Word recognition

Previously, we examined Cohort (Marslen-Wilson 1987), a mechanistic model of spoken word recognition.

Psychologists are also interested in visual word recognition, i.e. reading.

Both relate to questions of lexical access discussed by Jurafsky (1996).

Recurring themes: top-down vs. bottom-up processing, frequency effects.

Today: a Bayesian view of lexical access.

Cohort model was designed in light of evidence that

- word candidates that are inconsistent with context are active early in recognition (bottom-up activation).
- recognition is faster for contextually appropriate words (early selection).

However, Cohort

- cannot explain effects of frequency or neighborhood density.
- fails to recognize words out of context or in noise.
Bayesian approach

Step away from mechanistic explanations, consider why frequency and context affect recognition as they do.
- Hypothesis: word recognition is an optimal Bayesian decision process.
- Frequency and context affect the prior distribution over words.
Norris (2006) explores this hypothesis for visual word recognition.

Frequency effects

Psychologists 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 identification.
- Effects found in both spoken and visual recognition.
- Log frequency (or rank frequency) correlate much better with RT than raw frequency.

Norris (2006)

Basic idea (also see Jurafsky 1996): RT is inversely related to the posterior probability of word $W_i$ given the observed input data $I$:

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

- 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, lighting) increases $P(W_i | I)$.

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 character (phoneme) from the target word.

Effects of neighborhood density in visual 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.
Norris’s model represents words as points in a multi-dimensional space.

- BOY
- OAT
- FAT
- CAT

Input data is assumed to consist of discrete points, normally distributed around the true word.

- At each time step, a single data point is observed.
- Goal of recognition: identify word, i.e. estimate mean of distribution.
- As more samples accumulate, estimate will improve, $P(I|W_i)$ will become low for all but the true word.
Norris models recognition in isolation, so computes $P(W_i)$ based on frequency counts. However, mentions other possibilities:
- Number of different contexts word occurs in.
- Age of acquisition.
Also, notes that frequencies may differ in experimental situations.

- Implemented using a neural network (other methods possible).
- Each letter is represented as a 26-dimensional vector, words as concatenations or letters.
- Realistically large vocabulary with corpus frequency counts.
- Input samples accumulate, one per unit time.
- Simulated response occurs when $P(W_i | I) > .95$ (or .99).

**Key insight:** lexical decision does not require identifying any particular word.

$$P(wd | I) \propto P(I | wd) P(wd)$$

In experiment, $P(wd) = .5$. To compute $P(I | wd)$, sum over hypotheses:

$$P(I | wd) = \sum_{i=1}^{n} P(I | wd, W_i) P(W_i | wd)$$

$$= \sum_{i=1}^{n} P(I | W_i) P(W_i | wd)$$

$P(I | non-wd)$ can be computed similarly.
**Intuition**

**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 $P(wd)$ is higher.
- Therefore, larger $N$ speeds “yes” decision, slows “no” decision.

**Spoken word recognition**

Most effects are similar to visual recognition, but in lexical decision, larger $N$ slows “yes” response.

Speculation:
- Spoken recognition is more basic/ecologically valid.
- Lexical decision is not very natural.
- Speech system is adapted for identification, cannot “turn off” identification system.
- Reading system is less highly adapted, more flexible for different tasks.

But danger of post-hoc explanations.

**Discussion**

- Model correctly predicts frequency and neighborhood effects on RT in identification and lexical decision tasks and explains previously puzzling opposite effects of $N$.
- Model incorporates top-down (prior) and bottom-up (likelihood) information, but does not suggest bottom-up activation.
- Additional predictions, not yet tested:
  - Context can affect recognition both positively and negatively (through prior).
  - Degraded input will slow recognition – quantitative predictions.
- What about spoken word recognition?

**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.
- Qualitative predictions for context effects and degraded input are sensible, quantitative predictions are untested.
- Problems reconciling with spoken word recognition.
<table>
<thead>
<tr>
<th>References</th>
</tr>
</thead>
</table>