# Zipf's 1st Law

Gerald Penn

CSC 401
University of Toronto

http://www.cs.toronto.edu/∼gpenn/csc401

# Types vs. Tokens

*The* cat in *the* hat

- Token: instance of word (the: 2)
- Type: “kind” of word (the: 1)
- Not clear in other cases:
  - *run* vs. *runs*
  - *happy* vs. *happily*
  - *frágment* vs. *fragmént*
  - *email* vs. *e-mail*
  - *hat* vs. *hat,*

# Corpus (*pl.* Corpora)

A corpus is a collection of text(s) or utterances

- $10^6$: tiny
- $10^9$: reasonable
- $10^{15}$: current feasible limit for unannotated data

# Lexicon

A collection of word-types: like a dictionary, but not necessarily with meanings

# Frequency Statistics

(Term) Frequency

$$TF(w,S) = \ \# \text{ tokens of } w \text{ in corpus } S$$

Relative Frequency:

$$F_S(w) = \frac{TF(w,S)}{|S|}$$

What happens to $F_S(w)$ as $|S|$ grows?
Answer: $F_S(w)$ converges to $p(w)$
This is the *frequentist* view of probability theory.

# Frequency Statistics

(Term) Frequency

$$TF(w, S) = \text{ \# tokens of } w \text{ in corpus } S$$

Relative Frequency:

$$F_S(w) = \frac{TF(w, S)}{|S|}$$

What happens to $F_S(w)$ as $|S|$ and lexicon $|V|$ grow?

# Frequency Statistics

(Term) Frequency

$$TF(w, S) = \text{ \# tokens of } w \text{ in corpus } S$$

Relative Frequency:

$$F_S(w) = \frac{TF(w, S)}{|S|}$$

What happens to $F_S(w)$ as $|S|$ and lexicon $|V|$ grow?
Answer: Average rel. freq. converges to 0.
That means that there are more and more infrequent words.
Not at all unusual for a word to have prob. $10^{-7}$.

# Frequency Statistics

(Term) Frequency

$$TF(w, S) = \text{ \# tokens of } w \text{ in corpus } S$$

Relative Frequency:

$$F_S(w) = \frac{TF(w, S)}{|S|}$$

What happens to $F_S(w)$ as $|S|$ and lexicon $|V|$ grow?
But rel. freq. itself stabilizes — surprise!
Let $N = |S|$:

$$\log(F_r)_V + \log N \approx H_N - B_N \log(\frac{r}{|V|})$$

# The Zipf-Mandelbrot Equation

$$\log(F_r)_V + \log N \approx H_N - B_N \log(\frac{r}{|V|})$$

Line up all of the word types by (rel.) frequency:

| TF(w) | 3000 | 2900 | 1750 | 1700 | ... |
|---|---|---|---|---|---|
| w | the | and | a | to | ... |
| r | 1 | 2 | 3 | 4 | ... |

$r$: rank
$F_r$: the rel. freq. of the $r^{\text{th}}$ ranked word.

$H_N \longrightarrow 0$ because lowest rank word should occur with rel. freq. $\frac{1}{N}$ (*hapax legomenon* — often typos)
But when $B_N \longrightarrow B \neq 0$, then we say that the population is *Zipfian*.

(This assumes $N$ and $|V|$ grow independently.)

# Signficance

1. There are LOTS of infrequent words. For English:
   - top 31: 36%
   - top 150: 43%
   - top 256: 50%

   For Hungarian: top 4096: 50%. (why?)
2. There are distributions "in the world" that are hyperbolic:
   - Zipf (prob. thought $B = 1$)
   - Pareto distributions
   - Yule's Law: $B = 1 + \frac{g}{s}$
   - Champernowne's Ergodic Wealth Distribution Model

# Linguistic Signficance

1. There are distributions "in the world" that are hyperbolic:
   - Simon's discourse model (1956):
     - people imitate rel. freq's of word-types they hear
     - people innovate new words with small but constant prob.
   - ... but Mandelbrot's monkey model (1961) cast doubt on that.
   - Many other connections: age, polysemy, length, etc. of words