Corpus methods for research in pragmatics

Judith Degen, Stanford University

15/08/2016

ESSLLI 2016, Bolzano

Course goals

1. Corpora in linguistics / pragmatics

2. Steps involved in a corpus pragmatics project

3. Hands-on experience with corpus search / annotation / analysis / visualization

4. Example phenomenon: scalar implicature

5. Novel project: projection behavior of factive

Organizational

Website:
https://thegricean.github.io/esslli2016_corpuspragmatics

Schedule

Technical infrastructure (login credentials, tools)

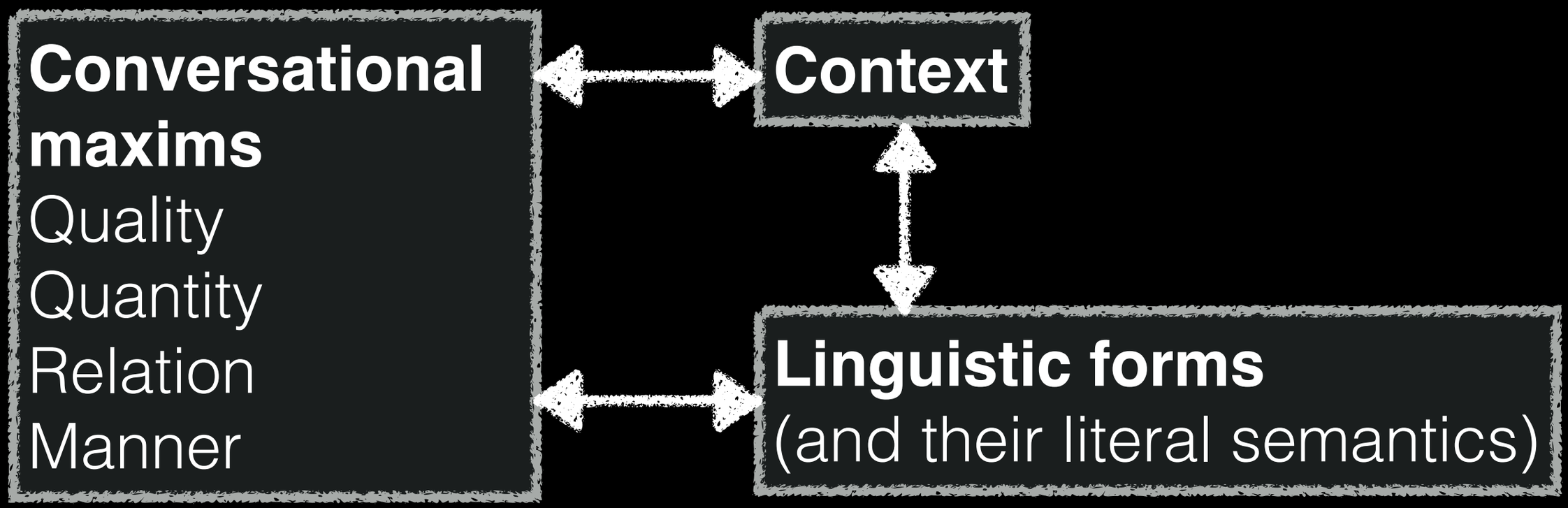
Today

- corpora in linguistics / pragmatics
- corpus vs other empirical methods in pragmatics
- challenges
- example: variation and context-dependence in scalar implicature Degen 2015

Building pragmatic theories

Cooperative Principle

"Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." — Grice 1975



Building pragmatic theories

Do scalar implicatures have a special GCI status?

Do scalar implicatures survive context shifts?

(1) Ann: Was the exam easy?
Tom: Some of the students failed.
Some, but not all, of the students failed.
The exam was hard.

(2) Ann: How is the teacher doing?
Tom: Some of the students failed.
Some, but not all, of the students failed.
The teacher isn't doing great.

Introspective judgments

Advantages
Fast and easy to obtain

Disadvantages
Small number of judgments (usually 1)
Small number of items
Items hand-selected by “experimenter”

Introspective judgments

Advantages
Fast and easy to obtain

Disadvantages
Small number of judgments (usually 1)
Small number of items
Items hand-selected by “experimenter”

} **BIAS**

Introspective judgments

Advantages
Fast and easy to obtain

Disadvantages
Small number of judgments (usually 1)
Small number of items
Items hand-selected by “experimenter”

} **BIAS**

Good as one source of data in theory-building, often a great starting point, but should never be the only source

de Marneffe & Potts 2014; Gibson et al 2011

Data in semantics/ pragmatics

- introspective judgments (*, ?, ??, #, ##)
- data from controlled psycholinguistic experiments
- naturally occurring linguistic data (corpora)

Psycholinguistic experiments

Advantages

Many participants

Many items

Many ways of measuring quality of interest

Disadvantages

Items often hand-selected by experimenter

Subjects exposed to unnatural distributions

Experiments can’t be arbitrarily long

Psycholinguistic experiments

Advantages

Many participants

Many items

Many ways of measuring quality of interest

Disadvantages

Items often hand-selected by experimenter

Subjects exposed to unnatural distributions

Experiments can’t be arbitrarily long

} **BIAS**

Psycholinguistic experiments

Advantages

Many participants

Many items

Many ways of measuring quality of interest

Disadvantages

Items often hand-selected by experimenter

Subjects exposed to unnatural distributions

Experiments can't be arbitrarily long

} **BIAS**

Great for targeted testing for an effect of interest, but uncertainty about naturalness of experimental context a huge concern in pragmatics

Degen & Tanenhaus 2015

Corpora

Advantages
Lots of naturally occurring data

Disadvantages
Unbalanced data
Not always annotated with the necessary information

Corpora

Advantages
Lots of naturally occurring data

Disadvantages
Unbalanced data
Not always annotated with the necessary information

Corpus analyses can be combined in very fruitful ways with introspective judgments & psycholinguistic data!

Artificial example:
Some of my friends came to the party.
~> Some, but not all, of my friends came to the party

Real example, observable:
Today we're gonna discuss some of the causes of grief.
~> some but not all of the causes?

posit

infer/annotate

Unobservable:
speaker intentions
common ground
listener inferences
world knowledge

Combining corpora and crowd-sourced experiments

corpora

experiments

Combining corpora and crowd-sourced experiments

corpora

experiments

annotate with linguistically untrained speakers' judgments on Mechanical Turk

Combining corpora and crowd-sourced experiments

corpora

experiments

```
var runModel = function(speake
  var speakerERP = speakerMode
  return Enumerate(function()
    var utt = sample(speakerER
    factor(params.speakeroptim
    return utt;
  });
};
```

models

annotate with linguistically untrained speakers' judgments on Mechanical Turk

How to pick a corpus

- Modality — spoken or written?
- Genre — news corpora, casual dialogs between friends, professional dialogs between co-workers, fiction, the Bible, child-directed speech
- Language
- Size (big difference between 200,000 and 20,000,000,000 tokens)
- Available annotation (typically: size-annotation tradeoff)
- Accessibility

Types of annotation

S-SIDE

Syntactic

Part-Of-Speech (POS) tags

Syntactic parses

Semantic

Co-reference

Information status (givenness)

…

SOCIAL

Speaker gender

Speaker age

Dialect

…

P-SIDE

Prosodic

Pitch accents

Boundary tones

Phonetic

Word duration

Syllable duration

Phonological

Phonemes

Number of syllables in word

…

General steps in a project

1. formulate research question(s)

2. decide on linking function(s)

3. conduct corpus search (iteratively develop search patterns and build database of variables of interest)

4. add annotation? (expert, crowd-sourced)

5. conduct data analysis and visualization

An example: variation and context-dependence in scalar implicatures from “some” to “not all”

Degen 2015

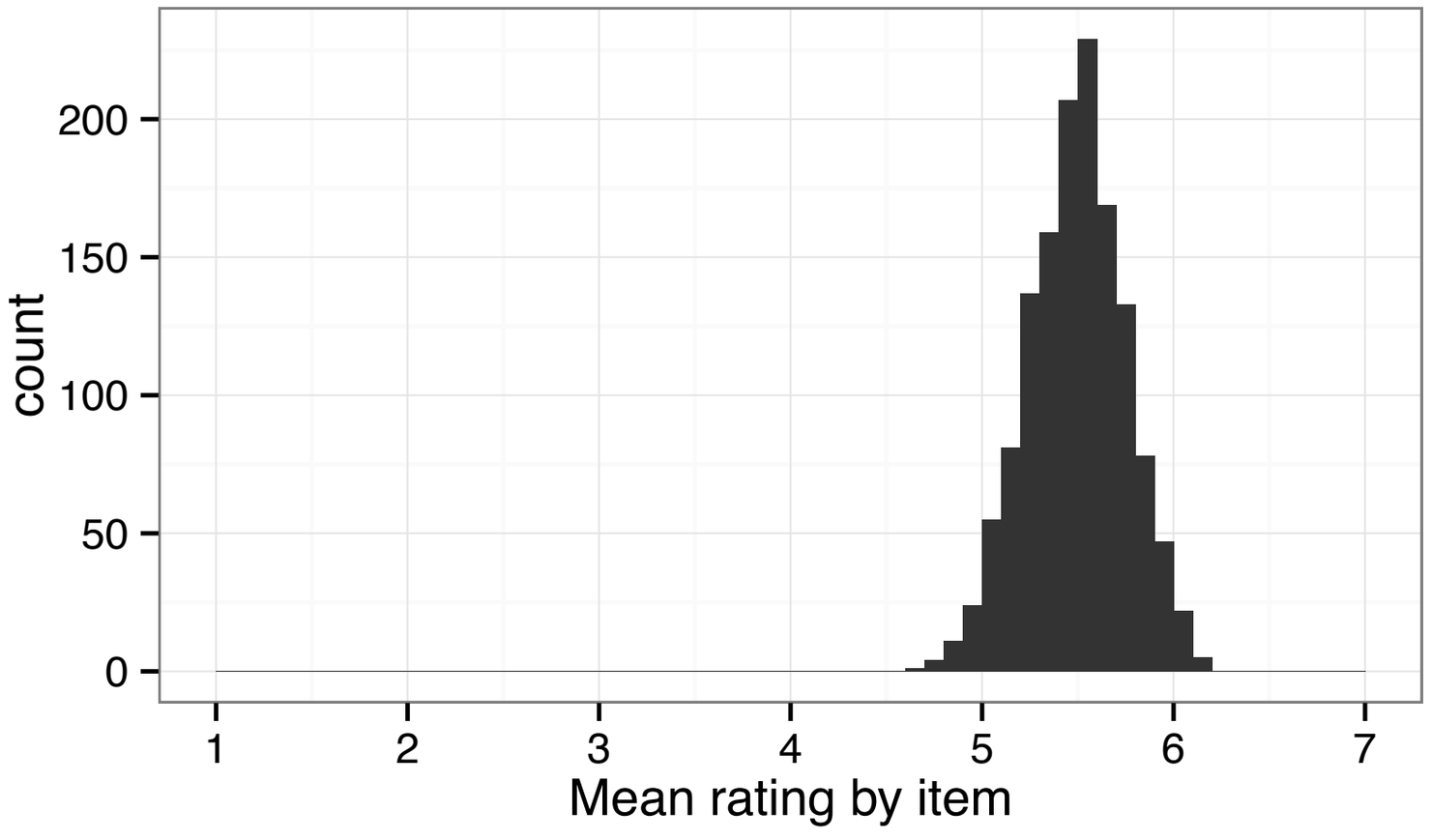
1. Research question

1. Do scalar implicatures from *some* to *not all* constitute a homogeneous class of inferences? Grice 1975; Gazdar 1979; Horn 1984; Levinson 2000

2. If there is variation in implicature strength, is it random or systematic? Russell 2012, Goodman & Stuhlmüller 2013, Degen, Franke & Jäger 2013, Franke & Jäger 2016

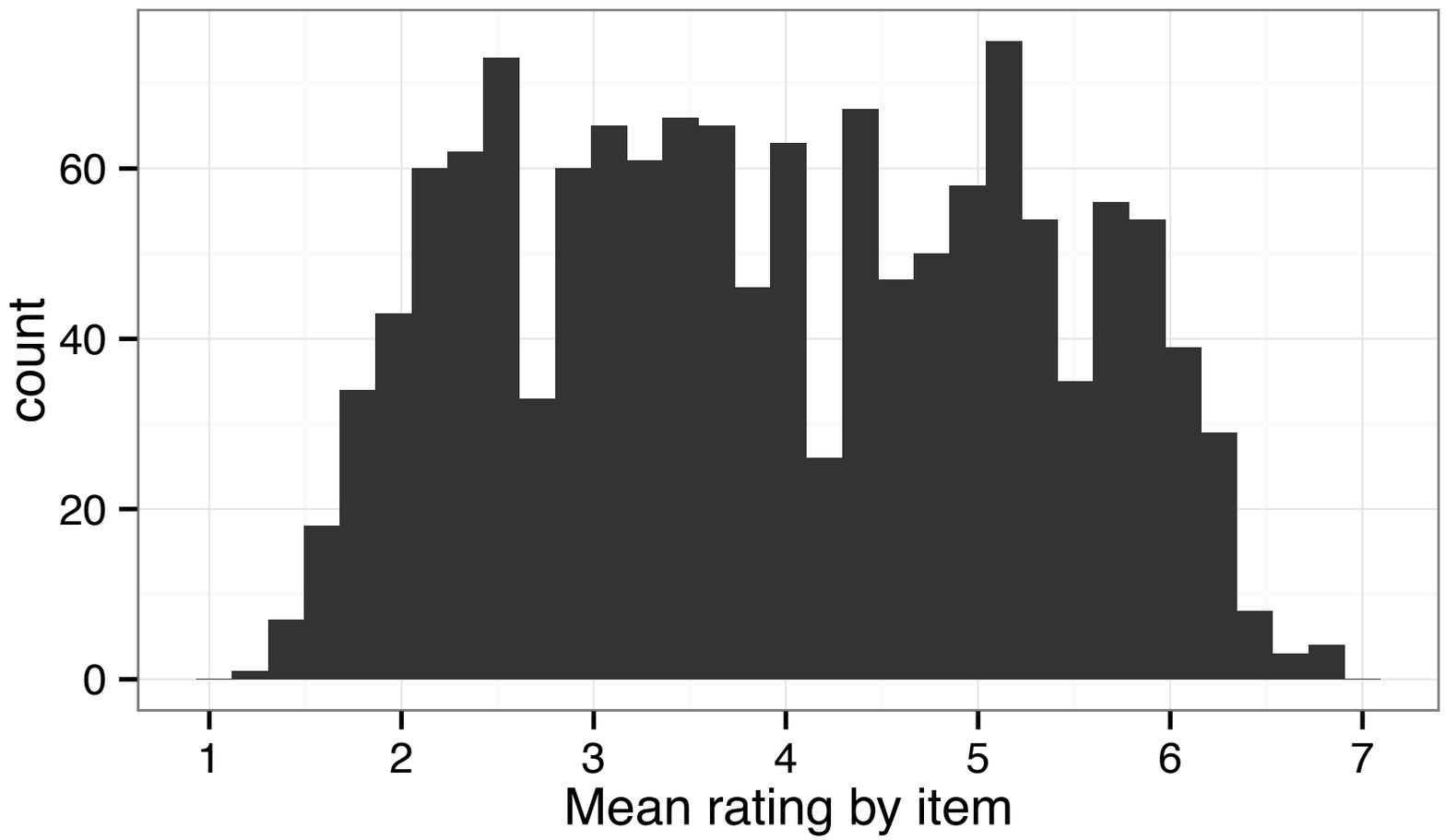
2. Linking function

Invariable implicature strength:



2. Linking function

Variable implicature strength:



Methodology

- Corpus search
  - extract instances of “some”
  - extract information about syntactic/semantic/pragmatic context
- Web-based experimentation
  - collect implicature strength judgments
- Visualization and statistical data analysis

3. Corpus search

Switchboard Corpus
Godfrey & McDaniel 1992, Calhoun et al. 2010

- spoken American English
- telephone dialogs between strangers about pre-defined topics
- ~ 800,000 tokens
- POS-tagged, syntactically parsed; information status annotation for ~ 23% of NPs Nissim et al. 2004

Extract data from corpus

use tgrep2 Rohde 2005 and the TGrep2 Database Tools (TDT) Degen & Jaeger 2011 to construct a database of 1749 “some” utterances

automatically extract available features:

- partitive
- linguistic mention
- grammatical function
- frequency of NP head
- singular/plural
- speaker
- …

SHOW DATABASE

4. Additional annotation

- implicature judgments (crowd-sourced)

- information status completion (expert, following Nissim et al 2004)

Collecting implicature strength ratings

Collecting implicature strength ratings

- Amazon's Mechanical Turk crowd-sourcing service
- for each item, collected similarity rating on 7-point Likert scale
- blocks of 20 items, 10 ratings per item (243 participants)
- 2 practice items

https://www.hlp.rochester.edu/mturk/jdegen/7_qpsome/output/qp.html?assignmentId=foo&list=3

5. Data analysis / visualization

Exclusion

- Cases where NP head is sg count noun (359):

  (1) She stuck my name on **some list**.

  * She stuck my name on some, but not all, list.

  (2) John kicked **some cat** off the street.

  ? John kicked some, but not all, cat off the street.

- Cases where entire NP consists of **some** (26):

  (3) **Some** say that coffee is healthy.

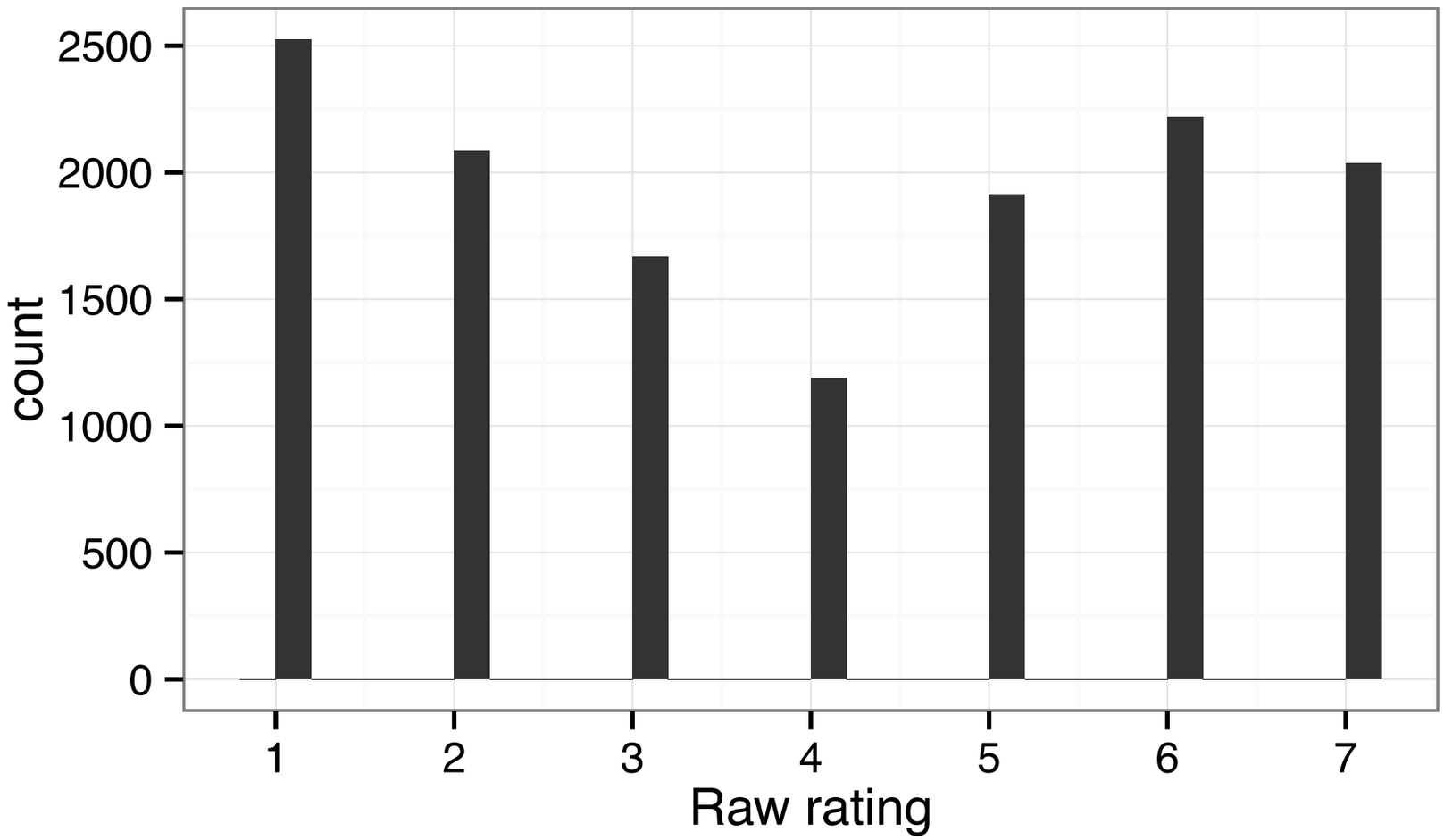
- Leaves 1363 cases

Analysis

1. Do scalar implicatures from *some* to *not all* constitute a homogeneous class of inferences?

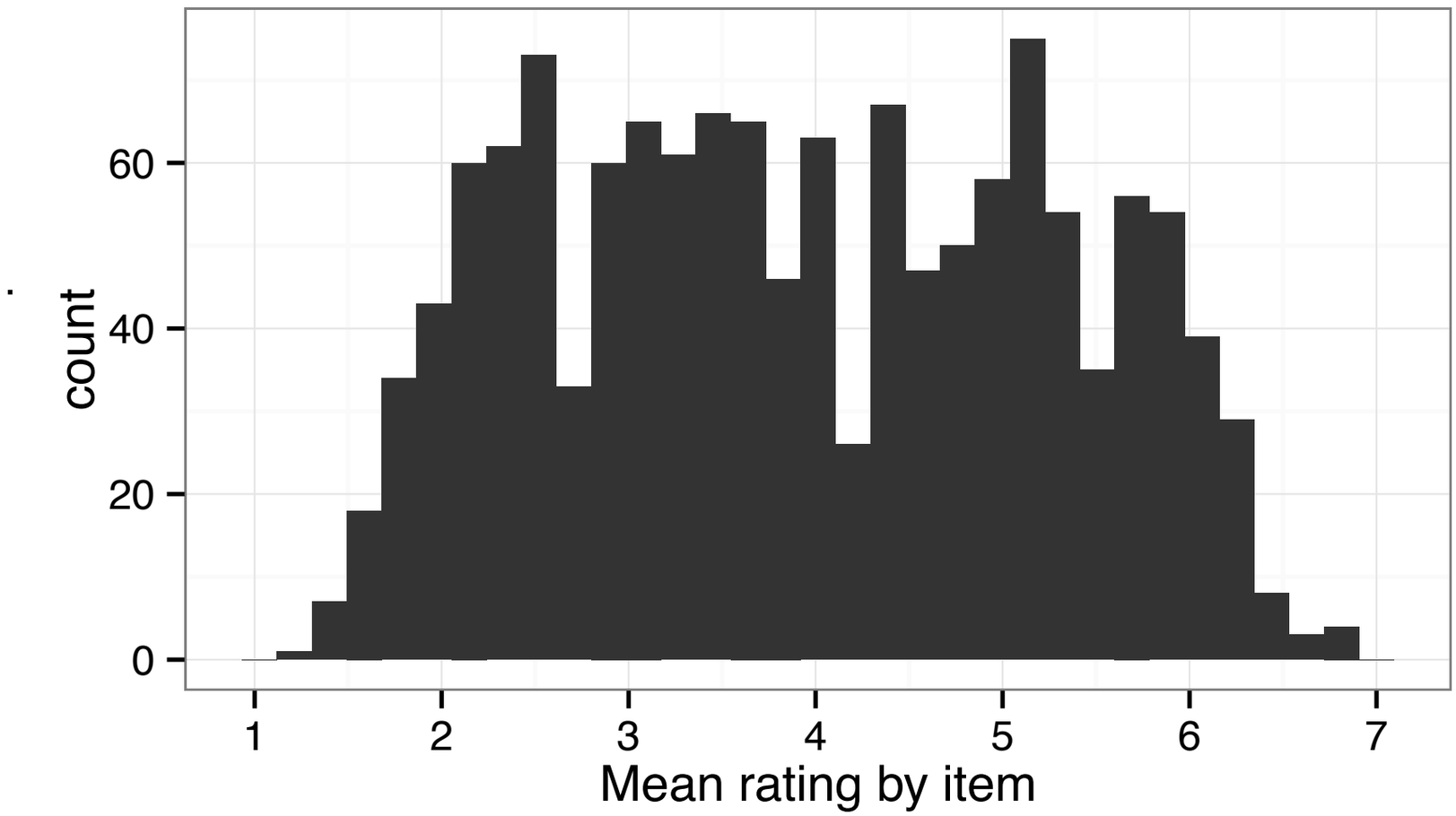
Analysis

1. Do scalar implicatures from *some* to *not all* constitute a homogeneous class of inferences?



Analysis

1. Do scalar implicatures from *some* to *not all* constitute a homogeneous class of inferences?

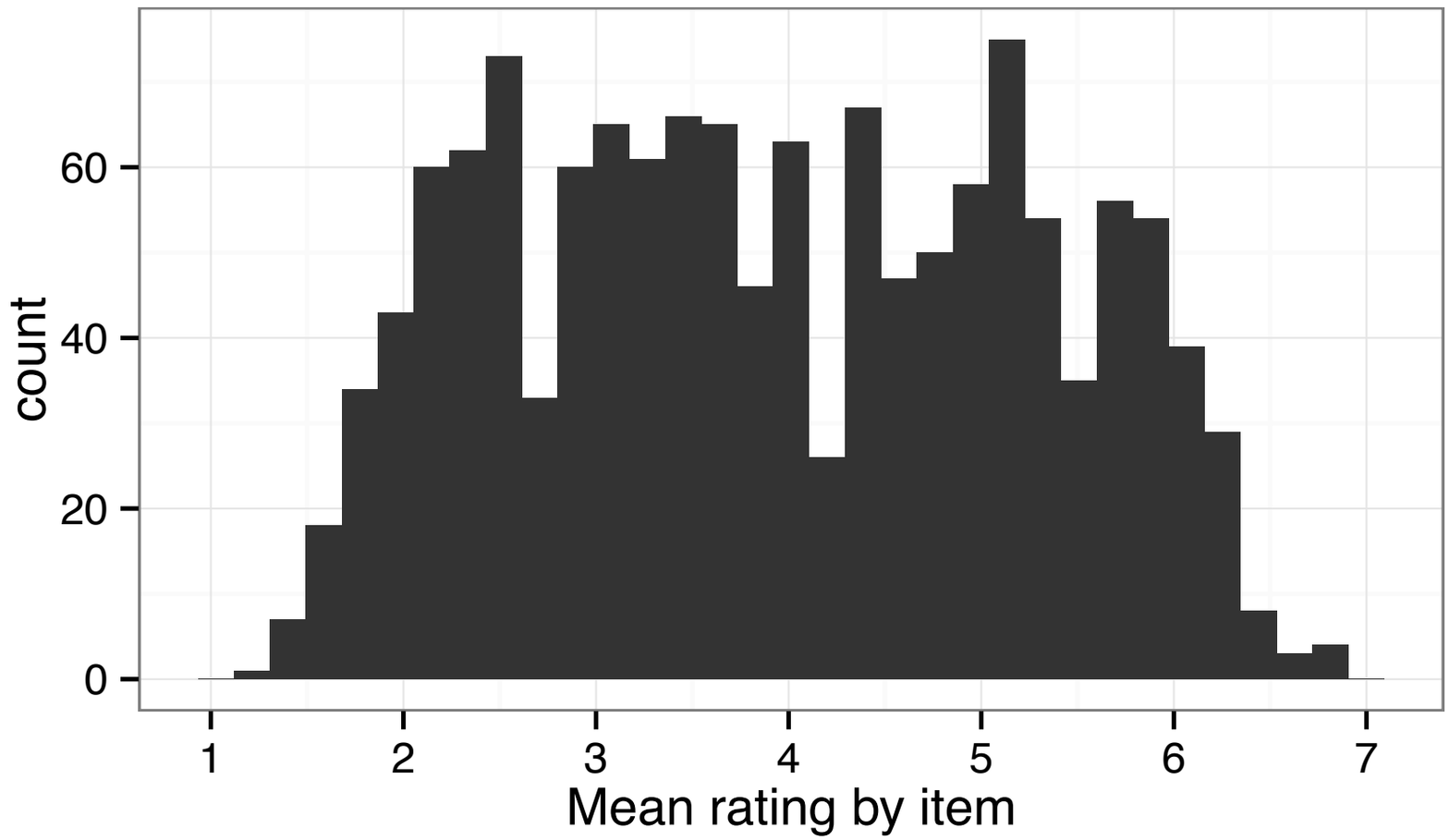


Analysis

2. If there is variation among implicatures, is it random or systematic?

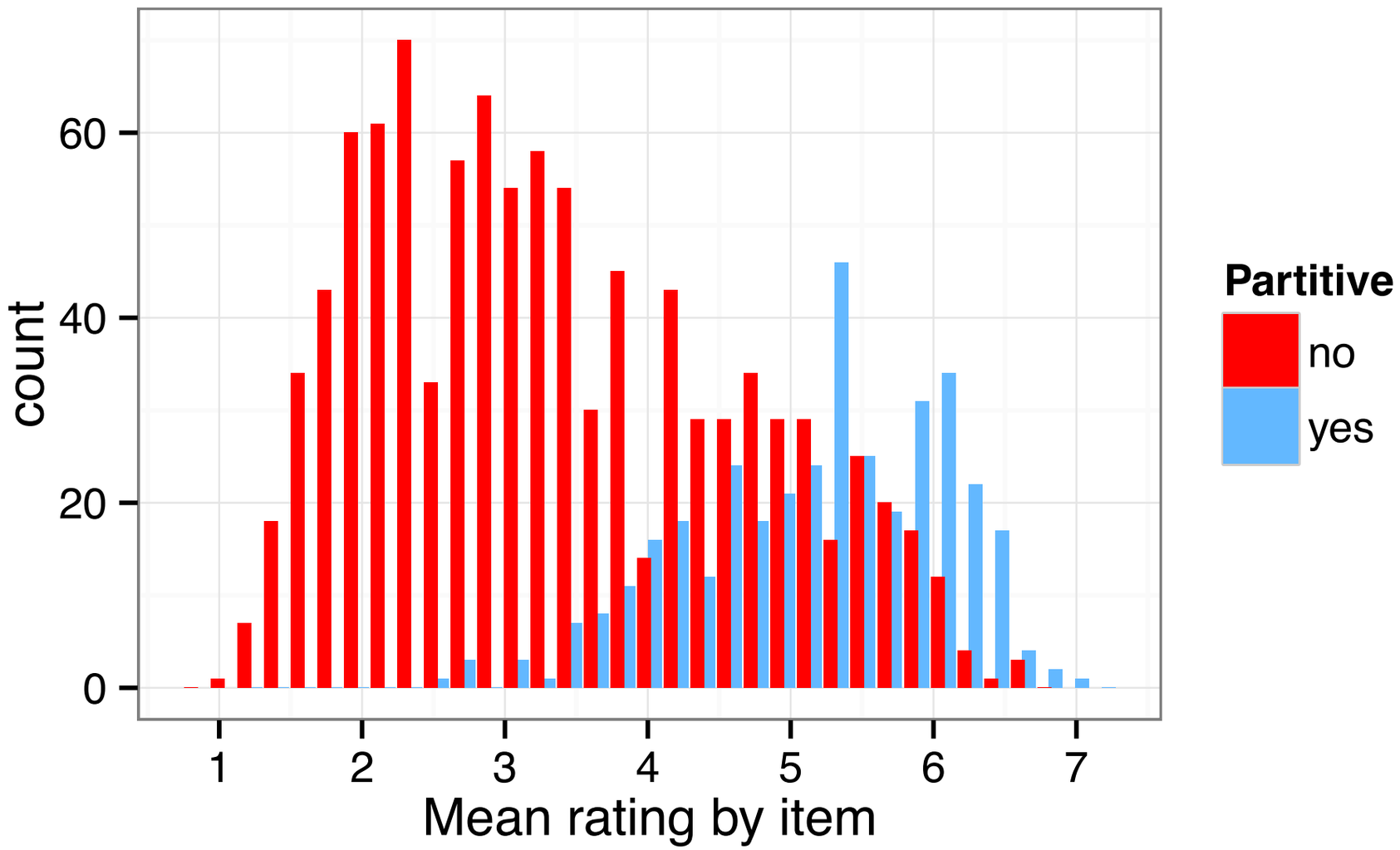
Analysis

2. If there is variation among implicatures, is it random or systematic?



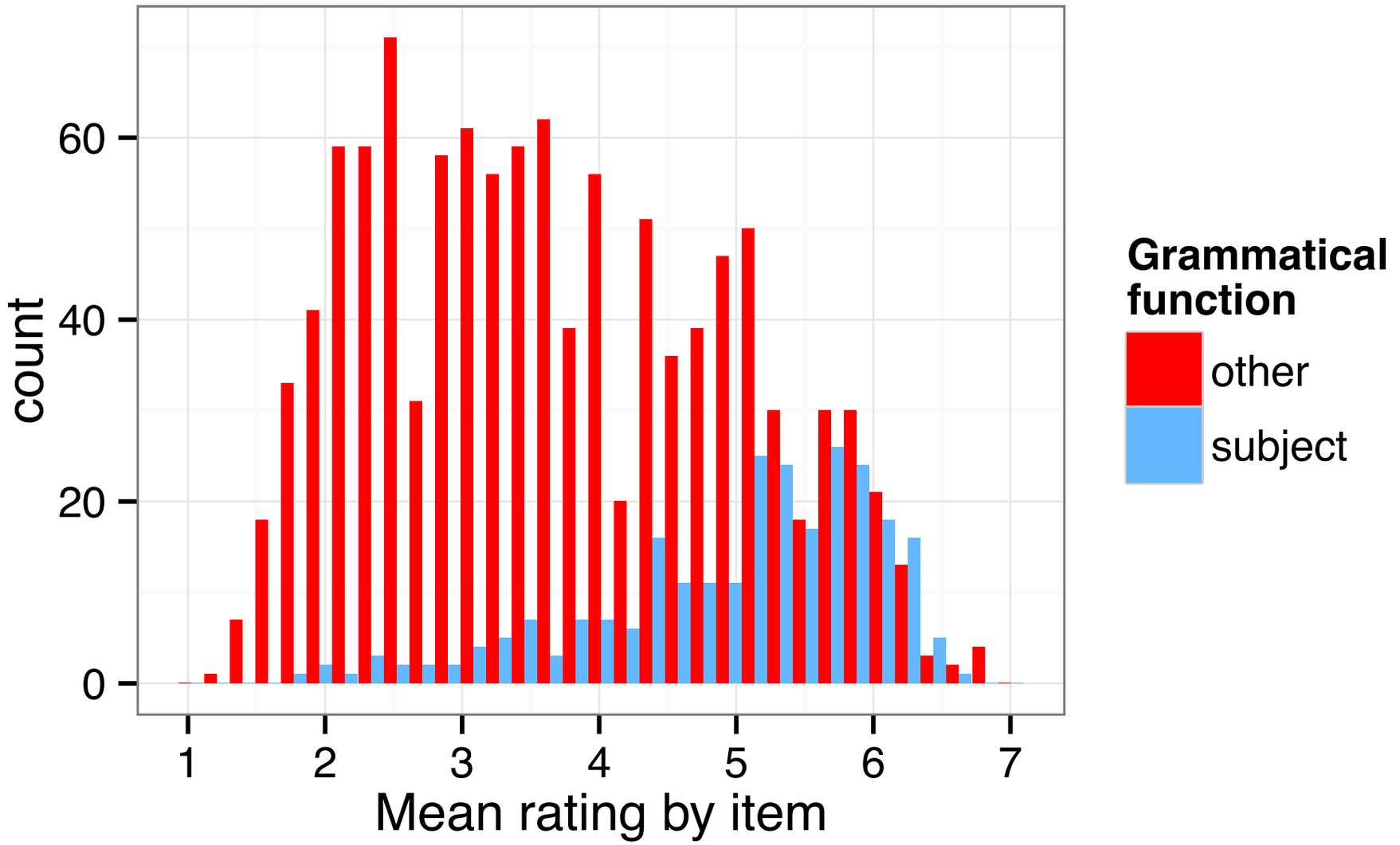
Analysis

2. If there is variation among implicatures, is it random or systematic?



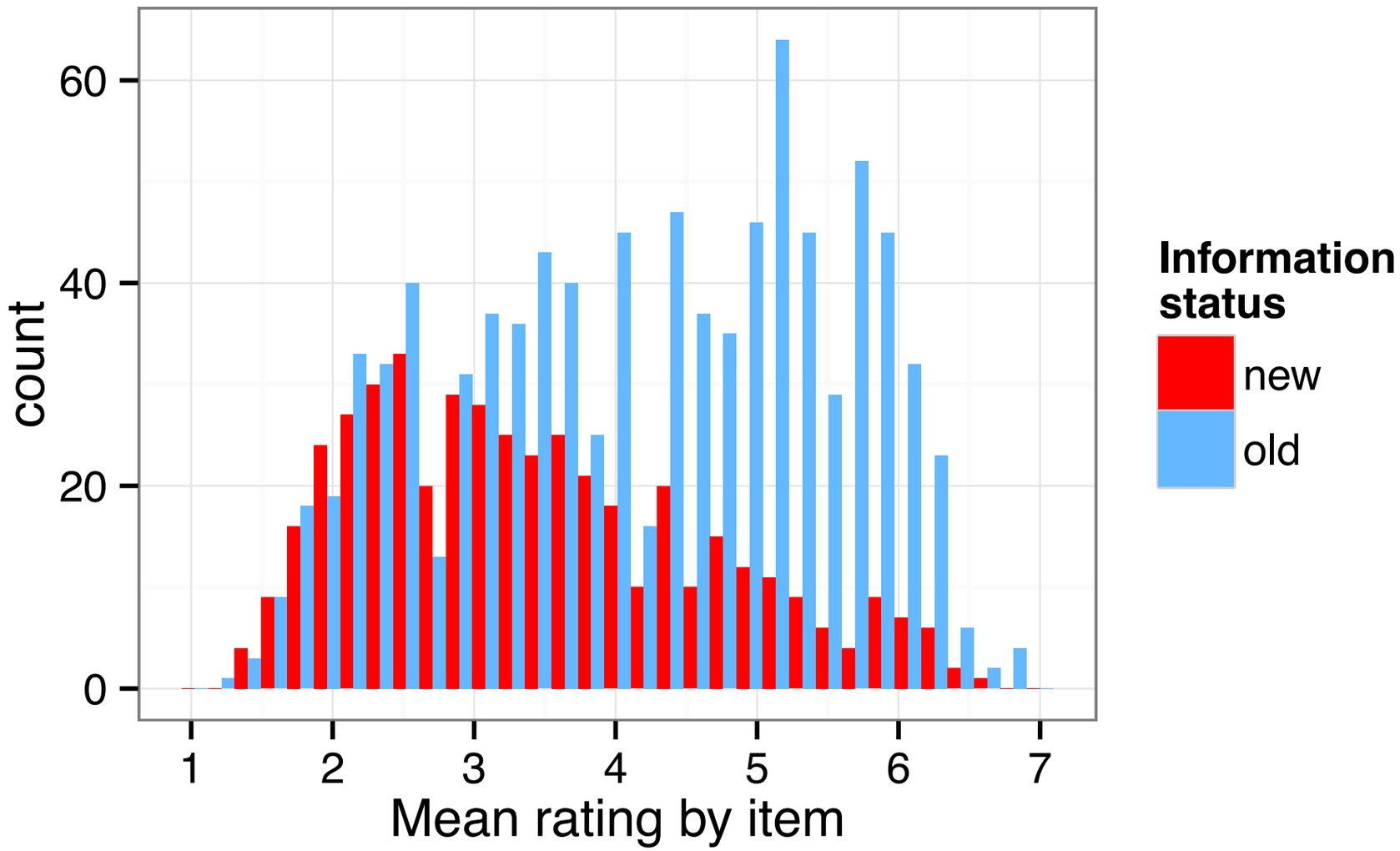
Analysis

2. If there is variation among implicatures, is it random or systematic?

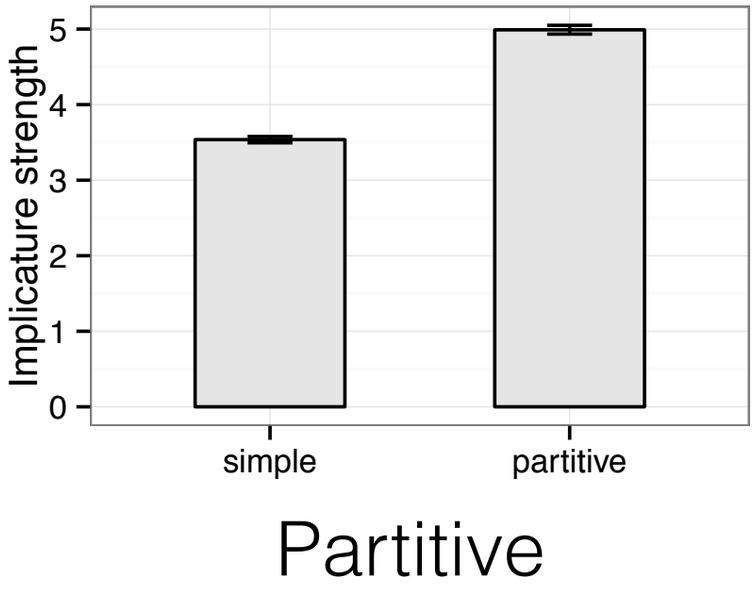


Analysis

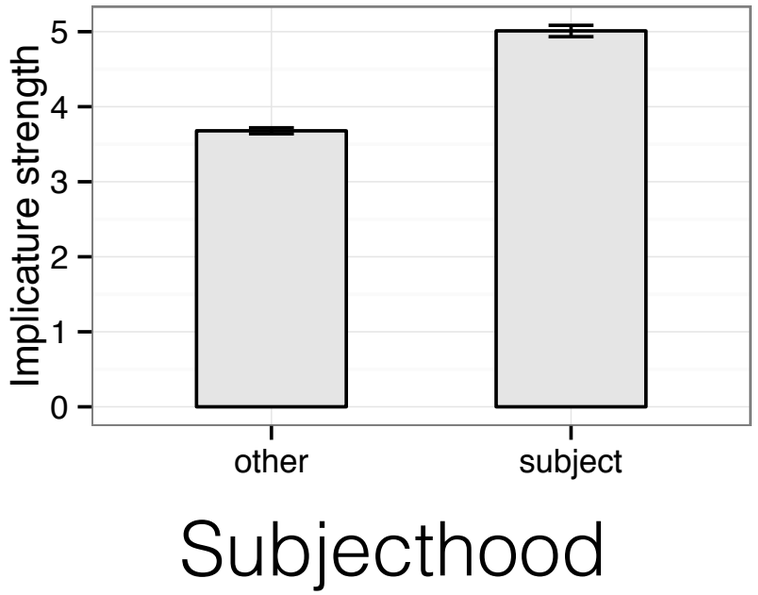
2. If there is variation among implicatures, is it random or systematic?



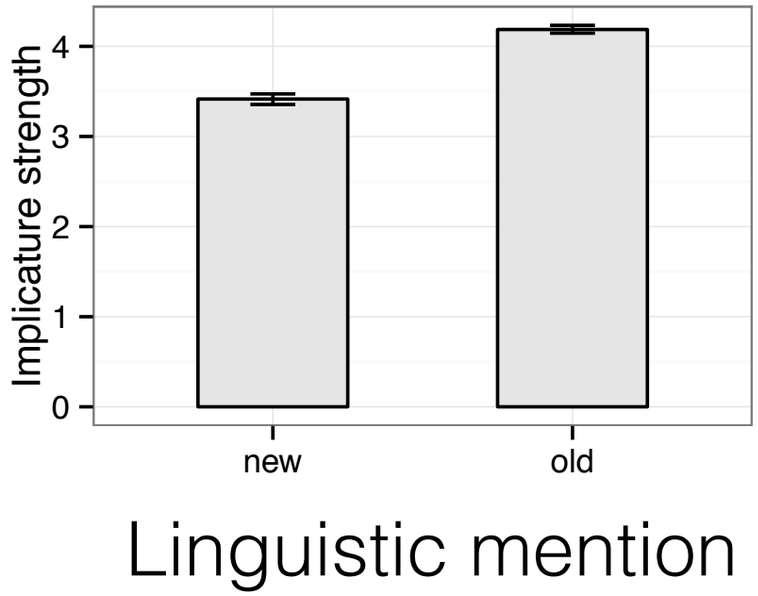
Analysis



Partitive



Subjecthood



Linguistic mention

Is it all the same effect? Eg, discourse accessibility?
Or are the effects independent?

Mixed effects linear regression

$$\text{Rating}_{ij} = \beta_0 + \beta_1 \text{Partitive}_{ij} + \beta_2 \text{GrammaticalFunction}_{ij} + \beta_3 \text{InfoStatus}_{ij} + b_j + \epsilon_{ij}$$

noise ~ $\mathcal{N}(0, \sigma_\epsilon)$

by-participant differences ~ $\mathcal{N}(0, \sigma_b)$

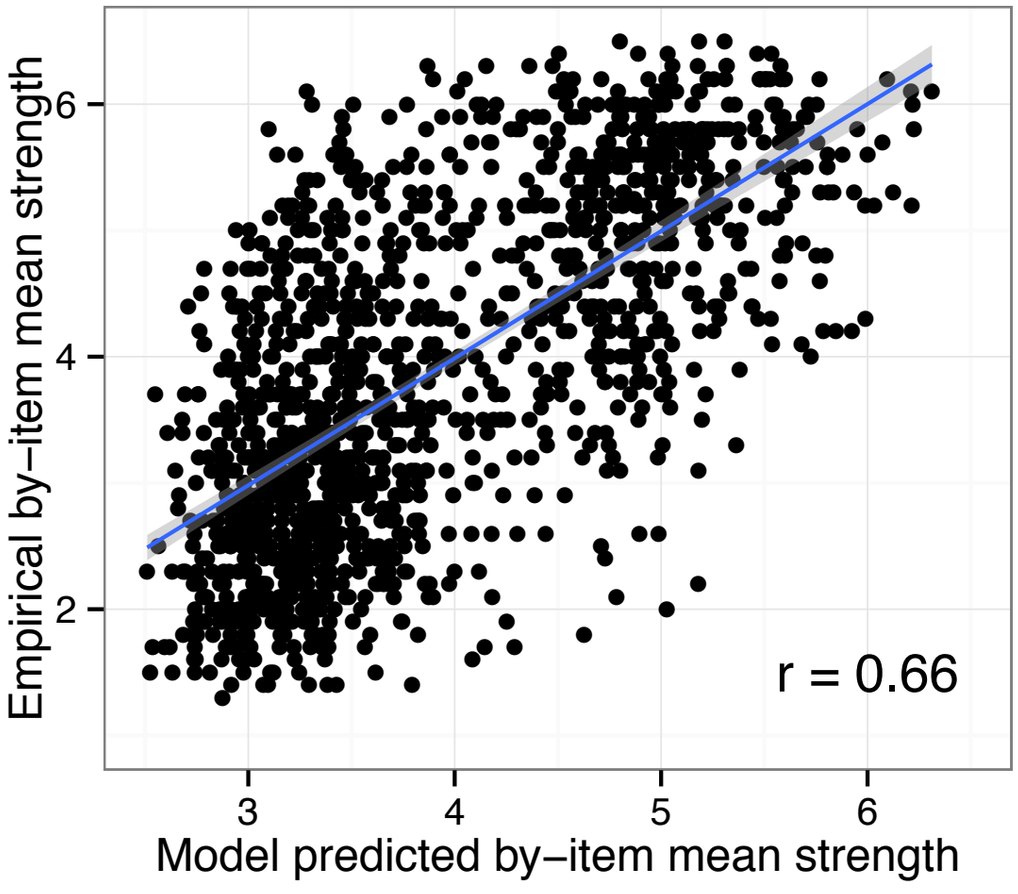
```
m = lmer(Rating ~ cPartitive + cGrammaticalFunction +
    cInfoStatus + (1|workerid), data=centered)
summary(m)
```

```
> summary(m)
Linear mixed model fit by REML
Formula: Rating ~ cPartitive + cGrammaticalFunction +
cInfoStatus + (1 | workerid)
   Data: centered
   AIC   BIC logLik deviance REMLdev
 56405 56450 -28197    56375   56393
Random effects:
 Groups   Name        Variance Std.Dev.
 workerid (Intercept) 0.47074  0.68611
 Residual             3.55331  1.88502
Number of obs: 13630, groups: workerid, 243

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          3.96828    0.04989   79.55
cPartitive           1.16780    0.03861   30.25
cGrammaticalFunction 0.85315    0.04396   19.41
cInfoStatus          0.41245    0.03564   11.57
```

Model evaluation

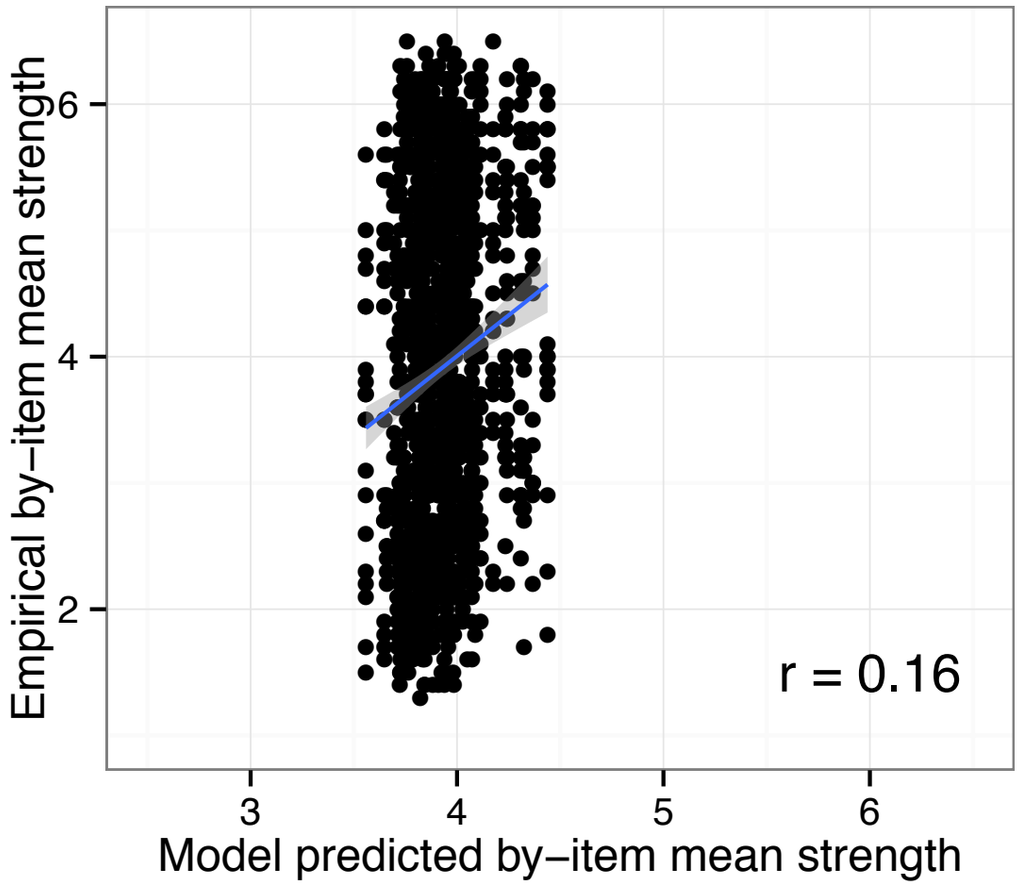
Full model



$$R^2_{\text{marginal}} = .16$$

$$R^2_{\text{conditional}} = .27$$

Subject variability



$$R^2_{\text{marginal}} = 0$$

$$R^2_{\text{conditional}} = .09$$

Conclusions

1. Do scalar implicatures from *some* to *not all* constitute a homogeneous class of inferences?

No.

2. If there is variation among implicatures, is it random or systematic?

The variation is systematic: implicature strength is dependent on various contextual features. But there is residual variation to be explained!

Tools used

- extracting data from corpus

  tgrep2 / TDT

- setting up mturk experiment

  javascript / HTML / mturk command-line tools

- data analysis & visualization

  R (especially lmer and ggplot)

- general pre- and post-processing

  python / bash

Tools used

tomorrow

- extracting data from corpus

  tgrep2 / TDT

- setting up mturk experiment

  javascript / HTML / mturk command-line tools

- data analysis & visualization

  R (especially lmer and ggplot)

- general pre- and post-processing

  python / bash

Tools used

tomorrow

- extracting data from corpus

  tgrep2 / TDT

Wednesday

- setting up mturk experiment

  javascript / HTML / mturk command-line tools

- data analysis & visualization

  R (especially lmer and ggplot)

- general pre- and post-processing

  python / bash

Tools used

tomorrow

- extracting data from corpus

  tgrep2 / TDT

Wednesday

- setting up mturk experiment

  javascript / HTML / mturk command-line tools

- data analysis & visualization

  R (especially lmer and ggplot)

Thursday/Friday

- general pre- and post-processing

  python / bash

Tomorrow

- TGrep2 tutorial
- hands-on project intro: the projection behavior of factive verbs Beaver 2010; Tonhauser 2016; Simons et al to appear

Website:
https://thegricean.github.io/esslli2016_corpuspragmatics

Github:
https://github.com/thegricean/esslli2016_corpuspragmatics