To Introspect or not: Invitation to Experimental Syntax

2015/5/20 @ ICU
(and subsequent occasions)

Shigeto Kawahara

Keio University

The Institute of Cultural and Linguistic Studies

Introduction

Grammatical judgment in syntax: Lasnik & Saito (1984: 266-270)

* Why$_i$ do you believe the claim that John left t_i?

?*What$_i$ do you believe the claim the John bought t_i?

??Who$_i$ thinks that who$_j$ won the election?

?Who$_i$ t_i thinks that for that reason$_j$ John left t_j?

Who$_i$ do you believe t_i to be intelligent?

A bit of self-introduction

- I got interested in generative linguistics during my undergraduate time at ICU (around 1999).
- I was first interested in syntax, thanks to Prof. Tomo Yoshida.
- I converted to a phonologist during my exchange study at UC, Santa Cruz.
- The reason, partly, was because I couldn't follow the judgment patterns in English…or, in all honesty, some of those in Japanese.
- I wrote a senior thesis on phonological theory, which is, with hindsight, surprisingly good (which I even came back to in 2014).

Continued

- But that BA thesis used any kind of "sound-related pattern" as phonological.
- In 2001, when I presented the theory at a conference, Prof. Haraguchi warned me against that attitude, although I did not take it very well.
- During my graduate time at UMass, I converted again to an experimentalist. I started appreciating that advice.
- During my professorial time at Rutgers, I was interacting with Julien Musolino and Kristen Syrett, who were concerned about the quality of syntactic and semantic judgments on scopal interactions.
- I now think that examining the quality of evidence for generative studies in general is crucial (Kawahara 2015, *Linguistic Vanguard*).

Grammatical judgment in syntax: Lasnik and Saito (1984: 266-270)

* Why$_i$ do you believe the claim that John left t_i?

?*What$_i$ do you believe the claim the John bought t_i?

??Who$_i$ thinks that who$_j$ won the election?

?Who$_i$ t_i thinks that for that reason$_j$ John left t_j?

Who$_i$ do you believe t_i to be intelligent?

Cited and discussed in Pullum 2013

A Japanese example

Taroo-wa[[Hanako-ga nani-o katta]ka] kikimashita ka ?

- Did Taro ask Hanako what she bought? (embedded scope)

-?? What is x, such that Taro asked Hanako whether she bought x? (matrix scope)

-There has been a huge debate about whether the embedded wh-element can take a matrix scope (i.e. whether Japanese has a wh-island effect or not).

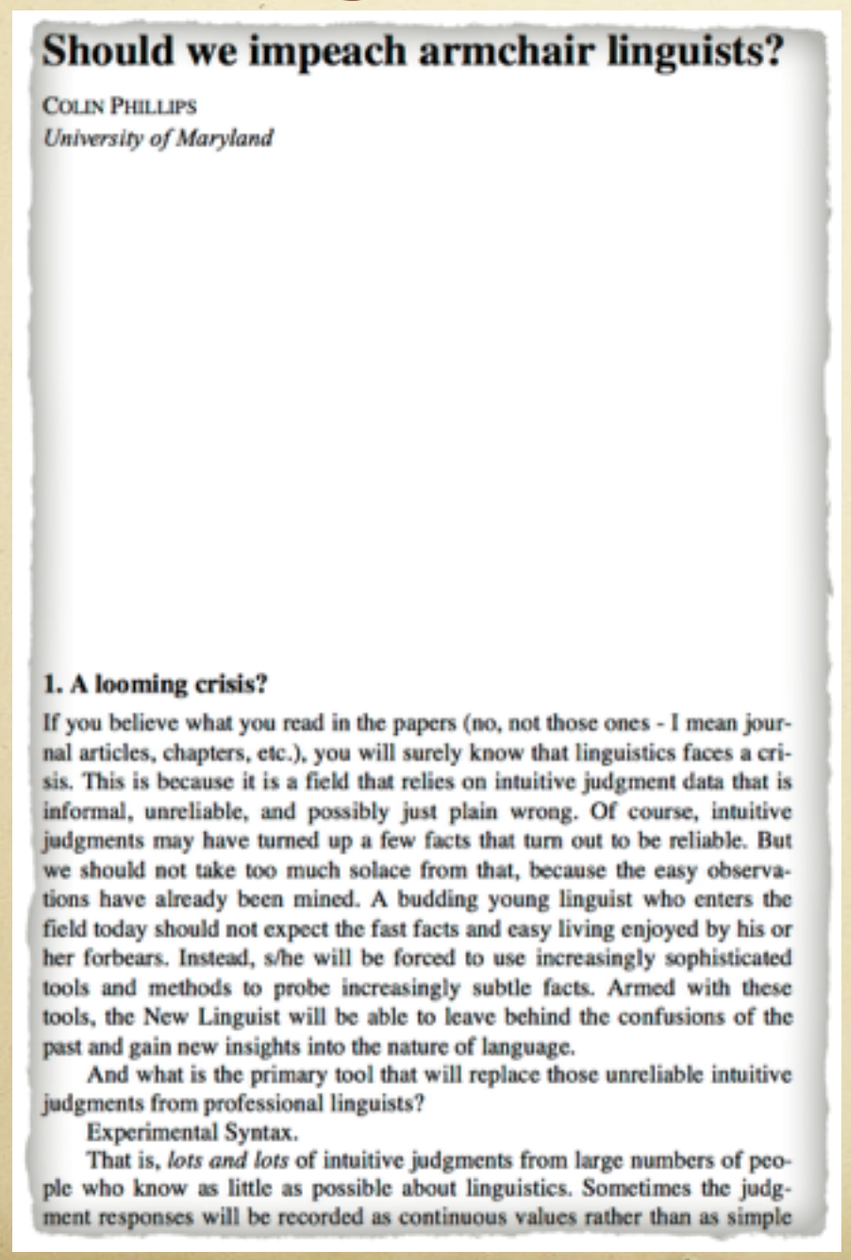
My position

- To be clear, I am not an anti-generative linguist. It may probably be safe to say that I (sometimes/often/from time to time) study languages from the generative perspective.

- But I do think that the methodology—or attitude, if I may—of the generative linguistics should be better justified and defended.

- I am not even claiming that the generative methodology is wrong. We need quantitative justification of our methodology (even if that is for a sociological reason).

Recommended (short) background reading



Should we impeach armchair linguists?

COLIN PHILLIPS
University of Maryland

1. A looming crisis?

If you believe what you read in the papers (no, not those ones - I mean journal articles, chapters, etc.), you will surely know that linguistics faces a crisis. This is because it is a field that relies on intuitive judgment data that is informal, unreliable, and possibly just plain wrong. Of course, intuitive judgments may have turned up a few facts that turn out to be reliable. But we should not take too much solace from that, because the easy observations have already been mined. A budding young linguist who enters the field today should not expect the fast facts and easy living enjoyed by his or her forbears. Instead, s/he will be forced to use increasingly sophisticated tools and methods to probe increasingly subtle facts. Armed with these tools, the New Linguist will be able to leave behind the confusions of the past and gain new insights into the nature of language.

And what is the primary tool that will replace those unreliable intuitive judgments from professional linguists?

Experimental Syntax.

That is, *lots and lots* of intuitive judgments from large numbers of people who know as little as possible about linguistics. Sometimes the judgment responses will be recorded as continuous values rather than as simple

Available online

Some criticisms 1 (from the assigned reading: Phillips 2009)

"Generative theories appear to rest on a weak empirical foundation, due to the reliance on informally gathered grammaticality judgments." (Ferreira 2005, p. 365)

"Judgments are inherently unreliable because of their unavoidable meta-cognitive overtones, because grammaticality is better described as a graded quantity, and for a host of other reasons." (Edelman & Christianson 2003, p. 60)

Some criticisms 2

"One might in fact conclude that we have not yet developed a means to evaluate empirical bases for hypotheses in generative grammar that is compelling enough to the majority of the practitioners. An evaluation of a given hypothesis thus tends to have an arbitrary aspect to it, influenced by such factors as whether or not the terms and concepts utilized are taken from a theory currently in fashion .." (Hoji & Ueyama 2007, p. 2)

Some criticisms 3

"Unfortunately, the findings of the experimentalists in linguistics very rarely play a role in the work of generative grammarians. Rather, theory development tends to follow its own course, tested only by the unreliable and sometimes malleable intuitions of the theorists themselves. The theories are consequently of questionable relevance to the facts of language." (Wasow & Arnold 2005, p. 495)

Some criticisms 4

- “Chomsky’s policy that the subject matter of linguistics is psychological in nature does not provide any reason for assuming that the purported facts that linguists have hitherto adduced as evidence for or against particular analyses are psychological in nature, nor even that they are strictly speaking facts (McCawley 1986: 28).”

Box 1. Cognitive biases and linguistic judgments.

There are at least three types of unconscious cognitive biases [8,9] that can adversely affect the results of intuitive judgments, given the way that they are currently typically gathered in the syntax/semantics literature:

1. Confirmation bias on the part of the researcher: researchers will often have a bias favoring the success of the predicted result, with the consequence that they will tend to treat data that do not support the hypothesis as flawed in some way (e.g. from a not quite native speaker, or from a speaker of a different dialect).
2. Confirmation bias on the part of the participants: individuals that the researcher asks to provide a judgment on a linguistic example – including the researcher him/herself – might be biased because they understand the hypotheses. When faced with complex materials, they could then use these hypotheses to arrive at the judgment.
3. Observer–expectancy effects (the "clever Hans" effect): individuals that the researcher asks to provide a judgment could be biased because they subconsciously want to please the researcher and are consequently affected by the researcher's subtle positive/negative reactions.

Gibson & Fedorenko 2010

Summary of the issues

- Generative grammarians generally rely on introspection-based data.
- This methodology is often frowned upon by psychologists and cognitive scientists (and phoneticians).
- Not quantitative (statistically supported?); unreliable; cherry-picking; biased…
- Again, I am not saying that these criticisms are valid outright. I am saying that there's not enough constructive conversation.

Part I: Some examples of experimental syntax

Superiority condition

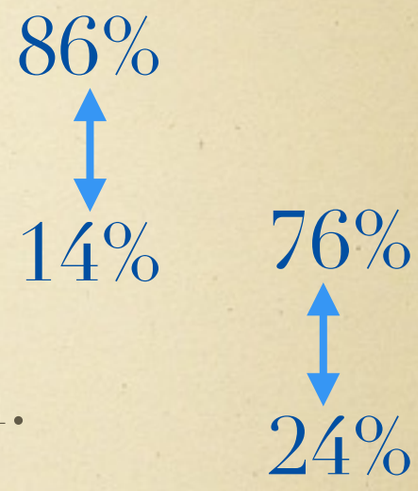
- Superiority condition:
  - I'd like to know who_i t_i hid where.
  - *I'd like to know $where_i$ who hid it t_i.
- Superiority condition ameliorated with an additional wh-element (Kayne 1983)?
  - ?I'd like to know $where_i$ who hid it when t_i.

Clifton et al. (2006)

a. I'd like to know who hid it where.

b. (*)I'd like to know where who hid it.

c. (?)I'd like to know where who hid it when.



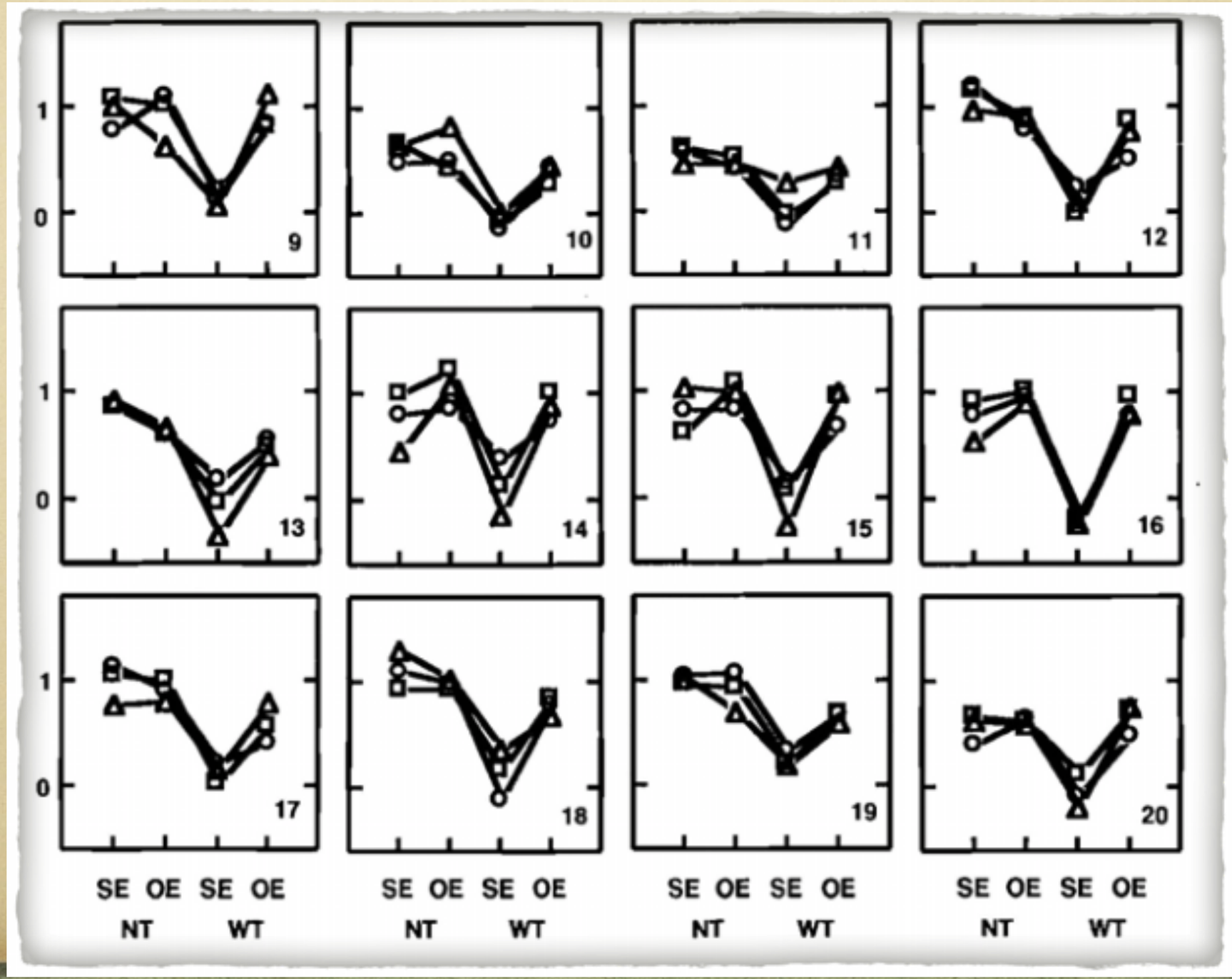
Comparing (b) and (c), this grammatical judgment experiment shows that the extra wh-element does not really "rescue" the superiority violation, contra Kayne (1983).

That-trace effect

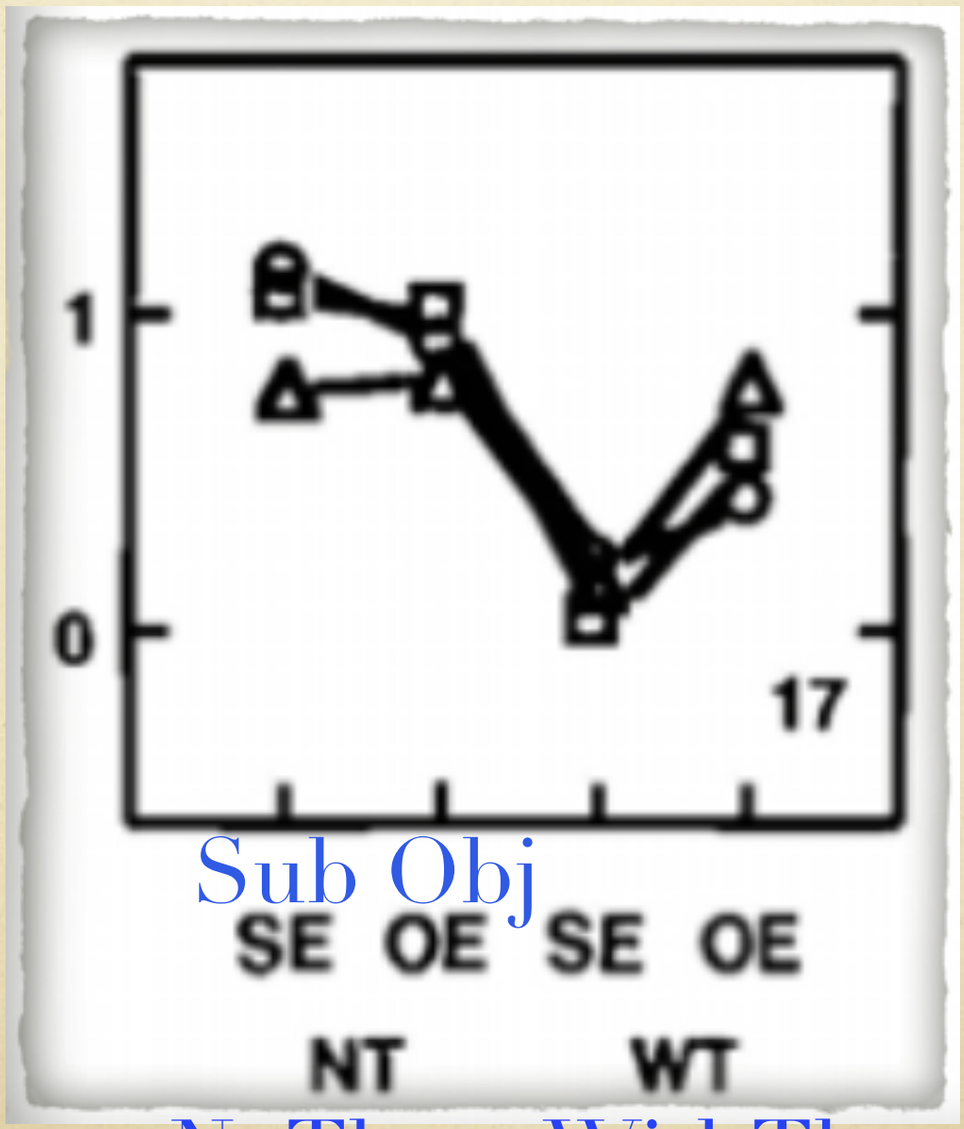
- Who$_i$ do you think t_i would win?
- *Who$_i$ do you think **that** t_i would win?

- Who$_i$ do you think John would hit t_i?
- Who$_i$ do you think that John would hit t_i?

Cross-speaker consistency: Cowart (1997)



Zooming in



The message

- Some syntactic effects are replicable with experimentation (the that-trace effect).
- Some effects are not (the amelioration of the superiority violation).
- What should we do when intuition-based data and experimental data disagree?

Availability of Pair-list answers (Achimova et al., NELS)

Table 1. Availability of pair-list answers for subject questions with object quantifiers*

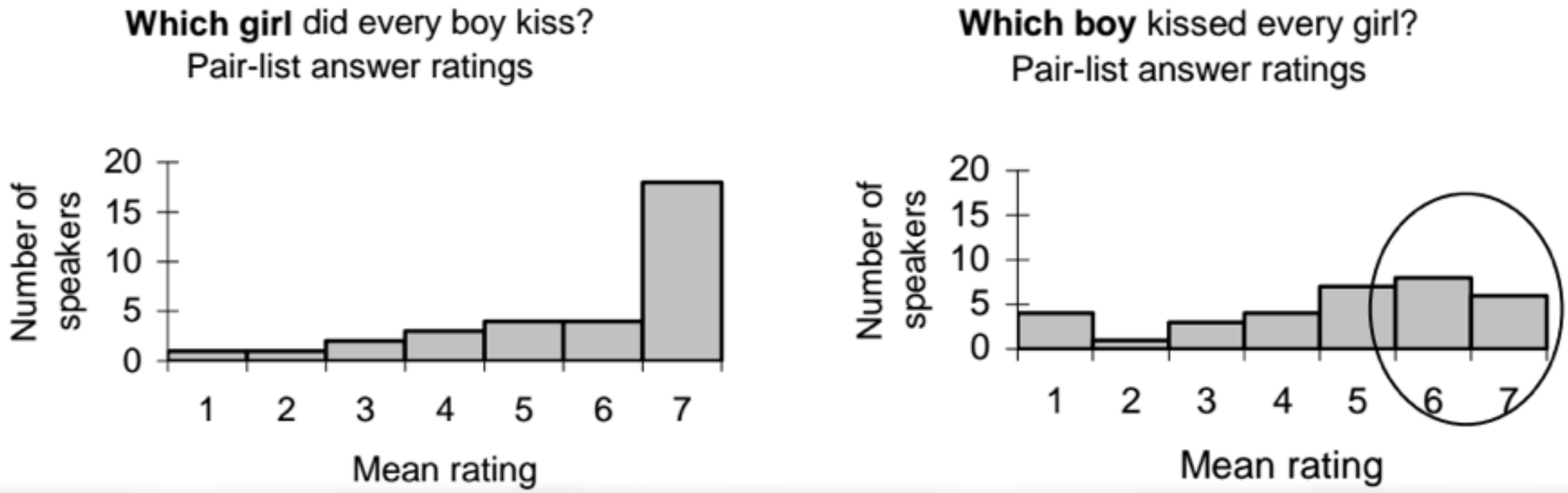
Subject questions	**May (1985)**	**Chierchia (1993)**	**Beghelli (1997)**	**Agüero-Bautista (2001)**
Who kissed every girl?	-	+	-	+
Which boy kissed every girl?	-	-	-	-
Which boys kissed every girl?	-	+	-	-
Which boy kissed each girl?	+		+	+

* Plus signs indicate that PLA are possible and minus signs that they are unavailable.

subject/object asymmetry?

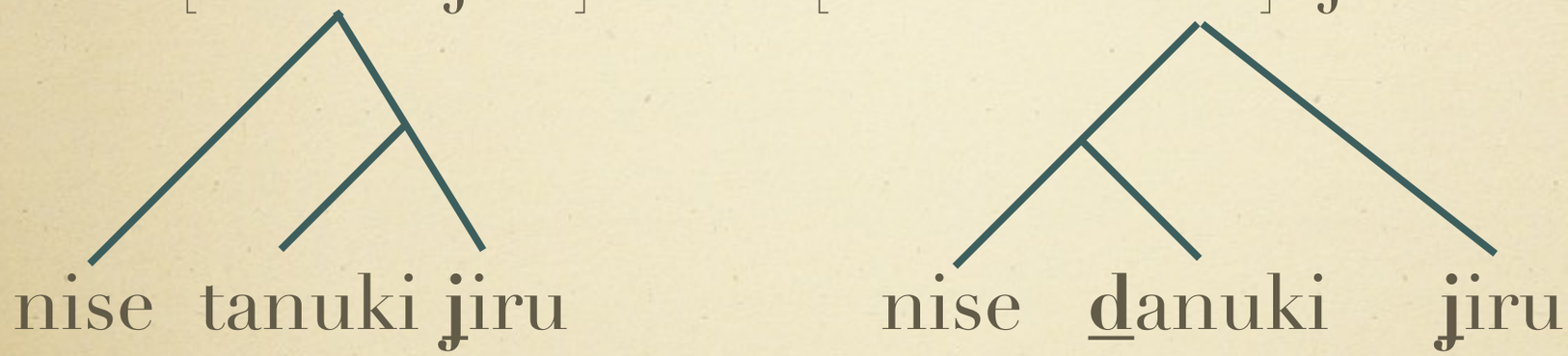
- They found a subject/object asymmetry.
- and between-speaker variability at the same time!

Figure 2. Distribution of ratings (averages across 4 items of a given type)



Right Branch Condition and rendaku (Otsu 1980)

nise [tanuki-jiru] vs. [nise danuki] jiru



- Rendaku only applies to elements on the right branch.
- A celebrated example in intro to Japanese linguistics class. The role of c-command in phonology? (and much subsequent work by Ito and Mester)

Kumagai (2014)

- The status of Right Branch Condition (RBC) has been debated (Vance 1980; Kubozono 2005), but it was more or less taken for granted in theoretical phonology.
- In particular, Ito and Mester's theory of rendaku often developed around RBC (1986, 2003, 2007).
- Kumagai's (2014) experiment shows that only a small set of naive native speakers are sensitive to RBC. Many of them could not choose the "right" rendaku forms, given contexts.

Kumagai's results

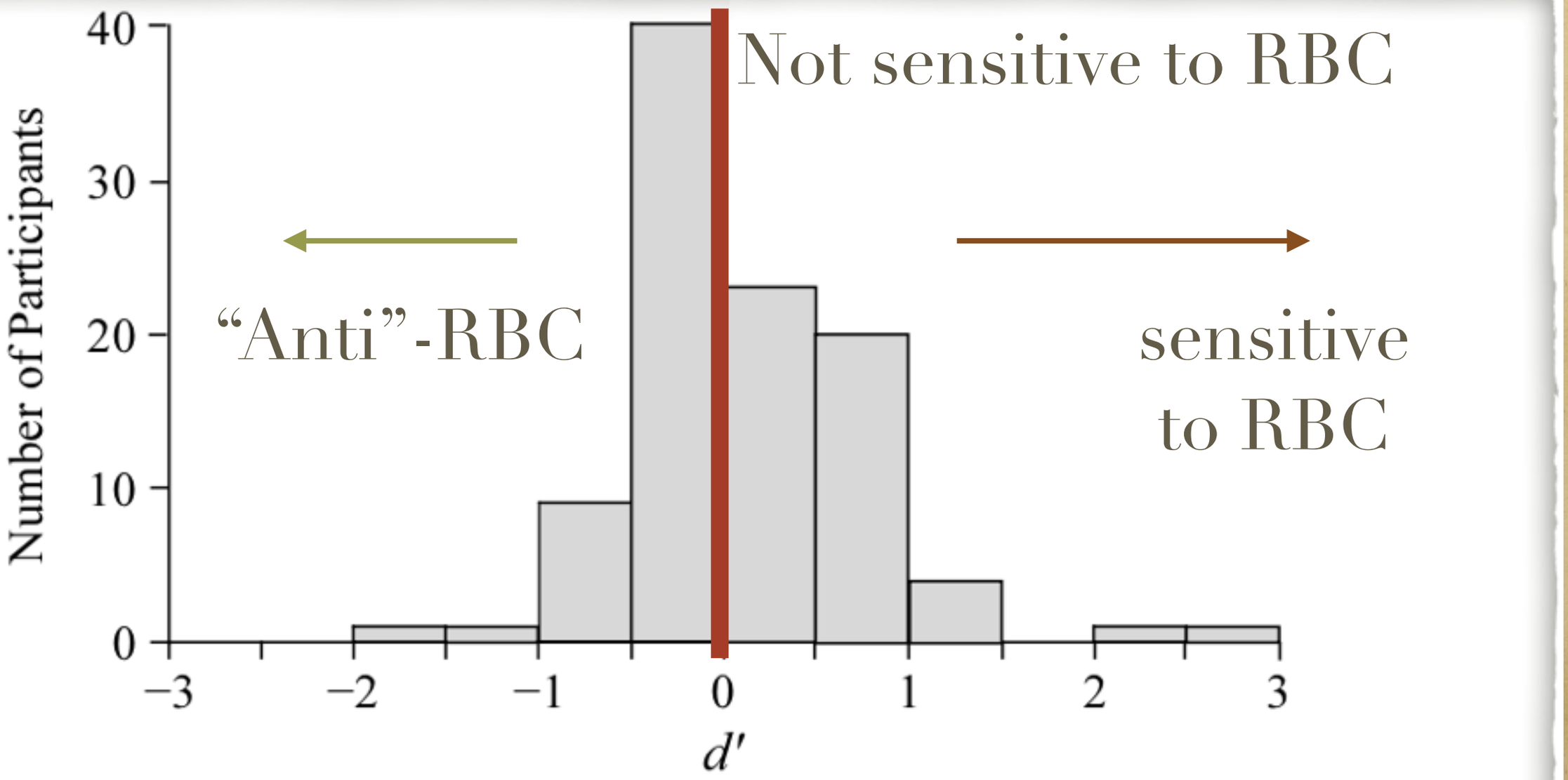


figure 1. Distribution of d' values

From an anti-generative perspective

PROCEEDINGS OF THE ROYAL SOCIETY B

CrossMark click for updates

FirstCite e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2012.1741
Published online

Review

How hierarchical is language use?

Stefan L. Frank[1,*], Rens Bod[2] and Morten H. Christiansen[3]

[1]*Department of Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way, London WC1H 0AP, UK*
[2]*Institute for Logic, Language and Information, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*
[3]*Department of Psychology, Cornell University, 228 Uris Hall, Ithaca, NY 14853-7601, USA*

It is generally assumed that hierarchical phrase structure plays a central role in human language. However, considerations of simplicity and evolutionary continuity suggest that hierarchical structure should not be invoked too hastily. Indeed, recent neurophysiological, behavioural and computational studies show that sequential sentence structure has considerable explanatory power and that hierarchical processing is often not involved. In this paper, we review evidence from the recent literature supporting the hypothesis that sequential structure may be fundamental to the comprehension, production and acquisition of human language. Moreover, we provide a preliminary sketch outlining a non-hierarchical model of language use and discuss its implications and testable predictions. If linguistic phenomena can be explained by sequential rather than hierarchical structure, this will have considerable impact in a wide range of fields, such as linguistics, ethology, cognitive neuroscience, psychology and computer science.

Keywords: language structure; language evolution; cognitive neuroscience; psycholinguistics; computational linguistics

"A heresy article" for generative linguistics?

GG's general concern (cf. Pesetsky's LSA plenary talk)

Against hierarchical structure?

"In this paper, we question this practice [of using hierarchical structures], not so much for language analysis but for the description of language use (p. 2)."

"This is not to say that hierarchical operations are non-existent, and we do not want to exclude their possible role in language comprehension or production. However, we expect that evidence for hierarchical operations will only be found when the language user is particularly attentive, when it is important for the task at hand (e.g. in meta-linguistic tasks) (p. 8)"

- Maybe it would be fine that RBC is visible when the language users “are particularly attentive”, as long as RBC exists
- RBC is competence; the null results in the experiments were an issue of performance.
- (But performance should not be used as a “convenient trash box”)

The Japanese wh-island effect (or lack thereof) revisited

Taroo-wa[[Hanako-ga nani-o katta]ka] kikimashita ka ?

- Did Taro ask Hanako what she bought?

-?? What is x, such that Taro asked Hanako whether she bought x?

The availability of the matrix scope reading has been controversial (Nishigauchi 1990 vs. Lasnik & Saito 1984),

Yoshida 1998

- Dare-ga katta ndai?
- *Taroo-ga katta ndai?
  - “ndai” requires a wh-reading.
- *John-wa [Mary-ga nani-o katta ka] shitteiru ndai?
- The matrix reading should be impossible!

But wait! Isn't prosody relevant?

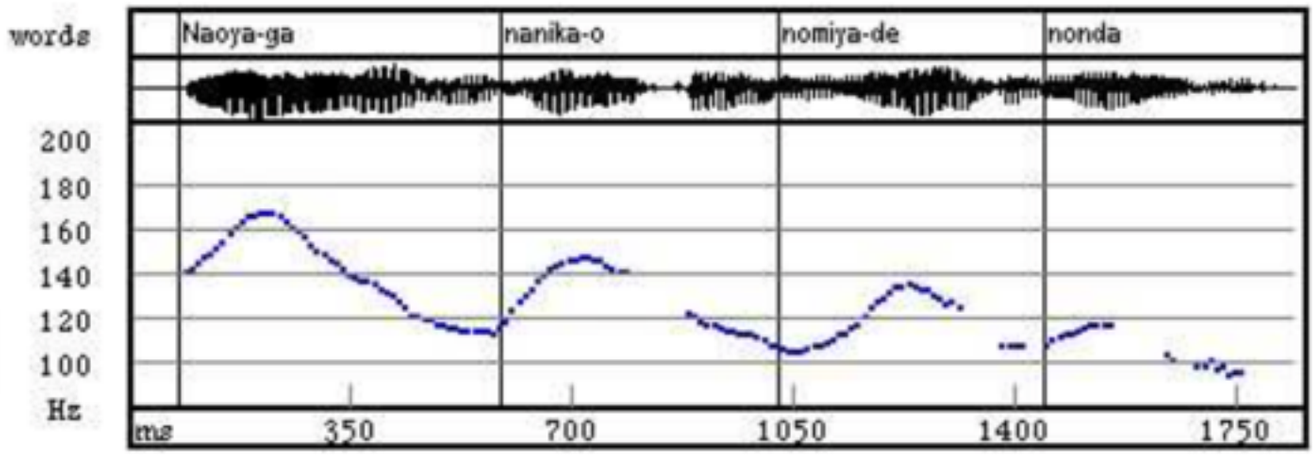
- For a long time, there has been an observation that focal accent on the embedded wh-element can yield the matrix reading (also acknowledged in Yoshida 1998; Tomioka 1997).

› John-wa [Mary-ga NANI-o katta ka] kikimashita ka

- Maybe the intonation is crucial? (Deguchi & Kitagawa 2002; Ishihara 2003 *et seq*.)

Ishihara (2003)

(28a): Non-interrogative sentence



(28b): *Wh*-question

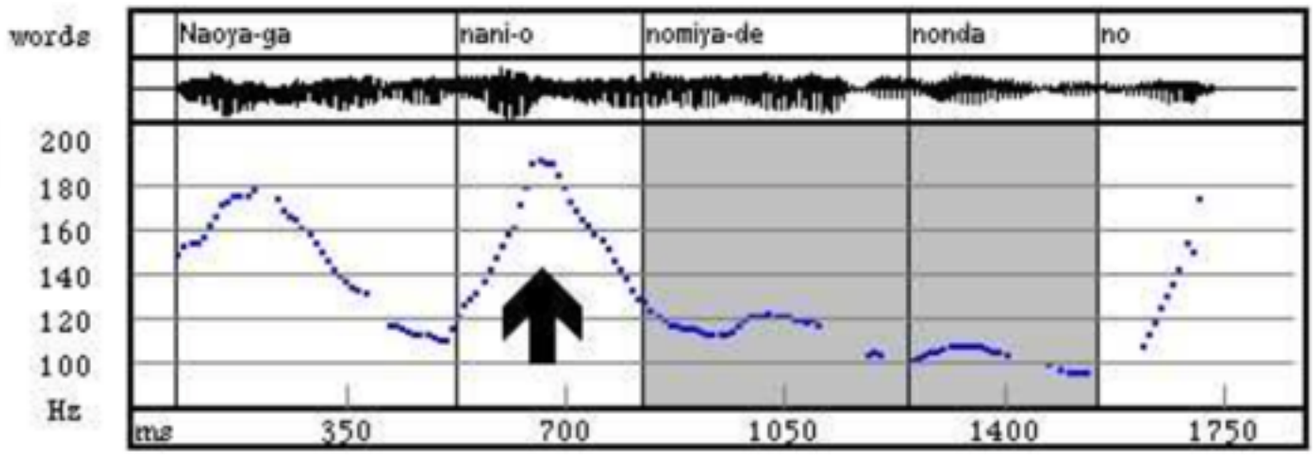


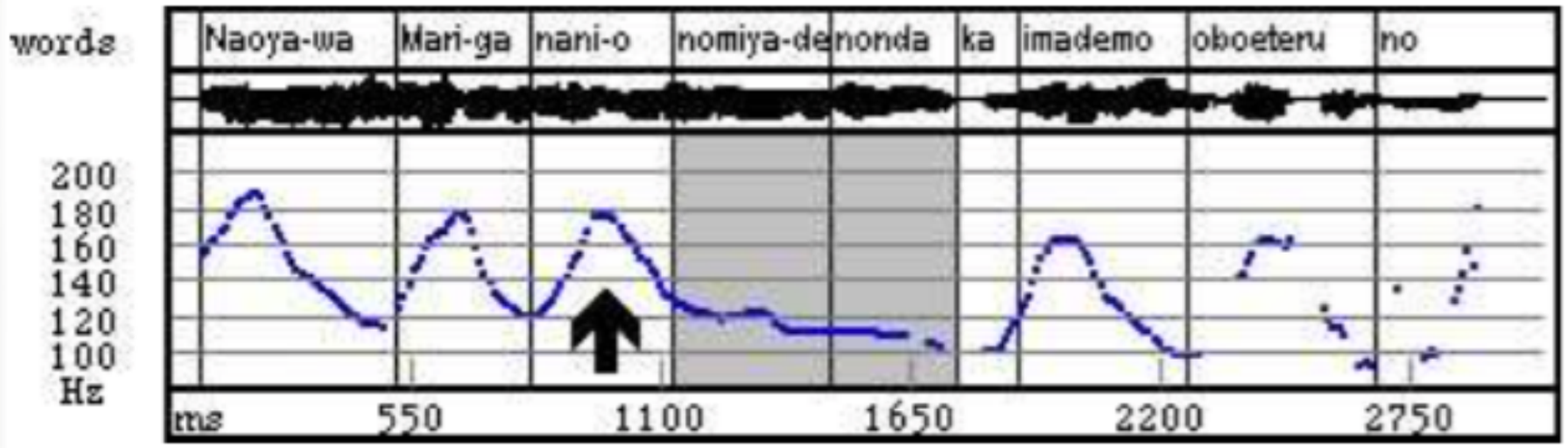
Figure 3-1: Single *wh*-question

Focal boost on wh-elements.

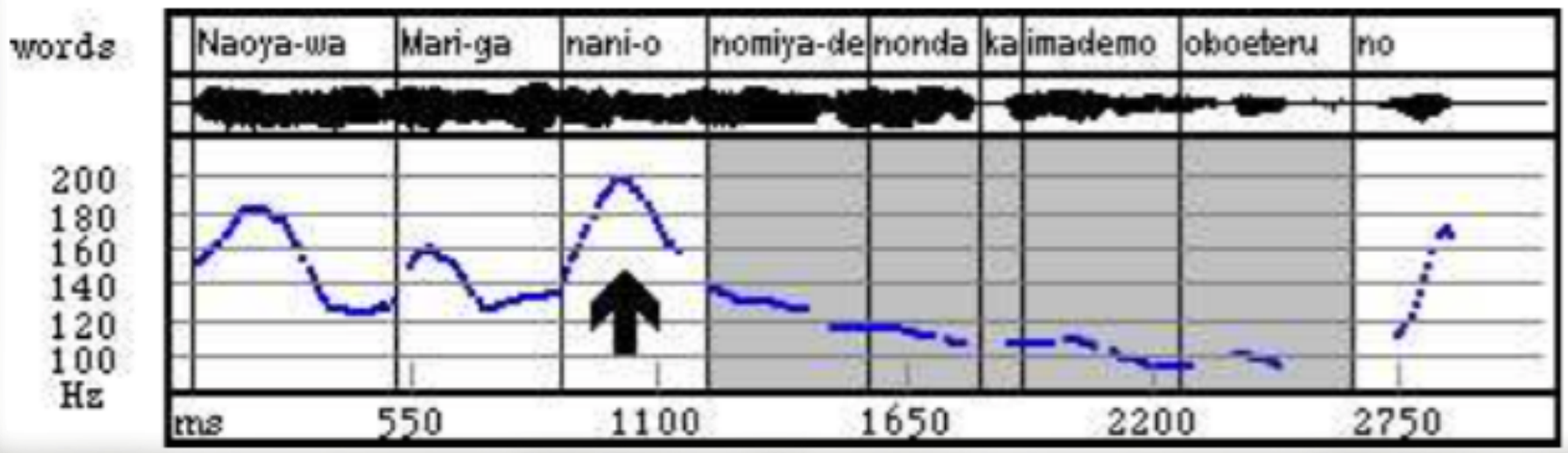
Pitch compression effect after wh-elements (focused elements in general)

Ishihara (2003)

(34a): 'Does Naoya still remember what$_i$ Mari drank t_i at the bar?'



(34b): 'What$_i$ does Naoya still remember whether Mari drank t_i at the bar?'



The pitch compression domain = the scope of the wh-element.

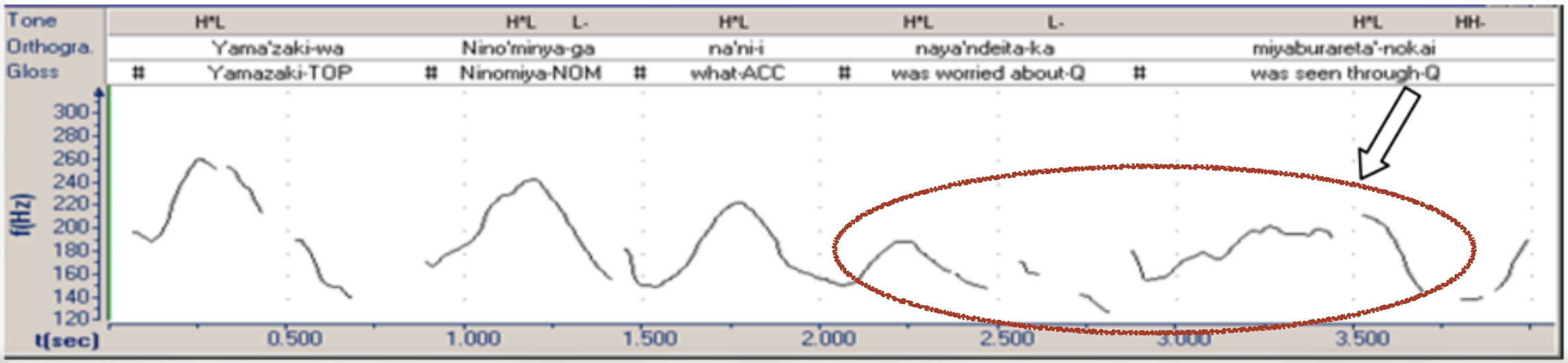
=> further developed by N. Richards as the universal condition on licensing wh-elements.

Richard's Uttering Trees

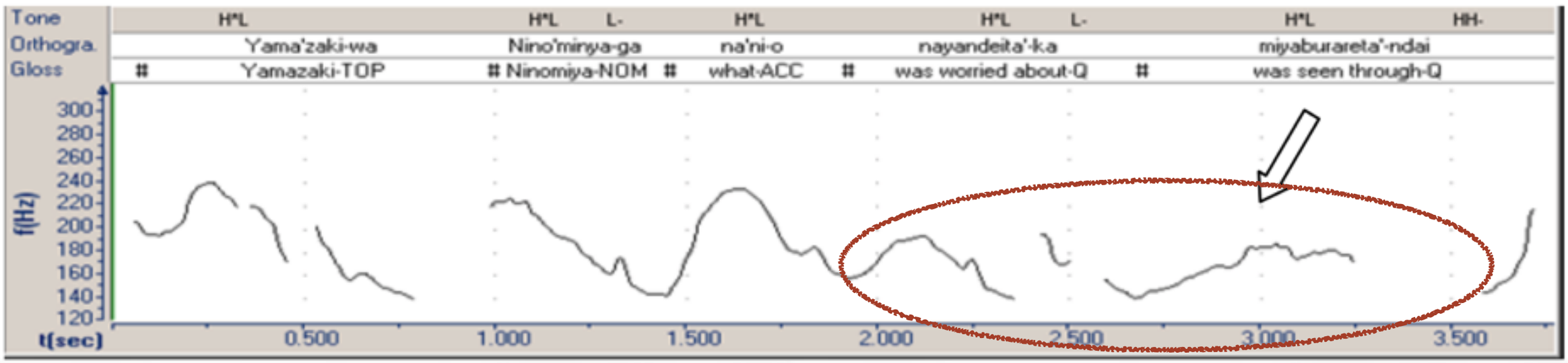
- Richards's second proposal, "Beyond Strength and Weakness," is an attempt to predict, for any given language, whether that language will exhibit overt or covert wh-movement. Richards argues that we can predict whether or not a language can leave wh in situ by investigating more general properties of its prosody. This proposal offers an explanation for a cross-linguistic difference—that wh-phrases move overtly in some languages and covertly in others—that has hitherto been simply stipulated. In both these areas, it appears that syntax begins constructing a phonological representation earlier than previously thought; constraints on both word order and prosody begin at the beginning of the derivation.

Hirotani (2005): Production

山崎は　飲み屋が　何に　悩んでいたか　見破れたのかい？



山崎は　飲み屋が　何を　悩んでいたか　見破れたんだい？



Hirotani (2005): Comprehension

TABLE 3.1. % choice of embedded (yes/no) question reading and standard deviation for each condition (Experiment 4)

	Embedded Reading Response	
	Frequency (%)	*SD*
A: Boundary, Unscrambled	84	26
B: No boundary, Unscrambled	43	35
C: Boundary, Scrambled	84	23
D: No boundary, Scrambled	42	36

- It is not (so much) a matter of production, but comprehension? (viz. is it a matter of competence at all?).
- What if the stimuli did not make sense to the participants, and they could only hear focused NANI?

Relegating everything to performance?

- The reductionist approach: attributing "grammatical effects" to non-linguistic, cognitive issues.

- A prominent example is wh-island effects; one could argue that these sentences are considered as "bad", because they are hard to process (e.g. Kluender 2005).

- Sprouse et al. (2012) show that there are no correlations between each individual's working-memory capacity and ratings of island-violating sentences.

What's crucial is not the conclusion *per se*

- There is an actual conversation here, between "reductionists" and "generativists" on the same footing.

- I find this line of research very promising. There's the same kind of tension between phoneticians and phonologists (cf. J. Ohala's work).

- Recommended reading: "Colin Phillips (2013) Some arguments and non-arguments for reductionist accounts of syntactic phenomena. Language and Cognitive Processes 28:156-187."

Some messages

- Experiments may not support "linguistic effects" outright (e.g. RBC). But that does not mean that those linguistic effects are not real.
- Experiments may reveal some complicated issues, such as a difference b/w linguists and non-linguists, b/w production and perception, b.w competence and performance, etc.
- Experiments allow generativists to constructively respond to "reductionism".

Part II: Addressing the criticism empirically

Recent responses to the criticisms against generative methodology

- Some conceptual responses (e.g. Marants 2005; Phillips 2009).

- An additional way to be even more constructive: Experimentally testing that "the introspection-based data out there" is in fact reliable.

- Jon Sprouse (UConn) has been developing a notable research program along this line.

Sprouse, Jon, Carson T. Schütze, & Diogo Almeida (2013) A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua* 134: 219-248.

From the abstract

- “We tested 296 data points from the approximately 1743 English data points that were published in Linguistic Inquiry between 2001 and 2010. We tested this sample with 936 naïve participants using three formal judgment tasks (magnitude estimation, 7-point Likert scale, and two-alternative forced-choice) and report five statistical analyses. The results suggest a convergence rate of 95% between informal and formal methods, with a margin of error of 5.3–5.8%."

Sample pair-wise comparisons

- ?? Thomas tried to have stopped the thief.
- Thomas tried to stop the thief.

(Martin 32.1)

- ??What did the students claim at that time that the teacher said?
- At that time, what did the students claim that the teacher said?

(Boskovic 34.4)

Targets

- 150 such pairs, one member of which was deemed unacceptable. Randomly sampled from LI (2001-2010).
- Only US English data. (A hint for future studies!)
- Tested these 300 pairs of sentences with 8 different lexical compositions (300*8=2400).
- Collected responses (online) from participants who are not linguists.

Directionality analysis

	Predicted diretion	Opposite direction	Total
ME (MagEst)	146 (99%)	2	148
LS (LikSc)	143 (97%)	5	148
FC (FrcdCh)	144 (97%)	4	148

Statistical analyses (sample)

Table 3
Categorized results of statistical tests for ME. Significant *p*-values are defined at $p < .05$ in each direction; marginal *p*-values are defined at $p \leq .1$ in each direction. Significant Bayes factors are defined at BF > 3 in each direction; marginal Bayes factors are defined at BF > 1 in each direction.

	One-tailed	Two-tailed	LME	Bayes factor
Significant in the opposite direction	–	2	2	2
Marginal in the opposite direction	–	0	0	0
Non-significant in the opposite direction	–	0	0	0
Non-significant in the predicted direction	10	9	16	13
Marginal in the predicted direction	1	1	3	2
Significant in the predicted direction	137	136	127	131

Sprouse, Jon & Diogo Almeida (2012) Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48: 609-652.

The abstract

- “In this paper we empirically assess this claim by formally testing all 469 (unique, US-English) data points from a popular syntax textbook (Adger 2003) using 440 naive participants, two judgment tasks (magnitude estimation and yes–no), and three different types of statistical analyses (standard frequentist tests, linear mixed effects models, and Bayes factor analyses). The results suggest that the maximum discrepancy between traditional methods and formal experimental methods is 2 %”

A brief look into the content

	Frequentist	Linear mixed effects	Bayes factors
Significant in the opposite direction	0	0	0
Non-significant	2	3	0
Marginal	1	0	2
Significant in the predicted direction	112	112	102
Replication failure rate	2.6%	2.6%	1.9%

Table 5

Counts of the replications and failures for the ME experiments. The failure rate includes marginal results as replication failures to derive a maximum failure rate.

Summary: Sprouse (2015)

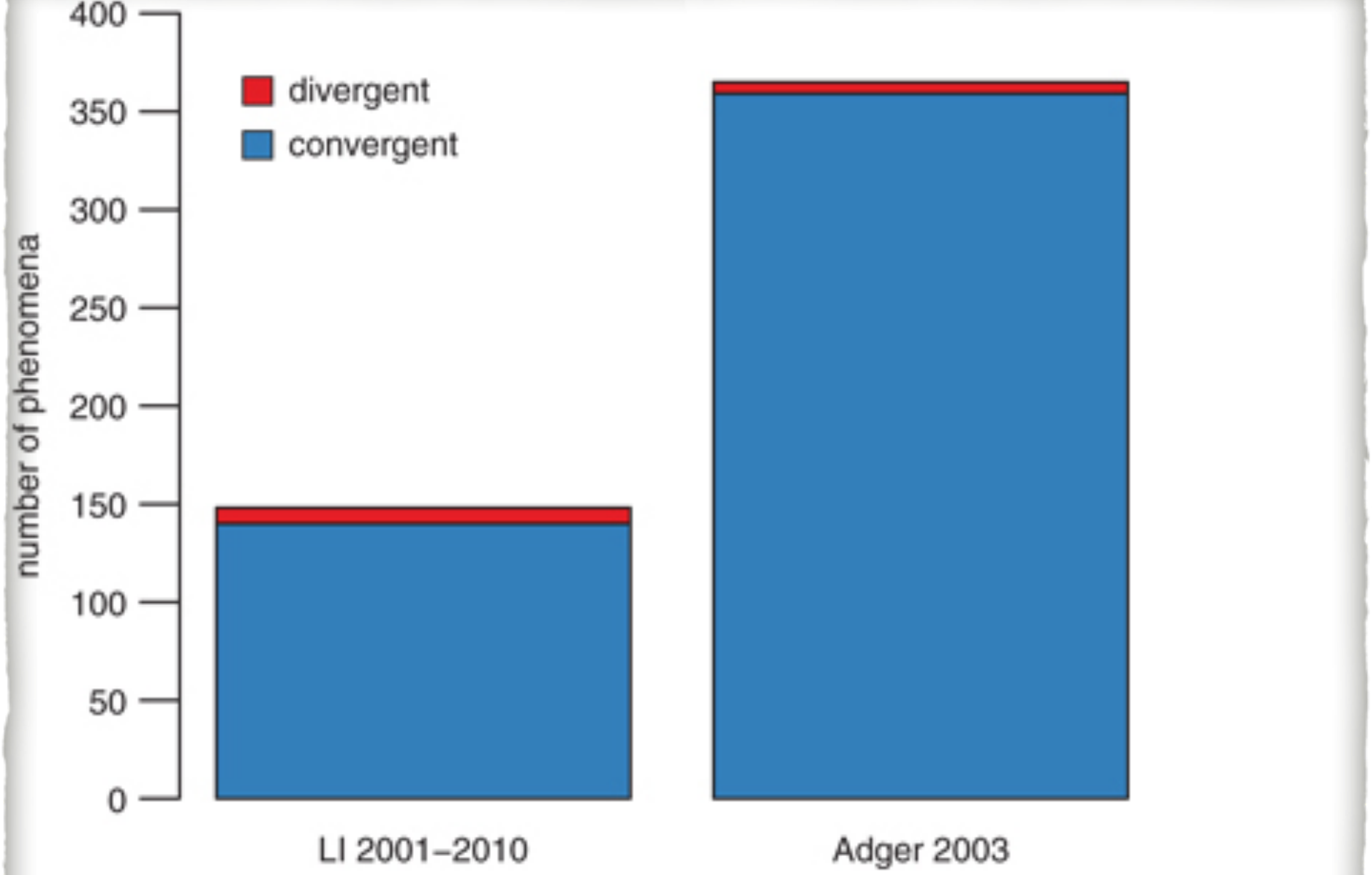


Figure 1

Convergent and divergent phenomena in *Linguistic Inquiry 2001–2010* and Adger (2003)

- I wonder if we can do a similar kind of assessment for the Japanese data out there.

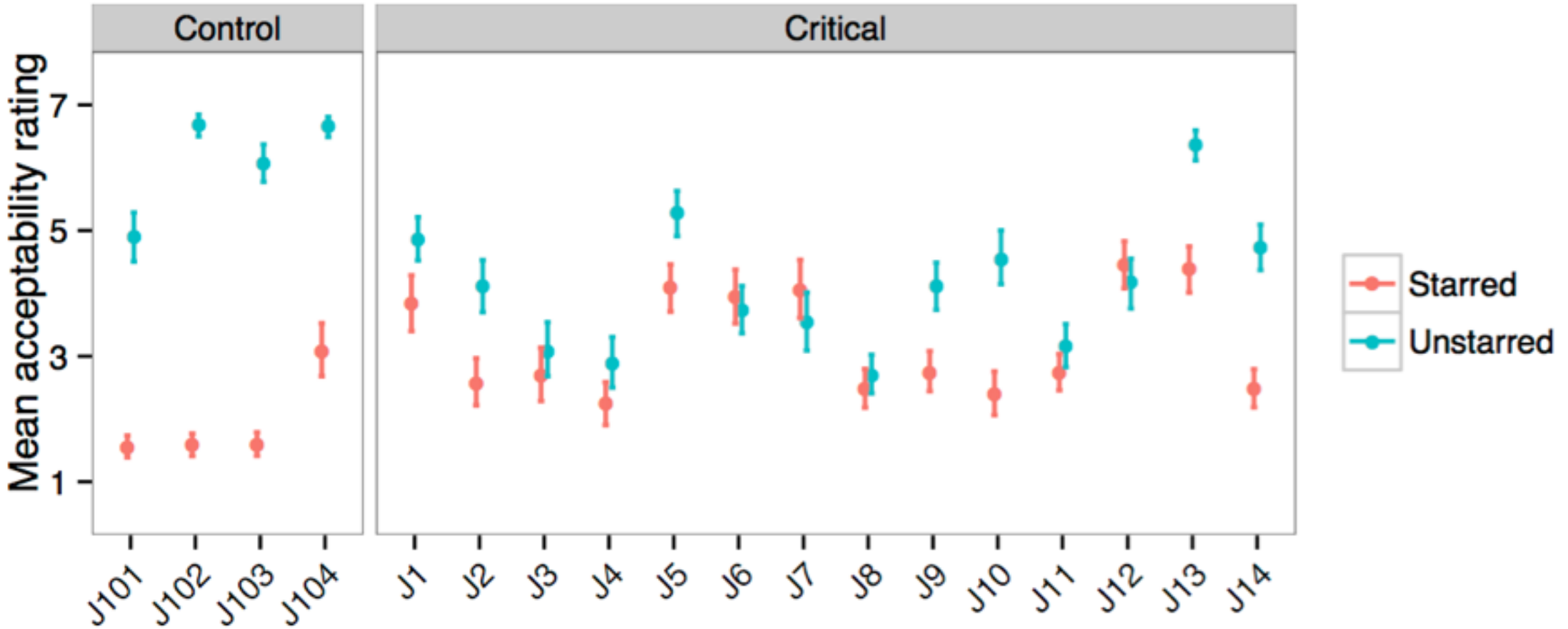
- A great topic for your research! A similar study (perhaps with a smaller scale) can be a great topic for your BA or MA research.

- See Linsen & Oseki (2015) below.

Linzen & Oseki (2015)

- The reliability of syntactic acceptability judgments has come under criticism in recent years. Studies conducted in response have shown that the vast majority of published judgments in English are robust. We present two judgment collection experiments, in Hebrew and Japanese, which investigated whether this holds for less widely spoken languages as well. Between a third and a half of the judgments we deemed questionable did not replicate. We argue that English judgments are more robust because of an informal peer review process; to extend this process to other languages, we propose an online platform that would enable vetting judgments or expressing concern about them.

The Japanese results



(b)

Figure 1: Results of the experiments: (a) Hebrew and (b) Japanese. Error bars represent bootstrapped 95% confidence intervals.

(28) J5 (Tada, 1992):

a. Taro-wa migime-dake-o tumur-e-ru.
Taro-Top right.eye-only-Acc close-can-Pres
'Taro can wink his right eye.' (can > only)

b. *Taro-wa migime-dake-ga tumur-e-ru.
Taro-Top right.eye-only-Nom close-can-Pres
'Taro can wink his right eye.' (can > only)

(29) J6 (Sakai, 1994):

a. Mary$_i$-no [kanozyo$_i$-ga kik-anakat-ta] hihan
Mary-Gen she-Nom hear-Neg-Past criticism
'Mary's criticism that she did not hear'

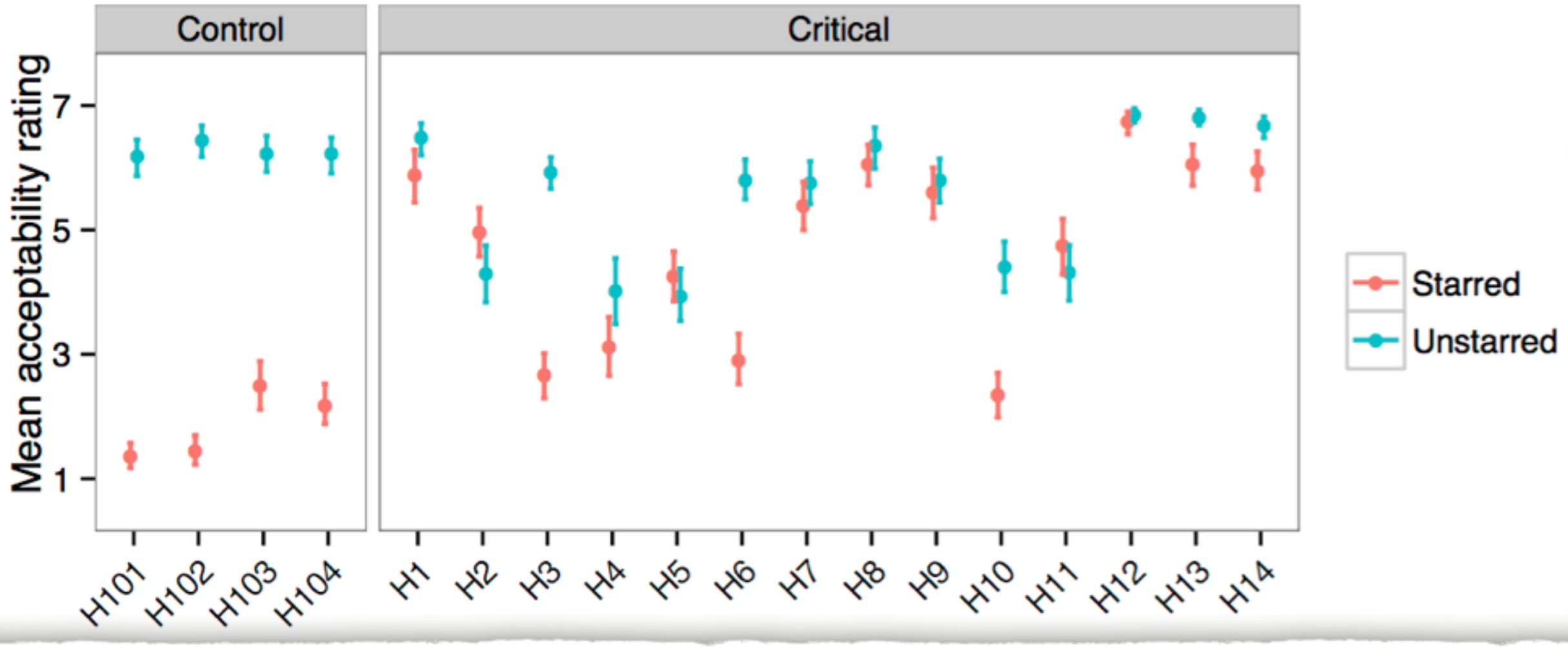
b. *Mary$_i$-no [kanozyo$_i$-no kik-anakat-ta] hihan
Mary-Gen she-Gen hear-Neg-Past criticism
'Mary's criticism that she did not hear'

(30) J7 (Oku, 1998):

a. Taro-wa zibun-no gakusei-o home-ta. Ziro-wa home-nakat-ta.
Taro-Top self-Gen student-Acc praise-Past Ziro-Top praise-Neg-Past
'Taro praised Taro's student. Ziro did not praise Ziro's student.'

b. *Taro-wa kuruma-o teineini arat-ta. Ziro-wa araw-anakat-ta.
Taro-Top car-Acc carefully wash-Past Ziro-Top wash-Neg-Past
'Taro washed cars carefully. Ziro did not wash cars carefully.'

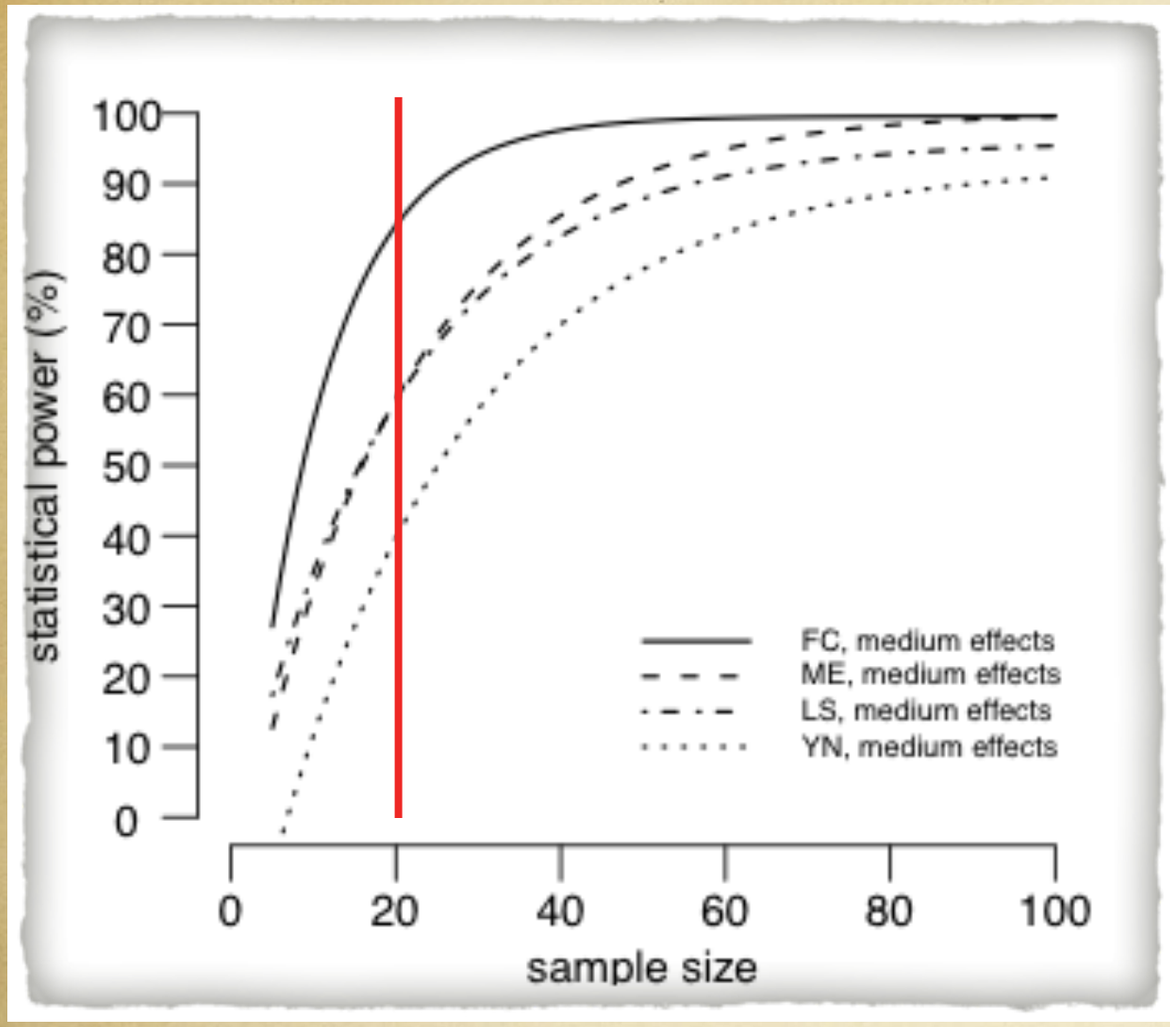
The Hebrew data

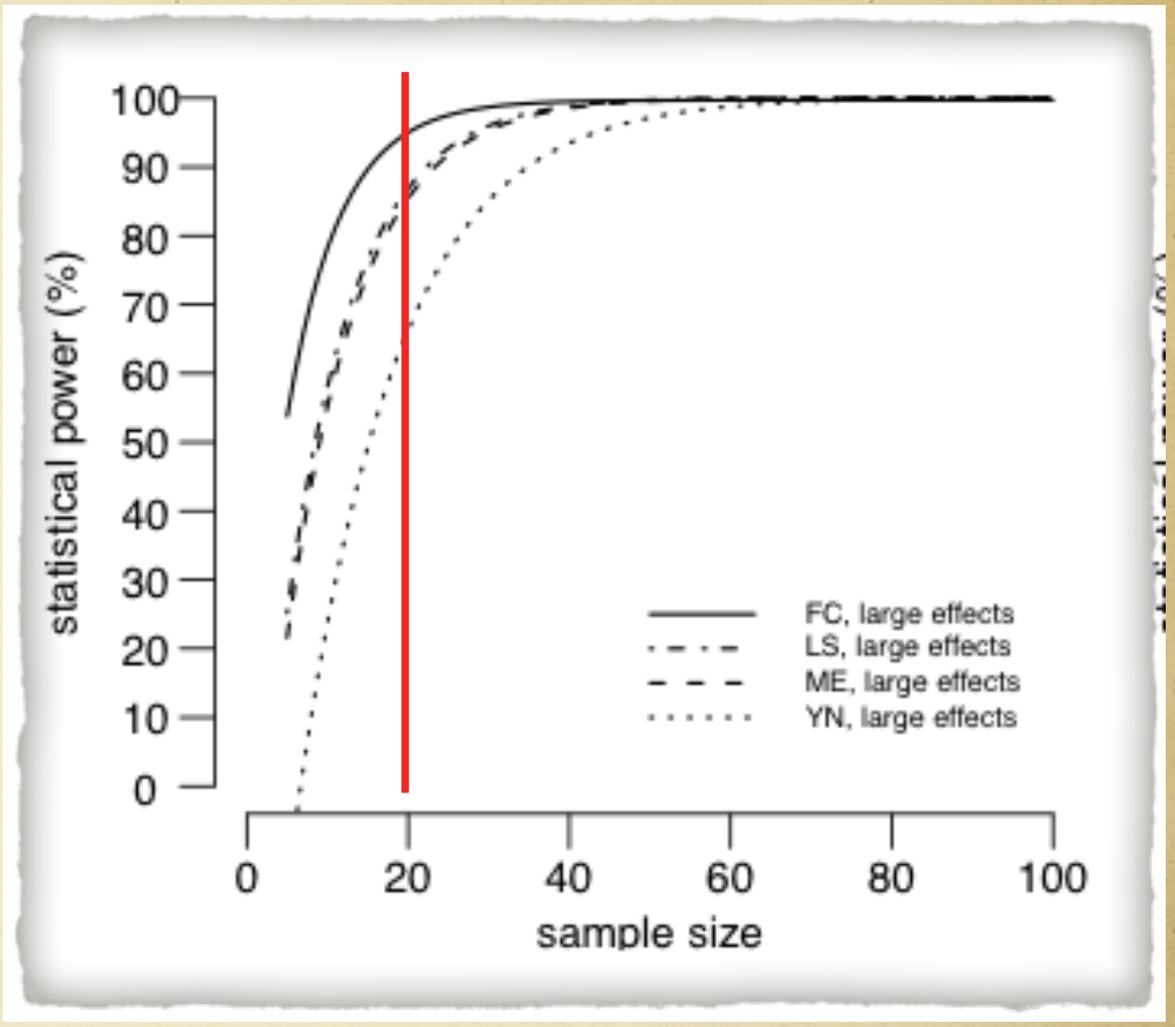


Some remaining issues I: Task effects

Several experimental formats

- A Lickert-scale experiment (LS)
- A yes/no judgment (YS)
- Magnitude estimation (ME)
- A head-to-head experiment (a.k.a. 2-alternative forced choice: 2AFC or simply, FC)





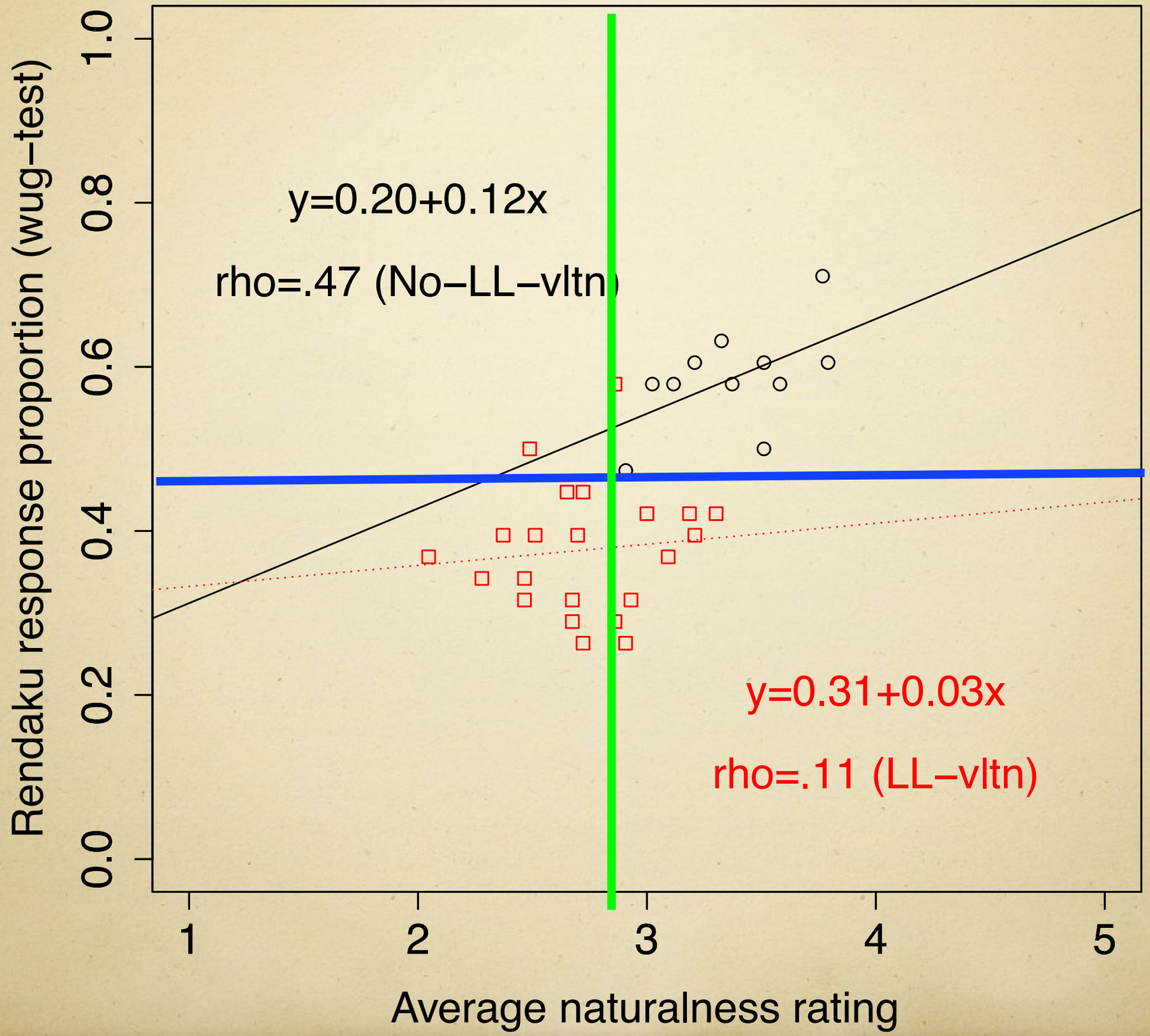
Statistical power: How likely would we find a statistically significant difference.

Based on Sprouse & Almeida 2012

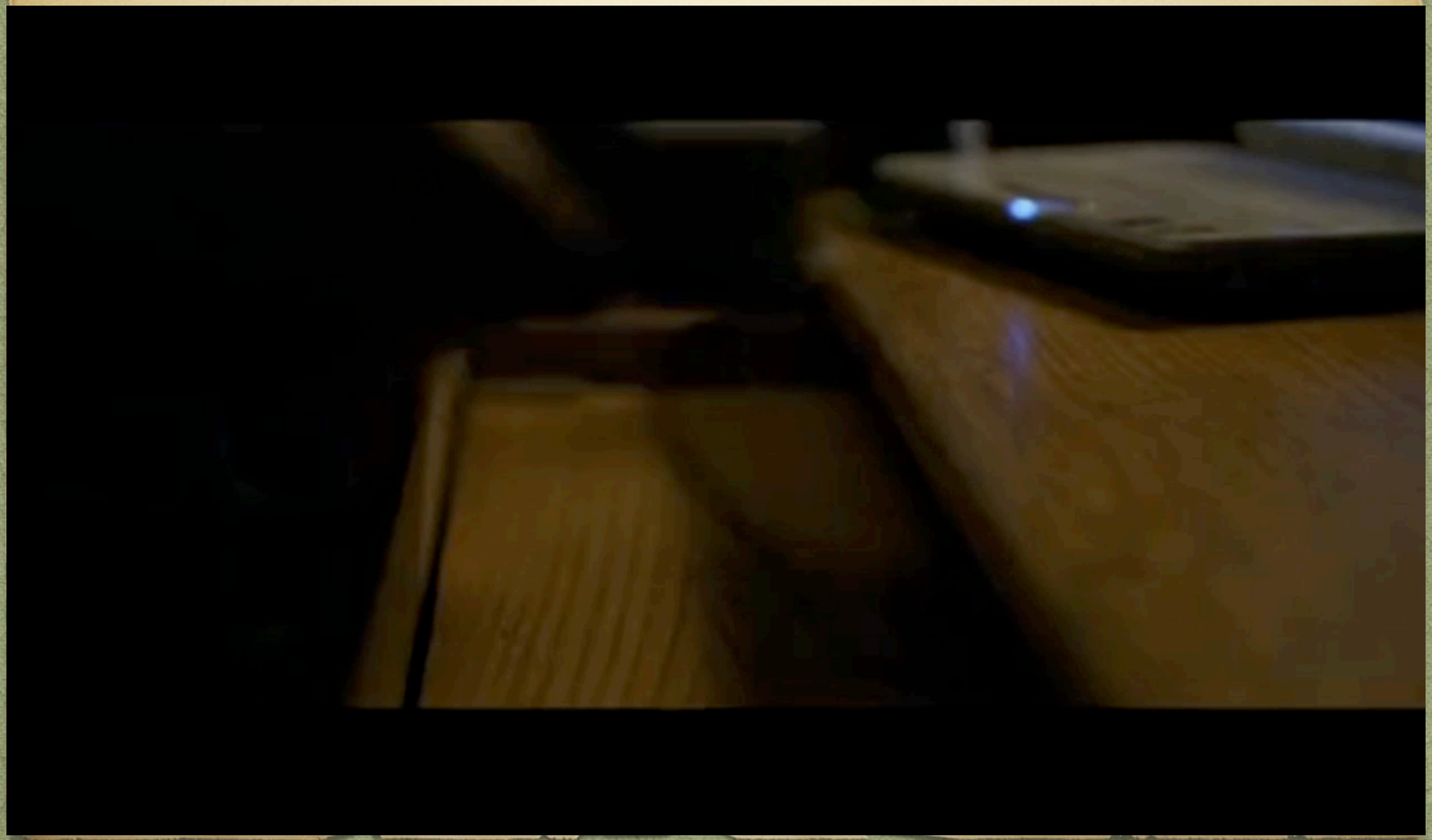
Kawahara (2015) on Lyman's Law

- Rendaku is blocked when there is another voiced obstruent already.
- [takara+kuji], [oo+tokage]
- A LS experiment (Kawahara 2012) vs. a FC experiment (Kawahara and Sano 2014)

Separated by LL–violation



Mark Zuckerberg knew…



Some remaining issues II: The issue of gradience

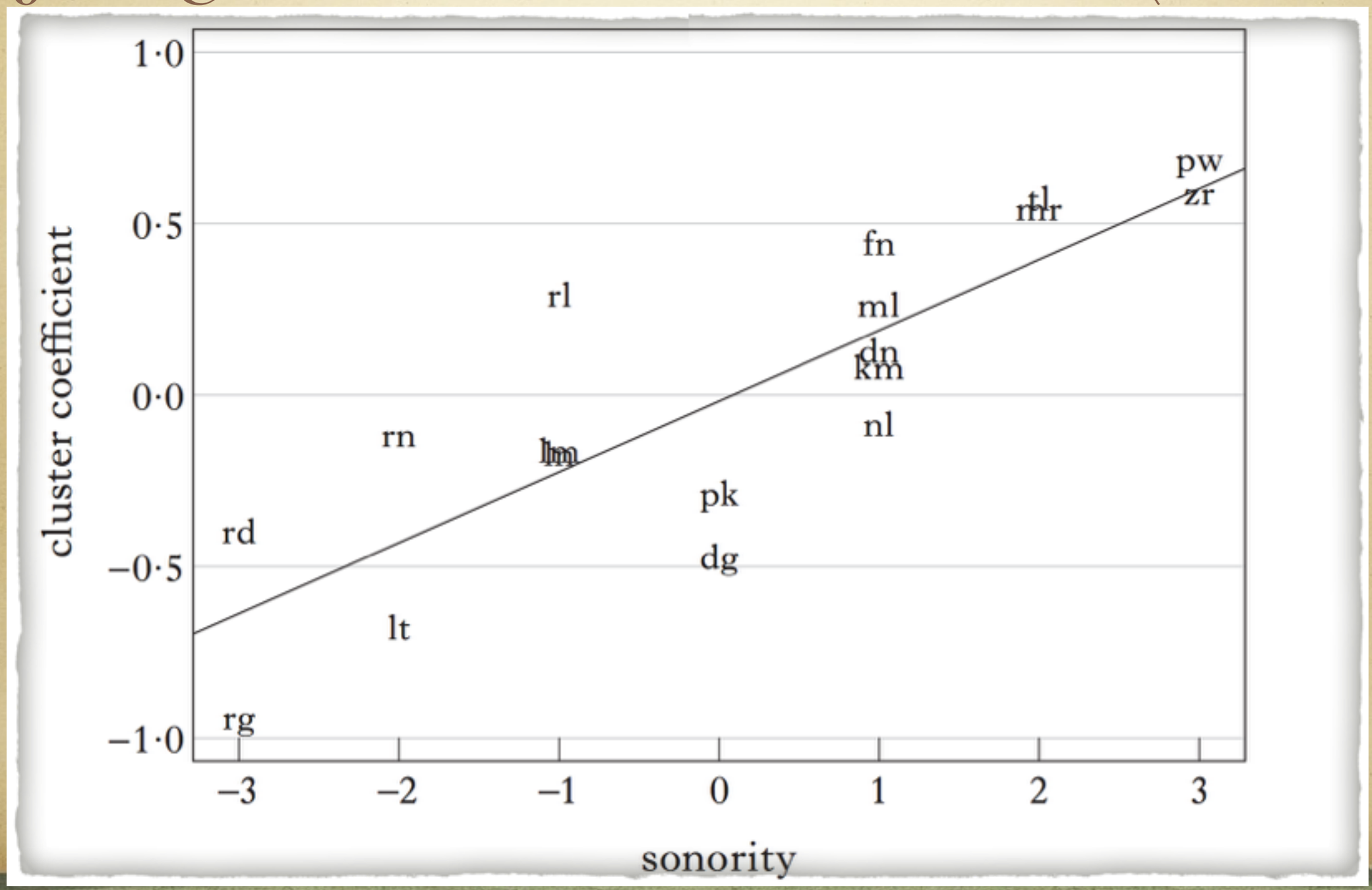
Grammatical-ungrammatical dichotomy

- Every sentence is either generatable or non-generatable by the grammar (Chomsky 1957).

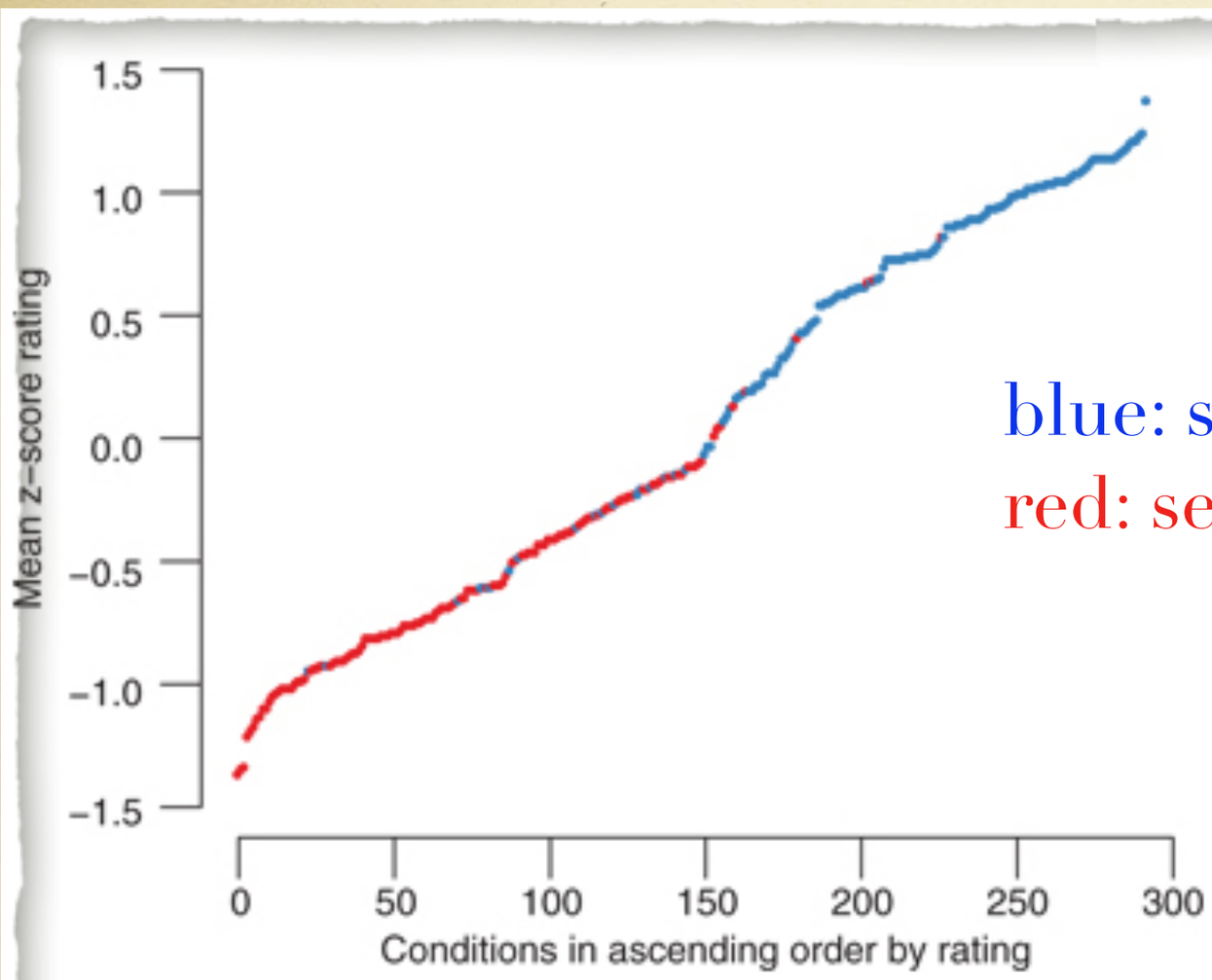
- Every word is either permitted by grammar or not: *blick*, *brick* vs. *bnick* (Halle 1973).

- There should be a dichotomy between grammatical structures and ungrammatical structures (both in phonology and syntax).

Gradient phonotactic judgment: Daland et al. (2011)



Gradient sentence judgment: Sprouse (2015)



blue: sentences w/o diacritic

red: sentences w/ diacritic

Figure 5

Acceptability for 300 sentence types randomly sampled from *Linguistic Inquiry 2001–2010* plotted in ascending order (Sprouse et al. 2013). Red dots indicate sentence types that were given a diacritic (*,?, or a combination) in LI. Blue dots indicate sentence types that were given no diacritic

A way out?

- Grammar is dichotomous; it is our cognitive system (a.k.a. performance) that "gives a gradience flavor" to any concept.
- In fact, mathematically dichotomous concepts can yield gradient judgments:
  - Odd number vs. even number: 7952 vs. 8
  - Female vs. male: nurse, mum, police woman, cow girl
  - Existing vs. non-existing: unicorn vs. round triangle

Table 1. *Categories, category exemplars, and exemplariness ratings for prototype and well-defined categories*

Prototype categories				Well-defined categories			
fruit				even number			
apple	1.3	orange	1.1	4	1.1	2	1.0
strawberry	2.1	cherry	1.7	8	1.5	6	1.7
plum	2.5	watermelon	2.9	10	1.7	42	2.6
pineapple	2.7	apricot	3.0	18	2.6	1000	2.8
fig	5.2	coconut	4.8	34	3.4	34	3.1
olive	6.4	olive	6.5	106	3.9	806	3.9
sport				odd number			
football	1.4	baseball	1.2	3	1.6	7	1.4
hockey	1.8	soccer	1.6	7	1.9	11	1.7
gymnastics	2.8	fencing	3.5	23	2.4	13	1.8
wrestling	3.1	sailing	3.8	57	2.6	9	1.9
archery	4.8	bowling	4.4	501	3.5	57	3.4
weight-lifting	5.1	hiking	4.6	447	3.7	91	3.7
vegetable				female			
carrot	1.5	peas	1.7	mother	1.7	sister	1.8
celery	2.6	spinach	1.7	housewife	2.4	ballerina	2.0
asparagus	2.7	cabbage	2.7	princess	3.0	actress	2.1
onion	3.6	radish	3.1	waitress	3.2	hostess	2.7
pickle	4.8	peppers	3.2	policewoman	3.9	chairwoman	3.4
parsley	5.0	pumpkin	5.5	comedienne	4.5	cowgirl	4.5
vehicle				plane geometry figure			
car	1.0	bus	1.8	square	1.3	square	1.5
boat	3.3	motorcycle	2.2	triangle	1.5	triangle	1.4
scooter	4.5	tractor	3.7	rectangle	1.9	rectangle	1.6
tricycle	4.7	wagon	4.2	circle	2.1	circle	1.3
horse	5.2	sled	5.2	trapezoid	3.1	trapezoid	2.9
skis	5.6	elevator	6.2	ellipse	3.4	ellipse	3.5

From

Armstrong et al. (1983)

- Even mathematically dichotomous distinctions can be perceived with gradience.
- But what's the evidence to assume that grammaticality is dichotomous? (Sprouse 2007)
- Or should we admit that grammatical judgments are gradient? How should the grammatical theory look like then? e.g. a MaxEnt grammar (Hayes & Wilson 2008).
- Recall that this issue is relevant to phonological theories as well (very much so!).

REPRINT, cite as Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49 (4)*, 243-256.

Standards for Internet-Based Experimenting

Ulf-Dietrich Reips

Experimental and Developmental Psychology, University of Zürich, Switzerland

Abstract. This article summarizes expertise gleaned from the first years of Internet-based experimental research and presents recommendations on: (1) ideal circumstances for conducting a study on the Internet; (2) what precautions have to be undertaken in Web experimental design; (3) which techniques have proven useful in Web experimenting; (4) which frequent errors and misconceptions need to be avoided; and (5) what should be reported. Procedures and solutions for typical challenges in Web experimenting are discussed. Topics covered include randomization, recruitment of samples, generalizability, dropout, experimental control, identity checks, multiple submissions, configuration errors, control of motivational confounding, and pre-testing. Several techniques are explained, including "warm-up," "high hurdle," password methods, "multiple site entry," randomization, and the use of incentives. The article concludes by proposing sixteen standards for Internet-based experimenting.

Key words: Internet-based experimenting, Web experiment, standards, experiment method, psychological experiment, online research, Internet research, Internet science, methodology

Psychological Research on the Net

Psychological Research On the net

Sponsored by the Hanover College Psychology Department

Below you will find links to known experiments on the internet that are psychologically related. They are organized by general topic area with the topic areas listed chronologically with the most recently added at the top. **If you wish to add a study please check this page.**

Maintained by John H. Krantz, Ph.D.

Recently Added Studies

(This list will hold studies added in the past ~30 days)

- In Sexuality: Virginity and Guilt Differences between Men and Women (**MEN ONLY**)(01/28/15)
- In Social: Does engaging in extra- curricular and pro-social activities improve your well being? (01/27/15)
- In Forensic: Offending Behaviour, Attitudes to Crime, and Personality Traits (01/27/15)
- In Forensic: Philosophy of punishment (01/27/15)
- In Health: An Examination of Trauma and Mental Health Outcomes (01/26/15)
- In Cognition: What are people afraid of, and why? (01/26/15)
- In I/O: Do You Behave Differently? Investigating Workplace Behaviors (01/26/15)
- In Mental Health: An investigation into the Relationship between Substance Use, Depression & Quality of Life (01/25/15)
- In Personality: Personality and Romantic Relationship Break-Ups (01/25/15)
- In Cyber: Psychological attributes relating to social networking use (01/23/15)
- In Personality: Personality & Attitudes About Bullying (01/23/15)
- In Personality: Guilt Proneness and Prosocial Behavior (01/23/15)
- In Cognition: Memory while playing online games (01/23/15)
- In Relationships: Social network effects on human pair choice (01/19/15)
- In Cognition: Memory judgments experiment (01/19/15)
- In Social: Differences in attitudes towards life among perfectionists and non-perfectionists: Looking for the perfectionist experience (01/14/15)
- In Social Cognition: Perceptions of healthcare professionals (01/14/15)
- In General: Investigating the relationship between attitudes towards animals and perceived level of mental capacity (01/14/15)
- In Relationships: Social Functions of Parasocial Relationships with Media Figures (01/14/15)
- In Relationships: Study for women: Environment and mate choice (01/13/15)
- In Mental Health: Stressful Life Events and Well-Being (01/13/15)
- In Mental Health: A study about depression, social relationships and recovery (01/12/15)
- In Personality: Personality and responses to videos (01/12/15)
- In Mental Health: Factors underlying the experience of social anxiety (01/09/15)
- In Relationships: Sex in Relationships (01/08/15)
- In Health: Exploring the links between 'distressed' personality type and subjective health complaints (01/08/15)
- In Relationships: Personality and Experiences in Relationships (01/08/15)
- In Relationships: Perceptions of Relationships (01/07/15)
- In Social: Experiences of Research Participation (01/07/15)
- In Sexuality: Labiaplasty: Attitudes and Predictors behind the Female Cosmetic Surgery (01/07/15)
- In Social: The effect of prior contact and perceived status on perceptions of contact (British UK-nationals)(01/07/15)
- In Relationships: Success and Personality (1/05/15)
- In Cross-Cultural: Examination of cross cultural differences in psychosocial functions of autobiographical memory (12/30/14)

Some concluding remarks

- I believe that the most constructive way to respond to the "anti-introspection-data criticism" is to show, quantitatively, that the introspection-data (a.k.a. "our method") are reliable.
- Maybe more can be done for Japanese?
- Experiments are not that hard these days.

Recommended readings

Criticisms and responses:

- Gibson, E. & E. Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14: 233–234.

- Marantz, A. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22: 429–445.

Overview papers:

- Myers, J. 2009. Syntactic judgment experiments. *Language and Linguistics Compass* 3. 406–423.

- Sprouse, J. 2015. Three open questions in experimental syntax. *Linguistic Vanguard.*

Two very famous monographs:

- Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

- Schütze, C. T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.