Agreeing with Google: We are Sensitive to the Relative Corpus Frequency of n-grams.

Cyrus Shaoul & Chris Westbury
University of Alberta, Dept. Of Psychology
Harald Baayen
University of Alberta, Dept. Of Linguistics

Why a psychology of n-grams?

- There may be parallels between the morpheme/word relationship and the word/n-gram relationship.
- Storage is ubiquitous (for inflected and derived words, and perhaps for some n-grams too).
- May allow us to better understand the process of lexicalization.
- May offer a better way of understanding semantic processing of sentences.

Classical Orthographic Freq.

- For words: a very strong predictor of speed and accuracy in word comprehension and production.
- If we are so sensitive to a word's frequency, why not to an n-gram's frequency?
- Subjective Frequency is related to Objective/ Corpus Frequency. Both are estimates of our experience with words.

Extending Subjective Frequency to n-grams

- Collected ratings on the subjective frequency of n-grams.
- groups of 150 undergraduates to rate 120 n-grams each.
- Measured the mean rating and the standard deviation of the ratings for each n-gram.

Subjective n-gram Frequency Survey

Second: please rate how frequently the phrases below are used. A rating of *almost never* means that the phrases are used very rarely. A rating of *very often* means that the phrase is used very frequently.

	almost never			about average			very often
be able	○	○	○	○	○	○	○
can be used	○	○	○	○	○	○	○
umbilical cord	○	○	○	○	○	○	○
is part of the	○	○	○	○	○	○	○
serve as a guide to	○	○	○	○	○	○	○

	almost never			about average			very often

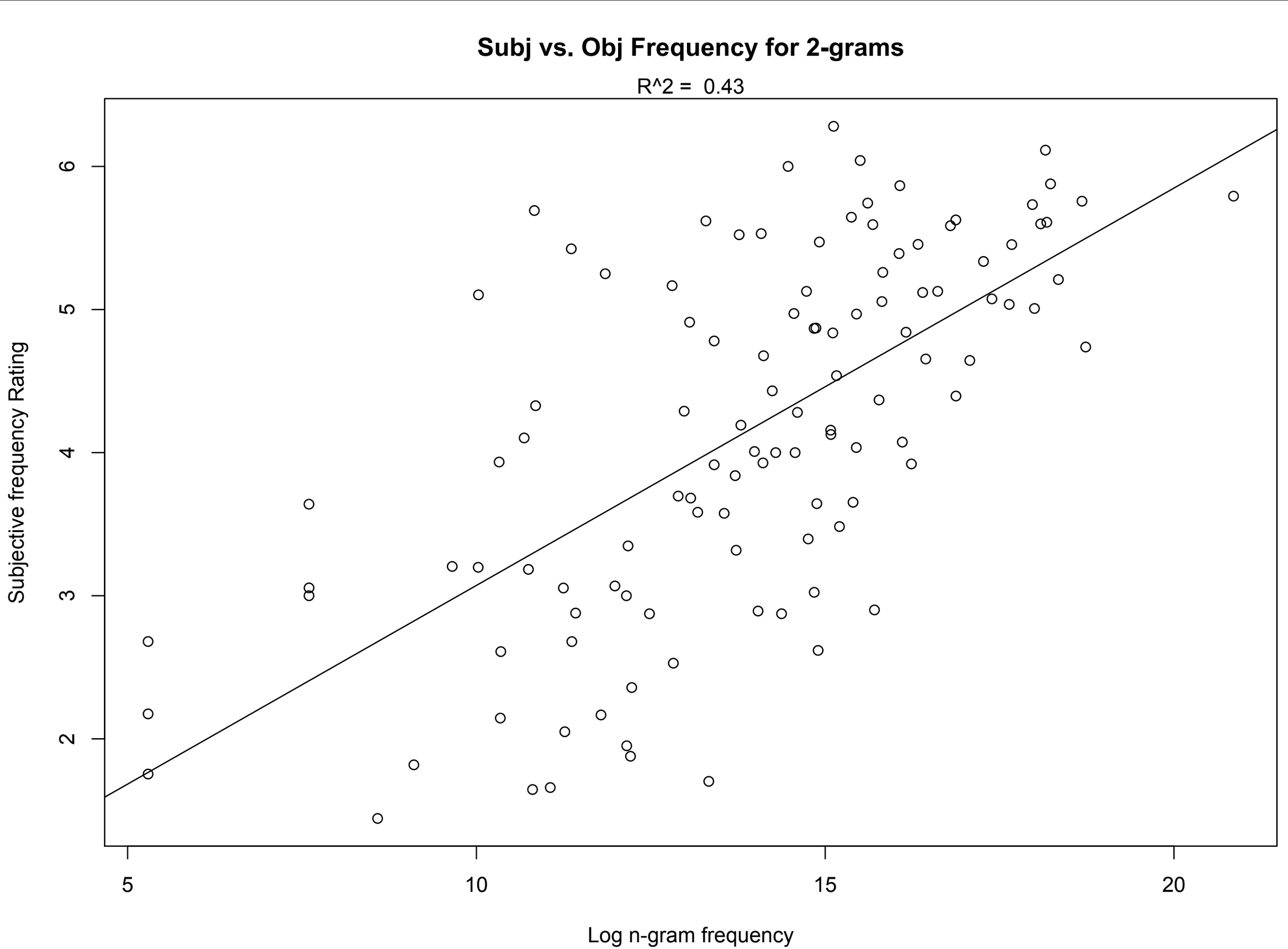
| hip and stylish | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| safer to keep it here | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Sanity Check

- Inter-rater variability was within reason for each item (1-2 points standard deviation for all items).
- Zero-frequency n-grams were rated appropriately.

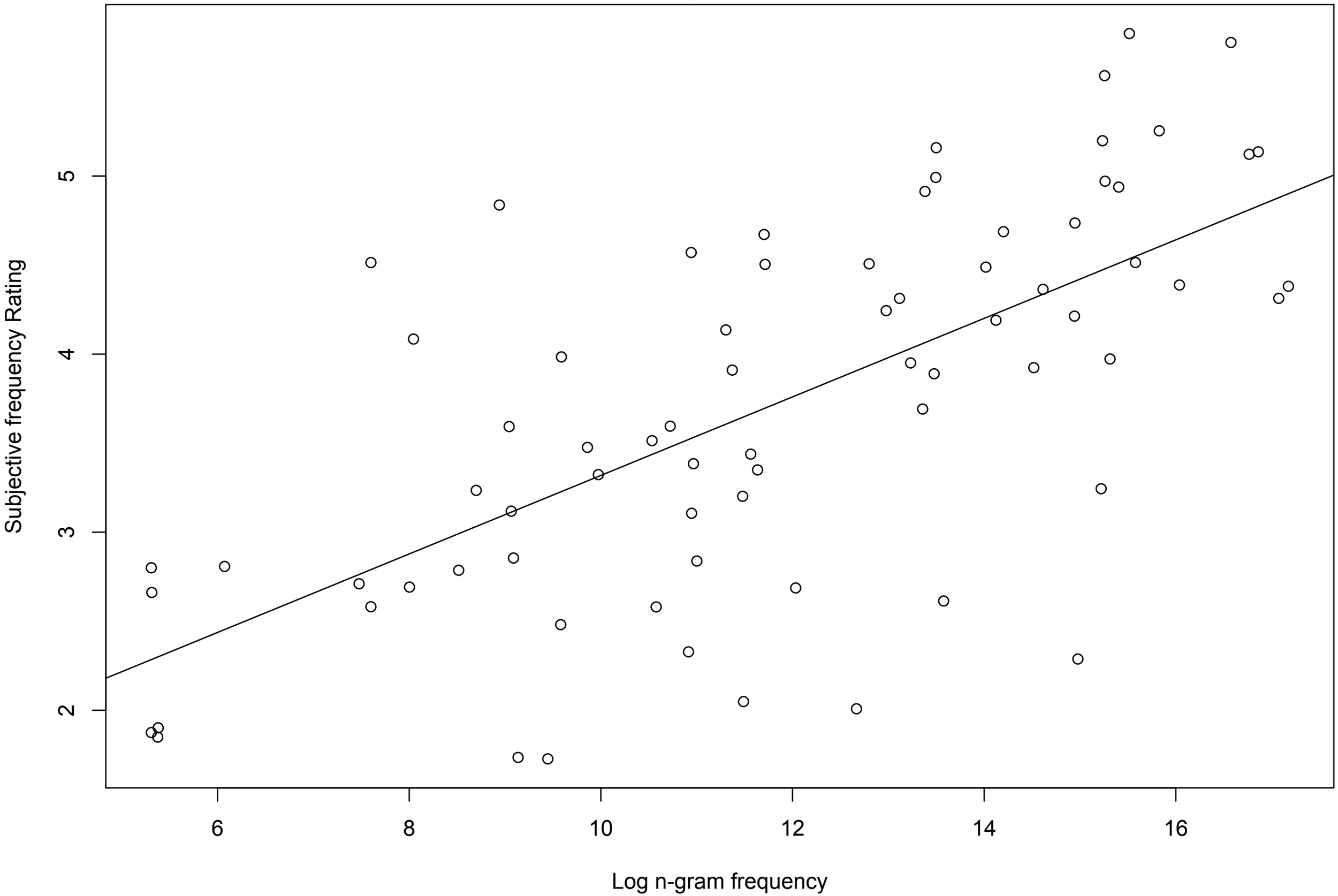
Subjective Vs. Objective Frequency

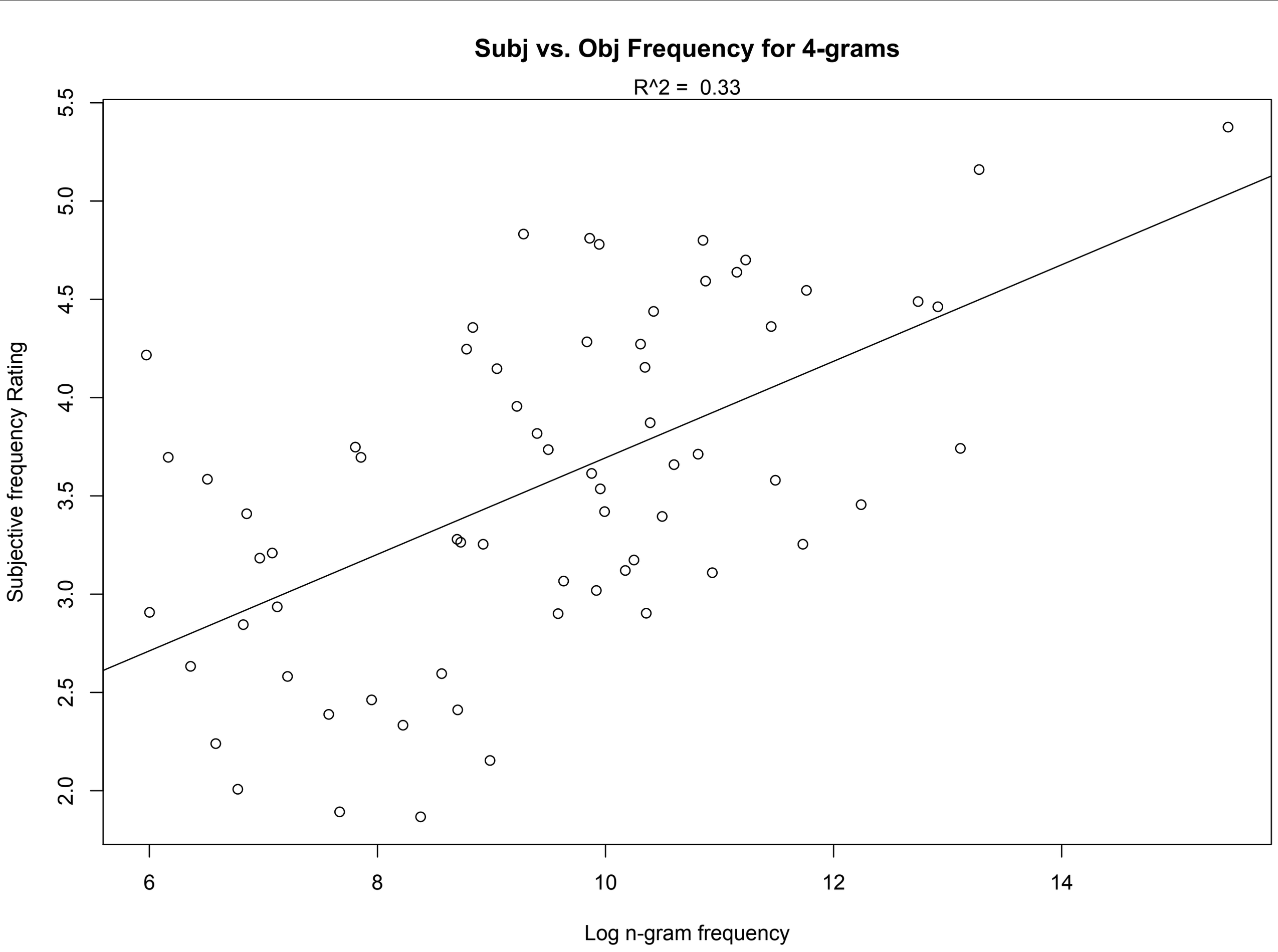
- Is subjective n-gram frequency (familiarity) correlated with objective n-gram frequency?
- Previous work with single words is compelling (Balota, Pilotti & Cortese, 2001): Log Freq and meaningfulness were correlated with familiarity. (Celex vs. Subjective Familiarity, r=0.83)

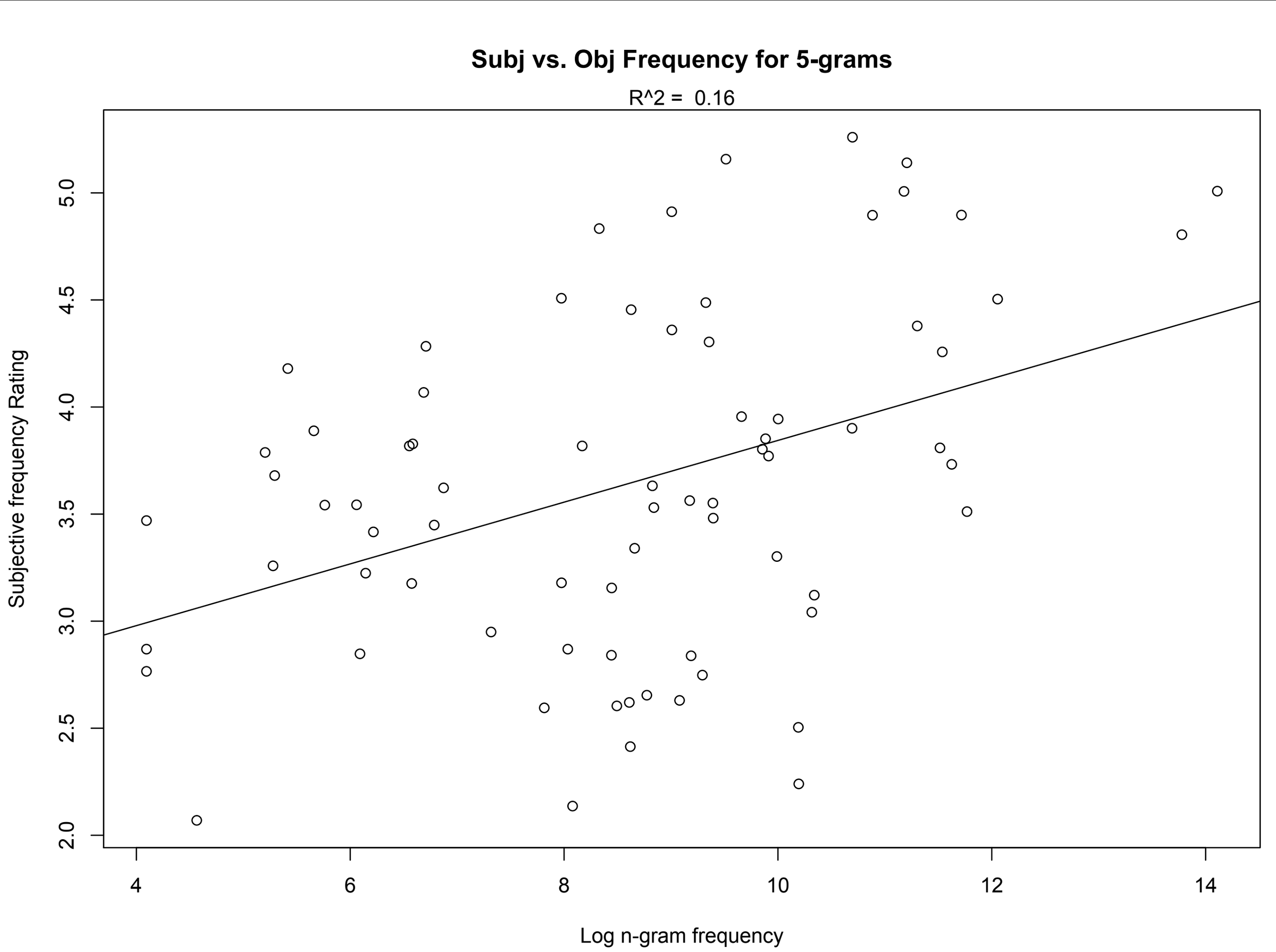

Subj vs. Obj Frequency for 2-grams
R^2 = 0.43
Subjective frequency Rating
2
3
4
5
6
5
10
15
20
Log n-gram frequency

Subj vs. Obj Frequency for 3-grams

R^2 = 0.44




Subj vs. Obj Frequency for 4-grams
R^2 = 0.33
Subjective frequency Rating
Log n-gram frequency
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6
8
10
12
14


Subj vs. Obj Frequency for 5-grams
R^2 = 0.16
Subjective frequency Rating
Log n-gram frequency
2.0
2.5
3.0
3.5
4.0
4.5
5.0
4
6
8
10
12
14

Getting at implicit frequency effects: the n-gram comparison task

➡ Hypothesis: The ratio of the frequencies of the two n-grams will influence the ability of subjects to predict the n-gram's Google frequency. The larger the ratio, the easier it will be to detect the difference, and therefore the more accurate the decisions will be.

First up: Unigrams

- Stimuli: 120 pairs of words, matched on OLD20 (Yarkoni & Balota, 2008) and length (4, 5 or 6 letters), with a even spread across the range of frequency ratios.
- 33 right handed undergraduates from U of A Psychology Dept. Research Pool, all native English speakers.

Statistical Inference

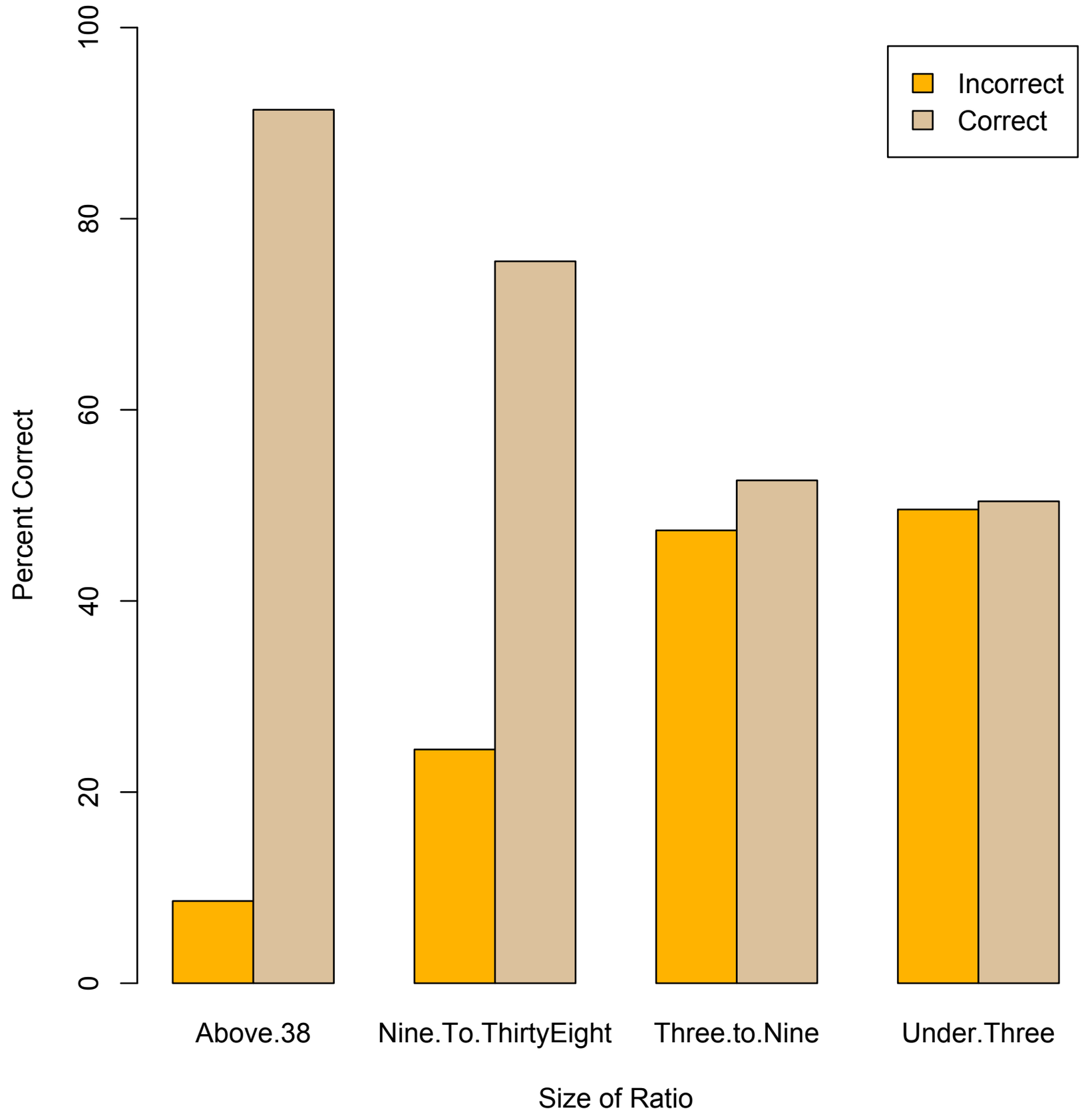
- Due to the nature of the design (within subjects, fully crossed items) we used Linear Mixed Effects Models to model the accuracy (lme4 package in R using a generalized linear mixed model for the binomial dependent variable).

tooth

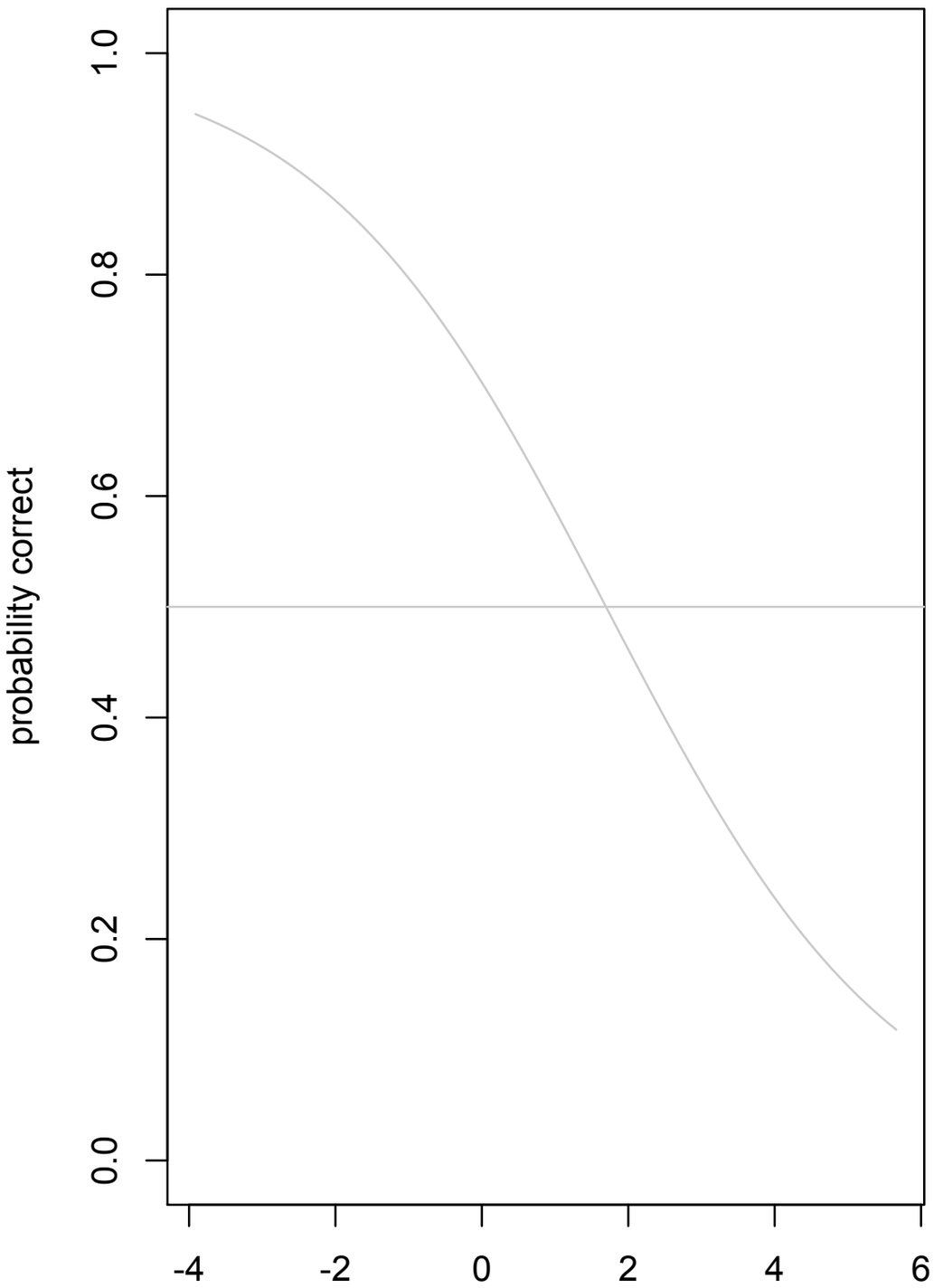
+

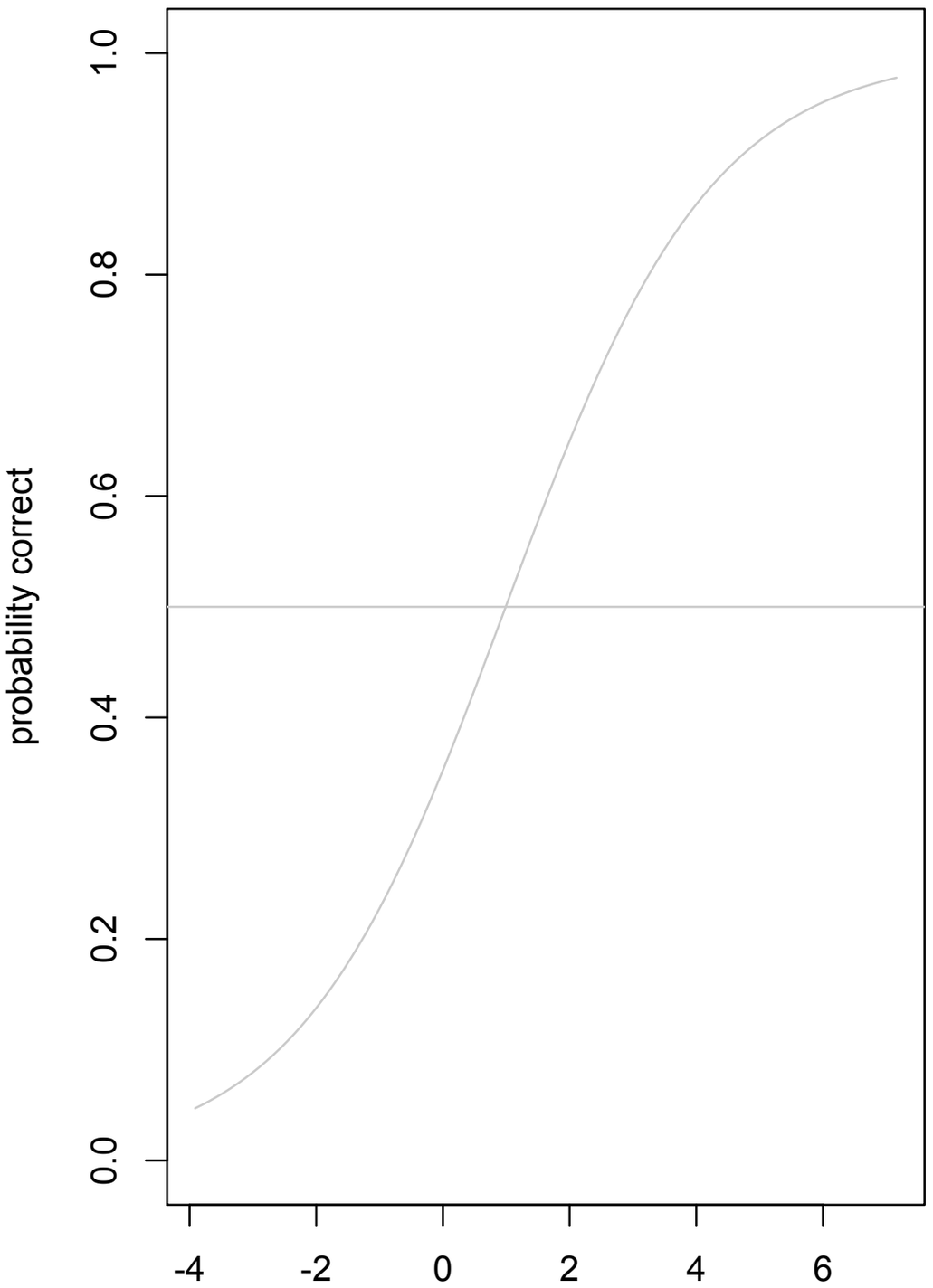
alert

Item accuracy for 1-grams



Plots of relationship from Linear Mixed Effects model for Google Web1T frequencies





Next: 2,3,4 and 5-grams

- Stims: 2, 3, 4 and 5-grams sampled from the Google Web1T data set based on n-gram frequency.
  - Pairs were matched on the geometric mean of the individual word frequencies.
  - Distributed across a broad range of geometric means and n-gram frequency ratios.
  - Subject Variables: Age, Education, Gender, Reading Speed, Vocabulary Size
- Participants: 49 right handed undergraduates from U of A Psychology Dept Research Pool, all native English speakers.

Sample 2-gram Stimuli

N-gram	n-gram Freq	Word 1 Freq	Word 2 Freq	Geom. Mean
metric tons	0.61	4.38	8.68	**1.5**
inner workings	0.26	11.48	1.22	**1.5**
N-gram Freq Ratio	**2.3**			

More sample stimuli

3-gram:

dubious scientific value vs. long curly hair

(N-gram Frequency Ratio= 41.8)

4-gram:

played a central role vs. making false statements in

(N-gram Frequency Ratio= 10.7)

5-gram:

the first step in the vs. and can be used for

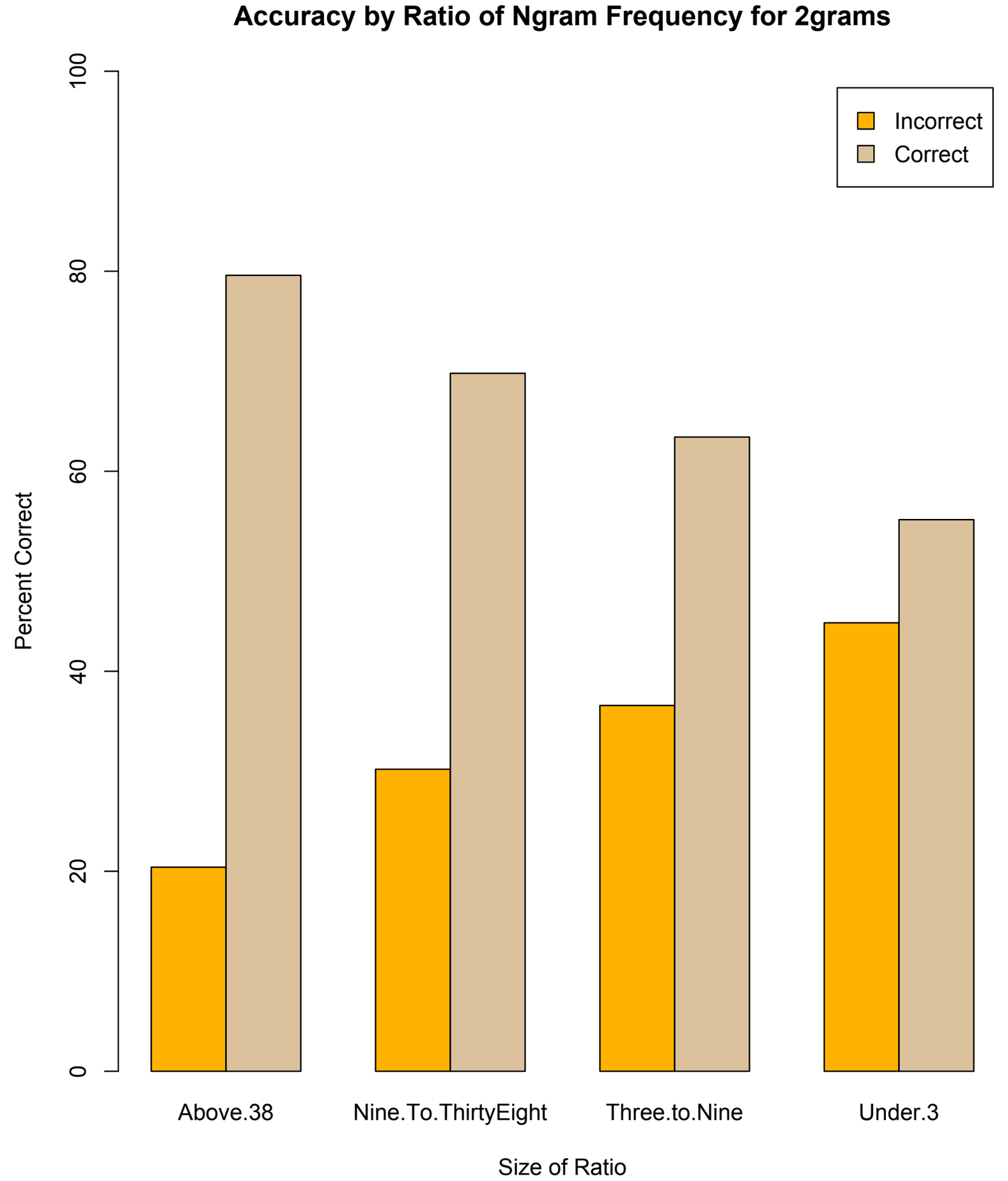
(N-gram Frequency Ratio= 0.8)

motor vehicle

+

heart disease

Item accuracy for 2-grams

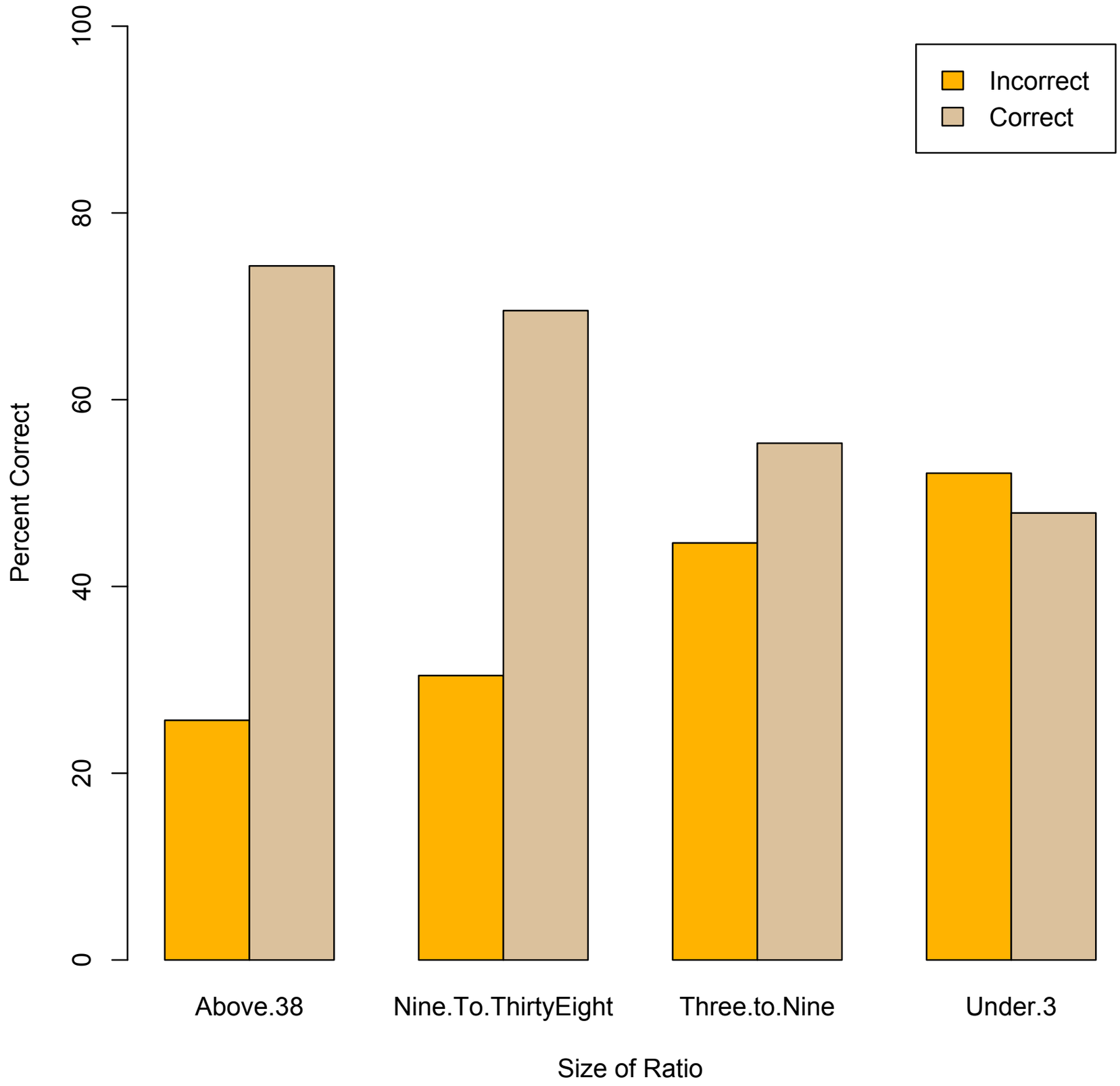


you will find

+

feel free to

Item accuracy for 3-grams

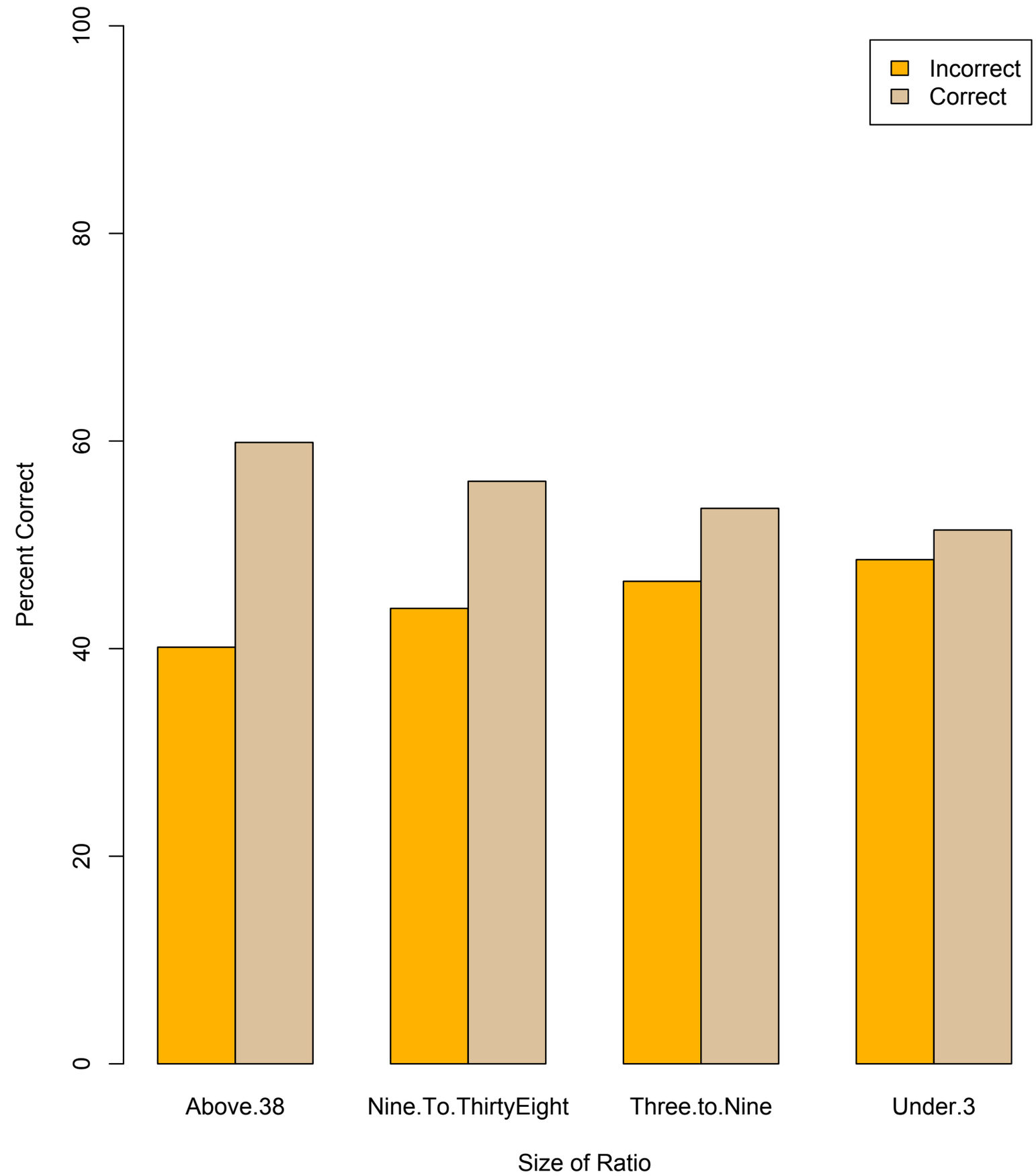


in court was a

+

starting in the new

Item accuracy for 4-grams

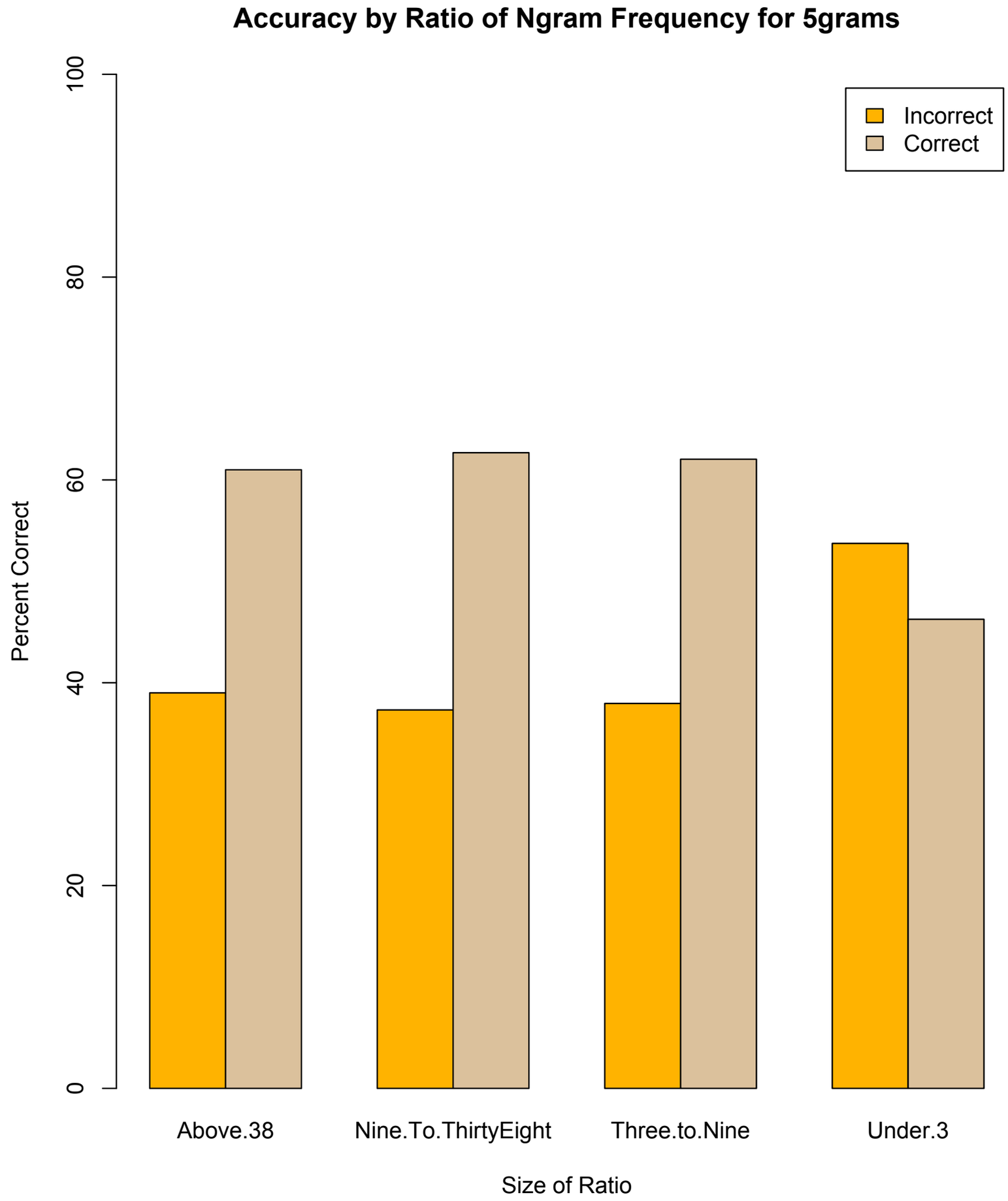


gave birth to a beautiful

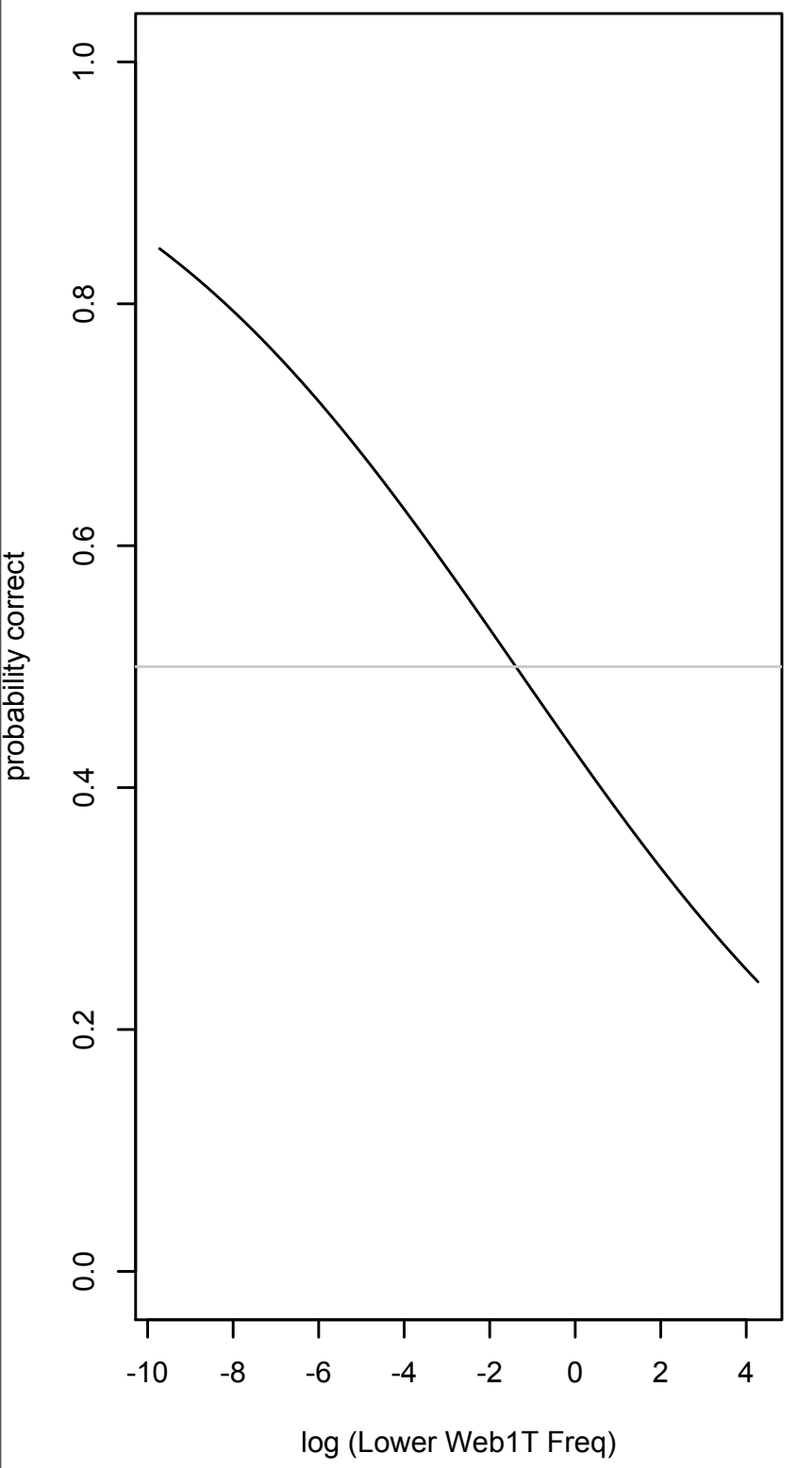
+

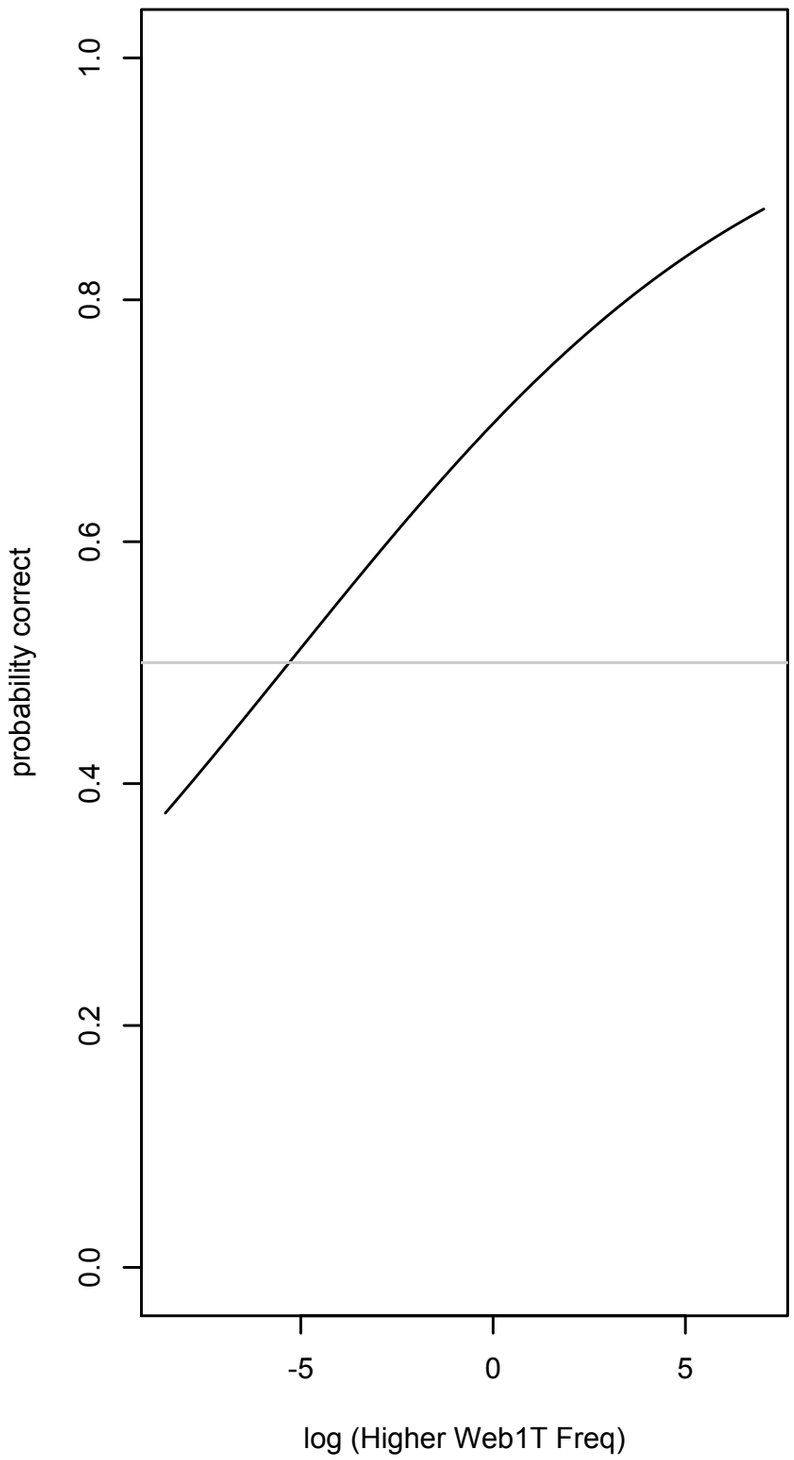
help you organize your home

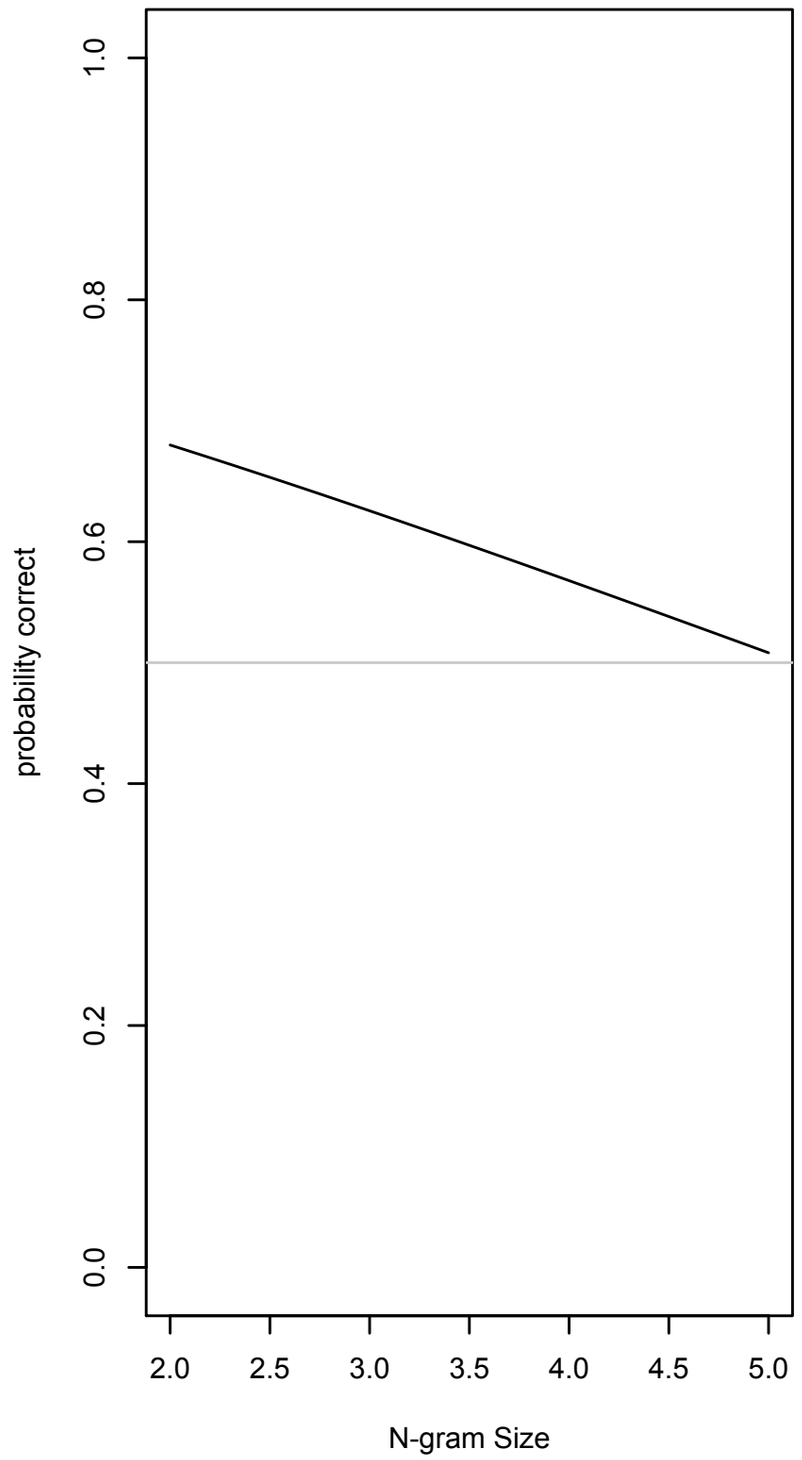
Item accuracy for 5-grams



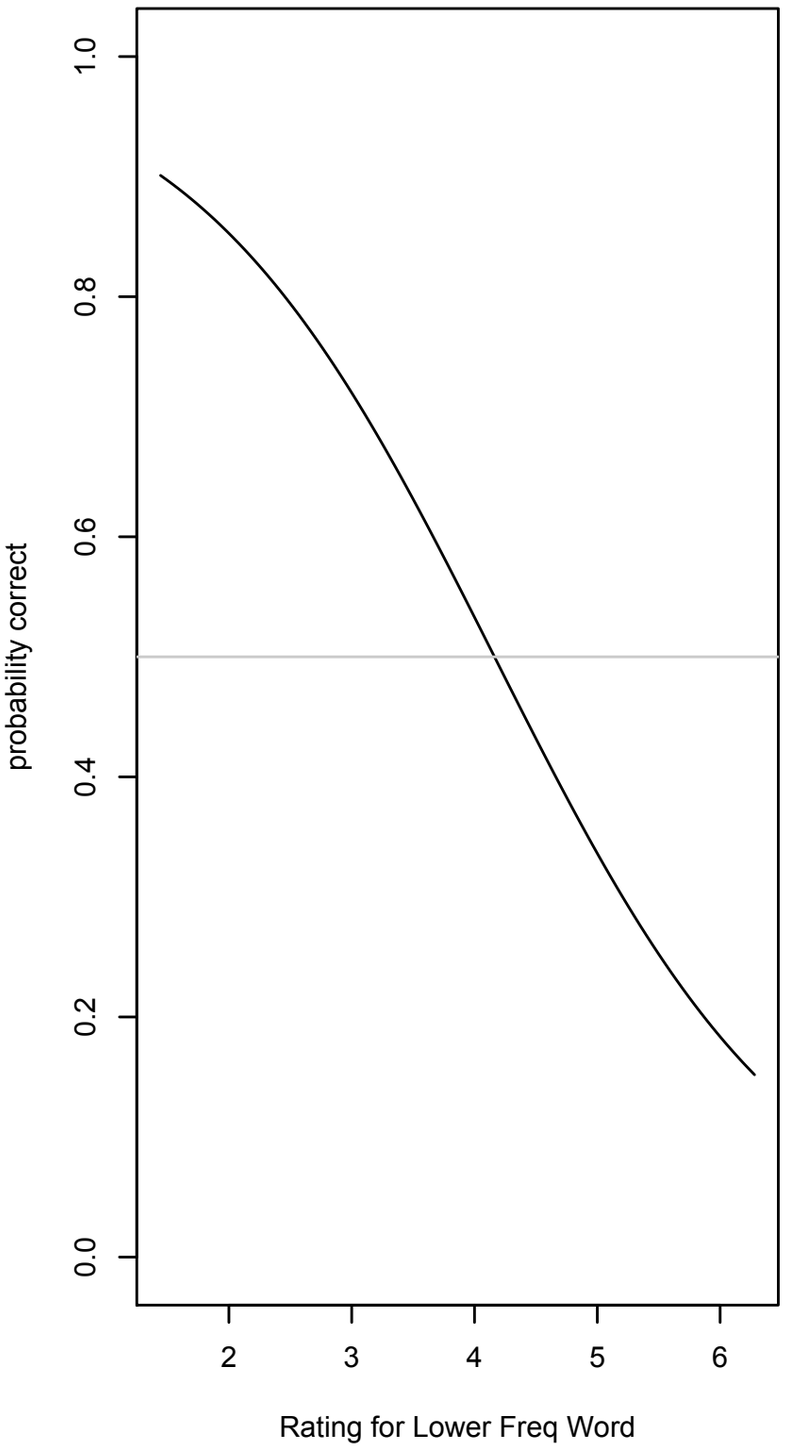
Plots of relationship from Linear Mixed Effects model for Google Web1T n-gram frequencies

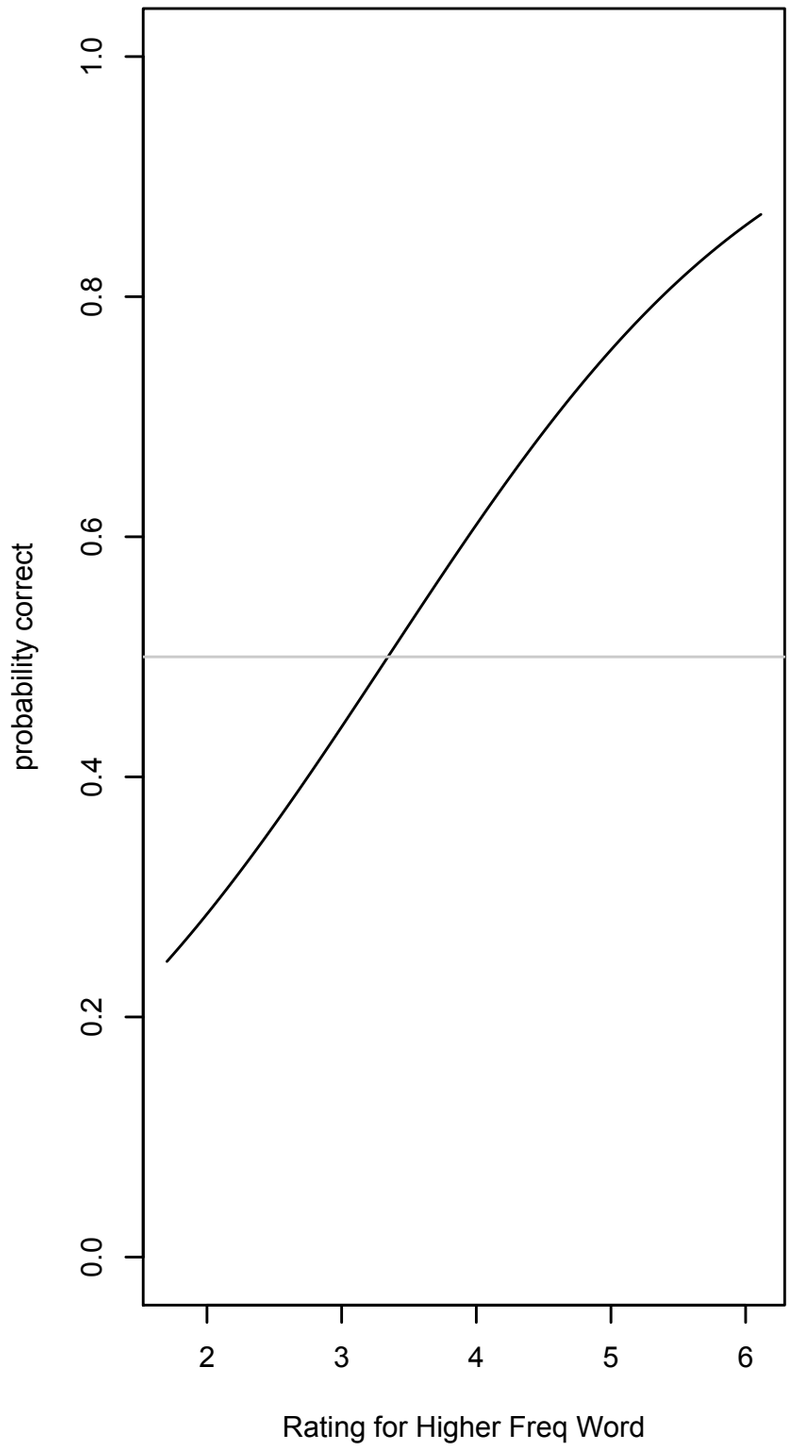


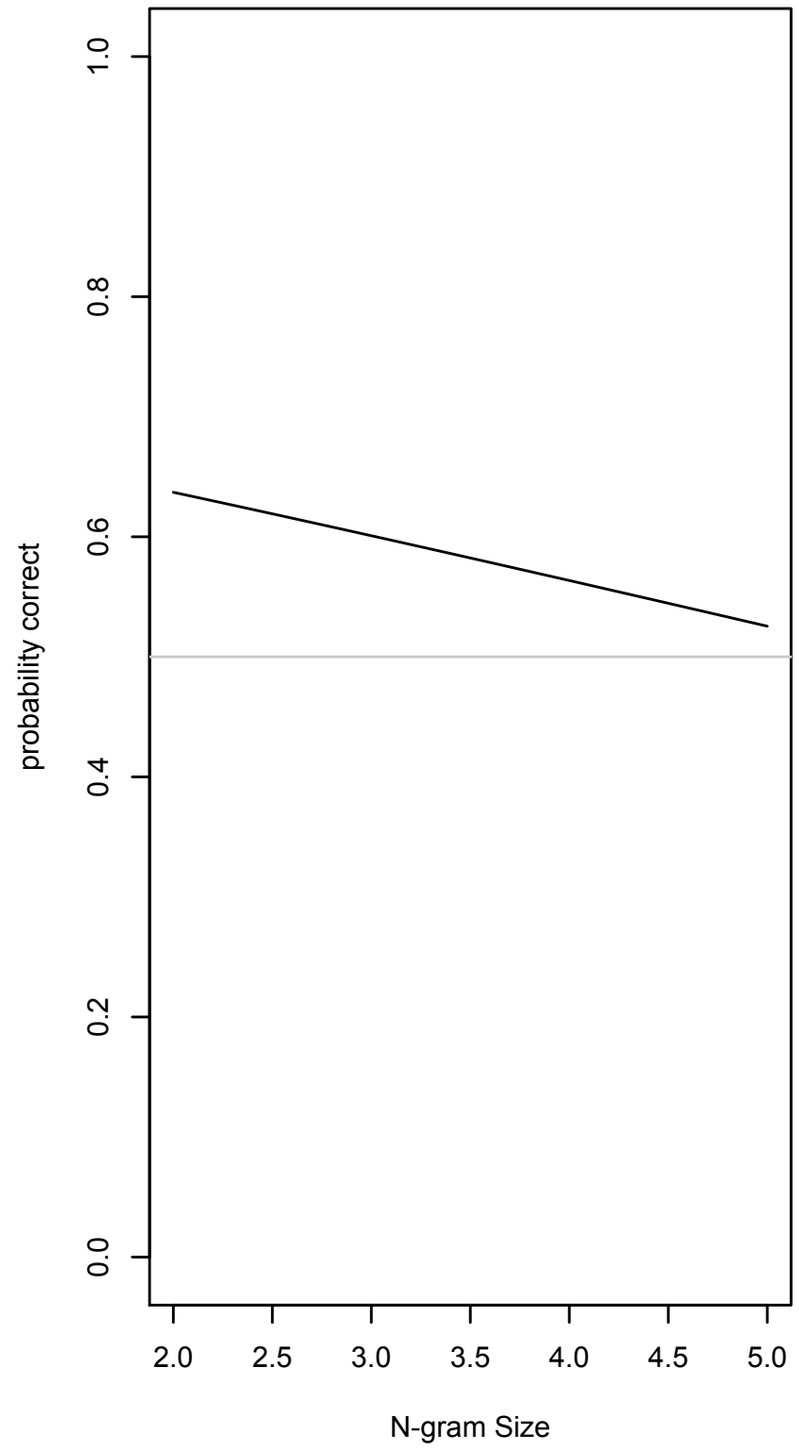




Plots of relationship from Linear Mixed Effects model for subjective ratings







Conclusions

- The subjective corpus frequency of n-grams can predict the likelihood of choosing the higher frequency n-gram.
- Lexical frequency of words in the n-gram are not driving performance.
- Implicit knowledge of the relative frequency of n-grams exists, and it is correlated with corpus frequency.

Questions to look into...

- How might n-gram frequency be represented?
- Are n-grams similar to words in other ways besides frequency effects?
- How are zero frequency n-grams processed?
- What impact does n-gram frequency have on production tasks?

Thank you!