

(t-score)是計算語言學中常用的統計顯著性的檢定(statistical significance test)，也是 Ken Church (1991)與他的同事率先運用在計算語言學，通常與互見訊息搭配一起使用。T-值與標準差和信賴區間(confidence interval)密切相關。當語料庫夠大時，而 T-值大於 1.65 時表示有 95%的信心證明差異是存在。

T 值的計算可以採用下列簡化的公式。

$$t \approx \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}}$$

三、熵(entropy):

熵(entropy)也是訊息理論中的基本概念，1960 年代被用於資訊檢索，1990 後被廣泛運用到計算語言學中。熵可視為是測量驚奇，不確定性，或訊息的量。公式如下。

$$\text{entropy} = -k \sum p_i \log(p_i)$$

四、N 連詞(ngram)語言模型(language model)是是統計式計算語言學中最簡單也最常用的語言模型。N 連詞語言模型假設一個句子第 N 個詞的機率可以由前面的 N-1 個詞決定。雖然這個假設過於簡化語言的複雜性，但在語音辨識與其它應用上不失為一個有效的方法。最常用的是二連詞(bigram)與三連詞(trigram)模型。