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
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 accord
aduchi <- read_csv("aduchi.csv")

aduchi <- aduchi %>%
  select(community, file, participant, dyad, SES, stem, animacy, nullovert, reference_turn, person, numb
  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="="oth
  ) %>% 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 =
  ) %>% mutate(
    #split primary dataset into input and output
    inputoutput = ifelse(participant == "CHI", "output", "input")
  )
```

2. 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: $P(overt|reference)$ from the likelihood $P(reference|overt)$ and the prior $P(overt)$.

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 $P(overt)$ compared to Paraguayan input. Thus, children with more Rioplatense input will acquire a lower prior $P(overt)$ 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, $P(overt|reference) = \frac{P(reference|overt) \times P(overt)}{\sum_{pro=\{null,overt\}} P(reference|pro) \times P(pro)}$. The function below returns the posterior $P(overt)$, given a community and a reference context. The posterior $P(null)$ for that context would simply be $1 - P(overt)$.

```
#This is the function for calculating a posterior from the 'priors' table and 'ref_likelihoods' tibbles

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$ref == r & ref_lik
  lik_ref_n <- ref_likelihoods$freq[ref_likelihoods$community == c & ref_likelihoods$ref == r & ref_lik

    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 N_{overt} overt pronouns and N_{null} null pronouns in any given context is equal to $P(N_{overt}|P_{predicted}(overt)) \times P(N_{null}|P_{predicted}(null))$. For example, in same-reference contexts, the predicted probability of overt and null pronouns for Mexico City kids is $P_{predicted}(overt) = 0.056$ and $P_{predicted}(null) = 1 - 0.056 = 0.944$, respectively, and the observed count of overt and null pronouns is $N_{overt} = 44$ and $N_{null} = 603 - 44 = 559$, respectively. Thus, the probability of the observed data is

$P(N_{overt} = 44|P_{predicted}(overt) = .056) \times P(N_{null} = 559|P_{predicted}(null) = .944) = 0.056^{44} \times 0.944^{559} = 8.504 * 10^{-70}$

To get the probability of the whole dataset, we then calculate $P(N_{overt}|P_{predicted}(overt)) \times P(N_{null}|P_{predicted}(null))$ 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.

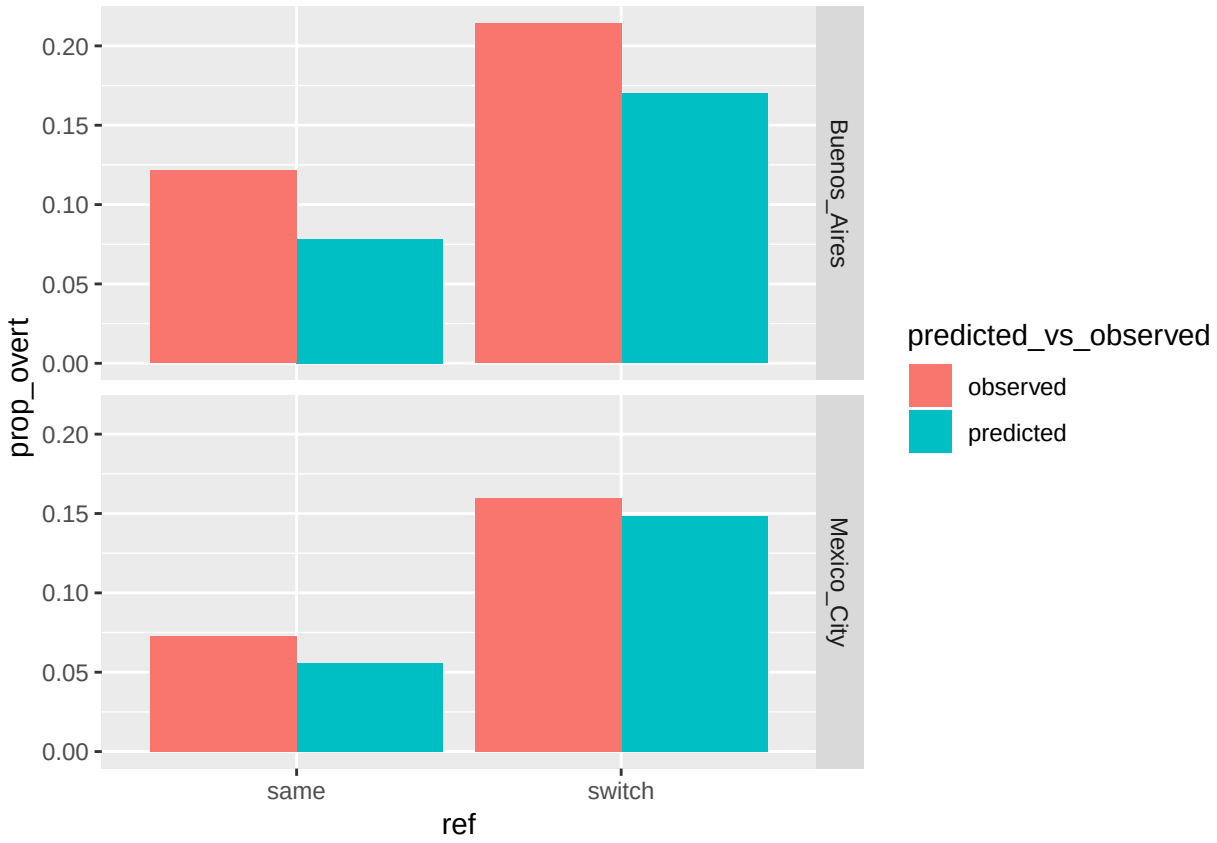
```
loglik <- predictions %>%
  transmute(
    community = community,
    ref = ref,
    N_overt = overt,
    N_null = tokens - overt,
    predicted_overt = predicted,
```

```
    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.