SLT_paper

AC

2020-03-06 (substantive version), latest minor edits 2020-07-17

Contents

History:

- 2020-08-05 final first version

Read data in

```
# read dataset composed with python
data_all <- read.csv("final_classifications_PU_zoon.csv")

#remove the word mixed that takes up space and is unnecessary
data_all$Zoon_classif=factor(gsub("Mixed_","",as.character(data_all$Zoon_classif),fixed=T))

#relevel the factor so that it's easier to read
data_all$Zoon_classif=factor(data_all$Zoon_classif, levels=c("Canonical","Non-Canonical",
                                                    "Crying","Laughing","Junk",levels(data_all$


# create lab column with easier to read correspondance
data_all$lab<-as.character(data_all$Major_Choice)
data_all$lab[data_all$lab=="Non-canonical syllables"]<-"Non-Canonical"
data_all$lab[data_all$lab=="Canonical syllables"]<-"Canonical"
data_all$lab[data_all$lab %in% c("Don't mark","None")]<-"Junk"
data_all$lab=factor(data_all$lab,levels=levels(data_all$Zoon_classif))
#apply same factor levels as zooniverse so that we can do symmetrical confusion matrices
```

Correspondence between lab & zooniverse annotation at the level of segments

Here we look at to what extent zooniverse and lab annotations match at the level of individual segments. Each data point is one segment (one “vocalization”).

```r
table(data_all$lab)
```

```
##
##                        Canonical                    Non-Canonical
##                              258                             2532
##                           Crying                         Laughing
##                               51                               49
##                             Junk               Laughing_Canonical
##                              904                                0
##                  Laughing_Crying           Laughing_Non-Canonical
##                                0                                0
## Laughing_Non-Canonical_Crying              Non-Canonical_Crying
##                                0                                0
## Non-Canonical_Laughing_Crying
##                                0
```

```r
table(data_all$Zoon_classif)
```

```
##
##                        Canonical                    Non-Canonical
##                              226                             2535
##                           Crying                         Laughing
##                               94                              130
##                             Junk               Laughing_Canonical
##                              625                                2
##                  Laughing_Crying           Laughing_Non-Canonical
##                                3                               76
## Laughing_Non-Canonical_Crying              Non-Canonical_Crying
##                                3                               99
## Non-Canonical_Laughing_Crying
##                                1
```

```r
mycf=confusionMatrix(data_all$lab, data_all$Zoon_classif, dnn = c("Lab","Zooniverse"))

conf_tab=mycf$table

# this package uses sensitivity & specificity
#Sensitivity=recall
#Specificity=precision

mycf
```

```
## Confusion Matrix and Statistics
##
##                                  Zooniverse
## Lab                               Canonical Non-Canonical Crying Laughing Junk
##   Canonical                              93           122      3        8   22
##   Non-Canonical                          51          2057     60       49  193
##   Crying                                  0            17     13        2    4
##   Laughing                                0             5      2       26    6
##   Junk                                   82           334     16       45  400
##   Laughing_Canonical                      0             0      0        0    0
##   Laughing_Crying                         0             0      0        0    0
##   Laughing_Non-Canonical                  0             0      0        0    0
##   Laughing_Non-Canonical_Crying           0             0      0        0    0
```

```
##   Non-Canonical_Crying                   0               0       0         0     0
##   Non-Canonical_Laughing_Crying          0               0       0         0     0
##                                Zooniverse
## Lab                              Laughing_Canonical Laughing_Crying
##   Canonical                                       0               1
##   Non-Canonical                                   0               1
##   Crying                                          0               1
##   Laughing                                        1               0
##   Junk                                            1               0
##   Laughing_Canonical                              0               0
##   Laughing_Crying                                 0               0
##   Laughing_Non-Canonical                          0               0
##   Laughing_Non-Canonical_Crying                   0               0
##   Non-Canonical_Crying                            0               0
##   Non-Canonical_Laughing_Crying                   0               0
##                                Zooniverse
## Lab                              Laughing_Non-Canonical
##   Canonical                                           6
##   Non-Canonical                                      51
##   Crying                                              0
##   Laughing                                            8
##   Junk                                               11
##   Laughing_Canonical                                  0
##   Laughing_Crying                                     0
##   Laughing_Non-Canonical                              0
##   Laughing_Non-Canonical_Crying                       0
##   Non-Canonical_Crying                                0
##   Non-Canonical_Laughing_Crying                       0
##                                Zooniverse
## Lab                              Laughing_Non-Canonical_Crying
##   Canonical                                                  0
##   Non-Canonical                                              1
##   Crying                                                     1
##   Laughing                                                   0
##   Junk                                                       1
##   Laughing_Canonical                                         0
##   Laughing_Crying                                            0
##   Laughing_Non-Canonical                                     0
##   Laughing_Non-Canonical_Crying                              0
##   Non-Canonical_Crying                                       0
##   Non-Canonical_Laughing_Crying                              0
##                                Zooniverse
## Lab                              Non-Canonical_Crying
##   Canonical                                         3
##   Non-Canonical                                    69
##   Crying                                           13
##   Laughing                                          0
##   Junk                                             14
##   Laughing_Canonical                                0
##   Laughing_Crying                                   0
##   Laughing_Non-Canonical                            0
##   Laughing_Non-Canonical_Crying                     0
##   Non-Canonical_Crying                              0
##   Non-Canonical_Laughing_Crying                     0
```

```
##                                 Zooniverse
## Lab                              Non-Canonical_Laughing_Crying
##   Canonical                                                  0
##   Non-Canonical                                              0
##   Crying                                                     0
##   Laughing                                                   1
##   Junk                                                       0
##   Laughing_Canonical                                         0
##   Laughing_Crying                                            0
##   Laughing_Non-Canonical                                     0
##   Laughing_Non-Canonical_Crying                              0
##   Non-Canonical_Crying                                       0
##   Non-Canonical_Laughing_Crying                              0
##
## Overall Statistics
##
##                Accuracy : 0.6824
##                  95% CI : (0.6673, 0.6972)
##     No Information Rate : 0.6682
##     P-Value [Acc > NIR] : 0.0322
##
##                   Kappa : 0.3773
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Canonical Class: Non-Canonical Class: Crying
## Sensitivity                   0.41150               0.8114      0.138298
## Specificity                   0.95376               0.6227      0.989730
## Pos Pred Value                0.36047               0.8124      0.254902
## Neg Pred Value                0.96239               0.6212      0.978360
## Prevalence                    0.05957               0.6682      0.024776
## Detection Rate                0.02451               0.5422      0.003426
## Detection Prevalence          0.06800               0.6674      0.013442
## Balanced Accuracy             0.68263               0.7171      0.564014
##                      Class: Laughing Class: Junk Class: Laughing_Canonical
## Sensitivity                 0.200000      0.6400                 0.0000000
## Specificity                 0.993723      0.8410                 1.0000000
## Pos Pred Value              0.530612      0.4425                       NaN
## Neg Pred Value              0.972230      0.9221                 0.9994729
## Prevalence                  0.034265      0.1647                 0.0005271
## Detection Rate              0.006853      0.1054                 0.0000000
## Detection Prevalence        0.012915      0.2383                 0.0000000
## Balanced Accuracy           0.596861      0.7405                 0.5000000
##                      Class: Laughing_Crying Class: Laughing_Non-Canonical
## Sensitivity                       0.0000000                       0.00000
## Specificity                       1.0000000                       1.00000
## Pos Pred Value                          NaN                           NaN
## Neg Pred Value                    0.9992093                       0.97997
## Prevalence                        0.0007907                       0.02003
## Detection Rate                    0.0000000                       0.00000
## Detection Prevalence              0.0000000                       0.00000
## Balanced Accuracy                 0.5000000                       0.50000
```

```
##                       Class: Laughing_Non-Canonical_Crying
## Sensitivity                                       0.0000000
## Specificity                                       1.0000000
## Pos Pred Value                                          NaN
## Neg Pred Value                                    0.9992093
## Prevalence                                        0.0007907
## Detection Rate                                    0.0000000
## Detection Prevalence                              0.0000000
## Balanced Accuracy                                 0.5000000
##                       Class: Non-Canonical_Crying
## Sensitivity                                0.00000
## Specificity                                1.00000
## Pos Pred Value                                 NaN
## Neg Pred Value                             0.97391
## Prevalence                                 0.02609
## Detection Rate                             0.00000
## Detection Prevalence                       0.00000
## Balanced Accuracy                          0.50000
##                       Class: Non-Canonical_Laughing_Crying
## Sensitivity                                       0.0000000
## Specificity                                       1.0000000
## Pos Pred Value                                          NaN
## Neg Pred Value                                    0.9997364
## Prevalence                                        0.0002636
## Detection Rate                                    0.0000000
## Detection Prevalence                              0.0000000
## Balanced Accuracy                                 0.5000000
```

Precision

Precision means: If a segment was called X by zooniverse coders, what proportion of the time was it called X by lab coders?

```
pdf("precision_with_mixed.pdf",height=10,width=10)
prop_cat=data.frame(conf_tab/colSums(conf_tab)*100) #generates precision because columns
prop_cat$id=paste(prop_cat$Lab,prop_cat$Zooniverse)
colnames(prop_cat)[3]<-"pr"

data.frame(conf_tab)->stall
stall$id=paste(stall$Lab,stall$Zooniverse)
stall=merge(stall,prop_cat[c("id","pr")])



ggplot(data = stall, mapping = aes(y = Lab, x=Zooniverse)) +
 geom_tile(aes(fill= rescale(pr)), colour = "white") +
  geom_text(aes(label = paste(round(pr),"%")), vjust = -1,size=2) +
  geom_text(aes(label = Freq), vjust = 1,size=2) +
  scale_fill_gradient(low = "white", high = "red", name = "Percentage") +
     theme(legend.position = "none") +
  xlab("Zooniverse") + ylab("Lab") +
  ggtitle("Precision")+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
dev.off()
```

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

Recall

Recall means: If a segment was called X by lab coders, what proportion of the time was it called X by zooniverse coders?

```
pdf("recall_with_mixed.pdf",height=10,width=10)
prop_cat=data.frame(conf_tab/rowSums(conf_tab)*100)  #generates recall because rows
prop_cat$id=paste(prop_cat$Lab,prop_cat$Zooniverse)
colnames(prop_cat)[3]<-"rec"

data.frame(conf_tab)->stall
stall$id=paste(stall$Lab,stall$Zooniverse)
stall=merge(stall,prop_cat[c("id","rec")])

ggplot(data = stall, mapping = aes(y = Lab, x=Zooniverse)) +
 geom_tile(aes(fill= rescale(rec)), colour = "white") +
  geom_text(aes(label = paste(round(rec),"%")), vjust = -1,size=2) +
  geom_text(aes(label = Freq), vjust = 1,size=2) +
  scale_fill_gradient(low = "white", high = "red", name = "Percentage") +
     theme(legend.position = "none") +
  xlab("Zooniverse") + ylab("Lab") +
  ggtitle("Recall")+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))


dev.off()
```

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

Child level descriptors

Although there may be errors at the level of the segment, what we really care about is whether Zooniverse annotations give a reliable image of the child's individual development. This is what we look at in this section.

```
#given results above, we map the mixed
data_all$Zoon_classif[data_all$Zoon_classif=="Laughing_Canonical"]<-"Canonical"
data_all$Zoon_classif[data_all$Zoon_classif=="Laughing_Non-Canonical"]<-"Non-Canonical"
data_all$Zoon_classif[data_all$Zoon_classif=="Laughing_Non-Canonical_Crying"]<-"Non-Canonical"
data_all$Zoon_classif[data_all$Zoon_classif=="Laughing_Crying"]<-"Crying"
data_all$Zoon_classif[data_all$Zoon_classif=="Non-Canonical_Crying"]<-"Non-Canonical"
data_all$Zoon_classif[data_all$Zoon_classif=="Non-Canonical_Laughing_Crying"]<-"Non-Canonical"

#and reset the factors for cleanliness
data_all$Zoon_classif=factor(data_all$Zoon_classif)
data_all$lab=factor(data_all$lab)

#get the ns by child, then calculate the linguistic ratio & canonical ratio, separately for zooniverse &
ztab=table(data_all$ChildID,data_all$Zoon_classif)
z_lr=rowSums(ztab[,c("Canonical","Non-Canonical")])/rowSums(ztab[,-which(colnames(ztab) %in% c("Junk"))]
z_cr=ztab[,c("Canonical")]/rowSums(ztab[,c("Canonical","Non-Canonical")])
```

```
ltab=table(data_all$ChildID,data_all$lab)
l_lr=rowSums(ltab[,c("Canonical","Non-Canonical")])/rowSums(ltab[,-which(colnames(ztab) %in% c("Junk"))
l_cr=ltab[,c("Canonical")]/rowSums(ltab[,c("Canonical","Non-Canonical")])

#put all the ratios together
if(sum(rownames(ztab)==rownames(ltab))==dim(ztab)[1]) ratios=cbind(rownames(ztab),z_lr,z_cr,l_lr,l_cr)
colnames(ratios)[1]<-"ChildID"

#add age
ages=aggregate(data_all$Age,by=list(data_all$ChildID),mean) #this is a weird way of adding ages, since

merge(ratios,ages,by.x="ChildID",by.y="Group.1")->ratios
colnames(ratios)[dim(ratios)[2]]<-"Age"

#cbinding results in text, so we numerize the ratios
for(thisvar in c("z_lr","z_cr","l_lr","l_cr")) ratios[,thisvar]=as.numeric(as.character(ratios[,thisvar]
summary(ratios)
```

```
##     ChildID       z_lr             z_cr              l_lr
##  1111_1 :1   Min.   :0.7625   Min.   :0.02473   Min.   :0.8219
##  1151_1 :1   1st Qu.:0.8976   1st Qu.:0.03569   1st Qu.:0.9387
##  1801_1 :1   Median :0.9303   Median :0.06496   Median :0.9659
##  2881_1 :1   Mean   :0.9120   Mean   :0.09207   Mean   :0.9523
##  3021_1 :1   3rd Qu.:0.9535   3rd Qu.:0.12083   3rd Qu.:0.9833
##  3041_1 :1   Max.   :0.9678   Max.   :0.23267   Max.   :1.0000
##  (Other):4
##       l_cr              Age
##  Min.   :0.01429   Min.   :11.83
##  1st Qu.:0.06697   1st Qu.:23.11
##  Median :0.07990   Median :43.78
##  Mean   :0.10326   Mean   :35.49
##  3rd Qu.:0.11944   3rd Qu.:46.27
##  Max.   :0.23944   Max.   :53.26
##
```

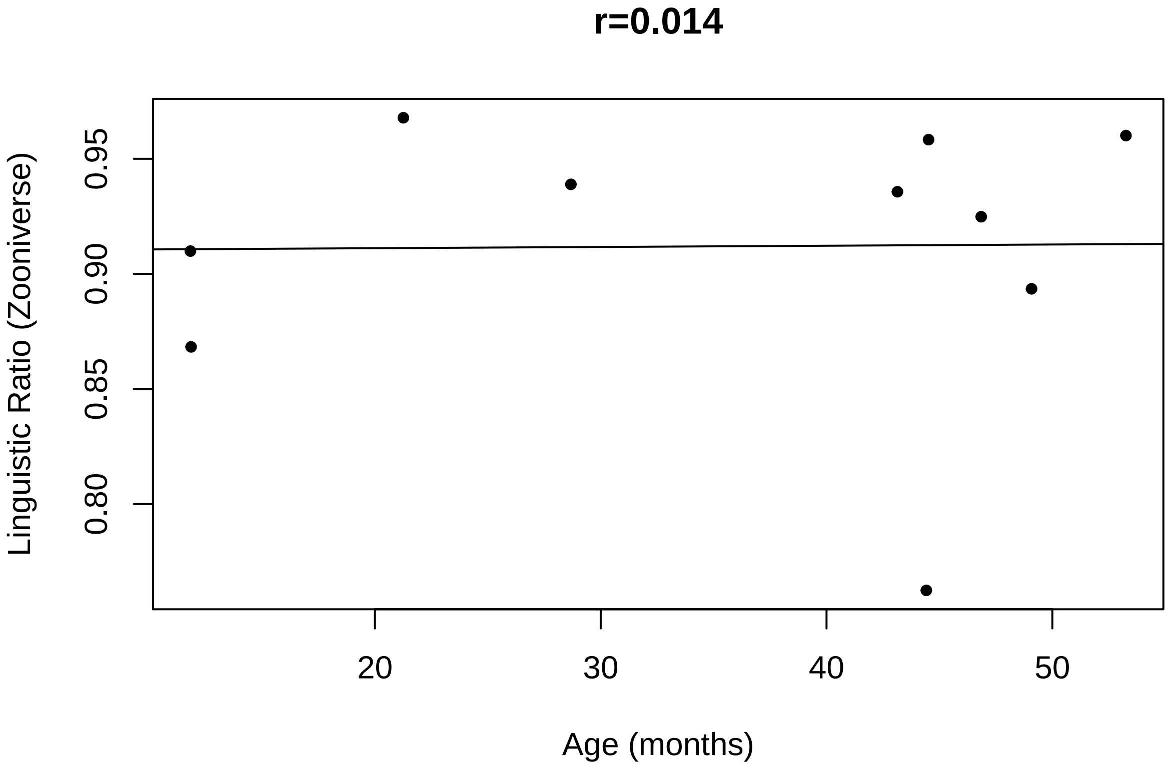
We first look generally at two measures that have been found to relate to age:

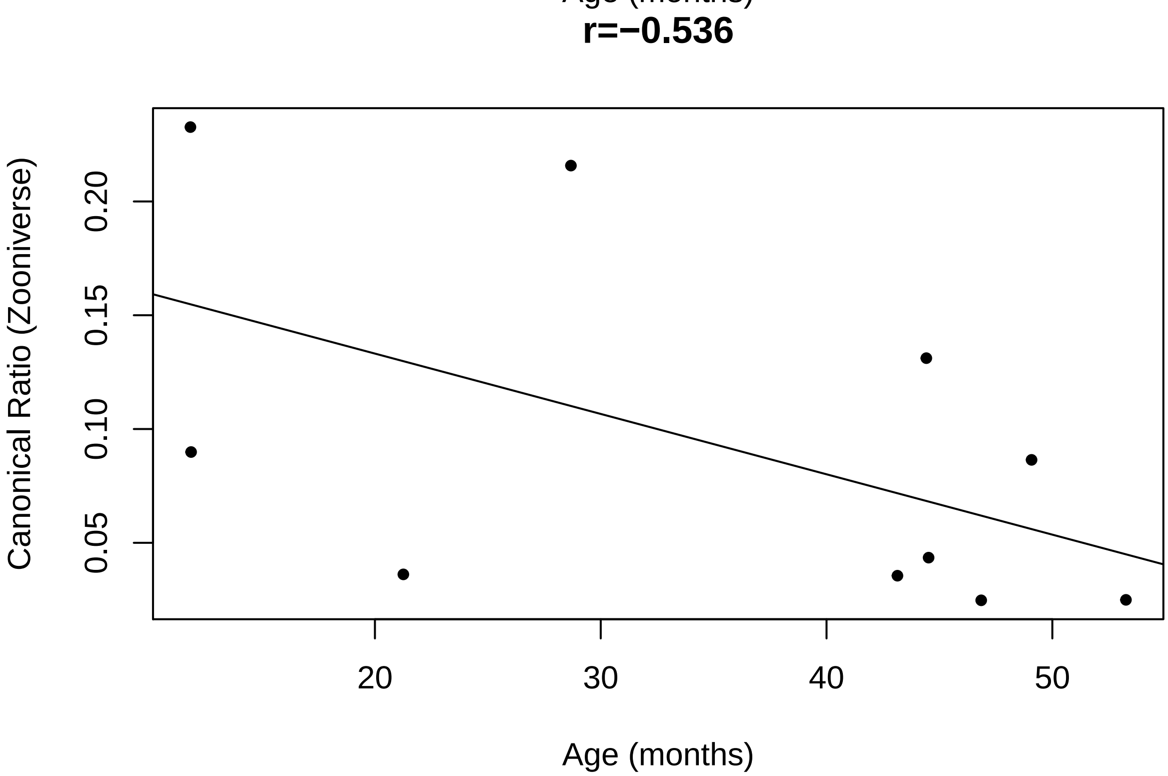
- linguistic ratio = ("Canonical"+"Non-Canonical")/"All vocalizations" (i.e. we remove junk)
- canonical ratio = "Canonical"/("Canonical"+"Non-Canonical") (i.e. we remove junk + non-linguistic vocalizations)

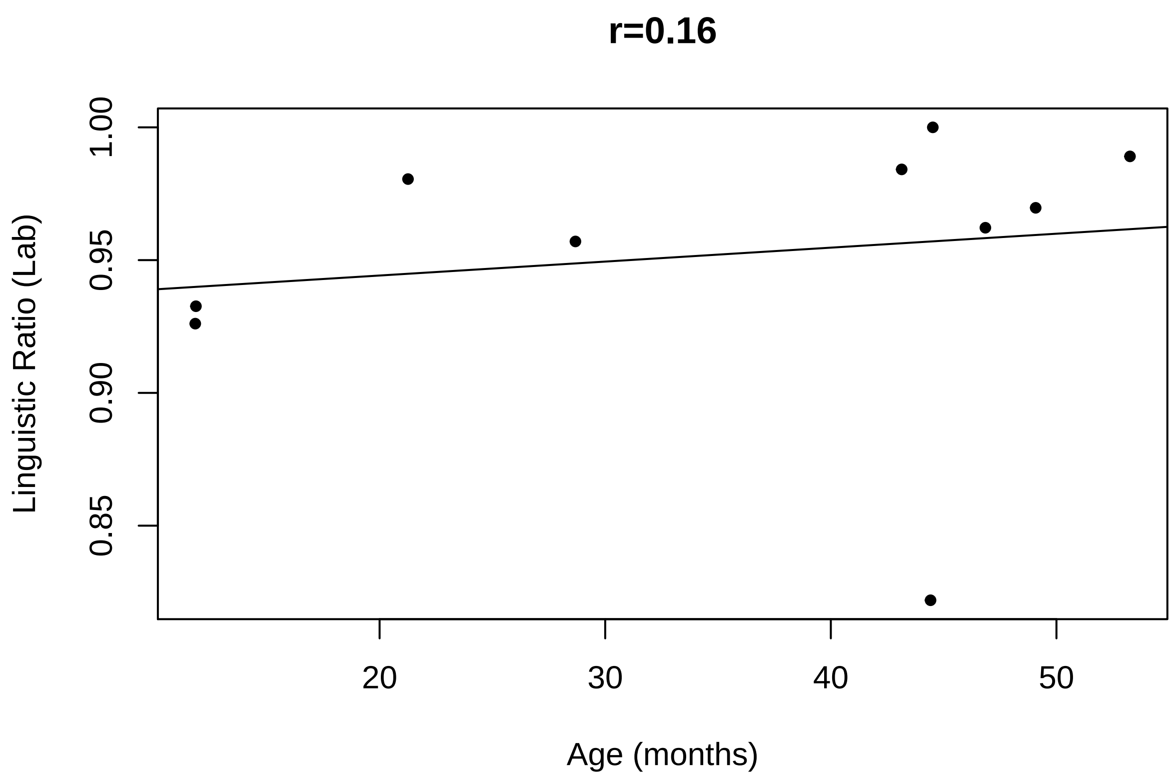
As expected, linguistic ratio goes up with age.

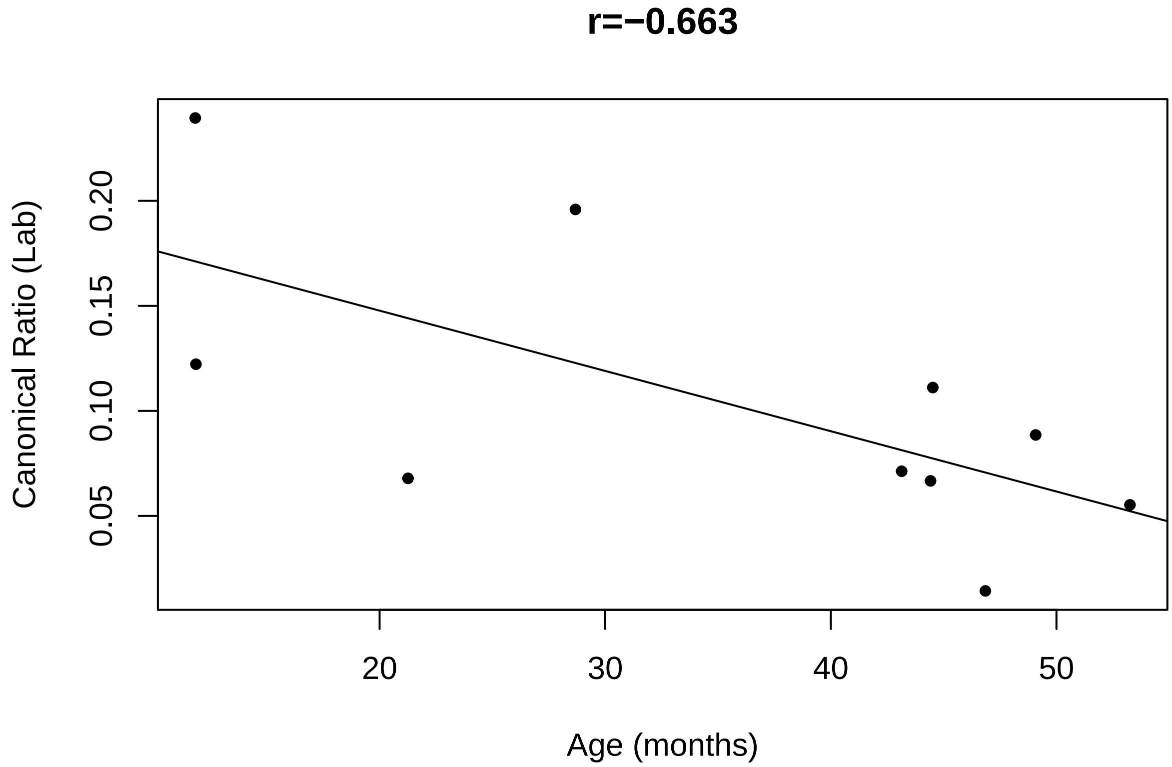
Surprisingly, canonical ratio goes DOWN with age.

```
prettynames=c("Linguistic Ratio (Zooniverse)","Canonical Ratio (Zooniverse)",
             "Linguistic Ratio (Lab)","Canonical Ratio (Lab)" )
names(prettynames)<-c("z_lr","z_cr","l_lr","l_cr")
for(thisvar in c("z_lr","z_cr","l_lr","l_cr")) {
  myr=round(cor.test(ratios[,thisvar],ratios$Age)$estimate,3)
  plot(ratios[,thisvar]~ratios$Age, pch=20,xlab="Age (months)",ylab=prettynames[thisvar],main=paste0("r
  abline(lm(ratios[,thisvar]~ratios$Age))
  }
```


r=0.014
Linguistic Ratio (Zooniverse)
0.80
0.85
0.90
0.95
20
30
40
50
Age (months)


r=−0.536
Canonical Ratio (Zooniverse)
0.05
0.10
0.15
0.20
20
30
40
50
Age (months)





But the key thing for us: Are Zooniverse annotations describing children similar to lab annotations? The answer is clearly yes.

```
#Ling ratio
pdf("ling_rat_z_vs_l.pdf",height=5,width=5)
lims=range(c(ratios[,"z_lr"],ratios[,"l_lr"]))
```

```
  myr=round(cor.test(ratios[,"z_lr"],ratios[,"l_lr"])$estimate,3)
  plot(ratios[,"z_lr"]~ratios[,"l_lr"], pch=20,xlab=prettynames["l_lr"],ylab=prettynames["z_lr"],main=pa
       xlim=lims,ylim=lims)
  abline(lm(ratios[,"z_lr"]~ratios[,"l_lr"]))
  lines(c(0,1),c(0,1),lty=2,col="darkgray")
dev.off()
```

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

```
  #CR
pdf("can_rat_z_vs_l.pdf",height=5,width=5)
lims=range(c(ratios[,"z_cr"],ratios[,"l_cr"]))
    myr=round(cor.test(ratios[,"z_cr"],ratios[,"l_cr"])$estimate,3)
  plot(ratios[,"z_cr"]~ratios[,"l_cr"], pch=20,xlab=prettynames["l_cr"],ylab=prettynames["z_cr"],main=pa
       xlim=lims,ylim=lims)
  abline(lm(ratios[,"z_cr"]~ratios[,"l_cr"]),col="darkgray")
    lines(c(0,1),c(0,1),lty=2,col="darkgray")
dev.off()
```

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