Analyses for JSLHR version

2020-10-17

Contents

History:

- 2020-08-05 final first version
- 2020-10-10 (rehaul),
- latest minor edits 2020-10-17

TODO:

- pipeline is still not transparent
- there are duplicate files across raw and derived data
- there are a bunch of files with similar names
- README is old
- a note said read demo data created by AC from info in paper - should be replaced with real demo data
- add code to print out the results paragraphs
- corage: - adjust margins and remove title repetition?
- chunk-based analyses for age

Read data in

```r
# read datasets

data_ang <- read.csv("../Derived_Data/classifications_PU_zoon_final17.csv",header=T,sep=",")
data_td <- read.csv("../Derived_Data/classifications_PU_zoon_final.csv")
data_all<-rbind(data_ang, data_td)


#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",
# 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

#add binomials for linguistic proportion
data_all$lab_ling=ifelse(data_all$lab %in% c("Canonical","Non-Canonical"),1,0)
data_all$zoo_ling=ifelse(data_all$Zoon_classif %in% c("Canonical","Non-Canonical"),1,0)
data_all$lab_ling[data_all$lab=="Junk"]<-NA
data_all$zoo_ling[data_all$lab=="Junk"]<-NA

#add binomials for canonical proportion
data_all$lab_can=data_all$zoo_can=NA
data_all$lab_can[data_all$lab=="Canonical"]<-1
data_all$lab_can[data_all$lab=="Non-Canonical"]<-0
data_all$zoo_can[data_all$Zoon_classif=="Canonical"]<-1
data_all$zoo_can[data_all$Zoon_classif=="Non-Canonical"]<-0


demo_data=read.csv("../Derived_Data/demo-data.tsv",sep="\t")
#add filenames to demo data, to be used later
demo_data_fn <- demo_data %>%
     left_join(select(data_all, filename, ChildID), by = c("ChildID"))
```

```
## Warning: Column `ChildID` joining factors with different levels, coercing to
## character vector
```

```
demo_data_fn<-unique(demo_data_fn)
```

Data post-processing

We collected a total of 169,628 judgments provided for 33,880 500-ms chunks, corresponding to 11,984 LENA segments. Nearly a fifth of chunks did not have at least 3 labels in agreement out of the 5 Zooniverse labels (N = 6,585, 19% of all chunks). Of the chunks without a majority agreement, 4341 (66%) contained one or two Junk judgements (out of 5), 6523 (99,9%) had at least two matching judgements (the threshold used for lab-annotated segments), and only 61 (0,01%) had 5 different judgements. Future work may explore different ways of setting the minimal requirement for convergence, but for further analyses here, we focused on the 81% of chunks that did have at least 3 labels in agreement; this represented 136,703 labels for 27,295 chunks, corresponding to 11,593 LENA segments. As the segments average 1.12 seconds in length, this means about 3.8 hours of audio data were annotated by 8 different annotators (3 in the laboratory, 5 on Zooniverse).

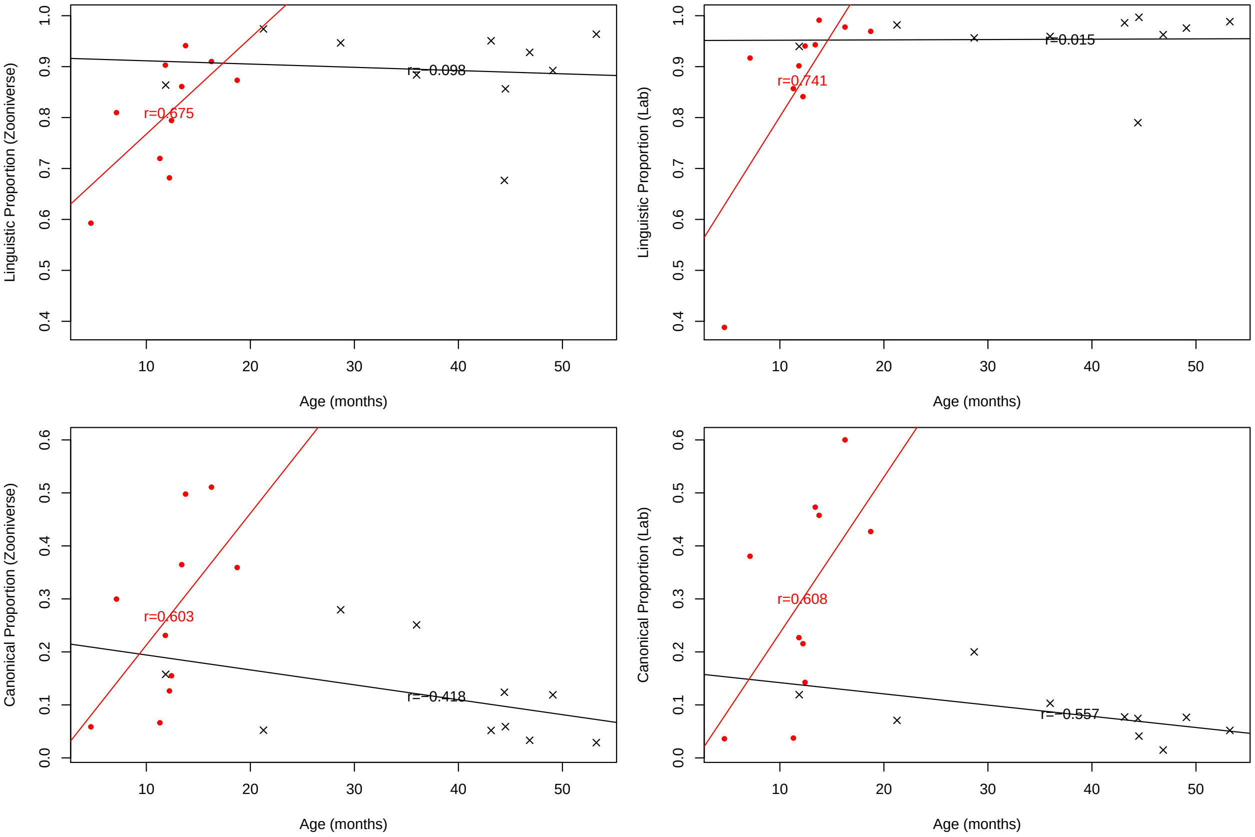
```
##                      filename       z_lp             z_cp
##  20180206_110905_009463: 1   Min.   :0.5927   Min.   :0.02882
##  20180419_111712_022875: 1   1st Qu.:0.8058   1st Qu.:0.05891
##  20180530_180405_024879: 1   Median :0.8785   Median :0.14065
##  20180530_181101_022875: 1   Mean   :0.8511   Mean   :0.19128
##  20180808_111325_024882: 1   3rd Qu.:0.9313   3rd Qu.:0.28454
##  20180906_133011_022875: 1   Max.   :0.9740   Max.   :0.51084
##  (Other)               :14
##       l_lp             l_cp             z_junk          ChildID
##  Min.   :0.3880   Min.   :0.01488   Min.   :0.04329   Length:20
##  1st Qu.:0.9130   1st Qu.:0.06609   1st Qu.:0.09418   Class :character
##  Median :0.9578   Median :0.11115   Median :0.12230   Mode  :character
##  Mean   :0.9131   Mean   :0.19129   Mean   :0.13302
##  3rd Qu.:0.9787   3rd Qu.:0.26529   3rd Qu.:0.16368
##  Max.   :0.9968   Max.   :0.60000   Max.   :0.27706
##
##             Diagnosis  Sex         Age        median_junk
##  AngelmanSyndrome:10   F: 7   Min.   : 4.67   Length:20
##  Low-RiskControl :10   M:13   1st Qu.:12.13   Class :character
##                               Median :17.50   Mode  :character
##                               Mean   :25.04
##                               3rd Qu.:43.46
##                               Max.   :53.26
##
```

Results

Descriptive analyses

In this section, we provide descriptive analyses of our dataset. According to lab annotators, 15% of segments were canonical, 56% non-canonical, 2% laughing, and 5% crying, with the remaining 22% being categorized as "Don't code". Zooniverse data revealed a similar distribution: 15% canonical, 60% non-canonical, 4% laughing, 9% crying, 13% junk. Next, we inspected the relationship between age and child-level derived metrics, of which we had two: i) Linguistic proportion = ("Canonical"+"Non-Canonical")/"All vocalizations" (i.e., we remove junk), and ii) Canonical proportion = "Canonical"/("Canonical"+"Non-Canonical") (i.e. we remove junk + non-linguistic vocalizations). See Figure XX for results.

Figure XX. Correlations between child-level descriptors and age as a function of metric (linguistic ratio in the top row, canonical ratio in the bottom row), annotation method, and child group (Red as TD, and black as AS)



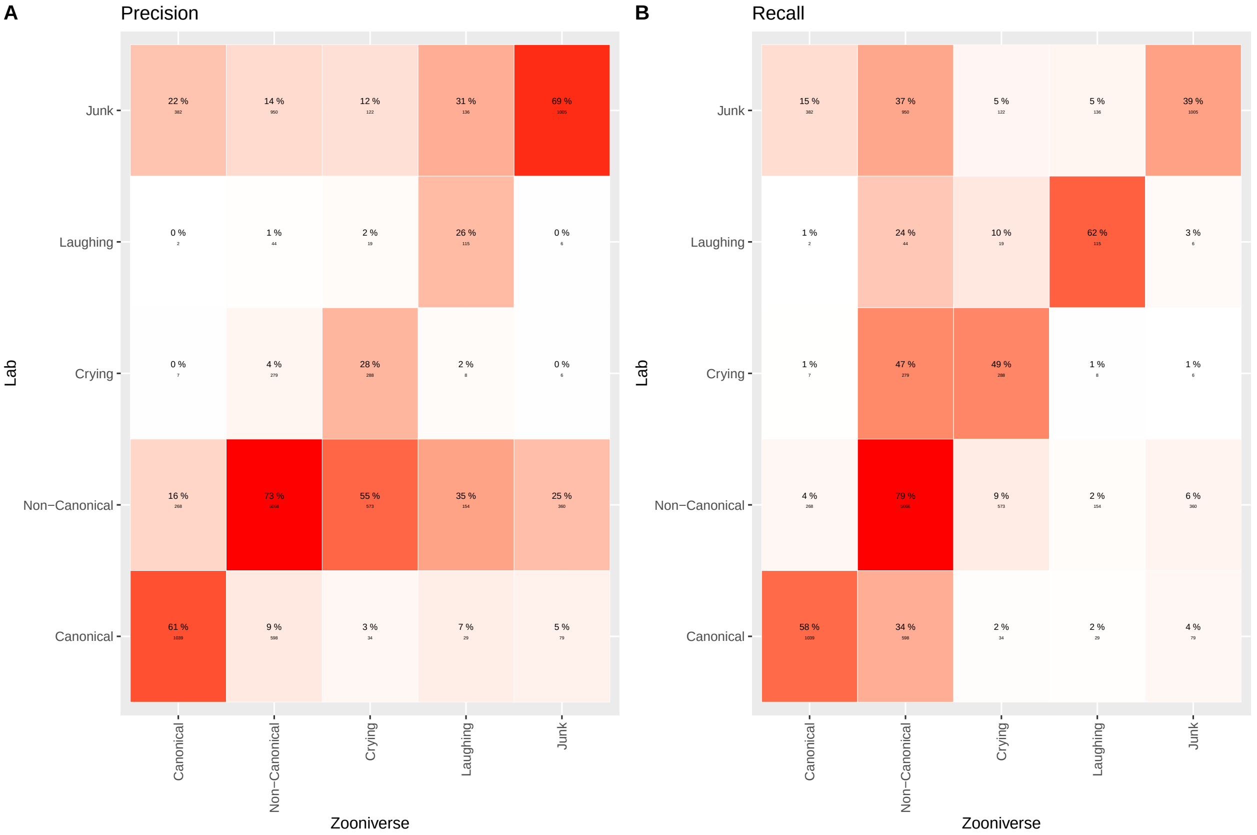
Descriptive analyses on the laboratory annotations showed that correlations between the Linguistic proportion and age differed across the groups. There was a near-zero relationship among the older children diagnosed with Angelman Syndrome r(8) = 0.015, CI [-0.621,0.638], p=0.968]; and a significant association among younger low-risk control children r(8) = 0.741, CI [0.209,0.935], p=0.014]. The Canonical proportion exhibited non-significant developmental decreases among older children diagnosed with Angelman Syndrome r(8) = -0.557, CI [-0.879,0.112], p=0.094]; and marginal developmental increases among low-risk control r(8) = 0.608, CI [-0.035,0.895], p=0.062].

Using the Zooniverse annotations, we found that the association with age was very weak for children diagnosed with Angelman Syndrome r(8) = -0.098, CI [-0.685,0.567], p=0.788]; whereas younger low-risk control children showed a significant increase with age r(8) = 0.675, CI [0.079,0.916], p=0.032]. Similarly, there were non-significant developmental decreases in the Canonical among children with Angelman Syndrome r(8) = -0.418, CI [-0.829,0.288], p=0.23]; and marginal developmental increases among low-risk control children, r(8) = 0.603, CI [-0.043,0.893], p=0.065].

Main analyses

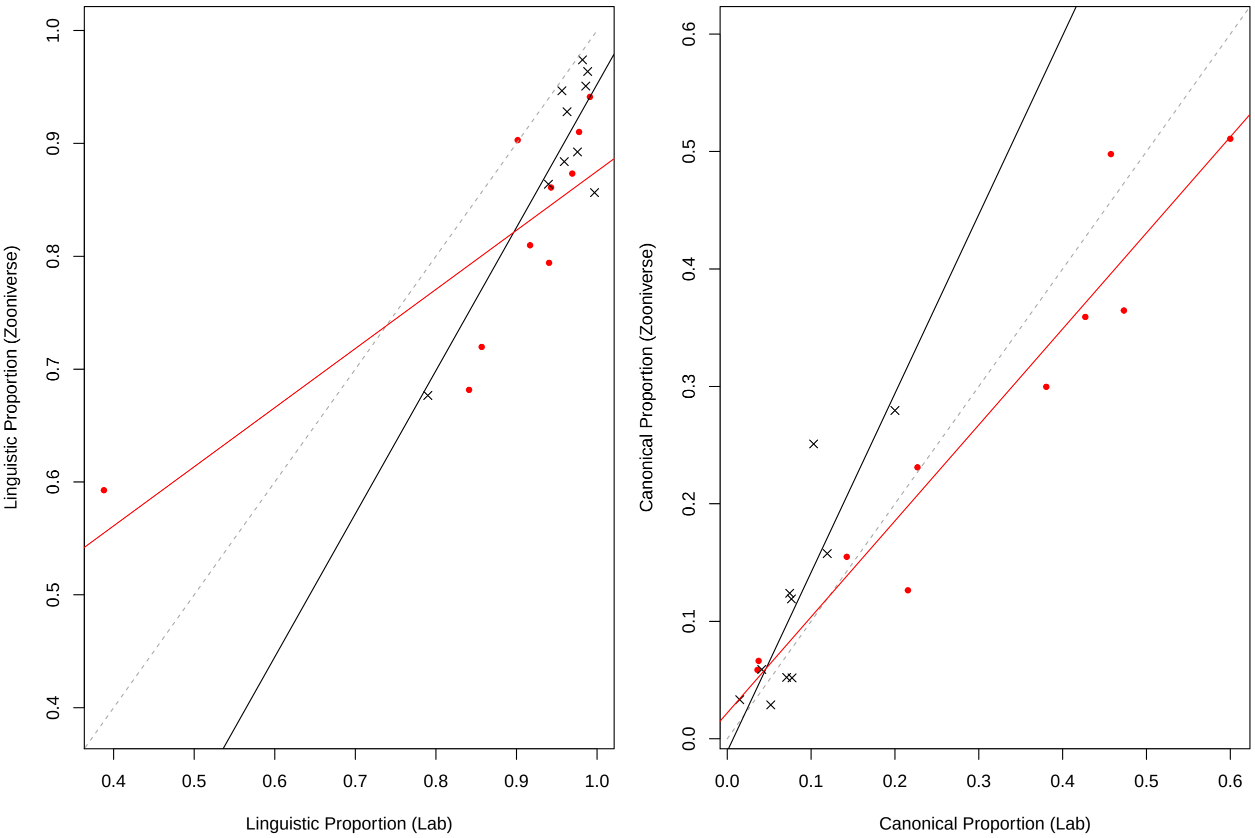
Next, we discuss the correspondence between citizen science classifications and the laboratory gold standard, at the level of individual clips. Results were visualized and assessed with a confusion matrix. We report a Precision plot (Fig. 1) and a Recall plot (Fig. 2): the diagonal elements show the number of correct segment-level classifications for each class while the off-diagonal elements show non-matching classifications.

```
##     Canonical Non-Canonical        Crying      Laughing          Junk
##     0.9213192     2.8808609     0.5824913     0.6608792     1.6885824
```



From both visualizations, it appears that performance is moderate to good, with an overall accuracy of 65%, CI = [64,66], a kappa of 0.426, and a Gwet's AC1 coefficient of 0.587, CI = [0.576,0.597].

Child level descriptors



Although the classification at the clip level is only moderately accurate, what we are ultimately interested in is whether citizen scientists' classifications are able to provide a reliable snapshot of childrens' individual development. Looking at all 20 children together, we found a strong positive correlation r(18) = 0.811, CI [0.575,0.923], p=0] between Linguistic proportion by child from the Zooniverse and the lab annotators' data. When we split by participant group, correlations remain high: for Angelman Syndrome r(8) = 0.877, CI [0.552,0.971], p=0.001]; low-risk control r(8) = 0.823, CI [0.401,0.957], p=0.003].

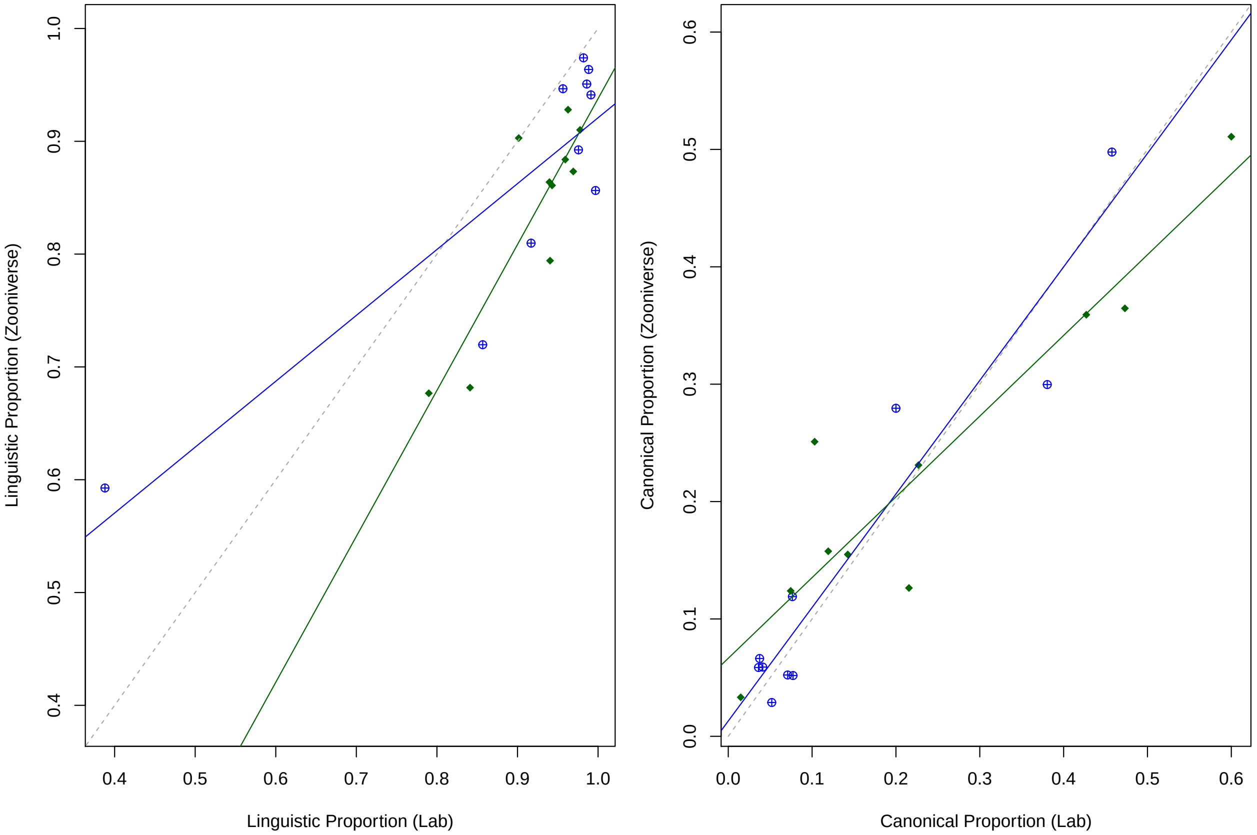
Similarly, a strong positive correlation is found in the Canonical Proportion r(18) = 0.937, CI [0.844,0.975], p=0]. When we split by participant group, correlations remain high although we do note they are somewhat smaller for the older children with Angelman Syndrome: r(8) = 0.858, CI [0.498,0.966], p=0.001]; than the low-risk control children r(8) = 0.96, CI [0.833,0.991], p=0.

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00742676 (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00215928 (tol = 0.001, component 1)
```

Additional analyses

We additionally explored under what conditions Zooniverse judgments more closely aligned with laboratory judgments. In previous work using a similar method, for instance, data from all three children from one dataset were often labeled as "Junk" (i.e., not a child's vocalization), and the data points from this corpus stood out when the authors attempted to integrate results with other corpora [@cychosz]. A high proportion of "Junk" may indicate that automated segmentation was errorful for those children, and may be a sign that the rest of the data could be compromised as well.



We investigated this hypothesis by calculating the proportion of their data labeled as “Junk” for each individual child. There was no significant difference in the proportions of their data labeled as “Junk” for older children with Angelman Syndrome (M = 0.126) compared to the younger low risk children (M = 0.14): Welch’s t(13.3654471) = -0.516, p = 0.614. We therefore collapsed across groups for this exploratory analysis, and split the 20 children using a median split on the proportion of their data labeled as “Junk”. Results were similar across these two post-hoc subgroups. For Linguistic Proportion, the correlation across lab and Zooniverse data for the lower junk group was r(8) = 0.871, CI [0.534,0.969], p=0.001]; and for the higher junk group it was r(8) = 0.872, CI [0.539,0.97], p=0.001. For Canonical Proportion, the correlation across lab and Zooniverse data for the lower junk group was r(8) = 0.957, CI [0.822,0.99], p=0]; and for the higher junk group it was r(8) = 0.931, CI [0.729,0.984], p=0. Thus, it does not seem that a higher proportion of “Junk” judgments is an index of low quality data.

Above, we concluded that derived metrics seem more promising than segment-level data. Notice that derived metrics do not require matching of clips (500 ms presented to Zooniverse participants) to segments (the original LENA segments presented to laboratory participants). As a result, there was one stage in our pre-processing that may not have been necessary, whereby we collapsed judgments across chunks associated to the same segment. We therefore repeated our analyses but deriving our proportions for the Zooniverse data not from the segment-level composite, but rather the individual chunk-level annotations.

Regarding correlations with age, we found that the Linguistic proportion was not strongly associated with age in the Angelman Syndrome group: r(8) = -0.0377654, CI [-0.651891,0.606277], p=0.917506], but increased with age in the low-risk control group: r(8) = 0.7156353, CI [0.1565459,0.9273939], p=0.0199552. Similarly, the Canonical proportion did not exhibit the same pattern across the groups, with developmental decreases found among children with Angelman Syndrome r(8) = -0.3804163, CI [-0.8148658,0.327701], p=0.2781704]; and developmental increases among low-risk control r(8) = 0.5730755, CI [-0.0884744,0.8838045], p=0.0833178.

As for correlations between Zooniverse and laboratory-derived metrics, we observe very similar levels of correlation: linguistic proportion overall r(18) = 0.845, CI [0.644,0.937], p=0] between Linguistic proportion by child from the Zooniverse and the lab annotators’ data. When we split by participant group, correlations remain high: for Angelman Syndrome r(8) = 0.845, CI [0.459,0.962], p=0.002]; low-risk control r(8) = 0.841, CI [0.45,0.962], p=0.002]. As for canonical proportion overall r(18) = 0.963, CI [0.906,0.985], p=0; Angelman Syndrome: r(8) = 0.866, CI [0.519,0.968], p=0.001]; low-risk control infants r(8) = 0.975, CI [0.892,0.994], p=0.

Since all of these results are very similar to those obtained when first matching clips to segments, we conclude that in the future this step may not be necessary. Instead, researchers can derive linguistic and canonical proportions directly from citizen scientists' clip level judgements [which was in fact what @cyhosz did].

```
#we start from the data base that has all the info
nsegs=data.frame(table(data_all$ChildID))
colnames(nsegs)<-c("ChildID","nsegs")

merge(nsegs,demo_data)->nsegs
nseg_t=t.test(nsegs$nsegs~nsegs$Diagnosis)

resampling=matrix(NA,nrow=50,ncol=6)
colnames(resampling)<-c("lp_r_all","lp_r_as","lp_r_lr","cp_r_all","cp_r_as","cp_r_lr")
for(i in 1:50){
  sel_seg=NULL
  for(thischi in levels(data_all$ChildID)) sel_seg<-c(sel_seg,as.character(data_all$segmentId_DB[sample

  sampling=data_all[as.character(data_all$segmentId_DB) %in% as.character(sel_seg),]

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

  ltab=table(sampling$ChildID,sampling$lab)
  l_lp=rowSums(ltab[,c("Canonical","Non-Canonical")])/rowSums(ltab[,-which(colnames(ztab) %in% c("Junk")
  l_cp=ltab[,c("Canonical")]/rowSums(ltab[,c("Canonical","Non-Canonical")])

  thistab=cbind(demo_data,z_lp,z_cp,l_lp,l_cp)

  resampling[i,"lp_r_all"]=cor.test(thistab$z_lp,thistab$l_lp)$estimate
  resampling[i,"lp_r_as"]=cor.test(thistab$z_lp[thistab$Diagnosis=="AngelmanSyndrome"],thistab$l_lp[thi
  resampling[i,"lp_r_lr"]=cor.test(thistab$z_lp[thistab$Diagnosis!="AngelmanSyndrome"],thistab$l_lp[thi

   resampling[i,"cp_r_all"]=cor.test(thistab$z_cp,thistab$l_cp)$estimate
  resampling[i,"cp_r_as"]=cor.test(thistab$z_cp[thistab$Diagnosis=="AngelmanSyndrome"],thistab$l_cp[thi
  resampling[i,"cp_r_lr"]=cor.test(thistab$z_cp[thistab$Diagnosis!="AngelmanSyndrome"],thistab$l_cp[thi
 }
myr=colMeans(resampling)
mycortab=rbind(mycortab,cbind(myr[c("lp_r_as" ,"lp_r_lr","lp_r_all")],myr[c("cp_r_as" ,"cp_r_lr","cp_r_a
rownames(mycortab)[7:9]<-c("100 seg AS","100 seg LR","100 seg all")
round(mycortab,3)->mycortab
colnames(mycortab)<-c("Ling. Prop.","Can. Prop.")
resampling=data.frame(resampling)
```

Next, we looked at whether having more segments from each child may lead to more reliable metrics. We observed a non-significant trend [t(17.923429)=527.4, 631.9] for lower number of segments in the Angelman Syndrome group (mean 527.4, range 326-924) than among low-risk control infants (mean 631.9, range 351-1074), likely because of the lower volubility of the former children. Moreover, these numbers of segments are much larger than those used in previous work relying on citizen science classifications [CITE MEG].

We therefore asked whether the correlations between laboratory and Zooniverse child-level estimates are affected by how much data is extracted from each child by sampling 100 segments from all children [following CITE MEG]. We randomly sampled 100 segments from each child's data, and recalculated the linguistic and canonical proportions based only on these associated judgments. We repeated this process 50 times, to assess the extent to which the association between laboratory- and Zooniverse-derived metrics varied in each random sample. The mean correlations across 50 runs were lower than those recovered using all segments from each child (see Table XX), particularly for the low-risk younger infants, and for linguistic proportion. This may indicate that, particularly in some groups of infants, 100 segments may not be sufficient to capture the child's linguistic and canonical proportions. We also observed quite a bit of variance across the 50 runs. For linguistic proportion, the range of correlations found for all infants was [0.6293868,0.8781453]; for the Angelman Syndrome group [0.6360604,0.937198]; for the low-risk group [0.2175744,0.902997]. For canonical proportion, the range of correlations found for all infants was [0.807683,0.9586492]; for the Angelman Syndrome group [0.559855,0.9663119]; for the low-risk group [0.7473822,0.972202].

```r
kable(mycortab)
```

	Ling. Prop.	Can. Prop.
All seg AS	0.877	0.858
All seg LR	0.823	0.960
All seg all	0.811	0.937
Chunks AS	0.845	0.866
Chunks LR	0.841	0.975
Chunks all	0.845	0.963
100 seg AS	0.851	0.846
100 seg LR	0.764	0.892
100 seg all	0.773	0.907