How good are leading theories of bridge verbs?

An experimental evaluation

Nick Huang[1], Diogo Almeida[2], Jon Sprouse[2]

National University of Singapore[1], NYU Abu Dhabi[2]

WCCFL 2022

Starting point: restrictions on wh-extraction

1. *Who ate the sandwich and pickles?*
2. **<u>What</u> did Jo eat the sandwich and _ ?*

A standard analysis of **island constraints**:

Wh-extraction is subject to structural (syntactic) constraints.

(Ross 1967, among many others)

But some restrictions are not as amenable to a syntactic analysis

3. *Who did Kim say that Jo saw _?*
4. *??Who did Kim stammer that Jo saw _?*

Say: bridge verb
Stammer: non-bridge

Difference is lexical, not syntactic.

Call this variation in acceptability “**bridge effects**”.

Why do bridge effects exist?

Dean 1967; Erteschik-Shir 1973; Ambridge & Goldberg 2008; Kothari 2008; Dąbrowska 2008, 2013; Liu et al. 2019, 2021; Richter & Chaves 2020

Three non-syntactic approaches for understanding bridge effects

3. *Who did Kim* ***say*** *that Jo saw _?*
4. ??*Who did Kim* ***stammer*** *that Jo saw _?*

1. **Information structure**

(Erteschik-Shir 1973; Ambridge & Goldberg 2008, etc.)

2. **Frequency**

(Kothari 2008; but see Liu et al. 2019; 2021 and Richter & Chaves 2020)

3. Prototype effects: **Semantic similarity** to *say/think*

(Dąbrowska 2008; 2013; also Ambridge & Goldberg 2008, etc.)

Success with bridge effects → a stronger case for non-syntactic theories of wh-extraction constraints

	Island effects	Bridge effects
Information structure account "No extraction from non-dominant/ backgrounded constituents"	✓	✓
Syntactic accounts	✓	??

Further consequences for autonomy of syntax, poverty of the stimulus, learnability, etc.

This talk: An empirical evaluation of these theories of bridge effects

- Overview of existing theories
    1. Information structure
    2. Frequency
    3. Semantic similarity
- The logic of our experiments
- Results and discussion
- Conclusion: **we need empirically stronger theories.**

Bridge effects = relative acceptability

5. *What did Kim* ***say / stammer*** *[that Jo saw _]?* “Long” extraction from clause

6. *Who _* ***said / stammered*** *[that Jo saw the robber]?* “Short” extraction

Bridge effects

= How much worse is long extraction, i.e. extracting from the complement clause?

= Penalty for long extraction = acceptability of (6) - acceptability of (5)

Theory 1. Information structure

- “No extraction from **non-dominant** constituents.” (e.g. Erteschik-Shir 1973)
- “No extraction from **backgrounded** clauses” (Ambridge & Goldberg 2008)

7. *Kim* ***said*** *[that Jo saw the robber].*

    Verbs like *say* foreground/focus the complement clause.

8. *Kim* ***stammered*** *[that Jo saw the robber].*

    Stammer draws attention to the act of stammering, not the clause; the clause is backgrounded.

Theory 2. Frame frequency

Bridge effects track how often a verb takes a finite complement clause.

9.	*What did Kim **say** that Jo saw?*	*Say*+clause very frequent
10.	*??What did Kim **stammer** that Jo saw?*	*Stammer*+clause rare

Independent psycholinguistic evidence that low-frequency structures create processing difficulties. (e.g. Hale 2001; Levy 2008)

Kothari 2008, Dąbrowska 2013, but see Liu et al. 2019, Richter & Chaves 2020

Theory 3. Semantic similarity / prototype effects

We hear many instances of extraction with *say* or *think*, e.g.:

*What did you **say** they will do?*
*Where do you **think** they went?*

For language processing purposes, we create “templates” based on prototypical questions.

Say template: $\boxed{\text{WH}}$ *do you say* $\boxed{\text{S-GAP}}$?

Think template: $\boxed{\text{WH}}$ *do you think* $\boxed{\text{S-GAP}}$?

(Replace $\boxed{}$ with a suitable constituent.)

Dąbrowska 2008, 2013; Verhagen 2005: see also Ambridge & Goldberg 2008

Theory 3. Semantic similarity / prototype effects

9. *What did Kim **say** that Jo saw?*

→ Use the existing *say* template.

10. ??*What did Kim **stammer** that Jo saw?*

→ No *stammer* template; modify existing templates instead.

→ Bridge effects reflect cost of modifying a template, which **decreases** with **semantic similarity** to *say* / *think*.

Dąbrowska 2008, 2013; Verhagen 2005: see also Ambridge & Goldberg 2008

Problem #1: No clear consensus from prior experiments testing these theories

E.g.

- Ambridge & Goldberg 2008, Dąbrowska 2013: experimental results **supporting information structure theory**.
- Liu et al. 2021: **failed to replicate results**.

depict elect suppose suggest profess endorse spot smell emphasize prefer dislike find proclaim approve disce
guarantee implore request deplore investigate hate deem ordain pretend presuppose advise swear say demand cha
praise criticize illustrate dread intercept assess stress attest pronounce wish specify recommend showcase make out c
out warrant postulate believe decree exclaim fear yelp envisage admire philosophize generalize recollect desp
estimate murmur preach deduce feel claim anticipate hope counsel sense decide announce hint reckon confirm obse

Problem #2: data quality

English has hundreds of verbs that appear with finite clauses.

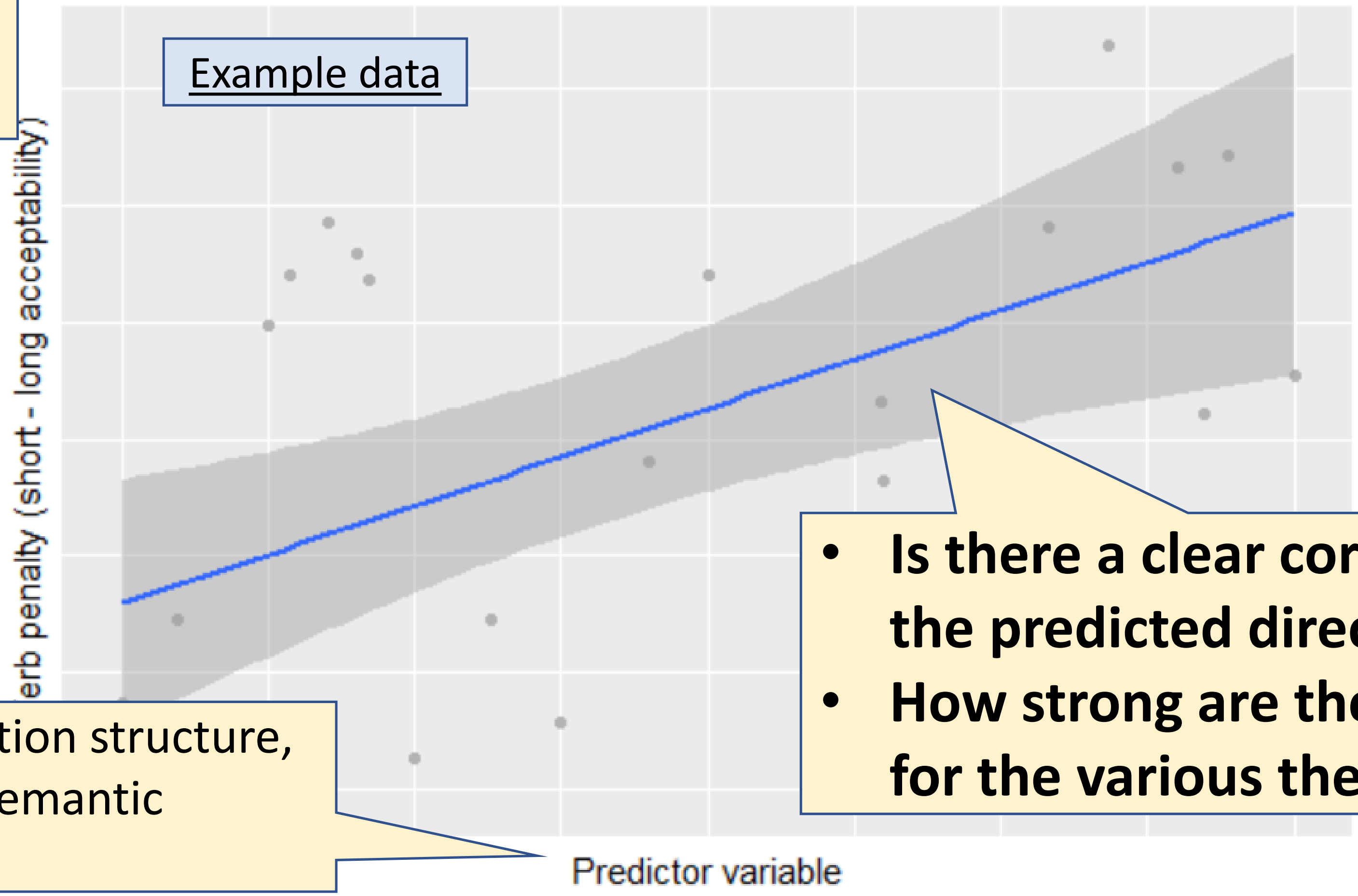
Current generalizations / experiments have small samples (~12-75 verbs).

Are these findings robust?

testify let slip justify figure jest stammer out clarify articulate trumpet predict disclose forgive report flaunt reason c
propose vote communicate submit moan gabble gesture cheer croon tell recite desire insist quip fuss judge comme
convince take for granted consent bark recount cry rave ramble twitter grieve sign urge teach notice burst out rese
conceal screech uncover put out brag beg reply find out stammer love doubt relish fume blubber speculate narrate pie

Our contribution: exhaustive (640 verbs), experimental evaluation of these theories

Bridge effects / relative acceptability



e.g. information structure, frequency, semantic similarity

- **Is there a clear correlation in the predicted direction?**
- **How strong are the correlations for the various theories?**

Quantifying bridge effects

5. *What did Kim* ***say / stammer*** *[that Jo saw _]?* “Long” extraction from clause

6. *Who _* ***said / stammered*** *[that Jo saw the robber]?* “Short” extraction

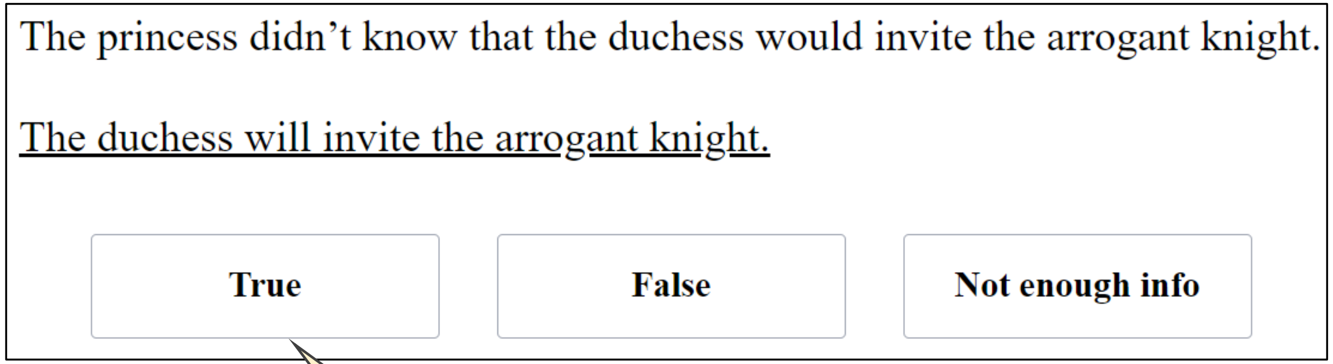
- Collect 60 sets of ratings per verb, for sentences like (5) and (6) on Amazon Mechanical Turk/CloudResearch (~9,600 participants).
- Calculate relative acceptability (“penalty”) for each verb.
- Analyse only the 484 verbs where “short extraction” sentences (6) are relatively OK (z-scored acceptability > 0)

We adopted the predictor measure proposed by advocates of each theory

1. **Information structure**

"Negation test" (4,800 AMT participants).

(Ambridge & Goldberg 2008)


The princess didn't know that the duchess would invite the arrogant knight.

The duchess will invite the arrogant knight.

True | False | Not enough info

True: more backgrounded

False: less backgrounded

2. **Frequency**

Frequency of Verb-*that* combinations + clause bias, from COCA

3. **Semantic similarity**

Off-the-shelf word-embeddings from Latent Semantic Analy
English Wikipedia; WordNet word classes. (Ştefănescu et al. 2014;

- Calculate **cosine similarity** (word-embeddings) and **hierarchical distance** (WordNet) to *say* and *think*

We adopted the predictor measure proposed by advocates of each theory

1. **Information structure**

"Negation test".

(Ambridge & Goldberg 2008)

2. **Frequency**

Frequency of Verb-*that* combinations + clause bias, from COCA. (Davies 2020)

3. **Semantic similarity**

Off-the-shelf word-embeddings from Latent Semantic Analysis and Global Vectors applied to English Wikipedia; WordNet word classes. (Ştefănescu et al. 2014; Fares et al. 2017; Fellbaum 1998)

- Calculate **cosine similarity** (word-embeddings) and **hierarchical distance** (WordNet) to *say* and *think*

We adopted the predictor measure proposed by advocates of each theory

1. **Information structure**

"Negation test".

(Ambridge & Goldberg 2008)

2. **Frequency**

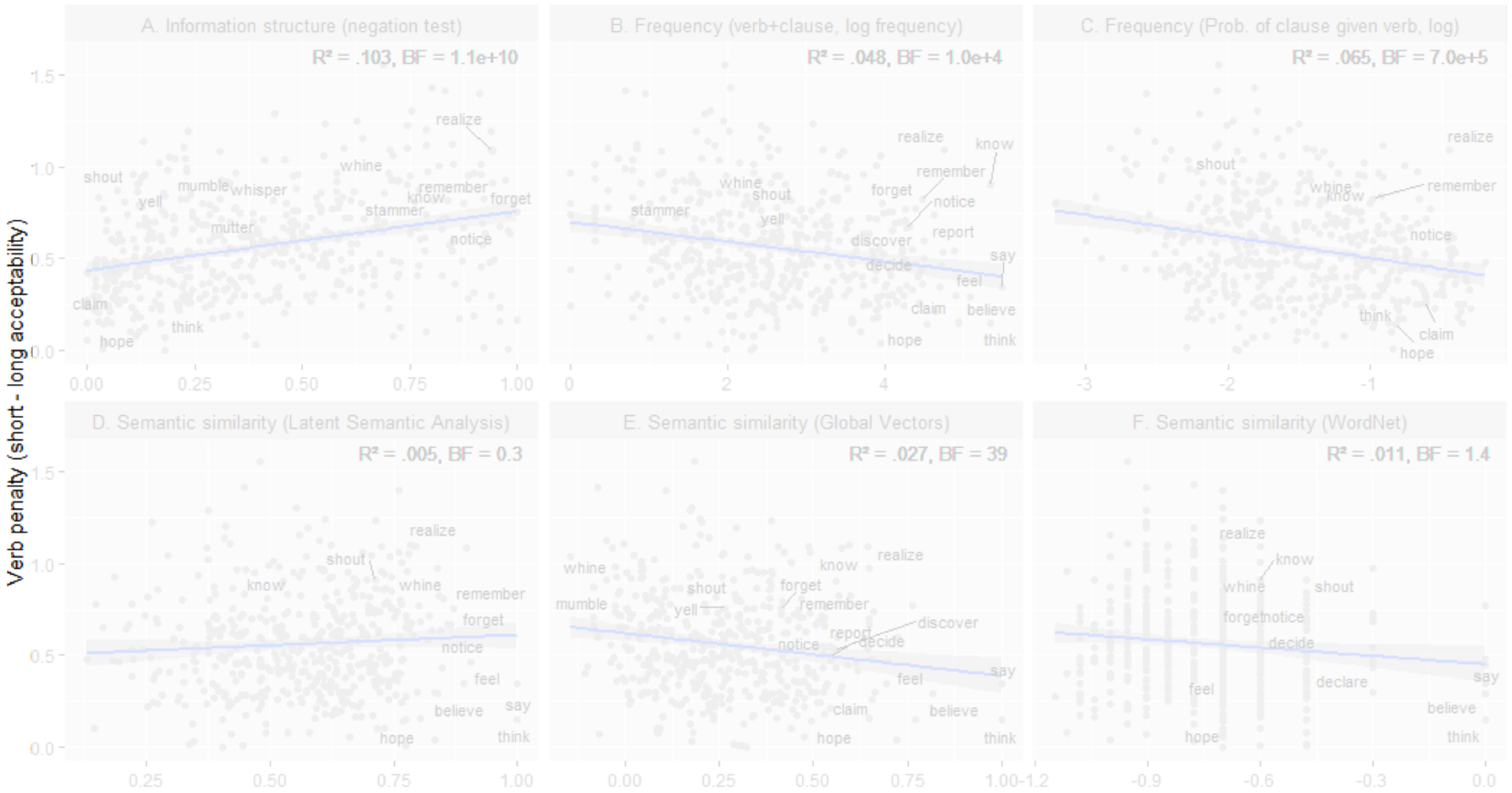
Frequency of Verb-*that* combinations + clause bias, from COCA. (Davies 2020)

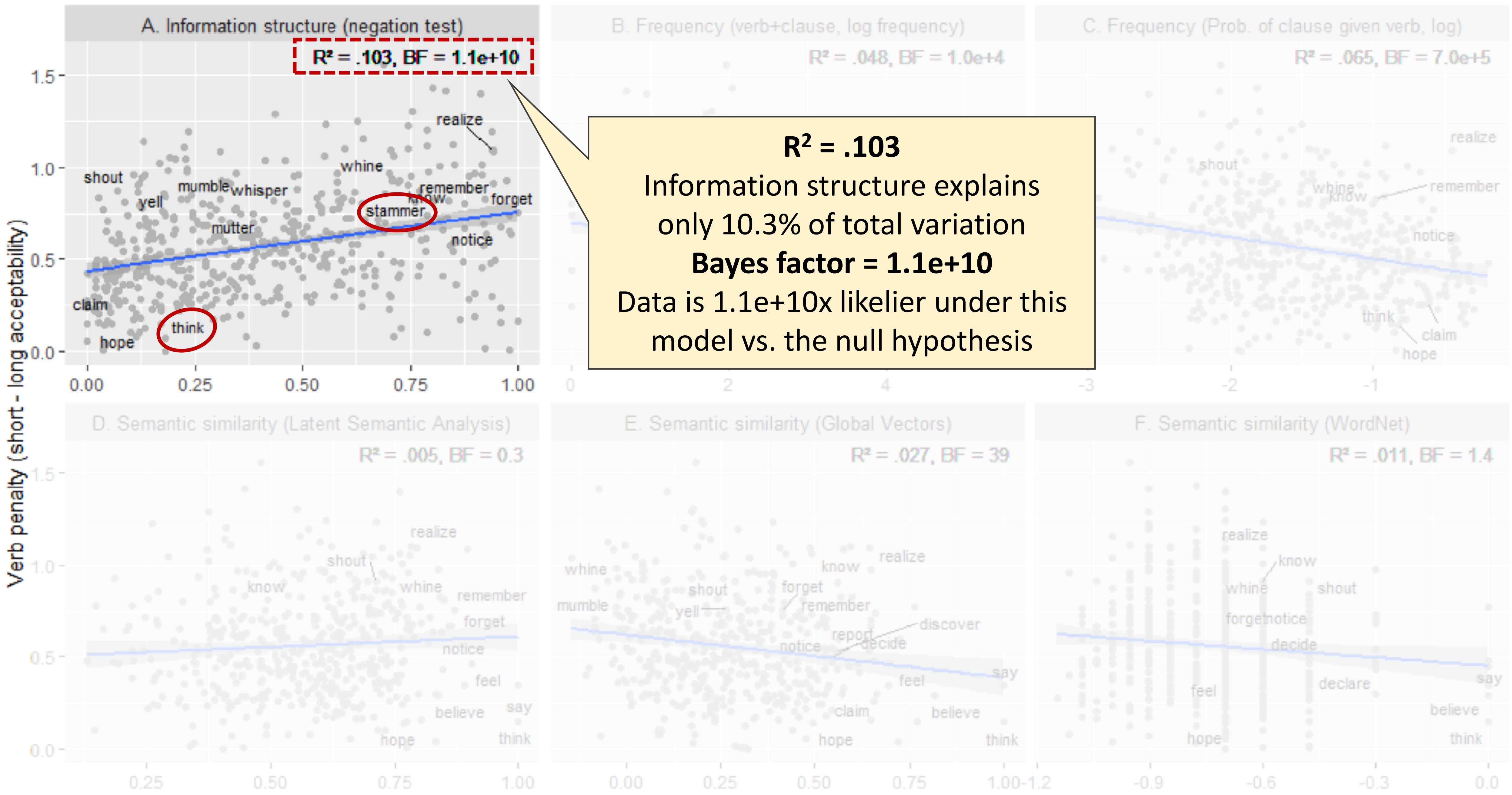
3. **Semantic similarity**

Off-the-shelf word-embeddings from Latent Semantic Analysis and Global Vectors applied to English Wikipedia; WordNet word classes. (Ştefănescu et al. 2014; Fares et al. 2017; Fellbaum 1998)

- Calculate **cosine similarity** (word-embeddings) and **hierarchical distance** (WordNet) to *say* and *think*

Six predictor variables in total


A. Information structure (negation test)
R² = .103, BF = 1.1e+10
B. Frequency (verb+clause, log frequency)
R² = .048, BF = 1.0e+4
C. Frequency (Prob. of clause given verb, log)
R² = .065, BF = 7.0e+5
D. Semantic similarity (Latent Semantic Analysis)
R² = .005, BF = 0.3
E. Semantic similarity (Global Vectors)
R² = .027, BF = 39
F. Semantic similarity (WordNet)
R² = .011, BF = 1.4
Verb penalty (short - long acceptability)

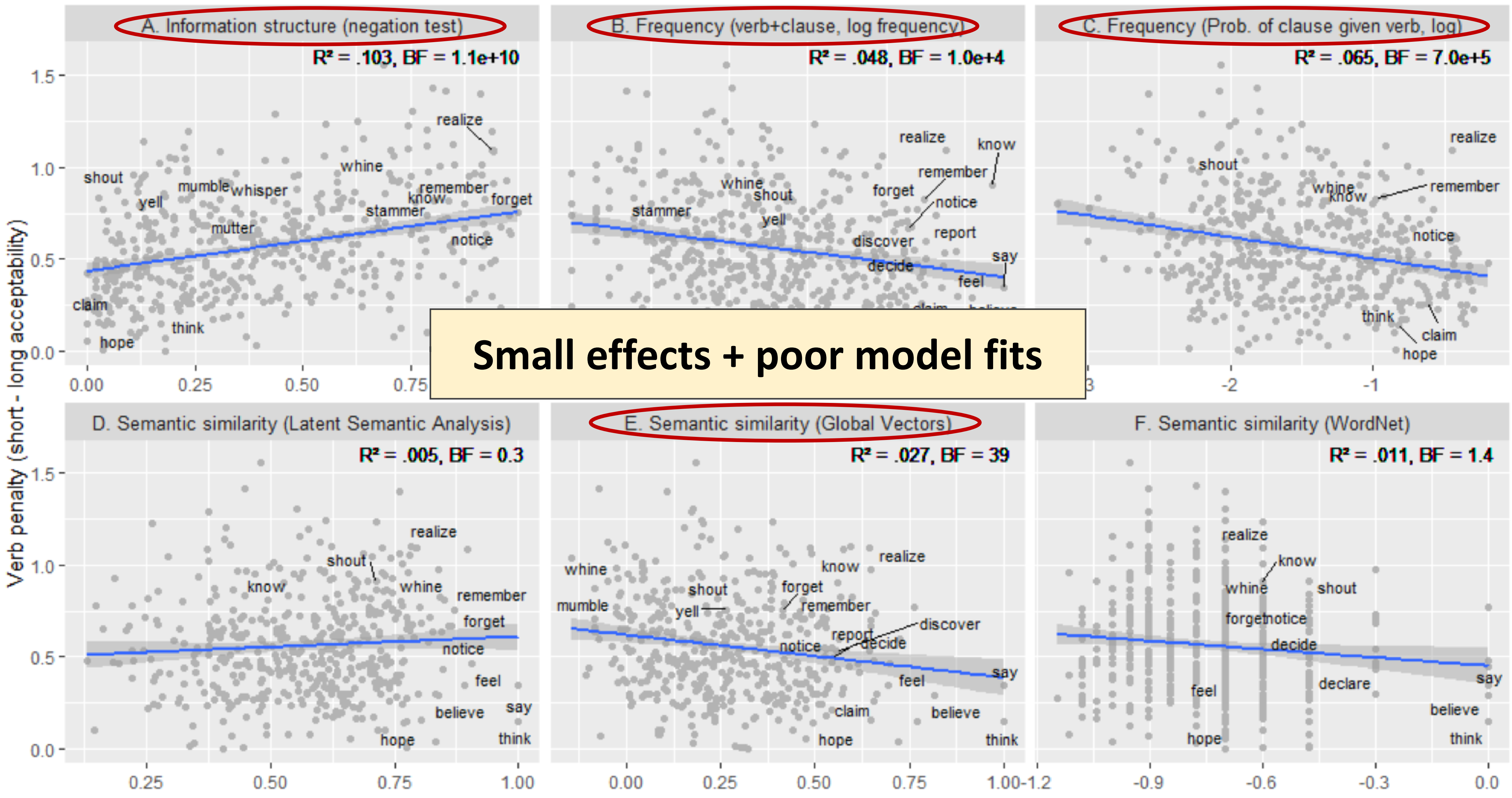


R^2 = .103

Information structure explains only 10.3% of total variation

Bayes factor = 1.1e+10

Data is 1.1e+10x likelier under this model vs. the null hypothesis

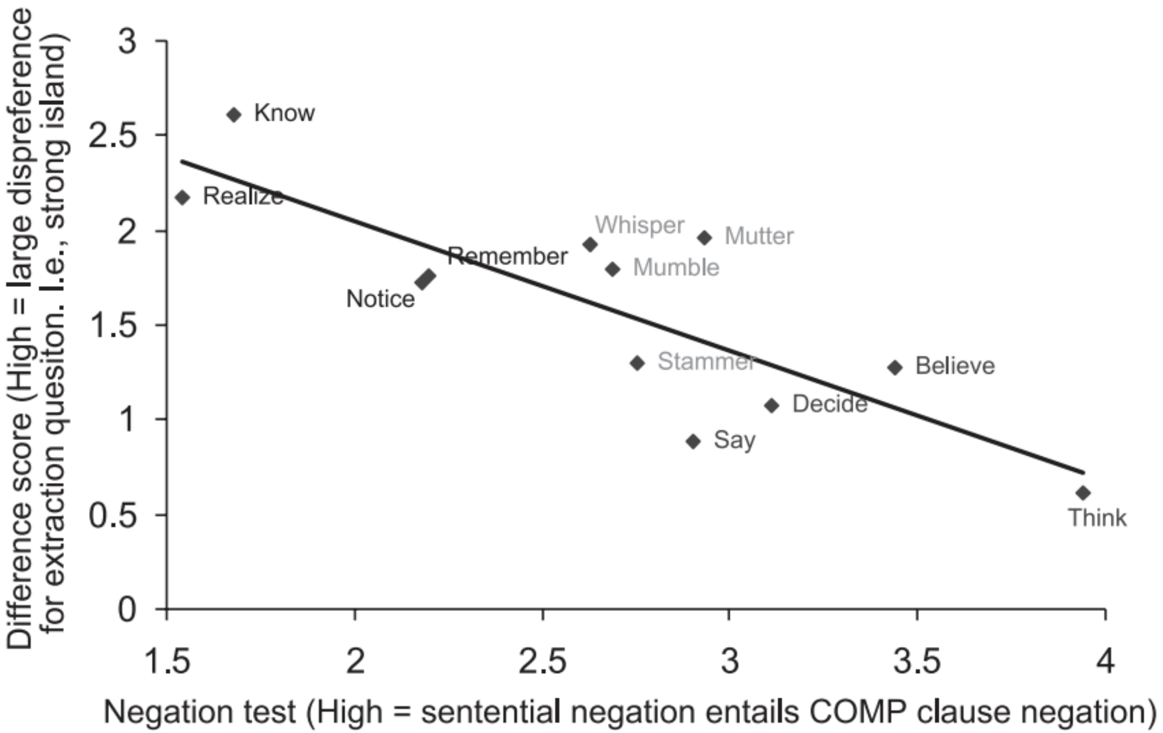


Small effects + poor model fits

Replicating Ambridge & Goldberg 2008: much clearer correlation on subset of 12 verbs

A&G 2008, fig. 3

$R^2 = .69$



"Foregroundedness"

Same 12 verbs

$R^2 = .48$

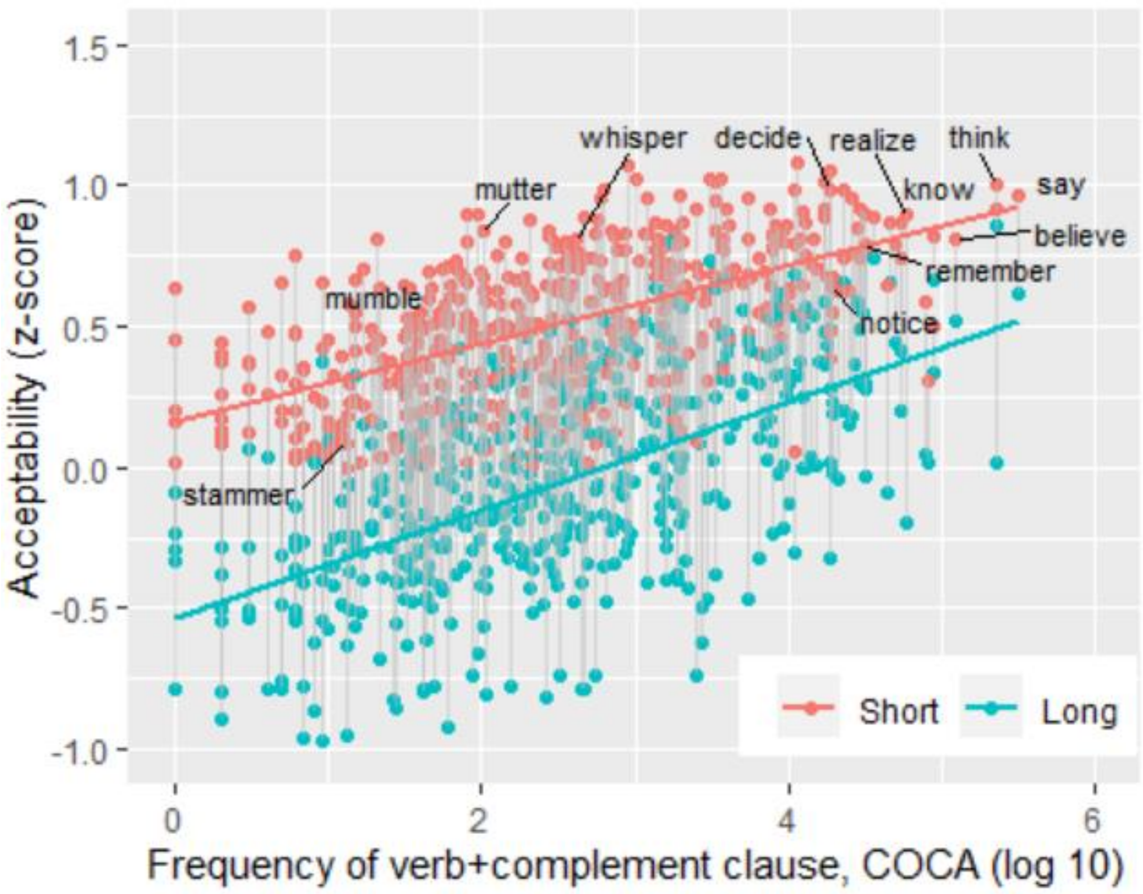
Re-examining a recent claim that there are no bridge effects

Liu et al. (2019, 2021): Relative acceptability doesn’t vary with frequency.

- Model acceptability for 48 verbs, with **only main effects** for **verb+clause frequency** and **short/long extraction.**

Not supported by an analysis of our full set of verbs (right).

- We find **an interaction between frequency and extraction** (*b*=0.05, *t*=5.92, *p*<.01).
- Graphically represented by **non-parallel** trend lines in scatterplot.



Summarising

Prior experimental studies have verb sample issues.

Analysing a full set of verbs, we find that even the best-performing theory of bridge effects (information structure) is **empirically weak**.

→Factors like information structure, frequency, semantic similarity contribute to bridge effects, but unlikely to be the only driver of bridge effects.

Results call for better theories of bridge effects

Our inspection of our results suggest that **verb classes** matter: verbs that allow nonfinite complement clauses (*believe*/*expect* NP *to* VP; *claim to* VP) tend to have higher relative acceptability (point-biserial correlation = .40, $p<.01$).

Further questions:

1. Is the verb class fact due to verb **semantics**, **pragmatics**, or even **syntax**?
2. **Cross-linguistic variation**: Some languages lack long-distance wh-questions (e.g. Polish, some German varieties). Why?

See also prior discussion by Erteschik-Shir 1973; Ambridge & Goldberg 2008; Fodor 1992

Thank you

- Various aspects of this research was in part supported by a NUS Postdoctoral Fellowship.
- Earlier versions of this project was supported by NSF Grant #DGE-1449815.