- MLE refers to a model that assigns the highest probability to the data of all models in a family. RF is the MLE for unigram and other n-gram models.
- MLE is bad for unseen (too low) and rare words (too hight) when used to predict the future.
- Laplace (add-one, or generally add-k) avoids zero probabilities, but performs poorly.
- Good-Turing gives a good estimate for unseen words and rare words, but can't be used directly on common words.
- GT can be made practical with SGT, using regression analysis to fit a log-log line.
- Alternately, GT can be made practical by using [an adapted form of] it up to some k.
- Alternately, combine language models with back-off.

Alternately again, take k language models and interpolate them:

$$P_{LI}(w) = \sum_{j=1}^{k} \lambda_j P_j(w)$$

Example: combining uni-, bi-, and trigrams (book version):

$$P_{LI}(w_n \mid w_{n-2}w_{n-1}) = \lambda_1 \cdot P_{MLE}(w_n \mid w_{n-2}w_{n-1}) + \lambda_2 \cdot P_{MLE}(w_n \mid w_{n-1}) + \lambda_3 \cdot P_{MLE}(w_n)$$

Example: combining unigram MLE and constant (compare Laplace):

$$P_{LI}(w) = \lambda \cdot P_{MLE}(w) + (1 - \lambda)\frac{1}{v}$$

Example: combining MLE and GT (compare Katz's cut off):

$$P_{LI}(w) = \lambda \cdot P_{MLE}(w) + (1 - \lambda) \cdot P_{GT}(w)$$