Introduction

Jorge Molina Mejía & Andrés Grajales Ramírez
Universidad de Antioquia – Colombia

"*Digital Humanities, Corpus and Language Technology: a look from diverse case studies*" is a title that takes up, in an innovative way, three fields of knowledge that are combined in this research book, which is the result of a joint editing work between the University of Antioquia and the University of Groningen. It is important to note that in the present time and context, it is of utmost importance to elaborate works that have interdisciplinary studies as a north and, in this sense, the work that we present below has the vocation to address current works in these three aspects, always with a view from the computer science and its application in the field of human and social sciences, and all this from an inter-university perspective. We have also decided to present the different chapters of this compendium in Spanish and English, so that they can be consulted by students and researchers who speak both languages. All this is based on the fact that the book we present here has been produced between two institutions in which the most widely used languages are Spanish and English. Nevertheless, from a global perspective, our intention is that the chapters published here will reach a large part of the researchers who use either of these two languages in their research and teaching process.

This book presents several case studies where the relationship between Digital Humanities and Language Technology and its application in linguistic corpora is evident. As previously anticipated, Digital Humanities can contribute to the creation and analysis of linguistic corpora thanks to the use of new technologies and tools that allow greater efficiency and precision in Natural Language Processing. On the other hand, the study of corpora can help to discover patterns and trends in linguistic data that would be difficult to detect using traditional methods, which benefits the Digital Humanities. New technologies and digital tools allow today to complement each other, through greater efficiency and precision in the processing and understanding of human languages. From this moment, it can be glimpsed that the future of these disciplines is highly promising, as they have begun to play an important role in research and studies, and is expected to continue to grow. As the current era advances and new developments emerge, language technologies

become more sophisticated, so there will be new opportunities, but also new challenges in these fields.

Currently, it is common for work related to these topics to be focused on fields such as literature, history, linguistics, sociology, etc. However, it is expected that, in the future, the Digital Humanities and the analysis of linguistic corpora will be able to extend their applications to even more diverse disciplines, such as digital anthropology, computational archaeology, cultural studies or music. This will make it possible to address and investigate a wide range of human phenomena from a digital approach. This is quickly evidenced by the recent advancement of artificial intelligences and machine learning, with which Natural Language Processing and corpus analysis are expected to become even more accurate. This will open new possibilities for linguistic, philological, and other studies, allowing researchers to perform more in-depth analysis, with more subtle pattern detection. Similarly, access to corpora of texts and data is expected to become increasingly easier, as with the rise of digital libraries, data repositories, and information gathering and storage tools, researchers will have access to an ever-increasing number of digital resources to analyze, which will greatly expand research possibilities.

In summary, the future of Digital Humanities, Corpus Studies, and Language Technology, all put together, demonstrates an inevitable expansion of their application in various disciplines, whereby the advancement of natural language processing techniques and access will be ever-increasing. These advances promise an exciting future within these disciplines, giving them a major role in future research, especially in the study of the Humanities in the digital environment. The possibilities and applications of these disciplines are just beginning to be visualized, but there will be more to come and explore. A revolution that is now focused on the "awakening" of AI, but that in the future may be something we did not see coming.

This book is therefore subdivided into three main parts, the first of which is devoted to Digital Humanities and the use of new technologies for different aspects of the human and social sciences. The second part deals with research works related to the compilation, characterization, or construction of linguistic corpora. Finally, the third part explores projects based on corpus analysis and natural language processing. All the chapters presented here have been rigorously evaluated by academic peers, experts in some of the fields of knowledge mentioned here. We will now present each of the parts and their respective chapters.

In the first part of this work, we can find four chapters, which deal with topics about digital humanities such as: visual arts, online libraries, relational databases for the study of classical Greek and Latin, and the use of Python in epistolary analysis.

Chapter I has been co-written by Professors John Roberto and Brian Davis and is entitled "*Understanding Outsider Art in the context of Digital Humanities*". This chapter presents the Outsider Art project, which aims to present a group of very innovative artists who are called "outsiders", who are usually marginalized aesthetically and socially due to their psychiatric condition, as well as homeless people, prison inmates, people with disabilities, migrants, and ethnic minorities. This is how this project arises, which aims to propose an automatic discovery of the semantic limits of outsider art in the context of digital humanities. Methodologically, this proposal is based on three tasks: a) the collection of a corpus of outsider art; b) generate a large dataset of digital images about this type of art; and c) build the first ontology of this art.

Chapter II deals with "*The Virtual Library of Spanish Philology (BVFE) and its Hispanic-American heritage*", and has been co-written by professors Jaime Peña Arce and María Ángeles García Aranda. This work has a double objective: on the one hand, to publicize the Library of Spanish Philology, which is a portal that gathers a large number of linguistic works related to Spanish, which can be accessed freely and free of charge. Secondly, the authors seek to investigate the Hispanic American component of its collection, with the purpose of reflecting on all that has been done and what still remains to be done.

In **Chapter III**, "*From two relational databases to an XML database. The COMREGLA project*", co-written by a group of researchers attached to higher education centers in Spain: Eveling Garzón Fontalvo, Berta González Saavedra, José Ignacio Hidalgo González, Iván López Martín, Alberto Pardal Padín, Guillermo Salas Jiménez and Cristina Tur. In this chapter the authors present a series of modifications and adaptations made on two relational bases of the REGLA project (REction and Complementation in Ancient Greek and Latin) whose emphasis is on verbal predications. It is important to emphasize that the purpose of the changes introduced is to make the information contained in the database compatible with other automatic language processing tools and to provide analyses that go beyond the nuclear and basic predications, that is, towards full texts. In order to enable the respective analyses, the researchers have created a new annotation standard that allows to reflect the richness of morphological, syntactic, semantic and lexical information; all this allows to account for the very recursion of language and to enrich the analysis with labels for linguistic components not studied before.

In **Chapter IV**, Santiago Alejandro Ortiz Hernández proposes the work called "*Analysis of the correspondence of Colonel Anselmo Pineda with Python: a look at the collector project and the territory from social networks and machine learning*". This chapter analyzes the collecting of Colonel Anselmo Pineda during the nineteenth century in Colombia,

based on his voluminous epistolary preserved in the National Library of Colombia. To this end, the author proposes a mixed methodology that combines the traditional close reading and a distant reading carried out from the machine thanks to techniques of data science and geographic information systems implemented thanks to the Python language. This approach has two main objectives: a) to discover the colonel's method of collecting documents by examining the composition of his network of collaborators reconstructed through his personal correspondence, all based on digital humanities and digital history; and b) to explore the spatial scope of this network of collaborators, which should make it possible to evaluate the spatial dimension in the formation of the Pineda library within the civilizing project of the nascent republic in New Granada.

The second part has to do with corpus linguistics, in this sense, six chapters were received, in which important topics such as: linguistic atlas corpora, the study of multimodal corpora applied to the Brazilian oral language, the study of Mexican indigenous languages, lie detection and credibility assessment based on corpora specially designed for this purpose, linguistic corpora that allow the study of colloquial German language, and a corpus of learners of English as a Foreign Language.

Chapter V, entitled "*Development of a corpus of linguistic atlases*", is a proposal by Professor Carolina Julià Luna. In this chapter, the author presents some characteristics and functionalities of this type of computer tools, in which data from various regional linguistic atlases of European Spanish are stored. The purpose of all this is to conserve the linguistic heritage, to serve as a source for the dissemination of variation and richness in the language and, finally, to help complement the data from textual corpora and lexicographic works that help to expand research on linguistic change and the history of the Spanish language.

Chapter VI deals with "*The C-ORAL-BRASIL proposal for the treatment of multimodal data in corpus: the pilot project of the BGEST corpus*", a work proposed by Professors Camila Barros and Heliana Mello. According to the authors, this chapter discusses methodological issues associated with the collection and processing of multimodal data, especially those related to the predominant role of action. The main objective of the chapter is to connect the organization of the structure of information, based on the union of the Theory of Language in Action and the concept of spatial-motor packaging. At the end, the authors will show us the crucial role of prosody in the informational categories of L-AcT and its impact on the interpretation of gestures.

Chapter VII, co-written by Antonio Reyes Pérez and Antonio García Zúñiga, is entitled "*Language technologies and indigenous Mexican languages: constitution of an Amuzgo-Span-*

ish parallel corpus". This proposal describes the particularities of the construction of the first Amuzgo-Spanish parallel corpus, which represents a real source of data for scientific research in the field of language, as well as for the development of resources and tools for languages that are scarcely represented and in danger of disappearing.

Chapter VIII deals with the "*Methodological Bases: the construction of a corpus for the detection of lies and the evaluation of credibility*" and is the work of Pedro Eduardo Hernández Fuentes. In this chapter it is possible to access the meta-analytical approaches that show that verbal information is a reliable indicator that allows to identify lies or to evaluate the credibility of a testimony. For this purpose, the author shows a work based on a linguistic corpus that has been developed thanks to a transdisciplinary perspective between linguistics and psychology.

In **Chapter IX**, "*Türkisch für Anfänger: proposal of a corpus of modern colloquial German, exemplified from routine phrases for greetings*", Karen Baquero Castro builds a specific corpus of German from more than 12,000 lines of dialogue from the German television series Türkisch für Anfänger. The aim of this corpus is to optimize the process and accompaniment in the teaching and learning of German as a foreign language. In order to exemplify its usefulness and use, the corpus focuses on the formulas used in the series, more precisely on the greeting formulas. These are analyzed by the author from a didactic perspective and appealing to the analysis of linguistic corpora that consider the context in order to favor the teaching-learning process by means of authentic texts.

Finally, among these works on corpus construction, we have **chapter X** "*CLEC - Colombian Learner English Corpus: first learner corpus of written production in English online in Colombia*", which deals with the study of Professor M. Victoria Pardo and Professor Antonio Tamayo, both Colombians, on the constitution of a corpus called CLEC. This would be the first corpus on English learners, based on written texts produced by the learners themselves, from Colombia, and accessible through the website of the TNT research group of the University of Antioquia. It is a corpus of more than 200,000 words that is fully labeled to classify the types of errors made by learners, as well as the level of the learner. The chapter shows the criteria used for the collection of CLEC, respecting the guidelines of corpus linguistics and learner corpus. Thus, in this corpus, learners' errors can be consulted, and this phenomenon can be studied by teachers and researchers, who can contribute new texts, as well as by those interested in learning and studying English as a foreign language.

The third and last part also deals with works in the field of corpus linguistics, but from a perspective more related to analysis and its methods, in which computational linguistics

and Natural Language Processing (NLP), as well as statistical analysis, are often used. This section is made up of five chapters.

Thus, **Chapter XI**, entitled "*Pronunciation of consonant clusters in Spanish speakers based on the Czech read speech corpora*", and written by Czech researchers Kateřina Pugachova and Jitka Veroňková, presents a study that aims to determine which Czech consonant clusters are difficult to pronounce for Spanish speakers and which are the most frequent sound changes due to differences in syllable structure between these two languages. A set of 26 consonant clusters in initial, middle, and final positions of words was selected. Seventy-five words containing the target consonant clusters were included in a coherent text written in Czech (of 838 words). The study provides useful information for improving the teaching of Czech to native speakers of Spanish.

Continuing with the analyses on specific corpora, in **Chapter XII**, "*Relating qualitative and quantitative analysis. A predictive statistical model proposal to complete the complex description of cognitive verbs*", M. Amparo Soler Bonafont (Spain) presents a proposal for a predictive statistical model to complete the complex description of cognitive verbs, specifically performative forms. The model designed allows us to recognize, with a high degree of explanatory power, the meanings, and pragmatic functions of polysemous and polyfunctional units such as "creo". Moreover, the model can be replicated in other texts and genres in which similar epistemic units may appear.

In **Chapter XIII**, "*Use of Bayesian networks for the analysis of corpus of local problems related to the Sustainable Development Goals*", Caro Piñeres and Moreno García, from the University of Córdoba (Colombia), present a sentiment analysis study based on Bayesian networks in a corpus related to social problem solving. It exemplifies the use of Bayesian networks for data analysis, modeling, and decision support in various domains. The need for techniques and tools that automatically construct Bayesian networks from massive text or bibliographic data is discussed, especially in relation to the United Nations-led Sustainable Development Goals (SDGs). The paper also discusses the collection and analysis of textual information to build Bayesian networks, as well as the limitations and challenges associated with this technique. The objective is to describe the process of collecting, organizing, annotating, and validating a corpus of more than 3,000 descriptions of problems related to SDG compliance in three regions of Colombia. The main outcome of the study was the creation of a large digital corpus of descriptions of problems related to SDG compliance in these three regions. In addition, the potential of the corpus was evaluated through the application of a Bayesian network algorithm, which produced a high rate of correct answers.

Chapter XIV welcomes us to the study on the correlation between the orientational metaphor BUENO ES ARRIBA / MALO ES ABAJO and positive/negative polarity in Spanish verbs. This study, entitled "*Correlation between the orientational metaphor GOOD IS UP / BAD IS DOWN and positive/negative polarity in Spanish verbs: a study with corpus statistics*" and conducted by colleagues from the Pontificia Universidad Católica de Valparaíso (Chile), seeks to test the relationship between vertical orientation and polarity in Spanish orientational metaphors. Ten Spanish verbs with 'up'/'down' meaning were selected and their association was measured in corpus concordances with lexical units with 'positive'/'negative' meaning, labeled by means of a polarity lexicon. The results of the study indicate that there is a relationship between vertical orientation and positive or negative polarity in real contexts of use of the units of analysis. This makes it possible to test empirically and by means of corpus statistical methods the orientational metaphor on a linguistic level. With this it can be stated, with a high degree of certainty, that verbs with a sense of 'up' will tend to be part of sentences in which a 'positive' sense will be expressed, and verbs with a sense of 'down' will tend to be included in sentences with a 'negative' sense.

Finally, a different and innovative study in the field of language processing is the work of José Luis Pemberty, accompanied and advised by J. Molina Mejía, editor of this volume. This **Chapter XV**, "*UnderRL Tagger: a free software for Under-Resourced Languages POS tagging*", presents a free software that allows morphologically annotating (POS) under-resourced languages (Under-Resourced Languages). With this model, the process can be performed manually, but the algorithm can also be trained to gradually automate it. The output format uses the EAGLES tags in XML, with the intention of making it possible to process big data. This would provide a valuable computing resource for languages with few native speakers or poorly studied languages.

Introducción

Jorge Molina Mejía & Andrés Grajales Ramírez
Universidad de Antioquia – Colombia

"*Humanidades Digitales, Corpus y Tecnología del Lenguaje: una mirada desde diversos casos de estudio*" es un título que retoma, de una manera innovadora, tres campos del conocimiento que se conjugan en el presente libro de investigación, el cual es fruto de un trabajo conjunto de edición entre la Universidad de Antioquia y la Universidad de Groningen. Es importante constatar que en la época y el contexto actuales resulta de suma importancia elaborar obras que tengan como norte los estudios interdisciplinarios y, en este sentido, la obra que presentamos a continuación tiene por vocación abordar trabajos actuales en estos tres aspectos, siempre con una mirada desde la informática y de su aplicación en el campo de las ciencias humanas y sociales, y todo ello desde una perspectiva interuniversitaria. Hemos decidido, además, que los diferentes capítulos que hacen parte del presente compendio se presenten en español y en inglés, esto con el fin de que puedan ser consultados por estudiantes e investigadores hablantes de ambas lenguas. Todo esto se fundamenta en el hecho de que el libro que aquí presentamos se ha realizado entre dos instituciones en las que las lenguas de mayor uso son el español y el inglés. No obstante, desde una perspectiva global, nuestra pretensión es que los capítulos aquí publicados lleguen a una gran parte de los investigadores que emplean alguna de estas dos lenguas en su proceso investigativo y de docencia.

El libro presenta diversos casos de estudio donde la relación de las Humanidades Digitales con la Tecnología del Lenguaje y su aplicación en corpus lingüísticos es evidente. Como se anticipó anteriormente, las Humanidades Digitales pueden aportar en la creación y análisis de corpus lingüísticos gracias a la utilización de nuevas tecnologías y herramientas que permiten una mayor eficiencia y precisión en el Procesamiento del Lenguaje Natural. Por otro lado, el estudio de corpus puede ayudar a descubrir patrones y tendencias en los datos lingüísticos que serían difíciles de detectar mediante métodos tradicionales, lo cual beneficia a las Humanidades Digitales. Las nuevas tecnologías y herramientas digitales permiten hoy en día complementarse, mediante mayor eficiencia y precisión en el tratamiento y comprensión de los lenguajes humano. Desde este instante, se puede vislum-

brar que el futuro de estas disciplinas es altamente prometedor, pues han empezado a desempeñar un papel importante en las investigaciones y los estudios, y se espera que siga creciendo. A medida que se avanza y surgen nuevos desarrollos en la era actual, las tecnologías del lenguaje se tornan más sofisticadas, por lo cual habrá nuevas oportunidades, pero también nuevos desafíos en estos campos.

Actualmente, es común que los trabajos relacionados con estas temáticas se centren en campos como la literatura, la historia, la lingüística, la sociología, etc. Sin embargo, se espera que, en el futuro, las Humanidades Digitales y el análisis de corpus lingüísticos puedan ampliar sus aplicaciones en disciplinas aún más diversas, tales como la antropología digital, la arqueología computacional, los estudios culturales o la música. Lo cual va a permitir abordar e investigar una amplia gama de fenómenos humanos desde un enfoque digital. Esto rápidamente se evidencia en el reciente avance de las inteligencias artificiales y el aprendizaje automático, con lo que se espera que el Procesamiento del Lenguaje Natural y el análisis de corpus se vuelvan aún más precisos. Esto abrirá nuevas posibilidades para los estudios lingüísticos, filológicos y demás, permitiendo que los investigadores realicen análisis a más profundidad, con detección de patrones más sutiles. De igual manera, se espera que el acceso a corpus de textos y datos sea cada vez más fácil, pues con el incremento de las bibliotecas digitales, los repositorios de datos y las herramientas de recolección y almacenamiento de información, los investigadores tendrán acceso a una cantidad cada vez mayor de recursos digitales para analizar, lo cual ampliará enormemente las posibilidades de investigación.

En resumen, el futuro de las Humanidades Digitales, el estudio de Corpus y la Tecnología del lenguaje, todo puesto en relación, demuestra una inevitable expansión de su aplicación en diversas disciplinas, por lo que el avance de las técnicas de procesamiento del lenguaje natural y el acceso será cada vez mayor. Estos avances prometen un futuro emocionante dentro de estas disciplinas, otorgándoles un papel principal en las investigaciones venideras, sobre todo, en cuanto al estudio de las Humanidades en el entorno digital. Las posibilidades y aplicaciones de estas disciplinas apenas se empiezan a visualizar, pero habrá más por llegar y explorar. Una revolución que ahora tiene puesto el foco en el "despertar" de las IA, pero que en el futuro puede tratarse de algo que no veníamos venir.

El presente libro se encuentra subdividido, por lo tanto, en tres grandes partes, la primera dedicada al tema de las humanidades digitales y la utilización de las nuevas tecnologías para diferentes aspectos de las ciencias humanas y sociales. En la segunda parte, se abordan trabajos de investigación que tienen que ver con la compilación, caracterización o construcción de corpus lingüísticos. Finalmente, la tercera propende por explorar pro-

yectos que tienen como punto de apoyo el análisis de corpus y el procesamiento del lenguaje natural. Todos los capítulos aquí presentados, han sido rigurosamente evaluados por pares académicos, expertos en alguno de los campos de conocimiento aquí mencionados. Pasaremos, a continuación, a presentar cada una de las partes y sus respectivos capítulos.

En la primera parte de la presente obra podemos encontrar cuatro capítulos, los cuales versan sobre temas acerca de las humanidades digitales tales como: las artes visuales, las bibliotecas en línea, las bases de datos relacionales para el estudio del griego y el latín clásicos, y el empleo de Python en el análisis epistolario.

El capítulo I ha sido coescrito por los profesores John Roberto y Brian Davis, y lleva por título "*Entender el Arte Outsider en el contexto de las Humanidades Digitales*". En este capítulo se presenta el proyecto de Arte *Outsider*, el cual tiene como objetivo presentar a un grupo de artistas muy innovadores que son los denominados "outsiders", los cuales normalmente se encuentran marginados a nivel estético y social debido a su condición psiquiátrica, también de ser personas sin hogar, reclusos carcelarios, personas con discapacidad, migrantes y minorías étnicas. Es así como surge este proyecto que tiene como finalidad proponer un descubrimiento automático de los límites semánticos del arte *outsider* en el contexto de las humanidades digitales. Metodológicamente, esta propuesta se fundamenta en tres tareas: a) la recopilación de un corpus de arte *outsider*; b) generar un gran conjunto de datos de imágenes digitales sobre este tipo de arte; y c) construir la primera ontología de este arte.

El capítulo II versa sobre "*La Biblioteca Virtual de la Filología Española (BVFE) y su acervo hispanoamericano*", y ha sido coescrito por los profesores Jaime Peña Arce y María Ángeles García Aranda. En este trabajo parte de un doble objetivo, por un lado, dar a conocer la Biblioteca de la Filología Española, la cual se constituye como un portal que recoge una gran cantidad de obras lingüísticas relacionadas con el español, a las que se puede acceder de forma libre y gratuita. En segundo lugar, los autores buscan indagar en el componente hispanoamericano de su acervo, con el propósito de recapacitar sobre todo aquello que se ha hecho y lo que aún queda por hacerse.

En **el capítulo III**, "*De dos bases de datos relacionales a una base de datos XML. El proyecto COMREGLA*", coescrito por un grupo de investigadores adscritos a centros de educación superior de España: Eveling Garzón Fontalvo, Berta González Saavedra, José Ignacio Hidalgo González, Iván López Martín, Alberto Pardal Padín, Guillermo Salas Jiménez y Cristina Tur. En este capítulo los autores presentan una serie de modificaciones y adaptaciones efectuadas sobre dos bases relacionales del proyecto REGLA (REcción y complementación en Griego Antiguo y Latín) cuyo énfasis se encuentra en las predicaciones

verbales. Resulta importante destacar que la finalidad de los cambios introducidos se enmarcan en el proyecto COMREGLA conduce a que la información contenida dentro de la base de datos sea compatible con otras herramientas de tratamiento automático del lenguaje y que provea análisis que vayan más allá de las predicaciones nucleares y básicas, es decir, hacia las de textos completos. Con el fin de permitir los respectivos análisis, los investigadores han creado un nuevo estándar de anotación que permite reflejar la riqueza de la información morfológica, sintáctica, semántica y léxica; todo ello permite dar cuenta de la propia recursividad del lenguaje y enriquecer el análisis con etiquetas para componentes lingüísticos no antes estudiados.

En **el capítulo IV**, el profesor Santiago Alejandro Ortiz Hernández propone el trabajo denominado "*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*". En dicho capítulo se analiza el coleccionismo del coronel Anselmo Pineda durante el siglo XIX en Colombia, a partir de su voluminoso epistolario conservado en la Biblioteca Nacional de Colombia. Para tal fin, el autor propone una metodología mixta que combina la tradicional lectura cercana y una lectura distante efectuada a partir de la máquina gracias a técnicas propias de la ciencia de datos y los sistemas de información geográfica implementados gracias al lenguaje Python. Esta manera de proceder busca dos grandes objetivos: a) poder descubrir el método de recopilación de documentos del coronel al examinar la composición de su red de colaboradores reconstruida mediante su correspondencia personal, todo ello basado en las humanidades digitales y la historia digital; y b) explorar el alcance espacial de esa red de colaboradores, lo que debería posibilitar la evaluación de la dimensión espacial en la conformación de la biblioteca Pineda al interior del proyecto civilizatorio de la naciente república en Nueva Granada.

La segunda parte tiene que ver con la lingüística de corpus, en este sentido se recibieron seis capítulos, en los cuales se abordan temas tan importantes como: los corpus de atlas lingüísticos, el estudio de corpus multimodales aplicados a la lengua oral brasileña, el estudio de lenguas indígenas mexicanas, la detección de mentiras y la evaluación de la credibilidad a partir de corpus especialmente diseñados para tal fin, corpus lingüísticos que permiten el estudio del alemán coloquial, y un corpus de aprendices de inglés como lengua extranjera.

El capítulo V, que lleva por título "*Desarrollo de un corpus de atlas lingüísticos*", es una propuesta de la profesora Carolina Julià Luna. En este capítulo, su autora presenta algunas características y funcionalidades de este tipo de herramientas informáticas, en la que se almacenan datos provenientes de diversos atlas lingüísticos regionales del español europeo.

Todo ello, tiene como finalidad que se pueda conservar el patrimonio lingüístico, que puedan servir como fuente de divulgación de la variación y la riqueza en el lenguaje y, finalmente, que ayuden a complementar los datos procedentes de corpus textuales y de obras lexicográficas que ayuden a ampliar las investigaciones sobre el cambio lingüístico y la historia de la lengua española.

En **el capítulo VI** se aborda "*La propuesta del C-ORAL-BRASIL para el tratamiento de datos multimodales en corpus: el proyecto piloto del corpus BGEST*", un trabajo propuesto por las Profesoras Camila Barros y Heliana Mello. Según las autoras, en este capítulo se discuten cuestiones metodológicas asociadas a la recopilación y al tratamiento de datos multimodales, especialmente a aquellos ligados al papel preponderante de la acción. El objetivo principal del mismo es el de conectar la organización de la estructura de la información, a partir de la unión de la Teoría de la lengua en Acto y el concepto de empaquetado espacio-motor. Al final, las autoras nos mostrarán el papel crucial que adquiere la prosodia en las categorías informacionales de la L-AcT y su impacto en la interpretación de los gestos.

El capítulo VII, coescrito por Antonio Reyes Pérez y Antonio García Zúñiga, lleva por título "*Las tecnologías del lenguaje y las lenguas indígenas mexicanas: constitución de un corpus paralelo amuzgo-español*". En esta propuesta se describen las particularidades de la construcción del primer corpus paralelo amuzgo-español, el cual representa una fuente de datos reales para la investigación científica en el campo del lenguaje, particularmente, así como en lo que respecta al desarrollo de recursos y de herramientas para lenguas escasamente representadas y en peligro de desaparición.

El capítulo VIII tiene que ver con las "*Bases metodológicas: la construcción de un corpus para la detección de mentiras y la evaluación de la credibilidad*", y es obra de Pedro Eduardo Hernández Fuentes. En este capítulo es posible acceder a los acercamientos metaanalíticos que muestran que la información verbal es un indicador confiable que permite identificar mentiras o evaluar la credibilidad de un testimonio. Para ello, el autor muestra un trabajo fundamentado en un corpus lingüístico que ha sido desarrollado gracias a una perspectiva transdisciplinaria entre lingüística y psicología.

En el **capítulo IX**, "*Türkisch für Anfänger: propuesta de un corpus del alemán coloquial actual, ejemplificado a partir de las fórmulas rutinarias de saludo*", Karen Baquero Castro construye un corpus específico de alemán a partir de más de 12 000 líneas de diálogo de la serie de televisión alemana *Türkisch für Anfänger*. El objetivo de este corpus es optimizar el proceso y el acompañamiento en la enseñanza y aprendizaje del alemán como lengua extranjera. Se centra entonces, para ejemplificar su utilidad y uso, en las fórmulas de tra-

tamiento allí presentes, más precisamente en las fórmulas de saludo. Estas son analizadas por la autora desde una perspectiva didáctica y apelando al análisis de corpus lingüísticos que tengan en cuenta el contexto para favorecer la enseñanza-aprendizaje por medio de textos auténticos.

Tenemos, por último, dentro de estos trabajos sobre construcción de corpus, **el capítulo X** "*CLEC - Corpus Colombiano de Aprendices de Inglés: primer corpus de producción escrita de aprendices de inglés en Colombia disponible en línea*", en el cual se aborda el estudio de la profesora M. Victoria Pardo y el profesor Antonio Tamayo, ambos colombianos, sobre la constitución de un corpus llamado *CLEC*. Este consistiría en el primer corpus sobre aprendientes de inglés, el cual se basa en textos escritos producidos por los mismos aprendientes, provenientes de Colombia, y accesible por medio de la web del grupo de investigación TNT de la Universidad de Antioquia. Es un corpus de más de 200 000 palabras que se encuentra totalmente etiquetado para clasificar los tipos de errores que cometen los aprendientes, así como también el nivel del estudiante. El capítulo muestra los criterios que se utilizaron para la recolección de *CLEC*, respetando las pautas de la lingüística de corpus y de corpus de aprendientes. Es así como en este corpus se pueden consultar los errores de los aprendientes y estudiar este fenómeno tanto profesores e investigadores, que pueden aportar textos nuevos, como interesados en aprender y estudiar el idioma inglés como lengua extranjera.

La tercera y última parte aborda también trabajos en el campo de la lingüística de corpus, pero desde una perspectiva más relacionada con el análisis y sus métodos, en el que a menudo se valen de la lingüística computacional y el procesamiento del lenguaje natural (PLN), como también del análisis estadístico. Esta sección se encuentra constituida por cinco capítulos.

De esta manera, **el capítulo XI**, titulado "*La pronunciación de los grupos de consonantes en hispanohablantes basándose en el corpus oral leído checo*", y escrito por los investigadores checos Kateřina Pugachova y Jitka Veroňková, presenta un estudio que tiene como objetivo determinar qué grupos de consonantes del checo son difíciles de pronunciar para los hablantes de español y cuáles son los cambios de sonido más frecuentes debido a las diferencias en la estructura silábica entre estos dos idiomas. Se seleccionó un conjunto de 26 grupos de consonantes en posiciones iniciales, medias y finales de palabras. Se incluyeron 75 palabras que contenían los grupos de consonantes objetivo en un texto coherente escrito en checo (de 838 palabras). El estudio proporciona información útil para mejorar la enseñanza del checo a los hablantes nativos de español.

Continuando con los análisis en corpus específicos, 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*", M. Amparo Soler Bonafont (España) nos presenta una propuesta de modelo estadístico predictivo para completar la descripción compleja de los verbos cognitivos, específicamente las formas performativas. El modelo diseñado permite reconocer con un elevado grado de explicatividad ante qué significados y funciones pragmáticas de unidades polisémicas y polifuncionales como "creo" nos encontramos. Además, el modelo es replicable en otros textos y géneros en los que pueden aparecer unidades epistémicas similares.

En **el capítulo XIII**, "*Uso de redes Bayesianas para el análisis de corpus de problemas locales relacionados con los Objetivos de Desarrollo Sostenible*", Caro Piñeres y Moreno García, de la Universidad de Córdoba (Colombia), presentan un estudio de análisis de sentimiento basado en redes bayesianas en un corpus relacionado con resolución de problemas sociales. Este ejemplifica el uso de redes bayesianas para el análisis de datos, modelado y apoyo a la toma de decisiones en varios dominios. Se discute la necesidad de técnicas y herramientas que construyan automáticamente redes bayesianas a partir de textos masivos o datos bibliográficos, especialmente en relación con los Objetivos de Desarrollo Sostenible (ODS) liderados por las Naciones Unidas. El documento también aborda la recopilación y análisis de información textual para construir redes bayesianas, así como las limitaciones y desafíos asociados con esta técnica. El objetivo es describir el proceso de recopilación, organización, etiquetado y validación de un corpus de más de 3 000 descripciones de problemas relacionados con el cumplimiento de los ODS en tres regiones de Colombia. El resultado principal del estudio fue la creación de un gran corpus digital de descripciones de problemas relacionados con el cumplimiento de los ODS en estas tres regiones. Además, se evaluó el potencial del corpus mediante la aplicación de un algoritmo de red bayesiana, que produjo una alta tasa de respuestas correctas.

El capítulo XIV nos da la bienvenida al estudio sobre la correlación entre la metáfora orientacional BUENO ES ARRIBA / MALO ES ABAJO y la polaridad positiva/negativa en verbos del español. Este estudio, titulado "*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*" y realizado por los colegas de la Pontificia Universidad Católica de Valparaíso (Chile), busca comprobar la relación entre la orientación vertical y la polaridad en las metáforas orientacionales del español. Se seleccionaron 10 verbos del español con significado 'subir'/ 'bajar' y se midió su asociación en las concordancias del corpus con unidades léxicas con significado 'positivo'/'negativo', etiquetadas mediante un lexicón de

polaridad. Los resultados del estudio indican que existe una relación entre la orientación vertical y la polaridad positiva o negativa en contextos reales de uso de las unidades de análisis. Esto permite comprobar empíricamente y mediante métodos de estadística de corpus la metáfora orientacional en un nivel lingüístico. Con ello se puede afirmar, con un grado elevado de certeza, que los verbos que presenten un sentido de 'subir' tenderán a formar parte de frases en las que se expresará un sentido 'positivo', y los verbos con sentido 'bajar' tenderán a estar incluidos en frases con sentido 'negativo'.

Por último, un estudio diferente e innovador en el ámbito del tratamiento del lenguaje es el trabajo de José Luis Pemberty, acompañado y asesorado por J. Molina Mejía, editor de este volumen. **Este capítulo XV**, "*UnderRL Tagger: un software libre para etiquetar POS en Under-Resourced Languages*", se presenta un software libre que permite anotar morfológicamente (POS) lenguas de pocos recursos (Under-Resourced Languages). Con este modelo se puede realizar de manera manual el proceso, pero, además entrenar el algoritmo para paulatinamente ir automatizándolo. El formato de salida utiliza las etiquetas EAGLES en XML, con la intención de que sea posible el tratamiento de grandes datos. De este modo, se les aportaría un valioso recurso informático a lenguas de pocos hablantes nativos o lenguas poco estudiadas.