CS674, Spring 2002
Lecture Handout, 2/20/02

1. "Lack of evidence is not evidence of lack."

- 14% of the word trigrams in a 350 million word corpus occur only once (Brown et al., 1992).
- 58% of the types in the Brown corpus occur only once.
- 95% of a standard prepositional-phrase attachment test set is unseen with respect to the corresponding training set (Collins and Brooks, 1995).

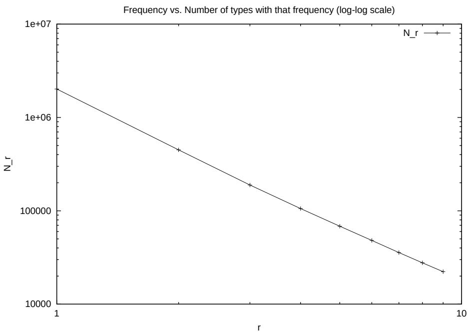
2. Good-Turing re-estimation of word bigram counts, from 22 million words of AP newswire. The data comes from Church and Gale (1991).

Recall that $N_r = \sum_{w \in \Sigma} \delta(\#(w) = r)$.

2**(a)** Figure 6.10 of Jurafsky and Martin (2000) (version here produced by slightly editing `http://www.cs.colorado.edu/~martin/SLP/Figures/fig06.10.ps` to match our notation).

r (MLE)	N_r	r^* (GT)
0	74,671,100,000	0.0000270
1	2,018,046	0.446
2	449,721	1.26
3	188,933	2.24
4	105,668	3.24
5	68,379	4.22
6	48,190	5.19
7	35,709	6.21
8	27,710	7.24
9	22,280	8.25

2**(b)** The N_r data plotted on a log-log scale. Note: this is *not* a strict Zipf's-law rank vs. frequency plot.



3. Notation for deleted estimation.

$\mu_r = p_{(r)} N_r$: total probability mass for count-r words

heldcount$_r = \sum_w \delta(\#_t(w) = r) \cdot \#_h(w)$: total count in the *held-out* data for words that occur r times in the *training data*.