DISEÑO Y ELABORACIÓN DEL SISTEMA GESTOR DE CONTENIDOS PARA LOS CORPUS LINGÜÍSTICOS DEL INSTITUTO CARO Y CUER….

Chapter · November 2017

CITATIONS
0

READS
19

5 authors, including:

Ruth Yanira Rubio López
Caro y Cuervo Institute
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Julio Alexander Bernal Chávez
Caro y Cuervo Institute
13 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Johnatan Estiven Bonilla
Caro y Cuervo Institute
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Estudios Lingüísticos, ISBN 978-959-7152-37-8

DISEÑO Y ELABORACIÓN DEL SISTEMA GESTOR DE CONTENIDOS PARA LOS CORPUS LINGÜÍSTICOS DEL INSTITUTO CARO Y CUERVO

Ruth Yanira Rubio
ruth.rubio@caroycuervo.gov.co
Andrea Lizeth Llanos
andrea.llanos@caroycuervo.gov.co
Julio Alexander Bernal Chavez
julio.bernal@caroycuervo.gov.co
Johnatan Estiven Bonilla
johnatan.bonilla@caroycuervo.gov.co
Daniel Eduardo Bejarano
daniel.bejarano@caroycuervo.gov.co
Grupo de Investigación de Lingüística de Corpus - Instituto Caro y Cuervo

Resumen

Esta ponencia presenta el proceso de diseño, elaboración y consolidación del Sistema Gestor de Contenidos (SGC) para los Corpus Lingüísticos del Instituto Caro y Cuervo desarrollado por el Grupo de Investigación de Lingüística de Corpus y el grupo de las TIC del Instituto. El proyecto tiene por objetivo el desarrollo de una plataforma lo suficientemente flexible para la sistematización, divulgación y explotación de corpus de diversas fuentes y registros (audio, vídeo, imagen, texto). Para ello, definiremos algunos conceptos básicos sobre los corpus y los SGC; presentaremos los materiales producto de las investigaciones del Instituto y las muestras base para el desarrollo de la plataforma. Finalmente, describiremos el proceso de diseño y elaboración del sistema y su estructura actual.

Palabras clave
Sistema gestor de contenidos, corpus, investigación, muestras.

1. Introducción

El Instituto Caro y Cuervo (ICC) tiene como misión proponer y ejecutar políticas para documentar, consolidar y enriquecer el patrimonio idiomático de la Nación. El ICC ha recopilado un acervo de datos de diferentes fuentes y registros producto de varias décadas de investigación sobre el español y otras lenguas de Colombia. Estos

Estudios Lingüísticos, ISBN 978-959-7152-37-8

materiales se compilaron bajo diversos criterios metodológicos y se encuentran almacenados en diferentes formatos. Por lo cual, el acceso, uso y salvaguarda de estos archivos es una tarea compleja y necesaria. En este sentido, el proyecto tiene como objetivo el desarrollo de un Sistema Gestor de Corpus Lingüísticos (SGCL) lo suficientemente flexible para la sistematización, divulgación y explotación de los corpus del ICC; este desarrollo se encuentra a cargo del Grupo de las TIC y el Grupo de Lingüística de Corpus del ICC.

El presente documento inicia con la definición de algunos conceptos básicos sobre los corpus y los SGC. Continúa con un resumen de la naturaleza de los materiales y corpus producto de las diversas investigaciones realizadas en el Instituto. Posteriormente, presenta un esbozo del proceso metodológico del diseño del SGCL. Por último, se expone la estructura del sistema de búsqueda y navegación del SGCL que a la fecha permite la consulta de archivos sonoros de investigaciones representativas de la lengua oral del español de Colombia.

2. Marco conceptual

2.1. Corpus

La Lingüística de Corpus es la encargada de sistematizar y analizar con herramientas computacionales conjuntos extensos de datos de una o varias lenguas bajo criterios lingüísticos, sociales, culturales y literarios, entre otros, denominados Corpus. Con el objetivo de comprender mejor el término, se define un corpus como:

> Una colección de datos lingüísticos, ya sea de textos escritos o de transcripciones de habla grabada, los que pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificación de hipótesis acerca de una lengua. (Crystal, 1991: 32).

Los Corpus se caracterizan por ser una muestra representativa de la lengua, estar formados por datos producidos en situaciones de comunicación, tener criterios explícitos de organización, ser de naturaleza computacional y, por ende, de fácil acceso y explotación. Los corpus pueden dividirse en subcorpus, es decir, divisiones que se efectúan dentro del corpus en general y corresponden a un conjunto de datos con características similares.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

La tipología de los corpus puede establecerse desde diferentes criterios: su diseño, los métodos utilizados para su constitución, las características formales, el medio de producción de los textos, el número de lenguas, etc. En el presente proyecto nos atañe la clasificación de corpus catalogados de acuerdo al medio de producción de los textos, en donde encontramos corpus escritos, orales, mixtos y multimodales.

Los corpus escritos están compuestos por muestras de lengua escrita, su recolección es sencilla y requieren de procesos de escaneo si aún no se encuentran digitalizados. Los corpus orales tienen como objetivo caracterizar desde un punto de vista lingüístico la lengua hablada, se constituyen por transcripciones ortográficas y fonéticas, señales de voz o, en algunos casos, por las grabaciones con sus respectivas transcripciones. Los corpus mixtos combinan las dos modalidades. Finalmente, están los corpus multimodales constituidos por datos textuales, grabaciones sonoras, fílmicas o imagen.

2.2. Sistema Gestor de Contenidos (SGC)

Un SGC es una herramienta de software que permite a usuarios, especialistas y no, almacenar, crear, editar, gestionar y publicar una variedad de tipos de contenidos digitales a un público objetivo. De acuerdo con Boiko (2001 citado por Osuna y Cruz, 2010) los SGC están compuestos por varios subsistemas que se comunican entre ellos: la colección que se encarga de la creación o adquisición de la información; la gestión que maneja y controla los datos, usuarios y los otros subsistemas; y la publicación que se ocupa de los productos finales de información digital.

Los SGC se caracterizan por garantizar que el sistema funcione correctamente; brindan seguridad en cuanto al contenido, la autentificación, los privilegios, etc.; suministran herramientas de soporte que ayudan al usuario con la resolución de problemas o inquietudes; facilitan la ejecución de tareas gracias a las funcionalidades que presenta, favoreciendo el rendimiento de las mismas en relación a los medios que se disponen; contienen un área para la administración de la plataforma, concretamente para la gestión de registros, programación y edición de contenido, administración de plantillas, entre otras; finalmente, se caracterizan por su interoperabilidad y su flexibilidad. (Centro de Apoyo Tecnológico a Emprendedores y Fundación Parque Científico y Tecnológico de Albacete, 2012: 14).

Estudios Lingüísticos, ISBN 978-959-7152-37-8

Teniendo en cuenta las características mencionadas anteriormente, se piensa en un SGC como una plataforma óptima para el manejo de los materiales producto de las investigaciones del ICC, ya que permite organizar de manera eficiente la variedad de contenidos que estas presentan: grabaciones de lenguas indígenas, del español de Bogotá y de aprendientes de lenguas extranjeras; fotografías de algunas localidades de Colombia; grabaciones para estudios fonéticos y fonológicos; y textos del español antiguo de Colombia. El SGC favorece el almacenamiento de estos contenidos en secciones y categorías, facilitando así su navegabilidad y permitiendo una estructura sólida, ordenada y sencilla para los administradores.

La plataforma solventa la necesidad de garantizar la preservación de estos materiales en formatos que no caduquen con el paso de los años; de igual manera, asegura la centralización de estos archivos en un espacio donde se encuentran sistematizados bajo criterios lingüísticos estándar que facilitan el acceso y el aprovechamiento del material para futuras investigaciones. En relación con los usuarios, el sistema cuenta con niveles de acceso que varían de acuerdo al usuario, es decir, dicho acceso será diferente tanto para el administrador, como para el editor o el creador de contenidos. Todos los SGC comparten una estructura básica que permite determinar la pertinencia de la información publicada por un usuario, para que el editor o administrador permita o deniegue su aparición en la plataforma.

3. Materiales y corpus producto de las investigaciones del ICC

Entre los materiales recopilados por el ICC se encuentran muestras y corpus de la siguiente naturaleza (Página Web Instituto Caro y Cuervo):

1. Documentos para la historia lingüística de Colombia, siglos XVI a XIX: Tiene como objetivo proponer un corpus diacrónico de referencia que permita indagar sobre nuestra variedad de habla a través de la historia.
2. Corpus Oral ASLEC – EURP: Tienen a cargo la sistematización de un corpus oral obtenido en la localidad de Ciudad Bolívar y la aplicación de una herramienta creada para la recolección de datos sociolingüísticos en Medellín.
3. Corpus para Estudios fonéticos y fonológicos/Propiedades fonéticas de los estilos de habla en el español bogotano: Se encarga de estudiar la pronunciación del español que se habla en bogotá teniendo en cuenta los siguientes estilos de habla: hiperarticulada, leída, narrada y conversaciones.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

Su objetivo es identificar variaciones dadas por las diversas situaciones comunicativas.

4. Corpus de tradición oral y lenguas indígenas (Tradición oral de los indígenas Pijao del sur del Tolima): Recopila información de tradición oral relacionada con “Cuentos, mitos y leyendas tradicionales de los indígenas pijao del sur del Tolima”, además de temáticas como la medicina tradicional, el menaje, el ajuar doméstico y la vivienda.
5. Corpus recogidos por los estudiantes para el desarrollo de tesis: Materiales y muestras recogidos por los estudiantes de la Maestría en Lingüística del Instituto Caro y Cuervo en sus trabajos finales de tesis. Se espera que estos materiales también queden almacenados y para consulta en el SGCL.
6. Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CAELE2): permite el aprovechamiento de muestras reales de lengua que han resultado de los procesos de aprendizaje de estudiantes de español como lengua extranjera y segunda lengua en el contexto colombiano.
7. Repositorio de grabaciones realizadas en investigaciones anteriores sobre lenguas indígenas y aprendientes de lenguas extranjeras: Grabaciones de lenguas indígenas de Colombia y de aprendientes de lenguas extranjeras de investigaciones hechas en el ICC que están digitalizadas y deben ser sistematizadas y añadidas a la plataforma.
8. Corpus Oral del Atlas Lingüístico-Etnográfico de Colombia (ALEC): Está compuesto por aproximadamente 600 horas de grabaciones de muestras recogidas en 264 localidades del país en el marco de la investigación para la realización del ALEC.
9. Corpus Oral del Habla Culta de Bogotá (HCB): se compone de 400 horas de grabación de encuestas realizadas en la ciudad de Bogotá en el proyecto vinculado al Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica.
10. Corpus Oral del Español Hablado en Bogotá (EHB): El corpus recoge 234 grabaciones de narraciones semilibres y 242 grabaciones de encuestas fonéticas realizadas en 60 barrios de Bogotá. (

Los últimos tres corpus orales que se nombran: ALEC, HCB y EHB son los usados para realizar el pilotaje de la herramienta.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

4. Diseño del Sistema de Gestión de Corpus Lingüísticos

Desde el año 2013, el Grupo de Lingüística de Corpus del ICC inició el proceso de sistematización y almacenamiento de las muestras del ALEC, el HCB y el EHB con el fin de elaborar un corpus oral a partir de estas investigaciones. Para la construcción del corpus se realizaron procesos como: la digitalización de las muestras; definición de metadatos; sistematización y almacenamiento de las grabaciones; e ingreso de metadatos a archivos de Excel.

El archivo de excel se dividió en 4 hojas vinculadas con la tipología de los datos de la grabación: la hoja 1 contenía datos generales y técnicos de las grabaciones (id, comprensibilidad, etc.); la hoja 2 los metadatos de los informantes (edad género, nivel educativo, etc.); la hoja 3 información específica sobre la encuesta (fecha, encuestador, etc.); y la hoja 4 datos sobre las partes de la encuesta y para la anonimización de los audios (tiempo de inicio y finalización del cabezotes, ubicación, etc.). Los metadatos organizados en este archivo se planearon para facilitar su conversión a archivos de texto plano separados por comas (CSV) fáciles de migrar a una base datos.

Al finalizar estos procesos y con la ausencia de una plataforma funcional para la consulta del corpus, teniendo en cuenta los diversos materiales producto de las investigaciones del ICC que necesitaban un espacio de almacenamiento y divulgación se consideró fundamental el trabajo en el SGCL. El desarrollo de este sistema se está llevando a cabo en un trabajo conjunto entre el Grupo de las TIC y el Grupo de Lingüística de Corpus del ICC.

Para el diseño, pruebas y desarrollo del SGCL se tomaron como base las tablas de metadatos ya establecidas y los tres corpus orales (ALEC, HCB, EHB). Asimismo, se hicieron diversas reuniones con los investigadores del ICC para presentar el proyecto, hacer revisión de las investigaciones existentes, obtener sugerencias para el desarrollo del sistema, y plantear los metadatos más generales para el diseño de la base de datos.

5. Descripción del sistema para los Corpus Lingüísticos del Instituto Caro y Cuervo

El SGCL está compuesto por tres componentes principales: la base de datos, la interfaz administrativa y la interfaz del usuario.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

5.1. Base de datos

El desarrollo de la base de datos del SGCL es relacional y se está llevando a cabo en MySQL y PHP. La base de datos está organizada de acuerdo con los metadatos técnicos de las muestras; los metadatos de los informantes; y los metadatos de sesiones. Las llaves primarias tienen tablas como: encuestador, informantes, archivos-audios, corpus, usuarios, entre otros.

5.2. Interfaz administrativa

La interfaz administrativa permite una gestión administrativa flexible, modular e integral. Para esto hay dos tipos usuarios administrativos que tienen permisos de acceso, administración y manejo de datos específicos:

- El administrador que es el encargado de manejar todo el sistema y tiene a su cargo módulos como la gestión de usuarios, gestión de investigadores, gestión de corpus, y gestión de publicaciones y noticias.

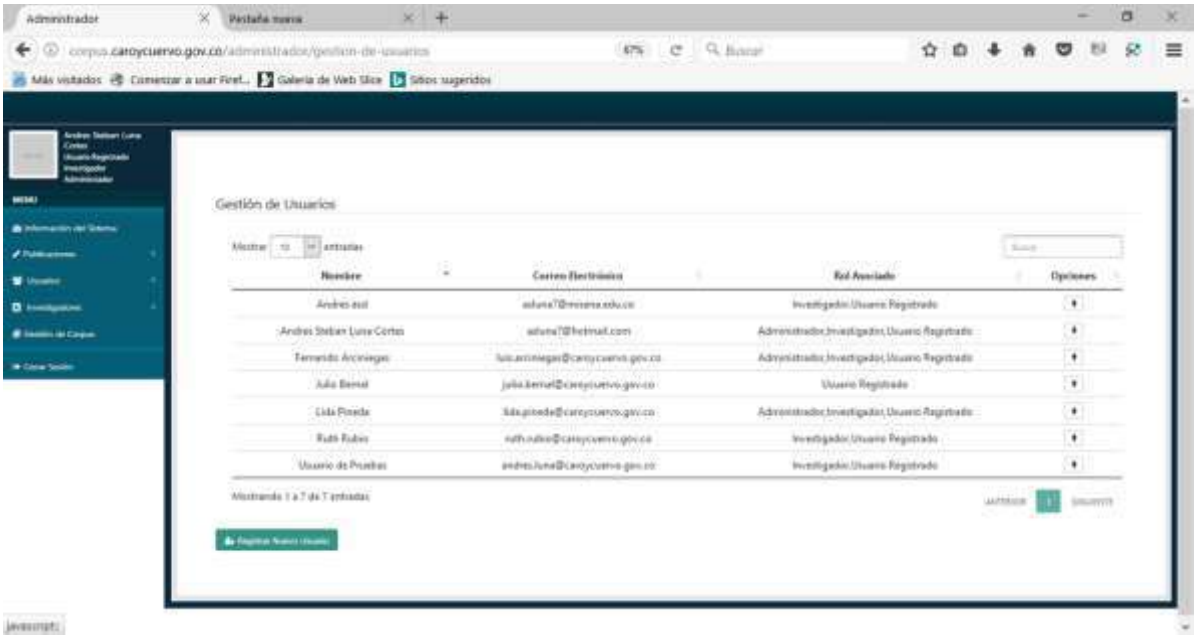

Figura 1. Interfaz administrativa. Módulos del administrador.

- El investigador que se ocupa del ingreso y manejo de nuevos corpus. Para esto tiene a su cargo módulos de registro de corpus, formularios de consulta, manejos y actualización de información asociada a sus corpus.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

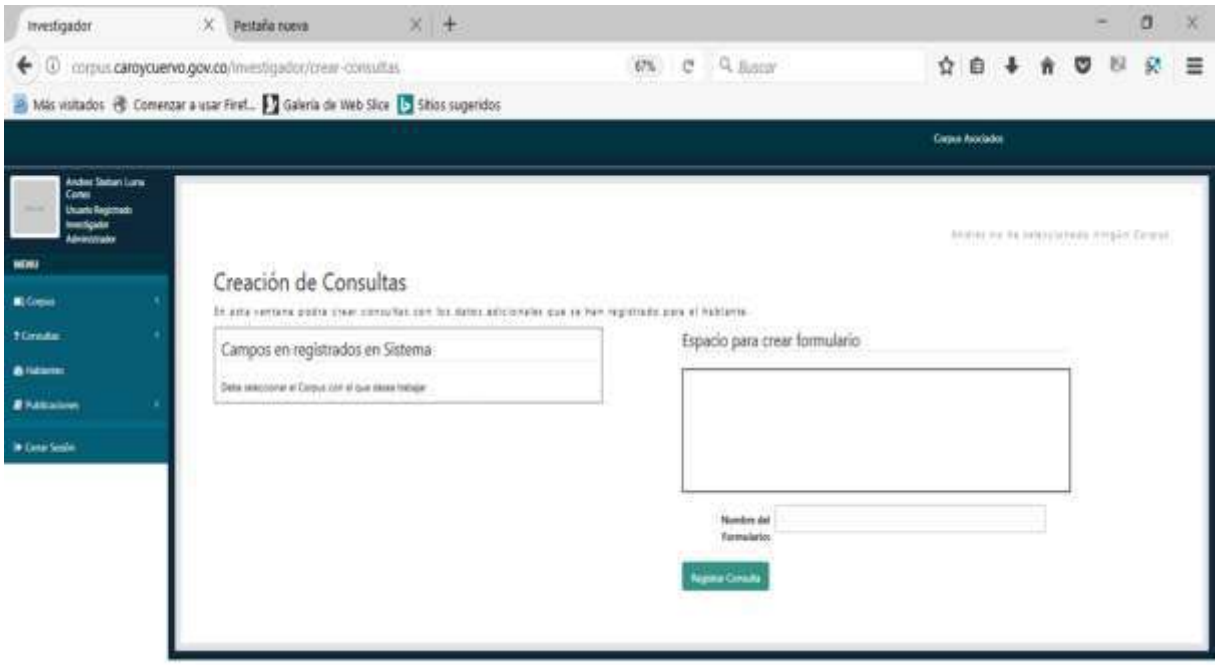

Figura 2. Interfaz administrativa. Módulos del investigador.

5.3. Interfaz de usuario final

La interfaz del usuario presenta una página de inicio y varias páginas que dan acceso a los corpus del Instituto. Asimismo, cumple con las plantillas y los símbolos distintivos del ICC. La interfaz de usuario cuenta con las siguientes pestañas de acceso:

- Inicio: página de inicio que estará conectada con las publicaciones y noticias del sitio central del ICC
- ¿Qué es CLICC?: pestaña de presentación de objetivos y presentación del proyecto.
- Corpus: pestaña de acceso a cada uno de los corpus de la plataforma. Presentación y búsquedas rápidas.
- Recursos: página con información de interés sobre corpus y programas de análisis lingüístico.
- Publicaciones: página de publicaciones relacionada con los corpus.
- Contacto: Información de contacto y envío de mensajes al grupo de lingüística de corpus.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

Figura 3. Interfaz gráfica de usuario. Páginas que la componen.

Con respecto a la consulta de corpus. En la pestaña de corpus es posible acceder a los corpus que se encuentran actualmente en la plataforma: EHB, HCB y ALEC. Al dar clic en uno de los corpus, en la parte izquierda de la página se encuentra la información general y en la parte derecha los dos tipos de búsqueda del corpus que se permiten.

Figura 4. Interfaz gráfica de un corpus.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

Cada corpus tiene acceso a dos tipos de búsqueda: la búsqueda rápida y la búsqueda avanzada. La primera puede ser consultada por cualquier usuario que ingrese a la plataforma. Esta búsqueda se puede hacer por palabra o por tres metadatos básicos: edad, género y lugar.

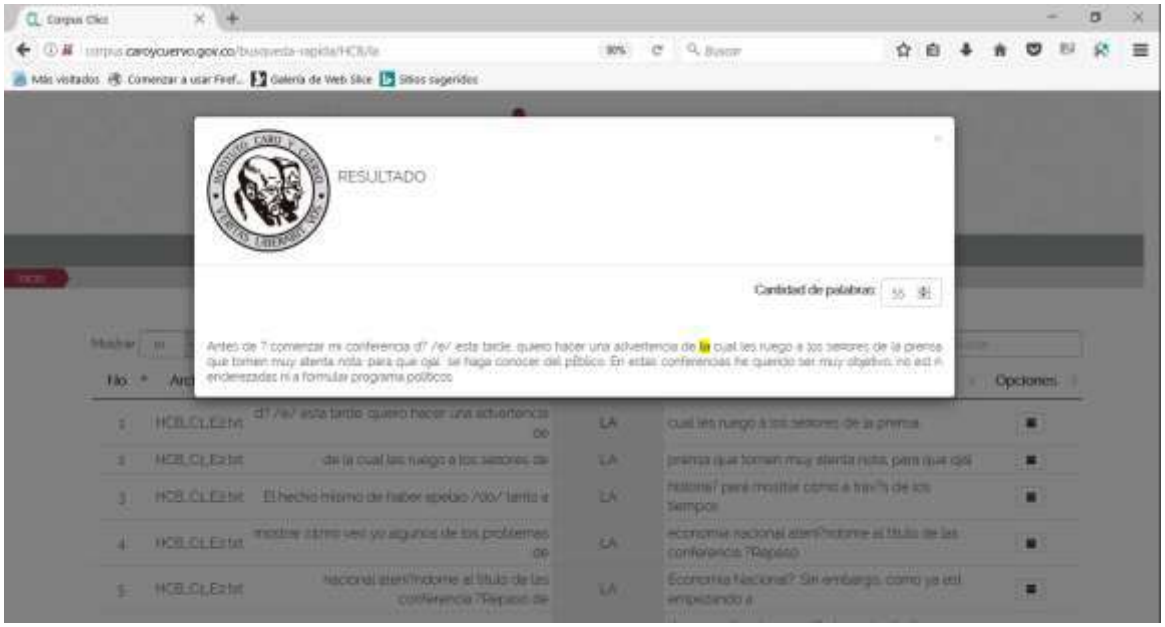

Figura 5. Ejemplo de búsqueda rápida por palabra

La búsqueda avanzada sólo puede ser realizada por usuarios registrados que hayan solicitado acceso al corpus que deseen consultar. Esta consulta permite hacer búsquedas con la mayoría de metadatos que tiene registrado cada corpus. Además, permite descargar el audio y la transcripción en los formatos permitidos por cada corpus.

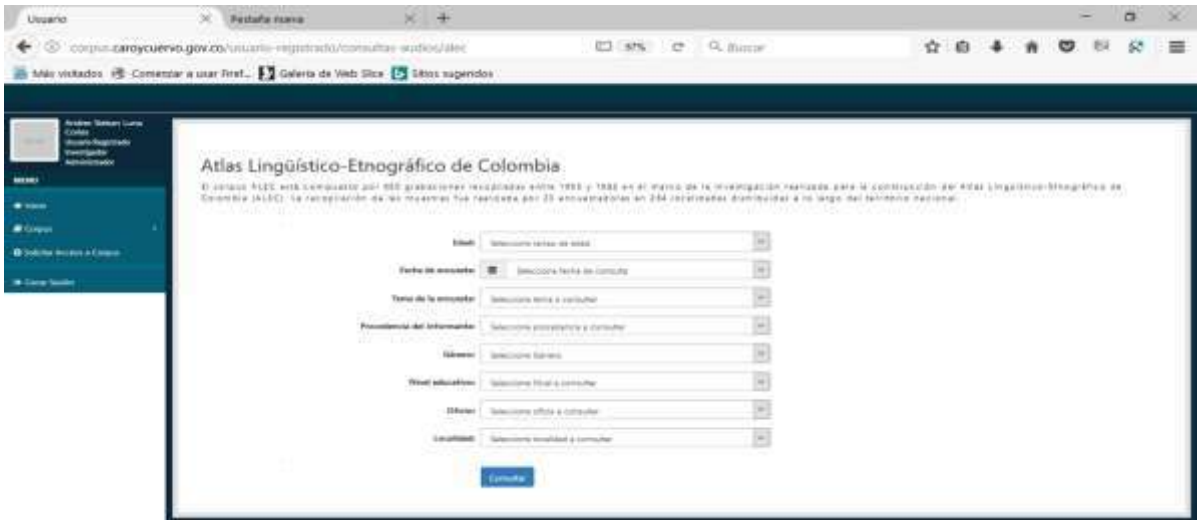

Figura 6. Consulta avanzada de usuario registrado

Estudios Lingüísticos, ISBN 978-959-7152-37-8

6. Conclusiones

Para iniciar, es importante resaltar la necesidad que tiene el ICC de contar con una plataforma que permita almacenar, sistematizar y divulgar los materiales producto de las investigaciones del Instituto, ya que esto facilitará el manejo de estos datos y asegurará su preservación y uso. Igualmente, permitirá la inclusión de tecnologías acordes con las tendencias actuales de investigación.

Hasta el momento el SGCL ha resultado ser una herramienta que se adecúa a los requerimientos para la inclusión de los corpus que han servido de base para el diseño y desarrollo de la plataforma. Se destaca la facilidad de manejo que ofrece y, por ende, el amplio rango de tipos de usuarios y búsquedas que permite. Adicionalmente, teniendo en cuenta la diversidad de la información que será ingresada al SGCL se ha hecho énfasis especial en diseñar una base de datos flexible y que incluya la mayor cantidad de metadatos posible. La base de datos es relacional lo que permite que no haya pérdida de datos, fallas en el sistema y duplicación de información.

Por otro lado, es necesario aclarar que la herramienta aún se encuentra en su versión alfa de desarrollo, se deben realizar varias pruebas con grandes cantidades de datos, mejorar las relaciones de la base de datos, e implantar pruebas iniciales de usabilidad para responder a las exigencias de los usuarios y garantizar su fácil uso y navegación. Además, aún falta desarrollo en los módulos de cada rol que componen el sistema y en la interfaz de investigador para el ingreso y administración de nuevos corpus.

Dentro de la proyección a dos años se propone el perfeccionamiento del diseño y desarrollo tanto de la interfaz administrativa como de la interfaz de usuario. Se planean asesorías y trabajo con los grupos de investigación para el paso de los materiales de sus investigaciones a la base de datos y posterior trabajo con el SGC; así como un asesoramiento constante para el desarrollo del mismo. Para llevar a cabo estos procesos se inició la revisión teórica y construcción de protocolos de transcripción ortográfica y de ingreso de nuevos corpus a la plataforma por parte de los investigadores. Esto permitirá el manejo y sistematización de datos siguiendo los estándares más utilizados actualmente.

Estudios Lingüísticos, ISBN 978-959-7152-37-8

7. Bibliografía

Crystal, D. (1991). *The Cambridge Encyclopedia of Language*. Cambridge, Cambridge University Press.

Centro de Apoyo Tecnológico a Emprendedores, Fundación Parque Científico y Tecnológico de Albacete. (2012) *Estudio de los sistemas de gestión de contenidos web. Análisis de las mejores soluciones del mercado*. Albacete, España.

Instituto Caro y Cuervo. (2017). Descripción de proyectos. Recuperado de: http://www.caroycuervo.gov.co/Investigacion/

Osuna, M. y De la Cruz, E. (2010). "Los sistemas de gestión de contenidos en Información y Documentación", en *Revista General de Información y Documentación*, Vol. 20 (2010). Editorial Universidad Complutense de Madrid, Castilla - La Mancha, pp. 67-100.

Villayandre, M. (2010). "Aproximación a La Lingüística Computacional." Tesis doctoral. Universidad de León. Web.