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 demo data created by AC from info in paper
demo_data=read.csv("./Data/demo-data.tsv",sep="\t")

# read dataset composed with python
data_all <- read.csv("./Data/new_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").

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

```
##
##                     Canonical                 Non-Canonical
##                           271                          2669
##                        Crying                      Laughing
##                            54                            52
##                          Junk            Laughing_Canonical
##                           958                             0
##               Laughing_Crying        Laughing_Non-Canonical
##                             0                             0
## Laughing_Non-Canonical_Crying          Non-Canonical_Crying
##                             0                             0
```

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

```
##
##                     Canonical                 Non-Canonical
##                           262                          2654
##                        Crying                      Laughing
##                           100                           138
##                          Junk            Laughing_Canonical
##                           614                             3
##               Laughing_Crying        Laughing_Non-Canonical
##                             5                            93
## Laughing_Non-Canonical_Crying          Non-Canonical_Crying
##                             7                           128
```

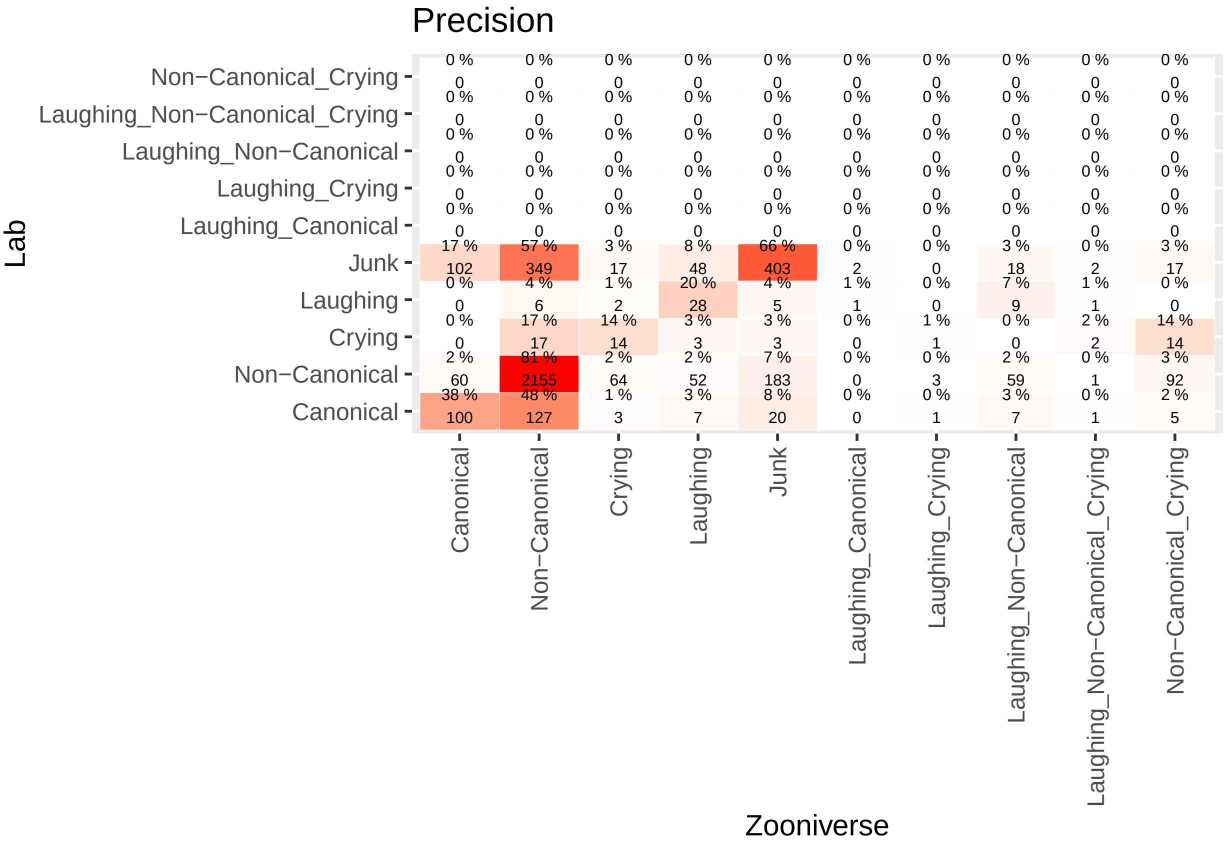
```
mycf=confusionMatrix(data_all$lab, data_all$Zoon_classif, dnn = c("Lab","Zooniverse"))

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                           100           127      3        7   20
##   Non-Canonical                        60          2155     64       52  183
##   Crying                                0            17     14        3    3
##   Laughing                              0             6      2       28    5
##   Junk                                102           349     17       48  403
##   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
##                                Zooniverse
## Lab                             Laughing_Canonical Laughing_Crying
##   Canonical                                      0               1
##   Non-Canonical                                  0               3
##   Crying                                         0               1
##   Laughing                                       1               0
##   Junk                                           2               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
##                                Zooniverse
## Lab                             Laughing_Non-Canonical
##   Canonical                                          7
##   Non-Canonical                                     59
##   Crying                                             0
##   Laughing                                           9
##   Junk                                              18
##   Laughing_Canonical                                 0
##   Laughing_Crying                                    0
##   Laughing_Non-Canonical                             0
##   Laughing_Non-Canonical_Crying                      0
##   Non-Canonical_Crying                               0
##                                Zooniverse
## Lab                             Laughing_Non-Canonical_Crying
##   Canonical                                                 1
##   Non-Canonical                                             1
##   Crying                                                    2
##   Laughing                                                  1
##   Junk                                                      2
##   Laughing_Canonical                                        0
##   Laughing_Crying                                           0
##   Laughing_Non-Canonical                                    0
##   Laughing_Non-Canonical_Crying                             0
##   Non-Canonical_Crying                                      0
##                                Zooniverse
## Lab                             Non-Canonical_Crying
##   Canonical                                        5
##   Non-Canonical                                   92
##   Crying                                          14
##   Laughing                                         0
##   Junk                                            17
##   Laughing_Canonical                               0
##   Laughing_Crying                                  0
##   Laughing_Non-Canonical                           0
##   Laughing_Non-Canonical_Crying                    0
##   Non-Canonical_Crying                             0
##
## Overall Statistics
##
##                Accuracy : 0.6743
##                  95% CI : (0.6596, 0.6888)
##     No Information Rate : 0.6628
##     P-Value [Acc > NIR] : 0.06382
##
##                   Kappa : 0.3692
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Canonical Class: Non-Canonical Class: Crying
## Sensitivity                   0.38168               0.8120      0.140000
## Specificity                   0.95430               0.6193      0.989754
## Pos Pred Value                0.36900               0.8074      0.259259
## Neg Pred Value                0.95660               0.6262      0.978228
## Prevalence                    0.06543               0.6628      0.024975
## Detection Rate                0.02498               0.5382      0.003497
## Detection Prevalence          0.06768               0.6666      0.013487
## Balanced Accuracy             0.66799               0.7156      0.564877
##                      Class: Laughing Class: Junk Class: Laughing_Canonical
## Sensitivity                 0.202899      0.6564                 0.0000000
## Specificity                 0.993792      0.8363                 1.0000000
## Pos Pred Value              0.538462      0.4207                       NaN
## Neg Pred Value              0.972166      0.9307                 0.9992507
## Prevalence                  0.034466      0.1533                 0.0007493
## Detection Rate              0.006993      0.1006                 0.0000000
## Detection Prevalence        0.012987      0.2393                 0.0000000
## Balanced Accuracy           0.598345      0.7463                 0.5000000
##                      Class: Laughing_Crying Class: Laughing_Non-Canonical
## Sensitivity                        0.000000                       0.00000
## Specificity                        1.000000                       1.00000
## Pos Pred Value                          NaN                           NaN
## Neg Pred Value                     0.998751                       0.97677
## Prevalence                         0.001249                       0.02323
## Detection Rate                     0.000000                       0.00000
## Detection Prevalence               0.000000                       0.00000
## Balanced Accuracy                  0.500000                       0.50000
##                      Class: Laughing_Non-Canonical_Crying
## Sensitivity                                      0.000000
## Specificity                                      1.000000
## Pos Pred Value                                        NaN
## Neg Pred Value                                   0.998252
## Prevalence                                       0.001748
## Detection Rate                                   0.000000
## Detection Prevalence                             0.000000
## Balanced Accuracy                                0.500000
##                      Class: Non-Canonical_Crying
## Sensitivity                              0.00000
## Specificity                              1.00000
## Pos Pred Value                               NaN
## Neg Pred Value                           0.96803
## Prevalence                               0.03197
## Detection Rate                           0.00000
## Detection Prevalence                     0.00000
## Balanced Accuracy                        0.50000
```

Precision

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

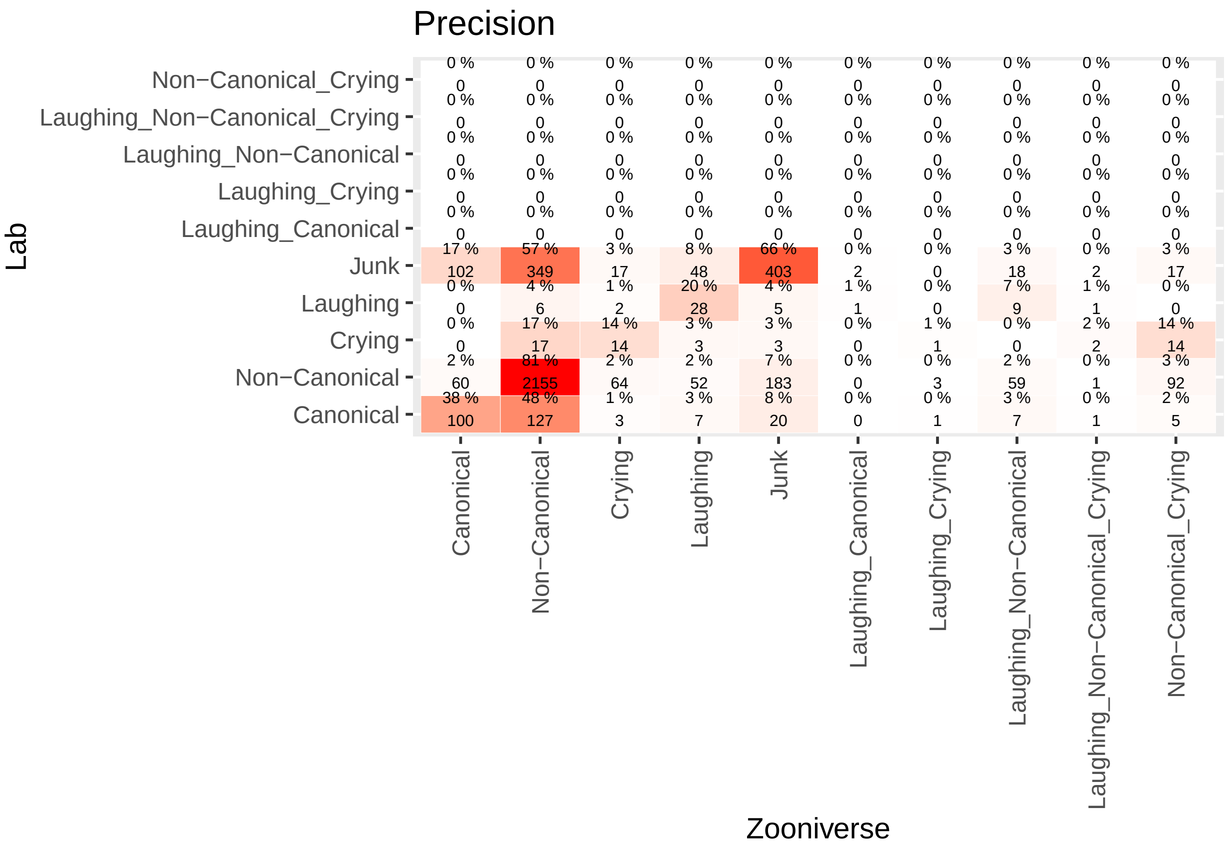
```
prop_cat=data.frame(conf_tab/colSums(conf_tab)*100) #generates precision because columns
prop_cat$id=paste(prop_cat$Lab,prop_cat$Zooniverse)
colnames(prop_cat)[3]<-"pr"

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))
```



Recall

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

```
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))
```



repeat collapsing

```
#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)

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           100           140      4        7   20
##   Non-Canonical        60          2307     67       52  183
##   Crying                0            33     15        3    3
##   Laughing              1            16      2       28    5
##   Junk                104           386     17       48  403
##
## Overall Statistics
##
##                Accuracy : 0.7125
##                  95% CI : (0.6982, 0.7265)
##     No Information Rate : 0.7198
##     P-Value [Acc > NIR] : 0.8503
##
##                   Kappa : 0.3989
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: Canonical Class: Non-Canonical Class: Crying
## Sensitivity                   0.37736               0.8005      0.142857
## Specificity                   0.95427               0.6774      0.989997
## Pos Pred Value                0.36900               0.8644      0.277778
## Neg Pred Value                0.95580               0.5693      0.977215
## Prevalence                    0.06618               0.7198      0.026224
## Detection Rate                0.02498               0.5762      0.003746
## Detection Prevalence          0.06768               0.6666      0.013487
## Balanced Accuracy             0.66581                0.7389      0.566427
##                        Class: Laughing Class: Junk
## Sensitivity                 0.202899      0.6564
## Specificity                 0.993792      0.8363
## Pos Pred Value              0.538462      0.4207
## Neg Pred Value              0.972166      0.9307
## Prevalence                  0.034466      0.1533
## Detection Rate              0.006993      0.1006
## Detection Prevalence        0.012987      0.2393
## Balanced Accuracy           0.598345      0.7463
```

```r
pdf("./Results/precision.pdf",height=10,width=10)
colsums=colSums(conf_tab)
my_conf_tab=conf_tab
for(i in 1:5) my_conf_tab[,i]=my_conf_tab[,i]/colsums[i]
colSums(my_conf_tab)
```

```
##     Canonical Non-Canonical        Crying      Laughing          Junk
##             1             1             1             1             1
```

```r
prop_cat=data.frame(my_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=8) +
  geom_text(aes(label = Freq), vjust = 1,size=8) +
  scale_fill_gradient(low = "white", high = "red", name = "Proportion") +
     theme(legend.position = "none") +
  xlab("Zooniverse") + ylab("Lab") +
  ggtitle("Precision")+theme(text = element_text(size=20),
        axis.text.x = element_text(angle=90, hjust=1))
dev.off()
```

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

```r
pdf("./Results/recall.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=8) +
  geom_text(aes(label = Freq), vjust = 1,size=8) +
  scale_fill_gradient(low = "white", high = "red", name = "Proportion") +
     theme(legend.position = "none") +
  xlab("Zooniverse") + ylab("Lab") +
  ggtitle("Recall")+theme(text = element_text(size=20),
        axis.text.x = element_text(angle=90, 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. In all of these graphs, red points correspond to children diagnosed with Angelman Syndrome, black for low-risk control.

```
#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
#improvement: now we merge with a demo data tab, but note this is merged with child id, so the problem


merge(ratios,demo_data,by="ChildID")->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.   :0.7667   Min.   :0.02872   Min.   :0.8289
##  1151_1 :1   1st Qu.:0.8967   1st Qu.:0.03864   1st Qu.:0.9383
##  1801_1 :1   Median :0.9321   Median :0.06747   Median :0.9650
##  2881_1 :1   Mean   :0.9120   Mean   :0.10022   Mean   :0.9524
##  3021_1 :1   3rd Qu.:0.9547   3rd Qu.:0.14691   3rd Qu.:0.9840
##  3041_1 :1   Max.   :0.9710   Max.   :0.23474   Max.   :1.0000
##  (Other):4
##       l_cr                    Diagnosis Sex        Age
##  Min.   :0.01701   AngelmanSyndrome:9   F:3   Min.   :11.83
##  1st Qu.:0.06450   Low-RiskControl :1   M:7   1st Qu.:23.11
##  Median :0.07924                              Median :43.78
##  Mean   :0.10246                              Mean   :35.49
##  3rd Qu.:0.11905                              3rd Qu.:46.27
##  Max.   :0.23529                              Max.   :53.26
##
```

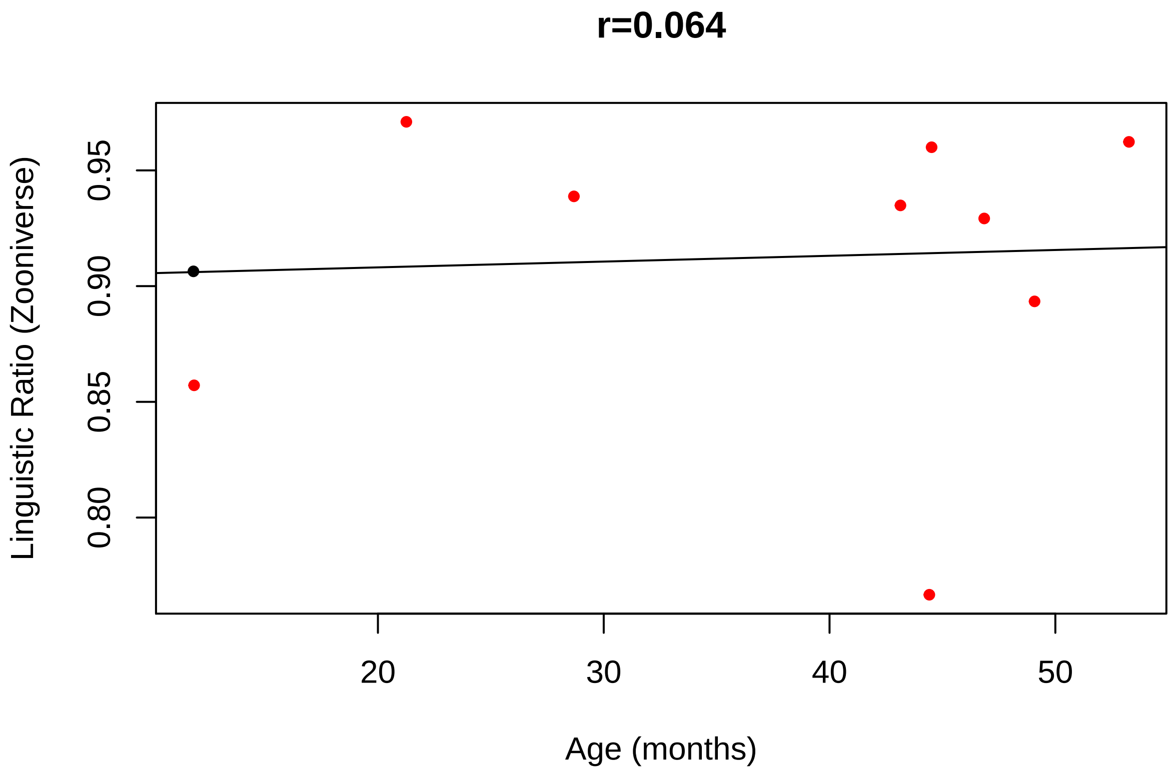
We first look generally at two measures that have been found to relate to age:

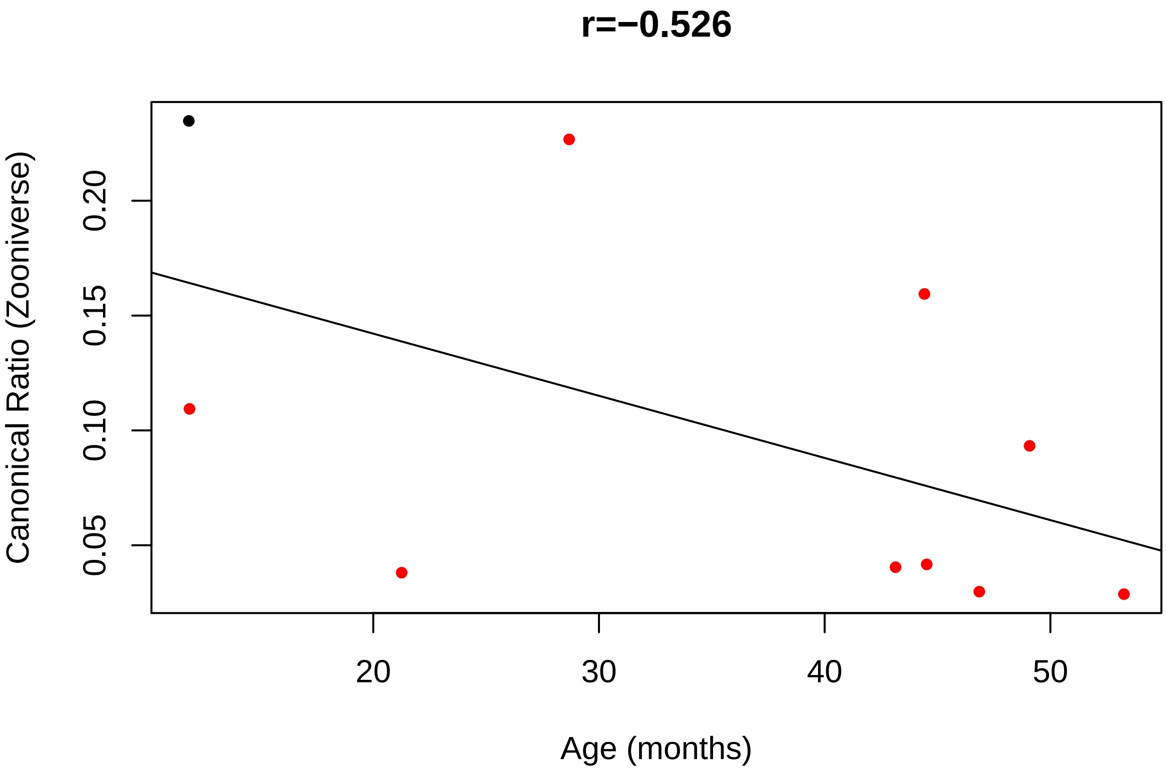
- linguistic ratio = (“Canonical”+“Non-Canonical”)/“All vocalizations” (i.e. we remove junk)
- canonical ratio = “Canonical”/(“Canonical”+“Non-Canonical”) (i.e. we remove junk + non-linguistic vocalizations)

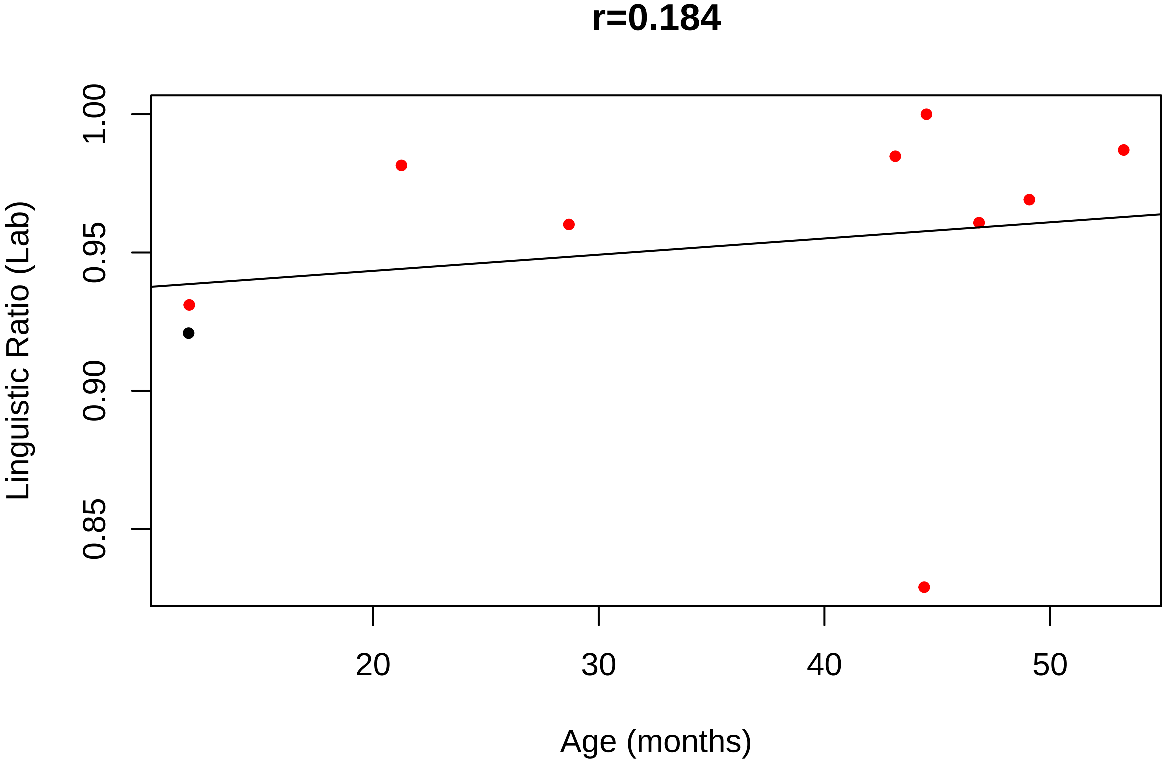
As expected, linguistic ratio goes up with age.

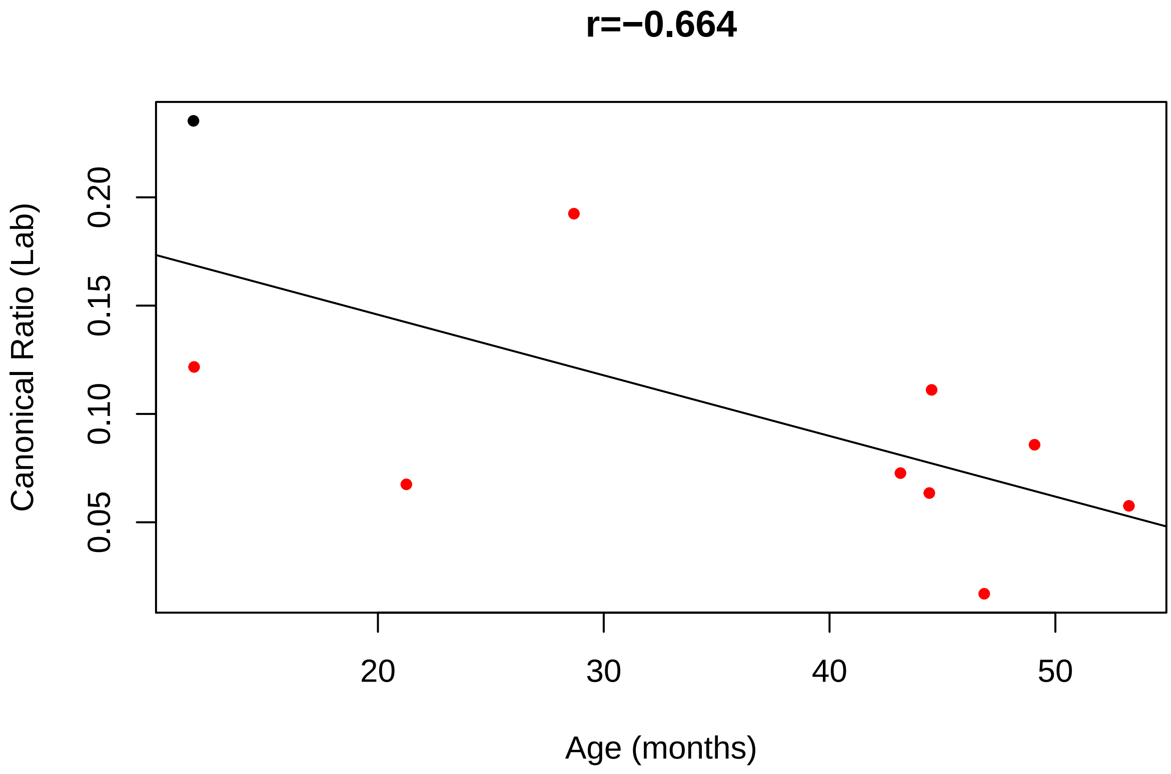
Surprisingly, canonical ratio goes DOWN with age.

```
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")

#!!! this is working the opposite than it should! but note that to get angsynd kids to come out in red,
mycols=c("red","black")
names(mycols)<-c("Low-RiskControl","AngelmanSyndrome")

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$Age, pch=20,xlab="Age (months)",ylab=prettynames[thisvar],main=paste0("r
       col=mycols[ratios$Diagnosis])
  abline(lm(ratios[,thisvar]~ratios$Age))
  }
```




r=−0.526
Canonical Ratio (Zooniverse)
0.05
0.10
0.15
0.20
20
30
40
50
Age (months)


r=0.184
Linguistic Ratio (Lab)
0.85
0.90
0.95
1.00
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("./Results/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"])$estimate,3)
  plot(ratios[,"z_lr"]~ratios[,"l_lr"], pch=20,xlab=prettynames["l_lr"],ylab=prettynames["z_lr"],main=pa
       xlim=lims,ylim=lims,
       col=mycols[ratios$Diagnosis])
  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("./Results/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"])$estimate,3)
  plot(ratios[,"z_cr"]~ratios[,"l_cr"], pch=20,xlab=prettynames["l_cr"],ylab=prettynames["z_cr"],main=pa
       xlim=lims,ylim=lims,
       col=mycols[ratios$Diagnosis])
  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
```

```
#COMBINED to save space
pdf("./Results/combined.pdf",height=5,width=5)
lims=range(c(ratios[,"z_lr"],ratios[,"l_lr"]),c(ratios[,"z_cr"],ratios[,"l_cr"]))
  #myr=round(cor.test(ratios[,"z_lr"],ratios[,"l_lr"])$estimate,3)

  plot(ratios[,"z_lr"]~ratios[,"l_lr"],xlab="Laboratory annotations",ylab="Zooniverse annotations",
       xlim=lims,ylim=lims,
       pch=20,col=mycols[ratios$Diagnosis])
    points(ratios[,"z_cr"]~ratios[,"l_cr"], pch=2, col=mycols[ratios$Diagnosis])
  abline(lm(ratios[,"z_cr"]~ratios[,"l_cr"]))
  abline(lm(ratios[,"z_lr"]~ratios[,"l_lr"]),lty=3)
 # lines(c(0,1),c(0,1),lty=2,col="darkgray")
dev.off()
```

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