

# Computational Cognitive Science

## Lecture 11: Word Recognition

Chris Lucas

(Slides adapted from S. Goldwater & F. Keller's)

School of Informatics  
University of Edinburgh  
clucas2@inf.ed.ac.uk

24 October, 2017

## Background

Word Recognition

Experimental Tasks

Psychological Findings

## Bayesian Reader

Word Identification

Lexical Decision

Discussion

Reading: Norris (2006).

# Word Recognition

Word recognition is easy:

- ▶ for a familiar language, we perceive a continuous speech stream as discrete words;
- ▶ in reading, we effortlessly segment sequences of letters into words.

Word recognition is hard:

- ▶ each word is an arbitrary mapping betw. sound and meaning;
- ▶ there are no word boundaries in the speech signal, and even for written language, there are sometimes no spaces;
- ▶ every time a word is pronounced, it sounds differently;
- ▶ written words can vary a lot (font, handwriting).

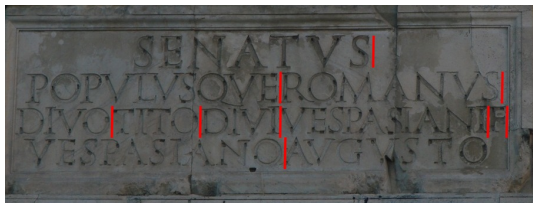
How do we recognize words, either in context or in isolation?

# Word Recognition

The input for spoken word recognition looks like this:



The input for visual word recognition can look like this:



During reading, the input is chopped up into incomplete (or redundant) bits that the brain has to assemble.

## Experimental Tasks: Reading

Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair, from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the find, thousands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

# Experimental Tasks: Reading

Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair, from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the find, thousands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

# Experimental Tasks: Reading

Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair, from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the news, the sands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

# Experimental Tasks: Gating

## Gating task:

- ▶ participants listen to increasingly long word prefixes;
- ▶ they respond by saying the word they think they hear as soon as they can;
- ▶ reaction time is recorded and the response is scored;
- ▶ the length of the prefix informative about word recognition.

Demo



# Experimental Tasks: Lexical Decision

Lexical decision task:

- ▶ either word or non-word is presented to participants;
- ▶ they have to decide as quickly as possible whether the stimulus is a word by pressing the right button;
- ▶ reaction time and accuracy are recorded;
- ▶ the stimulus can be presented either auditorily or visually.

slog

# Experimental Tasks: Lexical Decision

Lexical decision task:

- ▶ either word or non-word is presented to participants;
- ▶ they have to decide as quickly as possible whether the stimulus is a word by pressing the right button;
- ▶ reaction time and accuracy are recorded;
- ▶ the stimulus can be presented either auditorily or visually.

trespass

# Experimental Tasks: Lexical Decision

Lexical decision task:

- ▶ either word or non-word is presented to participants;
- ▶ they have to decide as quickly as possible whether the stimulus is a word by pressing the right button;
- ▶ reaction time and accuracy are recorded;
- ▶ the stimulus can be presented either auditorily or visually.

snarp

# Experimental Tasks: Lexical Decision

Lexical decision task:

- ▶ either word or non-word is presented to participants;
- ▶ they have to decide as quickly as possible whether the stimulus is a word by pressing the right button;
- ▶ reaction time and accuracy are recorded;
- ▶ the stimulus can be presented either auditorily or visually.

mogature

# Experimental Tasks: Lexical Decision

Lexical decision task:

- ▶ either word or non-word is presented to participants;
- ▶ they have to decide as quickly as possible whether the stimulus is a word by pressing the right button;
- ▶ reaction time and accuracy are recorded;
- ▶ the stimulus can be presented either auditorily or visually.

tresk

# Incrementality

Word recognition is *incremental* Marslen-Wilson (1987): humans need not hear or read the full word before recognition occurs.

Evidence for incrementality:

- ▶ *gating task*: recognition occurs when the prefix heard uniquely identifies the word (e.g., trespass, orange): *recognition point*;
- ▶ *lexical decision task*: reaction time for non-words is approximately constant from first non-word phoneme or letter (e.g., tres**k**, orans**o**).

# Frequency Effects

Experiments find robust frequency effects in word recognition:

- ▶ frequent words are easier to recognize, as measured by reaction time (RT) and accuracy;
- ▶ effects found in many tasks, including lexical decision and word identification;
- ▶ also in reading, frequent words receive shorter fixations and are skipped more often;
- ▶ log frequency (or rank frequency) correlates better with RT than raw frequency.

# Neighborhood Effects

*Neighborhood density* ( $N$ ) is also an important predictor of RT:

- ▶ intuition: number of words that are similar to the target word;
- ▶ often defined as the number of words that differ by one letter from the target word.

Effects of neighborhood density in word recognition:

- ▶ identification: higher  $N \Rightarrow$  more difficulty (often described as *competition*);
- ▶ lexical decision: higher  $N \Rightarrow$  less difficulty for words, more difficulty for non-words.

Opposite effects in different tasks are difficult for many models.



# Context Effects

Word recognition is influenced by *context*: words can be recognized sooner in context than in isolation:

- (1) Do you want some fish and chi-
- (2) Did you give the toys to the chi-

Evidence for context effects in reading, gating, and lexical decision.

How do bottom-up (acoustic or visual) and top-down (contextual) information interact during the recognition process?

# Bayesian Reader

Bayesian Reader (Norris, 2006) aims to explain why context, frequency, and neighborhood density affect word recognition:

- ▶ recognition is broken down into word identification, lexical decision, and semantic categorization;
- ▶ hypothesis: word recognition is based on Bayesian inference;
- ▶ frequency and context are assumed to affect the *prior distribution* over words.

Norris, 2006 explores the predictions of this hypothesis for visual word recognition.

# Bayesian Reader

Basic idea: RT is inversely related to the posterior probability of word  $W_i$  given the observed input data  $I$ :

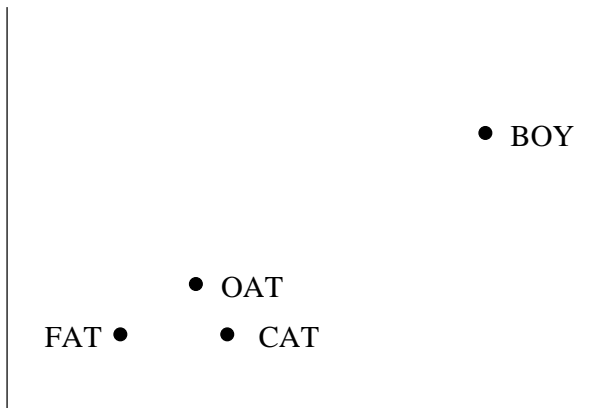
$$P(W_i|I) = \frac{P(I|W_i)P(W_i)}{P(I)}$$

It follows that:

- ▶ increasing  $P(W_i)$  (frequency, context) increases  $P(W_i|I)$ ;
- ▶ increasing  $P(I)$  (neighborhood density) decreases  $P(W_i|I)$ ;
- ▶ increasing  $P(I|W_i)$  (time available, lighting level, perceptual noise) increases  $P(W_i|I)$ .

# Representation

The model represents words as points in a multi-dimensional space.



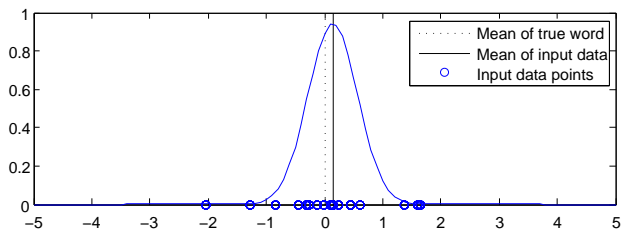
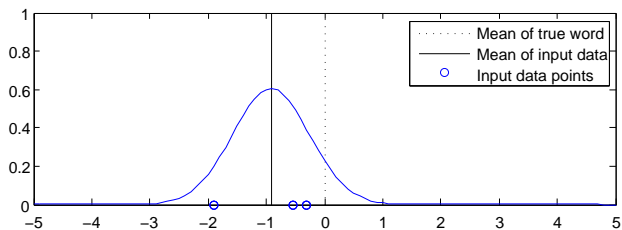
# Likelihood

Input data is assumed to consist of discrete points (perceptual samples), normally distributed around the true word:

- ▶ at each time step, a single sample is observed;
- ▶ the goal of recognition is to identify the word, i.e., estimate the mean of distribution;
- ▶ as more samples accumulate, the estimate will improve,  $P(I|W_i)$  will become low for all but the true word.

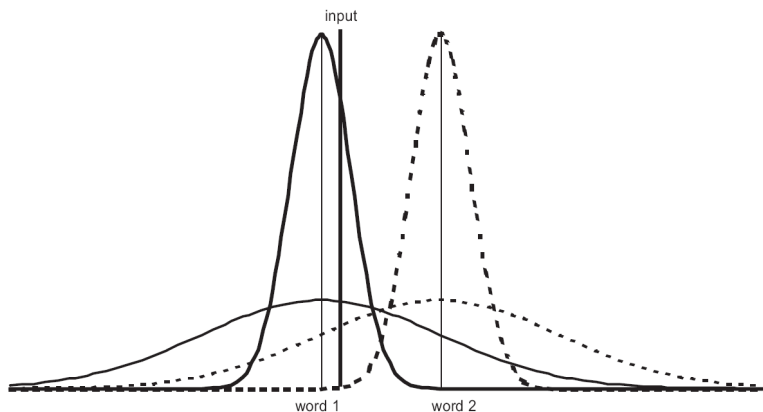
# Likelihood

Top: early in processing. Bottom: later in processing.



# Likelihood

When enough samples have accumulated, a decision can be taken:



# Prior

Bayesian Reader models word recognition in isolation, so the prior  $P(W_i)$  is computed based on frequency counts. However, Norris mentions other possibilities:

- ▶ could count the number of different contexts word occurs in;
- ▶ could use the age of acquisition of a word to estimate the prior (people can accurately estimate this; norming data available).

Also, word frequencies may differ across experiments, so maybe we shouldn't just use corpus frequencies.



# Implementation

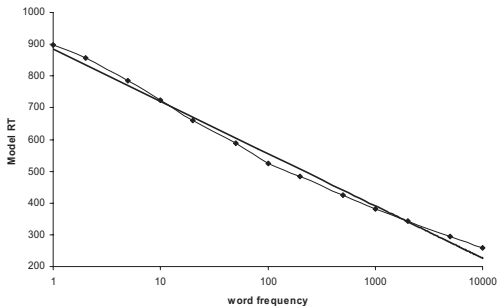
## Implementation of Bayesian Reader:

- ▶ computations implemented using a neural network (though Norris is not committed to this choice);
- ▶ each letter is represented as a 26-dimensional vector, words are concatenations of letters;
- ▶ realistically large vocabulary with corpus frequency counts;
- ▶ input samples accumulate, one per unit time;
- ▶ a simulated response occurs when  $P(W_i|I) > .95$ .

# Results

Posterior probability computed by Bayesian Reader predicts reaction time for word identification:

- ▶ reaction time correlates almost perfectly with log frequency;
- ▶ RT longer for words in larger neighborhoods (due to competition).



But: what about lexical decision?

## Lexical Decision

*Key insight:* lexical decision does not require identifying any particular word. Assume *word* indicates a word/non-word decision:

$$P(\text{word}|I) \propto P(I|\text{word})P(\text{word})$$

In lexical decision experiments,  $P(\text{word}) = 0.5$ . To compute  $P(I|\text{word})$ , sum over hypotheses:

$$\begin{aligned} P(I|\text{word}) &= \sum_{i=1}^n P(I|\text{word}, W_i)P(W_i|\text{word}) \\ &= \sum_{i=1}^n P(I|W_i)P(W_i) \end{aligned}$$

$P(I|\text{non-word})$  can be computed similarly.

# The Effect of Neighborhood Size $N$

## Word recognition:

- ▶ requires identifying a specific word hypothesis (MAP estimation);
- ▶ if many hypotheses cause similar input, more evidence is required to discriminate;
- ▶ therefore, larger  $N$  slows recognition time.

## Lexical decision:

- ▶ prediction does not require identifying any specific word hypothesis (sum over hypotheses);
- ▶ if many hypotheses cause similar input, higher probability that at least one of them is right, so  $P(\text{word})$  is higher;
- ▶ therefore, larger  $N$  speeds “yes” decision, slows “no” decision.



# Discussion

- ▶ Bayesian Reader correctly predicts frequency and neighborhood effects on RT in identification and lexical decision and explains previously puzzling opposite effects of  $N$ ;
- ▶ the model incorporates top-down (prior) and bottom-up (likelihood) information in word recognition;
- ▶ it makes additional predictions that haven't been tested yet:
  - ▶ context can affect recognition both positively and negatively (increased or decreased prior);
  - ▶ degraded input (lighting, visual noise) will slow recognition;
- ▶ but: the model can't explain why in lexical decision for *spoken* words, larger  $N$  slows "yes" decision.

# Summary

- ▶ Word recognition is affected by frequency and number of similar words;
- ▶ Bayesian model provides a rational explanation of frequency and neighborhood effects;
- ▶ assumptions: spatial representation of words, input accumulates over time;
- ▶ visual lexical decision does not require word identification;
- ▶ novel predictions for context effects and degraded input;
- ▶ problems reconciling with spoken word recognition.

# References

-  Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25, 71–102.
-  Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113(2), 327–357.