Visualing Free Response Data

Datasets provided:

main.items : items from the main conditions of the experiment, plus demographic information. Conditions: both, speaker, listener, none.

fillers : filler items, plus demographic information. Conditions: true, false.

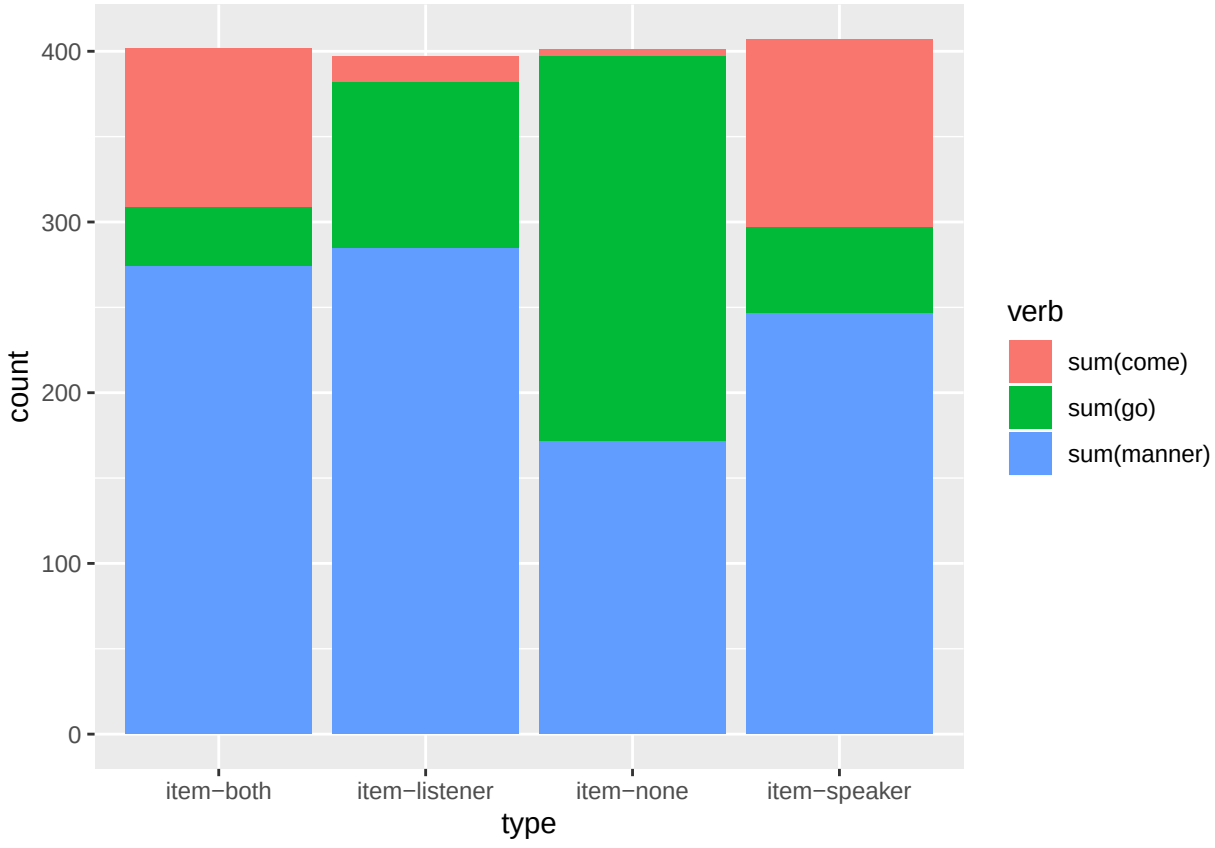
spatial.items : spatial control items. Conditions: left, between, close.

Main items

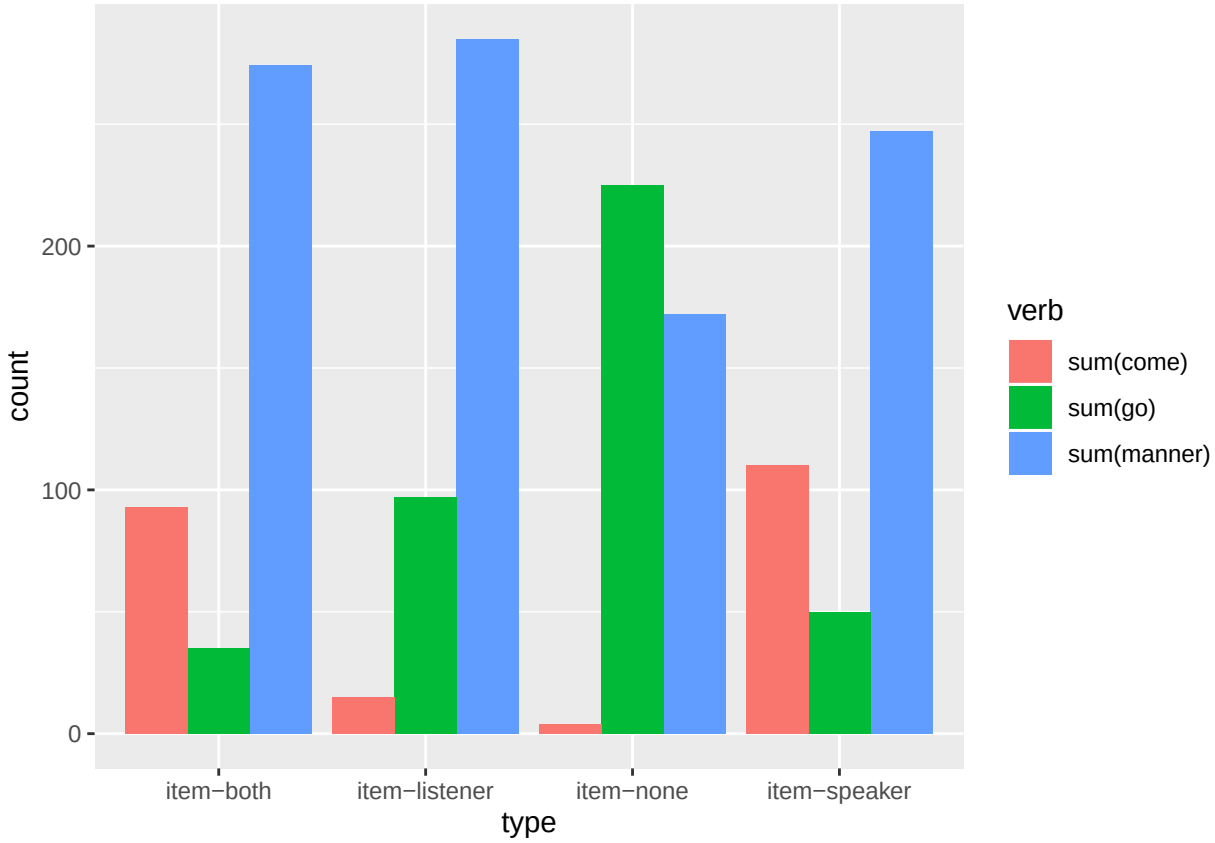
```
#Bar graph of participant means by condition
verb.counts <- main.items %>% group_by(type) %>% summarize(sum(come),sum(go),sum(manner))
```

```
## Warning: Grouping rowwise data frame strips rowwise nature
```

```
means <- melt(verb.counts,id.vars = c("type"))
colnames(means) <- c('type','verb','count')
means$verb <- as.factor(means$verb)
ggplot(means, aes(x=type,y=count,fill=verb)) + geom_bar(stat="identity")
```

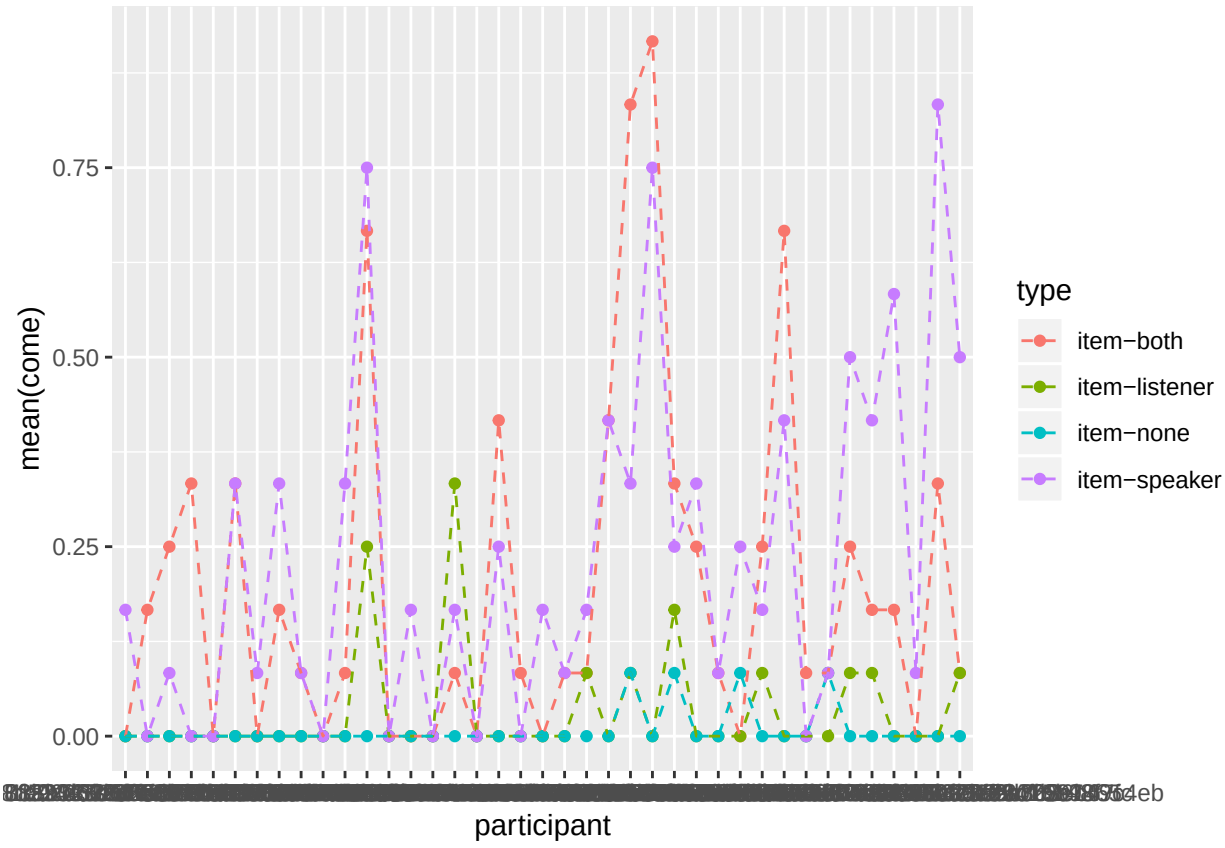


```
ggplot(means, aes(x=type,y=count,fill=verb)) + geom_bar(stat="identity", position = "dodge")
```

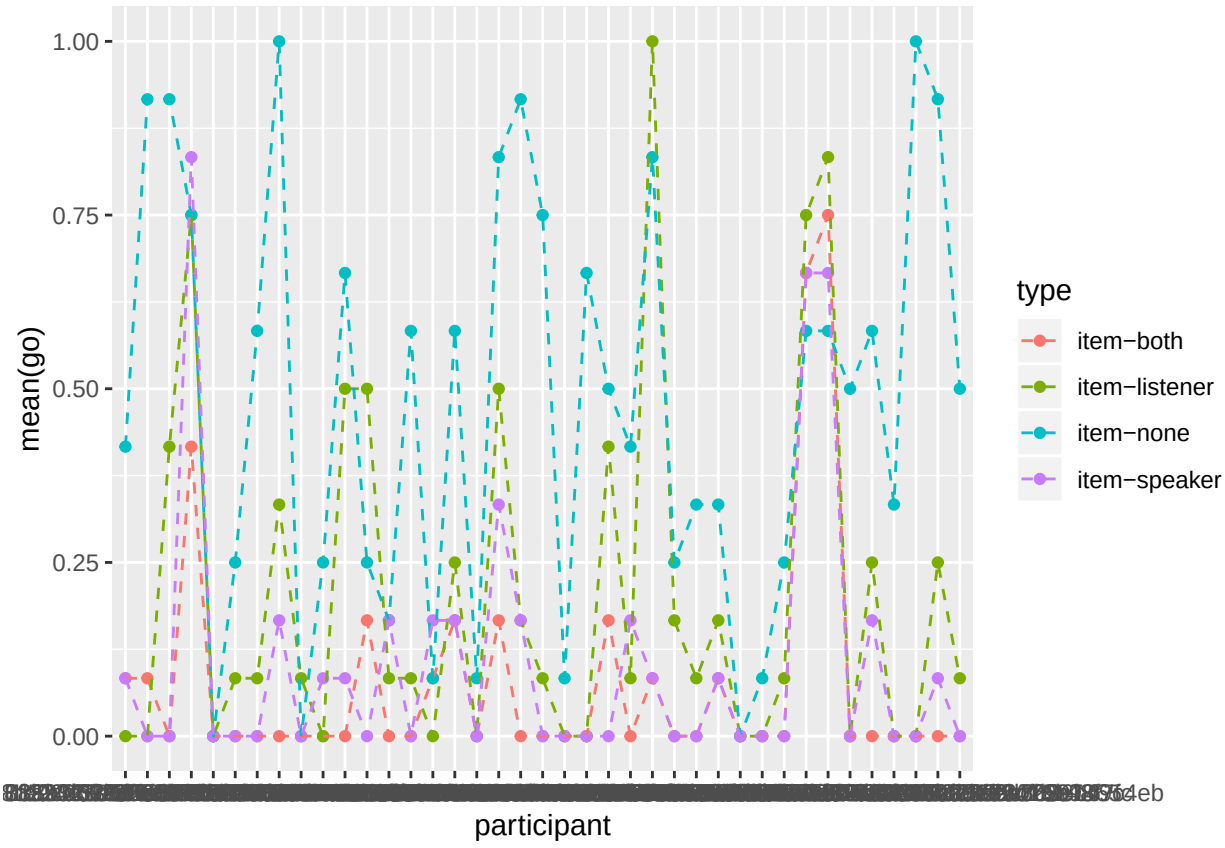


By participant analysis

```
#Plot value by participant
ggplot(subset(mean.by.participant,grepl("item",type,fixed=TRUE)), aes(x=participant,y=`mean(come)`,colou
```




```
ggplot(subset(mean.by.participant,grepl("item",type,fixed=TRUE)), aes(x=participant,y=`mean(go)`,colour=
```



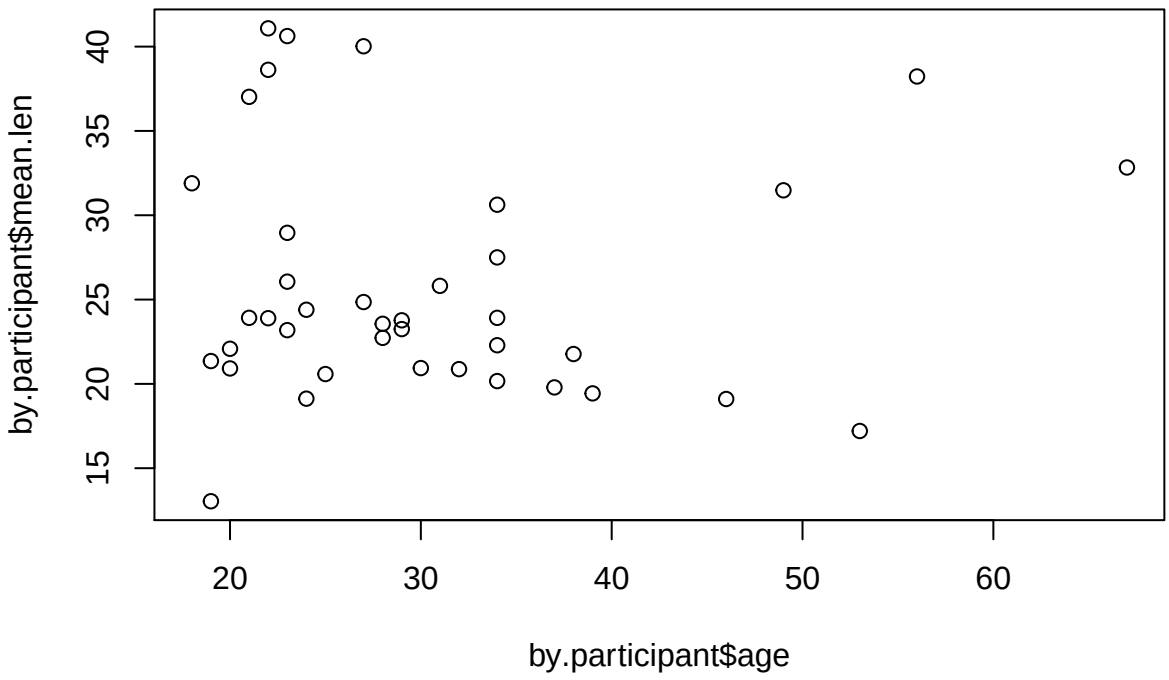
Demographic information

```
#Does one age group write longer answers?

by.participant <- main.items %>% group_by(participant) %>% summarize(mean.len = mean(nchar(answer)), ag
```

```
## Warning: Grouping rowwise data frame strips rowwise nature
```

```
plot(by.participant$age,by.participant$mean.len)
```

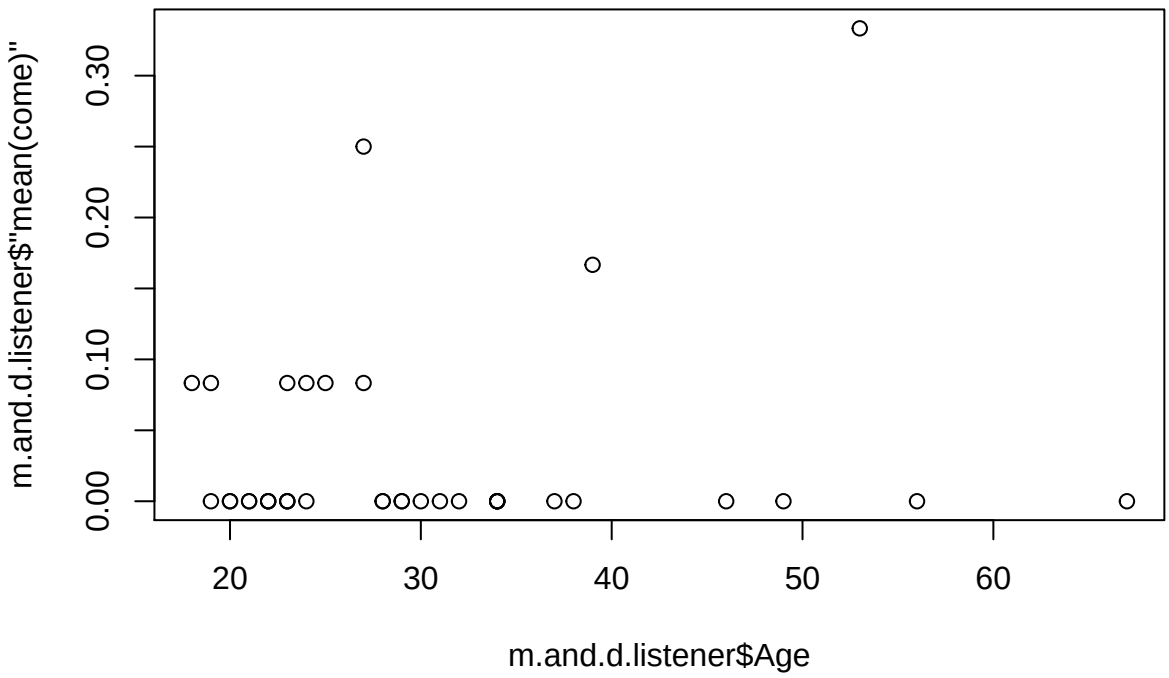


Pick another demographic question to query:

- Does age affect use of 'come' in the listener-only condition?
- Does the use of manner of motion verbs vary by age?
- Does the overall rate of 'come' responses vary by region?

```
#Does age affect use of 'come' in the listener only condition?

means.and.demos <- merge(mean.by.participant,db.cast,by="participant")
m.and.d.listener <- subset(means.and.demos,type=='item-listener')
plot(m.and.d.listener$Age,m.and.d.listener$'mean(come)')
```



More visualization using the quanteda package

```
main.corpus <- corpus(main.items, docid_field="doc_id", text_field = "answer") #create a corpus
main.dfm <- dfm(main.corpus) #create a document frequency matrix (show the dfm to class)
set.seed(100) #why do we need to set a seed?
textplot_wordcloud(main.dfm)
```

Let's make our word cloud neater by excluding stop words and punctuation.

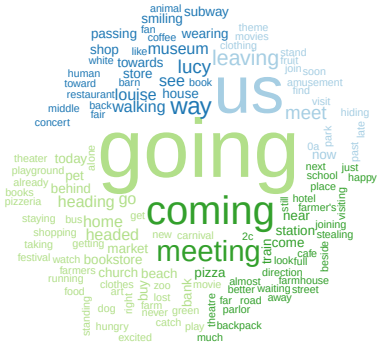
```
main.dfm <- dfm(main.corpus, remove = stopwords("English"), remove_punct = TRUE)
set.seed(100)
textplot_wordcloud(main.dfm,max_size = 8)
```

Plotting a comparison word cloud

```
compare.dfm <- dfm_group(main.dfm, groups = "type", fill = FALSE, force = FALSE)
textplot_wordcloud(compare.dfm,comparison = TRUE)
```

item−listener

item−both



item−none

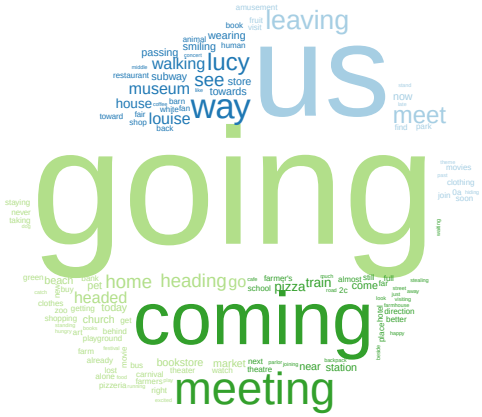
item−speaker

Experiment with scaling produce a better graphic

```
#Word cloud plotting can be misleading! How do you decide the right scale to display the words?
compare.dfm <- dfm_group(main.dfm, groups = "type", fill = FALSE, force = FALSE)
textplot_wordcloud(compare.dfm,comparison = TRUE,min_size = 0.25,max_size = 8)
```

item−listener

item−both



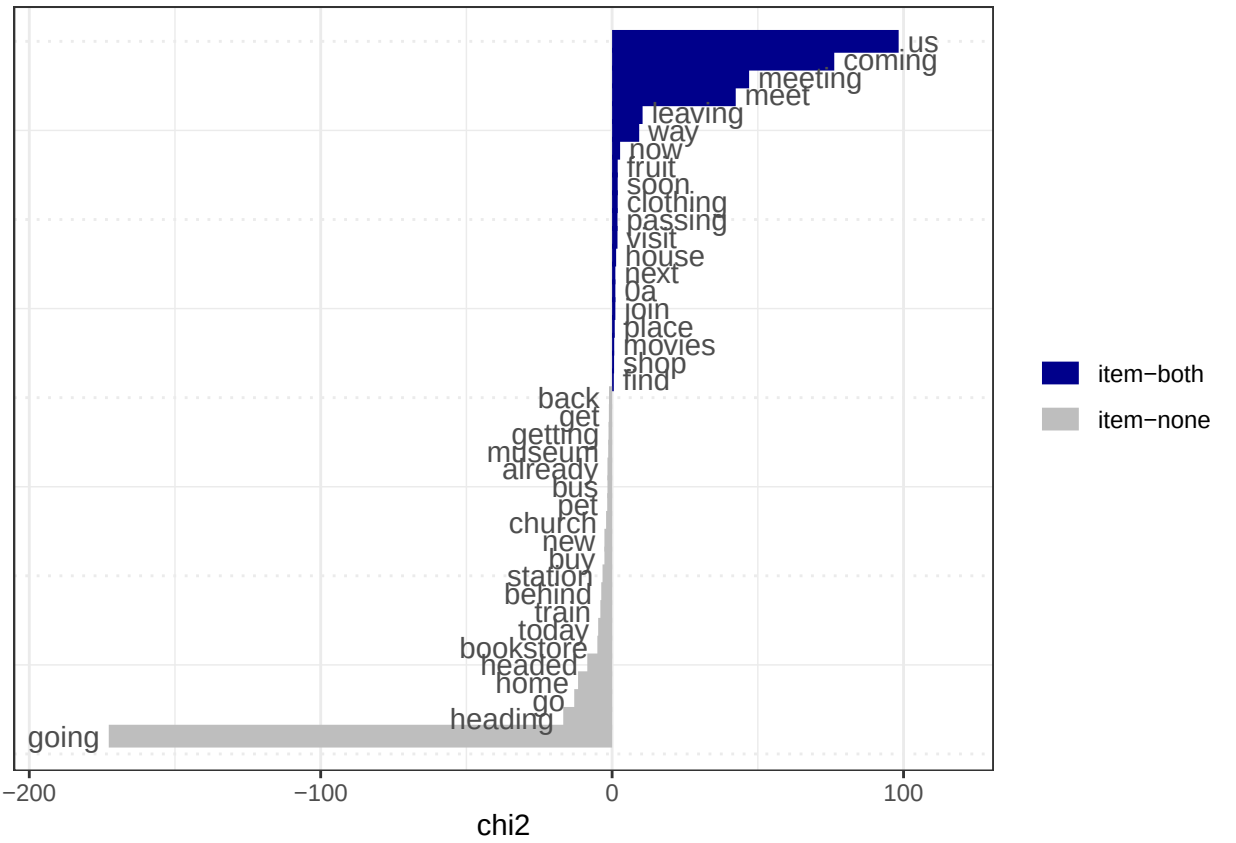
item−none

item−speaker

Compare "keyness" across document groups

Keyness is a measure of how distinguishing a word is as a feature of a particular document (or group of documents). There are several measures available in the quanteda package: Pearson's chi squared test, Fisher's exact test, likelihood ratio, and pointwise mutual information.

```
none.both.corpus <- corpus(subset(main.items,type=="item-none" | type=="item-both"), docid_field="doc_i
none.both.dfm <- dfm(none.both.corpus, remove = stopwords("English"), remove_punct = TRUE)
compare.none.both.dfm <- dfm_group(none.both.dfm,groups = "type", fill = FALSE, force = FALSE)
compare.none.both.keyness <- textstat_keyness(compare.none.both.dfm, measure='chi2')
textplot_keyness(compare.none.both.keyness)
```



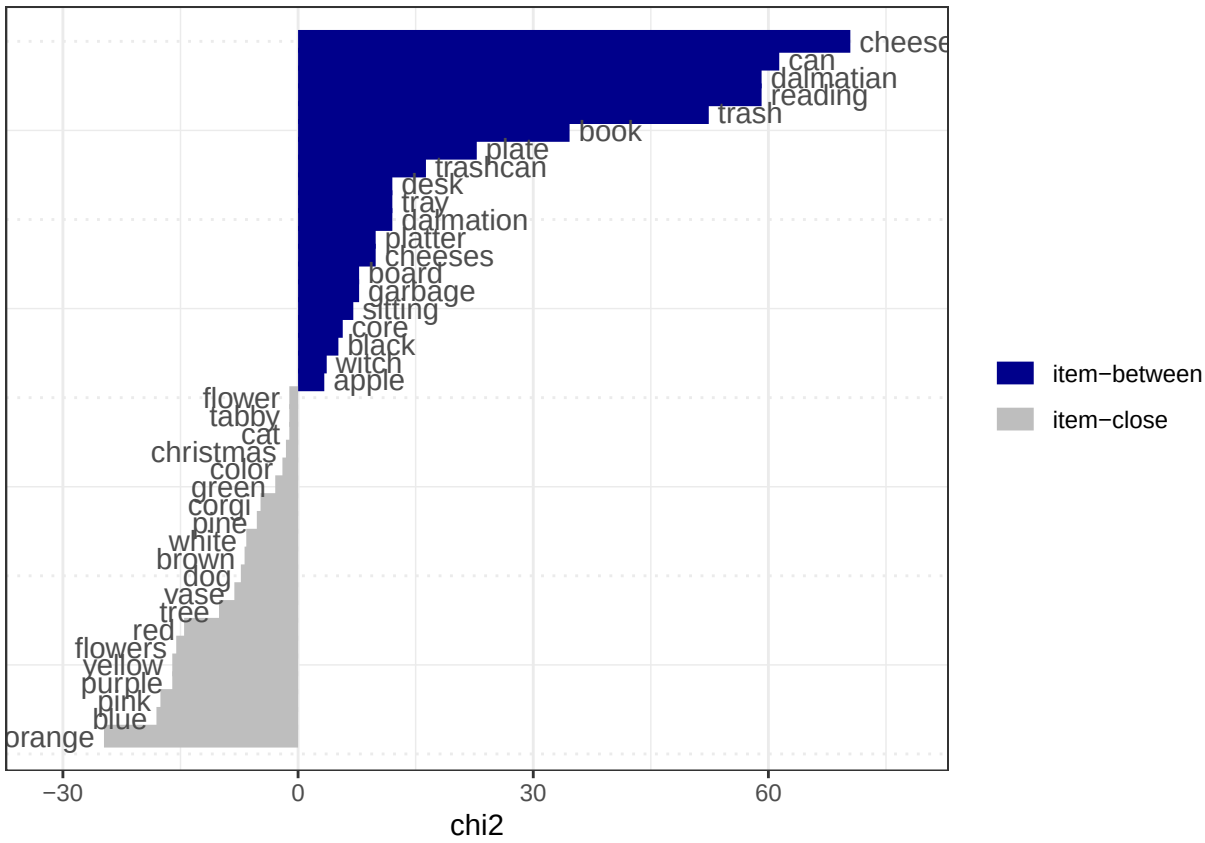
Here we are using chi squared, which is a measure of how different the observed distribution of a word is from the expected distribution if the distribution was the same between the two groups of documents. (So, the same as a chi squared test when used as a test of statistically significant differences between groups.)

Experiment with different measures of keyness: how does this qualitatively change the results?

Exercise: create a keyness visualization for another comparison of interest

- Male speakers versus female speakers
- False fillers versus true fillers
- Between spatials versus left spatials

```
spatial.corpus <- corpus(subset(spatial.items,type=="item-close" | type=="item-between"), docid_field="
spatial.dfm <- dfm(spatial.corpus, remove = stopwords("English"), remove_punct = TRUE)
spatial.dfm <- dfm_group(spatial.dfm,groups = "type", fill = FALSE, force = FALSE)
spatial.keyness <- textstat_keyness(spatial.dfm, measure='chi2')
textplot_keyness(spatial.keyness)
```



Which words are most closely associated with each other?

```
textplot_network(main.dfm,min_freq = 10)
```

```
## Registered S3 method overwritten by 'network':
##   method            from
##   summary.character quanteda
```

Play around with the frequency cut-off.