Word Distributions

• Words are not distributed evenly!
  – Same goes for letters of the alphabet (ETAOIN SHRDLU), city sizes, wealth, etc.

• Usually, the 80/20 rule applies
  – 80% of the wealth goes to 20% of the people or it takes 80% of the effort to build the easier 20% of the system
  – more examples coming up...
Introduction to NLP

Word Distributions
Shakespeare

- Romeo and Juliet:
  - And, 667; The, 661; I, 570; To, 515; A, 447; Of, 382; My, 356; Is, 343; That, 343; In, 314; You, 289; Thou, 277; Me, 262; Not, 257; With, 234; It, 224; For, 223; This, 215; Be, 207; But, 181; Thy, 167; What, 163; O, 160; As, 156; Her, 150; Will, 147; So, 145; Thee, 139; Love, 135; His, 128; Have, 127; He, 120; Romeo, 115; By, 114; She, 114; Shall, 107; Your, 103; No, 102; Come, 96; Him, 96; All, 92; Do, 89; From, 86; Then, 83; Good, 82; Now, 82; Here, 80; If, 80; An, 78; Go, 76; On, 76; I'll, 71; Death, 69; Night, 68; Are, 67; More, 67; We, 66; At, 65; Man, 65; Or, 65; There, 64; Hath, 63; Which, 60;
  - ...
  - A-bed, 1; A-bleeding, 1; A-weary, 1; Abate, 1; Abbey, 1; Abhorred, 1; Abhors, 1; Aboard, 1; Abound' st, 1; Abroach, 1; Absolved, 1; Abuse, 1; Abused, 1; Abuses, 1; Accents, 1; Access, 1; Accident, 1; Accidents, 1; According, 1; Accursed, 1; Accustom'd, 1; Ache, 1; Aches, 1; Aching, 1; Acknowledge, 1; Acquaint, 1; Acquaintance, 1; Acted, 1; Acting, 1; Action, 1; Acts, 1; Adam, 1; Add, 1; Added, 1; Adding, 1; Addle, 1; Adjacent, 1; Admired, 1; Ado, 1; Advance, 1; Adversary, 1; Adversity's, 1; Advise, 1; Afeard, 1; Affecting, 1; Afflicted, 1; Affliction, 1; Affords, 1; Affray, 1; Affright, 1; Afire, 1; Agate-stone, 1; Agile, 1; Agree, 1; Agrees, 1; Aim'd, 1; Alderman, 1; All-cheering, 1; All-seeing, 1; Alla, 1; Alliance, 1; Alligator, 1; Allow, 1; Ally, 1; Although, 1;

The BNC (Adam Kilgarriff)

Kilgarriff, A. Putting Frequencies in the Dictionary.
Stop Words

• Fact:
  – 250–300 most common words in English account for 50% or more of a given text.

• Example:
  – “the” and “of” represent 10% of tokens. “and”, “to”, “a”, and “in” – another 10%. Next 12 words – another 10%.

• Moby Dick Ch.1:
  – 859 unique words (types), 2256 word occurrences (tokens). Top 65 types cover 1132 tokens (> 50%).

• Token/type ratio:
  – 2256/859 = 2.63
Power-law Distribution

- **Power-law**
  - Many words with a small frequency of occurrence
  - A few words with a very large frequency
  - High skew (asymmetry)

- **Comparing to a normal distribution:**
  - Many people of a medium height
  - Almost nobody of a very high or very low height
  - Symmetry

Slide from Qiaozhu Mei
Scaling the Axes

- Long-tail on a linear scale - straight line on a log-log plot
Power Law Distribution

- The probability of observing an item of size ‘x’ is given by
  \[ p(x) = C x^{-\alpha} \]
  - normalization constant (probabilities over all \( x \) must sum to 1)
  - \( \alpha \) : scaling exponent, or power law exponent

- Straight line on a log-log plot

  \[ \ln(p(x)) = c - \alpha \ln(x) \]
Power Laws Are Seemingly Everywhere

- Moby Dick
- scientific papers 1981-1997 AOL users visiting sites ‘97
- bestsellers 1895-1965
- AT&T customers on 1 day
- California 1910-1992

Zipf's law is fairly general!

- Frequency of accesses to web pages
  - in particular the access counts on the Wikipedia page, with $s$ approximately equal to 0.3
  - page access counts on Polish Wikipedia (data for late July 2003) approximately obey Zipf's law with a slope $s$ about 0.5
- Words in the English language
  - for instance, in Shakespeare’s play Hamlet with $s$ approximately 0.5
- Sizes of settlements
- Income distributions amongst individuals
- Size of earthquakes
- Notes in musical performances

http://www.cut-the-knot.org/do_you_know/zipflaw.shtml
Another Way to Plot: Zipf’s Distribution

\[ p(k) \sim k^{-\alpha} \]
**Zipf's Law in Natural Language**

\[ \text{Rank} \times \text{Frequency} \approx \text{Constant} \]

- Constant \( \approx 0.1 \times \text{Length of collection (in words)} \)
- Not accurate at the tails, but accurate enough for our purposes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Freq.</th>
<th>Z</th>
<th>Rank</th>
<th>Term</th>
<th>Freq.</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>69,9</td>
<td>0.07</td>
<td>6</td>
<td>in</td>
<td>21,3</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>36,4</td>
<td>0.07</td>
<td>7</td>
<td>that</td>
<td>10,5</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>28,8</td>
<td>0.08</td>
<td>8</td>
<td>is</td>
<td>10,0</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>26,1</td>
<td>0.10</td>
<td>9</td>
<td>was</td>
<td>9,81</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>23,2</td>
<td>0.11</td>
<td>10</td>
<td>he</td>
<td>9,54</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Heaps’ Law

- Size of vocabulary: $V(n) = Kn^\beta$
- In English, $K$ is between 10 and 100, $\beta$ is between 0.4 and 0.6.

http://en.wikipedia.org/wiki/Heaps%27_law
Heaps’ Law (cont’d)

- Related to Zipf’s law: generative models
- Zipf’s and Heaps’ law coefficients change with language
