Validation results

Meg Cychosz

19 October 2020

```
# get total # of clips from each recording
complete2 <- complete %>%
  group_by(id) %>%
  distinct(file_name, .keep_all = T) %>%
  mutate(num_clips = NROW(Media)*2)

clips <- complete2 %>%
  select(id, num_clips) %>%
  distinct(id, .keep_all = T)

data <- merge(clips, random, by='id')

data2 <- rbind(data, complete2)
```

0.0.1 Language categories across random and full methods

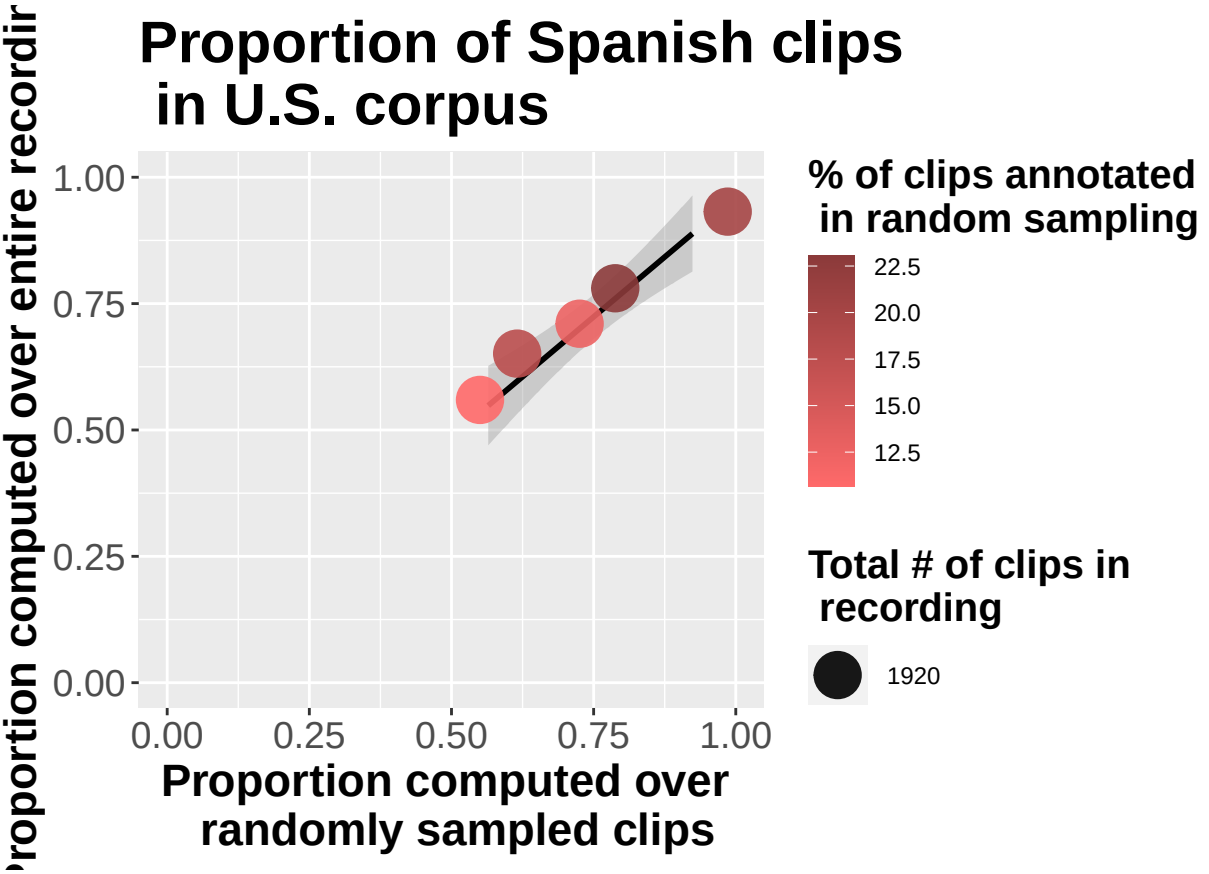
```
# get # and % of clips annotated/listened to (doesn't factor in repeats for random method)
# percen_anno = % of clips from total recording that were drawn
# speech_clips = # of clips containing identifiable speech (excluding child voc, media, and unsure)
# we may want a different denominator at some point - like the total number of clips drawn - but this c
data3 <- data2 %>%
  group_by(id, method) %>%
  mutate(percen_anno = (NROW(file_name)/num_clips)*100) # percen clips of total recording *listened* to

data_cts <- data3 %>%
 gather("addressee", "language", Adult2OtherChild, Adult2Others, Adult2TargetChild, Adult2unsure, Other
  filter(language=='Mixed' | language=='Spanish' | language=='English/Quechua') %>% # only clips w spee
  distinct_at(., vars(file_name, language), .keep_all = T) %>% # don't record multiple speakers speakin
  mutate(speech_clips = NROW(file_name))

que <- data_cts %>%
  group_by(id, method) %>%
  filter(language=='English/Quechua') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>% # irrespective of speaker/addressee; by-child only
  mutate(n_que=n()) %>%
  distinct_at(., vars(id, method), .keep_all = T) %>%
  mutate(percen_que = n_que / speech_clips) # compute que/eng percens

span <- data_cts %>%
  group_by(id, method) %>%
```

```
  filter(language=='Spanish') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_span = n())  %>%
  distinct_at(., vars(id, method), .keep_all = T) %>%
  mutate(percen_span = n_span / speech_clips) # compute span percens

mixed <- data_cts %>%
  group_by(id, method) %>%
  filter(language=='Mixed') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_mxd = n())  %>%
  distinct_at(., vars(id, method), .keep_all = T) %>%
  mutate(percen_mxd = n_mxd / speech_clips) # compute mixed percens
```

```
# compute ratio differently for the two corpora
#span$que2span_ratio <- que$n_que/span$n_span
#span$eng2span_ratio <- que$n_que/span$n_span

vars <- data_cts %>% colnames(.)
final_data <- span %>%
  merge(., data_cts, by=vars) %>%
  select(id, num_clips, age_YYMMDD, gender, location, method, percen_span, percen_anno, speech_clips)

final_data2 <-
  merge(final_data, que, by=c('id', 'gender', 'age_YYMMDD', 'location', 'method', 'num_clips', 'percen_a
  select(id, gender, location, method, percen_span, percen_que, num_clips, percen_anno, speech_clips)

plot_data <-
  merge(final_data2, mixed, by=c('id', 'gender', 'location', 'method', 'num_clips', 'percen_anno', 'spe
  select(id, gender, location, method, percen_span, percen_que, percen_mxd, num_clips, percen_anno, spe

# sanity check: calculate percen mixed + spanish + english/quechua
plot_data$total <- plot_data$percen_mxd + plot_data$percen_span + plot_data$percen_que
```

```
# for later
per_ann <- plot_data %>%
  filter(method=='random') %>%
  select(id, percen_anno)

us_plot <- plot_data %>%
  filter(location=='US') %>%
  select(-percen_que, -percen_anno, -percen_mxd, -speech_clips) %>%
  spread("method", "percen_span") %>%
  merge(., per_ann, by='id') %>%
ggplot(., aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_anno,2)),alpha=.9,position = position_jitter(width =
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
```

```
  ylab("Proportion computed over entire recording") +
  xlab("Proportion computed over \n randomly sampled clips") +
  ylim(0,1) +
  xlim(0,1)+
  #facet_wrap(~location, scales = "free") +
  labs(col='% of clips annotated \n in random sampling', title = 'Proportion of Spanish clips \n in U.S
 theme(title = element_text(size=18, face="bold"),
   axis.text=element_text(size=14),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=15)) +
      guides(size=guide_legend(title="Total # of clips in \n recording"))
us_plot
```

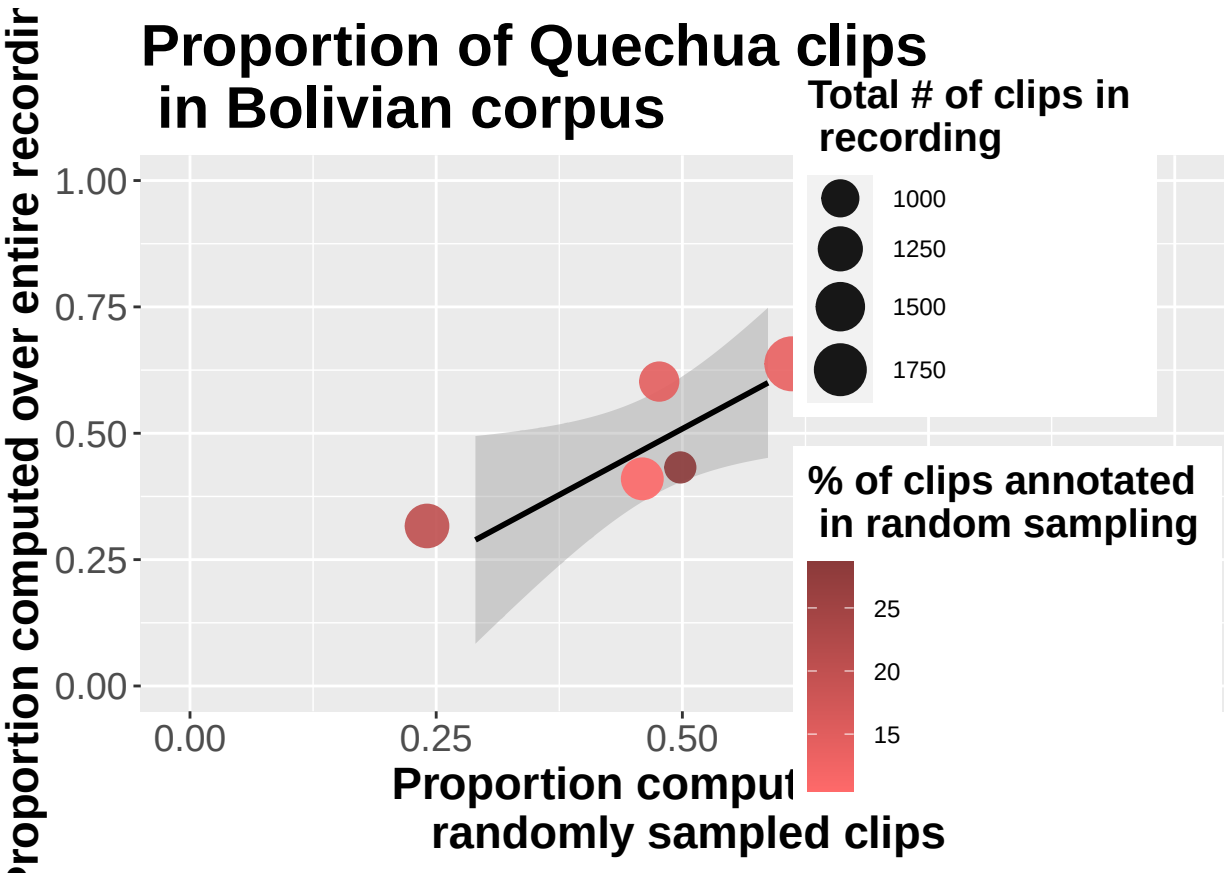



```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/us_plot.jpeg", height = 500, width = 60
us_plot
dev.off()
```

```
## pdf
##   2
```

```
bo_plot <- plot_data %>%
  filter(location=='Bolivia') %>%
  select(-percen_span, -percen_anno, -percen_mxd, -speech_clips) %>%
  spread("method", "percen_que") %>%
  merge(., per_ann, by='id') %>%
```

```
ggplot(., aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_anno,2)),alpha=.9,position = position_jitter(width =
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
  ylab("Proportion computed over entire recording") +
  xlab("Proportion computed over \n randomly sampled clips") +
  ylim(0,1) +
  xlim(0,1)+
  #facet_wrap(~location, scales = "free") +
  labs(col='% of clips annotated \n in random sampling', title = 'Proportion of Quechua clips \n in Boli
 theme(title = element_text(size=18, face="bold"),
   axis.text=element_text(size=14),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=15),
       legend.position = c(.8, .5)) +
      guides(size=guide_legend(title="Total # of clips in \n recording"))
bo_plot
```




```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/bolivia_plot.jpeg", height = 500, width
bo_plot
dev.off()
```

```
## pdf
##   2
```

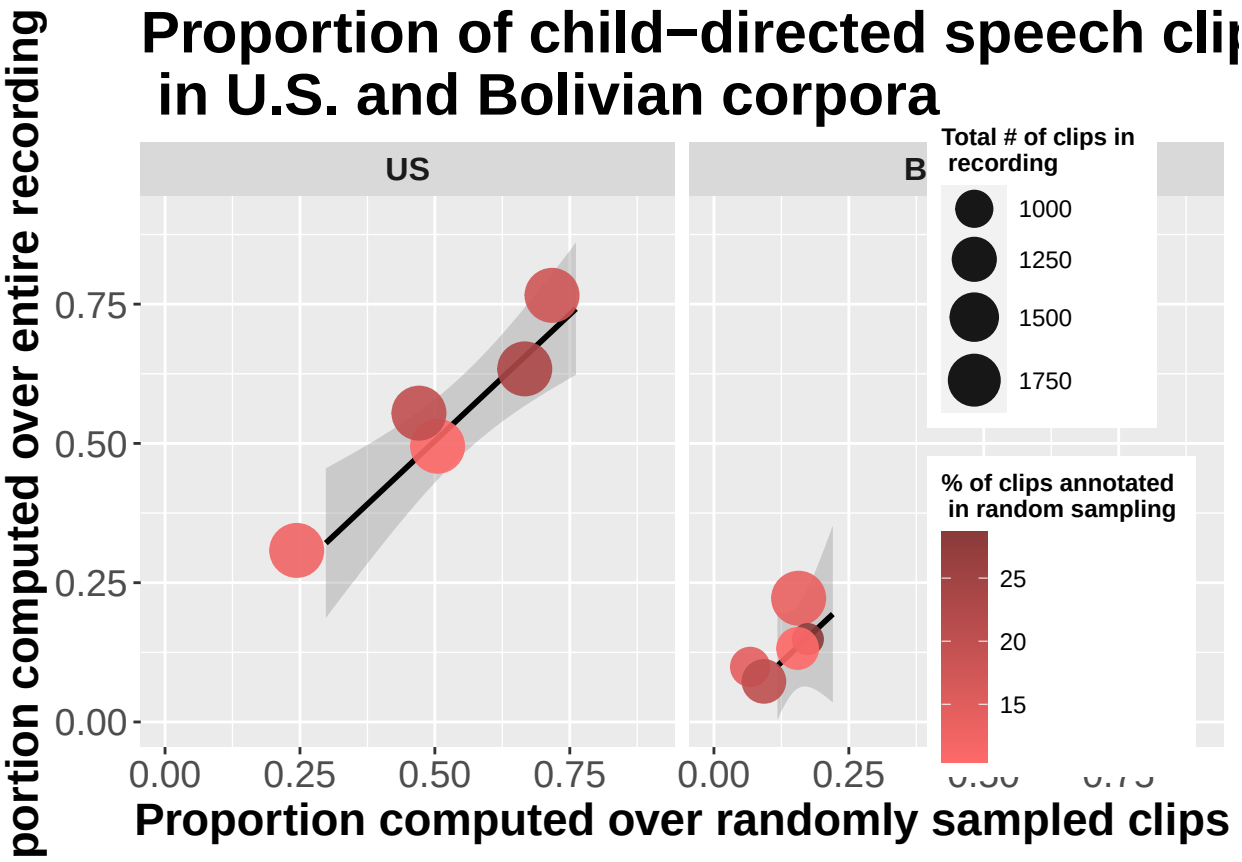
```r
us_cor <- plot_data %>%
  select(-percen_que, -percen_anno, -percen_mxd, -speech_clips) %>%
  spread("method", "percen_span") %>%
  filter(location=='US') %>%
  summarize(., cor(complete, random))

bo_cor <- plot_data %>%
  select(-percen_span, -percen_anno, -percen_mxd, -speech_clips) %>%
  spread("method", "percen_que") %>%
  filter(location=='Bolivia') %>%
  summarize(., cor(complete, random))
```

0.0.2 Chid-directed speech across random and full methods

```r
# target child-directed speech:all registers, irrespective of speaker
cds_cts <- data3 %>%
  select(-Adult2unsure, -Otherchild2unsure) %>% # remove unsure register
  group_by(id, method) %>%
  gather("addressee", "language", Adult2TargetChild, Adult2Others, Adult2OtherChild, Otherchild2TargetCl
  filter(language == 'English/Quechua' | language =='Mixed' | language=='Spanish' | language=='Unsure')


cds <- cds_cts %>%
  group_by(id, method) %>%
  filter(addressee=='Adult2TargetChild' | addressee=='Otherchild2TargetChild') %>%
  distinct(file_name, .keep_all = T) %>%
  mutate(n_cds = n()) %>% # # of CDS clips
  distinct_at(., vars(id, method), .keep_all = T)

ads <- cds_cts %>%
  filter(addressee=='Adult2Others' | addressee=='Otherchild2adults') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_ads = n()) %>% # # of ADS clips
  distinct_at(., vars(id, method), .keep_all = T) %>%
  select(id, method, n_ads)

o_child <- cds_cts %>%
  filter(addressee=='Adult2OtherChild' | addressee=='Otherchild2OtherChild') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_ods = n()) %>% # # of ODS clips
  distinct_at(., vars(id, method), .keep_all = T) %>%
  select(id, method, n_ods)

vars <- cds_cts %>% colnames(.)



o <- cds %>%
  merge(., cds_cts, by=vars) %>%
```

```
  select(id, num_clips, age_YYMMDD, gender, location, method, percen_anno, n_cds)
o2 <- merge(o, ads, by=c('method', 'id'))
o3 <- merge(o2, o_child, all = T)
o3[is.na(o3)] <- 0 # one child doesn't have any ODS

o3$total_reg_clips <- o3$n_ods + o3$n_cds + o3$n_ads

o3$percen_cds <- o3$n_cds / o3$total_reg_clips
o3$percen_ads <- o3$n_ads / o3$total_reg_clips
o3$percen_ods <- o3$n_ods / o3$total_reg_clips

# sanity check
o3$total <- o3$percen_ods + o3$percen_ads + o3$percen_cds
```

```
# for later
percen_cds_df <- o3 %>%
  distinct_at(., vars(id, method), .keep_all = T) %>%
  filter(method=='random') %>%
  select(id, percen_anno) # get the % of clips annotated for each id and method


bo_plot_data <- o3 %>%
  #filter(location=='Bolivia') %>%
  select(id, gender, location, num_clips, method, percen_cds) %>%
  spread("method", "percen_cds") %>%
  merge(., percen_cds_df, by='id')

# reorder location variable
bo_plot_data$location <- factor(bo_plot_data$location, levels = c("US", "Bolivia"))

bo_cds_plot <- ggplot(bo_plot_data, aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_anno,2)),alpha=.9,position = position_jitter(width =
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
  ylab("Proportion computed over entire recording") +
  xlab("Proportion computed over randomly sampled clips") +
  ylim(0,0.9) +
  xlim(0,0.9)+
  facet_wrap(~location, scales = "fixed") +
  labs(col='% of clips annotated \n in random sampling', title = 'Proportion of child-directed speech cl
 theme(title = element_text(size=18, face="bold"),
   axis.text=element_text(size=14),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=9),
      legend.position = c(.85, .55),
      strip.text.x = element_text(size=12, face="bold")) +
      guides(size=guide_legend(title="Total # of clips in \n recording"))
bo_cds_plot
```




```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/bolivia_cds_plot.jpeg", height = 500, w
bo_cds_plot
dev.off()
```

```
## pdf
##   2
```

```
cds_cors <- bo_plot_data %>%
  group_by(location) %>%
  summarize(., cor(complete, random))
```