Frequency of words given their function in a sentence

Mazunki Hoksaas

2023-09-30

Introduction

This analysis uses the "perry_winter_2017_iconicity.csv" dataset to explore the relationship between word frequency and part of speech using R and ggplot2.

Our intention here is to prove, given the provided dataset, that verbs in sentences are disproportionally iconic compared to other parts of speech.

Load Libraries and Data

```r
library(tidyverse)  # let's us use data pipes :)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
head(read.csv("perry_winter_2017_iconicity.csv"))
```

By checking the first rows of the data we can see we have columns for the words, their part of speech, their frequency and their iconicity. This is enough for what we need.

Additionally, we can see some other columns for the sensory experience (SER), their imageability (how easy it is to form an image of the word in our head), their systematicity (how strong their sound-meaning link is). I'm guessing the Conc column stands for Concreteness (as opposed to abstractness).

```r
# we can use the word column as the index since they are unique anyway

data <- read.csv("perry_winter_2017_iconicity.csv", row.names=1) %>%
        select(POS, Freq, Iconicity)

verbs <- filter(data, POS == "Verb")
other <- filter(data, POS != "Verb")

data <- bind_rows(
    verbs %>% mutate(Group = "Verb"),
    other %>% mutate(Group = "Non-verb")
  ) %>% arrange(Group)  # there's not many verbs compared to other POS
```

We start by taking an overview at what the most frequent words are, followed by which words are the most iconic

```r
print(head(verbs[order(-verbs$Freq, -verbs$Iconicity), ], n=10))
```

```
##       POS   Freq  Iconicity
## is   Verb 459663 -0.1428571
## have Verb 314232 -0.2666667
## do   Verb 312915  0.8461538
## be   Verb 293085  0.3846154
## know Verb 291780  0.7692308
## was  Verb 288391 -0.8333333
## are  Verb 265672 -0.9000000
## get  Verb 233772 -0.5833333
## go   Verb 193445  1.4545455
## come Verb 160190  0.2142857
```

```r
print(head(other[order(-other$Freq, -other$Iconicity), ], n=20))
```

```
##               POS    Freq  Iconicity
## you   Grammatical 2134713 -0.4000000
## I     Grammatical 2038529  3.1818182
## the   Grammatical 1501908  0.4285714
## to    Grammatical 1156570 -0.4166667
## a     Grammatical 1041179  0.4615385
## it    Grammatical  963712  1.0000000
## that  Grammatical  719677 -0.0625000
## and   Grammatical  682780  0.5625000
## of    Grammatical  590439  0.2307692
## what  Grammatical  501965  0.1428571
## in    Grammatical  498444  1.4615385
## me    Grammatical  471339  0.6000000
## we    Grammatical  459607  1.4285714
## this  Grammatical  406915  0.1333333
## he    Grammatical  389497  1.0588235
## on    Grammatical  354742  0.9166667
## for   Grammatical  351650 -1.4000000
## my    Grammatical  344899  1.5000000
## your  Grammatical  328715  0.0000000
## no   Interjection  304549  2.8125000
```

```r
print(head(data[order(-data$Iconicity, -data$Freq), ], n=10))
```

```
##               POS Freq Iconicity    Group
## humming      Verb  251  4.466667     Verb
## click        Verb  327  4.461538     Verb
## hissing      Verb   73  4.461538     Verb
## gurgle       Verb   12  4.416667     Verb
## mushy   Adjective   77  4.384615 Non-verb
## beep         Noun  332  4.357143 Non-verb
## screech      Noun  318  4.333333 Non-verb
## buzzing      Verb  221  4.333333     Verb
## zigzag       Noun   23  4.300000 Non-verb
## squeak       Verb  121  4.230769     Verb
```

An interesting observation is how predominant grammatical words are in the data set, but otherwise we don't see too much interesting stuff in these tables just yet.

```r
print(summary(verbs))
```

```
##      POS                  Freq           Iconicity
##  Length:557         Min.   :     1   Min.   :-2.1000
##  Class :character   1st Qu.:    62   1st Qu.: 0.4286
##  Mode  :character   Median :   373   Median : 1.2308
##                     Mean   : 10534   Mean   : 1.3804
##                     3rd Qu.:  2758   3rd Qu.: 2.2857
##                     Max.   :459663   Max.   : 4.4667
```

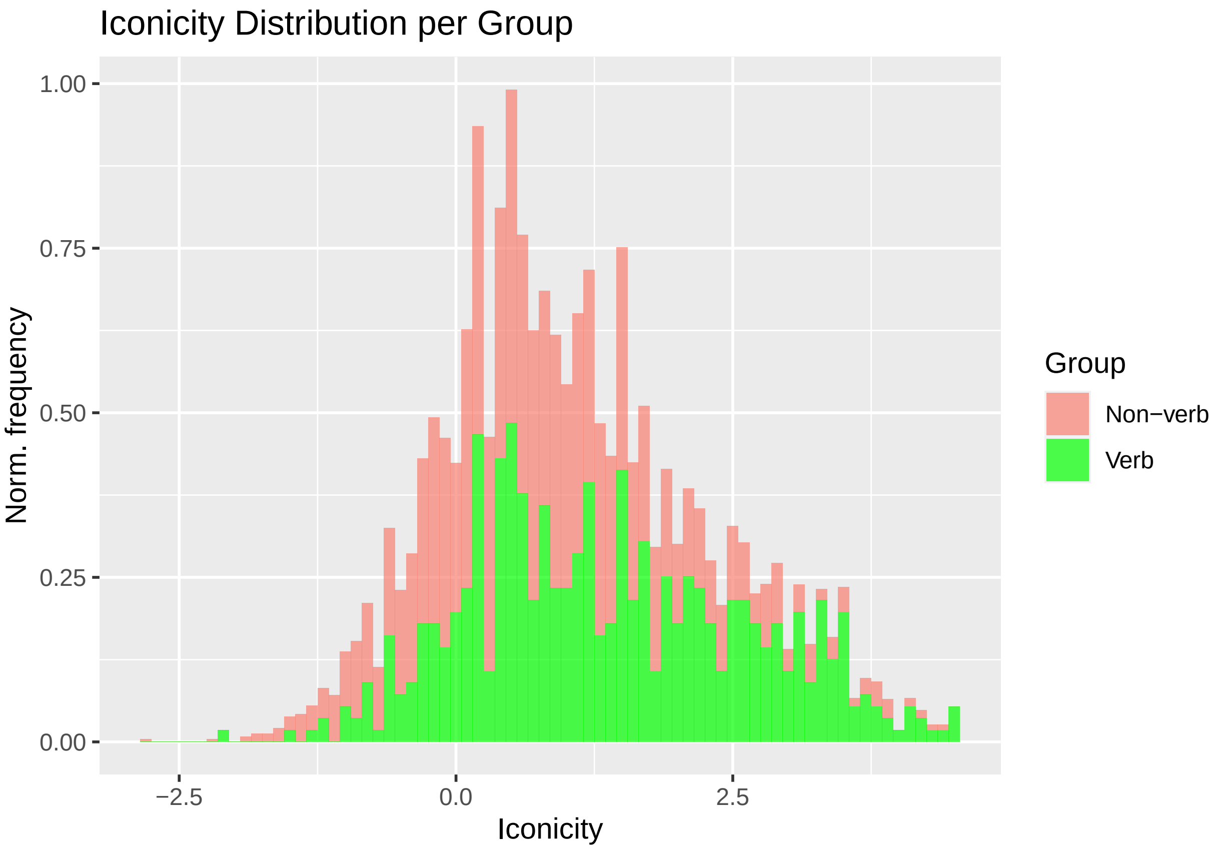
```r
print(summary(other))
```

```
##      POS                  Freq              Iconicity
##  Length:2390        Min.   :      1.0   Min.   :-2.80000
##  Class :character   1st Qu.:     75.2   1st Qu.: 0.09091
##  Mode  :character   Median :    354.5   Median : 0.70000
##                     Mean   :  11216.6   Mean   : 0.80580
##                     3rd Qu.:   1511.8   3rd Qu.: 1.42559
##                     Max.   :2134713.0   Max.   : 4.38462
```

The summary tells us that the frequency of words is very disparate, where some words are used extremely often, while other words are barely used at all. This is the case for both groups.

To continue, we display the distribution of the iconicity level of words according to their frequency.

```r
ggplot(data, aes(
    y=after_stat(density),  # density to normalize the frequency values
    x=Iconicity,
    fill=Group)
  ) +
  geom_histogram(binwidth=0.1, alpha=0.7) +
  scale_fill_manual(values=c("Verb"="green", "Non-verb"="salmon")) +
  ggtitle("Iconicity Distribution per Group") +
  xlab("Iconicity") +
  ylab("Norm. frequency")
```



After accounting for the frequency variation across groups by normalizing their values, we see similar graphs for both groups. For both groups we see a normal bell curve, where the most frequent words sit around 0.7 for the non-verbs, but slightly higher for the verbs, at 1.23.

The most iconic words fall well under the 0.25 distribution on their frequency for both groups. While it seems verbs do have a more linear distribution than their contrasted group, I believe this is best explained by the outliers of the grammatical words.

Overall, it seems that verbs are slightly more iconic than other groups, but the variance may be better explained through other factors.