LING/C SC 581:
Advanced Computational Linguistics
Lecture Notes
Feb 8th
Adminstrivia

- The Homework Pipeline:
  - Homework 4 + Homework 5
- No classes next week:
  - I'm out of town on business
  - No new homework assigned this week
  - I sorta lied: optional homework 6
    (if you choose to do, due Monday 19th midnight: > 1 week)
Today's Topics

1. Homework 5 review
2. Zipf's Distribution
   • Read the review paper:
     *Zipf’s word frequency law in natural language: A critical review and future directions, Piantadosi, S.*
   • Optional Homework 6
     • Using Python and our ptb WSJ corpus, compute and graph frequency against rank
     • Use pyplot
     • In your opinion, does our WSJ corpus obey Zipf's Law?
Zipf's Law

Let $r$ be the rank of some unit, e.g. a word, in a corpus
Let $freq(r)$ be the frequency associated with the $r$-th ranked word

Zipf:
- $freq(r) \propto \Phi 1/r^\alpha$  \hspace{1cm} $\Phi=1/10; \alpha \approx 1$

Zipf-Mandelbrot:
- $freq(r) \propto 1/(r+\beta)^\alpha$ \hspace{1cm} $\alpha \approx 1; \beta \approx 2.7$

There "should be" a power law relation between $r$ and frequency
- read cited paper for discussion on "should be"
Zipf's Law

Typically plotted as log-log

http://hyperphysics.phy-astr.gsu.edu/hbase/Math/logplot.html

A straight line with slope = 2 implies

\[ y \alpha t^2 \]
Zipf's Law: ANC

- ANC = American National Corpus
Zipf's Law: PTB Brown corpus

Frequency of PTB POS tags:
1. NN
2. IN
3. DT
4. JJ
5. NNP
6. NNS
7. RB
8. PRP
Zipf's Law: PTB Brown corpus

- Most common tokens for each POS tag

Fig. 7

Frequency distribution of words within several syntactic categories from the Penn Treebank: determiners (DTs), prepositions or subordinating conjunctions (INs), modals (MDs), nouns (NNs), past participle verbs (VBNs), third-person singular present verbs (VBZs). These plots represent a post-hoc selected subset of all syntactic categories.
Free Relatives

• Taken from Caponigro, I. (proceedings of WECOL 2002).
• Some background:

(1)  a. I appreciate \([_{FR} \text{what you did for me}]\).
    a’. I appreciate \([_{DP} \text{your help}]\).

    b. \([_{FR} \text{Who couldn’t sleep enough}]\) felt tired the following morning.
    b’. \([_{DP} \text{The insomniacs}]\) felt tired the following morning.

    c. You can’t smoke \([_{FR} \text{where the kids are sleeping}]\).
    c’. You can’t smoke \([_{DP} \text{there}]\).

    d. He opened the door \([_{FR} \text{when I was about to knock}]\).
    d’. He opened the door \([_{DP} \text{then}]\).
Free Relatives

• Free vs. Headed relatives:

(2) a. FR: I like \([\text{FR } \underline{\text{what}} \underline{\text{you bought}}]\).
   b. = HR: I like \([\text{HR } \underline{\text{the thing(s) which you bought}}]\).

   HEAD

(4) a. I like \([\text{FR } \underline{\text{what}} \underline{\text{you bought}}]\).
   a'. = I like \([\text{the thing(s) you bought}}]\).
   b. I wonder \([\text{wh-Q } \underline{\text{what you bought}}]\).
   b'. = I wonder \([\text{which thing(s) you bought}}]\).
Free Relatives

9.2.3 Free ("headless") relatives

1. General. A free or headless relative is defined as any relative clause that lacks a head. Free ("headless") relatives are labeled SBAR-NOM.

```
(S (SBAR-NOM-SBJ (WHNP-1 What)
 (S (NP-SBJ *T*+1)
   (VP is
     (PP-PRD of
       (NP (ADJP much more)
         importance)
     (PP to
       (NP the Colombian economy))
     (PP than
       (NP the supposed
         benefits)
     (PP of
       (NP laundered drug
         money))))))

(VP is
  (NP-PRD (NP higher prices)
    (PP for
      (NP (NP Columbia ’s)
        legitimate products)))

.)
```
Homework 5 Review

- /\^SBAR-NOM/ 775
Homework 5 Review

- /^SBAR-NOM/ 775
- FR what(ever) 542
Homework 5 Review

- /^SBAR-NOM/ 775
- FR what* 542
- FR where(verb) 21
Homework 5 Review

- \(^{\text{SBAR-NOM}}/\) 775
- FR *what* 542
- FR *where* 21
- FR *when(ever)* 14
Homework 5 Review

- /^SBAR-NOM/ 775
- FR what* 542
- FR where* 21
- FR when* 14
- FR who(ever) 26
  whose/whom
Homework 5 Review

- \(^{\text{SBAR-NOM}}/ \) 775
- \(\text{FR what}^*\) 542
- \(\text{FR where}^*\) 21
- \(\text{FR when}^*\) 14
- \(\text{FR who}^*\) 26
- \(\text{FR how}\) 104
Homework 5 Review

- /^SBAR-NOM/ 775
- FR what* 542
- FR where* 21
- FR when* 14
- FR who* 26
- FR how 104
- FR whether 15
Homework 5 Review

- /^SBAR-NOM/ <=, /*Which*/
- FR what* 542
- FR where* 21
- FR when* 14
- FR who* 26
- FR how 104
- FR whether 15
- FR which(ever) 13

"three of which"
Homework 5 Review

- /^SBAR-NOM/ 775
- FR what* 542
- FR where* 21
- FR when* 14
- FR who* 26
- FR how 104
- FR whether 15
- FR which* 13
- FR why 8
Homework 5 Review
Homework 5 Review

- /\^SBAR-NOM/ 775
- FR \textit{what}* 542 ADVP 4
- FR \textit{where}* 21 0
- FR \textit{when}* 14 0
- FR \textit{who}* 26 0
- FR \textit{how} 104 0
- FR \textit{whether} 15 0
- FR \textit{which}* 13 0
- FR \textit{why} 8 0
Homework 5 Review

- /\SBAR-NOM/ 775
- FR *what* 553  ADVP 4  QP 3
- FR *where* 21  0
- FR *when* 14  0
- FR *who* 26  0
- FR *how* 104  0
- FR *whether* 15  0
- FR *which* 13  0
- FR *why* 8  0
Homework 5 Review

[Diagram of tree structures with highlighted 'Inconsistent annotation']
Homework 5 Review

- /^SBAR-NOM/ 775
- /^SBAR-NOM/ <<, (/^WHNP/ << /[Ww]hat/) 554 ADVP/QP 7
- /^SBAR-NOM/ <<, (/^WH/ << /[Ww]here/) 21
- /^SBAR-NOM/ <<, (/^WH/ << /[Ww]hen/) 14
- /^SBAR-NOM/ <<, (/^WH/ << /[Ww]ho/) 26
- /^SBAR-NOM/ <<, (/^WH/ << /[Hh]ow/) 118
- /^SBAR-NOM/ <<, (/^IN/ << /[Wh]ether/) 15
- /^SBAR-NOM/ <<, (/^WH/ << /[Ww]hich/) 19
- /^SBAR-NOM/ <<, (/^WH/ << /[Ww]hy/) 8

782 total due to overlaps, e.g. coordination
Homework 5 Review

Mistake?
Homework 5 Review