# Intro

## Our goal is to evaluate the theory and find improvements for it

# Kennedy & McNally Theory

## Use examples from before

## Throw the chart in

## Mention the small set of direct predictions that we’re working from

# Related Works

## Experimental Syntax

## Clustering work

## Syrett & Lidz

## Look over my Lit Review again

# Experiment

## Intro

### Testing the theory on acceptability directly

## Experiment 1

### Methods

#### Model this section after papers on experimental syntax

#### Set of most common adjectives and adverbs narrowed down

#### Randomly picked 100 out of that set and tested everyone on those

#### Ran it with 50 subjects

### Results

#### Y subjects failed the catch trials, but the rest were used

#### We zscored the results between subjects

#### .9666 similarity between natural and grammatical conditions

##### collapsed data for analysis

## Experiment 2

### Methods

#### Picked a dataset that they make direct claims on

#### Took 49 x 7 words, iterated on it 7 times, randomized within the study

#### 340 subjects

### Results

## Discussion

### Anchoring effect from how/well

### Effects may have come from accessibility, not naturalness

#### May be easy to imagine circumstances using it… embedded with qualifying phrases

# Vector-Space Model

## Intro

### Pose the usage of corpus stats as a possible extension/replacement for human judgment

### This is NOT an intended use for the theory

### “corpus frequency should reflect acceptability but nosily”

## Methods

### Describe dataset, different clustering methods

## Results

### Graph of different correlations with humans

#### Be sure to comment on the ceiling performance

### Clustering generally wasn’t informative

#### Coclustering did find interesting groupings on adverbs

## Discussion

### Using a larger dataset in general hurt performance

### Smoothing necessary to deal with 0 counts

### CR wipes these out appropriately

#### Works for reasons slightly different than intended

#### Don’t actually want to flatten too much

# Results

## Removed “well” and “how” from most analysis because of its big weight

## Summary statistics

## Forced choice agreement between all 3 models

## Graphs

### Get trends across the adjective categories

# Discussion

## Specific observations from data

## Note things that went really well

## Correlation between experiment and model is hard b/c of sparsity

## Dependency parse may pull relative adjectives out of their context

### Support w/ a # for poor performance

## Relative adj w/ max standard mod is okay w/ people, but not with NYT

### People may be using it as hyperbole

## Half is a bad predictor

### Low agreement for closed in human ratings

## Slightly isn’t working very well

## See notes

# Conclusion