Franco Rodriguez, Camila (2023) Computer-based Reading Recall on Sociolinguistic Research: Towards a Cross-disciplinary Understanding of Bilingualism. Journal of Computer-Assisted Linguistic Research 7: 1-21. ISSN: 2530-9455.

# Computer-based Reading Recall on Sociolinguistic Research: Towards a Cross-disciplinary Understanding of Bilingualism

Camila Franco Rodriguez*
Temple University, USA
* Corresponding author: tul02251@temple.edu

Received: 21 April 2023 / Accepted: 14 July 2023 / Published: 18 December 2023

***Abstract***
Global bilingual communities are a fascinating phenomenon that has received constant attention from different angles and disciplines. Sociolinguistic research has also turned interest towards what motivates change in these globalized settings, as well as psycholinguistic research has wanted to focus on the cognitive aspects of L2 speakers. With the widespread use of computer-based methods, it seems natural to add them to contemporary research as a way of understanding variation and change to a deeper level. Through the data I have collected, I debate in this article the importance of including computer-based tests as part of traditional variationist research. I argue that the traditional separation of methods and data collection has influenced the research process to a point where some new behaviors could be overlooked. In this article I report the relationship between cognitive adaptation and social experiences in the Colombian in the Philadelphia bilingual community, which becomes more proficient not only because of age and time of L2 learning, but also because of how welcoming their social circles are, as well as how diverse their friendships and workplaces are.

***Keywords:*** Bilingualism, Sociolinguistics, Psycholinguistics, Variationist, Syntax

## 1. Introduction

According to the U.S. 2010 Census, the South American population in Pennsylvania is approximately 719,660. Philadelphia has seen a steady Colombian population growth, primarily around the northeast Philadelphia area, where several shops and traditional restaurants have opened to cater to the expanding community. The available 2010 Census information counted this portion of the population as U.S.-born nationals with Colombian descent and/or ancestry, which means that an additional population that had not established itself in the country may

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2023.18547
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA. 

not have been included in the census.

Spanish-English contact in the Colombian bilingual U.S. community shares aspects with other communities in New York, Florida, Texas, and California (states with a higher Latinx population) with multiple generations of Hispanics established for years, whereas others are more recent arrivals. The diaspora disparity results in first- and second-language variations at the phonological and morphosyntactic levels. Pronominal subject variation is a well-studied morphosyntactic feature among Spanish-English bilinguals. It has received a great amount of interest because subject pronoun expression (SPE) or relative frequency of overt vs. null subject pronoun in both languages is constrained differently from the grammatical category of the subject to its realization frequency. However, most studies on SPE variation in Spanish in the United States have been conducted in Dominican, Puerto Rican and Mexican speech communities.

I want to reconcile the idea that bilingual speakers can be socially and linguistically competent in both of their languages. This project is comprised of a sociolinguistic interview for collecting SPE variation in Spanish and a reading test for assessing the cognitive basis of the bilinguals´ unified syntax (see section 2.2). I collected naturalistic data of pronominal subject variation for different bilingual groups through bilingual sociolinguistic interviews. Participants completed a demographic questionnaire to describe their age, gender, education level, language background, age of arrival, etc. I used a second questionnaire to gather information about their social networks among other common independent variables. While taking part in the reading experiment participants were required to recall an L2-word (English for non-native speakers and Spanish for heritage speakers) that appears on the screen for 200ms (basic reading speed) in different grammatical and agrammatical positions.

We will observe how SPE shows independent Spanish variation. Specifically possible correlations with the percentage of word recall on the reading test with speakers who report higher proficiency in English. According to Hartsuiker et al. (2004), Declerck et al. (2020), and Helasvuo (2004), speakers should be able to recall more than 60% of the grammatical words they see on the screen, and lower proficiency speakers around 40-50% recall of grammatical words. We will use a mixed-effects logistic regression model to generate reproducible predictions in inferential statistics. Analyzing SPE data under an inferential lens will contribute to the understanding of the influence of factors such as neighboring communities, immigration motivators, socioeconomic background, and social networks.

## 2. Theoretical Background

### 2.1. Subject pronoun expression in North and South America

Sociolinguistic aspects of SPE variation steadily became relevant to studies comparing different Hispanic community settings and Spanish varieties such as the ones spoken in the coastal and mainland regions. Given the higher percentage of overt pronouns in coastal Spanish vis-à-vis mainland Spanish, Shin and Erker (2015) called SPE the "showcase variable" for syntactic variation studies because it shows a stable sociolinguistic pattern across Spanish-speaking

countries. SPE has been studied across Latin America and Spain, and among immigrant communities in the United States. Variationists have been trying to understand the relative frequency of overt and null pronominal subjects in their social and linguistic context describing their appearance and structural constraints.

Lastra and Butragueño's (2015) study on pronominal subject variation in Mexico addressed important questions about null subject variation in Spanish. As in previous studies, the authors hypothesized that internal or linguistic factors play a stronger role than social or external factors in SPE. Through logistic regression and corpus analysis, the authors discussed how age presents a significant influence on SPE. Additionally, they commented on the importance of verbal constraints in Mexican Spanish variation. Mexico City, specifically, showed a lower SPE, which agrees with other findings on mainland Latin American Spanish. However, overt subject expression is higher in monolingual coastal Spanish and Spanish in the United States, with younger generations shifting toward a higher use of an overt subject expression.

Orozco (2015) found similar SPE trends in Colombian Costeño Spanish. He proposed that internal constraints are more robust than social ones and have a greater statistical significance. Like other researchers, Orozco concluded there was a significant association between age and gender, on the one hand, and linguistic variation on the other. As found in other SPE studies mentioned in this section, internal constraints held significant influence, specifically on switch reference, tense-mood-aspect, lexical meaning, and subject-verb constraints such as the tense-mood-aspect, semantics, and inflection, among others. Once again, overt subject expression was more common in Colombian Costeño Spanish, reaching 50% alternation with the null subject. Such a high percentage of overt subjects in Colombian Costeño Spanish and in Caribbean Spanish more generally, is consistent with findings on SPE in bilingual communities.

Although there are common linguistic trends across coastal Spanish-speaking communities, the geographical and cultural differences shaping the dialect continuum in Spanish Latin America also manifest themselves sociolinguistically in SPE variation. In this respect, Claes (2011) reviewed the dialect differences between the Spanish of Barranquilla, Colombia, and San Luis, Mexico, finding significant sociolinguistic variation across internal and external factors between them, which underlines the importance of speech community identity in drawing conclusions in comparative sociolinguistic research. These dialectal differences will become relevant when studying Spanish-speaking immigrant communities in other regions, mostly in the United States. First, they indicate that multicultural communities with different dialectal backgrounds coexist in spaces with new social structures and values. Second, Hispanic settlements open the possibility for a bilingual upbringing and for communities to have native (NS) and non-native Spanish (NNS) speakers sharing the same space.

The variation of SPE between NS and NNS of Spanish in the United States can be studied through the internal constraints of Spanish, an inflected language with number and person markers and English. Because English, unlike Spanish, is a non-pro-drop language, it is more common to observe a higher usage of overt pronouns in English. Geeslin et al. (2008) documented that highly proficient NNS of Spanish had null subject usage similar to NS speakers but tended to use more contextual references around non-specific constructions.

Referencing when it is unnecessary by context or adding referential cues are some of the linguistic strategies that NNS speakers can use to realize null subjects, considering these additional strategies can also contribute to the overall description of SPE variation across multilingual communities. The relevance that language acquisition, dialect, and context have on SPE variation can be seen in the foundational work of Otheguy et al. (2007) on the study of sociolinguistic change in English-Spanish bilingual speech communities of New York City. Their work was critical in describing language variation and change in the development of Spanish-speaking communities. Their methodology shaped subsequent studies, and their work continues to influence null-subject sociolinguistic research.

Otheguy et al. (2007) analyzed each interview with the help of the Statistical Package for the Social Sciences (SPSS) to obtain percentages and supported their predictors through multivariate (VARBRUL) and bivariate (correlation) statistical analysis. The mixed-method aspect of the study provided reliable predictors that are still in use today in SPE research. They studied the influence of ten linguistic variables: genre of discourse, person and number of the verb, tense-mood-aspect of the finite verb, reflexive or non-reflexive use of the verb, specific or nonspecific reference, discourse connection between verbs, type of lexical content of the verb, clause type where the verb appears, the appearance of the verb in a set phrase, and the section of the interview where the clause is found. The authors also considered five independent social variables (age, sex, education, social class, and socioeconomic status) and ten sociolinguistic variables: national origin, regional origin, age of arrival in NYC, years spent in NYC, level of Spanish skills, level of English skills, use of Spanish with types of interlocutors (e. g., father, mother, siblings, spouse, etc.), and use of Spanish in various domains (e. g., home, work, etc.), use of Spanish with speakers from the same geographic origin and from different geographic origins.

Through this robust corpus codification, Otheguy et al. (2007) were able to observe a dialectal shift that balanced the SPE of coastal and mainlander Spanish by moving both acrolects toward a lower overt pronoun rate among coastal speakers and a higher pronoun rate among mainlanders. Otheguy et al. (2007) concluded that Latino communities of NYC were creating a system that incorporated linguistic features from coastal and mainland Spanish, implying that the SPE change was not a mere simplification in that it allowed various accommodation patterns that resembled the grammars of English and Spanish.

The impact of a second language should be considered for a better understanding of the SPE phenomena, another reason for assessing language contact and accommodation in sociolinguistic research. Following Otheguy et al. (2007), Barrera-Tobón and Raña-Risso (2016) also investigated NYC's variation through corpus-based sociolinguistics. Their findings suggest that English proficiency and time of exposure are two key variables that influence null subject variation. Also, they mentioned the importance of the process of overt subject usage, where there must be an initial shift in preverbal and pronoun use before the overt pronoun rate became higher. Like Otheguy et al. (2007) and Orozco (2018), they concluded that the difference between mainland and coastal macro dialects influences null subject variation, which means that dialect and place of origin must be methodologically considered in sociolinguistics research

with bilingual communities as both aspects carry social value and can influence SPE differences. Accommodation plays a key role in bilingual communities. Orozco (2018) reached similar results on SPE comparing Barranquilla's and NYC's population. In NYC, the overt pronoun rate for Spanish/English bilinguals was like that of coastal Spanish. Furthermore, these new findings question previous studies that explained increasing overt pronoun frequencies in bilinguals as solely by English contact, with further sociolinguistic implications for social mobility, identity formation and community development.

Flores-Ferran (2004) is the first sociolinguist to offer empirical evidence supporting other factors besides English contact conditioning subject pronoun rates in Puerto Rican Spanish. Even though her findings still show English influence in the null vs. overt pronoun variation, they also suggest the relevance of social factors such as gender, age, and socioeconomic status in SPE. Variationist studies on SPE in Spanish/English bilinguals do not correlate with age, gender, or socioeconomic status. Rather, internal constraints like pronoun type (person, number), switch reference, lexical verb class, and tense-mood-aspect are the main linguistic features that affect SPE. And even though the overt pronoun rate increases among bilinguals with the length of U. S. residency, it remains to be understood what occurs inside bilingual communities.

Otheguy et al. (2007) documented a significant proficiency decline in second-generation Spanish speakers as some parents preferred to use only English with their children. Variable language input will have an impact on the maintenance of the Spanish language, which will depend on the individual and collective efforts within the community to transmit it to the second generation. Parents' motivation to avoid using Spanish at home may have to do with prestige, their resistance to their children grow up speaking English with a foreign accent, and Spanish-induced grammatical features that could result in discrimination, loss of career opportunities, and overall hardship in the United States.

Other researchers who focused on syntactic and cognitive aspects of bilingualism have evaluated other constraints on SPE. This line of research has looked for differential contextual factors in the speaker's upbringing, social interactions and language use while considering the linguistic structure of the Spanish and English spoken by heritage speakers that arise in bilingual contexts. Speakers' ideologies and identities play important roles in language behavior and language shift. Ramos (2014) showed that different cultural groups presented divergent values and interests in their Spanish language preservation. These sociolinguistic behavior patterns seemed motivated by the previous generation's ideologies around what "proper" or normative Spanish means. The fear of speaking broken Spanish instilled among heritage speakers inside and outside the community created a feeling of inadequacy. As a result, heritage Spanish speakers either avoided using Spanish altogether or adopted a standard variety that changed their perceptions toward other Latino groups.

Spanish heritage speakers acquire grammatical features that correlate with bilingual development. In describing the emergence of morphosyntactic competence in children, Erker (2015) stated its dependence on the type, time, and quality of exposure. Thus, it may be that the null subject required higher syntactic competence that could only be achieved through high exposure or language study. Montrul (2015) supported this hypothesis as she followed two

children from the same family who had different Spanish exposure levels. During her study, the author noted that the child who received higher parental instruction in Spanish had a more consistent and accurate use of null subjects. The younger child with less exposure to Spanish favored greater use of overt pronouns but did not follow the same internal constraints that benefit null pronominal subjects.

### 2.2. Unified syntax and sociolinguistics

The concept of a unified or shared syntax stems from the theory that there is a shared syntax across bilingual speakers' languages. It means that both languages are readily available at any given time, which allows the bilingual speaker to read, speak, and comprehend either language. Hartsuiker et al. (2004) laid out the methodological frame for studying shared syntax through an experiment in which they provided speakers with L1 priming in certain verb tenses that later were shown to influence L2 verb production. English/Spanish bilinguals participated in a description task where they produced oral picture descriptions in English based on cards provided by the researchers. The participants worked alongside a "confederate" (who acted as another participant) who described pictures in Spanish to them. The confederate was provided with pre-written descriptions that primed the participant with passive sentences also providing filler sentences in active, intransitive, and OVS sentences (Hartsuiker et al. 2004). They were shown images with a passive voice description in English. Participants who received the passive voice stimuli in English would more frequently describe their provided pictures using the passive voice in Spanish. However, when participants were presented with active, intransitive, or OVS sentences, they would use different sentence structures. Through this study, Hartsuiker concluded that for Spanish sentences to prime English sentences there had to be a shared syntactic activation process that included both languages as it would be impossible for separate systems to influence each other if they were not active at the same time.

Shin (2012), Helasvuo (2004), and Declerck et al. (2020) investigated other cognitive aspects of shared syntax activation by measuring the reading and response time of bilingual subjects when prompted with either of their two languages. They all concluded that there must be a shared cognitive space for language because both grammars were readily accessible. To explain the phenomenon, they proposed that the parallel model hypothesis and sentence superiority hypothesis could contribute to showing how shared syntax affects language acquisition. Both hypotheses derived from formal approaches in bilingualism that put forward its transversal influence variably across the phonology, morphology, and syntax of one or both languages. Depending on the speakers' frequency of use in the two languages, it may be possible for bilinguals to compartmentalize their L1 and L2 to carry out different tasks. For example, some speakers described their first language to be more akin to home or familiar spaces while their second language helped them communicate and grow in a globalized society. The bilingual mind also reflects an interaction between the two grammars, which can cause differences in reading time, voice onset time, grammatical structure, and syntactic categories.

The parallel model hypothesis proposes that language is accessed through parallel activations in the brain. This means that an auditory stimulus is processed parallel to cognitive understanding and lexical knowledge, which finally allows for processing the sentence that has

been heard. A similar process occurs when the speaker is about to produce a sentence (Traxler 2011). According to Shin (2012) and Helasvuo (2004), the parallel model hypothesis extends to L2 processing. In particular, they found that the response time for words and descriptions implies that any stimuli will be processed in parallel in both languages allowing the speaker to recognize the separate word and its meaning even if the language is switched mid-sentence.

The sentence superiority hypothesis proposes that readers will have difficulty recalling an ungrammatical word that appears in any presented sentence. This is attributed to the way that our brains process information; they tend to process in clusters instead of individual elements. The presence of "odd" elements in the sentence such as incorrect spelling, meaning, or syntactic position will cause them to be ignored or filled in with "appropriate" information. Through this process, Snell and Grainger (2017) have proposed that the sentence has priority in cognitive interpretation and that when readers are presented with a word in an expected space, they will show higher recall percentages if the "odd" word follows syntactic cues even if it is a different language from the rest of the sentence. This hypothesis favors a parallel sentence processing that activates syntax before L1 or L2 are considered.

The two hypotheses support the model of a competent speaker who can perform multiple multilingual tasks. This cognitive view opposes the common belief that sometimes describes bilinguals as "confused," "losing their language," or even "not real bilinguals" when they do not perform perfectly. From a sociolinguistic standpoint, one should not equate bilingualism with nativeness as a criterion for social acceptance or social prestige. In Otheguy et al. (2007), and Rourke and Potowsky's (2016) –two cross-generational studies conducted in different Latino bilingual communities,– speakers adapted to what their bilingual context required, forming a speech community that benefited multiple dialects according to their prestige; this meant that Mexican communities would accommodate toward Dominican or Puerto Rican groups (e.g., closer to what covert prestige demanded) to achieve a higher level of social acceptance in their bilingual groups.

Sociolinguistic research has demonstrated that several factors influence language variation, including time of residency, social network, age, gender, education, and age of exposure to the language (Traxler 2011). Adding the concept of unified syntax to those sociolinguistic variables might open a new outlook connecting sociolinguistic behavior with psycholinguistic behavior by allowing us to understand the bilingual speaker as both cognitively and socially proficient, and to explore whether those two related behaviors influence each other or not.

Current approaches to language acquisition, particularly in bilingual studies, have close associations with psycholinguistic research. Zyzik (2017) described the shift from generativist theories to social theories in language acquisition and highlighted the importance of this paradigmatic change for applied linguistics. Some researchers, like Showstack (2018), focused their research on what happens inside the American school system, which includes classroom settings, power inside institutions and language ideologies. These spaces are usually oriented toward the creation of a speech community that encourages and nurtures the students as they approach a second language. Although it cannot replicate what happens when language is taught naturally it is one of the most common ways to approach L2 learning. However, Belleti

et al. (2007) and Sorace and Serratrice (2009) have been more interested in developmental language cognitive features of bilinguals that focus on acquisition as an ability and is usually measured through testing and experiments, while also considering the social aspect of acquisition both from parental figures and teachers. The articulation point between social and cognitive constraints on language development is crucial for this study because available tools and experimental testing techniques may prove beneficial for understanding pronoun subject expression in Philadelphia's Colombian Spanish.

## 3. OpenSesame: Using computer-based software for testing

OpenSesame (Mathôt, Schreij and Theeuwes 2011) is an open-source software created with the intent of offering visual help to create experiments in psychology, linguistics, neuroscience, and experimental economics. OpenSesame provides a graphical interface and basic functions in Python and JavaScript to create data collection tests (Figure 1). It is equipped with a drag-and-drop function for visual scripting of commonly used experiments, and it also allows integration of other testing equipment such as eye-tracking devices, EEG etc.

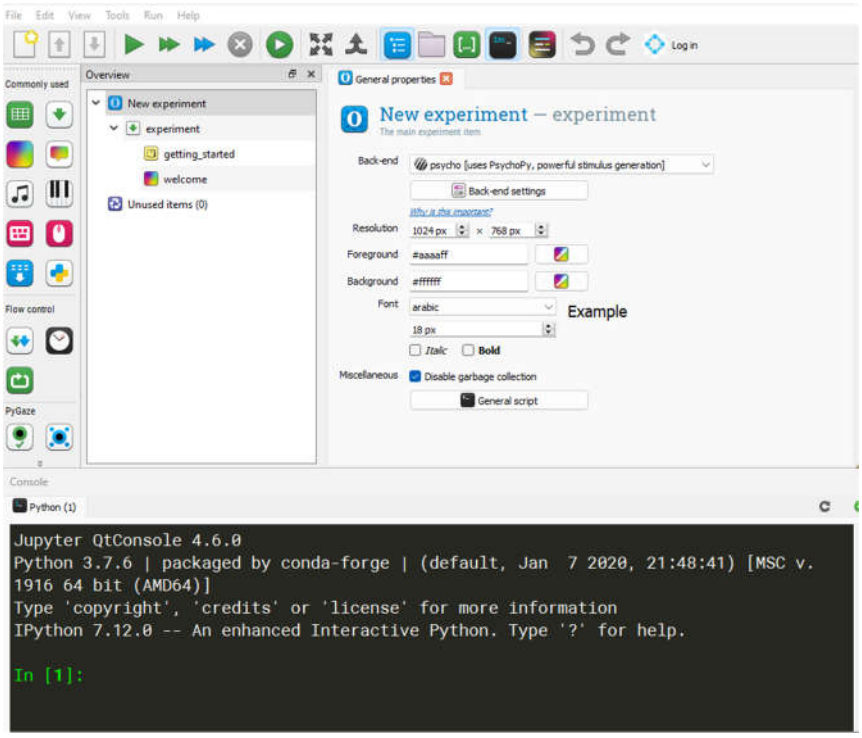


FIGURE 1. OPENSESAME

This article's Rapid Parallel Visual Presentation (RPVP) test was run using Joshua Snell's (2017) experiment, with some code changes for smoother processing and clearer randomization. This change was done by replacing a variable declaration system that hard-coded the pool data with a randomization of the pool, selecting sentences based on their grammatical order and reducing processing time (Figure 2). Instructions and data pool were translated from French to Spanish, while keeping the English sentences intact. Following Snell's previous work in sentence superiority effects I deemed the experiment helpful to account for bilingual variation. Here,

participants are required to type their answers in a box after being presented with a sentence, and they get visual feedback response of green (correct) or red (incorrect), while their accuracy is automatically calculated and given to the participant at the end of the test in a percentile.

```
path=exp.get_file('sentences1.txt')
with open(path) as file:
    stimuli1= file.readlines()
path=exp.get_file('sentences2.txt')
with open(path) as file:
    stimuli2= file.readlines()
path=exp.get_file('sentences3.txt')
with open(path) as file:
    stimuli3= file.readlines()
path=exp.get_file('sentences4.txt')
with open(path) as file:
    stimuli4= file.readlines()

for x in range(0,len(stimuli1)):
    stimuli1[x]=stimuli1[x].split('..')
    stimuli2[x]=stimuli2[x].split('..')
    stimuli3[x]=stimuli3[x].split('..')
    stimuli4[x]=stimuli4[x].split('..')

order = random.sample(range(0,49),25)

for x in range(len(stimuli1)):
    if x not in order:
        disord.append(x)
        random.shuffle(disord)

```

FIGURE 2. RANDOMIZING SEQUENCE

This type of software is very helpful for researchers who are interested in collecting psycholinguistic data but might lack coding experience. The visual scripting function creates a friendly learning environment. Additionally, it allows editing without affecting the functionality of the software. OpenSesame can also be run in a browser, which reduces stress among participants, and solves compatibility and participant availability issues.

## 4. Methodology

### 4.1. Data Collection

The current corpus is made up of two parts: Part 1. Sociolinguistic data with 17 recorded hours of sociolinguistic interviews and 32 successfully completed Bilingual Language Profiles and Social Network Questionnaires, and Part 2 with 32 Rapid Parallel Visual Presentation (RPVP) tests. This study has a total of 32 participants (14 Male and 18 Female), all required to be Colombian of birth or second-generation Colombian American. This is to account for different Colombian dialects that have differential SPE usage depending on the region of origin: Central regions have lower overt SPE and Coastal Regions have higher overt SPE. At this stage of the research, 11 interviews have been annotated and processed into the mixed effects model.

| Data Type | Sample Size | Type of annotation | Collection Method |
|---|---|---|---|
| Recorded Interviews. Verbal, Spanish | 17 hours | Broad transcription with SPE and Independent Linguistic Variable (Table 1) annotation | Sociolinguistic Interviews (Semi-structured) |
| Bilingual Language Profile. Written, Spanish | 32 Answered Profiles | Score from -200 to 200 | Online questionnaire completion |
| Social Network Questionnaire Written, Spanish | 32 Answered Questionnaires | Yes/No answers. Family composition (Family in the US, Family abroad) | Researcher wrote down the answers while asking the participant |
| Rapid Parallel Visual Presentation (RPVP) test | 32 tests | Recall percentage | Participants took the test in-person using a computer |

TABLE 1. CORPUS STRUCTURE

Participants were assigned to four types of groups: A, B, C and D. Groups A, B and C are based on each speaker's time of arrival to the United States (1-4 years, 5-10, years, 10+ years, respectively) and Group D is considered the baseline group consisting of eight participants from Colombia who have not resided in the United States or any other country but have received English as a second language education. Although not intentional, there is an apparent age separation in these groups; A : 20-30 years old, B 31-45 years old and C 50+ years old. D has a more variable population with ages from 21 to 35 years.

Participants also completed two questionnaires. The Bilingual Language Profile (University of Texas, 2012) quantifies the self-perceived proficiency and usage of each language on a bilingual speaker. This provides each speaker with a three-digit score from -200 to 0 and 0 to 200;. The closer a speaker is to 0, the more *balanced* their language use; on the other if the speakers scored negative numbers, their L2 is more prominent and if they score closer to positive number their L1 is more prominent. Alongside the Bilingual Language Profile, participants also completed a social network questionnaire that described which language they used and their social ties to other people in their families, neighborhoods, and workplaces.

All sociolinguistic interview data was recorded in Spanish using Zoom for online interviews or a basic recorder for in-person interviews. It follows a semi-structured interview with 19 questions where speakers were free to add any information aside from prepared questions about their arrival and adjustments to the United States, their family, work experiences in the

United States and in Colombia, and other topics that facilitated the production of informal language. Some of the questions were: *¿Cuál ha sido su experiencia en Philadelphia? ¿Tiene anécdotas interesantes de su primera llegada a los Estados Unidos? ¿Ha tenido trabajos en la ciudad últimamente?*. These types of questions usually prompt the speaker to talk about their own experiences, feelings and thoughts, which prompt different discourse moods (indicative, subjunctive or imperative) and different references. Some prototypical answers elicited during the sociolinguistic interviews were as follows:

Question: *¿Cuál ha sido su experiencia en Philadelphia?*

Female Speaker, Group A.
*Bueno, la verdad es que Filadelfia para mi es medio fea ¿Si me entiendes? Es una ciudad sucia y la gente es como no sé, grosera, no te hablan bien y no te saludan. Nunca pensé que viviría en una ciudad así.*

Male Speaker, Group A.
*A mí me encanta Filadelfia, es una ciudad muy chévere y tengo muchos amigos en el North. Yo hablo mucho con gente de dominicana y boricuas y me encanta su comida. Yo a veces me voy a nueva york con ellos y me gusta mucho, yo quiero vivir allá.*

Female Speaker, Group B.
*Por ejemplo, en el trabajo pues me ha gustado porque pues conozco gente latina de toda clase de eh de América Latina, pero siempre muchas personas que no son profesionales o educadas o no tanto educadas, sino formadas formadas que es diferente formadas.*

Male Speaker, Group C.
*La verdad es que no salgo mucho de mi casa porque no me gusta la gente, trabajo remoto desde la pandemia, me gusta salir a los parques de la ciudad.*

Unlike Groups A, B and C, the interviews to Group D had16 questions, specifically removing questions about international cities.

Question:*¿Se considera un buen hablante en ambas lenguas?*

Female speaker Group D
*No, A pesar que he estado constantemente estudiando y todo eso, a mí me pasa que si no soy muy constante con algo se me olvida. Entonces, después de lo de Open English duré como casi seis cuatro meses sin haber tomado otra vez una clase de esas. Entonces eso ya de por sí me ayudó a que hubiera muchas palabras que se me olvidarán.*

After transcribing the data, we need to obtain the SPE percentage, each sentence is annotated to count SPE occurrences and other surrounding aspects of the sentence, sentences are delimited

by subject pronouns, nominalizations are excluded as well as non-specific subjects. SPE is annotated until the interview is finished.

Example of annotation:

*Yo a veces me voy a nueva york con ellos*

*OvertSPE*
*First Person Singular*
*No Reflexive*
*Indicative mood*

*Quiero que trabaje remoto desde la pandemia*

*NullSPE*
*First person Singular*
*No reflexive*
*Subjunctive mood*

The Rapid Parallel Visual Presentation test (Section 4.2) is used in this study to avoid overt syntactic priming in favor of the sentence superiority hypothesis (Declerck 2020). Following previous works on the sentence superiority hypothesis, I presented participants with multiple four-word-long grammatical (*you ojos son green*) and ungrammatical (*your son green ojos*) bilingual sentences at 200 ms and asked them to recall a random word under,. For a participant to recall a sentence, they need to access not only their memory but infer the position of the word based on syntactic constraints and their own knowledge of the language. RPVP tests provide accuracy percentages that range from 20% to 71% recall for each speaker.

To account for these multiple effects, I proposed using a mixed effect logistic regression model (Section 4.2) expecting to observe significant results that relate the reading recall percentage variable to time of residency, age, region, and social networks, as well as the Bilingual Language Profile score. This is mostly attributed to the opportunities the speaker will have to interact in both languages and the time that they have had to adapt and learn from other speakers. Regarding linguistic variables, I expect that higher complexity aspects such as switch reference, genre of discourse, and tense-mood-aspect have higher significance when in the presence of the reading recall test since those variables have proven to be related to language knowledge and proficiency. Since unified syntax appears to be related to linguistic development and cognitive learning from individuals, we expect to find a correlation to higher-complexity linguistic variables and time-related sociolinguistic variables as speakers will need more time of exposition to experiment and modify their syntactic and social behavior.

### 4.2. Mixed-effect Logistic Regression Model

A mixed effect logistic regression model can find predictors of change for binary dependent variables in both categorical and random variables, this means that both fixed (gender, time of arrival, time of residency, TMA, etc.) and random (Reading Recall Percentage, BLP scores) can be mixed in one model to see the effects on a binary dependent variable (Overt Subject, vs Null Subject). The mixed effect model is better suited for semi-structured interviews, as it can interpret data with different entries for each speaker and it fits the participant's linguistic behavior more naturally than just logistic regression since mixed effects can interpret significance in unbalanced data.

| **Dependent Variables** | **Independent Sociolinguistic Variables** | **Independent Linguistic variables** | **Random Variables** |
|---|---|---|---|
| Subject Pronoun Expression: Overt vs. Null | Age | Genre of discourse | Reading Recall Percentage |
| | Gender | Person and singular/plural of verb | |
| | Time of Arrival to the US | Tense-Mood-Aspect | |
| | Time of residency | Reflexive/non-reflexive | |
| | Social Networks | Priming | |
| | Occupation | Discourse connection between verbs (switch reference/same reference) | Bilingual Language Profile (BLP) Score |
| | English and Spanish Self-Evaluation | Type of lexical content of the verb (Lexical) | |
| | Group (A, B, C or D) | Appearance of the verb in a set phrase | |
| | Region | Section of interview where it appears | |

TABLE 2. TOTAL ANALIZED VARIABLES FOR MIXED EFFECTS MODEL

Table 2 has the classification of every annotated variable included in the corpus and by using inferential statistics one can calculate each Independent Variable (IV) significance based on their p-value and probability (logodds) of creating change on the Dependent Variable (DV). Significance is achieved when p-value < 0.05. The model is run in R, using the rbrul tool (Johnson 2008) which uses the following libraries: rlang, Rcpp, lme4, lmerTest, MuMIn, readr, shiny, tools and lattice. This tool is commonly used in variationist linguistic research, and the

shiny package also provides visual scripting in R (Figure 3) which is very helpful for conference presentations and quick data analysis.


FIGURE 3. SHINY RBRUL

Categorical variables are square; round variables are continuous; and random variables are round and transparent. The DV is situated in the response column, predictors in the center and potential predictors in the potential column. The model is updated in real-time. Interactions can be created by overlapping variables. The tool can also be run in regular R code.

### 4.3. Rapid Parallel Visual Presentation (RPVP) or Reading Recall Test

Rapid Parallel Visual Presentation (RPVP) or Reading Recall, is the main cognitive test used in this study. Participants were asked to read different sentences at 200ms and then were asked to immediately recall a specific highlighted word that might be presented in either grammatical or ungrammatical positions. Every participant took the test by sitting in front of a computer; the sentences are presented through an OpenSesame program that provides them with around 120 sentences in random order at 200ms. Sentences are divided into four groups of 30: Group 1, unified syntax sentences, which have L2 words in grammatical positions; Group 2, sentences with L2 words in agrammatical positions; Group 3, L1 sentences with grammatical positions; and Group 4, L1 sentences with agrammatical positions.

After 200ms the sentence is replaced by hash signs (#) and a dot on top of one of the words (Figure 4). Participants will have to recall this word and type it again; speed of recall is not part of the test. I found higher recall in Groups 1 and 3, and lower recall in Groups 2 and 4.

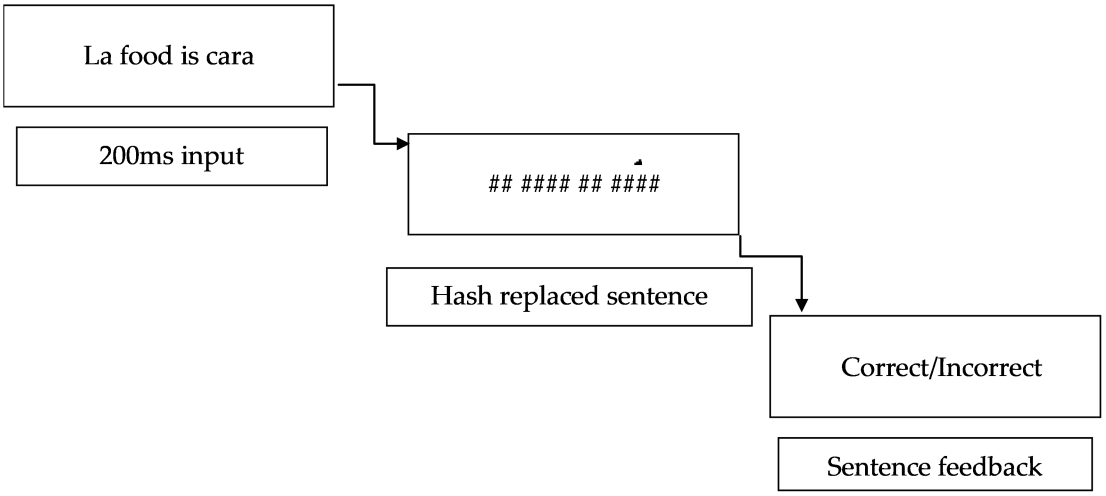


FIGURE 4. READING RECALL PROCEDURE

Incorporating reading recall accuracy into SPE variation seems to be a natural step after observing the development of SPE research, especially in bilingual territories like the United States. A globalized world where bilingual policies become widespread requires additional instruments to account for unified syntax in bilingual speakers. A computer-based test can provide additional random and categorical data to strengthen statistical predictors.

## 5. Results

### 5.1. Overview

To test the model accuracy, I conducted a mixed effects model run with 12 participants, which yielded 1237 tokens of variating use of the null vs overt subject. The most stable model run is the following:

model formula: nullvsovert~ SPANFriendshipContact + Gender + personandnumber + SocialNetworkType + semanticverbtype + Tense + TimeOfResidency + (ReadingRecall | dominancescore)

All the IVs presented in the formula were deemed significant by the model, with values of $p < 0.05$. As a preliminary run, we see promising findings in Table 3. Considering that Spanish is classified as a pro-drop language, my main category was Null subject, which was then compared to the behavior of Overt subjects. In Rbrul, data is presented with the classifiers logodds, n, proportion, and factor weight, described below:

- Logodds refers to a probability coefficient that represents the likelihood of an event occurring (in this case, the variation between null and overt pronouns), divided by the probability of said variation not occurring, positive values mean that the variable favors

the variation from null to overt, 0 means neutral effect and negative values mean that the variable does not favor variation from null to overt.

- N is the total number of tokens.
- Proportion is the stress shift for each factor level (meaning at what percentage the category shifts from null to overt).
- Factor weight is a probability from 0 to 1 of the exponential effect of each category in the null vs overt variation.

***Null/ (null+ Overt)***

| **Friendship Contact** | **logodds** | **n** | **proportion** | **factor.weight** |
|---|---|---|---|---|
| *Semanal* | 2.233 | 573 | 60% | 0.903 |
| *Mensual* | -0.162 | 257 | 66% | 0.46 |
| *Diario* | -2.071 | 274 | 52% | 0.112 |
| **Gender** | **logodds** | **n** | **proportion** | **factor.weight** |
| *Masculino* | 2 | 208 | 70% | 0.888 |
| *Femenino* | -2 | 896 | 57% | 0.112 |
| **Person and number** | **logodds** | **n** | **proportion** | **factor.weight** |
| *PPP* | 3.3 | 86 | 81% | 0.966 |
| *TPP* | 2.9 | 80 | 73% | 0.95 |
| *PPS* | 2.1 | 767 | 59% | 0.895 |
| *TPS* | 1.6 | 135 | 48% | 0.84 |
| *SPS* | 1.09 | 33 | 30% | 0.749 |
| **Reading Recall** | **logodds** | **n** | **proportion** | **factor.weight** |
| *60to80* | 0.91 | 313 | 65% | 0.713 |
| *40to60* | 0.04 | 433 | 62% | 0.512 |
| *20to40* | -0.95 | 358 | 51% | 0.277 |
| **Social network** | **logodds** | **n** | **proportion** | **factor.weight** |
| *americana* | 2.4 | 483 | 56% | 0.917 |
| *Mayorialatina* | -0.2 | 292 | 62% | 0.444 |
| *mixta* | -0.2 | 329 | 62% | 0.101 |
| **Semantic verb type** | **logodds** | **n** | **proportion** | **factor.weight** |
| *other* | 0.23 | 600 | 64% | 0.559 |
| *motion* | 0.02 | 133 | 63% | 0.507 |
| *cognitive* | -0.26 | 371 | 50% | 0.434 |
| **Tense** | **logodds** | **n** | **proportion** | **factor.weight** |
| *gerundio* | 1.304 | 19 | 84% | 0.787 |
| *perfecto* | 0.692 | 23 | 82% | 0.666 |
| *preterito* | -0.05 | 269 | 68% | 0.688 |
| *presente* | -0.555 | 687 | 55% | 0.365 |
| *futuro* | -0.65 | 5 | 60% | 0.343 |
| *imperfecto* | -0.741 | 101 | 55% | 0.323 |
| **Time Of Residency** | **logodds** | **n** | **proportion** | **factor.weight** |
| *GroupB 5- 10 years* | 2.148 | 477 | 52% | 0.895 |
| *GroupA 1- 4 years* | 0.992 | 399 | 66% | 0.729 |
| *Group C 10+ years* | -3.14 | 228 | 61% | 0.0415 |
| **BLP Dominance Score + Reading Recall** | **intercept** | **n** | | **proportion** |
| *std deviation* | 0.2 | 1104 | | 59% |

TABLE 3. MIXED EFFECT MODEL

In this model, 50% proportion is considered traditional SPE usage, since previous studies have stated that Spanish is a pro-drop language with a 50% distribution on null and overt subjects.

Significance was found in traditional linguistic and sociolinguistic variables for SPE: Tense, Semantic Verb type, Person and Number and Gender. Additionally, there was significance in Social Networks, Time of Residency, Friendship Contact and Reading Recall. When adding the random slope of BLP Dominance Score + Reading Recall results, we see a standard deviation of 0.2, which is promising, but still low for the expected data prediction. This is probably due to the small sample size used for this first run and will probably increase as more data is fed to the model.

### 5.2. Sociolinguistic and linguistic results

During the semi-structured interviews, participants were prompted to talk about their social circle and to comment on their observed social networks (mainly American, mainly Hispanic, or mixed) as well as how often they contacted their friends, family, and coworkers. From these variables, only friendship contact was significant, with more overt pronoun usage occurring when participants had less contact with their friends. Additionally, both mixed and mainly Hispanic social networks are relevant in the appearance of overt subjects. However, American social networks have little effect on SPE variation. Gender was also significant, with women presenting a more conservative approach to subject pronoun variation with a 57% proportion, and men using more overt pronouns with a 70% proportion; Time Of Residency is also significant, with a U slope behavior, with higher overt SPE in Group A and Group C and standard behavior for Group B. For these two last variables, more data is needed to see how gender affects SPE variation, as data is currently unbalanced with some categories having more tokens than others. Since Group D (baseline) was not added to this model, a future run will also include these results.

Time of Residency has a particular U-shaped distribution curve. This shape suggests that there is a collective event in Philadephia that happened 5-10 years ago that changed the use of overt pronouns.to better understand this variable, it is necessary to look at the collective experience of each immigrant group. Because the current data is only comprised of 11 out of 32 participants, it is possible that this variable's behavior will change in the future.

However, the significance of friendship contact and social networks in SPE may be supported by anecdotal experiences provided by participants. Participants had positive, negative, or neutral thoughts about Philadelphia and its inhabitants Participants with a negative perception of Hispanic or American groups would consciously avoid either fully English-speaking or fully Spanish-speaking groups, which are logged in the Mainly Hispanic/American groups. Motives were usually associated with cultural perception and previous positive or discriminatory experiences with either group. For example, if a participant had been discriminated against due to their accent, skin color, or heritage, they would often isolate themselves from the offending group. Positive experiences resulted in more willingness to engage with other cultures, creating a mixed friendship group.

Significant linguistic variables were Tense, Person and Number, and Semantic Verb Type. These variables have been found to be significant across the sociolinguistic literature on SPE variation. These variables tend to be generally recognized as significant since overt pronoun usage can help avoid referential ambiguity and null pronoun usage is highly context-dependent in naturalistic discourse.

### 5.3. Reading Recall and BLP Dominance Score Results

Reading recall accuracy scores were significant in the mixed-effect model. I intentionally recoded the variable since the sample size was relatively small and p values were very small ($p^{-10}$). Recoding reduces the variability in cells but can still successfully predict significance. The three reading recall groups were separated from the lowest score to the highest score in 20 percentage point intervals. I observed an increase in proportion values as participants scored higher in their tests. This variable also interacts with each participant's BLP score, with a slight increase in overt subject pronoun usage as the participant considers themselves more confident in their bilingual language use.

When reading recall is incorporated as a random variable that interacts with BLP dominance scores, the model estimates a slight slope that slowly increases the usage of overt pronouns as

Reading recall is higher despite BLP scores. BLP scores seem to require more variation to create a more accurate slope. It would be tempting to say that higher recall equates to higher overt pronoun usage due to the syntactic integration of a pro-drop and non-drop language. The data suggest that this cognitive adaptation is more complex than just a combination of two systems. I believe that Reading Recall exhibits promising predicting behavior and it would be very interesting to observe reading recall accuracy as a dependent variable in future analysis, this time, with a multilinear model to find predictors for non-binary categories.

It is worth noting that older participants struggled with understanding the functionality of the reading recall test, so its accessibility must be considered when interpreting data. Future iterations of the test should have longer trial loops.

### 5.4. Discussion

Current data suggests that there is a strong relationship between social interactions, bilingual confidence, and cognitive recall accuracy. Most importantly, data suggests that this variation is not necessarily linked to English exposure, but rather, to multicultural and multidialectal exposure. Participants that had bilingual or Hispanic friendship groups increased their use of overt pronouns, while participants that had mostly monolingual English-speaking groups were more likely to exhibit traditional SPE distribution. So far, this suggests that both English and Spanish remain separate codes for sociolinguistic variation, all while participants also experience the sentence superiority effect result of bilingual language acquisition.

Spanish variation in the U.S. that is independent from English has also been analyzed and theorized by other authors such as Flores-Ferran (2004) and Orozco (2018), who have also made convincing arguments about the variability of Spanish and its vitality in the United States. This article contributes to the field since it debates first language attrition as an inevitable result of

immigration and proposes a more social and flexible view of language change in Spanish-speaking immigrants in the U.S. I also present trends that were previously observed in Orozco (2018), where Colombian Spanish variation in the U.S. behaves similarly to SPE rates of Caribbean Spanish. The influence of Reading Recall in overt SPE variation shows that it is possible to acquire cognitive strategies to successfully recognize bilingual patterns while also learning and acquiring new dialectal patterns, by looking at other data descriptors we can safely assume that this change has no direct correlation to English exposure.

## 6. Conclusions

This study indicates a relationship between Null vs Overt Pronouns, sociolinguistic variables and reading recall test results as observed in the following aspects of the participants as language users :

(1) Participants who engage in positive bilingual or Spanish-Monolingual social interactions more often (especially in informal contexts) will see a 10-17% increase of the proportional use of overt pronouns. This is specific to participants who describe their social circle as mainly Hispanic or Mixed.
(2) Participants who have been in the U.S. for less than 5 years or more than 10 years will see an increase in their overt pronoun use. Participants who have been in the U.S. for 5-10 years will see no change in their overt pronoun usage.
(3) The interaction between higher dominance scores and reading recall still has to be observed. However, it is an encouraging finding in both its relationship to SPE variation and its behavior for predicting change as a random variable in mixed-effect models. Further analysis is needed.
(4) SPE variation behaves similarly to other Hispanic communities in New York, with an increase of overt subject pronouns. The influence of English contact is not conclusive at this point.

In conclusion, this study highlights the importance of positive social interactions and social context in shaping overt pronoun usage among bilingual individuals. The duration of stay in the U.S. also plays a significant role in language behavior. Furthermore, the preliminary findings suggest potential relationships between BLP dominance score, reading recall, and speech variation, which warrant further investigation. Future studies should incorporate larger datasets and more comprehensive data analysis techniques to provide a deeper understanding of these dynamics.

## References

Barrera-Tobón, Carolina and Rocío Raña-Risso. 2016. "A Corpus-based Sociolinguistic Study of Contact-induced Changes in Subject Placement in the Spanish of New York City Bilinguals". In *Spanish Language and Sociolinguistic Analysis*, edited by Sandro Sessarego and Fernando Tejedo-Herrero, 323-342. Amsterdam: John Benjamins. https://doi.org/10.1075/ihll.8.14bar

Belletti, Adriana, Elisa Bennati, and Antonella Sorace. 2007. "Theoretical and Developmental Issues in the Syntax of Subjects: Evidence from near-Native Italian." *Natural Language and Linguistic Theory* 25 (4): 657-689. https://doi.org/10.1007/s11049-007-9026-9.

Declerck, Mathieu, Yun Wen, Joshua Snell, Gabriela Meade, and Jonathan Grainger. 2020. "Unified Syntax in the Bilingual Mind." *Psychonomic Bulletin and Review* 27 (1): 149–154. https://doi.org/10.3758/s13423-019-01666-x.

Flores-Ferrán, Nydia. 2004. "Spanish Subject Personal Pronoun Use in New York City Puerto Ricans: Can We Rest the Case of English Contact?" *Language Variation and Change* 16 (1): 49–73. https://doi.org/10.1017/S0954394504161048.

Hartsuiker, Robert J., Martin J. Pickering, and Eline Veltkamp. 2004. "Is Syntax Separate or Shared between Languages? Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals." *Psychological Science* 15 (6): 409–414. https://doi.org/10.1111/j.0956-7976.2004.00693.x.

Helasvuo, Marja Liisa. 2004. "Shared Syntax: The Grammar of Co-Constructions." *Journal of Pragmatics* 36 (8): 1315-1336. https://doi.org/10.1016/j.pragma.2004.05.007.

Johnson, Daniel Ezra. 2008. "Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis." *Language and Linguistics Compass* 3 (1): 359–383. https://doi.org/10.1111/j.1749-818x.2008.00108.x.

Lagrou, Evelyne, Robert J. Hartsuiker, and Wouter Duyck. 2015. "Do Semantic Sentence Constraint and L2 Proficiency Influence Language Selectivity of Lexical Access in Native Language Listening?" *Journal of Experimental Psychology: Human Perception and Performance* 41 (6): 1524-1538. https://doi.org/10.1037/a0039782.

Lastra, Yolanda, and Pedro Martín Butragueño. 2015. "Subject Pronoun Expression in Oral Mexican Spanish." In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, edited by Ana M. Carvalho, Rafael Orozco, and Naomi Lapidus Shin, 39-58. Georgetown University Press.

Mathôt, Sebastiaan, Daniel Schreij, and Jan Theeuwes. 2011. "OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences." *Behavior Research Methods* 44 (2): 314–324. https://doi.org/10.3758/s13428-011-0168-7.

O'Rourke, Erin, and Kim Potowski. 2016. "Phonetic Accommodation in a Situation of Spanish Dialect Contact: Coda /s/ and /ℝ/ in Chicago." *Studies in Hispanic and Lusophone Linguistics* 9 (2): 355-399. https://doi.org/10.1515/shll-2016-0015.

Orozco, Rafael. 2015. "Pronominal Variation in Colombian Costeño Spanish." In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, edited by Ana M. Carvalho, Rafael Orozco, and Naomi Lapidus Shin, 17-38. Georgetown University Press.

Orozco, Rafael. 2018. *Spanish in Colombia and New York City*. Vol. 46. IMPACT: Studies in

Language and Society. Amsterdam: John Benjamins. https://doi.org/10.1075/impact.46.

Otheguy, Ricardo, Ana Celia Zentella, and David Livert. 2007. "Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community." *Language* 83 (4): 770-802. https://doi.org/10.1353/lan.2008.0019.

Potowski, Kim. 2007. "Characteristics of the Spanish Grammar and Sociolinguistic Proficiency of Dual Immersion Graduates." *Spanish in Context* 4 (2): 187–216. https://doi.org/10.1075/sic.4.2.04pot.

Shin, Jeong Ah, and Kiel Christianson. 2012. "Structural Priming and Second Language Learning." *Language Learning* 62 (3): 931-964. https://doi.org/10.1111/j.1467-9922.2011.00657.x.

Shin, Naomi Lapidus, and Daniel Erker. 2015. "The Emergence of Structured Variability in Morphosyntax: Childhood Acquisition of Spanish Subject Pronouns." In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, edited by Ana M. Carvalho, Rafael Orozco, and Naomi Lapidus Shin, 169-190. Georgetown University Press.

Showstack, Rachel. 2018. "Spanish and Identity among Latin@s in the U.S." In *The Routledge Handbook of Spanish as a Heritage Language*, edited by Kim Potowski, 106-120. London: Routledge https://doi.org/10.4324/9781315735139-7.

Snell, Joshua, and Jonathan Grainger. 2017. "The Sentence Superiority Effect Revisited." *Cognition* 168: 217–221. https://doi.org/10.1016/j.cognition.2017.07.003.

Sorace, Antonella, and Ludovica Serratrice. 2009. "Internal and External Interfaces in Bilingual Language Development: Beyond Structural Overlap." *International Journal of Bilingualism* 13 (2): 195-210. https://doi.org/10.1177/1367006909339810.

Wen, Yun, Joshua Snell, and Jonathan Grainger. 2019. "Parallel, Cascaded, Interactive Processing of Words during Sentence Reading." *Cognition* 189: 221-226. https://doi.org/10.1016/j.cognition.2019.04.013.