CS 585 Natural Language Processing
Fall 2004
Homework 1

Due: Thu, September 23, 2004

1 Parts of speech

What are the parts of speech of the words in the following paragraph?

The lemon is an essential cooking ingredient. Its sharply fragrant juice and tangy rind is added to sweet and savory dishes in every cuisine.

2 Noun-noun compounds

Come up with four examples of noun-noun compounds.

3 Sentence ambiguity

The italicized phrases in the following sentences are examples of attachment ambiguity. What are the two possible interpretations?

- Mary saw the man on the hill *with a telescope*.
- The company experienced growth in classified advertising *and preprinted inserts*.

4 Get you hands dirty with real data

To review or familiarize yourself with some Unix tools for text manipulation, read the first 19 pages of "Unix for Poets":

http://www.research.att.com/~kwc/tutorials/unix_for_poets.ps

Download the book David Copperfield from Project Gutenberg:

http://www.gutenberg.net/dirs/etext96/cprfd10.txt

Note that the file starts with the copyright and information about Project Gutenberg. Don't remove them; just use the file in its entirety.

To make sure everyone gets the same answer for this question, do the following conversions to the file:

- Convert all characters to lowercase. Assuming you saved the file as cprfd10.txt, invoke the command

```
$ tr '[A-Z]' '[a-z]' cprfd10.txt > cprfd10.lower
```

- Put each space-separated token on a line by itself (note the immediate newline after /\):

```
$ sed 's/ /\
/g' cprfd10.lower > cprfd10.singles
```

- Remove non-alphanumeric characters on the left and right ends of each token. Pipe this result to a temporary file named cprfd10.noPunct

```
$ sed 's/^[^a-z0-9]*//g' cprfd10.singles | sed 's/[^a-z0-9]*$//g'
```

- Remove empty lines.

```
$ sed '/^$/d' cprfd10.noPunct > cprfd10.onePerLine
```

Using Unix shell commands such as **sort**, **uniq**, **sed**, and **wc**, answer the following questions:

- Word counts. For each word, count the number of times it occurs in this dataset (you don't need to turn in this list). Hand in a list the top 15 terms that occur in this dataset in descending order, along with their counts. Your list should look like:

count_1 word_1
count_2 word_2
...
count_15 word_15

- Exploring Zipf's Law. How many distinct tokens are there in the book after the preprocessing steps above? Remove those tokens that occur only once in the book. How many distinct tokens are left now? Answer the same question after you remove tokens that occur twice or fewer, then those that occur thrice or fewer. How do you relate these numbers to Zipf's Law?

5 Chomsky normal form

Convert the following context-free grammar to Chomsky normal form. Show your work.

$$\begin{aligned} S &\rightarrow ASAB \\ A &\rightarrow CB \mid B \\ B &\rightarrow xA \mid y \\ C &\rightarrow Sz \end{aligned}$$

6 Bottom-up parsing

Consider the following CFG:

$$\begin{aligned} S &\rightarrow NP\ VP \mid Aux\ NP\ VP \mid VP \\ NP &\rightarrow Det\ NOM \\ NOM &\rightarrow Noun \mid Noun\ NOM \\ VP &\rightarrow Verb \mid Verb\ NP \\ Det &\rightarrow this \mid that \mid a \mid the \\ Noun &\rightarrow bank \mid flight \mid man \mid robbery \mid meal \\ Verb &\rightarrow hears \mid talking \mid bank \\ Aux &\rightarrow does \end{aligned}$$

Using the above grammer, show the bottom-up (shift reduce) parse of this sentence:

The man hears the bank robbery.

Show the stack, input remaining, and action at each step of the parse, as done in the lecture notes.