Validation results

Meg Cychosz

01 December 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)
```

```
data3 <- data2 %>%
  group_by(method, id) %>%
  mutate(num_clips_drawn = (NROW(file_name))) %>%
  mutate(percen_ofallclips_drawn=(NROW(file_name)/num_clips)*100) # sanity check - complete method shou
```

```
data_annon <- data3 %>%
 gather("addressee", "language", Adult2OtherChild, Adult2Others, Adult2TargetChild, Adult2unsure, Other
  filter(language=='Mixed' | language=='Spanish' | language=='English/Quechua' | language =='Unsure') %
  group_by(id, method) %>%
  distinct_at(., vars(file_name, language), .keep_all = T) %>% # don't record multiple speakers speakin
  mutate(total_annotations = NROW(file_name)) # N of annotations made; distinct from N of speech clips

# separately, calculate the num and % of annotated clips
data_annon_cts <- data_annon %>%
  group_by(id, method) %>%
  distinct(file_name, .keep_all = T) %>%
  mutate(speech_clips = NROW(file_name)) %>% # N of unique clips annotated - NOT the # of annotations
  mutate(percen_ofallclips_annon=(NROW(file_name)/num_clips)*100) %>% # % of total clips annotated
  select(speech_clips, percen_ofallclips_annon, id, method, file_name, num_clips_drawn, percen_ofallclip


for_speech_clips <- data_annon_cts %>%
  select(id, method, speech_clips) %>%
  distinct_at(., vars(id, method), .keep_all = T)

# calculate the num and % of all clips available for annotation
data_annon$Childsleep <- as.factor(data_annon$Childsleep)
data_avbl <- data3 %>%
  group_by(id, method) %>%
  distinct(file_name, .keep_all = T) %>% # two, for random and complete
  mutate(voc = if_else(percents_voc > 0, "1", "0")) %>% # turn percents_voc binary
  filter(sleeping=='1' | PID == '1' | researcher_present == '1' | voc == '0') %>%
  count() %>%
  rename(not_avl_clips = n) %>%
  merge(., data_annon, by=c('id', 'method')) %>%
  mutate(avbl_clips = num_clips - not_avl_clips) %>% # clips that were *available* for annotation
  merge(., for_speech_clips, by=c('id', 'method')) %>% # N of unique clips annotated - NOT the # of ann
  mutate(percen_avl_annon = (speech_clips / avbl_clips)*100) %>% # the % of available clips that were a
  distinct_at(., vars(id, method), .keep_all = T) %>%
  group_by(method) %>%
  mutate(avbl_clips = paste(speech_clips, "(",round(percen_avl_annon,2),"%)")) %>%
  ungroup()%>%
  select(avbl_clips, id, method) %>%
  pivot_wider(names_from=method, values_from=c("avbl_clips"))
```

```
percen_tbl <- data_annon_cts %>%
  select(-file_name) %>%
  distinct_at(., vars(id,method), .keep_all = T) %>%
  mutate(clips_drawn = paste(num_clips_drawn,"(",round(percen_ofallclips_drawn,2),"%)")) %>%
  mutate(clips_annon = paste(speech_clips,"(",round(percen_ofallclips_annon,2),"%)")) %>%
  select(-num_clips_drawn, -percen_ofallclips_annon, -speech_clips, -percen_ofallclips_drawn) %>%
  relocate(c(id, method, clips_drawn, clips_annon)) %>%
  pivot_wider(names_from=method, values_from=c("clips_drawn", "clips_annon")) %>%
  merge(., data_avbl, by=c('id'))

percen_tbl$id <- plyr::mapvalues(percen_tbl$id,
                                from=c('267-12mo', '261-8mo', '199', '198-9mo', '179', '1081', '1077',
             to=c('Spanish-English (267)', 'Spanish-English (261)','Spanish-English (199)',
  'Spanish-English (198)', 'Spanish-English (179)', 'Quechua-Spanish (1081)', 'Quechua-Spanish (1077)',

# actually decided to split this table and move part to the appendix
clip_annon_tbl <- percen_tbl %>%
  select(id, clips_annon_random, clips_annon_complete) %>%
  arrange(desc(id))

knitr::kable(clip_annon_tbl, caption = 'Number of clips annotated by child and annotation method.',
             booktabs=T,
             row.names = FALSE,
             col.names = c("Corpus (ID)", "Random", "Complete")) %>% # "
  kable_styling() %>%
  add_header_above(c(" " = 1, "# of clips annotated (% of total clips)" = 2)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

\begin{table}[!h]

\caption{(#tab:% drawn and annotated table)Number of clips annotated by child and annotation method.}

	# of clips annotated (% of total clips)	
Corpus (ID)	Random	Complete
Spanish-English (267)	101 (5.26 %)	274 (14.27 %)
Spanish-English (261)	92 (4.79 %)	294 (15.31 %)
Spanish-English (199)	118 (6.15 %)	467 (24.32 %)
Spanish-English (198)	81 (4.22 %)	302 (15.73 %)
Spanish-English (179)	120 (6.25 %)	633 (32.97 %)
Quechua-Spanish (1081)	92 (7.5 %)	285 (23.25 %)
Quechua-Spanish (1077)	83 (7.23 %)	355 (30.92 %)
Quechua-Spanish (1075)	81 (8.69 %)	199 (21.35 %)
Quechua-Spanish (1060)	111 (10.51 %)	405 (38.35 %)
Quechua-Spanish (1032)	97 (5.05 %)	372 (19.38 %)

\end{table}

```
clip_drawn_avbl_tbl <- percen_tbl %>%
  select(-clips_annon_random, -clips_annon_complete) %>%
  relocate(id, clips_drawn_random, clips_drawn_complete, random, complete) %>%
  arrange(desc(id))

knitr::kable(clip_drawn_avbl_tbl, caption = 'Number of clips drawn and number of clips annotated, by chi
             booktabs=T,
             row.names = FALSE,
             col.names = c("Corpus (ID)", "Random", "Complete", "Random", "Complete")) %>% # "
  kable_styling() %>%
  add_header_above(c(" " = 1, "# of clips drawn (% of total clips)" = 2, "# of clips annotated (% of ava
  kableExtra::kable_styling(latex_options = "hold_position")
```

\begin{table}[!h]
\caption{(#tab:% drawn and annotated table)Number of clips drawn and number of clips annotated, by child and annotation method.}

	# of clips drawn (% of total clips)		# of clips annotated (% of available clips)	
Corpus (ID)	Random	Complete	Random	Complete
Spanish-English (267)	345 (17.97 %)	960 (50 %)	101 (5.81 %)	274 (20.49 %)
Spanish-English (261)	290 (15.1 %)	960 (50 %)	92 (5.06 %)	294 (19.32 %)
Spanish-English (199)	192 (10 %)	960 (50 %)	118 (6.37 %)	467 (30.95 %)
Spanish-English (198)	284 (14.79 %)	960 (50 %)	81 (4.52 %)	302 (20.54 %)
Spanish-English (179)	192 (10 %)	960 (50 %)	120 (6.36 %)	633 (37.08 %)
Quechua-Spanish (1081)	249 (20.31 %)	613 (50 %)	92 (8.16 %)	285 (30.25 %)
Quechua-Spanish (1077)	137 (11.93 %)	574 (50 %)	83 (7.33 %)	355 (32.84 %)
Quechua-Spanish (1075)	267 (28.65 %)	466 (50 %)	81 (9.69 %)	199 (26.39 %)
Quechua-Spanish (1060)	154 (14.58 %)	528 (50 %)	111 (10.66 %)	405 (40.91 %)
Quechua-Spanish (1032)	263 (13.7 %)	960 (50 %)	97 (5.38 %)	372 (25.92 %)

\end{table}

0.0.1 Language categories across random and full methods

```
lang_annon <- data_annon %>%
  filter(language=='Mixed' | language=='Spanish' | language=='English/Quechua') %>% # only clips where
  group_by(id, method) %>%
  distinct_at(., vars(file_name, language), .keep_all = T) %>% # don't record multiple speakers speakin
  mutate(total_lang_annotations = NROW(file_name)) # N of language annotations made; distinct from N of

que <- lang_annon %>%
  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 / total_lang_annotations) # compute que/eng ratio

span <- lang_annon %>%
  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 / total_lang_annotations) # compute span ratio

mixed <- lang_annon %>%
  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 / total_lang_annotations) # compute mixed ratio
```

```
vars <- data_annon_cts %>%
  select(percen_ofallclips_drawn, id, method) %>%
  colnames(.)

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

final_data2 <-
  merge(final_data, que, by=c('id', 'method', 'percen_ofallclips_drawn', 'gender', 'location', 'num_clip
  select(id, gender, location, method, percen_span, percen_que, num_clips, percen_ofallclips_drawn, spe

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

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

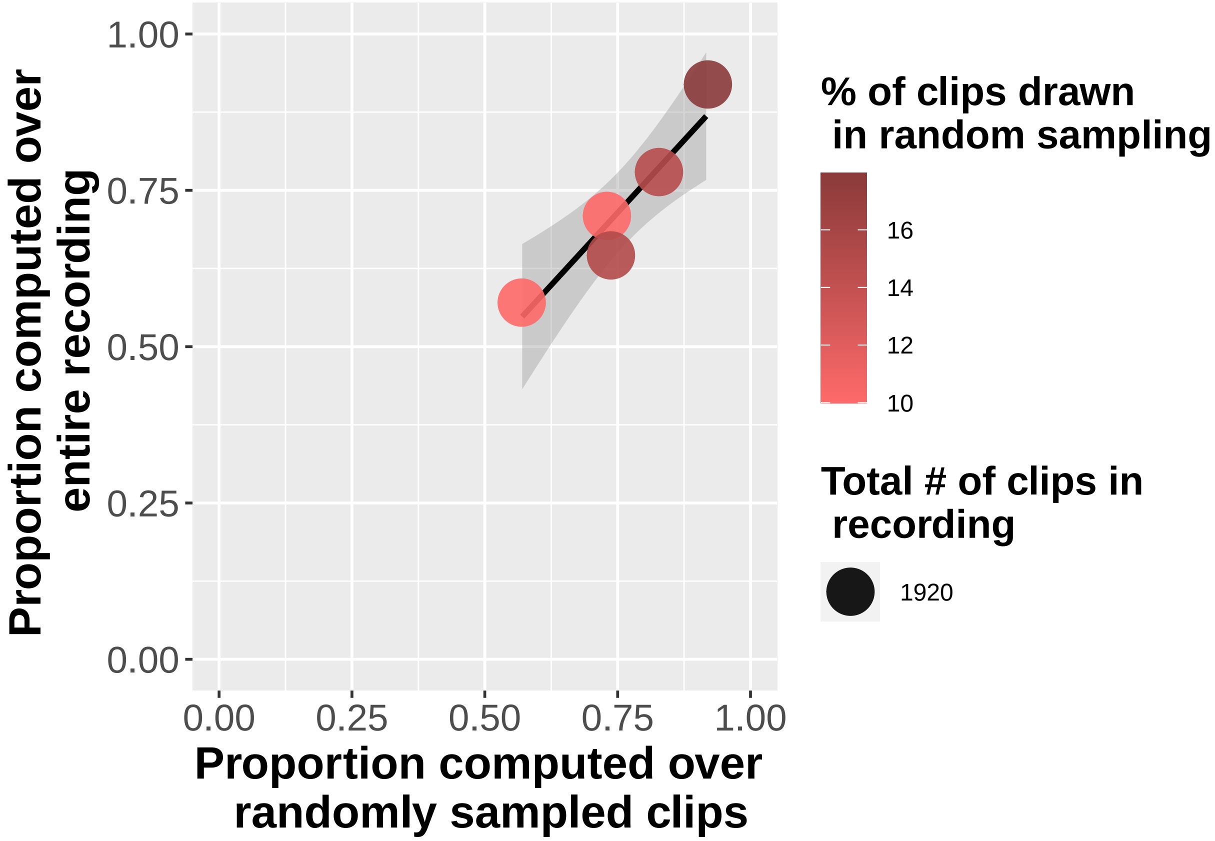
```r
# compute correlations
us_cor <- plot_data %>%
  distinct_at(., vars(method, id), .keep_all = T) %>%
  select(method, id, percen_span, location) %>%
  spread("method", "percen_span") %>%
  filter(location=='US') %>%
  summarize(., paste("r=",round(cor.test(complete, random)$estimate,2),",","p=",round(cor.test(complete

bo_cor <- plot_data %>%
  distinct_at(., vars(method, id), .keep_all = T) %>%
  select(method, id, percen_que, location) %>%
  spread("method", "percen_que") %>%
  filter(location=='Bolivia') %>%
  summarize(., paste("r=",round(cor.test(complete, random)$estimate,2),",","p=",round(cor.test(complete

# compute avg. %s of target lang categories
us_lang_tbl <- plot_data %>%
  filter(location=='US') %>%
  group_by(method) %>%
  summarize(avg=round(mean(percen_span),2),
            sd=round(sd(percen_span),2)) %>%
  mutate(stats=paste(avg,"(",sd,")")) %>%
  select(-avg, -sd) %>%
  spread(key='method', value = "stats")

bo_lang_tbl <- plot_data %>%
  filter(location=='Bolivia') %>%
  group_by(method) %>%
  summarize(avg=round(mean(percen_que),2),
            sd=round(sd(percen_que),2)) %>%
  mutate(stats=paste(avg,"(",sd,")")) %>%
  select(-avg, -sd) %>%
  spread(key='method', value = "stats")

# calculate relative errors
us_rel_error <- plot_data %>%
  filter(location=='US') %>%
  group_by(method, id) %>%
  summarize(avg=mean(percen_span)) %>%
  spread(key='method', value='avg') %>%
   mutate(relative_error = ((abs((random - complete)) / complete)*100),
         avg_rel_error = round(mean(relative_error),2),
         sd_rel_error = round(sd(relative_error),2)) %>%
  mutate(rel_error_stats=paste(avg_rel_error,"(",sd_rel_error,")")) %>%
  distinct(rel_error_stats)

bo_rel_error <- plot_data %>%
  filter(location=='Bolivia') %>%
  group_by(method, id) %>%
  summarize(avg=mean(percen_que)) %>%
  spread(key='method', value='avg') %>%
  mutate(relative_error = ((abs((random - complete)) / complete)*100),
         avg_rel_error = round(mean(relative_error),2),
         sd_rel_error = round(sd(relative_error),2)) %>%
  mutate(rel_error_stats=paste(avg_rel_error,"(",sd_rel_error,")")) %>%
  distinct(rel_error_stats)

# add correlations to table - will make pretty below
us_lang_tbl <- cbind(us_lang_tbl, us_cor) %>%
  cbind(., us_rel_error) %>%
  mutate(Corpus = "Spanish-English (Spanish)") %>%
  relocate(c(Corpus, random, complete))

bo_lang_tbl <- cbind(bo_lang_tbl, bo_cor) %>%
  cbind(., bo_rel_error) %>%
  mutate(Corpus = "Quechua-Spanish (Quechua)") %>%
  relocate(c(Corpus, random, complete))

lang_tbl <- rbind(us_lang_tbl, bo_lang_tbl)
```

```r
knitr::kable(lang_tbl, caption = 'Minority language estimates by corpus and annotation method.',
             booktabs=T,
             row.names = FALSE,
             col.names = c("Corpus (language)", "Random", "All-day", "Correlation between estimates", "
  kable_styling() %>%
  add_header_above(c(" " = 1, "Annotation Method" = 2, " " = 2)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 1: (#tab:generate lang tables)Minority language estimates by corpus and annotation method.

	Annotation Method			
Corpus (language)	Random	All-day	Correlation between estimates	Average relative error (SD)
Spanish-English (Spanish)	0.75 (0.13)	0.69 (0.12)	r= 0.96 , p= 0.01	5.36 (4.82)
Quechua-Spanish (Quechua)	0.48 (0.11)	0.5 (0.12)	r= 0.9 , p= 0.04	11.02 (4.28)

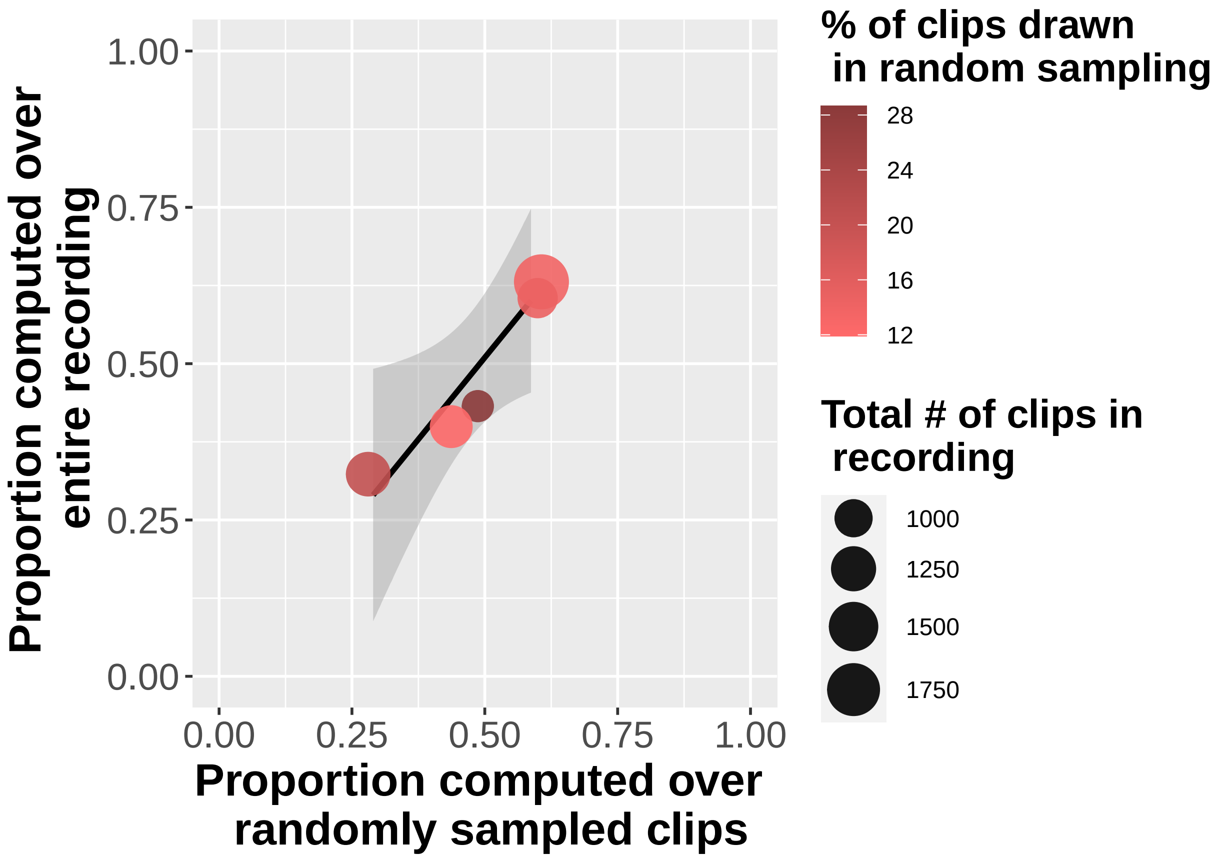
```r
# for later
per_ann <- plot_data %>%
  filter(method=='random') %>%
  select(id, percen_ofallclips_drawn)

us_plot <- plot_data %>%
  filter(location=='US') %>%
  distinct_at(., vars(method, id), .keep_all = T)%>%
  select(-percen_que, -percen_ofallclips_drawn, -percen_mxd, -speech_clips, -total) %>%
  spread("method", "percen_span") %>%
  merge(., per_ann, by='id') %>%
  distinct(id, .keep_all = T) %>%
ggplot(., aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_ofallclips_drawn,2)),alpha=.9,position = position_j
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
  ylab("Proportion computed over \n 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 drawn \n in random sampling') +
       #title = 'Proportion of Spanish clips \n in U.S. corpus') +
 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') %>%
  distinct_at(., vars(method, id), .keep_all = T) %>%
  select(-percen_span, -percen_ofallclips_drawn, -percen_mxd, -speech_clips, -total) %>%
  spread("method", "percen_que") %>%
  merge(., per_ann, by='id') %>%
  distinct(id, .keep_all = T) %>%
ggplot(., aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_ofallclips_drawn,2)),alpha=.9,position = position_ji
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
  ylab("Proportion computed over \n 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 drawn \n in random sampling') +
       #title = 'Proportion of Quechua clips \n in Bolivian corpus') +
 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
```

0.0.2 Chid-directed speech across random and full methods

```
reg_annon <- data_annon %>%
  filter(addressee=='Adult2TargetChild' | addressee=='Otherchild2TargetChild' | addressee=='Adult2Others
  group_by(id, method) %>%
  distinct_at(., vars(file_name, addressee), .keep_all = T) %>% # don't record multiple speakers speaki
  mutate(total_reg_annotations = NROW(file_name))# N of register annotations made; distinct from N of s

cds <- reg_annon %>%
  group_by(id, method) %>%
  filter(addressee=='Adult2TargetChild' | addressee=='Otherchild2TargetChild') %>%
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_cds = n()) %>% # # of CDS clips
  distinct_at(., vars(id, method), .keep_all = T) %>%
  mutate(percen_cds = n_cds / total_reg_annotations) %>%
  select(id, num_clips, age_YYMMDD, gender, location, method, percen_cds, n_cds, percen_ofallclips_drawn

ads <- reg_annon %>%
  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) %>%
  mutate(percen_ads = n_ads / total_reg_annotations) %>%
  select(id, num_clips, age_YYMMDD, gender, location, method, percen_ads, n_ads, percen_ofallclips_drawn

o_child <- reg_annon %>%
  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) %>%
  mutate(percen_ods = n_ods / total_reg_annotations) %>%
  select(id, num_clips, age_YYMMDD, gender, location, method, percen_ods, n_ods, percen_ofallclips_drawn

o2 <- merge(cds, ads, all=T)
o3 <- merge(o2, o_child, all = T)
o3[is.na(o3)] <- 0 # one child doesn't have any ODS

# 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_ofallclips_drawn) # get the % of clips annotated for each id and method

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

# compute correlations
cds_cors <- cds_plot_data %>%
  group_by(location) %>%
  summarize(., paste("r=",round(cor.test(complete, random)$estimate,2),",","p=",round(cor.test(complete

#reg_tbl <- o3 %>%
#  group_by(method, location) %>%
#  summarize(avg=round(mean(percen_cds),2),
#            sd=round(sd(percen_cds),2)) %>%
#  mutate(stats=paste(avg,"(",sd,")")) %>%
#  select(-avg, -sd) %>%
#  spread(key='method', value = "stats")

# calculate relative errors
cds_rel_error <- o3 %>%
  group_by(id) %>%
  #summarize(avg=mean(percen_cds)) %>%
  select(id,method,percen_cds,location) %>%
  spread(key='method', value='percen_cds') %>%
  mutate(relative_error = round(((abs(random - complete) / complete)*100),2)) %>%
  #mutate(avg_rel_error = round(mean(relative_error),2),
  #       sd_rel_error = round(sd(relative_error),2),
  #       rel_error_stats=paste(avg_rel_error,"(",sd_rel_error,")")) %>%
  distinct(relative_error, .keep_all = T)




# add correlations to table - will make pretty below
final_reg_tbl <- merge(cds_rel_error, cds_cors, by='location')

final_reg_tbl$location <-
  plyr::mapvalues(final_reg_tbl$location,
                  from = c("Bolivia", "US"),
                  to =c("Quechua-Spanish", "Spanish-English"))

final_reg_tbl2 <- final_reg_tbl %>%
  mutate(random = round(random,2),
         complete = round(complete,2)) %>%
  mutate(corpus_id = paste(location,"(",id,")")) %>%
  select(-location, -id) %>%
  relocate(corpus_id, random, complete)



knitr::kable(final_reg_tbl2, caption = 'Child-directed speech estimates by child and annotation method.
             booktabs=T,
             row.names = FALSE,
             col.names = c("Corpus (ID)", "Random", "All-day", "Relative error", "Within-corpus correla
  #column_spec(2, width = "4cm") %>% # force column headers onto two rows
  #column_spec(3, width = "3cm") %>%
  column_spec(5, width = "5cm") %>%
  kable_styling() %>%
  add_header_above(c(" " = 1, "Annotation Method" = 2, " " = 2)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 2: (#tab:cds proportion stats)Child-directed speech estimates by child and annotation method.

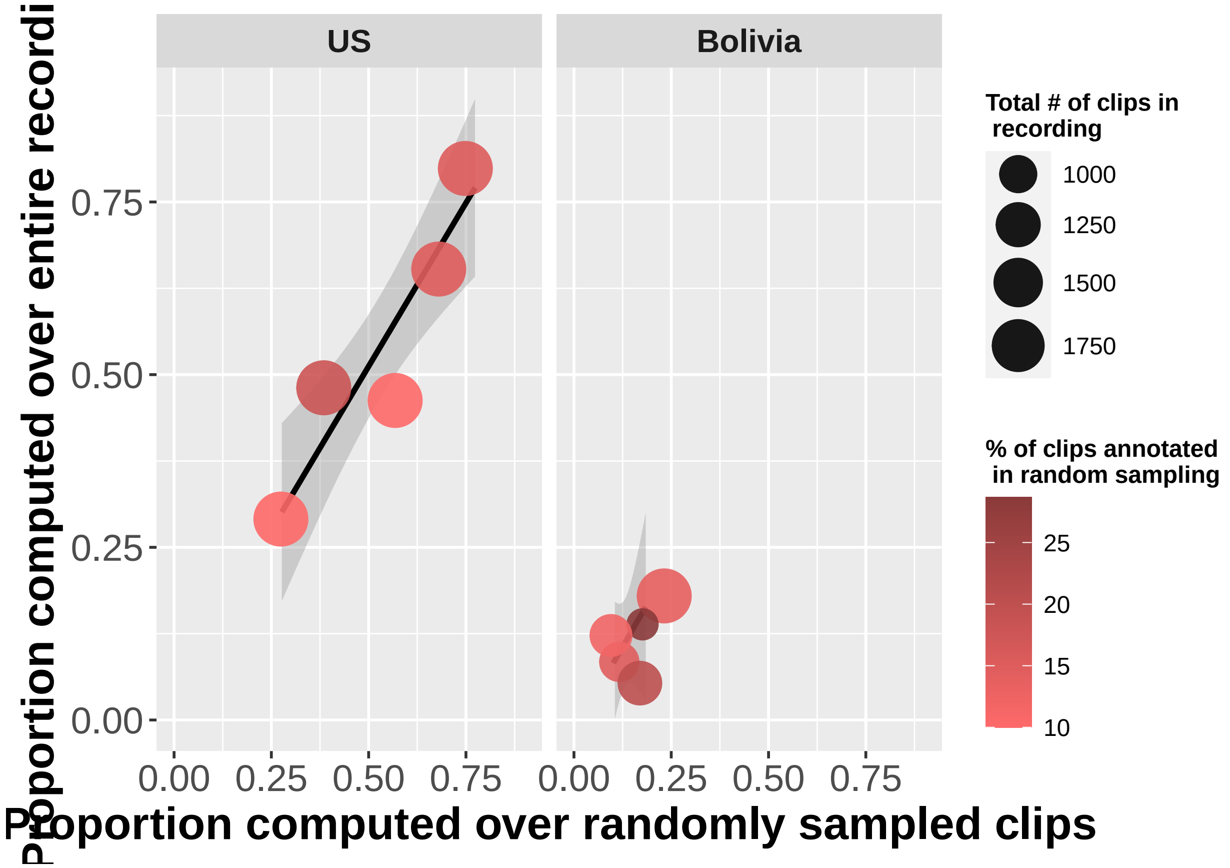
	Annotation Method			
Corpus (ID)	Random	All-day	Relative error	Within-corpus correlation between estimates
Quechua-Spanish (1032)	0.18	0.17	5.95	r= 0.63 , p= 0.26
Quechua-Spanish (1060)	0.12	0.07	55.09	r= 0.63 , p= 0.26
Quechua-Spanish (1075)	0.12	0.14	12.50	r= 0.63 , p= 0.26
Quechua-Spanish (1077)	0.10	0.11	10.48	r= 0.63 , p= 0.26
Quechua-Spanish (1081)	0.13	0.05	145.09	r= 0.63 , p= 0.26
Spanish-English (179)	0.52	0.46	12.61	r= 0.97 , p= 0.01
Spanish-English (198-9mo)	0.65	0.66	0.54	r= 0.97 , p= 0.01
Spanish-English (199)	0.28	0.30	6.24	r= 0.97 , p= 0.01
Spanish-English (261-8mo)	0.77	0.79	2.32	r= 0.97 , p= 0.01
Spanish-English (267-12mo)	0.40	0.47	15.23	r= 0.97 , p= 0.01

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

# compute correlations
ads_cors <- ads_plot_data %>%
  group_by(location) %>%
  summarize(., paste("r=",round(cor.test(complete, random)$estimate,2),",","p=",round(cor.test(complete

reg_tbl <- o3 %>%
  group_by(method, location) %>%
  summarize(avg=round(mean(percen_ads),2),
            sd=round(sd(percen_ads),2)) %>%
  mutate(stats=paste(avg,"(",sd,")")) %>%
  select(-avg, -sd) %>%
  spread(key='method', value = "stats")

# calculate relative errors
ads_rel_error <- o3 %>%
  group_by(method, location, id) %>%
  summarize(avg=mean(percen_ads)) %>%
  spread(key='method', value='avg') %>%
  group_by(id) %>%
  mutate(relative_error = ((abs(random - complete) / complete)*100)) %>%
  ungroup() %>%
  group_by(location) %>%
  mutate(avg_rel_error = round(mean(relative_error),2),
         sd_rel_error = round(sd(relative_error),2),
         rel_error_stats=paste(avg_rel_error,"(",sd_rel_error,")")) %>%
  distinct(rel_error_stats)

# add correlations to table - will make pretty below
final_reg_tbl <- merge(reg_tbl, ads_cors, by='location')
final_reg_tbl2 <- merge(final_reg_tbl, ads_rel_error, by='location')
final_reg_tbl$location <-
  plyr::mapvalues(final_reg_tbl$location,
                  from = c("Bolivia", "US"),
                  to =c("Quechua-Spanish", "Spanish-English"))

knitr::kable(final_reg_tbl2, caption = 'Average adult-directed speech estimates by corpus and annotatio
             booktabs=T,
             row.names = FALSE,
             col.names = c("Corpus", "Random", "All-day", "Correlation between estimates", "Average rela
  kable_styling() %>%
  add_header_above(c(" " = 1, "Annotation Method" = 2, " " = 2)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 3: (#tab:ads proportion stats)Average adult-directed speech estimates by corpus and annotation method.

	Annotation Method			
Corpus	Random	All-day	Correlation between estimates	Average relative error (SD)
Bolivia	0.45 (0.13)	0.4 (0.09)	r= 0.84 , p= 0.07	11.56 (12.11)
US	0.27 (0.17)	0.27 (0.2)	r= 0.96 , p= 0.01	16.75 (12.45)

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

cds_plot <- ggplot(cds_plot_data, aes(random, complete)) +
  geom_smooth(method = "lm", color="black") +
  geom_jitter(aes(size=num_clips,color=round(percen_ofallclips_drawn,2)),alpha=.9,position = position_j
  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 clips \n in U.S. and Bolivian corpora') +
 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"))
cds_plot
```



```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/cds_plot.jpeg", height = 500, width = 50
cds_plot
dev.off()
```

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

0.0.3 Part III: language across random and questionnaire methods

```
# enter questionnaire estimates
ques <- data.frame("id"=c("179", "198-9mo", "199", "261-8mo", "267-12mo"),
                   "ques_est"=c(".71", ".57", ".94", ".69", ".87"))

ques_tbl <- plot_data %>%
  filter(location=='US') %>%
  merge(., ques, by='id') %>%
  distinct_at(., vars(method, id), .keep_all = T) %>%
  select(-percen_ofallclips_drawn, -percen_mxd, -percen_que, -speech_clips, -total, -gender, -location,
  mutate(percen_span = round(percen_span,2)) %>%
  spread("method", "percen_span") %>%
  relocate(id, random, complete, ques_est)
```

```
# compute correlations
ques_random_cors <- ques_tbl %>%
  mutate(ques_est = as.numeric(ques_est)) %>%
  summarize(., paste("r=",round(cor.test(ques_est, random)$estimate,2),",","p=",round(cor.test(ques_est
ques_complete_cors <- ques_tbl %>%
  mutate(ques_est = as.numeric(ques_est)) %>%
  summarize(., paste("r=",round(cor.test(ques_est, complete)$estimate,2),",","p=",round(cor.test(ques_e

# create table
knitr::kable(ques_tbl, caption = 'Spanish language estimates in U.S. corpus, by child and estimation me
             booktabs=T,
             row.names = FALSE,
             col.names = c("Child ID", "Random", "All-day", "Parental Questionnaire")) %>%
  kable_styling() %>%
  add_header_above(c(" " = 1, "From daylong recording" = 2, " " = 1)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 4: (#tab:make table for questionnaire method)Spanish language estimates in U.S. corpus, by child and estimation method.

	From daylong recording		
Child ID	Random	All-day	Parental Questionnaire
179	0.57	0.57	.71
198-9mo	0.87	0.78	.57
199	0.76	0.70	.94
261-8mo	0.69	0.65	.69
267-12mo	0.92	0.92	.87

```
# we also want to know what the results are for the combination of CDS*Spanish, not just Spanish
reg_annon <- data_annon %>%
  filter(addressee=='Adult2TargetChild' | addressee=='Otherchild2TargetChild') %>% # only CDS clips
  group_by(id, method) %>%
  distinct_at(., vars(file_name, addressee), .keep_all = T) %>% # don't record multiple speakers speaki
  mutate(total_cds_annotations = NROW(file_name))#

span_cds_tbl <- reg_annon %>%
  group_by(id, method) %>%
  filter(addressee=='Adult2TargetChild' | addressee=='Otherchild2TargetChild' & location=='US') %>% # o
  merge(., ques, by='id') %>%
  filter(language=='Spanish') %>% # only Spanish clips
  group_by(method) %>%
  distinct(file_name, .keep_all = T) %>%
  group_by(id, method) %>%
  mutate(n_span_cds = n()) %>% # # of CDS clips where Spanish was spoken
  distinct_at(., vars(id, method), .keep_all = T) %>%
  mutate(percen_span_cds = round(n_span_cds / total_cds_annotations,2)) %>%
  select(method, percen_span_cds, id, ques_est) %>%
  spread("method", "percen_span_cds") %>%
  relocate(id, random, complete, ques_est)
```

```
# compute correlations
cor.test(as.numeric(span_cds_tbl$ques_est), span_cds_tbl$complete)
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.numeric(span_cds_tbl$ques_est) and span_cds_tbl$complete
## t = 1.022, df = 3, p-value = 0.382
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6781348  0.9600192
## sample estimates:
##       cor
## 0.5081637
```

```
cor.test(as.numeric(span_cds_tbl$ques_est), span_cds_tbl$random)
```

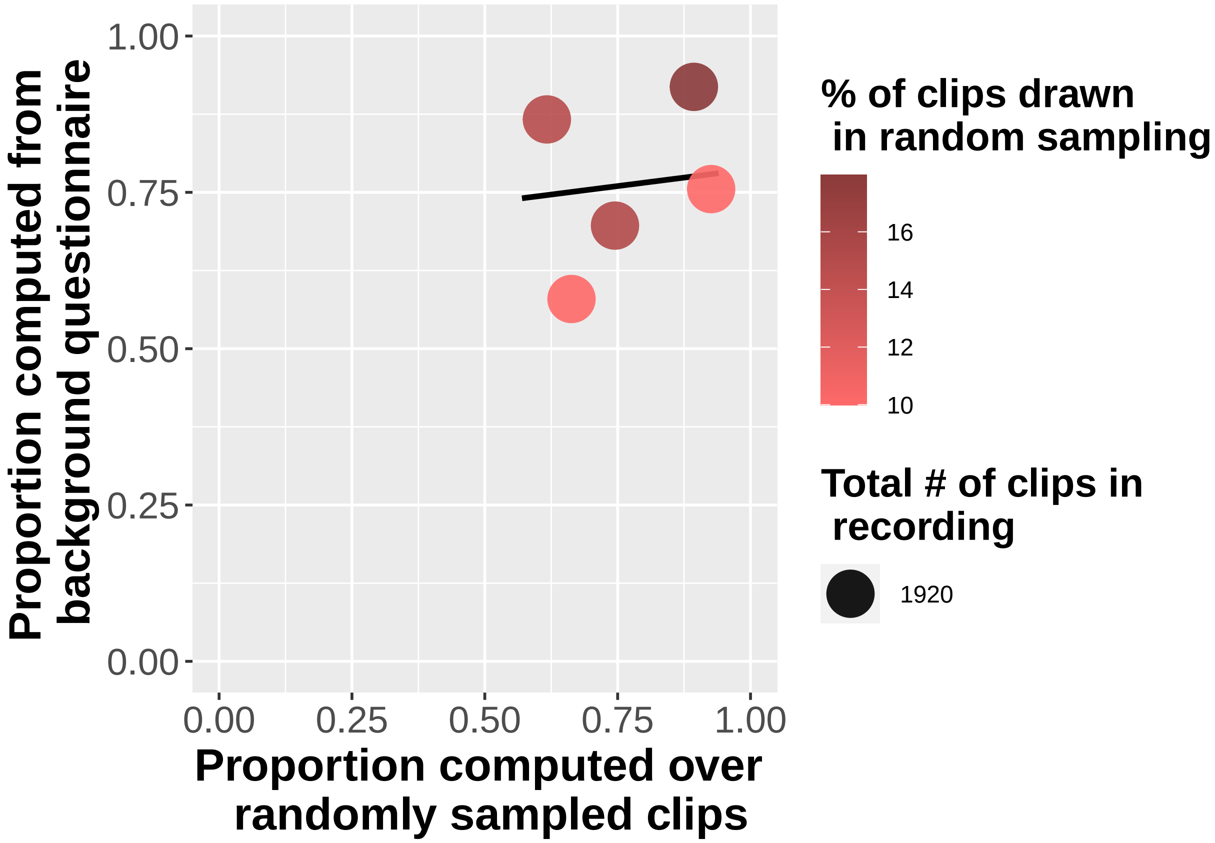
```
##
##  Pearson's product-moment correlation
##
## data:  as.numeric(span_cds_tbl$ques_est) and span_cds_tbl$random
## t = 0.12188, df = 3, p-value = 0.9107
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8656838  0.8969149
## sample estimates:
##       cor
## 0.0701952
```

```
# create table
knitr::kable(span_cds_tbl, caption = 'Spanish language in child-directed speech \n estimates in U.S. co
             booktabs=T,
             row.names = FALSE,
             col.names = c("Child ID", "Random", "All-day", "Parental Questionnaire")) %>%
  kable_styling() %>%
  add_header_above(c(" " = 1, "From daylong recording" = 2, " " = 1)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 5: (#tab:make table for questionnaire method)Spanish language in child-directed speech estimates in U.S. corpus, by child and estimation method.

	From daylong recording		
Child ID	Random	All-day	Parental Questionnaire
179	0.53	0.52	.71
198-9mo	0.78	0.64	.57
199	0.64	0.66	.94
261-8mo	0.55	0.48	.69
267-12mo	0.82	0.87	.87

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

ques_plot <- plot_data %>%
  filter(location=='US') %>%
  merge(., ques, by='id') %>%
  distinct_at(., vars(method, id), .keep_all = T) %>%
  select(-percen_que, -percen_ofallclips_drawn, -percen_mxd, -speech_clips, -total) %>%
  spread("method", "percen_span") %>%
  select(-complete) %>%
  merge(., per_ann, by='id') %>%
  distinct(id, .keep_all = T) %>%
ggplot(., aes(as.numeric(ques_est), random)) +
  geom_smooth(method = "lm", color="black", se=FALSE) +
  geom_jitter(aes(size=num_clips,color=round(percen_ofallclips_drawn,2)),alpha=.9,position = position_ji
  scale_size_continuous(range = c(5, 9)) +
  scale_colour_gradient(low='indianred1', high = 'indianred4') +
  ylab("Proportion computed from \n background questionnaire") +
  xlab("Proportion computed over \n randomly sampled clips") +
  ylim(0,1) +
  xlim(0,1)+
  labs(col='% of clips drawn \n in random sampling') +
       #title = 'Proportion of Spanish clips \n in U.S. corpus: random sampling and background question
 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"))
ques_plot
```



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

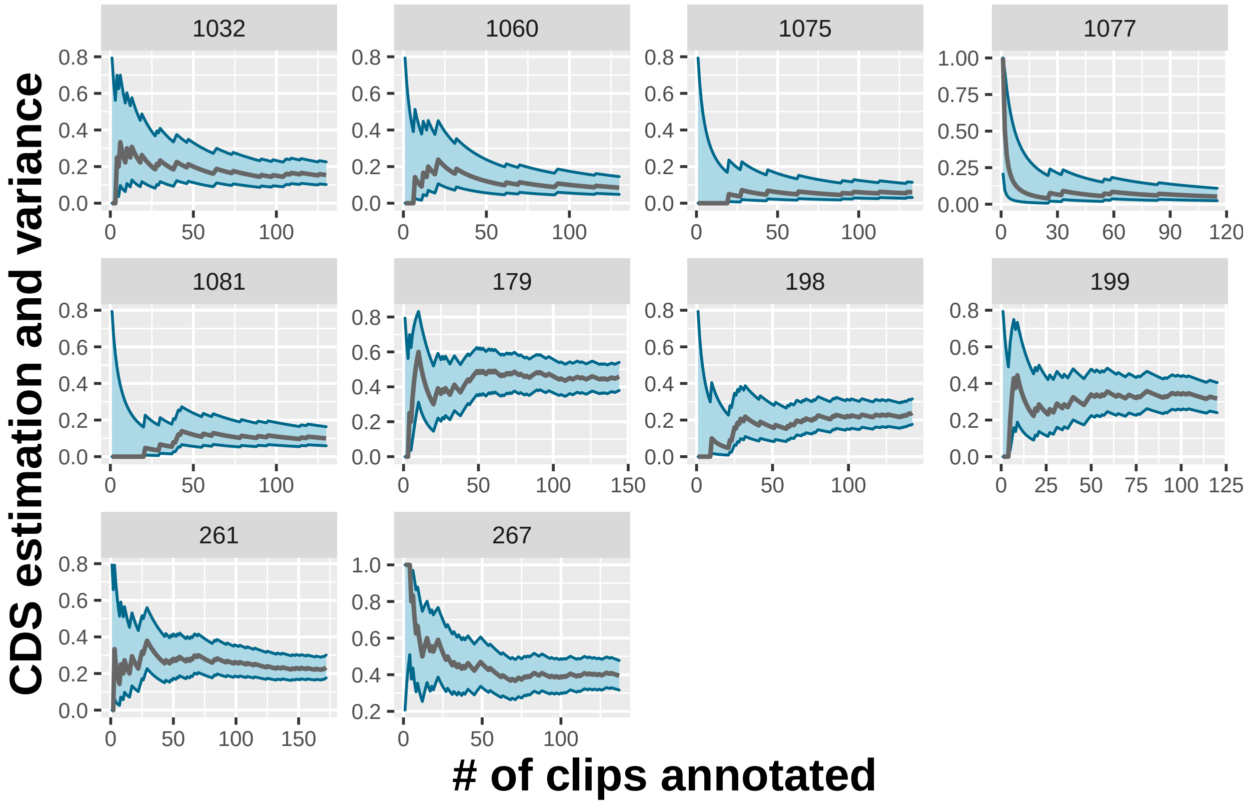
```
## pdf
##   2
```

0.0.4 Part I: Running variance

```r
random$id <- plyr::mapvalues(random$id,
                from=c("198-9mo", "261-8mo", "267-12mo"),
                to=c("198", "261", "267"))

# only doing for CDS first - filter for other languages for language
cds_var <- random %>%
  group_by(id) %>%
  mutate(total=n()) %>% # total clips drawn & listened to
  filter(researcher_present!='1' & sleeping!='1' & percents_voc>0) %>% # criteria for draw, but don't l
  distinct(file_name, .keep_all = T) %>%
  mutate(annotation_num = as.numeric(1:n())) %>% # total clips annotated for lang/reg/childvoc/media, n
  select(-Otherchild2OtherChild, -Otherchild2adults, -Otherchild2unsure, -Adult2OtherChild, -Adult2Othe
  gather("addressee", "language", Adult2TargetChild, Otherchild2TargetChild) %>%
  distinct_at(., vars(file_name, timestamp_HHMMSS), .keep_all = T) %>% # CDS only gets counted 1x/clip;
  select(-addressee)
```

```
cds_var$cds_cts <- plyr::mapvalues(cds_var$language,
                from=c("Categorize language to target child", "English/Quechua", "Mixed", "Spanish", "Ur
                to=c("0", "1", "1", "1", "1")) # where 'cat lang...' are ADS or OCDS
cds_var$cds_cts <- as.numeric(cds_var$cds_cts)
cds_var$total <- as.numeric(cds_var$total)

cds_rolling <- cds_var %>%
  group_by(id) %>%
  mutate(cds_running_cts = as.numeric(cumsum(cds_cts))) %>%
  mutate(roll_prop_cds = cds_running_cts / annotation_num,
         roll_mean_cds = rollmean(roll_prop_cds, k=10, fill = NA),
         roll_sd_cds = rollapply(roll_prop_cds, width=10, FUN=sd, fill=NA))

# running binomial confidence interval (wilson)
cds_rolling2 <- cds_rolling %>%
  group_by(id, annotation_num) %>% # group by id and sample size
  summarize(cis = binom.confint(cds_running_cts, annotation_num, methods = 'wilson', conf.level = .95))
  merge(., cds_rolling, by = c('id', 'annotation_num'))

# for models, compute binomial confidence interval in 5-clip batches
#cds_batches <- cds_rolling %>%
#  group_by(id) %>%
#  mutate(five_clip_batch = as.integer(gl(n(), 5, n())) * 5,
#         five_clip_batch = replace(five_clip_batch,  ave(five_clip_batch, five_clip_batch, FUN = lengt
#  ungroup %>%
#  fill(five_clip_batch) #%>%

#cds_batches2 <- cds_batches %>%
#  group_by(id, five_clip_batch) %>%
#  summarize(five_cis = binom.confint(cds_running_cts, 5, methods = 'wilson', conf.level = .95)) %>%
#  merge(., cds_batches, by = c('id', 'five_clip_batch'))
```

```
cds_var_plot <- cds_rolling2 %>%
#filter(roll_sd_cds!='NA') %>% # remove rows where variance wasn't estimated
mutate(mean_ci = cis$mean,
       upper_ci = cis$upper,
       lower_ci = cis$lower) %>%
ggplot(., aes(annotation_num, roll_prop_cds)) +
  #geom_line(aes(y=rollapply(roll_prop_cds, 10, FUN=sd, fill=NA))) +
  geom_ribbon(aes(ymax=upper_ci, ymin=lower_ci), fill='lightblue', color='deepskyblue4') +
    geom_line(aes(y=mean_ci), color='gray40', size=.8) +
  xlab("# of clips annotated") +
  ylab("CDS estimation and variance") +
  facet_wrap(~id, scales = "free") +
  #title = 'Variance in child-directed estimation as a function of clips annotated') +
 theme(title = element_text(size=12),
   axis.text=element_text(size=8),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=15)) +
  labs(caption = "Number of clips annotated refers to those annotated for language, speech register, chi
cds_var_plot
```



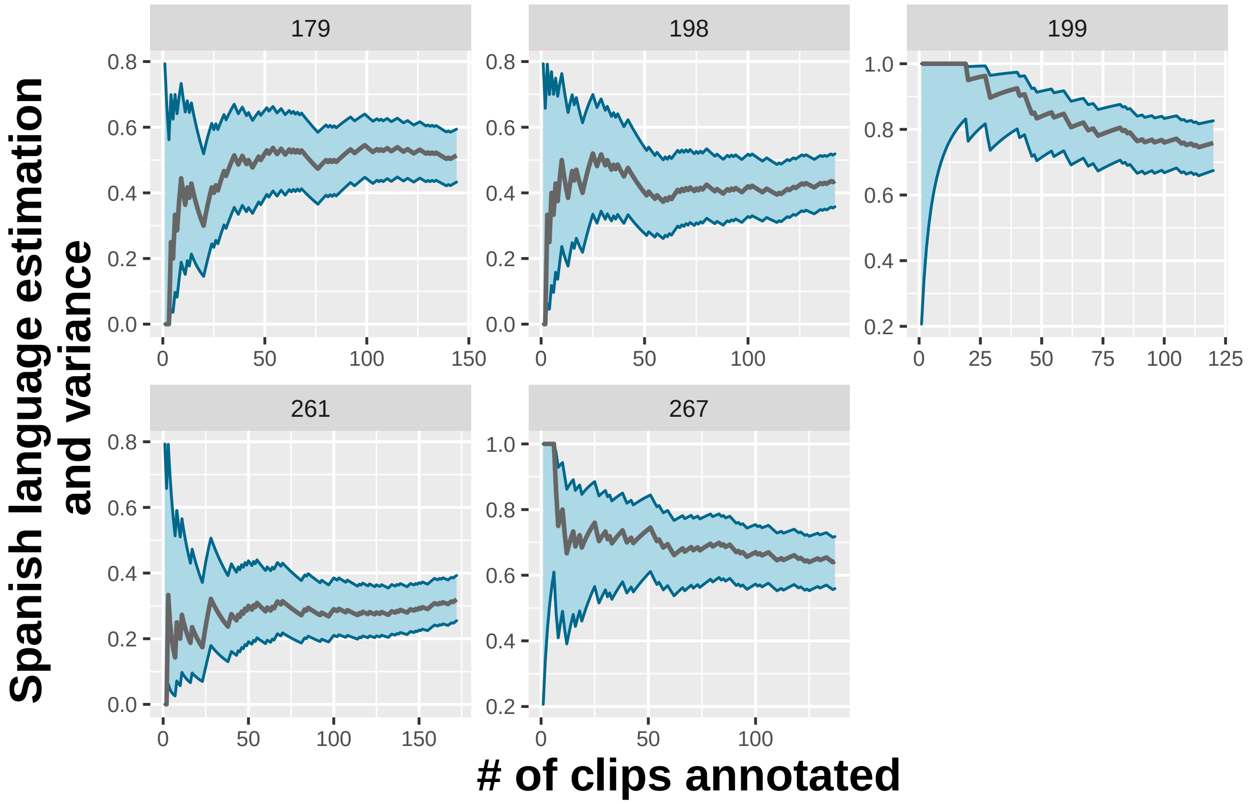
clips annotated refers to those annotated for language, speech register, child vocalizations, and/or media.

```r
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/cds_CI_var_plot.jpeg", height = 450, wi
cds_var_plot
dev.off()
```

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

```r
# now calculate rolling variances for US (Spanish)
span_var <- random %>%
  group_by(id) %>%
  mutate(total=n()) %>% # total clips drawn
  filter(researcher_present!='1' & sleeping!='1' & percents_voc>0) %>% # criteria for draw, but don't l
  distinct(file_name, .keep_all = T) %>%
  mutate(annotation_num = as.numeric(1:n())) %>% # total clips annotated for lang/reg/childvoc/media, n
  gather("addressee", "language", Adult2TargetChild, Otherchild2TargetChild, Otherchild2OtherChild, Othe
         Otherchild2unsure, Adult2OtherChild, Adult2Others, Adult2unsure) %>%
  distinct_at(., vars(file_name, timestamp_HHMMSS, language), .keep_all = T) %>% # each unique 'languag
  select(-addressee)

span_var$span_cts <- plyr::mapvalues(span_var$language,
                from=c("Categorize language to adults", "Categorize language to other adults",
                       "Categorize language to other child(ren)",
                       "Categorize language to someone unknown",
                       "Categorize language to target child",
                       "Unsure",
                       "None", "English/Quechua", "Mixed", "Spanish"),
                to=c("0","0","0","0","0","0","0","0", "0", "1"))

span_var2 <- span_var %>%
  distinct_at(., vars(file_name, span_cts), .keep_all = T) %>%
  mutate(span_cts = as.numeric(span_cts),
         total = as.numeric(total)) %>%
  group_by(file_name, timestamp_HHMMSS) %>%
  add_count() %>%
  filter(!(n==2 & span_cts==0)) %>% # when spanish and another category are marked, only count spanish
  group_by(file_name) %>%
  distinct_at(., vars(annotation_num, language), .keep_all = T) %>% # remove 1 count of spanish (it get
  select(-n)

span_rolling <- span_var2 %>%
  filter(location=='US') %>%
  group_by(id) %>%
  arrange(annotation_num) %>%
  mutate(span_running_cts = as.numeric(cumsum(span_cts))) %>%
  mutate(roll_prop_span = span_running_cts / annotation_num,
         roll_mean_span = rollmean(roll_prop_span, k=10, fill = NA),
         roll_sd_span = rollapply(roll_prop_span, width=10, FUN=sd, fill=NA))

# running binomial confidence interval (wilson)
span_rolling2 <- span_rolling %>%
  group_by(id, annotation_num) %>% # group by id and sample size
  arrange(annotation_num) %>%
  summarize(cis = binom.confint(span_running_cts, annotation_num, methods = 'wilson', conf.level = .95))
  merge(., span_rolling, by = c('id', 'annotation_num'))
```

```
span_var_plot <- span_rolling2 %>%
#filter(roll_sd_span!='NA') %>% # remove rows where variance wasn't estimated
    mutate(mean_ci = cis$mean,
         upper_ci = cis$upper,
         lower_ci = cis$lower) %>%
ggplot(., aes(annotation_num, roll_prop_span)) +
  geom_ribbon(aes(ymax=upper_ci, ymin=lower_ci), fill='lightblue', color='deepskyblue4') +
  geom_line(aes(y=mean_ci), color='gray40', size=.8) +
  xlab("# of clips annotated") +
  ylab("Spanish language estimation \n and variance") +
  facet_wrap(~id, scales = "free") +
  #title = 'Variance in Spanish language estimation as a function of clips drawn: US corpus') +
 theme(title = element_text(size=12),
   axis.text=element_text(size=8),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=15))  +
  labs(caption = "Number of clips annotated refers to those annotated for language, speech register, ch
span_var_plot
```



clips annotated refers to those annotated for language, speech register, child vocalizations, and/or media.

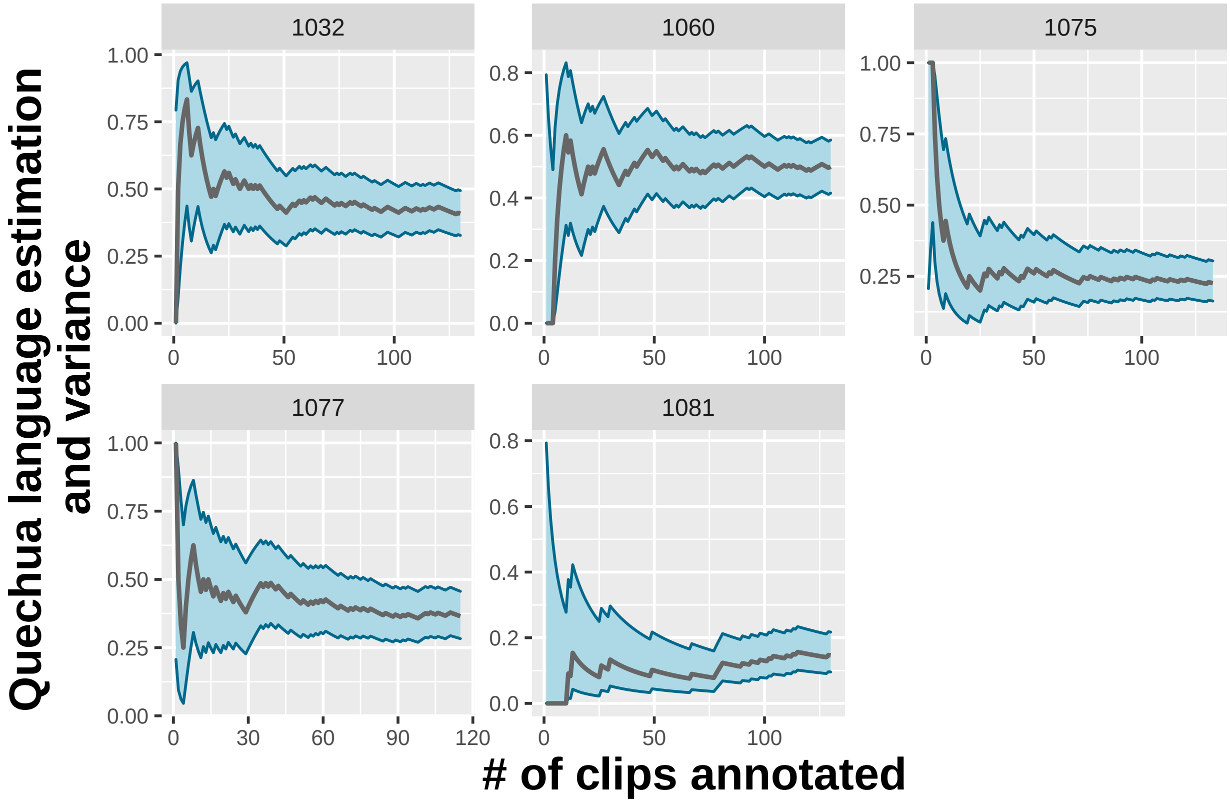
```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/span_CI_var_plot.jpeg", height = 450, wi
span_var_plot
dev.off()
```

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

```
que_var <- random %>%
  group_by(id) %>%
  mutate(total=n()) %>% # total clips drawn
  filter(researcher_present!='1' & sleeping!='1' & percents_voc>0) %>% # criteria for draw, but don't l
  distinct(file_name, .keep_all = T) %>%
  mutate(annotation_num = as.numeric(1:n())) %>% # total clips annotated for lang/reg/childvoc/media, n
  gather("addressee", "language", Adult2TargetChild, Otherchild2TargetChild, Otherchild2OtherChild, Othe
         Otherchild2unsure, Adult2OtherChild, Adult2Others, Adult2unsure) %>%
  distinct_at(., vars(file_name, timestamp_HHMMSS, language), .keep_all = T) %>% # each unique 'languag
  select(-addressee)

que_var$que_cts <- plyr::mapvalues(que_var$language,
                                   from=c("Categorize language to adults", "Categorize language to oth
                                          "Categorize language to other child(ren)",
                                          "Categorize language to someone unknown",
                                          "Categorize language to target child",
                                          "Unsure",
                                          "None", "English/Quechua", "Mixed", "Spanish"),
                                   to=c("0","0","0","0","0","0","0","1", "0", "0"))
```

```
que_var2 <- que_var %>%
  distinct_at(., vars(file_name, que_cts), .keep_all = T) %>%
  mutate(que_cts = as.numeric(que_cts),
         total = as.numeric(total)) %>%
  group_by(file_name, timestamp_HHMMSS) %>%
  add_count() %>%
  filter(!(n==2 & que_cts==0)) %>% # when quechua and another category are marked, only count quechua
  group_by(file_name) %>%
  distinct_at(., vars(annotation_num, language), .keep_all = T) %>% # remove 1 count of quechua (it get
  select(-n)

que_rolling <- que_var2 %>%
  filter(location=='Bolivia') %>%
  group_by(id) %>%
  arrange(annotation_num) %>%
  mutate(que_running_cts = as.numeric(cumsum(que_cts))) %>%
  mutate(roll_prop_que = que_running_cts / annotation_num,
         roll_mean_que = rollmean(roll_prop_que, k=10, fill = NA),
         roll_sd_que = rollapply(roll_prop_que, width=10, FUN=sd, fill=NA))

# running binomial confidence interval (wilson)
que_rolling2 <- que_rolling %>%
  group_by(id, annotation_num) %>% # group by id and sample size
  arrange(annotation_num) %>%
  summarize(cis = binom.confint(que_running_cts, annotation_num, methods = 'wilson', conf.level = .95))
  merge(., que_rolling, by = c('id', 'annotation_num'))
```

```
que_var_plot <- que_rolling2 %>%
#filter(roll_sd_que!='NA') %>% # remove rows where variance wasn't estimated
    mutate(mean_ci = cis$mean,
         upper_ci = cis$upper,
         lower_ci = cis$lower) %>%
ggplot(., aes(annotation_num, roll_prop_que)) +
  geom_ribbon(aes(ymax=upper_ci, ymin=lower_ci), fill='lightblue', color='deepskyblue4') +
  geom_line(aes(y=mean_ci), color='gray40', size=.8) +
  xlab("# of clips annotated") +
  ylab("Quechua language estimation \n and variance") +
  facet_wrap(~id, scales = "free") +
  #title = 'Variance in Quechua language estimation as a function of clips drawn: Bolivia corpus') +
 theme(title = element_text(size=12),
   axis.text=element_text(size=8),
      axis.title=element_text(size=17,face="bold"),
      legend.title = element_text(size=15)) +
  labs(caption = "Number of clips annotated refers to those annotated for language, speech register, ch
que_var_plot
```



clips annotated refers to those annotated for language, speech register, child vocalizations, and/or media.

```
jpeg("/Users/megcychosz/Google Drive/biling_CDS/results/figures/que_CI_var_plot.jpeg", height = 450, wi
que_var_plot
dev.off()
```

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

```
# report CI ranges at 80-clip mark and when annotation stopped, by child
que_cis_table <- que_rolling2 %>%
  group_by(id) %>%
  filter(annotation_num==80 | annotation_num==NROW(id)) %>% # get values at 80-clip mark and cut-off
  mutate(ci_range = cis$upper - cis$lower)

lang_cis_table <- span_rolling2 %>%
  group_by(id) %>%
  filter(annotation_num==80 | annotation_num==NROW(id)) %>% # get values at 80-clip mark and cut-off
  mutate(ci_range = cis$upper - cis$lower) %>%
  rbind(., que_cis_table) %>%
  select(id, annotation_num, ci_range) %>%
  mutate(ci_range = round(ci_range,2)) %>%
  mutate(timept = if_else(annotation_num==80, '80-clip_lang', 'Cut-off_lang')) %>%
  select(-annotation_num) %>%
  spread("timept", "ci_range")


final_cis_table <- cds_rolling2 %>%
  group_by(id) %>%
  filter(annotation_num==80 | annotation_num==NROW(id)) %>% # get values at 80-clip mark and cut-off
  mutate(ci_range = cis$upper - cis$lower) %>%
  select(id, annotation_num, ci_range) %>%
  mutate(ci_range = round(ci_range,2)) %>%
  mutate(timept = if_else(annotation_num==80, '80-clip', 'Cut-off')) %>%
  select(-annotation_num) %>%
  spread("timept", "ci_range") %>%
  merge(., lang_cis_table, by='id')

knitr::kable(final_cis_table, caption = 'Confidence interval range for Spanish/Quechua and child-directe
             booktabs=T,
             row.names = FALSE,
             col.names = c("Child ID", "80-clip", "Cut-off", "80-clip", "Cut-off")) %>% # "
  kable_styling() %>%
  add_header_above(c(" " = 1, "Language" = 2, "Child-directed speech" = 2)) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 6: (#tab:report CI ranges)Confidence interval range for Spanish/Quechua and child-directed speech estimation, by child, after annotating 80 clips and at annotation cut-off.

	Language		Child-directed speech	
Child ID	80-clip	Cut-off	80-clip	Cut-off
1032	0.16	0.12	0.21	0.17
1060	0.13	0.10	0.21	0.17
1075	0.10	0.08	0.18	0.14
1077	0.11	0.09	0.21	0.17
1081	0.14	0.10	0.14	0.12
179	0.21	0.16	0.21	0.16
198	0.18	0.14	0.21	0.16
199	0.20	0.16	0.17	0.15
261	0.19	0.13	0.19	0.14
267	0.21	0.16	0.20	0.16

```
# cds model
cds_model_data <- cds_rolling2 %>%
  group_by(id) %>%
  arrange(annotation_num) %>%
  mutate(halfrow = as.numeric(n()/2)) %>% # for a sanity check
  filter(row_number() > n()*.50) # get the top 10% of rows from each group

cds_model <- cds_model_data %>%
  #filter(roll_sd_cds!='NA') %>%
  filter(location=='US') %>%
  mutate(ci_range = cis$upper - cis$lower) %>%
  lmer(ci_range~annotation_num + (1|id), data = .) %>%
  summary()

# spanish model
# redo data to get the Bolivia corpus at the same time (more power for stats)
```

```
span_rolling_all <- span_var2 %>%
  group_by(id) %>%
  arrange(annotation_num) %>%
  mutate(span_running_cts = as.numeric(cumsum(span_cts))) %>%
  mutate(roll_prop_span = span_running_cts / annotation_num,
         roll_mean_span = rollmean(roll_prop_span, k=10, fill = NA),
         roll_sd_span = rollapply(roll_prop_span, width=10, FUN=sd, fill=NA))

# running binomial confidence interval (wilson)
span_rolling_all2 <- span_rolling_all %>%
  group_by(id, annotation_num) %>% # group by id and sample size
  arrange(annotation_num) %>%
  summarize(cis = binom.confint(span_running_cts, annotation_num, methods = 'wilson', conf.level = .95)
  merge(., span_rolling_all, by = c('id', 'annotation_num'))

# fit the spanish models
span_model_data <- span_rolling_all2 %>%
  group_by(id) %>%
  arrange(annotation_num) %>%
  mutate(halfrow = as.numeric(n()/2)) %>% # for a sanity check
  filter(row_number() > n()*.50)

span_model <- span_model_data %>%
  #filter(roll_sd_span!='NA') %>%
  mutate(ci_range = cis$upper - cis$lower) %>% # get the variance
  lmer(ci_range~annotation_num + (1|id), data = .) %>%
  summary()
```