January 23, 2018

N**:** Text length, i. e. number of tokens

$V(N)$**:** Vocabulary size, i. e. number of types

$V(i, N)$**:** Number of types occurring i times

1 Measures that use sample size and vocabulary size

$$\text{type-token ratio} = \frac{V(N)}{N}$$

$$\text{Guiraud's } R = \frac{V(N)}{\sqrt{N}}$$

$$\text{Herdan's } C = \frac{\log(V(N))}{\log(N)}$$

$$\text{Dugast's } k = \frac{\log(V(N))}{\log(\log(N))}$$

$$\text{Maas' } a^2 = \frac{\log(N) - \log(V(N))}{\log(N)^2}$$

$$\text{Dugast's } U = \frac{\log(N)^2}{\log(N) - \log(V(N))}$$

$$\text{Tuldava's } LN = \frac{1 - V(N)^2}{V(N)^2 \log(N)}$$

$$\text{Brunet's } W = N^{V(N)^{-a}} \text{ with } a = -0.172$$

$$\text{Carroll's } CTTR = \frac{V(N)}{\sqrt{2N}}$$

$$\text{Summer's } S = \frac{\log(\log(V(N)))}{\log(\log(N))}$$

2 Measures that use part of the frequency spectrum

$$\text{Honoré's } H = 100\frac{\log(N)}{1 - \frac{V(1,N)}{V(N)}}$$

$$\text{Sichel's } S = \frac{V(2,N)}{V(N)}$$

$$\text{Michéa's } M = \frac{V(N)}{V(2,N)}$$

3 Measures that use the whole frequency spectrum

$$\text{Entropy} = \sum_{i=1}^{N} V(i,N)\left(-\log(\frac{i}{N})\right)\frac{i}{N}$$

$$\text{Yule's } K = 10^4\left(-\frac{1}{N} + \sum_{i=1}^{N} V(i,N)\left(\frac{i}{N}\right)^2\right)$$

$$\text{Simpson's } D = \sum_{i=1}^{V(N)} V(i,N)\frac{i}{N}\frac{i-1}{N-1}$$

$$\text{Herdan's } V_m = \sqrt{-\frac{1}{V(N)} + \sum_{i=1}^{V(N)} V(i,N)\left(\frac{i}{N}\right)^2}$$

$$\text{McCarthy and Jarvis' } \textit{HD-D} = \sum_{i=1}^{V(N)} \frac{1}{42}\left(1 - \frac{\binom{i}{0}\binom{N-V(i,N)}{42-0}}{\binom{N}{42}}\right) = \sum_{i=1}^{V(N)} \frac{1}{42}\left(1 - \frac{\binom{N-V(i,N)}{42}}{\binom{N}{42}}\right)$$