Revista Internacional DE COOPERACIÓN & DESARROLLO
International Journal of Cooperation & Development

Editorial

La lingüística computacional y de corpus al servicio de la investigación en Lingüística y Educación: un esbozo en Colombia

Computational Linguistics and Corpus Linguistics at the Service of Research in Linguistics and Education: An Overview in Colombia

Gabriel Quiroz Herrera[1] y Carlos Mario Pérez-Pérez[2]

Tipo de artículo: Editorial
Recibido: noviembre de 2023
Revisado: diciembre de 2023
Aceptado: diciembre de 2023

Autores
1 Doctor en Lingüística Aplicada de la Universidad Pompeu Fabra. Profesor Titular de la Escuela de Idiomas y el Departamento de Lingüística de la Universidad de Antioquia. Coordina el Grupo Traducción y Nuevas Tecnologías. Sus principales intereses de investigación son traducción especializada, terminología, lingüística de corpus y aspectos lingüísticos del Procesamiento de Lenguaje Natural.
Correo electrónico: gabriel.quiroz@udea.edu.co
Orcid: https://orcid.org/0000-0003-4940-3022

2 Doctor en Lingüística por la Universidad de Antioquia. Profesor Asistente de la Facultad de Comunicaciones y Filología, de la Escuela de Idiomas y líder del Programa Institucional de Español Académico de la Universidad de Antioquia. Desarrolla su docencia e investigación en escritura académica, lingüística aplicada, lingüística de corpus, tecnologías del lenguaje y traducción.
Correo electrónico: carlosm.perez@udea.edu.co
Orcid: https://orcid.org/0000-0001-6792-322X

Cómo citar:
Quiroz Herrera, G. y Pérez-Pérez, C. M. (2023). La lingüística computacional y de corpus al servicio de la investigación en Lingüística y Educación: un esbozo en Colombia. *Revista Internacional de Cooperación y Desarrollo.* 10(2), 4-09
DOI: 10.21500/23825014.6954

> "Although the analytical methods employed differ considerably across corpus linguistic studies, the main research goal remains constant: to learn more about language and the way it is used with the help of a corpus."
> (Larsson et al., 2022, p. 1).

En el marco de la publicación del número 2 del volumen 10 de la *Revista Internacional de Cooperación y Desarrollo* de la Universidad de San Buenaventura de Cartagena, la lingüística de corpus se constituye como el eje que integra las diversas reflexiones presentadas en torno a la descripción del lenguaje en contextos de investigación, tanto lingüísticos como educativos, permitiendo la observación de diversidad de fenómenos en compilaciones de información a lo largo y ancho de la geografía colombiana. Investigaciones que aportan a la comprensión de las realidades en nuestro entorno local con proyección al impacto internacional en la consolidación de la educación como centro del desarrollo de las sociedades y en sincronía con el lenguaje como fuente de la expresión de las realidades humanas, a partir de múltiples perspectivas, como la semiótica, que permiten abordarlo desde sus dimensiones, tanto biológica como social. Y, por ende, el surgimiento de nuevas áreas de investigación lingüística entre las que se cuenta la lingüística de corpus, la lingüística computacional y la ingeniería lingüística, gracias al acceso, el abaratamiento y el aumento de las prestaciones (capacidad) de los computadores y, en consecuencia, su utilización en diferentes áreas del conocimiento humano.

La lingüística de corpus tiene que ver con los principios y prácticas para estudiar la lengua a partir de la observación de textos en lengua natural, más o menos representativos, almacenados electrónicamente (corpus) y analizados con herramientas computacionales (analizadores de texto, alineadores, sistema de ex-

plotación de datos, extractores terminológicos, etc.) capaces de seleccionar, ordenar, contar y calcular los datos lingüísticos. Consideramos textos en lengua natural todos aquellos discursos orales transcritos o escritos producidos por hablantes nativos, preferiblemente en situaciones auténticas de comunicación. El propósito principal de un corpus es verificar una hipótesis sobre el lenguaje o parte de él, por ejemplo, para determinar cómo el uso de un sonido, palabra, construcción sintáctica u otro elemento lingüístico en particular varía verticalmente (entre los diferentes géneros o registros) (Vásquez, en prensa) y horizontalmente (entre las diferentes áreas del conocimiento) (Crystal, 1992, p. 85; Pérez-Pérez, 2023). Se argumenta que la hipótesis central que hay detrás de la lingüística del corpus radica en que las diferencias del lenguaje no son cualitativas sino cuantitativas.

De esta manera, un corpus no es solo una simple colección de archivos electrónicos de textos que se almacenan en un computador. Antes de esto, un corpus tiene que haber sido objeto de un proceso de marcaje, de acumulación de informaciones estructurales, textuales y lingüísticas (marcajes gramatical y sintáctico) que permiten formalizar tanto las distintas subunidades en que se estructuran dichos textos como las informaciones lingüísticas y textuales (categoría gramatical, función sintáctica, oración, párrafo, etc.) que permitirán, por ejemplo, localizar entre millones de palabras aquellos contextos que contienen una determinada combinación sintagmática o palabra (Quiroz, 2007, pp. 135-136). Es importante tener en cuenta que no todos los corpus tienen los mismos objetivos ni necesitan la misma profundidad de marcaje, pero por leve que este sea, se necesitan herramientas informáticas para (semi) automatizarlo. Tanto la lingüística computacional como la ingeniería lingüística tienen aportaciones que hacer en este terreno: analizadores y etiquetadores de varios tipos, corpus de entrenamiento, desambiguadores lingüísticos, estadísticos. etc. (De Yzaguirre, 1996, pp. 69-71).

En cuanto al uso y aplicación de corpus y herramientas computacionales, estos han sido utilizados, por ejemplo, para la elaboración de diccionarios como el Oxford (Stevenson, 2010), MacMillan (Rundell, 2005), Lema (Fernández, 2003), Diccionario de Uso del Español de América y España (Vox, 2004), entre otros. De igual modo, en la elaboración de gramáticas como *The Logman Grammar of Spoken and Written the English* (Biber et. al, 2000), gramática realizada por un prestigioso grupo de lingüistas y gramatólogos encabezado por Douglas Biber y Stig Johanson, se utilizó el corpus de Longman (más de 40 millones de palabras en varios géneros textuales). En los estudios de traducción, el uso de los corpus se hace desde los años 90 en la alineación de textos, enseñanza de la traducción, entre otros (Baker, 1995, 1998, 1999; King, 1997; Kenny, 1998). En terminología, los corpus son un recurso y una herramienta fundamental para la extracción y compilación de terminologías (Campo et al., 2002), entre otros.

Un ejemplo más cercano a la realidad colombiana, lo constituye la creación del más reciente manual de terminología denominado *Terminología del español: el término* (2024), editado por Quiroz, Burgos & Zuluaga y confeccionado completamente con el Corpus para la Variación Horizontal y Vertical (VaTeHoVe), compilado a partir de artículos de investigación en distintos campos del saber (Quiroz et al., 2024) y diversidad de diccionarios y datos lexicográficos recuperados de investigaciones previas que antecedieron al manual. Igualmente, De Jacobi (2002, p. 1) resume otros usos de los corpus en la elaboración de materiales para la enseñanza de lenguas extranjeras (Johns, 1991, 2002; Aston, 1995, 2000; McEnery & Wilson, 1996, 2000; Lewis, 1997; Mindt, 1997, 2002; Hunston & Francis, 2000; Johns et al., 2008); en estudios sobre lenguaje e ideología (Stubbs, 1994; Fairclough, 2003, 2013; Gerbig, 2008; Flowerdew, 2009, 2017); en estilística (Barnbrook, 1996), entre otros. La lingüística de corpus y la ingeniería lingüística nos proporcionan métodos, prácticas y herramientas para solucionar, en parte, el problema planteado antes.

En la lingüística de corpus, el lenguaje se describe sobre la base de cómo se usa en los textos a diferencia de la intuición del lingüista, es decir, la introspección. A pesar de que la descripción lleva un componente cuantitativo importante, la descripción

cualitativa de los datos juega un papel central. Más que una oposición entre empirismo y racionalismo lingüístico, la lingüística de corpus nos permite complementar, ampliar las observaciones de muchas gramáticas y métodos de enseñanza con datos reales y, en muchos casos, observar fenómenos no descritos antes o descritos pobremente. En ese sentido, fiel al uso de la lingüística de corpus como una metodología, este número acoge una amplia amalgama de trabajos a partir de los cuales el uso de corpus no solo se sintoniza con la vanguardia investigativa, a partir de grandes cantidades de datos lingüísticos, sino que a su vez proporciona un espacio de encuentro de temáticas que versan en los ámbitos de la escritura, la lectura, la argumentación, el currículo, las estrategias para una sana convivencia, junto con pesquisas alrededor de las fórmulas de tratamiento pronominales, la conciencia fonológica y las normas ortográficas del español. A continuación, se describe cada artículo presentado:

En el artículo 1, "Análisis sociolingüístico del discurso académico argumentativo: un enfoque metodológico desde la lingüística de corpus", Vargas propone un análisis de la producción escrita en un ámbito universitario basado en criterios tanto lingüísticos como en variables sociales en la consolidación de textos argumentativos. Sus resultados plantean la necesidad de enfatizar en estos elementos al igual que introduce la noción del tiempo empleado en la producción final de cara a una correcta orientación de las prácticas escriturales.

En el artículo 2, "Tratamientos pronominales en Bogotá, Cali y Medellín, ¿hacia una ampliación de la solidaridad?", Barrero presenta una compilación de artículos relacionados con las formas de tratamiento pronominal en las tres principales ciudades capitales de Colombia: Bogotá, Cali y Medellín, con la finalidad de identificar posibles cambios en el uso de estas formas teniendo como base elementos de orden diacrónico y social. Los resultados apuntan a un cambio en proceso de las formas analizadas, siendo la *solidaridad* la forma de mayor crecimiento gracias a una posible disminución de la inequidad en estas ciudades por factores como la disminución de la pobreza.

En el artículo 3, "Consideraciones conceptuales, históricas y educativas sobre la conciencia fonológica y las normas ortográficas en el español", Díaz aborda el papel que juega la conciencia fonológica en el surgimiento de la norma ortográfica. Sus hallazgos plantean la relación entre la norma casual y los contextos de reacción inmediata, mientras que la norma formal se manifiesta en escenarios académicos o en presencia de roles de autoridad. De este modo, el autor aboga por una mayor interiorización de ambas normas según las variaciones descritas para un mayor desarrollo de la conciencia ortográfica.

En el artículo 4, "El meme de Internet como recurso pedagógico para el fortalecimiento de la lectura crítica", Anaya, Acevedo & Cediel abordan el uso del *meme* como una estrategia para potenciar la lectura crítica en la escuela. Sus hallazgos indican una relación entre la mejora de la perspectiva crítica, comprensiva y analítica a partir de la implementación de la estrategia basada en problemas de comprensión e interpretación lectora, en la mejora de los resultados en las pruebas nacionales estandarizadas cuando se proponen análisis que versan alrededor de textos discontinuos.

En el artículo 5, "Análisis sobre la producción y percepción del simbolismo sonoro corpóreo e imitativo en adolescentes bilingües del Montessori British School de Bogotá a partir de encuestas virtuales", Achipis da cuenta del simbolismo sonoro corpóreo e imitativo en un contexto de trabajo bilingüe con adolescentes en edad escolar. Sus resultados apuntan a la caracterización del fenómeno tanto en inglés como español. Así, según la autora, en el simbolismo sonoro imitativo los hombres suelen recurrir a formas no duplicadas con mayor frecuencia, mientras que las mujeres tienden a formas monosilábicas, independiente de la lengua utilizada.

En el Capítulo 6, "La reflexión y el diálogo como instrumentos para fomentar una sana convivencia", Pardo presenta la metodología y el análisis de datos empleados para el desarrollo de un modelo pedagógico basado en la Pastoral Educativa. Los resultados establecen la ruta de trabajo para el desarrollo y fortalecimiento de las habilidades socioemocionales so-

bre las cuales el modelo propuesto busca promover el diálogo intercultural e interreligioso, al fomentar el respeto por la diversidad y la convivencia pacífica.

En el artículo 7, "Fortalecimiento de la lectura crítica en estudiantes de secundaria a través del uso de la intertextualidad", González, Gómez y Cogollo abordan la intertextualidad como un elemento clave en el fortalecimiento de la lectura crítica en estudiantes en etapa de formación escolar. Sus hallazgos resaltan el valor pedagógico de la intertextualidad en las prácticas lectoras, al potenciar el reconocimiento del diálogo y la postura crítica entre el texto y el lector y, en consecuencia, un impacto positivo en el desempeño de las y los estudiantes en las pruebas nacionales e internacionales.

En el artículo 8, "Representaciones de las concepciones del currículo y su gestión en las prácticas pedagógicas de maestros de instituciones educativas de Córdoba, Colombia", Doria, Pérez y Salgado reflexionan sobre la concepción que del currículo tiene el maestro y la relación con sus prácticas pedagógicas en la cotidianidad. Los resultados permiten observar la interacción entre las teorías críticas y dialógicas del currículo y la concepción curricular reflexiva bajo la cual los maestros, objeto del presente estudio, procuran una constante mejora de sus acciones educativas a partir de la comprensión de la realidad educativa en la cual están inmersos.

En suma, el actual volumen presenta a la comunidad científica una importante compilación de artículos que contribuyen con la agenda de investigación de nuestros contextos. Trabajos que desde lo lingüístico y lo pedagógico permiten observar nuevas tendencias de análisis que impactan de manera directa las formas de entender las realidades actuales y proyectar las futuras indagaciones en lo educativo. Investigaciones que gracias a las bondades de la lingüística computacional y de corpus emplean menor tiempo en abarcar grandes cantidades de información y, por ende, una optimización de los recursos disponibles para los propósitos trazados. Una amalgama de posibilidades en las cuales el lenguaje se configura como un elemento transversal en la expresión de los saberes humanos.

Finalmente, un reconocimiento a la editora del presente volumen, Mg. Ibelis Blanco, por permitir una reflexión desde la lingüística computacional y de corpus en ámbitos del lenguaje y la educación en procura de una observación preliminar de nuestro contexto colombiano. Asimismo, un agradecimiento a las autoras y los autores de cada artículo, el tiempo empleado en cada investigación contribuye a la cultura investigativa de nuestro país. Sus escritos, sin duda, abren caminos de investigación para las actuales y futuras generaciones de investigadoras e investigadores.

Referencias

Aston, G. (1995). Corpora in Language Pedagogy: matching theory and practice. En C. Guy & B. Seidlhofer. (Eds.), *Principle and Practice in Applied Linguistics* (pp. 257-270). Oxford University Press.

Aston, G. (2000). Learning English with the British National Corpus. En P. Battaner & C. López. (Eds.), *VI Jornadas de corpus lingüístics: corpus lingüístics i ensenyament de lenguas* (pp. 25-40). Instituto Universitario de Lingüística Aplicada.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, *7*(2), 223-243.

Baker, M. (1998). Translation studies. En M. Baker. (Ed.), *Routledge encyclopedia of translation studies* (pp. 277-280). Routledge.

Baker, M. (1999). The role of corpora in investigating the linguistic behaviour of professional translators. *International journal of corpus linguistics*, *4*(2), 281-298.

Barnbrook, G. (1996). *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh Univeristy Press.

Campo, A., Martínez, M., Polo, N., de Yzaguirre, L., Tebé, C., & Cabré, M. T. (2002). La utilización de corpus paralelos alineados en la docencia de la traducción y de los lenguajes de especialidad. En L. Iglesias, S. Doval & M. Gómez. (Eds.), *Studies in contrastive linguistics.: proceedings of the 2nd International Contrastive Linguistics Conference, Santiago de Compostela, October, 2001* (pp. 71-82). Universidad de Santiago de Compostela.

Crystal, D. (1992). *An Encyclopedic Dictionary of Language and Languages*. Oxford University Press.

De Jacobi, C. (2002). Computadores, corpora y la enseñanza de español en cursos de letras. *Anuario brasileño de estudios hispánicos*, (12), 29-44.

De Yzaguirre, L. (1996). Ingeniería lingüística y terminología. *Terminómetro, Monográfico: La terminología en España*, 69-71.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. Psychology Press.

Fairclough, N. (2013). *Critical discourse analysis: The critical study of language*. Routledge.

Flowerdew, J. (2017). Corpus-based approaches to language description for specialized academic writing. *Language Teaching*, *50*(1), 90-106.

Flowerdew, J. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International journal of corpus linguistics*, *14*(3), 393-417.

Gerbig, A. (2008). *Language, people, numbers: Corpus linguistics and society*. Rodopi.

Hunston, S., & Francis. G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. John Benjamins Publishing Company.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. En C. Huo, L. Jiao & W. Wang. (Eds.), *An Empirical Study of Corpus Application in the Teaching of English Writing* (pp. 1-16). Creative Education.

Johns, T. (2002). Data-driven learning: The perpetual challenge. En B. Kettemann & G. Marko. (Eds.), *Teaching and learning by doing corpus analysis* (pp. 105-117). Brill.

Johns, T., Hsingchin, L., & Lixun, W. (2008). Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning*, *21*(5), 483-506.

Kenny, D. (1998). Corpora in translation studies. En M. Baker. (Ed.), *Routledge encyclopedia of translation studies* (pp. 50-53). Routledge.

King, M. (1997). Evaluating translation. En C. Hauenschild & S. Heizmann. (Eds.), *Machine translation and translation theory* (pp. 251-263). Mouton de Gruyter.

Larsson, T., Egbert, J., & Biber, D. (2022). On the Status of Statistical Reporting Versus Linguistic Description in Corpus Linguistics: A Ten-year Perspective. *Corpora*, *17*(1), 137-157.

Lewis, M. (1997). Pedagogical implications of the lexical approach. En J. Coady & T. Huckin. (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 255-270). Cambridge University Press.

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

McEnery, T., & Wilson, A. (2000). Linguistic Córpora and Language Teaching: Corpus Based Help for Teaching Grammar. En P. Battaner & C. López (eds.), *VI Jornadas de corpus lingüístics: corpus lingüístics i ensenyament de lenguas* (pp. 65-77). Instituto Universitario de Lingüística Aplicada.

Mindt, D. (1997). Complementary distribution, gradience and overlap in corpora and in ELT: Analysing and teaching the progressive. *Language and Computers*, *19*, 227-238.

Mindt, D. (2002). A corpus-based grammar for ELT. En B. Kettemann & G. Marko. (Eds.), *Teaching and learning by doing corpus analysis* (pp. 91-104). Brill.

Pérez-Pérez, C. M. (2023). *El término en español: hacia una ampliación del concepto a partir de su caracterización lingüística en cinco áreas de especialidad* [Tesis Doctoral, Universidad de Antioquia].

Quiroz, G. (2007). Preparación y procesamiento de un corpus para la creación de materiales en la clase de español para propósitos específicos. En M. González & M. Valmaseda (Coords.), *Actas del X Congreso Brasileño de Profesores de Español* (pp. 131-150). Embajada de España en Brasil – Consejería de Educación.

Quiroz, G., Burgos, D., y Zuluaga, F. (2024). *Terminología del español: el término*. Routledge.

Quiroz, G., Pérez-Pérez, C. M., y Vásquez, D. (2024). Metodología para un proyecto terminológico basado en corpus. En G. Quiroz, D. Burgos y F. Zuluaga (Eds.), *Terminología del español: el término*. Routledge.

Stubbs, M. (1994). Grammar, text, and ideology: computer-assisted methods in the linguistics of representation. *Applied linguistics*, *15*(2), 201-223.

Vásquez, D. (En prensa). *Caracterización lingüística de las unidades terminológicas para determinar niveles de especialización textual* [Tesis Doctoral, Universidad de Antioquia].

Recursos lexicográficos

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman grammar of spoken and written English*. Longman.

Fernández, F. (2003). *Lema. Diccionario de la Lengua Española*. Vox.

Rundell, M. (2005). *Macmillan English Dictionary for Advance Learners of American English*. Palgrave Macmillan.

Stevenson, A. (2010). *Oxford Dictionary of English*. Oxford University Press.

Vox. (2004). Diccionario de Uso del Español de América y España. McGraw-Hill.