An experimentalist's perspective on corpus data

Maik Thalmann
maik.thalmann@uni-goettingen.de

Corpus meeting

July 11, 2022

RTG 2636

Form-meaning mismatches

UNIVERSITÄT GÖTTINGEN

The problem

Corpus data are often used like experimental data, to answer the same kinds of questions.

But the properties of the data types are very different in a number of respects. All of which I will argue to be important enough to affect the valid inferences (with or without statistics).

I will also argue that these differences are meaningful even in abstraction from other » **unsolvable problems** inherent to (historical) corpus linguistics.

- » low power and infinite variation
- » power laws, and the problem of minimal pairs
- » markedness and generative processes
- » preference/access (see Thalmann & Panizza 2019)
- » ungrammaticality/negative evidence

Please take my code

Both the presentation and the R code for my plots are available online.

If you're considering a switch to incorporating R into your workflow, feel free to have a look here:

For those of you who know German, there is also a website with my materials for an introductory course to R here: `https://mkthalmann.github.io/inferenz/`.

A quick note

I will not criticize the quantitative approach most corpus studies take. But note:

> Tests come with assumptions, often in the form that dependent variables (or their errors) are
>
> - ⚠ normally distributed (like IQ scores),
> - ⚠ consist of independent observations (a count in one category does not influence the count in another; previous observations do not have an effect on later ones), and
> - ⚠ homogeneous in terms of variance across conditions (the spread around the mean is the same).

Arguably, **none of these assumptions hold**. In turn, none of the "classical" experimental tests (t-test, ANOVA, etc.) are valid models for hypothesis testing, though at times for different reasons.

In addition, χ^2, a favorite among corpus linguists, and the binomial test (also quite popular; but assumes "bag of words"), are not adequate to model corpus findings.

Illustration: Normality

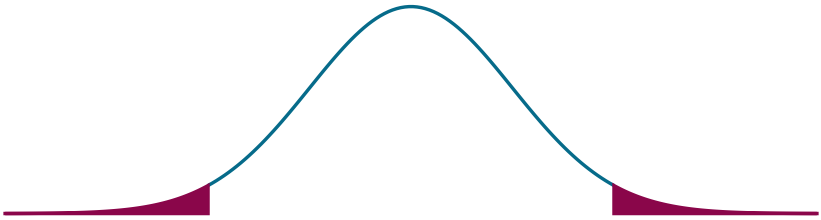
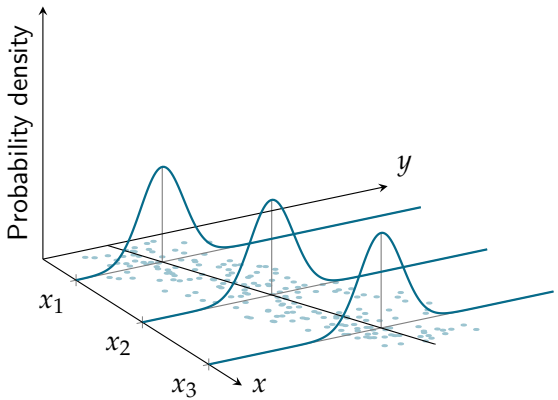
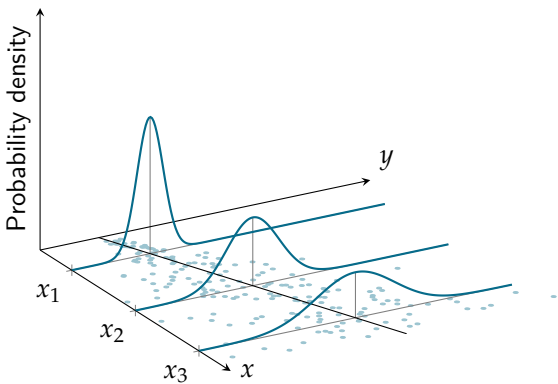


Illustration: Variance homogeneity for continuous x





Headaches around negative evidence

How not to treat data:

> The generalisations [of the present investigation; MT] are then necessarily tentative; nevertheless, I will assume that what has not yet been found [in corpora; MT] is ungrammatical: if one doesn't try and bet, one will never know. (Benincà 2004: 247)

(Thanks to Andrea Matticchio for pointing out this quote to me.)

This type of treatment is suspiciously similar to interpreting a p-value $> .5$. To add some further mileage to an already-overused dictum: "Absence of evidence is not evidence of absence."

Power Laws and Markedness

Corpora may never show a phenomenon. Even with balanced corpora, or incredibly large ones (say the ones available for German or English at `http://sketchengine.eu`).

The crucial feature of a generative system that produces infinite sentences from finite means is just that this option is expected. Certainly more so with lexical pairings, but to a more limited extent also with structures.

Consider the availability of a very highly marked structure, made possible by the interplay of two highly disprefered feature settings.

In such a case, **speakers may never find themselves in a situation where the marked option wins** over its unmarked competitors (unless super-licensing occurs, a property that for most phenomena is not well understood, see Weskott et al. 2011).

Yet, once prompted, speakers will confidently tell you that yes, this is a grammatical structure, though there has never been any direct evidence in either their input or their output. Such is the creativity of language.

Richness and paucity

In experiments, we control every aspect of human language that is reasonably controlled for within the constraints of our hypothesis.

Generally speaking, the gold standard is to only vary two things: lexicalizations (simply so that the items don't all look the same; *error/typo* and *glossary/diary* below) and the critical factor(s), ±DEF:

(1) a. There are **some** errors in the glossary. (+DEF)
b. There are **all** errors in the glossary. (−DEF)

(2) a. There are **some** typos in the diary. (+DEF)
b. There are **all** typos in the diary. (−DEF)

This is done since experimentalists need to make sure that the only driver of variance is the critical manipulation beforehand, because statistics cannot separate noise from systematic variation if the random component is also only present in one factor setting.

Noisy corpora

Corpora, by contrast, are full of noise. Being observational data, the distinction between systematic and random (from the perspective of our hypothesis) variation is moot. The data are what they are.

On the flipside, the noisiness of the data is what makes corpora collections of naturalistic data, rather than abstractions. In an ideal world, every utterance in a corpus is exactly as its producer wanted it to be given the contextual settings.

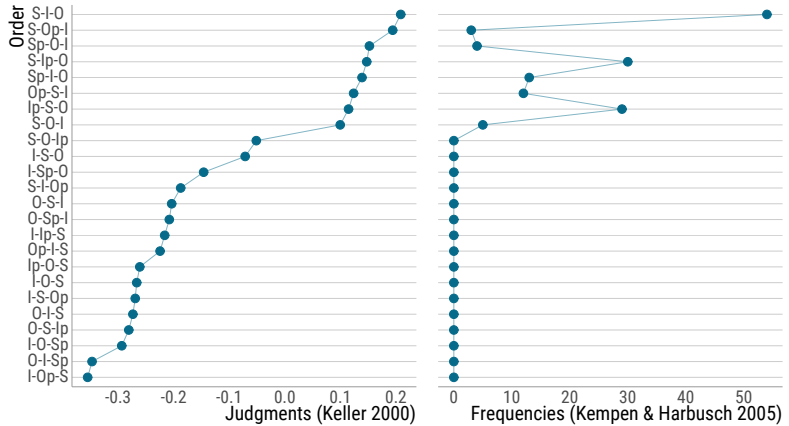
Of course, nothing like this could ever be said about experimental stimuli.

To sum up, we seem to have a complimentary distibution of the **noisiness/richness** property and the property of being **immediately accessible for variance partitioning** — a technique that all statistical tests rely on more or less obviously.

While I am certain that most linguists are aware of this, it seems to go unheeded.

Corpus-experiment correspondence

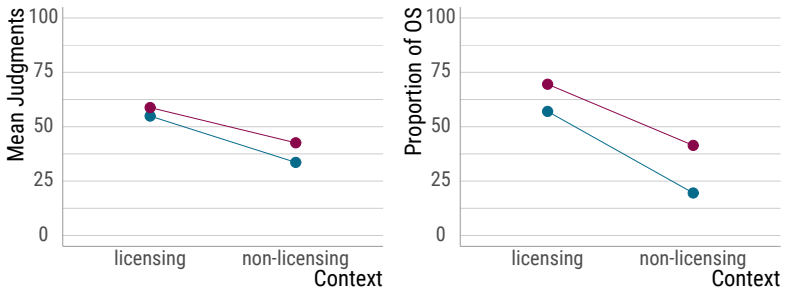
Even for a language like German (with humungous corpora of good quality and searchability) the correspondence between the two methods is partial at best, exemplified here with word order permutations in embedded clauses:



Effect sizes and Choices

Larger effect sizes for acceptability judgments compared to Forced Choice paradigms (Verhoeven & Temme 2017) — exemplified here using object- vs. subject-initial orders in German. See also Sprouse & Almeida (2017) for English (and further experimental paradigms).

Choices enlarge contrasts, presumably because not all alternatives are equally salient. This then creates a feedback loop that further **dampens the chances for marked orders**. Also, proportions in this case are necessarily dependent: choosing one variant has an immediate effect on the frequency of the other.



Sidebar: Gradience is not that interesting

Keller (2000) argues for grammaticality to be reconceptualized as a gradient property. In effect, grammar (rather than some third factor, Chomsky 2005) will have to contain some **function that assigns a weight to grammatical processes**, as well as another that **tallies up the costs incurred** over the course of a derivation (as assumed, a.o., by Gerbrich, Schreier & Featherston 2020). And, to be sure, we find contrasts between ungrammatical sentences:

(3) a. * Maury sleeped.
b. * Maury slept the moon.
c. * Maury sleeped the moon.
d. * Sleeped the moon Maury.

But the crucial thing to remember is this: Numbers will give you gradience sometimes even when you expect discreteness. And this might just be (Bayesian) uncertainty, rather than a necessary consequence of the system you are trying to capture theoretically. But see Bunk (2020), who does take differences between ungrammatical sentences (different variants of German V3) as evidence for V3 generally.

Sidebar: Gradience is not that interesting

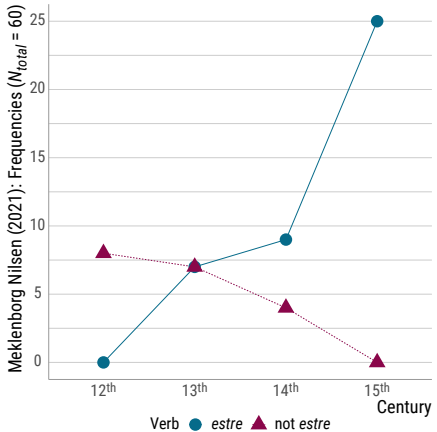
Plus, as acknowledged by Gerbrich, Schreier & Featherston (2020), people will prefer (4a) over (4b), for no reason situated within grammar (read: people are lazy). Given this, I believe more needs to be said to support the grammar-is-gradient property.

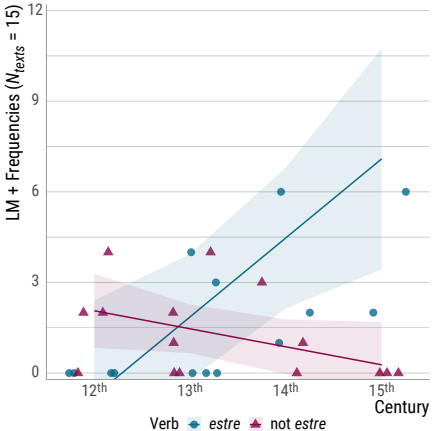
(4) a. Cows eat grass.
b. The seven cows in the field happily eat grass all afternoon.

NB: Armstrong, Gleitman & Gleitman (1983) show that people have gradient judgments on the well-defined concepts like ODD NUMBER and EVEN NUMBER. I am not inclined to think that is informative for the kinds of semantics we should assume for *odd*/*even* or—worse— *number*. For an in-depth discussion, see this blog post by Omer Preminger (2020).

Frequency tables

si topicalization in Old French apparently first started with **all verbs** and thematic roles, later only subjects and **only *estre*** were possible. The data are consistent with this, but I do not think they show it.





Frequency tables and temporal shifts

Now, certainly it could well be that the only active features for the French topicalization data are the theta criterion and the verb. But, again, while the data is consistent with this claim, they do not show it to be the case.

Rather than showing frequency shifts (of a phenomenon that is attested only 60 times over 400 years no less), and taking them to be sufficient without argument, visualizing which linguistic parameters are active in a more wholesale fashion appears to be a much more solid foundation to base a generalization on.

As it is, we are left to wonder whether other reasons might have played a role. And, though this is yet again an unsolvable issue, whether the described pattern is spurious or not.

But is there a better way?

Yes, instead of **ignoring richness and attempting to do experimental linguistics using corpora**, I believe that **harnessing the richness is a more advantageous approach**.

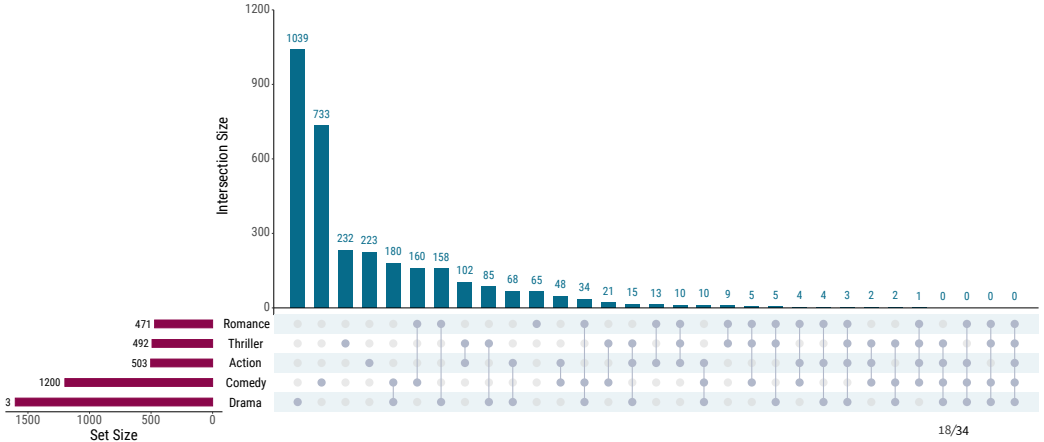
This boils down to presenting possible features that underly a phenomenon.

In the absence of speaker intuitions and negative judgments (the situation historical linguists generally face), the adaptation to contextual settings is a way of utilizing the richness of corpus data to arrive at a generalization.

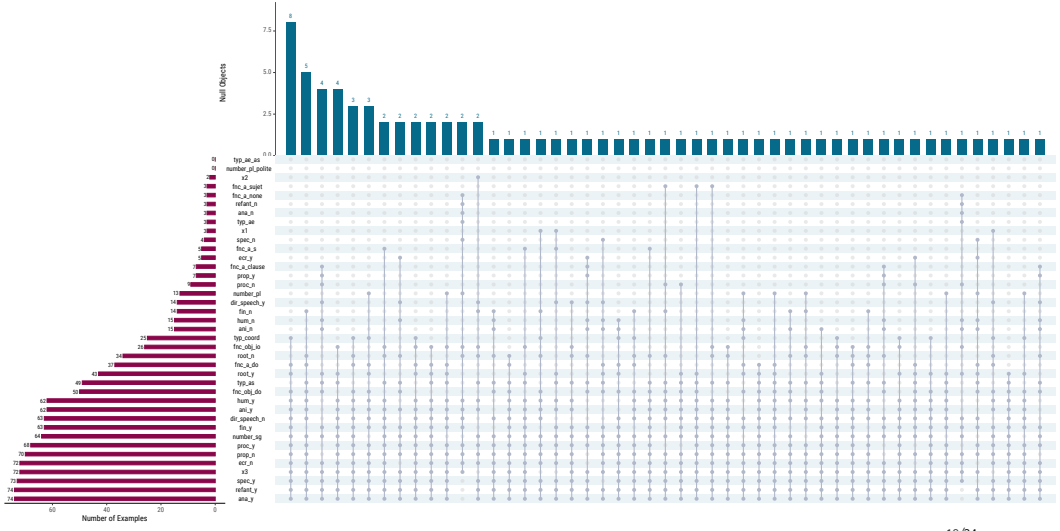
We will stay with Old French to show one way of doing this, but we will switch the phenomenon and the researcher.

Enter: Upsets as representations of feature matrices

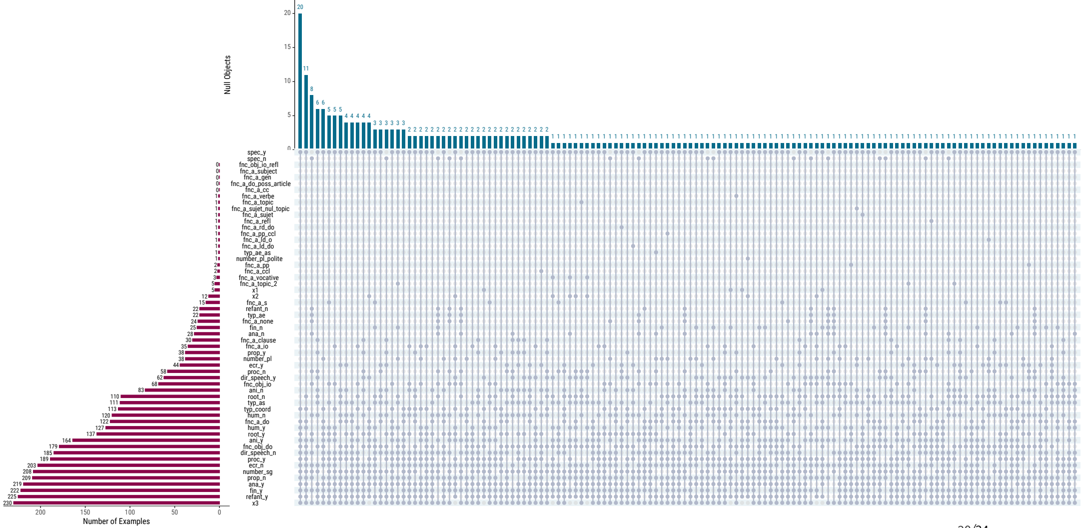
Before we do so let me briefly introduce upset plots — I use the `UpSetR` R package (Gehlenborg 2019). In upset plots, we visualize both the **size of sets of features** and the **size of the intersection** between them. (Imagine creating a Venn-Euler diagram for the data below!)



Prudence exploiting data richness



But wait, she's not done



Caveats and brain-like alternatives

Due to the sheer amount of work involved to produce these detailed descriptions (and the complexity in making some of the calls), this approach is, of course, not feasible for much larger data sets.

In these cases, people will often use the ever-mystifying neural networks and their cousins, where data is fed into a black box of sorts, and language or some classification of the input comes out.

But this leaves us with the opposite problem, in a way at least, summarized beautifully by Gilliam Ramchand in a recent blogpost:

> But I come away with the suspicion that while BERT and his descendents are getting better and better at performing, their success is like the equivalent of getting the answer 42 to meaning of Life the Universe and Everything. It still does not help if we don't know what exactly their version of the question was. (Ramchand 2022)

On time and its pitfalls

Time is just one meta datum, and one which is likely confounded with several others.

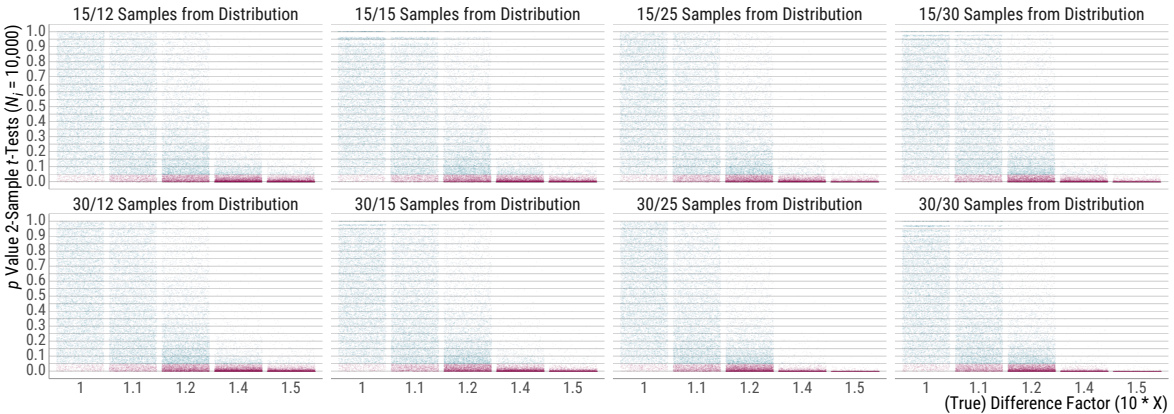
Continuous developments are an illusion in small corpora, again confounded with other variables.

Especially in historical copora consisting of relatively few texts, observations are clustered around certain time points (when a work is believed to be published), with large gaps inbetween.

None of the commonly used techniques to deal with time—(5)—handle this case well. And these problems become more severe as N decreases in size.

(5) a. Binning a time series to produce a totally ordered set of time windows
b. Treating time series as a genuine temporal expansion and using statistical models that are sensitive to the mathematical peculiarities of such distributions

Simulations about effect size



On what multiple occurrences mean

More occurrences from the same authors should not make you more sure about something forming a pattern.

While we certainly expect some variation in the way speakers phrase their utterances, we do not expect noise (errors) in the same way that we do in experiments.

Further, while speakers bias themselves in experiments as well (which is why we include speaker-sensitive error terms), **bias is likely to be much stronger in corpora—call it style**.

On what multiple occurrences mean: repeated measures

That is, **subsequent occurrences in a corpus have a different status from similar ratings in one condition across an experiment**. They are dependent in the first, and (largely) independent in the latter.

In effect, what we want for corpora is as many authors as we can possibly investigate. The need for repeated measures (the gold standard in experiments), on the other hand, is absent because we can be reasonably sure that speakers that produce their own texts will unsystematically employ spurious utterances (read: errors) to do so.

In a nutshell, in corpora we hit **diminishing returns** and produce wrong results if we employ methods that do not account for this type of data (see Gries 2015).

Repeated measures will also solve the problem that lowering the α-criterion is attempting to solve (see also Trafimow et al. 2018).

On variance and data presentation

Note that point estimates—think total frequencies—are not valid data summaries. To adequately interpret the data, both corpus researchers and the audience expected to understand that researcher's work need more than frequency tables.

Finally, since inferential statistics often works by partitioning variance in some way (such as contrasting expected with unexpected variance components), your test will need this information too.

Inferential statistics

Kilgarriff (2005): Corpus data is never, ever random. With sufficient amounts of data, hypothesis testing is flawed because language data do not come from a random distribution at all. And hypothesis testing does not tell us anything about associativity hypotheses. So a significant result will only tell us what we already knew: The null hypothesis is not randomly distributed.

In essence, language will give you false positives too readily.

And this problem increases as a number of N. As the number of observations increases, any random difference between means (which will almost necessarily differ from the underlying population mean) has a higher chance of reaching significance.

Also, if your corpus includes all the data that is available for a (stage of a) given language, statistical inference is meaningless. You did not sample. You tested the entire available population.

Inferential statistics

Gries (2005), on the other hand, argues that null hypothesis approaches can be used productively, once » **additional measures** are taken into account:

- » confidence intervals
- » effect sizes
- » corrections for multiple comparisons (if applicable)

Note that the countermeasures proposed by Gries are not applicable if, indeed, corpus data cannot be understood as random samples.

I will not pursue this point here, but I believe that this question deserves serious consideration.

Corpora as random samples

Koplenig (2019): But could we not just think of language as an imaginary population and assume further that all texts in copora are just randomly drawn from this imaginary population?

This would allow us to generalize from corpus findings to this imaginary population.

But there is no evidence that this is the case. Researchers would have to show that corpus data are indeed random samples, with the mathmatical properties associated with data of this type.

Note also that **inference from sample to population requires representativeness**. I will not pursue this point either, but corpus linguists should, in my view, be prepared to defend the position that, yes, the corpus they use is (demonstrably) representative of language X, whenever they apply inferential methods.

(Nothing changes when a corpus is the source of sampling, as is commonly done. While the inference to the entire corpus is naturally valid, any further inferences face the objections above.)

Please talk about dispersion

Say you find a pattern of some sort, across a wide variety of subcorpora or texts. In line with what I said above, it seems more valueable to have widely dispersed findings than clumped ones.

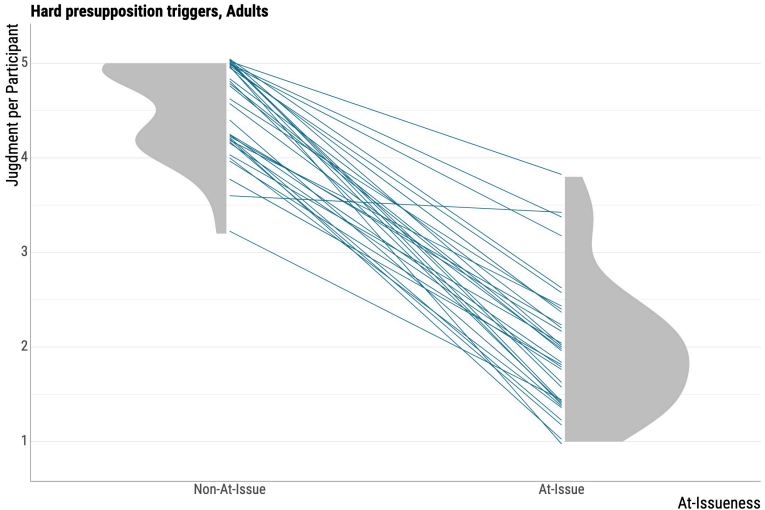
One way of addressing this is with measures of dispersion, or with mixed models, which may take into account greater or smaller than average errors (for pointers, see Gries 2015).

Regardless of how this is addressed in terms of measure, it is important to treat clumped data as qualitatively different from widely dispersed data, such that that the latter is preferable to the first.

To compare, imagine a study where all participants judged two conditions exactly identically, expect for one person, who consistently (across all items) rated them differently. If my test showed a significant difference between conditions and I took this at face value, I would get murdered during peer review.

Also, by-subcopus (with measures of dispersion/uncertainty) plots would be nice.

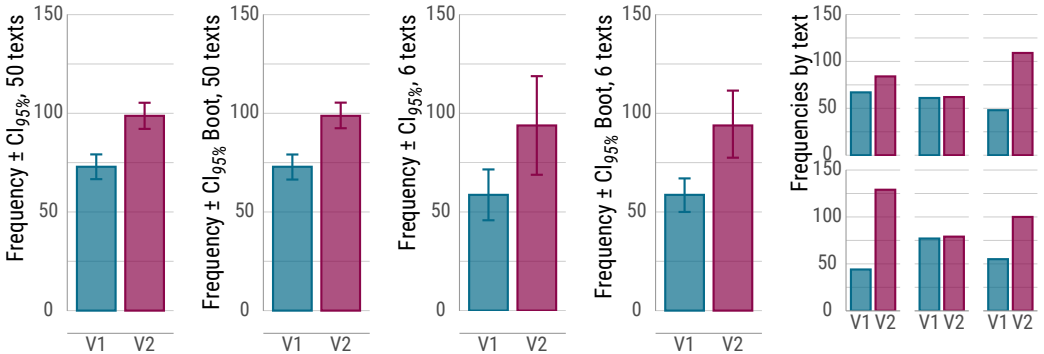
(A version of) By-participant plots with ggcorset



Dispersion: Confidence Intervals

Ideally, some version of this information should be present in all corpus studies (please do not use bar plots).

When you share your research, your audience should know accurate your findings are. Confidence intervals are great for this.



What to do

Reproducibility (we get the same results across multiple samples; your results generalize in an expected fashion to new lexicalizations and speakers — with an error rate ~ alpha)

Replicability (we can get the same results you got with the same data; essentially, this means that you should do all of your data-related work in scipted programming languages, and share the script with people upon publication of your results.)

> Actually, I see it as part of my job to inflict R on people who are perfectly happy to have never heard of it. Happiness doesn't equal proficient and efficient. In some cases the proficiency of a person serves a greater good than their momentary happiness. Patrick Burns (2005)

Plus, you would be able to use other people's annotations! Inter-annotator agreement for free. Richness captured for free.

References

Armstrong, Sharon L., Lila R. Gleitman & Henry Gleitman. 1983. What some concepts might not be. *Cognition* 13(3). 263–308.

Benincà, Paola. 2004. The left periphery of Medieval Romance. *Studi Linguistici e Filologici Online* 2(2). 243–297.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36(1). 1–22.

Gehlenborg, Nils. 2019. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. R package version 1.4.0.

Gerbrich, Hanna, Vivian Schreier & Sam Featherston. 2020. Standard items for English judgment studies: Syntax and semantics. In Sam Featherston, Robin Hörnig, Sophie von Wietersheim & Susanne Winkler (eds.), *Experiments in focus*, 305–328. Berlin: de Gruyter.

Gries, Stefan T. 2005. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2).

Gries, Stefan T. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93–117.

Keller, Frank. 2000. *Gradience in grammar experimental and computational aspects of degrees of grammaticality*. University of Edinburgh dissertation.

Kempen, Gerard & Karin Harbusch. 2005. The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence*, 329–350. Berlin: de Gruyter.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2).

Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.

Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Volume 2* 2(1).

Thalmann, Maik & Daniele Panizza. 2019. Present to the eye, away from the mind. Dissociating online comprehension and offline judgments of indirect scalar inferences. In Megan M. Brown & Brady Dailey (eds.), *Proceedings of 43rd annual Boston University Conference on Language Development*, 667–678. Somerville, MA: Cascadilla Press.

Trafimow, David et al. 2018. Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology* 9.

Verhoeven, Elisabeth & Anne Temme. 2017. Word order acceptability and word order choice. In Sam Featherston (ed.), *Linguistic evidence 2017 online proceedings*.

Weskott, Thomas, Robin Hörnig, Gisbert Fanselow & Reinhold Kliegl. 2011. Contextual licensing of marked OVS word order in German. *Linguistische Berichte* 225. 3–18.