AUSTRALIAN JOURNAL OF LINGUISTICS

Australian Journal of Linguistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cajl20

Making Voices Count: Corpus Compilation in Bilingual Communities

Catherine E. Travis [a] & Rena Torres Cacoullos [a]

[a] Australian National University, The Pennsylvania State University

To cite this article: Catherine E. Travis & Rena Torres Cacoullos (2013) Making Voices Count: Corpus Compilation in Bilingual Communities, Australian Journal of Linguistics, 33:2, 170-194

To link to this article: http://dx.doi.org/10.1080/07268602.2013.814529

PLEASE SCROLL DOWN FOR ARTICLE

Australian Journal of Linguistics, 2013
Vol. 33, No. 2, 170–194, http://dx.doi.org/10.1080/07268602.2013.814529

Making Voices Count: Corpus Compilation in Bilingual Communities*

CATHERINE E. TRAVIS AND RENA TORRES CACOULLOS

Australian National University, The Pennsylvania State University

Corpus compilation is of great relevance in linguistics today, with growing appreciation of studies based on spontaneous speech, in particular for minority communities. This paper puts forward a model for corpus compilation in bilingual communities, illustrated through the New Mexican Spanish–English Bilingual corpus, in which the same speakers—from a long-standing minority community in the United States—use both Spanish and English in the same conversations, smoothly alternating between their languages. We advocate community-based fieldwork for the collection of speech data by community members; the formation of a corpus which comprises recordings of spontaneous interactions and is thus of widespread usability (rather than being tied to any particular set of research questions or elicited linguistic features); scrupulously compiling information about the demography and the linguistic history of the participants that may shape their patterns of language use; and comprehensive transcription of the data taking into account prosodic aspects and making considered decisions about how to responsibly represent the speech. This community-based approach yields linguistic data situated in its social context and amenable to systematic

*Though the order of the authors is not alphabetical, both contributed equally to this work, which received funding from the National Science Foundation (BCS 1019112/1019122 [2010–2013] 'Collaborative Research: Evaluating convergence via code-switching: Cross-linguistic priming, rates and the structure of subject expression'; cf. http://nmcode-switching.la.psu.edu/). Thanks also to the Research School of Humanities of the Arts and the School of Language Studies of the Australian National University for support to Torres Cacoullos as a Visiting Fellow in November 2012, which made this article possible, and to Alexandra Aikhenvald, Jane Simpson and Nick Thieberger for valuable feedback on the paper. The research would have been impossible without help from a team of workers, and we gratefully acknowledge the contribution of Jenny Dumont (as UNM project manager, 2010–2011) and Colleen Balukas (at Penn State), as well as the research assistants from UNM (Daniel Abeyta, Rubel Aguilar, Raúl Aragón, Jason Gonzales, Aubrey Healey, Leah Houle, Rebeca Martínez, Ana Medina Murillo, Amanda Ortiz, Andrés Sábogal, Lillian Sanchez, Lizeth Trevizo, Kamie Ulibarrí, Víctor Valdivia Ruiz), and from Penn State (Nicole Benevento, Yolanda Gordillo, Everardo Tapia). We also thank Garland Bills and Neddy Vigil, for their own inspiring work on New Mexican Spanish, and other colleagues for their collaboration (Sonia Balasch, Evan Kidd, Chris Koops, Enrique Lamadrid and Damian Vergara Wilson).

quantitative analysis, which allows for confronting the many claims about language contact with the facts of bilingual usage.

Keywords: Code-switching; Bilingualism; Minority Communities; Corpus Compilation; Transcription Methods; Spanish

1. Introduction

Corpus compilation is of great relevance in linguistics today, with growing recognition that the insights that studies of spontaneous speech have to offer are unavailable from elicited data. The imperative for such studies is even more apparent in research on minority communities, including those in multilingual settings, where social attitudes may inhibit the occurrence of stigmatized features (e.g. perceived non-standard forms or code-switching) in formal environments such as those imposed by elicitation and experimental tasks.

Here, we offer an approach to corpus compilation, illustrated through an overview of our work with a long-standing minority bilingual community in the United States (US) to constitute the New Mexican Spanish–English Bilingual corpus (NMSEB: Torres Cacoullos & Travis in preparation). We advocate community-based fieldwork for the collection of speech data by community members; the formation of a corpus which comprises recordings of spontaneous interactions and is thus of widespread usability (rather than being tied to any particular set of research questions or elicited linguistic features); scrupulously compiling information about the demography and the linguistic history of the participants that may shape their patterns of language use; and comprehensive transcription of the data taking into account prosodic aspects and making considered decisions about how best to represent the speech. Building on the foundation of sociolinguistic corpus compilation (see, especially, Poplack 1989), the resulting NMSEB is a corpus of bilingual interactions that is—as far as we know—unique in recording spontaneous speech in two languages from the same speakers in the same setting, which can serve as a model for bilingual corpora.

2. The Speech Community as the Unit of Linguistic Study

There is increasing understanding that grammar (or linguistic 'competence') is the cognitive organization of our linguistic experience—the frequency and context of occurrence of linguistic forms (e.g. Bybee 2010). From this it follows that studies that depend on anecdotal observations or haphazardly collected examples do not allow for a reliable accounting of actual usage or 'the regular patterns that characterize natural exchanges in the speech community' (Poplack 1993: 263). Rather, usage is best observed in large corpora of spontaneous speech amenable to systematic quantitative analysis drawn from a well-defined speech community.[1]

It has been observed that linguistic patterns differ from community to community, and this is no less so for bilingual phenomena. Poplack's (1998) comparison of two bilingual communities (Spanish/English bilingual Puerto Ricans in New York and French/English bilinguals in Ottawa-Hull, Canada) revealed marked differences between the two communities in the ways in which they combine the two languages, despite typological similarities. While the bilingual Puerto Ricans were found to alternate smoothly and frequently between Spanish and English multi-word sequences, using such smooth code-switching as a discourse mode (1998: 46), the Ottawa-Hull bilinguals did not, and instead switched to English to fulfil certain rhetorical functions, such as deploying the appropriate lexical item or making metalinguistic commentary (1998: 51–52). Other multilingual communities may have restrictions on when and with what languages switching is appropriate. For example, Aikhenvald remarks that, in the Vaupés basin in northwest Amazonia in Brazil, switching between Tariana and Tucano is only considered acceptable when it is used to meet specific functional needs, such as to mark direct speech, to quote animals and evil spirits in narratives, for metalinguistic commentary and to exclude a participant who does not know the language (2002: 190–191).

The first step in corpus compilation, then, is to delimit a speech community, which has 'well-defined limits, a common structural base, and a unified set of sociolinguistic norms' (Labov 2007: 347).

2.1. New Mexican Spanish: An Endangered Variety

Northern New Mexico is home to the oldest variety of a European language spoken in the US (Bills & Vigil 1999: 43), and 'arguably the oldest continually spoken variety of Spanish anywhere in the Americas that has not been updated by more recent immigration' (Lipski 2008: 193). Spanish has been spoken in the region for over 400 years and, in the state overall, is still spoken today at home by over a quarter of the population (28.7%, significantly higher than the national average of 12.6%).[2] Further, New Mexico is the state with the highest proportion of Hispanics in the US (close to one million people, or 46.7% of the state, compared with the national average of 16.7%).[3] This large Spanish-speaking population stems from two distinct sources: immigrants from Mexico, as in the rest of the US, and descendants of the original Spanish colonizers. The two groups are separated geographically to some degree, with the former residing primarily in the south of the state toward the border

[1] Different issues may arise in the study of languages in situations of obsolescence, where few speakers remain (see, e.g. Dorian 2010).

[2] American Community Survey 5-Year Estimates—2007–2011, available at http://factfinder2.census.gov, last accessed 12 December 2012.

[3] US Census bureau 2011 population estimates, available at http://quickfacts.census.gov/qfd/states/35000.html, last accessed 12 December 2012.

with Mexico and in the urban centre of Albuquerque, and the latter residing primarily in the north of the state (Bills & Vigil 2008: 5; García-Acevedo 2000: 226–229; Gonzales-Berry & Maciel 2000: 3–4; Lipski 2008: 192).

Due to the remoteness of the region, and the arid and mountainous terrain, following settlement in 1598 from New Spain (what today is Mexico), New Mexico existed as a relatively isolated colony of the New World. Spanish speakers in the northern section of the state had minimal contact with speakers of other varieties of Mexican Spanish (Gonzales-Berry & Maciel 2000: 4; Lipski 2008: 195, 202), developing over time their own distinct variety of Spanish which we refer to here as 'Traditional New Mexican Spanish', following Bills and Vigil (2008: 8). Traditional New Mexican Spanish is said to be generally Mexican grammatically but distinguishable lexically, as richly mapped in the recently published linguistic atlas (Bills & Vigil 2008). An example is given in (1) below. Another oft-cited characteristic is the use of 'archaisms' such as *cuasi* 'almost' (standard *casi*), *trujo* 'he brought' (standard *trajo*), and *vide* 'I saw' (standard *vi*) (listed in the pioneering work of Aurelio M. Espinosa in the early 1900s (1930, 1946, among others)). Morphological features include the use of *ha* for the first-person present perfect instead of standard *he* (e.g. *ha comido/hamos comido* vs. *he comido/hemos comido* 'I/we have eaten' (Bills & Vigil 2008: 145–151)). Among phonological features is variable aspiration of syllable-initial /s/, which has received some empirical attention (Brown 2005a, 2005b; Lipski 2008: 204–207).

(1) Different words for 'pliers' in the Spanish of Mexico and New Mexico

Dolores *.. I never heard the word,*
pinzas.
Clara *[(THROAT)].*
Dolores *[like] growing up,*
it was just --
(H),
Clara *pliers.*
Dolores *% --*
pliers,
or tenazas,
we'd call them tenazas,
tenacitos.
but not --
we never .. said like pinzas.
Clara *mhm.*
Dolores *and Mexicans use like,*
.. pinzas. [22 Farolitos: 0:41:14–0:41:26][4]

[4] All examples given are from NMSEB (Torres Cacoullos & Travis in preparation). Examples are reproduced verbatim from the transcripts (transcription conventions are presented in Appendix III). Within brackets is the recording number, name and time stamp. In examples where Spanish and English are used, the original appears on the left, and the translation on the right, and speech that was originally produced in English appears in italics. Though this is evident for multi-word sequences, identifying just what language single-word segments of speech belong to is less than straightforward, requiring evaluation of their patterning (Poplack & Meechan 1998); see Torres Cacoullos and Aaron (2003) for evidence that in this community single English-origin nouns (e.g. *policeman* in example (10) below) behave as borrowings and not code-switches.

Traditional New Mexican Spanish came into contact with English before it came back into contact with Mexican Spanish (or varieties thereof). In 1848, Mexico lost the Mexican–American war, and with it one half of its land, including New Mexico, which was declared a territory of the USA. Anglophone settlers poured into the greater Southwest region, though far less into New Mexico than neighbouring areas, partly because it did not have the same level of natural resources (such as copper, gold and good land for cotton growing (Bills & Vigil 2008: 2)). As a result, English speakers were in the minority in New Mexico for longer than in the surrounding region: in 1890, 70% of the population of New Mexico could not speak English, a figure which dropped to 50% in 1900 and 33% in 1910 (Fernández-Gibert 2010: 48). However, New Mexico was awarded statehood in 1912, and English increasingly displaced Spanish in the educational system (cf. Gonzales-Berry 2000). Students were punished for speaking Spanish in schools, as poignantly recounted in one of the NMSEB narratives produced by a 58-year-old participant, given in example (2).

(2) Punishment at school

Pedro	*...(1.4) pero me acuerdo que,*	‘...(1.4) but I remember that,
	when we were in elementary,	*when we were in elementary,*
	. . . (1.0) if you got caught uh=,	*. . . (1.0) if you got caught uh=,*
	.. s- --	*.. s- --*
	uh speaking anything but English,	*uh speaking anything but English,*
	... (1.1) uh=,	*... (1.1) uh=,*
	you had to pay a price.	*you had to pay a price.*
	... @@	*... @@*
	.. (H) ... [@]	.. (H) ... [@]’
Ricardo	*[qué precio]?*	‘[what price]?
	... (0.8) @@@@	... (0.8) @@@@’
Pedro	*... yeah=.*	‘... yeah=.
	... estaba en el jarit- --	... you were in the ca- --
	.. la jarita.	.. the cane.’
Ricardo	*... la jarita.*	‘... the cane.’
Pedro	*... usaban la jarita or you had to go out and get a load of wood.*	‘... they would use the cane *or you had to go out and get a load of wood.*
	... @@[@@@@]	.. @@[@@@@]’
Ricardo	*[@@@]*	‘[@@@]’
Pedro	*.. (H) @@@ (H)*	‘.. (H) @@@ (H)’
Ricardo	*always bringing wood no?*	*‘always bringing wood no?’*

[10 El timbre Portátil: 0:11:00–0:11:24]

From the 1900s (with the Mexican Revolution, 1910–1920), immigration from Mexico has brought New Mexican Spanish into increasing contact with contemporary varieties of Mexican Spanish. While this of course helps the retention of Spanish overall, Traditional New Mexican Spanish is stigmatized in comparison with monolingual varieties, as indicated by the speaker in example (3), and thus in fact, the spread of Mexican Spanish threatens the future of Traditional New Mexican Spanish (Bills & Vigil 1999: 56).

(3) Prestige of Mexican Spanish

Trinidad	*la mayor de mis hijas,*	'my oldest daughter,
	... eh te- tenía= amigas en colegio de --	... um had friends at school from --
	en New Mexico State,	at New Mexico State,
	... que= hablaban --	... that spoke --
	que eran de México,	that were from Mexico,
	... de modo que ella aprendió muy bien,	... so she learned very well,
	a hablar.	to speak.'
Jake	*.. hm[=].*	'.. hm[=].'
Trinidad	*[l]o hablaba muy bonito,*	' [she] spoke very nicely,
	como los de México.	like people from Mexico.
	... porque aprendió más por ella,	... because she learned more on her own,
	que por nosotros.	than from us.'

[21 Actividades: 0:03:21–0:03:40]

A further threat to Traditional New Mexican Spanish is the Spanish that is taught in the schools as a foreign or second language, where the target is generally the 'educated standard', which also achieves prestige status over the traditional variety (Gonzales-Berry 2000). This disparagement of the local variety is displayed in example (4), about the speaker's granddaughter having her homework—which the speaker had helped her with—marked as wrong. In this way, contact with English as well as contact with Mexican Spanish and the educated standard are all said to be playing a role in the demise of Traditional New Mexican Spanish (Bills & Vigil 2008: 313).[5]

(4) Prestige of Educated Standard

Inmaculada	*.. they c- called it proper Spanish.*
Lucy	*mhm.*
Inmaculada	*o=r,*
	whatever,
	it was called,
	(H) but it wasn't our Spanish.
Lucy	*hm.*
Inmaculada	*so she got everything wrong.*

[14 Calcetines, medias y mallas: 0:26:25–0:26:32]

NM Spanish can therefore be considered an endangered variety. This is an important issue because, as Wolfram notes, the loss of a dialect, even one of a widely spoken language such as Spanish, represents 'the loss of significant linguistic information and linguistic–cultural identity parallel to the loss incurred when a language dies' (2000: 23). Traditional New Mexican Spanish has been documented in terms of its lexicon through the New Mexico Colorado Spanish Survey (Bills & Vigil 2008, inter alia), which includes recordings, made in the early 1990s, of older

[5] For recent overviews of the linguistic history of New Mexico, see Bills and Vigil (2008) and Travis and Villa (2011).

(near-)monolingual speakers of Traditional New Mexican Spanish. For students of language contact, the locus of which, as stressed by Weinreich (1968), is the bilingual speaker, the remaining speakers of Traditional New Mexican Spanish and English represented in NMSEB provide a precious window into bilingual speech phenomena.

2.2. A Community-based Speaker Sample

Members of a speech community share 'a set of values that are correlated with their use of linguistic variables', for example, members agree on the negative evaluations of stigmatized features, regardless of how frequently they themselves use these, as shown in style shifting or matched guise tests (Labov 2005: 9). While data are, of course, collected from individual speakers, 'individuals are not the final units of linguistic analysis, but the components that are used to construct models of our primary object of interest, the speech community' (Labov 2001: 34). We thus leave behind not only studies based on one or two individuals (who may not be constrained by group norms) but also generalizations from larger numbers of assorted speakers or participants of unknown social characteristics.

Established practice for collecting community-based data is through a stratified random sample, that is, a quota of participants who represent the specific social characteristics of interest to the study at hand, such as age, sex, socioeconomic status, ethnicity (Tagliamonte 2012: 23–34) or, in the study of bilingual areas, neighbourhood demography (Poplack 1989: 413–414). However, in the case of NMSEB, given ongoing shift to English in the New Mexican Hispanic community and our interest in the synchronic study of language contact, our primary criterion for inclusion was speaker bilingualism, with secondary concern for sample distribution by age and sex.

Speakers comprising NMSEB are bona fide New Mexicans who are bilingual. First, participants had to be no less than third-generation *Nuevomexicanos* 'New Mexicans' (applying what we refer to as the *abuelitos* test—the grandparents and parents of the participants must have been born in New Mexico). Second, NMSEB participants had to be speakers who regularly use both languages with the same interlocutor in the same domain, 'the appropriate code for the Hispano community' (Gonzales Velásquez 1995: 29). That is, rather than administering tests of proficiency, we rely on the criterion of regular use of both languages, as observed by the fieldworkers (and subsequently confirmed in the course of the interview) (cf. Poplack 1993: 254). Consonant with meeting this usage requirement was speakers' own self-rating, such that upon being asked to rate their Spanish and English on a scale of one to five, with one being minimally and five maximally good, all NMSEB participants rated themselves as either a four or a five in each language. Of some 60 interviews of between 30 minutes and one-hour long initially collected, we selected 31 with copious switching for inclusion in this corpus, for a total of 30 hours of speech from 41 speakers.

Figure 1 gives the distribution of Hispanics across the different counties in New Mexico, which range from 13.6% to 79.7% Hispanic. The major birthplaces of the

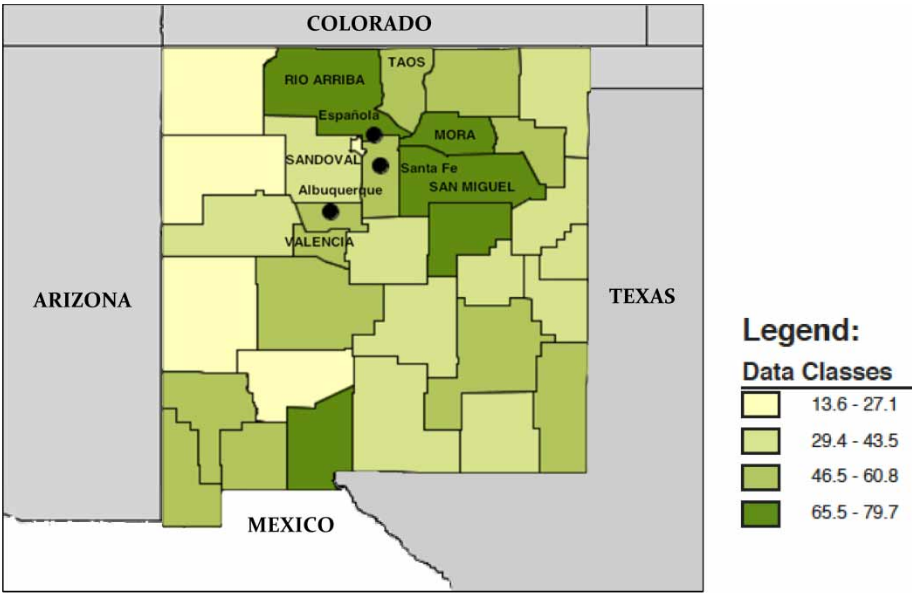

Figure 1 Percent of people who are Hispanic or Latino in the state of New Mexico (2007–2011 American Community Survey 5-Year Estimates), and major birthplaces of NMSEB participants (COUNTIES and Cities)

NMSEB participants are marked, counties in upper case and cities in lower case (see Appendix I for more details of the speakers). As can be seen, all are from northern New Mexico, and most are from counties with a high proportion of Hispanics (in particular, Rio Arriba (12 of 41 participants) and Taos (10/41)), as well as the city of Española (8/41) and a few from the urban centre of Albuquerque (only two born in Albuquerque, six residing there at the time of the interview).

Table 1 depicts the distribution of the speakers by sex and age (see Appendix I). At first glance we can see that, while the distribution by sex is fairly even, 61% (25/41) of the speakers are over the age of 60 and only six speakers are between the ages of 18 and 44 at the time of recording (2010–2011). This is consistent with the loss of intergenerational transmission of Traditional New Mexican Spanish, especially following the 1940s, with the acceleration of northern New Mexico villagers' contact with English (Gonzales Velásquez 1995: 20). As one speaker succinctly put it,

Table 1 Age and sex: NMSEB participants

Age	Female	Male	Total	%
18–44	3	3	6	15
45–59	7	3	10	24
60–74	10	3	13	32
75–89	3	9	12	29
Total	23 (56%)	18 (44%)	41	

(5) Loss of Spanish across generations

Adriana	... *(1.0) (TSK) y=,*	'... (1.0) (TSK) and,
	de ellos?	what about them?
	... *todos hablan español?*	... do they all speak Spanish?
	... *[o=],*	... [or=],'
Rocío	... *[(TSK)] (H) uh=,*	'... [(TSK)] (H) um,
	.. *mis hijas,*	.. my daughters,
	todas.	all of them.'
Adriana	... *mhm.*	'... mhm.'
Rocío	.. *mis nietos,*	'.. my grandchildren,
	no.	don't.'

[05 Las Tortillas: 0:49:09–0:49:11]

3. Data Collection Procedures

What kind of data offer insights on patterns of language use within the speech community? Intuitions about variable phenomena, especially ones involving stigmatized features, are known to be unreliable, as are generalizations based on 'overheard' language use, due to the problem of categorical perception, which might make a few unsystematically collected examples seem like an overall tendency (Poplack 1993: 253; Torres Cacoullos 2012: 233). Likewise, the formality imposed by controlled elicitation and experimental methods renders such methods particularly unsuitable to the collection of production data for non-standard varieties (cf. Sankoff 1988: 145–146). Rather, the most appropriate data collection method is direct observation of what speakers actually do, which requires recordings of large amounts of spontaneous everyday speech, or the *vernacular.*

3.1. The Sociolinguistic Interview

The *vernacular* is defined as the form of language that is acquired first and that is used with friends and family. Crucially, it is the style that provides the most systematic data for linguistic analysis as it is where the least attention is paid to speech and thus where unreflecting use of linguistic forms, least affected by self-monitoring, may be studied (Labov 1972: 112). This of course creates a paradox, in that we aim to observe how people talk when they are not being observed (Labov 1972: 113). One way of getting around this is through recordings of spontaneous conversation between friends and acquaintances. While very natural data can be collected in this way, a common limitation is that the friends and acquaintances belong to a restricted set of social circles that the researcher has direct access to, and it therefore can be difficult to obtain a stratified sample, essential if we are to understand how language is used across the community. Further, such an approach may detract from comparable data, given great variability in spontaneous conversation in terms of, for example, topic and relationship between the interlocutors. And finally, information on language attitudes (relevant to the interpretation of the data) is unlikely to come up naturally in spontaneous conversation.[6]

In order to circumvent these problems, Variationists work with the *sociolinguistic interview*, which, despite its name, is not an interview as such but rather a semi-directed dialogue, centred around a loosely structured set of topics that are considered to be of interest to the participant (Labov 1984: 32–42). Interviews may be conducted with one or more participants, and topics covered include what can be considered 'outsider' topics, of general interest and relevance across communities, such as family, work, school and childhood, as well as 'insider' topics, related to the culture and lifestyle of the community (in New Mexico, roasting green chile, for example). Topics about language are also included (for NMSEB, experiences with Spanish, English and bilingualism), through which information about the speaker's linguistic history and their language attitudes is collected, information which can inform the analysis (as we discuss in Section 4 below). What is most important is that through these topics, we seek to elicit narratives of personal experience for which participants, not the researcher, are the indisputable experts and during which monitoring of speech is minimized. Note that the sociolinguistic interview thus conducted consists of both more monologic and more interactional discourse, illustrated in examples (6) and (7) respectively.

(6) Narrative of personal experience collected via the sociolinguistic interview

Sandra	*... (H) he lit my cigarette,*	*'... (H) he lit my cigarette,*
	.. like a gentleman,	*.. like a gentleman,'*
Adriana	*@@*	'*@@*'
Sandra	*and then he lit his own.*	*'and then he lit his own.*
	... and there we were,	*... and there we were,*
	you know con la pata cruzada,	*you know* with our legs crossed,
	just más suave,	just real easy,
	smoking away.	*smoking away.*
	.. (H) y aquí viene la Betsy and she said,	and here comes Betsy *and she said,*
	... (1.0) (H) < VOX ah=,	... (1.0) (H) < VOX oh =,
	yo le voy a decir a mi daddy VOX >.	= I'm going to tell my daddy VOX >.
	yo le voy a decir a mi daddy que ustedes estaban chupando	I'm going to tell my daddy that you were sucking.
	... (1.4) no fumando.	not smoking.
	.. chupando.	sucking.
	... (H) a = nd Matt jumped down,	*... (H) a = nd Matt jumped down,*
	he says,	*he says,*
	no no no mijita,	no no no sweetheart,
	no le vas a decir al daddy.	you're not going to tell daddy.
	.. (H) déjame te llevo --	.. (H) let me take you --
	< X where X > there was a little store up here,	*< X where X > there was a little store up here,*
	by the highschool?	*by the highschool?*
	not very far.	*not very far.*
	... (THROAT)... (0.7) and he says,	*... (THROAT)... (0.7) and he says,*

[6] On problems associated with collecting and working with conversational data, see also Clyne *et al.* (2001: 236).

	.. mira,	look,
	te voy a llevar a la tiendita,	I'm going to take you to the store,
	te voy a comprar lo que tú quieras.	I'm going to buy you whatever you want.
	(H) so he went to his room and got all his ... (H) life savings,	*(H) so he went to his room and got all his ... (H) life savings,*
	< @ you know @ > ?	*< @ you know @ > ?*'
Adriana	*@@@*	'*@@@*'
Sandra	*@@*	'*@@*
	and he took her to the little store,	*and he took her to the little store,*
	llegó with a coke and,	'he arrived *with a coke and,*
	... (H) and with um,	*... (H) and with um,*
	... all kinds of candy and,	*... all kinds of candy and,*
	.. (H) gum.	*.. (H) gum.*
	bubble gum,	*bubble gum,*
	y quién sabe qué.	and who knows what.'
		[03 Dos Comadres: 0:15:47–0:16:35]

(7) Conversation collected via the sociolinguistic interview

Molly	*.. estuvo en = .. la mina in Gallup,*	'.. he was in the mine in Gallup,
	y luego,	and then,
	... (1.2) (H) y luego se fue pa',	... (1.2) (H) and then he went to,
	... pa' el estado,	to the state,
	no?	you know?'
Fabiola	*... oye pues,*	'... hey,
	... qué tú nunca pudiste aplicar por la,	... you didn't ever apply for the,
	... ese dinero que les están dando a los =,	... that money that they are giving the,
	.. mineros?	miners?'
Molly	*... (0.7) no le he hecho el try nunca no.*	'... (0.7) I haven't tried you know.
	.. pero,	.. but,'
Fabiola	*.. how come?*	'*.. how come?*'
Molly	*... (0.7) I don't remember cuando fue cuando trabajó,*	'*... (0.7) I don't remember* when it was when he worked,'
	como en seventy algo,	like in seventy something,
	.. creo no.	.. I think.'
Fabiola	*.. pero yo creo que ellos tienen records si aplicabas.*	'.. but I think that they have records if you applied,
	... lo buscaban.	... they'd look for them.
	... (1.0) no tienes check stubs de lo que trabajó él allá?	... (1.0) you don't have check stubs from when he worked there?'
Molly	*... yo creo que los tiré porque ya no =,*	'... I think that I threw them out because,
	... estaban muy viejos no pero,	... they were very old you know but,'
Fabiola	*... (1.7) pero even though si los --*	... (1.7) but *even though* if –
	even though que los tiraste,	even though you threw them out,
	yo creo que todavía podías.	I think you still could.
	... (1.4) you should try it.	*... (1.4) you should try it.*'
		[09 La Salvia: 0:11:14–0:11:54]

3.2. In-group Fieldworkers

As the vernacular is not used with outside observers, data collection should be carried out by in-group members (cf. Clyne *et al.* 2001: 235–236; Poplack 1993: 260). In the collection of the NMSEB corpus, we worked with a team of Hispanic New Mexican students of the University of New Mexico as research assistants who interviewed family members and acquaintances. The recordings making up the corpus were collected by eight students (seven undergraduate and one MA), all minimally-third-generation bilingual *Nuevomexicanos*,[7] four males and four females, between the ages of 21 and 30. All spoke some Spanish in the home but many had also studied Spanish at school and at the University of New Mexico (in the Heritage Language Program, specific for students with at least some home background in Spanish), and many had travelled in Spanish-speaking countries.[8] We gratefully acknowledge the invaluable contributions of these eight interviewers, for most of whom this was their first research experience: Daniel Abeyta, Rubel Aguilar, Raúl Aragón, Cheryl Conway, Jason Gonzales, Amanda Ortiz, Lillian Sánchez and Kamie Ulibarrí.

Most of the participants lived in northern New Mexico (where most of the interviewers were from), which meant that the interviewers travelled from Albuquerque back to their hometowns to make the recordings. In summer 2011, four students took a weekend fieldtrip to northern New Mexico and in this intensive effort collected 14 of the 31 interviews that make up the corpus. The following excerpt, reproduced from the diary of the MA student who led the field trip, voices the value for the fieldworkers of being able to engage in conversations with their community that they had never had before.

> … Although many of the people we interviewed were family friends, for many of us, it was the first time we had intentionally taken the time to sit down, have a real conversation, and hear their life stories. We even had the pleasure of meeting distant family members for the first time. It was convicting to think that we probably never would have met these people if it had not been for this summer job. Our eyes were opened to parts of our own communities we were previously oblivious to, mainly due to believing that we were too busy or had more important things to do.… On our journey, we discovered treasures such as friendship, laughter, proverbs, history, and food. Each of us found ourselves feeling part of something bigger than ourselves, connected to our community in a deep and meaningful way.… (Daniel Abeyta, June 2011)

4. Sociolinguistic Profile

An essential accompaniment to the linguistic data in the corpus is the demographic characterization of the speakers and the community they represent, including information about their linguistic history and their language attitudes, because

[7] We were less strict in applying this criterion in the case of the interviewers—two of our interviewers had one parent of Mexican and the other of New Mexican, origin, while all of our participants had both parents and all grandparents of New Mexican origin.

[8] Due to this mixed background, we do not include the speech of our interviewers in our analyses based on the corpus (a total of five hours (51,000 words) of our 30 hours of transcribed data are produced by the interviewers). Nor do we include the speech of people appearing in the recordings who do not meet our criteria for inclusion (a total of 18 minutes (3,000 words) from 14 such speakers).

material of unknown provenance is not only questionable but ultimately uninterpretable as data.

There are different ways of gathering pertinent information. One is through a questionnaire, which is useful for straightforward questions (such as the year and place of birth of the speaker). Another is through a content analysis of the sociolinguistic interview, through systematic extraction of any comments arising during the recording that are relevant to language attitudes and the participants' linguistic history (Poplack 1993: 271; Poplack *et al.* 2006: 191). We applied both of these in our compilation of the NMSEB corpus.

Subsequent to the sociolinguistic interview, a short questionnaire was conducted (Appendix II), which gathered information on participants' relative ratings of their own Spanish and English, as well as domains of use, and contact with, the two languages. Figure 2 presents a summary of the responses to questions regarding domains of use, and shows that NMSEB participants report that they use both languages not just in the home, but also with friends and at work, corroborating the bilingual nature of the community in which these speakers live. English is preferred for TV as well as for reading books, and reading newspapers is the only domain where Spanish is not used. Radio music is the only domain where Spanish is preferred. The different preferences for radio and TV may be explained by the existence of both local Spanish and local English radio stations that play New Mexican music, in contrast to the lack of local Spanish content on the TV (except for the news, Spanish programmes are from (inter)national networks).

The questionnaire also provided a measure of the degree of language contact. While as an artefact of our inclusion criteria all participants are bilingual, they vary in terms of how much contact they have with English. As an initial way to quantify this,

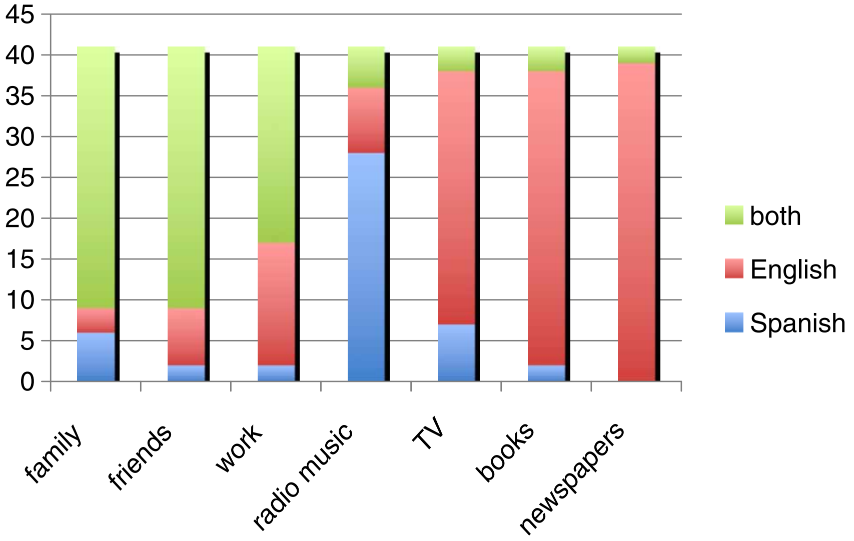

Figure 2 Preferred language across domains: NMSEB participants ($N = 41$)

we established a 'Contact with English' (CWE) Index by assigning a score to each speaker's response to 12 items in the questionnaire related to first and preferred language; where English was learned (home, school or both); the level of education attained;[9] self-rating of English relative to Spanish; language(s) spoken with family, among friends and at work; the preferred language on the radio and TV and for reading newspapers and books. Assigning each factor a score of 3 to indicate more contact with English, 1 to indicate more contact with Spanish, and 2 when a response of 'both' was given, yields an overall average CWE score for the 41 speakers of 2.09—at just about the midpoint between the two languages—with a range from 1.67 to 2.75 (Cronbach's $\alpha = 0.6697$ indicates that this index approaches a reliable scaling) (see Appendix I for the scores for each speaker).

Various tests can be conducted to examine correlations between index factors as well as with other variables, such as birthplace or region of current residence (for an example of Principal Components Analysis, see Hoffman and Walker (2010)). This provides, then, a measure against which to compare usage patterns, and enables us to directly test the hypothesis that participants with a greater contact with English will show (more) grammatical convergence (as has been widely implied (e.g. Thomason 2001)). More generally, the incorporation of such measures is crucial for evaluating claims about the linguistic consequences of contact, some of which may be pertinent to situations of stable (long-term) bilingualism and others to situations of shift over the course of one or two generations (as with most US immigrant groups) (cf. Silva-Corvalán 1994).

A limitation of such questionnaires is that they only reveal information about topics devised by the researcher. A content analysis of the sociolinguistic interviews, on the other hand, allows the analyst to more fully characterize the attitudes and experiences of a speech community, not only because it provides illustrations of tendencies in the questionnaire responses, but because it brings out issues and attitudes that are relevant to the community, without imposing pre-determined categories as attitude questionnaires do. We have seen several examples that were extracted during our content analysis above (on metalinguistic awareness of differences between New Mexican and Mexican Spanish (example (1)), punishment at school for use of Spanish (example (2)), the prestige of Mexican Spanish (example (3)) and of the educated standard (example (4)), and loss of Spanish across generations (example (5))). A compilation of such comments, systematically extracted, can provide a sociolinguistic profile of the community. (For a model of such content analysis, see Poplack *et al.* (2006: 196–207).)

[9] As education takes place in English-medium institutions, the higher the level of education, the greater the degree of contact with English.

5. Transcription

In order to create a corpus that, beyond the cherry-picked example, can be used for accountable quantitative analysis, it is necessary to transcribe the audio data. As noted above, of the roughly 60 interviews initially collected, one half of them were selected to form part of the transcribed corpus. This was due in some cases to poor audio quality, and in others because we prioritized those recordings in which both languages were used sufficiently to permit systematic quantitative analysis of both of the bilingual speaker's languages. Though the amount of Spanish and English varies in the recordings and the quantification of language contact phenomena awaits a series of studies, this promises to make NMSEB an invaluable bilingual corpus, with a roughly even distribution of speech produced in Spanish and English, and smooth code-switching throughout.

Transcriptions were done in *Elan*, a software programme that aligns the audio file to the transcription (Lausberg & Sloetjes 2009). To ensure accuracy, the transcriptions went through five rounds of revisions. All recordings comprising the corpus were transcribed initially by a Spanish–English bilingual (from New Mexico or from Mexico and currently living in New Mexico; in some cases, the interviewer him/herself), and then revised by a second bilingual who was dominant in the language distinct from that of the first transcriber. This was essential for these highly bilingual data, as the second transcriber was often able to catch some of the material that was produced in their dominant language that the first transcriber had misheard.

The third revision took place as the transcripts were anonymized: transcribers reviewed the transcripts for all identifiers (names of people, places, institutions, businesses) and modified these in the transcription, and in a fourth revision sound files were acoustically blurred by low-pass filtering with a *Praat* script.[10] Finally, the transcriptions, exported into *Excel*, were carefully read for content and any anomalies that were found were checked in the sound file and modified where appropriate.

We would estimate that, on average, for two-party conversations, close to 50 hours was spent on transcription for each hour of recorded data, and longer for those with more participants.[11] Note that this is comparable to other studies using a similar level of meticulous care in corpus preparation for similar two-party conversation (the Ottawa-Hull French project (Poplack 1989) and the Santa Barbara Corpus of Spoken American English (John Du Bois, PC and Robert Englebretson, PC)). Although this is an enormous investment of time and human resources, we see each step as absolutely essential to produce a quality transcription that represents as accurately as

[10] The script for anonymizing the sound files was written by Chris Koops.

[11] This is based on approximate calculations of 18 hours to prepare the first draft of the transcription; 15 hours to revise that; nine hours to do the third revision, including removing all identifiers and substituting appropriate pseudonyms (both of names and places, for the latter using neighbouring towns of a similar size) (see Section 5.2 below); two hours to anonymize the sound file which constituted another check on the transcription; and a further four hours to undertake the content revision and make required corrections. Multi-party conversations took longer than one-on-one interviews between the fieldworker and a single speaker due partly to a much higher occurrence of overlapping speech.

possible the content of what is said and, at the same time, includes prosodic information essential to the understanding of the transcript.

5.1. Intonation Units for Transcribing Speech

The data have been transcribed in accordance with the approach developed at the University of California, Santa Barbara (cf. Du Bois *et al.* 1993). This transcription method includes the annotation of features such as pauses, false starts, laughter, overlap, inhalation and so on, which may play a role in the interaction between the interlocutors and which must be considered if questions of bilingual cognitive cost are to be investigated in natural speech data (e.g. Dumont 2010). Fundamental to this approach is the notion of the Intonation Unit (IU), that is, 'a stretch of speech uttered under a single coherent intonation contour' (Du Bois *et al.* 1993: 47), each of which is represented on a distinct line in the transcription, and is followed by punctuation which represents the prosodic contour of that IU. Transcribed, our 30 hours of recorded speech add up to approximately 97,000 Intonation Units and 320,000 words.

IUs closely interact with syntax, such that material that is realized in the one IU often shows a tighter syntactic relationship than material that is spread across IUs. As an illustration, consider example (8) below, from our data, but not transcribed in IUs. Of particular interest to us is the first-person singular subject *yo*, as one project underway with this corpus concerns patterns of subject expression (Torres Cacoullos & Travis 2010, 2011). Without an indication of the prosody it is impossible to know if the *yo* (in bold) is a post-verbal subject on *dije* ('said I'), or a preverbal subject on *no puedo* ('I cannot'). However, once IUs are marked, as in example (9), it becomes clear that *yo* is a post-verbal subject on *dije*, while *puedo* has an unexpressed subject. (On the relevance of IUs to the study of morphosyntax, see Chafe (1994: Ch. 9), Ono and Thompson (1995: 233–246), Sánchez-Ayala (2001) and Torres Cacoullos and Travis (Forthcoming).)

(8) Ivette: *dije* ***yo*** *no no puedo estar yendo pa' atrás y pa' adelante.*
'.. I said I can't be going backwards and forwards.'

(9)	Ivette	*.. dije yo,*	'.. I said,
		No.	No.
		no puedo estar yendo pa' atrás y pa' adelante.	(I) can't be going backwards and forwards.'

[05 Las Tortillas: 0:43:54–0:49:11]

Thus, prosodically-based transcription can be crucial for both interpretation and analysis of morphosyntactic variables. In bilingual discourse, furthermore, IUs may play a role in the patterning of code-switching sites (cf. Durán Urrea 2012; Shenk 2006) and hence offer promise of deeper understanding of the ways in which languages are combined.

5.2. The Social Responsibility of the Linguist

As noted by England (1992: 33–34) 'a description of a language provides part of a social description of the people who speak that language'. This must be borne out both in preparing the corpus and reporting from it, in particular in relation to the representation of speech, or the orthography chosen; the removal of all identifiers from the corpus; the type of examples chosen to illustrate the points under consideration; and granting access to the corpus. While this is relevant for all linguists, it is particularly crucial for those working with speakers from minority communities and stigmatized varieties who may be more likely to be the recipients of negative stereotyping.

We begin with consideration of the orthography, which clearly is an issue for languages without a standardized writing system, but it is also not straightforward for languages where there is an accepted standard, such as English and Spanish.

A full phonetic transcription of a corpus this size is impractical, and this alone would detract from searchability and concordancing. As general practice, it is therefore best to use standard orthography, leaving phonetic transcription to be undertaken as a separate step of relevant analyses of the corpus. Not only will the attempt to capture phonetics with the alphabet fail, but, perhaps more importantly, non-standard spellings have social meanings attached to them, and can create negative perceptions (cf. Jaffe & Walton 2000).

However, given that variation is a significant linguistic fact to be confronted, it is also not desirable to obliterate all the variation exhibited in the data. We therefore follow Poplack (1993: 265–266) in using standard orthography for all phonetic or phonological variation, but non-standard spellings for morphological variation. For example, despite widespread aspiration and elision of /s/ in Traditional New Mexican Spanish, we transcribed an *s* in all cases where it was clear from the context that standard morphology or lexicon would predict an *s*, such as with second-person singular verbs, plural nouns and adjectives or in monomorphemic words that in the standard have an *s* (such as *así* 'like this'), regardless of how the form was pronounced (thus, always *los muchachos* 'the boys', and never *loh muchachoh* or *lo muchacho*, for example). Likewise, the variable New Mexican nasalization in *muncho* 'much' and *ansina* 'like this' was not transcribed, and instead we consistently used *mucho* and *asina* (though note that *asina* is itself a non-standard form, the standard being *así*). Such phonetic features can be fruitfully analyzed in the future with appropriate phonetic transcription of the relevant segments as dictated by the research question. On the other hand, we capture morphological forms such as the first-person plural person suffix *-nos* (instead of standard *-mos*, e.g. *estábanos* vs. *estábamos* 'we were'), and other verbal forms such as Imperfect *traíba* ('I brought', instead of *traía*) and Preterit *trujo* (instead of *trajo*). Further, some cases of accepted abbreviations were used including, in English, *cause* (for *because*), *gonna, wanna* and *kinda* and in Spanish *pa'* (for *para* 'for/toward', and likewise *pa'ca* for *para acá* 'toward here').[12]

Another key issue in terms of protecting the participants relates to obtaining their consent and strictly ensuring their confidentiality through the use of pseudonyms, at least when that is the desire of the participants (cf. Bowern 2010: 903). Surreptitious recordings are outright rejected for both practical and ethical reasons—they are likely to be of poor sound quality, and, even if the participant is asked to grant permission following the recording, in the long term this kind of deception will damage contacts with the community (Labov 1984: 51). For NMSEB, we secured consent—prior to any recording—on a written form approved by the Institutional Review Boards of the universities to which the researchers were affiliated.[13]

Speaker confidentiality is particularly relevant for small, close-knit communities such as this one, where participants can readily be identified, not just by their names, but by life details as well. We have thus anonymized names of people and locations, as well as nicknames, some jobs held, places of military postings, and so on, where these might identify the participant. For example, in the following example, we did not use a pseudonym for *navy* as sufficient members of the community have been in the navy that this would not allow for the participant to be identified; *tailor* and *policeman*, on the other hand, are 'pseudo-professions' for professions that may be specific to this referent.

(10)	Gabriel	.. *then he joined the Navy.*	'.. *then he joined the Navy.*
		... *(1.3) (TSK) y cuando regresó he was a tailor,*	... (1.3) (TSK) and when he came *back he was a tailor,*
		... *(H) and um =,*	... *(H) and um =,*
		... *y luego se hizo policeman.*	... and then he became a policeman.'
			[18 Las Minas: 0:35:44–0:35:54]

Selection of examples to illustrate linguistic phenomena is another area where the linguist must be socially responsible, by choosing examples that illustrate the linguistic point but do not contain other information that may create a mistaken perception or reinforce stereotypes of the community. For example, the inclusion of several examples related to drunkenness and abuse can create the impression that these are norms of the community, and thus, regardless of how well they illustrate the particular linguistic point under consideration, alternative examples should be given instead. The use of stigmatized features in bilingual communities should equally be

[12] Transcribers uniformly followed a Transcribing Protocol, viewable at the project website (http://nmcode-switching.la.psu.edu/) under the Tools tab. For discussion of related issues, see Otheguy and Zentella (2012: 40–41).

[13] University of New Mexico IRB #10-295 for Travis and Penn State University IRB #34265 for Torres Cacoullos.

treated with caution. If all examples contain other-language-origin words, that will reinforce an inflated impression of the frequency of lexical borrowing in the community under study (on the frequency of borrowing, see Poplack and Dion (2012)). Instead, a mix of examples should be sought, reflecting distributions observed in the corpus. In making a call for consideration of such issues, England notes that

> a request to use additional selection criteria for examples [is not] unscientific tampering with the data; it is instead a plea for sensitivity in the presentation of the data, and in many cases it is a plea for more accurate reporting in the data. (1992: 31)

A final issue regarding the social responsibility of the linguist concerns the accessibility of the corpus, both to the community (see Wolfram *et al.* (2008), on a variety of ways to engage with the community) and beyond. While it might seem desirable to make the data as widely accessible as possible (e.g. freely available on the web), some thought should be given to potential, if unintended, misuse of the data. It cannot be assumed that even all linguists, let alone others, will exercise the kind of caution prescribed above in extracting from the corpus, and this may in fact not be possible for those lacking familiarity with the speech community. Particular care should be exercised for minority communities speaking non-standard varieties, for which the predilection for spotlighting the expedient example (rather than doing the hard work of systematic quantitative analysis) may be exacerbated. Thus, we recommend some oversight in terms of who has access to the data and what it is to be used for, for example in the form of a written application, including an outline of the proposed project, as is currently being done for NMSEB.

6. Conclusion

Work remains to be done. Prominent among future tasks is broadly reporting the systematic structure of the linguistic varieties of this and other bilingual communities as revealed through corpus analysis. As well as advancing linguistic research, this can also serve the social goal of promoting greater public awareness of the realities of minority languages among teachers, journalists and policy makers, among others. Thus, for example, NMSEB will allow us to examine the putative role of code-switching in contact-induced structural change (Gumperz & Wilson 1971) and enable a community-based test of the hypothesis that certain kinds of words trigger code-switching (Clyne 2003), as well as serve more generally to scientifically characterize what is controversially termed ‘Spanglish’ (see, e.g. Otheguy & Stern 2010). For endeavours such as the NMSEB corpus to contribute to the maintenance as well as documentation of linguistic diversity, corpus materials can be incorporated into the development of pedagogical tools, providing an essential descriptive (rather than prescriptive) basis for heritage language teaching at the high school and university levels (cf. Wolfram *et al.* 1999).

We hope to have contributed an appreciation of the feasibility and fruits of community-based corpus constitution. The fruits bear stressing: faithful speech corpora constitute lasting linguistic and cultural archives in a way that data elicited to address particular questions or theories of the moment cannot. The community-based approach, as an object and means of study, yielding linguistic data situated in its social context and amenable to systematic quantitative analysis, allows for confronting the many claims about language contact with the facts of bilingual usage.

References

Aikhenvald AY 2002 *Language Contact in Amazonia* Oxford: Oxford University Press.

Bills GD & NA Vigil 1999 'Ashes to ashes: the historical basis for dialect variation in New Mexican Spanish' *Romance Philology* 53(1): 43–66.

Bills GD & NA Vigil 2008 *The Spanish Language of New Mexico and Southern Colorado: a linguistic atlas* Albuquerque: University of New Mexico Press.

Bowern C 2010 'Fieldwork and the IRB: a snapshot' *Language* 86(4): 897–904.

Brown EL 2005a 'New Mexican Spanish: insight into the variable reduction of "*la ehe inihial*" (/s-/)' *Hispania* 88(4): 813–824.

Brown EL 2005b 'Syllable-initial /s/ in Traditional New Mexican Spanish: linguistic factors favoring reduction *ahina*' *Southwest Journal of Linguistics* 24(1–2): 13–30.

Bybee J 2010 *Language, Usage and Cognition* Cambridge: Cambridge University Press.

Chafe W 1994 *Discourse, Consciousness and Time: the flow and displacement of conscious experience in speaking and writing* Chicago: University of Chicago Press.

Clyne M 2003 *Dynamics of Language Contact* Cambridge: Cambridge University Press.

Clyne M, E Eisikovits & L Tollfree 2001 'Ethnic varieties of Australian English' in D Blair & P Collins (eds) *English in Australia* Amsterdam: John Benjamins. pp. 223–238.

Dorian NC 2010 *Investigating Variation: the effects of social organization and social setting* New York: Oxford University Press.

Du Bois JW, S Schuetze-Coburn, S Cumming & D Paolino 1993 'Outline of discourse transcription' in J Edwards & M Lampert (eds) *Talking Data: transcription and coding in discourse* Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 45–89.

Dumont J 2010 'Testing the cognitive load hypothesis: repair rates and usage in a bilingual community' *Studies in Hispanic and Lusophone Linguistics* 3(2): 329–352.

Durán Urrea E 2012 *A community-based study of social, prosodic, and syntactic factors in code-switching* Unpublished PhD dissertation, The Pennsylvania State University State College.

England NC 1992 'Doing Mayan linguistics in Guatemala' *Language* 68(1): 29–35.

Espinosa AM 1930 *Estudios sobre el español de Nuevo Méjico, Parte I: fonética* Buenos Aires: Universidad de Buenos Aires.

Espinosa AM 1946 *Estudios sobre el español de Nuevo Méjico, Parte II: morfología* Buenos Aires: Universidad de Buenos Aires.

Fernández-Gibert A 2010 'From voice to print: language and social change in New Mexico, 1880–1912' in S Rivera-Mills & DJ Villa (eds) *Spanish of the US Southwest: a language in transition* Madrid: Iberoamericana. pp. 45–62.

García-Acevedo MR 2000 'The forgotten diaspora: Mexican immigration to New Mexico' in E Gonzales-Berry & DR Maciel (eds) *The Contested Homeland: a Chicano history of New Mexico* Albuquerque: University of New Mexico Press. pp. 215–238.

Gonzales Velásquez MD 1995 'Sometimes Spanish, sometimes English: language use among rural New Mexican Chicanas' in K Hall & M Bucholtz (eds) *Gender Articulated: language and the socially constructed self* New York: Routledge. pp. 421–446.

Gonzales-Berry E 2000 'Which language will our children speak? The Spanish language and public education policy in New Mexico, 1890–1930' in E Gonzales-Berry & DR Maciel (eds) *The Contested Homeland: a Chicano history of New Mexico* Albuquerque: University of New Mexico Press. pp. 169–189.

Gonzales-Berry E & DR Maciel (eds) 2000 *The Contested Homeland: a Chicano history of New Mexico* Albuquerque: University of New Mexico Press.

Gumperz J & R Wilson 1971 'Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India' in D Hymes (ed.) *Pidginization and Creolization of Languages* Cambridge: Cambridge University Press. pp. 151–167.

Hoffman MF & JA Walker 2010 'Ethnolects and the city: ethnic orientation and linguistic variation in Toronto English' *Language Variation and Change* 22(1): 37–67.

Jaffe A & S Walton 2000 'The voices people read: orthography and the representation of non-standard speech' *Journal of Sociolinguistics* 4(4): 561–587.

Labov W 1972 'Some principles of linguistic methodology' *Language in Society* 1(1): 97–120.

Labov W 1984 'Field methods of the project on linguistic change and variation' in J Baugh & J Sherzer (eds) *Language in Use: readings in sociolinguistics* Englewood Cliffs, NJ: Prentice Hall. pp. 28–53.

Labov W 2001 *Principles of Linguistic Change: social factors* Vol 2 Oxford: Blackwell.

Labov W 2005 'Quantitative reasoning in linguistics' in U Ammon, N Dittmar, KJ Mattheier & P Trudgill (eds) *Sociolinguistics/Soziolinguistik: an international handbook of the science of language and society, 1* Berlin: Mouton de Gruyter. pp. 6–22.

Labov W 2007 'Transmission and diffusion' *Language* 83(2): 344–387.

Lausberg H & H Sloetjes 2009 'Coding gestural behavior with the NEUROGES-ELAN system' *Behavior Research Methods, Instruments, & Computers* 41(3): 841–849 Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands Available at http://tla.mpi.nl/tools/tla-tools/elan/.

Lipski JM 2008 *Varieties of Spanish in the United States* Washington, DC: Georgetown University Press.

Ono T & SA Thompson 1995 'What can conversation tell us about syntax?' in PW Davis (ed.) *Alternative Linguistics: descriptive and theoretical modes* Amsterdam: John Benjamins. pp. 213–271.

Otheguy R & N Stern 2010 'On so-called Spanglish' *International Journal of Bilingualism* 15(1): 85–100.

Otheguy R & AC Zentella 2012 *Spanish in New York: language contact, dialectal leveling, and structural continuity* Oxford: Oxford University Press.

Poplack S 1989 'The care and handling of a mega-corpus: the Ottowa-Hull French project' in R Fasold & D Schiffrin (eds) *Language Change and Variation* Amsterdam: John Benjamins. pp. 411–451.

Poplack S 1993 'Variation theory and language contact: concepts, methods and data' in DR Preston (ed.) *American Dialect Research* Amsterdam: John Benjamins. pp. 251–286.

Poplack S 1998 'Contrasting patterns of code-switching in two communities' in P Trudgill & J Cheshire (eds) *The Sociolinguistics Reader: multilingualism and variation, 1* London: Arnold Publishers. pp. 44–65.

Poplack S & N Dion 2012 'Myths and facts about loanword development' *Language Variation and Change* 24(3): 279–315.

Poplack S & M Meechan 1998 'Introduction: how languages fit together in codemixing' *International Journal of Bilingualism* 2(2): 127–138.

Poplack S, JA Walker & R Malcolmson 2006 'An English "like no other"?: language contact and change in Quebec' *Canadian Journal of Linguistics/Revue canadienne de Linguistique* 51(2): 185–213.

Sánchez-Ayala I 2001 'Prosodic integration in Spanish complement constructions' in A Cienki, BJ Luka & MB Smith (eds) *Conceptual and Discourse Factors in Linguistic Structure* Stanford, CA: Center for the Study of Language and Information. pp. 201–213.

Sankoff D 1988 'Sociolinguistics and syntactic variation' in F Newmeyer (ed.) *Linguistics: The Cambridge survey. Vol. 4: Language: the socio-cultural context* Cambridge: Cambridge University Press. pp. 140–161.

Shenk PS 2006 'The interactional and syntactic importance of prosody in Spanish–English bilingual discourse' *International Journal of Bilingualism* 10(2): 179–205.

Silva-Corvalán C 1994 *Language Contact and Change: Spanish in Los Angeles* Oxford: Clarendon Press.

Tagliamonte SA 2012 *Variationist Sociolinguistics: change, observation, interpretation* Oxford: Wiley-Blackwell.

Thomason SG 2001 *Language Contact: an introduction* Washington, DC: Georgetown University Press.

Torres Cacoullos R 2012 'A milestone study: structured variability as the key to unraveling (contact-induced) language change' *Bilingualism: Language and Cognition* 15(2): 233–236.

Torres Cacoullos R & JE Aaron 2003 'Bare English-origin nouns in Spanish: rates, constraints and discourse functions' *Language Variation and Change* 15(3): 289–328.

Torres Cacoullos R & CE Travis 2010 'Variable *yo* expression in New Mexico: English influence?' in S Rivera-Mills & DJ Villa (eds) *Spanish of the US Southwest: a language in transition* Madrid: Iberoamericana. pp. 185–206.

Torres Cacoullos R & CE Travis 2011 'Using structural variability to evaluate convergence via code-switching' *International Journal of Bilingualism* 15(3): 241–267.

Torres Cacoullos R & CE Travis in preparation 'New Mexico Spanish/English Bilingual (NMSEB) corpus, National Science Foundation 1019112/1019122' Available at http://nmcode-switching.la.psu.edu/.

Torres Cacoullos, R & CE Travis Forthcoming 'Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation' *Journal of Pragmatics.*

Travis CE & DJ Villa 2011 'Language policy and language contact in New Mexico: the case of Spanish' in C Norrby & J Hajek (eds) *Uniformity and Diversity in Language Policy: global perspectives* Bristol: Multilingual Matters. pp. 126–140.

Weinreich U 1968 *Languages in Contact* The Hague: Mouton.

Wolfram W 2000 'Endangered dialects and social commitment' in JK Peyton, P Griffin, W Wolfram & R Fasold (eds) *Language in Action: new studies of language in society. Essays in honor of Roger W. Shuy* Cresskill, NJ: Hampton Press. pp. 19–39.

Wolfram W, J Reaser & C Vaughn 2008 'Operationalizing linguistic gratuity: from principle to practice' *Language and Linguistics Compass* 2(6): 1109–1134.

Wolfram W, C Temple Adger & D Christian 1999 *Dialects in Schools and Communities* Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix I

Table A1 NMSEB speaker characteristics

Speaker	Year born	Sex	Educ. level	Current residence	Birthplace	Occupation	CWE
Marta	1964	F	college	Río Arriba	Albuquerque	Guest services manager	2.75
Cristina	1973	F	college	San Miguel	San Miguel	Self-employed	2.75
Carmela	1978	F	college	Sandoval	Española	Teacher	2.58
Pedro	1953	M	college	Taos	Río Arriba	School administrator	2.5
Aurora	1962	F	college	Sandoval	Española	Teacher	2.5
Francisco	1963	M	high	Río Arriba	Río Arriba	Miner	2.42
Rocío	1945	F	high	Santa Fe	Santa Fe	Retired school teacher aid	2.33
Inmaculada	1952	F	college	Albuquerque	San Miguel	Social worker	2.33
Fabiola	1954	F	college	Taos	Taos	Secretary	2.33
Dolores	1963	F	college	Río Arriba	Española	School secretary	2.33
Susan	1934	F	high	Albuquerque	Albuquerque	Stay at home mom	2.25
Anita	1941	F	high	Albuquerque	San Miguel	Executive director	2.25
Trinidad	1938	F	high	Taos	Española	Substitute teacher	2.25
Leroy	1935	M	college	Río Arriba	Río Arriba	Retired government officer/rancher	2.25
Ivette	1946	F	high	Albuquerque	Valencia	Factory worker	2.17
Samuel	1922	M	college	Taos	Taos	School coach	2.17
Tomás	1989	M	high	Río Arriba	Río Arriba	Unemployed	2.17
Rubén	1925	M	college	Valencia	Río Arriba	Retired financial administrator	2.17
Sandra	1943	F	college	Española	Española	Retired	2.08
Diana	1941	F	high	Taos	Taos	Dry cleaner	2.08
Carlos	1993	M	high	Río Arriba	Española	Auctioneer	2.08
Neddy	1968	M	college	San Miguel	Mora	Car salesman	2.08
Mariana	1944	F	high	Taos	Taos	Mom/volunteer	2
Dora	1953	F	unknown	Río Arriba	Río Arriba	Housewife	2
Clara	1985	F	college	Río Arriba	Española	Editor for Univision	2
Bartolomé	1928	M	middle	Albuquerque	Sthn Colorado	Retired fire-fighter	1.92
Molly	1939	F	middle	Taos	Taos	Retired school cook	1.92
Mónica	1941	F	high	Albuquerque	Taos	Factory worker/ school custodian	1.92
Javier	1936	M	high	Taos	Taos	Rancher and janitor	1.92
Enrique	1933	M	middle	Taos	Taos	Miner/forest service	1.92
Benita	1941	F	high	Sandoval	Rio Arriba	Home maker	1.92
Alfredo	1941	M	high	Sandoval	Sandoval	Retired state highway department	1.92
Victoria	1959	F	college	Río Arriba	Española	Retired schoolteacher/ counsellor	1.83
Eduardo	1935	M	middle	Río Arriba	Río Arriba	Store owner/general contractor	1.83
Miguel	1944	M	middle	Valencia	Valencia	Labourer	1.75
Betty	1925	F	high	Sandoval	Río Arriba	Retired	1.75
Manuel	1954	M	middle	Río Arriba	Rí Arriba	Electrician/rancher	1.75
Marco	1941	M	middle	Taos	Taos	Miner	1.67
Leandro	1931	M	middle	Río Arriba	Taos	Miner	1.67
Norma	1940	F	high	Río Arriba	Río Arriba	Retired bank employee/B&B owner	1.67
Víctor	1928	M	high	Valencia	Río Arriba	Rancher	1.67

Appendix II: NMSEB Questionnaire

Date and time of interview ______________________________

Place ______________________________

Interviewer name ______________________________

Recording duration ______________________________

	Participant 1	**Participant 2**
IRB consent form #		
Name		
Age		
Occupation		
Level of education		
Where was speaker born?		
Where were speakers' parents born?		
Type of home (house, apt., etc.) and ownership		
First language		
Second language		
Preferred language		
Self-rating of English (scale 1–5)		
Self-rating of Spanish (scale 1–5)		
How was English learned?		
Language(s) spoken:		
… with family		
… with friends		
… at work		
Preferred language for:		
… radio		
… other music		
… TV		
… reading newspapers		
… reading books		

Appendix III: Transcription Conventions (Du Bois *et al.* 1993)

Carriage return	new Intonation Unit	@	one syllable of laughter
.	final intonation contour	<@ @ >	speech produced while laughing
,	continuing intonation contour	X	unclear syllable
?	appeal intonation contour	<X X >	unclear speech; transcriber's best guess at content
--	truncated intonation contour		
-	truncated word	%	glottal stop
=	lengthened syllable	(TSK)	click
..	short pause (0.5 secs)	(H)	in breath
. . .	medium pause (0.5–0.7 secs)	(THROAT)	throat clearing
. . .( )	timed pause (over 0.7 secs)	<VOX VOX >	marked speech quality
[ ]	overlapped speech		