Ling 105 Problem Set 6

Lev Kruglyak

Due: October 25, 2024

Problem 1. Do the speakers in the file appear to have the cot-caught merger? A simple way to tell is to compare the pronunciation (**Surface**) of words spelled with **-aught** to those spelled with **-ot**; is the vowel generally the same or different? (For this and the next question, don't worry about coding the whole solution; you can just pull up the relevant words and inspect them by hand.)

The speakers in the file seem to all use the vowel "Q" for words ending with "aught" and "ot". Of course, this excludes words such as "shoot" or "pilot".

Problem 2. How many times is the word **probably** pronounced with one vs. two vs. three syllables?

The word "probably" is pronounced with one syllable 1 time, with two syllables 13 times, and with three syllables 4 times.

Problem 3. Which speaker has the highest mean duration of word tokens?

The speaker with the highest mean token duration is Speaker 3, who has a mean duration of 311 ms.

Problem 4. What is the median duration of the word **the** in ms?

The median duration is 94 ms.

Problem 5. What is the minimum and maximum duration of a word token ending with orthographic **-th**? Filter using `str_detect()`.

The minimum duration is 41 ms, and the maximum duration is 756 ms.

Problem 6. Which word class has more syllables (**SylNSurface**) on average in this corpus, nouns or verbs? For nouns, **POS** begins with "N," and for verbs, **POS** begins with "V."

On average, nouns have 1.795 syllables per word, and verbs have 1.267 syllables per word. Nouns thus have more syllables per word on average.

Problem 7. What are the five most frequent words in the corpus?

The five most common words are “i”, “and”, “the”, “to”, and “that”.

Problem 8. What are the five most frequent past participles (**POS** is VBN) in the corpus?

The five most common past participles are “been”, “done”, “based”, “come”, and “born”, although the fifth place is tied with “exposed”, “gone”, “gotten”, “married”, and “taught”.

Problem 9. Which full part of speech (**POS**, full tag) is the most frequent, and what percentage of word tokens in the corpus does it account for?

The most common part of speech is a noun, accounting for 2092 tokens, or about 11.98%.

Problem 10. The number of syllables per word (as pronounced) is given by **SylNSurface**. It has been claimed that on average, syllables are shorter in words with more syllables, a principle known as polysyllabic shortening. Show that polysyllabic shortening holds by giving the mean duration per syllable in words of one vs. two vs. three vs. four vs. five vs. six syllables. (Report one mean per word size; do not report a mean for each position separately.) Hint: feel free to get the mean whole-word durations and then do the division by hand.

Number of syllables	Mean duration of syllable (ms)
1	206.45
2	196.02
3	183.68
4	168.78
5	163.18
6	154.44

Here is the code used for all problems:

```
library(tidyverse)
x = read.csv("words12.txt", sep="\t")

# Problem 1
print(x |> filter(str_detect(Word, "aught\$")) |> select(Speaker, Word, Surface))
print(x |> filter(str_detect(Word, "ot\$")) |> select(Speaker, Word, Surface))

# Problem 2
print(x |> filter(Word == "probably") |> select(SylNSurface) |> pull() |> table())

# Problem 3
print(x |> group_by(Speaker) |> summarize(avg_duration = mean(Duration, na.rm = TRUE)))

# Problem 4
print(x |> filter(Word == "the") |> summarize(avg_duration = median(Duration, na.rm = TRUE)))

# Problem 5
print(x |> filter(str_detect(Word, "th\$"))
  |> summarize(min_duration = min(Duration), max_duration = max(Duration)))
```

```
# Problem 6
print(x |> filter(str_detect(POS, "^N")) |> summarize(noun_mean_syllables = mean(SylNSurface)))
print(x |> filter(str_detect(POS, "^V")) |> summarize(verb_mean_syllables = mean(SylNSurface)))

# Problem 7
print(x |> group_by(Word) |> summarize(n = n()) |> arrange(desc(n)))

# Problem 8
print(x |> filter(POS == "VBN") |> group_by(Word) |> summarize(n = n()) |> arrange(desc(n)))

# Problem 9
print(x |> group_by(POS) |> summarize(n = n(), frequency = n / nrow(x) * 100) |> arrange(desc(n)))

# Problem 10
print(x |> group_by(SylNSurface)
  |> summarize(avg_syl_length = mean(Duration) / unique(SylNSurface)))
```