The Sociolinguistic Speech Corpus of Chilean Spanish (COSCACH)

A socially stratified text, audio and video corpus with multiple speech styles

Scott Sadowsky
Catholic University of Chile | Max Planck Institute for the Science of Human History

This paper presents the Sociolinguistic Speech Corpus of Chilean Spanish (COSCACH) v1.0, a 9.3-million-word corpus containing transcribed, lemmatized and morphologically tagged text, audio recordings and videos from 1,237 L1 speakers of Chilean Spanish, as well as a control sample of 21 non-Chilean L1 Spanish speakers. The COSCACH is the first freely available corpus of spoken Chilean Spanish of substantial size, as well as one of the largest speech corpora of any variety of Spanish. Following a review of other Chilean speech corpora, I describe how the COSCACH was constructed, covering corpus design, speaker recruitment and metadata collection, speech elicitation and recording, transcription, lemmatization and morphological tagging, and corpus compilation. I thereby aim to provide a blueprint for creating modern, large-scale speech corpora suitable for phonetic, sociophonetic and sociolinguistic research, in addition to traditional inquiry into semantics, lexis, grammar, pragmatics and discourse.

Keywords: speech corpora, Chilean Spanish, corpus design and construction, phonetics, sociolinguistics

1. Introduction

The Sociolinguistic Speech Corpus of Chilean Spanish is a spoken corpus consisting of transcribed text, audio and video from 1,237 L1 speakers of Chilean Spanish, as well as a control sample of 21 non-Chilean L1 Spanish speakers. Known as the COSCACH, its acronym in Spanish (from Corpus Oral Sociolingüístico del Castellano de Chile), it contains a total of 9,288,301 tokens, 68,705 types and

https://doi.org/10.1075/ijcl.19103.sad | Published online: 31 January 2022
International Journal of Corpus Linguistics 27:1 (2022), pp. 93–125. ISSN 1384-6655 | E-ISSN 1569-9811

1,061,711 utterances[1] derived from 83,002 minutes of audio recordings. It is the first publicly available Chilean Spanish speech corpus of substantial size, as well as one of the largest spoken corpora of any variety of Spanish.

The COSCACH's sociolinguistic orientation is embodied by the broad selection of social variables which are at its core. Speaker samples are structured around six such variables: locality, socioeconomic status, sex, age/generation, ethnicity and lingualism (status as a monolingual or bilingual speaker). It also includes five social variables derived from speaker locality, which are not used in speaker selection: urbanness, locality size, region, distance from Santiago (the country's capital) and travel time from Santiago. The non-Chilean speakers in the control sample are categorized using three social variables: country of origin, generation/age and sex. All speakers participated in diverse elicitation tasks that produced a broad range of speech types, from highly controlled to spontaneous, in order to make possible the systematic study of style in the Labovian sense of differing levels of attention paid to speech (Labov, 2006).

The COSCACH was designed from the ground up for phonetic research, in addition to more traditional types of corpus studies. Audio recordings were made with professional-grade microphones and recorders, the highest practical quantization frequency and bitrate, and a lossless file format. These characteristics make the corpus suitable for even the most fine-grained phonetic analysis, unlike most general-purpose speech corpora. Furthermore, the array of speech styles elicited means that both traditional phonetic inquiry (which tends to favor the use of controlled reading passages) and sociophonetic investigation (which prefers naturalistic speech, with or without reading tasks as a point of comparison) are accommodated. The corpus can be accessed at https://corpora.pro.

This paper describes the design and creation of the COSCACH v1.0, with the dual objectives of documenting the corpus itself and providing a blueprint for building large-scale speech corpora suitable for phonetic and sociolinguistic research. Section 2 reviews other spoken corpora of Chilean Spanish and shows why the COSCACH is necessary. Section 3 covers the corpus's design, detailing the speaker samples it contains and the social, geographic and demographic variables that were used to structure them. Section 4 describes data and metadata collection procedures, including speaker recruitment, sociodemographic information gathering, speech elicitation, and recording techniques and technologies. Section 5 details the process of turning the recordings into a corpus, which involves transcription, redaction of personal information, text extraction, lemmatization, morphological tagging, and the creation of a queryable database.

1. In developing the corpus, we defined 'utterance' operationally as speech produced between two pauses of at least 250 ms each.

Section 6 explains how to access the COSCACH and which of its components are publicly available. Finally, Section 7 presents my conclusions and thoughts on future directions for the COSCACH.

2. Other Chilean Spanish speech corpora

In this section I examine existing Chilean Spanish speech corpora, focusing on their size, coverage and availability. I then show why the creation of the COSCACH was necessary.

2.1 ESECH and PRESEEA-SA

The Estudio Sociolingüístico del Español de Chile corpus (ESECH) (San Martín & Guerrero, 2015) and the Proyecto para el Estudio Sociolingüístico del Español de España y América: Santiago corpus (PRESEEA-SA) (San Martín et al., 2016) are complementary speech corpora with a common origin. Both were developed between 2005 and 2012 using the same elicitation tasks and interview script. In both, recordings were made by undergraduate students as a class assignment. The main difference between them is that the ESECH stratifies speakers using a bona fide socioeconomic classification with four strata, while the PRESEEA-SA uses three educational levels as a proxy for socioeconomic status (SES).

The ESECH contains 2,004,853 words from 192 speakers, while the PRESEEA-SA contains 1,118,352 words from 108 speakers, all of whom hail from Santiago. However, 35 speakers are shared between them, so together they total 265 unique speakers and approximately 2.76 million words (A. San Martín, personal communication, June 21, 2019). Public access is limited to 18 PRESEEA-SA interviews. The remainder exist as XML and MP3 files and are not publicly accessible, although access might be granted to qualified academics upon request and after signing an agreement (A. San Martín, personal communication, July 28, 2019).

These corpora are similar to the COSCACH in many regards. They differ significantly in four aspects. The first is size: they contain approximately 30% of the tokens and 21% of the speakers that the COSCACH does. Second is their geographical coverage, which is limited to Santiago. Third is the type of research they permit. The use of consumer voice recorders and mobile phones, plus the lossy MP3 format, make them unsuitable for phonetic research. Likewise, the use of educational level instead of SES by PRESEEA-SA limits its usefulness in sociolinguistic research. Fourth, they lack lemmatization and morphological tagging, as

well as a queryable interface (with the exception of the 18 recordings mentioned above).

2.2 King-ASR-290

The King-ASR-290 corpus (Li et al., 2015) contains recordings of 300 Chilean Spanish speakers reading a single set of written prompts. Speakers are categorized by sex, age group and region. Recordings were made using 13 different types of mobile phone. This corpus is designed for training automatic speech recognition systems. The fact that it contains only a reading-based elicitation task means that it cannot be used for traditional corpus linguistic research. For this same reason, and also because it includes scant information about the speakers, it has very limited usefulness in sociolinguistic inquiry. In any case, these considerations will be moot for most scholars, as access to the corpus costs USD 36,000.00 (SpeechOcean, personal communication, December 2, 2021).

2.3 Additional speech corpora

The first Chilean Spanish speech corpus was developed as part of the Estudio del Habla Culta de Santiago de Chile project (Rabanales, 1995). Its 100 hours of recordings were made between 1970 and 1972, and over the next two decades orthographic transcriptions of 40 hours of interviews involving 89 Santiago speakers were published in two volumes (Rabanales & Contreras, 1979, 1990). Although the authors speak of "educated standard" Spanish, the selection criteria they used required speakers to have a university degree, know at least one foreign language and have traveled abroad (Fernández de Molina Ortés, 2017). In the Chile of the early 1970s, this essentially limited the sampling frame to the local oligarchy, making it somewhat unrepresentative. The fact that the corpus exists only in printed form means that it cannot be used in modern corpus studies.

The Corpus Oral de Lenguaje Adolescente – Santiago de Chile corpus (COLAs) (Jørgensen, n.d.) is a collection of conversations between adolescents from Santiago. It can be used by academics after presenting a research plan and obtaining authorization from its administrators. This provides access to 35 HTML files containing what I calculate to be 64,117 words of transcribed speech (after metadata, links and tags are removed), plus the associated recordings.

The GRUPES corpus (Fant & Harvey, 2008; Gille, 2015) consists of face-to-face group interactions between students at a Santiago university which were recorded on video and then transcribed orthographically. The 25 hours of recordings yielded some 150,000 words of text. It is not publicly available.

Finally, the Spanish Royal Academy has developed two publicly accessible corpora with spoken components: Corpus de Referencia del Español Actual (CREA) (Real Academia Española, n.d.-a) and Corpus del Español del Siglo XXI (CORPES XXI) (Real Academia Española, n.d.-b). They contain 611,084 and 36,415 words of Chilean speech, respectively (G. Rojo Sánchez, personal communication, September 8, 2019).

2.4 Justification for the COSCACH

As we have seen, the COSCACH is not the first speech corpus of Chilean Spanish. It is, however, the only one to fulfill a number of criteria. The following characteristics make the corpus a unique and necessary linguistic resource which opens up possibilities for innovative research:

i. Significant size by modern standards
ii. Designed with phonetic research in mind
iii. Representative of the country's geographic and socioeconomic diversity
iv. Takes into account speaker ethnicity
v. Contains both monolingual speakers and speakers who are bilingual in Spanish and an indigenous language
vi. Lemmatized and morphologically tagged
vii. A queryable corpus rather than a set of loose files
viii. Freely available to researchers

3. Corpus design and speaker sampling

The majority of the COSCACH consists of Chilean Spanish speakers, as described in Section 3.1. However, a small control sample of non-Chilean Spanish speakers was also collected. This component is detailed in Section 3.2.

3.1 Chilean speaker samples

The first priority in designing the COSCACH was to collect large, diverse samples of L1 Chilean Spanish speech in order to achieve the greatest possible representativity, flexibility in research design, and generalizability of results. The second was to facilitate phonetic, phonological and sociolinguistic research *in addition to* the traditional lexical, semantic, grammatical, pragmatic and discourse studies that make up the bulk of corpus linguistic inquiry. This was made possible by four design decisions.

First, rich sociodemographic metadata was collected from all speakers (see Section 3.1.1), and additional variables were derived from speaker locality (see Section 3.1.2). This metadata makes it possible to study potential correlations between linguistic phenomena and speakers' social characteristics, which is the bread and butter of variationist sociolinguistics, while the locality-based variables allow theories of language change to be tested. Second, multiple elicitation tasks were used in order to obtain a broad range of speech styles, making research on phonetic and phonological stylistic variation possible (see Section 4.5). Third, most speakers participated in a 16-question language attitudes interview (see Section 4.5.6). The information it provides can be used to potentially better understand sociolinguistic phenomena, or it can be studied as speech in and of itself. And fourth, audio recordings were made with professional-grade equipment and techniques and are suitable for fine phonetic analyses (see Section 4.6.1).

Table 1 provides an overview of the structure and contents of the COSCACH's speaker samples. The rest of this section describes these samples in detail.

3.1.1 *Speaker inclusion variables*

The COSCACH's Chilean samples are designed around speakers' locality, ethnicity, lingualism, sex and socioeconomic status. Speaker generation was used in a limited fashion: complete fixed-quota samples were required for generation 2 (16–24 years) in all localities, and also for generations 3+4 (25–49 years) and 5+6 (50+ years) in Arica. Otherwise, speaker generation was not a selection criterion. Year of recording was used only in the case of Concepción, where samples were taken in 2009 and 2017.

Cells of six speakers (five in Arica) were established for each combination of sex and socioeconomic status that was present in the target population in sufficient numbers (see 3.1.1.F for an explanation of the socioeconomic status classification system). Whenever possible, all six SESs were included, which produced balanced samples of 72 speakers each (6 speakers per cell × 2 sexes × 6 SESs). In the case of Mapuche and rural Hispano-Chilean samples, however, we could find no speakers who belonged to the two highest SESs, and virtually none who belonged to the third highest. We therefore included only the lowest three SESs (Cb, D and E) in these cases, which produced balanced samples of 36 speakers (6 speakers per cell × 2 sexes × 3 SESs).

Note that the fixed-quota samples were a *minimum* requirement. If fieldworkers completed all quotas in a given locality earlier than planned, they proceeded to record additional speakers opportunistically until it was time to move on to the next site.

Table 1. COSCACH speaker samples, ordered by locality from north to south*

Locality	Ethnicity	Lingualism	Gen.	Dates	Speakers	Words	MWPS
Arica	HC	Monolingual	2	2016–17	61	454,564	7,452
	HC	Monolingual	3–4	2016–17	65	507,732	7,811
	HC	Monolingual	5–6	2017	59	456,854	7,743
Antofagasta	HC	Monolingual	2	2017	74	552,078	7,461
La Serena	HC	Monolingual	2	2016	71	488,872	6,886
Santiago	HC	Monolingual	2	2016–17	89	703,147	7,901
	HC	Monolingual	5–6	2016–17	5	45,912	9,182
	Mapuche	Monolingual	2	2016–17	46	334,709	7,276
Curicó	HC	Monolingual	2	2016	72	524,312	7,282
Concepción	HC	Monolingual	2	2009	101	624,799	6,186
	HC	Monolingual	2	2017	72	576,159	8,002
Tirúa	HC	Monolingual	2	2016	36	231,297	6,425
	Mapuche	Monolingual	2	2016	39	230,724	5,916
Temuco	HC	Monolingual	2	2013, 2016	88	742,707	8,440
	HC	Monolingual	3–4	2013, 2016	3	25,666	8,555
	Mapuche	Monolingual	2	2013, 2016	41	332,309	8,105
	Mapuche	Monolingual	4–5	2013, 2016	4	46,180	11,545
Melipeuco	HC	Monolingual	2	2016–17	36	225,399	6,261
	Mapuche	Monolingual	2	2016–17	42	263,007	6,262
Wallmapu	Mapuche	Bilingual	3–6	2016–17	39	325,375	8,343
Valdivia	HC	Monolingual	2	2017–18	76	531,372	6,992
Chiloé	HC	Monolingual	2	2017–18	80	576,176	7,202
	Mapuche	Monolingual	2	2018	38	251,483	6,618
Non-Chileans	N/A	Monolingual	2–5	2017–18	21	237,468	11,308
Total					**1,258**	**9,288,301**	**7,383**

* 'MWPS' is mean words per speaker; 'HC' is Hispano-Chilean; 'Gen.' is speaker generation; 'Dates' refers to the year(s) in which recordings were made.

A. *Locality*

Speaker samples were collected in a wide range of localities that cover most of Chile's populated territory, as illustrated in Figure 1. Nine of the COSCACH's sampling localities were chosen to maximize geographic coverage. The sparsely

populated north is represented by Arica, Antofagasta and La Serena. The next locality is the nation's capital and only megacity, Santiago, located at the approximate center of Chile's populated territory. South of Santiago, five roughly equidistant locations were selected: Curicó, Concepción, Temuco, Valdivia and the island of Chiloé. Sampling stopped at Chiloé because the territories further south, while geographically vast, contain only 1.53% of the country's population.

Figure 1. Map of speaker sample localities. The Wallmapu territory is centered on Temuco

Two additional sampling localities were chosen to ensure adequate inclusion of Spanish-monolingual speakers of Mapuche ethnicity: Tirúa, which is on the Pacific coast, and Melipeuco, which is nestled in the Andes. Together with Temuco, which lies between them, they also provide coverage of three different Mapudungun geolectal regions, which could potentially aid in researching language contact.

The final locality is not a city or a village, but a large area that corresponds to the core Mapuche territory. Known unofficially as Wallmapu, it is centered on Temuco and extends outward from it for several hundred kilometers. The sample of Mapudungun-Spanish bilingual speakers was taken in 34 different rural locations within this area. The final make-up of the COSCACH's speaker sample is detailed in Table 2, while the names of each sampling locality and the municipalities they contain are provided in Table 3.

Table 2. Chilean speaker samples by locality

Locality	Samples	Speakers	Words
Arica	3	185	1,419,150
Antofagasta	1	74	552,078
La Serena	1	71	488,872
Santiago	3	140	1,083,768
Curicó	1	72	524,312
Concepción	2	173	1,200,958
Tirúa	2	75	462,021
Temuco	4	136	1,146,862
Melipeuco	2	78	488,406
Wallmapu	1	39	325,375
Valdivia	1	76	531,372
Chiloé	2	118	827,659
Total	**23**	**1,237**	**9,050,833**

In order to be included in the sample for a given locality, speakers were required to have lived there after their fifth birthday for all but one year of their lives per generation (starting with generation 2). Thus, members of generation 2 (16–24 years of age) were permitted to have lived outside their locality for up to one year after they turned five; members of generation 3 (25–34 years of age) for up to two years; and so on. This criterion was applied to ensure that the speakers from each locality had experienced as little linguistic influence as possible from other localities.

Table 3. Speaker sample localities and the municipalities each includes

Locality	Municipalities
Arica	Arica
Antofagasta	Antofagasta
La Serena	La Serena, Coquimbo
Santiago	Cerrillos, Cerro Navia, Conchalí, El Bosque, Estación Central, Huechuraba, Independencia, La Cisterna, La Florida, La Granja, La Pintana, La Reina, Las Condes, Lo Barnechea, Lo Espejo, Lo Prado, Macul, Maipú, Ñuñoa, Padre Hurtado, Pedro Aguirre Cerda, Peñalolén, Pirque, Providencia, Pudahuel, Puente Alto, Quilicura, Quinta Normal, Recoleta, Renca, San Bernardo, San Joaquín, San José de Maipo, San Miguel, San Ramón, Santiago, Vitacura
Curicó	Curicó
Concepción	Concepción, Chiguayante, Hualpén, Penco, San Pedro de la Paz, Talcahuano
Tirúa	Tirúa, Cañete
Temuco	Temuco, Padre Las Casas
Melipeuco	Melipeuco, Lonquimay
Wallmapu	N/A
Valdivia	Valdivia
Chiloé	Ancud, Castro, Chonchi, Dalcahue, Quemchi, Quinchao

B. *Ethnicity*

Speaker ethnicity was included as a variable in response to a small but growing body of research (B. Rogers, 2016; Sadowsky, 2020; Sadowsky & Aninao, 2019) which suggests that certain features of Chilean Spanish may have arisen from contact with Mapudungun, the language of the Mapuches – Chile's largest indigenous group. Samples of Mapuche speakers of Chilean Spanish were collected in six localities: Santiago, Tirúa, Temuco, Melipeuco, Chiloé and Wallmapu (see Table 4). The speakers in the first five samples are all monolingual in Spanish, while those in the Wallmapu sample are Mapudungun-Spanish bilinguals (Section 3.1.1.C covers the bilingual sample in depth).

Table 4. Chilean speaker samples by ethnicity

Ethnicity	Speakers	Words
Hispano-Chilean	988	7,267,046
Mapuche	249	1,783,787
Total	**1,237**	**9,050,833**

In order to be included in one of the Mapuche samples, speakers were required to fulfill two criteria: (i) they had to self-identify as Mapuche, and (ii) at least one of their parents' combined four surnames (one paternal and one maternal each, as per the Hispanic custom) had to be Mapuche. The first criterion served to create an initial speaker pool, while the second sought to remove false positives from it (e.g. non-Mapuche speakers claiming to be Mapuche out of solidarity or affinity). These samples, consisting of a minimum of 36 speakers, include only the lowest three SESs, which reflects the demographic reality of Chile's Mapuche population.

The Hispano-Chilean samples consist of speakers who did not self-identify as Mapuche and whose parents had no Mapuche surnames. Note that virtually all Chileans are of mixed European and Amerindian ancestry, with a mean indigenous DNA admixture of 44.7% (Eyheramendy et al., 2015). Thus, the variable 'speaker ethnicity' should be interpreted as representing two clusters on a bi-ethnic continuum rather than two discreet population groups.

C. *Lingualism*

The lingualism variable encodes whether speakers are monolingual in Spanish or bilingual in Spanish and Mapudungun. There is one sample of bilinguals in the corpus, made up of 39 speakers from Wallmapu (see Table 5). Its purpose is to further facilitate research on the possible Mapuche influence on Chilean Spanish.

Table 5. Chilean speaker samples by lingualism

Lingualism	Speakers	Words
Spanish monolingual	1,198	8,725,458
Spanish-Mapudungun bilingual	39	325,375
Total	**1,237**	**9,050,833**

The collection of the bilingual sample was a painstaking process that required over five months of full-time fieldwork. The difficulties stemmed from three main factors: the paucity of Mapudungun speakers – Zúñiga (2007) estimates their number at 144,000, but our experience both here and in the Sound Comparisons project (Heggarty et al., 2019) indicates that the actual number is significantly lower; the difficulty of accessing the scattered and remote locations where most speakers live; and the fact that less than 10% of self-identified speakers agreed to participate in a brief language competency screening, which involved describing a number of images in Mapudungun.

To make the collection of the bilingual sample viable, we applied no age restrictions. Of the 39 speakers who participated, only three (7.6%) were under 35. Their mean age was 52.5 years, and the oldest speaker was 84. This pattern

reflects the fact that Mapudungun is barely spoken by the younger generations (Gundermann et al., 2009). For this same reason, no SES quotas were used (see 3.1.1.F for information on the SES classification system). In the resulting sample, 91.9% of speakers belonged to the D or E SESs. All speakers acquired Mapudungun at home before learning Spanish.

D. *Age/Generation*

The COSCACH groups speakers into five generations, or age cohorts, as shown in Table 6. The youngest, generation 2, consists of 16- to 24-year-olds, who were the main focus of fieldwork (generation 1 (13–15 years) was not included in the corpus). This was due to the fact that the speakers of generation 2 are somewhat easier to access than those of other age groups, since they can typically be found in large numbers in secondary or tertiary educational institutions. Other generations were sampled opportunistically except for Arica, where two 60-person samples of older speakers (generations 3+4 and 5+6) were collected.

Table 6. Chilean speaker samples by generation

Generation	Age range (years)	Speakers	Words
2	16–24	1,063	7,651,062
3	25–34	38	309,099
4	35–49	51	413,201
5	50–64	72	563,544
6	65+	13	113,927
Total		**1,237**	**9,050,833**

This strategy made it possible to obtain speaker samples for generation 2 that met or exceeded all locality, ethnicity, lingualism, sex and SES quotas, in spite of the significant financial and time limitations the project faced. The disadvantage of this strategy is that older speakers are underrepresented in the corpus.

E. *Sex*

All samples were designed with equal numbers of female and male speakers, and these quotas were met in virtually every case (see Table 7).

F. *Socioeconomic status*

Chilean speakers were socially stratified with the EMIS system, which uses educational level, occupation and a special matrix to determine SES (Sadowsky, 2021). The socioeconomic structure of the COSCACH is presented in Table 8.

One of the most noteworthy aspects of the COSCACH is that it includes complete samples of the full spectrum of socioeconomic strata. The extreme upper

Table 7. Structure of Chilean speaker samples by sex

Sex	Speakers	Words
Female	621	4,597,577
Male	616	4,453,256
Total	**1,237**	**9,050,833**

Table 8. Chilean speaker samples by socioeconomic status (SES)

SES	Descriptor	Speakers	Words
A	Extreme upper	141	1,124,374
B	Upper	150	1,155,882
Ca	Upper-middle	182	1,325,045
Cb	Lower-middle	243	1,812,521
D	Lower	286	2,028,387
E	Extreme lower	235	1,604,624
Total		**1,237**	**9,050,833**

(A) and upper (B) SESs, which are the country's economic elite, place great stock in their privacy, and are reticent to interact with outsiders, making them virtually impenetrable. We were able to access these groups by going through the exclusive private schools their children attend and then, when possible, using the snowball or network technique (Milroy, 1987) to move upward in the age hierarchy. Even so, this strategy required dogged persistence on the part of fieldworkers, as up to 95% of the private schools in each locality refused to authorize any sort of access whatsoever. In spite of these difficulties, we managed to recruit 291 speakers from these two SESs.

At the other end of the socioeconomic hierarchy are the lower (D) and extreme lower (E) SESs, whose members have virtually never been included in studies of Chilean Spanish. The reasons for this appear to be twofold. First, Chilean society is plagued by classism, profound inequality and rampant socioeconomic discrimination (Bengoa, 2018; Garretón & Cumsille, 2002; Ruiz-Tagle, 2016), which have engendered intensely negative attitudes toward the speech of the lower socioeconomic strata. Such attitudes, combined with the intense prescriptivism that is still rampant, seem to have led local scholars to shun these groups. Second, accessing speakers who belong to these strata can be a difficult or even dangerous undertaking, as they tend to live, work and/or go to school in neighborhoods with high crime rates. In light of these challenges, the fact that

we were able to include 286 D speakers and 235 E speakers is significant. There was no special strategy or technique involved in collecting these speaker samples beyond fieldworkers' extreme tenacity.

G. *Year of recording*
Most speaker samples were recorded between 2016 and 2018. In the case of Concepción, however, two samples were taken, one in 2009 and one in 2017. While eight years is not a particularly long period in the evolution of a language, it is possible that real-time differences will become apparent when comparing these two samples.

3.1.2 *Derived variables*

The COSCACH includes five additional variables which were not used to structure speaker samples. All are derived from locality and incorporate additional geographic or demographic information. One of the main reasons for including them is to allow testing different models of language variation and change, such as the gravity (Trudgill, 1974) or cascade (Labov, 2001) model, the wave model, the tree model and others (Heggarty et al., 2010).

A. *Region*
The 'region' variable divides Chile into five geographic zones from north to south (see Table 9). It reflects a series of non-linguistic phenomena which could shed light on the development of the Spanish spoken in the country. These include the direction of Spanish colonization, internal migration patterns, the location of indigenous populations current and past, and contact with neighboring countries.

B. *Urbanness*
The 'urbanness' variable indexes the percentage of each locality that is classified as urban by Chile's 2017 census (Instituto Nacional de Estadísticas, 2018). These values are the averages of the municipalities that make up each locality. Wallmapu is a special case, as it is not a coherent demographic entity, but an ad hoc macro-region for which Census data does not exist. As all speakers included in this sample lived in isolated rural areas, they were assigned an urbanness level of 0%. In the corpus, urbanness has been grouped into six categories[2] (see Table 10).

2. These categories are U1 (0–24%), U2 (24–49%), U3 (50–74%), U4 (75–89%), U5 (90–94%) and U6 (95–100%).

Table 9. Speaker regions and the localities each contains

Region	Code	Locality
Far North	R1	Arica
		Antofagasta
North	R2	La Serena
Center	R3	Santiago
		Curicó
South	R4	Concepción
		Tirúa
		Temuco
		Melipeuco
		Wallmapu
Far South	R5	Valdivia
		Chiloé

Table 10. Chilean speaker localities by percentage of urbanness

Locality	% Urbanness	Category
Arica	92.6	U5
Antofagasta	97.9	U6
La Serena	92.5	U5
Santiago	99.3	U6
Curicó	88.9	U4
Concepción	99.0	U6
Tirúa	35.9	U2
Temuco	86.1	U4
Melipeuco	40.2	U2
Wallmapu	0.0	U1
Valdivia	93.2	U5
Chiloé	63.3	U3

C. *Locality size*

The 'locality size' variable captures the population of each locality using data from the 2017 census (Instituto Nacional de Estadísticas, 2018). Absolute population

was converted into 10 categories whose cut-off points correspond largely to natural breaks in the raw population numbers (see Table 11).[3]

Table 11. Chilean speaker localities by population

Locality	Population	Category
Arica	221,364	P4
Antofagasta	361,873	P4
La Serena	448,784	P5
Santiago	6,227,944	P9
Curicó	149,136	P3
Concepción	732,209	P6
Tirúa	10,417	P1
Temuco	358,541	P4
Melipeuco	16,389	P1
Wallmapu	N/A	P0
Valdivia	166,080	P3
Chiloé	21,310	P1

In the case of two localities, population was handled in a special fashion. For Chiloé, the *average* population of the six municipalities from which speakers came was used, rather than the sum, as was done in other localities. This was necessary because Chiloé is a large, predominantly rural island whose municipalities do not form a single conurbation, unlike other multi-municipality localities. In the case of Wallmapu, for which no population data exists, category P0 was assigned, reflecting the small, rural villages or settlements from which speakers were recruited.

D. *Distance and travel time from Santiago*

Chile is a hyper-centralized nation, with all forms of power – economic, political, administrative, media, cultural, military, ecclesiastical and so on – concentrated in the capital, Santiago (Sadowsky & Aninao, 2019). Anecdotal observations likewise suggest that linguistic innovations tend to begin in Santiago and radiate outward. In order to facilitate the study of such phenomena, the COSCACH includes two measures of distance from Santiago (see Table 12).

3. The population categories are P0 (Under 5,000), P1 (5,000–24,999), P2 (25,000–99,999), P3 (100,000–199,999), P4 (200,000–399,999), P5 (400,000–599,999), P6 (600,000–799,999), P7 (800,000–999,999), P8 (1,000,000–1,499,999) and P9 (1,500,000 and over).

The first is driving distance, measured in kilometers. These values were determined by using Google Maps' "directions" function, which provides a more accurate estimation of the distance speakers would actually traverse when traveling between localities than linear measurements would. The second is travel time using the fastest bus available, rounded to the nearest quarter hour. This was also calculated using Google Maps. Bus travel was chosen because it is the most common mode of transportation for most of the population.

Table 12. Distance and travel time from each locality to Santiago

Locality	Distance (km)	Distance category	Time (hours)	Time category
Arica	2,036	D7	24.00	T8
Antofagasta	1,337	D5	19.75	T7
La Serena	472	D2	7.00	T3
Santiago	0	D0	0.00	T0
Curicó	196	D1	3.25	T1
Concepción	500	D2	7.25	T3
Tirúa	704	D3	10.50	T5
Temuco	679	D3	9.75	T4
Melipeuco	771	D3	11.25	T5
Wallmapu	825	D4	10.50	T5
Valdivia	849	D4	12.25	T6
Chiloé	1,200	D5	18.25	T7

The values for Wallmapu were calculated by using the approximate center of the sampling area and then adding 100 km to the distance and 1.5 hours to the time, to account for the additional travel required to reach speakers in scattered and typically hard-to-reach locations. The values for Chiloé are based on the distance to Castro, the municipality in which most speakers lived. Absolute distance and time values were then converted into categories.[4]

4. Distance categories correspond to the following ranges of kilometers: D0 (0–99), D1 (100–299), D2 (300–599), D3 (600–799) D4 (800–1,099), D5 (1,100–1,399), D6 (1,400–1,799) and D7 (1,800 and over). Time categories correspond to the following range of hours, rounded to the nearest quarter hour: T0 (0.00–1.75), T1 (2.00–3.75), T2 (4.00–5.75), T3 (6.00–7.75), T4 (8.00–9.75), T5 (10.00–11.75), T6 (12.00–15.75), T7 (16.00–19.75) and T8 (20.00 and over).

3.2 Non-Chilean control sample

A sample of non-Chilean L1 Spanish speakers was collected for the sole purpose of determining the degree to which various linguistic phenomena are exclusive to Chile. This is best achieved by applying identical elicitation tasks and recording techniques to suitable non-Chilean speakers, rather than by relying on other similarly organized corpora, which are few in number, methodologically heterogeneous and often inaccessible. The relatively small size of this sample means that it is only suitable for use as a control group, and not for research into non-Chilean Spanish speech in general.

The speakers in the control sample (see Table 13) are categorized by country of origin (which is treated as their locality), sex and age/generation. As SES is highly culture-bound, it was not possible to calculate it for non-Chilean speakers using the EMIS system. Similarly, ethnicity and lingualism are not included for these speakers because in the COSCACH these concepts are defined in terms of the presence or absence of Mapuche ethnicity and the Mapudungun language, respectively, and these are not present outside of Chile (with the exception of certain areas of Argentina). Finally, the control sample does not contain the derived variables detailed in 3.1.2, as these are also specific to Chile.

Table 13. Structure of the non-Chilean speaker sample

Country	Speakers	Words
Argentina	2	10,316
Bolivia	1	9,354
Colombia	3	34,465
Cuba	4	51,341
Mexico	5	50,601
Paraguay	1	14,974
Peru	1	12,312
Venezuela	4	54,105
Total	**21**	**237,468**

All speakers in this sample were recorded in Santiago. In addition to meeting the criteria set forth in Section 4.3.3, they were required to have resided in Chile for less than one year in order to minimize the effects of accommodation and adaptation to local speech.

4. Data collection

This section details the COSCACH's data collection process. After specifying the timeframe in which samples were collected, I explain how speaker recruitment was performed and how socio-demographic metadata was obtained. I then describe the various elicitation tasks that were used to obtain different types of speech, followed by the recording equipment and techniques that were employed.

4.1 Timeframe

The COSCACH's speaker samples were recorded between 2009 and 2018. The first of the two Concepción samples was recorded in 2009, and served as a pilot study for fine tuning the methods used in creating the remainder of the corpus. The Temuco samples were recorded in 2013 and 2016. All other recordings were made between 2016 and 2018, as detailed in Table 1.

4.2 Fieldworkers

Fieldworkers were responsible for finding speakers suitable for inclusion in the COSCACH, obtaining their informed consent and sociodemographic information, performing and recording elicitation sessions, and obtaining authorization when recruitment was performed in an institutional setting, such as a school or place of work. Fieldworker candidates underwent approximately 15 hours of initial training, after which they carried out a series of test elicitation sessions which were evaluated and used to improve their technique. Those judged to be capable of performing the work to a high standard were then sent into the field, paired with a more experienced worker whenever possible.

A total of 25 fieldworkers participated in the creation of the corpus, with just six of them accounting for 85.4% of all recordings, and an additional six accounting for a further 12.9%. The small number of workers involved minimized variation in the elicitation process, which was particularly important to ensure the consistency of the conversational and language attitudes interviews.

4.3 Speaker recruitment

This section details the process used to recruit speakers for the COSCACH. The techniques employed by fieldworkers to find speakers are first presented, followed by ethics considerations and then speaker exclusion criteria.

4.3.1 *Recruitment procedures*

Upon arriving at a locality, fieldworkers initiated speaker recruitment by searching out institutions such as secondary schools, universities, adult education centers, neighborhood councils and clubs. This has two advantages: such places offer efficient access to large numbers of speakers, and they can often provide a physical space in which recordings can be made. Additional speakers were found by going door to door in residential areas, by approaching people in high-traffic public areas, and by applying the snowball technique to previously-recruited speakers. Potential participants who provided informed consent (Section 4.3.2) then filled out a sociodemographic questionnaire (Section 4.4) with the information necessary to determine whether they met the COSCACH's inclusion criteria (Section 3.1.1), did not present any characteristics which would lead to their exclusion (Section 4.3.3), and belonged to a cell which required additional speakers.

4.3.2 *Informed consent and institutional review board (IRB) approval*

In order to ensure that all speakers participated voluntarily, understood the nature and purpose of the activities, and were aware of their rights, adult speakers were asked to sign an informed consent form. Minors were requested to sign an informed assent form, and their parents or guardians an informed consent form on their behalf. Speakers who did not wish to do so were excluded.

The Concepción 2009 and Temuco samples were recorded when the author was affiliated with the Universidad de Concepción and the Universidad de La Frontera, respectively. These samples predate the establishment of IRBs at the universities in question, and are therefore exempt from such requirements. The research project under which the remainder of the COSCACH was created was authorized by the Catholic University of Chile's IRB.

4.3.3 *Further criteria for exclusion of speakers*

A set of criteria for excluding atypical speakers was established in order to ensure the representativity of the corpus. Accordingly, even if speakers met all the requirements set forth in Sections 3.1 and 3.2, they were nonetheless excluded if they reported having been diagnosed with any sort of speech, language or hearing impairment or disorder at any point in their lives; if they wore daytime retainers or other obstructive dental appliances which could interfere with articulation (braces without elastic bands were allowed); if they had one or more tongue or lip piercings, as these could also interfere with articulation; if they were native speakers of any language except Spanish (with the exception of Mapudungun in the Spanish-Mapudungun bilingual speaker sample), given that bilingualism has the potential to produce speech patterns not found in the overwhelming major-

ity of the population, which is monolingual; if one or both of their parents were not Chilean, in order to avoid potential linguistic influences from other varieties of Spanish and other languages; if they worked in or studied language-related fields such as linguistics, speech and language pathology, journalism or language teaching, as the linguistic prescriptivism these fields often inculcate favors artificial speech; or if they demonstrated a notorious degree of status incongruence (Labov, 2001: 114).

4.4 Socio-demographic questionnaire

The COSCACH includes a rich variety of socio-demographic metadata for each speaker, as detailed in Sections 3.1, 3.2 and 4.3.3. As the accuracy and completeness of this information is of the utmost importance, the written questionnaire used to obtain it was applied to each speaker by a fieldworker who went through the items one by one, explained exactly what information was desired, and answered any questions speakers had. In order to increase the accuracy of this metadata, fieldworkers reviewed each speaker's questionnaire before their elicitation session and made a mental note of any answers that were unclear or incomplete. Then, during the interview, they worked in questions that would clarify these responses, and subsequently corrected or amended the questionnaires as needed.

4.5 Elicitation tasks

The COSCACH makes use of a series of different elicitation tasks in order to obtain a wide variety of speech styles. These range from openly artificial tasks which seek to produce highly controlled speech, to maximally naturalistic ones which aim to elicit the most spontaneous speech possible (see Table 14).

In our experience, speakers are invariably at their most uneasy and tense at the beginning of an elicitation session, as it is an unusual activity carried out in the presence of a stranger. This emotional state favors self-monitoring, hypercorrection, excessive formality and other linguistically unrepresentative behavior. We therefore begin each session with the activity which seeks the most controlled and self-conscious speech, the sustained pronunciation of vowels, as it is highly compatible with this emotional state. Subsequently, tasks of increasing naturalness and decreasing focus on speech production are performed. By the time speakers reach the conversational interview, they are more accustomed to the situation, the recording equipment and the fieldworker, and thus speak more naturally than at the beginning of the session – typically much more so.

The sole exception to this ordering by increasing naturalness is the language attitudes interview. As a question-driven activity it is *less* natural than the conver-

Table 14. Elicitation tasks, in the order in which they were performed

Elicitation task	Type	Chileans		Non-Chileans		Total	
		Speakers	Words	Speakers	Words	Speakers	Words
Vowels	Reading	1,008	5,039	20	100	1,028	5,139
Minimal pairs	Reading	810	53,769	20	1,305	830	55,074
Other word lists	Reading	101	161,607	0	0	101	161,607
Meaningful sentences	Reading	1,236	1,021,857	20	16,977	1,256	1,038,834
Meaningful texts	Reading	1,200	1,947,756	20	33,027	1,220	1,980,783
Conversational interview	Interview	1,211	5,209,462	21	146,213	1,232	5,355,675
Language attitudes interview	Interview	930	651,343	19	39,846	949	691,189
Total			**9,050,833**		**237,468**		**9,288,301**

sational interview, which is completely freeform, and so would be expected to precede it. However, the topics of the language attitudes interview focus participants' attention on how they and others speak, which favors self-monitoring and other linguistically artificial behavior. To prevent this from negatively affecting the naturalness of the conversational interview, the language attitudes interview was done last.

As Table 14 makes clear, the COSCACH includes not just interviews and conversations, as are traditional in oral corpora, but also speech elicited by means of reading-based tasks. Some 34.9% of the corpus's running text (3,241,437 words) comes from such tasks, while 65.1% (6,046,864 words) comes from interviews. This design decision facilitates a broad range of sociophonetic and sociophonological research involving speech style, as well as more traditional phonetic and phonological research based on highly controlled (read) speech. The downside, of course, is that approximately one third of the COSCACH is not suited for traditional corpus studies.

In the following subsections, each individual elicitation task is described in detail. While most were applied to the majority of speakers, there are exceptions – some due to time constraints during fieldwork, others due to minor changes in corpus design.

4.5.1 *Sustained pronunciation of isolated vowels*

In this task, speakers pronounce each of the five vowels of Spanish in a sustained fashion for up to 10 seconds. This is designed for the study of voice quality (jitter, shimmer, harmonics-to-noise ratio, etc.) rather than the linguistic characteristics of the vowels, for which it is too artificial to be useful.

4.5.2 *Reading of minimal pairs or other word lists*

These tasks, which focus on isolated words, seek to elicit highly controlled speech. In the Concepción 2009 sample, word lists consisting of both bare words and words embedded in carrier phrases were read. In most subsequent samples, these tasks were replaced by a list of 64 minimal pairs. In addition to all vowels, the minimal pairs contain the 17 consonant phonemes of Spanish plus /t͡ɾ̪/[5] in all of the following positions (if they can occur in them): word-initial, intervocalic, word-internal coda and word-final coda.

4.5.3 *Reading of meaningful sentences*

This task consists of 50 sentences of meaningful text containing a total of 732 words (the Concepción 2009 version is slightly shorter). These sentences contain all Spanish phonemes in a variety of positions, but they are not balanced in any way. Their purpose is to obtain continuous speech that is moderately controlled.

4.5.4 *Reading of meaningful texts*

This task involves reading eight meaningful texts consisting of a total of 83 sentences or 1,340 words (the Concepción 2009 version contained two fewer texts). Since speakers' attention is focused on the stories being told, they tend to monitor their speech production less, which favors a low-control style. This is further aided by the distraction caused by the light-hearted or humorous nature of several of the texts.

4.5.5 *Conversational interview*

The conversational interview is the longest task – it ranges from 30 to 50 or more minutes. It aims to elicit the most naturalistic and spontaneous speech possible, and every aspect of its design seeks to contribute to achieving this goal. The interview is carried out after speakers have spent 15 to 25 minutes doing reading-based tasks, allowing them to become accustomed to the situation. It immediately follows a humorous and somewhat absurd text which seeks to put them further at

5. In Chilean Spanish, /t͡ɾ̪/ behaves as a phoneme with some eight allophones, rather than as a consonant cluster. See Sadowsky and Salamanca (2011) for more information.

ease. It furthermore eschews the common sociolinguistic practice of using a set of predefined questions or topics, which is a format that does not exist in daily life. Instead, fieldworkers are instructed to imagine that the speaker is someone they have just met at a party and want to get to know. This way of conceptualizing the conversational interview, which mimics a real-life situation in which complete strangers may naturally establish near-instant rapport, has proven extremely successful at producing relaxed, informal and often very personal conversations that are seldom free of laughter.

At the same time, fieldworkers are trained in techniques to obtain better recordings which are not part of normal conversation. These include transmitting more backchannel information than usual via non-verbal cues (e.g. head nodding and eyebrow movements) in order to avoid "talking over" the speaker; occasionally waiting in silence longer than is usual when speakers are producing mostly short utterances, in order to get them to go into greater depth on whatever topic they are speaking about; and always having several "emergency questions" prepared in case a potentially awkward silence occurs.

4.5.6 *Language attitudes interview*

The language attitudes interview consists of 16 questions about speakers' ideas and beliefs about the Spanish spoken in different countries, throughout Chile, and by the speakers themselves. It typically lasts five to ten minutes. Unlike the other tasks, it does not seek to elicit a given speech style, but to obtain specific information that can be investigated in and of itself, and can also potentially aid in interpreting the results of the sociolinguistic studies that are performed with the COSCACH. Nevertheless, it can also be analyzed as speech in its own right.

4.6 Recording

As one of the goals of the COSCACH was to permit exacting phonetic analyses, great care was taken to obtain audio recordings of the highest possible quality. Section 4.6.1 details the equipment and techniques used to meet this goal. Section 4.6.2 recounts the post-processing that was applied to these recordings. Video recordings were also made, and are detailed in Section 4.6.3. In all, the COSCACH contains 83,002 minutes of audio recordings, which is equivalent to 1,383 hours, or 173 eight-hour work days. Because some 20% of the video recordings turned out be unusable for various reasons, such as the speaker moving out of frame excessively, only about 66,000 minutes of video are available.

4.6.1 *Audio equipment and configuration*

The recording quality requirements of the COSCACH ruled out the use of mobile phones as audio recording devices, as even the most expensive models have low-quality microphones with irregular frequency responses. Such devices also tend to apply various digital signal processing techniques which make for more pleasant phone conversations and reduced bandwidth costs for telephony firms, but which corrupt the audio signal in the process. Such techniques include noise cancelation, low- and high-pass filtering, automatic gain control and lossy compression.

After extensive testing, Audix HT5 small-condenser head-worn microphones (Audix Microphones, 2017) were chosen for audio recording.[6] These devices have a range of 20 Hz to 20,000 Hz, an almost perfectly flat frequency response, and a fast response time to transient sounds such as stop bursts. Furthermore, the fact that the HT5's capsule is placed at a constant distance of approximately 2 cm from the speaker's mouth means that recordings have an outstanding signal-to-noise ratio even in sub-optimal environments.

The microphones were fed into Fostex FR2-LE audio recorders. Although it is a relatively old design, and is bigger and heavier than most recorders on the market, it offers two decisive advantages over similar devices. First, it can run off of 7.2 V NiMH remote-control vehicle batteries. A 5,000 mAh battery of this type allows for approximately 20 hours of autonomous operation. And second, the FR2-LE's preamps are extraordinarily quiet, producing a minimum of self-noise, and therefore cleaner recordings.

Audio recordings were made in uncompressed WAV format at 24 bits and 48 kHz, single channel, with peak input levels set between −24 dB and −12 dB. The extra headroom that 24 bits provides makes it possible to record at a substantially lower volume while still maintaining a great distance from the noise floor. This prevents clipping from occurring when speakers' loudness increases after recording levels have been set.

4.6.2 *Audio post-processing*

Audio recordings underwent a minimal amount of post-processing. Using *Audacity* (Audacity Development Team, 2018), extremely loud non-linguistic sounds (coughs, laughs, slamming doors, etc.) were carefully selected and their volume was reduced to the approximate level of the surrounding recording. Each recording's volume was then normalized *as a whole* to −1 dB peak, and DC correction was applied. This brought the overall volume up to a comfortable level for tran-

6. The first speaker sample, Concepción 2009, used a Studio Projects C-1 large-condenser microphone, an M-Audio FastTrack Pro interface, and a laptop computer running Audacity (Audacity Development Team, 2018).

scription and impressionistic analysis, while not altering the relative intensity of the linguistic information.

4.6.3 *Video recording equipment and procedures*

Video recordings of speakers were made with Sony HDR-CX405 cameras at 1920×1080 resolution and a framerate of 29.97 fps.[7] The massive size of the raw MPEG-2 files (about 10 GB per speaker) made it necessary to transcode them to H.264 format, using *HandBrake* (HandBrake Team, 2019). This reduced the size of the COSCACH's video recordings from approximately 12 TB to 800 GB, at only a small cost in quality. The purpose of these videos is to study articulatory phenomena – specifically, labial and dental articulations. For this reason, they focus tightly on speakers' mouths; hand gestures and most body movements are not visible.

5. Transcription, text processing and corpus compilation

Turning the recordings and metadata into an actual corpus involved transcribing the audio, redacting personally identifiable information, extracting the transcriptions, annotating them with additional linguistic information and speaker metadata, and then compiling them into a database, as is detailed below.

5.1 Transcription

Audio recordings were segmented at the utterance level and transcribed orthographically using *Praat* (Boersma & Weenink, 2018). In order to standardize this process as much as possible, transcribers followed a 35-page protocol, available as part of the general COSCACH protocol at https://www.corpora.pro. Among other things, it covers the placement of utterance boundaries; the proper handling of incomplete words and unintelligible speech; a standardized spelling for rarely-written exclamations, onomatopoeias, foreign loanwords and non-lexical backchannels (e.g. *hm*, *eh*, *a-ha*); and the spelling of verbs conjugated in the Chilean *voseo* paradigm, the sometimes low-prestige 2nd person singular form which is ignored by the educational system.[8]

Fieldworkers' speech during the two interviews is only barely audible due to the characteristics of the head-worn microphones used by speakers. At the same

7. The Concepción 2009 sample used a Sony DCR-SR47 camera at 720×480 resolution and 29.97 fps.

8. Such forms include *hablái*, *hablís*, *hablarái*, *hablarai*, *hablaríai* and *hablabai*.

time, it is of little interest for phonetic and phonological research, as it consists of the same individuals talking in dozens or hundreds of interviews. For these reasons, as well as due to severe budget constraints, fieldworkers' speech was not transcribed. The corpus thus consists exclusively of monologic speech. For the same reason, the identification of who is speaking at any given moment in a recording is 100% accurate.

5.2 Anonymization and protection of speakers' privacy

Every effort has been made to protect speakers' anonymity and privacy. After the relevant metadata was extracted from the sociodemographic questionnaires, they were shredded. Speaker names were replaced by anonymous codes in all instances, and personally identifying information was redacted in the transcriptions. In addition, all people who worked on the project signed non-disclosure and confidentiality agreements. Finally, in order to be completely sure that speakers' anonymity and privacy is protected, a second anonymization and redaction process was commissioned using a different team.

5.3 Text extraction and annotation

Once the transcription and redaction processes were complete, text was extracted using a customized version of *MaSCoT-R* (Sadowsky, 2017) and saved in UTF-8 text files. It was then lemmatized and morphologically tagged with *FreeLing's* (Padró & Stanilovsky, 2012) Chilean Spanish version (Sadowsky, 2016), which includes the 10,000 or so headwords of the *Diccionario de uso del español de Chile* (Academia Chilena de la Lengua, 2010), Chilean *voseo* allomorphs for all verbs, nearly 3,000 local toponyms and demonyms, entries for hundreds of commonly misspelled words, and a series of other items characteristic of Chilean Spanish. This version of *FreeLing* uses a slightly modified EAGLES tag set that assigns the value "V" to the last slot of pronoun and verb tags that correspond to Chilean *voseo* forms, as otherwise their codes would be identical to those of the *tuteo* verb forms.

5.4 Corpus compilation and use

In order to turn the collection of tagged texts into a proper corpus, the *IMS Open Corpus Workbench* (Evert & Hardie, 2011) was employed. This software package compiles the raw text, lemmas and morphological tags, as well as all metadata, into a single database that forms the core of the system. *CQPweb* (Hardie, 2012) is used to provide a powerful yet easy-to-use web interface to the COSCACH. It

allows users to perform complex queries, carry out various types of analyses, and visualize results in different ways.

6. Availability and access

The full text of the COSCACH is available at https://corpora.pro. Accessing it requires nothing more than creating an account using a valid academic e-mail address.[9] As funding permits, we plan on redacting personally identifying and sensitive information from the audio recordings so that they may be shared with other researchers. Anonymizing and redacting the videos, on the other hand, is an extremely laborious process that will require many thousands of hours of work. This is unlikely to be possible in the near term, and thus the videos will have to be examined on site for the foreseeable future. The entire 143-page protocol which governed the creation of the COSCACH is available on the same website.

7. Conclusions and future directions

As a large-scale, openly accessible speech corpus that is richly-annotated with linguistic information and speaker metadata, the COSCACH opens up unprecedented opportunities for innovative research on Chilean Spanish. By providing access to large and well-structured samples of Chilean speakers of both sexes and all six standardized SESs, sociolinguistic research of previously unseen depth, breadth and precision will be possible. The impact of the Mapuches and their language on Chilean Spanish can be empirically examined by researchers from all linguistic sub-disciplines, without the need to fund and perform grueling – and potentially unsuccessful – fieldwork. The same holds true for the examination of rural vs. urban speech. The longstanding assertion that Chilean Spanish shows little if any regional variation can finally be put to the test in a meaningful way. At the same time, massive phonetic and phonological studies with exacting standards can now be performed in a fraction of the time previously required. Furthermore, the COSCACH has the potential to contribute to the creation of a dictionary and a descriptive grammar of Chilean Spanish, neither of which cur-

9. Requests from other types of e-mail address will be considered on a case-by-case basis.

rently exist.[10] These possibilities are in addition to the wide variety of traditional corpus linguistic research which it facilitates.

The next stage of the corpus's development, which will give rise to the COSCACH v2.0, will focus on growth. As funding permits, we plan on transcribing fieldworkers' speech, increasing the size of the non-Chilean control sample, adding samples of speakers from generations 3 through 6 to existing localities, and possibly incorporating several new localities, such as Iquique, Talca and Punta Arenas.

Funding

The creation of the COSCACH was partially funded by CONICYT Chile through FONDECYT grant #11150900 (~ € 120,000).

Acknowledgements

I would like to express my deep gratitude to all the people who generously shared part of their lives by agreeing to be recorded for the COSCACH. The fieldwork involved in the creation of a speech corpus is a massive and often brutally demanding undertaking that is all-too-often glossed over in the literature. I would like to publicly recognize the highly dedicated team of fieldworkers, some of whom spent over two years on the road in order to obtain the recordings that make up the COSCACH. Those who recorded over five speakers are, in descending order, María José Aninao, Beatriz Yáñez-Valdenegro, Sebastián Zepeda-Pallero, Ruth Contreras, Bárbara Galdames, Camila Aedo, Tiare Araya, Edson Salgado, Lorena Perdomo-Pinto and Viviana Vergara-Fernández. Likewise, I would like publicly recognize the work of the team of transcribers who turned the recordings into usable text. Those who made over five transcriptions are Belén Solís-Román, Ignacia Fuentes, Francisco Beltrán, Francisco Martínez, María José Zanetta, Sebastián Zepeda-Pallero, Mareba Torres, Andrea Noria, Ruth Contreras, Bárbara Galdames, Paola Vega and Roby Delgado. Special thanks are due to the project's post-processor and quality control supervisor, Danitza Matus. Finally, I would like to extend my deep gratitude to Sebastián Zepeda-Pallero, who was in charge of the three Arica speaker samples.

10. While a certain number of Chilean dictionaries do exist, they are without exception differential in nature, i.e. dictionaries of Chileanisms. Though of interest to linguists, such works are virtually useless to speakers.

References

Academia Chilena de la Lengua (Ed.). (2010). *Diccionario de uso del español de Chile (DUECh)* [Dictionary of Chileanisms (DUECh)]. MN Editorial / Asociación de Academias de la Lengua Española / Gobierno de Chile / Consejo Nacional de la Cultura y las Artes.

Audacity Development Team. (2018). *Audacity: Free Audio Editor and Recorder* (2.3.0) [Computer software]. http://www.audacityteam.org/

Audix Microphones. (2017). *Audix HT5 Spec Sheet, version 4.1*. https://web.archive.org/web/20150329040350/http://www.audixusa.com/docs_12/specs_pdf/HT5.pdf

Bengoa, J. (2018). *La comunidad fragmentada: Nación y desigualdad en Chile* [The Fragmented Community: Nation and Inequality in Chile]. Editorial Catalonia.

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* (6.0.42) [Computer software]. http://www.praat.org/

Evert, S., & Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*. https://eprints.lancs.ac.uk/id/eprint/62721/1/Paper_153.pdf

Eyheramendy, S., Martinez, F.I., Manevy, F., Vial, C., & Repetto, G.M. (2015). Genetic structure characterization of Chileans reflects historical immigration patterns. *Nature Communications*, *6*, 6472. https://doi.org/10.1038/ncomms7472

Fant, L., & Harvey, A. (2008). Intersubjetividad y consenso en el diálogo: Análisis de un episodio de trabajo en grupo estudiantil [Intersubjectivity and consensus in dialog: Analysis of a student group work session]. *Oralia*, *11*, 307–322.

Fernández de Molina Ortés, E. (2017). Estudio contrastivo de la norma culta de tres ciudades peninsulares. Análisis del campo semántico de la vivienda [A contrastive study of educated speech in three Spanish cities: Analysis of the semantic field of housing]. *Onomázein*, *37*, 90–111. https://doi.org/10.7764/onomazein.37.09

Garretón, M.A., & Cumsille, G. (2002). Las percepciones de la desigualdad en Chile [Perceptions of inequality in Chile]. *Revista Proposiciones*, *34*, 1–9.

Gille, J. (2015). On the development of the Chilean Spanish discourse marker "cachái." *Revue Romane*, *50*(1), 3–29. https://doi.org/10.1075/rro.50.1.01gil

Gundermann, H., Caniguan, J., Clavería, A., & Faúndez, C. (2009). Permanencia y desplazamiento, hipótesis acerca de la vitalidad del mapuzugun [Persistence and displacement: A hypothesis on the vitality of Mapudungun]. *Revista de Lingüística Teórica y Aplicada*, *47*(1), 37–60. https://doi.org/10.4067/S0718-48832009000100003

HandBrake Team. (2019). *HandBrake* (1.2.0) [Computer software]. https://handbrake.fr/

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Heggarty, P., Maguire, W., & McMahon, A. (2010). Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1559), 3829–3843. https://doi.org/10.1098/rstb.2010.0099

Heggarty, P., Shimelman, A., Abete, G., Anderson, C., Sadowsky, S., Paschen, L., Maguire, W., Jocz, L., Aninao, M. J., Wägerle, L., Appelganz, D., Pheula do Couto e Silva, A., Lawyer, L. C., Câmara Cabral, A. S. A., Walworth, M., Michalsky, J., Koile, E., Runge, J., & Bibiko, H.-J. (2019). Sound Comparisons: A new online database and resource for research in phonetic diversity. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia 2019* (pp. 280–284). Australasian Speech Science and Technology Association. http://intro2psycholing.net/ICPhS/papers/ICPhS2019_Proceedings.pdf

Instituto Nacional de Estadísticas. (2018). 1.2 Población total por sexo y área urbana-rural, según grupos de edad [1.2 Total population by sex and urban/rural provenance, by age group]. In *Segunda Entrega de Resultados Censo 2017* [Second Report on the Results of the 2017 Census]. Instituto Nacional de Estadísticas. http://resultados.censo2017.cl/download/1_2_POBLACION.xls

Jørgensen, A. M. (n.d.). *Corpus Oral de Lenguaje Adolescente (COLA)* [Adolescent Spoken Language Corpus (COLA)]. Retrieved December 23, 2021, from https://blogg.hiof.no/colam-esp/

Labov, W. (2001). *Principles of Linguistic Change, vol. 2: Social Factors*. Blackwell.

Labov, W. (2006). *The Social Stratification of English in New York City* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511618208

Li, M., Song, Q., Li, K., Hao, Y., & Chen, X. (2015). *Definition of corpus, scripts, standards and specifications of recording device, environment/speaker coverage for Spanish language, version 1.1* (Technical Report King-ASR-290). SpeechOcean China.

Milroy, L. (1987). *Language and Social Networks* (2nd ed.). Blackwell.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*.

Rabanales, A. (1995). El estudio del habla culta de Santiago de Chile (1967–1993) [The study of educated speech in Santiago, Chile (1967–1993)]. *Thesaurus*, *50*(1–3), 51–68.

Rabanales, A., & Contreras, L. (1979). *El habla culta de Santiago de Chile: Materiales para su estudio, tomo I* [Materials for Studying Educated Speech in Santiago, Chile, vol. 1]. *Anejo N° 2 del Boletín de Filología*. Editorial Universitaria.

Rabanales, A., & Contreras, L. (1990). *El habla culta de Santiago de Chile: Materiales para su estudio, tomo II* [Materials for Studying Educated Speech in Santiago, Chile, vol. 2]. Instituto Caro y Cuervo.

Real Academia Española. (n.d.-a). *Corpus de Referencia del Español Actual (CREA)* [Contemporary Spanish Reference Corpus (CREA)]. Retrieved August 28, 2019, from https://www.rae.es/recursos/banco-de-datos/crea

Real Academia Española. (n.d.-b). *Corpus del Español del Siglo XXI* [Corpus of 21st Century Spanish]. Retrieved August 28, 2019, from https://www.rae.es/recursos/banco-de-datos/corpes-xxi

Rogers, B. (2016). *When Theory and Reality Collide: Exploring Chilean Spanish Intonational Plateaus* [Ph.D. dissertation, University of Minnesota]. The University of Minnesota Digital Conservancy. http://conservancy.umn.edu/handle/11299/181656

Ruiz-Tagle, J. (2016). La persistencia de la segregación y la desigualdad en barrios socialmente diversos: Un estudio de caso en La Florida, Santiago [The persistence of segregation and inequality in socially diverse neighborhoods: A case study from Santiago's La Florida municipality]. *EURE (Santiago)*, *42*(125), 81–108. https://doi.org/10.4067/S0250-71612016000100004

Sadowsky, S. (2016). *FreeLing_es-CL: Chilean Spanish version of the FreeLing tagger* [Computer software]. https://github.com/Linguista/FreeLing-es_CL

Sadowsky, S. (2017). *MaSCoT-R: The Massive Speech Corpus Tool, Recursive Version* (3.2) [Computer software]. https://github.com/Linguista/MaSCoT-R

Sadowsky, S. (2020). Español con (otros) sonidos araucanos: La influencia del mapudungun en el sistema vocálico del castellano chileno [Spanish with (other) Araucanian sounds: The influence of Mapudungun on the Chilean Spanish vowel system]. *Boletín de Filología*, *55*(2), 33–75. https://doi.org/10.4067/S0718-93032020000200033

Sadowsky, S. (2021). EMIS: Sistema de estratificación socioeconómica para la investigación lingüística [EMIS: A socioeconomic stratification system for linguistic research]. In B.M.A. Rogers & M. Figueroa Candia (Eds.), *Lingüística del castellano chileno: Estudios sobre variación, innovación, contacto e identidad* [Chilean Spanish Linguistics: Studies on Variation, Innovation, Contact, and Identity] (pp. 367–396). Vernon Press. https://vernonpress.com/book/606

Sadowsky, S., & Aninao, M.J. (2019). Internal Migration and Ethnicity in Santiago. In A. Lynch (Ed.), *The Routledge Handbook of Spanish in the Global City* (pp. 277–311). Routledge. https://doi.org/10.4324/9781315716350-10

Sadowsky, S., & Salamanca, G. (2011). El inventario fonético del español de Chile: Principios orientadores, inventario provisorio de consonantes y sistema de representación (AFI-CL) [The phonetic inventory of Chilean Spanish: guiding principles, provisional consonant inventory and system of representation (AFI-CL)]. *Onomázein*, *24*(2), 61–84. http://onomazein.letras.uc.cl/Articulos/24/3_Sadowsky.pdf

San Martín, A., & Guerrero, S. (2015). Estudio Sociolingüístico del Español de Chile (ESECH): Recogida y estratificación del corpus de Santiago [Sociolinguistic Study of Chilean Spanish (ESECH): Collection and stratification of the Santiago Corpus]. *Boletín de Filología*, *50*(1), 221–247. https://doi.org/10.4067/S0718-93032015000100009

San Martín, A., Guerrero, S., & Rojas, C. (2016). *PRESEEA-SA: Corpus de Santiago de Chile. Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA)* [PRESEEA-SA: The Santiago, Chile Corpus. Project for the Sociolinguistic Study of Iberian and American Spanish (PRESEEA)]. Universidad de Chile.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, *3*, 215–246. https://doi.org/10.1017/S0047404500004358

Zúñiga, F. (2007). Mapudunguwelaymi am? '¿Acaso ya no hablas mapudungun?' [Mapudunguwelaymi am? 'By chance do you not speak Mapudungun anymore?']. *Estudios Públicos*, *105*, 9–24.

Address for correspondence

Scott Sadowsky
Departamento de Ciencias del Lenguaje
Facultad de Letras
Pontificia Universidad Católica de Chile
Avenida Vicuña Mackenna 4860
Macul
Santiago, CP 7820436
Chile

ssadowsky@gmail.com
https://orcid.org/0000-0001-7797-8435

Publication history

Date received: 9 September 2019
Date accepted: 20 October 2021
Published online: 31 January 2022