Experiencia Piloto en el Levantamiento de un Corpus Especial de Lengua Oral para la Descripción Lingüística del *Creole* de San Andrés

Bryan Steven Loaiza Camacho

Universidad Nacional de Colombia
Facultad de Ciencias Humanas, Departamento de Lingüística
Bogotá, Colombia
2018

Experiencia Piloto en el Levantamiento de un Corpus Especial de Lengua Oral para la Descripción Lingüística del *Creole* de San Andrés

Bryan Steven Loaiza Camacho

Trabajo de grado presentado como requisito para optar al título de:
Lingüista

Directora:
Raquel Sanmiguel Ardila, Ph.D.
Docente Asociada

Línea de Investigación:
Educación, Lengua y Cultura

Grupo de Investigación:
Estado y Sociedad

Universidad Nacional de Colombia
Facultad de Ciencias Humanas, Departamento de Lingüística
Bogotá, Colombia
2018

Dedicatoria

Quisiera dedicar este trabajo a mis padres, Elvia y Carlos, por su amor y su apoyo constante durante todos estos años de vida, en los días buenos y en los malos, y por enseñarme que no se construye un hogar solo con ladrillos; a mi hermano, Andrés, por su buen humor siempre; a mi hermana, Claudia, por mostrarme que al amor no le importa la distancia ni el tiempo; a mis sobrinos, por siempre inspirarme a hacer de este mundo algo un poquito mejor;

a mis mejores amigos: a Camilo, por discutir conmigo acerca del queso y la cuajada; a Juliana, por siempre estar llena de color y al pelo; a Josué, por saber cómo hacerme reír inmediatamente después de hacerme enojar;

a Erika, quien me impulsó a seguirme retando y gracias a quien tuve la oportunidad de llevar a cabo este trabajo;

a Aslan, por siempre alegrarme el día al llegar a casa;

a Felipe, por su eterna paciencia, su incesante luz, su inmarcesible optimismo, su incandescente sonrisa y su infinito amor.

Agradecimientos

Este trabajo no se pudo llevar a cabo sin la ayuda de incontables personas.

En primer lugar, mis más profundos agradecimientos a la profesora Raquel Sanmiguel, por su constante apoyo y motivación, por creer en este proyecto y por sus invaluables aportes en este proceso.

En la Sede Bogotá, a la profesora Ana María Ospina, quien facilitó mi primer acercamiento a esta labor que se estaba gestando. A la profesora María Emilia Montes, por su consejo en los meses que me acercaban por este camino. Al profesor George Dueñas Luna, por permitir que me diera cuenta del mundo de herramientas a mi alcance.

En el Instituto Caro y Cuervo, a Julio Bernal Chávez, por tomarse el tiempo de responder algunas preguntas relacionadas con los corpus electrónicos, y a Alejandro Correa, cuyas enseñanzas aportaron en gran medida al desarrollo de este trabajo.

En la Sede Caribe, agradezco enormemente la colaboración del equipo de trabajo con el que se realizó el piloto; sin ellas, no habríamos logrado tanto ni habríamos aprendido tanto como lo hicimos: a Maureen Hooker, a Emma Forbes y a Penny Bryan, muchísimas gracias por su constancia y su dedicación.

Un agradecimiento para Roy Salmo Suárez, por prestarme su voz por unos minutos y quien facilitó mi primer acercamiento a la lengua *creole*. No fue un curso intensivo, pero fue un buen comienzo.

En el Centro Cultural del Banco de la República en San Andrés, a Andrés Steel, por ofrecer su ayuda en conseguir recursos bibliográficas que apoyaron este trabajo, y a Irma Bermúdez, por mostrarme las instalaciones del Centro de Memorias Orales, los equipos a disposición de este y el conocimiento contenido en esas memorias.

Agradezco a los profesores del grupo Dah Wi! por su cordialidad y la acogida que nos brindaron a mí y a mis compañeros de la Sede Bogotá durante nuestra estadía y a la profesora Luz Amparo Sanabria, por su ayuda y por el *syrup cake*.

A los integrantes del grupo de investigación Estado y Sociedad por su interés en este proyecto.

Resumen

Los estudios del *creole* de San Andrés no han sido sistemáticos ni continuos. Los datos primarios sobre los que se han basado las descripciones son escasos, inaccesibles y están desfasados. Por ello, la Sede Caribe de la Universidad Nacional propone recolectar un corpus de la lengua. El presente trabajo se basa en un piloto llevado a cabo durante el semestre 2018-I para evaluar la viabilidad de dicho proyecto. Con un grupo de colaboradores locales, se hicieron talleres de formación y se recopiló y se transcribió un corpus menor. Aunque no se lograron todas las metas propuestas, el resultado del piloto fue positivo. El análisis del corpus recolectado permite observar ciertos procesos lingüísticos y formular hipótesis, todos dignos de estudio. Una posterior implementación del proyecto debe tener en cuenta los retos y los logros que surgieron durante este piloto. Esperamos que este trabajo también pueda informar otros proyectos de alcance similar en Colombia, donde la lingüística de corpus es un campo aún incipiente, especialmente en materia de lenguas indígenas y criollas.

Palabras clave: *creole* de San Andrés, lingüística de corpus, documentación lingüística

Abstract

Studies on Islander Creole English have not been systematic nor continuous. The primary data that the analytical resources of the language draw upon are scarce, inaccesible, and outdated. Because of this, the National University Caribbean Branch has proposed building a corpus of the language. This paper draws upon a pilot experience executed during the 2018-I semester in order to assess the plausibility of such a project. With a group of local collaborators, training sessions were held and a minor corpus was compiled and transcribed. Although not all goals were met, the outcome of the pilot was positive. An analysis of the collected corpus allows us to observe certain linguistic processes and formulate hypotheses, all of which are worthy of study. A subsecuent implementation of the project must take into account the challenges and the achievements that we came across during the pilot. We hope that this paper can also inform other projects of similar scope in Colombia, where corpus linguistics is still a fledgling discipline, especially regarding indigenous and creole languages.

Keywords: Islander Creole English, corpus linguistics, language documentation

Tabla de Contenido

1. Introducción

El panorama multilingüe en el Archipiélago de San Andrés, Providencia y Santa Catalina ofrece un terreno vasto para la investigación lingüística.

La configuración lingüística de esta región, ubicada a ciento ochenta (180) kilómetros de la costa occidental de Nicaragua y a cuatrocientos ochenta (480) kilómetros de la costa atlántica de Colombia continental, es el resultado de una serie de dinámicas históricas y sociales que han venido transformando la realidad de estas islas.

El primer aporte a la situación lingüística actual se da por medio del establecimiento de ingleses puritanos en la isla de Providencia por vía de las Bermudas en el siglo XVIII y la llegada de colonos jamaiquinos que traían consigo a personas esclavizadas transportadas desde las costas de África Occidental a principios del siglo siguiente (Parsons, 1985). Como evidencia de ello, hoy en las islas se habla una variante caribeña del inglés estándar y un criollo de base léxica inglesa (en adelante, *creole* [1]), el cual tiene un parentesco estrecho con las demás lenguas criollas de base léxica inglesa que se hablan a lo largo del Caribe occidental (García León, 2014).

Posteriormente, el Imperio Español obtiene soberanía sobre el Archipiélago en 1786 por medio de la firma del Tratado de Versalles en Londres y ordena el desalojo de los colonos ingleses que residían en las islas; sin embargo, por intervención del capitán irlandés Thomas O'Neill, se les concedió a estos el permiso de permanecer en el territorio, con la condición de que le rindieran pleitesía a la Corona Española.

Las islas se convirtieron en parte de la recién fundada República de Colombia en 1822, reconociendo la Constitución de Cúcuta, la cual proclamaba que las islas se incorporarían al territorio nacional (Dittmann, 2013).

Pese a estar en manos de diferentes mandatos de lengua mayoritariamente española desde el siglo XVIII, la población de las islas había logrado defender su relativa autonomía y mantener su lengua y sus costumbres durante la mayor parte de su historia. Esto cambia por completo en 1953, con

[1] La lengua ha recibido diversos nombres. En inglés, se ha denominado como *Abacoan speech* (Edwards, 1970) y *Wende*, aunque este último término se considera arcaico (Decker y Keener, 2001); en español se la ha conocido como *criollo sanandresano* (Dittmann, 1992) y como *criollo isleño*. En el estándar ISO 639-3, el cual clasifica cada lengua con un código de tres (3) letras, la lengua se codifica como `icr`, abreviatura de *Islander Creole English* (https://iso639-3.sil.org/code_tables/639/data). Aquí decidimos utilizar el término *creole*, en consonancia con O'Flynn de Chaves (1990), quien lo adopta por el valor positivo connotado.

la visita del entonces presidente, Gustavo Rojas Pinilla, y su proclamación de San Andrés como puerto libre.

Con la conformación del puerto libre y la construcción del aeropuerto en la parte norte de la isla, comenzó una serie de procesos y de dinámicas que transformaron el panorama del territorio a un ritmo acelerado. Incrementó la población de manera precipitada con la llegada masiva de migrantes provenientes de Colombia continental y este hecho, junto con una mayor presencia del Estado, ha provocado una serie de tensiones que en buena medida han afectado el panorama lingüístico de San Andrés en los últimos decenios.

En efecto, se ha podido observar un proceso de desplazamiento lingüístico, en el que la lengua española ha ido ganando predominancia sobre la variedad de inglés caribeño de la isla y el *creole* en el interior de la comunidad raizal, situación reflejada en las tasas de monolingüismo y bilingüismo en la isla: mientras que el monolingüismo y el bilingüismo en inglés y *creole* están en disminución y se restringen a las generaciones de mayor edad, ha habido un aumento en el bilingüismo *creole*-español y en el monolingüismo en español en las generaciones más jóvenes (Andrade Arbeláez, 2006). Es de esperarse que, en estas condiciones, eventualmente el *creole* caiga en desuso y sufra el destino de atrición lingüística (Crystal, 2000).

Como parte de los esfuerzos por asegurar la vitalidad de la lengua *creole*, es necesario estudiarla y conocer su estructura, así como los procesos que sufre en los diferentes niveles (fonológico, morfológico, sintáctico, léxico, etc.) al estar en contacto con otros sistemas lingüísticos, como lo son el inglés de variedad caribeña y el español. Los estudios sobre la lengua, sin embargo, han avanzado de manera ralentizada, una situación que se ha visto exacerbada por el estigma y el rechazo hacia esta lengua tanto adentro como afuera de la comunidad raizal, por un lado (O'Flynn de Chaves, 1990), la falta de consecución de recursos en pos de la lengua por parte de instituciones estatales, por otro, y por el recelo que mantienen ciertos sectores de la comunidad raizal contra las personas exteriores que quieren estudiar la lengua.

Consecuentemente, los estudios descriptivos de la lengua son contados. Entre ellos, se destacan los trabajos de Edwards (1970), O'Flynn de Chaves (1990), Dittmann (1992) y Bartens (2003). Cada uno aborda diferentes temas relativas a la estructura de la lengua desde perspectivas y paradigmas distintos, lo cual permite examinarlos en conjunto y así obtener una visión más íntegra de la lengua que la que brindaría una lectura aislada de uno u otro. Nos parece importante, en este punto, reseñar brevemente cada uno de estos trabajos y mencionar los puntos en los se destacan y asimismo aquellos en los que hace falta una mayor profundidad.

Edwards (1970) examina la variación lingüística del habla en la isla de San Andrés desde su dimensión social. Logra una descripción fonológica de los sociolectos presentes en la sociedad isleña que los individuos apropian en la forma de sus propios ideolectos. La variabilidad de estos ideolectos, explica, depende en gran medida del punto de la escala social en la que se encuentran los individuos. Este es un aporte importante al estudio sociolingüístico de la isla; sin embargo, el trabajo es de hace

casi cinco decenios y, por los cambios que se han evidenciado en la isla en este periodo de tiempo, es notoria la necesidad de datos más recientes. Por otro lado, Edwards (1970) únicamente examina la variación presente en el nivel fonológico y no hace referencia a la estructura morfosintáctica sino en relación con la fonología. Por estas cuestiones, es evidente la necesidad de un trabajo de revisión histórica y de profundización en los diferentes niveles de la lengua.

La obra de O'Flynn de Chaves (1990) responde en cierta medidad a dicha necesidad. Esta obra examina los niveles de la lengua que no se habían estudiado antes. El estudio que realizó podría considerarse como la primera aproximación sistemática hacia la lengua. En este libro, expone la fonología de la lengua sin examinar la variación que puede presentar y da un esquema conciso de los procesos fonológicos y de la fonotaxis de la misma. En el nivel de la sintaxis, brinda un listado de las clases de palabras presentes en la lengua, así como una descripción de la estructura del sintagma nominal y verbal y una caracterización de la predicación. En estas cuestiones dedica un espacio breve, ya que su interés primordial es estudiar la forma en que se marca el tiempo, el aspecto y la modalidad en esta lengua. El énfasis que se les coloca aquí se debe a la tendencia de las lenguas criollas de usar marcadores preverbales para denotar estos rasgos (Patiño Rosselli, 2000), una característica presente en *creole* y cuyas sutilezas logra sintetizar la autora. Esta obra se aproxima a la variedad basilectal presente en la isla de San Andrés; sin embargo, no se examinan otras cuestiones de tipo variacionista, hecho que constata la misma autora al mencionar que filtró aquellas muestras de habla que se aproximaban demasiado al inglés. Por otro lado, a partir de esta obra, se muestra imperioso el estudio sobre las clases de palabras de la lengua y una discusión más profunda acerca de las mismas.

El trabajo de Bartens (2003) complementa esta carencia de discusión sobre las clases de palabras del *creole*. En esta obra, caracteriza cada una de las clases de palabras que identificó a partir del análisis de la estructura gramatical del *creole* y una comparación con las estructuras gramaticales del inglés y del español. Ahonda, además, en otras cuestiones de sintaxis, como la formación de frases adverbiales y preposicionales y la formación de oraciones simples y complejas. En cuestión de léxico, se brinda un apéndice en el que se sistematizan algunos africanismos presentes en *creole*. Es de destacar la crítica que hacen Forbes y Kouwenberg (2005) sobre esta obra: afirman que, para los estudios lingüísticos, la descripción es algo superficial, ya que se dirige a maestros de lengua sin demasiados conocimientos sobre lingüística, y aún así requiere que se sepa algo de esta disciplina. Por otro lado, al enfatizar en la comparación con el inglés y el español, hay poca discusión sobre construcciones y rasgos que no tienen equivalentes en estas dos lenguas, si es que los hay. Añadimos también que, pese a que sí se tratan ciertas cuestiones de variación, por la misma naturaleza de la obra no se permite un examen minucioso de ellas.

Por último, el trabajo de Dittmann (1992) es un referente para los estudios sociolingüísticos del territorio. Esboza un panorama de la cultura isleña, el cual aborda la realidad de las tres (3) lenguas presentes en el territorio en relación con las funciones que cumplen y los dominios y espacios en los que se utilizan. La descripción que realiza de la lengua es somera y concisa y propone ser un referente para la elaboración de currículos y materiales educativos. Caracteriza la lengua en el

contexto del libro, razón por la cual las observaciones que realiza no tienen un alcance tan amplio como las obras anteriores; sin embargo, es de resaltar que se presta una atención particular a la variación lingüística en los diferentes niveles de la lengua y se examina esta variación en relación con factores sociales y con la movilidad en la escala social. Esta obra, en cierta medida, adolece de lo mismo que la obra de Edwards (1970): al ser un trabajo de hace casi tres decenios, se debe apuntar a obtener datos más recientes. Esta misma autora expande sobre varios aspectos de la lengua en una descripción más actualizada, aunque corta, en Dittmann (2013).

Basándonos en el panorama de los trabajos sobre la lengua que hemos presentado hasta aquí, nos parece evidente la necesidad de continuar adelantando estudios descriptivos y sincrónicos del *creole* que permitan expandir sobre el conocimiento que ya se tiene sobre él; que logren dar cuenta de la variación de la lengua al interior de la misma comunidad raizal, puesto que cualquier gramática de una lengua criolla debe "tener un alcance 'polilectal' " (Patiño Rosselli, 2000, p.133) (i.e. debe dar cuenta de la variación en el rango de los ideolectos intermedios que hay entre el *creole* y el inglés de variedad caribeña), y debe permitir que se examinen los fenómenos de cambio que se producen en el habla como resultado del contacto entre el *creole* y el español.

El presente trabajo propone aproximarse a la necesidad de ampliar los estudios acerca de la lengua *creole* y los alcances de los mismos. Para ello, se plantea en el enfoque de la lingüística de corpus. Las investigaciones lingüísticas enmarcados en esta perspectiva se basan en la observación de muestras reales de habla y permiten aunar sus resultados con datos tangibles y cuantificables. Exploraremos asimismo la posibilidad de enmarcar esta propuesta en el desarrollo de corpus electrónicos, los cuales suelen constar de una ingente cantidad de datos y cuyo tratamiento por medio de herramientas informáticas permite una exploración sistemática y amplia.

Expondremos a lo largo de este trabajo la experiencia de haber llevado a cabo un piloto para el proyecto de recolección de un corpus oral de la lengua *creole* de San Andrés en el semestre 2018-I. Pasaremos revista al enfoque teórico fundamentado en el trabajo con corpus electrónicos, a la métodología de campo empleada, a los resultados del piloto, a los retos que se presentaron en el camino y a las recomendaciones derivadas de esta experiencia, las cuales, esperamos, sirvan para el desarrollo posterior y continuo del proyecto.

Este trabajo no solo propone ser un insumo importante para los estudios descriptivos del *creole* de San Andrés, sino también un aporte a los estudios lingüísticos en Colombia, país en el que la aplicación de herramientas de lingüística de corpus es aún muy incipiente, especialmente en lo que se refiere a lenguas indígenas y criollas [2].

[2] J. Bernal Chávez, comunicación personal, 11 de mayo de 2018.

2. Antecedentes y Justificación

El presente trabajo se enmarca en el proyecto de investigación "Educación, lengua y cultura en el contexto plurilingüe y multicultural Caribe del Archipiélago de San Andrés, Providencia y Santa Catalina", el cual surge del grupo Estado y Sociedad de la Sede Caribe de la Universidad Nacional de Colombia en la línea *Educación, Lengua y Cultura*.

El grupo Estado y Sociedad ha venido trabajando para adelantar acciones investigativas que contribuyan al entendimiento de la conformación social, cultural y lingüística actual de San Andrés. Por medio de estas tentativas propone ubicar la realidad sanandresana en el contexto del Caribe occidental y del Gran Caribe, con los cuales mantiene lazos no solamente históricos y geográficos, sino también culturales y lingüísticos, entre otros.

Desde la línea *Educación, Lengua y Cultura*, se enfatiza en la necesidad de contextualizar los currículos educativos de los niveles de primaria y secundaria a la realidad lingüística y cultural de la isla y del Gran Caribe, puesto que se ha podido evidenciar que en la gran mayoría de casos, estos currículos no logran articular conocimientos del territorio en sus planes de estudio. Para facilitar y aportar a este proceso de contextualización, la línea se propone crear materiales de referencia (diccionarios y gramáticas) para consulta de los profesores y los coordinadores de los institutos educativos (Sanmiguel Ardila, 2017).

Entre marzo de 2011 y marzo de 2013, a partir del Proyecto de Política y Planeación Nacional realizado en la Sede Caribe, se identificaron una serie de hechos:

1. es necesario formular acciones investigativas que contribuyan a avanzar en el conocimiento y la conceptualización sobre la naturaleza, el uso y la función identitaria de la lengua *creole*,
2. entre los habitantes de la isla hay desinformación, falta de consenso y actitudes disímiles con respecto a la lengua y
3. desde la Universidad Nacional no se han adelantado estudios académicos sobre el *creole* de forma sistemática y constante, debido en gran medida a la carencia de profesionales de ciencias del lenguaje que permanezcan de forma permanente en la isla.

Con base en esto, en 2014 se articula un proyecto enfocado hacia la investigación de la lengua *creole* (Sanmiguel Ardila, Schoch, y Pelufo, 2014). Con el propósito de formular el proyecto, se realizó una revisión de las investigaciones existentes para así conocer el estado de los estudios sobre la lengua. En esa medida, se hizo una revisión sobre dos campos específicos: la educación y la descripción lingüística. Aquí nos enfocaremos en la segunda.

La observaciones principales que surgieron en el transcurso de la formulación del proyecto fueron:

- las últimas investigaciones de enfoque descriptivo sobre la lengua son de hace más de quince (15) años,
- en varias de estas investigaciones no se menciona de manera explícita la manera en que se obtuvieron los datos ni los participantes que tomaron parte en ellas y
- los datos primarios sobre los que se basaron estas investigaciones son escasos, de difícil acceso y/o están desfasados y requieren actualización (Sanmiguel Ardila y cols., 2014).

Basándose en estas observaciones, el proyecto de investigación se propuso servir como un insumo para recopilar datos lingüísticos de la lengua que sirvan de base para adelantar nuevos estudios descriptivos y sincrónicos de la lengua. Por la dificultad del acceso a datos de investigaciones previas y el obstáculo que ello presenta para llevar a cabo nuevos estudios, se adujo además que el corpus producto del proyecto se ha de almacenar en un repositorio que permita el acceso posterior de investigadores de diversa índole, de manera que ellos puedan basar sus propios trabajos investigativos sobre los datos que obtengan de su uso.

Para ello, se sugirió que la mejor metodología en la que se puede apoyar esta recopilación es la lingüística de corpus y el enfoque del proyecto habría de ser de carácter empírico y cuantitativo, aunque no se ha de ignorar los enfoques de corte cualitativo. Por el avance que ha tenido la informática y el consecuente mejoramiento de las técnicas de recolección de datos, se propuso adelantar esta labor en el esquema de los corpus electrónicos y la lingüística computacional y se evaluó la viabilidad de trabajar con un corpus oral y con un corpus escrito (Sanmiguel Ardila y cols., 2014).

Con el objetivo de acercar el Departamento de Lingüística de la Sede Bogotá al proyecto de investigación y de lograr trazar una ruta común, en septiembre de 2017 se invitó a las profesoras Ana María Ospina Bozzi y María Emilia Montes Rodríguez, lingüistas de larga trayectoria en el país y docentes asociadas de ese departamento, a participar en las actividades de la Mesa Creole. Este fue un espacio en el que se reunieron miembros de la comunidad con interés en la lengua y diferentes investigadores con experiencia en metodologías de compilación de datos lingüísticos de lenguas minoritarias y de tradición oral. En este espacio se abordaron y se hicieron sugerencias acerca de diversos temas, como: los criterios y las herramientas para la compilación de datos, los criterios para la selección de consultores de lengua, la representatividad de las muestras, el tipo de encuestas que se pueden aplicar y diferentes temas gramaticales, léxicos y de tipología textual que pueden ser dignos de estudio. El resultado de la Mesa Creole fue positivo y demostró la viabilidad de que se adelante la labor de recopilación de un corpus [1].

Para finalizar este punto, el proyecto de investigación actualmente vigente y en la que se enmarca este trabajo retoma la labor realizada en el 2014 y, a partir de ella, se formula como una vía para el acercamiento de estudiantes de pregrado del Departamento de Lingüística en Bogotá con el objeto de contribuir con este proyecto por medio de pasantías, prácticas académicas y/o monografías

[1] María Emilia Montes, comunicación personal, 6 de septiembre de 2017.

investigativas (Sanmiguel Ardila, 2017).

2.1. Repositorios de la lengua creole

En la actualidad hay pocos repositorios que contengan datos lingüísticos del *creole*.

En relación con los datos primarios en los que se basaron las investigaciones reseñadas en la Introducción, únicamente los de Dittmann (1992) fueron puestos en acceso general por voluntad de la investigadora. Los documentos escritos y las grabaciones que obtuvo durante el trabajo en campo se encuentran despositados en la Hemeroteca del Departamento de Idiomas de la Universidad del Valle (Sanmiguel Ardila y cols., 2014).

Por lo demás, los datos utilizados en las otras investigaciones no están disponibles de forma pública.

En algunos repositorios en línea hay datos lingüísticos del *creole*; sin embargo, estos aun son reducidos.

En el APiCS Online (*The Atlas of Pidgin and Creole Language Structures Online*), hallamos una entrada enciclopédica de la lengua realizada por Angela Bartens. En ella, se reseña el trasfondo sociohistórico de la lengua, su situación sociolingüística, la fonología, la gramática y el léxico de la misma. Estos rasgos lingüísticos y sociolingüísticos se resumen en una base ofrecida por la misma página. En cuanto a datos primarios, solo se encuentra en el sitio Web un texto en *creole* (una historia de Anaansi) con glosa interlineal de cuatro (4) páginas [2].

Hallamos en el OLAC (*Open Language Archives Community*) otro repositorio pequeño del *creole*. En él, están depositadas dos (2) grabaciones en formato WAV puestas a disposición por el investigador Derek Bickerton. Una de las grabaciones fue tomada en una fecha desconocida, aunque se presume que fue grabada en la década de los 60, dura cuartenta y cuatro (44) minutos y contiene las voces de una mujer de edad avanzada y un hombre que reporta tener dieciocho (18) años quienes relatan diferentes aspectos de la vida y la historia en San Andrés. La segunda grabación no se encuentra tan bien documentada, ya que no se reporta ni de quienes son las voces en la grabación ni se documenta el contenido de la grabación o la época en la que pudo haber tomado lugar; en lo poco que se documenta en el repositorio, se indica que en la grabación hay datos de dos (2) lenguas: el *creole* de San Andrés y el criollo guyanés de base léxica inglesa [3].

En el contexto nacional, hallamos que el Centro Cultural del Banco de la República en San Andrés adelanta un proyecto para la conformación de un Centro de Memorias Orales (Banco de la República, 2017), el cual propone "recolectar, preservar y divulgar la memoria oral" con la participación

[2]http://apics-online.info/surveys/10
[3]http://www.language-archives.org/language/icr

de miembros de la comunidad raizal. El proyecto comenzó en el año 2015 y actualmente cuenta con cincuenta (50) grabaciones de historias de vida, cuentos, cantos, testimonios, entre otros productos orales derivados de las vivencias y las experiencias de personas raizales. Cada una de estas grabaciones se encuentra disponible para su consulta en línea. Las grabaciones no tienen una duración prefijada, pero se estipula que no pueden superar los noventa (90) minutos. En el repositorio del Centro de Memorias, hay veintiséis (26) grabaciones que contienen muestras de la lengua *creole*. Es un trabajo en curso; sin embargo, los productos del Centro son un insumo valioso para la documentación en el Archipiélago y pueden ser un aporte para los estudios descriptivos de la lengua, motivo por el cual es factible pensar en aunar esfuerzos para proyectos comunes entre la Universidad Nacional y el Banco de la República [4].

2.2. Corpus electrónicos en Colombia

La importancia de los corpus lingüísticos en el estudio y la descripción de las lenguas se ha reconocido desde el siglo XVIII, cuando se recopilaban grandes conjuntos de textos escritos con la finalidad de estudiar lenguas muertas, como el latín y el griego antiguo, así como anotar y comentar textos bíblicos (Villayande Llamazares, 2010, pp. 295–296). Esta tendencia también se ve reflejada en los estudios lexicográficos y la compilación de diccionarios, e.g., el *Oxford English Dictionary* en Inglaterra y *An American Dictionary of the English Langauge* en Estados Unidos. Ambos trabajos fueron resultado del análisis de grandes recolecciones de citas representativas del inglés de la época (Bolaños Cuellar, 2015, pp. 40–41).

En el contexto colombiano, hay un paralelo con la labor documental y lexicográfica de este periodo con la publicación de los primeros tomos del *Diccionario de construcción y régimen de la lengua castellana*, los cuales se basaron en el trabajo de recolección del filólogo Rufino José Cuervo, quien había consultado citas de los escritores más afamados en lengua española hasta ese momento (Bolaños Cuellar, 2015, p. 41).

Fue en los comienzos del siglo XX, con el estructuralismo estadounidense, cuando se empezaron a sentar las bases de lo que hoy se llegaría a conocer como *lingüística de corpus* y a considerarse los corpus de lengua como herramienta fundamental para la lingüística al ser un suministro importante de datos reales y tangibles que apoyan la investigación (Villayande Llamazares, 2010).

En Colombia, las metodologías inspiradas en los trabajos de corpus de lengua de los estadounidenses apoyaron labores de documentación no solo lexicográfica, sino también dialectológica y sociolingüística, como se evidencia en la recopilación de muestras para el *Atlas Lingüístico-Etnográfico de Colombia* (ALEC) y los trabajos sobre el español hablado en Bogotá, todos productos del Instituto Caro y Cuervo (ICC) (Lozano Ramírez, 2012).

[4] http://www.banrepcultural.org/centro-de-memorias-orales

En un principio el trabajo con corpus lingüísticos se hacía de forma manual y era lento en comparación con el potencial que ofrecen los corpus electrónicos actuales; sin embargo, en las últimas tres (3) décadas del siglo XX, luego de superar las críticas del generativismo chomskyano, los avances tecnológicos en informática posibilitaron la recopilación de muestras comparativamente más grandes de lengua desde diversas fuentes y su organización en lo que ahora conocemos como corpus electrónicos (Villayande Llamazares, 2010). Los proyectos surgidos en este periodo, aunque no posean el tamaño ni el nivel de desarrollo de los corpus electrónicos de hoy, aún informan muchos de los principios de estos: el carácter informatizado de los documentos recopilados, el equilibrio en el tamaño de las muestras y la representatividad de cada una de las variedades lingüísticas que se estudien.

Los corpus electrónicos más grandes suelen ser de lenguas globales, e.g., del inglés, del español, del francés, etc. Estos corpus lingüísticos, por lo general, apuntan a ser representativos de una lengua o de una determinada variedad de lengua; compárese en este sentido el *International Corpus of English* (ICE) [5] , el cual se propone ser un corpus general de la lengua inglesa en veinte (20) países donde se habla como primera o segunda lengua, con el *Corpus of Contemporary American English* (COCA) [6] , el cual recoge muestras de más de 160 000 textos de Estados Unidos producidos entre 1990 y 2017 y que propone ser un corpus representativo de las variedades de lengua inglesa de dicho país.

Un corpus electrónico del español actual notable es el que se recoge en el contexto del *Proyecto para el estudio sociolingüístico de España y de América* (PRESEEA) [7]. Este proyecto busca crear un corpus general representativo del español, no solo en sus dimensiones dialectales y geográficas, sino también sociales. En estos momentos está disponible la consulta de muestras de catorce (14) ciudades del mundo hispano.

En Colombia el campo de la lingüística de corpus está aún en una etapa incipiente. En el I Congreso Internacional de Lingüística Computacional y de Corpus, celebrado en mayo de 2017 en la sede del ICC en Bogotá, se observó que varias de las iniciativas nacionales se construyen en el contexto de un proyecto investigativo más amplio y responden a una problemática o una pregunta determinada, e.g., sobre lenguaje farmacéutico, médico, legal, etc. Por otro lado, no existe un corpus general del español de Colombia, aunque sí se ha sugerido la necesidad de construir uno (Davies, 2017).

Los esfuerzos que se han realizado para construir corpus electrónicos del español en el país suelen enfocarse en un dialecto, en la variedad lingüística de una ciudad, en un género o en una temática determinada. En este sentido, se destacan las labores del Grupo de Investigación en Traducción y Nuevas Tecnologías (Grupo TNT) [8] y del Grupo de Estudios Sociolingüísticos [9] , ambos de la

[5]http://ice-corpora.net/ice/
[6]https://corpus.byu.edu/coca/
[7]http://preseea.linguas.net/
[8]http://www.udea.edu.co/wps/portal/udea/web/inicio/investigacion/grupos-investigacion/humanidades/tnt
[9]http://www.udea.edu.co/wps/portal/udea/web/inicio/investigacion/grupos-

Universidad de Antioquia. El primer grupo ha levantado un corpus que evidencia la descripción de la pobreza en los medios de comunicación colombianos (POLAME Corpus) [10] y otro de artículos periodísticos generales de tres (3) de los diarios de mayor difusión en Colombia (TNT Corpus) [11] , ambos completamente anotados por clases de palabra y lematizados; el segundo grupo levantó el Corpus Sociolingüístico de Medellín [12] , proceso para el cual se utilizaron métodos de campo útiles para informar el levantamiento de nuevos corpus en el país. La construcción de este corpus hace parte de la inclusión de muestras del español de Medellín en el PRESEEA [13] . También se tiene planeado incluir en este proyecto muestras del español de Tunja (Calderón Noguera, 2008) [14] y de Bogotá [15] .

En la actualidad, el ICC adelanta esfuerzos para digitalizar los corpus lingüísticos orales recopilados para el ALEC y las investigaciones sobre el *Español Hablado en Bogotá* y el *Habla Culta de Bogotá* (Bernal Chávez, Bonilla, Rubio, Llanos Chávez, y Bejarano Bejarano, 2018). Desde 2013 se han llevado a cabo diferentes fases para convertir los audios originales en archivos digitales, documentar los metadatos correspondientes a cada sesión con los consultores de lengua y hacer la transcripción y alineación de los audios en formato digital. En 2017 el ICC articuló el proyecto "Corpus Lingüísticos del Instituto Caro y Cuervo" (CLICC). Uno de los objetivos de este proyecto es desarrollar un Sistema Gestor de Contenidos (SGC) para "el almacenamiento, la organización, la consulta y la explotación de los corpus del ICC"[16]. Está planeado que se permita el libre acceso al proyecto en el segundo semestre de 2018 [17].

Es de notar que, aunque el panorama lingüístico de Colombia es diverso, no se han explotado los avances en lingüística de corpus para desarrollar corpus electrónicos de las lenguas indígenas y criollas en el país. En efecto, no se encuentran corpus en línea para estas lenguas [18]; hay algunos repositorios para la documentación lingüística, como el del Centro Colombiano de Estudios de Lenguas Aborígenes (CCELA), el cual surge de la Maestría en Etnolingüística de la Universidad de los Andes (Colciencias, 2006), y el Centro de Documentación Palabra y Memoria, del Departamento de Lingüística de la Universidad Nacional de Colombia [19].

Por lo anterior, consideramos que la lingüística de corpus es una metodología cuyos potenciales no se han explorado ni refinado lo suficiente para el estudio de la diversidad lingüística en Colombia y, por ello, la construcción de corpus electrónicos de las lenguas indígenas y criollas puede

investigacion/humanidades/estudios-sociolinguisticos
[10] http://grupotnt.udea.edu.co/polame-corpus/
[11] http://grupotnt.udea.edu.co/corpustnt/
[12] http://comunicaciones.udea.edu.co/corpuslinguistico/
[13] http://preseea.linguas.net/Equipos/Medell%C3%ADn.aspx
[14] http://preseea.linguas.net/Equipos/Tunja.aspx
[15] http://preseea.linguas.net/Equipos/Bogot%C3%A1.aspx
[16] http://caroycuervo.gov.co
[17] J. Bernal Chávez, comunicación personal, 11 de mayo de 2018.
[18] J. Bernal Chávez, comunicación personal, 11 de mayo de 2018
[19] http://www.humanas.unal.edu.co/linguistica/laboratorios-y-centros-de-documentacion/centro-de-documentacion-palabra-y-memoria/

beneficiar no solamente a las comunidades involucradas, sino también la labor lingüística en el país.

2.3. Corpus electrónicos de lenguas criollas

La criollística, al ser un campo interdisciplinar, colinda con la lingüística de contacto. En este sentido, es un campo orientado en gran medida al análisis de datos, el cual se puede beneficiar de las metodologías de la lingüística de corpus (Mello, 2014).

No están disponibles, sin embargo, muchos corpus electrónicos de variedades de contacto y, entre estos, menos de lenguas criollas. Entre los que existen, destacamos el *Corpus of Northern Haitian Creole*, de la Universidad de Indiana [20] ; el *Chavacano Language Corpus Project*, de la Universidad Ateneo de Zamboango (Filipinas) (Tardo, 2006) y los corpus de San Tome y de los criollos del Golfo de Guinea, de la Universidad de Lisboa (Hagemeijer, Hendrickx, Amaro, y Tiny, 2012), entre otras propuestas de proyección pequeña.

Un corpus cercano al que se plantea en este trabajo es el *Corpus of Written British Creole*, de la Universidad de Lancaster, el cual se propone estudiar textos literarios de autores jamaiquinos y de habla criolla radicados en Gran Bretaña (Sebba y Dray, 2007). Este corpus se propuso como un experimento para observar los retos y el potencial de compilar corpus de lenguas sin estandarizar. Entre los retos con los que se enfrentó el proyecto estuvieron obtener las debidas licencias de los autores y las editoriales para incluir sus textos (por motivos económicos e ideológicos, entre otros), no disponer del tiempo ni de los recursos para incluir y anotar los textos y también determinar la representatividad del corpus, un reto que se vio exacerbado por la distinción ocasionalmente borrosa entre el criollo jamaiquino y el inglés estándar, especialmente en un formato escrito.

Una de las problemáticas más relevantes en la compilación de este corpus fue la relación *type-token* [21]; en una lengua sin estandarizar como el criollo jamaiquino, es difícil determinar si ciertos *tokens* se deben relacionar a un mismo *type* o a *types* diferentes. Esto afecta en gran medida la anotación del corpus. Pese a sus limitaciones, la aplicación del corpus ha contribuido al análisis de ciertos rasgos gramaticales como la expresión de la modalidad y además ha permitido examinar la lengua en uso para así distinguir entre lo que es lengua criolla y lo que no lo es. Esto, a su vez, es un insumo importante para la estandarización y la escolarización en el criollo jamaiquino.

Esperamos que el piloto descrito en el presente trabajo complemente estos desarrollos de corpus electrónicos de lenguas criollas y asimismo pueda ahondar y contribuir al conocimiento en cuanto a los retos y el potencial de levantar datos de lenguas sin estandarizar.

[20] http://www.indiana.edu/ creole/index.shtml

[21] En términos simples, los *tokens* son cada una de las palabras que aparecen en un texto y los *types* son todas las palabras *diferentes* que aparecen en él. Normalmente, son más los *tokens* que los *types* en un texto cualquiera, ya que rara vez se encuentra que cada palabra en un texto ocurra una sola vez (Manning y Schütze, 1999).

3. Objetivos

Para la formulación del piloto, se enunciaron los siguientes objetivos.

Objetivo general

- Llevar a cabo una experiencia piloto que contribuya al levantamiento de un corpus especial de la lengua *creole* de San Andrés por medio de la formación de un equipo local de colaboradores con el que se pueda adelantar actividades de recolección y transcripción de datos lingüísticos.

Objetivos específicos

- Establecer los parámetros específicos que guiarían la recolección de los datos lingüísticos y en los cuales se basaría la transcripción y anotación del corpus.
- Llevar a cabo talleres que abordarían metodologías de lingüística de corpus y entrenar al equipo local de colaboradores en técnicas de recolección en campo.
- Supervisar la transcripción y anotación de los datos lingüísticos recogidos, de forma que se les pudiera hacer un tratamiento preciso y eficaz.

4. Marco Teórico

Puesto que el proyecto de recolección de un corpus del *creole* de San Andrés está enfocado tanto hacia la descripción como la documentación de una lengua criolla, el piloto se enmarca teóricamente entre dos (2) campos: la *documentación lingüística* y la *criollística*. La *lingüística de corpus*, la cual determina la metodología y el enfoque del proyecto, guarda amplia relación con la documentación lingüística y ambas, aunque con frecuencia buscan fines divergentes, se encuentran ligadas en varios aspectos fundamentales. Por otro lado, la criollística aparece como un campo que presenta diferentes retos para las dos.

Haremos aquí un barrido general por los principales postulados conceptuales que atañen a este piloto.

4.1. La Lingüística de Corpus

Es de notar que, entre los campos de las ciencias del lenguaje, la *lingüística de corpus* no se puede considerar conceptualmente *per se* una teoría, sino una metodología y un enfoque que informa otros campos y cuya aplicación puede beneficiar diferentes ramas de investigación lingüística (Gries, 2009). Por lo tanto, es importante que los resultados tomados de una investigación con corpus se interpreten a la luz de alguna teoría (Baker, 2010).

En lo que sigue de este apartado, daremos algunas definiciones conceptuales acerca de dos (2) de los elementos más importantes en este campo: los *corpus* como tal y la *anotación de corpus*, con énfasis especial en la anotación gramatical.

4.1.1. Los corpus lingüísticos

El término *corpus* puede denotar diferentes conceptos. La definición cotidiana reza que es un "conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación" [1] , definición común entre las ciencias humanas y naturales y de la cual surgen términos como "corpus de conocimiento" para referirse al compendio conceptual de una profesión [2].

En el sentido en el que se utiliza en la lingüística, al cual nos adscribimos aquí, se distingue de esa definición y se refiere específicamente a los *corpus lingüísticos*. En esencia, "cualquier texto o

[1] http://dle.rae.es/?id=AwTBMcs
[2] https://www.regulacioninformatica.org/wiki/index.php?title=Corpus_de_Conocimiento

colección de textos se puede concebir teóricamente como un corpus" (Baker, 2010, p. 95) [3] ; sin embargo, aquí entendemos como corpus lingüístico "una colección de piezas de una lengua que se seleccionan y se ordenan de acuerdo a criterios lingüísticos explícitos para ser utilizadas como una muestra de dicha lengua" (Sinclair, 1996, p. 4) [4].

Por virtud de los avances tecnológicos que han impactado en la compilación de corpus lingüísticos, el sentido de *corpus* se ha extendido hasta incluir aquellos corpus lingüísticos que tienen un soporte informático. Hacemos una distinción con respecto a los primeros al referirnos a estos como *corpus electrónicos*.

Como se puede ver, hay un mayor nivel de especificidad con cada denominación: el corpus lingüístico es una clase de corpus y el corpus informático es un tipo de corpus lingüístico. En lo subsiguiente y a lo largo de este trabajo, usaremos los tres términos indistintamente, a menos que sea necesario hacer la distinción entre corpus lingüístico y corpus electrónico, en cuyo caso haremos énfasis en dicha distinción.

Por sí mismas, estas características no son suficientes para definir un corpus. Los corpus, según McEnery y Wilson (2001), suelen seguir una serie de pautas:

- los textos de un corpus suelen ser muestras de una variedad lingüística, i.e., puesto que raramente se pueden analizar todos los enunciados de una lengua, se debe ser selectivo en la recopilación de los textos o fragmentos de lengua que se incluyan en un corpus;
- los corpus suelen ser representativos, i.e, apuntan a ser lo más diverso posible en cuanto a sus fuentes y a sus dominios de uso, para así develar patrones lo más generales posibles de una determinada variedad lingüística;
- los corpus son de tamaño finito, puesto que no es factible recolectar todas las muestras posibles de una variedad lingüística y la recolección desmedida de muestras puede tener implicaciones negativas en lo que se refiere a la representatividad;
- los corpus suelen estar informatizados y esto permite que se analicen con un mayor poder de procesamiento que si se analizaran de forma manual.

Partiendo de lo expuesto anteriormente, podemos definir un corpus electrónico como

> un *conjunto de textos* de *lenguaje natural* e *irrestricto*, almacenados en un *formato electrónico homogéneo*, y *seleccionados* y *ordenados* de acuerdo con *criterios explícitos* para ser utilizados como *modelo* de un estado o nivel de lengua determinado, en estudios o apliaciones relacionados en mayor o menor medida con el *análisis lingüístico* (Santulla, 2005, citado en Villayande Llamazares, 2010, p. 325).

[3] Traducción propia.
[4] Traducción propia.

Conocemos ya lo que caracteriza un corpus electrónico como el que se plantea en este trabajo. Con esto en mente, podemos comenzar a teorizar acerca de la tipología específica del corpus de nuestro piloto. Hay distintas formas de clasificar los corpus. Entre ellas, destacamos:

- **Por la modalidad o el medio de producción de los textos:** podemos contar principalmente dos (2) tipos de corpus: *escritos* y *orales*. Los corpus escritos contienen muestras de lengua escrita y su recolección suele ser sencilla, gracias a la cantidad de textos digitalizados y disponibles para su estudio. Por otro lado, los corpus orales recogen muestras de lengua hablada y suelen incluir transcripciones de estas muestras, grabaciones o ambas [5].

- **Por el tamaño y la especificidad de los textos:** podemos distinguir entre *corpus generales*, los cuales suelen ser representativos de una lengua o una variedad particular y requieren de un trabajo largo de compilación, y los *corpus especializados*, los cuales reúnen textos de un tipo particular de lengua (o *sublenguaje*) y son útiles, e.g., para describir una jerga (Villayande Llamazares, 2010; Bernal Chávez y Hincapié Moreno, 2018).

- **Por el número de lenguas:** los corpus que contienen textos en un sola lengua o variedad se les conoce como *corpus monolingües* [6] ; los que contienen textos en dos o más lenguas se denominan *corpus multilingües*. Los corpus multilingües se dan únicamente de dos maneras (Sinclair, 1996): en la forma de *corpus paralelos* y *corpus comparables*. Un corpus comparable es un conjunto de textos similares en más de una variedad lingüística; la cuestión que determina la similitud aún es objeto de debate pero usualmente contienen textos producidos en circunstancias similares de comunicación [7]. Un corpus paralelo contiene textos que han sido traducidos en una o más variedades lingüísticas [8] ; este tipo de corpus suele ser útil para los estudios traductológicos y la traducción automática.

[5] Es de notar que, en la tipología del Grupo EAGLES (Sinclair, 1996), cualquier corpus que implique la intervención del lingüista o del investigador en las interacciones diarias de los hablantes se puede considerar un *corpus especial*. En este sentido, los corpus orales se pueden clasificar de dos (2) formas: 1) los *corpus orales*, los cuales contienen muestras de tipo experimental (i.e. especial) y son para el uso del gremio de fonetistas e ingenieros lingüísticos interesados en el síntesis de habla y 2) los *corpus de lengua oral*, los cuales pueden o no ser especiales (ya que no implican por necesidad la intervención de un investigador) y cuyo uso es mayoritariamente para el gremio de lingüistas interesados en la descripción gramatical y léxica de muestras reales y la documentación.

[6] Es de notar que los textos de un corpus monolingüe son de producción mayoritaria en una lengua. Lo que queremos decir con esto es que, si los textos de un corpus están internamente producidos en una sola lengua, e.g., en inglés, y contiene citas, fragmentos, préstamos, alternancia, etc., en otra lengua, como el español, aún así, no se puede dejar de considerar como un corpus monolingüe, puesto que no son textos *comparables*. Sin embargo, esta noción se vuelve problemática al considerar ciertas situaciones sociolingüísticas, como se indica en 5.2.3.

[7] Consíderese el ICE. Cf. 2.2

[8] Un ejemplo de un corpus paralelo se puede observar en el sitio Web Linguee (https://www.linguee.com/). En la página, se pueden realizar búsquedas de un término o una frase específica en una lengua y las equivalencias de estas se presentan en otra lengua. La búsqueda asimismo muestra la concordancia (Cf. 5.3.4) de las búsquedas y sus equivalencias en pequeñas citas recopiladas de otros sitios o documentos en línea, los cuales se han traducido desde o hacia la lengua meta. En la concordancia, las palabras claves se encuentran resaltadas.

La clase de corpus que nos propongamos recolectar depende en gran medida de su proceso de construcción, por lo cual definiremos la tipología del corpus en la sección 5.2.3.

4.1.2. La anotación de corpus

Los corpus pueden existir de dos (2) formas: en su estado original sin ningún tipo de tratamiento (i.e. *en bruto*) o mejorados con algún tipo de información lingüística (i.e. *anotados*) (McEnery y Wilson, 2001, p. 32).

La *anotación* es la práctica de añadir infomación *lingüística e interpretativa* a un corpus en bruto y así enriquecerlo para su explotación posterior (Leech, 1997, p. 2).

Es de notar que tanto los corpus en bruto como los corpus anotados son útiles para la investigación lingüística y el uso de cualquiera puede dar resultados importantes. La ventaja de los corpus anotados, sin embargo, es que los análisis que se realizan sobre una cantidad reducida de texto se pueden realizar de manera sistemática sobre un número ingente de datos.

Las ventajas que presenta trabajar con corpus anotados son notables, puesto que este tipo de recurso permite que se extraiga información del corpus de forma más eficaz que de un corpus en bruto y, por los elementos analíticos explícitos en el corpus, puede beneficiar su uso posterior por parte de investigadores interesados en el corpus ahorrándoles tiempo de análisis, así sea con fines distintos al de los recopiladores que hicieron la anotación. Por lo tanto, podemos decir que los corpus anotados son *multifuncionales y reutilizables* (McEnery y Hardie, 2012).

Hay diferentes tipos de anotación (Villayande Llamazares, 2010): gramatical, semántica, léxica, etc. La forma más común, y en la que nos enfocamos, es la anotación gramatical. En la anotación gramatical, una *etiqueta* se asocia con un *token* por medio de diferentes recursos textuales (e.g. un guion bajo [9]) e indica la categoría gramatical de este. La ventaja que presenta frente a otros tipos de anotación es su versatilidad: a partir de un número limitado de etiquetas, se puede hacer la anotación de cada uno de los *tokens* de un corpus, lo cual sirve para distinguir entre homógrafos y posteriormente hacer análisis y anotación sintáctica y semántica. Además, es la anotación más fácil de automatizar.

Hay tres (3) maneras de hacer anotación gramatical: 1) de forma totalmente automática, 2) de forma automática con correción manual posterior y 3) de forma totalmente manual (McEnery y Hardie, 2012). La forma 1) es la que permite un mayor nivel de procesamiento de los datos, ya que logra anotar una cantidad enorme de datos de forma muy eficaz y con un margen mínimo de error. Sin embargo, en ocasiones, por falta de herramientas informáticas, se debe optar por la opción 3), la cual, aunque posiblemente tenga menor margen de error, es lenta y costosa (McEnery y Hardie, 2012).

[9] E.g, **desde_PREP**, donde **PREP** es una etiqueta que indica la clase gramatical *preposición* que se asocia al *token* **desde**.

Por las posibilidades que ofrece la anotación gramatical, nos parece de gran utilidad para nuestro piloto, pese a la falta de recursos informáticos para el *creole*. Por otro lado, para la documentación lingüística puede ser útil, puesto que, como indica Woodbury (2011, p. 183), no es común entre los lingüistas documentales añadir anotación gramatical a los productos de una documentación, a pesar de ser esta una de las metas típicas en un proyecto documental.

4.2. La Documentación Lingüística

La recopilación de corpus es parte esencial de la labor de la documentación lingüística, por lo que es de interés revisar algunos de los postulados teóricos de esta.

La documentación lingüística se puede definir como un campo de acción que se encarga principalmente de la "creación, anotación, preservación y diseminación de registros trasparentes de una lengua" (Woodbury, 2011, p. 159) [10], especialmente de datos primarios de ella.

El objetivo para una documentación es que sea *duradera* y *versátil*, duradera puesto que apunta a ser de utilidad para usuarios futuros a largo plazo y versátil puesto que apunta a beneficiar a un espectro amplio de usuarios, tales como la misma comunidad de habla e investigadores de diferentes disciplinas, más allá de los objetivos específicos que pueda tener un proyecto documental (Himmelmann, 2006, pp. 1–2).

Puesto que un proyecto de documentación ha de ser versátil, debe ser exhaustivo y recolectar la mayor cantidad de datos lingüísticos que pueda. Sin embargo, la mayor parte del tiempo, esta meta es inalcanzable, por razones tanto materiales como éticas. Por ello, la recolección de datos en un proyecto debe ser sistemática y teorizada de tal forma que responda a las necesidades específicas del proyecto e incluya tantos aspectos de una lengua como estén al alcance.

En la perspectiva de las ciencias del lenguaje, hay tres (3) razones fundamentales (aunque no únicas) para documentar una lengua, según Himmelmann (2006, pp. 3–5):

1. debido al creciente número de lenguas en peligro en el mundo, es necesario empezar a documentarlas, no solamente para la investigación lingüística, sino porque también esto es un elemento de interés para el mantenimiento lingüístico y puede apoyar los esfuerzos de recuperación de una lengua;

2. la "economía" investigativa y los recursos investigativos irían en aumento, i.e., si se decidiera comenzar a trabajar sobre una lengua ahora y más adelante en el tiempo otro investigador o grupo de investigadores decidiera continuar ese trabajo, se beneficiaría más de datos primarios de la lengua disponibles en un repositorio que de algunos pocos textos y fragmentos

[10] Traducción propia.

disponibles [11] , así como el trabajo que resultase del segundo, aunque cumpliera unos propósitos específicos, beneficiaría a otros investigadores o grupos posteriores y así sucesivamente, haciendo que de esa manera una lengua poco documentada, por vía aditiva, obtenga una documentación amplia que fortalezca las disciplinas que lleven a cabo su estudio;

3. permitiría que el análisis lingüístico sea *verificable*, i.e., cualquier afirmación o estudio disciplinar que se haga sobre una lengua puede ser objeto de escrutinio.

Para entender mejor la conceptualización de la documentación lingüística, los registros que constituyen los productos de una documentación suelen fijarse por medio de una representación lexicogramatical de forma, como la *escritura*, o por un medio que preserve los datos de habla, como las *grabaciones* (de audio y/o de video); para ser transparentes, los registros deben contener lo que se conoce en filología como un *aparato* (en adelante, *metadatos*) que registre de manera sistemática la información sobre el origen del registo y el evento que representa, acompañados por una traducción del registro en una lengua de comunicación amplia, generalmente en forma de anotación (Woodbury, 2011, p. 160).

El conjunto de registros sobre una lengua se conoce como un *corpus documental de lengua* (o simplemente *corpus* [12]) y el conjunto de ideas que guían la recolección de un corpus documental de lengua se conoce como la *teorización del corpus* (Woodbury, 2011, p. 161).

Encontramos aquí un punto de encuentro entre la lingüística de corpus y la documentación lingüística: a medida que avanzan las innovaciones informáticas, la lingüística de corpus puede informar las metodologías de la documentación lingüística; asimismo, la documentación lingüística puede informar los propósitos y la teoría que acompañan la recopilación de un corpus. No se debe pensar, sin embargo, que hay un hilo conector constante entre ambas disciplinas.

El enfoque empírico de la documentación lingüística, el cual comparte con la lingüística de corpus, es que conocer el uso es conocer la lengua, por lo cual abroga por la creación de "grandes cantidades de textos" (Boas, 1917, citado en Woodbury, 2011, p. 163) [13] para estudiar todos los usos de lengua posibles.

En la documentación lingüística, hay dos (2) tipos de datos que son de interés: los *datos primarios* y el *conocimiento metalingüístico* (Himmelmann, 2006, pp. 7–11).

Los datos primarios son instancias del *comportamiento lingüístico observable*, lo que incluye cómo las personas hablan entre ellas en su vida cotidiana. Es deseable documentar todos los tipos de situaciones comunicativas como sea posible, pero en muchas ocasiones por cuestiones pragmáticas

[11] Lo cual, como mencionamos en 2 y 2.1, es la situación general del *creole*.

[12] Sin embargo, mantenemos el término *corpus documental de lengua*, el cual se compone de *todo* el material general sobre una lengua incluyendo diccionarios y gramáticas, para distinguirlo de la taxonomía de los corpus presentados en 4.1.1.

[13] Traducción propia.

solo se puede tomar una muestra de ellas.

El conocimiento metalingüístico no es un tipo de dato que se pueda observar con tanta facilidad como los datos primarios. Con este término se entiende la "habilidad de los hablantes nativos de facilitar interpretaciones y sistematizaciones de unidades y eventos lingüísticos" (Himmelmann, 2006, p. 8) [14]. Puesto que el conocimiento metalingüístico no aparece con tanta frecuencia en eventos comunicativos cotidianos como los datos primarios, suelen usarse métodos de *elicitación lingüística* para obtenerlo.

Es de notar que, en este sentido, los diccionarios y las gramáticas no son datos primarios ni contienen el conocimiento metalingüístico de los hablantes, sino que son *formatos analíticos* o *recursos descriptivos* [15] que formalizan el conocimiento metalingüístico. Estos, sin embargo, sí constituyen una clase de datos dentro de la documentación lingüística.

Por otro lado, los *metadatos* han de acompañar cada registro y son la documentación en sí *acerca* de los datos primarios: las personas que participan en la captación de datos, el momento en que se recolectaron, dónde y por cuánto tiempo, entre otras cuestiones que ayudan a identificar cada registro de forma individual para su manejo futuro. Esta documentación facilita que las personas que accedan al corpus de lengua documental puedan rastrear los registros de forma rápida y sencilla.

Los datos primarios suelen clasificarse en grabaciones y transcripciones (Good, 2011, pp. 214–216), pero se debe aclarar que las transcripciones, aunque son en efecto una representación de los eventos comunicativos, implican un análisis lingüístico extenso, por lo que las transcripciones suelen pertenecer más al ámbito de la descripción (de los eventos comunicativos) y las grabaciones más al ámbito de la documentación.

Puesto que se espera que un corpus de lengua documental pueda ser de acceso general, es imperativo no confundir los datos ni la estructura de los datos con la presentación de los datos. Para asegurar que los datos se puedan presentar ante la comunidad académica y la comunidad de habla, los datos se debe *codificar* de manera adeucada. Cada una de estas formas de representar los datos se conoce como una *implementación* y depende en gran medida del próposito para el que se representa la información. Esto es importante para la elección de las herramientas utilizadas para la documentación y el formato que se escoge para el almacenamiento, ya que, si se concentra excesivamente en la presentación de los datos más que en establecer la relación entre ellos, esto puede dificultar la *reutilización* de los datos, una de las máximas de la documentación lingüística, lo mismo que si no se escoge un formato adeucado para la distibución, el acceso y el manejo de los datos.

Por último, es indispensable pensar *desde el comienzo* de un proyecto de documentación en cómo se archivarán los datos para permitir su acceso y cómo se publicarán y distribuirán los materiales que resulten del proyecto.

[14] Traducción propia.
[15] También: *recursos analíticos*.

En la documentación lingüística el énfasis se coloca sobre los datos lingüísticos y no tanto sobre los productos secundarios que surgen de ellos. Por ello, aunque el interés primordial de nuestro proyecto es desarrollar materiales de referencia, consideramos que los postulados de esta disciplina son un insumo importante para nuestro trabajo. En primer lugar, porque es a partir de los datos que se pueden crear estos materiales; en segundo lugar, porque es también de nuestro interés documentar la lengua *creole*, tanto para su preservación como para su estudio posterior.

4.3. La Criollística

Es importante aquí hacer algunas aclaraciones conceptuales acerca de las lenguas criollas, para luego proceder a enunciar los posibles retos que pueden presentar frente a la documentación y su estudio por medio de la lingüística de corpus.

Para entender lo qué es una lengua criolla, es necesario primero entender qué es una lengua pidgin. Una *lengua pidgin* es en esencia "una lengua reducida que resulta del contacto prolongado entre grupos de personas que no hablan una lengua común" (Holm, 1988, p. 5) [16] . Los grupos en contacto suelen adaptar rasgos de sus propias lenguas y de las lenguas con las que se encuentran en contacto. Estos códigos de comunicación suelen perder cualquier tipo de inflexión morfológica que puedan tener las lenguas en contacto, por lo cual se pueden considerar *lenguas analíticas*. Las lenguas pidgin pueden ganar estabilidad social (i.e. se usan de forma estable entre diferentes comunidades con propósitos específicos) y entonces se las puede considerar un *pidgin extendido*.

Una *lengua criolla* deriva de una lengua pidgin. La principal diferencia con respecto a estas es que la lengua criolla pasa por un proceso de *nativización*; es decir, los niños comienzan a aprender la lengua pidgin como lengua materna hasta el punto de expandir sus funciones. Con el paso del tiempo, el sistema de la lengua se vuelve más o menos estable. En este punto, se vuelve una lengua criolla.

Pese a que las lenguas criollas no poseen una filiación genética con su lengua lexificadora (generalmente la lengua de *superestrato*), se las suele clasificar de acuerdo a esta (García León, 2011). Por ello, en lo que se refiere al *creole* de San Andrés, se suele decir que es una lengua criolla de base léxica inglesa. Estos criollos surgieron en gran parte en la región caribeña y tienen relación entre sí; justamente el *creole* se encuentra estrechamente emparentado con otras variedades criollas de base léxica inglesa habladas en Jamaica (en la cual encuentra su origen), Bocas del Toro (Panamá), Bluefields y Corn Islands (Nicaragua). (Patiño Rosselli, 2000). De hecho, se afirma que sus comunidades de habla respectivas conformaron "una comunidad de habla antes de que fueran separad[a]s por las fronteras nacionales de los países a los que pertenecen sus hablantes" (García León, 2014, p. 202).

[16] Traducción propia.

Los posibles retos de las lenguas criollas frente a los corpus y la documentación

Ya se advirtió en 2.3 que pocos trabajos se han hecho con respecto a la recopilación de corpus electrónicos de lenguas criollas y que apenas se han explorado algunas problemáticas relacionadas a su construcción.

Nos referimos aquí en particular al trabajo de compilación del *Corpus of Written British Creole English* y el *Corpus of Written Jamaican English* (Sebba y Dray, 2007), experiencias las cuales ayudan a informar nuestro trabajo.

En primer lugar, es de notar la enorme variabilidad que se encuentra al interior de las lenguas criollas. En las situaciones en las que la lengua criolla está en contacto con la lengua lexificadora, tiende a existir lo que se denomina un *continuo poscriollo*: la variedad "más criolla" (o *basilecto*) forma junto con la lengua estándar (o *acrolecto*) un continuo en el que se ubican los diferentes ideolectos de los hablantes (o *mesolectos*), los cuales contienen formas intermedias entre los dos (Sebba y Dray, 2007, pp. 210–212). Esta variabilidad afecta principalmente cuestiones de selección, transcripción y anotación.

En primer lugar, la selección del corpus puede ser una tarea compleja, puesto que plantearse recolectar un corpus de una lengua implica determinar *qué* pertenece a la lengua criolla y qué no, especialmente en situaciones en las que suele haber alternancia de código en el continuo. En San Andrés, no hay un contacto tan estrecho con el inglés como a principios del siglo XX y de hecho se afirma que no sería apropiado hablar de un continuo hoy en día (Bartens, 2002), pero aún hay huellas de dicho continuo presentes en la actualidad y esto puede afectar la inclusión de muestras en el corpus.

El mayor contacto es con el español y es factible lograr distinguir las formas de una y otra lengua; sin embargo, ello complica otra parte de la selección de textos, ya que se debe decidir sobre si se incluyen aquellas interacciones que contienen alternancia con el español o no. En caso de que no, se debe ser cuidadoso en hacer un *muestreo* de los textos que contienen casi exclusivamente instancias de *creole*.

En segundo lugar, la variabilidad de la lengua se puede ver reflejada en la ortografía. No se ha estandarizado el *creole* y, aunque existe una ortografía propuesta (Cf. Anexo **A-1**), no se ha difundido masivamente (Moya, 2006). Por otro lado, aunque se lograra implementar la ortografía para un proyecto, la variabilidad se complejiza con aquellas formas que se pueden considerar tanto criollas como estándar y asimismo con aquellas que se consideran más estándar que criollas. Como se mencionó anteriormente en 4.2, la transcripción requiere un nivel de análisis previo, por lo cual en varias ocasiones recaerá sobre la persona que transcribe determinar si una forma es de una lengua o de otra.

Esto también afecta la relación *type-token* (Cf. 2.3), puesto que, para un *type* que puede ser tanto

criollo como estándar, se debe determinar si cuenta como un *type* separado en cada lengua o si se cuenta como uno solo. Por otro lado, puesto que el *creole* no está estandarizado, en varios casos resulta vago determinar qué cuenta como un *type* en la transcripción.

Por último, la anotación gramatical se complica con respecto a la recopilación de un corpus del *creole*. Por un lado, no se disponen de recursos para hacer una anotación automática de los textos, razón por la cual toda la anotación debe ser, en un principio, manual. Por otro lado, debido a la variabilidad mencionada arriba y lo problemático de la relación *type-token*, no se logra determinar con precisión si la anotación se debe hacer para un solo código o para varios.

Planteadas ya las posibles dificultades que supone la recopilación de un corpus de lengua criolla, intentaremos responder a cada una.

En la documentación lingüística, se anima a que un corpus sea lo más diverso posible y que contenga todas las interacciones comunicativas de una comunidad de habla como sea posible. El interés en este campo no es principalmente hacia la descripción, sino hacia la documentación de datos lingüísticos. Incluso hay quiénes plantean que la meta de un corpus de lengua documental debe estar basado en una *comunidad de habla* y no necesariamente en un único código (Himmelmann, 1996, citado en Woodbury, 2011, p. 169). Por ello, consideramos que, si el interés de un proyecto es en la documentación de una lengua, debe apuntar a recolectar la mayor cantidad de textos, independientemente de que haya variación o no. Incluso, si se presenta variación, esto puede ser bastante útil para una descripción de las prácticas comunicativas de la comunidad.

La cuestión acerca de determinar qué es *creole*, y qué no, se puede resolver a partir del mismo estudio de un corpus. En Sebba y Dray (2007), los investigadores se enfrentaron al problema de determinar qué formas pertenecían al inglés y qué formas al criollo jamaiquino; sin embargo, al final, pese a las limitaciones del corpus, lograron obtener una perspectiva más comprensiva acerca de la lengua que cuando comenzaron y lograron identificar mejor cuáles formas eran criollas y cuáles estándar. Por ese mismo principio, podemos afirmar que, luego de recopilar y estudiar un corpus, podremos obtener una perspectiva más informada e íntegra acerca de las estructuras del *creole*.

En cuanto a la anotación gramatical, por la falta de recursos informáticos, es cierto que, al menos en un principio, debe ser manual. Sin embargo, con una cantidad suficiente de textos anotados, se torna posible automatizar la tarea por medio de un anotador gramatical. Es de notar que hay dos (2) formas de programar un anotador (Villayande Llamazares, 2010): 1) por medio de reglas y 2) por medio de probabilidades estocásticas. La opción 1) requiere un conocimiento profundo de la gramática de una lengua; sin embargo, como se mencionó en la Introducción, no hay una gramática lo suficientemente amplia del *creole* para lograr esto. Por lo tanto, la opción 2) es la más adecuada en estos momentos. Luego de anotar una cantidad suficiente de textos en *creole*, se puede determinar en qué momento es factible entrenar un anotador de forma probabilística con base en el corpus ya anotado. Pero para ello primero se debe hacer la labor manual.

5. Marco Metodológico

En la documentación lingüística es importante el contacto continuo y la participación activa de miembros de la comunidad de habla en el proyecto de documentación en cada una de sus etapas: formulación de los propósitos, recolección de datos, transcripción, etc. En la tradición documental, la teorización del corpus siempre ha tenido en cuenta a la comunidad en el *diseño de proyectos* (Woodbury, 2011) y siempre se ha propuesto entrenar a hablantes nativos como documentadores (Cf. 4.2).

Por ello, en la formulación de la metodología, hubo un constante interés en involucrar a hablantes nativos de la lengua y en entrenarlos para que hicieran parte íntegra del proyecto y asimismo se formaran como documentadores.

Las principales preguntas que surgen en una teorización del corpus y el diseño de proyectos tienen que ver con (Woodbury, 2011, p. 162):

- los participantes
- los propósitos
- los agentes interesados

Cabe mencionar que no profundizamos sobremanera en la formulación de estos aspectos del proyecto de documentación, puesto que aún es un trabajo en curso. Los tomamos en cuenta en la medida en que son pertinentes para el piloto, a partir del cual se deben seguir buscando caminos posteriores para continuar con un proyecto más grande y más ampllio.

En 5.1, procedemos a elaborar sobre estos aspectos del proyecto; en 5.2, exponemos cuáles fueron los mayores factores que se tomaron en cuenta para diseñar el corpus y en 5.3, mostramos cuáles fueron los instrumentos principales y las herramientas informáticas que se utilizaron para llevar a cabo el piloto en el semestre 2018-I.

5.1. Teorización del corpus y diseño del proyecto

5.1.1. Los agentes interesados

Nos pareció importante, desde el punto de vista documental, hacer un acercamiento a la comunidad raizal de San Andrés y escuchar sus opiniones, así como preguntarles cuáles sienten que son las mayores necesidades al interior de su comunidad con respecto a su lengua. Esto surge de nuestra

concepción de que, aunque desde la academia se identificó una serie de carencias (Cf. 2), son los miembros de la comunidad los principales interesados en cualquier proyecto de documentación y para quienes se formula.

Por otro lado, nuestro interés en acercarnos a la comunidad también surge de la necesidad de involucrar a diferentes personas al equipo de trabajo, ya que, por el momento somos pocos participantes y de un número contado de disciplinas. El éxito de muchos proyectos de esta índole reside precisamente en que se componen de individuos de diferentes trasfondos (Woodbury, 2011, p. 176). Un proyecto documental de estas dimensiones debe ser inclusivo.

Quisimos exponer los beneficios que puede presentar un proyecto de este tipo y las maneras en qué puede beneficiar a la comunidad un registro sistemático de su lengua.

Hay una diversidad de opiniones en el interior de la comunidad con respecto a la documentación y el estudio descriptivo del *creole*, algunas de oposición y algunas de apoyo. Aquí hacemos un recuento de ellas:

- el *creole* no es digno de estudiarse ya que es una lengua desprestigiada: esta posición suele apoyarse, aunque no necesariamente, en la noción de que la lengua es una variante degenerada del inglés y que, por oportunidades económicas, es mejor solo estudiar inglés y/o español;
- el *creole* es meritorio de ser estudiado y de ser valorado, pero no por parte de instituciones estatales como la Universidad Nacional, sino únicamente por miembros de la comunidad;
- el *creole* es meritorio de ser estudiado y de ser valorado y el aporte de instituciones estatales como la Universidad Nacional que aúnan fuerzas con la comunidad puede ser valioso para esta labor.

El panorama realmente presenta muchas más matices que las que enumeramos aquí. En general, podemos decir que no se ha logrado un acercamiento demasiado cercano con la mayor parte de la comunidad, pero sí con una porción significativa (los que son más cercanos a la tercera posición), quienes muestran interés en apoyar el proyecto. Los cercanos a la primera y la segunda posición, aunque tienen posturas contrarias, se muestran medianamente permisivos con respecto a las acciones investigativas de la Universidad, así no las apoyen por completo.

5.1.2. Los participantes

Al proyecto, desde la Universidad Nacional, están vinculados un (1) estudiante de lingüística (el autor del presente trabajo) y la profesora asociada Raquel Sanmiguel, quien apoyó el contacto inicial con la comunidad y gracias a quien se logró citar a los participantes del piloto.

En cuanto a los participantes miembros de la comunidad con quien se realizó el piloto, inicialmente se contactó a ocho (8) personas, las cuales todas afirmaron tener interés en el proyecto y se mostraron dispuestas a fungir como colaboradores para recolectar datos. Estas personas son todas

educadoras y tienen interés en las acciones que se realicen con respecto al *creole*. De estas personas, quedaron cinco (5) quienes asistieron a las primeras sesiones de formación.

Por cuestiones de tiempo, algunos colaboradores del grupo con el que inicialmente se planteó el piloto no pudieron seguir asistiendo a las sesiones y, al final, este se realizó con tres (3) colaboradoras, quienes estuvieron presentes de manera constante durante el desarrollo de las sesiones de formación y la recolección de datos.

Las integrantes del grupo de colaboradoras con quienes llevamos a cabo el piloto fueron:

- Maureen Hooker
- Penny Bryan
- Emma Forbes

5.1.3. El propósito

Es importante tener en cuenta que los propósitos de un proyecto documental pueden ser diversos: la preservación, la revitalización o incluso el estudio descriptivo de una lengua (Woodbury, 2011, p. 162).

Para nosotros, desde la academia, se reconoció la necesidad de adelantar estudios descriptivos de la lengua (Cf. 2). Fue evidente en la revisión de los estudios previos que hay varias áreas en las que se deben profundizar. Pero, para este piloto, decidimos enfocarnos en la documentación de ciertos eventos comunicativos más que en estudiar un fenómeno concreto. Esto fue con el propósito de conocer cómo se pueden explotar las herramientas de corpus para llevar a cabo el piloto y de qué forma se puede formular el proyecto de documentación en el futuro. Esto también está acorde con el espíritu de la documentación lingüística, según el cual los recursos descriptivos son productos que surgen *de* los datos primarios; por ende una documentación no debería apuntar exclusivamente *hacia* la creación de estos recursos.

Con las participantes del proyecto, se discutieron los posibles propósitos del piloto. Para ellas, lo esencial era estudiar cómo las personas utilizan la lengua hoy en día en una situación de contacto con el inglés y el español, así como conocer las prácticas comunicativas de personas mayores. Esto nos pareció propicio para comenzar a estructurar la recolección del corpus.

5.2. Diseño del corpus

En el diseño del corpus que se recolectaría para el piloto, se tomaron en cuenta ciertos criterios de clasificación interna y externa. Procedemos a enumerar los criterios que guiaron la construcción del corpus y los motivos que justificaron estas decisiones (Sinclair, 1996; Villayande Llamazares, 2010).

5.2.1. Criterios externos

En primer lugar, se tomaron en cuenta aquellos criterios que no están relacionados con el contenido lingüístico que contendrían los textos, sino que refieren al estado material de los textos que formarían parte del corpus.

Origen

Nos referimos aquí a la persona quien produce un texto y datos relevantes acerca de ella. Para el piloto, se decidió que el corpus contendría textos producidos por personas entre los cuarenta y cinco (45) y los sesenta y cinco (65) años de edad, sin distinguir entre sexo (i.e. podían ser tanto hombres como mujeres), cuyo trasfondo lingüístico incluyera el *creole* como lengua nativa o segunda lengua, que hubiera vivido en San Andrés de forma continua gran parte de su niñez y su adolescencia (no más de seis (6) años por fuera en total) y que estuviera residiendo en la isla de forma continua durante los últimos dos (2) años.

Estado

Este criterio se refiere al estado físico de los textos y asimismo al soporte de los mismos, i.e. si los textos son de soporte oral o escrito. La decisión aquí fue no incluir textos escritos, porque hay muy pocos textos escritos en lengua *creole* y no son muy diversos (Ramírez-Cruz, 2017, pp. 118–121): hay diecinueve (19) libros de textos para primaria que fueron publicados por la Universidad Cristiana y algunos textos religiosos traducidos al *creole* (incluyendo la totalidad del Nuevo Testamento); otros textos disponibles son algunas antologías de poetas locales, revistas, ensayos y cuentos para niños, las cuales tampoco suman una cantidad considerable. Además, por cuestiones de tiempo, se determinó que no era factible conseguir las respectivas licencias y permisos de estos textos para incluirlos dentro de un corpus. Por ello, se decidió que los textos fueran primordialmente de soporte oral, con el debido consentimiento de los hablantes que los produjeran.

Objetivos

El objetivo se refiere al fin con el que la persona produce el texto y la audiencia a la que quiere llegar. En nuestro caso, la audiencia de cada texto sería la colaboradora que lo recolectaría (i.e. la persona que entrevistaría al hablante) y el fin con el que el entrevistado produciría el texto sería para contarle a la colaboradora detalles acerca de su vida temprana.

5.2.2. Criterios internos

Con criterios internos nos referimos a criterios basados en la organización lingüística interna de los textos. El énfasis aquí es en la estructura textual como tal y no tanto en fenómenos a nivel de palabra o de oración.

Tema

Este criterio responde al ámbito o dominio al que pertenece un texto. Es un criterio complejo de definir, ya que implica la clasificación de los diferentes sentidos de los textos en áreas del conocimiento.

Baste decir aquí que, para el tema de los textos, se decidió que serían de temática personal, que tocarían elementos acerca de la crianza de las personas entrevistadas durante su infancia en un espacio determinado: *di yaad*, el patio, un elemento central en la vida social de San Andrés.

Estilo

El estilo es el modo de lengua de los textos. En nuestro caso, por la naturaleza del tema y del método de recolección, podemos decir que el estilo de los textos sería informal y privado, ya que se producirían en entrevistas semidirigidas con las colaboradoras.

5.2.3. La clasificación del corpus

Con lo anterior en mente, podemos comenzar a clasificar nuestro corpus de acuerdo a la tipología expuesta en 4.1.1.

En cuanto a la modalidad de nuestro corpus, podemos decir que sería un *corpus oral*, ya que, aunque posteriormente se realizaría una trascripción de los textos recolectados, el soporte y el medio originario de los textos sería primordialmente el habla. Adicionalmente, aunque se componga de textos de entrevistas semidirigidas, estos no serían de tipo experimental, por lo cual lo consideramos un *corpus de lengua oral* (Cf. Nota al pie 4.5).

Por el tamaño y la especificidad de los textos, no se podría decir que nuestro corpus oral sea general ni especializado. Esto es debido a que no se puede decir que el corpus sea representativo de una variedad lingüística. Por cuestiones de tiempo, las muestras que se planearon tomar serían cortas y pocas, por lo que no sería general. Por otro lado, como no nos enfocamos a describir un sublenguaje específico ni una jerga, tampoco sería un corpus especializado. En este punto, sin embargo, afirmamos que el corpus recolectado sería un *corpus especial* (Cf. Nota al pie 4.5), puesto que, por la naturaleza de los textos, habría una intervención explícita de las colaboradoras, las cuales guiarían hasta cierto punto el curso de los textos.

Con relación al número de lenguas, se esperaba que el corpus fuera monolingüe; es decir, que solo contendría textos en *creole*. No descartamos los textos que puedan contener alternancia en español, ya que ello puede ser un insumo importante para la investigación; en ese caso, seguiría siendo un corpus monolingüe. Un posible punto de tensión sería si un texto en particular tiene contenido acrolectal (Cf. 4.3) y es demasiado cercano al inglés. Frente a un texto de tal naturaleza, si es que surgiera, nuestra recomendación fue incluirlo en el corpus, ya que podría darnos una perspectiva acerca de la variación lingüística del *creole* en la comunidad de habla raizal. Lo importante es recopilar y estudiar el corpus, no esforzarse en que adhiera a una tipología determinada.

Por esa misma razón, sería apresurado afirmar que nuestro corpus es *comparable* o no, ya que solo los textos que resulten del piloto nos permitirán afirmar si contienen más de una variedad lingüística.

Lo anterior nos permite afirmar que el corpus que nos proponemos recolectar es un *corpus especial de lengua oral*.

5.3. Instrumentos para el piloto

Hechas ya las precisiones acerca de las características del corpus, procedemos a indicar las fases en las que se llevaría a cabo el piloto, las cuales podemos dividir en cuatro:

1. Recolección de datos lingüísticos
2. Transcripción de los datos
3. Anotación de los datos
4. Análisis preliminar de los datos

Procedemos a enumerar los instrumentos metodológicos usadas en cada fase.

5.3.1. Recolección de datos lingüísticos

En esta fase, cada una de las colaboradoras contactaría a una persona que cumpliera con las características personales que se mencionaron en 5.2.1.

Una vez hecho el contacto con estos participantes, procedierían a concertar alguna cita con ellos para hacer la entrevista semidirigida y asimismo grabar la interacción en formato digital. Luego tranferirían el archivo de audio resultante a una computadora y anotarían los metadatos pertinentes.

Equipo de grabación de audio

Para la recolección de los datos, se buscó que los archivos de audio se tomaran con equipos de grabación digitales y portátiles. Para una calidad óptima, sugerimos conseguir equipos con las siguientes especificaciones:

- Formato de grabación: sin pérdidas, preferiblemente WAV
- Respuesta de frecuencia: 80–11 000 Hz
- Frecuencia de muestreo: 44 100 Hz / 22 050 Hz
- Profundidad de bit: 16–24 bits
- Relación señal/ruido: 80dB
- Micrófono omnidireccional

Por costo, se sugirió que el equipo más adecuado para nuestros propósitos era el Sony ICD-PX470 [1]. El equipo cumple con todos los requerimientos, excepto por el del micrófono: es estéreo y no omnidireccional; sin embargo, es posible adquirir un micrófono omnidireccional como accesorio, el cual se puede conectar a la grabadora y, si es un micrófono de solapa, se podría reducir el nivel de intrusión que podría experimentar el hablante. Asimismo, el equipo posee un conector USB propio, lo cual permite la fácil transferencia de datos a un computador.

Por cuestiones administrativas, sin embargo, no fue posible conseguir este equipo. Entonces, cada colaboradora decidió utilizar un teléfono celular para tomar los datos. A continuación relacionamos los equipos utilizados y sus respectivas versiones de sistema operativo:

- Un (1) Alcatel Touch Pixi, Modelo 4003A – Android 4.4.2
- Un (1) Sony Xperia Z5, Modelo E6603 – Android 7.1.1
- Un (1) iPhone 6, Modelo MG3A2CL/A – iOS 11.0.3

Entrevistas

Para cada sesión, se creó un formato de consentimiento (Cf. Anexo D). La forma en que se decidió registrar que el entrevistado daba su consentimiento para ser grabado era por medio de la grabación misma. Esto se decidió con el propósito de que, si el entrevistado tuviera alguna objeción o alguna solicitud específica con respecto al proyecto, pudiera expresarlas ante el micrófono y así quedaran grabadas. Este formato se decidió escribir en inglés para ser más cercano a la lengua meta.

No se realizó un cuestionario para la entrevista, sino que se decidió que cada colaboradora fuera realizando las preguntas acorde al tema a medida que fuera avanzando la entrevista, de manera que, si el entrevistado se desviara por algún punto relacionado con la vida en el patio, la colaboradora tendría la posibilidad de enfocarse por esta vía para así no limitarse a una sola temática.

Para cada entrevista, se acordó que el tiempo aproximado debía ser de treinta (30:00) minutos.

Metadatos

Es necesario llevar un registro estructurado de todas las sesiones que se realicen en un proyecto, esto con el propósito de que su organización y su búsqueda posterior se pueda hacer de forma manejable. Si no se hace esto, el aprovechamiento futuro de los datos resultaría caótico.

Por ello, se estableció que cada sesión, y los productos derivados de ella (archivo de audio, transcripción, formato de metadatos), tendría un nombre que codificaría cierta información sobre la misma: en el nombre de la sesión primero se codificaría el barrio en el que tomó lugar con una

[1] https://www.sony.com.co/electronics/grabadores-voz/icd-px470

abreviación (Cf. Anexo B); segundo, se colocaría la edad del entrevistado en números arábigos [2] ; tercero, se codificaría el sexo del hablante por medio de ⟨m⟩, para masculino, y ⟨f⟩, para femenino; por último, a esto se le agrega un guion bajo (⟨_⟩) y se codifica, con números abábigos en el formato DDMMYY, la fecha en que tomó lugar la sesión. Entonces, e.g., si una sesión tomó lugar en San Luis con un hablante masculino de veintitrés (23) años el 7 de marzo de 2018, el nombre del archivo de audio resultante se codificaría de la siguiente manera:

sl23m_070318.wav

Para complementar las sesiones, también se decidió crear un formato de metadatos que rellenaría cada colaboradora. Este formato tiene campos para el nombre de los hablantes, su procedencia, su trasfondo lingüístico y otras cuestiones que podrían resultar de interés para estudios variacionales (Cf. Anexo 5.3.1).

5.3.2. Transcripción de los datos lingüísticos

Hay diferentes propuestas con respecto a la transcripción de corpus de lengua oral. Para el Grupo EAGLES, lo recomendado es trabajar en cuatro (4) niveles (Llisterri, 1996):

- **Nivel I** — representación ortográfica con puntuación mínima y sin información interaccional,
- **Nivel II** — representación ortográfica mejorada con información básica sobre la identidad de los hablantes, turnos de habla y elementos no verbales,
- **Nivel III** — contiene la misma información que el Nivel II, con marcación en las fronteras tonales y otra información fonética,
- **Nivel IV** — contiene la misma información que el Nivel III, a la que se agrega codificación adicional sobre entonación e información acústica.

Los dos (2) últimos niveles deben ser transcritos por fonetistas profesionales.

Por las limitaciones de personal (no hay un fonetista profesional en el equipo de trabajo), se decidió no hacer una transcripción en los Niveles III y IV [3].

Por otro lado, a partir de las necesidades del proyecto de documentación, se debió establecer qué hay que transcribir y de qué forma, puesto que no todas las propuestas de transcripción son útiles para todos los proyectos (Adolphs y Knight, 2010).

[2] Si por algún motivo se desconociera la edad del hablante, entonces se colocaría el múltiplo de diez del rango de edad (aproximado) junto con la letra ⟨s⟩; e.g., si se presume que un entrevistado está en sus cuarenta, esto se codificaría como **40s**.

[3] Esto demuestra lo importante que es identificar las habilidades de los participantes de un proyecto documental antes de comenzar con él.

Puesto que el interés es en los entrevistados, de los cuales solo habría uno en cada sesión (y puesto que se decidió no transcribir la porción de las colaboradoras en las sesiones), no se torna necesario incluir información del Nivel II acerca de turnos de habla. Por esto, y por cuestiones de tiempo, se decidió únicamente trabajar en el Nivel I de transcripción. Los demás niveles pueden ser objeto de enriquecimiento en una etapa posterior del proyecto.

Para la transcripción ortográfica, se decidió utilizar la ortografía del *creole* desarrollada por la Islander Spelling Committee, con el apoyo de la Universidad Cristiana (Cf. Anexo A para la explicación y discusión sobre el sistema de escritura), con uso mínimo de signos de puntación, como recomienda el Grupo EAGLES (Llisterri, 1996).

La escritura de esta ortografía resulta problemática por su falta de estandarización; sin embargo, se decidió adoptarla para esta proyecto por ser una escritura fonemática y porque las grafías que utiliza se pueden codificar sin problema en prácticamente cualquier computadora, ya que todos los caracteres forman parte de la codificación ASCII.

Para que las colaboradoras se pudieran familiarizar con esta escritura, se colocaron a su disposición dos (2) textos de la Universidad Cristiana: 1) un *Glossary* bilingüe inglés estadounidense-*creole*, compilado por Dulph Mitchell y Ronald C. Morren, y adaptado a la ortografía desarrollada por el Islander Spelling Committee por Eran McGowan y Ronald G. Metzger en el 2004 (Mitchell y Morren, 2004); 2) una *Limited Word List*, para uso escolar (*Islander Creole English Limited Word List*, 2006).

Estos materiales, sin embargo, no son exhaustivos ni contienen todas las formas léxicas que ocurren en la lengua; asimismo, cabe la posibilidad de que una palabra presente en cualquiera de los dos (2) materiales ocurra como una variante ideolectal en una sesión. Por estos motivos, se decidió que, basándose en la escritura, las colaboradoras adaptaran las formas ya documentadas para acomodarlas mejor a la pronunciación de los entrevistados o simplemente representaran las formas sin documentar, siempre ateniéndose a la ortografía.

Se decidió que cada colaboradora transcribiría las mismas sesiones que había recogido, ya que tendría la mayor familiaridad con ellas.

La transcripción de las sesiones se decidió hacer por medio del programa ELAN.

5.3.3. Anotación de los datos lingüísticos

Para la anotación del corpus se tuvo en cuenta dos (2) hechos: 1) que en la documentación lingüística, pocos son los proyectos que realmente emplean anotación y 2) la anotación de los corpus ofrece varias ventajas frente a los corpus en bruto (Cf. 4.1.2).

Por ello, se decidió que el corpus sería un *corpus anotado* y que, en los textos, a cada *token* se le asignaría una etiqueta, la cual indicaría su categoría gramatical.

No se contó con una persona experta en procesamiento del lenguaje natural en el equipo quien pudiera asistir en programar un anotador basado en reglas. Además, como ya se mencionó anteriormente (Cf. 4.3), al no contar con una gramática exhaustiva del *creole*, cualquier anotador basado en reglas contaría con datos insuficientes.

Entonces se decidió que la anotación de los textos se haría de forma manual, por parte de cada una de las colaboradoras, sobre los textos que habrían transcrito. Una vez creada la lista de etiquetas (Cf. Anexo E para la explicación y discusión sobre las etiquetas), se explicó el uso de cada una a las colaboradoras y se les entregó a las colaboradoras la lista para que se fueran familiarizando con ella.

La anotación gramatical se decidió hacer, así como la transcripción, por medio del programa ELAN.

5.3.4. Análsis preliminar de los datos lingüísticos

Nos propusimos aquí caracterizar de manera breve los textos producidos por los entrevistados al analizarlos desde tres (3) conceptos principales (Baker, 2010):

- **Frecuencias:** El número de aparición de los *tokens* o frases gramaticales. Lo deseado aquí es determinar cuáles son los *tokens* con mayor frecuencia de aparición; asimismo, junto con las etiquetas, determinar cuáles son las categorías gramaticales más frecuentes. Las frecuencias se calcularían asimismo por medio de porcentajes. En este punto, también interesaría analizar la variación de los *tokens* en cuanto a ortografía, para determinar aquellas formas en las que hay variabilidad lingüística y/o discrepancia en la transcripción ortográfica, así como la variabilidad que podrían mostrar ciertos *tokens* con respecto a su etiqueta gramatical. Otro punto interesante sería observar las alternancias con el español y los *tokens* sugerentes de algún proceso de innovación léxica.
- **Concordancias:** Los contextos de ocurrencia de determinados *tokens*. Aquí se buscaría determinar tanto los contextos de ocurrencia de determinados *tokens* que pueden ser de interés como de las clases gramaticales. Por medio de las concordancias, también es posible estudiar particularidades de la estructura sintáctica de la lengua, así como su variabiidad.
- **Colocaciones:** La coocurrencia estadísticamente significativa de *tokens*. El análisis en este punto buscaría determinar cuáles son las colocaciones más frecuentes en el corpus de lengua oral en *creole*.

El análisis de estos conceptos dentro del corpus se realizaría por medio del programa AntConc. Para el análisis en AntConc, las transcripciones anotadas por medio de ELAN se exportarían como un archivo de texto plano (i.e. de extensión `.txt`) y, por medio de expresiones regulares en algún editor de texto (e.g. Notepad++[4]), se asociaría cada etiqueta con su respectivo *token*. Por esta razón era importante tratar de asegurarse de que todos los *tokens* en las transcripciones estuvieran anotados.

[4]https://notepad-plus-plus.org/

Descripción del programa ELAN

ELAN [5] es una anotador lingüístico para archivos de audio y video. Es desarrollado por el Instituto Max Planck de Psicolingüística. Está disponible para sistemas Windows 7 en adelante, Mac OS X 10.6 en adelante y Linux/Unix. Actualmente está en su versión 5.2.

Para hacer las anotaciones, primero se cargan los archivos a un nuevo proyecto de ELAN. Este proyecto se almacena con la extensión `.eaf`. La interfaz gráfica del proyecto se divide en dos (2) secciones: una superior, en la que se ubican los controles del proyecto, las cuales controlan aspectos como la velocidad y el volumen de reproducción de los archivos de audio y el video, y una inferior, en la que se realizan las anotaciones como tal (Cf. Figura **5-1**).

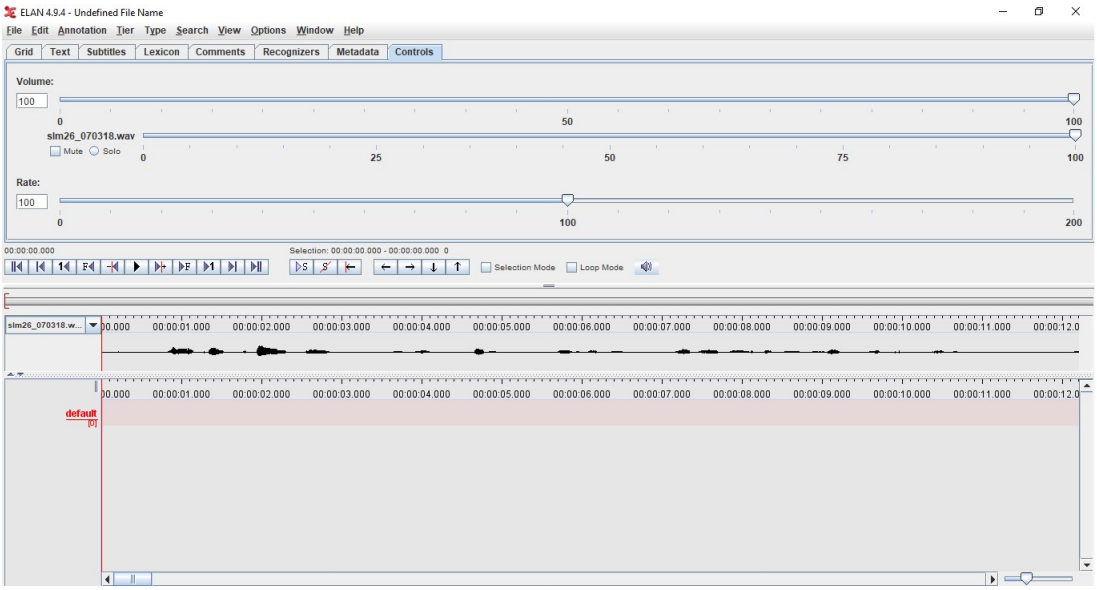


Figura 5-1.: Interfaz gráfica en ELAN

La sección de anotación se divide en líneas, las cuales se pueden modificar de acuerdo a las necesidades del usuario. Estas líneas pueden formar dependencias entre sí, de suerte que el contenido y la estructura de una línea dependiente dependan del contenido de la línea madre de aquella. Para formar estas dependencias, a cada línea se le asigna un tipo lingüístico con un determinado estereotipo, el cual determina la manera en que las líneas se relacionan entre ellas.

En las líneas madres, se puede comenzar a anotar el archivo por medio de segmentaciones. Estas segmentaciones se pueden hacer arrastrando el cursor a lo largo de la línea en el modo *Anotación* o se pueden hacer presionando la tecla Enter a medida que se reproduce el archivo en el modo *Segmentación*. En el primer caso, las anotaciones se agregan haciendo doble clic sobre la segmentación y se deben hacer una por una; en el segundo caso, hechas las segmentaciones, se pueden hacer las

[5] https://tla.mpi.nl/tools/tla-tools/elan/

anotaciones en el modo *Transcripción*, el cual permite el trabajo sobre varias segmentaciones. El resultado de esta anotación y las dependencias entre las líneas se ejemplifican en la Figura **5-2**.

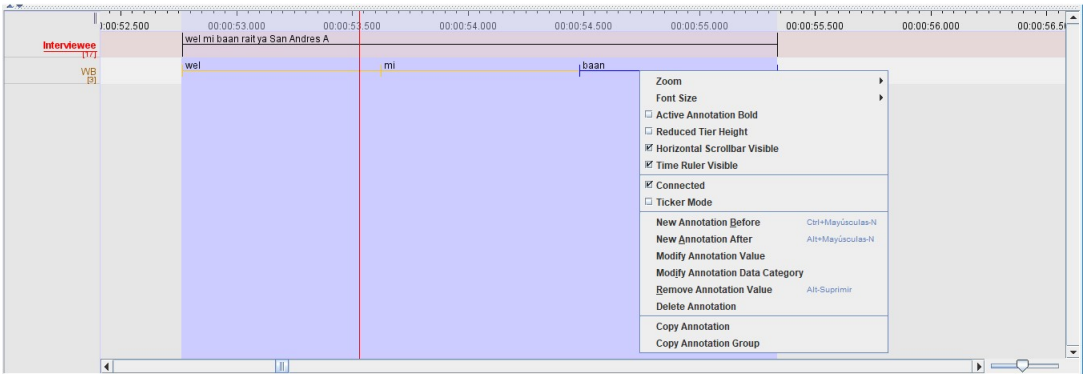


Figura 5-2.: Dependencias de línea y anotaciones

Aplicación en el proyecto

Para nuestro piloto, se decidió que, para cada sesión en ELAN, habría cuatro (4) líneas:

1. **Interviewee** — La línea madra. Las anotaciones dentro de esta línea contendrían la transcripción ortográfica del entrevistado. El tipo lingüístico se llamaría **tx** y no tendría estereotipo.

2. **WB** — Dependiente de la línea madre **Interviewee**. En esta línea se tokenizaría la transcripción ortográfica en **Interviewee** (i.e., se separaría cada *token* de la línea madre en anotaciones separadas). El tipo lingüístico para esta línea se llamaría **wb** y su estereotipo sería *symbolic subdivision*.

3. **WB-POS** — Dependiente de la línea **WB**. En esta línea se asignaría la etiqueta de clase gramatical a cada *token* segmentado en **WB**. Para ello se utilizaría un *vocabulario controlado*, el cual permite que, en una línea determinada, los valores de las anotaciones estén limitadas a una lista cerrada (Cf. Figura **5-3**). De esta forma se evitaría que hubiera discrepancia en las etiquetas. El tipo lingüístico de esta línea se llamaría **pos** y su estereotipo sería *symbolic association*.

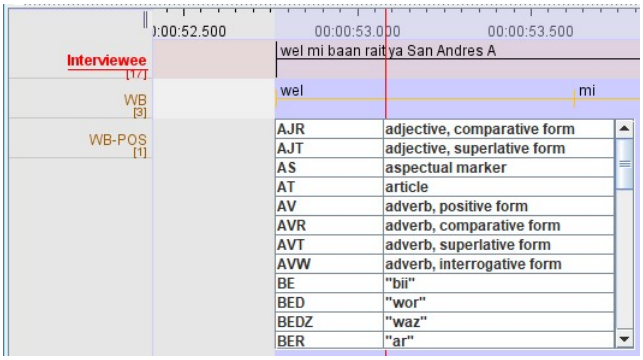


Figura 5-3.: Vocabulario Controlado en **WB-POS**

4. **FreeTranslation** — Dependiente de la línea madre **Interviewee**. En esta línea se introduciría una traducción libre al inglés (con signos de puntuación) de la transcripción en **Interviewee**. El tipo lingüístico se llamaría **ft** y su estereotipo sería *included in*.

Los productos esperados al final del piloto, entonces, tendrían una apariencia similar a la de la Figura **5-4**.

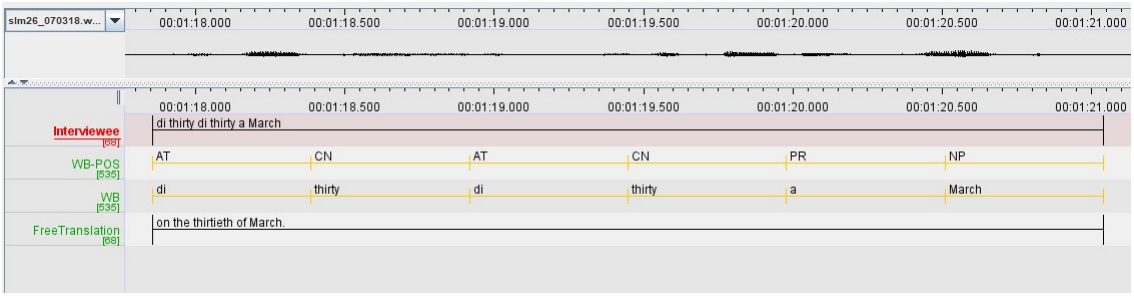


Figura 5-4.: Proyecto completado

Descripción del programa AntConc

AntConc [6] es un programa para análisis de corpus electrónicos. Es desarrollado por Laurence Anthony, de la Universidad de Okayama, Japón. Está disponible para sistemas Windows 98 en adelante, para Mac OS y Linux/Unix. Actualmente está en su versión 3.5.7.

Para realizar análisis, el programa admite archivos de texto en formato `.txt`, `.html` y `.xml`. La interfaz gráfica se compone de tres (3) secciones principales: en la sección izquierda, se encuentra el listado de todos los archivos que se están analizando en una sesión determinada, i.e. el corpus; en la sección inferior, se muestran los resultados de las diferentes herramientas de búsqueda que permite hacer el programa y, en la sección superior, se encuentran varias pestañas, entre las cuales se puede alternar para diferentes tipos de análisis:

[6]http://www.laurenceanthony.net/software/antconc/

- **Concordance** — permite hacer la búsqueda de un *token* particular en todo el corpus y lo muestra en sus diferentes contextos de uso;
- **Concordance Plot** — permite visualizar la frencuencia de ocurrencia de los *tokens* en el corpus de forma gráfica: cada texto se representa como un rectángulo y la frecuencia y ubicación del *token* en el texto se representa como una barra vertical;
- **File View** — permite visualizar el contenido completo de un texto y, si se está buscando un *token*, este se resalta dentro del mismo;
- **Cluster/N-Grams** — permite visualizar los elementos que acompañan un determinado *token*, ya sea a la izquiera o a la derecha;
- **Collocates** — permite visualizar los elementos que acompañan un determinado *token*; la diferencia con **Cluster/N-Grams** es que esta opción también muestra la frecuencia de los elementos acompañantes a la izquiera y a la derecha y asimismo muestra la probabilidad de ocurrencia de cada uno;
- **Word List** — muestra el listado de todos los *tokens* del corpus, ordenados desde el más frecuente hasta el menos frecuente, junto con la frecuencia de cada uno;
- **Keyword List** — genera un listado de *keywords* (palabras clave) con base en el *word list* del corpus en comparación con el *word list* de otro corpus de referencia; i.e., los *tokens* se ordenan de acuerdo a la variación comparativa de su frecuencia.

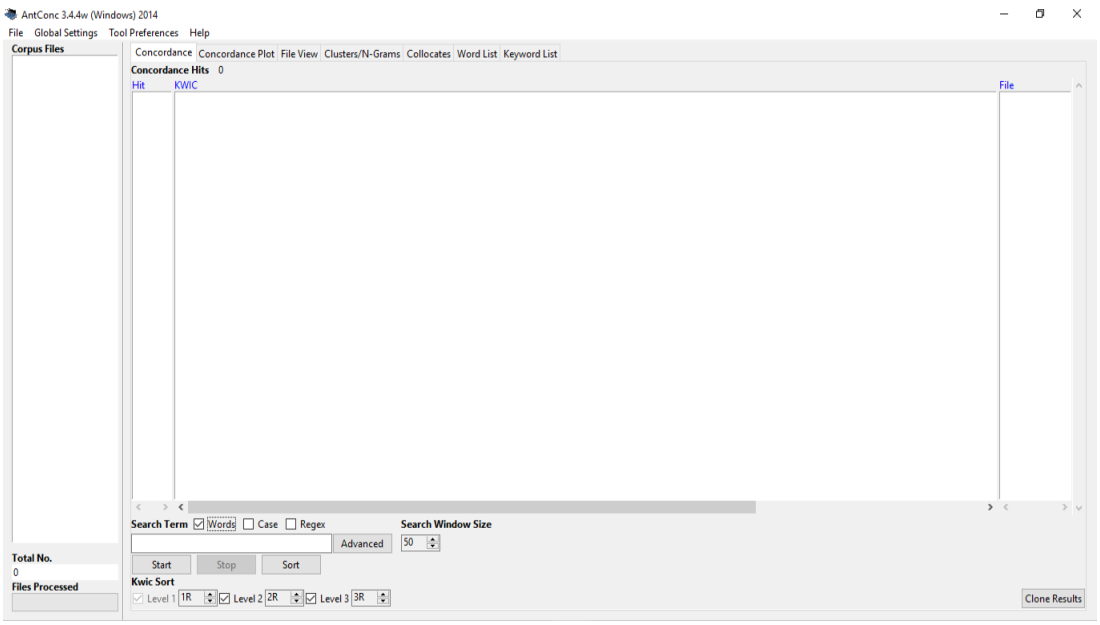


Figura 5-5.: Interfaz gráfica de AntConc

AntConc también acepta corpus anotados. Por este hecho, y por el potencial que ofrecen sus diferentes herramientas, se decidió utilizar este *software* para el análisis de nuestro corpus.

6. Resultados

6.1. Los productos del piloto

Durante el piloto, las colaboradoras llevaron a cabo con las personas entrevistadas un total de seis (6) sesiones de trabajo, dos por cada colaboradora. Estas sesiones y los nombres asignados a ellas se encuentran ilustrados en la Tabla **6-1**.

	Nombre de sesión
1	`mhf57_270418`
2	`phm54_280418`
3	`Lolia_Yard`
4	`Pabli_Yard`
5	`macf61_280418`
6	`maslf52_210418`

Tabla 6-1.: Sesiones de trabajo

Obtuvimos de estas sesiones de trabajo los siguientes productos:

- cuatro (4) archivos de audio en formato WAV
- dos (2) archivos de audio en formato MPEG-4 Parte 14 (`.m4a`)
- cinco (5) transcripciones realizados en ELAN (`.eaf`)
- una (1) transcripción realizada en Microsoft Word (`.docx`)
- seis (6) formatos de metadatos rellenados

En las diferentes fases del piloto, se presentaron dificultades y asimismo hubo logros en conseguir los productos esperados.

6.2. Recolección de los datos lingüísticos

Como mencionamos en 5.3.1, se propuso que cada sesión de trabajo fuera de treinta minutos (30:00) aproximadamente. Con esto en mente, si cada colaboradora realizaba dos (2) sesiones con los entrevistados, el corpus esperado habría sido, en total, de tres horas (3:00:00). Sin embargo, ninguna sesión de trabajo alcanzó la meta propuesta y el tiempo total de las sesiones en el corpus

fue inferior a lo propuesto, ya que es de un poco más de una hora (1:00:00). La duración de cada sesión se ilustra en **6-2**.

	Nombre de sesión	**Duración**
1	`mhf57_270418`	8:08
2	`phm54_280418`	6:22
3	`Lolia_Yard`	6:39
4	`Pabli_Yard`	4:45
5	`macf61_280418`	21:27
6	`maslf52_210418`	20:26
Total de duración		1:07:47

Tabla 6-2.: Duración de sesiones

El tiempo medio de duración de cada sesión es de once minutos y diecisiete segundos con una desviación típica de siete minutos y treinta y tres segundos (11:17±7:33); solo hubo dos sesiones que se acercaron al tiempo propuesto: `macf61_280418` y `maslf52_210418`. Esto muestra una clara desproporción con respecto a las muestras y esto afecta, a su vez, los principios de representatividad y de muestreo que debe tener el corpus.

Por otro lado, no es el caso que todos los archivos de audio estén en formato WAV, puesto que dos de ellos (`mhf57_270418` y `phm54_280418`) están en formato M-PEG Parte 14 (`.m4a`). Este formato de audio comprime los datos de audio [1], lo cual afecta la calidad de la grabación. Se desaconseja este tipo de formatos para la documentación y el trabajo en campo, ya que no son ideales para estudios fonéticos (Bowern, 2008). Las sesiones se tomaron en este formato debido al equipo de grabación que usó una de las colaboradoras.

En cuanto a la sistematización de las sesiones por metadatos, no hubo mayores inconvenientes, ya que se logró registrar toda la información requirida en cada una de las sesiones. Los metadatos nos permitieron comprobar que, en todos los casos, se entrevistó a hablantes nativos de *creole* en el rango de edad entre cuarenta y cinco (45) y sesenta y cinco (65) años. Lo único problemático fueron las convenciones que se establecieron para nombrar las sesiones (Cf. 5.3.1), ya que en dos (`Lolia_Yaad` y `Pabli_Yaad`), no se aplicaron [2].

Como se mencionó en 5.3.1, el consentimiento de los entrevistados se registraría de forma oral en las sesiones; sin embargo, en dos sesiones (`mhf57_270418` y `phm54_280418`), esto no se hizo, aunque la colaboradora reporta que los entrevistados dieron su consentimiento de forma oral antes de la respectiva sesión con cada uno.

[1] https://es.wikipedia.org/wiki/MPEG-4_Parte_14

[2] Téngase en cuenta, sin embargo, que, con el conocimiento del lugar y la fecha de la sesión, se pueden renombrar para cumplir con las convenciones.

Retos

La temática

Las colaboradas del piloto expresan que el mayor impedimento para cumplir con el tiempo requerido en las sesiones fue la temática, ya que sienten que al delimitarla a la temática de la crianza en el patio, no hay suficiente información que pueden dar los entrevistados en este respecto.

Experiencia en trabajo de campo

Otro reto con el que se enfrentaron las colaboradores fue la inseguridad frente al trabajo de campo, ya que algunas de ellas reportan no tener la experiencia ni los conocimientos suficientes para hacer entrevistas semidirigidas.

Por otro lado, expresan que, junto con esta inseguridad en el trabajo de campo, se suma la falta de un cuestionario de posibles preguntas que se pueden aplicar a los entrevistados para lograr un diálogo fluido. Esto resultaba en que no se lograban formular preguntas pertinentes a la temática y/o que las preguntas que se formulaban fueran cerradas, lo cual propiciaba a que los entrevistados se limitaran a dar respuestas binarias de "sí/no".

Equipos de grabación

Puesto que no se contaron con los equipos físicos adecuados y las colaboradoras se vieron obligadas a utilizar sus teléfonos celulares para llevar a cabo las sesiones, esto puede haber afectado la recolección de los datos. Al tener que alternar la posición del teléfono para asegurarse de capturar las respuestas de los entrevistados, esto puede resultar intrusivo e intimidante para los mismos, propiciando que estén muy conscientes de la situación y que se limiten a dar respuestas cortas y concisas.

Esto se ve reflejado también en que, según reportan las colaboradoras, una vez hubieran terminado las sesiones y dejaran de grabar, en ocasiones los entrevistados comenzaban a dar más información relacionada con la temática y mantenían una conversación con ellas por un tiempo prolongado.

Restricción del tiempo

Algunas colaboradoras expresan que, puesto que el piloto mismo proponía que cada una hiciera una transcripción posterior, simplemente no llevaron a cabo sesiones de treinta minutos (30:00) porque estaban conscientes de que no dispondrían de suficiente tiempo para transcribirlas en la fase siguiente.

6.3. Transcripción de los datos lingüísticos

Para cada sesión de trabajo se logró hacer una trascripción ortográfica. En la Tabla **6-3** se puede apreciar el número de *tokens* de cada transcripción.

	Nombre de sesión	Número de *tokens*
1	`mhf57_270418`	863
2	`phm54_280418`	915
3	`Lolia_Yard`	569
4	`Pabli_Yard`	321
5	`macf61_280418`	144
6	`maslf52_210418`	2193
Total de tokens		5005

Tabla 6-3.: Número de tokens de transcripciones

El promedio de *tokens* en el corpus es de ochocientos treinta y cuatro por cada transcripción, con una desviación típica de setecientos treinta (835±730). Basándonos en un cálculo hecho durante los talleres de formación, se asumió que habría un número promedio aproximado de cien (100) *tokens* por minuto transcrito. Por eso, si se esperaba que el total de duración de las sesiones fuera de tres horas (3:00:00), entonces era razonable esperar que el corpus resultado de las transcripciones tendría un número aproximado de 18 000 *tokens*.

Puesto que no se cumplió con lo primero, entonces se hizo una nueva aproximación. Con una hora y ocho minutos (1:08:00) aproximadamente de sesiones de trabajo, el corpus tendría un número esperado de seis mil ochocientos (6800) tokens.

El corpus es más pequeño que eso; sin embargo, se debe tener en cuenta que la transcripción de `macf61_280418` se hizo sobre una porción de aproximadamente cuatro minutos (4:00). Por ello, se hizo un tercer cálculo: si la duración total de las sesiones transcritas es de aproximadamente cincuenta y dos minutos (52:00), entonces el número esperado de *tokens* es de cinco mil doscientos (5200). El número observado de *tokens* es de cinco mil cinco (5005), menos de lo esperado. Sin embargo, si se normaliza la desviación del número esperado y el número observado de *tokens*, el valor resultante (−0,02) se acerca a cero (0), por lo cual podemos decir que hay un número de *tokens* similar al número esperado (Lijffijt y Gries, 2012) [3].

Retos

Herramientas informáticas

Las colaboradoras reportan que el impedimento más grande al momento de transcribir era el uso del programa ELAN. En ocasiones surgían elementos inesperados, se les dificultaba hacer las selec-

[3]Esto se calcula de la siguiente manera:

$$DP_{norm} = \frac{DP}{1-s} = \frac{195}{1-5005} \approx -0{,}02$$

donde DP_{norm} es la normalización, DP es el valor absoluto de la diferencia entre el número observado y el número esperado de *tokens* y s es el número de *tokens* del corpus más pequeño (i.e. el observado).

ciones dentro de la línea madre o el programa les mostraba un error. Ellas expresan que lo mejor, para evitar esto, es hacer que la formación en el programa sea más práctica, puesto que los talleres de formación tuvieron un carácter más explicativo. La justificación detrás de esto es que no se contaba casi con archivos de audio para una formación más aplicada.

A pesar de estos escollos, la mayoría logró hacer todas las transcripciones en el formato `.eaf` de ELAN, excepto una colaboradora a quien el programa le mostró errores al cargar el archivo de audio con formato `.m4a` de la sesión `mhf57_270418`. Por este motivo, ella hizo la transcripción en un documento de Microsoft Word. Se desaconseja esto, puesto que el formato histórico de Microsoft Word (`.doc`) es un formato propietario, i.e. está diseñado para ser utilizado exclusivamente con los productos de Microsoft Office (Good, 2011, pp. 223-224). Las nuevas versiones de Microsoft Word utilizan un formato abierto (`.docx`) [4] , el mismo con el que se hizo la transcripción; sin embargo, tampoco se aconseja el uso de esta, ya que muchas herramientas de análisis de corpus no lo aceptan y es más sencillo usar un formato abierto de todas formas (Good, 2011, p. 224). Por este motivo, se decidió convertir esta transcripción a un formato de texto plano (`.txt`).

Familiaridad con la ortografía

La lengua *creole* es de tradición mayoritariamente oral y, pese a que hay una ortografía desarrollada para la lengua (Cf. Anexo A), no es de difusión masiva y no se ha dado a conocer entre la población (Moya, 2006). Incluso las colaboradoras, al integrarse al piloto, reportaron que no habían tenido contacto con la ortografía antes, o al menos no de forma directa ni formal.

La literacidad en San Andrés se podría caracterizar como *multigráfica*, en la que hay una "coexistencia de dos o más códigos escritos para una lengua o una variedad" (Lüpke, 2011, p. 316) [5]. Los dos códigos ortográficos predominantes en San Andrés son el del español y el inglés; la mayor parte de las publicaciones y la producción editorial en la isla es en estas dos lenguas. En la medida en que la enseñanza de la literacidad en el colegio es bilingüe inglés-español, las personas raizales recurren a la ortografía de una u otra cuando desean escribir en *creole*. Esto se evidencia, e.g., cuando se utiliza el grafema ⟨j⟩ para representar el fonema /h/ [6]. Considérense los siguientes ejemplos de unas carteleras hechas por estudiantes PEAMA de la Universidad Nacional Sede Caribe para una marcha:

(1) **Today** fi di Island,
Tomorrow fi yu[7,8]

(2) Witout **insumos** wi no gat
one dignify health[7,8]

[4] https://es.wikipedia.org/wiki/Office_Open_XML

[5] Traducción propia.

[6] Considérese la escritura de influencia española en ⟨¿jou iu de?⟩ '¿cómo estás?' en comparación con la escritura desarrollada por la Universidad Cristiana: ⟨how yu deh?⟩ (Diario de Campo de Bryan Steven Loaiza Camacho, Cuaderno A, 3 de abril de 2018).

[7] Diario de Campo de Bryan Steven Loaiza Camacho, Cuaderno B, 18 de abril de 2018.

[8] Negrilla propia.

En (2), ⟨insumos⟩ se trata de un préstamo del español. Los demás elementos resaltados son influencia de la ortografía del inglés.

En las transcripciones, hay instancias de esta misma interferencia:

1. ⟨c⟩ en lugar de ⟨k⟩

2	ahn evrybady go houm til di neks die Ahn wehn da	**crab**	taim now wehn crab staat kom doun wi go ahn w	maslf52_210418

1	miinwail dem deh sorv dat dem gat wan pat deh mek	**krab**	suup or somting fi wi iit nat a ting nat a	phm54_280418

2. ⟨x⟩ para el grupo consonántico /ks/

1	yuuztu get alang wii yuuztu skip wii yuuztu plie	**jaks**	wii yuuztu plie stik haas wii yuuztu draiv bo	mhf57_270418

3	di gyrls fi plie jax an ef wii non gat non	**jax**	wii tek a yong laim or bitansuit ariinj an tek	Lolia_Yaad

3. Palabras ortográficas, e.g. ⟨once⟩ en lugar de ⟨wans⟩

1	ak den di aktivitys dem waz a lat difarent dan de	**once**	dem tideh bikaaz now-a-diez wel az yu rialaiz	Pabli_Yaad

Restricción de tiempo

Una colaboradora expresa que, por restricciones de tiempo, no pudo completar la transcripción de `macf61_280418` y por ello se limitó a unos cuatro minutos (4:00) aproximados.

6.4. Anotación de los datos lingüísticos

Por restricciones de tiempo, no se pudo adelantar en el transcurso de este piloto la anotación gramatical de las transcripciones. En los talleres de formación, sin embargo, se les mostró a las colaboradoras la lista de etiquetas y la manera en que se puede hacer este tipo de anotación por

medio de la tokenización y el uso de un vocabulario controlado en ELAN.

A pesar de que no se pudo adelantar esta fase, las colaboradoras mostraron interés en adelantar esta labor, ya que es una oportunidad no solo de avanzar en los propósitos del proyecto de documentación, sino también para que ellas mismas puedan estudiar de forma sistemática la gramática de su lengua. Por ello, se dejó la posibilidad abierta de que, en algún momento futuro, se adelante la anotación gramatical con las mismas colaboradoras.

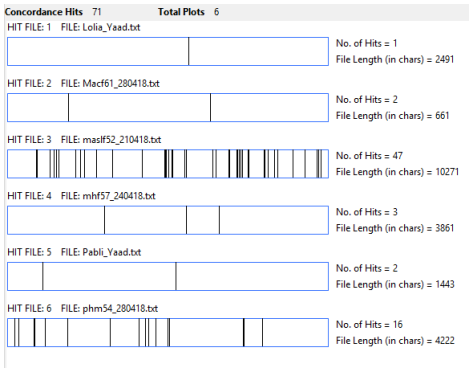
6.5. Análisis preliminar de los datos lingüísticos

Hacemos aquí un breve análisis sobre el corpus y mostramos cómo se pueden explotar las herramientas informáticas y los conceptos de frencuencia, colocación y concordancia para estudiar diferentes aspectos de la lengua y la comunidad de habla raizal.

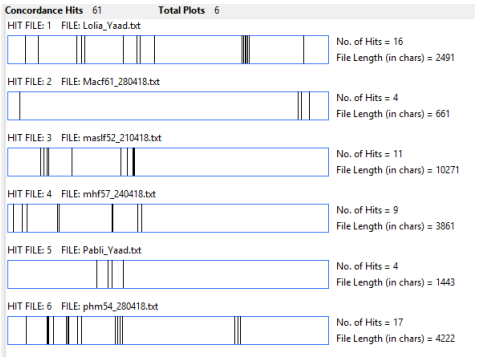
Puesto	**Frecuencia**	**Token**
1	230	wi
2	191	di
3	139	ahn
4	134	dem
5	124	an
6	95	a
7	94	yuuztu
8	81	fi
9	81	so
10	78	now
11	72	wii
12	71	go
13	68	ai
14	61	plie
15	55	tu
16	55	yu
17	54	deh
18	44	taim
19	44	yo
20	40	wehn

Tabla 6-9.: Los veinte (20) tokens más frecuentes

En la Tabla **6-9**, se aprecian las veinte (20) palabras más frecuentes en el corpus, ordenadas de más frecuente a menos frecuente. Podemos ver que en esta lista se cumple el principio general de que los *tokens* más frecuentes en un corpus son las *palabras funcionales* (Manning y Schütze, 1999, p. 21):



(a) El *token go*



(b) El *token plie*

Figura 6-1.: *Concordance Plots*

pronombres, artículos, conjunciones, preposiciones y marcadores de tiempo. El *token* en la posición 18, *taim*, además de ser sustantivo léxico, se une a ciertas palabras funcionales para formar frases adverbiales (Bartens, 2003, p. 74) [9], lo cual explicaría su alta ocurrencia.

Hay dos (2) *palabras léxicas* que se destacan en la lista: el *token* en la posición 12, *go*, y el *token* en la posición 14, *plie*.

Un primer vistazo a las graficaciones de concordancia de ambos *tokens* en las Figura **6-1** daría la impresión de que, con respecto a *go*, su alta aparición se debe a una desproporción de unos textos en comparación con otros y de que *plie* está más balanceado en todos los textos; de hecho, no hay ninguna desproporción en los dos casos.

Cálculese la desviación de la proporción para *go* y para *plie*. Este cálculo se aprecia en las Tablas **6-10** y **6-11**, respectivamente (Lijffijt y Gries, 2012). Una vez hecho esto, se puede notar que el valor resultante para *plie* (0.282) y para *go* (0.266) son similares y al normalizarlos, los resultados siguen siendo cercanos.

En el caso de *go*:

$$DP_{norm} = \frac{DP}{1 - min(s)} = \frac{0{,}266}{1 - 0{,}288} \approx 0{,}274 \tag{6-1}$$

En el caso de *plie*:

$$DP_{norm} = \frac{DP}{1 - min(s)} = \frac{0{,}283}{1 - 0{,}288} \approx 0{,}291 \tag{6-2}$$

Por lo anterior, podemos afirmar que no hay una desproporción entre los textos con respecto a

[9] E.g. en la frase adverbial *wen taim*.

Paso 1	**Paso 2**	**Paso 3**		
% Esperado	% Observado	Diferencias abs.	Suma de diferencias abs.	División por 2
0.028	0.028	0.001		
0.064	0.028	0.036		
0.113	0.140	0.099		
0.172	0.042	0.130	0.533	0.266
0.182	0.225	0.042		
0.438	0.661	0.223		

Tabla 6-10.: Dispersión de la proporción de *go*

Paso 1	**Paso 2**	**Paso 3**		
% Esperado	% Observado	Diferencias abs.	Suma de diferencias abs.	División por 2
0.028	0.066	0.036		
0.064	0.066	0.001		
0.113	0.262	0.149		
0.172	0.147	0.025	0.565	0.282
0.182	0.278	0.096		
0.438	0.180	0.258		

Tabla 6-11.: Dispersión de la proporción de *plie*

estos dos *tokens*.

Por otro lado, hay una fuerte correlación entre el número de *tokens* en los textos y la frecuencia de *plie* y *go*: en ambos casos, un mayor número de *tokens* está correlacionado con una mayor frecuencia (Kendall Tau-B test, $p < 0{,}05$).

Puesto	**Freq.**	**Freq. (Iz.)**	**Freq. (Der.)**	**Prob.**	**Token**
1	13	12	1	1.99436	wi
2	9	9	0	2.75474	yuuztu
3	9	4	5	2.19039	ahn
4	7	3	4	3.16540	tu
5	6	1	5	5.72437	baibl
6	5	4	1	2.67998	yu
7	5	0	5	3.87637	sii
8	5	4	1	4.55444	evrybady
9	5	0	5	4.65398	da
10	4	0	4	4.81748	out
11	4	2	2	6.13941	lang
12	4	1	3	2.85401	in

13	4	4	0	1.79956	fi
14	3	3	0	2.63691	wan
15	3	0	3	3.91701	huom
16	3	0	3	3.40244	dah
17	3	1	2	0.77017	an
18	2	0	2	4.13941	skuul
19	2	0	2	3.33205	roun
20	2	1	1	0.98966	go

Tabla 6-12.: Colocaciones de *go*

Una posible explicación de la alta frecuencia de *go* se puede encontrar en sus colocaciones, las cuales se pueden apreciar en la Tabla **6-12**. Como verbo, funciona como el núcleo de un sintagma verbal y lo suelen acompãnar varias palabras funcionales. Sin embargo, también sirve para indicar una intención al preceder otro verbo, e.g. en "*go sii*". Por otro lado, la gran variedad de colocados que acompañan a *go* también incluye sustantivos, como *baibl*, *huom* y *skuul*. Ello se debe a que, en *creole*, en ocasiones se puede coordinar una frase verbal y una frase nominal sin necesidad de una preposición (Bartens, 2003, p. 35).

Por otro lado, la alta frecuencia de *plie* y su dispersión balanceada pueden ser indicativos de otra clase de fenómeno, quizás de tipo discursivo. Es posible que, en cuanto a la temática, los miembros de la comunidad de habla raizal asocien de forma recurrente la crianza en el patio con el juego, más que con otros ámbitos. Esta hipótesis se ve respaldada en que otros *tokens* relacionados con diferentes aspectos del patio no ocurren con tanta frecuencia. Compárese, e.g., *mama* 'mamá', *fren* 'amigo' y *frenz* 'amigos':

Puesto	**Frecuencia**	**Token**
63	15	mama
606	1	fren
607	1	frenz

Tabla 6-13.: Frecuencias

Esto se ve asimismo reflejado en las colocaciones de *plie*, las cuales se muestran en la Tabla **6-14**. La mayoría se refiere a nombres de juegos o juguetes específicos. Esto muestra que el corpus es útil para estudiar el léxico de la lengua en este dominio semántico, además del discurso de la comunidad de habla en torno a la crianza en el patio.

Puesto	**Freq.**	**Freq. (Iz.)**	**Freq. (Der.)**	**Prob.**	**Token**
1	2	1	1	7.35842	marbl
2	1	0	1	6.35842	wit
3	1	0	1	6.35842	stik

4	5	3	2	6.35842	ring
5	1	0	1	6.35842	liberty
6	1	0	1	6.35842	lata
7	1	0	1	6.35842	jaks
8	3	0	3	6.35842	gig
9	1	0	1	6.35842	flags
10	1	0	1	6.35842	dag
11	2	0	2	6.35842	biesbaal
12	1	0	1	6.35842	basketbaal
13	4	0	4	6.03649	haid

Tabla 6-14.: Colocaciones de *plie*

Cabe aclarar, en este punto, que las generalizaciones que hacemos aquí se aplican a un corpus limitado por su tamaño y no pretenden ser definitivos. Hay posibilidad de que la alta frecuencia de *plie* se deba a las preguntas hechas por las colaboradoras. En cualquier caso, un corpus más grande en torno a la temática de la crianza en el patio podrá confirmar o falsear nuestra hipótesis.

Otros posibles procesos lingüísticos dignos de estudiarse, pero que salen del alcance de este análisis, son:

- La variación lingüística

1	liip ina mai breda hous bikaa ihn gat di muoto di	**televizhan**	si tingz in deh so dah dong deh Ai de	phm54_280418

1	unshain nait bikaaz dem deh taim dem haadly osea	**telivizhan**	wos neva di distrokshan iina dem muome	maslf52_210418

- Alternancia de código y préstamos léxicos

1	wan a di prablem dem now dem hav a lat a	**aparatos**	dem now tu kiip dem ak-yopai wi didnt hav	mhf57_270418

2	uol barrio sevral a dem yuuztu kom fram farda	**barrio**	an sidong an plie wi yuuztu go an plie	phm54_280418

- Cambio lingüístico e innovación léxica [10]

1	aalweiz wi aalwaiz aal laik di taim dis mont now	**espesifik**	dis April mont yu nuo dah crab taim	maslf52_210418

1	lrait so bot aftaword now shii yuuztu yuuz ihn	**estrategy**	pan mi wehn taim da Friday iivnin now	maslf52_210418

[10] Con una colaboradora, se ha sugerido que estas palabras han pasado por un proceso de cambio histórico, el cual consiste en una épentesis inicial del fonema /e/ por analogía del español (Diario de Campo de Bryan Steven Loaiza Camacho, Cuaderno B, 11 de mayo de 2018). Un punto que soporta esta hipótesis es el hecho de que la fonotaxis del *creole* permite los grupos consonánticos /str/ y /sp/ en inicio de palabra (O'Flynn de Chaves, 1990, p. 40); sin embargo, haría falta recopilar un corpus más grande para comprobar esta hipótesis y conocer sus posibles alcances.

7. Conclusiones

En cuanto a la comunidad de habla raizal, hay diversas opiniones con respecto al estudio de la lengua *creole* de San Andrés. Hay un recelo latente de algunos sectores frente a este tipo de acciones investigativas, mientras hay miembros a quienes sí les interesa la documentación de la lengua. Con algunas de estas personas se desarrolló un piloto durante el semestre 2018-I en el se recopiló y transcribió un corpus especial de lengua oral. Este piloto hace parte de un proyecto de investigación más amplio en el interior de la Universidad Nacional Sede Caribe y espera sentar las bases para darle continuidad a dicho proyecto.

Los textos que se recopilaron para el corpus fueron de individuos raizales entre los cuarenta y cinco (45) y los sesenta y cinco (65) años de edad, quienes compartieron algunas de sus experiencias con respecto a la crianza en el patio, un espacio socializador en el contexto de la isla de San Andrés.

En total, se logró llevar a cabo seis (6) sesiones de trabajo. El corpus recolectado no cumplió con todas las metas que se habían propuesto; sin embargo, se logró obtener uno que pudiera dar cuenta de ciertas particularidades dentro de la comunidad de habla raizal. Este corpus también permitió formular una serie de hipótesis a partir de las cuales se pueden adelantar mayores estudios de la lengua a partir de la recopilación de corpus electrónicos.

Los mayores retos enfrentados durante el piloto fueron: 1) la inseguridad frente al trabajo en campo y las técnicas de elicitación; 2) la falta de equipo físico especializado para las sesiones; 3) la disponibilidad de tiempo de las participantes del piloto, problemática reflejada en el hecho de que no se pudo llevar a cabo la fase de anotación gramatical del corpus; 4) la implementación del sistema de escritura del *creole* y 5) la inseguridad frente a las herramientas informatáticas para la transcripción y la sistematización de los datos lingüísticos.

En general, se demostró que el uso de herramientas informáticas permite realizar determinados tipos de análisis de forma sistemática y altamente eficaz y asimismo puede beneficiar futuros proyectos de documentación y descripción lingüística.

Podemos afirmar que los resultados del piloto fueron satisfactorios y en su mayor parte positivos, lo cual aboga por la posibilidad de que se continúe con el proyecto, siempre con los retos y los logros obtenidos durante este periodo en mente.

Esperamos que esta experiencia informe otros proyectos de alcance similar y aporte al avance de la lingüística de corpus en Colombia, especialmente con miras a las lenguas criollas e indígenas.

8. Recomendaciones

Es importante incentivar el acercamiento del proyecto con la comunidad de habla raizal y abrir el diálogo para definir las metas y los propósitos para los que se hace la documentación de la lengua *creole* en el contexto de este proyecto. En la medida en que los miembros de la comunidad son agentes interesados en su propia lengua, hay que buscar un vínculo más estrecho con ellos. Como se indicó en 5.1.1, hubo un acercamiento inicial con algunos miembros de la comunidad; sin embargo, por la amplia variedad de posturas adoptadas al interior de la misma, es importante contar con un alcance mayor en la comunidad, especialmente con los líderes.

Hay que vincular a individuos de diferentes trasfondos al proyecto, ya que este tipo de iniciativas necesita apoyo humano y realmente se puede beneficiar del aporte de profesionales de diferentes áreas, como la antropología, la sociología, las ciencias de la computación, la bibliotecología y la archivística.

Por otro lado, es importante tratar de contactar a diferentes entidades que puedan estar interesadas en el trabajo que se realiza o en algún aspecto de ello. La colaboración entre instituciones puede propiciar que se lleven a cabo trabajos investigativos conjuntos, en los que cada entidad se encarga y se especializa en un aspecto determinado, e.g., la recolección o el almacenamiento de datos.

Es necesario conseguir los recursos suficientes para darle continuidad al proyecto. Eso implica conseguir los equipos físicos adecuados para la grabación de sesiones de trabajo, los computadores necesarios para llevar a cabo los talleres de formación y el espacio de almacenamiento para archivar los datos que se recolecten.

Conocida la situación de multigrafía de San Andrés (Cf. 6.3), hay que reflexionar acerca de cómo lograr que se implemente el sistema ortográfico adoptado y de cómo los colaboradores pueden aprenderlo. En este sentido, sería quizás provechoso llevar a cabo talleres de escritura con algún miembro de la comunidad conocedor de la escritura al mando.

Lo ideal es que el proyecto de documentación sea lo más diverso posible. Por lo tanto, las temáticas de las sesiones de trabajo deben ser espontáneas. La descripción lingüística se beneficiaría así al ponerse su enfoque sobre los datos y no los datos sobre el enfoque de la descripción. Pese a esto, es útil formular algún cuestionario guía con el que se pueda llevar el flujo de las sesiones de trabajo.

Para una correcta implementación del proyecto, se debe planificar de manera sistemática las fases del mismo y fijar metas, siempre atentos tanto al tiempo de los participantes en general y a los

recursos. Para esto, sugerimos siete (7) fases:

1. **Fase de diseño:** En esta fase, se determina el tipo de texto que se ha de recoger (habla espontánea, controlada, semicontrolada, etc.). Luego se diseñan los intrumentos metodológicos con los que se va a trabajar (formato de metadatos, cuestionario); se consigue el equipo físico para el proyecto (grabadora, micrófono, computadora); se decide sobre las especificaciones técnicas de los productos (`.wav`, `.txt`, etc.) y se decide sobre el modelo de transcripción de los datos.

2. **Fase de entrenamiento:** El entrenamiento de colaboradores implica conocer de antemano sus perfiles, sus intereses y el tiempo del que disponen para el entrenamiento y las reuniones. Es importante tener como recompensarlos por su tiempo. Es importante hacer los talleres de entrenamiento prácticos, para que los colaboradores aprendan a medida que realizan el trabajo. Esta fase se da en una serie de etapas:
   a) Acerca de las generalidades del proyecto (qué se busca y para qué, cómo se busca)
   b) Instrumentos metodológicos
   c) Convenciones de transcripción y sistematización de los datos
   d) Programas o soportes informáticos
   e) Categorías gramaticales para la anotación
   f) Anotación de los textos por medio del método informático seleccionado

3. **Fase de recolección:** Es clave fijar con los colaboradores las fechas en que se llevarán a cabo las sesiones de trabajo, de acuerdo a la cantidad requerida y al tiempo del que disponen los mismos. Una persona debe encargarse de recolectar los datos recogidos y sistematizados por todos los colaboradores y colocarlos en un archivo central. Otro aspecto importante es hacer una revisión de cada grabación para determinar que son pertinentes al proyecto.

4. **Fase de transcripción:** Es recomendable que los colaboradores trabajen sobre las sesiones de trabajo que ellos mismos llevaron a cabo, puesto que están más familiarizados con estos. Una vez que se hagan las transcripciones, estas pasan por dos (2) revisiones: una (1) por otro colaborador y otra (1) por una persona en la gerencia del proyecto.

5. **Fase de anotación:** Esta es la fase que requiere mayor atención y minucia. Esta debe tener el mismo seguimiento y proceso de revisión que la Fase 4.

6. **Fase de codificación:** Esta fase se lleva a cabo por dos (2) o tres (3) personas, las cuales agregan la información paralingüística y metalingüstica que sea necesaria. Estos elementos pueden verse reflejadas en la Fase 4. por medio de signos de puntuación; sin embargo, en esta etapa estas marcas se debe codificar por medio de etiquetas de algún estándar como el de los lineamientos TEI (TEI Consortium, 2018).

7. **Fase de procesamiento:** En esta fase, se relacionan los textos con los metadatos de las sesiones y los participantes por medio de alguna base de datos relacional, e.g. MySQL, para así hacer búsquedas de manera adecuada, rápida y amable a un público general.

8. **Fase de archivación:** Una vez esté recopilado el corpus, se debe decidir sobre la forma en que se almacenará para garantizar su preservación y su acceso futuro. Los aspectos de esta fase, a pesar de ser la última, son de los primeros que se deben definir en la Fase 1.

A. Anexo A: Ortografía de la Universidad Cristiana y Discusión

Las convenciones ortográficas que decidimos adoptar para este proyecto se basan en gran medida en la propuesta desarrollada por el Islander Spelling Committee, con el apoyo de la Corporación Universidad Cristiana, para la escritura del *creole*.

La idea de desarrollar un sistema de escritura para el *creole* surgió por la dificultad a la que se enfrentaban los profesores de lengua con respecto a la manera en que la lengua se escribía antes de la propuesta, la cual tenía influencia del sistema de Belice. Así, la propuesta que surgió apuntaba a ser mucho más fonémica, i.e., a tener una correspondencia uno-a-uno entre cada fonema de la lengua y un símbolo grafemático dado. Entonces el fonema /a/ se escribiría siempre como ⟨a⟩ y nunca ⟨e⟩; el fonema /k/ se escribiría siempre como ⟨k⟩ y nunca como ⟨c⟩, etc. Esto le concede cierto grado de transparencia al sistema ortográfico, en el sentido de que las palabras siempre se "escriben como suenan".

Reproducimos aquí las convenciones de las equivalencias como fueron propuestas por el Islander Spelling Committee:

Fonema	**Grafema**	**Ejemplo**
/a/	⟨a⟩	bad 'malo'
/aː/	⟨aa⟩	baan 'nacer'
/b/	⟨b⟩	bii 'abeja'
/ʧ/	⟨ch⟩	chiip 'barato'
/d/	⟨d⟩	daag 'perro'
/e/	⟨e⟩	bed 'cama'
/f/	⟨f⟩	fish 'pez'
/g/	⟨g⟩	guot 'cabra'
/h/	⟨h⟩	hit 'golpe'
/i/	⟨i⟩	bit 'morder'
/iː/	⟨ii⟩	tiit 'diente'
/ʤ/	⟨j⟩	brij 'puente'
/k/	⟨k⟩	kiek 'pastel'
/l/	⟨l⟩	fuul 'tonto'
/m/	⟨m⟩	muma 'madre'
/n/	⟨n⟩	nait 'noche'

/ŋ/	⟨ng⟩	sing ‘cantar’
/ɲ/	⟨ny⟩	nyam ‘comer’
/o/	⟨o⟩	brok ‘romper’
/p/	⟨p⟩	paip ‘cañería’
/r/	⟨r⟩	ries ‘carrera’
/s/	⟨s⟩	suop ‘jabón’
/ʃ/	⟨sh⟩	shied ‘sombra’
/t/	⟨t⟩	tuo ‘remolcar’
/u/	⟨u⟩	fut ‘pie’
/uː/	⟨uu⟩	truu ‘verdad’
/v/	⟨v⟩	vuot ‘votar’
/w/	⟨w⟩	wiet ‘esperar’
/j/	⟨y⟩	yaam ‘ñame’
/z/	⟨z⟩	briiz ‘brisa’
/ʒ/	⟨zh⟩	okiezhan ‘ocasión’
/ai/	⟨ai⟩	Baibl ‘Biblia’
/ie/	⟨ie⟩	tiebl ‘mesa’
/ou/	⟨ou⟩/⟨ow⟩	hous ‘casa’/kow ‘vaca’
/uo/	⟨uo⟩	ruop ‘soga’
/iu/	⟨yu⟩	yunifaam ‘uniforme’

Tabla A-1.: Correspondencias del sistema ortográfico

Adicionalmente, junto a estas correspondencias, hay algunas consideraciones que se deben tomar en cuenta:

1. No es el caso que este sistema de escritura sea cercano a la ortografía del inglés estándar, ya que muchas ocurrencias del inglés escrito se evitan. La reduplicación de letras no ocurre en la gran mayoría de palabras que sí tienen este rasgo en su contraparte inglesa, e.g., ⟨mis⟩ ‘señorita’ vs. ⟨miss⟩ en inglés estándar. La reduplicación sí ocurre, sin embargo, cuando es contrastable, como en el caso de la vocales: ⟨ful⟩ ‘lleno’ contrasta con ⟨fuul⟩ ‘tonto’.

2. Es el caso, sin embargo, que se tomaron ciertas decisiones cuestionables con respecto al tratamiento de nombres propios y numerales. En estos casos, el Islander Spelling Committee ha decidido que estos se han de escribir igual que en inglés estándar y no usando las correspondencias fonémicas expuestas arriba. Aquí entran el nombre de las personas (⟨John⟩ y no *⟨Jon⟩ [1]), los lugares (⟨Providence Island⟩ y no *⟨Pravidens Ailant⟩), los días de la semana (⟨Monday⟩ y no *⟨Mondie⟩), los meses del año y los números (⟨two⟩ y no *⟨tuu⟩, etc.). La razón aducida para esto es que estas funcionan como “palabras para la vista”, las cuales pueden ser más fáciles de decodificar para estudiantes con conocimiento del inglés.

[1] Utilizamos el símbolo asterisco (*) para indicar aquellas formas escritas cuyo uso se juzgaría incorrecto desde las convenciones del Islander Spelling Committee.

3. Otras convenciones que no entran en esta categoría, sin embargo, son el uso de ⟨w⟩ para representar /u/ a final de palabra (de manera que, en vez de *⟨kou⟩, /kou/ 'vaca' se escribiría ⟨kow⟩) y el uso de ⟨y⟩ para representar /i/ a final de palabra [2]. No se ha dado respuesta de por qué la regla se estableció así.

4. El uso del grafema ⟨h⟩ es bastante problemático más allá de los casos en los que forma parte de un dígrafo (como en ⟨ch⟩, ⟨sh⟩, ⟨zh⟩) o en los representa una fricativa glotal. En especial, en su uso por "motivos estéticos" en palabras como ⟨weh⟩ (pronombre relativo), ⟨deh⟩ (cópula locativa) y ⟨dah⟩ (cópula ecuativa), donde no posee estatus fonémico. Esta situación se ve exacerbada por la falta de consistencia en su uso al interior de los materiales publicados por la Organización Universidad Cristiana [3].

En tanto que el sistema ortográfico no está en una etapa avanzada de desarrollo, aún hay vacíos que no se han tocado. Adicionalmente, puede que aún haya dudas acerca de la mejor forma de escribir la lengua. Reseñamos aquí cómo adoptamos este sistema ortográfico para este proyecto y algunas de las formas en las que se pueden disipar estas dudas.

1. La propuesta del Islander Spelling Committee es un esfuerzo invaluable para la preservación de una ecología lingüística en la que el *creole* pueda prosperar y, por ello, no la debemos ignorar. Por el contrario, las fortalezas de este sistema de escritura sobrepasan sus problemáticas y son el motivo por el cual lo adoptamos con el propósito de documentar y describir la lengua.

2. No debemos ignorar los materiales y los recursos que se han producido como resultado de esta propuesta y de los esfuerzos de la Organización Cristiana. Si por algún motivo, hay dudas acerca de la escritura correcta de una palabra determinada, cabe la posibilidad de que se encuentre entre los recursos desarrollados por esta comunidad.

3. Por otro lado, si por algún motivo una palabra o un término no está presente en los recursos existentes o una variante es simplemente *muy* diferente del estándar propuesto, no se debe temer en presentar una noción propia acerca de cómo se debe representar. El punto aquí no es solo revisar laboriosamente los materiales existentes, sino expandir sobre ellos y sobre nuestro conocimiento de la lengua.

4. Puesto que estamos trabajando sobre un corpus de lengua oral y nuestro foco está sobre estructuras gramaticales y palabras léxicas, y no sobre cualidades paralingüísticas específicas,

[2] La excepción a esta regla, sin embargo, es que nunca ocurre cuando /i/ forma parte del diptongo /i/; así, /bwai/ 'niño' siempre se escribe como ⟨bwai⟩ y nunca como *⟨bway⟩. Hemos notado una restricción más: /i/ a final de palabra nunca se representa como ⟨y⟩ si la palabra es monosilábica. En publicaciones que utilizan esta escritura, no hemos visto que el complementador *fi* se represente como *⟨fy⟩, ni que se utilice este grafema en pronombres personales, como *mi* o *wi*. Únicamente lo hemos visto en palabras tales como ⟨sity⟩ 'ciudad' y ⟨apartyunity⟩ 'oportunidad', que son multisilábicas.

[3] Hay un uso de ⟨h⟩ al que quisieramos llamar la atención: en el dígrafo ⟨hn⟩, se supone que representa nasalización vocálica. En el caso de determinados pares mínimos que contrastan únicamente en la cualidad nasal de vocales, como en ⟨ihn⟩ 'él/ella' vs. ⟨ih⟩ 'ello', puede ser un recurso útil. Sin embargo, solo quisimos usarlo en esos casos en los que hay en efecto un par mínimo, puesto que el estatus general de la nasalidad vocálica en *creole* aún no se ha estudiado en profundidad.

evitaremos el uso de signos de puntación durante el proceso de transcripción de la lengua. Esta decisión la tomamos con el propósito de evitar aquellas situaciones que pueden surgir en las que no hay seguridad en la manera de colocar puntuación sobre un texto basado en el habla, lo cual de por sí constituye un reto. De esta forma, evitamos complicar demasiado el proceso de anotar los textos con etiquetas gramaticales.

5. Además, con el ánimo de mantener el proceso de anotación gramatical tan sencillo como sea posible, proponemos que la escritura de los números sea completamente grafémica, e.g., ⟨nineteen⟩ en lugar de ⟨19⟩. Adicionalmente, proponemos que los dígitos pertenecientes a un solo número se unan con un guion (⟨-⟩), e.g., ⟨nineteen-ninety-one⟩ en lugar de ⟨nineteen ninety-one⟩.

6. De ninguna forma proponemos que estos cambios sean adoptados por la comunidad general, ni que los textos que se produzcan con estas modificaciones sean un intento de estandarizar la lengua. En la medida en que estas propuestas benefician nuestro proyecto de investigación, las adoptamos. Cualquier cambio permanente en el sistema ortográfico de una lengua se debe debartir al interior de la comunidad misma y nosotros participaremos en ese debate solamente en la medida en que la comunidad, como es su derecho, lo considere productivo.

B. Anexo B: Códigos de Barrios

Las abreviaturas para los barrios se basan en la lista disponible en la página de Datos Abiertos Colombia. [1]

Para cada una de las abreviaturas, se procuró que no se sobrepasaran de dos caracteres. En 6.1, la forma en que se codificaron los barrios en que tomaron lugar las sesiones de trabajo fue de tal forma que en algunas el barrio se codificó con más de dos caracteres. La razón detrás de ello es que este listado se generó luego de la recolección de datos. En su momento, se iba creando una codificación *ad hoc* para cada sesión; sin embargo, debido a lo poco práctico que eso resultaba, se decidió sistematizar cada barrio de modo que se pudiese aplicar esta propuesta en el futuro.

Nombre de Barrio	**Abreviatura**
5 de Noviembre	5n
Abraham Hole	ah
Almendros	al
Amigo	am
Angula	an
Atlantico I Etapa	a1
Atlantico II Etapa	a2
Back Road	br
Back Road Parte Alta	bz
Back Road Parte Baja	bj
Bailey Boat	bb
Barker	be
Barker's Hill	bh
Barrack	bk
Barrio de los Profesores	bq
Barrio Obrero	bo
Battle Ally	ba
Big Fig Tree	bf
Big Point	bp
Big Twuestick Tree	bt
Bight 1	b1
Bight 2	b2

[1] https://www.datos.gov.co/w/ng2u-6iaj/dneh-mcp2?cur=Pc5kuwTny9U&from=root

Bill Taylor	bx
Bina	bi
Black Dog	bd
Bob Ground	bg
Bob Red Ground	bu
Bottom Ground	bn
Bottom Side	bs
Bowie Bay	by
Brisa Cantera	bc
Brook's Bottom	bm
Brook's Hill	bl
Buenos Aires	bw
Cabana Altamar	ca
Campo Hermoso	co
Captain Ground	cn
Carpinter Yard	cy
Cartagena Alegre	ce
Cassion Call	cc
Cesar Gaviria	cg
Chain Ground	cd
Choco Mar	cm
Clay Mounth	ch
Cliff	cf
Cocal	cl
Coco Plum Bay	cb
Constans Spring	cs
Cotton Tree	ct
Courth House	cu
Cove Hill	cv
Cove Sea Side	cz
Daddy Williams	dw
David Hill	dh
Dumo Rock	dr
Durna Pond	dp
Elsy Bar	eb
Esperanza	ez
Flower's Hill	fh
Forbes Landing	fl
Francis	fr
Free Town	ft

Galan	gl
Goat Head	gh
Grace Piece	gp
Green Hill	gn
Ground Road	gr
Guinea Hen	ga
Haines Bight	hb
Hell Gate	hg
Hill Well	hw
Hoffie	hf
Hooker Bight	hr
Horn Landing	hl
Jack Pond	jp
Jenny Bay	jb
Jhon Bernard	jd
Jhon Thyme Hill	jt
Jhon Well	jl
Jhonny Well	jy
Jim Pond	jn
Joe Wood Point	jw
Jones Ground	js
Jungla	jg
Kitty	kt
La Jaiba	lj
La Paz	lp
La Union	lu
Las Gaviotas	lt
Las Palmas	la
Laureles	lr
Lever South End	ls
Lime Hill	lh
Linval	lv
Little Hill	li
Long Ground	lo
Los Guangaros	lg
Los Milagros	lm
Low Bight	lb
Lox Bight	lx
Macca Ground	mg
Man Off War Tree	mw

Manuel Ground	md
Mariah Hill	mh
Massally	ms
Massamy Hill	ml
Mattina Hill	mt
May Mounth	mm
Michel Hill	mc
Miss Rose	mr
Modelo I Etapa	m1
Modelo II Etapa	m2
Morris Landing	mn
Natania I Etapa	n1
Natania II Etapa	n2
Natania III Etapa	n3
Natania IV Etapa	n4
Natania V Etapa	n5
Natania VI Etapa	n6
Natania VII Etapa	n7
Natania VIII Etapa	n8
New Castle	nc
New Town	nt
Nixon Hill	nh
Nixon Point	np
Nuevo Mexico	nm
Old Harbor Hill	oh
Old Hill	ol
Old Plantain Walk	ow
Orange Hill	or
Paraiso	pd
Perry Hill	ph
Plat Form	pf
Platanal	pa
Pleasant Point	pp
Polly Hill	pl
Pox Hole	px
Punkin Peace	pc
Putty Hill	pt
Raile Fence	rf
Red Ground	rg
Relleno Oriental	ro

Roack	rk
Rock Hole	rh
Rock Rock	rr
Sagrada Familia	sf
San Francisco de Asis	sa
San Luis	sl
Santa Ana	sn
Sarie Bay	sy
Savannah	sv
Schonner Bight	sb
School House	sc
Seaser Smith	ss
Serranilla	sr
Shingle	sg
Simpson Well	sw
Simth Channel	sm
Sisbet Red Ground	se
Slave Hill	sh
Sound Bay	so
Sprat Bight	sp
Swaloo Point	si
Swamp Ground	sd
Tablitas	tb
Terminal Maritimo	tr
Tolima	tl
Tom	tm
Tom Hooker	th
Urbanizacion El Bight	ub
Vietnam	vt
Villa Caribe	vc
Villa Modelia	vm
Yazzy Landing	yl
Zapadilly Tree	zt
Zarabanda	zb
Zigle	zg
Zona de Terminal	zr
Zona Deportiva de Swamp Ground	zd
Zotas	za

Tabla B-1.: Abreviaturas para los barrios de San Andrés

C. Anexo C: Formato de Metadatos

SESSION METADATA

SESSION NAME	
RECORDED BY	
DURATION	
SPEAKER	
LANGUAGE	
OTHER LANGUAGES IN RECORDING	
DATE (DD/MM/YYYY)	
PLACE	
DESCRIPTION	
ADDITIONAL COMMENTS	

PARTICIPANT METADATA

NAME	
OTHER NAMES	
PSEUDONYM	
AGE	
SEX	
PLACE OF RESIDENCE	
OTHER/PREVIOUS PLACES OF RESIDENCE (YYYY-YYYY)	
OCCUPATION	
EDUCATION	
MARITAL STATUS	
LANGUAGE BACKGROUND	
ADDITIONAL COMMENTS	

D. Anexo D: Cuestionario de Consentimiento

This is part of a research project that intends to collect samples of the *Creole* language that will allow us to study it and thus contribute to its preservation. In the future, we would like to use these samples as a basis to develop other reference and educational materials.

It is important that I get *permission* from everyone working with me. That is, we need to record that you allow us to do this work, and we want to be clear about any objections that you want to put on it. We need to know so we can respect your wishes. I will ask you some questions about our work. Please interrupt me at any time if you have any doubts.

Is it alright if we record you?

Participant Identification

- Is it alright for me to tell other people that you are working with me?
- If not, should I use a nickname?

Permission to Disseminate Materials

- Can I show my students, colleagues, and other researchers the work we are doing?
- Can I put copies of everything in an archive in case anything happens to my copies?

Permission to Use Raw Materials in Other Linguistic Projects

- Is it alright for us to make use of our work with you as a sample within articles and books we may write?
- Can we use this material in the future for other purposes, such as writing a dictionary or developing educational materials?
- Are you okay with other researchers accessing the material in the future for further studies?

E. Anexo E: Lista de Etiquetas

Las etiquetas gramaticales propuestas aquí están inspiradas por diferentes *tagsets* usados en diferentes corpus. Entre estos, mencionamos el UCREL Claws5 Tagset [1], el Brown Corpus Tagset [2] y el Treetagger Tagset [3].

Este *tagset* está basado en las descripciones de Bartens (2003) y O'Flynn de Chaves (1990). Pese a que intenta ser exhaustivo, puede que aún hayan clases gramaticales sin reconocer que aporten una mirada valiosa a la lengua y a su estrcutura morfológica. Está abierto a revisión y cualquier sugerencia al *tagset* es bienvenido.

En el *tagset* aquí propuesto, hay cuatro (4) clases léxicas principales: adverbios, adjetivos, sustantivos y verbos. La forma base de cada clase se limitó a dos caracteres: **AJ** para adjetivos, **AV** para adverbios, **NN** para sustantivos y **VB** para verbos. Partiendo de ahí, los caracteres que suceden estas secuencias pretenden mostrar algún rasgo de inflexión morfológica: en el caso de adjetivos y adverbios, se añadió **R** para indicar una forma comparativa y **T** para indicar una forma superlativa.

En cuanto a sustantivos, no parecen sufrir inflexión morfológica de forma regular; sin embargo, se acuerdo a nuestro conocimiento personal, en algunas variedades del *creole*, surge una forma plural del sustantivo. Para indicar esto, se añadió **Z** a la etiqueta **NN** (i.e. **NNZ**). Sin embargo, el mecanismo principal con el se marca la pluralidad es por medio de un pronombre posnominal.

En cuanto a verbos copulares, tenemos dos etiquetas diferentes para las dos cópulas prinicipales: *deh* locativo y *dah* ecuativo. Sin embargo, en ciertas variedades del *creole*, otras formas más cercanas al inglés se presentan, de manera que hemos añadido la etiqueta **BE** para la forma infinitiva de la cópula y otras etiquetas que informan de su inflexión.

En lo que respecta a los marcadores TMA, hemos añadido etiquetas separadas para distinguirlos.

Adicionalmente, añadimos la etiquta **FW** para indicar entranjerismos. Esta etiqueta se puede unir mediante un guion (⟨-⟩) a prácticamente cualquier clase léxica abierta. Por el momento, solo hemos identificado préstamos léxicos en la clase de los sustantivos; sin embargo, si se hallasen otros en las demás clases léxicas, la aplicación de esta etiqueta se puede expandir.

[1] http://ucrel.lancs.ac.uk/claws5tags.html
[2] http://www.hit.uib.no/icame/brown/bcm.html
[3] https://courses.washington.edu/hypertxt/csar-v02/penntable.html

Etiqueta	**Atributo**	**Ejemplo**
AJ	adjetivo, forma positiva	*likl* en "di **likl** bwai ron huom"
AJR	adjetivo, forma comparativa	*beta/wo(r)s*
AJT	adjective, forma superlativa	*besties/worsara*
ASP	marcador aspectual	*don* en"a **don** finish"
		yuuztu en "Ai **yuuztu** liv in Leaky Hill"
		deh en "dehn **de** wiet pan mi"
AT	artículo	*di* en "**di** gud ting bout ih"
		wan en "A hav **wan** buk"
AV	adverbio, forma positiva	*alwiez* en "**alwiez** Ai liv rait deh ina San Luis"
AVR	adverbio, forma comparativa	*faasa* en "ihn staat swim muo **faasa** out iina di diip"
AVT	adverbio, forma superlativa	*wors* en "him da di wan we du di **wors**"
AVW	adverbio, forma interrogativa	*wen* en "A rimemba **wen** A yuuztu bii smaal"
BE	"bii"	"A rimemba wen A yuuztu **bii** smaal"
BED	"wor"	—
BEDZ	"waz"	—
BER	"ar"	—
BEZ	"iz"	"di ting **iz** dem did rekognaiz wi"
CC	conjunción	*an* en "mai muma **an** mai sista"
		if en "**if** A gat di aportunity"
CN	número cardinal	*twenty-six* en "A **twenty-six** yierz uol"
DAH	cópula ecuativa "dah"	"mai niem **dah** Salmo"
DEH	cópula locativa "deh"	"A gwain **deh** rait ya"
DT	determinante	*dis* en "A de get tu twenty-seven **dis** yier"
DTW	determinante, forma interrogativa	*wich* en "**wich** entity yu de work wid?"
FI	complementador de pronombre y verbo "fi"	"naiz **fi** miit yu"
		"disya buk dah **fi** mi"
FI-N	complementador de pronombre y verbo, no "fi"	"Anaansi begin **tu** ron"
FOC	marcador focal "dah"	"**dah** huu di hel deh mari now-a-diez"
FW-NN	palabra extranjera, sustantivo	*buela* en "mai **buela** kom an biit mi"

GW	"gwain"	"A **gwain** staat tiich"
		"A **gwain** skuul"
IJ	interjección	*wel* in "**wel** mi baan rait ya"
MEK	marcadores exhortativos "beg," "mek," y "les"	"**les** go"
		"**mek** mi sii"
		"**beg** no du ih"
MOD	marcador modal	*waahn* en "wi **waahn** tek wan foto wid unu"
NEG	partícula de negación	*no* en "mi **no** gat **no** prablem"
		neva en "A **neva** rimemba di niem"
NN	sustantivo	*laif* en "mos a mi **laif**"
NNZ	sustantivo, plural	*yierz* en "six **yierz** ago"
NP	nombre propio	"A baan rait ya **San Andres**"
OD	número ordinal	*fos* en "di **fos** pliez Ai yuuztu liv"
PN	pronombre personal	*wi* en "**wi** staat sit dong"
PNF	pronombre reflexivo	*wiself* en "wi hafy difend **wiself**"
PNI	pronombre indefinido	*somting* en "unu gat **somting** fi du"
PNL	pronombre relativo	*we* en "dat dah di gyal **we** mi nuo"
PNS	pronombre posesivo	*mai* en "**mai** muma an **mai** sista"
		mi en "mi an **mi** frend"
PNW	pronombre interrogativo	*huu* en "**huu** di hel deh mari"
PR	preposición	*iina* en "rait deh **iina** di biich"
QT	cuantificador	*mos* en "**mos** a mi laif"
TM	marcador temporal	*wehn* en "dem **wehn** yong"
VB	verbo, forma base	*stody* en "Ai **stody** Sociology"
VBD	verbo, forma anterior	*gaan* en "wi **gaan** plie"

Bibliografía

Adolphs, S., y Knight, D. (2010). Building a spoken corpus: What are the basics? En A. O'Keeffe y M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 38–52). Abingdon: Routledge.

Andrade Arbeláez, J. (2006). Estudio sociolingüístico de San Andrés, isla: un aporte a la cultura sanandresana. *Cuadernos del Caribe*(8), 42–55. Descargado de `https://revistas.unal.edu.co/index.php/ccaribe/article/view/41704`

Anthony, L. (2011). *Laurence Anthony's Website: AntConc.* Descargado 2018-03-08, de `http://www.laurenceanthony.net/software/antconc/`

Anthony, L. (2014). *AntConc Tutorial (Ver. 3.4.0).* Descargado de `https://www.youtube.com/playlist?list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS{\_}TZj`

Baker, P. (2010). Corpus Methods in Linguistics. En L. Litosseliti (Ed.), *Research Methods in Linguistics* (pp. 93–113). New York and London: Continuum Intl Pub Group.

Banco de la República. (2017). *Centro de Memorias Orales.* Descargado 24 de mayo de 2018, de `http://www.banrepcultural.org/centro-de-memorias-orales`

Baquero, J. (2010). *Lingüística computacional aplicada.* Bogotá: Universidad Nacional de Colombia, Facultad de Ciencias Humanas, Departamento de Lingüística.

Bartens, A. (2002). Another Short Note on Creoles in Contact with Non-Lexifier Prestige Language. *Journal of Pidgin and Creole Languages*, *17*(2), 273–278.

Bartens, A. (2003). *A Contrastive Grammar Islander - Caribbean Standard English - Spanish.* Helsinki: Academia Scientiarum Fennica.

Bartens, A. (2013a). San Andres Creole English. En S. M. Michaelis, P. Maurer, y M. Haspelmath (Eds.), *The survey of pidgin and creole languages. Vol. 1. English-based and Dutch-based languages.* (Oxford Uni ed.). Oxford. Descargado de `http://apics-online.info/surveys/10`

Bartens, A. (2013b). San Andres English structure dataset. En S. M. Michaelis, P. Maurer, y M. Haspelmath (Eds.), *Atlas of Pidgin and Creole Languages Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Descargado de `http://apics-online.info/contributions/10`

Bernal Chávez, J. A., Bonilla, J. E., Rubio, R., Llanos Chávez, A. L., y Bejarano Bejarano, D. E. (2018). Atlas Lingüístico-Etnográfico de Colombia Geolinguistic Corpus. En *V International Linguistics and Language Conference* (pp. 128–141). Istanbul: Eastern Mediterranean Academic Research Center.

Bernal Chávez, J. A., y Hincapié Moreno, D. A. (2018). *Lingüística de corpus*. Bogotá: Instituto Caro y Cuervo.

Bolaños Cuellar, S. (2015). La lingüística de corpus: Perspectivas para la investigación lingüística contemporánea. *Forma y Función*, *28*(1), 31–54. doi: 10.15446/fyf.v28n1.51970

Bowern, C. (2008). *Linguistic Fieldwork: A Practical Guide*. Basingstoke & New York: Palgrave Macmillan. doi: 10.1057/9780230590168

Brigham Young University. (2017). *Corpus of Contemporary American English*. Descargado 2018-05-24, de `https://corpus.byu.edu/coca/`

Calderón Noguera, D. (2008). El corpus del español hablado en Tunja. *Cuadernos de Lingüística Hispánica*, *12*, 17–30.

Colciencias. (2006). Estudio, conservación y práctica de las lenguas aborígenes. En ITEMS Ltda. y R. Polo (Eds.), *75 maneras de generar conocimiento en colombia* (pp. 84–85). Bogotá: Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología Francisco José de Caldas Colciencias.

Crystal, D. (2000). *Language Death*. Cambridge, New York & Port Melbourne: Cambridge University Press.

Datos Abiertos. (2018). *Barrios de San Andrés Isla*. Descargado 24 de mayo de 2018, de `https://www.datos.gov.co/w/ng2u-6iaj/dneh-mcp2?cur=Pc5kuwTny9U{\&}from=root`

Davies, M. (2017). Why Size Alone Is not Enough: The Importance of Historical, Genre-Based, and Dialectal Variation in Language. En *Congreso Internacional de Lingüística Computacional y de Corpus (ponencia)*. Bogotá: Instituto Caro y Cuervo.

Decker, K., y Keener, A. (2001). *A Report on the English-Based Creole of San Andres and Providence Islands, Colombia*.

Dittmann, M. (1992). *El criollo sanandresano: lengua y cultura*. Universidad del Valle.

Dittmann, M. (2013). English in the Colombian Archipelago of San Andres. En T. Hopkins y K. Decker (Eds.), *World Englishes: Vol. III* (cap. 7). London & New York: Bloomsbury.

Edwards, J. D. (1970). *Social Linguistics on San Andrés and Providencia Islands, Colombia* (Tesis Doctoral no publicada). Tulane University.

Forbes, M., y Kouwenberg, S. (2005). Review: A Contrastive Grammar Islander-Caribbean Standard English-Spanish by Angela Bartens. *Language*, *81*(3), 769–770.

García León, D. L. (2011). Las lenguas criollas del Caribe: orígenes y situación sociolingüística. Una aproximación. *Forma y Función*, *24*(2), 41–67. Descargado de `https://revistas.unal.edu.co/index.php/formayfuncion/article/view/38470`

García León, D. L. (2014). Reflexiones en torno a la situación sociolingüística de las lenguas criollas de base léxica inglesa del Caribe. *Forma y Función*, *27*(1), 199–232. doi: http://dx.doi.org/10.15446/fyf.v27n1.46952

Good, J. (2011). Data and Language Documentation. En P. K. Austin y J. Sallabank

(Eds.), *The Cambridge Handbook of Endangered Languages* (pp. 212–234). New York: Cambridge University Press.

Gries, S. T. (2009). What Is Corpus Linguistics? *Language and Linguistics Compass*(3), 1–17. Descargado de http://www.linguistics.ucsb.edu/faculty/stgries/research/2009{_}STG{_}CorpLing{_}LangLingCompass.pdf doi: 10.1111/j.1749-818x.2009.00149.x

Grupo TNT. (s.f.). *Polame Corpus.* Descargado 24 de mayo de 2018, de http://grupotnt.udea.edu.co/polame-corpus/

Grupo TNT. (2018). *Corpus TNT.* Descargado 2018-05-24, de http://grupotnt.udea.edu.co/corpustnt/

Hagemeijer, T., Généreux, M., Hendrickx, I., Mendes, A., Tiny, A., y Zamora, A. (2014). The Gulf of Guinea Creole Corpora. En N. Calzolari y cols. (Eds.), *Proceedings of the Ninth Internatiional Conference on Language Resources and Evaluation (LREC '14).* Reykjavik: European Language Resources Association (ELRA).

Hagemeijer, T., Hendrickx, I., Amaro, H., y Tiny, A. (2012). A Corpus of Santome. En *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMILS8/AfLaT2012).* Istanbul.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? En J. Gippert, N. P. Himmelmann, y U. Mosel (Eds.), *Essentials of Language Documentation* (pp. 1–30). Berlin.

Ho, D. (2016). *Notepad++.* Descargado 28 de mayo de 2018, de https://notepad-plus-plus.org/

Holm, J. (1988). *Pidgins and Creoles, Vol. I, Theory and Structure.* Cambridge & New York: Cambridge University Press.

Infowiki. (2009). *Corpus de Conocimiento.* Descargado 2018-05-24, de https://www.regulacioninformatica.org/wiki/index.php?title=Corpus{_}de{_}Conocimiento

Instituto Caro y Cuervo. (s.f.). *Investigación.* Descargado 28 de mayo de 2018, de https://www.caroycuervo.gov.co/Investigacion/

Islander Creole English Limited Word List. (2006).

Leech, G. (1997). Introducing Corpus Annotation. En R. Garside, G. Leech, y T. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Oxon & New York: Routledge.

Lijffijt, J., y Gries, S. T. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", International Journal of Corpus Linguistics, 13:4 (2008). *International Journal of Corpus Linguistics*, *17*(1), 147–149.

Llisterri, J. (1996). *EAGLES Preliminary recommendations on Spoken Texts/EAG-TCWG-SPT/P* (Inf. Téc.). Expert Advisory Group on Language Engineering Standards.

Llisterri, J. (1999). Transcripción, etiquetado y codificación de corpus orales. *Revista Española de Linguistica Aplicada*, *Extra 1.*, 53–82.

Lozano Ramírez, M. (2012). Breves notas sobre la investigación lingüística en Colombia. *Cuadernos de Lingüística Hispánica*(19), 13–22.

Lüpke, F. (2011). Orthography development. En *The Cambridge Handbook of Endangered Languages* (pp. 312–336). Cambridge.

Manning, C. D., y Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge & London: Massachusetts Institute of Technology.

McEnery, T., y Hardie, A. (2012). *Corpus Linguistics.* Cambridge: Cambridge University Press.

McEnery, T., y Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2.ª ed.). Edinburgh: Edinburgh University Press.

Mello, H. (2014). What Corpus Linguistics can offer Contact Linguistics: the C-ORAL_BRASIL corpus experience. *PAPIA*, *24*(2), 407–427.

Mitchell, D., y Morren, R. C. (2004). *Glossary - U.S. English to Creole* (E. McGowan y R. G. Metzger, Eds.).

Moya, S. (2006). Fi Wii News: A Creole Writing Experience. *Cuadernos del Caribe*(8), 89–96.

Office Open XML. (2018). Descargado 24 de mayo de 2018, de `https://es.wikipedia.org/wiki/Office{\_}Open{\_}XML`

O'Flynn de Chaves, C. (1990). *Tiempo, aspecto y modalidad en el criollo sanandresano.* Bogotá: Colciencias-Universidad de los Andes.

Open Language Archive. (s.f.). *OLAC resources in and about the Islander Creole language.* Descargado 2018-05-24, de `http://www.language-archives.org/language/icr`

Parsons, J. J. (1985). *San Andrés y Providencia: Una geografía histórica de las islas colombianas del Caribe* (3.ª ed.). Bogotá: El Áncora.

Patiño Roselli, C. (2002). Sobre las dos lenguas criollas de Colombia. *Cuadernos del Caribe*(3), 13–18.

Patiño Rosselli, C. (2000). *Sobre etnolingüística y otros temas.* Bogotá: Instituto Caro y Cuervo.

PRESEEA. (s.f.). *Proyecto para el Estudio Sociolingüístico del Español de España y de América.* Descargado 24 de mayo de 2018, de `http://preseea.linguas.net/`

Ramírez-Cruz, H. (2017). *Ethnolinguistic Vitality in a Creole Ecology: San Andrés and Providencia* (Tesis Doctoral no publicada). University of Pittburgh.

Real Academia Española. (2001). *Corpus2.* Descargado 24 de mayo de 2018, de `http://dle.rae.es/?id=AwTBMcs`

Sanmiguel Ardila, R. (2017). *Educación, lengua y cultura en el contexto plurilingüe y multicultural Caribe del Archipiélago de San Andrés, Providencia y Santa Catalina.* San Andrés.

Sanmiguel Ardila, R., Schoch, M., y Pelufo, A. d. M. (2014). *Proyecto de Investigación sobre el creole (lengua criolla de base inglesa) del Archipiélago de San Andrés y Providencia.* San Andrés.

Sebba, M. S., y Dray, S. (2007). Developing and Using a Corpus of Written Creole. En J. C. Beal, K. P. Corrigan, y H. L. Moisl (Eds.), *Creating and Digitalizing Language Corpora. Vol 1: Synchronic Databases* (pp. 181–204). New York: Palgrave Macmillan. doi: 10.1057/9780230223936

SIL International. (2017). *ISO 639-3.* Descargado 28 de mayo de 2018, de `https://iso639-3.sil.org/`

Sinclair, J. (1996). *EAGLES Preliminary recommendations on Corpus Typology/EAG-TCWG-CTYP/P* (Inf. Téc.). Expert Advisory Group on Language Engineering Standards.

Sony. (2018). *Grabadora de voz Sony ICD-PX470.* Descargado 24 de mayo de 2018, de `https://www.sony.com.co/electronics/grabadores-voz/icd-px470`

Tardo, D. S. (2006). Developing the Chavacano Reader Project from the Chavacano Corpus. En *International Conference on Austronesian Linguistics.*

TEI Consortium. (2018). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.*

The Ice Project. (2016). *International Corpus of English.* Descargado 24 de mayo de 2018, de `http://ice-corpora.net/ice/`

The Language Archive. (s.f.). *ELAN.* Descargado 2018-03-08, de `https://tla.mpi.nl/tools/tla-tools/elan/`

Universidad de Antioquia. (2017a). *Grupo de Estudios Sociolingüísticos.* Descargado 2018-05-24, de `http://www.udea.edu.co/wps/portal/udea/web/inicio/investigacion/grupos-investigacion/humanidades/estudios-sociolinguisticos`

Universidad de Antioquia. (2017b). *Grupo de Investigación en Traducción y Nuevas Tecnologías - TNT.* Descargado 24 de mayo de 2018, de `http://www.udea.edu.co/wps/portal/udea/web/inicio/investigacion/grupos-investigacion/humanidades/tnt`

Universidad Nacional de Colombia. (2017). *Centro de Documentación Palabra y Memoria.* Descargado 24 de mayo de 2018, de `http://www.humanas.unal.edu.co/linguistica/laboratorios-y-centros-de-documentacion/centro-de-documentacion-palabra-y-memoria/`

University of Indiana. (2017). *IU Creole Institute.* Descargado 24 de mayo de 2018, de `http://www.indiana.edu/{~}creole/index.shtml`

Villayande Llamazares, M. (2010). *Aproximación a la lingüística computacional* (Tesis Doctoral no publicada). Universidad de León.

Woodbury, A. (2011). Language Documentation. En P. K. Austin y J. Sallabank (Eds.), *The Cambridge Handbook of Endangered Languages* (pp. 159–186). Cambridge University Press.

Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books. Descargado de `http://ota.ox.ac.uk/documents/creating/dlc/`