word$_1$, for word$_2$, for word$_3$, for each of these 2×2 tables, you compute an association measure and then when you have all those measures, you can rank all the words by their association to that construction. As we've seen in a ton of literature that has been using this collostructional approach, a lot of times you'll find that the words that like to occur in certain constructions, they share certain semantic characteristics, sometimes they share certain information-structural characteristics, and other kinds of things and so they allow us to interpret semantics of constructions for instance.

One big issue of discussion for many years now has been, in corpus linguistics actually for decades, is what association measure to use. There's a lot of different measures that can be applied to a deceptively simple 2×2 table like this. Many of you may have heard of a chi-squared test, for instance, as a test that is routinely used for tables like this. But the chi-squared test, for instance, makes some assumptions that a lot of times are violated with these kinds of data. So people have come up with literally many dozens of association measures that can be used to quantify the strength of association in these tables.

As I say here, much of the discussion, however, is actually purely academic because many of the measures that are being used most of the time in corpus linguistic approaches are all just different ways of interpreting logistic regression results. So a logistic regression is a regression that tries to predict [or] model something binary, namely this construction in question (construction$_1$) or another one (other constructions) on the basis of a binary predictor, namely this word (word$_1$) or another one (other words). So actually, it would look like overkill, but you can, instead of doing a chi-squared test on a table like this, you can do a logistic regression on a table like this and the results will be typically at least very, very similar.

So once you adopt this regression-modeling perspective, then actually many of the different measures that people have been hitting each other over their head with are all very, very comparable. For instance, the most frequently-used measure probably is this one, G^2 (which some of you may know as the log-likelihood ratio), odds ratio, and log odds ratio: All of these values actually come from a logistic regression. If people sort of debate whether using this or this is better, they're actually not debating very much, in the sense that these will all be extremely highly correlated.

Some other measures cannot be derived from logistic regression, like the ones that I mentioned here, Fisher-Yates exact test (p_{FYE}), the test that a lot of people have been using in collostructional kind of analyses, or chi-squared. But they're still extremely, highly correlated with anything that comes out of a logistic regression. If you, in general, are interested in looking into this kind of issue more—because association is an important concept in usage-based



How should association be measured?

- The following considerations are relevant to choosing an AM
  - symmetry: is the AM supposed to be symmetric or not?
    - nearly all AMs are: p_{FYE}, *LLR*, X^2, *MI*, *t*, *z*, log odds ratio …
    - some are not: $p(y|x)$, ΔP, …
  - metric type: +effect −freq. vs +effect +freq
    - the former: log odds ratio, the asymmetric ones 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*
- probably best settings in an ideal world:
  - symmetry: no
  - metric type: +effect
  - (frequency: token+type)
- ideally dispersion would be included in some way
- let me suggest two measures for your consideration
  - log odds ratio
  - ΔP

FIGURE 3

linguistics—then learning something about binary logistic regression is probably time well spent, because it will help you understand all sorts of debates and all sorts of results that have been published on these kinds of questions.

Now, if this is how association usually measured, then how *should* it be measured? There's a bunch of characteristics is that you should pretty much always consider when you talk about, or when you consider, which association measure you think is best for your particular case study. The first one is this, namely the question of symmetry. Nearly all association measures are symmetric. All the ones that are listed here, Fisher-Yates exact test, log-likelihood ratio, chi-squared, Mutual Information, all these statistics basically are symmetric and by *symmetric*, I mean they quantify how much a verb and construction are attracted *to each other*. Another way of using symmetry or describing symmetry here would be that the association is bidirectional: The word likes the construction, and the construction likes the word to the same degree. That's what is meant by *symmetry* here.

But there are some measures ($p(y|x)$, ΔP,…) that are not symmetric, so that means these measures here. This would be a conditional probability: What is the probability of this construction given this verb? That would be different from what is the probability of this verb given this construction. So with these measures, you *can* distinguish between cases where a verb likes a construction a lot, but the construction doesn't like the verb a lot. You can keep those things apart. That's probably a useful thing, because it's not really obvious at all that associations that we form in our minds as part of a learning process would be symmetric. Usually, if only temporarily, we see something first, and then we see something else so chances are that that has at least some kind of impact on the degree [and direction] of association we form between these things.

The second important characteristic, and that one has been debated particularly hotly in collostructional analysis literature, namely, is the type of metric (+effect −freq. vs. +effect +freq.) that your association measure is. There's essentially two options, to simplify a little bit here. One of the two metrics is this, namely, the association measure reflects association (+effect), but it does not reflect frequency (+effect −freq.). Whereas the association measures that are mostly used are of the latter type, so they reflect association strength, but also frequency (+effect +freq.). So in a way, or one other way to look at it would be that, some metrics measure only one dimension, namely, how strong is the association and I don't care in how many data points I observe this. Other metrics measure an association, but also take into consideration the sample size, the number of items you have, and give you that back in one number. So measures that do not include frequency would be something like the log odds ratio or the asymmetric measures that I've mentioned here, conditional probabilities and Δ*P*. The most widely used ones, actually, like log-likelihood ratio, Fisher-Yates exact test, they react to both the association strength and the frequency with which something has been observed.

It's still an ongoing debate which of these two scenarios is better. I'll talk a little bit about what I'm thinking, but just to give you a heads-up already, this one [pointing to the latter type] is simpler to use, because for every word and construction pairing, [[for example,]] for *give* in ditransitive, for *tell* in ditransitive or something, it gives you one value and that value reflects both effect and frequency: It's easy if you're statistics-averse and you want just one little value to sort by.

However, like I've already indicated the other day when we talked about dispersion, there I said, conflating of frequency and dispersion into one adjusted-frequency value loses a lot of information. That, of course, happens here as well. The reason why one might consider something like this [pointing to the former type], a measure that only measures association strength but not frequency, is to keep your data clean. The statistic that you're reporting only looks at effect size, because frequency, you have that anyway, so that would be the second axis in a plot. That's something we'll look into later.

Then, third: frequency information. Pretty much all of the association measures that are widely used only use token frequency. That means, these 2×2 tables that you've seen before, you don't know how many *different* other constructions a word shows up in.

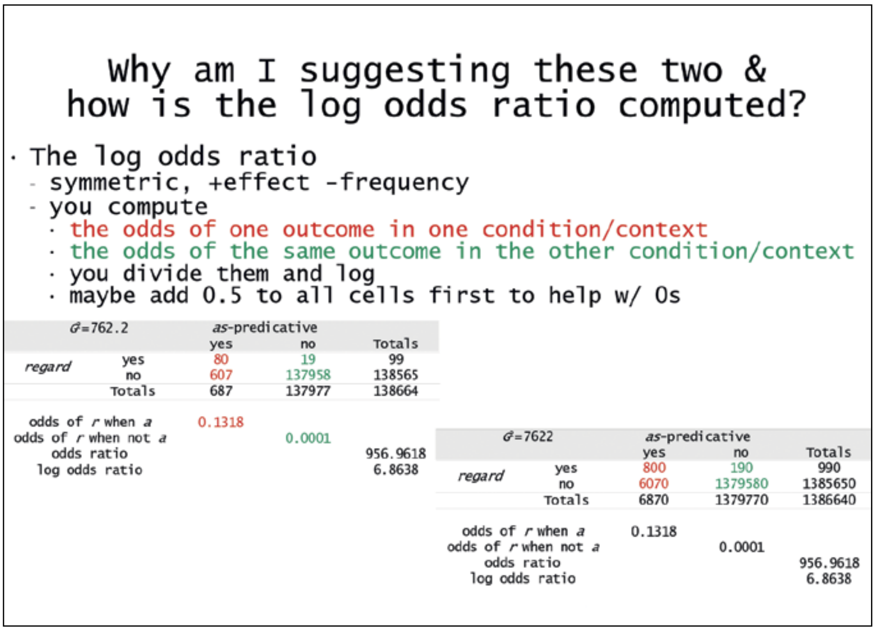
Let me go back a real quick to show you what I mean here. In this case [[Figure 2]], we know word$_1$ shows up 80 times with this construction, and we know it shows up 200 times with other constructions—but you don't know how many different constructions these 200 other ones are—could be 1 or could be 200, but we don't know. Pretty much all association measures but one that I know at least work like that: they just take the 200 and they do not consider how many other competing constructions are there. The only measure that does use type frequency as well is a measure that is hardly ever used, namely, lexical gravity *G*. Computing it is a little bit more involved, but theoretically, of course, it seems like a very useful idea in fact.

Now the best settings ideally would be, probably, to use a measure that is not symmetric so that you can distinguish cases from where the verb likes the construction, but the construction does not like the verb, or the other way around.

Second, probably, at least for cognitively, supposedly, realistic analysis, you probably want to use only an effect size [[+effect]] here, so that your results for association are not tainted also by frequency, but you keep those two things separate. Frequency, ideally, one would be able to use both token and type frequency, although no one has done that yet, for reasons that may become apparent later.

Then, ideally, you would also include dispersion because we've already seen that a co-occurrence frequencies like these can be very misleading depending on how high the dispersion is. If this corpus has 20 parts and all the 80 co-occurrences are in one of those 20 parts, whatever you're doing with that, you should know that, as opposed to these 80 being distributed all over the corpus being very representative. So, ideally, we would include dispersion here as well and I'll show you a little bit about how to do this at a later point. Now, obviously, if there are many dozens of association measures—the last overview paper that I've seen discussed 80—and since then at least one or two others have been developed as well, so there's more than 80 of those, what should you be using? Let me suggest two here for your first consideration at least. The first one is the log odds ratio, and the second one is ΔP. Let me now tell you why I think these two are useful and should be considered.

The log odds ratio is a symmetric measure. So, that's already kind of a downside. But it has another good thing, and that is that it only is a measure of association and it does not reflect frequency so it's cleaner than something like



Why am I suggesting these two & how is the log odds ratio computed?

- The log odds ratio
  - symmetric, +effect -frequency
  - you compute
    - the odds of one outcome in one condition/context
    - the odds of the same outcome in the other condition/context
    - you divide them and log
    - maybe add 0.5 to all cells first to help w/ 0s

G^2=762.2		*as*-predicative		
		yes	no	Totals
regard	yes	80	19	99
	no	607	137958	138565
	Totals	687	137977	138664
odds of *r* when *a*		0.1318		
odds of *r* when not *a*			0.0001	
odds ratio				956.9618
log odds ratio				6.8638

G^2=7622		*as*-predicative		
		yes	no	Totals
regard	yes	800	190	990
	no	6070	1379580	1385650
	Totals	6870	1379770	1386640
odds of *r* when *a*		0.1318		
odds of *r* when not *a*			0.0001	
odds ratio				956.9618
log odds ratio				6.8638

FIGURE 4

log-likelihood ratio or something like that. So how's it computed? Let's use this construction and this verb as an example. We're again looking at the *as*-predicative construction. So that is this construction, what did I say, whatever, *He was regarded as a very famous linguist*, that would be an example. *He saw himself as a very important linguist*, *He described himself as a very important linguist*, *He considered himself as a very important linguist*" this kind of construction. So, verb, a direct object, *as* and then something. *This attack was widely regarded as being out of the blue*, that would be another example.

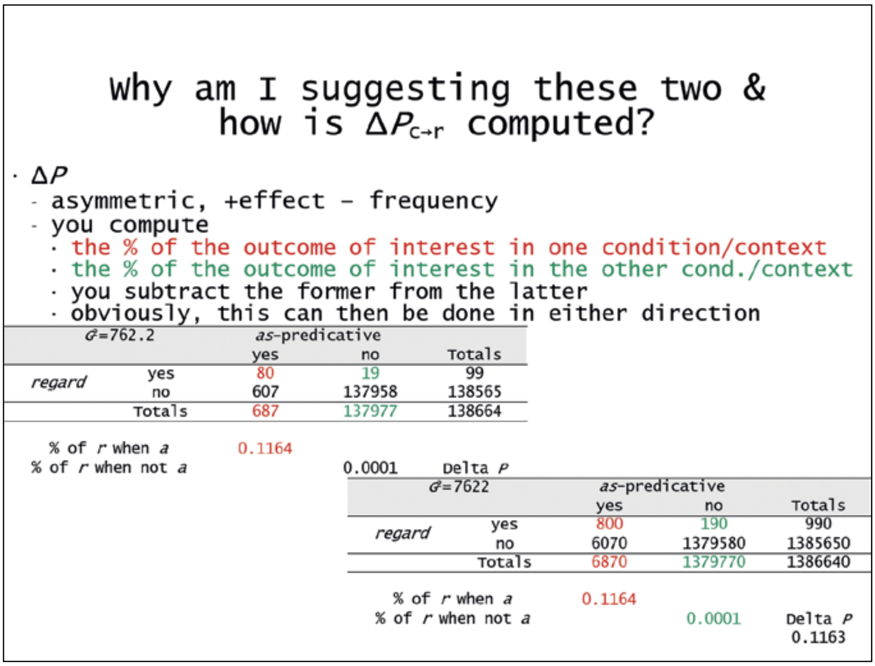
So we're looking at this construction, either that (*regard* in *as*-predicative) it's there or it's not, yes or no, and we're looking at this verb *regard*, which is either there or not. In the corpus that we're looking at, the corpus contains this many verbs. That's what we're using as a unit of sampling here. Of those verbs, 99 cases are *regard*. And of those 99 cases of *regard*, 80 are in this construction. So most, the vast majority, I mean 80 percent, pretty much, and 19 uses of *regard* are not in that construction. The construction has a frequency of 687, and 80 of those, so one eighth essentially kind of, are with *regard*, and then there are 607 others.

So how do you compute the log odds from this? I don't know how many of you bet on horses or something, but it's basically that type of odds. So you divide the number of times that something happens of interest, this number here [referring to 80], by the other option that [referring to 607]. So 80 divided by 607 is 0.1318. These are the odds of *regard* being used when the construction is an *as*-predicative. There are 687 *as*-predicatives, and the odds of *regard* are this (*regard* in *as*-predicative) versus this (*regard* not in *as*-predicative). This many cases of *regard* "yes", compared to this many cases of *regard* "no", when the construction is in fact an *as*-predicative. So those (0.1318) are the odds of *regard* when the construction is an *as*-predicative.

Then you compute the same thing here: what are the odds of *regard* when the construction is not *as*-predicative? And as you can see they're tiny: 19 divided by 137,958. I stopped here, I rounded it off at four decimals, but obviously it's very small. So then the odds ratio is this (0.1318) divided by that (19/137,958), which gives you this number (956.961) and then you log it. That's the log odds ratio (6.8638), and this is a pretty damn high value on that scale. Again: what are the odds of *regard* versus not when it's the construction of interest divided by what are the odds of *regard* compared to it's not when it's all other constructions, and then this divided by this, log—so relatively straightforward. You can do this with any spreadsheet, even if you wanted to do it with a pocket calculator, not particularly tricky. Sometimes, what you need to do is, if one of these numbers is zero, or in cases, some of these numbers are zero that you add 0.5 to every number first, and then you do the computation that I showed here.

The important thing here to realize: so, first, it's symmetric. This is the number that says how much the verb *regard* and the construction *as*-predicative like each other. Second, like I said, this does not include frequency information. This is counter to other measures. Here, actually, I'm not showing you how this is calculated, but this ($G^2 = 762.2$) is the log likelihood value for this table, 762.2 is also super high. Now, what happens if we pretend we had a corpus ten times as big as the one that we're using here? We are multiplying every one of these numbers by ten. What happens then is this: The odds ratio, the log odds, all stay the same. If this is ten times larger, and you divide it by this, which is now ten times larger, of course, you get the same number. But look at this ($G^2 = 762.2$). That value ($G^2 = 762.2$) went up by a factor of ten. So, this value doesn't separate corpus size and association—it conflates them into one value, whereas this one (log odds ratio = 6.8638) nicely only includes the association strength. So that's one of the potential selling points for this measure.

Then, what about the other one, ΔP? The thing about ΔP is that it also does not grow if the corpus becomes bigger. Just like the log odds ratio, that's what they have in common but ΔP is asymmetric: So ΔP can distinguish how much the verb likes the construction from how much the construction likes the verb. That's why here, I'm writing how is ΔP and then the c→r means column/row.



Why am I suggesting these two &
how is $\Delta P_{c\to r}$ computed?

- ΔP
  - asymmetric, +effect – frequency
  - you compute
    - the % of the outcome of interest in one condition/context
    - the % of the outcome of interest in the other cond./context
    - you subtract the former from the latter
    - obviously, this can then be done in either direction

G^2=762.2		*as*-predicative		
		yes	no	Totals
regard	yes	80	19	99
	no	607	137958	138565
	Totals	687	137977	138664

% of *r* when *a* 0.1164
% of *r* when not *a* 0.0001 Delta *P*

G^2=7622		*as*-predicative		
		yes	no	Totals
regard	yes	800	190	990
	no	6070	1379580	1385650
	Totals	6870	1379770	1386640

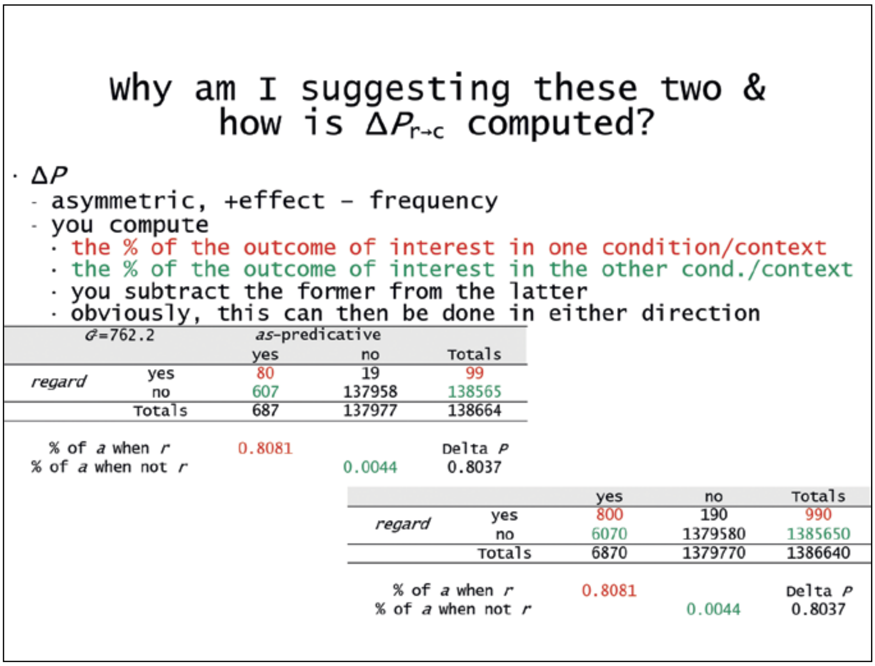
% of *r* when *a* 0.1164
% of *r* when not *a* 0.0001 Delta *P* 0.1163

FIGURE 5

So how much does ΔP from whatever is in the columns, the construction, like whatever is in the rows, the verb, how is that ($\Delta P_{c\to r}$) computed? We're looking at the same example. And it's actually kind of similar. So this is the same table as before.

What you compute is you compute the percentage of the outcome of interest in one condition. If there's an *as*-predicative, 687 times, how often in percent is that *regard*? We're computing how much is 80 out of 687? So it's 0.1164. This is really just saying 11.6% of the *as*-predicatives are with *regard*. Then we do the same: how often is a verb *regard* when it's *not* in the *as*-predicative? That's, of course, super rare. There's a buttload of verbs in general, but only 19 of those are *regard* because in general it's not a frequent word. Then you just subtract this (0.1164) minus this (0.0001), and that (0.1163) is the value. What this tells you is, how much does knowing that the construction is the *as*-predicative help you expect *regard*? When the construction is not the *as*-predicative, *regard* is super rare. But when the construction becomes the *as*-predicative, it's quite common, and ΔP is the difference between those two. So it ranges from minus one to plus one and the higher, the stronger the attraction.

Here we're going from columns to rows. We're predicting from the absence or the presence of a construction which verb is going to happen. Here too, this



Why am I suggesting these two & how is $\Delta P_{r \rightarrow c}$ computed?

- ΔP
  - asymmetric, +effect – frequency
  - you compute
    - the % of the outcome of interest in one condition/context
    - the % of the outcome of interest in the other cond./context
    - you subtract the former from the latter
    - obviously, this can then be done in either direction

G^2=762.2		*as*-predicative		
		yes	no	Totals
regard	yes	80	19	99
	no	607	137958	138565
	Totals	687	137977	138664
% of *a* when *r*		0.8081		Delta *P*
% of *a* when not *r*			0.0044	0.8037

		yes	no	Totals
regard	yes	800	190	990
	no	6070	1379580	1385650
	Totals	6870	1379770	1386640
% of *a* when *r*		0.8081		Delta *P*
% of *a* when not *r*			0.0044	0.8037

FIGURE 6

thing does not reflect frequency: If we multiply that whole table by ten again, Δ*P* is the same. So, [[it is]] very nice, keeping things separate. This is the example for from the construction to the verb.

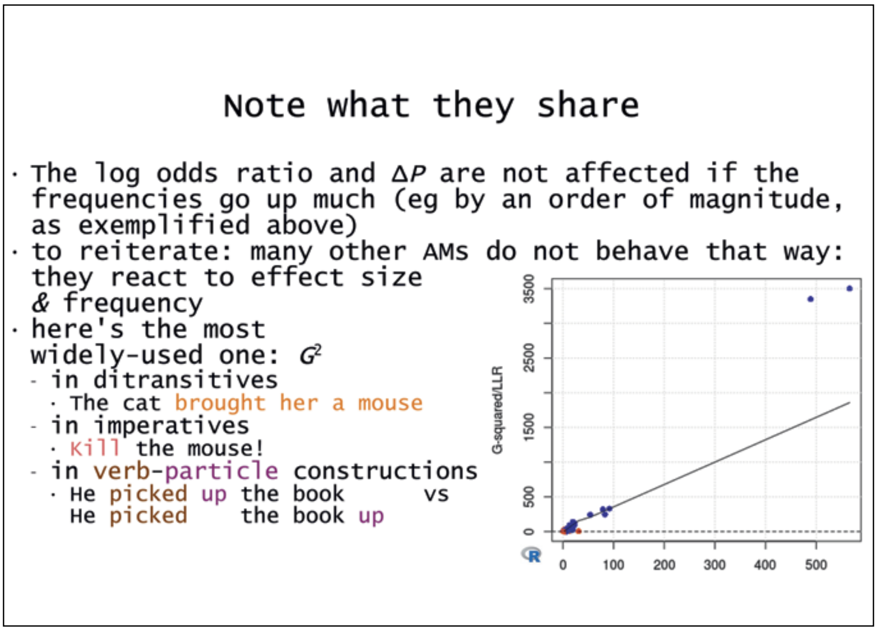
Obviously, since this is asymmetric, we now also need to look at from the verb to the construction. Somewhat confusingly, but there was no other way to show it in a parallel way in a spreadsheet. So now we're saying if the verb is *regard*, how often do we see *as*-predicative? This (80) out of this (99) is a staggering 80%. If you see *regard*, you can be pretty certain, it's in an *as*-predicative.

But then the other one is how often do you see an *as*-predicative if the verb is not *regard*? Not very much. Again, the Δ*P* value is the difference between the two. You can see this one (0.8037) is super high. That's what Δ*P* or any asymmetric measure buys you. This measure here can 'see' that, well, if the construction doesn't actually attract a verb that much, but the verb attracts the construction super strongly.

I'll come back to this a little bit later in a moment. But, since it fits here right now, let me mention it already. There's a lot of cases where corpus linguists and psycholinguists alike, they're very annoyed at the fact that sometimes corpus data don't match up nicely with experimental psycholinguistic data. But, for instance, I've seen one example, someone used corpus data and tried to correlate association measures from corpus data with the result from an association experiment, sort of, given one word, and then someone was supposed to give words that they associate with that word. Then the corpus-linguistic author basically correlated the psycholinguistic results with corpus associations of this type. But what she did is, she used a bi-directional association measure when of course the experimental task was totally directional. Namely, you get one word and you're supposed to go from that verb to somewhere else. So part of the mismatch of course could very well be that she used an association measure that actually is not compatible with the experimental task to which she is trying to compare the corpus data.

Same thing with the sentence completion task: If you do a sentence completion experiment, and the sentence fragment ends in a verb and then you look at how do people complete that sentence. That's a directional question. You give a verb and you expect to see a construction after that. So you need this: how much does a verb boost the appearance of a construction? Not the other way around, and not something symmetric.

That's one of these cases where psycholinguists and cognitive linguists always like, "well, corpus data, they are not really that great" and then they use their wrong measures and you're like, "yeah, of course they're not that great, if you don't do the math right". Here. too, again, this does not change, if the corpus size gets much, much bigger. So, quite an attractive measure.



Note what they share

- The log odds ratio and Δ*P* are not affected if the frequencies go up much (eg by an order of magnitude, as exemplified above)
- to reiterate: many other AMs do not behave that way: they react to effect size & frequency
- here's the most widely-used one: *G*^{2}
  - in ditransitives
    - The cat brought her a mouse
  - in imperatives
    - Kill the mouse!
  - in verb-particle constructions
    - He picked up the book vs He picked the book up

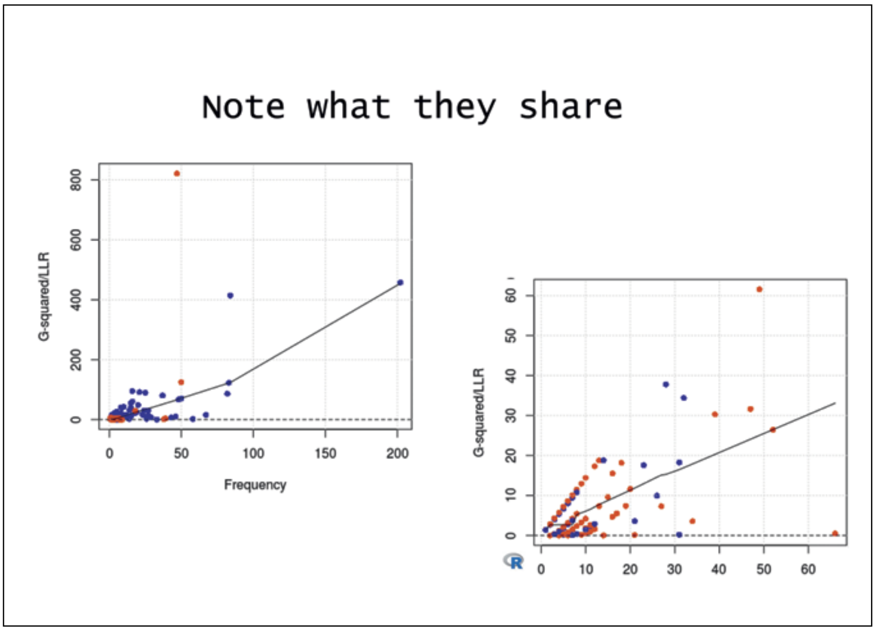
FIGURE 7

So what do they share? Again to recap, both of them are not affected by the corpus size increasing a lot, but actually increasing at all. If everything is as before, then the measures will stay the same. And that is not how most measures react: Many standardly-used measures, they do incorporate both. Again, I think there is an area of application for that: If you're interested in a simple-sorting kind of result, but for anything that aims to be cognitively realistic, I think this is not the way to go.

So let me show you three examples using the most widely used association measure, the log likelihood ratio. Here's an example: We're looking at the ditransitive constructions, like *The cat brought her a mouse*. This plot here, what it shows is the frequency with which a verb occurs in that slot and the association measure, so every blue or red point is one verb in that construction and you can see there's a relatively strong correlation. The regression line doesn't capture quite those here but you can see that on the whole, there's an upward trend: As the frequency goes up on the whole, the points go higher up, even if most of the values are clustered down here, because of the Zipfian distribution.

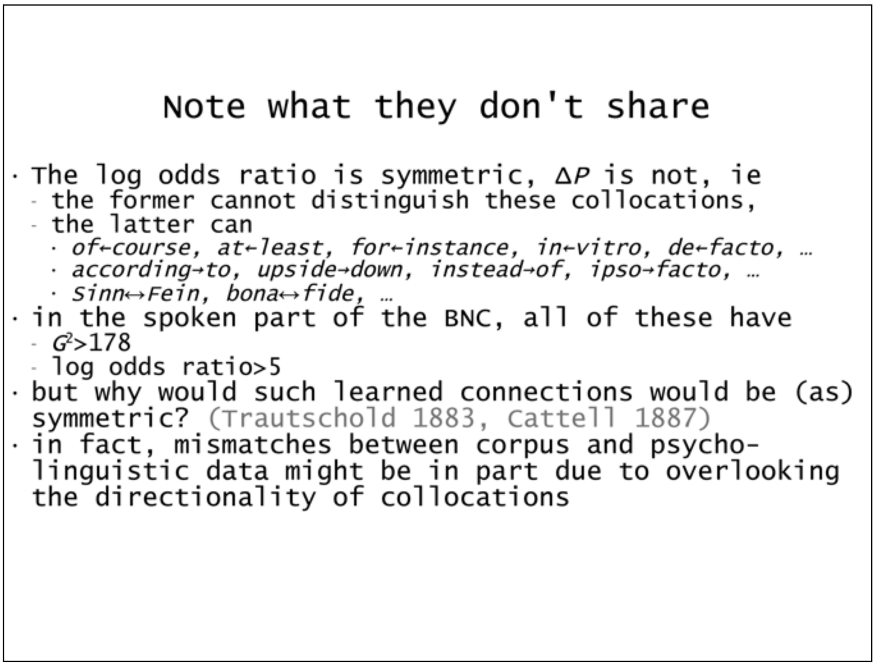
Same thing with imperatives, so which verbs go into this slot (*Kill the mouse!*)? Again, there are some outliers, but on the whole, there's a positive relation as frequency increases, so does this association measure.

Finally, for this one, that's another example that we'll discuss later, verb-particle constructions. So the two constructions in question are *picked up the*



Note what they share

FIGURE 8



Note what they don't share

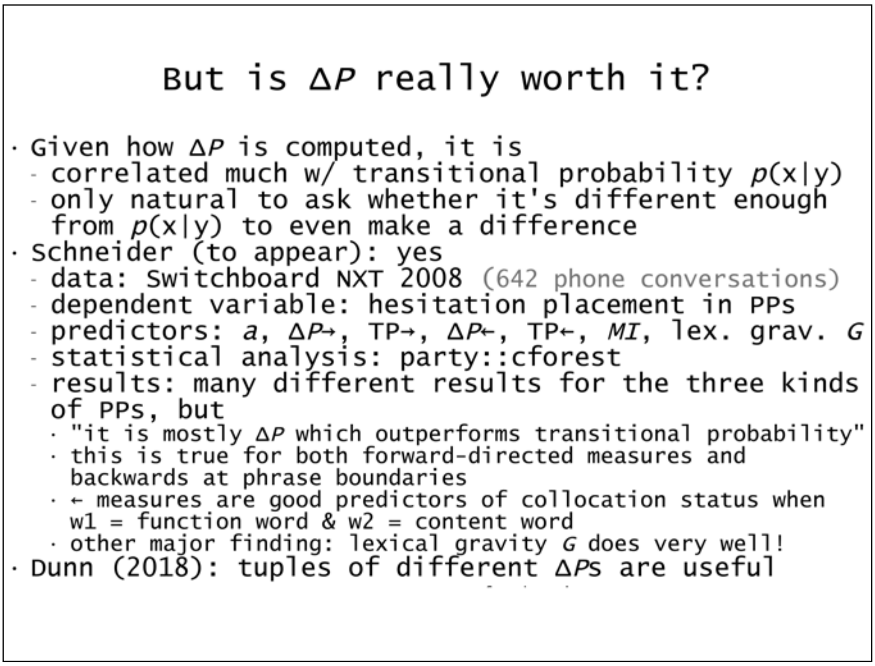
- The log odds ratio is symmetric, ΔP is not, ie
  - the former cannot distinguish these collocations,
  - the latter can
    - *of←course, at←least, for←instance, in←vitro, de←facto, …*
    - *according→to, upside→down, instead→of, ipso→facto, …*
    - *Sinn↔Fein, bona↔fide, …*
- in the spoken part of the BNC, all of these have
  - G^2>178
  - log odds ratio>5
- but why would such learned connections would be (as) symmetric? (Trautschold 1883, Cattell 1887)
- in fact, mismatches between corpus and psycholinguistic data might be in part due to overlooking the directionality of collocations

FIGURE 9

mouse and *pick the mouse up*. Again, there's a very clear correlation between the association measure and frequency here. So this is potentially a problem: if you're interested in keeping those dimensions separate, then this measure does not do that. Whatever you measure here, to a large extent, it actually also reflects this, you're not keeping them separate.

Now, what do they *not* share? Like I said, the log odds ratio is symmetric so it is somewhat less informative if you want to put it that way because the log odds ratio will not be able to distinguish certain kinds of collocations from each other. These are all collocations where the second word is highly predictive of the first. If I ask you what word might you expect in front of the word *instance*, of course you're going to say *for*. If I ask you what word you might expect in front of *least*, it's very likely that you would say *at*, maybe you would say *the*, *the least I can do* or something. Here this one is particularly nice, in front of *facto*, what's there going to be other than *de*?

But there are collocations where it's the other way around. I mentioned this example earlier, I think. If I ask you what comes after *according*, of course, you're going to say *to*; if I ask you what's after *instead*, of course, you're going to say *of*. But if I ask you what's in front *of*, chances are you give me a whole bunch of different things as well so that correlation is not that strong. And so here we



But is Δ*P* really worth it?

- Given how Δ*P* is computed, it is
  - correlated much w/ transitional probability *p*(x|y)
  - only natural to ask whether it's different enough from *p*(x|y) to even make a difference
- Schneider (to appear): yes
  - data: Switchboard NXT 2008 (642 phone conversations)
  - dependent variable: hesitation placement in PPs
  - predictors: *a*, Δ*P*→, TP→, Δ*P*←, TP←, *MI*, lex. grav. *G*
  - statistical analysis: party::cforest
  - results: many different results for the three kinds of PPs, but
    - "it is mostly Δ*P* which outperforms transitional probability"
    - this is true for both forward-directed measures and backwards at phrase boundaries
    - ← measures are good predictors of collocation status when w1 = function word & w2 = content word
    - other major finding: lexical gravity *G* does very well!
- Dunn (2018): tuples of different Δ*P*s are useful

FIGURE 10

have *facto* again, actually" if you start from *facto*, then you end up at *de*. But *facto* is also the thing that people would say occurs after *ipso*. And then there are cases which are completely predictable in both directions, at least in some corpora, *bona*↔*fide* or *Sinn*↔*Fein*, perfectly predict each other. And so the log odds ratio will treat them all the same. It would not distinguish between them whereas Δ*P* would establish these three groups.

Like I said, log odds ratio for all of those is greater than 5, pretty high. But there's no reason to assume that these kinds of associations are, in fact, symmetric because, like I said, if only because of time, there's always going to be one thing first, and then the other thing second. Then we might interpret correlation or co-occurrence but it's not obvious at all that it would be symmetric.

Now one question you might have though is whether the effect of Δ*P* is actually worth it. Do we really need to compute this? Because given how Δ*P* is computed, it's extremely highly correlated with transitional probability.

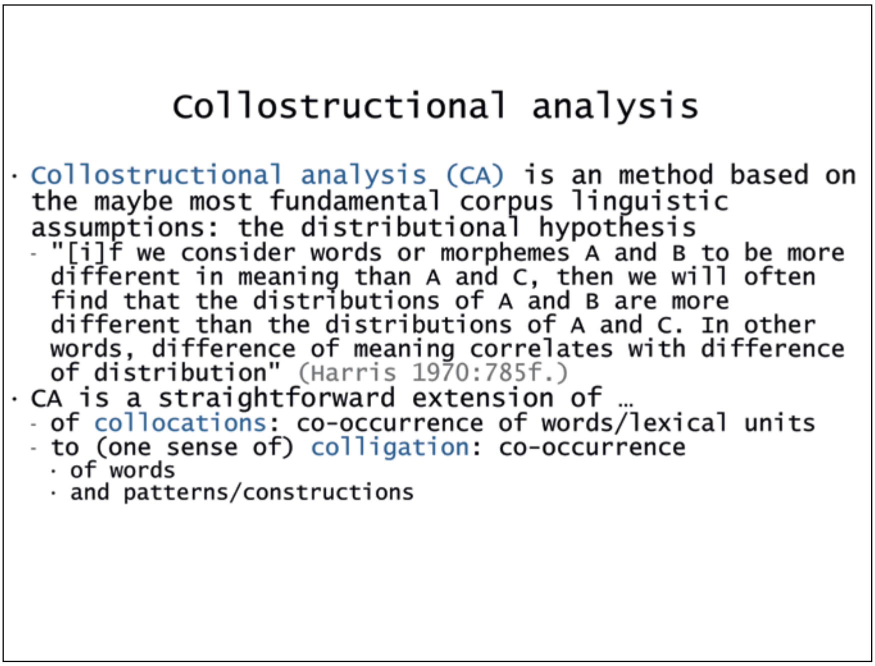
Let me actually show you that in the table again. Δ*P* is this (800) divided by that (990), that's one probability, minus this (6,070) divided by that (1,385,650), that's another. So obviously, these two (0.8081 and 0.0044) are transitional probabilities. And Δ*P* (0.8037) will be always be very highly correlated with this one (0.8081). Why? Because usually, the *d* cell here, which is not this construction and not this verb, is usually very high, just like here (i.e. 1,379,580). If you look at one word and one construction, then of course most of the corpus is going to be other things. So that number (i.e. 1,379,580) is always going to be very high, which makes this (i.e. 6070) divided by that (i.e. 1,385,650) very small. So Δ*P* will be very highly related to this (i.e. 0.8081). And so there have been people actually who suggested that just take this number (i.e. 0.8081), forget about the normalization, you know, minus this (i.e. 0.0044) to that (i.e. 0.8037).

So the question might be, is it even worth it to compute it like that? Maybe we can do it without it. But there have been some first studies now that show that it's worth it. So Schneider's (2018) book, as a part of her dissertation, did a study where she looked at data from the Switchboard corpus so phone conversations between strangers put together on a switchboard. She looked at the hesitation placement in prepositional phrases. Where do people become slower because they're entering into an area where there's production planning difficulties? And she compared a whole bunch of different predictors, co-occurrence frequency, Δ*P* in one direction, the transitional probability in that same direction, Δ*P* in the other direction, the transition probability in the other direction, other collocation measures, and so on, and then she did a random forest analysis on this dataset, really nice.

She did find a lot of different kinds of results for three different kinds of differently complex prepositional phrases but one of her main conclusions is this, namely, "it is mostly Δ*P* which outperforms transitional probability"; "both for forward-directed measures and backward-directed ones". So yes, it does come with some work, but there will be enough cases to make it worth your while, which is essentially what she's finding.

Then in another paper, I'm not sure it's still to appear. Just this morning I saw a reference to it on ResearchGate, and it's said at 2018, so a study in the *International Journal of Corpus Linguistics* where James Dunn look at a whole bunch of different Δ*P* values and he also found that it is an extremely useful concept. So yes, mathematically, there will be a high correlation, but that should not detract from the fact that, on the whole, the higher degree of precision of Δ*P* is better. If you're in the market for an association measure, so to speak, then Δ*P* is probably a pretty good one.

Now how this might be applied? This is where I do want to talk a little bit about collostructional analysis, even though it's kind of dated by now, but still a lot of people are using it. It's a method that, like everything else in corpus linguistics, is based on the distributional hypothesis, which I'm giving here again, "If 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 from A



Collostructional analysis

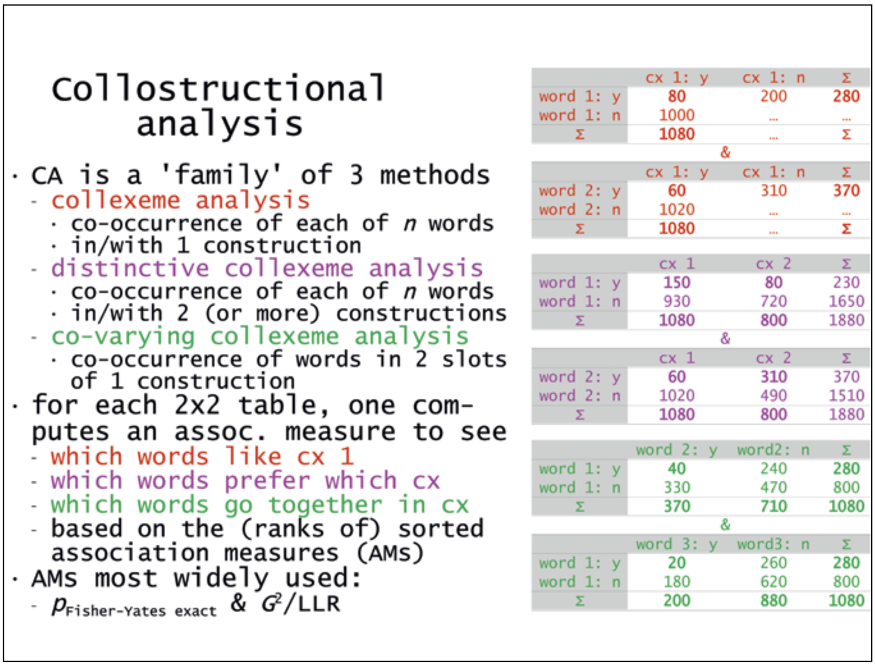
- Collostructional analysis (CA) is an method based on the maybe most fundamental corpus linguistic assumptions: the distributional hypothesis
  - "[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution" (Harris 1970:785f.)
- CA is a straightforward extension of …
  - of collocations: co-occurrence of words/lexical units
  - to (one sense of) colligation: co-occurrence
    - of words
    - and patterns/constructions

FIGURE 11

and B are more different than the distributions of A and C". So, in a sense, it's a straightforward extension of collocation work in corpus linguistics as it has been happening for decades. The only difference or the main difference being that instead of looking at words co-occurring together, we're looking at the co-occurrence of words in, or with, patterns or constructions.

Three different methods have been distinguished. The first one would be essentially the type that you've seen before. The first method is collexeme analysis. You're looking at one construction, which is in the columns here, construction$_1$ (yes versus no); construction$_1$ (yes versus no) and you're looking at a bunch of words, each of which occurs at least once in the construction. The construction$_1$ occurs 1,080 times, word$_1$ occurs in it 80 times, word$_2$ occurs in it 60 times, and so on. The idea is, for every word, you compute a measure of association: for this one, for that one, and for all other ones, kind of like what we discussed before.

The second possibility, maybe actually even more widely used, because it's simpler, is distinctive collexeme analysis. You have two, or theoretically more competing constructions—competing in the sense of they are functionally similar. For instance, they might constitute one of those famous argument



Collostructional analysis

- CA is a 'family' of 3 methods
  - collexeme analysis
    - co-occurrence of each of *n* words
    - in/with 1 construction
  - distinctive collexeme analysis
    - co-occurrence of each of *n* words
    - in/with 2 (or more) constructions
  - co-varying collexeme analysis
    - co-occurrence of words in 2 slots of 1 construction
- for each 2x2 table, one computes an assoc. measure to see
  - which words like cx 1
  - which words prefer which cx
  - which words go together in cx
  - based on the (ranks of) sorted association measures (AMs)
- AMs most widely used:
  - $p_{\text{Fisher-Yates exact}}$ & G^2/LLR

	cx 1: y	cx 1: n	Σ
word 1: y	80	200	280
word 1: n	1000	…	…
Σ	1080	…	Σ

&

	cx 1: y	cx 1: n	Σ
word 2: y	60	310	370
word 2: n	1020	…	…
Σ	1080	…	Σ

	cx 1	cx 2	Σ
word 1: y	150	80	230
word 1: n	930	720	1650
Σ	1080	800	1880

&

	cx 1	cx 2	Σ
word 2: y	60	310	370
word 2: n	1020	490	1510
Σ	1080	800	1880

	word 2: y	word2: n	Σ
word 1: y	40	240	280
word 1: n	330	470	800
Σ	370	710	1080

&

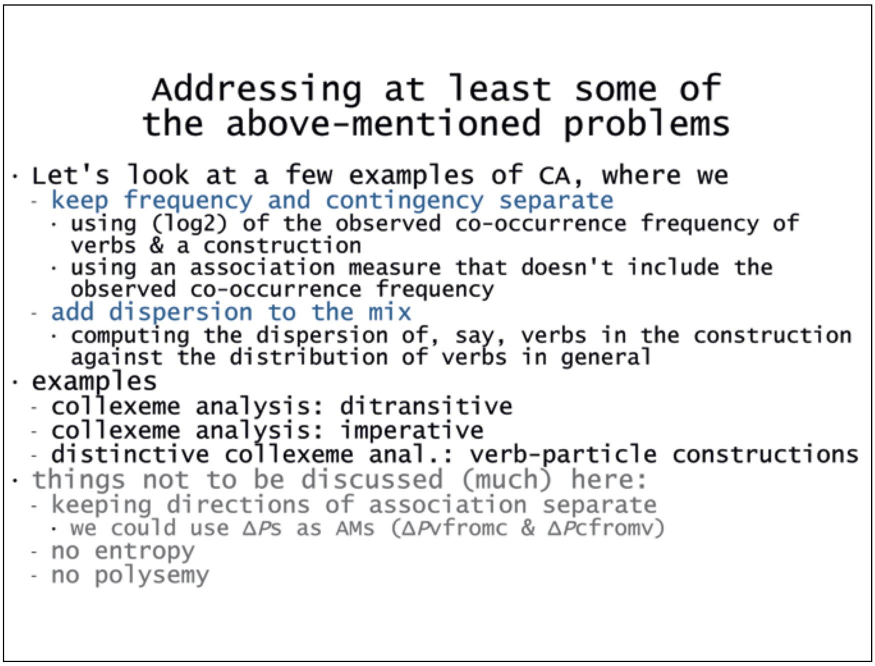
	word 3: y	word3: n	Σ
word 1: y	20	260	280
word 1: n	180	620	800
Σ	200	880	1080

FIGURE 12

structure alternations, or something like, like ditransitive vs. prepositional dative or something like that. So you find every instance of this construction, every instance of that construction, and then every word that shows up at least once in one of the two. So this would be a case where word_1 strongly prefers to occur in this construction (150) as opposed to that one (80). Whereas here we have a word_2 that very, very strongly prefers to occur in this one (310) as opposed to the other one (60). And so distinctive collexeme analysis would quantify that and would compare the two with each other.

Finally [co-varying collexeme analysis], not used that much, although it's also interesting sometimes. You have one construction with two slots in it, and you're looking at co-occurrences sort of depending on what happens in the first slot, what's going to happen in the other one. So it's word_1, yes or no, and word_2, yes or no, in the same construction and then you can quantify the preferences there.

If we want to do this, but also address some of the problems that I've mentioned before—so the fact that something like log-likelihood conflates frequency and effect size, the fact that all association measures do not take dispersion into consideration—then how can we do that? We're going to look at a few examples where we try to address at least some of these things. So we



Addressing at least some of the above-mentioned problems

- Let's look at a few examples of CA, where we
  - keep frequency and contingency separate
    - using (log2) of the observed co-occurrence frequency of verbs & a construction
    - using an association measure that doesn't include the observed co-occurrence frequency
  - add dispersion to the mix
    - computing the dispersion of, say, verbs in the construction against the distribution of verbs in general
- examples
  - collexeme analysis: ditransitive
  - collexeme analysis: imperative
  - distinctive collexeme anal.: verb-particle constructions
- things not to be discussed (much) here:
  - keeping directions of association separate
    - we could use Δ*P*s as AMs (Δ*P*vfromc & Δ*P*cfromv)
  - no entropy
  - no polysemy

FIGURE 13

will keep frequency and contingency or association separate. We're not going to use a measure that grows even if just a corpus size grows—we will use a measure that only grows if the effect becomes, in fact, stronger.

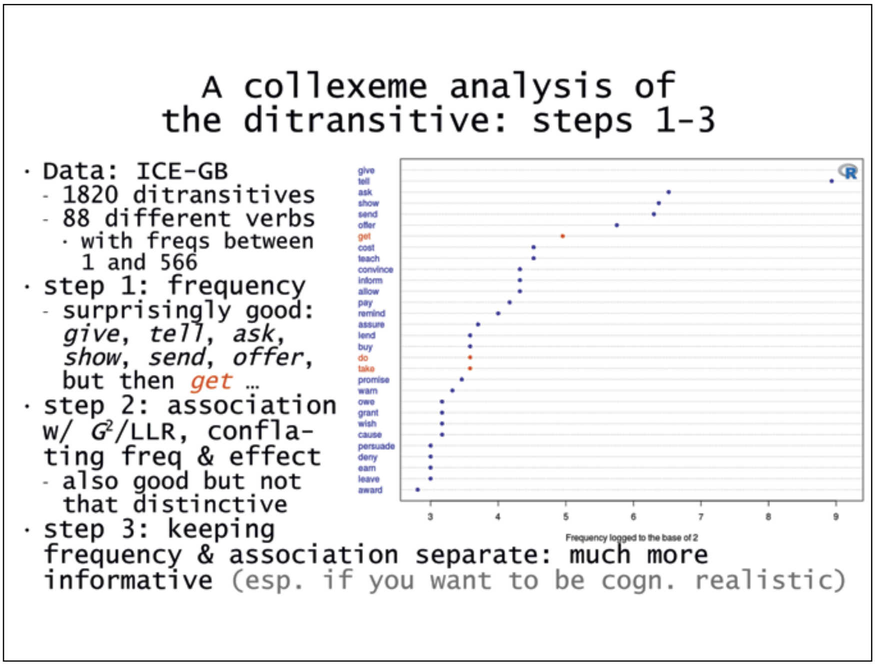
So the example that I'm going to use here is that for frequency, we're going to look at the log frequency, because in psycholinguistics research, most of the time, we find frequency effects on a log scale. Then, as an association measure, we're going to use the log odds ratio for now simply because we already have a variety of dimensions to juggle and I don't want to add two association measures: verb to construction, construction to verb, to the mix at the same time.

Then we'll add dispersion to the mix by looking at how evenly are the instances of the verb in the construction attested throughout the corpus. We want to avoid this example that I talked about yesterday: We want to avoid cases where verbs like *fold* or *process* score high on association strength, although they only show up in a construction in a single file—that's what we want to protect ourselves against.

We're going to look at three examples, namely the three constructions I showed you before on the slide (Figure 12). For log-likelihood ratio, [[first,]] we're going to do a quick look at a collexeme analysis of the ditransitive, obviously one of the most widely studied constructions out there. Second, we're going to look at the collexeme analysis of the imperative because that's a construction where yesterday we saw dispersion causes problems so we'll now going to check, can we handle that? Third, we're going to look at a distinctive collexeme analysis case, so do we find verbs that prefer the order where the particle comes before the verb and are there verbs that prefer the other order where the particle comes later? So are there verbs that prefer the construction, *He brought back the mouse*, and are there other verbs that prefer the construction, *He brought the mouse back*? In fact, you will see that there are quite strong tendencies.

Now we're not going to look at here at this point, simply because we don't have the time and the complexity quickly becomes quite daunting, we're not going to look at different directions of associations, and we're not going to look at entropy or polysemy at this point in time. I have data for at least this part and that part—polysemy, I haven't done yet myself. That awaits future research. Let's build this up step-wise.

We're going to first look at the ditransitive construction. I wrote a small script that basically gives us the number of ditransitives in the British component of the International Corpus of English. We find a pretty Zipfian distribution here as always: The 1,820 ditransitives, that's 88 different verbs showing up



A collexeme analysis of the ditransitive: steps 1-3

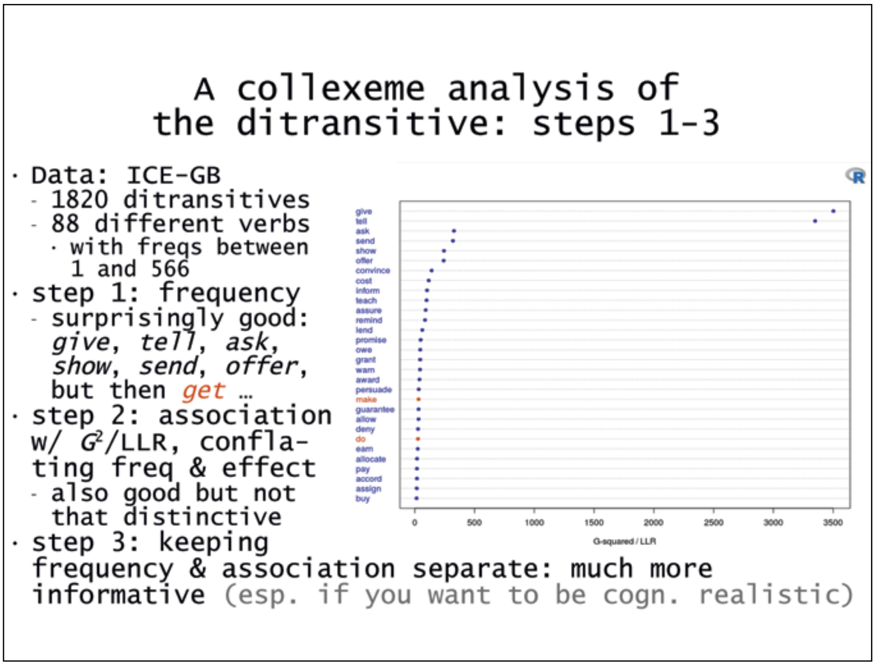
- Data: ICE-GB
  - 1820 ditransitives
  - 88 different verbs
    - with freqs between 1 and 566
- step 1: frequency
  - surprisingly good: *give*, *tell*, *ask*, *show*, *send*, *offer*, but then *get* …
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good but not that distinctive
- step 3: keeping frequency & association separate: much more informative (esp. if you want to be cogn. realistic)

FIGURE 14

in that construction and they have frequencies between 1 and 566. One verb actually is like nearly like 30% or something of all the instances but then there are also quite a few cases that show up only a single time.

Now, if you look at the frequencies with which verbs show up in that construction, then you can plot it like this [referring to the plot in Figure 14]. Here on the *x*-axis, we have the frequency, but again, it is logged, but you can see, *give* is the most frequent word and here is a dot that's behind this [pointing to the dot that represents the frequency of *give*]. That's the 566. Then *tell* is a little bit less frequent. Then there's a huge gap already, everything else is way below that. As before, actually, the first results are quite good, given the semantics that one, after decades of researching this construction to death, after decades of this, if you look at the verbs, *give*, *tell*, *ask*, *show*, *send*, *offer*, that's exactly what everyone has always been talking about in that construction. No big surprises there.

Again, I do want to point out though the verbs you see here in red, *get*, *do* and *take*, especially *get* shows up in the construction quite frequently, but actually less often than expected: *Get* is an extremely frequent word in general so while this is a high number, it's actually too low a number because given how frequent *get* is in general, you know, this number should be way higher than it is. Again, the ranking by frequency alone doesn't even tell you.



A collexeme analysis of the ditransitive: steps 1-3

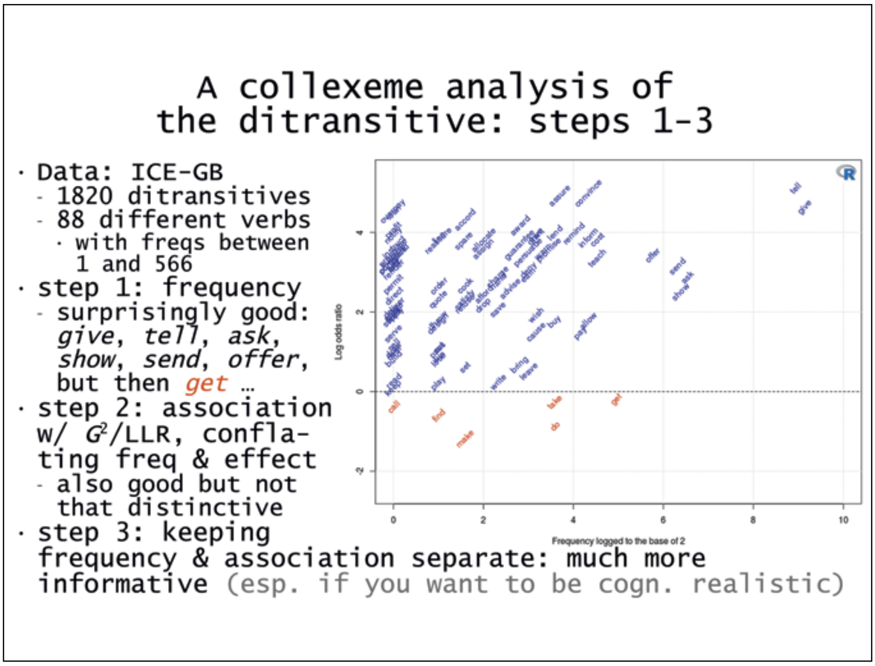
- Data: ICE-GB
  - 1820 ditransitives
  - 88 different verbs
    - with freqs between 1 and 566
- step 1: frequency
  - surprisingly good: *give*, *tell*, *ask*, *show*, *send*, *offer*, but then *get* …
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good but not that distinctive
- step 3: keeping frequency & association separate: much more informative (esp. if you want to be cogn. realistic)

FIGURE 15

Now let's compute an association measure. This time around, actually first, the one that is not that great, namely, the likelihood-ratio, because it conflates frequency and effect. You can see, again, *give* and *tell* win out, but this time by a huge margin. Because now, on top of the fact that they already lead in terms of frequency [referring to the graph in Figure 14], this distance is even made bigger because of association coming to the mix. So now the verbs that are repelled by the construction actually have very, very low associations. It's better in that sense—we see which words are repelled by the construction—but it's worse in the sense that there's a clear conflation of frequency and effect size.

Now what about keeping frequency and association separate? This would be one way to do this and it's actually kind of interesting for a reason, I'll discuss in a moment. So we have a frequency on the *x*-axis, again logged. This is 4, 16, 64 and so on [referring to the frequency represented on the *x*-axis]. You can see *give* is more frequent in the ditransitive than *tell*, [because] it's to the right of *tell* but *tell* is higher up. So the association of *tell* to the ditransitive is higher than that for *give*.

That's something that the normal kind of measure doesn't tell you. It doesn't tell you where a certain value comes from, whether it's the more frequent component or whether it's the more attraction component. Here you can see that



A collexeme analysis of
the ditransitive: steps 1-3

- Data: ICE-GB
  - 1820 ditransitives
  - 88 different verbs
    - with freqs between 1 and 566
- step 1: frequency
  - surprisingly good: *give*, *tell*, *ask*, *show*, *send*, *offer*, but then *get* …
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good but not that distinctive
- step 3: keeping frequency & association separate: much more informative (esp. if you want to be cogn. realistic)

FIGURE 16

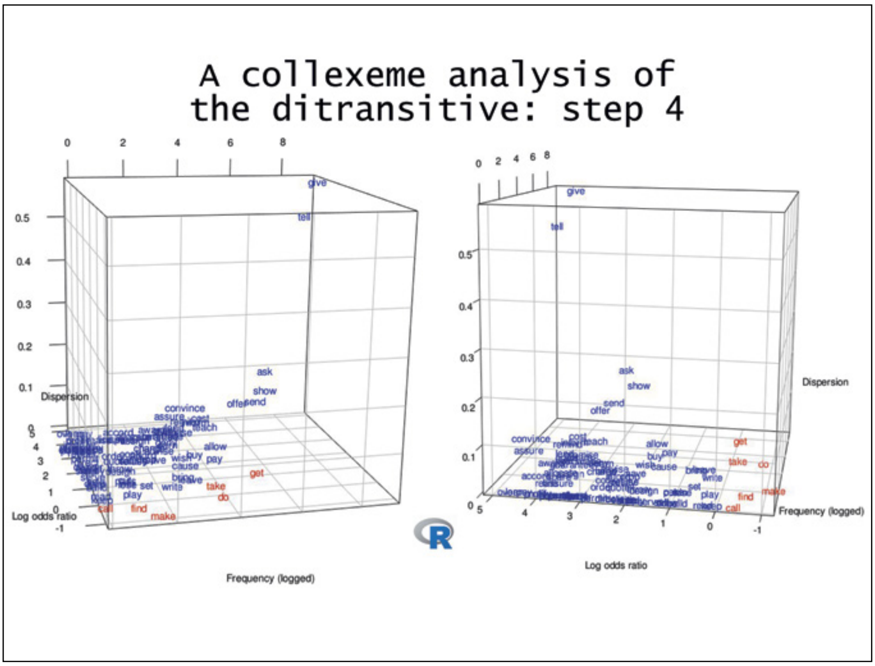
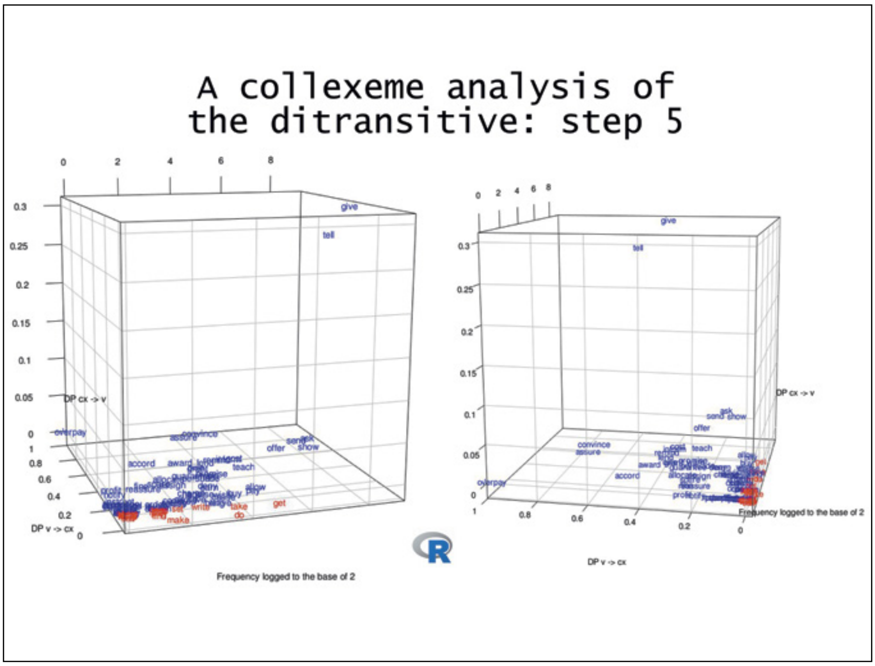


FIGURE 17

very well. You can see for all the words that are repelled, their log odds ratio is negative. And you can see up here that the correlation between frequency and associations are actually not that strong. It's not like we have a nice point cloud that goes up like this [referring to a linear relation] but we kind of have a pretty big mess here with maybe something going up slightly. What this clearly shows is that our keeping frequency and association separate works. You can't look at this axis [*x*-axis] here and completely clearly predict what's going to happen on this axis [*y*-axis]. Frequency and association are not the same. For cognitively *realistic* analyses, there's no good motivation to conflate them into one number, pretending they are the same and hoping for the best.

Now this [Figure 17] is what happens when we add dispersion. This one is a little bit hard to interpret. I am going to show you the interactive version of this plot. [[The explanation of the 3-D plot]] This axis here is frequency on a log scale. You see the frequency values here. It's always opposite of the legend: This is the label for this axis, and the axis is shown on the opposite side of the cube up and top. Then this is association, the log odds ratio. The red verbs are all the ones that are repelled, the blue words are all the ones that are attracted. The further to the right, the stronger the attraction. Then this axis is dispersion.

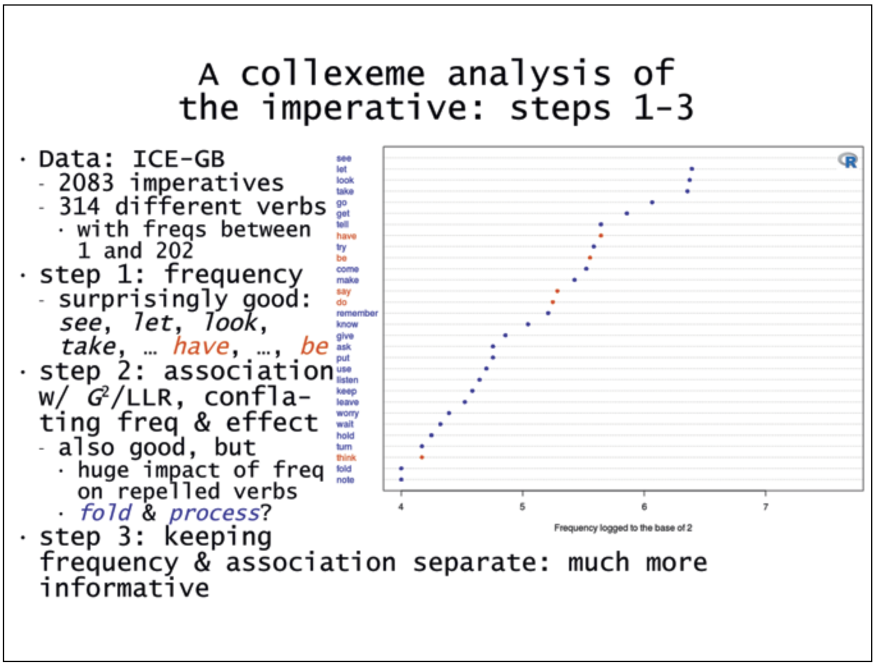


A collexeme analysis of the ditransitive: step 5

FIGURE 18

Now there's a variety of things you can see pretty clearly. One, for instance, is there's a bunch of words that are as attracted to the ditransitive as is *give*, verbs like *tell*, in fact, but then also things that are much less frequent like *convince*, *assure*, *reassure*. *accommodate*, *remind*, etc. They're all relatively similar to *give*. But of course, if we rotate this, we see obviously that *give* is way more frequent in the construction. Also, we see: the association of *tell* is stronger than *give*. If we rotate it like this, then *tell* is to the left of *give*, so it has a higher association value, but *give*, again beats *tell* to the point in terms of dispersion. It's more evenly distributed in the corpus in that construction. Basically, what we have here is a plot that talks exactly about the three dimensions that Ellis et al. (2016), and actually many other people, are talking about: frequency, recency, and contingency and association. This plot shows you all of these dimensions as opposed to dumbing it down into a single number, and then hoping that that works well.

A lot of insight that we can gain if we keep these things separate. Theoretically, I'm just mentioning this for the record. I'm currently writing this up, but there is one way in which you *can* actually get a single number from this as well per verb. I'm not actually recommending it right now, because there is one problem, but, just to mention it theoretically, and that is like this (3-D plot):



A collexeme analysis of the imperative: steps 1-3

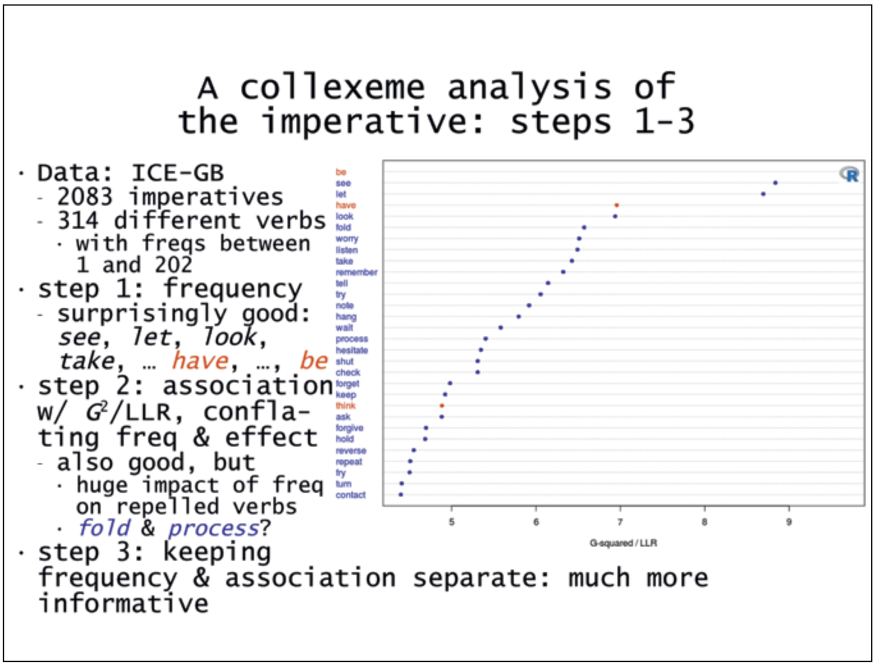
- Data: ICE-GB
  - 2083 imperatives
  - 314 different verbs
    - with freqs between 1 and 202
- step 1: frequency
  - surprisingly good: *see*, *let*, *look*, *take*, … *have*, …, *be*
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good, but
    - huge impact of freq on repelled verbs
    - *fold* & *process*?
- step 3: keeping frequency & association separate: much more informative

FIGURE 19

Look at this point here where the mouse is right now, that point is the origin of that cube, that's when every dimension is zero. It's when this dimension is zero, it's when this one is zero (you are at the bottom of the cube), and it's when this edge is zero. So if you want one number, IF, what you could do is, you can take that cube, so keep your eye on that point here, and so you measure the distance from the origin to where the verb is. You have a cube like that with the origin here, if a verb is located down here, up here, and then in the back, then that is the Euclidean distance from the origin to that point. And if you do it like this, that number will capture frequency, recency, and association. Again, there's a problem, which I'm not going to bother you with now, but theoretically, that's one way in which this could be done.

I think we at least agree that this is way more informative than putting it all in one number or just reporting a frequency.

What about the same for the imperative? This is the example where we hope that our approach can help deal with *fold* and *process*. We have about 2,083 imperatives, about 314 different verbs in them, with again, Zipfian-distributed frequencies between 1—hapaxes—and 202. If we sort it by frequency, again, it looks like this [referring to Figure 19]. Examples like, *see* and *let* and *look*, again it's up here, pretty good. But again, also, *have*, *be*, *say* and *do* are quite frequent



A collexeme analysis of
the imperative: steps 1-3

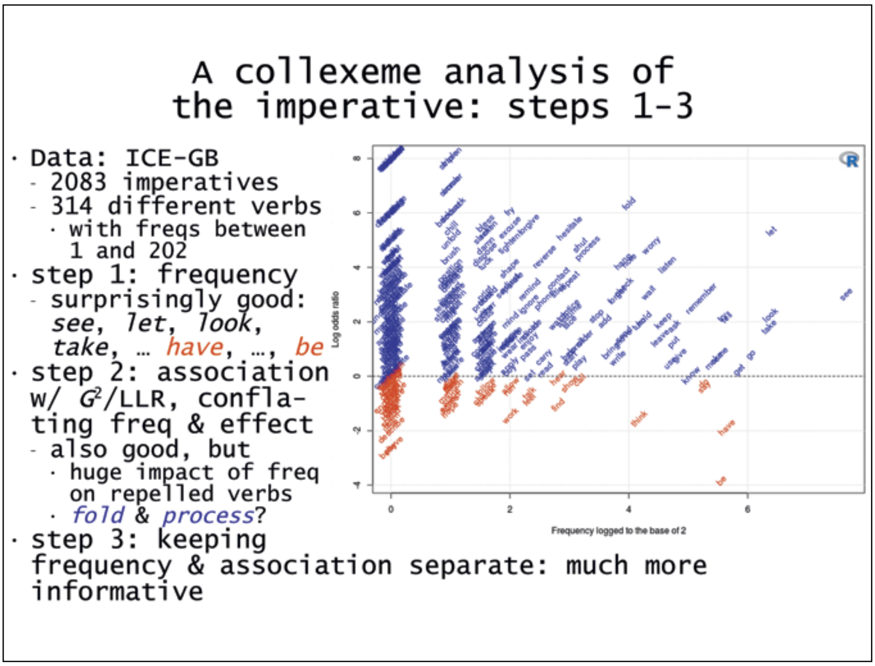
- Data: ICE-GB
  - 2083 imperatives
  - 314 different verbs
    - with freqs between 1 and 202
- step 1: frequency
  - surprisingly good: *see*, *let*, *look*, *take*, … *have*, …, *be*
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good, but
    - huge impact of freq on repelled verbs
    - *fold* & *process*?
- step 3: keeping frequency & association separate: much more informative

FIGURE 20

in the imperative, but less often than you would think given how frequent they are in general. But the frequency doesn't show you that. You only see that because I marked it with the color. If I didn't put the color in there, you wouldn't know that this is actually repelled by the imperative. And it's obvious, you don't often see *have*, *have* what, it's not a verb that lends itself for the imperative.

Here's the log-likelihood value. This one actually is really funny, because the highest log-likelihood value is scored for a verb that is repelled by the construction. You would have to make that negative actually so that it gets sorted at the bottom. But then we get *see*, *let*, *look*, and *fold*, there the problem verb is and here is *process*, the other problem verb that we hope a bigger approach, the better one, that will now not rank that highly. Again, *have* is very high up but actually repelled, the same with *think*, you don't use *think* in an imperative very much.

Again, what we want to do is first to keep frequency and associations separate. We have frequency here now. And so *see* wins very strongly, quite a bit of a gap till the next one. But actually, as you can see, the association is actually not that high—there's a ton of verbs that are more strongly attracted to the imperative. That means the fact that *see* here wins—it's the highest blue



A collexeme analysis of the imperative: steps 1-3

- Data: ICE-GB
  - 2083 imperatives
  - 314 different verbs
    - with freqs between 1 and 202
- step 1: frequency
  - surprisingly good: *see*, *let*, *look*, *take*, … *have*, …, *be*
- step 2: association w/ G^2/LLR, conflating freq & effect
  - also good, but
    - huge impact of freq on repelled verbs
    - *fold* & *process*?
- step 3: keeping frequency & association separate: much more informative

FIGURE 21

verb—the fact that *see* wins here, that's not due to a super high association strength, it's due to a higher frequency, which again, from looking at this value, you don't know that. You only know that if you keep them separate and see, what does it do here, what does it do here? Things like *let*, *worry*, *listen*, and *hesitate*, *don't hesitate* probably, *forgive* as in *forgive me* probably, all those have a higher degree of attraction to the imperative than this one. But actually, so does *fold*. Not that frequent, but very strongly attracted to the imperative. Even more than *hesitate* and *shut* (which is probably *shut up*).

Again, we want to keep it separate. We're getting this [referring to a 3-D plot similar to the one in Figure 22]. So here we have the same thing. Frequency is on this axis, log odds ratio on this axis, and dispersion on this one. The verbs that are [[blue are attracted]], the words that are shown in [[red are repelled]], if you apply a *post hoc* correction. And we can see, *see* seems to be the overall winner. How does that happen? First, now this is frequency, so it's more frequent than everything else. Secondly, it's not more attracted to it than everything else, but it's way more evenly dispersed in the imperative than everything else, that's why it has this marked position here up at the top.

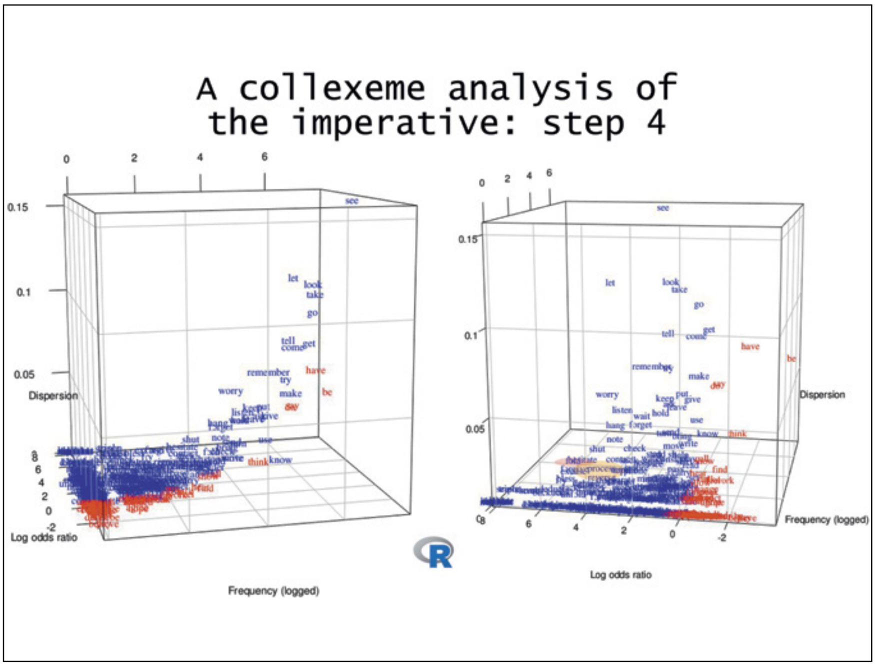
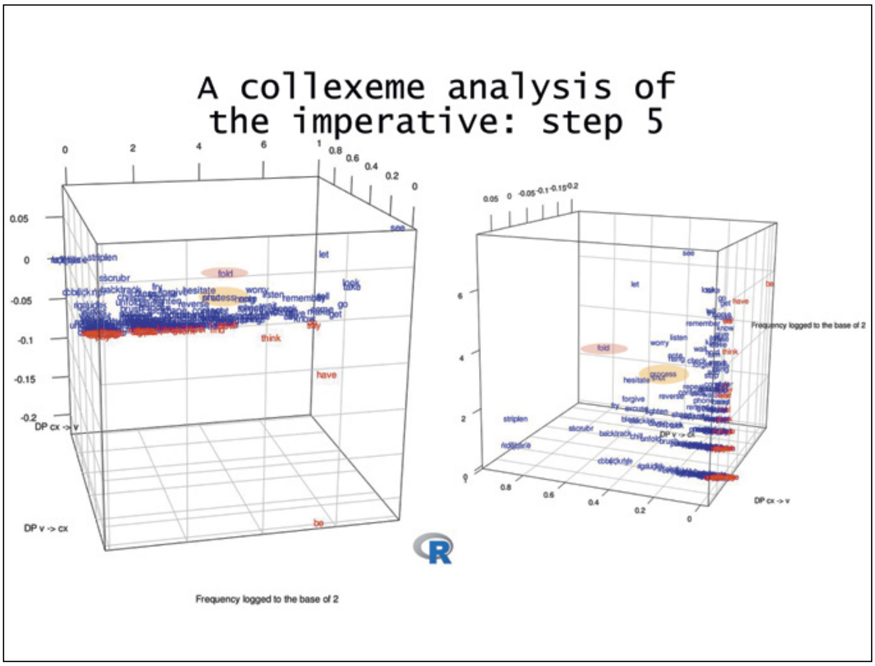

A collexeme analysis of the imperative: step 4

FIGURE 22

However, what this approach also allows you to do is to say, 'well, but I want to prioritize attraction'. What you could do is—mathematically, not physically—is you could say, 'I want to compute this origin distance again'. So the value for *see* will be higher than for *let*. What you can do mathematically is you can take this axis (log odds ratio) and pull it longer. So you take this whole cube, basically make it go till here. That means this one (the dispersion axis) doesn't change, the one in the back, the frequency doesn't change, but you're emphasizing association, and then this one will win. The problem that I mentioned earlier is that it's not clear how far you want to pull it and secondly, how do you motivate that theoretically? I mean, the pulling is motivated theoretically, but how far? If you pull it only a little bit, *see* will still win—if you pull it all here, *let* will win and so how do you now explain to your reader your well-reasoned decision to pull it till here? That's going to be tricky. So here the problem then is, we have a mathematical solution to a problem, but not a theoretical one, very annoying.

Those are essentially the same data. The only thing I changed here is to give you at least a hint: Now I changed, I used Δ*P*-values. Now these cubes actually are directional attraction. We have frequency here, then attraction from

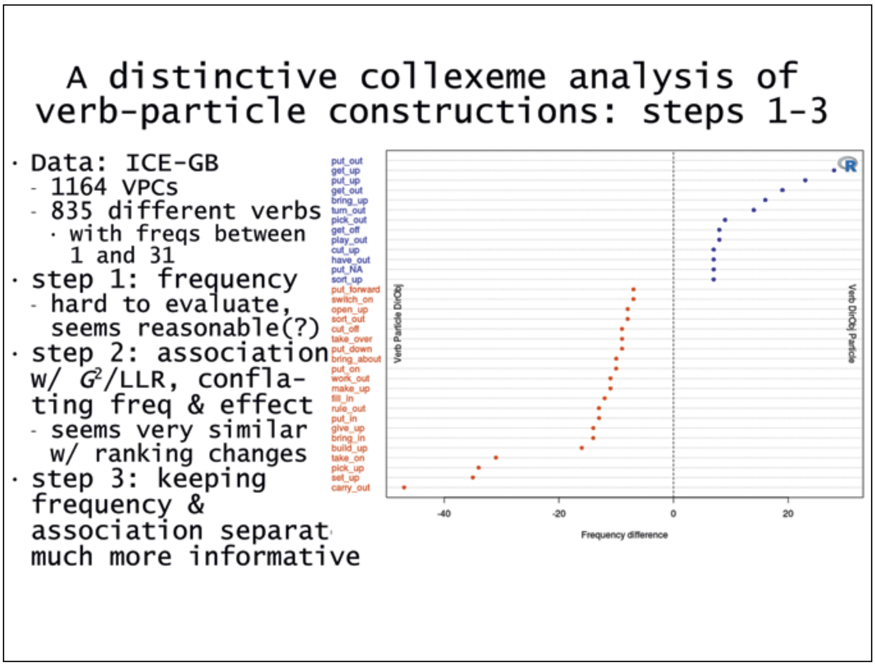


A collexeme analysis of the imperative: step 5

FIGURE 23

the verb to the construction in this axis, the one going into the back of the wall; and then construction to verb, that's the one going up. You can see, for instance, here's *see* and here's *let*. *See* is still scoring high on frequency, but here with this attraction, it's close to zero. So what *see* does with that construction is that—the construction likes *see*, but not the other way around. You know where there's a ton of verbs where it is the other way around. So here then we have *fold* and *process*, the ones whose dispersion puts down and downgraded, so that we didn't make a wrong conclusion there.

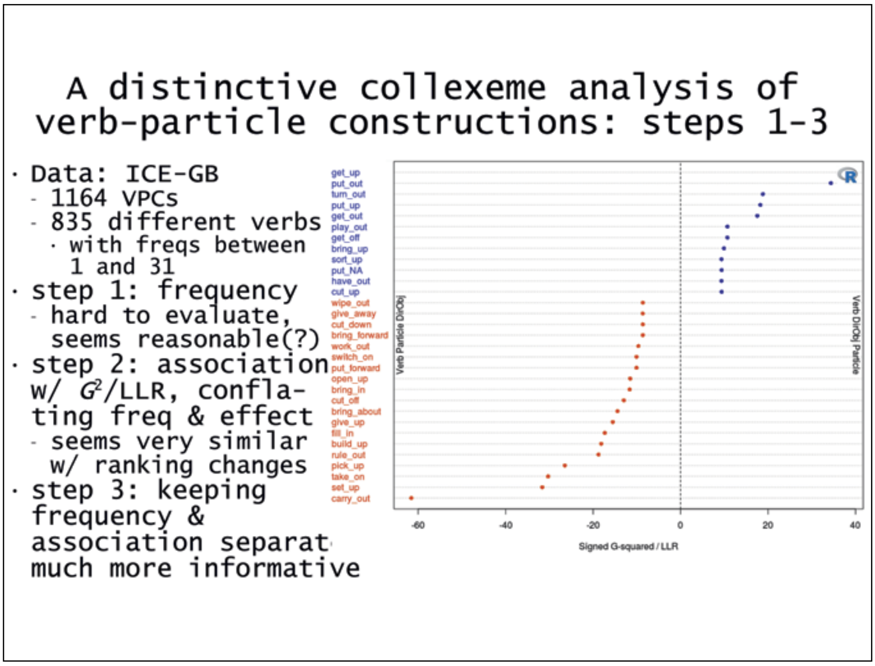
So, final example, maybe speeding this up a little bit in terms of time: We're looking at the verb-particle constructions so this is a distinctive collexeme analysis, the results look a little bit different. This dot chart now is organized from left to right. The verbs that score highly are the ones that like verb-direct object-particle, verbs like *put out*, *get up*, *put up*, *bring up*, they like the construction where the particle is at the end. And verbs here at the bottom with the negative values, they like the other order, namely, verb-particle-direct object. So, *carry out*, *set up*. These data say that you're more likely to say *pick up a book* than *pick a book up*. So, here in the middle, I omitted a whole bunch of verbs that are not distinctive, just to save space on the plot.



A distinctive collexeme analysis of verb-particle constructions: steps 1-3

- Data: ICE-GB
  - 1164 VPCs
  - 835 different verbs
    - with freqs between 1 and 31
- step 1: frequency
  - hard to evaluate, seems reasonable(?)
- step 2: association w/ G^2/LLR, conflating freq & effect
  - seems very similar w/ ranking changes
- step 3: keeping frequency & association separat much more informative

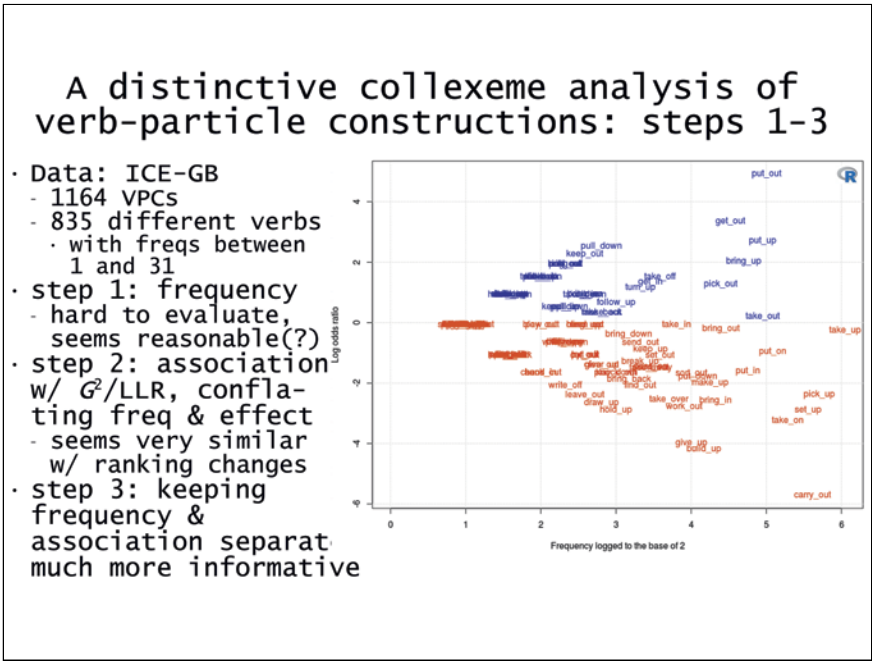
FIGURE 24



A distinctive collexeme analysis of verb-particle constructions: steps 1-3

- Data: ICE-GB
  - 1164 VPCs
  - 835 different verbs
    - with freqs between 1 and 31
- step 1: frequency
  - hard to evaluate, seems reasonable(?)
- step 2: association w/ G^2/LLR, conflating freq & effect
  - seems very similar w/ ranking changes
- step 3: keeping frequency & association separat much more informative

FIGURE 25



A distinctive collexeme analysis of verb-particle constructions: steps 1-3

- Data: ICE-GB
  - 1164 VPCs
  - 835 different verbs
    - with freqs between 1 and 31
- step 1: frequency
  - hard to evaluate, seems reasonable(?)
- step 2: association w/ G^2/LLR, conflating freq & effect
  - seems very similar w/ ranking changes
- step 3: keeping frequency & association separat much more informative

FIGURE 26

It doesn't change much actually, if you apply the likelihood-ratio test here, as the results are really very similar. We still have *get up*, *put out*, *put up*, *get out* here at the top, and we still have *carry out* here at the bottom, so not much of a change. That's in part of course because this is so much affected by frequency as well.

But if we look at frequency and associations separately, we get a very different picture: For instance, some of the most frequent verbs, the verbs on the right, actually have no strong attraction at all. They go equally well with both. *Take up* has no preference for either construction, but something like *put out* or *carry out*, they have very strong attractions, that's why they're very high up in the plot and very much further down in the plot. And again, not really a correlation here between frequency and association, but it's not like you would draw a line through this, and it would be clearly going down or something like that.

Then this would be the three dimensional representation again. Frequency is the back axis, this is association, either for this construction or for that construction, and this is dispersion. Here, those are the verbs that are most strongly attracted to this construction (verbs attracted to VPO are colored in red) and this is the verb that is most strongly attracted to that construction (verbs attracted to VOP are colored in blue).

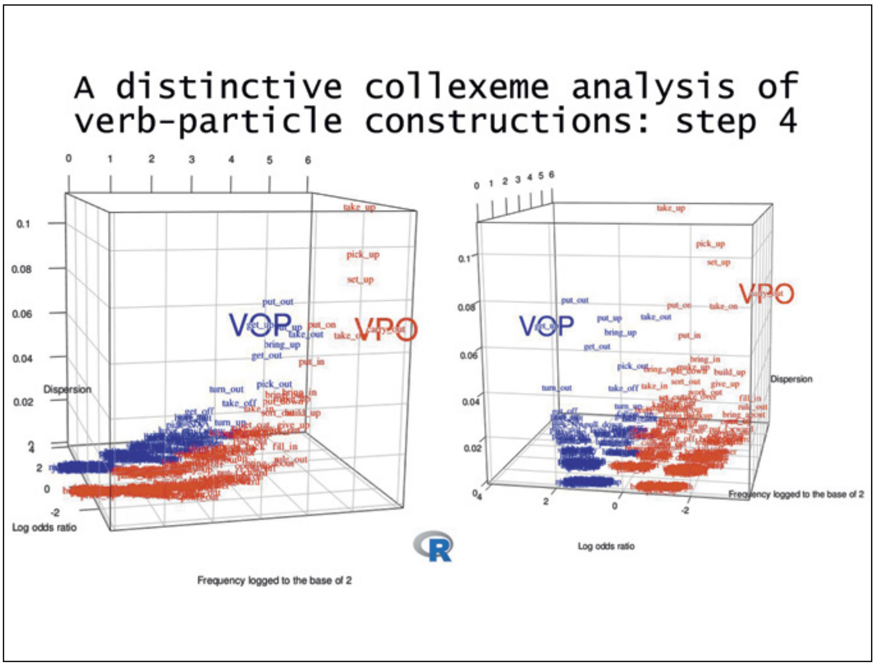
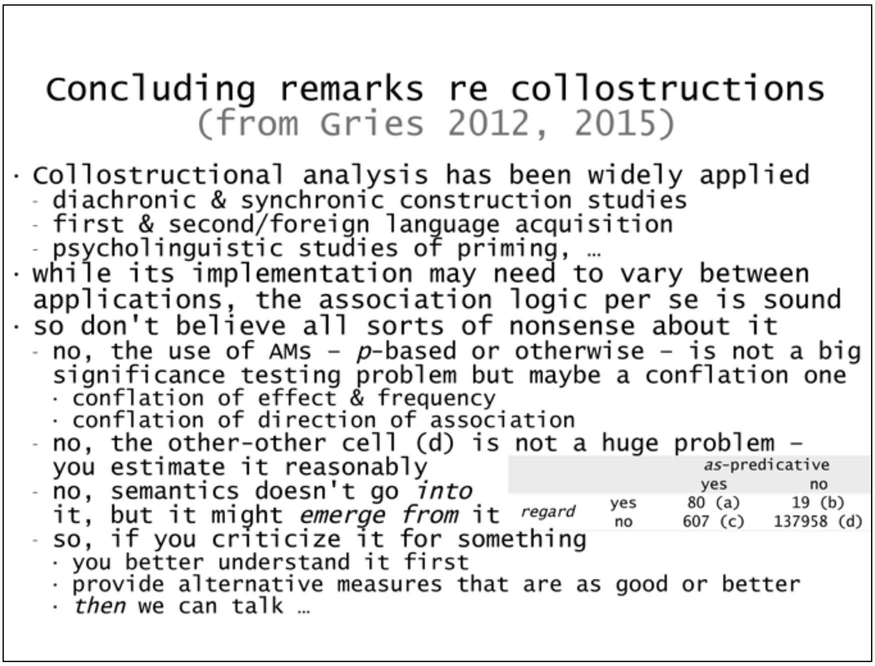

A distinctive collexeme analysis of verb-particle constructions: step 4

FIGURE 27

I'm not going to push it quite that far. But it is really, really tempting to use the word *prototype* somehow in that connection. Because in terms of frequency and association and dispersion, this is the verb that most goes with that. Again, I'm not going to push it quite that far, but it's tempting to at least think about that.

To wrap up, collostructional analysis as a method in general has been very widely used, no doubt about that. There have been diachronic studies, there have been synchronic studies, this method has been used in first and second, or foreign language acquisition, it has been used successfully in studies having to do with priming effects for native and non-native speakers, actually.

The exact implementation varies between applications. Not all applications use the same association measure and it's not obvious always which association measure to use. I give you two recommendations but depending on what exactly you have in mind for your study you might have different ideas about that. But the more important point is this, namely, that the logic of including association per se is sound. We can debate how we measure it but given the psycholinguistic, the psychological, and all sorts of linguistic literature itself, we do want to make sure that this is a dimension of information we do include. That means, you shouldn't believe all sorts of nonsense that you can



Concluding remarks re collostructions (from Gries 2012, 2015)

- Collostructional analysis has been widely applied
  - diachronic & synchronic construction studies
  - first & second/foreign language acquisition
  - psycholinguistic studies of priming, …
- while its implementation may need to vary between applications, the association logic per se is sound
- so don't believe all sorts of nonsense about it
  - no, the use of AMs – *p*-based or otherwise – is not a big significance testing problem but maybe a conflation one
    - conflation of effect & frequency
    - conflation of direction of association
  - no, the other-other cell (d) is not a huge problem – you estimate it reasonably
  - no, semantics doesn't go *into* it, but it might *emerge from* it
  - so, if you criticize it for something
    - you better understand it first
    - provide alternative measures that are as good or better
    - *then* we can talk …

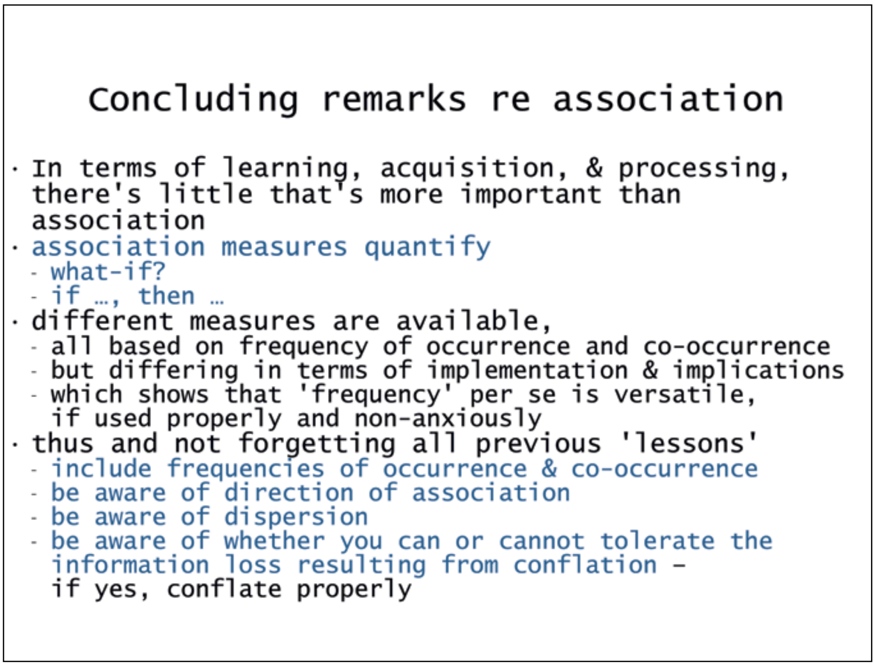
		as-predicative	
		yes	no
regard	yes	80 (a)	19 (b)
	no	607 (c)	137958 (d)

FIGURE 28

read about collostructional analysis in some not-to-be-named-here publications. For example, some people have harped on the fact that values like log likelihood or *p*Fisher-Yates exact test, that's like a huge problem because of all the null hypothesis significance testing issues that you run into—that's not really necessarily the case. If anything, the problem is a conflation one, namely, that you conflate frequency and effect size but the fact that the measure is based on the *p*-value per se can, in fact, be even corrected for.

Some other people have talked about how difficult it is to compute this number (cell *d*), all the instances that are not the verb in question and that are not the construction question. And again, at this time, I'm not even saying who said this, but this is just nonsense. You estimate that number on the basis of everything else that you have here. If you're looking at a construction and a *verb* slot, then obviously this number will not be determined by the number of *nouns* in the corpus, but by the number of *verbs*. Plus, if you do simulations, it doesn't even matter that much how high that number is so don't believe that part, like computing this cell is so difficult that you can run this.

Someone has criticized the analysis for that it disregards semantics. That is true—but only in the most trivial sense. The point is, semantics, that doesn't *go into* the analysis—it *comes out of* it. So once you've done the rankings of all



Concluding remarks re association

- In terms of learning, acquisition, & processing, there's little that's more important than association
- association measures quantify
  - what-if?
  - if …, then …
- different measures are available,
  - all based on frequency of occurrence and co-occurrence
  - but differing in terms of implementation & implications
  - which shows that 'frequency' per se is versatile, if used properly and non-anxiously
- thus and not forgetting all previous 'lessons'
  - include frequencies of occurrence & co-occurrence
  - be aware of direction of association
  - be aware of dispersion
  - be aware of whether you can or cannot tolerate the information loss resulting from conflation – if yes, conflate properly

FIGURE 29

the verbs in that construction, then typically at least you can interpret that in a semantic way. The idea of this is to *prepare* you for a semantic analysis and not to *presuppose* one. In other words, if you criticize the method, then you'd better understand it first, and provide measures that are as good as the ones that are being used to better. Then obviously, we can talk. I mean, at least some of these claims are just demonstrably false actually, even in the papers that criticize the method for it.

Then last slide. I would go so far as to kind of support Nick Ellis and colleagues here very much: In terms of learning, acquisition and processing, there is really little that's more important than association, because nothing happens without a context. Everything will be tied to some condition in some way. What association measures do is they quantify basically *what-if* or *if …, then …* scenarios: 'what is going to happen if this is the case?', or 'if this happens, then what will be the next corollary of that?'.

There's a whole bunch of different measures that are available, minimally 80. They all are based on frequency of occurrence and co-occurrence but they're different in terms of how exactly these frequencies are used and that also means they're different in terms of what you can take away from them. Again that means that frequency is a relatively versatile notion that goes way beyond just how often does something happened, *if* you use it properly and not in fear of any imperialists. So to not forget all the previous lessons, do look at frequencies of occurrence, do look at frequencies of co-occurrence, but then also be aware of direction of association: Don't always assume by default that something is bidirectional. Be aware of dispersion, you've seen in a 3-D plots how much of a difference it can make. Be very careful with conflating things into numbers, because a lot of times that information loss might kill exactly what you're interested in. And now this thing crashed. Thanks.