Disentangling syntax and semantics in acceptability and intelligibility ratings

Jana Häussler (University of Leipzig) & Tom Juzek (Nuance Communications)

Disentangling syntax and semantics in acceptability and interpretability ratings

& plausibility

Jana Häussler (University of Leipzig) & Tom Juzek (Nuance Communications)

Background

1) Acceptability vs. Grammaticality

2) Our own study on gradience

Acceptability: An expression of grammaticality plus extra-grammatical factors

Simple examples for factors:

→ length

→ complexity

→ …

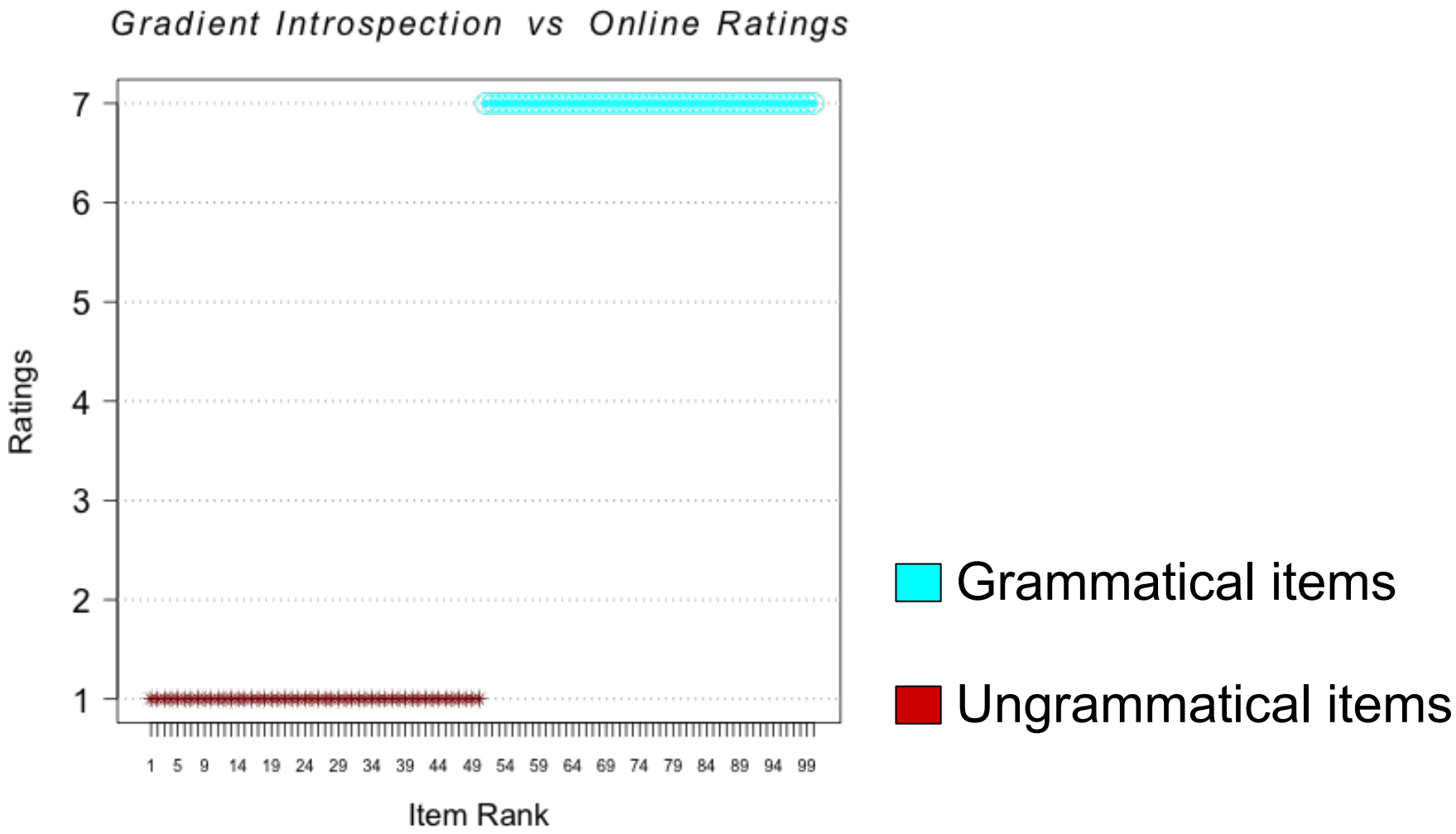
The relationship between A and G is not entirely clear.

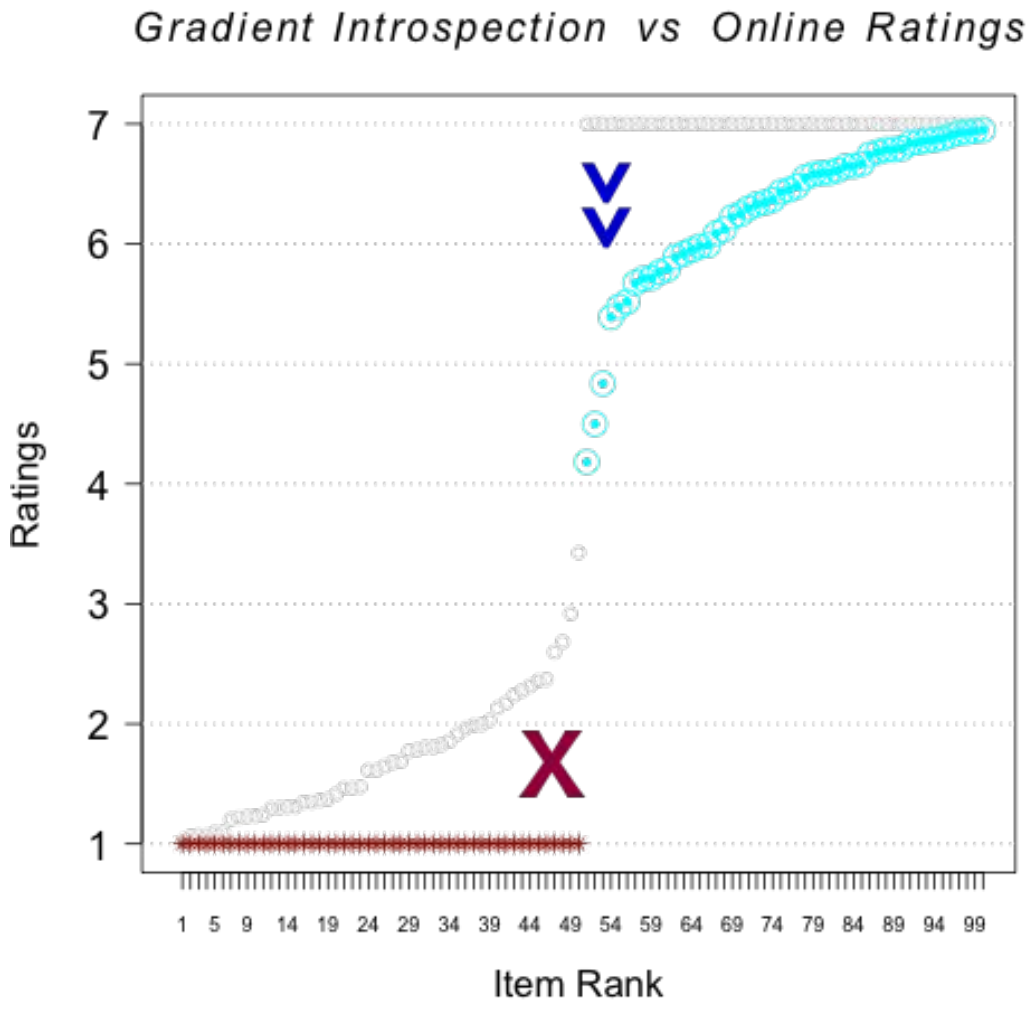
Factors affect acceptability in both directions:

Grammatical items degrade: center embedding, garden-path sentences, “colorless green ideas”, etc.

Ungrammatical items are ameliorated: grammatical illusions like missing-vp effect, etc.

Maybe another factor: intelligibility (“semantic goodwill”, Katz’s “semi-sentences”)

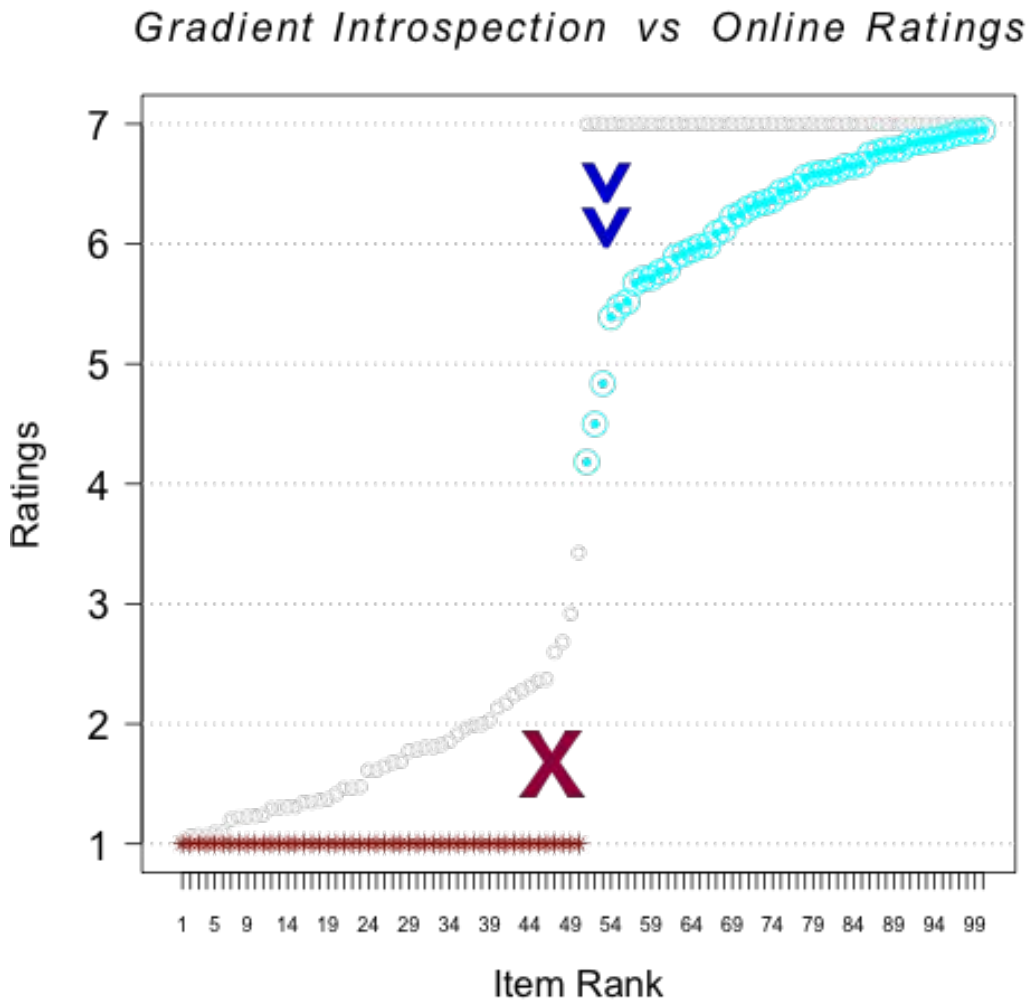

Gradient Introspection vs Online Ratings
Ratings
7
6
5
4
3
2
1
1 5 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99
Item Rank
Grammatical items
Ungrammatical items



Commonly observed: Penalty for grammatical items

Not as much observed /underresearched: Bonus for ungrammatical items (but: gramm. illusions)

Acceptability



“Colorless green ideas”

“Man bit dog” (Semi-sentence, Katz 1964).

Acceptability

Underresearched:

So far, there's little experimental research on the semantic impact, à la Katz's "semi-sentences" or sentences like Chomsky's "colorless green ideas".

Starting point for this project:

Our work on data reliability and data structure, in which we observed issues with reliability (→ SSA 2013) and were puzzled by the amount of observed gradience and its *quality*.

Gradience in acceptability has been acknowledged for quite some time, at least since Chomsky (1965). Featherston (2005) provided a nice illustration of gradience in acceptability.

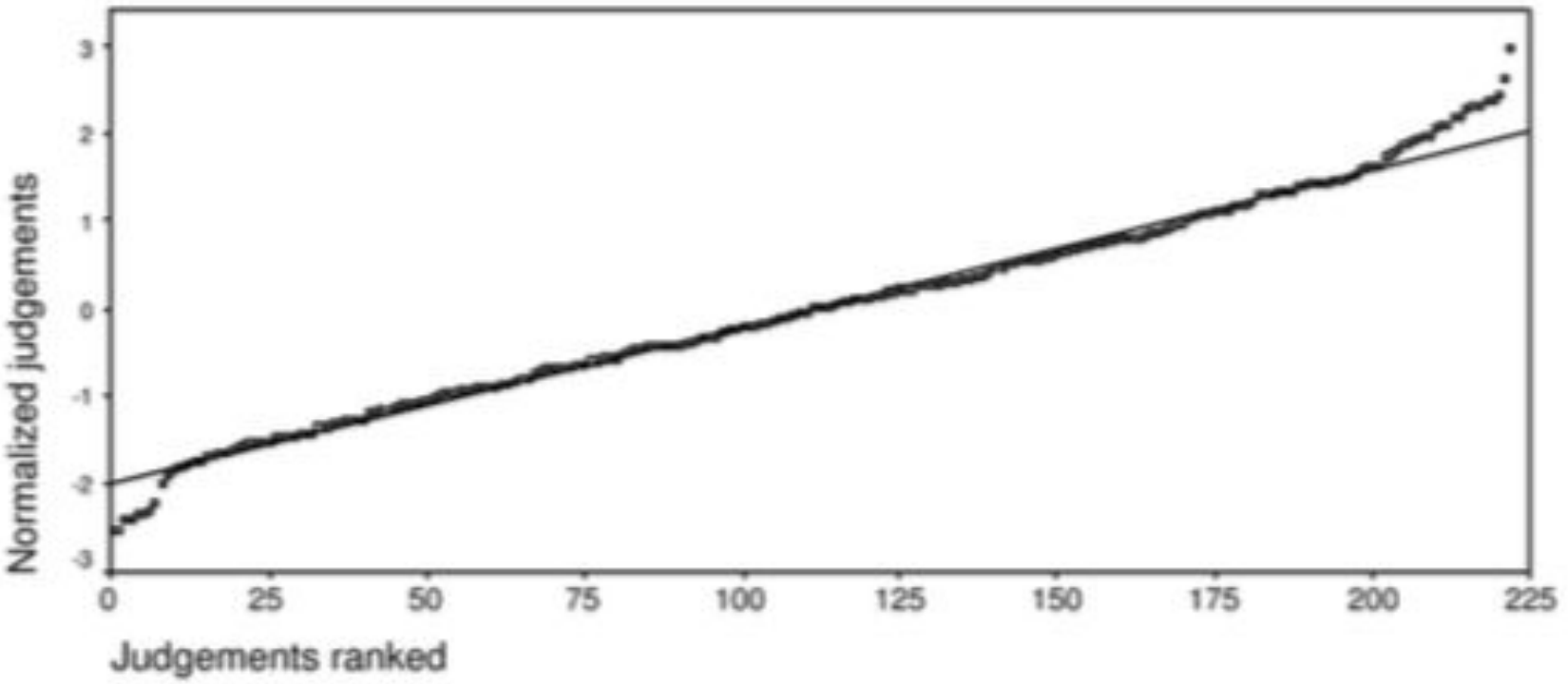


Figure 1: Judgements elicited under controlled conditions produce a linear pattern of well- formedness.

Featherston (2005)

Our previous study

CLS 2015, LE 2016

Our study compared:

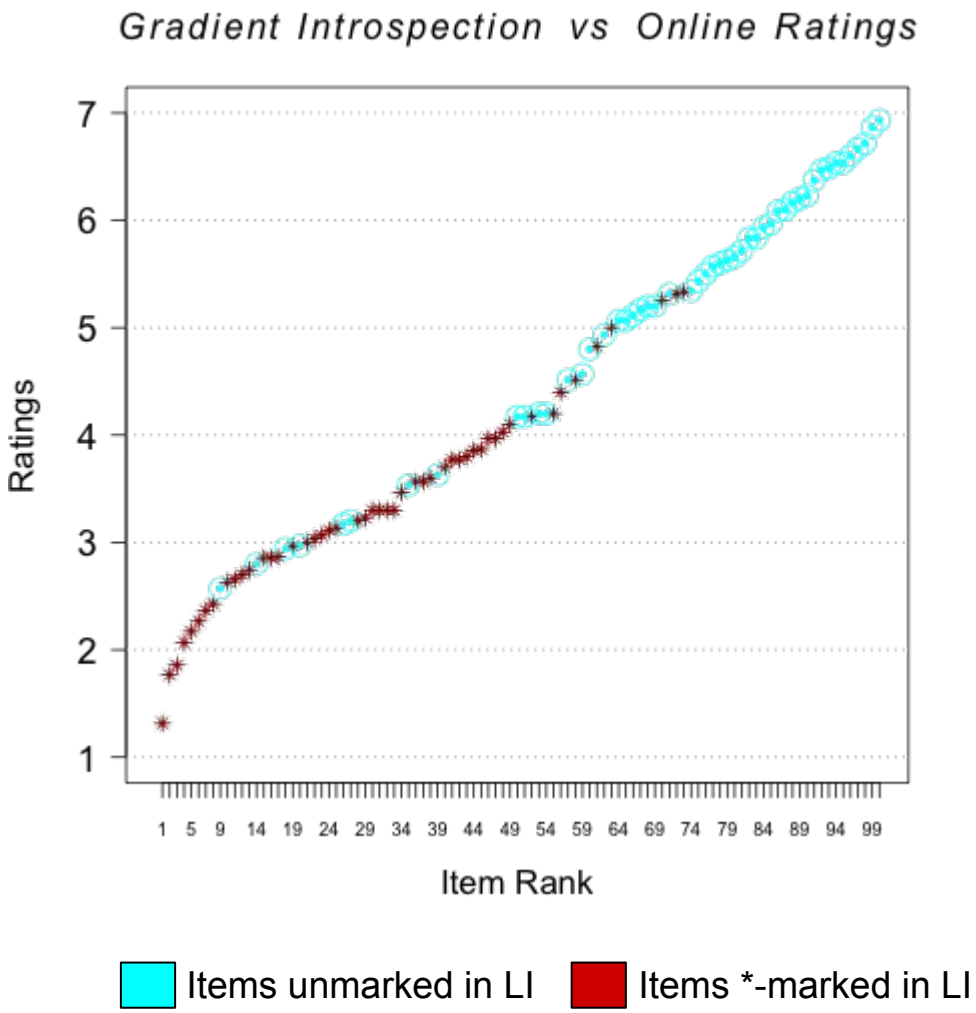
- author judgments taken from papers in Linguistic Inquiry

and

- experimental ratings

We discussed issues of:

- reliability
- and gradience



Our previous study

CLS 2015, LE 2016

Our study compared:

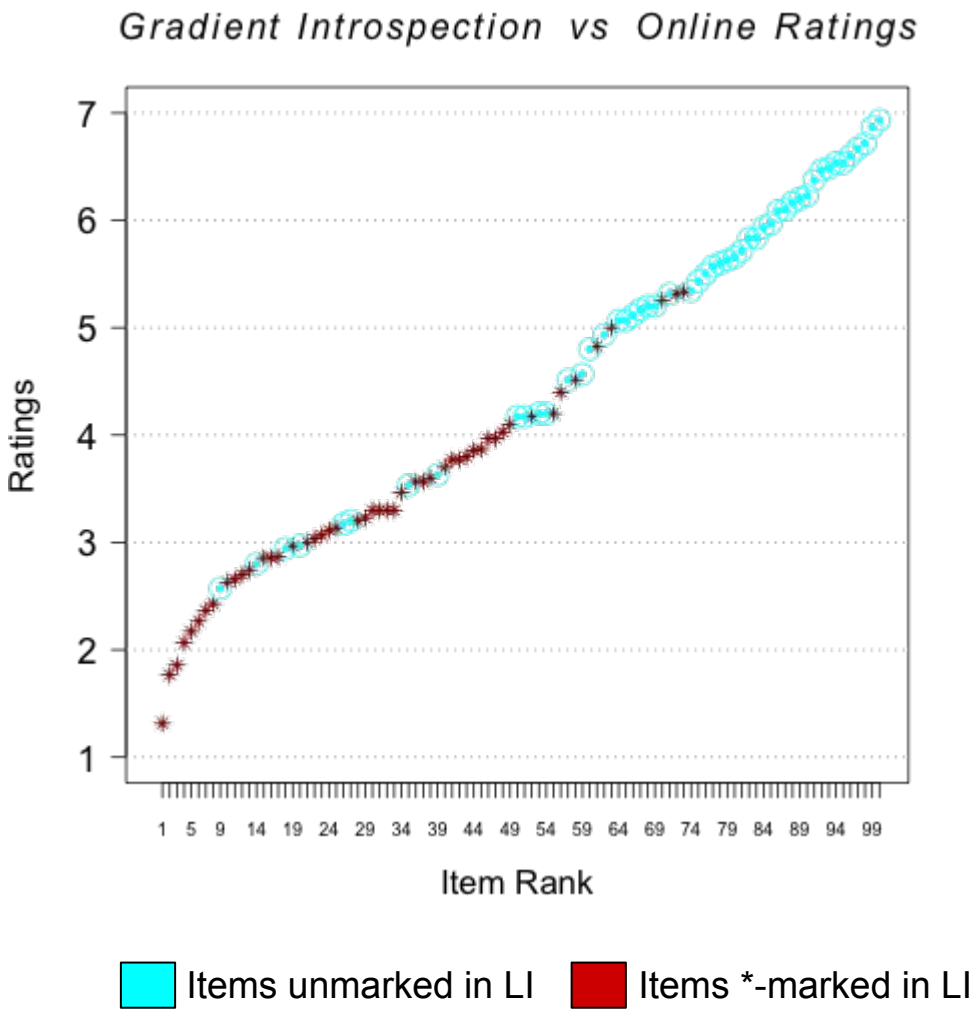
- author judgments taken from papers in Linguistic Inquiry

and

- experimental ratings

We discussed issues of:

- **reliability**
- and gradience



Our previous study

CLS 2015, LE 2016

Our study compared:

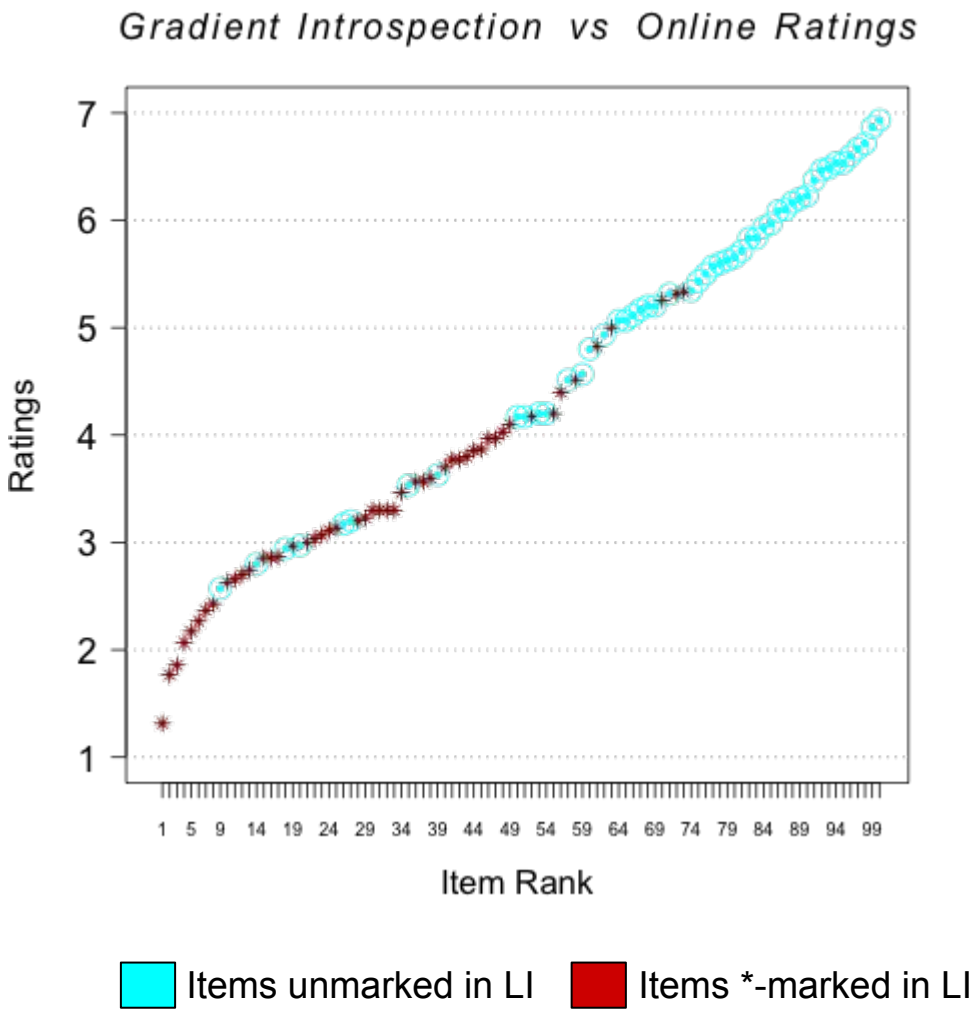
- author judgments taken from papers in Linguistic Inquiry

and

- experimental ratings

We discussed issues of:

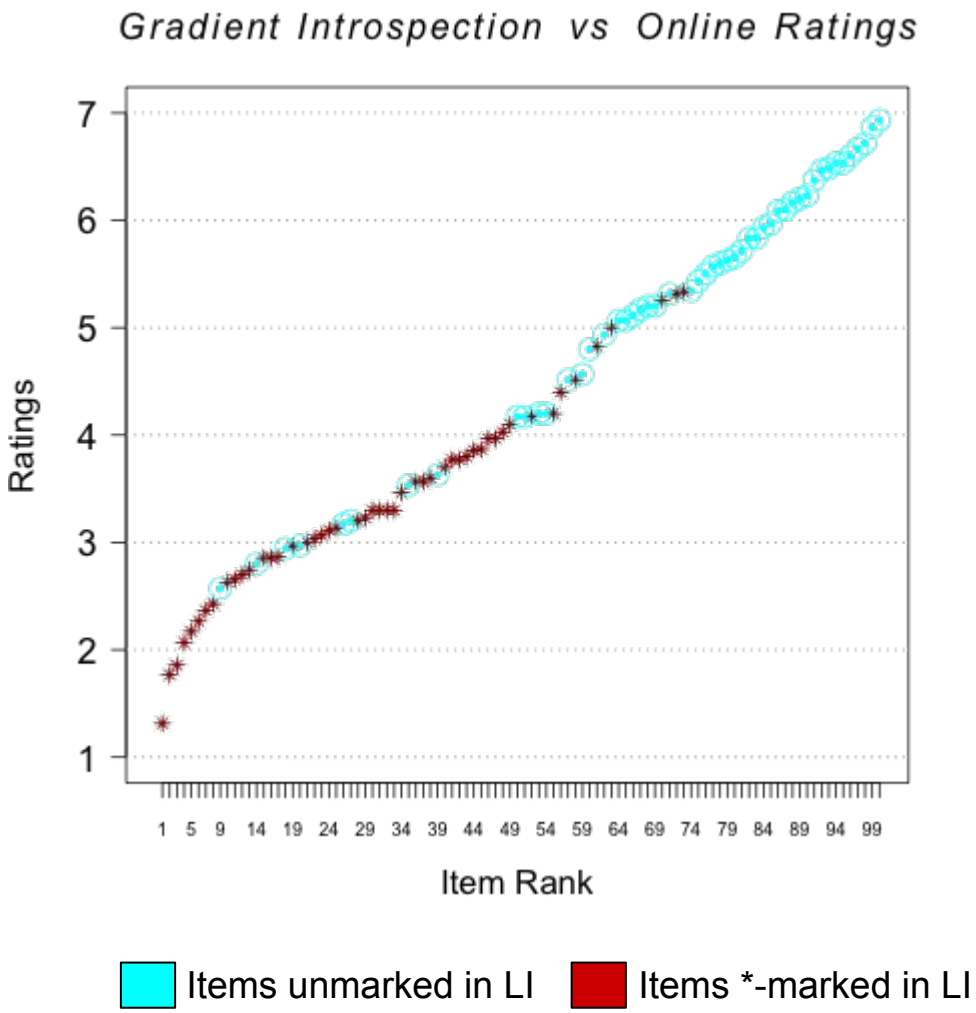
- reliability
- and **gradience**



Our previous study

Puzzle:

Why are so many
***-items** in the mid-bin?

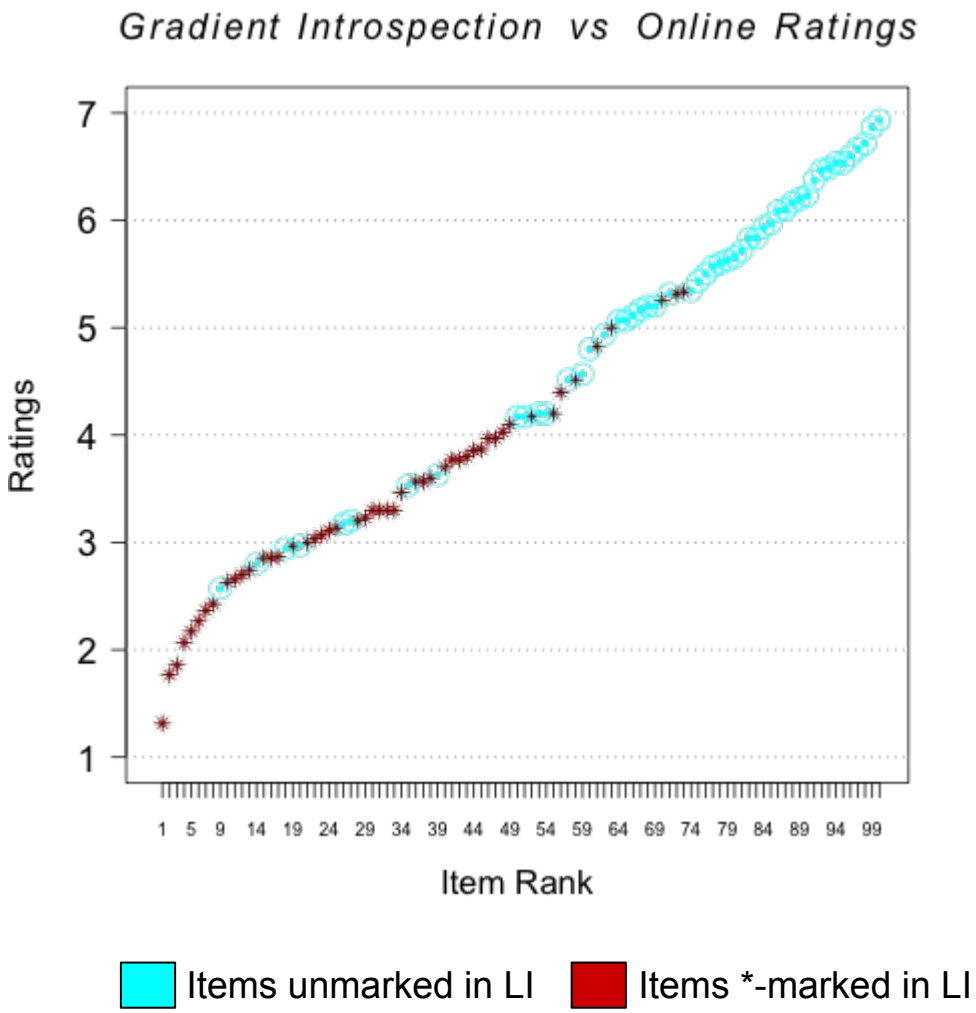


Our previous study

Puzzle:

Why are so many ***-items** in the mid-bin?

They include none of the known grammatical illusions.



Our previous study

Puzzle:

Why are so many
***-items** in the mid-bin?

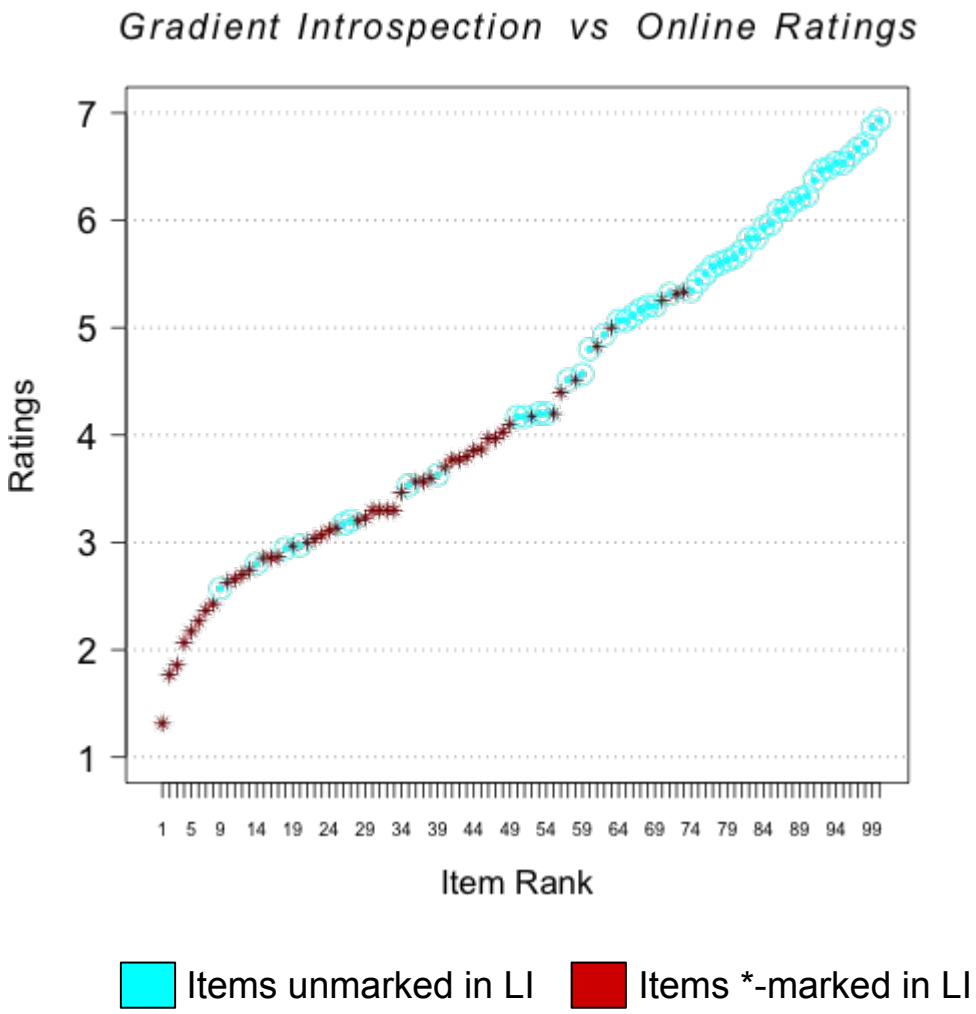
MO: Look at factors that could have caused this.

→ Not due to aggregation

→ Some scale effects

→ Processing effects unlikely

→ ???

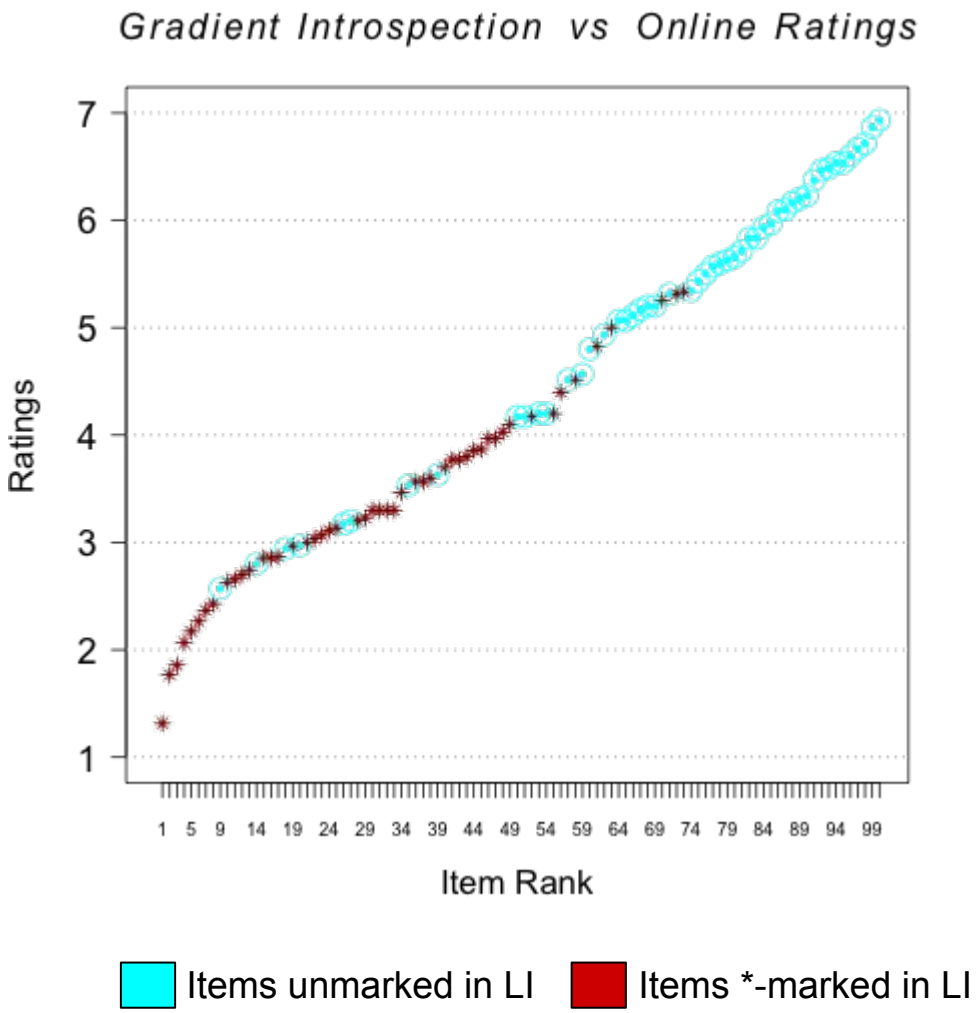


Our previous study

Puzzle:

Why are so many ***-items** in the mid-bin?

Does intelligibility drive the amelioration?



What follows is to some degree explorative.

→ Adjustments across the experiments

→ Consequences for statistical analyses

Experiment 1

Comparing syntactic and semantic ratings

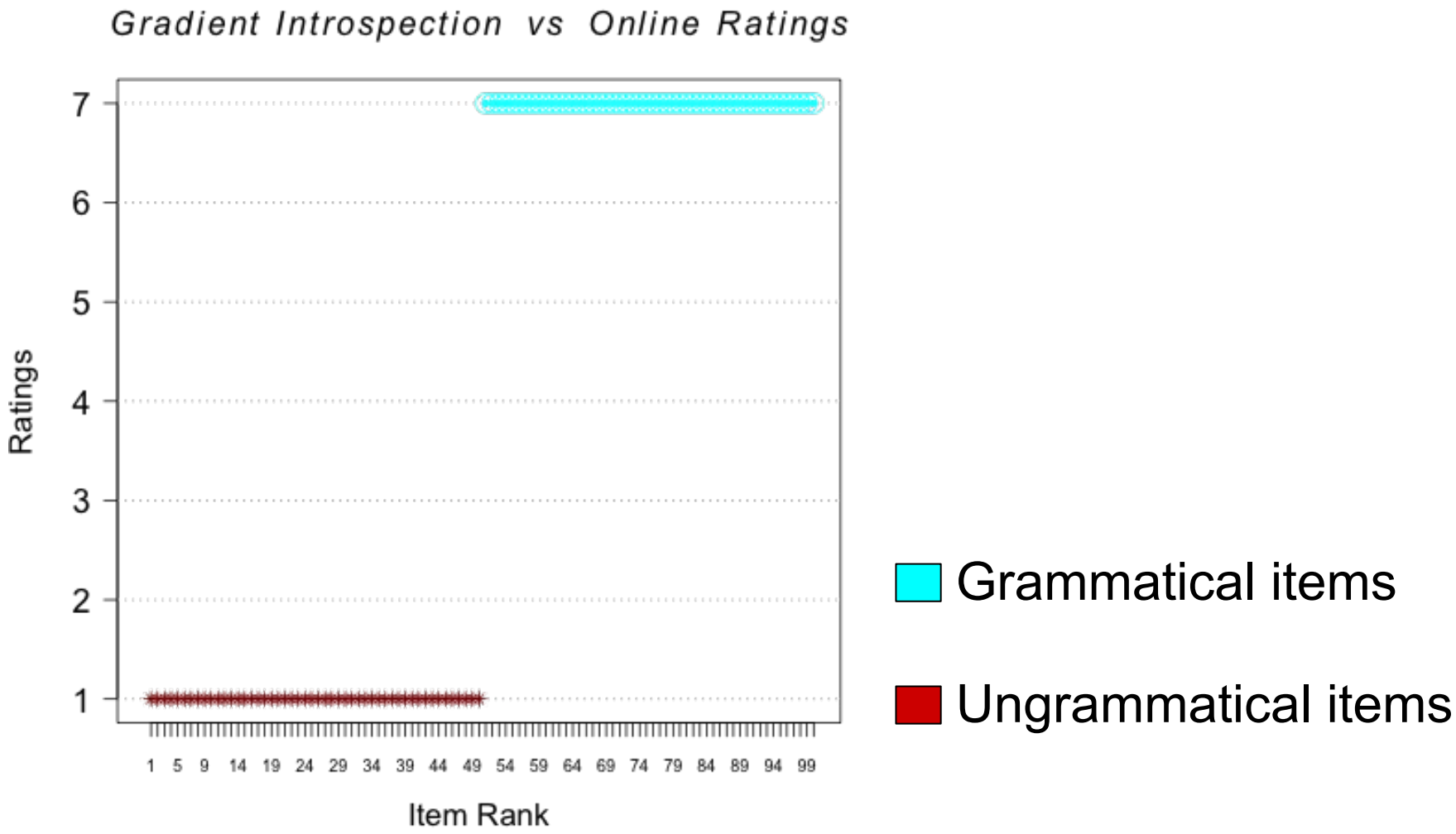
Experiment 1

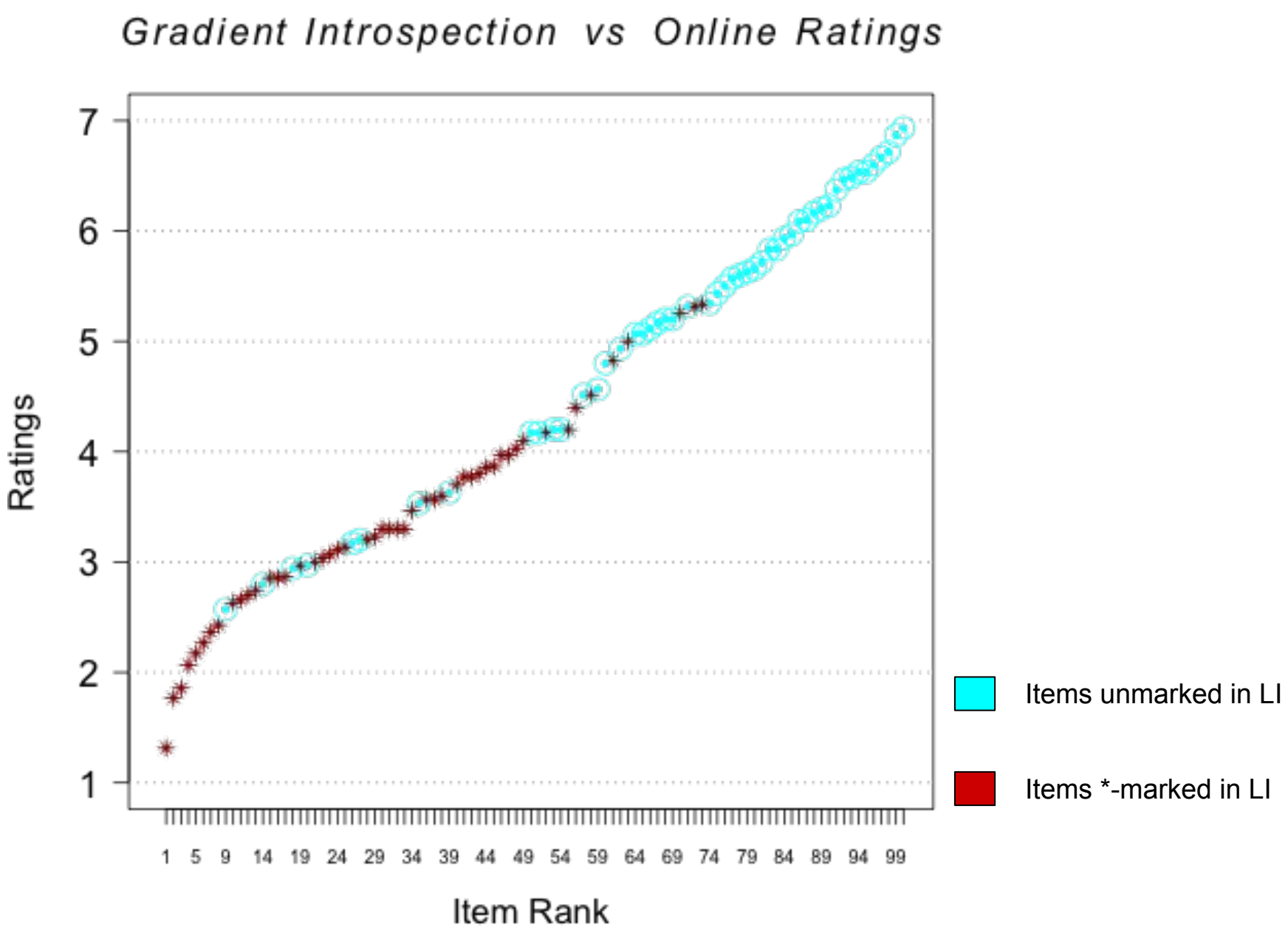
Experiment 1a: **acceptability** ratings

Experiment 1b: **intelligibility** ratings

Prediction:

If intelligibility increases syntactic acceptability, then intelligibility ratings for *-items with intermediate syntactic acceptability should higher than for *-items with low syntactic acceptability.


Gradient Introspection vs Online Ratings
Ratings
7
6
5
4
3
2
1
1 5 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99
Item Rank
Grammatical items
Ungrammatical items


Gradient Introspection vs Online Ratings
Ratings
7
6
5
4
3
2
1
1 5 9 14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 89 94 99
Item Rank
Items unmarked in LI
Items *-marked in LI

Experiment 1

Participants

- recruited via Prolific (https://www.prolific.ac/)

- exclusion criteria:
  - native language other than British English
  - incomplete results
  - failure on control items
  - response times indicating fast clicking through

No. of participants: 25 (Exp 1a) + 23 (Exp 1b)

Experiment 1

Materials:

100 sentences randomly selected from our *LI* corpus
(50 *-items + 50 OK-items)

(all of which were used in the previous study)

→ https://goo.gl/4xujyK

Experiment 1

Materials:

01) The teacher is Jenny.
02) Which man did you persuade to read which book?
03) John proved that Mary is sick.
etc.

51) *John offered Susan to leave.
52) *It seems to Naomi to have solved the problem.
53) *John appears to hit Bill right now.
etc.

LI Corpus

A few words about our item extraction.

LI Corpus

We extracted example sentences from articles in *Linguistic Inquiry* (2001-2010) and categorized them according to judgement type (cf. Sprouse et al. 2013).

Unlike Sprouse et al.(2013):

- at least one author is a native speaker of US-English
- no restriction to syntactic papers

→ 160 articles in total
→ 4334 judgments
→ of which ~2600 are standard acceptability judgments

LI Corpus

For standard acceptability judgments, we created four sub-corpora based on two criteria:

- Number of levels the author distinguishes in the paper
  - “binary papers” (only 2 levels: * and unmarked)
  - “gradient papers” (more than 2 levels)

- Marking of the sentence itself
  - “*-items” (items marked with a *)
  - “OK-items” (items left unmarked in the paper)

  (We ignored items with markings like ?, ??, ?* etc.)

Experiment 1

Procedure

- Part A: **syntactic** acceptability judgements
- Part B: **semantic** acceptability judgements

- 7-point scale
- online (separate website)
- warning mechanism (extreme RTs & “gotcha” items)

Experiment 1a Syntactic Rating

From the instructions:

*Important: Please do not be bothered with spelling differences between American English and British English or with punctuation. And crucially, please do not be bothered with meaning. For example, while "*Jack did his job goodly*" is meaningful and intelligible, it is also not fully grammatical. Thus, it should receive a low rating. On the other hand, "*Colourless green ideas sleep furiously*" is not really meaningful but fully grammatical. Thus, it should receive a high rating.*

Experiment 1b Semantic Rating

From the instructions:

*Important: Please do not be bothered with spelling differences between American English and British English or with punctuation. And crucially, please do not be bothered with grammaticality. For example, while "*Jack did his job goodly*" is ungrammatical, it is also meaningful and intelligible. Thus, it should receive a high rating. On the other hand, "*Colourless green ideas sleep furiously*" is fully grammatical but not meaningful/intelligible. Thus, it should receive a low rating.*

Experiment 1a/b

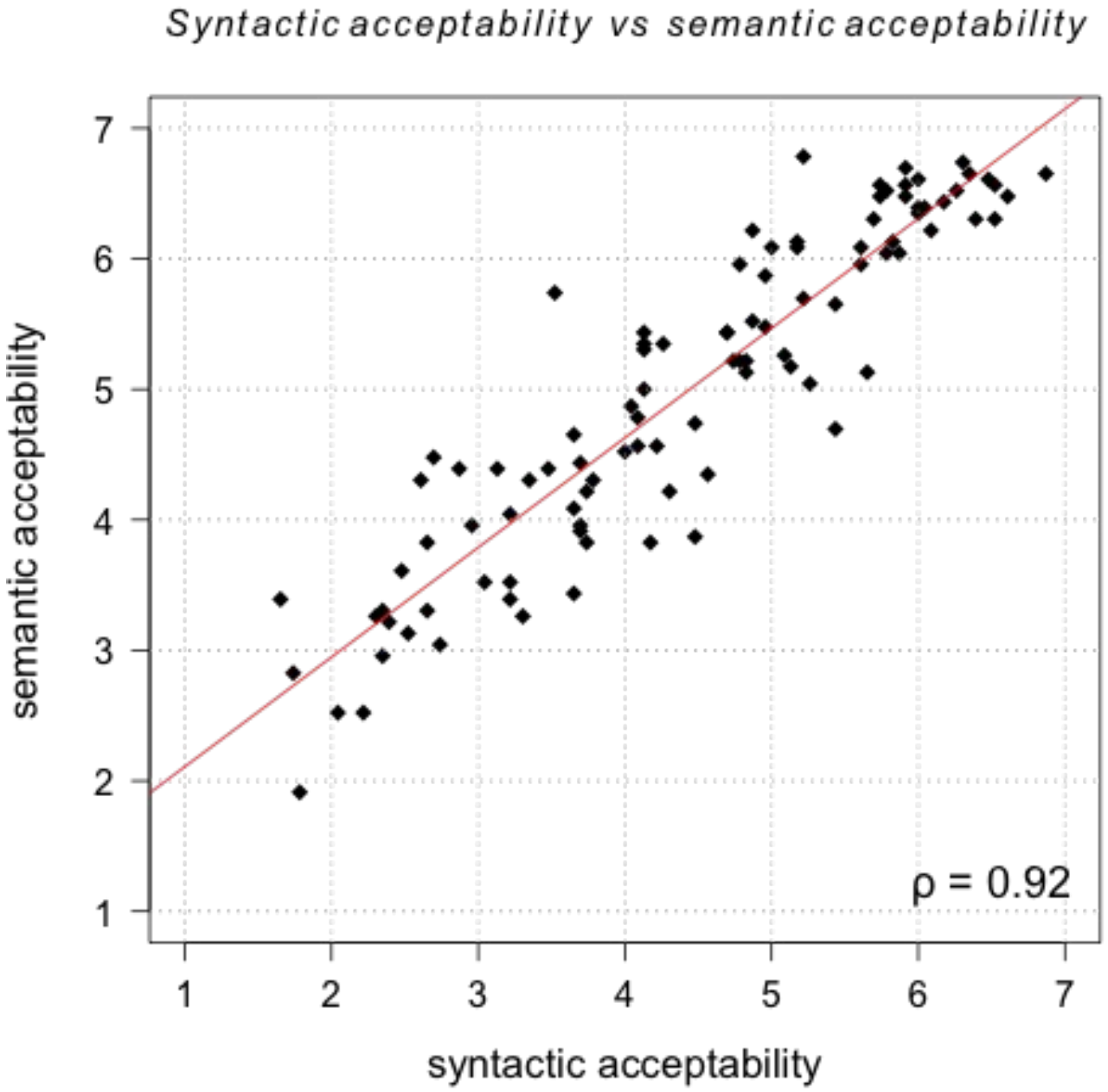
Syntactic/semantic interface

1/108

John was arrested.

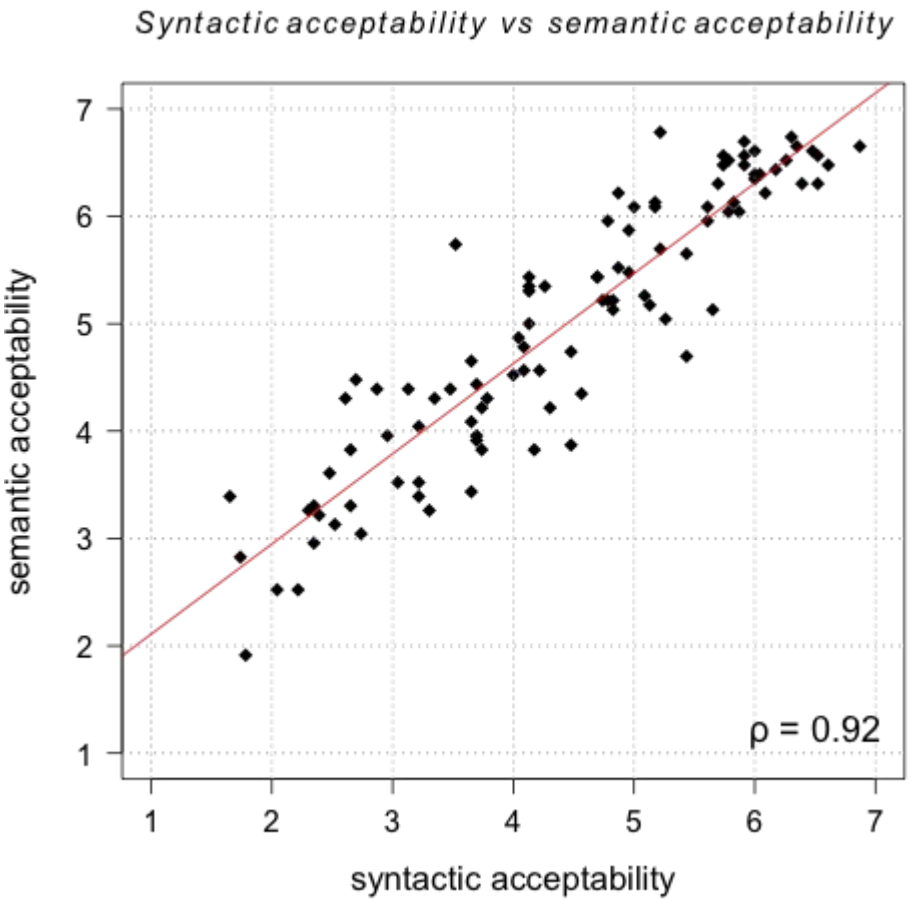
Experiment 1

Results


Syntactic acceptability vs semantic acceptability
semantic acceptability
syntactic acceptability
ρ = 0.92
1
2
3
4
5
6
7

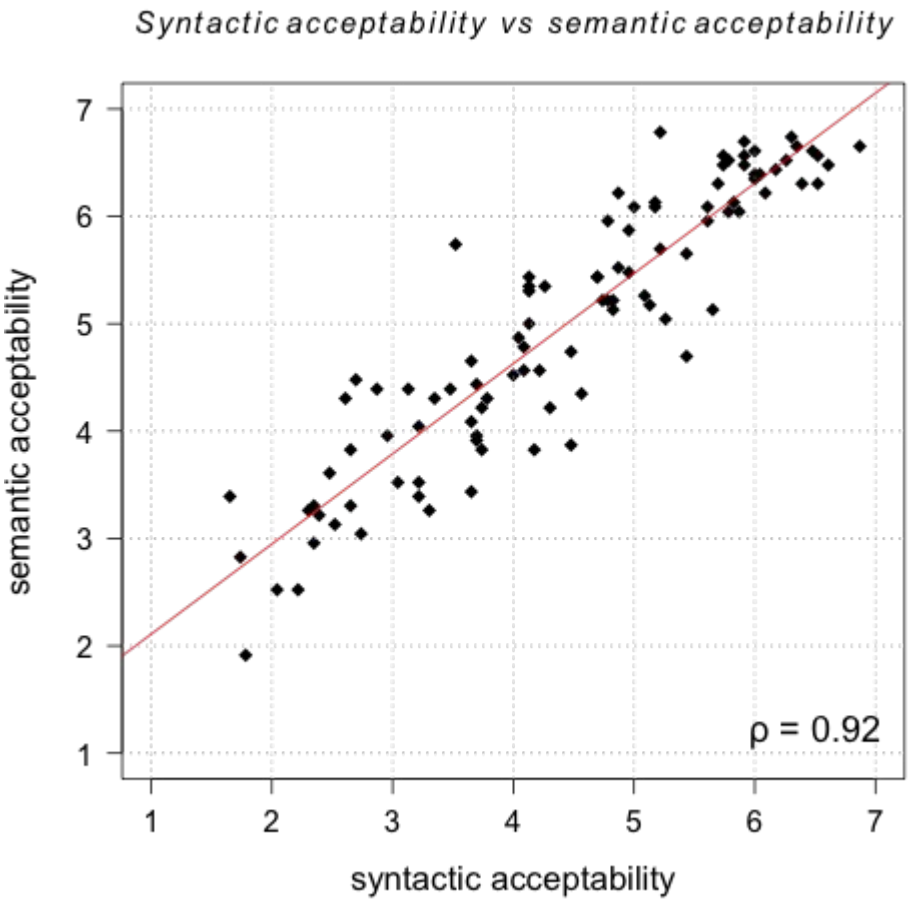
Experiment 1 Results

Syntactic and semantic ratings correlate strongly.

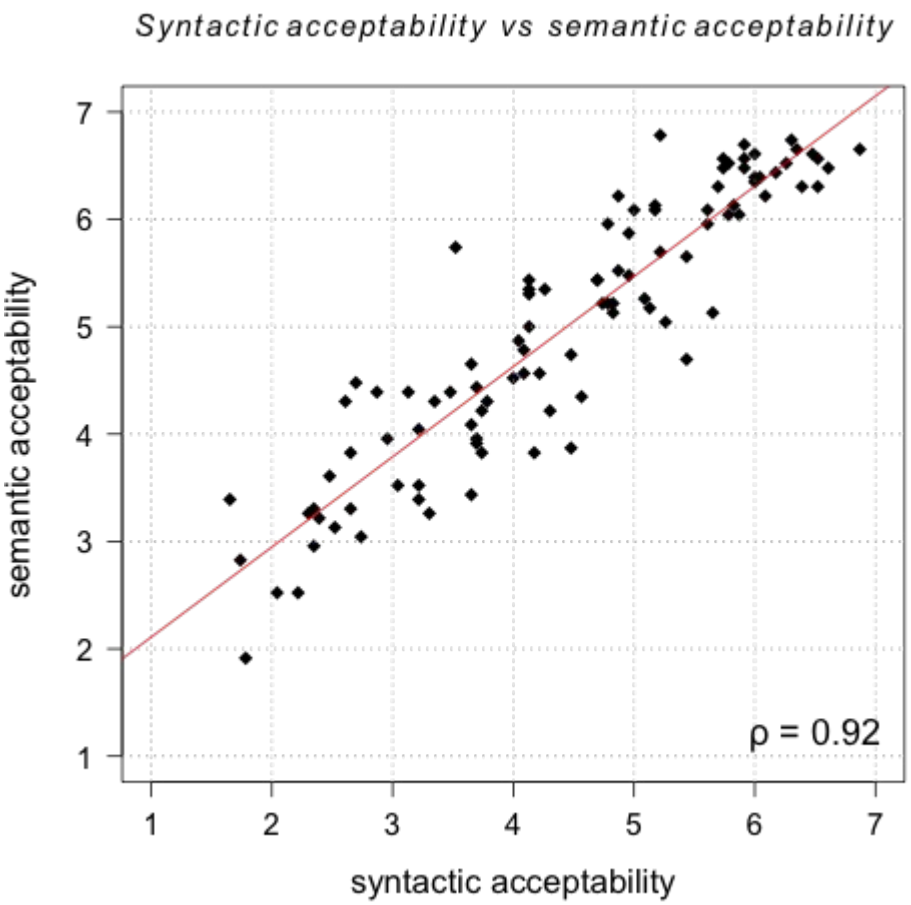


Experiment 1 Results

The middle is populated to a good degree.



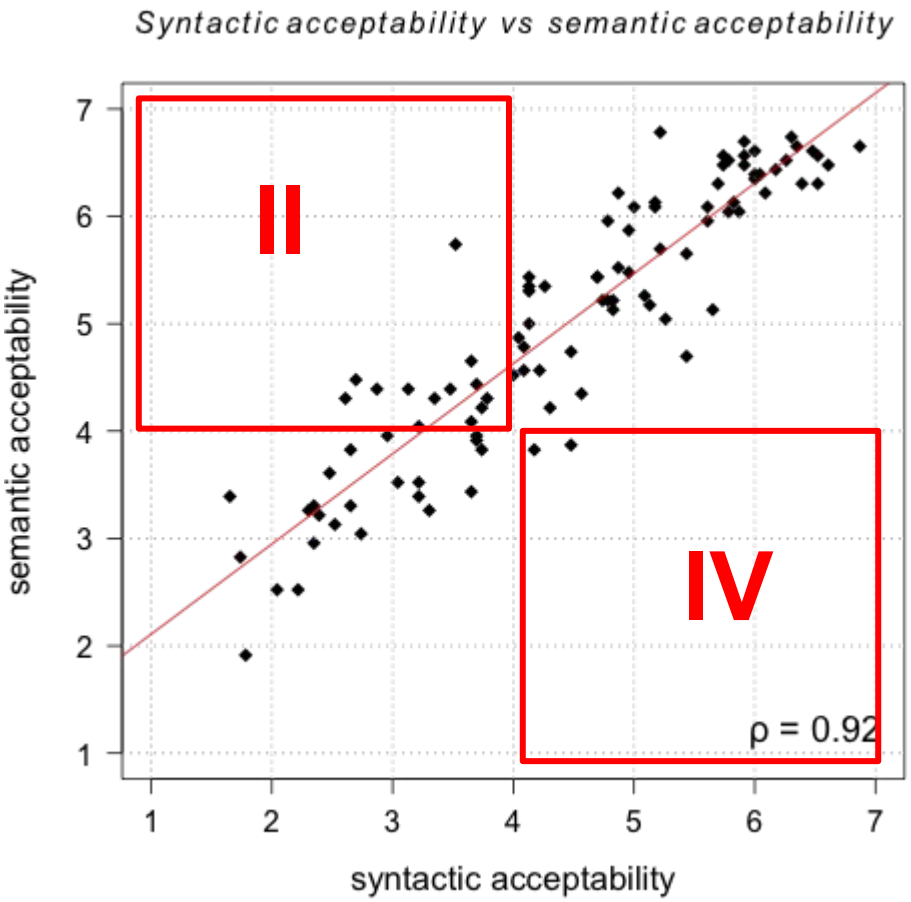
Should we conclude a bi-directional impact?



Should we conclude a bi-directional impact?

If so, more data points should occur 2nd and 4th quadrant.

Ungrammatical but intelligible items

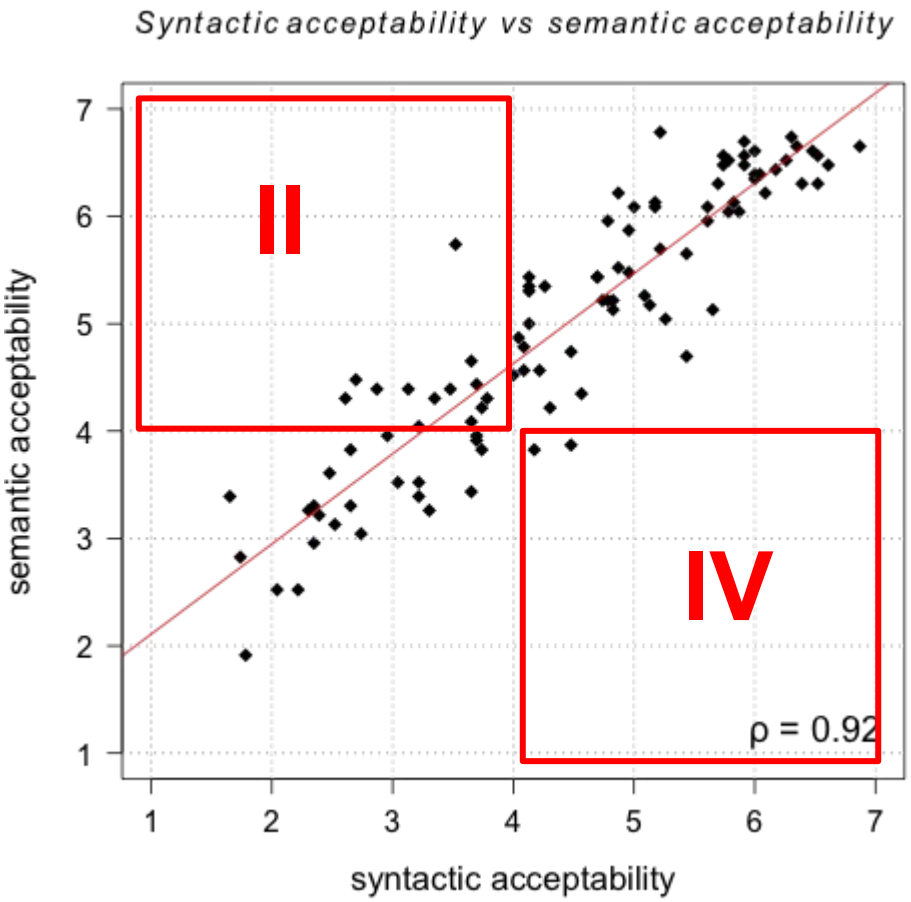


Grammatical but non-sense items

Should we conclude a bi-directional impact?

Apart from a conflation, it's hard to see any directed impact.

Ungrammatical but intelligible items



Grammatical but non-sense items

Should we conclude a bi-directional impact?

Potential confound here:

LI items are designed to illustrate syntactic violations.

Obvious semantic anomalies are avoided.

Items are supposed to be intelligible.

→ Experiments 2 & 3 test items that are intentionally manipulated wrt their semantics.

A suspicion:

Participants struggle to ignore meaning in acceptability ratings (even when asked to do so).

Ungrammatical items are hard to judge regarding semantic unmarkedness, because this requires a repair of the grammatical structure. The repair might not be obvious.

Or: They struggle to distinguish the tasks.

Experiment 2

Experiment 2

Got replaced by Experiment 3

Experiment 3

Disentangling syntactic and semantic factors

Experiment 3

Two basic comparisons:

- syntactic vs semantic acceptability rating
- syntactic vs. semantic anomalies

Additional comparison: two types of semantic anomalies

- contradictions
- violations of selectional restrictions

(To keep the design manageable, we decided to test only one type of syntactic violation, viz. agreement violations.)

Experiment 3

Factors

Between groups (Experiment 3a/3b)

- Rating task: syntactic rating vs. semantic rating

Within Items (and within participants)

- Syntax: +/- grammatical
- Semantics: +/- meaningful/plausible

Between items (but within participants)

- Type of semantic violation: contradiction vs. violation of selection restriction (see examples next slides)

Experiment 3

Syntactic anomaly: agreement violation

*Annie are my mother and she is my best friend.

*The two removers was inspecting the safe for ten minutes.

50% of the violations: singular subject and plural verb

50% of the violations: plural subject and singular verb

Additional violation types in fillers:

- island violations
- Subcategorization errors
- word order errors

Experiment 3

Semantic anomalies: Contradictions

1) #My sister Jane is married to a bachelor.
2) #Annie is my mother and she is my twin sister.
3) #San Diego is both entirely in California and not entirely in California.
4) #My new Volkswagen is emitting a lot of carbon dioxide, but it is not emitting any CO2.

Experiment 3

Semantic anomalies: licensing of durative *for*-PP

(Violations of selectional restrictions; verb semantics)

1) #All of the windows were breaking for at least three minutes.
2) #The two removers were dropping the safe for ten minutes.
3) #Two of the balloons were popping for about 20 minutes.
4) #Mary was winning the national lottery for three weeks.

Experiment 3

The factors Syntax and Semantics were fully crossed.

4 conditions within items:

[+grammatical,+meaningful/plausible]

[-grammatical,+meaningful/plausible]

[+grammatical,-meaningful/plausible]

[-grammatical,-meaningful/plausible]

2 conditions between items:

Contradictions vs. *for*-PPs

Experiment 3

Materials

16 items with +/- contradiction
16 items with *for*-PP (+/- licensed)

32 fillers with various syntactic and/or semantic violations

Experiment 3

Contradictions

(1a) Annie is my mother and she is my best friend.

(1b) *Annie are my mother and she is my best friend.

(1c) #Annie is my mother and she is my twin sister.

(1d) #*Annie are my mother and she is my twin sister.

Experiment 3

Violation of selectional restrictions:

Licensing of a durative *for*-PP

(2a) The children are staying for ten days.

(2b) *The children is staying for ten days.

(2c) #The children are arriving for ten days.

(2d) #*The children is arriving for ten days.

Experiment 3

Participants

- recruiting via Prolific
- exclusion criteria as before

Participants in analyses: 37 + 37

Experiment 3

Procedure: adapted from in Experiment 1 (7pt scales)

Changes

- labels on the scale (endpoints)
- different example in the instruction
- new calibration items
- new “gotcha” items

Experiment 3 Syntactic rating

John was arrested.

unnatural/ungrammatical <<< >>> natural/grammatical

Experiment 3

Semantic rating

John was arrested.

meaningless/implausible <<< >>> meaningful/plausible

Experiment 3

Examples in the instruction

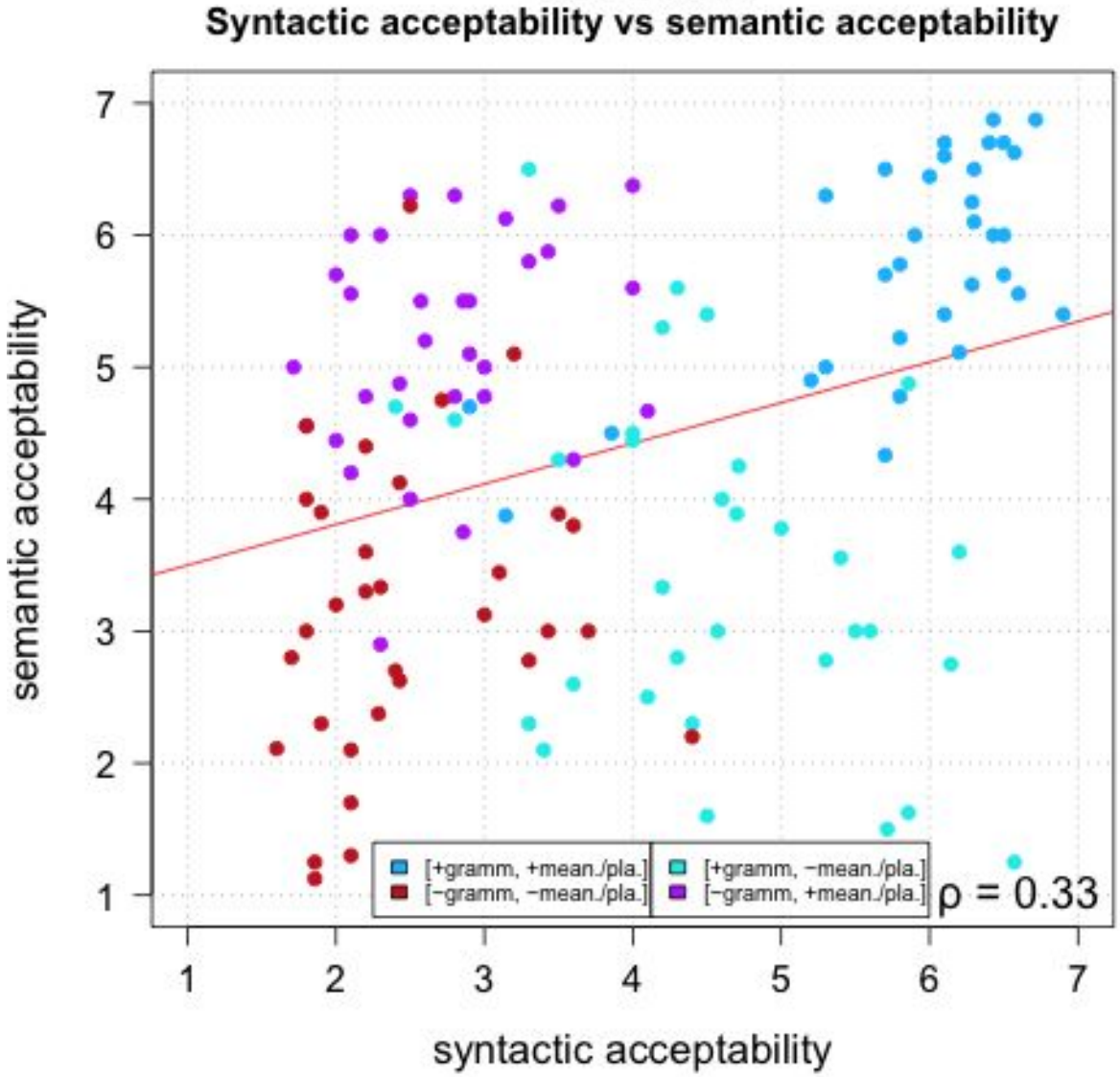
* John did his job goodly.

The storm intentionally broke the window.

Experiment 3

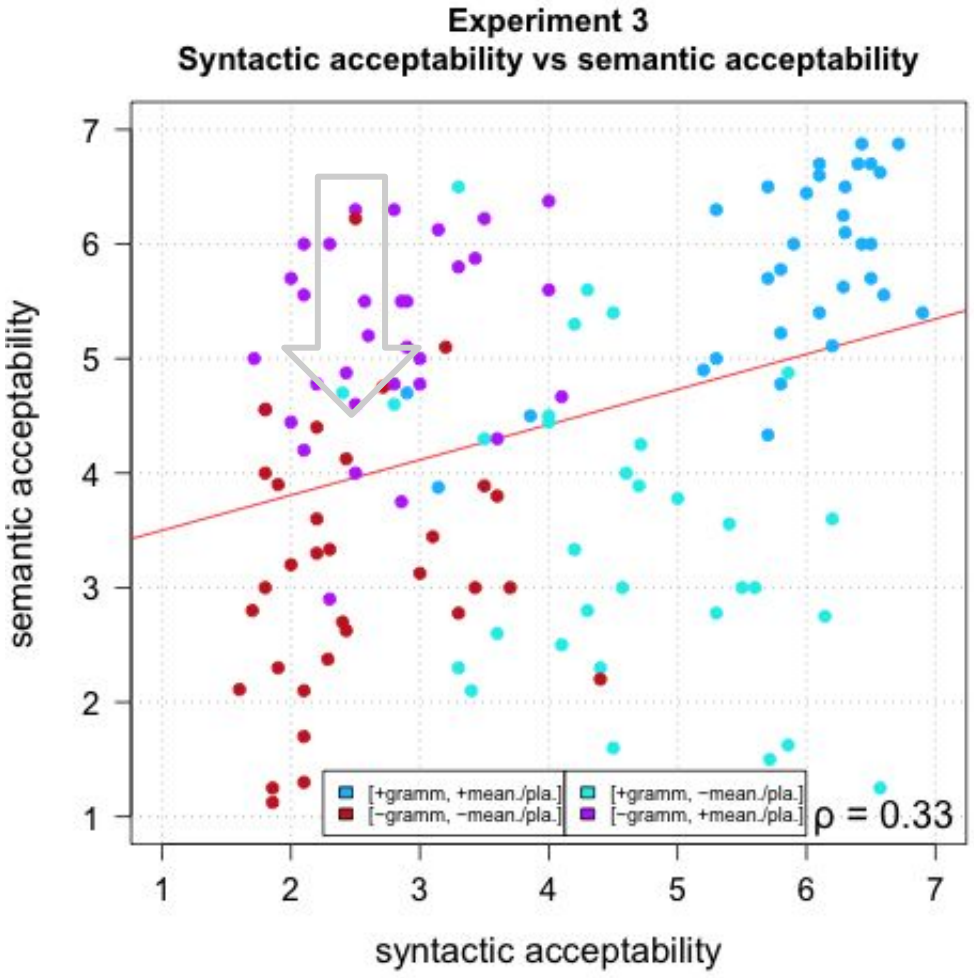
Results

Experiment 3 Results



Experiment 3

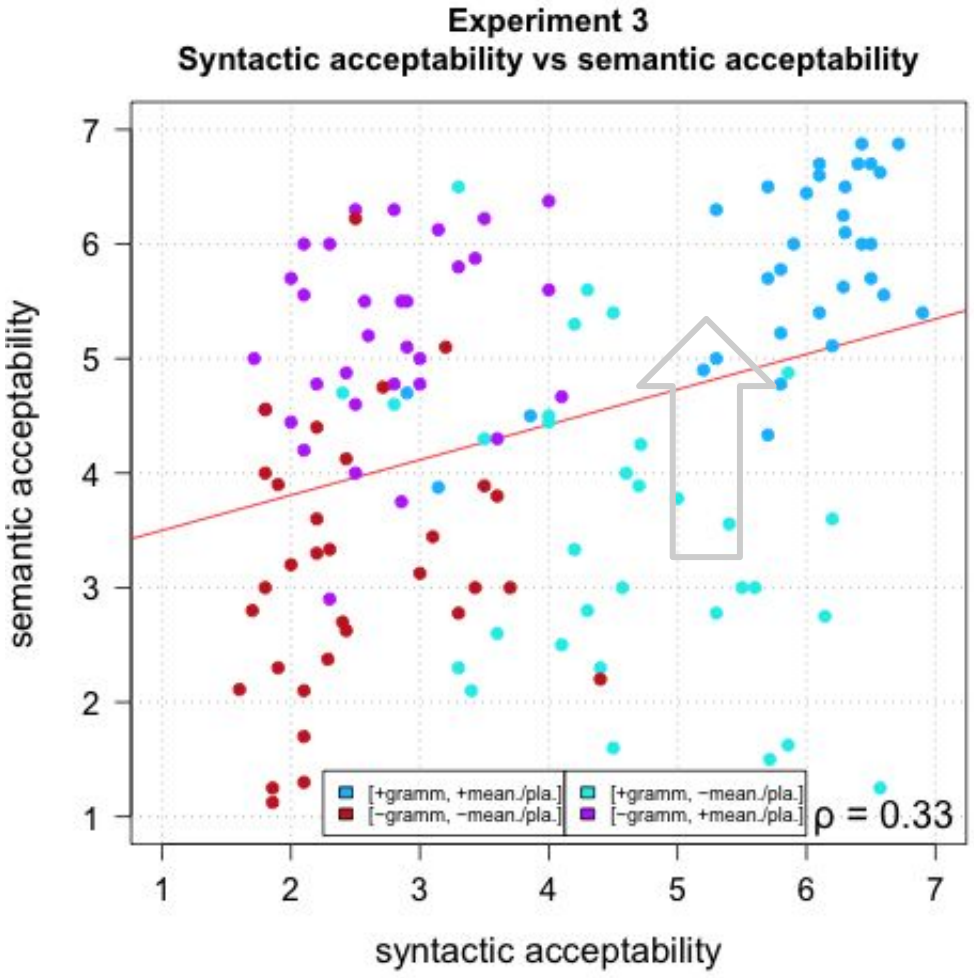
Results



No degrading effect of grammaticality on semantic ratings.

Experiment 3

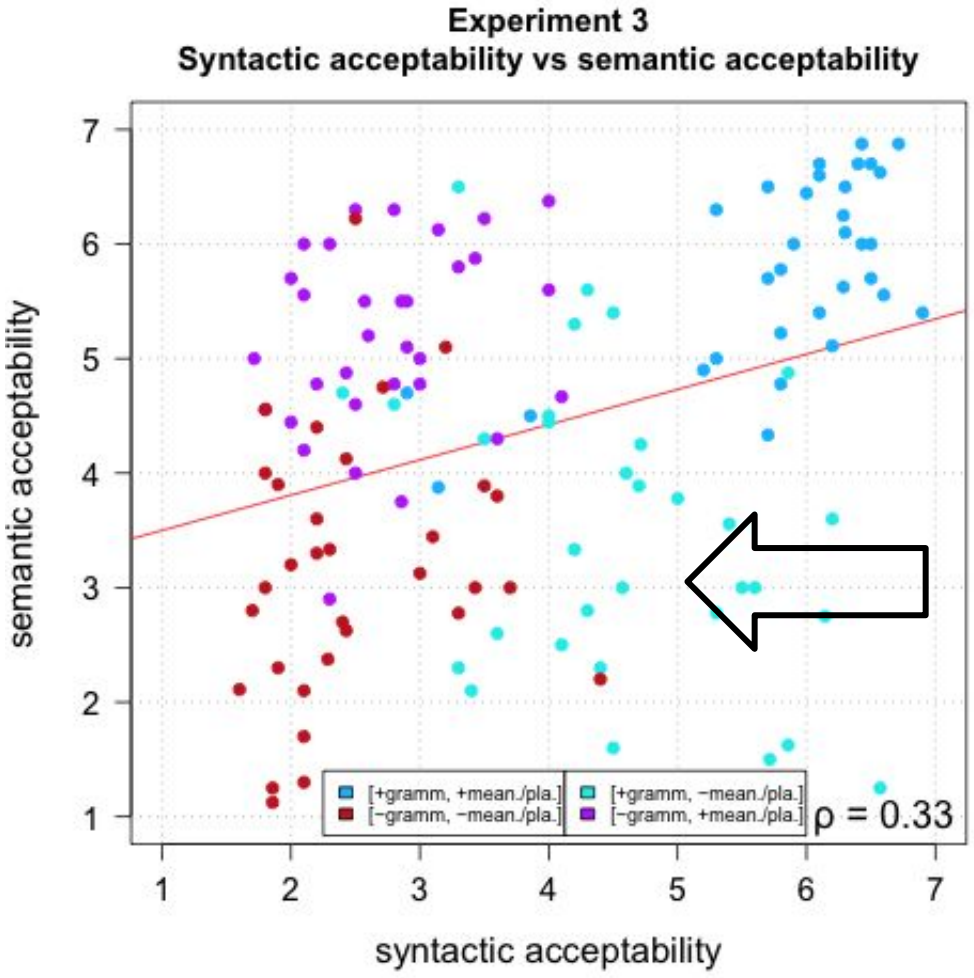
Results



Ameliorating effect of grammaticality on semantic ratings?

Experiment 3

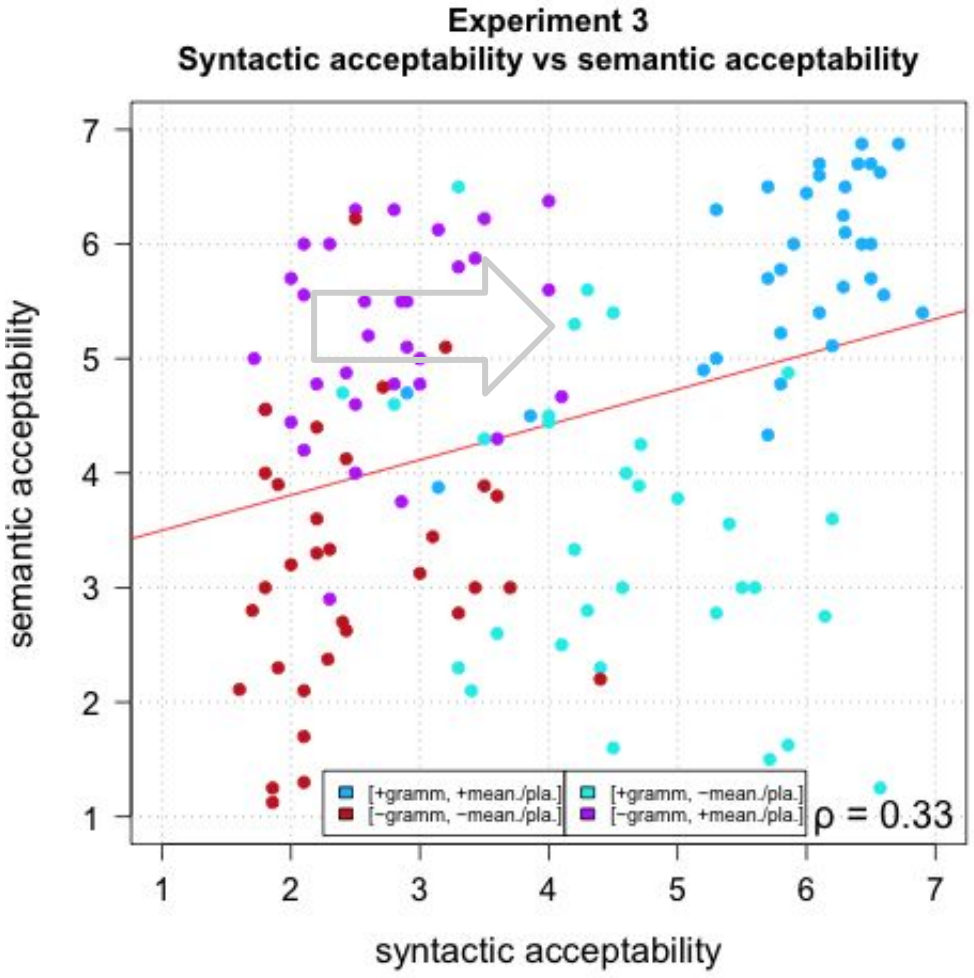
Results



degrading effect of plausibility on syntactic ratings

Experiment 3

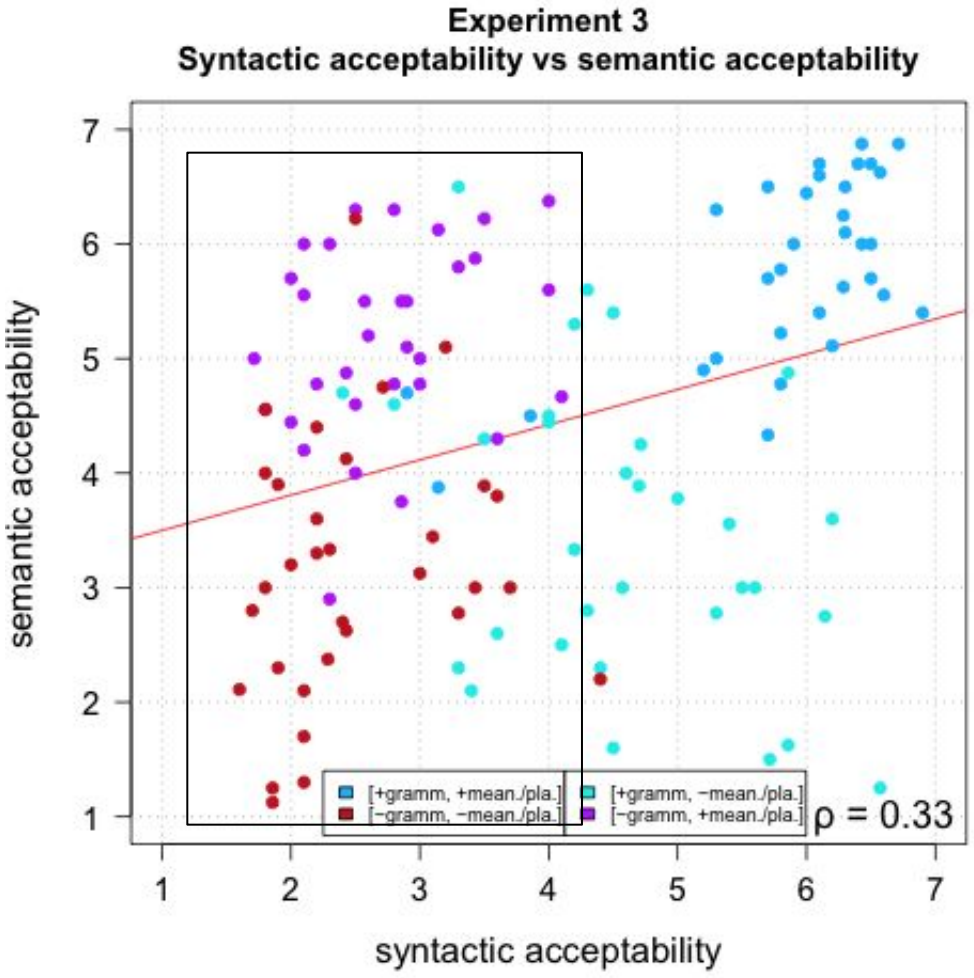
Results



No ameliorating effect of plausibility on syntactic ratings

Experiment 3

Results



No ameliorating effect of plausibility on syntactic ratings

[-gramm., +plaus.] & [-gramm., -plaus] cover the same on the syntactic scale

Experiment 3 Results

Participants can make the distinction.

At the same time, we also see degrading effects of plausibility on syntactic ratings.

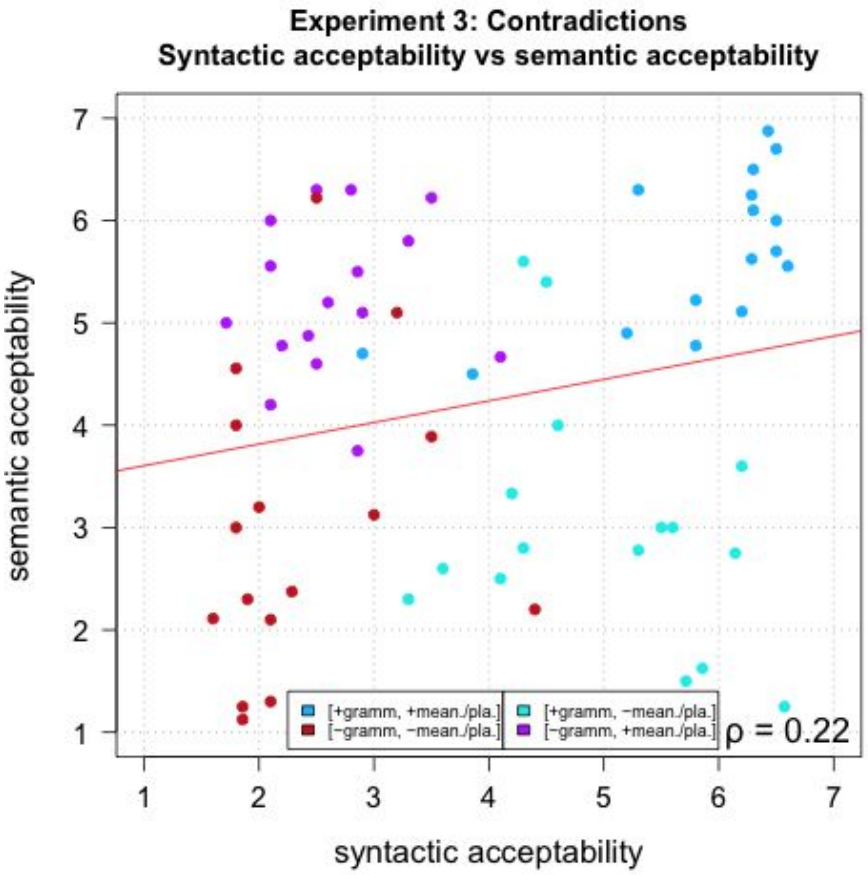
Semantic ratings seem to be less vulnerable to syntactic effects.

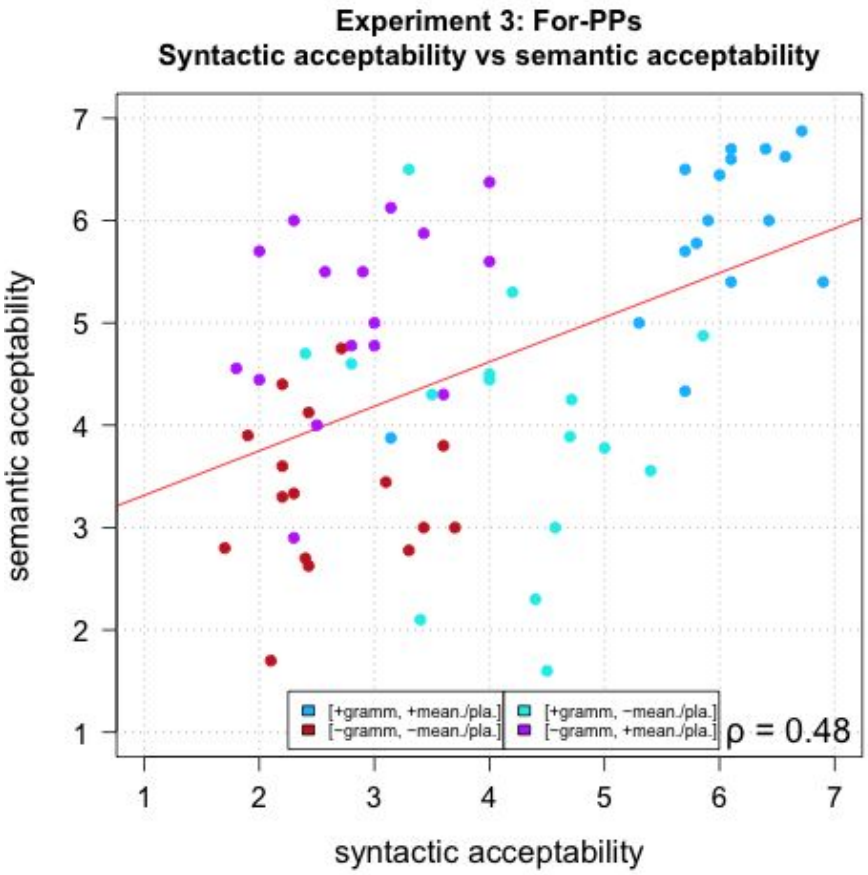
Caution: This could be due to the type of syntactic violation examined in Experiment 3 (agreement violations).

The impact might be different for other violations.

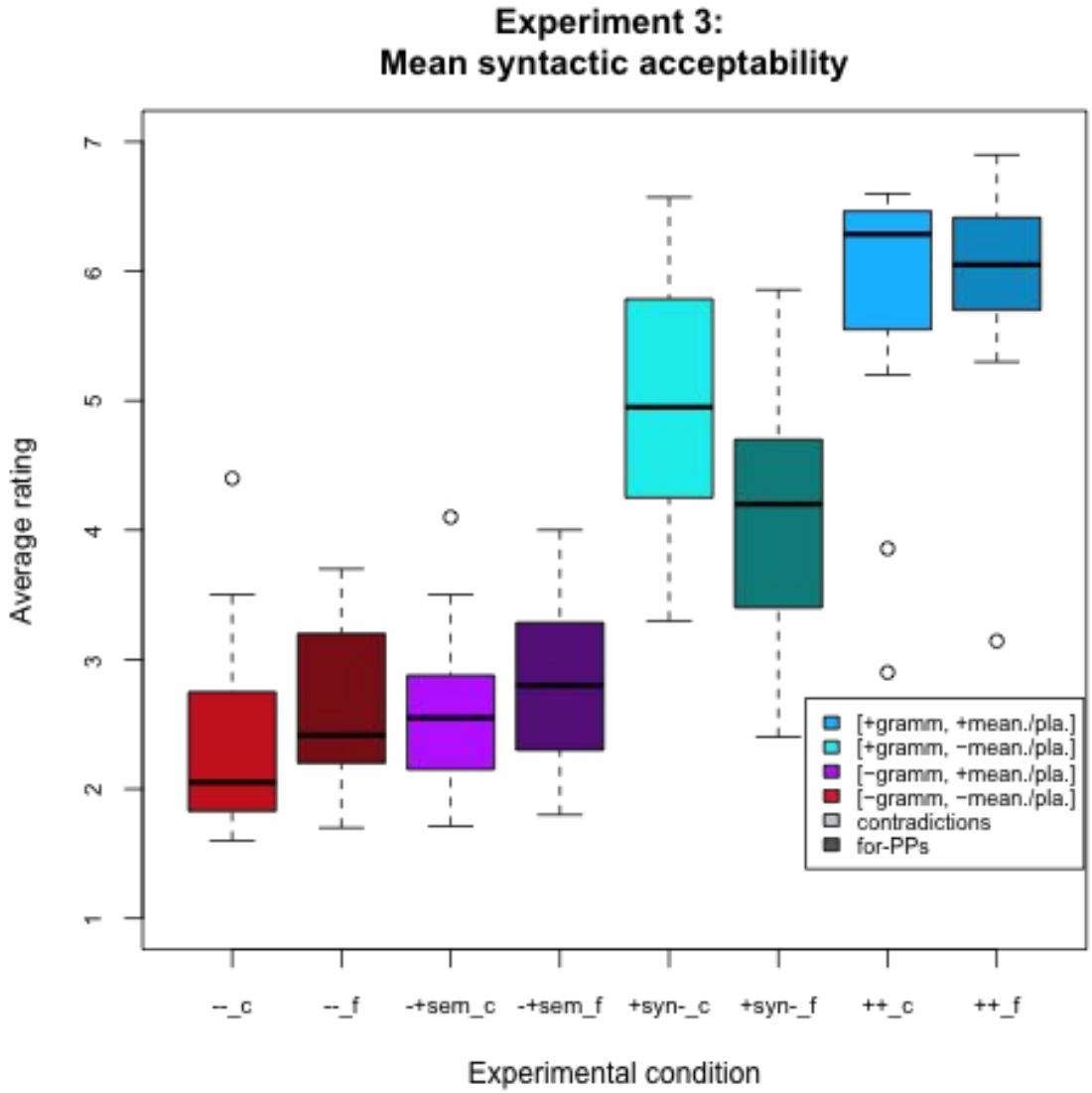
Comparing the two types of semantic anomalies

Comparing the two types of semantic anomalies

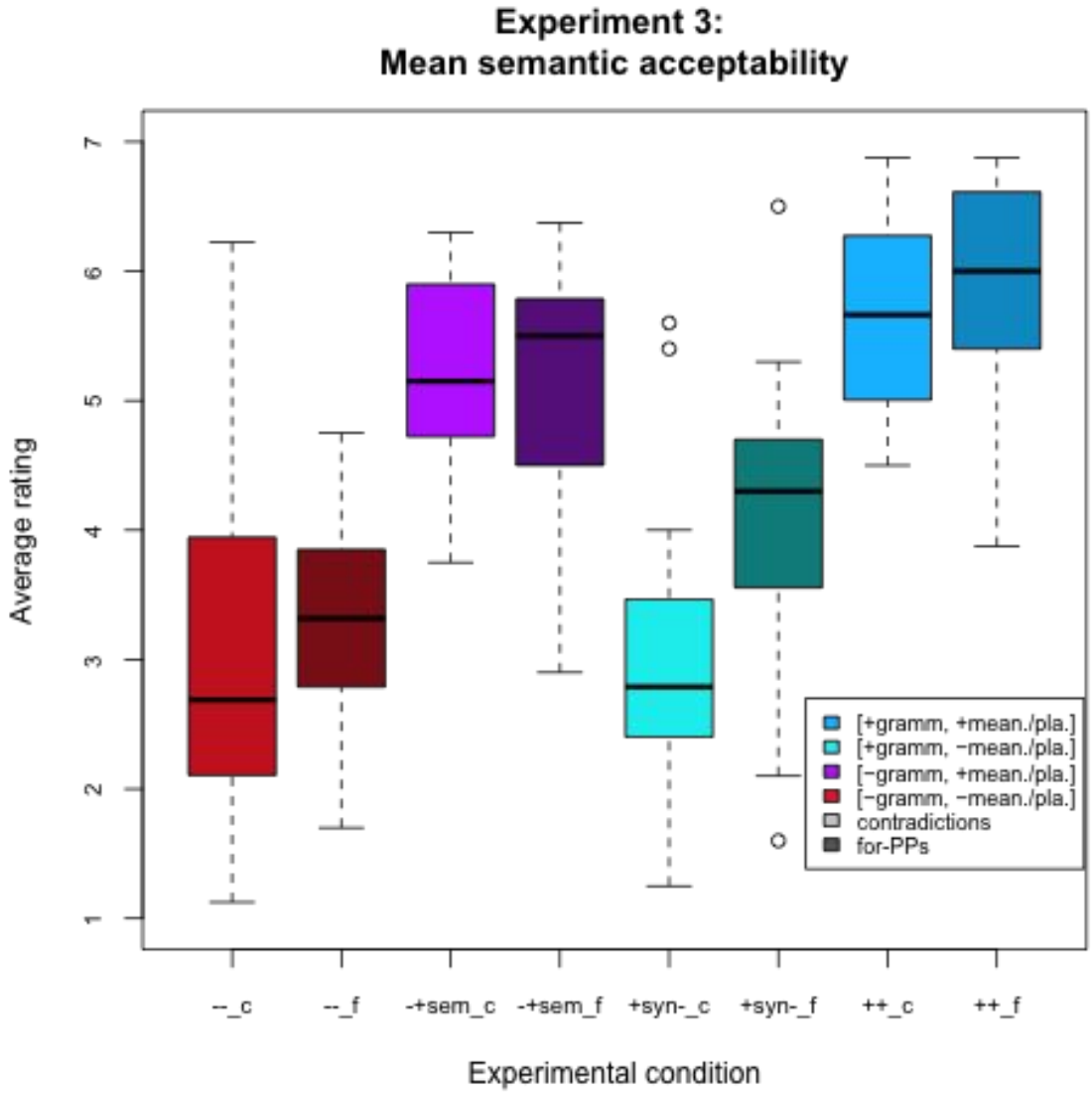




Comparing the two types of semantic anomalies



Comparing the two types of semantic anomalies



Comparing the two types of semantic anomalies

Violations of selectional restrictions on *for*-PP received intermediate ratings on both scales.

Compared to contradictions, …

- they receive a stronger penalty on the syntactic scale,
- and a weaker penalty on the semantic scale

Apparently, the violation is considered a matter of grammar and of semantics, and less severe than “pure” violations.

General Discussion

Theoretical impact:

- *for*-PPs challenge the syntax-semantics border

- Semantic effects explain only little of the gradience observed in Experiment 1 and the LE-2016-study

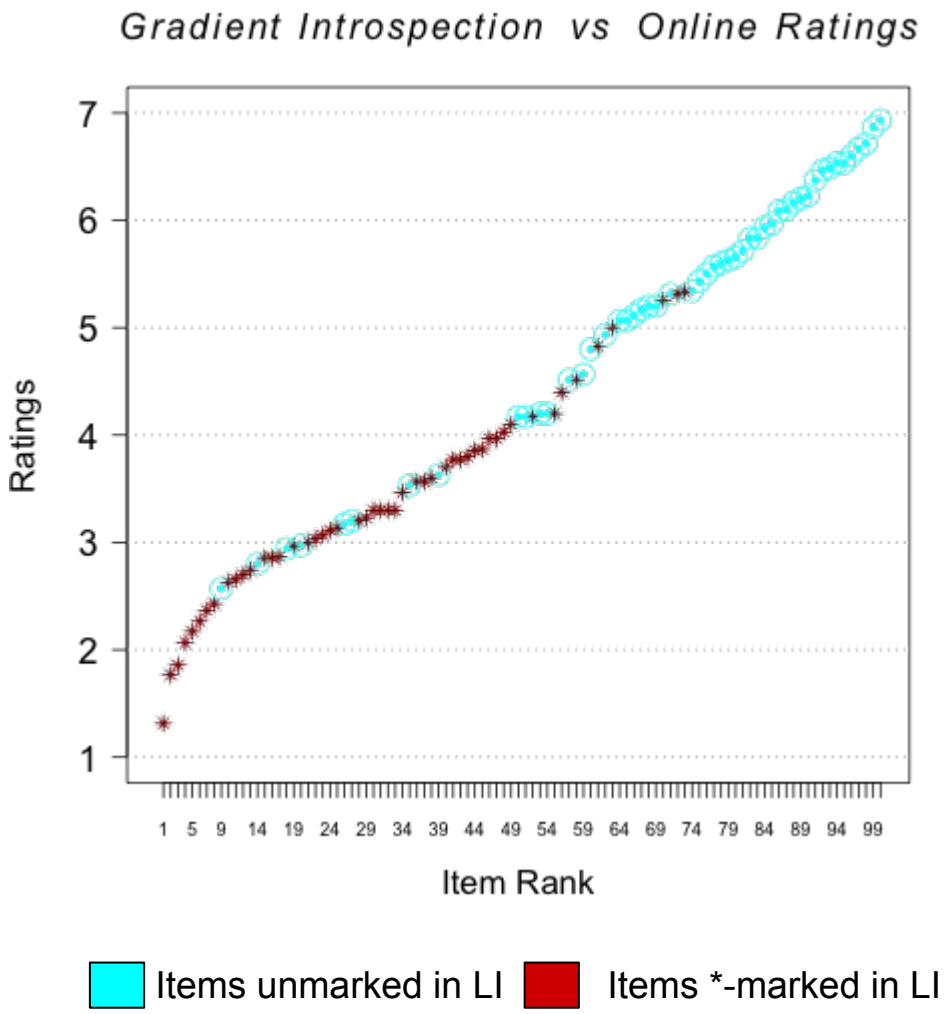
→ contributes a further piece of evidence for gradience rooted in grammar (in addition to other sources)

General Discussion

Back to our starting point:

Why are so many *-items in the mid-bin?

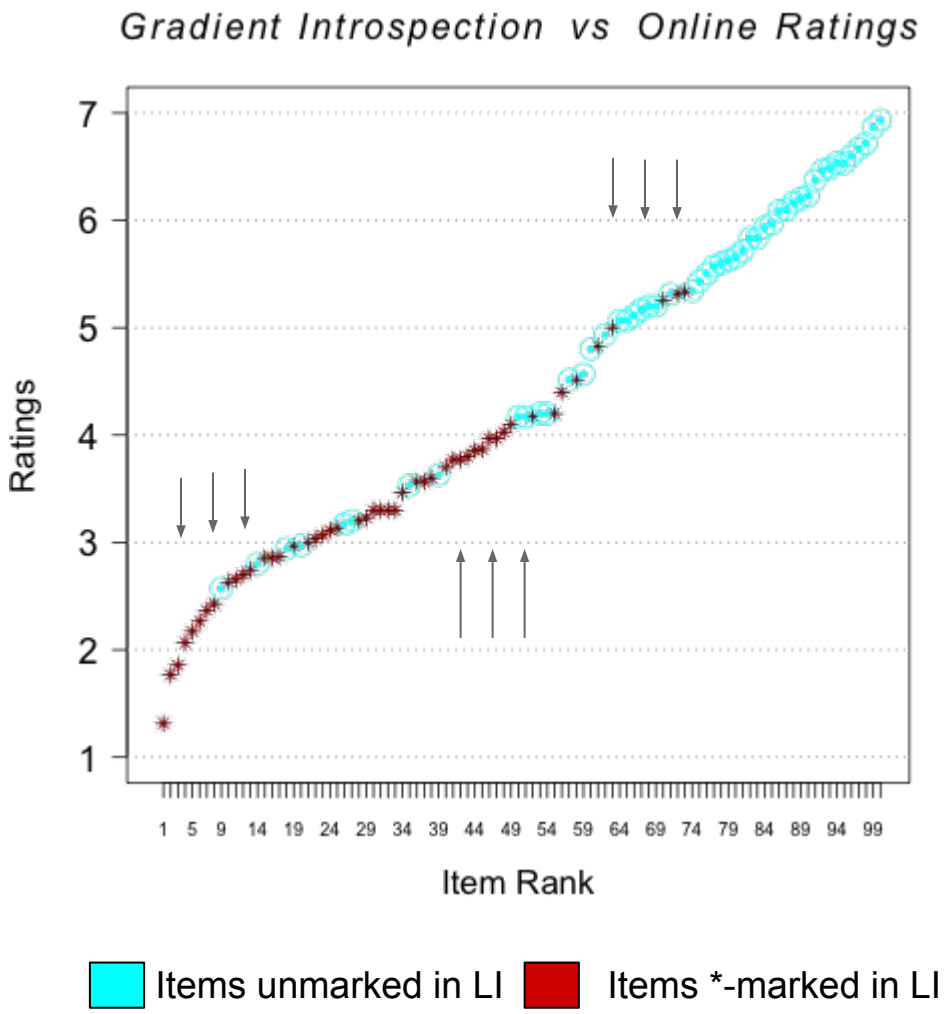
- some semantic effect
- some scale effects
- probably not due to performance factors (gramm. illusions)



General Discussion

If there are semantic effects, do they work in both directions?

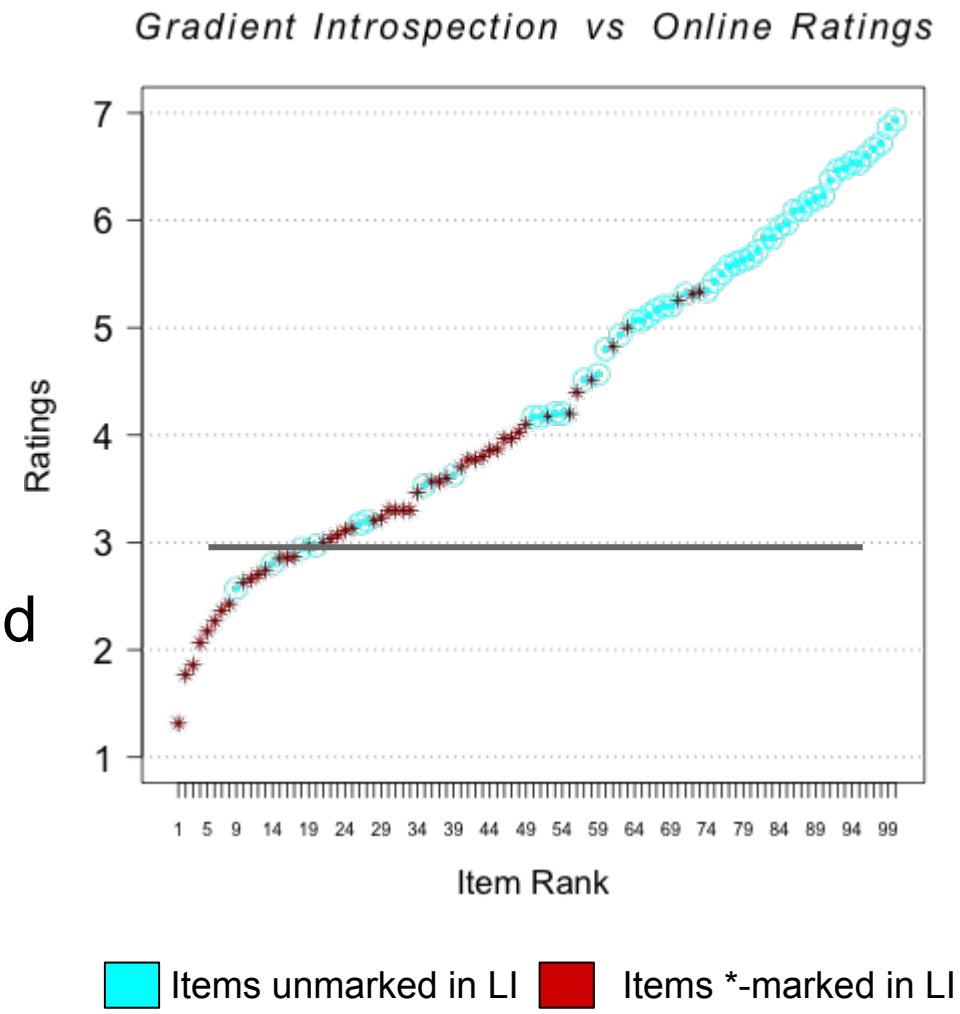
And how can we decide? (penalty or bonus)



General Discussion

Further question:

Is there gradience among ungrammatical items?
(in the grammar, modulated by extra-grammatical factors)



Open Questions

Do linguistic make sharper distinctions?

(cf. Culbertson & Gross, 2009 vs. Devitt 2014)

Will other violations have other effects?

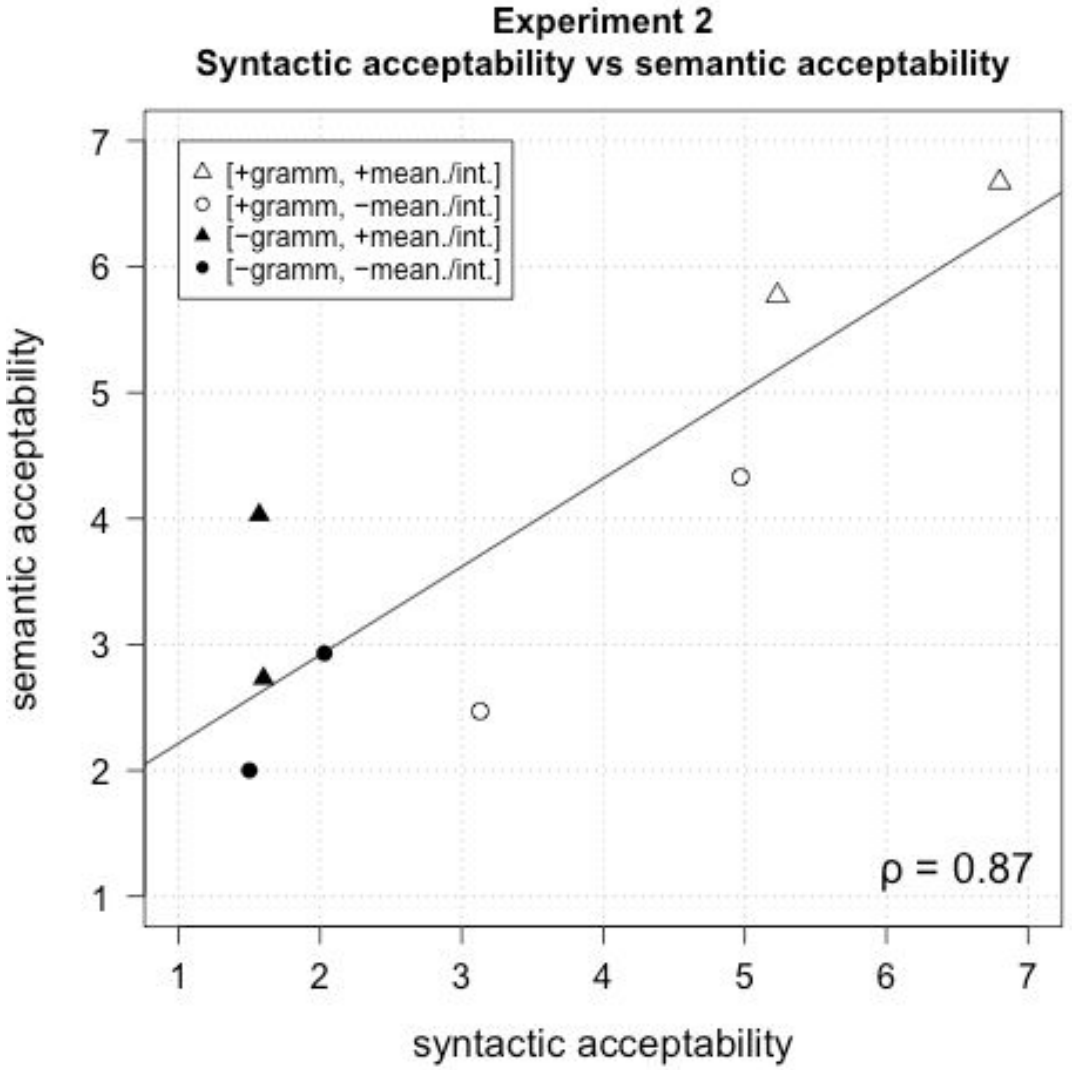
What explains the discrepancy btw Exp2 and Exp3?
(size/heterogeneity of item set? Scale labels?)

How to model various impact factors simultaneously?

Additive effects? Interactions?

What does “meaningful” actually mean?

Experiment 2



Pilot study

Participants: 30+30

Procedure: as in Exp1
(7-pt scale, no labels)

Materials: 8 items
(two in each condition)

Two factors:

- Grammaticality
- Meaningfulness

Experiment 2

[+gramm., +meaning]

(1a) *Mary was given a cake*.

(1b) *John tried to begin to eat his sandwich*

[+gramm., -meaning]

(2a) *Mary's twin brother has no biological sister*.

(2b) *I take my coffee with cream and dog*

[-gramm., +meaning]

(3a) *John goed home fastly*.

(3b) *Peter haven't heard that Mary called he*.

[-gramm., +meaning]

(4a) *John say to take good water of oneselves*.

(4b) *These might seeming mice to be in the number*.

Two side notes

- Note that our study is explorative. Tests will be posthoc.
  → Problem for publication?

- Prolific
- Growing pool participants
- Prescreening possible
- Ethical payment
- low rate of non-cooperative behaviour

→ good alternative to MTurk (especially outside the US)

https://www.prolific.ac/

Thank you!

Special thanks to Tom Wasow for advice and discussions

Thank you!

Special thanks to Tom Wasow for advice and discussions