Massive Auditory Lexical Decision: Going Big in the Auditory Domain

Benjamin V. Tucker and Daniel Brenner

Department of Linguistics, University of Alberta
{bvtucker,brenner}@ualberta.ca

Mental Lexicon 2016: Ottawa, 20 Oct. 2016

Acknowledgements

- R. Harald Baayen, D. Kyle Danielson, Danielle Fonseca, Catherine Ford, Matt Kelly, Pearl Lorentzen, Filip Nenadic, Katelynn Pawlenchuk, Michelle Sims
- With funding provided by:

Conseil de recherches en sciences humaines du Canada

Social Sciences and Humanities Research Council of Canada

Canada

UNIVERSITY OF ALBERTA

APhL Alberta Phonetics Laboratory

Background

'Megastudies' have several important advantages:

- statistical power
- minimization of strategic effects
- comprehensiveness
- multi-functionality
- complementing traditional small factorial experiments
- model development and testing

Megastudies

Visual Lexical Decision

For example:

- the English Lexicon Project (Balota et al., 2007): 40,000 words and non-words
- the French Lexicon Project (Ferrand et al., 2010): 38,000 words and non-words
- the Dutch Lexicon Project (Keuleers et al., 2010): 14,000 words and non-words
- the British English Lexicon Project (Keuleers et al., 2012): 28,700 words and non-words

Auditory megastudies

The only Auditory Lexical Decision megastudy: BALDEY (Ernestus Cutler, 2015):

- 5,541 words and 5,541 pseudo-words
- 10 female and 10 male listeners

Auditory megastudies

The only Auditory Lexical Decision megastudy BALDEY (Ernestus Cutler, 2015):

- 5,541 words and 5,541 pseudo-words
- 10 female and 10 male listeners

And now:

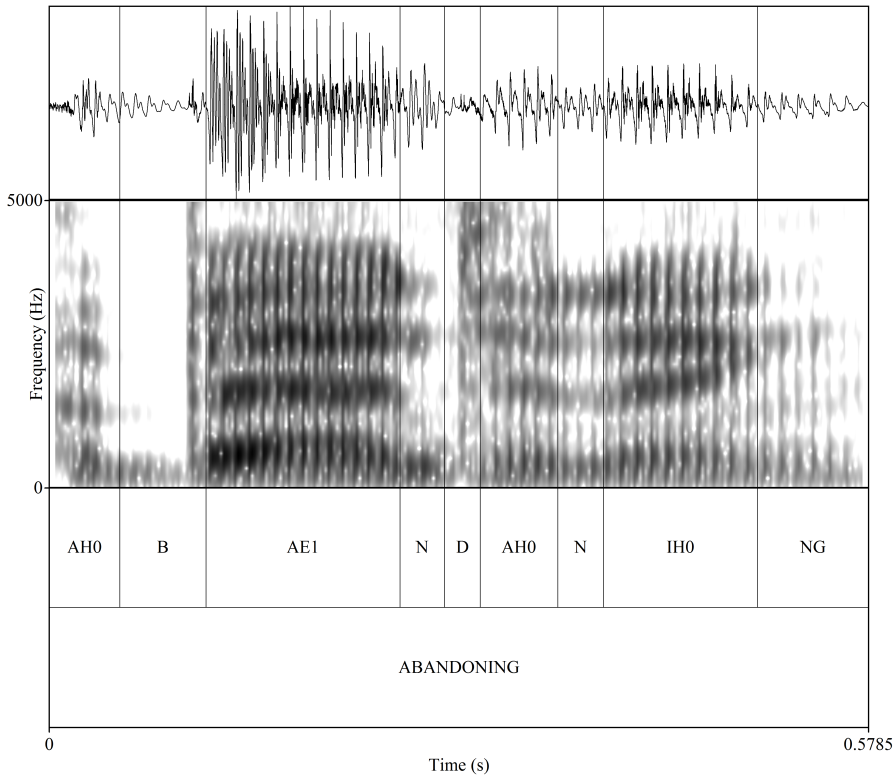
- MALD: Massive Auditory Lexical Decision

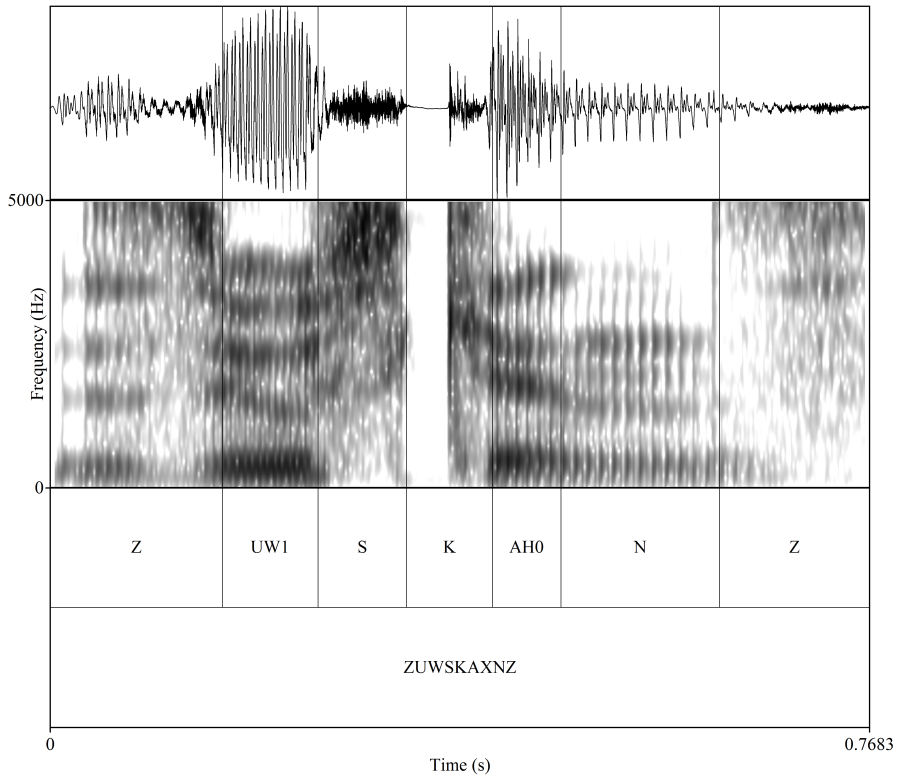
Items

- Male Western Canadian English speaker (age 32) recorded in a sound attenuated booth
  - 28,511 words
  - 11,400 non-words (wuggy, Keuleers Brysbaert, 2010b)
    - Words and non-words are morphologically complex
    - 1000 compound words and non-words
- About 2000 words/day or 800 non-words/day
- Items were extracted and mispronunciations removed leaving:
  - 26,800 English words
  - 9,600 pseudo-words
- All items provided with segmental level mark-up (p2fa, Yuan Liberman, 2008)

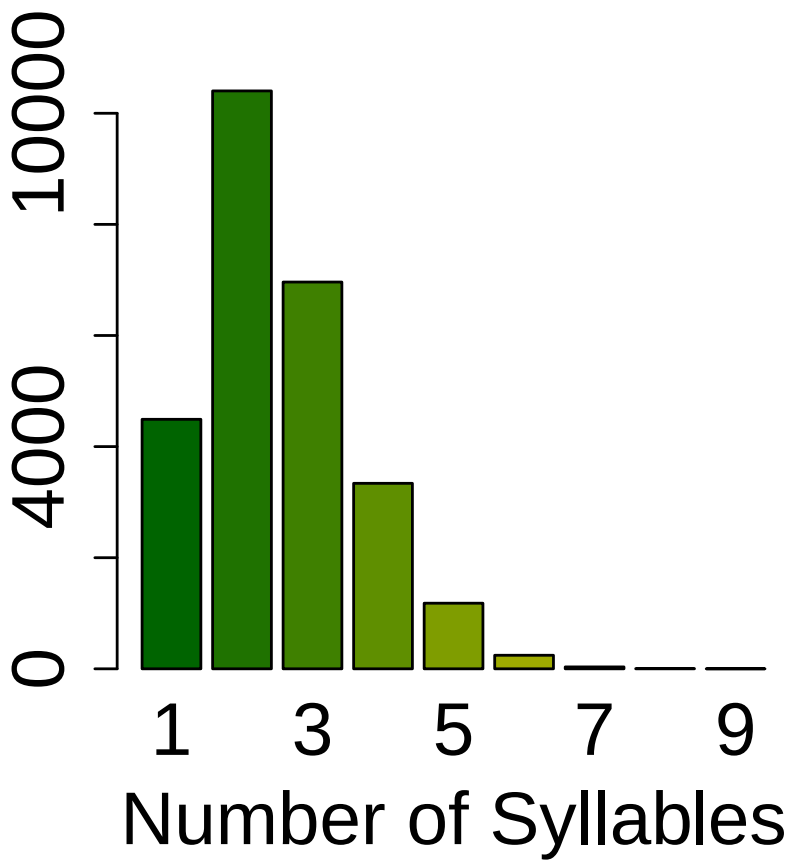
APhL
Alberta Phonetics Laboratory

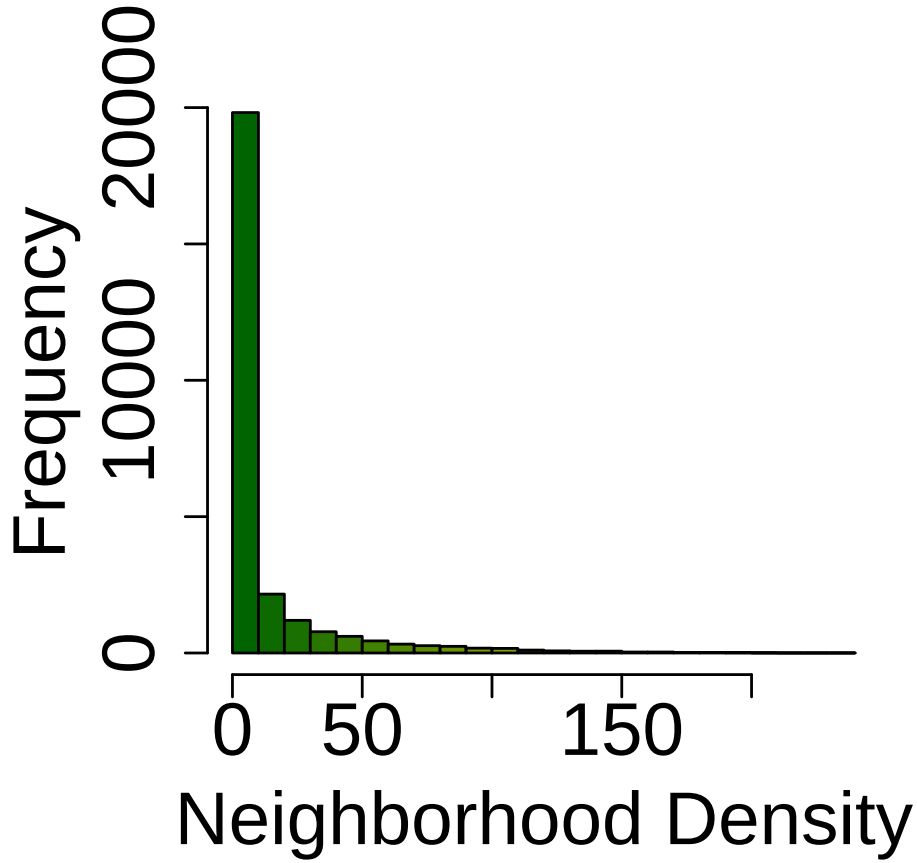
Item markup


5000
Frequency (Hz)
0
AH0
B
AE1
N
D
AH0
N
IH0
NG
ABANDONING
0
0.5785
Time (s)


5000
Frequency (Hz)
0
Z
UW1
S
K
AH0
N
Z
ZUWSKAXNZ
0
0.7683
Time (s)

Item information





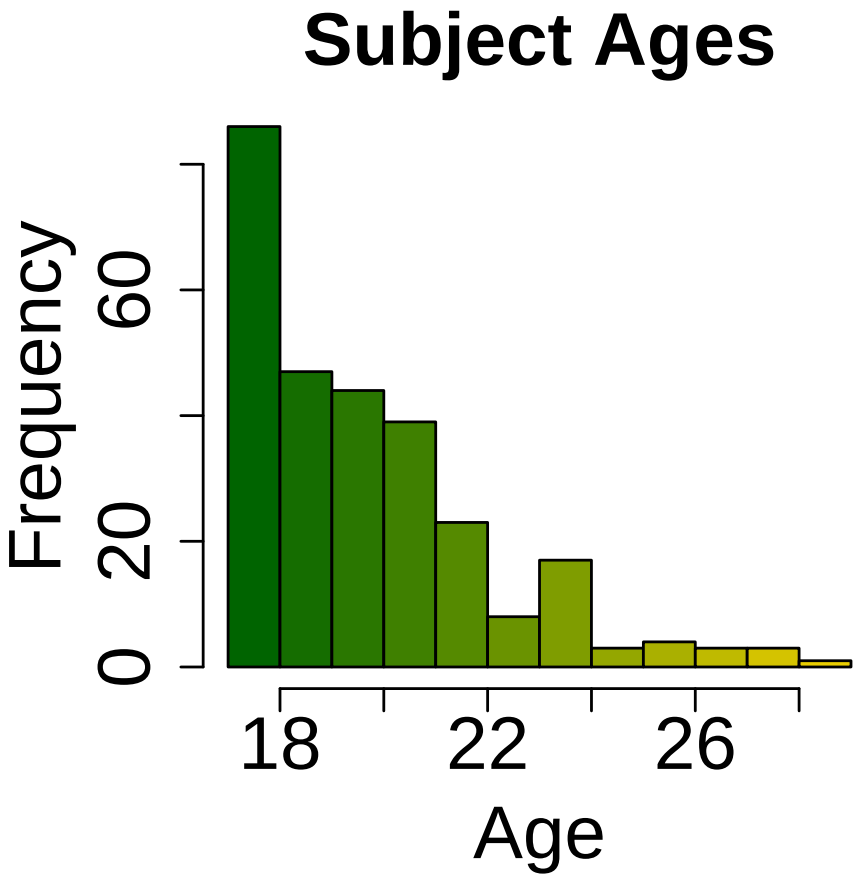
Procedure

- Hearing screening
- Auditory Lexical Decision task
- Session lasts approximately 25min
  - Goal: At least 4 responses per word (400 words/400 pseudowords per experiment)
- Participants could participate in up to three sessions

- 232 monolingual Canadian English participants
- 285 total experimental sessions

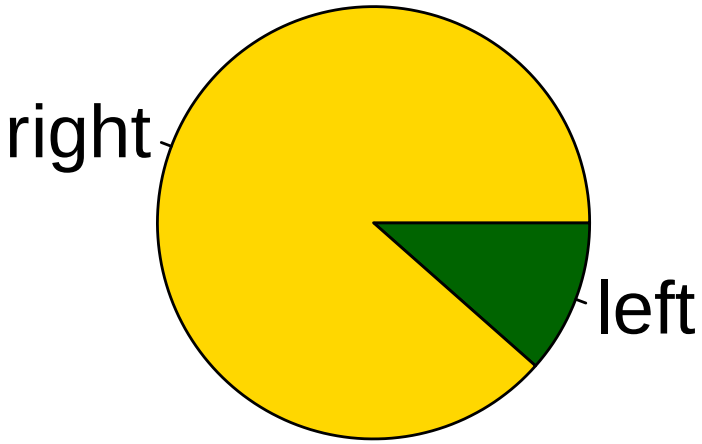
228,000 total button presses

Participants

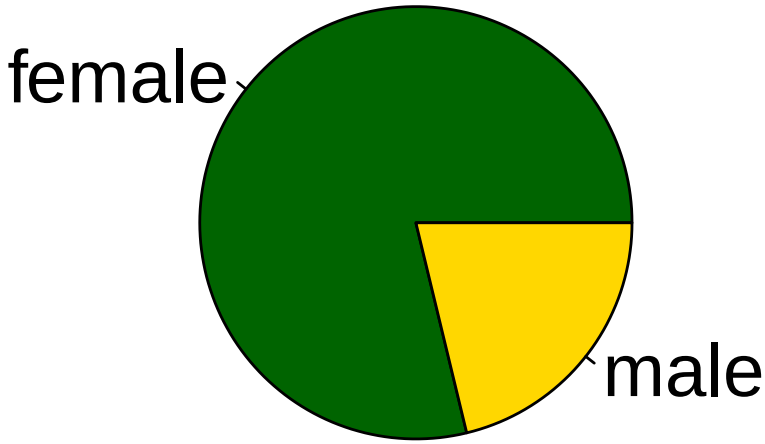


Participants

Subject Handedness



Subject Sex



Data Frame

Independent Variables (already in the data file):

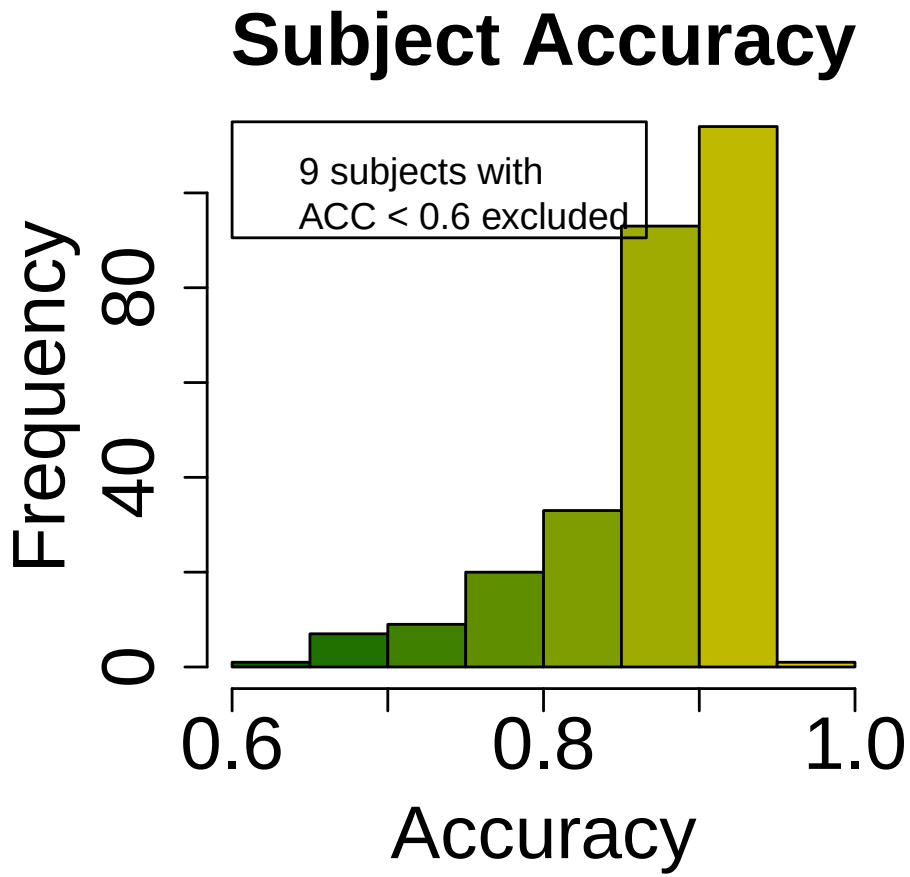
- Word Duration
- Education
- Neighborhood Density
- Frequency (COCA, COCA Spoken, Google nGram)
- Non-word characteristics (e.g. phonotoactic probability, Phonological Neighborhood Density)
- Age
- Sex
- Handedness
- Word Run Length

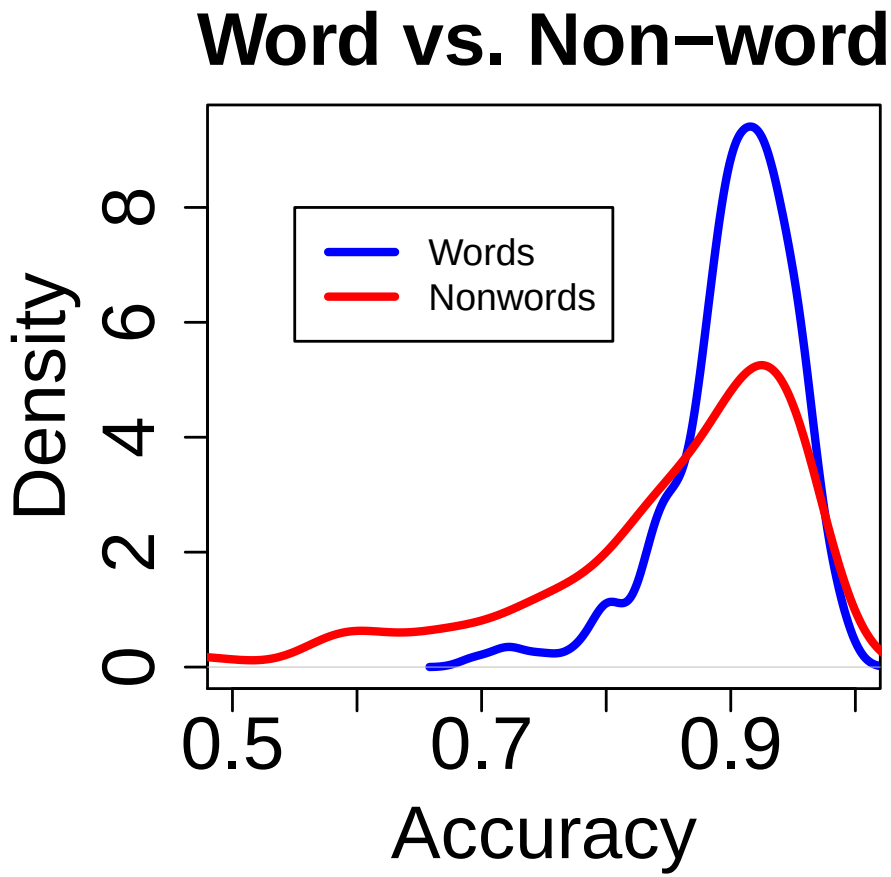
Data Frame

Dependent variables:

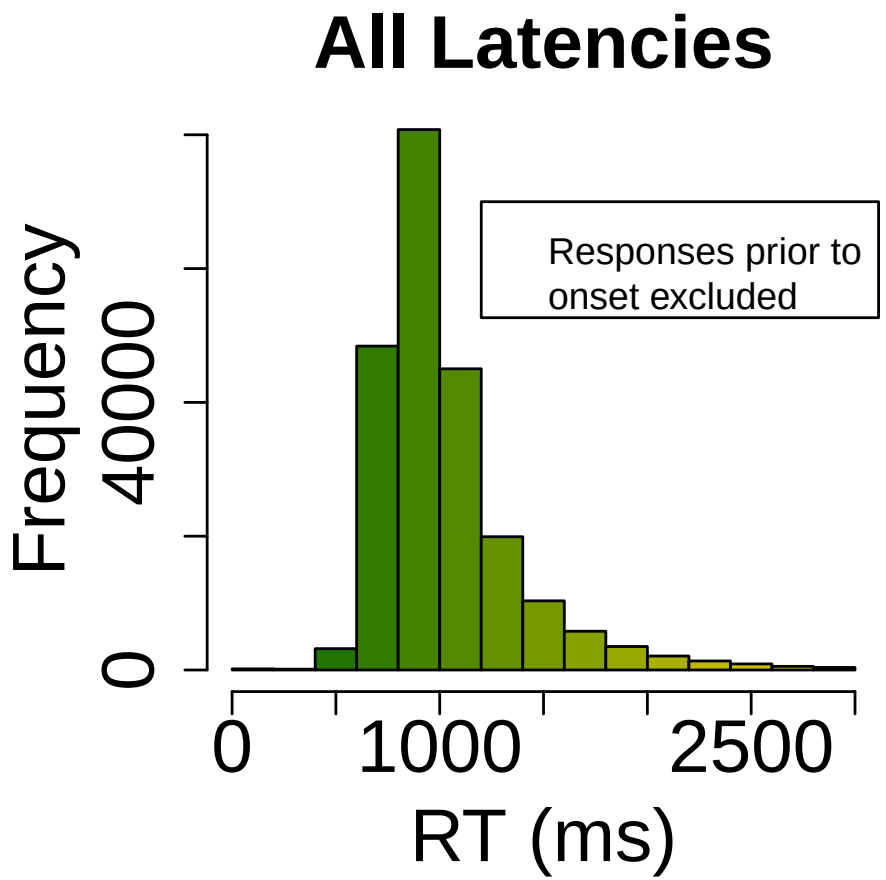
- Acoustic characteristics
- Response Latency
- Accuracy

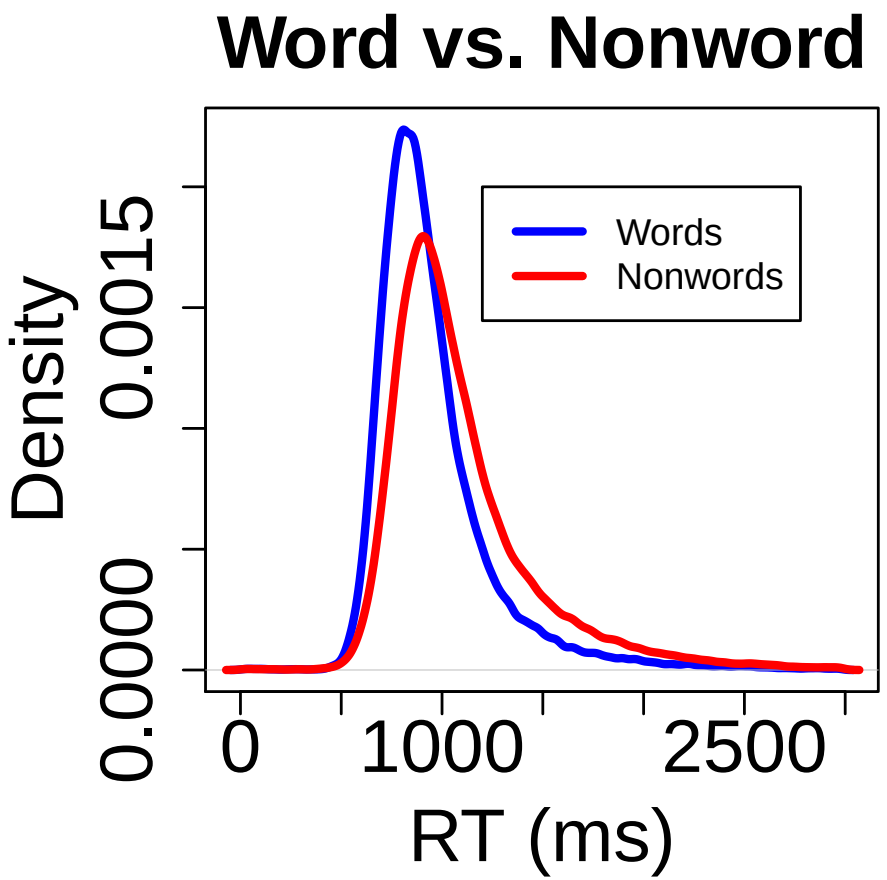
Accuracy


Subject Accuracy
9 subjects with
ACC < 0.6 excluded
Frequency
0
40
80
0.6
0.8
1.0
Accuracy


Word vs. Non−word
Words
Nonwords
Density
0
2
4
6
8
0.5
0.7
0.9
Accuracy

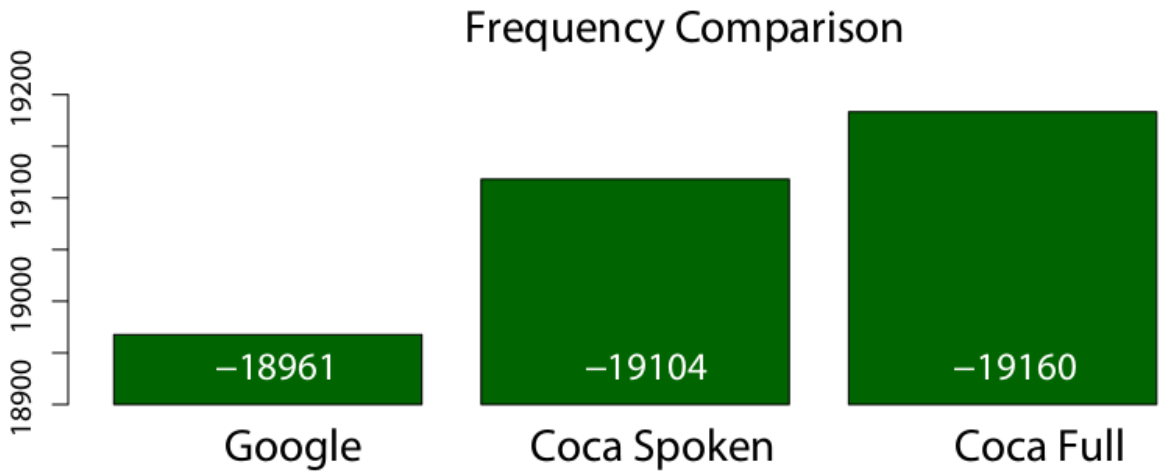
Response Latency





Frequency

Linear mixed effects regression (R core team, 2013; Bates et al., 2014)



- Counts based on all genres in COCA are better than spoken language
- Counts based on the Google Unigram corpus are less predictive than COCA

Thank you!

Stay tuned for the public release of the database:
https://aphl.artsrn.ualberta.ca/