R

# relativeFrequency 相対頻度について

## サンプルデータ　１００万語

http://www.thegrammarlab.com/?nor-portfolio=1000000-word-sample-corpora

http://micusp.elicorpora.info/

MICUSP Sample

http://www.thegrammarlab.com/?nor-portfolio=1000000-word-sample-corpora#

## 全体の高頻度語

## 1，１０，１００，１０００位の語と頻度

### ■ topfeatures() の結果の中身

・names 属性のついた数値ベクトル
　・要素番号のように、要素を指定すると、その数値が表示される。
・要素番号の指定もできる。（この場合、要素番号が頻度順位となる）

## チャンク内の検索と頻度

### ■ "the" の頻度分布

### ■ "the" の累積頻度分布

### ■１００語ずつ１００万語まで１万回サンプリングしてみるとほぼ直線的に増えていることがわかる。

### ■最初の１００語での出現頻度は、７回。これを１万倍したら 70000。実際は、68659 回。1.02 の精度。

## １０００位の sites について

### ■ sites の頻度分布

### ■ site の累積頻度分布

## 分散を見てみる

### ■ the はどこにおいても均一に出現している

### ■ sites の出現は、偏っている。

### ■ゆえに、sites については、その一部に基づき、相対頻度を出すことは、結果がゆがむ恐れが大きい。

## therefore について

### ■累積頻度