January 22, 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

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

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

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

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

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

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

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

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

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

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