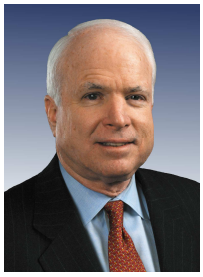
SCHWEIZER JUGEND FORSCHT
LA SCIENCE APPELLE LES JEUNES
SCIENZA E GIOVENTÙ
SCIENZA E GIUVENTETGNA

Analysis of US Political Speeches

The US Presidential Election'08

Main Findings



In Obama's Speeches

- Tends to overuse "McCain" or "Bush (administration)", or "Kennedy"
- Terms overused by Obama are "together" (plans for the future), "black", "energy", "change", "jobs", "union", "(yes we) can", or "investing"
- Other overused terms "Berlin", "Germany"
- Words like "election", "November", "dream" occur more often than usually
- *Latest overused terms*: "women", "solar", "fiscal", "economic", "Iranian" "childcare", and "patriotism"

In Political Speeches

- The main topics are related with the terms "war", "(our) country", "Iraq", "economy", and "healthcare"
- More traditional topics with "taxes", "fiscal", "people", "security"
- Compared to the Brown Corpus, politicians tend to underuse "she", "woman", "women", and in a lesser extent "he"
- Overused terms "I", "we", "will", "our", "my", "us", "must", as well as "care", "work", "family"
- *New topics* introduced in the last three months: "energy", "oil", "Israel" and reinforced the term "economy" or "jobs"

In McCain's Speeches

- Tends to overuse the word "(Barack) Obama"
- "Reform" is also overused together with "criminal", "federal", "judicial, "court", "nuclear (energy)"
- Other overused items are "friendship", and "Canada"
- The term "veteran" seems to be used by both candidates
- *Latest overused terms:* "business(es)" "Hispanic", "fuel", "foreign", "OPEC", "electric" or "NAFTA"

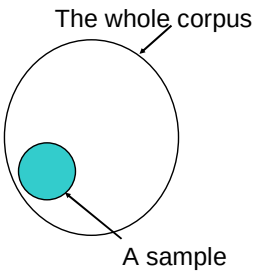
- Some statistics of our US political corpus from speeches extracted from the web site of the candidates
  - 466,292 word tokens (number of words)
  - 11,866 word types (distinct words)
  - 3,931 hapax (words occurring only once)
- About the frequency
  - the most frequent word "the" occurs 22,615 times (or 4.85%)
  - 30.04% of the corpus is composed by the first 15 most frequent word types

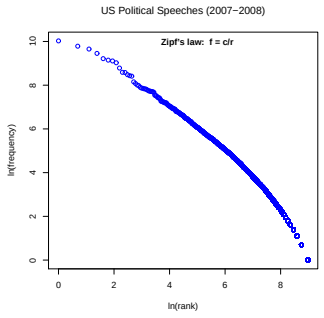
- A **`perl`** script example (removing the plural '-s')

```
#!/usr/bin/perl -w
while ($line = <>) {     # read line by line
   chomp($line);         # remove the newline char
   if (length($line) > 4) {    # only for word > 4
      if ($line=~ m/[^ea]ies$/) {
          $line =~ s/ies$/y/i;    # replace -ies with -y (plural form)
          } else {
             if ($line=~ m/[^eaoi]es$/) {
                $line =~ s/es$/e/i;    # replace -es with -e (plural form)
                }  }
      if ($line=~ m/[^eus]s$/) {
        $line =~ s/s$//i; }     # delete -s (plural form)
   } # end if (length()
} # end while
exit(0);
```

- Our statistical model
  - We have a (large) corpus (used as reference)
  - We have a small sample (*p* is the relative size of it)
  - Selecting (randomly) a term *t*, we can count its frequency in the whole corpus ($freq_{corpus}$) and its expected frequency in the sample
  - We can compare the frequency observed with the expected frequency and compute its Z-score

$$Z\ score = \frac{freq_{observed} - (p \cdot freq_{corpus})}{\sqrt{freq_{corpus} \cdot p \cdot (1-p)}}$$





Contact: Adrian Kronauer, Freies Gymnasium Zürich
Martina Rakaric, Kanti Wettingen
Prof. Jacques Savoy, Université de Neuchâtel

Université de Neuchâtel
Institut d'informatique