Subject Expression in Native and Non-Native Speakers of Spanish

Chandra Frank

Abstract

The project studied the differences in subject expression between native and non-native speakers of Spanish. It was hypothesized that as the non-native speakers gain greater proficiency, their subject expressions will look more like those produced by native speakers. To assess this, essays were collected from native speakers and three different levels of non-native speakers: beginner, intermediate, and advanced. The relative frequencies of four different ways of expressing subject in Spanish were measured in each essay. These relative frequencies were compared by type across the levels using two different methods of analysis: an analysis of variance for pairwise comparisons using a randomization test and a multinomial regression analysis within a Bayesian framework using a Dirichlet prior. The two analyses agreed that there was a difference between the native group and every other proficiency level for almost every type of subject expression. They disagreed over whether there was a difference for two different types of subject expressions between the beginner and intermediate groups and the beginner and advanced groups. The results suggest that when teaching non-native speakers, more emphasis should be placed on the different types of subject expression and when it is best to use each one.

1. Introduction

The current study seeks to answer the question: at what proficiency level do non-native speakers of Spanish produce subjects like native speakers? It was hypothesized that as the non-native speakers gain greater proficiency, their subject expressions will look more like those produced by native speakers. This study can help determine if non-native speakers are progressing towards a native level and can indicate whether subject expression needs to be

explicitly taught to non-native speakers of Spanish and at what level it should be introduced and emphasized.

2. Methods

To study these subject expressions, written essays were collected from students. The essays from the non-natives speakers were collected from the online corpus CEDEL2: Corpus Escrito del Español L2. Native speakers' essays were collected from two universities in Spain. The ages of students were 17-19 years old with no study abroad experience. The topic for the text was a recent trip or holidays. Age of the student when first exposed to Spanish as well as number of years studying Spanish were not restricted, as the important consideration was only the proficiency level. The number of essays collected for each level were as follows: 16 beginners, 25 intermediate, 25 advanced, and 23 natives.

Each conjugated verb was labeled according to the type of subject expression that appeared with it. The possible types of expressions were noun phrase, pronoun, other, and null (non-expressed). Examples of these four different types with the subject underlined are as follows (English translation in parentheses):

Noun phrase: Mi amiga habla español. (My friend speaks Spanish.)

Pronoun: Yo hablo español. (I speak Spanish.)

Other: Me gusta hablar español. (I like to speak Spanish.)

Null: Hablo español. (I speak Spanish.)

Note that there is no underlined subject in the null example because it is implied in the way the verb is conjugated and can be dropped.

The count of every type of subject expression in each essay was recorded along with the total number of subjects used. These counts were compared across the proficiency levels using two different methods of analysis. The first was a one-way analysis of variance for group and pairwise comparisons. For group comparisons an F statistic was used and for pairwise

comparisons a t-statistic was used with the Fisher LSD correction. All the alternative hypotheses in the group and pairwise comparisons were one-sided. A randomization procedure was used to avoid issues with lack of normality. In this analysis each subject type was considered separately, and the proficiency levels were randomized among the observed relative frequencies. This was repeated 5000 times. Then the proportion of iterations that resulted in an F or t-statistic more extreme than the observed statistic was calculated to produce a p-value. A decision-based approach with $\alpha = 0.05$ was used in the analysis.

The second analysis was a multinomial regression within a Bayesian framework using the following model:

$\underline{Y_{ij}} \sim \text{Multi}(\underline{P_j}, n_i)$

$\underline{P_j} \sim \text{Dir}(\underline{\alpha_j})$

where i = student and j = proficiency level.

In the multinomial likelihood, the vector $\underline{Y_{ij}}$ represents the counts of each subject expression type for student i. The parameter n_i corresponded to the total number of subjects that a student used which varied between students. The parameter P_j corresponded to the probability that a student of a certain proficiency level would use a specific subject expression. The sum of $\underline{P_j}$ within the vector for every proficiency level was one, in accordance with the multinomial distribution. A Dirichlet prior was put on $\underline{P_j}$. All parameters ($\underline{\alpha_j}$) in the Dirichlet were set equal to 1 in order to have a vague prior. The means from the posterior distributions were recorded for the proportion of each subject type used by every different proficiency level. For the pairwise comparisons each pair of proficiency levels was compared by subtracting the posterior distribution of the proficiency level with a smaller estimated proportion from the posterior of the proficiency level with a larger estimated proportion. The proportion of these differences that was less than 0 was calculated and recorded. A decision-based approach with $\alpha = 0.05$ was used in the analysis.

Finally, the p-values from the randomization approach were compared to the proportion of iterations less than 0 from the Bayesian analysis. Here an evidence-based approach was used. If both values were relatively small or relatively large, the approaches were said to agree that there was a difference or no difference between the proficiency levels. If a value was large in one approach but small in the other, they were said to disagree about whether there was a difference between the proficiency levels.

3. Results

3.1 Summary statistics

Level	Mean Percent Noun Phrases	Standard Deviation
Beginner	26.5	22.2
Intermediate	27.1	9.35
Advanced	23.3	7.91
Native	17.9	7.02

Table 1: Summary statistics for noun phrases by level

Level	Mean Percent Pronouns	Standard Deviation
Beginner	16.3	16.4
Intermediate	20.6	14.1
Advanced	16.0	11.5
Native	11.3	6.01

Table 2: Summary statistics for pronouns by level

Level	Mean Percent Other	Standard Deviation
Beginner	2.01	4.15
Intermediate	1.53	1.66
Advanced	2.22	2.47
Native	3.74	3.90

Table 3: Summary statistics for other expression type by level

Level	Mean Percent Null Subjects	Standard Deviation
Beginner	55.1	25.0
Intermediate	50.8	17.8
Advanced	58.5	14.0
Native	67.1	12.3

Table 4: Summary statistics for null subjects by level

Native speakers tend to use other (Table 3) and null subject (Table 4) expressions more frequently than non-native speakers, and they use pronouns (Table 2) and noun phrases (Table 1) less frequently than non-native speakers.

For all expression types except other the natives had a lower standard deviation than all levels of non-natives. This would suggest that the native speakers are more consistent with each other in their usage of certain subject expression types. Beginners had the highest standard deviation for all subject types, suggesting that they are the most inconsistent with each other.

In tables 1-4, the intermediate group had the biggest difference from the native group for every subject type, so it is expected that most of the statistically significant differences will occur between these two proficiency levels. The advanced speakers had the smallest difference from

the native speakers for every type, which is expected, and it is predicted that there will be fewer statistically significant differences between these two levels.

The intermediate group could have numbers that look the most different from the natives because they are at a lower level but there is less variation than the beginners. At the beginner level there were some instances of using almost 100% or 0% of a certain type of subject expression, which led to the high standard deviations and potentially skewed the means higher or lower. At the intermediate level there was less variation although they are still using different frequencies than native speakers.

3.2 Randomization Approach

Subject Expression	p-value
Noun phrase	0.0196*
Pronoun	0.0268*
Other	0.021*
Null	0.022*

Table 5: p-values for overall F randomization tests by subject type (* indicates significance)

The results of the overall randomization approach shown in Table 5 indicate that there is a difference between the proficiency levels for each type of subject expression. To find exactly where those differences are the results of the pairwise comparisons must be analyzed. These results also justify the use of the Fisher LSD adjustment for pairwise comparisons because differences were found for all group comparisons.

Subject Expression	Beg vs Inter	Inter vs Adv	Beg vs Adv	Nat vs Beg	Nat vs Inter	Nat vs Adv
Noun Phrase	0.4556	0.1344	0.1966	0.0142*	0.004*	0.0542
Pronoun	0.1424	0.0954	0.463	0.1058	0.0048*	0.0904
Other	0.3152	0.2238	0.4188	0.0416*	0.0072*	0.0474*
Null	0.2156	0.0538	0.2774	0.02*	0.0006*	0.0454*

Table 6: Proportion of iterations resulting in a more extreme value than the test statistic in a randomization approach (* indicates significance)

In both the other and null expression types, shown in the bottom two rows of Table 6, the non-natives did not have any differences across the proficiency levels and were different from the natives at all levels. This suggests that for these types of subject expressions there is no change or progression towards producing the same relative frequency of null and other subjects as natives across the proficiency levels. Non-native speakers are never producing these subjects in a way similar to the native speakers. Even if there is some change between the proficiency levels it is not a large enough difference to be greater than the natural variation seen between the levels.

For the noun subject types in the top row of Table 6, the non-natives were once again the same across all the proficiencies. However, when compared to the natives, there was no difference between the advanced speakers but there were differences between the natives and beginners and the natives and intermediates. This suggests that there is some progression between the intermediate and advanced levels that brings the advanced speakers to a native level. This is undetectable when comparing the non-natives to each-other because the changes could be gradual or slight and not greater than the variability, but they are enough to make the advanced produce noun phrases similar to native speakers

The pronoun expression was similar to the noun phrases as there were no differences across the non-native speakers, but also no difference between the natives and advanced.

However, the beginners were also the same as the natives, while the intermediates were not. This could indicate progression away from the natives or “backwards progression” between the beginner and intermediate levels, although it would need further research. A more likely explanation is that the larger variability in the beginner level is causing potential differences to go undetected because any change in the relative frequency of pronouns is small in comparison to the variability.

3.3 Bayesian Analysis

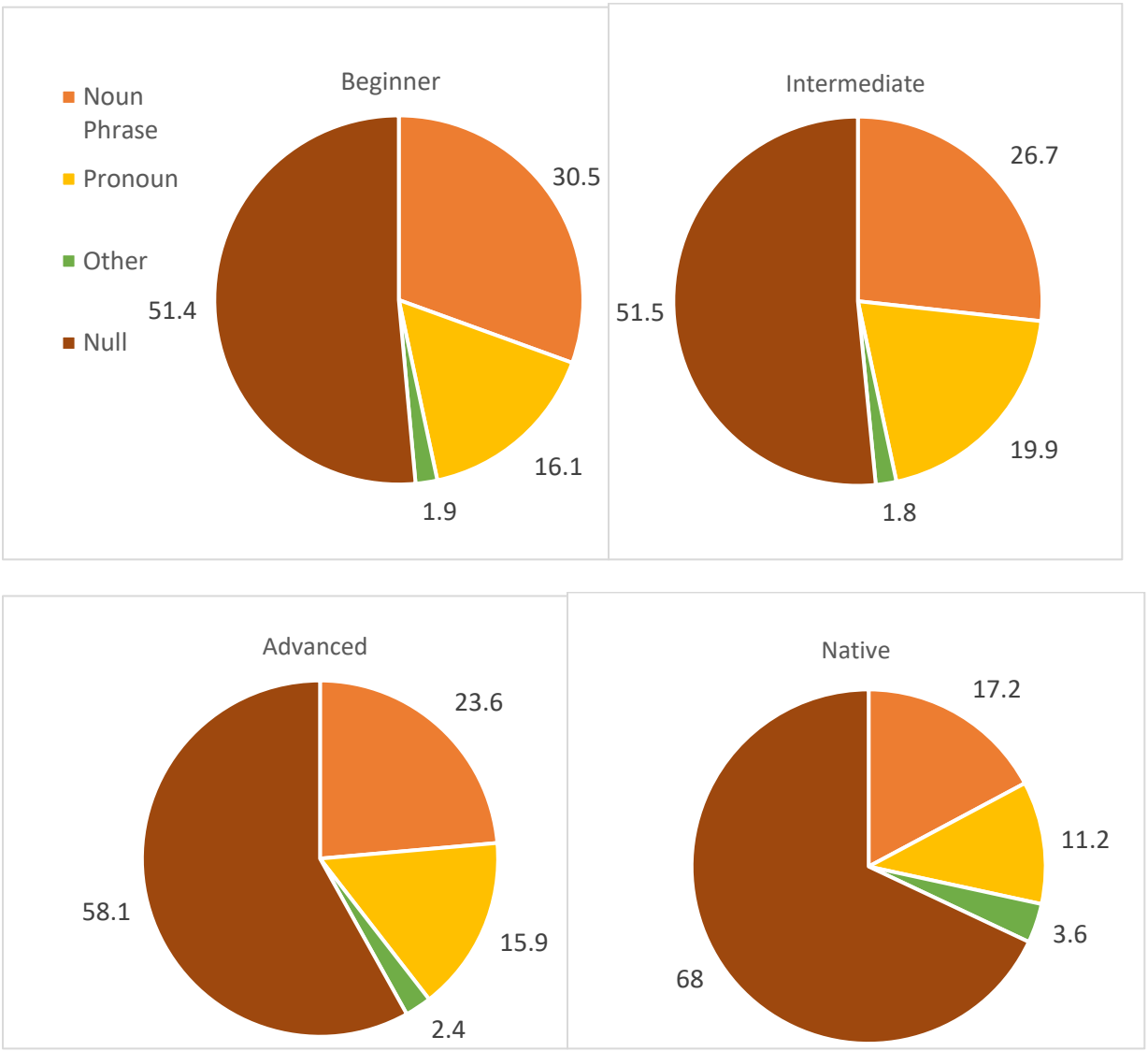

Figure 1: Means from posterior distributions for proportions of subject expressions.

The pie charts in Figure 1 show how the proportions of subject expressions are changing across the proficiency levels according to the Bayesian analysis. In general, advanced and native speakers use a higher proportion of null and other subjects and a lower proportion of noun phrases and pronouns than beginners and intermediates. The exception to this is between the beginner and intermediate levels as the intermediates use more pronouns than beginners and slightly fewer other subjects. This overall trend makes sense as the more complicated constructions contained in the other expression are better understood by advanced and native speakers, and the null subject is not a part of English grammar, so beginner and intermediate learners tend to use nouns and pronouns instead.

Subject Expression	Beg vs Inter	Inter vs Adv	Beg vs Adv	Nat vs Beg	Nat vs Inter	Nat vs Adv
Noun Phrase	0.0591	0.0273*	0.0016*	0*	0*	0.0001*
Pronoun	0.0320*	0.0022*	0.4529	0.0058*	0*	0.0006*
Other	0.4775	0.147	0.2520	0.0377*	0.0066*	0.0531
Null	0.4849	0.0003*	0.0059*	0*	0*	0*

Table 7: Proportion of iterations resulting in a difference less than 0 in a Bayesian analysis (* indicates significance)

The results in Table 7 show that for all subject types except for other, the Bayesian analysis found that the natives were different from every level of the non-natives. These results suggest that even if there is progression among non-native speakers, they are never achieving a level equal to that of natives, except for the other expression type.

In the other expression type, there were almost no differences, except for between the beginners and natives and the intermediates and natives. This indicates that the proportion of other subject expressions does not change much as proficiency increases but remains relatively

small, even up to the native level. There could be slight changes between the levels of non-natives that are undetectable when they are compared to each other but more noticeable when low proficiencies are compared to the native group. As the advanced group produces a proportion of other subjects that is greater than beginners and intermediates but less than natives, any changes are not noticeable when it is compared to the other levels.

Between the beginner and intermediate levels there are almost no changes, except in the pronouns. This most likely resulted in other much smaller changes in the other three expression types that were not large enough to be considered as differences.

Between the intermediate and advanced levels there are more changes as there were many differences between the advanced group and the other levels of non-natives. The exception to this is the pronoun expression type in which there are differences between the intermediate level and the other levels but not between the beginners and advanced. This could be a sign of “backward progression” from beginner to intermediate where they move farther away from a native level and then correct at the advanced level. However, it is more likely due to the high variability associated with the beginner level. This makes it harder to detect differences when comparing the beginners to other levels because there is so much uncertainty. It could be that each level of non-natives is progressing towards a native level, but those differences cannot be seen for certain comparisons.

3.4 Comparing the Approaches

Subject Expression	Beg vs Inter	Inter vs Adv	Beg vs Adv	Nat vs Beg	Nat vs Inter	Nat vs Adv
Noun Phrase	D	O	D	O	O	O
Pronoun	O	O	X	O	O	O
Other	X	X	X	O	O	O
Null	X	O	D	O	O	O

Table 8: Comparison of the results of the two analysis methods. “X” means approaches agree there is no difference, “O” means approaches agree there is a difference, “D” means approaches disagree.

The results in Table 8 show that native speakers were different than every proficiency level of non-native speakers. Despite some expression types showing progression across the levels, the non-natives are never achieving similar subject expressions as native speakers.

In the other expression type the approaches agree that all the proficiencies of non-natives look the same and they are all different from natives. This means that in this area there is little to no progress among the non-natives towards producing this subject type with the same relative frequency as a native.

For the pronoun type the approaches agreed that the advanced and beginners are the same, but the intermediates are different from everything else. This result could indicate that there is some sort of movement occurring across the levels, whether it be forwards or backwards, however there is nothing that brings the non-natives to a native level. The result could also be due to the high variability at the beginner level. This makes any changes that occur between the beginners and the other levels harder to detect because it is small in comparison to the variability.

In the noun and null expression types, the approaches agreed that the native, advanced, and intermediates are all different from one another. This indicates progression between the

intermediate and advanced levels, but that progression is insufficient to bring the advanced up to a native level.

In these same two expression types, the approaches disagree about what is happening at the beginner level. While they agree that the beginners are similar to the intermediates for the null subjects, it is unclear whether they are different from the advanced. For the noun phrases, neither approach agrees if the beginners are similar to the intermediates or the advanced levels. This could be due to the smaller sample size in the beginner group as well as the higher variability for that proficiency level across all subject types. This was especially true in the noun and null types where the approaches disagree about the conclusions.

4. Conclusion

The main finding of the study was that the natives are different than the non-natives across the proficiency levels for almost all the subject types. This is the case even when there is progression throughout the levels of non-natives. This suggests that subject expression may need to be introduced earlier to non-natives or taught more explicitly, especially at the advanced level.

The study also found that progression is occurring among the non-natives, mostly between the intermediate and advanced levels. This is true for all subjects except in the other expression type in which little to no progression is occurring. This could be because all non-natives are familiar with basic other type expressions such as “me gusta comer.” However, this type does not get introduced explicitly as a way to express subject until the advanced level. It is possible that the advanced speakers are not fully grasping the concept well enough to use it like a native speaker and are instead using the more basic other constructions like a beginner or intermediate would.

One limitation of this study was the smaller sample size for the beginner group, as well as higher standard deviations for that proficiency level across the subject types. The small sample

size was made necessary by the selection process, specifically in setting the age range from 17-19 years old. At this age many students have already been studying Spanish for several years and fall into the intermediate or advanced levels, so the number of beginners was limited.

Possible future research could include studying different curriculum or teaching methods to compare which ones are better at progressing non-native speakers closer to a native level. In addition, the pronoun expression type could be broken down into different types of pronouns and analyzed in a similar way to the current study. This could offer more insight into subject expression specifically for pronouns to understand at what level non-natives are comprehending and implementing the different types of pronouns. Age could also be used as a predictor, both for proficiency level and for the relative frequencies of subject expressions. Finally, other language combinations could be used in place of English and Spanish or native Spanish speakers who have different dialects could be compared using the same study design.

Acknowledgements

I would like to thank my primary mentor, Dr. Jared Studyvin, for guiding me during the analysis phase of the project and helping me edit during the final stage. I also want to express my gratitude to my Spanish mentor, Dr. Irene-Checa-Garcia, for her guidance in setting up the project in the initial stage and her help in identifying subject expressions in the essays.

Finally, I would like to thank the Arts and Sciences Board of Visitors for the grant that helped fund my research and allowed me to spend the summer collecting and labeling essays.