Peer speech project

Out of the lab corpora: Linaza, Vila, SerraSole, Marrero, Aguirre, OreaPine, Nieva, Ornat, Romero, Geneva, Pauline, Champaud, York, Leveille, Clark, Brown, Kuczaj, Providence, Sachs, Hall, Snow, Valian, Post, Gleason, Suppes, Braunwald, Bloom70, Caroline, Wagner, Rigol, Miller, Leo, Gaeltacht, ZhouDinner, TCCM, Beijing, LeeWongLeung, Utrecht, Wijnen, Stellenbosch, Thomas, Forrester, Wells, MPI-EVA-Manchester, Lara, Belfast, SCECL, Kovacevic, Tanja, Protassova, Antelmi, Calambrone, Klammler, Roma, Florianopolis, Santos, Jordina, Julia, MireiaEvaPascual, Avram, Ishii, Miyata, MiiPro, Hamasaki, Plunkett, Kari, Ringstad, Jiwon, Ryu, Jakarta, Demuth, Soto, Bodor, Reger, Vija, Beek, Korgesaar, Zupping, Argus, Kohler, Narasimhan, Doukas, Levy, BatEl, Ravid, BSF, Samadi, Family

Number of out of the lab corpora: 88

Role tags in all CHILDES : Target_Child, Father, Brother, Mother, Investigator, Adult, Unidentified, Observer, Sister, Child, Girl, Aunt, Playmate, Grandmother, Uncle, Family_Friend, Grandfather, Visitor, Cousin, Boy, Camera_Operator, Babysitter, Teenager, Toy, Environment, Non_Human, Student, Teacher, Sibling, Housekeeper, Media, Doctor, Group, Caretaker, Speaker, Nurse, Target_Adult

```
#finds corpora with peer speech
namePeerSpeech=list()
dataPeer=list()

#3 CHECK: all cousins, boys etc Nnot adults?
childSpeakers=c("Sister", "Brother", "Playmate", "Teenager", "Cousin", "Child", "Girl",
"Sibling", "Boy")

cSelectPeer=subset(cSelect, (cSelect$role %in% childSpeakers))
peerCorpusName=unique(cSelectPeer$corpus_name)
```

Out of lab CHILDES corpora with child speech (tags= Sister, Brother, Playmate, Teenager, Cousin, Child): Linaza, Vila, SerraSole, Marrero, Aguirre, Romero, Geneva, Pauline, Champaud, York, Clark, Brown, Kuczaj, Providence, Sachs, Hall, Valian, Post, Gleason, Suppes, Braunwald, Bloom70, Caroline, Wagner, Rigol, Miller, Leo, Gaeltacht, ZhouDinner, TCCM, Beijing, LeeWongLeung, Stellenbosch, Forrester, Wells, MPI-EVA-Manchester, Lara, Belfast, SCECL, Kovacevic, Calambrone, Santos, Jordina, MireiaEvaPascual, Ishii, Miyata, MiiPro, Hamasaki, Plunkett, Kari, Ryu, Jakarta, Demuth, Soto, Bodor, Reger, Vija, Korgesaar, Zupping, Argus, Kohler, Levy, BatEl, Ravid, BSF, Samadi, Family

Number of CHILDES corpora with peer speech: 67

```
# counts number of utterances per speaker for the selected out-of-lab corpora
#This chunk takes time to compute!
tablep=data.frame()

i=1
for (name in peerCorpusName[1:2]) {   #choosing only first 2 corpora to have it compile faster!
  cp<-get_utterances(corpus=name)
  tabletmp_<-cp %>% group_by(speaker_role) %>% summarise(no_rows = length(speaker_role))
  tablep<-rbind(tablep, tabletmp_)
  i=i+1}
tablep

nuttsSummary<-tablep %>% group_by(speaker_role) %>% summarise(no_rows = sum(no_rows))
nuttsSummary
```

Per corpus Number of utterances per speaker for CHILDES corpora with peer speech: (see table) Total Number of utterances per speaker for CHILDES corpora with peer speech: (see nuttsSummary)

```
# counts number of utterances per speaker for the selected out-of-lab corpora with wireless recordings

tablew=data.frame()

#4 CHECK if only these
wirelessCorpusName<-c("Wells", "Demuth", "Hall")

i=1
for (name in wirelessCorpusName) {
  cw<-get_utterances(corpus=name)
  tabletmp<-cw %>% group_by(speaker_role) %>% summarise(no_rows = length(speaker_role))
  tablew<-rbind(tablew, tabletmp)
  i=i+1}
tablew

nuttsWirelessSummary<-tablew %>% group_by(speaker_role) %>% summarise(no_rows = sum(no_rows))
nuttsWirelessSummary
```

Per corpus Number of utterances per speaker for CHILDES corpora with peer speech AND wireless recordings: (see tablew) Total Number of utterances per speaker for CHILDES corpora with peer speech AND wireless recordings: (see nuttsWirelessSummary)

```
lang<-"Sesotho"

#reads demuth corpus and counts utterances per speaker
demuth<-read.csv(file="/Users/lscpuser/Documents/peerproject/peerproject/sesotho_emilie_
CDI.csv", header=TRUE) # memory processing problems, can't load googlesheets library
sesotho_speakers<- demuth %>% group_by(role_raw) %>% summarise(no_rows = length(role_ra
w))
sesotho_speakers<- sesotho_speakers %>%   arrange(desc(no_rows))
sesotho_speakers<-as.data.frame(sesotho_speakers)
sesotho_speakers_input<-subset(sesotho_speakers, !(sesotho_speakers$role_raw=="Target_Ch
ild")) # speaker category  and n of utts per speaker
sesotho_speakers_input
```

```
##        role_raw no_rows
## 2        Mother   10283
## 3        Cousin    8265
## 4       Brother    5944
## 5  Investigator    5486
## 6   Grandmother    5152
## 7      Playmate    3640
## 8         Adult     996
## 9         Uncle     377
## 10       Sister     257
## 11       Father     180
## 12     Teenager     142
```

```
total_input<-sum(sesotho_speakers_input$no_rows)

#mother input:
sesotho_mother<-subset(sesotho_speakers_input, (sesotho_speakers_input$role_raw=="Mothe
r"))
sesotho_mother$no_rows/total_input
```

```
## [1] 0.2525171
```

```
#siblings input:
sesotho_siblings<-subset(sesotho_speakers_input, (sesotho_speakers_input$role_raw=="Sist
er" |sesotho_speakers_input$role_raw=="Brother" ))
sum(sesotho_siblings$no_rows/total_input)
```

```
## [1] 0.1522764
```

```
#other children input:
sesotho_peers<-subset(sesotho_speakers_input, (sesotho_speakers_input$role_raw=="Cousin"
|sesotho_speakers_input$role_raw=="Playmate" | sesotho_speakers_input$role_raw=="Teenage
r"   ))
sum(sesotho_peers$no_rows/total_input)
```

```
## [1] 0.2958352
```

```
#other adult input:
sesotho_adults<-subset(sesotho_speakers_input, ( sesotho_speakers_input$role_raw=="Grand
mother" | sesotho_speakers_input$role_raw=="Uncle" | sesotho_speakers_input$role_raw=="A
dult" |  sesotho_speakers_input$role_raw=="Father"    ))
sum(sesotho_adults$no_rows/total_input)
```

```
## [1] 0.164653
```

```
#annotated utterances by emilie
annotated<-demuth[1:36782,]
annotated_input<-subset(annotated, !(annotated$role_raw=="Target_Child"))
annotated_input$childdirected<-as.factor(annotated_input$childdirected)
total_annotated_input<-length(annotated_input$utterance_id)
annotated_table<- annotated_input %>% group_by(childdirected) %>% summarise(no_rows = le
ngth(childdirected))
```

```
## Warning: Factor `childdirected` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
annotated_table<-as.data.frame(annotated_table)
annotated_table<- annotated_table %>%   arrange(desc(no_rows))
annotated_table
```

```
##     childdirected no_rows
## 1               1   18013
## 2               ?     614
## 3               A      99
## 4           16434      60
## 5            <NA>      40
## 6           16779      39
## 7             MIX      38
## 8           16502      37
## 9           16455      34
## 10          16711      32
## 11          16631      31
## 12          16713      31
## 13              O      31
## 14          16773      30
## 15          16747      28
## 16          16588      27
## 17           SELF      27
## 18          16399      25
## 19          16460      25
## 20          16509      25
## 21          16775      25
## 22          16494      24
## 23          16551      24
## 24          16575      24
## 25          16730      24
## 26          16778      24
## 27          16784      24
## 28          16501      23
## 29          16478      22
## 30          16525      22
## 31          16401      21
## 32          16473      21
## 33          16486      20
## 34          16610      20
## 35          16422      19
## 36          16491      19
## 37          16569      19
## 38          16737      18
## 39          16437      17
## 40          16632      17
## 41          16673      17
## 42          16821      17
## 43          16405      16
## 44          16435      16
## 45          16710      16
## 46          16714      16
## 47          16731      16
## 48          16740      16
## 49          16453      15
## 50          16620      15
## 51          16677      15
## 52          16791      15
## 53          16429      14
## 54          16449      14
## 55          16472      14
## 56          16480      14
## 57          16584      14
## 58          16668      14
## 59          16748      14
## 60          16469      13
## 61          16511      13
## 62          16617      13
## 63          16822      13
## 64          16428      12
## 65          16451      12
## 66          16465      12
## 67          16524      12
## 68          16540      12
## 69          16573      12
## 70          16702      12
## 71          16709      12
## 72          16741      12
## 73          16745      12
## 74          16786      12
## 75          16818      12
## 76          16414      11
## 77          16443      11
## 78          16481      11
## 79          16500      11
## 80          16574      11
## 81          16623      11
## 82          16661      11
## 83          16676      11
## 84          16756      11
## 85          16856      11
## 86          16415      10
## 87          16461      10
## 88          16520      10
## 89          16612      10
## 90          16772      10
## 91          16820      10
## 92          16823      10
## 93          16431       9
## 94          16448       9
## 95          16518       9
## 96          16666       9
## 97          16671       9
## 98          16724       9
## 99          16781       9
## 100         16816       9
## 101         16433       8
## 102         16444       8
## 103         16459       8
## 104         16470       8
## 105         16586       8
## 106         16627        8
## 107         16646        8
## 108         16663        8
## 109         16670        8
## 110             0        7
## 111         16485        7
## 112         16532        7
## 113         16583        7
## 114         16639        7
## 115         16650        7
## 116         16698        7
## 117         16722        7
## 118         16754        7
## 119         16439        6
## 120         16440        6
## 121         16487        6
## 122         16495        6
## 123         16543        6
## 124         16552        6
## 125         16559        6
## 126         16606        6
## 127         16645        6
## 128         16649        6
## 129         16656        6
## 130         16717        6
## 131         16795        6
## 132         16825        6
## 133         16858        6
## 134         16410        5
## 135         16441        5
## 136         16488        5
## 137         16513        5
## 138         16562        5
## 139         16593        5
## 140         16609        5
## 141         16635        5
## 142         16640        5
## 143         16642        5
## 144         16742        5
## 145         16762        5
## 146         16792        5
## 147         16797        5
## 148         16402        4
## 149         16406        4
## 150         16432        4
## 151         16457        4
## 152         16467        4
## 153         16471        4
## 154         16477        4
## 155         16546        4
## 156         16567        4
## 157         16600        4
## 158         16604        4
## 159         16608       4
## 160         16629       4
## 161         16667       4
## 162         16675       4
## 163         16678       4
## 164         16691       4
## 165         16725       4
## 166         16744       4
## 167         16776       4
## 168         16806       4
## 169         16817       4
## 170         16830       4
## 171         16395       3
## 172         16403       3
## 173         16416       3
## 174         16483       3
## 175         16507       3
## 176         16516       3
## 177         16523       3
## 178         16535       3
## 179         16536       3
## 180         16560       3
## 181         16563       3
## 182         16580       3
## 183         16587       3
## 184         16603       3
## 185         16630       3
## 186         16644       3
## 187         16658       3
## 188         16680       3
## 189         16682       3
## 190         16684       3
## 191         16704       3
## 192         16729       3
## 193         16764       3
## 194         16769       3
## 195         16771       3
## 196         16807       3
## 197         16813       3
## 198         16851       3
## 199         16859       3
## 200     NTSOAKI !       3
## 201         16446       2
## 202         16447       2
## 203         16463       2
## 204         16466       2
## 205         16474       2
## 206         16490       2
## 207         16492       2
## 208         16510       2
## 209         16522       2
## 210         16547       2
## 211         16565       2
## 212         16572       2
## 213         16614       2
## 214         16648       2
## 215         16662       2
## 216         16665       2
## 217         16706       2
## 218         16716       2
## 219         16718       2
## 220         16726       2
## 221         16743       2
## 222         16758       2
## 223         16765       2
## 224         16766       2
## 225         16767       2
## 226         16808       2
## 227         16839       2
## 228         16849       2
## 229         16850       2
## 230         16854       2
## 231         17792       2
## 232             C       2
## 233         11678       1
## 234         16393       1
## 235         16404       1
## 236         16413       1
## 237         16417       1
## 238         16419       1
## 239         16436       1
## 240         16442       1
## 241         16452       1
## 242         16493       1
## 243         16504       1
## 244         16506       1
## 245         16515       1
## 246         16530       1
## 247         16544       1
## 248         16556       1
## 249         16566       1
## 250         16570       1
## 251         16579       1
## 252         16596       1
## 253         16598       1
## 254         16615       1
## 255         16616       1
## 256         16618       1
## 257         16622       1
## 258         16628       1
## 259         16636       1
## 260         16637       1
## 261         16638       1
## 262         16672       1
## 263         16692       1
## 264         16696       1
## 265         16701       1
## 266         16723       1
## 267         16728       1
## 268         16752       1
## 269         16759       1
## 270         16760       1
## 271         16770       1
## 272         16783       1
## 273         16787       1
## 274         16800       1
## 275         16802       1
## 276         16824       1
## 277         16827       1
## 278         16832       1
## 279         16840       1
## 280         16843       1
## 281         16844       1
## 282         16857       1
## 283         16868       1
## 284           DOG       1
```

```
#target  child directed utts
child_directed<-subset(annotated_input, (annotated_input$childdirected=="1"))
cddirected_utts<- as.data.frame(child_directed %>%select(utterance))
#write.table(cddirected_utts, file=paste0("~/Documents/peerproject/peerproject", lang,
 "childdirected"), row.names=F, col.names=T,  quote=F)
length(child_directed$utterance)/total_annotated_input #percentage of child directed spe
ech vs total annotated
```

```
## [1] 0.8560498
```

```
#matches addressee with speaker role (especially for non-child directed)
speaker_info<- as.data.frame(unique(demuth %>%select(speaker_id, role_raw)))
colnames(speaker_info)[colnames(speaker_info)=="role_raw"] <- "role_adressee"
annotated_speaker_info<-merge(x=annotated_input, y=speaker_info, by.x="childdirected", b
y.y="speaker_id", all.x=TRUE, sort=TRUE)

#adult directed utts
adult_directed<-subset(annotated_speaker_info, !(annotated_speaker_info$role_adressee==
"Playmate"| annotated_speaker_info$role_adressee=="Cousin"|annotated_speaker_info$role_a
dressee=="?"| annotated_speaker_info$childdirected=="SELF"| annotated_speaker_info$role_
adressee=="SELF"| annotated_speaker_info$role_adressee=="Brother"| annotated_speaker_inf
o$childdirected=="NA"| annotated_speaker_info$role_adressee=="Teenager"|annotated_speake
r_info$role_adressee=="Sister"|annotated_speaker_info$childdirected=="1"|annotated_speak
er_info$childdirected=="0"|annotated_speaker_info$childdirected=="O"))
addirected_utts<- as.data.frame(adult_directed %>%select(utterance))
write.table(addirected_utts, file=paste0("~/Documents/peerproject/peerproject", "Sesotho
adultdirected"), row.names=F, col.names=T,  quote=F)
length(adult_directed$utterance)/total_annotated_input #percentage of child directed spe
ech vs total annotated
```

```
## [1] 0.04148845
```

```
#na directed
na_directed<-subset(annotated_input, (annotated_input$childdirected=="NA" |annotated_inp
ut$childdirected=="?"  ))
nadirected_utts<- as.data.frame(na_directed %>%select(utterance))
#write.table(cddirected_utts, file=paste0("~/Documents/peerproject/peerproject", lang,
 "childdirected"), row.names=F, col.names=T,  quote=F)
length(na_directed$utterance)/total_annotated_input #percentage of child directed speech
vs total annotated
```

```
## [1] 0.02917974
```

```
#Sentence  type
sentence_type_child_annotated<- child_directed %>% group_by(sentence_type) %>% summarise
(no_rows = length(sentence_type))
directed_questions<-subset(sentence_type_child_annotated, (sentence_type_child_annotated
$sentence_type=="question"))
directed_questions$no_rows/sum(sentence_type_child_annotated$no_rows)
```

```
## [1] 0.4012102
```

```
sentence_type_adult_annotated<- adult_directed %>% group_by(sentence_type) %>% summarise
(no_rows = length(sentence_type))
adultdirected_questions<-subset(sentence_type_adult_annotated, (sentence_type_adult_anno
tated$sentence_type=="question"))
adultdirected_questions$no_rows/sum(sentence_type_adult_annotated$no_rows)
```

```
## [1] 0.2038946
```

```
#WELLS

lang<-"English"
wells<-read.csv(file="/Users/lscpuser/Documents/peerproject/ongoingwellsannotation/total
2.csv", header=TRUE)
wells_input<- wells[!grepl("Target", wells$ROLE),] #remove target child utts

#number of utterances per speaker
wells_input_speakers<- wells_input %>% group_by(ROLE) %>% summarise(no_rows = length(ROL
E))
wells_input_speakers<- wells_input_speakers %>%   arrange(desc(no_rows))
wells_input_speakers<- as.data.frame(wells_input_speakers )
wells_input_speakers  #number of utterances per speaker
```

```
##                      ROLE no_rows
## 1                  Mother    3667
## 2          Nicola  Sister     616
## 3           Rachel Sister     455
## 4         Richard Sibling     264
## 5          Rebecca Sister     247
## 6                  Father     221
## 7           Sarah  Sister     218
## 8           Louise Sister     194
## 9            Unidentified     179
## 10       Jonathan Brother     178
## 11           Lorna Sister     115
## 12         Adrian Sibling     109
## 13                 Sister      62
## 14         Christine Aunt      59
## 15            Hazel Child      54
## 16          Carol Visitor      39
## 17                  Child      31
## 18       Catherine  Child      30
## 19                  Adult      28
## 20           Claire Child      24
## 21    Helen Family_Friend      23
## 22           Kerry  Child      23
## 23             Tina Child      23
## 24                Visitor      23
## 25         Neighbor Adult      22
## 26            Naomi Child      19
## 27             Lee  Child      14
## 28                   Aunt      13
## 29           Kelly  Child      12
## 30          Nicole Sister      12
## 31           Sirka  Adult      12
## 32    Erika Family_Friend       8
## 33             Dean child       7
## 34         Isabelle Child       6
## 35                              5
## 36   Television Non_Human       5
## 37          Unidentified%       5
## 38      Lorraine Playmate       3
## 39  Suzanne Family_Friend       3
## 40                  Uncle       3
## 41            Grandmother       2
## 42            Dale  Child       1
## 43        Rachel Playmate       1
## 44          TVMan Visitor       1
```

```
total_winput<-sum(wells_input_speakers$no_rows) #number  of total input utterances

#mother input:
wells_mother<-subset(wells_input_speakers, (wells_input_speakers$ROLE=="Mother")) #numbe
r of utterances by mother
wells_mother$no_rows/total_winput # percentage of utterances by mother in total input
```

```
## [1] 0.5211768
```

```
#siblings input:
wells_siblings<- wells_input_speakers[grep("Sister|Brother|Sibling", wells_input_speaker
s$ROLE),]
sum(wells_siblings$no_rows)/total_winput
```

```
## [1] 0.3510517
```

```
#other children input:
wells_chi<- wells_input_speakers[grep("Child|Playmate", wells_input_speakers$ROLE),]
sum(wells_chi$no_rows)/total_winput
```

```
## [1] 0.03425242
```

```
#other adults input:
wells_adu<- wells_input_speakers[grep("Adult|Uncle|Grandmother|Family_Friend|Visitor|Aun
t|Father",wells_input_speakers$ROLE),]
sum(wells_adu$no_rows)/total_winput
```

```
## [1] 0.06495168
```

```
#ADRESSEE ANNOTATIONS PART
wells_annotated_input<-subset(wells_input, !(wells_input$DIRECTED=="")) #select utteranc
es annotated by Naomi Alex up to now
wells_annotated_value<- wells_annotated_input %>% group_by(DIRECTED) %>% summarise(no_ro
ws = length(DIRECTED))
wells_annotated_value<-as.data.frame(wells_annotated_value)
wells_total_annotated_input<-sum(wells_annotated_value$no_rows) #annotated adressees and
n of utts
wells_annotated_value
```

```
##     DIRECTED no_rows
## 1          ?     117
## 2          A       6
## 3          C       6
## 4        CAR      88
## 5        CHI     928
## 6        CHR       4
## 7        ERI       4
## 8        FAT      32
## 9        HEL       5
## 10       HMO       3
## 11       LOU      75
## 12       MIX       1
## 13       MOT     260
## 14       NIC      31
## 15         O       8
## 16       PET      22
## 17       RAC     226
## 18       REB     111
## 19      SELF      46
## 20       SIR      40
## 21       SUZ       7
## 22 TELEPHONE      29
## 23       TVM       1
## 24       VIS       9
```

```
#target  child directed
wells_annotated_CHI<-subset(wells_annotated_input, (wells_annotated_input$DIRECTED=="CHI")) #wells annotated child-directed corpus
length(wells_annotated_CHI$UTTERANCE)/ length(wells_annotated_input$UTTERANCE)    # n of child-directed utterances
```

```
## [1] 0.4507042
```

```
cddirected_utts<- as.data.frame(wells_annotated_CHI %>%select(UTTERANCE))
write.table(cddirected_utts, file=paste0("~/Documents/peerproject/", lang,"Wellschilddirected.txt"), row.names=F, col.names=T,  quote=F)


#adult directed
adult_directed<-subset(wells_annotated_input, !(wells_annotated_input$DIRECTED=="CHI" |wells_annotated_input$DIRECTED=="0" |wells_annotated_input$DIRECTED=="O" | wells_annotated_input$DIRECTED=="?" | wells_annotated_input$DIRECTED=="SELF"| wells_annotated_input$DIRECTED=="NIC"  | wells_annotated_input$DIRECTED=="TELEPHONE" | wells_annotated_input$DIRECTED=="MIX" | wells_annotated_input$DIRECTED=="LOU" | wells_annotated_input$DIRECTED== "REB" | wells_annotated_input$DIRECTED=="PET"| wells_annotated_input$DIRECTED=="RAC" | wells_annotated_input$DIRECTED=="CHR" | wells_annotated_input$DIRECTED=="NA"))
addirected_utts<- as.data.frame(adult_directed %>%select(UTTERANCE))
write.table(addirected_utts, file=paste0("~/Documents/peerproject/", lang, "Wellsadultdirected.txt"), row.names=F, col.names=T,  quote=F)

length(addirected_utts$UTTERANCE)/ length(wells_annotated_input$UTTERANCE)    # n of child-directed utterances
```

```
## [1] 0.2238951
```

```
#NA annotations
wells_annotated_na <- subset(wells_annotated_input,(wells_annotated_input$DIRECTED=="?"|wells_annotated_input$DIRECTED=="NA"|wells_annotated_input$DIRECTED==""|wells_annotated_input$DIRECTED==" "))
length(wells_annotated_na$UTTERANCE)/ length(wells_annotated_input$UTTERANCE)
```

```
## [1] 0.0568237
```