Quantifying cronuts: Predicting the quality of blends

Constantine Lignos
Children's Hospital of Philadelphia

Hilary Prichard
University of Pennsylvania

http://lignos.org/blends

I. Introduction

Calling all innovators

- Linguistic innovation is hard

Calling all innovators

- We don't have to make an entirely new word happen
- **Anyone** can blend existing ones

Blends in action

BRO + ROMANCE

BROMANCE

Blends in action

FRIEND + ENEMY

FRENEMY

Blends in action

HORSE + CORGI

HORGI

Blends in action

FRIENDS + FAMILY

FRAMILY???

Some blends are better than others

Questions to answer:

1. What makes some blends better than others?
2. How can we predict which blends people will understand and like?

Our approach:

1. Collect ratings of blends
2. Build a model of what people do
3. Identify the predictors that matter
4. (In progress) Extend to rating new blends

II. An ontology of blends

A working definition

For the purpose of this study, a blend:

1. Is a linear combination of two source words
2. Uses overlap and/or truncation at the point of blending

Non-blends:

1. Compounds without truncation: *manspreading*
2. Libfixes: *work-aholic, gamer-gate, lumber-sexual*

Blend classes

1. *Complete overlap:* Source words overlap in output, all of each source word appears
   alcoholiday   *guesstimate*   *mathlete*
2. *Partial overlap*: Source words overlap in output, but not all sounds are preserved
   affluenza   *brony*   *facon*   *sext*   *shitticism*
3. *No overlap*: No segmental overlap, but some truncation at combination point
   cosplay   *sharknado*   *shotchka*   *zonkey*

Selecting items

Chosen from:

- Wikipedia portmanteau list
- Thurner portmanteau dictionary
- Listening for everyday occurrences

Excluded:

- Brand names
- Unclear analysis (*keytar, murse*)

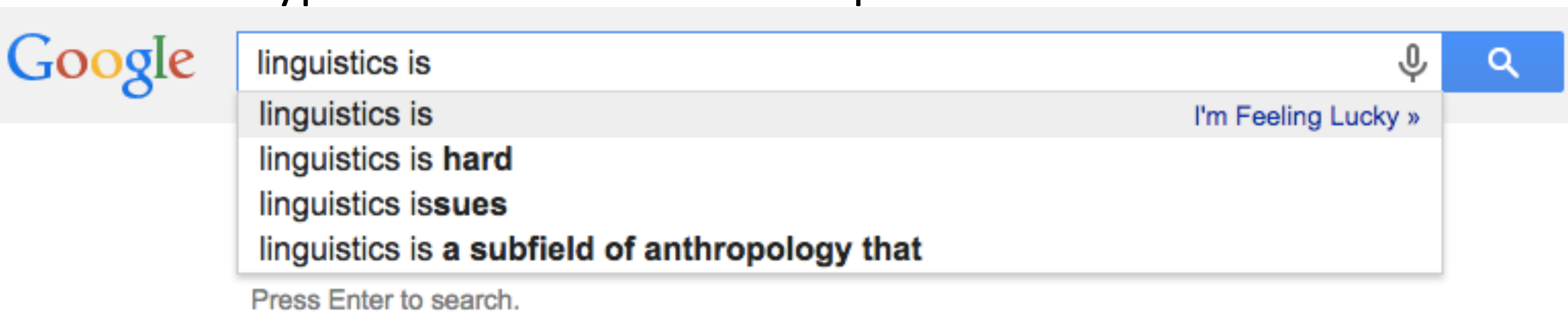
III. Quantifying blends

Defining the source-output relationship

1. Amount of phonological content present
    - e.g., Gries 2004
2. Phonological wellformedness
    - e.g., Kelly 1998
3. How easily the source words can be identified using the output content

Identifying source from partial content

- In speech processing: cohort effects (e.g., Marslen-Wilson 1987)
- In layperson's terms: autocomplete



Example*

affluence + influenza = affluenza

Segment content ratios (r1, r2):

affluen / affluence   fluenza / influenza

Identification probability (p1, p2):

p(affluence | affluen-)   p(influenza | -fluenza)

*All computations are over segments; orthography shown for convenience

Computing identification probability

(Fake) example:

language + debate = langbate

lang- = L AE NG
Competitors: *language* (.99), *languid* (.006), *languish* (.004)
p1 = .99

-bate = B EY T
Competitors: *bait* (.45)*, debate* (.43)*, masturbate* (.06),
rebate (.01)...
p2 = .43

Pronunciations from CMUdict (modified), SUBTLEX-US frequencies

Some harder to quantify factors

- Orthographic disambiguation: *fauxhawk* helped by *x* when written (not confusable with *focus, folk*, etc.)
- Semantic restrictions: *cronut, labradoodle* helped by restriction on what could possibly be combined
- Phonological problems: *coatigan* creates flapping context, *rawnola* creates stress clash


DARK CHOCOLATE CHIP
Chococonut®
5 CHEWY GRANOLA BARS
12 g WHOLE GRAINS | SOURCE OF FIBRE
PER 35 g SERVING
5 BARS x 35 g NET WT. 175 g

Some *even harder* to quantify factors

- Stress/metrical structure: surely contribute to choice of output form among alternatives, but it's not straightforward
  - How much does the stress on *tornado* improve *sharknado*?
  - If syllable structure is respected, where'd the *d* go in *frenemy?*
- Plausibility: does the blend make any sense? What's a *mirthquake*?

IV. Results

Survey design

- Chose 88 attested blends that were likely to be understood but varied in apparent quality
- Participants (n=34) rated each blend on two scales:
  1. Understandability: Is it easy to understand what words make up this blend?
  2. Naturalness: Does this combination of words sound natural to you?
- Could answer: "Didn't understand" or on scale:
  Terrible Poor Fair Good Excellent
- Expected high correlation between understandability and naturalness; our interest was in the outliers

Best and worst blends

Most understandable:

Blend	Source words	Average rating (1-5)
mathlete	math + athlete	4.8
sexpert	sex + expert	4.8
guesstimate	guess + estimate	4.8

Least understandable:

fozzle	fog + drizzle	1.8
mizzle	mist + drizzle	2.3
brinkles	bed + wrinkles	2.3

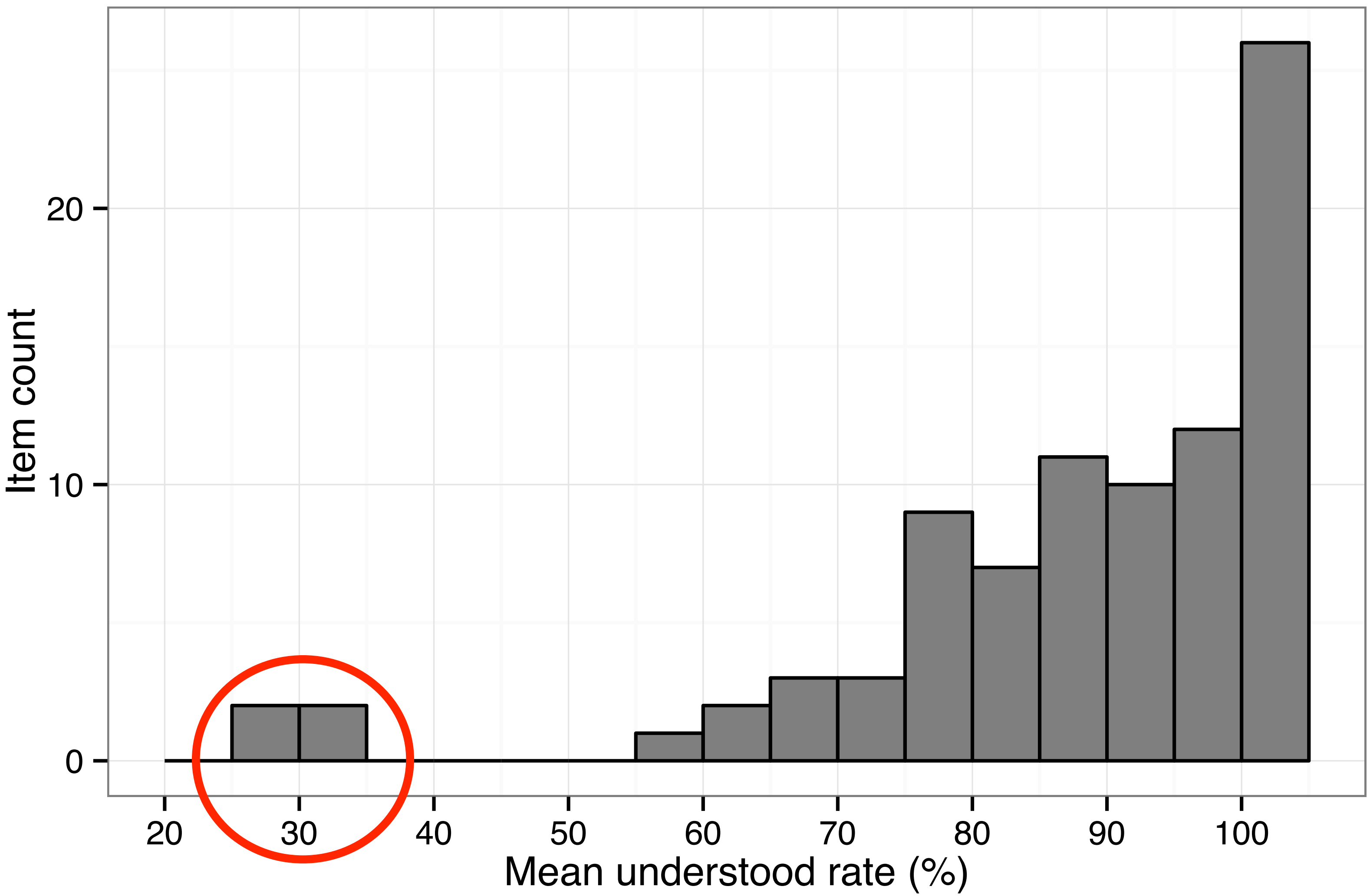
Most natural:

sexpert	sex + expert	4.8
mockumentary	mock + documentary	4.7
guesstimate	guess + estimate	4.7

Least natural:

dunch	dinner + lunch	2.1
nukemare	nuke + nightmare	2.2
rawnola	raw + granola	2.3

Item mean rates of understanding



Rarely-understood blends

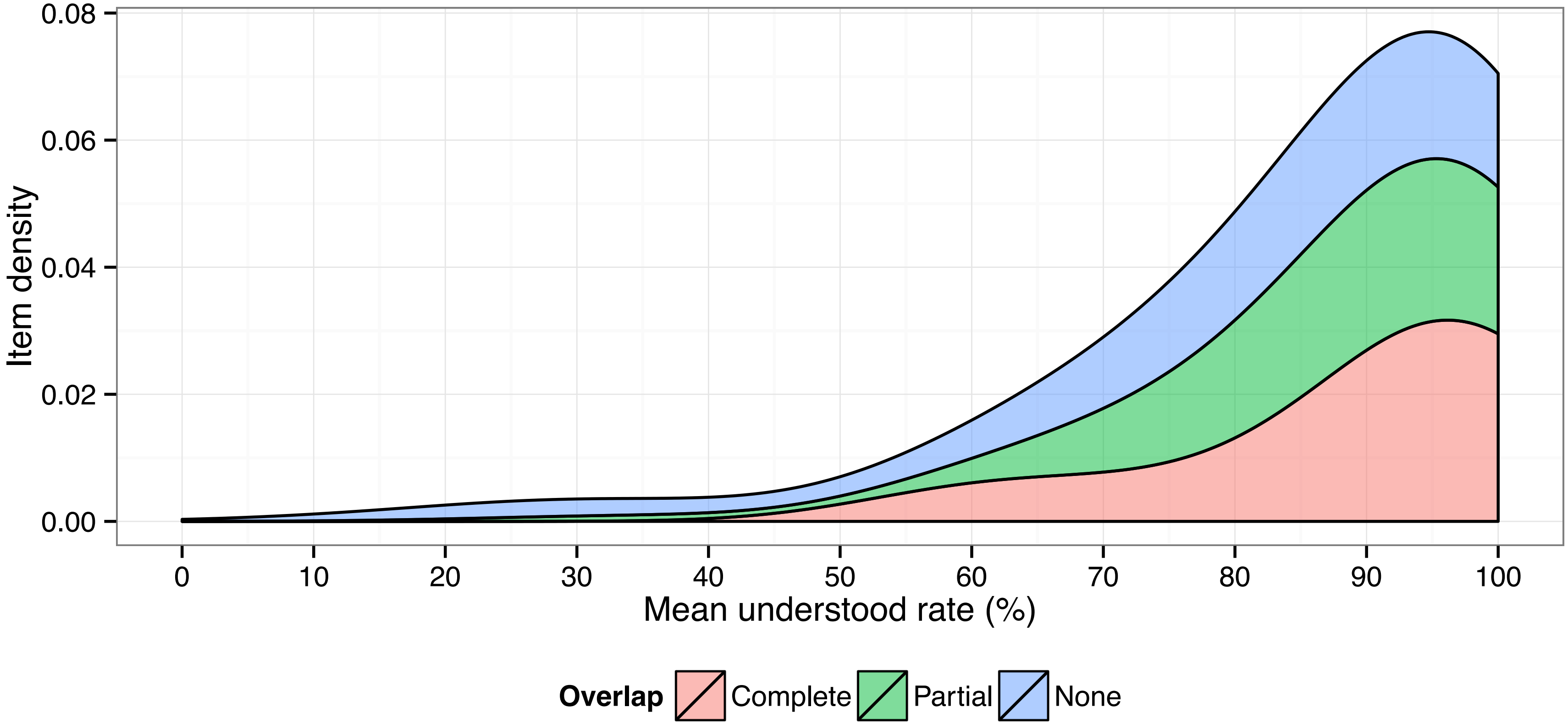
Least understood:

fozzle	fog + drizzle	26%
wonut	waffle + donut	28%
brinkles	bed + wrinkles	31%
mizzle	mist + drizzle	34%
wegotism	we + egotism	58%

Item types

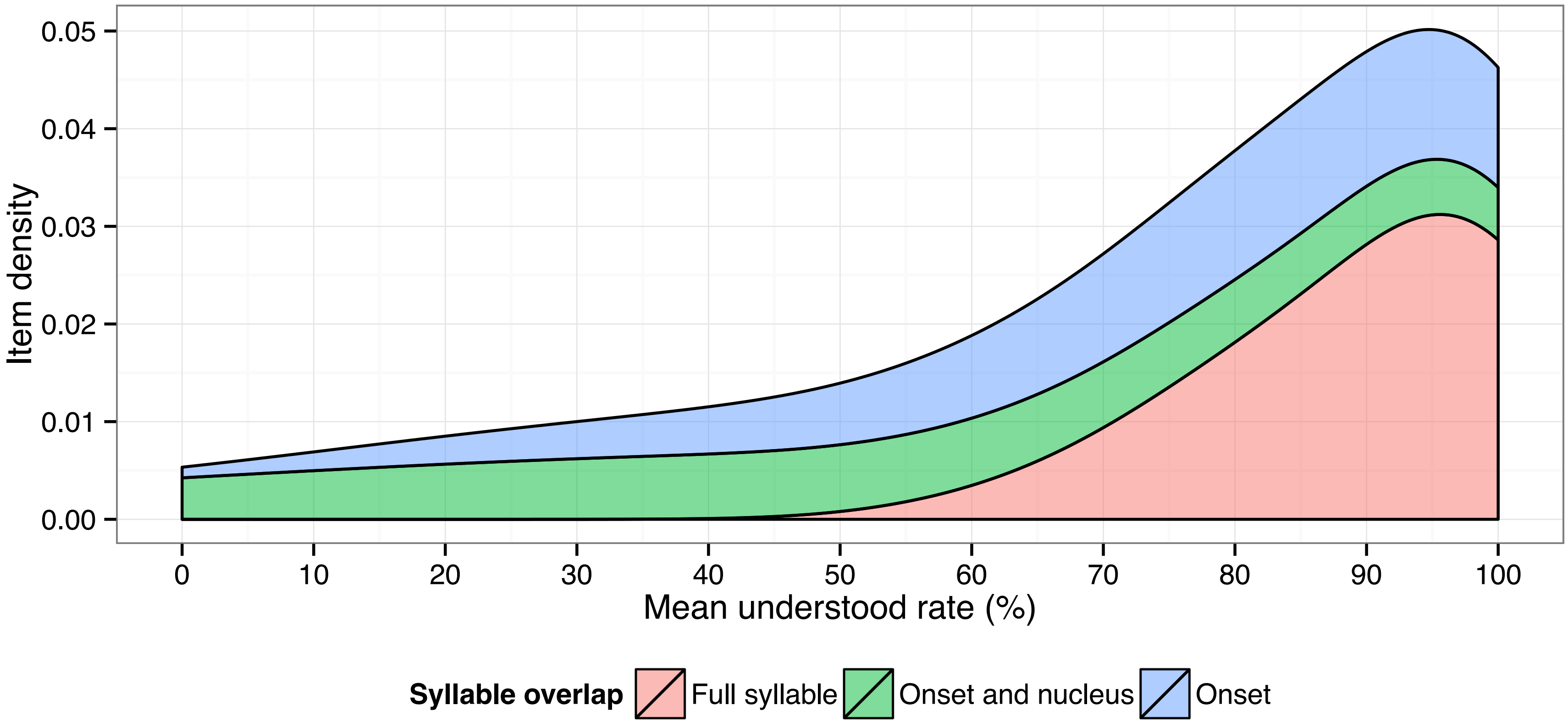
- Source word overlap (complete, partial, or none)
  - Hypothesis: complete overlap leads to the best blends
- First syllable overlap
  - Hypothesis: if there isn't enough of the syllable structure of the first word, it's hard to recover (above and beyond what segments tell us)
  - Levels of first syllable of first word present:
    - Onset
    - Onset and nucleus
    - Full first syllable

Overlap type



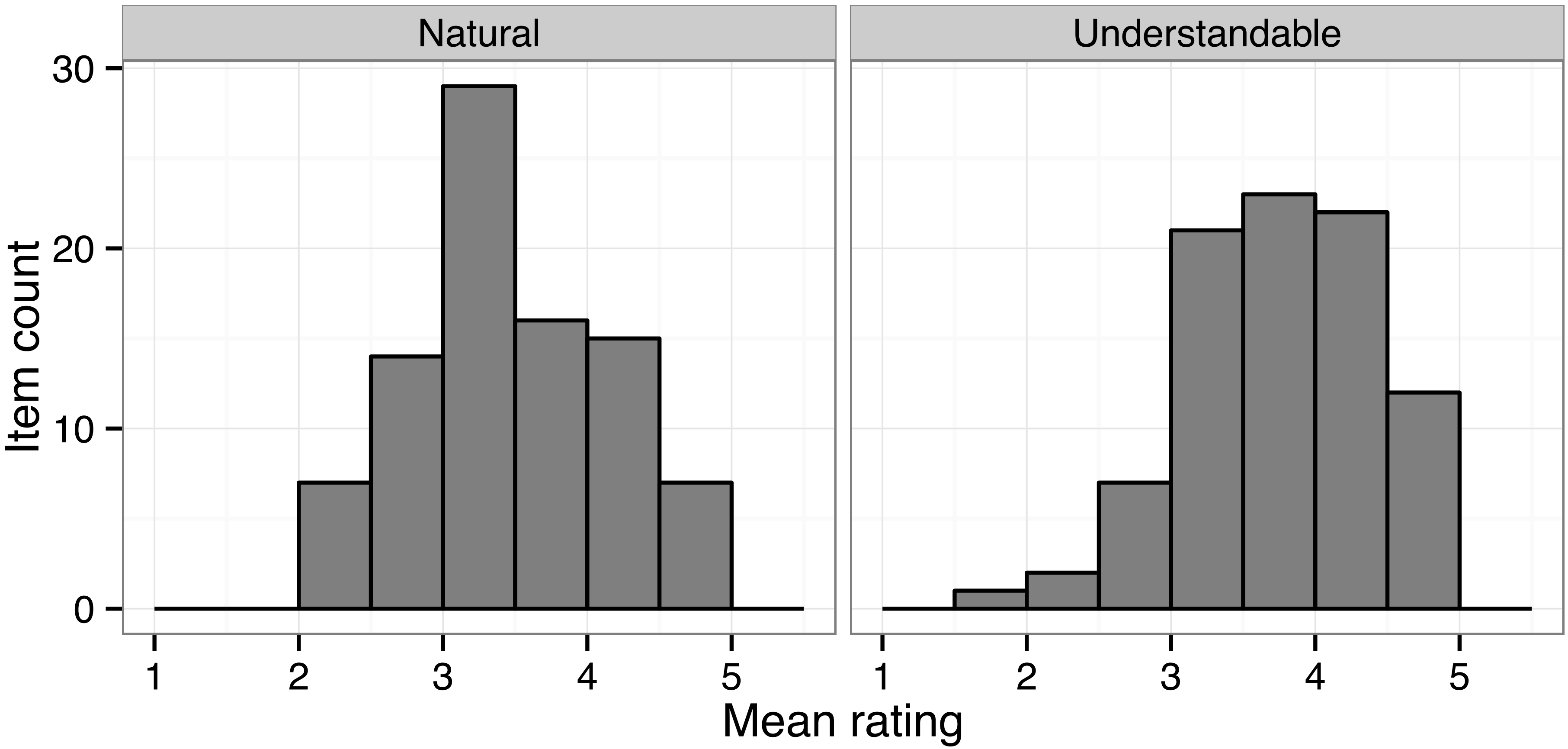
Overall little difference in understandability by overlap type

Syllable overlap type



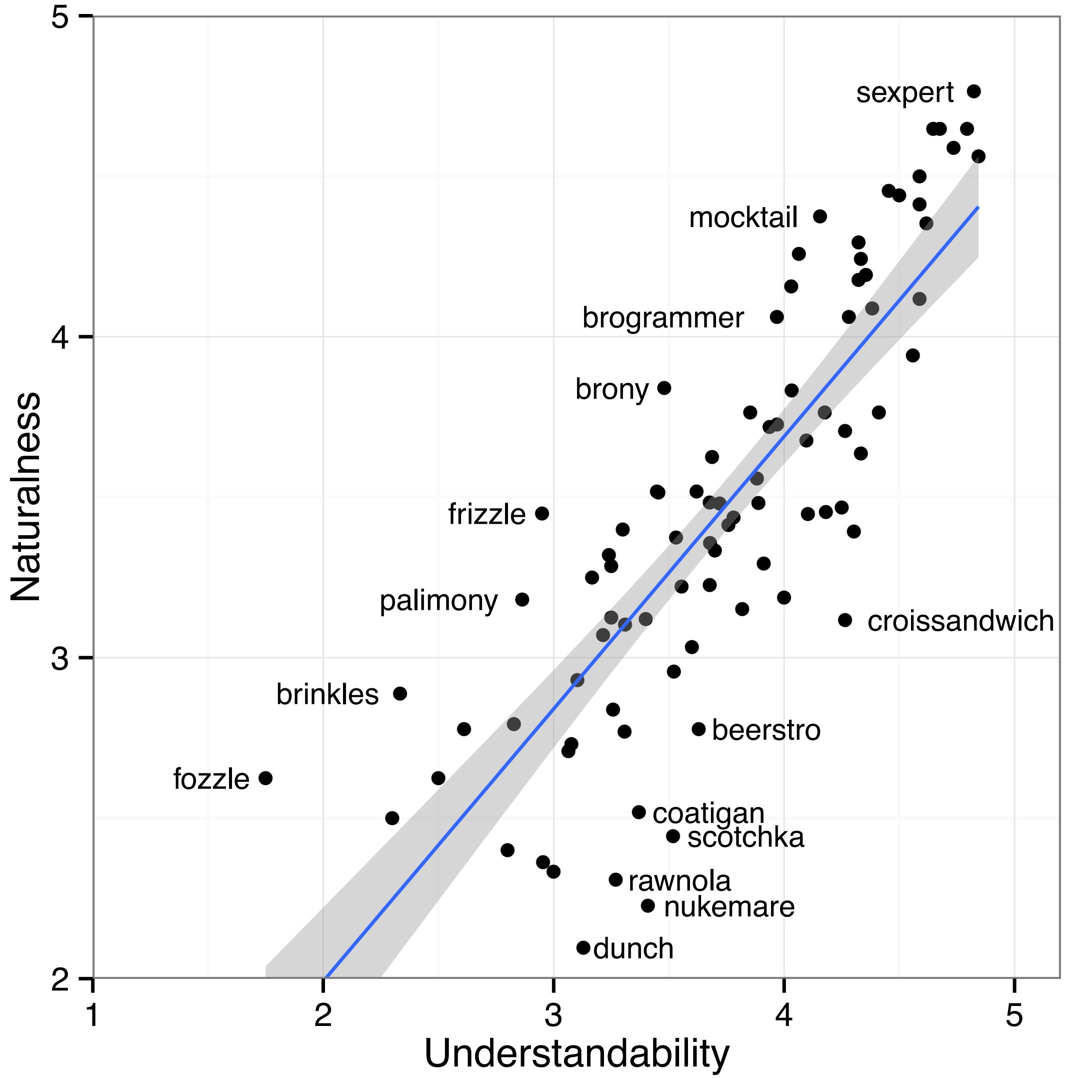
Full syllable overlap helps with understanding, but there's little differentiation between just onset and onset and nucleus

Item mean understandability/naturalness



Items rated more understandable than natural, understandable ratings skew high

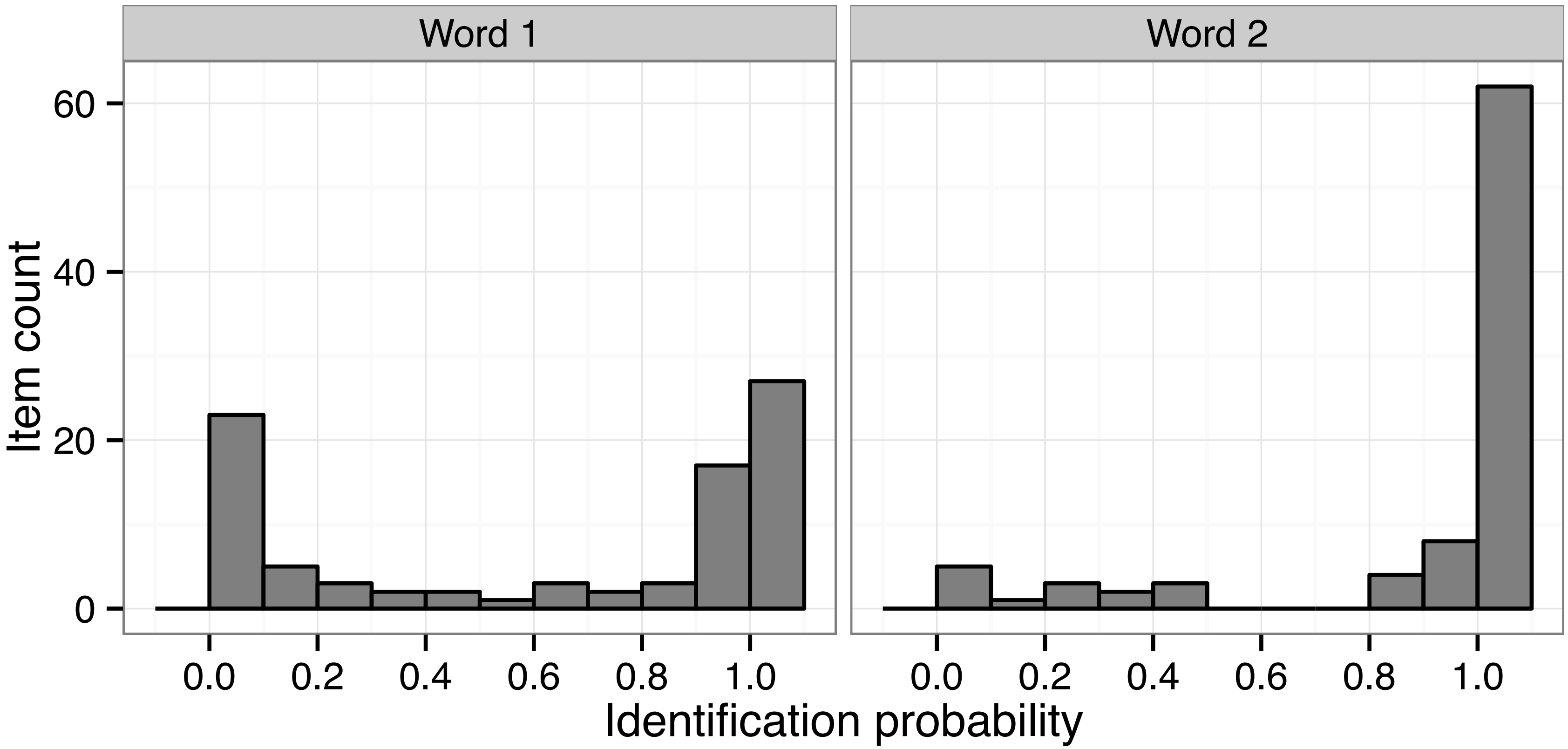
Ratings correlations



Above: more natural than understandable. Often minimal edits from real words.

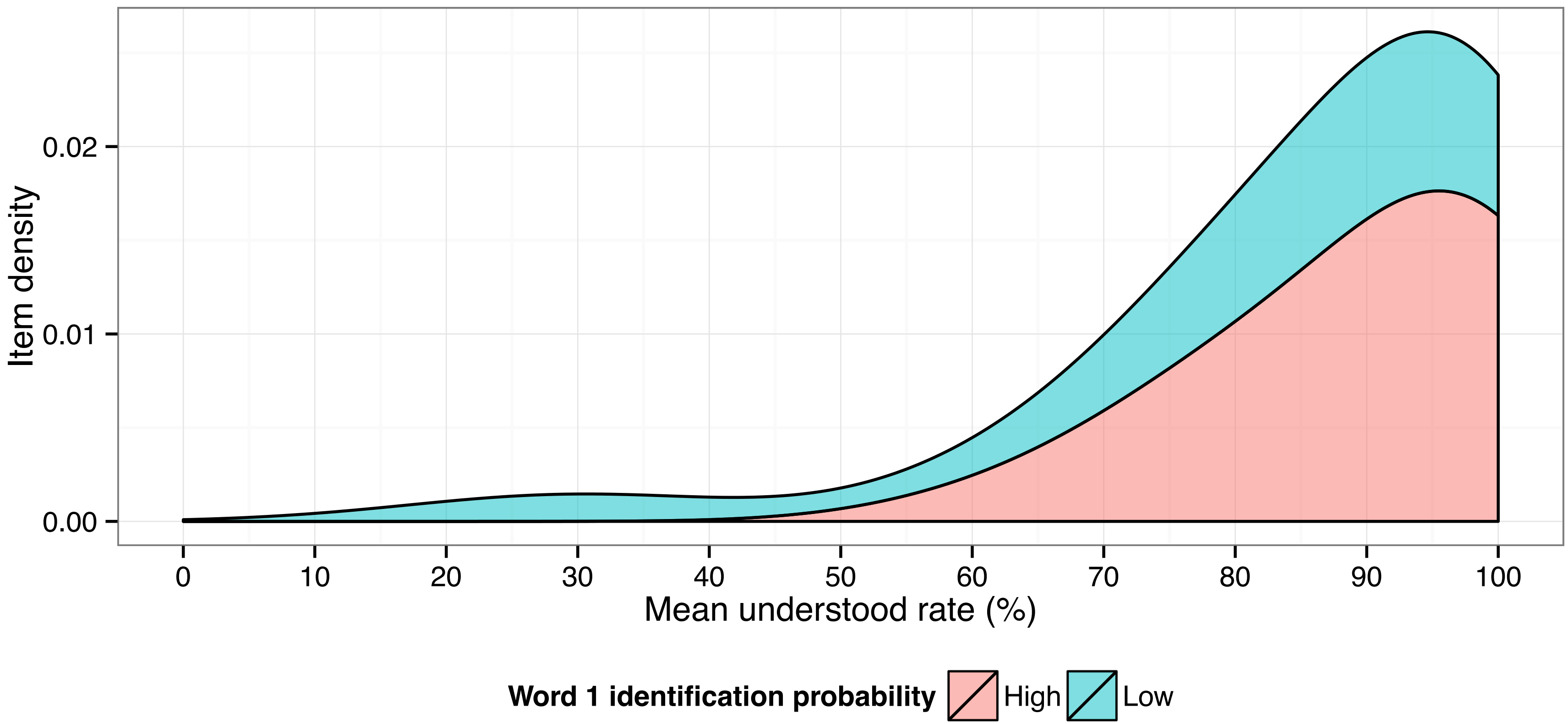
Below: can be understood, but unnatural. Marketing/branding:
beerstro
coatigan
croissandwich
rawnola

Distribution of identification probability



Word 1 has very high/low ID prob., Word 2 at ceiling

High and low identification probability

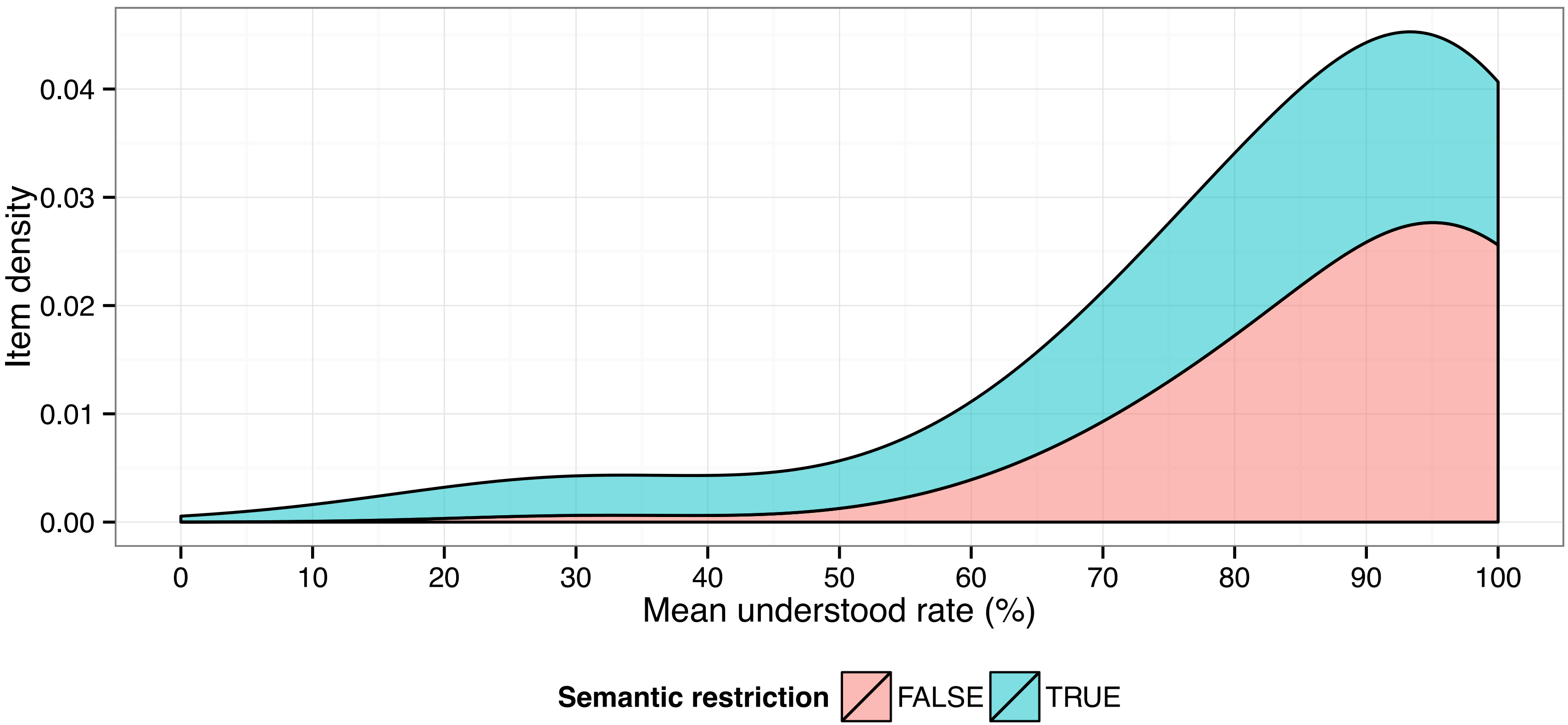


Low word 1 identification probability (< 0.5) is indicator of low rate of understanding

Modeling understandability

- Used cumulative-link mixed-effects models to model ratings, assess significance by Chisq. LL ratio test
  - No interval assumption or normality assumption
- Significant effect of first (p = 0.003) and second (p = 0.009) ID prob. on understandability
  - Each doubling of ID prob. → 15% chance of higher rating for word 1, 30% chance of higher rating for word 2
- ID probs. stronger predictor than segment ratios (better log-likelihood/AIC/BIC)
- Same pattern holds for modeling whether an item was understood as a binary response

Semantic domain restriction



Less-understood items are more likely to have a restriction (e.g., source words must be foods); this probably makes otherwise unacceptable blends tolerable.

V. Conclusion

Summary

- First human subjects study evaluating blend quality
- Identified properties of bad blends:
  - Less overlap
  - Little phonological content carried over from first word
- Good blends, however, come in all kinds
- Found reliable effect of identification probability on ratings
  - Suggests statistical processing effects on blend reconstruction
- Not yet able to model other blend domains
  - Personal names (*Kimye, Bennifer*), featural overlap (*hangry)*

Modeling blend choice

- We had subjects rate *attested* blends and modeled their ratings
  - Possible improvements: model continuous levels of semantic relatedness, part-of-speech matching between source words
- Next step is modeling the blend point of a given source word pair: why *frenemy* and not *frendemy* or *fenemy*?
- Proposal: model blend choice as binary classification
  - Positive examples are attested blends (*frenemy*)
  - Negative examples are unattested alternates (*frendemy*, *fenemy*)
  - Similar to Maxent OT or Harmonic Grammar

Further human subjects experiments

- Test impact of domain restriction on reconstruction ability
- Ask participants to give source words for a blend with or without the semantic domain
- Example:

What words are put together to make *fozzle*?
Hint: They're weather-related

Thanks!

References

Gries, S. Th. (2004). Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics* 42(3): 639–667.

Kelly, M. H. (1998). To "brunch" or to "brench": some aspects of blend structure. *Linguistics* 36(3): 579–590.

Blend sources:
http://en.wikipedia.org/wiki/List_of_portmanteaus

Thurner, D. (1993). *Portmanteau dictionary: blend words in the English language, including trademarks and brand names*. Jefferson, North Carolina: McFarland and Company.

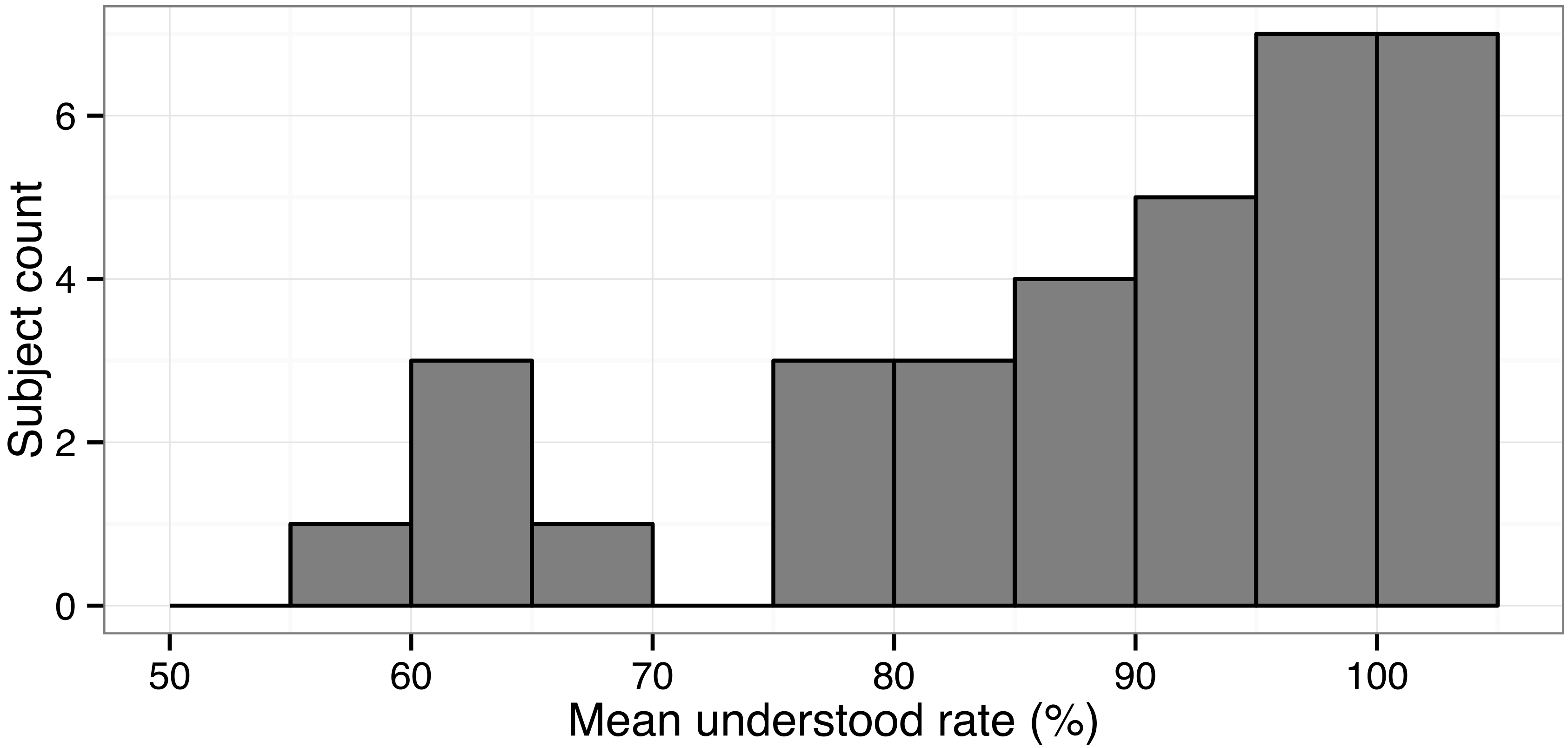
On libfixes vs. blends:

Zwicky:
http://arnoldzwicky.org/category/morphology/libfixes/

Gorman:
http://sonny.cslu.ohsu.edu/~gormanky/blog/defining-libfixes/

Additional slides

Subject variation



Cluster of five subjects said they didn't understand more than 30% of items