© Serie de Comunicación Social 2024-2026. ISBN: 978-959-7174-42-4
Centro de Lingüística Aplicada, Santiago de Cuba, 2026

LEONEL RUIZ MIYARES*
JOSÉ CRUZATA FERRER**
YAMILEYDIS VERANES VÁZQUEZ**
Centro de Lingüística Aplicada*
Empresa de Desarrollo de Aplicaciones, Tecnología y Sistemas (DATyS)**
Santiago de Cuba, Cuba
lcosmeruizmiyares@gmail.cu

El Corpus del español de Cuba (CorEsCu) y su procesamiento computacional[1]

Introducción

En Cuba, el desarrollo alcanzado en el Procesamiento del Lenguaje Natural (PLN) –joven disciplina científica en el país– ha posibilitado el momento de emprender proyectos superiores y de mayor complejidad.

El Centro de Lingüística Aplicada[2] (CLA) de la Agencia de Ciencias Sociales y Humanísticas del Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) de Santiago de Cuba concluyó en enero del 2001, luego de varios años de preparación, el primer etiquetador gramatical cubano, el ETIPROCT[3] –ETIquetador y PROcesador de Corpus Textuales– (Ruiz Miyares, 2001a). Se iniciaba de esta manera un largo, complejo, pero fructífero camino del Procesamiento del Lenguaje Natural en Cuba, donde esta Institución es una de las pioneras.

Casi con el surgimiento del ETIPROCT, a partir del 2000, y como parte de la creación de diccionarios electrónicos en el CLA (Alegría et al., 2006a; Alegría et al, 2006b; Alegría et al., 2006c; Heredia et al., 2011; Ruiz Miyares, 1997b; Ruiz Miyares, 1998; Tamayo Lozada y Ruiz Miyares, 2021), fue necesaria la construcción de diferentes herramientas para lograr ese objetivo. En ese momento el Grupo de investigación Ixa[4] de Procesamiento del Lenguaje Natural de la Facultad de Informática de la Universidad del País Vasco (UPV/EHU) comienza la colaboración científica con el CLA que dura hasta la actualidad.

Las herramientas de apoyo a los diccionarios electrónicos son:

- Conjugador verbal
- Divisor silábico
- Pluralizador
- Editor de diccionarios leXkit
- Entorno de lematización
    - Lematizador automático
    - Corrección de la lematización
- Elaboración en sí de los diccionarios

Se debe resaltar que el editor de diccionarios leXkit se ha empleado en cuatro ediciones (2005, 2008, 2009, 2014) del conocido *Diccionario básico escolar* (Miyares Bermúdez, 2014) con resultados satisfactorios.

Por otro lado, y también como parte del desarrollo del PLN en Cuba, desde el 2005 el Centro de Lingüística Aplicada establece nexos de colaboración científica con la Empresa de Desarrollo de Aplicaciones, Tecnologías y Sistemas (DATyS) elaborando otras herramientas de PLN dentro del macroproyecto *Suite de procesamiento de lenguaje natural (analizador morfológico, etiquetador, analizador sintáctico, corrector ortográfico y reconocedor de entidades)* (Colectivo de autores, 2007-2009), (Ríos García, 2013), (Arredondo Toledo et al., 2016).

El CLA había incursionado ya en la construcción de diversos corpus (Causse Cathcart y Ruiz Miyares, 2000), (Miyares Bermúdez, 2006), (Pérez Marqués et al. 2011), algunos de los cuales sirvieron en la producción de herramientas lingüísticas de amplio uso en la sociedad cubana[5].

[1] Trabajo publicado en inglés en **BuLAG 40** (Bulletin de Linguistique Apliquée et Générale) Nr. 40, **Languages Analysis, Comparison and Generation - Systems, Models and Applications. Homage to Peter Greenfield**, Sylviane CARDEY, François-Claude REY, Iana ATANASSOVA (eds.), Presses universitaries de Franche-Comté, Francia, pp. 215-252, 2022 con el título **Cuban Spanish corpus: computational processing** de Leonel Ruiz Miyares, José Cruzata Ferrer y Yamileydis Veranes Vázquez.

[2] http://www.cla.cu/clanuevo/es/

[3] En (Ruiz Miyares y Zamora Matamoros, 2000) se realiza una detallada descripción del ETIPROCT, donde se incluye su eficiencia.

[4] https://www.ixa.eus/?language=en

[5] El corpus del *Léxico activo-funcional del escolar cubano* fue fundamental en la confección de los primeros diccionarios escolares cubanos: *Diccionario escolar ilustrado* (1998, 2017) y *Diccionario básico escolar* (2003, 2008, 2009, 2014), dirigidos por la prestigiosa lexicógrafa Eloína Miyares Bermúdez. También dicho corpus se utilizó para elaborar, conjuntamente entre

Con toda esa amplia experiencia acumulada, tanto en el desarrollo de herramientas de PLN, como en la creación de diversos corpus textuales, el Centro de Lingüística Aplicada inició en el 2021 el proyecto de investigación titulado *Corpus del español de Cuba (CorEsCu)* con la participación de investigadores y técnicos del CLA y de la empresa DATyS.

La confección del Corpus del español de Cuba supone poseer una poderosa herramienta linguo-estadística donde lingüistas, periodistas, escritores, etc. tendrán acceso a un sinnúmero de resultados, entre los que se encuentran:

- Conocer con mayor precisión el léxico y las características de la variante cubana del español.
- Describir las peculiaridades del vocabulario en diferentes regiones del país.
- Realizar trabajos lexicográficos (verificar el uso de una determinada palabra en contexto, selección de ejemplos para diccionarios cubanos, etc.).
- Profundizar en la sintaxis del español de Cuba.
- Entrenar sistemas de PLN.
- Caracterizar la variedad del vocabulario en diferentes géneros de texto.

1. El sistema computacional y el procesamiento de los textos

Para llevar adelante este ambicioso proyecto se necesita un *software* con prestaciones adicionales que el ETIPROCT no incluye, como por ejemplo la presencia de la lematización, el reconocimiento de errores ortográficos y la ampliación del espectro del *tagset*, el cual adiciona etiquetas que se emplean para reconocer con mayor exactitud los elementos lingüísticos dentro del texto.

Si el ETIPROCT distingue 36 rasgos gramaticales (9 etiquetas para los sustantivos, 5 para los adjetivos, 7 para los pronombres, 7 para los verbos y 8 etiquetas para el resto de las categorías gramaticales (artículo, adverbio, preposición, conjunción, interjección, contracción, lexías complejas y siglas)), el nuevo sistema abarca 49 etiquetas, 13 más, donde se adicionan etiquetas con propiedades más detalladas de los verbos (transitividad, tiempo, persona, etc.), del artículo (género, número, etc.), unidades de medida (kg, cm, …), fecha u hora (11-11-2016, 18:50, …), identificadores (simposio@cla.cu, …), símbolos (%, $, &, …), lo que ayuda a obtener mayores y diversos resultados linguo-estadísticos para beneficio de los investigadores.

El sistema está compuesto por tres módulos:

- Módulo 1: Preprocesamiento de los textos
- Módulo 2: Procesamiento de los textos
- Módulo 3: Recuperación y visualización de información. Estadísticas

1.1. Preprocesamiento de los textos

En este módulo se introducen todos los documentos que forman parte del corpus y que son etiquetados posteriormente en el módulo 2.

Previamente, los lingüistas recopilan los textos a procesar a partir del diseño de la investigación siguiendo criterios ya establecidos en la teoría de la Lingüística de corpus (Biber, 1993), (Berber Sardinha, 2004); en el caso de CorEsCu la composición del corpus es la siguiente:

- Clasificación del corpus: Cerrado, rango específico
- Período: 1995-2014
- Corte: Sincrónico (etapa específica)
- Tipo: General, aunque incluye textos especializados
- Composición: Textos completos de la prensa (periódicos y revistas) (49%), fragmentos de libros de ficción y de no ficción (49%) y misceláneas (2%).
- Constitución: Mixta, 95% de la lengua escrita y 5% de la oral.
- Lengua: Español de Cuba (monolingüe)
- Anotación: Corpus con etiquetado gramatical automático, con revisión manual.

---

lingüistas del CLA e hispanistas del Consejo Nacional de Investigaciones de Italia, el *Diccionario ortográfico del español* (1999) y el *Vocabulario inverso y anagramas del español* (2001).

- Lematización: Automática, con revisión manual.
- Tamaño: Mediogrande, 5 millones de palabras de la variante cubana del español.
- Uso previsto: Lexicológico, lexicográfico, gramatical, general.
- Corpus de referencia: Ofrecerá información detallada sobre una lengua con información lingüística asociada.

Se puede considerar que el Corpus del español de Cuba será un corpus representativo (Berber Sardinha, 2004: 22) de la variante cubana del español, pues contendrá más de 5 millones de formas del período de 1995 al 2014[6] y recoge información de los siguientes textos:

- Periódicos nacionales (3): *Granma, Trabajadores* y *Juventud Rebelde.*
- Periódicos provinciales (15): *El Guerrillero*, Pinar del Río; *El Artemiseño*, Artemisa; *Tribuna*, La Habana; *Mayabeque*, Mayabeque; *Girón*, Matanzas; *5 septiembre*, Cienfuegos; *Vanguardia*, Villa Clara; *Escambray*, Sancti Spítirus; *Invasor*, Ciego de Ávila; *Adelante*, Camagüey; *Periódico 26*, Las Tunas; *Ahora*, Holguín; *La Demajagua*, Granma; *Sierra Maestra*, Santiago de Cuba y *Venceremos*, Guantánamo.
- Periódico *Victoria* del municipio especial Isla de La Juventud.
- Revistas no especializadas (5): *Bohemia, Mujeres, Somos Jóvenes, Zunzún y Pionero.*
- Revistas especializadas (7): *Alma Mater, Juventud Técnica, Mar y Pesca, Cine cubano, Tablas, La Gaceta de Cuba, Revista Cubana de Salud Pública.*
- Fragmentos de libros de escritores cubanos.
- Misceláneas: Correos electrónicos, invitaciones, citaciones, etc.

Se ha previsto la localización y recogida de la mayor cantidad posible de textos digitalizados del período seleccionado en los siguientes formatos .TXT, .RTF, .DOC, .DOCX, .HTML, .PDF.

El módulo 1 es una herramienta de escritorio para el desarrollo de tres tareas:

- Introducir los textos
- Introducir los metadatos
- Exportar los textos

La introducción de los textos consiste en la captación de las obras en cualquiera de los formatos descritos y se introduce la metainformación de cada una de ellas, a saber:

- Medio: periódico/revista/libro/misceláneas
- Nombre del medio
- Fecha de la publicación: dd/mm/aa
- Tipo de obra:
  - Ficción (novelas, cuentos, obras teatrales y guiones)
  - No ficción (según las cinco áreas temáticas: Ciencias exactas, tecnologías y salud; Ciencias sociales, creencias y pensamiento; Política, economía y justicia; Artes y cultura; Ocio y vida cotidiana.)
- Fuente: nacional o provincial
- Título
- Volumen
- Página
- Editorial
- Autor/es
- Sexo del autor: Femenino (F) / Masculino (M) / Sin datos
- Sección: Ciencia, Cultura, Deportes, …
- Temas: Ciencias exactas, tecnología y salud; Ciencias sociales, creencias y pensamiento; …

[6] Considerando que Cuba comienza a digitalizar los periódicos y revistas en 1999 y teniendo en cuenta el cuantioso trabajo que llevaría escanear cientos de periódicos y revistas –con el correspondiente cotejo del texto digitalizado resultante con el texto original–, se comenzaron a procesar los textos digitalizados, el resto se procesará más adelante.

- Subtemas: Biología, Botánica, Zoología, Paleontología, …
- Género: novela, cuento, teatro, guion

Luego, se exportan los textos con los metadatos asociados (Fig. 1) para su procesamiento en el módulo 2. El módulo 1 no permite la inclusión de textos duplicados.

Fig. 1. Muestra del módulo 1, introducción de la metainformación de un texto del periódico *Granma*.

1.2. Procesamiento de los textos

El módulo 2 se encuentra en ambiente web y también posee tres tareas:

- Importar datos
- Procesamiento de los documentos
- Exportar datos

En este módulo existe un administrador que controla las altas y bajas de los usuarios a trabajar en el mismo y es el único autorizado a exportar los documentos.

Luego de la importación de los textos provenientes del primer módulo, se realiza el procesamiento de los mismos que contiene los siguientes pasos:

- Segmentación
- Corrección ortográfica
- Etiquetación gramatical
- Etiquetación de nombre de entidad (persona, lugar, organización y evento)
- Lematización

Es importante destacar que todos esos pasos pertenecen a complejas herramientas de Procesamiento del Lenguaje Natural, donde sobresale la etiquetación gramatical automática de cada palabra de cada documento, la que se efectúa sobre la base de los modelos ocultos de Markov (HMM, sus siglas en inglés), con la utilización de bigramas y trigramas, además del empleo de diversas heurísticas y diccionarios. La explicación detallada de cada paso del módulo 2 se puede encontrar en García Moya (2008), Castro Castro et al. (2009), Colectivo de autores (2007-2009) y Viant Morán (2010).

Una vez procesados los textos, continúa la etapa más compleja del proyecto, la verificación de la etiquetación donde el lingüista revisa de manera meticulosa, principalmente, cada etiqueta asignada por el etiquetador a cada palabra con el objetivo de obtener un corpus lo más fiable posible.

Para este caso se crearon convenciones con el propósito de tener uniformidad de criterios entre los lingüistas-revisores y los lingüistas-expertos:

> Debido a la posible discrepancia en la clasificación de las palabras, o la inquietud o duda que no pocas veces surge al analizar un término en un contexto lingüístico dado, el CLA elaboró un conjunto de convenciones con el fin de esclarecer dicha taxonomía. De tal modo se evita una situación embarazosa al determinar la clase de palabra con la que nos enfrentamos. Dichas convenciones establecen normas

> para la etiquetación del Corpus y abarca un elevado número de clases de vocablos, perífrasis y locuciones de uso habitual en nuestra lengua. (Ocaña Dayar, 2023: 33)

Se necesita de mucha precisión y concentración en este arduo trabajo.

Durante la revisión, los documentos procesados poseen seis estados:

- Sin asignar: El administrador del sistema aún no ha asignado el texto a ningún lingüista.
- Asignado: El administrador del sistema ha asignado el texto a un lingüista.
- En revisión: Se está revisando el texto.
- Revisado: El lingüista considera listo el documento.
- Aprobado: El documento es aprobado por el lingüista-experto.
- Exportado: Se exporta el texto para el módulo 3.

La Fig. 2. muestra el filtraje de los textos asignados a uno de los investigadores del estudio.

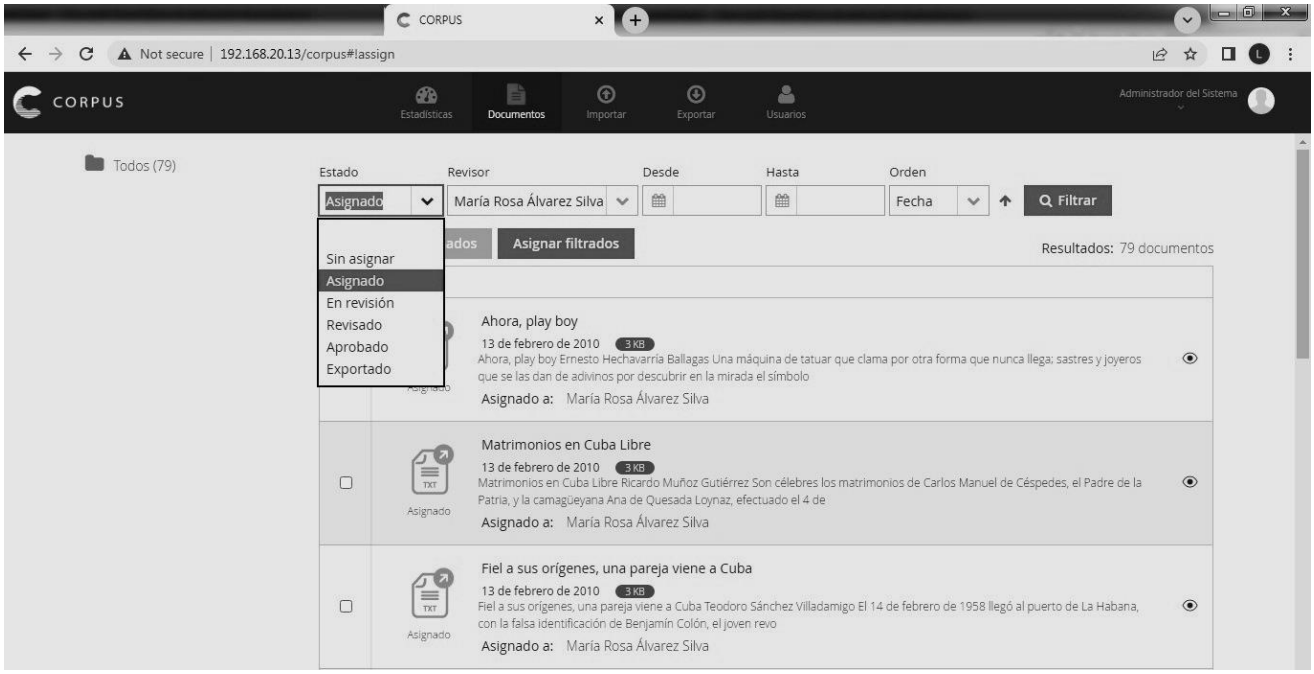

Fig. 2. Muestra del módulo 2. Filtraje de los textos asignados a uno de los investigadores del estudio.

En su ambiente web, el módulo 2 muestra:

- Texto etiquetado
- Texto plano o documento en .txt
- Fichero fuente o texto original
- Detalles o estadísticas: información estructural (cantidad de párrafos, oraciones y palabras del documento procesado), estadística de las categorías gramaticales, información sobre las entidades
- Historial
- Metadatos del texto que se revisa

Toda esta información es importante para el lingüista-revisor, pues puede verificar el contexto de uso de una palabra con la que se tenga alguna duda en su texto plano o en su fichero fuente, cuán rico es el documento en datos gramaticales, puede comprobar los cambios realizados por él mismo en el historial, etc.

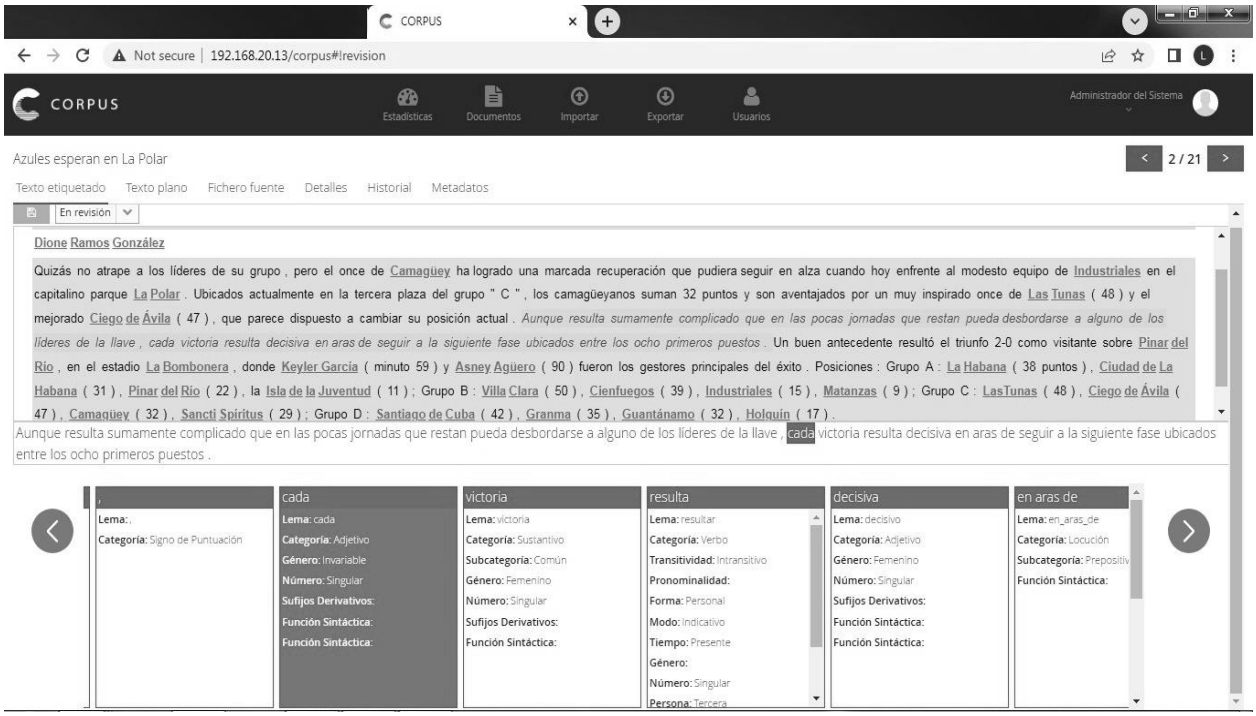

Fig. 3. Muestra del texto etiquetado del artículo *Azules esperan en La Polar* del periódico *Adelante*, Camagüey, del 2010 perteneciente al módulo 2.

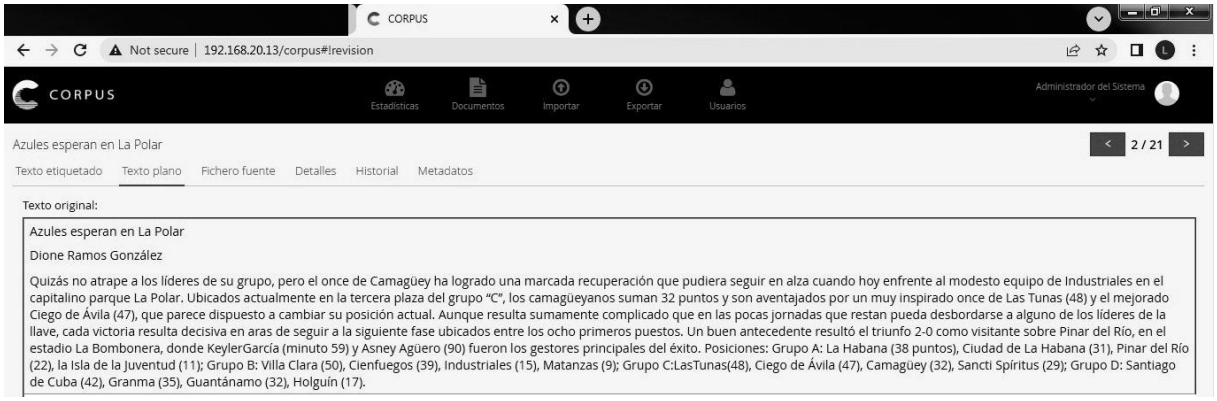

Fig. 4. Muestra del texto plano del mismo artículo de la figura anterior.

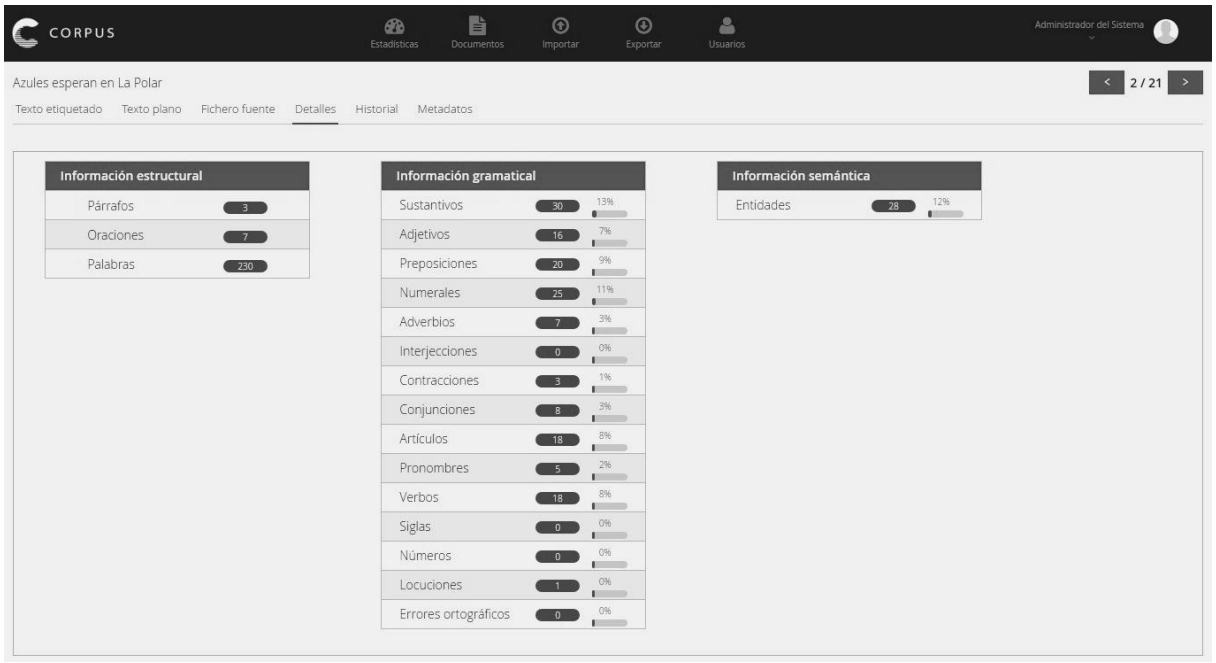

Fig. 5. Información estadística del documento mostrado en la figura 3.

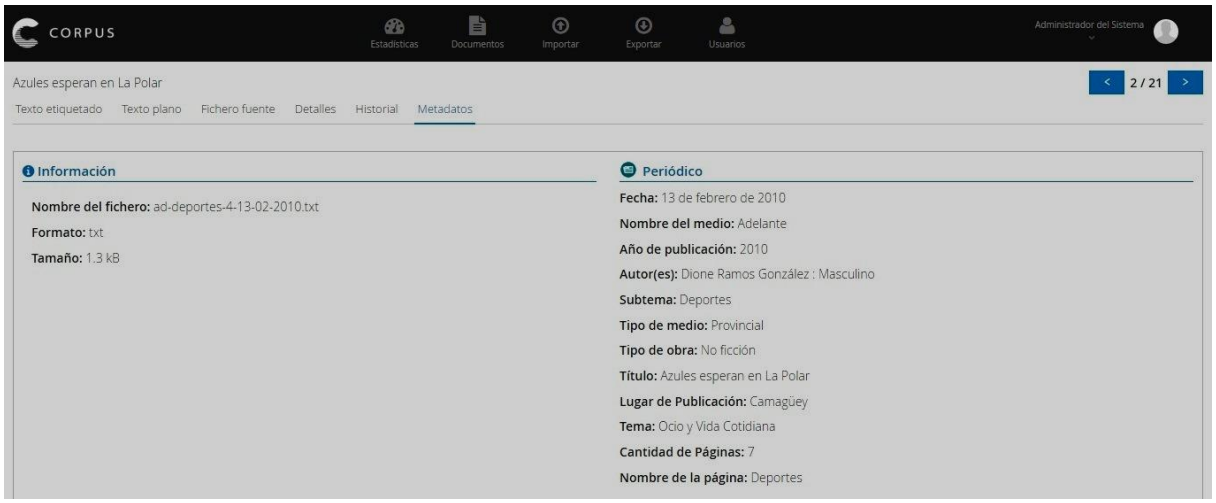

Fig. 6. Metadatos del documento mostrado en la figura 3.

En las figuras 3, 4, 5 y 6 se observan ejemplos de las salidas del módulo 2 para un mismo texto, donde la Fig. 3 es la más relevante, pues es allí donde el lingüista-revisor efectúa el control de calidad de la etiquetación y hace los cambios necesarios, entre los que se encuentran: insertar palabra, eliminar palabra, editar palabra, realizar la corrección ortográfica[7], crear o quitar una entidad y unir palabras. Obsérvese en la parte inferior de la Fig. 3 cómo se muestra la oración que el lingüista revisa y cada palabra posee su lema, la categoría gramatical, el género, número, según el caso, y es allí donde se realizan las correcciones, si es necesario.

Después de concluida la comprobación, el lingüista-revisor pasa el texto al estado *revisado* para que el lingüista-experto verifique el mismo, se deslinden las imprecisiones y dudas, –si las hubiera– y se pone el texto en el estado *aprobado*.

Cuando todos los textos estén en estado *aprobado* se procede a exportar cada documento para su análisis en el módulo 3.

Como se puede apreciar, la labor de los lingüistas en el módulo 2 supone agotadoras jornadas de análisis de miles de palabras, reflexión y mucha seriedad; consume mucho tiempo, es costoso, pero con rigor se obtendrá un resultado muy valioso.

[7] En este caso, el lingüista corrige el error ortográfico detectado por el sistema y este se registra para su posterior análisis en el módulo 3, como se verá más adelante.

1.3. Recuperación y visualización de información. Estadísticas

En este módulo los usuarios del sistema (lingüistas, periodistas, escritores, etc.) pueden realizar infinitas consultas y búsquedas en aras de ejecutar estudios lingüísticos específicos de la variante cubana del español.

Esta sección incluye:

- Búsqueda:
  - Simple
  - Contiene
  - Empieza en
  - Termina en
  - Frase
  - Búsqueda avanzada
- Listado de palabras
- Correcciones
- Estadísticas

Describiremos este módulo sucintamente debido a que en el próximo acápite se ofrecerán resultados concretos de los primeros aportes de este estudio y allí se mostrarán ejemplos de estas funcionalidades.

La búsqueda *simple* recupera aquellos términos que coincidan exactamente con la solicitud realizada, sin embargo la consulta *contiene*, recupera todos aquellos que admitan la petición. En las Fig. 7 y Fig. 8 se observa la diferencia de ambas demandas.

Fig. 7. Ejemplo de la búsqueda *simple* del vocablo *humano*.

Fig. 8. Ejemplo de la búsqueda *contiene* del vocablo *humano*.

La *búsqueda avanzada* permite realizar las pesquisas de manera opcional por tipo de medio, tipo de obra, autores, nombre del medio, fechas, etc. (Fig. 9)

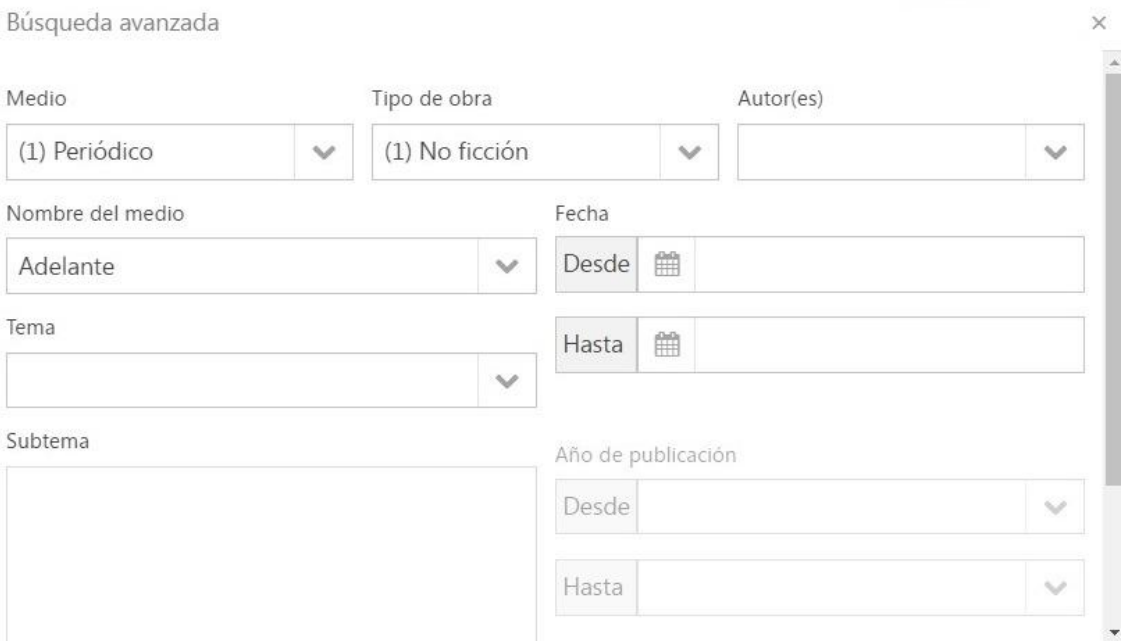

Fig. 9. Ejemplo de *búsqueda avanzada* del módulo 3.

Fig. 10. Resultado de la *búsqueda avanzada* solicitada en la Fig. 9.

En la Fig. 10 se puede ver la recuperación obtenida diferente a la que se ofrece en la Fig. 8 debido a los requerimientos realizados en la *búsqueda avanzada* de la Fig. 9.

2. Primeros resultados

El estudio del Corpus del español de Cuba comenzó con el análisis de la prensa del Oriente de Cuba. Se recopilaron de forma digital los periódicos provinciales desde Camagüey hasta Guantánamo del período 2001-2014 –cuando ya estaban en ese formato–, y se inició el procesamiento de *Adelante* y *Venceremos*, precisamente la prensa de las provincias extremas de esa región para realizar en un futuro algunas comparaciones léxicas, entre muchos otros análisis.

Se introdujeron en el módulo 1 todos los artículos de un periódico seleccionado al azar del primer semestre del año y de uno del segundo semestre del rango de años escogido.

En la Fig. 11 se refleja la distribución de los ensayos procesados en el módulo 1 de los periódicos del Oriente de Cuba, donde las cifras no pueden ser las mismas, pues el rango de años varía al igual que el número de artículos del ejemplar, como es lógico.

Provincia	Periódico	Cantidad de artículos
Camagüey	Adelante	668 (2003-2014)
Las Tunas	26	626 (2006-2014)
Holguín	Ahora	731 (2001-2014)
Granma	La Damajagua	754 (2002-2014)
Santiago de Cuba	Sierra Maestra	625 (2006-2014)
Guantánamo	Venceremos	512 (2006-2014)
Total	**6**	**3916**

Fig. 11. Distribución de los artículos procesados en el módulo 1 de los periódicos del Oriente de Cuba.

El equipo de trabajo del CorEsCu desde muy temprano –luego del procesamiento por el módulo 2 de los primeros textos y de su aprobación por los lingüistas-revisores–, estaba deseoso de apreciar cómo serían los resultados iniciales del estudio, por lo que en (Álvarez Silva y Ocaña Dayar, 2023) y en (Ocaña Dayar 2023) ya se describen los atisbos inaugurales de esta extensa e importante investigación. Todo lo anterior se ha logrado independientemente de lo complicada que es la revisión, la presencia de poco personal especializado para este tipo de investigación (bajas, licencias de maternidad, etc.), del equipamiento computacional limitado y obsoleto que posee el CLA, de la pandemia COVID-19 y de la grave crisis económica por la que atraviesa nuestro país en estos momentos.

A continuación se describirá lo obtenido hasta la fecha de hoy, como complemento de los dos artículos antes señalados.

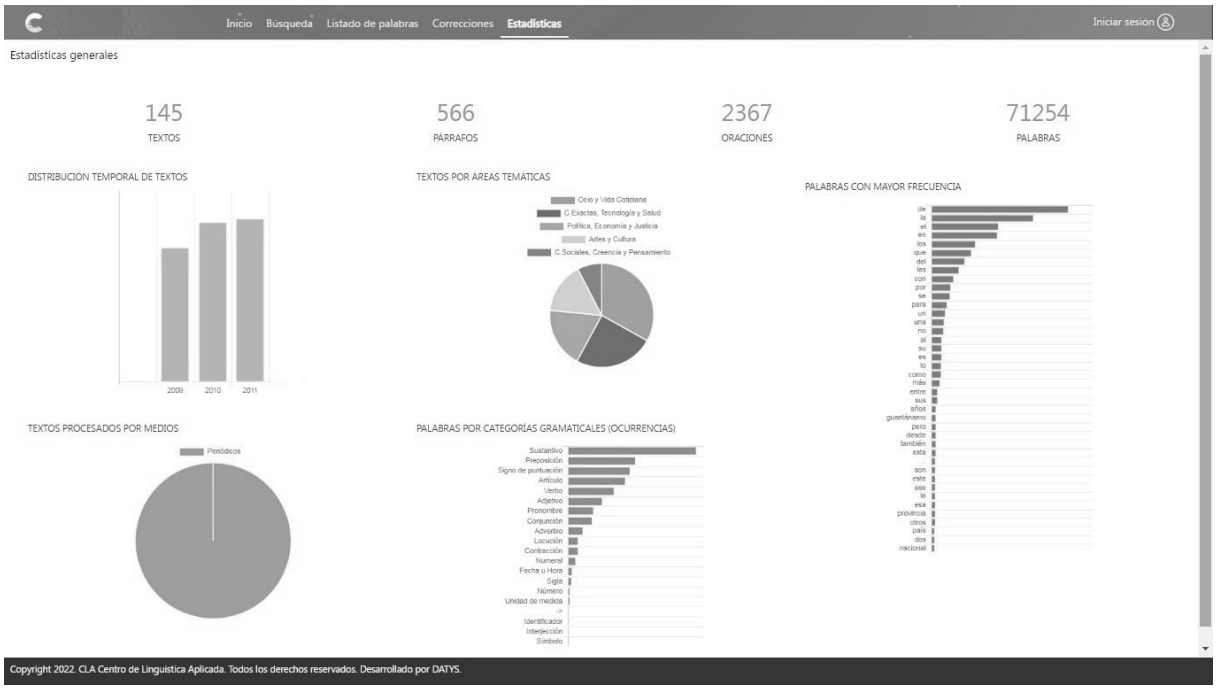

Fig. 12. Principales estadísticas del módulo 3 del CorEsCu.

Como se observa en la Fig. 12, ya se han procesado y aprobado 145 documentos que contienen 566 párrafos, 2367 oraciones y 71254 palabras.

En la tabla Nr. 1 se encuentra la distribución de las principales categorías gramaticales presentes en CorEsCu en el corte realizado hasta la fecha, según el número de ocurrencias. Resalta a la vista la gran cantidad de

sustantivos empleados, ocupando el primer lugar. Se puede entender este fenómeno al iniciarse este estudio con los medios de prensa, los periódicos específicamente, y el alto uso por los periodistas de disímiles sustantivos en la descripción de sus reportajes, reseñas, editoriales, etc.

Las quince palabras más frecuentes hasta ahora en el corpus son las que aparecen en la tabla Nr. 2. Como siempre sucede, encabezan la lista aquellas palabras con un alto rendimiento funcional, dígase preposiciones, conjunciones, contracciones, artículos, pronombres. El número 24 lo ocupa el primer sustantivo, *años*, con 124 ocurrencias.

Las figuras 13, 14 y 15 representan diferentes formas de búsqueda, donde en la 13 y 14 se puede observar la cantidad de casos encontrados, en cuántos documentos, la descripción gramatical de la palabra seleccionada –donde se advierte la lematización–, y en la 15 la cantidad de casos con la terminación *-ción* y en cuántos documentos.

Categoría gramatical	Frecuencia	%
Sustantivo	18706	26.2
Preposición	9777	13.7
Artículo	8276	11.6
Verbo	6599	9.2
Adjetivo	4928	6.9
Pronombre	3591	5
Conjunción	3456	4.8
Adverbio	2020	2.8
Locución	1355	1.9
Contracción	1318	1.8
Numeral	1005	1.4
Fecha u Hora	508	0.7
Sigla	372	0.5
Unidad de medida	97	0.1

Tabla Nr. 1. Distribución descendente de las principales categorías gramaticales en CorEsCu en el corte realizado hasta la fecha, según el número de ocurrencias.

Nr.	Palabra	Frecuencia
1	de	4215
2	la	3125
3	y	2154
4	el	2059
5	en	2021
6	los	1337
7	que	1212
8	del	1020
9	a	1009
10	las	830
11	con	668
12	por	570
13	se	562
14	para	469
15	un	403
…		
24	años	124

Tabla Nr. 2. Las 15 palabras más frecuentes del CorEsCu hasta la fecha.

En la descripción del módulo 3 se señaló de la existencia en el sistema del listado de las palabras presentes en el corpus. En la Fig. 16 se evidencia un fragmento de ese listado en los primeros resultados de CorEsCu, donde se incluyen las formas de las palabras lematizadas.

Fig. 13. Ejemplo de recuperación de una búsqueda.

Fig. 14. El mismo ejemplo anterior, con la descripción gramatical de la forma verbal *entendió*.

© Serie de Comunicación Social 2024-2026. ISBN: 978-959-7174-42-4
Centro de Lingüística Aplicada, Santiago de Cuba, 2026

Fig. 15. Fragmento de la recuperación de la búsqueda por *-ción*.

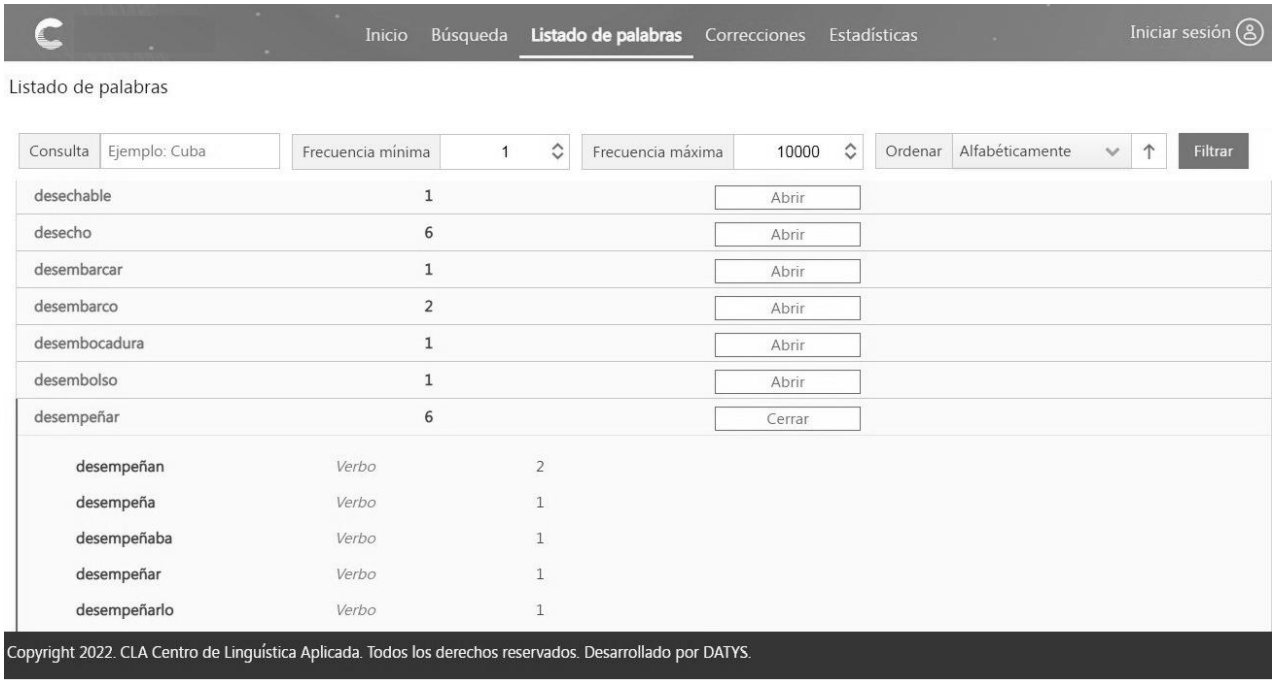

Fig. 16. Fragmento del listado de palabras donde se incluyen las formas empleadas en el corpus del verbo *desempeñar*.

Aunque su frecuencia de aparición no es alta, en algunos medios de prensa se manifiestan a veces errores ortográficos producidos por lapsus, falta de atención o mala revisión, lo que empaña el prestigio de los mismos. El sistema computacional de CorEsCu capta estos errores y en la Fig. 17 se señala un ejemplo de esos errores que incluye en qué periódico sucedió, para su posterior información al rotativo.

Fig. 17. Fragmento donde se muestran los errores ortográficos encontrados en los artículos periodísticos.

Cada resultado de búsqueda se puede guardar en ficheros de diferentes formatos (.XLSX, .PDF o .TXT) para análisis posteriores de los investigadores.

3. Conclusiones

El proyecto del Corpus del español de Cuba avanza, continúa la revisión de los textos con rigor, el sistema computacional aún se prueba y se le realizarán las mejoras pertinentes con el fin de obtener lo antes posible un *software* eficiente.

Como se ha señalado, la importancia de la confección de CorEsCu es fundamental para emprender múltiples investigaciones lingüísticas sobre la variante cubana del español, donde se podrán hacer estudios de diferentes vocabularios (periodísticos, de revistas, escritores, etc.), análisis diastráticos, diatópicos, de colocaciones y fraseología; análisis de estilística, sintácticos, de concordancia de una palabra o grupo de palabras; se podrán detectar neologismos y un gran etcétera, por lo que la implementación del sistema computacional y la creación del primer Corpus del español de Cuba representarán un salto cualitativo en los estudios lingüísticos en nuestro archipiélago.

4. Agradecimientos

Los autores de este trabajo desean agradecer a los lingüistas-revisores, a los lingüistas-expertos y a los informáticos que participan o participaron en este complejo proyecto: María Rosa Álvarez Silva, Alejandro Miyares Peña, Yoandra Chuen Gómez, Humberto Ocaña Dayar, Alex Muñoz Alvarado, Nancy Cristina Álamo Suárez, Javier Tamayo Lozada y Rolando Urrutia Cleger.

5. Bibliografía

Alegría, I.; Arregi, X.; Artola, X.; Astiz, M. y Ruiz Miyares, L. (2006a) *A Dictionary Content Management System* in Proceedings XII EURALEX International Congress, Vol. I, pp. 105-109, Turin, Italy.

Alegría, I.; Arregi, X.; Artola, X.; Astiz, M. y Ruiz Miyares, L. (2006b) *Building an Electronic Version of the Cuban Basic School Dictionary* in Proceedings XII EURALEX International Congress, Vol. I, pp. 243-250, Turin, Italy.

Alegría, I.; Arregi, X.; Artola, X.; Astiz, M. y Ruiz Miyares, L. (2006c) *Different issues in the design and development of the electronic Cuban Basic School Dictionary* in Linguistics in the Twenty First Century, edited by Eloína Miyares Bermúdez and Leonel Ruiz Miyares. *Cambridge Scholars Press*, Cambridge, United Kingdom, in cooperation with Centro de Lingüística Aplicada, Santiago de Cuba, Cuba, pp. 273-288.

Álvarez Silva, M.R. y Ocaña Dayar, H. (2023): *Los nombres propios en el proyecto "Corpus del español de Cuba". Primera aproximación*. Serie de Comunicación Social, 2022-2023. Centro de Lingüística Aplicada, Santiago de Cuba, pp. 78-81.

Arredondo Toledo, L.; Castro Castro, D.; Ruiz Miyares, L. (2016): *Detección de relaciones entre nombres de entidades para el español* en *VIII Conferencia Internacional de Ingeniería Eléctrica (CIIE 2016)*, Universidad de Oriente, Santiago de Cuba, pp.1-4. (Publicación electrónica)

Berber Sardinha, T. (2004): **Lingüística de Corpus**, Editora Manole Ltda., Brasil.

Biber, D. (1993): *Representativeness in Corpus Design*. Disponible en: http://otipl.philol.msu.ru/media/biber930.pdf

Bocorny Finatto, M.J.; Rodrigues Rebechi, R.; Sarmento, S.; Pereira Bocorny, A.E. (2018): **Linguística de corpus: perspectivas**. Instituto de Letras, UFRGS, Porto Alegre.

Castro Castro, D.; Lannes Losada, R.; Pons Porrata, A.; Ramírez Cruz, Y. (2009): *Construcción de un corrector ortográfico* en ACTAS del *XI Simposio Internacional de Comunicación Social*, Centro de Lingüística Aplicada, Santiago de Cuba, pp. 257-260.

Causse Cathcart, M; Ruiz Miyares, L. (2000): *Aplicación del etiquetador gramatical cubano a un corpus textual oral*. ACTAS de la *X Conferencia Internacional Lingüístico-Literaria*, Universidad de Oriente, Santiago de Cuba, pp. 72-80. (Publicación electrónica)

Colectivo de autores (2007-2009): *Construcción de herramientas para el Procesamiento del Lenguaje Natural: lematizador, desambiguador, reconocedor de nombres de entidades y analizador sintáctico*. Proyecto de investigación conjunto Centro de Lingüística Aplicada - Facultad de Matemática y Computación de la Universidad de Oriente, Santiago de Cuba.

García, L.; Pons, A.; Ruiz Miyares, L. (2007): *A proposal of a morphological tagger for Spanish based on Cuban corpora* en ACTAS del *RANLP 2007: International Conference on Recent Advances in Natural Language Processing*, Bulgaria, pp. 210-214.

García, L.; Pons, A.; Ruiz Miyares, L.; Cobos Castillo, Y. (2007): *Una propuesta de etiquetador morfosintáctico para el español* en ACTAS del *X Simposio Internacional de Comunicación Social*, tomo I, Centro de Lingüística Aplicada, Santiago de Cuba, pp. 500-504.

García Moya, L. (2008): *Un etiquetador morfológico para el español de Cuba*. Tesis de maestría. Universidad de Oriente, Santiago de Cuba.

Heredia González, R.; Ruiz Miyares, L.; Miyares Bermúdez, E. (2011): *La versión electrónica del Diccionario Escolar Ilustrado*. Comunicación Social en el siglo XXI. Centro de Lingüística Aplicada, Santiago de Cuba, pp. 1077-1080, Vol. II.

Miyares Bermúdez, E. (directora) (2006): **Léxico Activo-Funcional del Escolar Cubano**. Editorial Centro de Lingüística Aplicada, Santiago de Cuba.

Miyares Bermúdez, Eloína (2014, directora): **Diccionario básico escolar**. Editorial Oriente, Santiago de Cuba.

Miyares Bermúdez, E.; Artola Zubillaga, X.; Alegría Loinaz, I.; Arregi Iparragirre, X.; Ruiz Miyares L.; Álamo Suárez, C. y Pérez Marqués, C. (2010): *La segunda y tercera ediciones del Diccionario básico escolar* en Proceedings of the *XIV EURALEX International Congress*, Fryske Akademy, Afuk, Ljouwert, The Netherlands, pp. 164-171.

Miyares Bermúdez, E.; Artola Zubillaga, X.; Alegría Loinaz, I.; Arregi Iparragirre, X.; Ruiz Miyares L.; Álamo Suárez, C. y Pérez Marqués, C. (2012): *Las últimas ediciones del Diccionario básico escolar de Cuba* en **Avances de lexicografía hispánica (I)**, Universidad Rovira i Virgili, Tarragona, España, pp. 201-213.

Ocaña Dayar, H. (2023): *Variedad funcional de las clases de palabras según el contexto. Su ejemplificación en el Corpus del español de Cuba*. Serie de Comunicación Social, 2022-2023. Centro de Lingüística Aplicada, Santiago de Cuba, pp. 32-35.

Pérez Marqués, C.; Quintana Polando, M.; Ruiz Miyares, L. (2009): *El vocabulario escolar en Guamá tras la aplicación de los programas de la Revolución* en ACTAS del *XI Simposio Internacional de Comunicación Social*, Centro de Lingüística Aplicada, Santiago de Cuba, pp. 628-632.

Pérez Marqués, C.; Quintana Polando, M.; Ruiz Miyares, L. (2011): **Desarrollo léxico en escolares de primaria. Ejercicios para su perfeccionamiento.** Ediciones Centro de Lingüística Aplicada, Santiago de Cuba.

Pons Bordería, S. (2022): **Creación y análisis de corpus orales: saberes prácticos y reflexiones teóricas**. Peter Lang, Berlín.

Ríos García, J. (2013): *Herramientas para Análisis Sintáctico del idioma español*. Tesis de Maestría. Universidad de Oriente, Santiago de Cuba.

Ruiz Miyares, L. (1994): *Aplicación de la computación al Estudio del Vocabulario Básico del Escolar Cubano*, en Estudios de Comunicación Social, Editorial Academia, La Habana, pp.96-105.

Ruiz Miyares, L. (1996): *Algunas consideraciones sobre la Lingüística Computacional* en Ciencia en su PC (Revista Electrónica), Vol. 1, Nro. 1, Santiago de Cuba.

Ruiz Miyares, L. (1997a): *Versión avanzada de un sistema computacional aplicado a una investigación lexicológica* en Estudios de Comunicación Social. Editorial Academia, La Habana, pp.85-113.

Ruiz Miyares, L. (1997b): *Diccionario escolar computarizado. Primeros resultados* en Estudios de Comunicación Social. Editorial Academia, La Habana, pp.118-122.

Ruiz Miyares, L. (1998): *Informatización de diccionarios fraseolóxicos: o Diccionario Automatizado de Fraseología Cubana* en ACTAS del I Coloquio Galego de Fraseoloxía, Xunta de Galicia, Centro Ramón Piñeiro, Santiago de Compostela, pp.257-264.

Ruiz Miyares, L. (1999): *Primeros pasos de la etiquetación automática en Cuba*. ACTAS del *VI Simposio Internacional de Comunicación Social*, Santiago de Cuba. Ediciones Editorial Oriente, Centro de Lingüística Aplicada y el Consiglio Nazonale delle Ricerche, pp.710-714.

Ruiz Miyares, L. (2000): *Etiquetación automática en corpus textuales cubanos. Primeros resultados.* ACTAS del *JADT 2000, 5tas. Jornadas internacionales de análisis estadístico de corpus textuales*, Lausana, Suiza, pp.237-244.

Ruiz Miyares, L. y Zamora Matamoros, L. (2000): *Análisis estadístico del comportamiento del primer etiquetador cubano en tres corpus de la prensa*. ACTAS del *XVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN), Vigo, España, pp. 133-140.

Ruiz Miyares, L. (2001a): *Desarrollo de un modelo computacional para el procesamiento de corpus textuales basado en la etiquetación automática.* Tesis doctoral. Universidad de Oriente, Santiago de Cuba, Cuba - Universidad de Twente, Enschede, Países Bajos.

Ruiz Miyares, L. (2001b): *Experiencia cubana en el Procesamiento del Lenguaje Natural*. ACTAS del *SLPLT2, Segundo Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje*, Jaén, España, pp. 29-33.

Ruiz Miyares, L. (2015): *Dos aproximaciones de procesamiento computacional de la fraseología cubana* en la revista *Paremia*, 24, España, pp. 211-219.

Ruiz Miyares, L.; Cruzata Ferre, J. y Veranes Vázquez, Y. (2022): *Cuban Spanish corpus: computational processing* en **BuLAG 40** (Bulletin de Linguistique Apliquée et Générale) Nr. 40, **Languages Analysis, Comparison and Generation - Systems, Models and Applications. Homage to Peter Greenfield**, Sylviane CARDEY, François-Claude REY, Iana ATANASSOVA (eds.), Presses universitaries de Franche-Comté, Francia, pp. 215-252, 2022.

Sierra Martínez, G. (2015): **Introducción a los corpus lingüísticos**. Instituto de Ingeniería, Universidad Nacional Autónoma de México (UNAM).

Tamayo Lozada, J. y Ruiz Miyares, L. (2021): *El Diccionario básico escolar en móviles y tabletas con sistema operativo Android*, revista **Serie Científica** de la Universidad de las Ciencias Informáticas, 14 (7), pp. 54-66, (https://publicaciones.uci.cu/index.php/serie/article/ view/903).

Viant Morán, R.; Ruiz Miyares, L.; Acosta Arafet, C. (2008): *Propuesta de un analizador morfológico del idioma español basado en el modelo de dos niveles* en ACTAS *FIE-08*, *Conferencia Internacional*, 5ta. edición, Universidad de Oriente, Santiago de Cuba. (Publicación electrónica)

Viant Morán, R.; Ruiz Miyares, L.; Pons Porrata, A.; Acosta Arafet, C.; Cobos Castillo, Y. (2009): *Un analizador morfológico basado en el modelo de dos niveles para el idioma español* en ACTAS del *XI Simposio Internacional de Comunicación Social*, Centro de Lingüística Aplicada, Santiago de Cuba, pp. 321-327.

Viant Morán, R.A. (2010): *Un analizador morfológico para el español de Cuba*. Tesis de maestría. Universidad de Oriente, Santiago de Cuba.