Final Report

To: Lucy Grace, Dr. Katie Carmichael, Dr. Stephanie Pickle DeHart, and Dr. Jennifer Van Mullekom

From: Lanxin Xiang and Claudia Clinchard
Date: April 21, 2025

Subject: Summary of the analysis on pronoun dropping amongst bilingual speakers residing in Miami

Purpose: The purpose of this memo is to provide a summary of the project, the analysis conducted, the results, and future recommendations.

Introduction

Background: Bilingualism is prominent in the United States across generations. When speaking Spanish, some individuals use a subject pronoun before a verb ("Yo estoy hambre", or "I am hungry") whereas others do not ("Estoy hambre"). The presence of the subject pronoun does not affect the meaning of the sentence and is not required to be used, unlike in English. Previous research (Erker & Otheguy, 2020) showed a slight increase in subject pronoun usage when speaking Spanish in individuals born in the United States compared to native Spanish speakers born outside of the United States. The current project aims to explore whether the age of bilingual speakers is associated with the likelihood of the subject pronoun being dropped. It was expected that younger speakers would use pronouns less frequently than older speakers.

Research Question: How does age affect the presence of pronouns in the Spanish language for bilingual speakers?

Statistical Question: Is there any relationship between age and proportion of pronouns used?

Data

Data Source: All data was taken from the website Bangortalk-Miami[1].

Study Design: Stratified sampling was employed to ensure a balanced representation across demographic groups. A total of 20 individuals were sampled from the website Bangortalk – Miami (Erker & Otheguy, 2020), comprising 5 individuals from each of the following categories: young (< 35-year-old) females, young (< 35-year-old) males, older (≥35-year-old) females, and older (≥35-year-old) males. Token data were then extracted from the recorded conversations according to the following guidelines:

[1] Data is available from https://bangortalk.org.uk/speakers.php?c=miami and was pulled on March 19, 2025.

1. Only conversations involving two people were taken into consideration.
2. Only one speaker's tokens from each conversation were transcribed. The speaker was selected based on the number of tokens spoken (i.e., if there were at least 20 Spanish tokens in the conversation).
3. 20 tokens from each person were recorded. The verb used in the token, if a subject pronoun was used during the token, and if Spanish or English was used before the token were all recorded along with the speaker's age and gender.

Data Dictionary: The data is secondary data, and the information was coded from the audio recordings. The variables and descriptive statistics are presented in **Table 1.**

Table 1. Data Dictionary.

Variable	**Information**	**Level (subject vs. token)**	**Frequencies or Descriptive Statistics**
Token	Categorical; Indicates the token number	Token	1-20 for each speaker
ID	Categorical; Indicates the conversation ID	Subject	N = 20
Speaker	Categorical; Indicates the speaker	Subject	N = 20
Age (IV)	Continuous; Indicates the age of the speaker	Subject	Range: 11-78 years; Mean = 38.60; Median = 36.00; SD = 18.91; Quantile 1 = 22.75; Quantile 3 = 53.25;
AC (IV)	Categorical (binary "younger" versus "older"); indicates if the speaker is in the younger or older group (35 or older)	Subject	Younger: N = 10 Older: N= 10
Gender (IV)	Categorical (binary "male" versus "female"); indicates if the speaker is male or female	Subject	Female: N = 10 Male: N = 10
Pronoun (DV)	Categorical (binary "yes" versus "no"); indicates if the pronoun was used	Token	"0" (no): N = 309 "1" (yes): N = 91
CONJ (IV)	Categorical (6 categories: first person singular, second person singular, third person singular, first person plural, second person plural, or third person plural); indicates the conjugation of the verb used	Token	1p (1st person plural): 14 1s (1st person plural): 96 2p (2nd person plural): 0 2s (2nd person singular): 42 3p (3rd person plural): 56 3s (3rd person singular):192
Verb (IV)	Categorical; indicates the verb spoken in the token	Token	88 levels

Variable	Information	Level (subject vs. token)	Frequencies or Descriptive Statistics
Before (IV)	Categorical (binary "Spanish" or "English"); indicates if Spanish or English was spoken before the token	Token	Spanish = 365 English = 31 NA = 31
Sentence (IV)	The whole token/sentence that was spoken.	Token	394 levels

Note. IV = Independent Variable; DV = Dependent Variable.

Data Issues and/or Preparations:

Data cleaning was performed prior to analysis, with two main modifications based on data inspection:

1. According to our client input, the "NA" values in the variable "Before" represented tokens that appeared at the start of the sentence. To ensure the dataset could be fully utilized, all "NA" entries in this variable were recoded as "Start."
2. During the inspection, we observed that one category (1p) of the "Conjugation" variable was perfectly predictive of pronoun drop, and one category (2p) of "Conjugation" was absent from the dataset (which is discussed further in the Exploratory Data Analysis section). This missing information could cause numerical issues during the model fitting. To address this, the Conjugation variable was split into two components: person-tense (1^{st}, 2^{nd}, or 3^{rd}) and number (singular or plural).

The links to all code and the information is provided in Appendix D.

Exploratory Data Analysis

Exploratory data analysis was performed to examine the counts of the different ages (**Figure 1**) and the percentage of time in which pronouns were used across ages (**Figure 2**), gender (**Figure 3**), language used before (**Figure 4**), conjugation (**Figure 5**), whether first person, second person, or third person conjugations were used (**Figure 6**), and whether singular or third person conjugations were used (**Figure 7**). Of note, there was one potential outlier that can be seen in **Figure 2**. While the majority of individuals under the age of 40 did not use pronouns more than 20% of the time, this 12-year-old used pronouns 35% of the time. **Figure 8** presents the percentage of times that pronouns were used per individual speaker by descending age. **Figure 9** indicates shows the percentage of time that pronouns were used by both age and gender. The distribution of age and the relationship between age and percentage of time pronouns were used are shown in **Figure 1** and **Figure 2.** Pronouns tended to be used more frequently with older individuals, although the 12-year-old used pronouns more frequently (35% of the time) than others around their age. This pattern indicates that older bilingual speakers may be more likely to include subject pronouns.

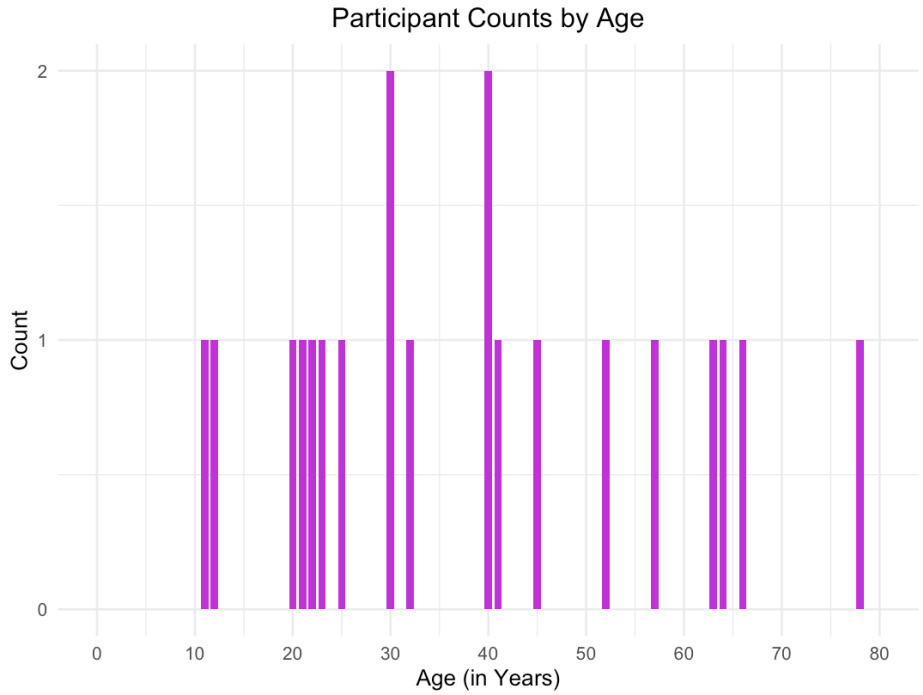

Figure 1. Count of ages in sample.

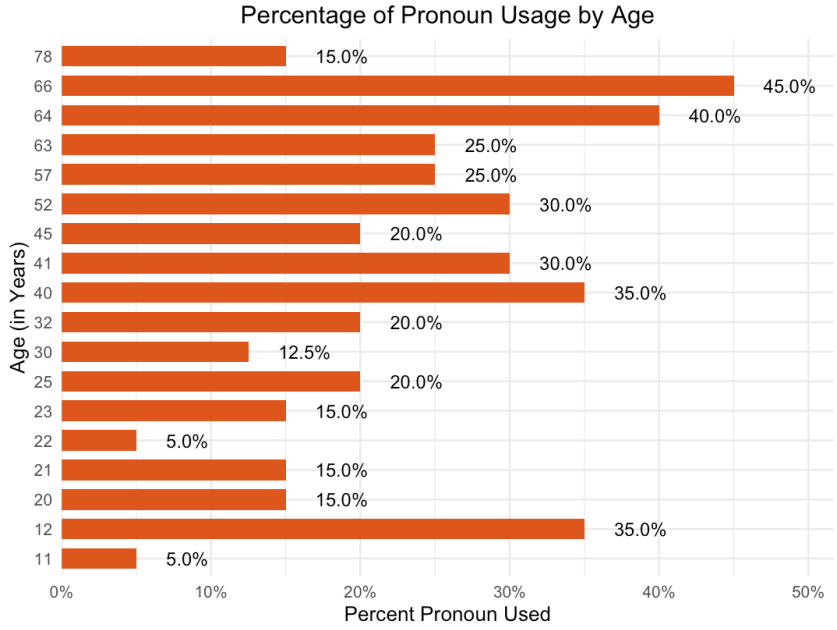

Figure 2. Average percentage of time pronouns were used by age. There were two 30-year-olds and two 40-year-olds, and the presented percentages were averaged across the individuals of the same age.

As shown in **Figure 3**, when examining gender independently, gender does not appear to have a meaningful effect on pronoun usage. The proportion of pronoun use is nearly identical between female (23.0%) and male (22.5%) speakers.

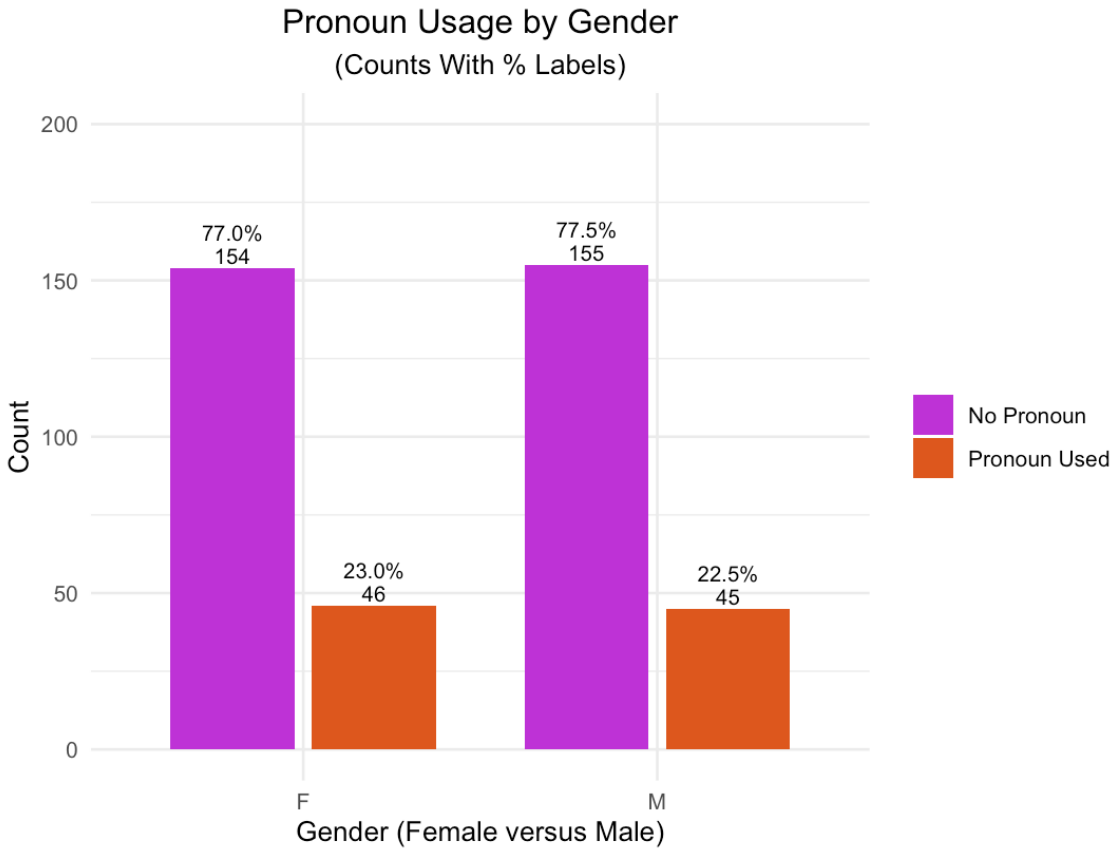

Figure 3. Count and percentage of time pronouns were used by gender. F represents female and M represents Male.

Figure 4 suggests pronouns were used more frequently when English was used in the preceding sentence. For tokens occurring at the start of a conversation, pronouns were used 25.0% of the time; however, this estimate is based on only four cases, so it should be interpreted with caution.

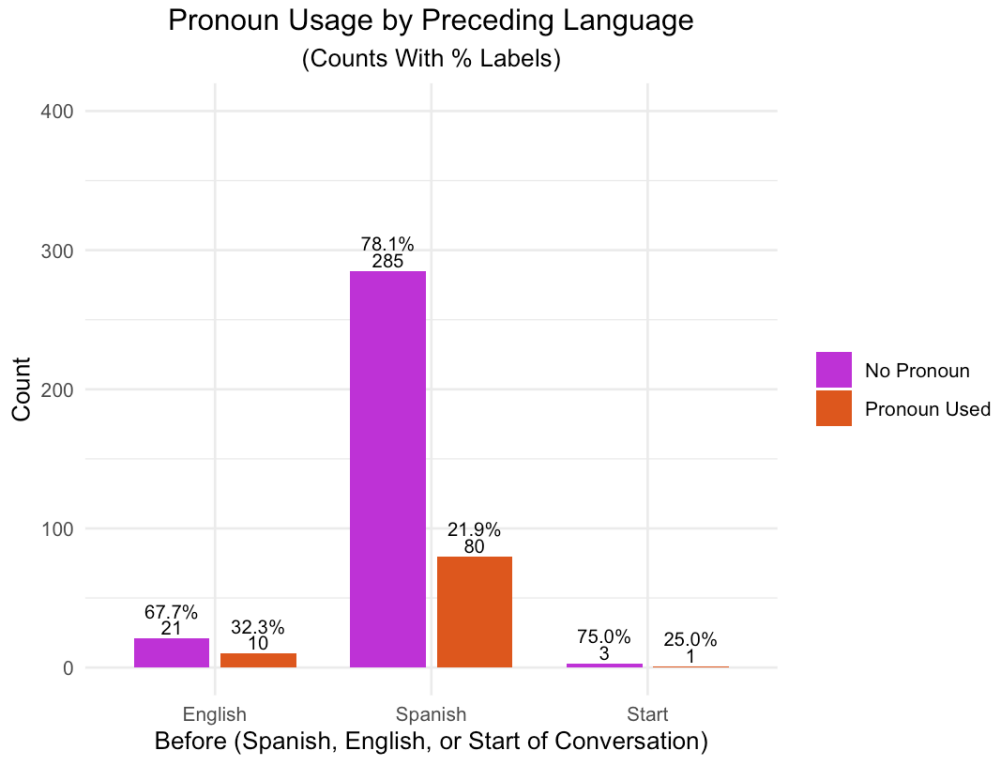

Figure 4. Count and percentage of time pronouns were used by language used before the token.

Figure 5 shows the proportion of subject pronoun usage across different verb conjugation categories. The most frequently used conjugation was third person-singular (3s), accounting for 192 observations, where pronouns were used 14.6% of the time. In contrast, second-person singular (2s) conjugations showed the highest rate of pronoun use, with 50% of tokens including a pronoun, despite a smaller sample size (N = 42). First-person singular (1s) also had a relatively high rate of pronoun usage at 39.6% (N = 96). First-person plural (1p) showed 0% pronoun use across 14 tokens.

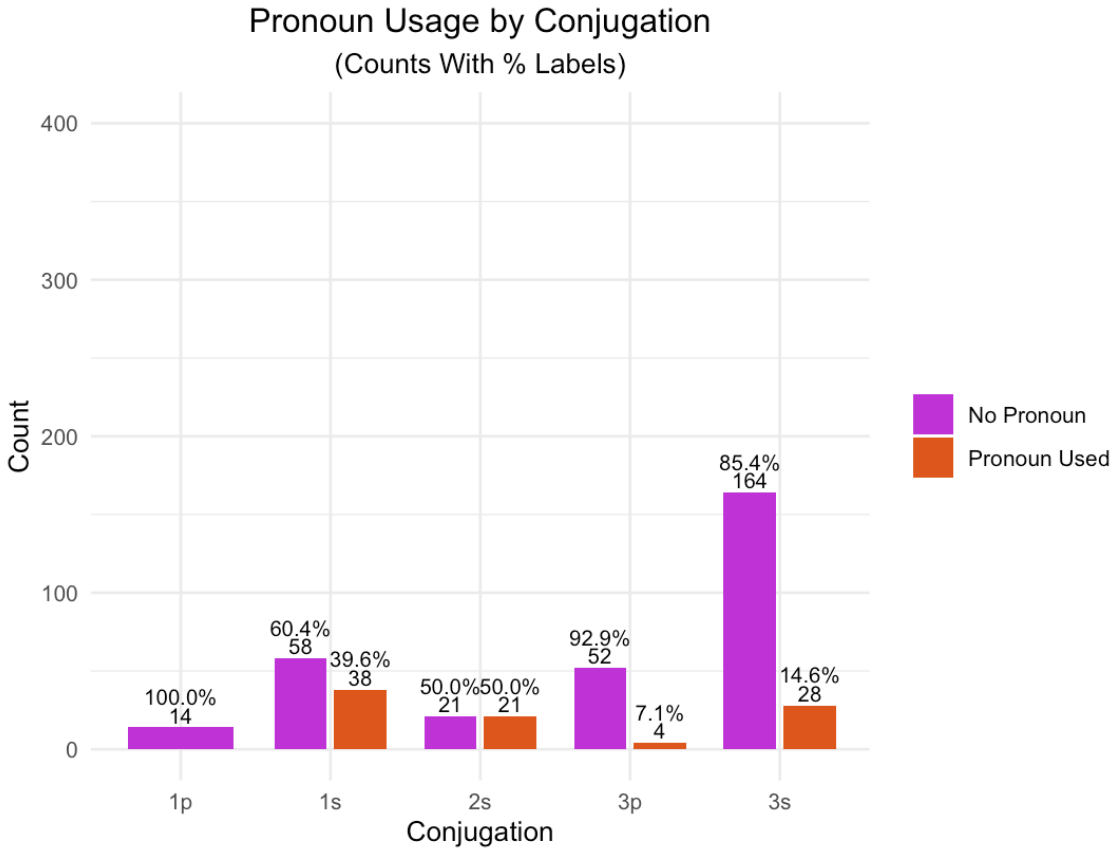

Figure 5. Count and percentage of time pronouns were used by conjugation. 1p = 1st Person Plural Conjugation; 1s = 1st Person Singular Conjugation; 2s = 2nd Person Singular Conjugation; 3p = 3rd Person Plural Conjugation; 3s = 3rd Person Singular Conjugation.

This distribution reveals the key issues described in data preparation. The 1p category exhibited zero pronoun use, which leads to a case of perfect prediction. This issue creates numerical instability or non-convergence when fitting logistic models, as the model cannot estimate a finite coefficient for a predictor that perfectly predicts the outcome. Similarly, the absence of certain combinations (like missing 1s tokens in parts of the dataset) contributes to sparsity, which can further compromise estimation.

Figure 6 displays that pronoun usage varies substantially across grammatical person. Pronouns were used most frequently with **second-person** conjugations (50.0%), followed by **first-person** (34.5%). In contrast, **third-person** conjugations had the lowest pronoun usage, at only 12.9%, despite representing the largest number of tokens (N = 248). This pattern suggests that subject pronouns are more likely to be included when the speaker refers to themselves or directly addresses someone else, and less likely when referring to others.

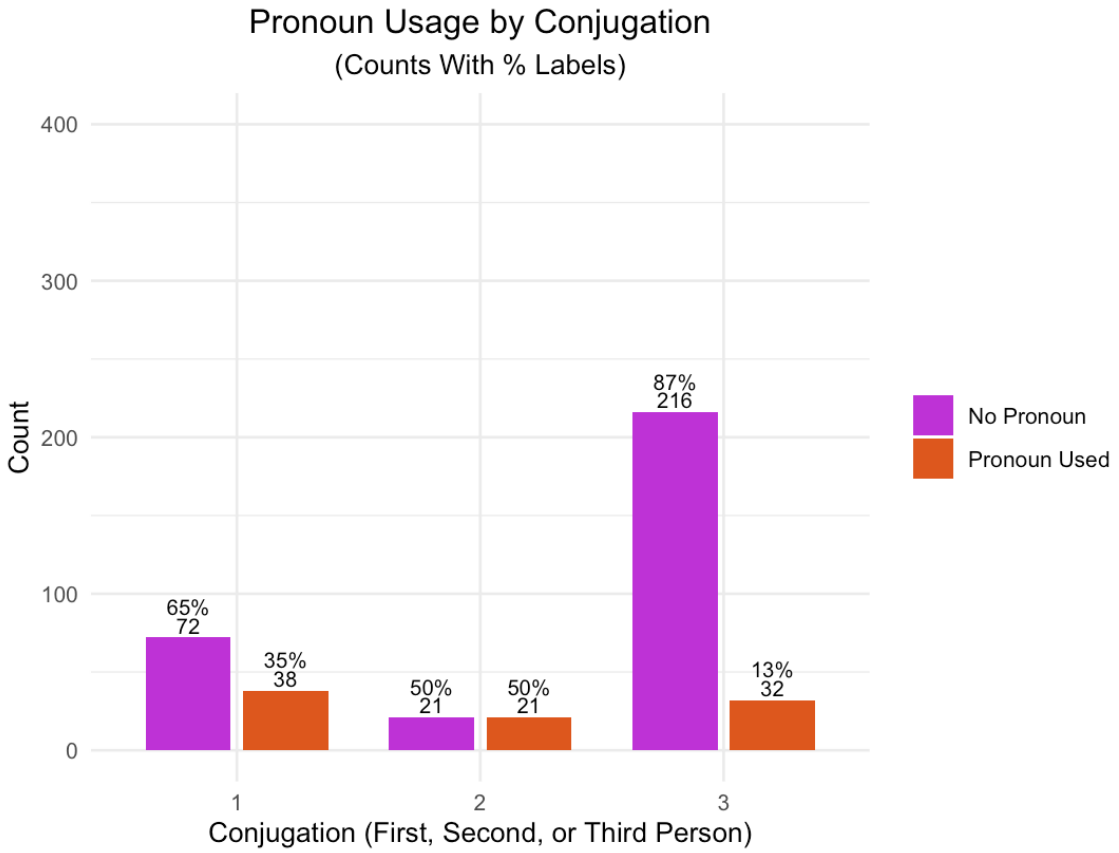

Figure 6. Count and percentage of time pronouns were used by first person, second person, or third person conjugations. 1 = First Person Conjugation; 2 = Second Person Conjugation; 3 = Third Person Conjugation.

Figure 7 shows that pronouns were used much more frequently with singular verb conjugations (26.4%) compared to plural ones (5.7%). Additionally, singular conjugations were far more common in the dataset (N = 330) than plural conjugations (N = 70), suggesting both a higher baseline usage and a stronger association with subject pronoun presence.

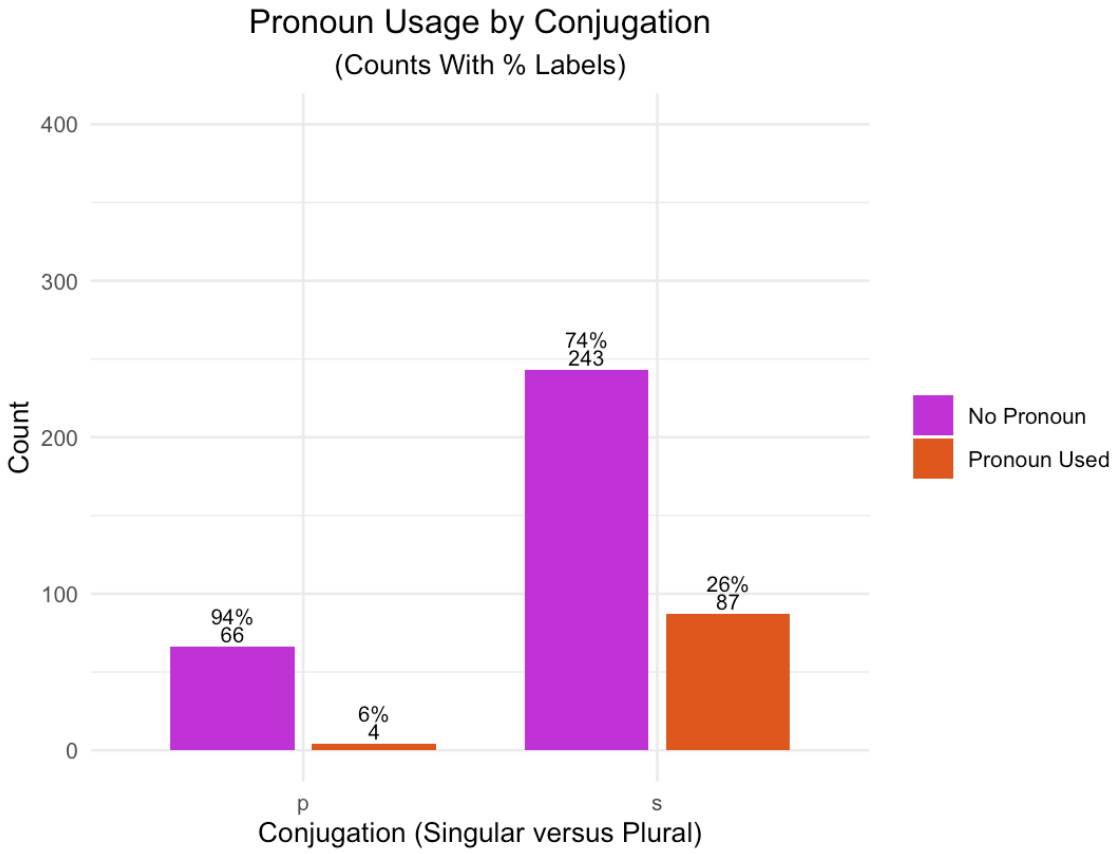

Figure 7. Count and percentage of time pronouns were used by singular or plural conjugations. P = Plural Conjugation; S = Singular Conjugation.

Figure 8 displays the proportion of pronoun use for each speaker, arranged by descending age. Each bar represents an individual speaker, with name and age shown on the x-axis. Speakers identified as female are shown in red, while male speakers are shown in blue. The chart reveals variability in pronoun use across both gender and age. For example, among older speakers, VIC (female, age 66) and JAC (male, age 41) show relatively high usage rates (around 0.45 and 0.45, respectively), while younger speakers like HER (female, age 22) and FEL (male, age 11) exhibit much lower rates. This suggests potential age- or gender-related patterns in pronoun usage that warrant further analysis.

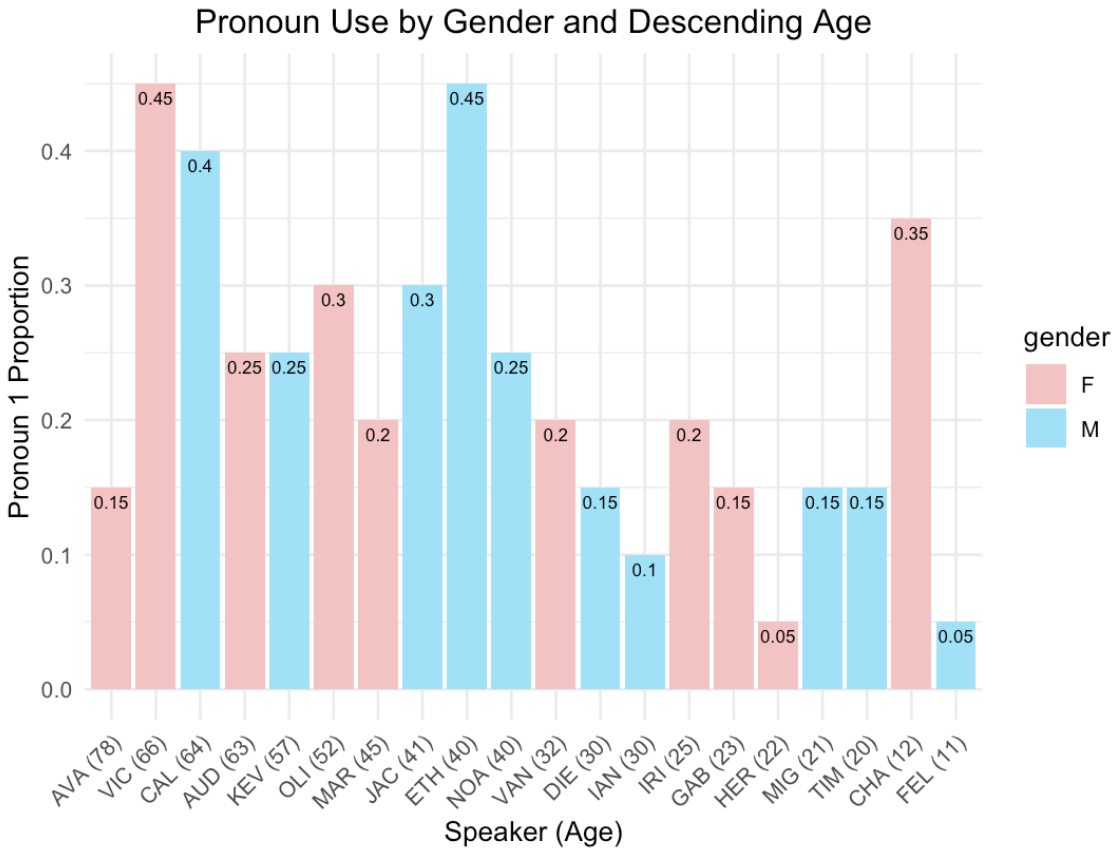

Figure 8. Percentage of time pronouns were used by descending age. Females are presented with red, and males are presented with blue. The three-character strings indicate the speaker. The ages are presented in parentheses along the x-axis. Females are presented with red, and males are presented with blue. The three-character strings indicate the speaker. The ages are presented in parentheses along the x-axis.

Figure 9 shows the percentage of pronoun use broken down by both gender (male vs. female) and age group (younger than 35 vs. 35 or older). There appears to be a potential interaction effect between age and gender. Older males used pronouns most frequently (33.0%), while younger males used them the least (12.0%). In contrast, the difference between older and younger females was smaller, with older females using pronouns 27.0% of the time and younger females 19.0%. This pattern suggests that the effect of age on pronoun usage may vary by gender.

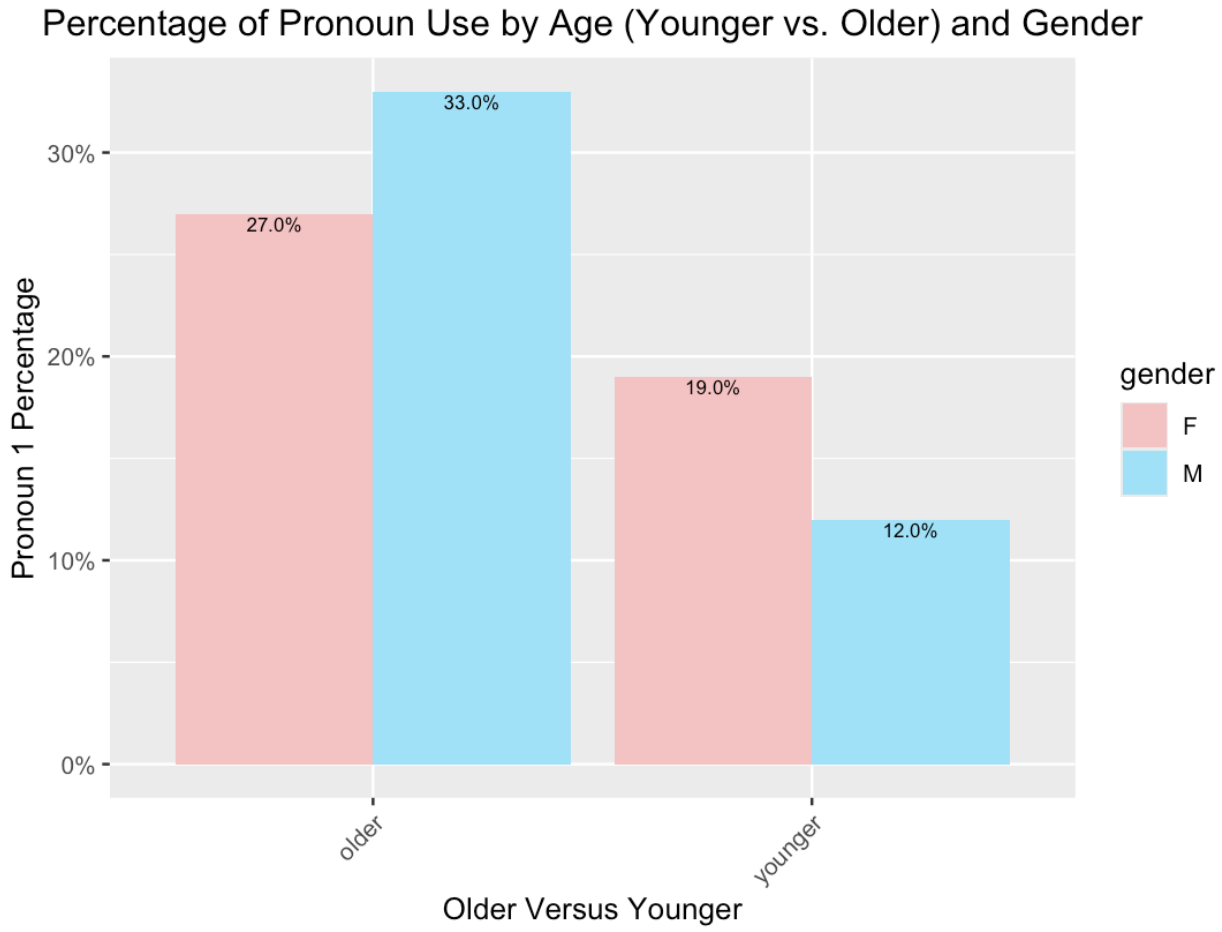

Figure 9. Percentage of pronoun use broken down by both gender (male versus female) and younger (younger than 35) or older (35 or older).

Data Analysis

A generalized linear mixed model (GLMM) with repeated measures was performed. The dependent variable was pronoun usage. The independent variables were gender (male versus female), age, "before" (if Spanish or English was used before the token), and conjugation.

There were no pronouns used for the first-person plural conjugation, and this did not allow the model to run. Thus, conjugation was divided into two variables: one variable that indicated whether the verb was singular or plural, and another variable that indicated if the verb was first person, second person, or third person. An interaction between age and gender was observed during exploratory data analysis, so this interaction term was included in the model as a fixed effect, along with gender, age, the preceding language ("before"), and the two new conjugation variables. A random effect for speaker was included to account for the repeated-measures design and account for tokens produced by the same individual potentially being more similar than those produced by different individuals. After fitting the model, multiple comparisons were conducted for variables with multiple levels that showed substantial effects. This allowed us to identify which specific levels differed significantly from one another.

Results

A generalized linear mixed-effects model was used to examine the likelihood of subject pronoun use among bilingual speakers. Model diagnostics indicated a good fit, confirming that the model results are valid and reliable. The model explained around 30% of the variance (conditional $R^2 = 0.290$), which is typical in research that examines pronoun usage (Erker & Otheguy, 2020). The conditional R^2 shows us how much variation in the outcome is explained by the model we used. In our model, it tells us that 29% of the variability in pronoun use can be explained by the combination of the predictors (age, gender, conjugations, language used before the token, and the random effects between the speakers. The model diagnostics indicated that the data fit the model with, which is presented in Appendix A. We present the key findings below. The odds ratios (OR) presented here indicate the proportion of odds that the pronoun was used. If the odds ratio is greater than 1, that indicates that the odds of a pronoun being used increased. If the odds ratio is less than 1, that indicates that the odds of a pronoun being used decreased. Additional findings are presented in Appendix B.

The key findings from the model include:

1. The interaction between age and gender was significant (OR = 1.05 (95% CI [1.01; 1.08], p = 0.004), indicating each additional year of age is associated with a 5% increase in the odds of using a pronoun. That is, older males were more likely to use pronouns than younger males, while age did not significantly affect pronoun use among females (see Figure 10). The main effect of age was not significant (OR = 1.01 (95% CI [0.99; 1.03], p = .400).

2. Gender had a significant main effect. Male speakers had 81% lower odds of using pronouns than female speakers (OR = 0.19 (95% CI [0.05; 0.75], $p = 0.018$).
3. Third-person verbs were associated with significantly lower odds of pronoun use compared to first-person (OR = 0.22 (95% CI [0.12; 0.40], $p < 0.001$). Multiple comparisons confirmed this finding, with both first- and second-person forms used with pronouns significantly more often than third-person forms.
4. Singular verb forms had significantly higher odds of pronoun use compared to plural forms (OR = 4.46 (95% CI [1.52; 13.04], $p = 0.006$).
5. There were no significant effects for the preceding language ("before") variable ($p = .818$ to $p = .999$) or for second-person vs. first-person conjugation ($p = .625$).

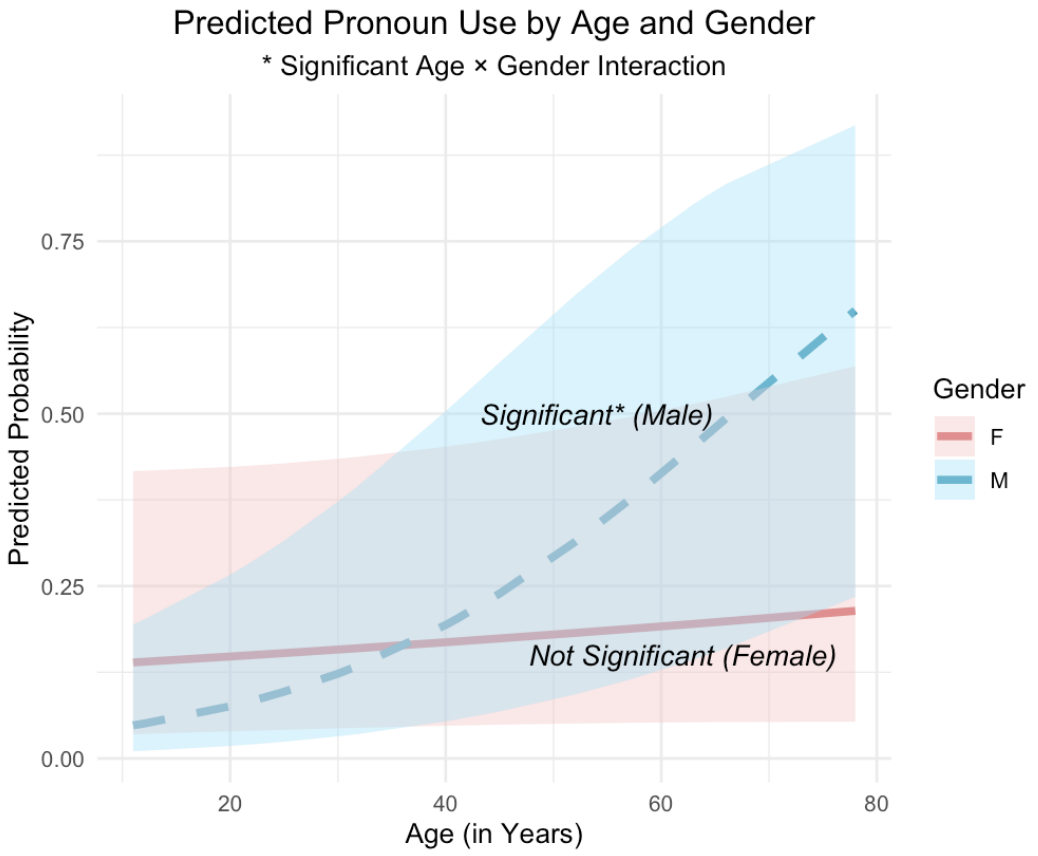

Figure 10. Predicted pronoun use from our GLMM based on both age and gender. F = Female; M = Male; The dashed line for males indicates the significant interaction for males specifically.

Additionally, we attempt to run the model without the 12-year-old who was a potential outlier. The random effects were small and the model did not converge.

Interpretation

The analysis revealed that subject pronoun use among bilingual Spanish-English speakers was influenced by both grammatical structure and speaker demographic characteristics.

One of the most notable findings involved the interaction between gender and age. In general, men used pronouns slightly less often than women, but this difference depended on age. For women, pronoun use remained relatively consistent across ages, with only a slight, not statistically significant, increase among older women. In contrast, age was a larger difference for men. Specifically, older men were noticeably more likely to use pronouns than younger men. This finding suggests that age plays a more important role in shaping pronoun use for male speakers than for female speakers. However, the lack of a significant age effect among females may be influenced by potential outliers in the data. The 12-year-old female speaker exhibited unusually high pronoun use compared to other females, which may have masked the overall age trend within the female group. Further investigation is needed to assess the impact of this potential outlier on the model results.

Grammatical structure played an important role in pronoun use. Pronouns were used much less often with third-person verb forms compared to both first- and second-person forms. Follow-up comparisons showed that speakers were about 4.5 times more likely to use pronouns with first-person verbs than with third-person verbs and about 6.5 times more likely to use pronouns with second-person verbs than with third-person verbs. There was little difference (i.e., not statistically significant) between first- and second-person usage. These patterns reflect how Spanish often omits third-person subject pronouns unless needed for clarity, while first- and second-person pronouns may be used more often to highlight the speaker or listener in conversation.

Whether singular or plural pronouns were used was another strong predictor. Pronouns were used much more often with singular verbs than with plural verbs, indicating that subject reference in singular form may more frequently favor pronoun inclusion.

Conclusions and Recommendations

Overall, the results from the existing data showed a significant interaction effect between age and gender, such that older males were more likely to use pronouns than younger males. This interaction effect was not significant for females, and there was not a significant main effect of age. The use of third person conjugations had lower odds of a pronoun being used compared to first and second person conjugations. Additionally, singular conjugations were associated with significantly higher odds of a pronoun being used as compared to plural conjugations. There was no significant effect of preceding language on the odds of a pronoun being used.

To build on the findings of this study, several directions for future research are recommended. First, it would be valuable to collect more data across a broader and more balanced set of verb conjugations. In the current dataset, not all speakers were exposed to the same range of conjugation types, which limits our ability to compare pronoun use across grammatical forms.

In addition, more data should be collected for specific verbs of interest. For example, the client expressed interest in the verb creer, which showed a high rate of pronoun use (80%). However, it only appeared 10 times in the dataset, which is too few to support strong conclusions. This limitation is common across other verbs as well; many appeared infrequently and thus cannot be reliably analyzed without further data collection. The breakdown of the verbs can be shown in **Appendix C** in **Table 4**, and shows the need for future analysis when more data is available.

Another useful direction for future research is to include information about the conversations themselves. In this study, each speaker came from a separate conversation, but no additional details about those conversations were included in modeling. Future studies could record who the speaker is talking to (e.g., a peer, family member, or interviewer), what role they play in the conversation (such as asking questions or responding), and where the token appears in the flow of the conversation. Adding variables like the position in the turn or the overall topic of the conversation could help explain why pronouns are used more in some situations than others. Further, there was the potential outlier of the 12-year-old who used pronouns 35% of the time. It would be beneficial to understand the nuances of the conversation, as the difference in pronoun use may lie in those details. Given that this one speaker changed the results of the main effect of age, it is important to examine where those differences originate.

It would also be beneficial to include the same number of possibilities for variables of interest, such as conjugation. For example, singular pronouns were used much more (N = 330) than plural pronouns (N = 70). Additionally, having the same number of sentences in which English was used before would help determine if this played a role in our findings.

Finally, increasing the number and diversity of speakers would enhance the generalizability of the findings. Including more variation in age, gender, and other variables would also allow for more robust modeling of speaker-level patterns. All of the speakers were in the Miami region, not allowing for the findings to extend beyond that region of the United States. Future work should examine different locations in the United States as well as different countries (e.g., Spain) in order to see how location may play a role in pronoun use.

References

Deuchar, M., Parafita Couto, M.D.C, Webb-Davies, P., & Donnelly, K. (March 19, 2025). [Data set]https://bangortalk.org.uk/speakers.php?c=miami .

Erker, D., & Otheguy, R. (2020). American myths of linguistic assimilation: A sociolinguistic rebuttal. *Language in Society,* 1-37. https://doi.org/10.1017/S00474045200000019

Hartig, F. (2024). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.4.7, <https://CRAN.R-project.org/package=DHARMa>.

Appendix

Appendix A: Model diagnostics

The R package DHARMa (Hartig, 2024) was used to examine the residuals and test the assumptions of our model. The plots (**Figure 11**) showed that the assumptions of the model were met. There were no obvious patterns or trends in the residuals.

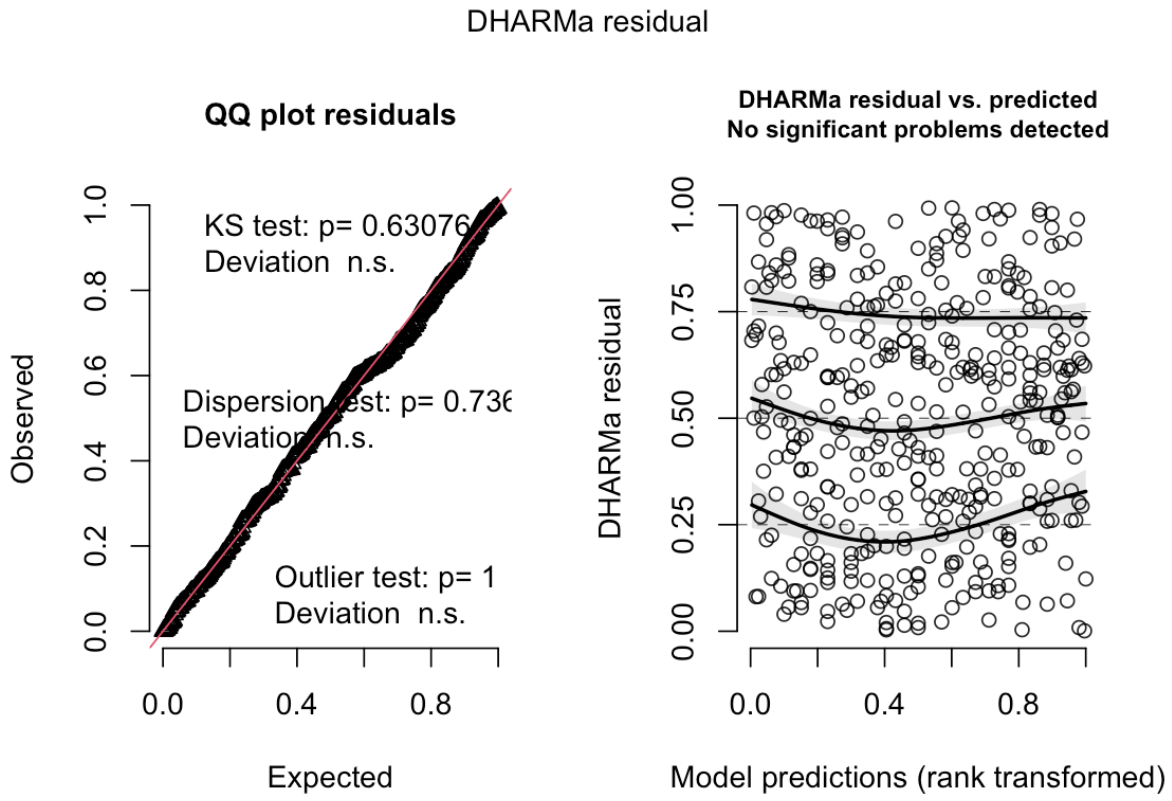

Figure 11. Residual plots for the Generalized Linear Mixed Model.

Appendix B: Model results

Table 2. Model results with odds ratios, confidence intervals, *p* values, random effect information, and model fit information.

	pronoun		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.15	0.03 – 0.71	**0.017**
age	1.01	0.99 – 1.03	0.400
gender [M]	0.19	0.05 – 0.75	**0.018**
before [spanish]	0.76	0.31 – 1.86	0.546
before [start]	1.03	0.07 – 14.86	0.984
person [2]	1.44	0.67 – 3.10	0.356
person [3]	0.22	0.12 – 0.40	**<0.001**
pr [s]	4.46	1.52 – 13.04	**0.006**
age × gender [M]	1.05	1.01 – 1.08	**0.004**
Random Effects			
σ^2	3.29		
$\tau_{00\ speaker}$	0.03		
ICC	0.01		
$N_{\ speaker}$	20		
Observations	400		
Marginal R^2 / Conditional R^2	0.284 / 0.290		

Table 3. Multiple Comparison for Conjugation (person).

Contrast	Odds Ratio	p-value	**Interpretation**
person1 – person2	0.7	0.6253	No significant difference
person1 – person3	4.55	<0.0001	Significant: 1st person uses pronouns more than 3rd person; 1st person has 4.5× higher odds of using a pronoun than 3rd person
person2 – person3	6.53	<0.0001	Significant: 2nd person uses pronouns more than 3rd person; 2nd person has 6.5× higher odds of using a pronoun than 3rd person

Appendix C: Verbs of Interest

Table 4. Times pronouns were used for verbs that appeared at least 5 times for all tokens, the total times the verb was used, and the percentage of time the pronoun for each verb.

Verb	Pronouns Used	Total Times Used	Percentage of Time Pronoun was Used
creer	8	10	80.0%
querer	7	10	70.0%
empezar	3	6	50.0%
gustar	2	5	40.0%
poder	4	10	40.0%
venir	2	7	28.6%
decir	5	19	26.3%
ser	15	59	25.4%
hacer	4	19	21.1%
llevar	1	5	20.0%
tener	5	32	16.6%
estar	6	39	15.4%
veer	1	8	12.5%
haber	1	9	11.1%
saber	1	12	8.3%
ir	2	29	6.7%
mirar	0	6	0.0%

Appendix D: Code

The code (ling_code_4.21.25.R) and instructions (Instructions and notes on R code.pdf) for using the code were provided in the data and code sharing repository on OneDrive.