Propuesta recibida en julio 2023 y aceptada para publicación en noviembre 2023.

Corpus lingüísticos del Instituto Caro y Cuervo (CLICC): una plataforma en línea para el almacenamiento, sistematización y consulta de corpus

Linguistic Corpus of the Caro y Cuervo Institute (CLICC): an online platform for corpus storage, systematization and consultation

Ruth Yanira Rubio López
Instituto Caro y Cuervo

Andrés Steban Luna Cortés
Instituto Caro y Cuervo

Nathalia Solano-Guzmán
Instituto Caro y Cuervo

Resumen

Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) es una plataforma en línea para el almacenamiento, sistematización, administración y consulta de corpus, que nació con el objetivo de contar con un espacio para la salvaguarda de los archivos producto de las investigaciones del Instituto y que, actualmente, está disponible para que investigadores, comunidades o personas interesadas puedan publicar sus corpus sobre las lenguas de Colombia. CLICC es un espacio de acceso libre dirigido a público general y especializado interesado en explorar y contribuir a la documentación de la diversidad lingüística y cultural de Colombia. En este documento se describen sus características, funcionalidades, consultas y perspectivas futuras. También se explican los diversos ajustes que se han hecho para garantizar la publicación de corpus de distintos tipos, el respeto por los permisos y singularidad de cada corpus, y el aprovechamiento futuro de los archivos con fines diversos.

Palabras clave

corpus, herramientas de gestión de corpus, lingüística de corpus, diversidad lingüística y cultural

Abstract

Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) is an online platform for corpus storage, systematization, administration, and query. CLICC was created with the objective of safeguarding the files produced by research conducted by the Institute. Now, it is available so that researchers, communities, or interested people can publish their corpora of Colombian languages. CLICC is a free access space open to the general and specialized public interested in exploring and contributing to the documentation of Colombia's linguistic and cultural diversity. This document describes its characteristics, functionalities, consultations, and future perspectives. It also highlights the various changes that have been made to guarantee the publication of different types of corpora, the permissions and singularity of each corpus, and the future use of the documents for various purposes.

Keywords

corpus, corpus management tools, corpus linguistics, linguistic and cultural diversity

1. Introducción

Hoy en día, los corpus son el insumo principal para multiplicidad de tareas y objetivos, entre los que podemos encontrar la investigación lingüística, la elaboración de diccionarios, la documentación, el procesamiento de lenguaje natural, que son posibles gracias a la divulgación, fácil acceso y reutilización de corpus. Como bien lo mencionan Torruella & Llisterri (1999, p. 29), "dado el esfuerzo económico y humano que supone la creación de un corpus, parece lógico pensar en que éste debe poder ser reutilizado por otros investigadores y para fines diferentes a los que fue concebido." De ahí la importancia de contar con plataformas que permitan la sistematización de los corpus, como también su divulgación y uso con fines diversos.

Algunas plataformas de este estilo son Sketch Engine (Kilgarriff et al., 2014), Gestor de Corpus (GECO) (Sierra et al., 2017), Linguistic Tools (Caminada et al., 2008), Susteainibility Platform for Linguistic Corpora and Resources (SPLICR) (Rehm et al., 2008), y la plataforma virtual que presenta este artículo, Corpus Lingüísiticos del Instituto Caro y Cuervo, en adelante CLICC.

DOI: 10.21814/lm.15.2.407

La plataforma CLICC[1] se creó con el objetivo inicial de contar con un espacio suficientemente flexible para la salvaguarda, divulgación y análisis de corpus recopilados en las distintas investigaciones realizadas en el Instituto Caro y Cuervo (Rubio et al., 2017). CLICC permite el almacenamiento, administración, análisis y publicación de corpus de distintos tipos, por parte de investigadores del Instituto. Próximamente, se espera que otros equipos de investigación, comunidades o cualquier persona interesada puedan cargar sus datos en la plataforma.

En este documento se presentan las razones de creación de CLICC y su estado actual: la estructura, funcionalidades, los corpus cargados hasta el momento, consultas, ajustes y mantenimiento que se ha realizado desde su desarrollo (2017). Además, se esclarecen algunas perspectivas para el mejoramiento de la plataforma.

2. ¿Por qué la creación de CLICC?

Tras varias décadas de investigación, el Instituto Caro y Cuervo (ICC) ha recopilado un amplio acervo de datos sobre las lenguas de Colombia (fotografías, grabaciones, manuscritos, entre otros). Muchos de estos materiales tienen potencial de convertirse en corpus, o directamente se han recopilado con los métodos de la lingüística de corpus. CLICC se creó, precisamente, para garantizar la salvaguarda de estos materiales, facilitar su almacenamiento y administración en un mismo espacio, y permitir la publicación y aprovechamiento de los corpus para fines diversos. Al respecto es importante tener en cuenta varios aspectos. Primero, CLICC funciona completamente en línea, almacena la información en los servidores del ICC, tiene una monitorización y mantenimiento por parte del grupo de las TIC, y respeta los permisos y manejo de la información de cada corpus. Así pues, sigue la misión del Instituto de salvaguardar el patrimonio lingüístico de Colombia, a la vez que ofrece un servicio gratuito con el que se disminuye el riesgo de pérdida de información valiosa. Segundo, el sistema permite en un mismo espacio virtual varias funciones que suelen encontrarse por separado en distintas herramientas: la creación y administración de los corpus, similar a *Sketch engine* (Kilgarriff et al., 2014); la adición de metadatos de las muestras y de sus hablantes, función propia de programas como *SayMore*;[2] hacer consultas de frecuencias, concordancias y colocaciones, como *AntConc* (Anthony, 2013) o *WordSmith* (Scott, 2008); y realizar consultas por los metadatos de cada corpus. Asimismo, para el análisis de corpus textuales se instaló parte del código de etiquetado morfosintáctico de *Freeling* (Padró & Stanilovsky, 2012) y *Treetagger* (Schmid, 1994) en el servidor de producción, y se creó un *script* para ejecutar estos archivos desde CLICC. Esto es lo que facilita la realización de consultas por palabra y colocaciones de corpus textuales u orales con transcripción. Tercero, la plataforma permite divulgar el corpus para el público general y hacer consultas generales y especializadas (en el usuario registrado) (Ver sección 3). Esto permite que corpus que se han recopilado para investigaciones específicas puedan ser consultados o utilizados para otras investigaciones o con otros fines; por ejemplo, para su uso en materiales de enseñanza de lenguas.

3. Estado actual de la plataforma CLICC

3.1. Estructura y funcionalidades

Para la estructuración y requerimientos iniciales de la plataforma, se tomaron como referentes los corpus orales del Atlas Lingüístico y Etnográfico de Colombia (ALEC), del Habla Culta de Bogotá (HCB) y del Español Hablado en Bogotá (EHB),[3] construidos a partir de los datos recopilados en investigaciones realizadas en el Instituto entre los años 50 y los 90 del siglo XX (Rubio & Bernal, 2019). La plataforma, entonces, estaba organizada para la publicación de estos corpus y estaba compuesta por la base de datos, la interfaz administrativa y la interfaz de usuario. La base de datos estaba organizada de acuerdo con los metadatos de informantes y de sesiones de las muestras, y las llaves primarias tenían tablas como encuestador, informantes, usuarios, entre otras (Rubio et al., 2017). Sin embargo, con el objetivo de que CLICC funcione para corpus de diversos tipos y con permisos de acceso diferentes, se añadió una interfaz para investigadores y una interfaz para usuario registrado. Actualmente, CLICC está compuesta por:

- una base de datos relacional desarrollada en MySQL y PHP;
- la interfaz de usuario general (IUG) en la que cualquier usuario puede conocer CLICC y consultar los corpus publicados;
- la interfaz de usuario registrado (IUR), que permite hacer consultas especializadas con ba-

[1] https://clicc.caroycuervo.gov.co

[2] https://software.sil.org/saymore/

[3] Para ver la lista de corpus creados y listos para cargue de datos en CLICC vea lo Cuadro 1.

TIPO	SIGLA	NOMBRE DEL CORPUS
Monolingües y orales sin transcripción	ASMYCU	Acervo de tradición oral afrocaucano "Manuel y Constanza USSA"
	CAELE/2 ORAL	Corpus del Español como Lengua Extranjera y Segunda - Oral
	EHB	Corpus del Español Hablado en Bogotá
	DDALN	Diplomado de documentación audiovisual de lenguas nativas
Monolingües y orales con transcripción	ALEC	Corpus del Atlas Lingüístico-Etnográfico de Colombia
	HCB	Corpus del Habla Culta de Bogotá
	EURP	Corpus de Espacios Urbanos de Restablecimiento Poblacional
	CLC	Corpus de habla leída y conversacional del español de Colombia
Monolingües y textuales	LVBC	Corpus Literatura de la Violencia Bipartidista en Colombia
	CLEC	Corpus Léxico del Español de Colombia
	CAELE/2 ESCRITO	Corpus del Español como Lengua Extranjera y Segunda-Escrito
	DHLC	Corpus de Documentos para la historia lingüística de Colombia
Bilingüe oral y textual	CLS	Corpus de la Lengua Sáliba

Cuadro 1: Corpus creados en CLICC

se en los metadatos de cada corpus y la descarga de archivos dependiendo de los permisos otorgados para cada corpus;

- la interfaz de usuario investigador (IUI) para el ingreso y administración de corpus;
- y la interfaz administrativa (IA) que permite la gestión de usuarios, corpus y accesos.

De la misma manera, CLICC cuenta con varios tipos de usuarios que tienen funcionalidades y permisos específicos, con el objetivo de garantizar un entorno colaborativo que respete las políticas de seguridad y las características de cada corpus, y que sea de uso intuitivo para cualquier tipo de usuario (experto o no). En cuanto a la base de datos, inició siendo relacional hasta la tercera forma, es decir, seguía unas guías de diseño más sencillas debido a las características similares que tenían los corpus orales de prueba. Sin embargo, la adición de nuevos corpus y metadatos que no compartían atributos comunes llevó a la necesidad de normalizarla hasta la sexta forma normal. Esto implicó identificar y eliminar todas las relaciones multivariadas en la base de datos para permitir el registro de una tabla que almacenara los atributos. Luego, la información se registró en otra tabla en la que se almacenarían las respuestas a los atributos previamente identificados y relacionados con la entidad de los nuevos corpus registrados.

3.2. Corpus creados en CLICC

Hasta el momento, en la plataforma se han creado 13 corpus de distintos tipos (Ver cuadro 1),que categorizamos siguiendo algunos parámetros de Torruella & Llisterri (1999). Actualmente, el investigador solicita la creación de su corpus y recibe capacitación por parte del Grupo de Lingüística de Corpus y Computacional (LICC) para realizar los siguientes procesos: sistematización de archivos, cargue y registro de metadatos y muestras; definición de metadatos de búsqueda general y especializada; ingreso de la información general del corpus; y, solicitud de la publicación del corpus. Los corpus creados en CLICC se encuentran en momentos diferentes de este proceso y, como se puede evidenciar, todos se centran en las lenguas de Colombia, por lo que la plataforma se consolida como un repositorio para la salvaguarda y estudio de la diversidad lingüística y cultural del país.

3.3. Ingreso y administración de corpus

En la IUI se ecuentran los menús para funciones vinculadas con el ingreso y administración de corpus. Primero, en la ventana principal de la IUI se encuentra el menú "Mis corpus" para escoger el corpus a trabajar. Al seleccionarlo, se pueden visualizar la cantidad y últimas sesiones consultadas por los usuarios registrados con acceso al corpus (Ver fig. 1). También, se encuentra la información general del corpus: la cantidad de hablantes, transcripciones, sesiones, etc., que se han registrado en el sistema. Esto permite visualizar el tamaño del corpus y también el avance en su construcción.

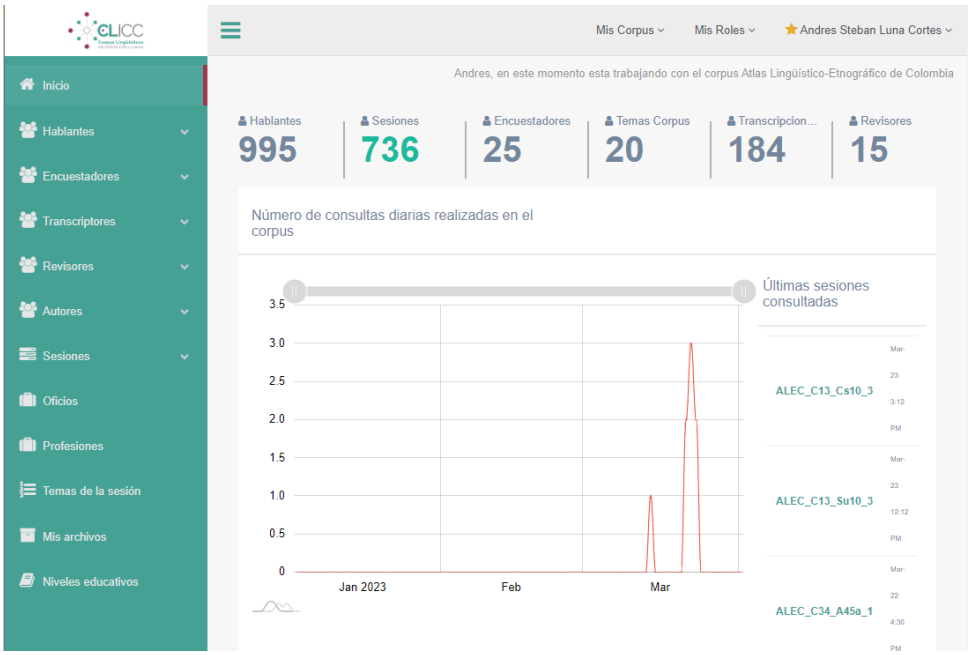

Figura 1: Interfaz de investigador CLICC

Luego, el sistema cuenta con un espacio para el almacenamiento de las muestras en distintos formatos que se encuentran en la pestaña "Mis archivos" (Ver figura 1). CLICC genera un conjunto de carpetas de acuerdo con el tipo de corpus; por ejemplo, uno oral cuenta con carpetas para los audios en formato MP3 y WAV, para las transcripciones en formato TXT, PDF, DOCx y EAF,[4] y para los anexos que incluyen una carpeta para los consentimientos informados. Los formatos tienen unos objetivos específicos dentro del sistema y no todos son obligatorios: el TXT para las consultas, los PDF que se visualizan en el usuario general, y los MP3 para agilizar la consulta de los audios, por su peso reducido. Los otros formatos pueden facilitar el aprovechamiento de las muestras de distintas maneras o con otros fines; por ejemplo, la transcripción en EAF puede facilitar el análisis lingüístico y exportación a otros formatos y el audio en WAV es más recomendado para la realización de estudios fonéticos (si el corpus es oral).

Por otro lado, en la parte izquierda de la interfaz están las distintas pestañas para el ingreso de metadatos de las muestras y de informantes (Ver figura 1). Estos metadatos se definieron a partir de los corpus iniciales de la plataforma. Sin embargo, no son obligatorios: si el corpus requiere otro tipo de metadatos se pueden definir y el sistema generará una plantilla de Excel para su registro. El ingreso de los metadatos se puede hacer diligenciando los espacios establecidos en la plataforma o a través de una plantilla de Excel, para facilitar la carga masiva de datos. Finalmente, para que las muestras, metadatos y todos los archivos de una sesión estén conectados, al terminar el ingreso de datos se realiza un proceso de registro de sesión que los vincula a todos, lo que permite las consultas del sistema. En cuanto a la administración de los corpus, al dar clic sobre el nombre de su usuario, el usuario investigador encontrará un menú que le permite:

1. gestionar los usuarios registrados que han solicitado acceso al corpus (aceptar o eliminar permisos);
2. solicitar el registro para iniciar la creación de un corpus nuevo;
3. solicitar acceso a otros corpus como usuario investigador;
4. ingresar la información de su corpus, que aparecerá en la IUG (descripción, metodología, equipo, cómo citar);
5. y seleccionar qué formatos y cuántas descargas diarias serán permitidas a los usuarios que tiene acceso al corpus (si aplica).

Teniendo en cuenta que para la creación del corpus puede haber varios usuarios involucrados y con distintas tareas, existe un usuario líder de corpus. Este será quién tenga todas las funcionalidades de administración: los numerales 1, 4 y 5 de la lista previa son exclusivos de este tipo de usuario. Este líder o líderes de corpus se marcan con una estrella amarilla al lado de su nombre (Ver Figura 1).

[4] Formato en el que quedan almacenadas las transcripciones realizadas con el programa ELAN.

Figura 2: Descripción general y consultas del corpus CorlexCO

3.4. Consultas de los corpus

La IUG es el espacio de consulta en línea para cualquier usuario interesado en conocer CLICC y consultar sus corpus. En la página web se presentan los objetivos, equipo, y guías de uso de la Plataforma. Asimismo, en la pestaña de cada corpus el usuario encontrará la descripción general del corpus, la sección de consultas, la metodología, el equipo y cómo citar (Ver Figura 2)

Para consultar cada corpus, están disponibles tres tipos de búsqueda: por palabra, también conocida como KWIC (*Key Word In Context*); por metadatos, que en la interfaz general suelen ser dos o tres datos relevantes de acuerdo con el tipo de corpus; y por colocaciones. (Ver Figura 3).

Las consultas aparecerán de acuerdo con la información que se haya ingresado del corpus; por ejemplo, el corpus ASMYCU, hasta el momento, cuenta con los audios sin su transcripción, por lo que se pueden consultar a partir de dos metadatos (temas y lugar de la encuesta). Si a futuro el corpus es transcrito se habilitarán los otros dos tipos de búsqueda, como en el corpus del ALEC.

En cuanto a los resultados, generalmente se muestra la cantidad, el identificador del archivo de la muestra, los resultados y un visor para escuchar y ver la transcripción o el texto. Cuando hay un texto o transcripción, suele ir acompañada en la parte inicial de una ficha de metadatos de la muestra.

Las consultas y resultados que hemos mencionado hasta ahora son las que puede hacer cualquier usuario interesado al ingresar a la plataforma. Sin embargo, en este espacio solo podrá visualizar la información. Para tener otras funcionalidades la persona se puede registrar en el sistema, lo que le permitirá:

- **Hacer consultas especializadas de metadatos:** el usuario registrado suele tener más metadatos disponibles para las consultas. Por ejemplo, en el corpus HCB el usuario registrado, además de los metadatos generales, puede buscar por tema, nivel educativo, o profesión de los informantes.
- **Descargar archivos:** de acuerdo con los permisos del corpus, se pueden descargar muestras o transcripciones en distintos formatos.
- **Visualizar la anotación POS automática de las muestras o transcripciones:** si el corpus tiene transcripción o es de tipo textual, el sistema hace la anotación con los etiquetadores de *Freeling* y *Treetagger*.

Al consultar una muestra el usuario registrado podrá elegir y visualizar la anotación con cualquiera de los dos etiquetadores (Ver Figura 4)

3.5. Seguridad del sistema

Para poder garantizar la seguridad de la información de los corpus, se descartó la opción de crear CLICC en un sistema de gestión de contenidos (CMS). En su lugar, se inició un proceso de desarrollo que siguió todas las fases de la ingeniería de sistemas y se implementó en servidores propios del Instituto para así poder brindar soporte y garantizar la actualización del sistema. Adicionalmente, se implementaron las siguientes medidas:

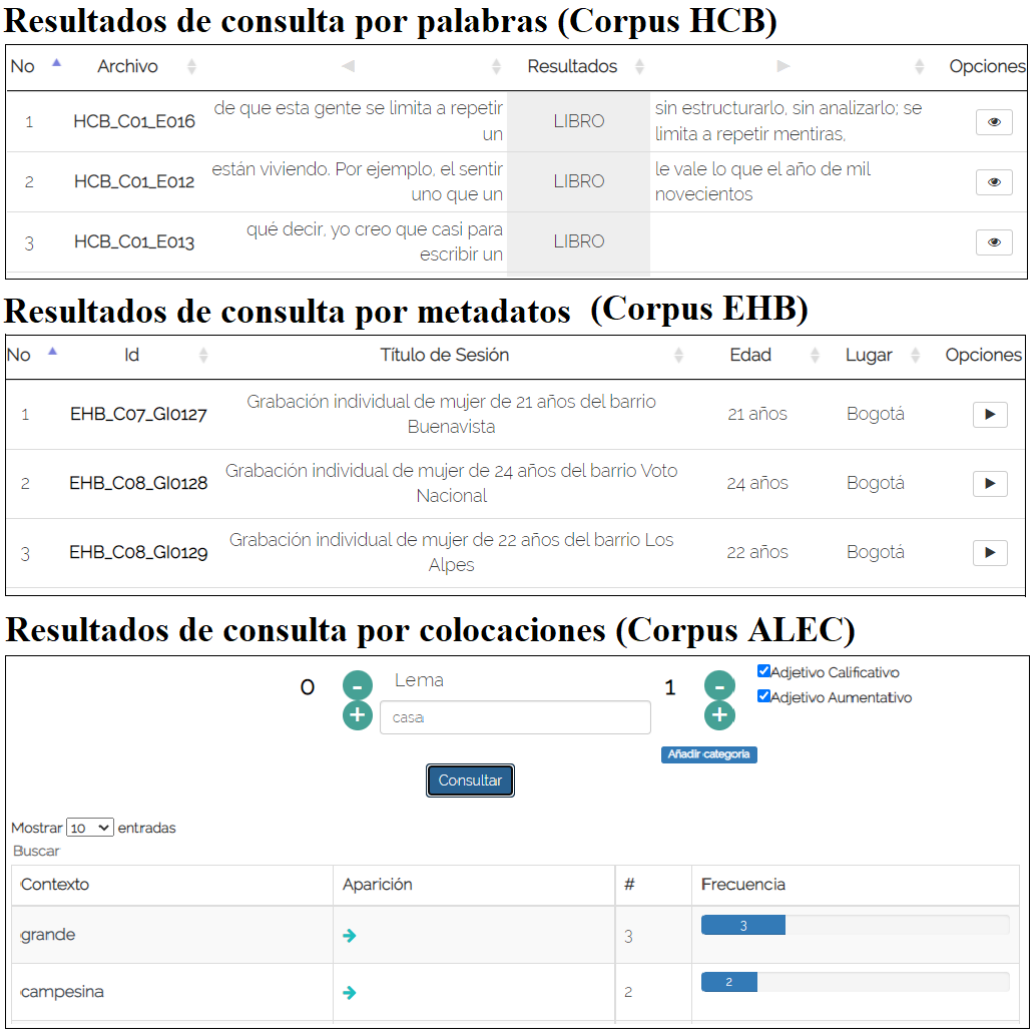

Figura 3: Resultados de consultas en la interfaz de usuario general en CLICC

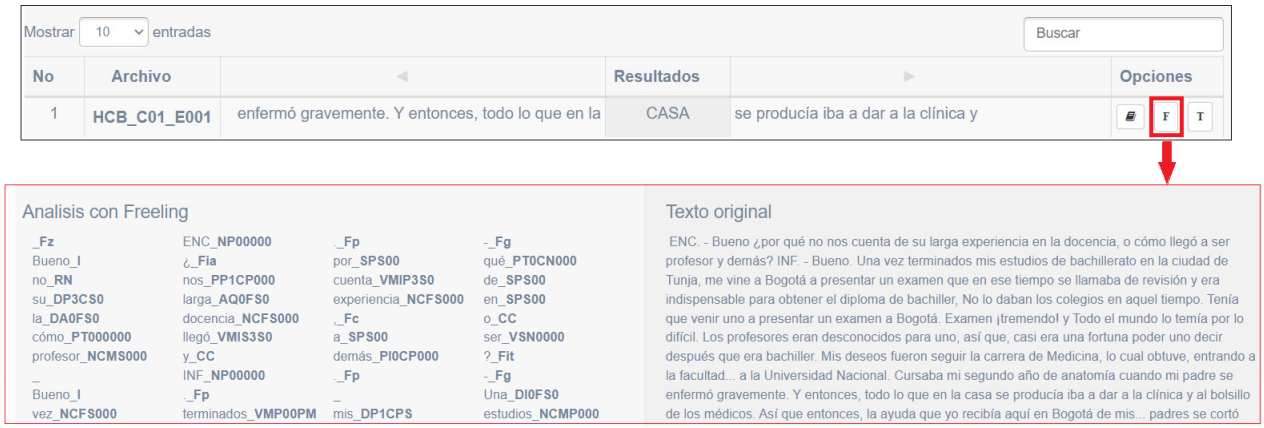

Figura 4: Visualización en IUR de opciones para consulta de etiquetado morfosintáctico y visualización de una muestra con Freeling

- **Contraseñas seguras:** Las contraseñas que se utilizan para acceder a cualquier tipo de usuario de CLICC son creadas por el propio usuario. Para asegurar la protección de la información, se requiere que las contraseñas contengan como mínimo 8 caracteres, combinando mayúsculas y minúsculas.
- **Permisos de usuario:** El sistema se encuentra modularizado, lo que permite la coexistencia de los 4 roles que se pueden asignar a cada usuario y, así, especificar a qué información tiene acceso. Adicional a esto, el sistema permite hacer el seguimiento de las acciones realizadas por cada usuario durante su sesión, ya que cuenta con un registro detallado de las mismas.
- **Verificación de archivos:** El módulo dedicado al cargue de los archivos de los corpus ha sido diseñado para permitir solamente la carga de extensiones de archivo previamente configuradas en el código, de manera que cualquier archivo malicioso será automáticamente rechazado por el servidor y no se permitirá su ingreso al sistema.

3.6. Políticas de seguridad y uso

En CLICC se ha implementado la norma ISO 27001 (ISO, 2022) con el fin de establecer políticas de seguridad robustas y efectivas para proteger los datos almacenados. Entre las políticas implementadas, destacan las siguientes:

- **Identificación y evaluación de riesgos:** Constantemente se monitorea el sistema para identificar cualquier amenaza o vulnerabilidad, lo que permite establecer controles de seguridad adecuados y reducir posibles riesgos.
- **Acceso y control de acceso:** El administrador del sistema es quien tiene la capacidad de permitirlo o rechazarlo. De esta manera, se asegura que solamente el personal autorizado pueda acceder e interactuar con el código, el servidor y la base de datos.
- **Control de acceso físico:** Se aplica una restricción del acceso a los puntos físicos de servidores y centros de datos con el fin de garantizar la protección de los recursos y la información almacenada.
- **Copias de seguridad y recuperación de desastres:** Las copias de seguridad se realizan en el servidor principal ubicado en la sede principal del instituto y, adicionalmente, se crea una copia de respaldo en los servidores alojados en la sede alterna para garantizar la disponibilidad y la recuperación de la información ante posibles contingencias.

Respecto a las políticas de uso, para poder registrarse en CLICC, los usuarios deben aceptar unas condiciones de uso que incluyen compromisos respecto a la confidencialidad del código, el uso de la información con fines no comerciales, entre otros. Para la publicación de los datos, se ha procurado que los corpus cuenten con consentimientos informados de los participantes, y se está avanzando en la definición de mecanismos legales para el cumplimiento de las leyes colombianas de tratamiento de datos y la protección de derechos de autor en la publicación de corpus antiguos y nuevos.

4. Perspectivas futuras

El mejoramiento de la plataforma está orientado actualmente a dos frentes: la adaptación de la plataforma para almacenamiento y consulta de corpus bilingües, y la creación e implementación de un *pipeline* para procesamiento del español. Respecto a los corpus bilingües, se partió de la sistematización de Corpus de Lengua Sáliba, siguiendo los flujos de trabajo con ELAN y FLEx de Gaved & Salffner (2014) y Pennington (2014), y procurando que queden disponibles tanto las transcripciones de audios alineadas en las dos lenguas (sáliba y español), como la información especializada que pueden ofrecer investigadores y sabedores de las lenguas nativas (como la glosa morfológica o la transcripción en diferentes sistemas de escritura). Se están realizando análisis de datos y pruebas con los archivos resultantes de dicha sistematización (txt, eaf y xml) para consolidar un protocolo de ingreso de corpus bilingües y el desarrollo a futuro de una consulta por palabras que ofrezca resultados alineados en las dos lenguas. Por otro lado, se está construyendo un *pipeline* de procesamiento para lematización, etiquetado de partes del discurso y parseado de dependencias, mediante el uso de una arquitectura avanzada de redes neuronales, que sustituirá a *Freeling* en CLICC. Esta tecnología se fundamenta en AnCora (Taulé et al., 2008), como modelo general del idioma español, y en un modelo propio calibrado para el español colombiano, construido a partir de experiencias previas con otros corpus como el Corpus Oral y Sonoro del Español Rural.[5] Estos corpus se están consolidando como referentes para el afinamiento de los modelos de procesamiento del lenguaje natural en español, los cuales suelen ser entrenados con español escrito (Bonilla et al., 2022).

[5] http://www.corpusrural.es

5. Conclusiones

CLICC es una plataforma en línea colaborativa para el almacenamiento, administración y consulta de corpus de las lenguas de Colombia. Cuenta con varios tipos de usuarios que tienen permisos específicos y funcionalidades de acuerdo a sus roles, lo que facilita el trabajo colaborativo y el manejo y privacidad de la información. El sistema está compuesto por una base de datos, una interfaz de usuario general (la web), una interfaz de usuario registrado (para consultas especializadas y descargas), una interfaz de investigador (para el ingreso y administración de corpus) y una interfaz de administración (para la gestión de usuarios y corpus). Hasta el momento se han creado 13 corpus de distintos tipos, de los cuales 6 ya están públicos para consulta en línea y su reutilización en otras investigaciones.

Como perspectivas a futuro, es necesario seguir trabajando en el desarrollo de las funciones para corpus bilingües paralelos y alineados, en la creación e implementación de herramientas especializadas para análisis y procesamiento de muestras del español, y en el mantenimiento y mejoramiento de la plataforma para el tratamiento de diferentes tipos de corpus. Todo con miras a que CLICC se siga consolidando como un espacio para que investigadores, grupos de investigación, comunidades y personas interesadas puedan administrar y publicar sus corpus de las lenguas de Colombia y así contribuir a la documentación, investigación, estudio y promoción de la diversidad lingüística y cultural del país.

Referencias

Anthony, Laurence. 2013. Developing AntConc for a new generation of corpus linguists. En *Corpus Linguistics Conference (Abstracts Book)*, 14–16.

Bonilla, Johnatan, Miriam Bouzouita & Rosa Segundo Díaz. 2022. La construcción del Corpus Oral y Sonoro del Español Rural-Anotado y Parseado (COSER-AP): avances en el etiquetado de partes del discurso. *Revista Internacional de Lingüística Iberoamericana* 20(2). 77–96. 10.31819/rili-2022-204006.

Caminada, Nuno, Violeta Quental & Milena Garrão. 2008. Linguistics Tools: uma plataforma expansível de funções de consulta a corpus. En *WebMedia: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, 364–368. 10.1145/1809980.1810067.

Gaved, Tim & Sophie Salffner. 2014. Working with ELAN and FLEx together: an ELAN-FLEx-ELAN teaching set. https://es.scribd.com/document/357359102/Working-with-ELAN-and-FLEx-together-pdf.

ISO. 2022. ISO/IEC 27001 Information security management systems. Organización Internacional para la Estandarización, https://www.iso.org/standard/27001#lifecycle.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vit Suchomel. 2014. The Sketch Engine: ten years on. En *Lexicography ASIALEX 1*, 7–36. 10.1007/s40607-014-0009-9.

Padró, Lluís & Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. En *Language Resources and Evaluation Conference (LREC)*, 2473–2479.

Pennington, Ryan. 2014. Producing time-aligned interlinear texts: Towards a SayMore–FLEx–ELAN workflow. Unpublised draft: https://www.sil.org/resources/archives/66553.

Rehm, Georg, O. Schonefeld, Andreas Witt, C. Chiarcos & T. Lehmberg. 2008. SPLICR: a sustainability platform for linguistic corpora and resources. En *KONVENS*, 86–96.

Rubio, Ruth & Julio Bernal. 2019. Corpus Oral del Instituto Caro y Cuervo: reestructuración, diseño y construcción. *Lexis* 43(1). 195–219.

Rubio, Ruth, Andrea Llanos, Julio Bernal, Johnatan Bonilla & Daniel Bejarano. 2017. Diseño y elaboración del sistema gestor de contenidos para los corpus lingüísticos del Instituto Caro y Cuervo. En *Estudios Lingüísticos*, Instituto de Literatura y Lingüística. Cuba.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *International Conference on New Methods in Language Processing*, 1–9.

Scott, M. 2008. Developing WordSmith. *International Journal of English Studies* 8(1). 95–106.

Sierra, Gerardo, Julian Solórzano & Arturo Curiel. 2017. GECO, un gestor de corpus colaborativo basado en web. *Linguamatica* 9(2). 57–72. 10.21814/lm.9.2.256.

Taulé, Mariona, Antonia Martí & Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. En *Language Resources and Evaluation Conference (LREC)*, s.p.

Torruella, Joan & Joaquim Llisterri. 1999. Diseño de corpus textuales y orales. En *Filología e informática. Nuevas tecnologías en los estudios filológicos*, 45–77. Editorial Milenio.