Language variation in U.S. Spanish: Social factors

Rena Torres Cacoullos & Grant M. Berry

Abstract

In this chapter, we survey research testing correlations of linguistic behavior with speakers' sociodemographic characteristics. The effects of social factors suggest stable variation in some cases, and changes in progress in others. In cases of stable variation, the patterns reported correspond to those found throughout the Spanish-speaking world. For example, women tend to use standard variants, such as sibilant rather than aspirated syllable-final /s/, more than men do among Puerto Ricans in Philadelphia. Change in progress in immigrant communities may be detected synchronically in apparent time and inferred from comparison with variation patterns in communities of origin. One scenario may be the receding of a phonological variant in both its overall rate and associations with social factors, as with the decline of hiatus-breaking [j̯] (*cayer* from *caer* 'to fall') in Salvadoran communities in Houston. Another may be the spread of a variant and the development of social conditioning, for example, the reduction of intervocalic ⟨ll⟩ (*eØa* from *ella* 'she/her') among Mexicans and Salvadorans living in the same neighborhoods in Houston. Contributing to the paucity of studies on social factors in U.S. Spanish has been the familiar problem that social characteristics of speakers are generally less well-defined than linguistic categories, particularly in minority language situations, and that social factors are often highly interdependent. A solution to this conundrum can be found in principal component analysis, as a heuristic for grouping speakers strictly on the basis of their linguistic behavior, demonstrated here in a community-based corpus of New Mexican Spanish.

1. The social context of Spanish in the U.S.: conceptual and methodological challenges

The neglect of sociolinguistic patterns of language variation in U.S. Spanish is due at least in part to the abiding preoccupation with contact-induced change and a methodological predilection for acceptability judgments, experimental tasks, or cherry-picked examples. This "Hispanic tradition" of language study as Bills (1975:vi-vii) characterized it nearly half a century ago is hampered by an "interest in the accumulation of speech fragments with little concern for linguistic or sociological context" and "almost exclusive interest in deviations from standard Spanish". Adherence to the analyst's idealizations as the benchmark for evaluations leaves working-class varieties of U.S. Spanish in a no-win situation, as pointed out by Ana Celia Zentella; for example, New Mexican Spanish is branded 'archaic' "porque se describe en referencia a la norma de otra comunidad [because it is described in reference to the norm of another community]," but at the same time "tampoco se vale ser innovador…al notar la reacción…en contra de…*lonche*…y otros préstamos [nor is it worthy to be innovative...noting the reaction...against...*lonche*... and other borrowings]" (Zentella 1990:157).

Methodological issues begin with data collection. Appropriate data come from "Labovian-type studies of Chicano speech in a natural setting," as Peñalosa (1981:7) asserted, holding that "experimental and interview settings…bias findings." This is because the **vernacular**—the unreflecting use of language in the absence of the observer, when minimum attention is paid to monitoring speech—is the style that is most regular in structure (Labov 1972:112). In contrast, when speakers of subordinate varieties are asked direct questions about their language, as is the case with acceptability judgments, their answers shift toward (or away from) the prestige variety in irregular and unforeseeable ways (Labov 1972:111). Whether data are gathered by an in- or out-group member has also been demonstrated to make a difference, for example, in rates of word-final nasal velarization (e.g., *fuero*[ŋ] *los)* by Salvadorans interviewed by a Mexican in Houston (Hernández 2011) or in the frequency and kinds of code-switching in a bilingual Puerto Rican community in New York (Poplack 1993:259).

When we turn our attention to continuity with Spanish varieties spoken across the Americas and to community-based samples of vernacular speech (e.g., Otheguy and Zentella 2012), social factors in tandem with linguistic constraints become important for diagnosing stability vs. change in U.S. Spanish varieties; these factors also serve to detect parallels vs. divergences vis-à-vis Spanish varieties outside the U.S. Moreover, even for assessing contact-induced change, it has become clear that neglect of the social context of bilingualism is risky, because once speakers are adequately characterized with respect to social factors, phenomena attributed either to majority language influence or to minority language loss may turn out to be conditioned by social class instead.

An instructive example concerns use of the subjunctive mood, presumed to be undergoing attrition among speakers of Spanish in the U.S. due to contact with English (e.g., Ocampo 1990). To test a parallel presumption about the French subjunctive in Canada, Poplack (1997) considered external measures of contact at the individual and community level. If contact with English is playing a role, speakers with higher English proficiency and those living in neighborhoods with a higher proportion of English speakers should show a lower rate of the subjunctive than speakers with lower indices of contact with English. Neither measure of contact correlated with subjunctive rate. Instead, after accounting for a strong lexical effect of the governing verb, systematic quantitative analysis of both internal and external constraints exposed the unsuspected effect of social class, with professionals displaying a proclivity for the subjunctive (Poplack and Levey 2010:402-404). Moreover, subjunctive use is characterized not by change, but by long-term stable variability, despite centuries of normative injunctions (Poplack, Lealess and Dion 2013).

In this chapter, we survey the scant number of reports on social factors in U.S. Spanish, first for **stable linguistic variables**—those with distribution patterns that persist across time and communities—and then for possible **changes in progress**—where age distributions are gradient. In the final section, we apply statistical procedures (principal component analysis and regression analysis) to a community-based corpus of New Mexican Spanish in order to infer and test social factors relevant to conditioning language variation.

2. Social class and gender in stable variation

The "central dogma" of sociolinguistics stated by Labov is that "the community is prior to the individual" (Labov 2006:5). Individual speaker behavior can be understood only once the community pattern is known, since individual linguistic behavior results from social histories and memberships. For Spanish-speaking communities in the U.S., the social factor most considered, though not always examined, has been speaker gender. Usually, gender is implicated in claims of changes in progress, with women being seen in some cases as conservative and in others as leaders of linguistic change. Hasty pronouncements of change can stem from equating language change with perceived departures from an idealized norm or even with ordinary variation; that is, failing to differentiate between situations of stable variation and situations of change in progress. Crucially, while all change implies the existence of variability in language, the converse is not true: "not all variability and heterogeneity in language structure involves change" (Weinreich, Labov and Herzog 1968:188).

The "gender paradox" is the pattern of gender differentiation whereby "women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not" (Labov 2001:292-293). This follows from two generalizations for the distinct scenarios of variation: women use stigmatized variants at lower rates than men for stable sociolinguistic variables, but adopt innovative forms earlier, both for prestige variants and for linguistic **changes from below** (i.e, changes from within the system which occur below the level of conscious awareness and consequently lack style shifting in initial stages) (Labov 2001:261-293).

A stable sociolinguistic variable in Spanish is variation in the realization of the forms of the copula *estar.* This dates back approximately half a millennium, judging from the recommendation of Juan de Valdés (1535) that the verb be written without *e-* to distinguish it from the demonstrative pronoun:

> "me ha parecido, por no hacer tropezar al letor, poner la e cuando son pronombres, porque el acento está en ella, y quitarla cuando son verbos, porque, estando el acento en la última, si miráis en ello, la primera e casi no se pronuncia, aunque se escriba [It has seemed a good idea to me, in order not to trip up the reader, to put the letter *e* before the words that are pronouns, because it is accented there, and not to include it in verbs, because, being that the accent is on the last syllable, if you look at it, the first *e* is almost not pronounced, even though it is written] ".

Drawing on 32 interviews recorded in San Antonio with Mexican-Americans raised in South Texas, Garcia and Tallon (2000) examined three variants of *estar*: *está*, *'stá* (with apheresis of the vowel) and *'tá* (N=1,025) in multivariate analysis. They find phonological conditioning by preceding segment. In accordance with the generalization for stable variation, female speakers favored "canonical" *está*, while males favored the *'tá* variant, leading the authors to suggest that the latter may be "the least formal variant" and "a marker of male speech" (García and Tallon 2000: 356-357).

Another variable showing social stratification is the alternation between *para* and *pa'*. This is conditioned by morphosyntactic and phonological, but also social, factors. A study based on data extracted from recordings with 171 speakers in San Antonio (token numbers not reported), indicates that males use the abbreviated form of the preposition at nearly double the rate of females (42% vs. 23%) (Lantolf 1982). What is more, gender interacts with education and occupation: although males display a higher rate of *pa'* across all occupation and education levels, the gender difference is as high as 48 percentage points for blue-collar workers and as low as 5 for professionals, with a 20-point difference within white-collar workers (the largest group sampled) (Lantolf 1982:172). Considering internal factors, the reduced form was favored in directional (locative) uses (Lantolf 1982:167).

The social and linguistic conditioning of *para ~ pa'* in San Antonio parallels that found outside the United States. Based on two analyses in Venezuelan corpora (48 speakers and 1599 tokens in one, 72 speakers and 2144 tokens in the other), Bentivoglio and Sedano (2011:169-171) report that expressions of directionality are favorable contexts for *pa'* (*me fui pa Nueva York* 'I went to New York') while *para* is preferred for purposives (*para terminar* 'in order to finish') and furthermore that following consonants promote the abbreviated form (as in *pa' comprar* 'in order to buy'), but that the strongest effect is that of socioeconomic level: the "low level" showed the highest rates of the reduced variant *pa'*.

A class-based account is also proposed for a higher Spanish subject pronoun rate among Colombians and possibly Cubans who have lived in New York City for more than five years as compared with newcomers from those same countries. Shin and Otheguy (2013:442-443) point to the high affluence rankings of these Latino national-origin groups in census data, offering the conjecture that affluent Latinos are susceptible to influence from English due to looser social networks and more interaction with speakers of English. However, among the Colombians and Cubans sampled (N=45), no effect is found for social class or education. Nevertheless, a "woman effect" is reported, which is most pronounced among those who were Latin-American born but had lived in NYC more than five years (Shin and Otheguy 2013:439). While the gender effect among Colombians and Cubans may be because, as suggested by Shin and Otheguy (2013:446), women have more contact with US-born children or friends than men do, it also appears to be the case that women have higher rates of pronominal subject expression than men in Colombia to begin with (Orozco 2015:30; see also Martín Butragueño and Lastra 2015:50).

Understanding variation in Spanish among Latino New Yorkers necessitates knowledge of their social context (Otheguy and Zentella 2012:149-150), in the same way that it is imperative to distinguish language contact settings due to immigration, as a result of conquest, or across national boundaries (Guadalupe Valdés 1982; cf. Poplack and Levey 2010: 396-397). The hypothesis of susceptibility to English or other-dialect influence would necessarily be tested by measures of degree of contact. One metric for degree of contact with English could be the proportion of Spanish vs. English speakers in neighborhoods of residence (see Poplack and Levey 2010:399, 402). A parallel measure

could be applied to ascertain participants' level of interaction with speakers of other dialects, as has been done for contact between Salvadorans and Mexicans in Houston (Hernández 2009:598-600; see below). Direct measures of contact may also be developed from sociolinguistic profiles culled from content analysis of recorded conversations constituting a corpus; for example, concerning time and location of acquisition of English, preferred or "most comfortable" language, language choice according to interlocutor, and general affect toward the bilingual situation (Poplack, Walker and Malcolmson 2006:196-207).

Poplack's (1979) dissertation with Puerto Ricans living in Philadelphia remains a model study that has yet to be repeated in a U.S. Spanish community. Based on 24 sociolinguistic interviews collected from a neighborhood block in Philadelphia over a period of one year (Poplack 1979:28-37), it was one of the first to apply rigorous statistical analysis (logistic regression and principal component analysis) to data on linguistic variation and show the effects of social predictors. Of the 24 participants, 15 were female, and most were working class or unemployed and had limited formal education and social mobility (Poplack 1979:38-43). Poplack tested gender, age, education, language proficiency, and geographic origin (1979:48-50) as they conditioned lenition of coda /s/ (N=19,284), /n/ (N=8,648), and /r/ (N=7,142) (1979:64, 108,143).

Poplack grouped coda /n/ and /r/ by morphemic status, separating verbal /n/ and infinitival /r/ from monomorphemic /n/ and /r/. While there was very limited social conditioning of /n/ lenition (cf. Poplack 1979:123,127) overall, for monomorphemic /r/, she found an increased lenition rate for males (Poplack 1979:165), while for infinitival /r/ she found a slight effect of education (Poplack 1979:172). Social factors were of particular importance with coda /s/ lenition. For plural /s/, only geographic origin and language proficiency were selected by the model (Poplack 1979:86). For monomorphemic and verbal coda /s/, however, Poplack found that each of the five social factors tested (age, speech style, education, geographic origin, and language proficiency) conditioned lenition in word-final position (1979:75,96). It is important to note that the speakers studied here were from an immigrant population with a relatively short history (less than fifty years) in Philadelphia, which raises the question of the role of social factors given varying degrees of community stability and geographic permanence.

3. Changes in progress

Linguistic change in Spanish in the U.S. is often proclaimed, though not as often demonstrated. Making a reasonable case for language change requires, first of all, a robust quantitative pattern, which is verified in the speech of a community-based sample of speakers selected in a principled manner (Poplack et al. 2012). Changes in progress can be detected synchronically in **apparent time**—the distribution of variant forms across age cohorts (Labov 1994:43-72).

For example, in her pioneering sociolinguistic study of Panama City in 1969-71, Henrietta Cedergren (1973) observed a process of deaffrication from [tʃ] to the fricative

[ʃ] in apparent time, with an inverse relationship between lenition and age. The lenited variant increased as age decreased, peaking in the second youngest age group of 27-32 year olds and slightly declining in the 15-26 group. Cedergren obtained data using the same sampling procedure in 1983 to see whether real-time evidence would suggest a genuine change in progress or **age-grading**—that is, change with age that repeats in each generation and results in stable community behavior in aggregate. In fact, her comparison revealed age-grading—the same pattern was followed across apparent time at each point in real time—but with lenition incrementally higher for all but the two youngest groups; this is interpreted to mean that [tʃ] lenition in this community had peaked (Labov 1994: 94-97).

Deaffrication of [tʃ] showed a correlation with age in Tomé, in the Río Abajo region of New Mexico, just south of Albuquerque. Excluding postnasal and postlateral cases (*pla<u>nch</u>ar* 'to iron', *<u>el ch</u>ile* 'the chili pepper'), which are categorically realized as affricates, Jaramillo and Bills (1982) give an apparent time interpretation to the distribution of the variants across age groups in a sample of 36 speakers (N=1029). They find a shift from the fricative variant to the affricate, as the rate of [ʃ] is nearly halved in the youngest (17-30) age group compared to approximately 80% in the older groups. The interpretation of a shift toward the more standard pronunciation is supported by considering the effect of education, operationalized as years of formal instruction. Since age and educational attainment partially overlap (a greater proportion of younger than older people had college education), Jaramillo and Bills (1982:161) cross-tabulated age and education, to show an independent effect for education. In fact, within the young group, 8 speakers with a college education had a lenition rate approximately four times lower than that of the other 4 young speakers. Speakers with more than two years of formal study of Spanish also tended to lenite less often. The researchers conclude that the "perceived change" away from the "long-established" fricative variant "appears to simply reflect a sociological change related to education" whereby some residents are "expanding their command of different varieties of Spanish" (Jaramillo and Bills 1982:163-164).

Change in progress may be inferred from comparison of variation patterns in communities of origin. In Salvadoran communities, sequences of front vowels in hiatus with other vowels (as in *vea* 'he/she/you(formal) sees (Subj)') alternate with a hiatus-breaking [j̯] variant (*veya*) (Lipski 1994:258). Hernández (2015) compared rates and conditioning of the hiatus-breaking [j̯] variant for the immigrants in Houston to comparable data from San Sebastián, El Salvador, the municipality of origin for most families. Hernández reports that the rate of hiatus-breaking [j̯] in Houston (6%, N=737) is less than a third of that in San Sebastián (20%, N=811), receding to 2% (N=288) in the second immigrant generation. While in San Sebastián the hiatus-breaking variant was favored by older speakers and disfavored by women and those with a secondary school education, in Houston, with the now overwhelming preference for the hiatus variant, none of the social factors investigated—education level, gender and age of the speaker—make a statistically significant contribution.

Contrariwise, speaker gender does appear to make a contribution to linguistic variation in Kennett Square, Pennsylvania, though in a diminished way. Matus-Mendoza (2004) analyzed variable assibilation of word-final /r/ to a voiceless retroflex sibilant (*deci*[ř] 'to say, tell') in a corpus of 83 sociolinguistic interviews with speakers in Moroleón, Guanajuato and Kennett Square, where many mushroom industry workers are from Moroleón. The linguistic conditioning of assibilated /r/ was the same in Moroleón (N=2,796 (Matus Mendoza 2004:21)) and Kennett Square (unknown N). As for extralinguistic factors, rates in Moroleón differed across locales, with more frequent assibilation in urban than in rural areas, and across genders, with women assibilating more than men (Matus-Mendoza 2004:20-22). Differences according to occupation and education level also indicate that assibilation is a prestige variant in Mexico (Matus-Mendoza 2004:26-27). In Kennett Square, the rates of assibilation increase with more schooling and among women, but the percentages are "extremely low…compared to…Moroleón" (6% among women in Kennett Square vs. 24% in Moroleón), suggesting an "equalizing situation" in the shared working environment (Matus-Mendoza 2004:27).

Contraction of a phonetic variant has also occurred in Houston, where Salvadorans live and work alongside Mexicans. Composition of neighborhoods of residence provides one measure to approximate degree of dialect contact. In Houston's Segundo Barrio, Hispanics make up 90% of the population and the ratio of Mexicans to Salvadorans is on the order of 10 to 1, while in Holly Spring, where Hispanics constitute 12% of the population, it is closer to 2 to 1 (Hernández 2011:55). Hernández (2011) capitalizes on this difference in examining variation between word-final alveolar and velar nasals, as in *los pueblos fuero*[ŋ] *los que sufriero*[ŋ] *más* 'the towns were the ones that suffered the most', among Salvadorans in Houston, where Mexicans generally do not participate in this alternation. The rate of nasal velarization declines in Houston compared with San Sebastián, El Salvador, the community of origin (23%, N=430), but more so in Segundo Barrio (3%, N=476) than in Holly Spring (14%, N=981) (Hernández 2011:66). On this basis, Hernández is able to propose that differences between the two Houston communities are explained by amount of exposure to speakers of Mexican Spanish (cf. Trudgill 1986:39). One scenario of possible change in progress in a U.S. Spanish immigrant community, then, is dropping an alternation that constitutes a linguistic variable in the community of origin (cf. Weinreich 1968:18-19).

A contrary development may be the spread of a new linguistic variable. This appears to be the case with intervocalic ⟨ll⟩ ([j̯]) deletion (e.g., *iba a ir el bos por eØa a Brownsville* 'the bus was going to go to Brownsville for her' vs. *estudiar a Matamoros con ella* 'to study in Matamoros with her') in the Segundo Barrio and Holly Spring neighborhoods in Houston. Hernández (2015) compared speech data from sociolinguistic interviews conducted among Salvadoran and Mexican immigrants. The participants were first generation, second generation or (in the case of the Mexican speakers) third generation, and most were from families from San Sebastián, El Salvador, or Matamoros, Mexico. The rate of intervocalic ⟨ll⟩ deletion in Houston is twice as high in the second and third than in the first generation among Mexicans (N=383) and three times as high in the second generation than in the first among Salvadorans (N=622) (Hernández 2015). This means that second generation Mexican and Salvadorans show a closer elision rate

(31% and 23%, respectively) than do their first generation counterparts (17% and 5%, respectively). Though not significant in multivariate analysis, there appears to be a tendency for higher elision rates among men than women, in both national origin groups, for this expanding phonological variable.

As indicated by the studies surveyed in this chapter, linguistic and social categories are linked, yet social factors remain understudied — particularly socio-economic status. Common belief would claim that linguistic patterns in U.S. Spanish are unaffected by speakers' socioeconomic status (Bills and Vigil 2008:250). Furthermore, some researchers assert that speakers' occupation or education should *not* be expected to correlate with the language of minority communities, since Spanish is not instrumental for success in the employment market (Garcia and Tallon 2000:358, n.1).

Contributing to the lack of studies of social factors is the problem of grouping speakers according to sociological characteristics. This is at least no less exacting in minority-language situations than elsewhere, as the appropriateness of the criteria must be independently established for particular communities. For example, a solution for immigrant communities is offered by Orozco (2007:105), who classified New York City Colombians into three groups by taking into consideration their occupations both in NYC and in Colombia: those who retained white-collar jobs, blue-collar workers before and after immigration, and blue-collar workers in NYC who held white-collar positions in Colombia.

But a remaining problem in general is that social categories, unlike linguistic categories, have no standard or agreed-upon methods of demarcation. An additional obstacle is that social groupings often correlate with one another, and as such it is disadvantageous to include them *in omnibus* in a statistical model. We now illustrate an alternative approach which can circumvent this problem by grouping speakers on the basis of their linguistic behavior to infer social grouping.

4. Predicting social variation with linguistic behavior: Clustering and stratification in New Mexican Spanish

Hints of the social conditioning of variable usage can already be discerned in the earliest linguistic study of Spanish in the U.S. Over a century ago, Espinosa (1911:10) in *The Spanish language in New Mexico and Southern Colorado* suggests the social evaluation of the aspirated variant of /s/, qualifying it as "widespread among the rural uneducated classes". Dating back to 16th-17th century settlement from New Spain (today, Mexico), Northern New Mexico is home to Traditional New Mexican Spanish. As Lipski (2000:2-4) has noted, New Mexican Spanish was deemed by Espinosa and contemporary linguists in Latin America and Spain to be nothing less than another national variety of the language.

In New Mexico, it is the speakers of English, not Spanish, who are (descendants of) immigrants. In 1850, the area became a U.S. territory, and in 1878, the railroad

arrived along with accelerating Anglo-American immigration. In 1912, New Mexico was admitted to the Union as the 47th state and English increasingly displaced Spanish in schools—even in northern, longstanding Spanish-speaking communities—by the 1940s. Today Spanish is taught as a foreign language and, while Hispanics represent as much as 80% of the population in some northern counties, there is a continued shift toward English (Bills and Vigil 2008 *inter alia*). The remaining speakers of Traditional New Mexican Spanish provide an invaluable window into Spanish language use in a native community.

The New Mexico Spanish-English Bilingual (NMSEB) corpus consists of spontaneous speech collected by in-group community members and exhaustively transcribed in prosodic units (Torres Cacoullos and Travis, In Preparation). Participants were selected to cover a range of demographic backgrounds to permit the assessment of extra-linguistic constraints on linguistic variation (Travis and Torres Cacoullos 2013, Torres Cacoullos and Travis 2015).

To identify which social factors may contribute a consistent effect on linguistic variation data, we cast a wide net by looking at the problem in reverse (Horvath and Sankoff 1987, Poplack 1979:190-223). We will use the linguistic behavior of speakers in NMSEB to cluster them via a principal component analysis (PCA), and then interpret the resultant configurations in terms of our extralinguistic knowledge of the speakers to identify the social characteristics that individuals within those clusters have in common.

Principal component analysis (PCA) is a data optimization method, used to partition a multidimensional space into a number of orthogonal components which reduce the dimensionality of that space; the dimensions that contribute toward partitioning the variance of that space are called the principal components. A PCA works best when there is a high amount of variance in the space; typically this is found when each row vector (in this case, speaker) has more than 10 numeric variables, or dimensions (in this case, linguistic features) (cf. Horvath and Sankoff 1987:186).

We focus on four phonetic variables: onset (syllable-initial) /s/ lenition, coda (syllable-final) /s/ lenition, intervocalic /d/ elision, and intervocalic ⟨ll⟩ lenition. Each of these has been studied and implicated as either characteristic of New Mexican Spanish or as a stable, socially-stratified variable in other dialects of Spanish (e.g., Espinosa 1909:72,75; Gutiérrez 1981; Lapesa 1968:354, 356; Lipski 2008:204, 2011:75-83; Samper Padilla 2011:105-114).

The dimensions for the PCA were based on the linguistic constraints for each of the four phonetic variables. Since onset /s/ lenition was most strongly favored by preceding non-high vowels—as in *ése* 'that one' or *la señora* 'the woman' (cf. Brown 2005a)—counts of onset /s/ in favorable phonetic contexts (preceding non-high vowels) were separated from counts of onset /s/ in other phonetic environments. Additionally, since complete elision of onset /s/ was rare, tokens were divided into full ([s]) and lenited variants and counts were included separately. Doing so produced counts of four variants:

full onset /s/ preceded by a non-high vowel, lenited onset /s/ preceded by a non-high vowel, full onset /s/ in other contexts, and lenited onset /s/ in other contexts.

For coda /s/, lenition was most strongly favored when the following phone was a voiced consonant, as in *desde* 'since/from' or *los viejitos* 'the old people' (cf. Brown 2005b). Unlike onset /s/, coda /s/ showed a mix of full ([s]), aspirated ([h]), and elided (ø) variants, so we considered counts of each separately. This produced six additional variants per speaker (each of the three variants followed by a voiced consonant and in other environments). For intervocalic /d/ elision, we take counts of intervocalic approximants against the number of elided intervocalic tokens (in which there was no perceptible frication as well as audible vowel coarticulation, e.g., *casado* [kasau]; *casada* [kasa:] 'married'). Finally, for ⟨ll⟩ lenition in words like *ellos* 'they/them', reduced and completely elided forms were grouped together.

In all, there were 14 variants across the four phonetic variables. Since onset and coda /s/ were further subdivided by phonetic environment, this produced six categories: onset /s/ in a favorable environment for lenition, onset /s/ in other environments, coda /s/ in a favorable environment for lenition, coda /s/ in other environments, intervocalic /d/, and intervocalic ⟨ll⟩. If a participant had fewer than 20 tokens in total for any of these categories (summing up all variants within those categories), their counts for all variants in that category were zeroed out to keep low token counts from warping the PCA output.

The principal components resulting from the PCA[1] were then plotted based on the amount of variance each principal component accounted for. Three principal components accounted for 78% of the total variance in the dataset. We then examined the associations of each of the 14 variants with each of these three principal components. Many of the variants showed moderate associations, or loadings (with magnitude greater than 0.3; $|PC_x|>0.3$) (cf. Horvath and Sankoff 1987:194), indicated by bolded text and cell shading in Table 1; variants with weaker associations ($0.25\leq |PC_x| <0.3$) are listed in bold without shading.

[1] The PCA was conducted in R (R Core Team 2015) using the prcomp() function. The counts for each column were scaled to account for different overall token counts for the distinct variables.

Table 1: Loadings of 14 Consonantal Variants in New Mexico (NMSEB) on Principal Components

Variable	Variant (Dimension)	PC1	PC2	PC3
Onset /s/	[s].Preceding NonHighV	0.24	**0.40**	-0.11
	[h]/ø.Preceding NonHighV	**-0.33**	-0.11	-0.01
	[s].Other environments	0.13	**0.38**	**-0.45**
	[h]/ø.Other environments	**-0.31**	-0.03	0.21
Coda /s/	[s].Following Voiced C	0.20	**0.34**	-0.15
	[h].Following Voiced C	**-0.30**	0.13	**-0.33**
	ø. Following Voiced C	**-0.28**	0.09	**-0.40**
	[s].Other environments	**0.28**	**0.30**	-0.04
	[h].Other environments	**-0.34**	0.05	**-0.26**
	ø .Other environments	**-0.28**	-0.09	**-0.36**
Intervocalic /d/	Approximant intervocalic /d/	-0.15	**0.43**	**0.26**
	Elided intervocalic /d/ (ø)	**-0.25**	**0.31**	0.21
Intervocalic ⟨ll⟩	Full intervocalic /j/	-0.23	**0.34**	0.22
	Lenited intervocalic /j/	**-0.30**	0.21	**0.32**
	Variance accounted for:	**46%**	**21%**	**11%**
	Interpretation:	Lenition (general)	Retention (general), except /d/	NM Spanish vs. other dialects

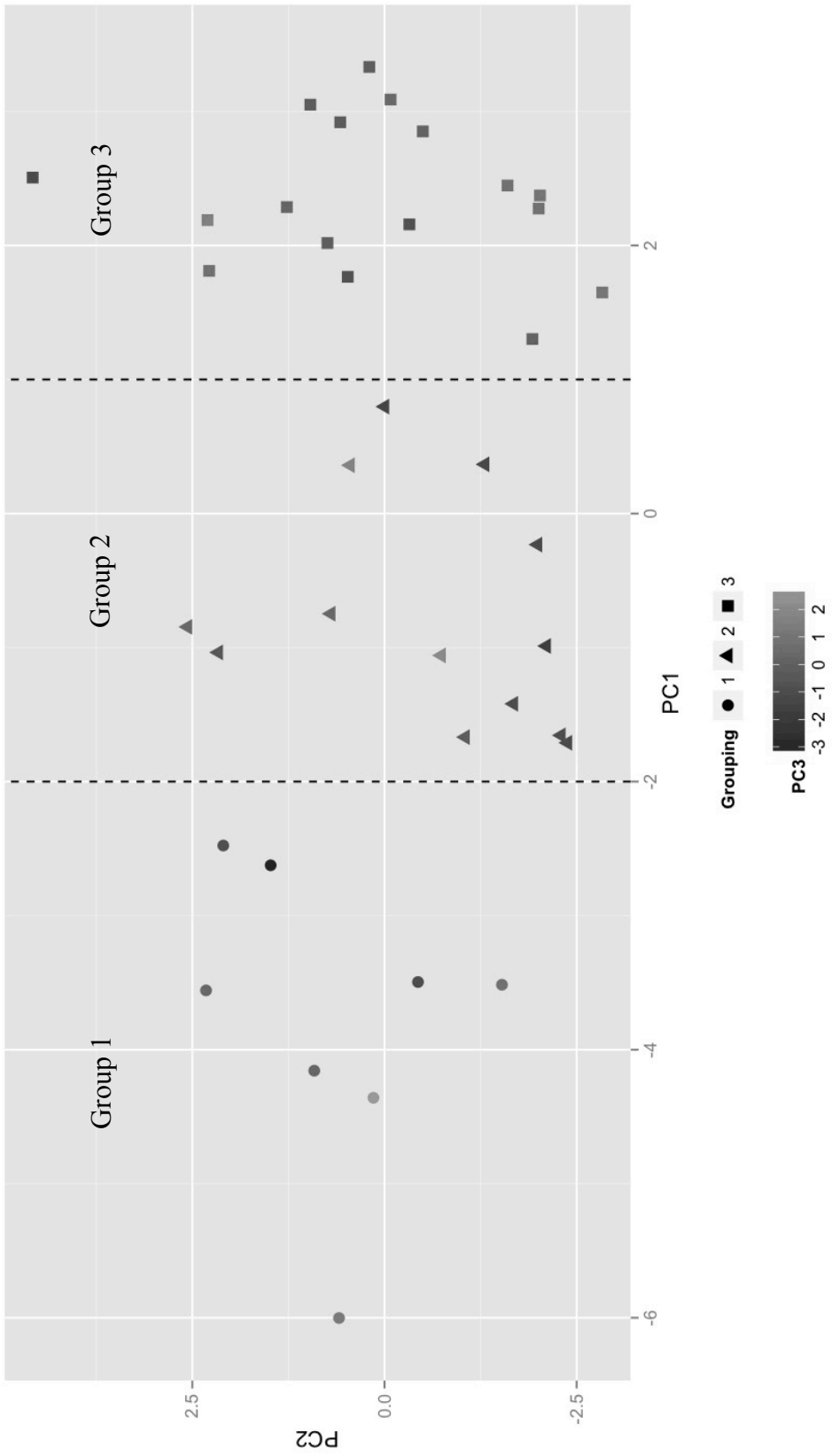
We interpret the loadings as follows. Principal Component 1 (PC1), which accounts for 46% of the variance, appears to represent lenition in general. That is, speakers who have (negative) associations with this component are more likely than other speakers to aspirate onset /s/ and will also tend to lenite (aspirate) coda /s/ and to use lenited intervocalic /j/ as well. Principal Component 2, accounting for another 21% of the variance, is largely the complement of PC1. Here retention of full variants is (positively) associated with PC2. However, we note that both approximant and elided intervocalic /d/ correlate positively with this component with similar magnitudes, indicating that the component makes no distinction between the two. In fact, the variants of intervocalic /d/ pattern similarly in each principal component, indicating that /d/ elision is not a phonetic variable which contributes much meaningful variance for grouping our speakers. Principal Component 3, which accounts for 11% of the total variance, is more complicated. Both onset /s/ retention in other than preceding non-high-vowel environments (i.e., in disfavorable contexts for /s/ aspiration) and coda /s/ lenition, especially ø, pattern in the same direction (negatively), and these are contrasted with intervocalic ⟨ll⟩ lenition and, though its association is marginal, onset /s/ aspiration in preceding non-high-vowel contexts. This component, then, groups more standard and general Spanish linguistic patterns, namely onset /s/ retention and coda /s/ lenition, in

opposition to traditional New Mexican variants, that is, intervocalic ⟨ll⟩ lenition and onset /s/ aspiration.

Effectively, the Principal Component Analysis has taken a 14-dimensional space representing each variant of our four variables and reduced it to a three-dimensional space where highly correlated items pattern together. This permits geographical representations of the data which can elucidate similarities in speaker behavior, but it also, crucially, illustrates the associations of the phonetic variants to one another. Through such an analysis, we are able to apprehend that participant groupings are strongly determined by patterns of lenition.

With a general linguistic interpretation of the principal components in mind, we then ask how individual speakers associate with each of the principal components. In Figure 1, by plotting each participant according to their loading on the first two Principal Components, and letting the shading indicate the third Principal Component, we are able to capture the results of the PCA visually and use this to cluster participants. In doing so, we observe that the participants naturally fall into three main groups, primarily delineated by PC1 (indicated by shape in Figure 1).

Figure 1: Grouping of NMSEB Speakers by Linguistic Behavior (from PCA)



Using these speaker clusters based on linguistic behavior, we compared sociodemographic characteristics of the speakers to assess what was shared among a majority of members. Group 1 mostly consists of miners, factory workers, or ranchers who are men with a middle or high school education. Group 2 is mainly constituted by middle or high school educated men and women, some in production (e.g., factory workers), and some in service (e.g., in dry cleaning) occupations. Group 3 is a more urban, predominately female group, in which we find most college-educated participants and professionals (e.g., teachers).

Based on these clusters, then, it appears that socioeconomic status (occupation and education), gender, and rural vs. urban locale should be considered as candidates for conditioning linguistic variation in NM Spanish. For a composite socioeconomic index based on occupation and education, we grouped speakers into 'production workers' (N=14), 'service employees' (N=15) and 'professionals' (N=9). There were 22 women and 16 men. As to locale, 'urban' were those participants from cities with 10,000 or more residents (Albuquerque, Española, Las Vegas, Los Lunas, Santa Fe) (N=11).

To provide additional evidence that these social factors may be predictors of variation, we determined whether they were distributed unevenly among the three speaker clusters, via Fisher's exact tests. According to these, gender ($p<0.05$) was disproportionately distributed among groups, with males being more common in Group 1 and females in Group 3. Subsequent Fisher's exact tests conducted pairwise indicated that occupation-education was differentially distributed across groups 1 and 3 ($p<0.05$), suggesting that social class as well as gender may be useful social categories for conditioning linguistic behavior. Though not reaching statistical significance, there seems to be a slight bias toward rural speakers in Group 1 (7/8 are rural). Thus, we include Rural vs. Urban locale as a social predictor, understanding that this characteristic may not be as robust as gender or socioeconomic status in distinguishing participants' linguistic behavior.

These three social predictors were considered together with linguistic factors in generalized linear mixed models (conducted using the lme4 package (Bates et al. 2014) in R (R Core Team 2015)).[2] A separate model was fit for each of the four phonetic variables, and this model was compared to models with only linguistic and only social predictors (via likelihood ratio tests). While both onset and coda /s/ lenition were primarily determined by linguistic factors (as also reported by Brown 2005a, 2005b), model fits for intervocalic /d/ elision and ⟨ll⟩ lenition improved with the inclusion of a combination of linguistic *and* social predictors.

Tables 2 and 3 show the results of generalized linear mixed models for intervocalic /d/ elision and ⟨ll⟩ lenition, respectively. The Intercept refers to the log-likelihood of a dependent variable at a given reference level (reference levels are listed below each table). Levels of each predictor are assigned a weighting (β-coefficient), or Estimate, with positive values indicating an increased likelihood and negative values

[2] With weighted effect coding, to account for imbalances in discourse data.

indicating a decreased likelihood for a given level (factor). For example, in the case of intervocalic /d/ elision, the positive Estimate for a preceding non-high vowel suggests that this phonetic context increases the likelihood of elision.[3] Also indicated in the model outputs is significance (determined by estimated p-values computed via a Wald test).

Intervocalic /d/ deletion is strongly affected by social class. In agreement with reports on Latin American varieties of Spanish (e.g., Panamanian (Cedergren 1973) and Venezuelan (D'Introno and Sosa 1986)), we find that intervocalic /d/ elision is favored in working class speech. Also replicating reported patterns, we see that men elide intervocalic /d/ more often than women. The primary factors conditioning intervocalic /d/ elision, however, are still linguistic. Phonetic context and participle status work together to vastly increase lenition rates with participles from the first conjugation (*-ado*) relative to non-participles when the preceding phone is a non-high vowel. Intervocalic ⟨ll⟩ lenition is also conditioned by a combination of social and linguistic factors. We find that while the strongest predictors of ⟨ll⟩ lenition—phonetic context—are linguistic, there is also an effect of occupation: speakers from production occupations lenite ⟨ll⟩ most often, followed by speakers from service and professional occupations.

[3] Because Occupation has three levels, we compared production workers to other workers as one contrast, and service to professional occupations for the second contrast. Thus, weightings are reflective of how the second group behaves with respect to the first. In /d/ elision, for example, the negative estimate for Production vs. Other indicates that speakers from non-production occupations elide less than speakers in production occupations.

Table 2: Social and Linguistic Factors Conditioning
Intervocalic /d/ Elision in NMSEB (N=3447)*

Factor	Estimate	Std. Error	Sig.	N	% elision
(Intercept)	-3.74	0.54	***		
Preceding Non-high V	3.08	0.57	***	2788	16%
(Preceding High V)				659	3%
Participle	1.65	0.59	**	471	34%
(Not a participle)				2976	10%
Production vs. Other Occupations	-1.28	0.42	**		
Service vs. Professional	-0.12	0.42			
(Production Occupation)				1753	18%
(Service Occupation)				1132	10%
(Professional Occupation)				444	9%
Rural Locale	-0.09	0.29		2376	14%
(Urban Locale)				1065	12%
Male Gender	0.60	0.29	*	1633	17%
(Female)				1808	11%
Preceding Non-high V:Participle	1.52	0.78	*	320	48%
(Prec. Non-high V, Non-participle)				753	14%
(Prec. High V, Participle)				150	3%
(Prec. High V, Non-participle)				288	1%

*Generalized linear mixed model, lme4 package (Bates et al. 2014) in R (R Core Team 2015)
Random Effects (SD): Speaker (0.55); Word (2.01)
Reference Level: /d/ Present, Prec High V, Non-participle, Urban, Female
| *** $p<0.001$ | ** $p<0.01$ | * $p\leq0.05$ |

Table 3: Social and Linguistic Factors Conditioning Intervocalic ⟨ll⟩ Lenition in NMSEB (N= 1335)*

Factor	Estimate	Std. Error	Sig.	N	%Lenition
(Intercept)	-0.15	0.53			
Preceding Front V	-0.04	0.86		600	74%
(Preceding Non-front V)				735	49%
Following Front V	0.26	0.56		26	46%
(Following Non-front V)				1309	61%
Asymmetry in Height	0.89	0.50		1222	59%
(Symmetry in Height)				113	74%
Production vs. Other Occupations	-0.8	0.35	*		
Service vs. Professional	0.13	0.43			
(Production Occupation)				784	67%
(Service Occupation)				365	54%
(Professional Occupation)				186	43%
Rural Locale	0.55	0.36		1000	63%
(Urban Locale)				335	51%
Male Gender	0.02	0.36		753	63%
(Female)				582	56%
Preceding Front V: Following Front V	-3.05	1.04	**	15	40%
(Prec. Front V; Foll. Non-front V)				585	74%
(Prec. Non-front V; Foll. Front V)				11	55%
(Prec Non-front V; Foll. Non-front V)				724	49%

*Generalized linear mixed model, lme4 package (Bates et al. 2014) in R (R Core Team 2015)
Random Effects (SD): Word (1.2), Speaker (0.76)
Reference Level: Full token, Preceding Non-Front V, Following Non-Front V, Height Symmetry, Urban Locale, Female Gender
| *** $p<0.001$ | ** $p<0.01$ | * $p\leq0.05$ |

5. Conclusion

Although social factors in U.S. Spanish have received inadequate attention to date, the few available reports confirm the need to account for social variation. As we have seen, where social factors *have* been tested, sociolinguistic patterns generally replicate those found across the Spanish-speaking world, revealing the systematic character of varieties of Spanish in the US.

A contributing factor to the paucity of studies has been the familiar problem that social characteristics of speakers are generally less well-defined than linguistic categories, particularly in minority language situations, and that social factors are often highly interdependent. A solution to this conundrum can be found by using a data

optimization method such as principal component analysis as a heuristic for grouping speakers strictly on the basis of their linguistic behavior, with the groups thus defined then interpreted according to social characteristics. We have illustrated one such analysis in a corpus of New Mexican Spanish. By applying PCA to counts of known phonetic variables, we determined that occupation-education, gender, and demographic locale were likely social factors of variation. This was confirmed via regression analysis for two phonetic variables. For intervocalic /d/ and ⟨ll⟩, the highest lenition rates are found in speakers with production occupations and, for /d/, among men. In each case, we observe social stratification common not only to many other dialects of Spanish, but to many language varieties in general.

The study of Spanish in the US can advance with data from community-based speech corpora that are constituted by participants of known sociodemographic characteristics sampled in an informed, principled way. These are, effectively, the principal components of accountable sociolinguistic research.

6. References

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-5. http://CRAN.R-project.org/package=lme4

Bentivoglio, Paola, & Sedano, Mercedes. 2011. Morphosyntactic variation in Spanish speaking Latin America. In Díaz-Campos, Manuel (ed.), *The Handbook of Hispanic Sociolinguistics*. Oxford: Blackwell. 168-186.

Bills, Garland D. 1975. Linguistic research on United States Hispanics: State of the art. In Teschner, Richard V., Garland D. Bills & Jerry R. Craddock (eds), *Spanish and English of United States Hispanos: A critical, annotated, linguistic bibliography*. Arlington, VA: Center for Applied Linguistics.

Bills, Garland D. and Vigil, Neddy A. 2008. *The Spanish language of New Mexico and Southern Colorado: A linguistic atlas*. Albuquerque: University of New Mexico Press.

Brown, Esther L. 2005a. Syllable-initial /s/ in Traditional New Mexican Spanish: Linguistic factors favoring reduction *ahina*. *Southwest Journal of Linguistics, 24*(1-2), 13-30.

Brown, Esther L. 2005b. New Mexican Spanish: Insight into the variable reduction of "La ehe inihial" (/s-/). *Hispania 88*(4), 813-824.

Cedergren, Henrietta. 1973. The interplay of social and linguistic factors in Panama. Cornell University dissertation.

D'Introno, Francesco, & Sosa, Juan Manuel. 1986. La elisión de la /d/ en el español de Caracas: Aspectos sociolingüísticos e implicaciones teóricas. In Núñez Cedeño, Rafael, Páez, Iraset & Guitart, Jorge (eds.), *Estudios sobre la fonología del español del Caribe*. 135-163. Caracas: La Casa de Bello.

Díaz-Campos, Manuel. 2011. *The handbook of Hispanic sociolinguistics*. Wiley-Blackwell.

Espinosa, Aurelio M. 1911. The Spanish language in New Mexico and Southern Colorado. Santa Fe, NM: Historical Society of New Mexico (No. 16).

Garcia, MaryEllen & Tallon, Michael. 2000. *Estar* in Mexican-American Spanish: Phonological or morphological variability? In Roca, Ana (ed.), *Research on Spanish in the United States: Linguistic issues and challenges*. Somerville, MA: Cascadilla Press. 348-359.

Gutiérrez, John. 1981. An analysis of the phoneme /s/ in New Mexico Spanish. In Elerick, Charles (ed.), *Proceedings of the Ninth Annual Southwestern Areal Language and Linguistics Workshop*, El Paso: University of Texas at El Paso. 234-239.

Hernández, José Esteban 2009. Measuring rates of word-final nasal velarization: The effect of dialect contact on in-group and out-group exchanges, *Journal of Sociolinguistics, 13*(5), 583-612.

Hernández, José Esteban. 2011. Measuring rates and constraints of word-final nasal velarization in dialect contact. In Ortiz-López, Luis A. (ed.), *Selected Proceedings of the 13th Hispanic Linguistics Symposium*.Somerville, MA: Cascadilla Proceedings Project. 54-69.

Hernández, José Esteban. 2015. *Yod* travels North: Measuring rates and constraints of two phonological processes in dialect contact. Invited talk, Spanish, Italian and Portuguese, Penn State University, January 26.

Horvath, Barbara, & David Sankoff. 1987. Delimiting the Sydney speech community. *Language in Society, 16*(2), 179-204.

Jaramillo, June A. & Bills, Garland D. 1982. The phoneme /ch/ in the Spanish of Tomé, New Mexico. In F. Barkin, E. A. Brandt., & J. Ornstein-Galicia (eds.), *Bilingualism and Language Contact: Spanish, English, & Native American Languages*. New York: Teachers College, Columbia University. 154-165.

Labov, William. 1972. Some principles of linguistic methodology. *Language in Society 1*(1), 97-121.

Labov, William. 1994. *Principles of linguistic change, Vol. 1: Internal factors*. Oxford: Blackwell.

Labov, William. 2001. *Principles of Linguistic Change, Vol. 2: Social Factors*. Malden, MA: Blackwell.

Labov, William. 2006. *The social stratification of English in New York City*. 2nd ed. Cambridge: Cambridge University Press.

Lapesa, Rafael. 1968 (1942). *Historia de la lengua española*. 7a edición. Madrid: Escelicer, S.A.

Lantolf, James P. 1982. Speaker sex and *para* reduction in Chicano Spanish. In F. Barkin, E. A. Brand, & J. Ornstein-Galicia (eds.), *Bilingualism and Language Contact: Spanish, English, & Native American Languages.* New York: Teachers College, Columbia University. 166-176.

Lipski, John M. 1994. *Latin American Spanish*. London: Longman.

Lipski, John M. 2000. Back to zero or ahead to 2001? Issues and challenges in U.S. Spanish research. In: Roca, Ana (ed.), *Research on Spanish in the United States: Linguistic issues and challenges.* Somerville, MA: Cascadilla Press. 1-41.

Lipski, John M. 2008. *Varieties of Spanish in the United States*. Washington, D.C.: Georgetown University Press.

Lipski, John M. 2011. Socio-phonological variation in Latin American Spanish. In Díaz-Campos, Manuel (ed.), *The handbook of Hispanic sociolinguistics*. Wiley-Blackwell. 72-97.

Martín Butragueño, Pedro & Lastra, Yolanda. 2015. Subject pronoun expression in oral Mexical Spanish. In Carvalho, Ana M, Orozco, Rafael & Lapidus Shin, Naomi (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective.* Washington, D.C.: Georgetown University Press. 39-57.

Matus-Mendoza, Maríadelaluz. 2004. Assibilation of /-r/ and migration among Mexicans. *Language Variation and Change* 16: 17-30.

Ocampo, Francisco. 1990. El subjuntivo en tres generaciones de hablantes bilingües. In Bergen, J. (ed.), *Spanish in the United States: Sociolinguistic Issues.* Washington, D.C.: Georgetown University Press. 39-45.

Orozco, Rafael. 2007. Social constraints on the expression of futurity in Spanish-Speaking urban communities. In Holmquist, Jonathan et. al (eds*.), Selected proceedings of the Third Workshop on Spanish Sociolinguistics*. Somerville, MA: Cascadilla Proceedings Project. 103-112.

Orozco, Rafael. 2015. Pronominal variation in Colombian Costeño Spanish. In Carvalho, Ana M, Orozco, Rafael & Lapidus Shin, Naomi (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*. Washington, D.C.: Georgetown University Press. 17-37.

Otheguy, Ricardo & Zentella, Ana Celia. 2012. *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. New York: Oxford University Press.

Peñalosa, Fernando. 1981. Some issues in Chicano sociolinguistics. In Duran, Richard P. (ed.), *Latino language and communicative behavior*. Norwood, N.J.: ABLEX, 3-18.

Poplack, Shana. 1979. Function and process in a variable phonology. University of Pennsylvania dissertation.

Poplack, Shana. 1993. Variation theory and language contact: Concepts, methods and data. In Preston, Dennis R. (ed.), *American dialect research.* Amsterdam: John Benjamins. 251-286.

Poplack, Shana. 1997. The sociolinguistic dynamics of apparent convergence. In Guy, Gregory R., Feagin, Crawford, Schiffrin, Deborah & Baugh, John (eds.) *Towards a social science of language. Papers in honor of William Labov: Social interaction and discourse structures*, v2. Amsterdam: John Benjamins. 285-309.

Poplack, Shana & Levey, Stephen. 2010. Contact-induced grammatical change. In Auer, Peter, Schmidt, Jürgen Erich (eds.), *Language and Space—An International Handbook of Linguistic Variation: Volume 1—Theories and Methods.* Berlin: Mouton de Gruyter. 391-419.

Poplack, Shana, Zentz, Lauren & Dion, Nathalie. 2012. What counts as (contact-induced) change. *Bilingualism: Language and Cognition, 15*(2), 247-254.

Poplack, Shana, Lealess Allison & Dion, Nathalie. 2013. The evolving grammar of the French subjunctive. *Probus, 25,* 139-195.

Samper-Padilla, José Antonio. 2011. Sociophonological variation and change in Spain. In: Díaz-Campos, Manuel (ed.), *The handbook of Hispanic sociolinguistics*. Wiley-Blackwell. 98-120.

Shin, Naomi Lapdius & Otheguy, Ricardo. 2013. Social class and gender impacting change in bilingual settings: Spanish subject pronoun use in New York. *Language in Society, 42*, 429-452.

Silva-Corvalán, C. (1994). *Language contact and change: Spanish in Los Angeles*. Oxford: Clarendon.

Torres Cacoullos, Rena & Travis, Catherine E. (In preparation). New Mexico Spanish-English Bilingual (NMSEB) corpus. National Science Foundation 1019112/1019122. http://nmcode-switching.la.psu.edu/.

Torres Cacoullos, Rena and Catherine E. Travis. 2015. Gauging convergence on the ground: Code-switching in the community. *International Journal of Bilingualism, 19(4)*, 365-386.

Travis, Catherine E. & Torres Cacoullos, Rena. 2013. Making voices count: Corpus compilation in bilingual communities. *Australian Journal of Linguistics, 33(2)*, 170-194.

Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.

Zentella, Ana Celia. 1990. El impacto de la realidad socio-económica en las comunidades hispanoparlantes de los Estados Unidos: Reto a la teoría y metodología lingüística. In Bergen, J. (ed.), *Spanish in the United States: Sociolinguistic Issues.* Washington, D.C.: Georgetown University Press. 152-166.

Valdés, Guadalupe.1982. Bilingualism in a Mexican border city: A research agenda. In Barkin, F., Brandt, E.A., & Ornstein-Galicia, J. (eds.), *Bilingualism and Language Contact: Spanish, English, & Native American Languages*. New York: Teachers College, Columbia University. 3-17.

Valdés, Juan de. 1535 (2010). *Diálogo de la lengua*. Biblioteca Virtual Miguel de Servantes, online edition based on José F. Montesinos (Madrid, Espasa-Calpe, 1976).

Weinreich, Uriel. 1968. *Languages in Contact: Findings and Problems*. The Hague: Mouton.

Weinreich, U., Labov, W., & Herzog, M. 1968. Empirical foundations for a theory of language change. In Lehmann, W. P., & Malkiel, Y. (eds.), *Directions for historical linguistics*. Austin, TX: University of Texas Press. 95–188.