Type と Token

R のパッケージ corpus を参考に、Type と Token の振る舞いを見てみる

http://corpustext.com/articles/corpus.html

■オズの魔法使いのテキストを取ってきて本文だけにする。

```
oz.text <- gsub("\\n", " ", text)
oz.text.nopunct <- gsub("\\W+", " ", oz.text)
oz.words <- strsplit(oz.text.nopunct, "\\W")
oz.words <- unlist(oz.words)
write(oz.words, file="ozWords.txt")
length(oz.words)
```

・39,456 語

```
> head(sort(table(oz.words), decreasing=T), 10)
oz.words
 the  and   to   of    a    I  was   in  you   he
2731 1593 1096  811  795  647  501  463  448  410
```

■ Type と Token の分布を見てみる

・394 行 2 列の行列、0 で初期化

```
> oztt <- matrix(0, nrow=394, ncol=2)
```

・100 語ずつ累積して 39,400 語までの Type と Token を見てみる。

```
i <- 1
y <- 0
while (i <= 394) {
  y <- i * 100
  tmp <- oz.words[1:y]
  oztt[i,1] <- length(tmp)
  oztt[i,2] <- length(unique(tmp))
  i <- i+1
}
```

・データフレーム化して、見出しをつける

```
> oztt <- as.data.frame(oztt)
> colnames(oztt) <- c("token","type")
> head(oztt)
  token type
1   100   63
2   200  122
3   300  167
4   400  209
5   500  248
6   600  280
```

```
>plot(oztt$token, oztt$type)
```

多様性を見てみる

■ TTR を追加する

```
> oztt$ttr <- oztt$type / oztt$token
```

```
> head(oztt)
  token type       ttr
1   100   63 0.6300000
2   200  122 0.6100000
3   300  167 0.5566667
4   400  209 0.5225000
5   500  248 0.4960000
6   600  280 0.4666667
```

・token の増加に伴う type の増加、および、TTR の減少をグラフにプロット

```
> plot(oztt$token, oztt$type)
> par(new=T)
> plot(oztt$token, oztt$ttr, col = "red")
```

■ Guiraud Index（ギロー・インデックス）を追加してみる

```
> oztt$gi <- oztt$type / sqrt(oztt$token)
> par(new=T)
> plot(oztt$token, oztt$gi, col = "blue")
```

・およそ 4000 語を越えたあたりからほぼ水平で意外と安定している。

1000 語までで見てみる

```
ozttt <- matrix(0, nrow=1000, ncol=2)
 i <- 1
 while (i <= 1000) {
   tmp <- oz.words[1:i]
   ozttt[i,1] <- length(tmp)
   ozttt[i,2] <- length(unique(tmp))
   i <- i+1
 }

ozttt <- as.data.frame(ozttt)
colnames(ozttt) <- c("token","type")


cor.test(ozttt$type, ozttt$token)

       Pearson's product-moment correlation

data:  ozttt$type and ozttt$token
t = 234.34, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9898566 0.9920780
sample estimates:
      cor
0.9910356


lm(ozttt$type ~ ozttt$token)
Call:
lm(formula = ozttt$type ~ ozttt$token)

Coefficients:
(Intercept)  ozttt$token
    43.8034       0.3761


#plot(ozttt$token, ozttt$type)
#abline(lm(ozttt$type ~ ozttt$token))

pred1000 <- predict(lm(ozttt$type ~ ozttt$token), interval = "prediction")

    結果を保存した pred1000 のデータををデータフレーム型に変更
pred1000 <- as.data.frame(pred1000)
```

```
    データをプロット
plot(oztttt$token, oztttt$type)

    フィット（回帰直線）を黒で描く
lines(oztttt$token, pred1000$fit, col = "black")

    上限値を赤で描く
lines(oztttt$token, pred1000$upr, col = "red")

    下限値を青で描く
lines(oztttt$token, pred1000$lwr, col = "blue")
```

2 千語までで見てみる

1 万語までで見てみる

```
oztt10t <- matrix(0, nrow=10000, ncol=2)
 i <- 1
 while (i <= 10000) {
   tmp <- oz.words[1:i]
   oztt10t[i,1] <- length(tmp)
   oztt10t[i,2] <- length(unique(tmp))
   i <- i+1
 }

oztt10t <- as.data.frame(oztt10t)
colnames(oztt10t) <- c("token","type")

plot(oztt10t$token, oztt10t$type)

oztt10t$ttr <- oztt10t$type / oztt10t$token

par(new=T)
plot(oztt10t$token, oztt10t$ttr, col = "red")

oztt10t$gi <- oztt10t$type / sqrt(oztt10t$token)
par(new=T)
plot(oztt10t$token, oztt10t$gi, col = "blue")
```