

# LING 388 HOMEWORK 1

DUE SEPTEMBER 25, 2009

## 1. REGULAR GRAMMARS

1.1. **a - 2 pts.** List the Regular Right-linear rules for the language  $a^*b^+c$ . This language contains strings such as

- bc
- abc
- bbbbbc
- aaaaaabc
- aabbbbbbbc

Remember that Right-linear rules can only have two forms: a Non-terminal on the left, and either a single terminal, or a single terminal followed by a single non-terminal on the left.

- $S \rightarrow a$
- $S \rightarrow aB$

1.2. **b - 4 pts.** Write a Python parser that accepts strings in  $a^*b^+c$ , and rejects all others.

## 2. UNIX AND WORD FREQUENCY

2.1. **4 pts.** Find a text of your choice on <http://www.gutenberg.org> and download the plain text version. Write a Unix command that finds the number of occurrences for each word in the text and sorts the results into a text file. Your text file should look something like the following:

```
25486 the
14272 of
10879 and
10429 in
7670 to
7373 a
6571 coffee
```

Use Excel or another program of your choice to plot a rank vs. frequency plot of the results. Rank is determined by how many occurrences a word has. For example the word 'the' above has rank 1 because it occurs most frequently in the text. The word 'of' has rank 2 and so on. If you plot your data using a logarithmic scale on both the rank and frequency axes, your plot should look similar to the following picture.

