Lesson 1 - What is Language?

Erin M. Buchanan

12/31/2018

What is this course about?

- Computational Linguistics
- Dealing with language (which is messy)
- Learning special analyses for language in *R*
- Make reports of your work in *Rmarkdown*

What will you learn?

- What is computational linguistics and language processing?
- How can we apply statistics to answer questions about qualitative data (i.e. language at any level)?
- What are the popular ways to measure language association and model human language?

Syllabus

- You should read the syllabus for course policies and other important information.
- You will use Moodle for all course related activities.
- Let's check those things out now.

Writing

- You will be expected to write reports with code and text embedded.
- You will want to embed or otherwise cite your sources for material you are referencing.
- Please use APA style on how to citations (search Purdue OWL for tips).

Human Language

Things to think about:

- What was the last thing you said to someone?
- . . . the last thing you wrote down?
- . . . the last thing you heard?
- How exactly did you do those things?

Parts to Human Language

- Biological: brain areas, mouth, tongue, larynx
- Cognitive: symbol systems, word order
- Social: knowledge of other users, social rules, attitudes

Language Purpose

- Communication
- Emotional expression
- Social interaction
- Thinking

Studying Language

- Linguistics: study of language
- Psycholinguistics: psychological processes involved in language and the individual (sometimes called cognitive linguistics)
- Computational linguistics: analysis of language through the lens of computer science
- . . . even more names, as we expand and cross over with other fields

What is Language?

- System of symbols and rules that enable us to communicate
- Some terms to know:
  - Semantics: study of meaning
  - Syntax/grammar: system of rules for language to be well formed
  - Morphology: study of words
  - Pragmatics: study of language use
  - Lexicon: mental dictionary, word storage

History of Studying Language

- Before/around 1900: Galton and Freud
- 1950s: Famous conference at Cornell, Dartmouth
- Chomsky and Skinner
- Influenced heavily by research on artificial intelligence, computing power increases, thinking about modeling language with computers

Basic Language Terminology

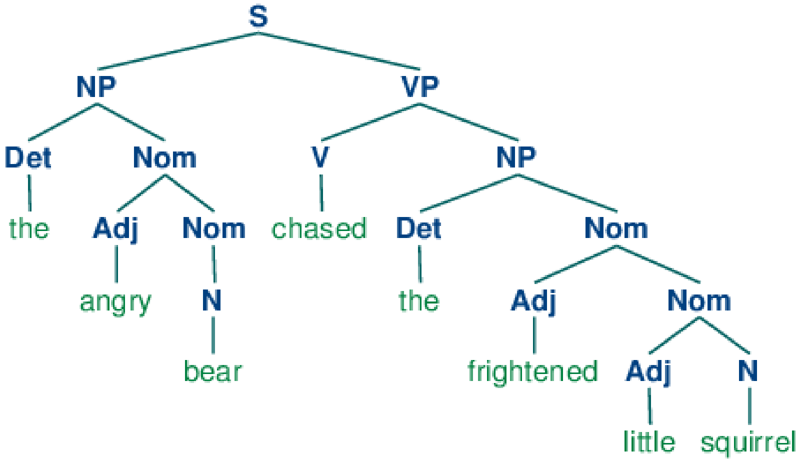
- Phoneme: basic unit of sound
  - The Many to Many Problem with English
  - Vowels and Consonants
- Syllables: rhythmic unit of speech
- Morphemes: smallest unit of meaning in a word
- Words: smallest element in isolation with meaning
  - Token: total number of words in a text
  - Types: number of distinct words

Basic Language Terminology

- Categories of words:
  - Nouns
  - Adjectives
  - Verbs
  - Adverbs
  - Determinants
  - Pronouns
  - Prepositions
  - Conjunctions

Basic Language Terminology

- Phrases: group of words forming a grammatical unit (noun versus verb)
- Allows you to make tree diagrams of sentences



Defining Human Language

Hockett's Feature Design: communalities between languages that define language as separate from other communication systems (i.e., animals)

- Semanticity: symbols are tied to meaning
- Arbitrariness: symbols are arbitrary (not tied to meaning)
- Discreteness: symbols can be broken down and recombined (morphemes)
- Productivity: users can create and understand novel text (creativity)

Applying Statistics to Language

- Originally, studying language was part of a qualitative skill set
- Statistics were simple percentages/means
- Language was considered innate -> so all humans had the same underlying system
- We just had to figure out what that system was . . .

Applying Statistics to Language

- However: statistical language learning and the interaction with the environment could not be ignored
- Language knowledge is shaped by language use
- As we learn and use a language, we are "intuitive statisticians" - this implies that language can be analyzed with statistics

Influence of Our Surroundings

- Frequency, frequency, frequency
- Cognitive mechanisms
  - Probabilistic structure of categories
- Social mechanisms
  - Representations of word meanings
  - New words in your lifetime
  - Slang

Language and Statistics Now

Examples:

- Model word choice
- Corpora!
- Behavioral profiles
- Semantic Vector models
- Along with experimental results relying on traditional statistics: t-tests, ANOVA, correlation, regression, etc.

Install the Packages

- You will need to know *R* for this course
- Be sure to have the newest version of *R* (3.5.2) and *RStudio* (1.1.463 or the dev version 1.2+)
- The package for the book is *Rling* - it's included online for you to download

```
install.packages(file.choose(), repos = NULL, type = "sourc

install.packages("modeest")

if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("genefilter", version = "3.8")
```

Load the Libraries

```r
library(Rling)
library(modeest)
data(ldt)
head(ldt)
```

```
##                Length Freq Mean_RT
## marveled            8  131  819.19
## persuaders         10   82  977.63
## midmost             7    0  908.22
## crutch              6  592  766.30
## resuspension       12    2 1125.42
## efflorescent       12    9  948.33
```

Basic Statistics (Continuous)

- Using the English Lexicon Project, what can we learn about word length and response latencies?
  - What is the ELP?
  - What is length?
  - What is response latency?

Basic Statistics (Continuous)

- Variables we can calculate:
  - `summary`: min, 1st quantile, median, mean, 3rd quantile, max
  - `mlv`: mode from *modeest* package
  - `sd`: standard deviation

Basic Statistics (Continuous)

```r
summary(ldt$Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00    8.00    8.23   10.00   15.00
```

```r
mlv(ldt$Length)
```

```
## [1]  8 10
```

```r
sd(ldt$Length)
```

```
## [1] 2.501939
```

Graphical Displays (Continuous)

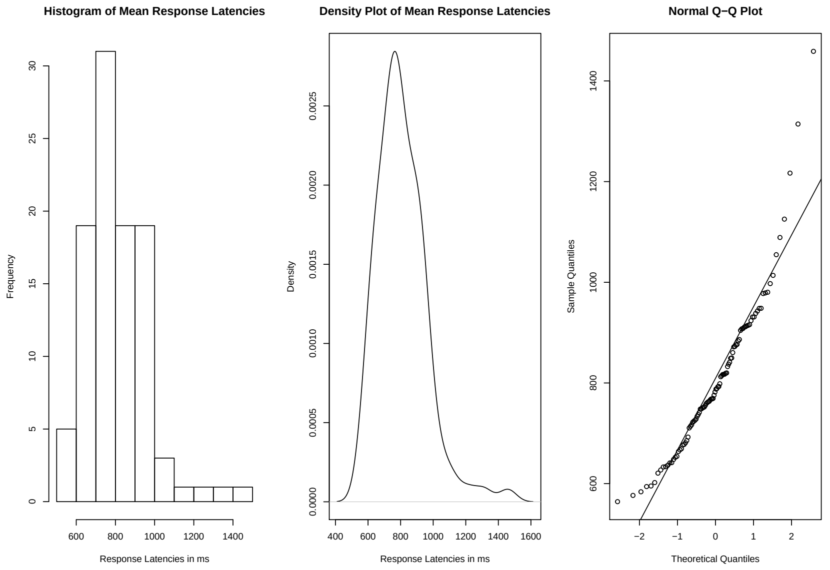
```
par(mfrow = c(1, 3))

hist(ldt$Mean_RT, main = "Histogram of Mean Response Latenc
     xlab = "Response Latencies in ms")

plot(density(ldt$Mean_RT), main = "Density Plot of Mean Res
     xlab = "Response Latencies in ms")

{qqnorm(ldt$Mean_RT)
qqline(ldt$Mean_RT)}
```

Graphical Displays (Continuous)



Graphical Displays (Continuous)

- Data appears to indicate a bit of skew and some outliers:
  - Use `boxplot` to view those outliers
  - We can view those outliers by using z-scores

```
summary(ldt$Mean_RT)
```

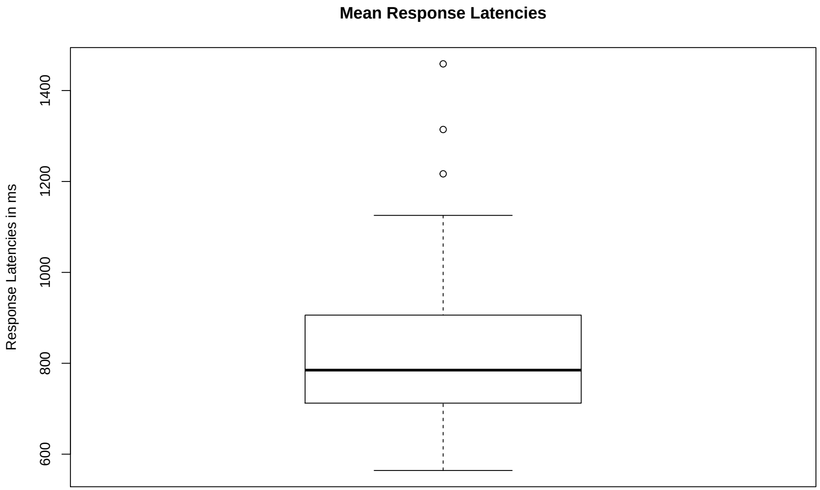
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   564.2   713.1   784.9   808.3   905.2  1458.8
```

```
ldt[abs(scale(ldt$Mean_RT)) > 3, ]
```

```
##               Length Freq Mean_RT
## dessertspoon      12   11 1314.33
## diacritical       11  162 1458.75
```

Graphical Displays (Continuous)

```
boxplot(ldt$Mean_RT, main = "Mean Response Latencies",
        ylab = "Response Latencies in ms")
```

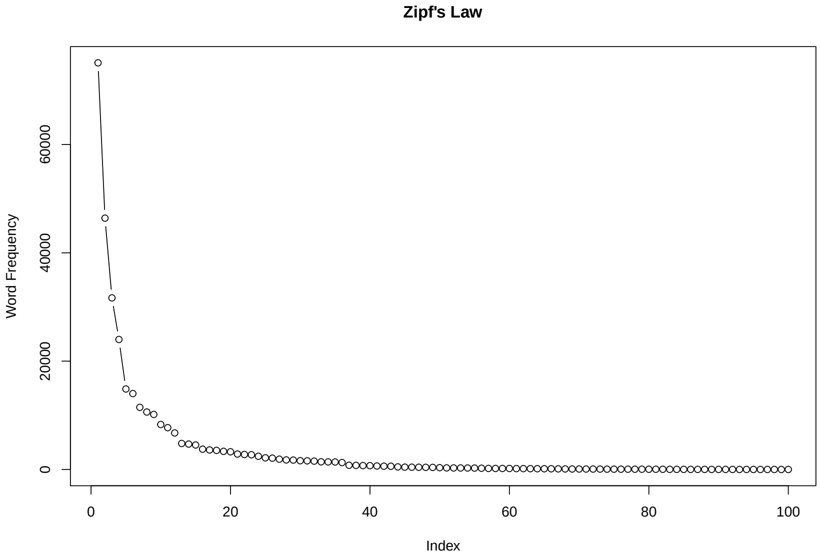


Zipf's Law (Continuous)

- Zipf's Law: word frequency is inversely related to its frequency rank.
  - Therefore the first word is twice as likely as the second word, three times as likely as the third word, etc.

Zipf's Law (Continuous)

```
plot(sort(ldt$Freq, decreasing = TRUE),
     type = "b", main = "Zipf's Law", ylab = "Word Frequenc
```



Basic Statistics (Categorical)

- What do we need to do differently to visualize/understand categorical data?

```r
data(sent)
head(sent)
```

```
##    clause  subj
## 1   Trans   Hum
## 2   Trans Abstr
## 3    Ditr Abstr
## 4   Trans   Hum
## 5 Intrans Abstr
## 6 Intrans   Hum
```

Basic Statistics (Categorical)

- Dataset contains 20 sentences marked with the types of verbs in each clause:
  - Intransitive: subject + no objects: He sneezed.
  - Transitive: subject + one object: The cat bit him.
  - Ditransitive: subject + two objects: He gave Mary ten dollars.

```
summary(sent$clause)
```

```
##    Ditr Intrans   Trans
##       2      10       8
```

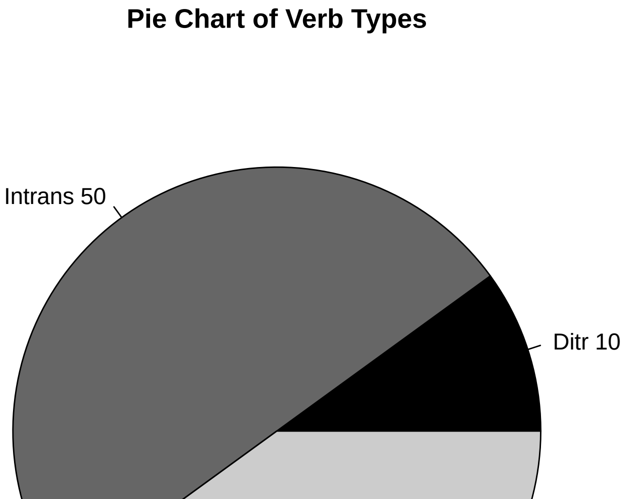
Basic Statistics (Categorical)

```
sent.t = table(sent$clause)
prop.table(sent.t)
```

```
##
##    Ditr Intrans   Trans
##     0.1     0.5     0.4
```

Graphical Displays (Categorical)

```
pie(sent.t,
    main = "Pie Chart of Verb Types",
    col = c("black", "grey40", "grey80"),
    labels = paste(names(sent.t), prop.table(sent.t)*100))
```



Basic Color Terms

- Berlin and Kay (1969) proposed a theory about our linguistic interpretations of colors, mainly that color vocabulary falls into universal categories:

$$\begin{bmatrix}\text{white}\\ \text{black}\end{bmatrix} < \begin{bmatrix}\text{red}\end{bmatrix} < \begin{bmatrix}\text{green}\\ \text{yellow}\end{bmatrix} < \begin{bmatrix}\text{blue}\end{bmatrix} < \begin{bmatrix}\text{brown}\end{bmatrix} < \begin{bmatrix}\text{purple}\\ \text{pink}\\ \text{orange}\\ \text{grey}\end{bmatrix}$$

Basic Color Terms

- Data is from the Corpus of Contemporary American English (COCA)
- Counts of adjective use of color terms
- Problems with simple frequency count is corpus size

```
data(colreg)
head(colreg)
```

```
##        spoken fiction academic press
## black   20335   41118    26892 73080
## blue     4693   22093     3605 21210
## brown    1185   10914     1201 11539
## gray     1168   12140     1289  6559
## green    3860   14398     4477 26837
## orange    931    3496      474  5766
```

Basic Color Terms

- We know the corpus size (number of words) from looking at COCA statistics
- We can calculate the deviation of those proportions (like standard deviation)

```
freqreg <- c(95385672, 90344134, 91044778, 187245672)

dev_prop = function (observed_count, expected_count){
  DP_value = sum(abs(prop.table(observed_count) - prop.tab
  DP_normal = DP_value / (1-min(prop.table(expected_count))
  return(DP_normal)
}
```

Basic Color Terms

- Values close to zero indicate spread is similar given corpus size
- Values close to one indicates one of the subsets is favored more strongly

```
dev_prop(colreg["black", ], freqreg)
```

```
## [1] 0.1356127
```

```
dev_prop(colreg["gray", ], freqreg)
```

```
## [1] 0.4707974
```

Summary

- In this lecture, you learned:
  - What is human language?
  - History of analyzing human language
  - Simple statistics for continuous and categorical variables
  - Simple graphs for continuous and categorical variables
- Next: Linear Regression + Frequency/Response Latencies