LSA.308 Computational Psycholinguistics, Class 4

Discussion of Norris (2006): *The Bayesian Reader*

17 July 2007

1 Introduction

The basic observation: Frequent words are recognized more quickly than infrequent words. True in (at least):

- lexical decision
- naming
- semantic classification
- perceptual identification
- eye-fixation in reading

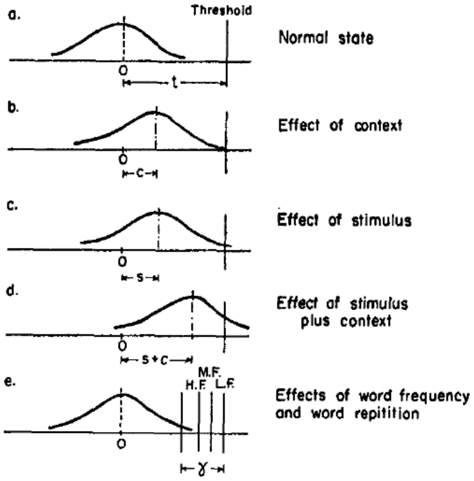
Other concomitant observations:

- It is log frequency, not absolute frequency, that seems to best explain recognition times
- *Neighborhood density* affects both decision and identification tasks, but in different ways
- Neighborhood density interacts with frequency

2 Previous Theories

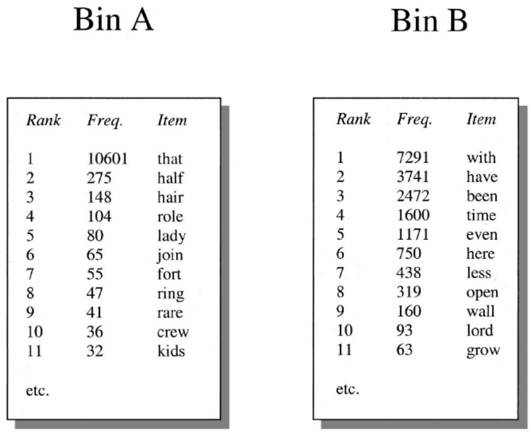
2.1 Logogen Model (Morton, 1969)

- Word representations have to be pushed beyond a certain threshold to be activated
- More frequent words have representations starting closer to threshold
- Could be thought of as a model of skill acquisition (invoking absolute frequency)



2.2 Search Models (Forster, 1976; Murray and Forster, 2004)

- The lexicon is organized into a set of *bins*
- Each bin has a set of words sharing superficial orthographic characteristics
- The words in a bin are ordered by *relative* frequency
- For a given stimulus, the bin is chosen via a hash code
- The bin is then searched serially for a match to the stimulus



Bin A

Rank	*Freq.*	*Item*
1	10601	that
2	275	half
3	148	hair
4	104	role
5	80	lady
6	65	join
7	55	fort
8	47	ring
9	41	rare
10	36	crew
11	32	kids
etc.		

Bin B

Rank	*Freq.*	*Item*
1	7291	with
2	3741	have
3	2472	been
4	1600	time
5	1171	even
6	750	here
7	438	less
8	319	open
9	160	wall
10	93	lord
11	63	grow
etc.		

Figure 1. Hypothetical contents at the top of two bins in a search model. Freq. = frequency.

2.3 Critiques of previous models

- *How* explanations versus *why* explanations
- "levels of analysis" (Marr, 1982; Anderson, 1990): computational, algorithmic, implementational
  - The *computational* level focuses on the goal of the computation

  - The *algorithmic* level focuses on the representations and procedures required to achieve the goal of computation
  - The *implementational* level focuses on how these representations and procedures can be physically realized

- *proximate* versus *ultimate* cause in biology (Mayr, 1961)

  Why does a warbler fly south from New Hampshire one day in late August?

  - *proximate* causes (*functional* biology): the warbler is sensitive to the number of hours of daylight in the day, and to the ambient temperature, which passed a critical threshold
  - *ultimate* causes (*evolutionary* biology): warblers need to eat insects and are thus under selective pressure to move to where insects are plentiful; and this selective pressure caused the bird's genetic constitution to evolve such that the bird migrates south in winter

2.4 Ideal Observer Analysis

- The "ideal observer" is one that makes optimal use of the available information
- Ideal observer analysis requires a precise specification of a task and the available perceptual input
- Bayes' rule is natural in ideal observer analysis because it specifies the result of the complete use of all available information
- "the primary objective of an ideal observer is to compute the probability of each possible true state of the environment given the stimulus" (Geisler and Kersten, 2002)
- Classic example: inference about probability of having a disease based on the outcome of a test, using
  - P(Test Outcome|Disease)
  - P(Test Outcome|¬Disease)
  - P(Disease)
- ***How do differential reaction times come about in the Bayesian Reader's ideal observer analysis?

  The observer is *uncertain* about the match between perceptual input and the correct word. The observer has the goal of correctly satisfying the task as quickly as possible but to some minimum accuracy threshold. The faster the threshold is reached, the faster the RT.

- ***How does the simple disease example relate to frequency effects in the Bayesian Reader?

  The frequency of ±disease corresponds to word/non-word frequency; the likelihood terms correspond to distance of wordforms from input vector

3 The Bayesian Reader

- Three main assumptions behind Norris's Bayesian Reader:
  - Each word in the lexicon W_i is represented as a point in a Euclidean space of dimension $n = 26L$
  - Perception involves repeated sampling from a symmetric Gaussian distribution centered on the true word
  - $P(I|W)$ can always be computed from an estimate of the input distribution's variance [n.b.—this may not be necessary]

- The Bayesian Reader infers posterior beliefs about the input word through Bayes rule:

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

  This is a form of *Bayesian hypothesis testing*!

  Given the other assumptions, we get:

$$P(I|W_i) = \prod_{j=1}^{t} \left(\frac{1}{2\pi\sigma_i^2}\right)^{n/2} \exp\left[-\frac{D(s_j, W_i)}{2\sigma_i^2}\right]$$

  where σ_i^2 is the Gaussian variance and $D(s_j, W_i)$ is the Euclidean distance between the j-th sample and the true position of W_i.

- The input I consists of a sequence of observation samples $s_{1\cdots t}$. Note that the resulting likelihood function $P(I|W_i)$ is **not** obviously what Norris uses in Appendix A! Can you spot the difference?

  Answer: In Appendix A, Norris writes the data likelihood as

$$P(\mu|W_i) = \big(\frac{1}{2\pi\sigma_M^2}\big)^{n/2} \exp\big[-\frac{D(\mu, W_i)}{2\sigma_M^2}\big]$$

  or equivalently (for inference of posterior word probabilities)

$$P(\mu|W_i) = \exp\big[- \frac{D(\mu, W_i)}{2\sigma_M^2} \big]$$

that is, paying attention only to the mean and standard error of the input sample. This seems different!

The link between what Norris does and the "proper" probabilistic inference can be made as follows. If we express the correct likelihood in log-space we have

$$L(I|W_i) \equiv \log(I|W_i) = \frac{nt}{2} \log\big(\frac{1}{2\pi\sigma_i^2}\big) - \sum_{j=1}^{t} \frac{D(s_j, W_i)}{2\sigma_i^2}$$

We can partition the sum of distances $D(s_j, W_i)$ as follows:[1]

$$\sum_{j=1}^{t} \frac{D(s_j, W_i)}{2\sigma_i^2} \quad = \quad \frac{1}{2\sigma_i^2} \left[t(\mu - W_i)^2 + \sum_{j=1}^{t} (s_j - \mu)^2 \right]$$

Plugging this back in to the likelihood gives us

$$P(s_{1\cdots t}|W_i) = \left(\frac{1}{2\pi\sigma_i^2} \right)^{nt/2} \exp\left[- \sum_{j=1}^{t} (s_j - \mu)^2 \right] \exp\left[- \frac{t(\mu - W_i)^2}{2\sigma_i^2} \right]$$

[1] The more expanded version of this is as follows:

$$\begin{aligned}
\sum_{j=1}^{t} D(s_j, W_i) &= \sum_{j=1}^{t} (s_j - W_i)^2 \\
&= \sum_{j=1}^{t} (s_j - \mu + \mu - W_i)^2 \\
&= \sum_{j=1}^{t} (s_j - \mu)^2 + 2(s_j - \mu)(\mu - W_i) + (\mu - W_i)^2 \\
&= t(\mu - W_i)^2 + \sum_{j=1}^{t} (s_j - \mu)^2 + 2(s_j - \mu)(\mu - W_i) \\
&= t(\mu - W_i)^2 + \sum_{j=1}^{t} (s_j - \mu)^2
\end{aligned}$$

The last term in line 4 can be dropped because the mean of $s_j - \mu$ is 0—think carefully about this if it's not obvious. You can also find this partitioning of the variance in many statistics books.

Now, we use the likelihood only for calculating the posterior probability of a word. In this case we always have likelihood terms in both the numerator and denominator of the expression. *If we assume that the variance σ_i is the same for all words W_i,* then the first two terms of 3 are independent of the choice of word. As a result, we can treat the likelihood as simply being

$$P(s_{1\dots t}|W_i) \propto \exp\left[-\frac{t(\mu - W_i)^2}{2\sigma_i^2}\right]$$

Finally, we can substitute in the "standard error of the mean" $\sigma_M = \sigma_i/\sqrt{t}$ for a sample of size t, giving us

$$P(s_{1\dots t}|W_i) \propto \exp\left[-\frac{(\mu - W_i)^2}{2\sigma_M^2}\right]$$

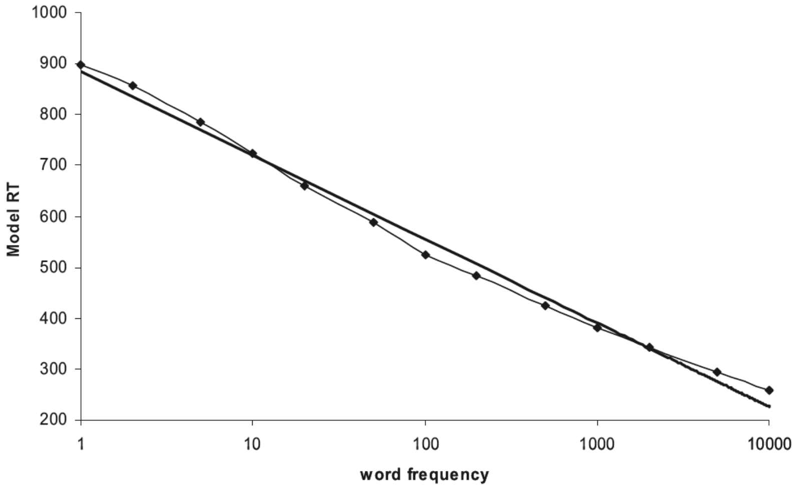
giving us the expression used by Norris. IMHO, Norris (2006) would have been clearer if it had included this reasoning in Appendix A.

- There is one more crucial parameter setting that Norris doesn't seem to include in the paper. Did you notice what it is?

  Answer: The standard deviation σ for $P(s_j|W_i)$ doesn't seem to appear in the paper.

3.1 Word frequency in identification

The relationship between word frequency and average identification time automatically comes out as logarithmic:



3.2 Word frequency in lexical decision

- What does Norris claim is the relevant posterior probability for lexical decision? How does it differ from identification?

  It's

  $$P(\text{Word}|I) \\ = \frac{\sum_i (W_i|I)}{\sum_i (W_i|I)+P(\text{NonWord}|I)}$$

  Crucially, this expression *marginalizes* over the words in the lexicon (as well as the non-words).

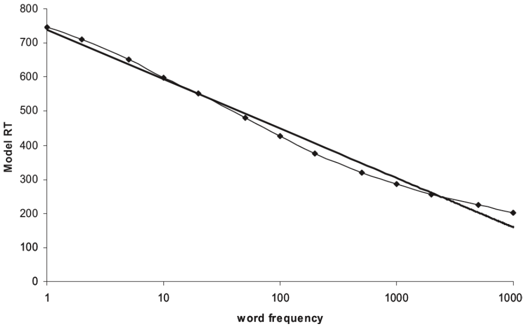
- What would make sense for the probability of non-word given the input? How does Norris define it?

  The simplest thing to do would just be to say

  $$P(\text{NonWord}|I) \propto \sum_i P(I|\text{NonWord}_i)P(\text{NonWord}_i)$$

  and to define a uniform distribution over all non-words in the lexicon. Norris does something convoluted instead: he divides probability mass from non-words into a "virtual word" that is nearby the input, and "background non-words" that are far away. IMHO he doesn't do a very good job justifying this division or clarifying exactly how background non-words work.

- Log-frequency explains RTs for lexical decision as well.



3.3 Neighborhood Effects

- Intuitively, how should effects of neighborhood density differ for identification and decision in the Bayesian Reader?

  High neighborhood density should facilitate decision, due to the marginization over words, and hinder identification

- How should frequency interact with neighborhood effects?

  For decision, neighborhood density should have stronger effects for low-frequency than for high-frequency words. The situation is less clear with identification.

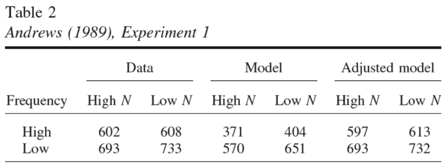


Table 2
Andrews (1989), Experiment 1

	Data		Model		Adjusted model	
Frequency	High *N*	Low *N*	High *N*	Low *N*	High *N*	Low *N*
High	602	608	371	404	597	613
Low	693	733	570	651	693	732

Figure 1: Lexical decision results

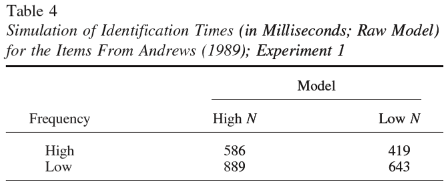
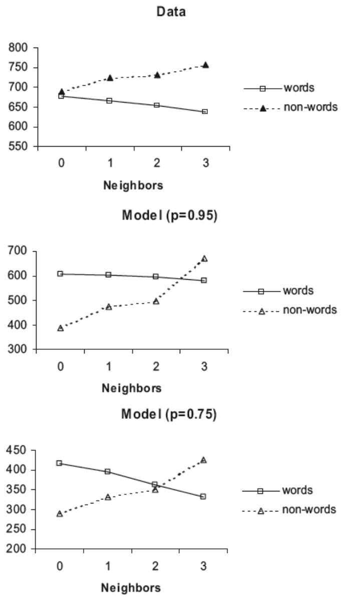


Table 4
Simulation of Identification Times **(in Milliseconds; Raw Model)**
for the Items From Andrews (1989); **Experiment 1**

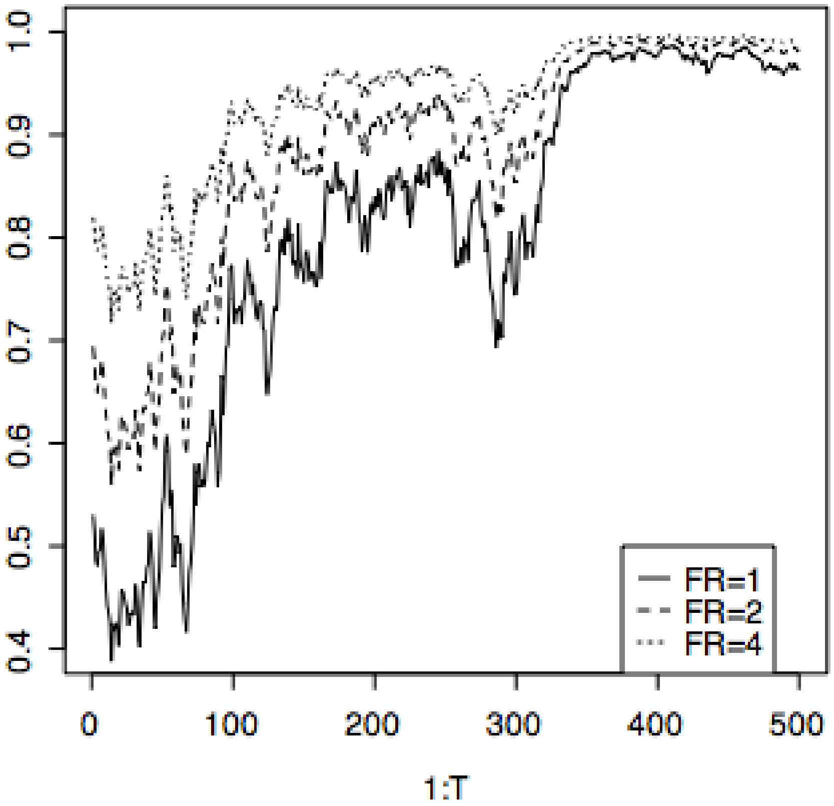
	Model	
Frequency	High *N*	Low *N*
High	586	419
Low	889	643

Figure 2: Identification results

- How should neighborhood density affect lexical decision for non-words?



4 A bit of play: why log frequency?



5 Conclusions & Discussion

- Norris's Bayesian Reader is a *computational* theory of visual word recognition that accounts for effects of frequency and neighborhood density in the context of task-specific goals
- The generative probabilistic model is extremely simple (though is it ever awkward?)
- Focuses on *discrimination* as the source of frequency-sensitive effects on reaction time—contrast with the story of surprisal for sentence processing?

References

Anderson, J. R. (1990). *The Adaptive Character of Human Thought*. Lawrence Erlbaum.

Forster, K. I. (1976). Assessing the mental lexicon. In Wales, R. and Walker, E., editors, *New approaches to language mechanisms*, pages 257–287. Amsterdam: North-Holland.

Geisler, W. S. and Kersten, D. (2002). Illusions, perception and bayes. *Nature Neuroscience*, 5(6):508–510.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mayr, E. (1961). Cause and effect in biology. *Science*, 134(3489):1501–1506.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2):165–178.

Murray, W. S. and Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111(3):721–756.