Optimal model of SPE by learners of monolingual and contact Spanish

#IMPORTANT: DO NOT USE plyr library. It will interfere with dplyr group by and summarise operations in code chunks 'priors' and 'likelihoods' below (even if plyr is called first).  
library(tidyverse)

This is a model of how learners calculate the posterior P(pronoun|reference) i.e., the probability of realizing an overt subject personal pronoun, given whether that pronoun is same-reference or switch-reference. This model assumes an OPTIMAL learner who accurately relates the posterior to P(reference|pronoun) and P(pronoun) i.e., the likelihood of an overt pronoun being same- or switch-reference and the prior probability of using an overt pronoun, both of which are learned from the input.

We will look at two populations: (i) WC children exposed to Mexico City Spanish, and (ii) WC children raised in Villa21 Buenos Aires, whoare exposed to both Paraguayan Spanish (parents) and Rioplatense Spanish (teachers, daycare workers, etc.– approximated here by the investigators’ speech).

# Assumptions about the input

1. Children learn SPE from

* animate subjects only
* pronouns only [TODO: check if the priors & likelihoods differ much if the contrast is between any overt animate (pronoun, DP, name, etc. =1) versus null animate (null pronoun =0)–inanimates are not coded for key characteristics like reference]
* NOT usted(es) [TODO: check if the priors & likelihoods differ much for usted(es) = 2s/p and usted(es) = 3s/p]
* within-turn reference chains (i.e., reference\_turn is defined) [TODO: check if the priors differs for reference\_turn undefined]

#put Mexico City and Buenos Aires together into a single aduchi dataset and format its variables according to the above assumptions.  
aduchi <- read\_csv("aduchi.csv")  
  
aduchi <- aduchi %>%   
 select(community, file, participant, dyad, SES, stem, animacy, nullovert, reference\_turn, person, number, tma, yr, mo) %>%  
 filter(  
 animacy == 1 &   
 nullovert %in% c(0, 1) &   
 reference\_turn %in% c("same", "switch") &  
 #exclude usted(es), following Shin 2016  
 person %in% c("1", "2", "3") &   
 number %in% c("s", "p", "S", "P") &  
 tma %in% c("cond", "fut=", "pas", "pres", "PRES", "pret", "PRET", "sub&pres")  
 ) %>% mutate(  
 #make the outcome variable numeric  
 SPE = as.numeric(paste(nullovert)),  
 #rename the reference variable  
 ref = reference\_turn,  
 #create personnum factor following Shin 2016  
 personnum = as.factor(paste0(person, tolower(number))),  
 #create TMA factor following Shin 2016  
 tma = recode(tma, "pres"="present", "PRES"="present","pret"="preterite", "PRET"="preterite", "pas"="imperfect", "sub&pres"="other", "cond"="other", "fut="="other")  
 ) %>% filter(  
 #filter out a coding mistake  
 personnum != "2p"  
 ) %>% mutate(  
 #relevel personnum and tma  
 personnum = fct\_relevel(personnum, "1s", "2s", "3s", "1p", "3p"),  
 tma = fct\_relevel(tma, "other", "imperfect", "preterite", "present")  
 ) %>% mutate(  
 #Add child ages and age groups.   
 child\_age\_mo = 12\*yr + mo,  
 child\_age\_yr = child\_age\_mo / 12,  
 #Note: use 2 age groups - must be split at 56 to prevent missing values  
 child\_age\_group = cut(child\_age\_mo, breaks = c(-Inf, 56, Inf), labels=c("younger", "older"), right = FALSE)  
 ) %>% mutate(  
 #split primary dataset into input and output  
 inputoutput = ifelse(participant == "CHI", "output", "input")  
 )

1. Children in Villa21 learn a single grammar from PS and RpS input combined. The degree to which each input type contributes to a child’s grammar is proportional to how much that child is exposed to each type. For the moment, we will assume that the proportions are roughly equal to the proportions in this dataset. [TODO: If you think children learn only from PS or only from RpS input, calculate priors and likelihoods separately for PS and RpS input and test child output against each one. If you think they learn a single grammar from both, but the proportions are different, then vary how much each separate value of likelihoods and priors contributes to the model for Villa21 children.]

## # A tibble: 2 x 3  
## speech\_type tokens proportion  
## <chr> <int> <dbl>  
## 1 ps\_input 2357 0.611  
## 2 rps\_input 1498 0.389

# The model

This model assumes accurate calculation of the posterior: from the likelihood and the prior .

## Priors

Priors are learned accurately from the input that the child receives. The last column is the prior probability of null (SPE=0) and overt (SPE=1) SPE in each community. For Villa21 Buenos Aires, Rioplatense input has a lower prior compared to Paraguayan input. Thus, children with more Rioplatense input will acquire a lower prior than children with less Rioplatense input. The prior reported below for Villa21 Buenos Aires reflects the roughly 40/60 Rioplatense/Paraguayan input profile in our sample, which we have assumed is representative of children in this community.

#Form a tibble with the prior of each form in each community  
priors <- aduchi %>%   
 filter(inputoutput == "input") %>%  
 group\_by(community, SPE) %>%  
dplyr::summarise(n = n()) %>%  
 mutate(  
 freq = n/sum(n)  
 )  
  
print(priors)

## # A tibble: 4 x 4  
## # Groups: community [2]  
## community SPE n freq  
## <chr> <dbl> <int> <dbl>  
## 1 Buenos\_Aires 0 3352 0.870  
## 2 Buenos\_Aires 1 503 0.130  
## 3 Mexico\_City 0 2589 0.890  
## 4 Mexico\_City 1 319 0.110

## Likelihoods

Likelihoods are also learned accurately from the input. The last column is the likelihood of each reference context, given an overt or null pronoun - calculated for each community. For Villa21, more Rioplatense input will not change the likelihoods, since we have found in previous work that both dialects condition pronoun realization on reference *to the same degree*.

ref\_likelihoods <- aduchi %>%   
 filter(inputoutput == "input") %>%  
 group\_by(community, SPE, ref) %>%  
 dplyr::summarise(  
 n = n()  
 ) %>%  
 mutate(  
 freq = n/sum(n)  
 )  
print(ref\_likelihoods)

## # A tibble: 8 x 5  
## # Groups: community, SPE [4]  
## community SPE ref n freq  
## <chr> <dbl> <chr> <int> <dbl>  
## 1 Buenos\_Aires 0 same 1530 0.456  
## 2 Buenos\_Aires 0 switch 1822 0.544  
## 3 Buenos\_Aires 1 same 130 0.258  
## 4 Buenos\_Aires 1 switch 373 0.742  
## 5 Mexico\_City 0 same 1149 0.444  
## 6 Mexico\_City 0 switch 1440 0.556  
## 7 Mexico\_City 1 same 68 0.213  
## 8 Mexico\_City 1 switch 251 0.787

## Posteriors

Posteriors are calculated accurately from the priors and likelihoods. Namely, . The function below returns the posterior , given a community and a reference context. The posterior for that context would simply be .

#This is the function for calculating a posterior from the 'priors' table and 'ref\_likelihoods' tibbles above.   
  
optimal <- function(c, r) {  
 #all the priors necessary to compute the posterior  
 prior\_o <- priors$freq[priors$community == c & priors$SPE == 1]  
 prior\_n <- priors$freq[priors$community == c & priors$SPE == 0]  
   
 #all the likelihoods necessary to compute the posterior  
 lik\_ref\_o <- ref\_likelihoods$freq[ref\_likelihoods$community == c & ref\_likelihoods$ref == r & ref\_likelihoods$SPE == 1]  
 lik\_ref\_n <- ref\_likelihoods$freq[ref\_likelihoods$community == c & ref\_likelihoods$ref == r & ref\_likelihoods$SPE == 0]  
   
 posterior <-   
 (lik\_ref\_o \* prior\_o)/  
 (lik\_ref\_o \* prior\_o +  
 lik\_ref\_n \* prior\_n)  
 posterior  
}

# Predicted versus observed proportion overt

The table below shows the observed rate of overt SPE in each reference context for each community, as compared to the predicted probability of overt SPE in the optimal model, and the error (observed - predicted).

observed <- aduchi %>%   
 filter(inputoutput == "output") %>%  
 group\_by(  
 community, ref  
 ) %>%  
 summarise(  
 overt = sum(SPE),  
 tokens = n(),  
 observed = overt/tokens  
 )   
predictions <- observed %>% rowwise() %>% mutate(  
 predicted = optimal(community, ref),  
 error = observed - predicted  
 )  
  
print(predictions)

## Source: local data frame [4 x 7]  
## Groups: <by row>  
##   
## # A tibble: 4 x 7  
## community ref overt tokens observed predicted error  
## <chr> <chr> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 Buenos\_Aires same 83 683 0.122 0.0783 0.0432  
## 2 Buenos\_Aires switch 128 598 0.214 0.170 0.0441  
## 3 Mexico\_City same 44 603 0.0730 0.0559 0.0171  
## 4 Mexico\_City switch 115 720 0.160 0.148 0.0113

# Model fit

The mean squared error of the optimal model, across same and switch-referenc contexts is below. It appears that the optimal model makes slightly better predictions overall for Mexico City kids (MSE= 0.0002) than for Villa21 Buenos Aires kids (MSE = 0.0019).

fit <- predictions %>% ungroup() %>%  
 group\_by(community) %>%  
 summarise(  
 MSE = mean(error\*error)  
 )  
print(fit)

## # A tibble: 2 x 2  
## community MSE  
## <chr> <dbl>  
## 1 Buenos\_Aires 0.00191   
## 2 Mexico\_City 0.000210

A second way to evaluate how well the model predicts children’s behavior is its log likelihood given the observed data. A model’s likelihodd given the data is calculated by first assuming that the model is true and then calculating the probability of observing the data that we have observed in our sample. If the model actually is true (or at least, a better hypothesis than other models), this probability should be high, but if it is untrue (or at least, a worse hypothesis than other models) the probability should be low.

Assuming that the optimal model is true, the probability of observing overt pronouns and null pronouns in any given context is equal to . For example, in same-reference contexts, the predicted probability of overt and null pronouns for Mexico City kids is and , respectively, and the observed count of overt and null pronouns is and , respectively. Thus, the probability of the observed data is

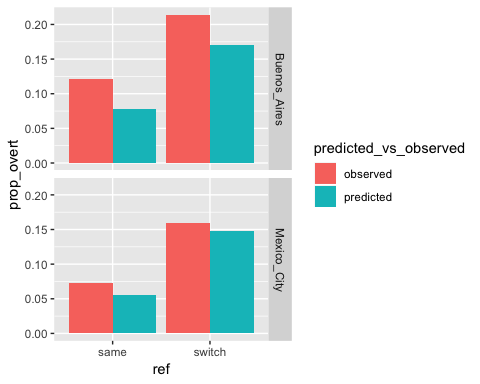
To get the probability of the whole dataset, we then calculate in switch-reference environments and multiply the two together. Since these probabilities get very small, especially for datasets with many observations, we typically take the logarithm.

loglik <- predictions %>%   
 transmute(  
 community = community,  
 ref = ref,  
 N\_overt = overt,  
 N\_null = tokens - overt,  
 predicted\_overt = predicted,  
 predicted\_null = 1 - predicted,  
 loglik = log(predicted\_overt)\*N\_overt + log(predicted\_null)\*N\_null  
 ) %>%   
 group\_by(community) %>%  
 summarise(  
 loglik = sum(loglik)  
 )  
  
print(loglik)

## # A tibble: 2 x 2  
## community loglik  
## <chr> <dbl>  
## 1 Buenos\_Aires -575.  
## 2 Mexico\_City -476.

Finally, let’s visualize the observed and predicted rates of overt SPE, to see where our model is over- or under-predicting pronoun rates.

obs\_v\_opt <- predictions %>%  
 select(-c(error, overt, tokens)) %>%  
 gather(key = "predicted\_vs\_observed", value = "prop\_overt", observed, predicted) %>%  
 arrange(community, ref)  
  
ggplot(obs\_v\_opt,  
 aes(x = ref, y = prop\_overt, fill = predicted\_vs\_observed)) +  
 facet\_grid(community~.) +  
 geom\_col(position = "dodge")



This model seems to be a better fit for the Mexico City cohort, with a lower MSE, a higher (=less negative) log likelihood and a visually better fit to the data in both same- and swith-reference contexts. Specifically, while the model under-predicts the frequency of overt SPE in both Mexico City and Villa 21, Buenos Aires, it is more severely under the mark for the latter group.