1. Presentación

Recolectar un corpus del español hablado en la comunidad de habla de Tunja, Boyacá, Colombia, es una investigación de base para la comprensión de los usos dialectales del español en el mundo hispánico. Este proyecto se enmarca dentro de los estudios sincrónicos del habla y se justifica en la necesidad de reconocer y explicar, de manera experimental, la estructura y funcionamiento sociolingüístico del español hablado en Colombia en general, y de Tunja en particular. La investigación tiene como finalidad recolectar un corpus del habla de Tunja, Boyacá, y caracterizar la comunidad de habla que sirvan de material de estudio para analizar los usos dialectales del español en sus variaciones sociofónicas, sociogramaticales, socioléxicas, sociodiscursivas y pragmáticas; teniendo en cuenta variables sociales preestratificadas de nivel de instrucción, edad, sexo y procedencia.

Tal como lo enuncia Vida Castro (2007:7): «*Un buen indicador del conocimiento sobre una lengua y una garantía de que su estudio se aborda con procedimientos rigurosos y actuales es la existencia de corpus lingüísticos extensos y variados que proporcionan una imagen a escala de la actuación lingüística comunitaria. Estos corpus tienen aplicaciones diversas y útiles (computación, e ingeniería lingüística, traducción, elaboración de diccionarios y gramáticas, síntesis del habla, etc); de entre ellas no es desdeñable la que consiste en saber cómo es la estructura subyacente en las múltiples manifestaciones efectivas de una lengua, más allá de la imagen ideada a partir de conjuntos limitados de datos, cuando no únicamente de los procedentes de la propia intuición lingüística del gramático*». Corpus que no ha sido recolectado hasta ahora evidenciando un precario estado del arte en la región, lo que trae como consecuencia estudios aislados, fragmentados y muchas veces impresionísticos, en los que, tal como se expresa en la cita, la intuición del gramático prevalece sobre la evidencia empírica que puede ofrecer el corpus.

Se entiende como corpus la colección digital de datos sociolingüísticos que registran el habla oral de una comunidad. Estos materiales se recogen, etiquetan y clasifican de manera sistemática, de tal manera que puedan ser analizados mediante el uso de programas informáticos. El corpus recogido en la comunidad de habla de Tunja está incluido en el macrocorpus PRESEEA[1], el cual es de carácter panhispánico y pretende obtener muestras representativas de la variación sociolingüística urbana de la lengua española en ciudades capitales de España y de América. Para este fin, en el levantamiento del corpus del español hablado en Tunja, Colombia, se emplea una metodología similar y los mismos parámetros para la recolección de la muestra de Proyecto PRESEEA[2], tal como se explica en los materiales y métodos.

El sistema de signos, que es la lengua, se realiza en el habla contextualizada en sociedades que tienen formas particulares de satisfacer sus necesidades. La lengua es un constructo mental modélico que tiene como finalidad asegurar la comprensión y la comunicación. Como la lengua no existe sin los hablantes, y los hablantes hacen uso aproximante de la lengua debido a que la función pragmática antecede a la función metalingüística, entonces el habla genera variación e incluso cambio en la lengua misma. Este carácter variante es analizado por la sociolingüística, interdisciplina que ha permitido identificar cómo los comportamientos sociales, determinados como variables de sexo, edad, nivel de instrucción, ingresos económicos, modos de vida, etc., influyen en la variación lingüística de un modo específico, en cada comunidad, mediante comportamientos lingüísticos concretos. El carácter complejo del lenguaje, sistematizado en la lengua y actualizado en el habla es definitorio en todos los idiomas del mundo, de ahí que desde mediados del siglo XX se venga revolucionando el enfoque de los estudios sociales en general, y lingüísticos en particular, dejando de lado el estudio de las normas compiladas en las gramáticas, cuya intención era la de homogenizar o establecer una forma única o prestigiosa, para darle paso a estudios del uso lingüístico que permiten conocer la riqueza expresada en la variedad. En esta perspectiva, y aún desde los estudios de la dialectología, los análisis de los usos lingüísticos tienen como finalidad reconocer el aporte de cada variación a la vigencia de la lengua misma.

En el caso específico del español hablado en el departamento de Boyacá, hasta ahora se están empezando a realizar estudios sistemáticos sobre el uso del español

---

[1] Cf. www.linguas.net/preseea

[2] PRESEEA es un proyecto para la creación de un corpus de lengua española hablada, representativo del mundo hispánico en su variedad geográfica y social. Esos materiales se reúnen atendiendo a la diversidad sociolingüística de las comunidades de habla hispanohablantes. PRESEEA agrupa a cerca de 40 equipos de investigación sociolingüística. Es el fruto del trabajo coordinado de investigadores comprometidos con una metodología común para reunir un banco de materiales coherente que posibilite su aplicación con fines educativos y tecnológicos.

en la región, sin embargo esta es aún una tradición insipiente. La ausencia de estudios del habla impide definir y caracterizar los usos lingüísticos de los hablantes de la región para comprender su aporte a la vigencia de la lengua española tanto en Colombia como en el mundo hispánico.

El corpus del habla de Tunja se convierte como material de base para estudios que a mediano plazo pueden afrontar los siguientes problemas de investigación:

¿Cuáles son las características y funcionalidades de los usos dialectales del español en el Departamento de Boyacá, específicamente en Tunja?

¿Cómo aportan al diasistema del español de Colombia los usos dialectales del español en el departamento de Boyacá, comunidad de habla de Tunja?

¿Cómo se comporta la variación sociolingüística en el español hablado en Tunja?

¿El uso dialectal del español en el departamento de Boyacá se puede segmentar diatópicamente?

¿Qué actitudes y creencias tienen los hablantes de Tunja sobre su comunidad de habla?

¿Inciden los usos dialectales en las competencias cognitivas y en la identidad regional?

Estas preguntas permiten planificar el desarrollo de investigaciones a corto y mediano plazo, abriendo un espacio de reflexión a investigadores, docentes y estudiantes de pregrado y postgrado, tanto nacionales como internacionales.

De los contextos sociolingüísticos: individuos, grupos, redes, mercados lingüísticos; las investigaciones en el mundo hispánico privilegian la comunidad de habla. Esta elección permite realizar estudios del lenguaje en complejos urbanos, donde a pesar de la diversidad de los hablantes, existen razones de uso lingüístico determinadas y valoradas socialmente. Rodríguez Cadena (2008:21), afirma que *«El concepto de comunidad de habla está en el centro de las investigaciones sociolingüísticas, pues no sólo es la fuente de los datos lingüísticos sino también el argumento esencial que motiva la búsqueda de la regularidad de la variación y el cambio»*. En acuerdo con la afirmación anterior, el corpus del español hablado en Tunja se levanta desde la comunidad de habla, entendida ésta como *«un conjunto de hablantes que comparten efectivamente, al menos una lengua, pero que, además comparten un conjunto de normas y valores de naturaleza sociolingüística: Comparten unas mismas actitudes lingüísticas, unas mismas*

reglas de uso, un mismo criterio a la hora de valorar socialmente los hechos lingüísticos, unos mismos patrones sociolingüísticos» Moreno Fernández (1988:20), concepto de comunidad que como ya he planteado antes: *«no se centra en los individuos sino en las relaciones comunicativas que estos realizan mediante sistemas de participación grupal o mediante redes estratificadas socialmente. Este comportamiento de la comunidad de habla la convierte en la unidad social más práctica para los estudios sociolingüísticos ya que en el juego de usos en el contexto social, de manera condensada o laxa, es posible reconocer el sistema de reglas y los juicios de valor lingüísticos frente a actitudes y creencias.»* Calderón Noguera (2004:156)

Como los estudios que se realizan con el corpus pueden ser abordados desde la sociolingüística, la sociología del lenguaje, la etnografía del habla o de la comunicación, o desde la etnolingüística; los criterios de recolección del corpus tienen en cuenta variables sociales preestratificadas tales como procedencia, género o sexo, generación y nivel de instrucción; para ser cruzadas con variables lingüísticas mediante el desarrollo de proyectos específicos.

La muestra recolectada en el área urbana de Tunja fue la siguiente:

Variables	Generación 1		Generación 2		Generación 3	
	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
Nivel de instrucción 1	3	3	3	3	3	3
Nivel de instrucción 2	3	3	3	3	3	3
Nivel de instrucción 3	3	3	3	3	3	3

Para un total de 54 hablantes, lo cual ubica la representatividad por encima de 1/ 25.000, tal como sucede con otros corpus del proyecto PRESEEA.

Las variables sociales se clasificaron de la siguiente manera:

Generación: generación 1: Edades comprendidas entre 20 y 34 años; generación 2: Entre 35 y 54 años; generación 3: Más de 55 años.

Nivel de instrucción: nivel de instrucción 1: Primaria, de 0-10 años; nivel de instrucción 2: Secundaria, de 10-14 años; nivel de instrucción 3: Superior, 15 años o más.

Género o sexo: Masculino, femenino.

Procedencia: Caracterizada teniendo en cuenta hablantes nacidos en Tunja que han permanecido en ella por largos períodos de tiempo (10-20 años); no nacidos en Tunja pero llegados a ella antes de los diez años y haber permanecido en ella largos períodos de tiempo (10 o más años).

La realización del proyecto, metodológicamente implicó: selección de hablantes o informantes, observación, construcción de notas y diarios de campo; grabación de entrevistas semidirigidas; trascripción de entrevistas usando el sistema TEI[3]; realización de pruebas lingüísticas tales como: lectura de textos, lista de palabras, pares mínimos, prueba léxica y prueba de formas de tratamiento. Análisis inicial de los materiales identificando variantes sociofónicas, sociogramaticales, socioléxicas, sociodiscursivas y pragmáticas que se configuran como hipótesis previas para estudios posteriores.

Las entrevistas del corpus recogido son semidirigidas, esto significa que se trabaja mediante centros de interés o asuntos generales que se comportan como marcos de temas para que el hablante produzca un discurso de estilo informal. Los centros de interés desde los cuales se motivaron las entrevistas, fueron: el clima, el barrio, los vecinos, la vivienda, la ciudad, la gente que vive en Tunja, problemas de la ciudad, familia y amistad, profesión y trabajo, esparcimiento, costumbres, deseo de mejora económica, narraciones. Estos marcos de temas tienen como propósito establecer relaciones de confianza para iniciar la entrevista; inducir a los hablantes a expresar discursos descriptivos, explicativos, expositivos, argumentativos, narrativos y dialogales; así como discursos hipotéticos y futuros. Con la intención de optimizar la entrevista, esta se graba en la casa del hablante tratando que haya privacidad. La entrevista es realizada por estudiantes de pregrado o postgrado formados mediante un taller de entrevista en ciencias sociales; al evento asisten dos investigadores de campo, uno que realiza la entrevista y otro que levanta notas de campo referidas a factores lingüísticos y sociales.

Las entrevistas tienen una duración promedio de 40 minutos, de tal manera que la duración de este corpus es de 540 minutos aproximadamente. La necesidad de que cada entrevista dure más de 40 minutos obedece a que, según las técnicas de entrevista, a partir de los 20 minutos, el informante puede acercarse a su habla vernácula, que es el objeto básico de estudio de la sociolingüística.

Las entrevistas grabadas son transliteradas o transcritas guardando fidelidad de la forma como se expresan los hablantes, con este fin, la transcripción literal de las

---

[3] The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. Cf. http://www.tei-c.org/index.xml

grabaciones ha sido realizada, esencialmente en ortografía ordinaria. Se prefirió que el entrevistador fuera quien transcribiera ya que él está más familiarizado con la entrevista. Las transcripciones fueron evaluadas tres veces por personas diferentes para garantizar su fidelidad. La transcripción del corpus oral no es una tarea fácil y requiere de un complejo sistema de herramientas de marcación que permiten representar, mediante el código escrito, las características propias del código oral. Esto por las razones que afirma Vida Castro (2007:38): *«los registros orales más espontáneos presentan una serie de características que hacen difícil el acomodo al código escrito: desviaciones de los usos normativos de la lengua utilizada, vacilaciones, presencia de pausas y silencios sin justificación sintáctica, aparición abundante de palabras cortadas y autocorrecciones, suspensiones y abandonos voluntarios de turno... presencia de interrupciones y solapamientos»*. Para minimizar la brecha entre oralidad y escritura se emplean convenciones y etiquetado de los textos mediante el uso de las normas internacionales de marcación textual del *Tex Encoding Initiative: TEI*, el cual es un sistema de transcripción estándar que asume los fundamentos del Standard Generalized Markup Language: SGML. Este sistema permitirá además el manejo computacional del corpus.

Tal como se planteó con anterioridad, el corpus, además de las entrevistas, está compuesto de otras pruebas que no son incluidas en el PRESEEA, y que lo hace aún más rico ya que facilita estudios posteriores en los difrentes niveles sociolingüísticos. Las pruebas adicionales son:

Lectura de textos: los hablantes leyeron el texto *El eclipse* de Augusto Monterroso, texto seleccionado por lo interesante del tema, su brevedad y fácil comprensión. El propósito de esta prueba es identificar la variación sociolingüística expresada en estilos formales escritos.

Lista de palabras y frases: prueba que tiene como finalidad identificar la variación sociofónica usando palabras y frases relacionadas con el entorno sociocultural. Específicamente se usaron nombres de barrios de Tunja, nombres de pueblos cercanos a Tunja y frases sobre tradiciones y cultura como: *Tibaná tierra de ruana y sombrero, Jenesano tierra de gente buena, crecen los feligreses de Boyacá Boyacá, expectativa por la excelente artesanía de Ráquira, triatlón deja exhaustos a los atletas en Monguí, volví a Villa Pinzón varias veces, chorizos, chicha y chicharrones en Chivatá;* entre otras.

Lista de pares mínimos: prueba que permite identificar la variación sociofónica, mediante la pronunciación de palabras con oposición o similitud de pares mínimos: Aptitud/actitud, apto/acto, acción/opción, espiar/expirar, huevo/cuervo, bienes/vienes, caza/cacería; entre otras.

Prueba léxica: se trata de una prueba que presenta 132 definiciones categorizadas por campos semánticos sobre vestido, alimentación, comportamientos sociales, y

transporte. La prueba permite conocer el vocabulario del hablante, reconocer cómo ha escuchado decir a otras personas y cómo evalúa esa forma mediante una escala de valores en la que puede responder: 1. Bien, 2. De mal gusto, 3. Gracioso, 4. Indiferente, 5. Nunca lo usaría. El vocabulario recolectado puede constituir un léxico de uso y puede servir para reconocer actitudes y creencias sobre el léxico usado por individuos no pertenecientes a la comunidad de habla.

Prueba de formas de tratamiento: las formas de tratamiento pueden constituirse como indicios, marcadores o estereotipos de una comunidad de habla, así por ejemplo, el uso del pronombre personal tú puede considerarse como muestra de afecto y cercanía en una comunidad, pero igualmente de irrespeto en otra. Los estudios de formas de tratamiento tienen una importante tradición en el mundo hispánico, en el caso específico del corpus del habla de Tunja, se toman como referencia los pronombres personales: usted, sumercé, sumerced, tú; y su uso se explica mediante razones sociales de: 1. Cercanía Afectiva, 2. Lejanía Afectiva, 3. Respeto, 4. Edad, 5. Estatus superior, 6. Estatus inferior. La prueba refiere al uso en contextos familiares y sociales más amplios tales como trabajo y comunidad. El corpus de formas de tratamiento también aporta información sobre normas de cortesía en eventos comunicativos puntuales tales como: saludos, despedidas, formas de invitación y de ofrecer condolencias.

Las pruebas de entrevistas, lecturas de textos, listas de palabras y frases y pares mínimos, fueron recogidas usando grabadoras de voz digital Sony® ICD-P620. Las demás pruebas fueron diligenciadas por escrito.

Resultados.

Los resultados de la investigación son:

a. El Corpus:

54 entrevistas semidirigidas, grabadas digitalmente.

54 entrevistas transcritas mediante el sistema TEI, que permite analizar datos lingüísticos mediante un software especializado.

50 pruebas grabadas de lectura de textos.

50 pruebas grabadas de listas de palabras y frases.

50 pruebas grabadas de pares mínimos.

54 pruebas léxicas escritas.

54 pruebas escritas de formas de tratamiento.

Las entrevistas en audio, las trasliteraciones, artículos de avances y resultados, y otras informaciones se encuentran disponibles en el enlace:

http://www.uptc.edu.co/facultades/f_educacion/maestria/linguistica/investigacion/preseea/

Este corpus ya está impactando en la comunidad académica de la Universidad Pedagógica y Tecnológica de Colombia, UPTC, tanto a nivel de pregrado, donde aproximadamente 50 estudiantes del programa de Licenciatura en Idiomas Modernos, han sido formados en técnicas de recolección de datos sociolingüísticos; como a nivel de posgrado, donde docentes y estudiantes de la Maestría en Lingüística están ejecutando proyectos de investigación. Esto permite aportar al estado del arte mediante la formación de talento humano.

b. Proyectos de investigación que se ejecutan a partir del corpus:

Los proyectos de investigación que se ejecutan a partir del corpus, son:

Caracterización Sociolingüística de la comunidad de habla de Tunja, por Blanca Nidia Durán Mendivelso de la Universidad Pedagógica y Tecnológica de Colombia de la UPTC.

La modalidad como marca de identidad, por Ofelia Amanda Cárdenas y Nubia Yanira Cárdenas, de la UPTC.

Variación léxica en el habla de Tunja, por Sibel Marcelo Salcedo Cely, de la UPTC.

Marcadores Discursivos en el habla de Tunja, por Fanny Carolina Ortiz Pulido, de la UPTC.

Uso verbal del español hablado en Tunja, por Donald Freddy Calderón Noguera, de la UPTC.

La metáfora en el habla de Tunja, por Donald Freddy Calderón Noguera, de la UPTC

Discurso y género en el habla de Tunja, por Lucía Bustamante Vélez, de la UPTC.

La Cortesía y formas de tratamiento en el habla de Tunja, por Gloria Smith Avendaño, de la UPTC.

Fórmulas de tratamiento en el español de Cundinamarca y Boyacá, trabajo que desarrolla la investigadora Luz Marcela Hurtado Cubillos, del Departament of foreign languages literatures and cultures de la Central Michigan University, EUA.

Este libro, resultado del proyecto *El español hablado en Tunja, Materiales para su Estudio*[4], se pone a disposición de la comunidad académica gracias a la gestión de la Dirección de Investigaciones de la Universidad Pedagógica y Tecnológica de Colombia.

[4] El proyecto es realizado por el grupo de investigación orporación Si Mañana Despierto para la Creación e Investigación de la Literatura y las Artes en su línea de investigación Tradición Oral y el Grupo para el Estudio Socolinguistico del Caribe e Hispanoamérica: GIESCAH