EDICIONES Universidad Valladolid

Andrés Grajales Ramírez, Jorge Molina Mejía y Pablo Valdivia Martin (eds.), *Humanidades Digitales, Corpus y Tecnología del Lenguaje. Una mirada desde diversos casos de estudio*. Facultad de Comunicaciones y Filología Universidad de Antioquia / Universidad de Groningen Press, 2023, 344 pp.

DOI: https://doi.org/10.24197/redd.7.2024.164-179

Resulta imposible obviar la transformación que han supuesto las Humanidades Digitales (HHDD) en el campo de la investigación. Las nuevas tecnologías y herramientas informáticas aplicadas a las Humanidades han abierto las puertas a múltiples de vías de estudio poco abordadas hasta la fecha. Estas tecnologías propician, además, el diálogo entre numerosas disciplinas, desde la psicología, el arte, la historia o, por supuesto, la lingüística. La aparición de *softwares* sostenibles y de libre acceso favorece la puesta en común de estudios con investigadores de distintas instituciones en cualquier parte del mundo, nuevas metodologías en docencia o, sencillamente, facilita el acceso a gran cantidad de información a cualquier público interesado. Parece innegable, por tanto, que nos encontramos ante un cambio de paradigma en las metodologías de estudio y de recopilación de información.

Sin embargo, tal y como apunta Pablo Valdivia en el prefacio del libro que nos ocupa (p. 11), aún no existe pleno consenso sobre la definición y por tanto la acotación de las HHDD. Nos permitimos tomar como válida la aproximación que se hace en esta obra, que sería la unión entre tecnología y las múltiples áreas dentro de las Humanidades y las Ciencias Sociales a las que esta puede aplicarse

Humanidades Digitales, Corpus y Tecnología del Lenguaje es el título, pues, del libro que tenemos entre manos, resultado de un proyecto conjunto de la Facultad de Comunicaciones y Filología de la Universidad de Antioquia y la Universidad de Groningen. Publicado en Países Bajos en 2023, ejercen como editores Andrés Grajales Ramírez y Jorge Molina Mejía, profesores del departamento de Lingüística y Filología de la Universidad de Antioquia, así como Pablo Valdivia Martin, profesor de Cultura Europea y Literatura en la Universidad de Groningen. La obra describe distintas propuestas de diseño, creación y análisis de corpus a través de quince artículos elaborados por

diversos autores de instituciones varias. En ella, se explica desde varios enfoques la relación entre las Humanidades Digitales con las Tecnologías del Lenguaje y su aplicación a los corpus digitales en áreas de investigación poco o nada exploradas, bien por la falta de datos para el análisis o por la gran inversión de recursos necesaria para llevar a cabo la búsqueda de estos. Las propuestas que presentan estos artículos sirven para facilitar considerablemente el procesamiento de grandes cantidades de datos, lo que hasta ahora no hubiese sido posible sin una inversión de tiempo y trabajo titánica.

El volumen consta de un prefacio, una amplia introducción al contenido del libro que se ofrece tanto en inglés como en español, y un total de 15 artículos separados por tres bloques diferenciados por temática. El primer bloque (pp. 31-119) está dedicado a las Humanidades Digitales aplicadas a diferentes disciplinas dentro de las propias Humanidades y las Ciencias Sociales, como el arte, las bibliotecas en línea o la historia. El segundo, que abarca las pp. 121- 242, se centra en la lingüística de corpus detallando distintas tipologías de corpus digitales. En la tercera y última parte (pp. 245-323), se explican los procesos de construcción de varios corpus a partir de la utilización de tecnologías de Procesamiento de Lenguaje Natural (PLN) y de diferentes softwares de análisis. Todos ellos poseen un evidente enfoque lingüístico, aunque aplicado en diferentes áreas de trabajo. Los capítulos están redactados indistintamente en español o inglés, pero cada uno contiene un resumen o *abstract* en ambos idiomas.

Esta agrupación por bloques de los artículos que componen el libro ayuda a la comprensión de la temática de cada capítulo. Y es que, a pesar de que todos los trabajos incluidos en la obra tienen como tema principal las Humanidades Digitales y la construcción de corpus con recursos computacionales, la información no viene expuesta de la misma forma ni se centra en la misma parte del proceso, como veremos más adelante.

El prefacio (pp. 11-13) lo firma Pablo Valdivia Martín, de la Universidad de Groningen, y en él reflexiona sobre las Humanidades Digitales, la lingüística de corpus, las tecnologías del lenguaje o la inteligencia artificial y lo que todo ello supone para la investigación y análisis en las distintas áreas de estudio de Humanidades. Hace un breve resumen de la estructura de la obra y el enfoque interdisciplinar que han tenido presente en los artículos que se han escogido, de modo que sirva como ejemplo de las múltiples posibilidades de estudio que ofrecen las nuevas herramientas de análisis aplicadas a este campo.

La introducción (pp. 15-30) viene a cargo de dos de los editores de la obra, Jorge Molina Mejía y Andrés Grajales Ramírez, ambos de la Universidad de Antioquia. Como mencionábamos anteriormente, está disponible tanto en inglés como en español. Uno de los motivos de esta decisión, según sus editores, es que la información contenida se pueda difundir fácilmente tanto a los investigadores que lo precisen como a los docentes. La introducción incluye también un breve resumen de cada bloque y los capítulos que lo componen.

El primer bloque comienza con el artículo "Entender el Arte Outsider en el contexto de las Humanidades Digitales" (pp.33-54) de John Roberto y Brian Davis, de la Universidad de la Ciudad de Dublín. Abordan el Arte *Outsider* (AO), un movimiento artístico especialmente creativo e innovador, creado por artistas que por su condición se encuentran en un contexto de marginación social. Asimismo, presentan el Proyecto de Arte Outsider dentro del marco de las HHDD. Se trata de un estudio transdisciplinario, que se nutre de exposiciones digitales ya presentes en museos virtuales, como OmekaS, para la construcción de un corpus que asocia texto e imágenes, además de una ontología con información sobre AO, en la que se jerarquizan los conceptos semánticos y que sitúa al artista como categoría central, de modo que se minimiza la dificultad que supone la heterogeneidad en los datos. Las imágenes se escanean y categorizan y los textos, de distinta tipología, que se han analizado con CATMA, un software de Procesamiento de Lenguaje Natural (PNL) y *Machine Learning*, de acceso abierto y basado en XML y TEI. El resultado es una fuente de información sobre AO inexistente hasta la fecha, con gran cantidad de datos analizados objetivamente, y que favorece la investigación sobre un tipo de arte fuera de lo normativo y que ha cobrado relevancia en los últimos años.

El capítulo II tiene como protagonista la "Biblioteca Virtual de la Filología Española (BVFE) y su acervo hispanoamericano" (pp. 55-72). Los autores son Jaime Peña Arce y M.ª Ángeles García Aranda, de la Universidad Complutense de Madrid. Este proyecto forma parte de la biblioteca de la UCM desde 2010 y se basa en los trabajos de investigación del llorado Manuel Alvar Ezquerra (1950-2020), a quien está dedicado. La BVFE está orientada a la investigación y facilita el acceso a obras lingüísticas sobre el español. Cuenta actualmente con más de 9300 registros divididos en diccionarios, gramáticas, fichas de autores o atlas lingüísticos entre otros materiales, además de obras y autores hispanoamericanos o incluso textos de lenguas amerindias que se hablan en territorio hispano. Las obras digitalizadas se localizan en búsquedas intensivas en diferentes repositorios y se almacenan

en una base de datos que, una vez realizadas las verificaciones necesarias, se depositan en un servidor, al que el usuario accede introduciendo parámetros de búsqueda. Aunque se enfrenta a inconvenientes en cuanto a cambios de URL, crecimiento exponencial de materiales digitalizados en línea, o incidencias en mantenimiento, es un proyecto muy consolidado con una labor ampliamente reconocida y líder de visitas.

El tercer capítulo "De dos bases de datos relacionales a una base de datos XML. El proyecto COMREGLA", (pp. 73-90) lo firman diferentes autores en una colaboración de la Universidad de Salamanca, Universidad Complutense de Madrid y el IES Sant Margal. Presenta COMREGLA, un trabajo de unificación en XML de dos bases de datos relacionales de REGLA, un proyecto de estructuras predicativas de los verbos del griego y el latín que se fundamenta en la Gramática Funcional de Dik (1997) y en el que participan miembros de distintas universidades españolas. COMREGLA nace para facilitar el acceso y compatibilidad de los datos de REGLA con otras herramientas de tratamiento automático del lenguaje, con XML como estándar de anotación. De esta forma se mejora la efectividad de ambas bases de datos y el análisis de estructuras complejas de coordinación y subordinación de las estructuras predicativa, lo que con REGLA no era posible. Para ello, se migran las bases de datos y se reorganizan y almacenan en XML. Al no analizar elementos sueltos sino el texto completo se contemplan las relaciones entre unidades básicas, se añade información léxica a los elementos y se observan los niveles y tipos de coordinación y subordinación de los predicados. Se solventan así varias carencias de REGLA sin renunciar a su análisis pormenorizado, mejorando la compatibilidad con otras herramientas y recursos.

El artículo, titulado "Análisis del epistolario del coronel Anselmo Pineda con Python: una mirada al proyecto coleccionista y al territorio desde las redes sociales y el aprendizaje automático" (pp. 91-119), por Santiago Alejandro Ortiz Hernández, de Red Humanidades Digitales en Colombia, cierra este bloque. En él perfila y destaca la figura de Anselmo Pineda, uno de los mayores coleccionistas del siglo XIX en este país. Su colección de documentos públicos se ha convertido en archivo de estado y una representación de la historia nacional. Este artículo hace una extensa descripción biográfica del coronel Pineda y detalla la creación de un corpus con las 3631 epístolas personales que se conservan. Para el análisis de los datos contenidos en las cartas, estas se han transcrito, geoetiquetado e incluido en una base de datos con información relevante como destinatario, remitente, fecha o tema. Mediante estadísticas descriptivas se observa la frecuencia la

red de contactos de Pineda, utilizando para ello mapas de distribución de lugar con software de redes como Network y Holoviews. Por otro lado, los atributos generados a partir de la minería de texto y el PLN sirven al algoritmo de aprendizaje automático (modelo Random Forest) para la posterior clasificación y disección del corpus epistolar. Este tipo de estudios suponen un gran aporte para la investigación con grandes cantidades de documentos y la experimentación con nuevas metodologías que pueden extrapolarse a distintos corpus documentales.

Este primer bloque hace una correcta introducción a los trabajos de corpus digitales y el análisis de datos con herramientas informáticas. No profundizan excesivamente en los aspectos más técnicos de construcción de corpus o del procesamiento de los datos, de modo que un lector poco experto puede formarse una idea general sobre el proceso de construcción de corpus que en los siguientes bloques de artículos será más específica, como veremos más adelante.

Por otra parte, los cuatro artículos de este bloque subrayan una de las premisas principales del libro, trabajar con información disponible solo parcialmente o que no era accesible de cara a la investigación. Resaltan el potencial que ofrecen las herramientas informáticas de analizar grandes cantidades de datos o minimizar los tiempos de búsqueda de información y ponen el foco en áreas de estudio que casi no se habían abordado hasta el momento dentro de la historia, la lingüística o el arte. Un buen ejemplo de ello se explica en el primer capítulo. El Arte Outsider es un movimiento artístico prácticamente desconocido hasta la fecha, a pesar de la influencia que ha supuesto para el arte más común (por conocido) que está presente en los museos tradicionales. Sin embargo, el procesamiento del material teórico y pictórico existente sobre este tipo de arte con un enfoque computacional favorece en gran medida el acceso por parte de cualquier investigador que quiera analizar este material. Podría decirse, en definitiva, que las Humanidades Digitales "democratizan" el acceso a la información.

El segundo bloque, dedicado a la construcción de corpus, comienza con el artículo de Carolina Julià Luna titulado "Desarrollo de un corpus de atlas lingüísticos" (pp. 123-142), más concretamente CORPAT, una herramienta digital de bases de datos que recopila atlas lingüísticos regionales del español europeo, con el fin de ampliar la investigación sobre el cambio lingüístico e historia de la lengua en España y preservar el patrimonio lingüístico y cultural de este país. Contiene una base de datos MySQL 5.6 relacional de código abierto, que une la información lingüística con la geográfica. Además de explicar el funcionamiento de esta herramienta, el texto recoge la historia de

los estudios realizados en materia de geografía lingüística desde sus inicios, sus variantes metodológicas y la evolución del proceso de automatización, desde su planteamiento en los años 70 por Manuel Alvar y Verdejo (1978) hasta los avances actuales con la aplicación de aplicaciones como Google My Maps o de ALPI (Atlas Lingüístico de la Península Ibérica) en 2010. CORPAT se encuentra aún en fase de desarrollo, pero facilita la difusión de materiales, ofrece una perspectiva histórica y actual de la geografía lingüística española y europea y almacena datos procedentes de diversas fuentes, como bibliotecas y centros de investigación.

En el capítulo VI "La propuesta del C-ORAL-BRASIL para el tratamiento de datos multimodales en corpus: el proyecto piloto del corpus BEST" (pp. 143-162), que firman Camila Barros y Heliana Mello, de la Universidad Federal de Minas Gerais, en Brasil, se aborda el tratamiento de datos multimodales en el corpus BGEST, que se engloba en el proyecto de investigación C-ORAL-BRASIL. BGEST es un corpus multimodal, aún en fase de optimización, que estudia la prosodia gestual en la mayor parte de lenguas romances. Maneja una gran cantidad de datos de diferente tipología e incorpora además datos gestuales, con el fin de impedir la pérdida de información en las transcripciones, al considerar el gesto como pico de energía con significado semántico. En este sentido, el proyecto intenta solucionar los inconvenientes de corpus multimodales que manejan una gran cantidad de información, invirtiendo más recursos y reduciendo los tiempos de tratamiento de datos sobre todo en la captura de gestos. Estos se analizan con ELAN, un software abierto de anotación multimodal que incorpora información sobre movimientos y dirección de los gestos. BGEST evidencia la necesidad de un sistema multimodal de cara al futuro para los proyectos de recopilación de datos en actos de habla.

El siguiente artículo, "Las tecnologías del lenguaje y las lenguas indígenas mexicanas: constitución de un corpus paralelo amuzgo-español" (pp. 163-183), supone otro ejemplo ilustrativo de cómo las Tecnologías del Lenguaje Humano (TLH) ayudan a disminuir la brecha tecnológica entre comunidades y prevenir la desaparición de lenguas indígenas por falta material de investigación y recursos. Antonio Reyes Pérez y H. Antonio García Zúñiga, de la Universidad Autónoma de Querétaro y el Instituto Nacional de Antropología e Historia respectivamente, ambos en México, presentan el primer corpus paralelo amuzgo-español. Recopila muestras reales del amuzgo, una lengua indígena mexicana, procedente de la familia otomangue, muy diversa y con variantes internas. Para el corpus se obtienen datos orales del amuzgo producidos en contexto natural con diferentes grupos

de hablantes. Estas muestras se registran, etiquetan y transcriben y se someten a un análisis acústico, tonal y fonológico con herramientas que ya hemos mencionado anteriormente, como ELAN. Para la fase paralela amuzgo-español, se realiza un proceso de glosado y categorización lingüística de cada segmento. Ello permite, con la ayuda de distintos softwares de alineación o traducción asistida, que las transcripciones se hagan con la mayor fidelidad posible y se puedan alinear las categorías gramaticales y sus significados. El corpus está aún en fase de implementación, pero ha sentado la base para futuros estudios de ampliación de datos en este campo.

En el capítulo VIII ("Bases metodológicas: la construcción de un corpus para la detección de mentiras y la evaluación de la credibilidad", pp. 185-200), a cargo de Pedro Eduardo Hernández Fuentes, de la Universidad Nacional Autónoma de México, se expone una propuesta de metodología interdisciplinar entre lingüística y psicología, del Language and Cognition Laboratory of the Cognitive Sciences Research Center (UAEM), con la construcción de un corpus de detección de la mentira y evaluación de la credibilidad en un testimonio. Parte de las investigaciones de DePaulo et al. (2003) y Vrij (2018), que señalan la importancia de la información verbal para evaluar la mentira. El proyecto ha creado una base de datos con registros individuales que ayudan a detectar patrones de mentira en el discurso, así como una plataforma de distribución de textos. Se han efectuado entrevistas en las que se obtienen narraciones tanto verdaderas como falsas, que se transcriben, etiquetan y analizan con el método doble ciego. Se detectan patrones en el léxico, así como en las pausas, negaciones, adverbios, tiempos verbales, etc. Este tipo de entrevistas se han modificado a lo largo de los años, y existen otras vías de análisis abiertas, como las micro expresiones faciales, pero se confirman como más efectivos los análisis conjuntos del comportamiento no verbal con el análisis lingüístico que por separado para la evaluación de la credibilidad en un hablante.

Karen Lorena Castro, de la Universidad de Salamanca, recoge en el capítulo IX ("*Türkisch für Anfänger*: propuesta de un corpus del alemán coloquial actual, ejemplificado a partir de las fórmulas rutinarias de saludo", pp. 201-216) un proyecto de corpus didáctico pionero para la enseñanza del alemán coloquial. Se trata de una base de datos que compila fórmulas rutinarias de saludo, incluidas como fraseolexemas del alemán, extraídas de la serie *Türkisch für Anfänger*. El artículo hace una revisión teórica sobre las características y clasificación de las fórmulas rutinarias, así como su complejidad y polifuncionalidad en la teoría de actos de habla, todo ello con un enfoque didáctico. Se han transcrito 52 capítulos de la serie, que se

caracterizan por la "oralidad fingida". La serie ilustra situaciones cotidianas actuales e interculturales, lo cual garantiza la variedad diatópica, diafásica y diastrática del alemán coloquial actual en el corpus. El análisis cuantitativo del corpus indica usos de estás fórmulas rutinarias por franjas de edad y permite descubrir patrones de lengua o realizar comprobaciones empíricas. Además, confirma que muchas de las fórmulas de saludo no corresponden con las que aparecen en los manuales de aprendizaje, ya que estos se nutren de norma escrita y no oral. Su uso orientado al aprendizaje facilita el uso contrastivo de estas microexpresiones en traducciones, y tiene como finalidad que los estudiantes de alemán incorporen estas fórmulas a su vocabulario en su contexto correcto.

El segundo bloque lo finaliza el artículo ("CLEC – Corpus Colombiano de Aprendices de Inglés: primer corpus de producción escrita de aprendices de inglés en Colombia disponible en línea", pp. 217-242), de M.ª Victoria Pardo Rodríguez y Antonio Jesús Tamayo Herrera, de la Universidad de Antioquia (Colombia) y el Instituto Politécnico Nacional de México, también dedicado a un corpus de uso didáctico, CLEC (Corpus Colombiano de Aprendices de Inglés), un proyecto de Traducción y Nuevas Tecnologías (TNT) de corpus en línea de producción escrita de aprendices de inglés en Colombia. Está basado en los principios de *Learner Corpora* (LC) y en el criterio de Granger (2002, p. 7) y Gilquin (2015, p. 1). Los datos se extraen de entrevistas a estudiantes de diferentes niveles inglés con español como lengua nativa, en las que responden con lenguaje propio a cuestiones de la vida cotidiana. La particularidad de este corpus radica en que se transcriben los errores y su contexto, que se etiquetan en un software de análisis, en base al *Manual of Error Tagging* de Louvain University (Dagneaux *et al*, 2005). Estos errores se categorizan y se interpretan, gracias también a los metadatos señalados en el material recogido. Aunque las interpretaciones pueden variar, obedecen a un mismo criterio de etiquetado. La extracción de datos y análisis estadísticos se realizan con Wordsmith (Scott, M., 2005) y se recurre a Lancsbox (Brezina et al, 2015) para el detalle de error y gráficos. Mediante la aplicación de búsqueda se pueden añadir, modificar y eliminar los errores, lo que lo convierten en un corpus revisable y modificable por estudiantes, profesores o lingüistas.

Los casos de estudio que contiene esta segunda parte ilustran el proceso de creación de corpus, como habíamos adelantado. Una de las características comunes en todos ellos es la recopilación y procesamiento de datos multimodales que sirven como material de trabajo para estudios lingüísticos o interdisciplinares. Se detalla más en profundidad el proceso de recogida de

datos para cada uno de los casos, dejando constancia de uno de los activos más importantes que ofrecen las HHDD: la obtención de nuevos materiales de uso didáctico y el acceso fuentes de datos no disponibles hasta el momento o poco accesibles para su análisis.

Cabe destacar la acertada elección de artículos para este bloque, que ejemplifican la amplia variedad de posibilidades de estudio. Desde una base de datos para análisis fonológicos que aporta no únicamente datos procedentes de transcripciones a partir de grabaciones, sino que integran imágenes con información gestual; la obtención de material de estudio para la conservación de lenguas en peligro de desaparición (véase el corpus paralelo amuzgo-español) y de geolocalización de regionalismos de una lengua, o los nuevos métodos de enseñanza de lenguas extranjeras mediante corpus digitales, que complementan la teoría de los manuales de texto con corpus multimodales con ejemplos de actos de habla y propician el uso en su contexto de expresiones coloquiales. No menos importante es la aplicación de estos corpus multimodales para aumentar la efectividad de análisis en estudios interdisciplinares. Todos estos ejemplos evidenciarían un cambio de paradigma en cuanto a metodologías de estudio, en las que las TIC (Tecnologías de la Información y la Comunicación) abarcan cada vez más terreno, no solamente de cara a la investigación, sino también para la docencia.

El tercer y último bloque, que incluye los cinco artículos a continuación, describe distintos casos de análisis de corpus y Procesamiento de Lenguaje Natural (PLN).

El primer artículo, titulado "La pronunciación de los grupos de consonantes en hispanohablantes basándose en el corpus oral leído checo" (pp. 247-272), a cargo de Kateřina Pugachova & Jitka Veroňková, de la Facultad de Artes de la Charles University en Rep. Checa, analiza la pronunciación de los grupos de consonantes en hispanohablantes basándose en el corpus oral leído checo. Se trata de un análisis perceptivo de la pronunciación de determinados grupos de consonantes del checo en hablantes de español, un total de 23, consonantes ya de por sí de difícil pronunciación para un hablante hispano. Toma como referente distintos trabajos previos sobre pronunciación de lenguas en hablantes de español que profundizan en los factores que influyen en la pronunciación para mejorar el acento. El corpus se compone de grabaciones de discursos en checo por hablantes hispanos, con español como L1, de distinta procedencia, grado de preparación en checo y de estancia en la República Checa. Se utilizan distintos softwares tanto para la selección de palabras que figuran en los textos que leen los participantes

como para los análisis posteriores. Finalmente, se obtiene el número de realizaciones de estos grupos de consonantes en distintas posiciones de la palabra y se detallan los diferentes factores y fenómenos que ocurren en la pronunciación. Este corpus proporciona una buena base de material para la investigación y suple en parte la escasez de libros de texto para la enseñanza de checo a españoles.

M.ª Amparo Soler Bonafont, de la Universidad Complutense de Madrid, presenta en el capítulo XII ("Relacionando los análisis cualitativo y cuantitativo. Una propuesta de modelo estadístico predictivo para completar la descripción compleja de los verbos cognitivos", pp. 273-290) un modelo estadístico predictivo para completar la descripción compleja de los verbos cognitivos, fruto de un análisis en profundidad de algunos resultados de su tesis doctoral (Soler, 2019). Se trata de una descripción sistémica que combina el análisis cuantitativo y cualitativo, con una aproximación cognitiva, lo que diferencia este enfoque de trabajos previos. Se toma como paradigma la forma verbal "creo", un verbo cognitivo de gran complejidad por su polisemia y polifuncionalidad semántico-pragmática. El corpus se construye con datos reales, procedentes de discursos parlamentarios y conversaciones coloquiales. A través de la herramienta STATA se realizan regresiones multinomiales para conocer los grados de explicatividad y diferenciar categorías en cuanto a significado y funcionalidad de estas formas complejas. Para el análisis cuantitativo se analiza el corpus recopilado con más de 700 ejemplos cualitativos y variables de análisis previos, añadiendo también datos de corpus adicional con datos de COLAm y CORPES XXI. Los resultados reconocen los tipos de variables que ya estaban presentes en estudios previos, pero además se detectan las diferencias existentes en formas verbales en cuanto a semántica y pragmática de las formas cognitivas, lo que demuestra la eficacia de este método de análisis combinado, extrapolable a otro tipo de textos y géneros.

En el capítulo XIII ("Uso de redes Bayesianas para el análisis de corpus de problemas locales relacionados con los Objetivos de Desarrollo Sostenible", pp. 291-306), Manuel Caro Piñeres y Ernesto Llerena García, de la Universidad de Córdoba en Colombia, detallan una modalidad de análisis que usa redes Bayesianas en el análisis de corpus de problemas locales relacionados con los Objetivos de Desarrollo Sostenible (ODS). El estudio emplea la *Design Science Research Methodology* (DSRM) (Hevner et al., 2007), que aplica tecnologías innovadoras al estudio del conocimiento humano y desarrolla sistemas aplicados al Procesamiento del Lenguaje Natural (PLN). El lexicón del corpus se compone de más de 3000

descripciones de problemas relacionados con las SDG. Estos se organizan por campos semánticos que relacionan hiperónimos e hipónimos de cada elemento, proceso que se realiza con Wordnet, un sistema de léxico digital asociado a teorías psicolingüísticas, de acceso abierto y disponible en línea. De este modo, una palabra se puede asociar a varias categorías. También aplica *Machine Learning* para incorporar un nuevo vocabulario relacionado con las SDG usando Inteligencia Artificial (IA). La clasificación se realiza online mediante el *Naïve Bayes mod*el (Schütze et al., 2008) que usa los datos del corpus para realizar cálculos probabilísticos y evaluar sus categorías. Los resultados de los análisis, una vez procesos los datos recogidos, tienen una gran precisión en cuanto a la correspondencia de los conceptos semánticos. A través de la metodología empleada y el uso de redes Bayesianas se obtiene un corpus amplio de descripción de problemas relacionados con el cumplimiento de las SDG.

El siguiente artículo ("Correlación entre la metáfora orientacional bueno es arriba/ malo es abajo y polaridad positiva/negativa en verbos del español: un estudio con estadística de corpus", pp. 307-323), a cargo de Benjamín López Hidalgo e Irene Renau y Rogelio Nazar, de la Pontificia Universidad Católica de Valparaíso en Chile, trata un nuevo caso de lingüística de corpus que analiza la correlación entre la metáfora orientacional "BUENO ES ARRIBA/ MALO ES ABAJO" y la polaridad positiva/negativa en verbos del español. Se pretende estudiar esta cuestión con suficiente evidencia empírica, ante la escasez de estudios conocidos hasta la fecha, en su mayoría experimentales. Para ello, se han seleccionado diez verbos del español con significado subir/bajar para medir la coocurrencia con un lexicón de polaridad positiva/negativa. Se trabaja con la versión española del corpus EsTenTen (Kilgarriff & Renau, 2013), ya etiquetado morfosintácticamente, y los diccionarios Battaner (2003) y RAE (2014). Igualmente, se recurre a herramientas como Jaguar para la preparación de muestras, al tiempo que con un script se miden las frecuencias de coocurrencia en el corpus con el vocabulario de polaridad. Las concordancias obtenidas se clasifican y se evalúa el grado de polaridad (positiva, negativa o neutra). De este modo, se consiguen resultados empíricos con un método de análisis estadístico de corpus y datos reales. Este método es aplicable a otros lexicones para análisis de expresiones metafóricas y puede ampliarse con otros algoritmos de aprendizaje automático.

El último capítulo ("UnderRL Tagger: un software libre para etiquetar POS en Under-Resourced Languages", pp. 325-344) cierra también el bloque análisis de corpus y está elaborado por José Luis Pemberty Tamayo y Jorge

Mauricio Molina Mejía, de la Universidad de Antioquía. Presenta un modelo de PNL adaptado a la lingüística de corpus a través de UnderRL Tagger, una herramienta de libre acceso para etiquetado semiautomático de POS (part-of-speech) en Under-Resourced Languages (URLa), una serie de lenguas que están en desventaja respecto a las más comunes, al no contar con suficientes recursos computacionales para el procesamiento de textos o corpus que sirvan de base para la construcción de herramientas de análisis. Tampoco tienen suficiente presencia en la web, lo que disuade a los lingüistas a la hora de trabajar con ellas. UnderRL Tagger usa el estándar EAGLES, de modo que los datos se pueden compartir de manera global y trabajar con una gran cantidad de textos, reduciendo tiempo y recursos. En su interfaz de navegación se crean, asignan y editan las etiquetas, que progresivamente van enriqueciendo un diccionario de etiquetas disponible para todos los textos con los que se trabaja, extrapolable a otras lenguas. Los textos se categorizan en base a EAGLES y se almacenan como XML con un identificador único. De este modo, el sistema de UnderRL Tagger se adapta al usuario y se facilita el proceso de etiquetado de las URL, que además está estandarizado y supone una ayuda tecnológica para la investigación y una forma de compartir la información con otros investigadores.

El bloque final de artículos del libro sirve como introducción a los procesos de análisis que se aplican a los corpus descritos en cada capítulo. Se explica con detalle la función de las herramientas escogidas para el etiquetado de textos, softwares de procesamiento de lenguaje humano y cómo éstos se aplican al tratamiento de los datos recogidos en los corpus. Las herramientas y demás recursos computacionales que se emplean son capaces de combinar tipos de análisis, como cualitativo y cuantitativo, miden coocurrencias cruzando dos bancos de datos que demuestren la relación entre variables, o verifican la correspondencia de campos semánticos comunes dentro de un grupo de diferentes discursos.

Nuevamente resulta un acierto la estructura escogida en esta obra en cuanto al contenido de artículos. En estos cinco últimos capítulos el elemento común es la descripción minuciosa de los aspectos computacionales que se aplican en cada estudio y cómo procesa la información cada herramienta de análisis. Son partes importantes y extensas de cada capítulo, lo que dificulta su comprensión por parte del lector si no existe una mínima base previa de conocimiento en este campo. Es, por tanto, adecuado que se estructuren en un único bloque diferenciado del resto y justificado que este bloque sea el último de los tres.

El conjunto del libro supone sin duda un activo para el campo de la investigación de Humanidades. Ofrece diferentes ejemplos de estudios, varios de ellos interdisciplinares, que utilizan herramientas computacionales gracias a las cuales tenemos acceso a volúmenes de datos que hasta ahora, si bien era posible, suponían una inversión de recursos y tiempo inabarcables por parte de los investigadores. Los trabajos que se han escogido hacen énfasis en los beneficios que aporta el trabajo con tecnologías a la hora de suplir la falta de datos en diferentes áreas de estudio dentro de las Humanidades, así como recuperar proyectos hasta ahora prácticamente inasumibles. Este es, precisamente, uno de los elementos más importantes a destacar en la obra, además del hilo conductor de los artículos que en ella se incluyen.

En cuanto a la forma, la estructura del libro es idónea para este tipo de recopilación, al facilitar la lectura tanto al investigador o docente como al lector interesado. El índice permite navegar por el contenido con mucha facilidad. Los artículos también están correctamente estructurados, y cumplen todos los criterios en cuanto a forma, citas, etc. Se han revisado cuidadosamente, de modo que no se encuentran casi erratas. Otro punto a su favor deriva de las lenguas escogidas para la redacción de los textos, español o inglés, lo que facilita enormemente su difusión, al igual que el hecho de redactar la introducción y los resúmenes de cada capítulo en ambos idiomas.

Todo ello hace de Digital Humanities, Corpus and Language Technology un trabajo ineludible en el campo de la investigación y la divulgación dentro de las Humanidades Digitales, pero también para el lector interesado en ampliar su formación teórica en un campo tan novedoso como lleno de posibilidades de estudio.

NAYRA SÁNCHEZ VERA
Universidad Complutense de Madrid
naysanch@ucm.es

Referencias bibliográficas

REFERENCIAS BIBLIOGRÁFICAS

Alvar, M. y Verdejo, M. (1978). Automatización de atlas lingüísticos. *Revista de Dialectología y Tradiciones Populares*, *34*, 23-39.

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer* (6.1.16) [Computer software]. http://www.praat.org/

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173. doi.org/10.1075/ijcl.20.2.01bre

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). *Error Tagging Manual Version 1.2*. Centre for English Corpus Linguistics, Université Catholique de Louvain.

DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to Deception. *Psychological Bulletin*, *129*(1), 74-118.

Dik, S. C. (1997). *The Theory of Functional Grammar* (K. Hengeveld (ed.); 2nd, rev. ed., Issues 20-21). Mouton de Gruyter.

García Mouton, P. (2010). El procesamiento informático de los materiales del *Atlas de la Península Ibérica* de Tomás Navarro Tomás. En G. Aurrekoetxea y J. L. Ormaetxea (Eds.), *Tools for linguistic variation* (pp. 167-174). Universidad del País Vasco/Euskal Herriko Unibertsitatea.

García Mouton, P. (2017). El *Atlas Lingüístico de la Península Ibérica* (*ALPI*) en línea. Geolingüística a la carta. *Estudis romanics*, *39*, 335-343.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9-34). Cambridge University Press.

Granger, S, Gilquin, G, & Meunier, F. (Eds.). (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press. doi.org/10.1017/CBO9781139649414

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and oreign Language Teaching* (pp. 3-33). John Benjamins Publishing Company.

Hevner, A. R. (2007). A three-cycle view of design science research. *Scandinavian journal of information systems,* 19(2), 4.

Kilgarriff, A. & Renau, I. (2013). EsTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, *95*, 12-19

Kövecses, Z. (2002). *Metaphor. A practical introduction*. Oxford University Press.

Kövecses, Z. (2008). Conceptual metaphor theory: some criticisms and alternative proposals. *Annual Review of Cognitive Linguistics*, *6*, 168-184.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lakoff, G. (1993). The contemporary theory of metaphor. En A. Ortony (Ed.), *Metaphor and thought* (2.a ed.) (pp. 202-251). Cambridge University Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh. The embodied mind and its challenge to western thought*. Basic Books.

Leech, G., & Wilson, A. (1996). *EAGLES recommendations for the morphosyntactic annotation of corpora*. Istituto di Linguistica Computazionale http://www.ilc.cnr.it/EAGLES96/annotate/node1.html

Nazar, R., Vivaldi, J. & Cabré, M. T. (2008). A suite to compile and analyze an LSP corpus. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, (pp.1164-1169). European Language Resources Association.

Pemberty Tamayo, J. L. & Molina Mejía, J. M. (2020). UnderRL Tagger: Concepción y elaboración de un sistema de etiquetado semiautomático para Under-Resourced Languages. In J. M. Molina Mejía, P. Valdivia Martin & R. A. Venegas Velásquez (Eds.), *Actas III Congreso Internacional de Lingüística Computacional y de Corpus - CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus - WoPATeC 2020* (pp. 78-81). Universidad de Antioquia.

Pemberty Tamayo, J. L.; Molina Mejía, J. M. & Marín Morales, M. I. (2020). *UnderRL Tagger* (Versión 1.0) [Software]. Corpus Ex Machina, Universidad de Antioquia.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Scott, M. (2005). *WordSmith*. Lexically. http://lexically.net/wordsmith/research/

Soler, M. A. (2019). *Semántica y pragmática de los verbos doxásticos en la interacción oral en espanol. Un estudio monográfico sobre la forma verbal* creo [Tesis doctoral. Universitat de Valencia]. RODERIC. https://roderic.uv.es/handle/10550/71798

Soriano, C. (2012). La metáfora conceptual. En I. Ibarretxe-Antunano & J. Valenzuela (Coords.), *Lingüística cognitiva* (pp. 97-121). Anthropos.

Sousa, X. (2020). Humanidades digitales y geografía lingüística: la edición digital del *Atlas Lingüístico de la Península Ibérica*. En A. Gallego & F. Roca (Eds.), *Dialectología digital*. Anexo de *Verba* (pp.139-158). Universidad de Santiago de Compostela.

Vrij, A. (2018). Deception and truth detection when analyzing nonverbal and verbal cues. *Applied Cognitive Psychology*, *33*(2), 160-167.