MAESTRO y SOCIEDAD e-ISSN 1815-4867

Volumen 22 Número 3 Año 2025
Artículo original

# La transformación lexicográfica: el papel de los corpus en el desarrollo histórico de los diccionarios

## The lexicographical transformation: the role of corpora in the historical development of the dictionaries

## Transformação lexicográfica: o papel dos corpora no desenvolvimento histórico dos dicionários

Lic. Yoandra Chuen Gómez[1]*, https://orcid.org/0009-0004-7214-4967
Dr.C. Alex Muñoz Alvarado[2], https://orcid.org/0000-0002-3374-4554
Dr. C. Alejandro Arturo Ramos Banteurt[3], https://orcid.org/0000-0002-9589-2077

[1,2] Centro de Lingüística Aplicada "Vitelio Ruiz Hernández-Eloína Miyares Bermúdez, Cuba.
[3] Universidad de Oriente, Santiago de Cuba, Cuba.

*Autor para correspondencia. email: yoandra.chuen@gmail.com

**Para citar este artículo:** Chuen Gómez, Y., Muñoz Alvarado, A. y Ramos Banteurt, A. (2025). La transformación lexicográfica: el papel de los corpus en el desarrollo histórico de los diccionarios. *Maestro y Sociedad*, 22(3), 2524-2533. https://maestroysociedad.uo.edu.cu

## RESUMEN

Introducción: La incorporación de los corpus electrónicos a la lexicografía ha originado un proceso transformativo de la disciplina que representa un acontecimiento relevante. En los tiempos anteriores al referido hito, los lexicógrafos se basaban en su pericia, saber especializado y sensibilidad lingüística para la creación de diccionarios, sin disponer de suficiente información documental del objeto descrito, por lo cual la subjetividad prevalecía en el oficio. Actualmente, gracias a las facilidades que brindan los corpus, los diccionaristas pueden respaldar su trabajo con una base sólida. El objetivo del presente artículo es examinar con una visión global el devenir de la lexicografía desde el momento de la asimilación por ella de los corpus lingüísticos, a fin de reconocer el impacto de estos en la conformación y la actualización de los diccionarios.

Materiales y métodos: Los métodos principales empleados en este estudio son el bibliográfico-documental y el histórico-lógico.

Resultados: La selección de los textos debe estar sujeta a la finalidad lexicográfica que se pretenda conseguir, y obedece a criterios establecidos con antelación. El trabajo con este tipo de fuentes es un apoyo para los creadores de diccionarios, les proporciona a estos herramientas y técnicas mediante las cuales podrán determinar los vocablos que formarán parte del lemario macroestructural y los que no.

Discusión: En contraste con las limitaciones anteriores, la lexicografía moderna está basada en datos empíricos obtenidos de corpus, lo que permite crear definiciones más precisas y relevantes, y proporciona ejemplos contextuales extraídos de usos reales del lenguaje; los diccionarios de nuestros días pueden actualizarse con mayor frecuencia, gracias al uso de corpus y herramientas digitales; con ello, se pueden agregar nuevas palabras y significados de manera más rápida, lo que mantiene a los diccionarios más a tono con la realidad que pretenden describir; las obras lexicográficas contemporáneas están disponibles en formato electrónico, lo cual permite un acceso más amplio y ágil; las versiones en línea y las aplicaciones de diccionarios permiten a los usuarios buscar y obtener información de manera instantánea.

Conclusiones: La lexicografía desarrollada a partir de la introducción de los corpus electrónicos ha ido superando cada vez más la subjetividad y la normatividad características de la práctica lexicográfica tradicional, poniendo en primer plano el enfoque descriptivo y objetivo y enriqueciendo notablemente sus fuentes de datos.

**Palabras clave**: Lexicografía, corpus lingüísticos, diccionarios, desarrollo histórico, transformación

## ABSTRACT

The incorporation of electronic corpora into lexicography has originated a transformative process in the discipline that represents a significant event. In the times prior to this milestone, lexicographers relied on their expertise, specialized knowledge, and linguistic sensitivity for the creation of dictionaries, without having sufficient documentary information on the described object, which is why subjectivity prevailed in the craft. Currently, thanks to the facilities provided by corpora, lexicographers can support their work with a solid foundation. The objective of this article is to examine, from a global perspective, the development of lexicography since its assimilation of linguistic corpora, in order to recognize their impact on the formation and updating of dictionaries. The main methods employed in this study are the bibliographic-documentary and the historical-logical.
**Keywords**: Lexicography, linguistic corpora, dictionaries, historical development, transformation

## RESUMO

Introdução: A incorporação de corpora eletrônicos à lexicografia levou a um processo transformador na disciplina que representa um desenvolvimento significativo. Antes desse marco, os lexicógrafos dependiam de sua expertise, conhecimento especializado e sensibilidade linguística para criar dicionários, carecendo de informações documentais suficientes sobre o objeto descrito. Consequentemente, a subjetividade prevalecia na profissão. Atualmente, graças às facilidades proporcionadas pelos corpora, os dicionaristas podem fundamentar seu trabalho com uma base sólida. O objetivo deste artigo é examinar, de uma perspectiva global, a evolução da lexicografia desde o momento em que assimilou os corpora linguísticos, a fim de reconhecer seu impacto na criação e atualização de dicionários. Materiais e métodos: Os principais métodos empregados neste estudo são o bibliográfico-documental e o histórico-lógico. Resultados: A seleção dos textos deve estar sujeita ao objetivo lexicográfico a ser alcançado e seguir critérios pré-estabelecidos. Trabalhar com esses tipos de fontes auxilia os dicionaristas, fornecendo-lhes ferramentas e técnicas para determinar quais palavras serão incluídas no léxico macroestrutural e quais não serão. Discussão: Em contraste com as limitações acima, a lexicografia moderna baseia-se em dados empíricos obtidos a partir de corpora, o que permite a criação de definições mais precisas e relevantes e fornece exemplos contextuais extraídos de usos da linguagem na vida real. Os dicionários atuais podem ser atualizados com mais frequência, graças ao uso de corpora e ferramentas digitais. Isso permite que novas palavras e significados sejam adicionados mais rapidamente, mantendo os dicionários mais sintonizados com a realidade que buscam descrever. Obras lexicográficas contemporâneas estão disponíveis em formato eletrônico, permitindo um acesso mais amplo e ágil. Versões online e aplicativos de dicionário permitem que os usuários pesquisem e recuperem informações instantaneamente. Conclusões: A lexicografia desenvolvida desde a introdução dos corpora eletrônicos tem se distanciado cada vez mais da subjetividade e da normatividade características da prática lexicográfica tradicional, priorizando uma abordagem descritiva e objetiva e enriquecendo significativamente suas fontes de dados.
**Palavras-chave**: lexicografia, corpora linguísticos, dicionários, desenvolvimento histórico, transformação.

Recibido: 15/4/2025 Aprobado: 2/7/2025

## Introducción

A lo largo de la historia, la lexicografía –disciplina de la lingüística dirigida a la elaboración de diccionarios y a la fundamentación teórica de este quehacer– ha sufrido cambios progresivos aparejados a cambios sociales, culturales y del desarrollo científico-técnico. En la actualidad, la incidencia que ha tenido el uso de las nuevas tecnologías ha originado una transformación importante en la disciplina en cuanto a la manera de recopilar y analizar los datos y de difundir el producto obtenido de ello.

El siglo XX implicó una verdadera revolución, la de índole tecnológica, que posibilitó la informatización de lo que hoy constituye una importante herramienta para la lexicografía, los corpus lingüísticos, al facilitar el procesamiento del lenguaje natural y sentar las bases, con ello, para un cambio de postura y de accionar metodológico respecto al estudio de las lenguas. Los corpus informatizados han propiciado un salto cualitativamente superior en la metodología y los resultados de la diccionarística, específicamente en la forma en que se crean, conforman y actualizan las obras lexicográficas. Los resultados alcanzados por la moderna lexicografía basada en corpus lingüísticos son superiores en comparación con los de la lexicografía tradicional intuitivista o basada en materiales insuficientemente sistematizados. Antes de que la lexicografía comenzara a sacar provecho de los corpus electrónicos, los redactores de diccionarios, generalmente, no fundaban su trabajo en el estudio de muestras discursivas organizadas al efecto; en casos esporádicos, se valían de la recopilación de materiales escritos de habla prestigiosa (obras literarias, fundamentalmente) para seleccionar las entradas y redactar los artículos lexicográficos, además de acudir a su intuición y experiencia como conocedores cultos de la lengua descrita. Sin embargo, con la disponibilidad de grandes colecciones de textos devenidos corpus

(representativos bien de la lengua escrita, bien de la lengua oral, o de ambas), los especialistas pueden basar sus decisiones en datos lingüísticos reales y verificables.

Los corpus son colecciones de textos o bases de datos cuyo tamaño puede variar; pueden estar conformados tanto por masivos volúmenes de información como por muestras más reducidas. Este aspecto es decidido por los investigadores responsables y está motivado por el propósito que se persiga con el proyecto. Dichas herramientas son representativas en la medida en que muestran una porción de la realidad lingüística a investigar. Los textos no son escogidos al azar, sino que responden a criterios específicos. Además, permiten realizar un análisis integral del fenómeno lingüístico que se desea examinar.

Los estudios basados en corpus resultan efectivos porque ofrecen una visión auténtica del uso real del lenguaje.

A lo largo del presente trabajo, se tratan varios puntos esenciales referidos a los corpus lingüísticos y su influencia en la lexicografía. En primer lugar, se presenta un somero acercamiento a los corpus lingüísticos en el que se consideran su definición, sus características principales, su clasificación y sus funcionalidades. Luego, se explican y evalúan las herramientas informáticas auxiliares más usadas en el trabajo con corpus y en la lexicografía moderna. Y, por último, se describe y valora el papel de los corpus en la práctica lexicográfica, su acción transformadora dentro de la disciplina, que ha mejorado significativamente la calidad y la relevancia de los recursos lexicográficos modernos.

El objetivo de este artículo es examinar de modo panorámico el desarrollo histórico de la lexicografía a partir de la introducción de los corpus informatizados, con la finalidad de reconocer la influencia de estos en la conformación y la actualización de los diccionarios.

## Materiales y Métodos

La investigación realizada constituye una revisión bibliográfica en la cual se emplearon fundamentalmente los métodos bibliográfico- documental e histórico- lógico, útiles para la recopilación de información y el análisis sobre los cambios acontecidos en la lexicografía con la aparición de los corpus lingüísticos. Se partió de la búsqueda de materiales bibliográficos especializados sobre el tema en cuestión, artículos científicos, libros y consulta de recursos digitales referidos a la lexicografía y la lingüística de corpus.

## Resultados

Con anterioridad al siglo XIX, los corpus, que entonces se basaban solo en fuentes escritas, fueron utilizados para la investigación de lenguas muertas, las cuales, al carecer de hablantes contemporáneos, no podían describirse más que por esa vía. Con posterioridad, y hasta principios del XX, la finalidad de dicho recurso se diversificó, destinándose, principalmente, al análisis comparativo de lenguas vivas a fin de establecer su origen y clasificarlas en familias, al estudio de la adquisición de la capacidad lingüística por los niños, a la sistematización léxica para la enseñanza de idiomas extranjeros, a la elaboración de convenciones ortográficas y de gramáticas descriptivas (Villayandre, 2008, p.330).

Cabe mencionar que en las épocas mencionadas las recopilaciones de textos que se incluirían en el corpus se realizaban manualmente, puesto que aún no había nacido la computación ni se habían inventado sus herramientas técnicas. El primer registro de la utilización de los ordenadores con este fin data de 1949, cuando el padre Robert Busa se apoya en un programa computarizado para transcribir la obra de Tomás de Aquino y de otros autores mediante tarjetas perforadas, estableciendo de esta forma un vínculo sostenido hasta nuestros días entre el trabajo con corpus y los ordenadores. Ya en los 60, con el desarrollo de las computadoras, se hizo posible el procesamiento y almacenaje de masivas cantidades de datos, lo que permitió la informatización de los textos.

Estos hechos, que sustentan la importancia de los corpus informatizados, no garantizaron que se reconociera conceptualmente el valor de su carácter digital durante largo tiempo. En este sentido, Francis, Kučera & Mackie (1982) plantean que un corpus es "[...] a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis"; mientras que Sinclair (1991) lo define como "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language". Las definiciones abordadas con anterioridad sugieren que inicialmente no se consideraba

primordial reconocer en las definiciones de corpus su esencia informatizada.

Si bien en sus inicios Sinclair no incluyó el carácter informatizado en su concepto de corpus, cuatro años después reconoció la relevancia de la informática para el desarrollo de investigaciones posteriores y propuso que " The notion of "corpus" refers to a machine – readable collection of ( spoken or written) texts that were produced in a natural communicative setting, and the collection of texts is compiled with the intention (1) to be representative and balanced with respect to a particular variey or register or genre and (2) to be analized linguistically" ( Sinclair, 1996, citado en Gries, 2009).

No puede dejarse de mencionar que, en esta misma etapa, Sinclair estableció una distinción entre lo que constituye un corpus, una colección de ejemplos y un archivo de textos, sentando las bases para el reconocimiento de los corpus como un recurso digital y marcando una pauta para investigaciones posteriores.

Actualmente no se piensa en la posibilidad de usar los corpus de forma manual. Las potencialidades que la informática ha brindado al procesamiento del lenguaje natural son innegables.

Las consideraciones de autores como McEnery & Wilson (1996), Santalla (2005), Rojo (2016) y Sierra (2015) guardan relación entre sí al señalar ciertas características de los corpus, el ser conjuntos de textos debidamente organizados y representativos de una realidad de la lengua y utilizarse para realizar estudios lingüísticos. Se distinguen, no obstante, en ciertos matices y en el énfasis puesto en algunas aristas. McEnery & Wilson (1996) destacan la idea de una muestra finita de textos, específicamente de carácter computacional, que maximiza la representatividad de una variedad lingüística. Santalla (2005), por su parte, amplía el concepto al incluir la homogeneidad del formato y los criterios explícitos de selección; además, resalta que el corpus actúa como modelo de un estado lingüístico determinado. Rojo (2014) introduce la noción de "fragmentos de textos" y hace hincapié en el propósito científico de su estudio, mientras que Sierra (2015) pone énfasis en la representatividad de la lengua en uso y la función práctica de los corpus para realizar análisis lingüísticos.

Los autores antes mencionados y otros coinciden en declarar la importancia de determinados elementos caracterizadores de los corpus y de puntualizarlos debidamente. Es pertinente señalar el valor especial que tiene cada uno de dichos elementos para la lexicografía, aspecto al que se le prestará atención seguidamente.

- Textos en lenguaje natural:

En la definición de corpus ofrecida por Sinclair (1991) se explicita una característica que retoma Rojo (2016) en su definición, referida a que los textos que conformen el corpus deben producirse en situaciones reales, y no ser creados por herramientas computacionales o inteligencia artificial. En el anterior planteamiento se muestra una de las bondades del empleo de los corpus, expresada en la posibilidad de plasmar en los diccionarios el uso de las palabras y frases en contextos verídicos. Esto propicia que las definiciones y los ejemplos sean más significativos y entendibles para los usuarios.

- Tamaño:

El tamaño de los corpus debe ser finito y acordarse por los investigadores antes de comenzar el proyecto, aunque existen corpus abiertos que admiten agregar nuevos textos para garantizar el incremento sostenido de la muestra lingüística.

La voluminosa información contenida en un corpus permite –a través de herramientas de procesamiento del lenguaje natural–constatar y documentar el cambio y la variación lingüísticos desde diversas perspectivas, lo que a su vez facilita la determinación de patrones de uso y el establecimiento de la frecuencia de unidades y fenómenos de la lengua o de las lenguas representadas.

- Representatividad:

Los textos escogidos deben ser una muestra representativa de la lengua o las lenguas que interese analizar o procesar. No se pretende que, con ellos, se generalicen aspectos del idioma dado, sino más bien que, a través de la verificación de datos reales, se pueda dar paso a la ejemplificación o a la validación de teorías. Para que un corpus sea representativo, debe mostrar una diversidad lingüística, incorporando distintos registros, géneros y situaciones de uso.

- Formato electrónico:

La digitalización de la información que contenga el corpus es esencial para los lingüistas o investigadores a cargo de él y para sus futuros usuarios. Esto facilitará la localización de elementos buscados, la obtención de

concordancias, el conteo de ocurrencias y casos, el acceso a los metadatos de cada uno de los textos fuente, entre otras operaciones. Las ventajas de tener un corpus en formato electrónico han hecho quela confección de fichas de papel haya perdido actualidad en la diccionarística moderna por las diferencias entre el uno y el otro soporte, a favor del primero, en cuanto a manipulación y procesamiento de la información, precisión de los datos obtenidos y transferencia de datos. Gracias a las amplias posibilidades que brinda la informática, los lexicógrafos pueden realizar búsquedas con eficiencia y rapidez, conformar listados de palabras, acceder a concordancias, metadatos, la cantidad de apariciones, entre otras.

- Selectividad:

La selección de los textos debe estar sujeta a la finalidad lexicográfica que se pretenda conseguir, y obedece a criterios establecidos con antelación. El trabajo con este tipo de fuentes es un apoyo para los creadores de diccionarios, les proporciona a estos herramientas y técnicas mediante las cuales podrán determinar los vocablos que formarán parte del lemario macroestructural y los que no.

3.2 Clasificación y uso de los corpus

Los corpus constituyen el eje central de una metodología enfocada en el procesamiento y análisis de datos; sin embargo, debido a la gran diversidad de corpus disponibles, se hace imprescindible contar con un marco de referencia preciso que ilustre la funcionalidad que dichas herramientas cumplen en las investigaciones lingüísticas. En esta dirección, resulta necesario entender cómo se clasifican los corpus.

Las clasificaciones de los corpus han sido abordadas por diversos autores a partir de los parámetros fundamentales para su comprensión. Villayandre (2008) clasifica estos bancos de datos atendiendo a: modalidad de la lengua, número de lenguas a que pertenecen los textos, límites del corpus, carácter general o especializado de los textos, período temporal que abarcan los textos, tamaño de los textos y tratamiento aplicado al corpus. Por otra parte, Sierra (2008) sugiere una tipología basada en: el origen de los datos, la espontaneidad del habla, la codificación y anotación, la especificidad de los elementos, la autoría de los elementos, la temporalidad de los elementos, el propósito de estudio, la lengua, la cantidad de texto, la distribución del tipo de texto, la accesibilidad, la documentación y la representatividad. Asimismo, Martín (2009) refiere que estos deben clasificarse como: generales o de referencia, especializados, diacrónicos, monitor, paralelos y de internet. Zapata (2015) se acoge a la tipología sugerida por Torruela & Llisterri (1999), quienes dividen los corpus en dos grandes grupos: orales y escritos, los cuales se subdividen según: el porcentaje y distribución de los textos, la especificidad de los textos, la cantidad de texto que se recoge en cada documento, la codificación y anotación y la documentación que acompaña a los textos. Finalmente, Hincapié & Bernal (2018) establecen una tipología según: el medio de producción de textos, el número de lenguas, la especificidad de los textos en general, la distribución de los textos, el tamaño de las muestras recogidas, la información extra de los textos, y la documentación que acompaña a los textos.

En la clasificación de Villayandre (2008), se abordan aspectos formales, de contenido y temporalidad, a la vez que se logra un elevado nivel de análisis lingüístico, y es esta razón la que conduce a los autores de este trabajo a estimar más acertada tal tipología (Tabla.1) y a adoptarla. Como podrá apreciarse a continuación, la referida clasificación permite analizar un corpus desde distintos puntos de vista; por ello resulta de vital importancia para su aplicación en el campo lingüístico.

**Tabla1. Clasificación concebida por Villayandre (2008)**

| Parámetros | Tipos de corpus |
|---|---|
| Modalidad de la lengua | Escritos: formados únicamente por muestras procedentes de la modalidad escrita de la lengua. |
| | Orales: recogen muestras de lengua hablada, que pueden ser transcripciones ortográficas de grabaciones, utilizadas sobre todo en lingüística de corpus, o grabaciones acompañadas de transcripciones ortográficas y/o fonéticas, más usadas en lingüística y tecnologías del habla. |
| | Mixtos: combinan ambas modalidades, aunque siempre favoreciendo la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, siempre requiere un proceso posterior de transcripción de las grabaciones. |

| | |
|---|---|
| El número de lenguas a que pertenecen los textos | Monolingües: están compuestos por textos en una sola lengua. Se recopilan con el objetivo de dar cuenta de una lengua o variedad lingüística en general (o de un subconjunto de la misma). |
| | Bilingües o multilingües: están formados por textos en dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros y sin compartir criterios de selección. |
| Los límites del corpus | Cerrados: constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número o límite, el corpus se da por finalizado (…) |
| | Abiertos o monitor: son corpus dinámicos, que se mantienen en constante crecimiento, normalmente mediante la introducción periódica de nuevas cantidades de textos según unas proporciones previamente definidas. |
| El carácter general o especializado de los textos | Generales: pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible (…) |
| | Especializados: recogen textos que puedan aportar datos para la descripción de un tipo particular de lengua (‘sublenguaje’). |
| El período temporal que abarcan los textos | Diacrónicos o históricos: incluyen textos de diferentes etapas temporales sucesivas con el fin de poder observar evoluciones de la lengua en un período largo, lo que lo diferencia de los corpus monitor, que no abarcan períodos temporales tan amplios. |
| | Sincrónicos: su finalidad es permitir el estudio de una o más variedades lingüísticas en un momento determinado del tiempo, pero sin prestar atención a su evolución. |
| El tamaño de los textos | De referencia: aquellos formados por fragmentos de textos, habituales en los corpus que quieren proporcionar una información lo más completa posible sobre una lengua y tienen que incluir textos de diferentes géneros, temáticas, etc. |
| | Textuales: aquellos que incluyen textos enteros, sin fragmentar. |
| El tratamiento aplicado al corpus | Simples, en bruto, no anotados o no codificados: consisten en textos guardados sin formato alguno y sin añadir ningún tipo de información adicional, como pueden ser códigos o anotaciones. |
| | Codificados o anotados: están formados por textos a los que se les han añadido, de forma manual o automática, determinadas informaciones. |

### 3.2 Herramientas informáticas usadas para el trabajo con corpus y su aplicación en lexicografía

En torno a la importancia de contar con aplicaciones específicas para el manejo de un corpus en provecho del trabajo lexicográfico, Kilgarrif & Kosem(2012) comentan lo siguiente: “To analyse corpus data, lexicographers need software that allows them to search, manipulate and save data, a ‘corpus tool’. A good corpus tool is key to a comprehensive lexicographic analysis – a corpus without a good tool to access it is of little use”.

El uso de herramientas informáticas aptas para el procesamiento analítico de los corpus ha supuesto una mejoría considerable en lexicografía, teniendo en cuenta todas las posibilidades que se abrieron paso con el inicio de la revolución tecnológica a finales del siglo XX. Las prestaciones que tales elementos de soporte lógico brindan se han convertido en una necesidad técnica y científica, ya que, sin acceso a ellos, se imposibilita la realización de una investigación lexicográfica exhaustiva. Entre sus ventajas, están propiciar análisis lingüísticos cualitativos y cuantitativos, ejecutar tareas relativamente sencillas, como el almacenamiento de grandes cantidades de datos, la búsqueda de información, la determinación de listados de palabras según su frecuencia o la etiquetación morfosintáctica automática (la cual requiere de revisión manual), y tareas más complejas, como la gestión de los metadatos y la clasificación de los textos o el análisis semántico. Se debe significar que, si bien las primeras no requieren de tecnología avanzada–dado que pueden ser programadas mediante

algoritmos sin gran sofisticación–, las últimas dan paso a procesos complejos de procesamiento del lenguaje natural, técnicas de inteligencia artificial y análisis estadísticos. Unas y otras herramientas se mezclan para, según el propósito que persigan, mejorar la eficacia y la calidad en el trabajo con corpus en lexicografía.

De las referidas aplicaciones informáticas, existen dos tipos especializados que se emplean para análisis lexicográficos: los sistemas gestores de diccionarios (SGS) y los sistemas de consulta de corpus (SCC). Los SGS, para Rubio, Estiven & Bernal (2021), son programas especializados en la creación, compilación, escritura, administración y publicación de diccionarios en distintos formatos; son capaces de automatizar procesos y facilitan el trabajo colaborativo. Por su parte, Abel (2012) refiere que los SCC permiten la consulta y el análisis de grandes cantidades de datos lingüísticos en corpus, facilitando la recuperación precisa y rápida de información relevante para la elaboración lexicográfica.

Actualmente ambos sistemas pueden funcionar integrados, potenciando significativamente la interacción entre la obra gestada y su banco fuente, lo cual tiene una repercusión inmediata en la lexicografía.

## Discusión

1. Potencialidades del uso de corpus en lexicografía

La lingüística de corpus, bajo el prisma de las revoluciones científicas, se desarrolla no por cambios conceptuales que impone el referente real, sino que se vincula al surgimiento y consecuente aplicación de nuevas herramientas que implican el descubrimiento de aspectos nuevos que deben ser explicados desde esa perspectiva (Dyson, 1997); pues, si bien es cierto que desde los años 50 –con proyección a los 60– se puede afirmar que existía la lingüística de corpus(LC), la recolección y sistematización exigía un gran despliegue de fuerza de trabajo. Los encargados de realizar esa labor tenían que hacerlo de manera manual, convirtiéndose en un proceso lento que implicaba contar palabra por palabra y analizar el resultado para su valoración semántica. A esto se asociaba la necesidad de contar disponer de con espacios físicos amplios para la acumulación en archivo de todo el material, que debía tener aseguradas las condiciones adecuadas y poseer un orden preciso que viabilizara la conformación del corpus lingüístico.

La aparición de los ordenadores y de los programas informáticos supuso un punto de ruptura y, por tanto, una verdadera revolución científica, puesto que permitieron superar los escollos de momentos anteriores, irrumpiendo la llamada lingüística de corpus computacional al ponerse en función las herramientas de procesamiento del lenguaje natural.

Desde que Sinclair publicó en 1987 el Collins COBUILD English Language Dictionary, se produjo un cambio notable en la lexicografía, al ser este el primer diccionario basado enteramente en corpus; es decir, este proyecto siguió los parámetros establecidos para cualquier diccionario basado en corpus; los lemas y los sentidos de las palabras fueron extraídos del corpus: no se definió ninguna palabra externa a él. Cabe mencionar también que, antes de la elaboración del Birmingham Corpus–este fue el corpus con el que se elaboró el COBUILD–, ya se habían elaborado el Brown Corpus y el LOB, aunque estos no tuvieron muy buena acogida, puesto que no eran corpus de gran tamaño, por lo que resultó imposible constatar una realidad lingüística significativa del funcionamiento de las palabras.

A finales de los 80 se logra publicar la primera edición del COBUILD; los especialistas en lexicografía lo utilizaron para la organización de las entradas, de manera tal que el significado más relevante fuera el primero en aparecer. Fueron capaces de redactar definiciones más precisas que reflejaron el uso contemporáneo de las palabras, proporcionaron ejemplos oracionales auténticos y, además, el uso del corpus les permitió determinar cuál información podría ser omitida y cuál no.

La repercusión de esta obra es evidente. “The impact of corpus data on lexicography since 1987 (the date of publication of COBUILD, the first corpus-driven dictionary') has been overwhelming. At last lexicographers have sufficient evidence to make the generalizations that they need to make with reasonable confidence” (Hanks,2009). El COBUILD marcó un hito antes y un después en el quehacer lexicográfico pues fue el primer diccionario del inglés elaborado en su totalidad a partir de un corpus.

En su publicación “Sobre la construcción de diccionarios basados en corpus”, Rojo (2009) plantea: “El objetivo de un proyecto lexicográfico basado en corpus es, con toda claridad, recoger las palabras que figuran en un corpus representativo de la lengua o variedad lingüística sobre la que se trabaja y reflejar los significados realmente presentes en los textos, incorporando las marcas de uso correspondientes en cada caso”.

Lo relevante en el uso de corpus en el proceso lexicográfico se da en el hecho de que se utilizan para la creación del diccionario solo las palabras contenidas en dichas bases de datos; no se buscan agentes externos de consulta, como otros diccionarios u otro medio. El corpus permite a los lexicógrafos:

- La creación de definiciones más precisas y relevantes a partir de la observación de unidades léxicas en contextos de uso reales.
- La determinación de frecuencias léxicas, parámetro que repercute en la conformación de la macroestructura de los repertorios al posibilitar distinguir los vocablos de uso común de los que son raros o desusados.
- La identificación y extracción de ejemplos auténticos del empleo de las palabras en contextos variados, favoreciendo así la comprensión multilateral, el conocimiento exacto y la adquisición de ellas por parte de los usuarios del diccionario.
- La constatación de neologismos o de nuevos sentidos de vocablos ya existentes.

Los beneficios de la utilización de corpus son indudables. Ello permite acceder a la información de manera expedita, ágil y sencilla; el proceso es ordenado y propicia la actualización de datos constantemente; lo contenido en ellos es representación de una realidad lingüística determinada; dichos recursos brindan la posibilidad de realizar consultasen línea y compartir la información recogida en ellos, entre otras potencialidades.

A diferencia de los corpus, que involucran su desarrollo con el auge de la revolución tecnológica y que forman parte de una disciplina relativamente joven; la lexicografía, disciplina milenaria, ha registrado no solo las transformaciones de la lengua, sino también los cambios acontecidos en la humanidad desde sus diversas áreas. En palabras de Tarp (2014),"dictionaries constitute a privileged mirror of social and cultural development during the past four thousand years, not only in terms of the development of languages, but also of handycraft, economic life, culture, education, natural and social sciences, humanities, sport, and even such exotic phenomena as entertainment, pastime, holiday (…)".

Los cambios que han venido aparejados con la incorporación al proceso lexicográfico de los corpus lingüísticos han sido muchos. La lexicografía tradicional se valía, fundamentalmente, de la consulta de obras, de las observaciones de los especialistas y de la intuición de estos; la actualización de los diccionarios tradicionales era menos frecuente, puesto que se necesitaba más esfuerzo y tiempo para recopilar y analizar datos. Los diccionarios tradicionales solo estaban disponibles en soporte papel. Esto limitaba su accesibilidad y capacidad de actualización.

En contraste con las limitaciones anteriores, la lexicografía moderna está basada en datos empíricos obtenidos de corpus, lo que permite crear definiciones más precisas y relevantes, y proporciona ejemplos contextuales extraídos de usos reales del lenguaje; los diccionarios de nuestros días pueden actualizarse con mayor frecuencia, gracias al uso de corpus y herramientas digitales; con ello, se pueden agregar nuevas palabras y significados de manera más rápida, lo que mantiene a los diccionarios más a tono con la realidad que pretenden describir; las obras lexicográficas contemporáneas están disponibles en formato electrónico, lo cual permite un acceso más amplio y ágil; las versiones en línea y las aplicaciones de diccionarios permiten a los usuarios buscar y obtener información de manera instantánea.

En definitiva, la lexicografía moderna ha sido transformada por el uso de tecnologías digitales y corpus lingüísticos, lo que ha llevado a diccionarios más precisos, actualizados y representativos del uso real del lenguaje.

## Conclusiones

Tras haber examinado –en la panorámica concebida– el desarrollo histórico de la lexicografía a partir de la incorporación en ella de los corpus lingüísticos, podemos concluir que este tipo de herramientas digitales ha propiciado una transformación trascedente en los métodos de trabajo, los recursos y los resultados de dicha disciplina y, hoy en día, tiene un rol determinante en el procesamiento del lenguaje natural y, en consecuencia, en la elaboración de las obras lexicográficas.

Como se ha visto, los corpus no son de índole general, sino que estos son creados y conformados con fines específicos, razón por la cual se hace necesaria una metodología interdisciplinaria basada en su clasificación para lograr un mejor uso de esa herramienta.

La lexicografía desarrollada a partir de la introducción de los corpus electrónicos ha ido superando cada vez más la subjetividad y la normatividad características de la práctica lexicográfica tradicional, poniendo en primer plano el enfoque descriptivo y objetivo y enriqueciendo notablemente sus fuentes de datos.

Los diccionarios creados según estos principios han asegurado una representación fiel de la realidad lingüística, mostrando definiciones y ejemplos lexicográficos comprobables y precisos. El aprovechamiento de los corpus en esta faena ha propulsado el avance del procesamiento del lenguaje natural y el desarrollo de recursos digitales para el perfeccionamiento de los análisis lingüísticos.

El progreso experimentado por la lexicografía desde el hito tecnológico abordado aquí hasta el momento presente ofrece perspectivas prometedoras para la creación de obras de referencia cada vez más adaptadas a las necesidades de los usuarios y más fidedignas en su reflejo del funcionamiento y el uso lingüísticos.

## Referencias bibliográficas


Alonso Ramos, M. (2009). Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionarios? (P. Cantos Gómez, & A. Sánchez Pérez, Edits.) 1191-1207.

Bolaños Cuéllar, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. Forma y Función, 28(1), 31-54.

Dyson, F. (1997). Imagined Worlds. Harvard University Press.

Edo Marzá, N. (2012). Lexicografía especializada y lenguajes de especialidad: fundamentos teóricos y metodológicos para la elaboración de diccionarios especializados. Lingüística, 27(1), 98-114.

Foucalt, M. (1992). Arqueología del saber. Editorial Siglo XXI.

Gelpí Arroyo, C. (2003). El estado actual de la lexicografía: los nuevos diccionarios. En A. M. Medina Guerra, Lexicografía española (págs. 307-327). Ariel .

Gries, Stefan. (2009). Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge. Londres.

Hanks, P. (2009). The impact of corpora. En P. Baker (Ed.), Contemporary Corpus Linguistics (págs. 114-236). Continuum.

Hanks, P. (2012). Corpus Evidence and Electronic Lexicography. En S. Granger, & M. Paquot (Edits.), Electronic Lexicographic (págs. 57-82). Oxford University Press.

Kilgarrif, A.,& Kosem, Iztok (2012). Corpus tools for lexicographers. En S. G. Paquot (Ed.), Electronic lexicography (págs. 31-56). Oxford University Press .

Martín Herrero, C. (2009). Aproximación a ciertas perspectivas en Lingüística de Corpus. (P. Cantos Gómez, & A. Sánchez Pérez, Edits.) A survey of corpus based research[en línea], 1020-1032.

McEnery, T., & Wilson, A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.

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

Nomdedeu, A., & Tarp, S. (s.f.). Introducción a la lexicografía en español: funciones y aplicaciones. Routledge.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. Revista de Lingüística Teórica y Aplicada, 46(1), 118-119.

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

Ramos, M. (2009). Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionarios? A survey of corpus based research, 1191-1207.

Rojo, G. (2014). Sobre la construcción de diccionarios basados en corpus. Revista Tradumàtica: Traducció i Tecnologies de la Informació i la Comunicació, 07, 1-7.

Rojo, G. (2015). Sobre los antecedentes de la lingüística de corpus. Studium grammaticae: homenaje al profesor José A. Martínez, 675-689.

Rojo, G. (2016). Los corpus textuales del español. (J. Gutiérrez-Rexach, Ed.) Enciclopedia lingüística hispánica, 285-296.

Rubio, R. Y., Estiven, J. & Bernal, J. A. (2021). Dictionary Writing Systems y otras herramientas informáticas para la elaboración, administración y publicación de diccionarios. Linguistica y leteratura, 80, 340-360.

Santalla del Río, M. P. (2005). La elaboración de corpus lingüísticos. En I. M. Palacios Martínez, M. Cal Varela, & P. Nuñez Pertejo (Edits.), Nuevas tecnologías en lingüística, traducción y enseñanza de lenguas (págs. 45-66). Servizo de Publicacións e Intercambio Científico, Santiago de Compostela.

Sierra, G. (2015). Introducción a los corpus lingüísticos. Ciudad de México: UNAM.

Sinclair, J. M. (1991). Corpus, concordance, collocation. Oxfordshire: Oxford University Press.

Tarp, S. (2014). «Dictionaries in the Internet Era: Innovation or Business as Usual? Revista Alicantina de Estudios Ingleses(27), 233-261.

Tarp, S. (2015). Excesos en el uso de corpus en la lexicografía:«pesca» de términos y definiciones. Revista de Lexicografía, 21, 145-163.

Tarp, S. (2017). ¿Adiós a los corpus con fines lexicográficos? En L. Ruiz Miyares (Ed.), Estudios de Lexicología y Lexicografía. Homenaje a Eloína Miyares Bermúdez. (págs. 55-75). Centro de Lingüística Aplicada.

Tognini- Bonelli, E. (1996). Corpus linguistics: a practical introduction.Edinburg University Press.

Villayandre, M. (2008). Lingüística con Corpus. Estudios de Historia de la Filología, 30, 329-349.

**Conflicto de intereses**

Los autores declaran no tener ningún conflicto de intereses.

**Declaración de responsabilidad de autoría**

Los autores del manuscrito señalado, DECLARAMOS que hemos contribuido directamente a su contenido intelectual, así como a la génesis y análisis de sus datos; por lo cual, estamos en condiciones de hacernos públicamente responsable de él y aceptamos que sus nombres figuren en la lista de autores en el orden indicado. Además, hemos cumplido los requisitos éticos de la publicación mencionada, habiendo consultado la Declaración de Ética y mala praxis en la publicación.

Yoandra Chuen Gómez y Alex Muñoz Alvarado : Proceso de revisión de literatura y redacción del artículo.

Alejandro Ramos Banteurt : Revisión y corrección de la redacción del artículo.