Digital Humanities, Corpus and Language Technology

Humanidades Digitales, Corpus y Tecnología del Lenguaje

Editors
Andrés Grajales Ramírez
Jorge Molina Mejía
Pablo Valdivia Martin

UNIVERSIDAD DE ANTIOQUIA
Facultad de Comunicaciones y Filología

University of Groningen Press

Digital Humanities, Corpus and Language Technology
Humanidades Digitales, Corpus y Tecnología del Lenguaje

Digital Humanities, Corpus and Language Technology

A look from diverse case studies

Humanidades Digitales, Corpus y Tecnología del Lenguaje

Una mirada desde diversos casos de estudio

Editors
Andrés Grajales Ramírez
Jorge Molina Mejía
Pablo Valdivia Martin

UNIVERSIDAD DE ANTIOQUIA
Facultad de Comunicaciones y Filología

University of Groningen Press

Published by University of Groningen Press
Broerstraat 4
9712 CP Groningen
The Netherlands

In co-edition with Facultad de Comunicaciones y Filología, Universidad de Antioquia (Colombia)

This book has been published open access thanks to the financial support of the Open Access Book Fund of the University of Groningen.

Additionally, we are grateful for the financial support of OSL (The Netherlands Research School for Literary Studies).

Cover design: Bas Ekkers
Typesetting: LINE UP boek en media bv | Mirjam Kroondijk

ISBN (print) 9789403430232
ISBN (ePDF) 9789403430249
DOI https://doi.org/10.21827/6458c72616bed

International Scientific-Editorial Committee

To the team that oversaw the academic and scientific evaluation of the chapters that make up this book: Thank you very much for your effort, willingness, and knowledge.

Comité Científico-Editorial Internacional

Al equipo que se encargó de evaluar académica y científicamente los capítulos que componen este libro: Muchas gracias por su esfuerzo, disposición y conocimientos.

Dra. Lirian Astrid Ciro. *Universidad del Valle, Colombia.*
Dr. Carlos A. Mayora Pernía. *Universidad del Valle, Colombia.*
Dra. Irina Kostina. *Universidad del Valle, Colombia.*
Dr. Jorge Mauricio Molina Mejía. *Universidad de Antioquia, Colombia.*
Dra. Ana María Agudelo Ochoa. *Universidad de Antioquia, Colombia.*
Dr. Ricardo Cedeño Montaña. *Universidad de Antioquia, Colombia.*
Dr. Juan David Martínez Hincapié. *Universidad de Antioquia, Colombia.*
Mg. María Isabel Marín Morales. *Universidad de Antioquia, Colombia.*
Mg. Laura M. Quintero Montoya. *Universidad de Antioquia, Colombia.*
Mg. Juan E. Hincapié Atehortúa. *Universidad de Antioquia, Colombia.*
Dr. George E. Dueñas Luna. *Universidad Nacional, Colombia.*
Dr. Fabio A. González Osorio. *Universidad Nacional, Colombia.*
Dr. Jhon Williams Montoya Garay. *Universidad Nacional, Colombia.*
Dra. Bell Manrique Losada. *Universidad de Medellín, Colombia.*
Dr. Andrés Lombana Bermúdez. *Pontificia Universidad Javeriana, Colombia.*
Dr. Sergio Jiménez Vargas. *Instituto Caro y Cuervo, Colombia.*
Dr. Pablo Valdivia Martin. *University of Groningen, Países Bajos.*
Mg. Juan Albá Durán. *University of Groningen, Países Bajos.*
Dr. René A. Venegas Velasquez. *Pontificia Universidad Católica de Valparaíso, Chile.*
Dr. Ricardo Martínez-Gamboa. *Universidad Diego Portales, Chile.*
Dr. Fernando M. Carranza. *Universidad de Buenos Aires, Argentina.*
Dr. César Antonio Aguilar. *Instituto de Investigaciones en Educación de la Universidad Veracruzana, México.*
Dr. Miguel Fuster Márquez. *Universitat de València, España.*
Dr. Diego A. Burgos Herrera. *Wake Forest University, Estados Unidos de América.*
Dra. Emmanuelle Esperança-Rodier. *Université Grenoble Alpes, Francia.*
Mg. Norman D. Gómez Hernández. *Johannes Gutenberg-Universität Mainz, Alemania*.

Series:

Data Science, Culture & Social Change

This collection is a joint editorial effort between the research groups Data Science, Culture and Social Change of the University of Groningen and the research incubator group Corpus ex Machina of the Universidad de Antioquia. The relationship between these universities has grown stronger in recent years and this collection aims to continue the production of knowledge from a modern, interdisciplinary and multicultural perspective. The 'Data Science, Culture and Social Change' series will provide a collaborative space for an international network working within and across different fields (digital humanities, educational innovation, cultural analytics, computational and corpus linguistics, discourse analysis, political science, computer science, etc.).

Table of Contents

Preface *11*

Introduction *15*

Introducción *23*

Part I Digital Humanities *31*

Chapter I

Understanding Outsider Art in the context of Digital Humanities *33*

Entender el Arte Outsider en el contexto de las Humanidades Digitales

— John Roberto & Brian Davis

Chapter II

La *Biblioteca Virtual de la Filología Española* (*BVFE*) y su acervo hispanoamericano *55*

The Biblioteca Virtual de la Filología Española (BVFE) and its Hispanic American heritage

— Jaime Peña Arce & M. Ángeles García Aranda

Chapter III

De dos bases de datos relacionales a una base de datos XML. El proyecto COMREGLA *73*

From two relational databases to an XML one. Project COMREGLA

— Eveling Garzón Fontalvo, Berta González Saavedra, José Ignacio Hidalgo González, Iván López Martín, Alberto Pardal Padín, Guillermo Salas Jiménez & Cristina Tur

Chapter IV

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 *91*

Analysis of Colonel Anselmo Pineda's epistolary with Python: a glance to the collecting project from the study of the territory and social networks

— Santiago Alejandro Ortiz Hernández

Part II Corpus construction *121*

Chapter V
Desarrollo de un corpus de atlas lingüísticos *123*
Development of a corpus of linguistic atlases
— Carolina Julià Luna

Chapter VI
The C-ORAL-BRASIL proposal for the treatment of multimodal corpora data: the BGEST corpus pilot project *143*
La propuesta del C-ORAL-BRASIL para el tratamiento de datos multimodales en corpus: el proyecto piloto del corpus BGEST
— Camila Barros & Heliana Mello

Chapter VII
Las tecnologías del lenguaje y las lenguas indígenas mexicanas: constitución de un corpus paralelo amuzgo-español *163*
Human language technology and the indigenous languages in Mexico: the Amuzgo-Spanish parallel corpus
— Antonio Reyes Pérez & H. Antonio García Zúñiga

Chapter VIII
Methodological bases: the construction of a corpus for the detection of deception and credibility assessment *185*
Bases metodológicas: la construcción de un corpus para la detección de mentiras y la evaluación de la credibilidad
— Pedro Eduardo Hernández Fuentes

Chapter IX
Türkisch für Anfänger: propuesta de un corpus del alemán coloquial actual, ejemplificado a partir de las fórmulas rutinarias de saludo *201*
Türkisch für Anfänger: proposal of a corpus of modern colloquial German, exemplified from routine phrases for greetings
— Karen Lorena Baquero Castro

Chapter X
CLEC - Colombian Learner English Corpus: first learner corpus of written production in English online in Colombia *217*

CLEC - Corpus Colombiano de Aprendices de Inglés: primer corpus de producción escrita de aprendices de inglés en Colombia disponible en línea

— María Victoria Pardo Rodríguez & Antonio Jesús Tamayo Herrera

Part III Corpus analysis and Natural Language Processing *245*

Chapter XI
Pronunciation of consonant clusters in Spanish speakers based on the Czech read speech corpora *247*

La pronunciación de los grupos de consonantes en hispanohablantes basándose en el corpus oral leído checo

— Kateřina Pugachova & Jitka Veroňková

Chapter XII
Relacionando los análisis cualitativo y cuantitativo. Una propuesta de modelo estadístico predictivo para completar la descripción compleja de los verbos cognitivos *273*

Relating qualitative and quantitative analysis. A predictive statistical model proposal to complete the complex description of cognitive verbs

— M. Amparo Soler Bonafont

Chapter XIII
Use of Bayesian networks for the analysis of corpus of local problems related to the Sustainable Development Goals *291*

Uso de redes Bayesianas para el análisis de corpus de problemas locales relacionados con los Objetivos de Desarrollo Sostenible

— Manuel Caro Piñeres & Ernesto Llerena García

Chapter XIV

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 *307*

Correlation between the orientational metaphor GOOD IS UP / BAD IS DOWN and positive/negative polarity in Spanish verbs: a study with corpus statistics

— Benjamín López Hidalgo, Irene Renau & Rogelio Nazar

Chapter XV

UnderRL Tagger: a free software for Under-Resourced Languages POS tagging *325*

UnderRL Tagger: un software libre para etiquetar POS en Under-Resourced Languages

— José Luis Pemberty Tamayo & Jorge Mauricio Molina Mejía

Preface

Pablo Valdivia Martin
University of Groningen – Netherlands

When discussing with colleagues and students about the change in the paradigm that we are witnessing in the Humanities, we often find it challenging to define the fundamental elements of our discussion. In this regard, it is more important than ever to find common ground and a baseline for starting the dialogue in the Humanities from wherever we, terminologically, are. One of the goals of this book is to provide a shared territory where it will be easier to move, get inspired, and move forward together. Therefore, we must ask ourselves critical questions and offer tentative working frameworks. Despite commonly and regularly using the term Digital Humanities, it sometimes seems difficult to agree on what we call Digital Humanities. Thus, under the context of this volume, I suggest a working definition of Digital Humanities as an interdisciplinary field that applies computational methods and tools to study human culture and society. It encompasses various disciplines, such as literature, history, art, music, linguistics, philosophy, and more. Digital Humanities aims to enhance our understanding of human expression and experience through analyzing, visualizing, and preserving digital data.

Additionally, when I refer to the term Corpus Studies, also crucial in this book, I opt for a broad definition encompassing a large and structured collection of texts or other forms of data that are representative of a language or a domain. Corpus Studies is essential for Digital Humanities because it provides the raw material for various types of analysis, such as text mining, sentiment analysis, topic modeling, stylometry, and more. Corpus Studies can also help us discover new patterns, trends, and insights not readily observable in individual texts or sources.

Furthermore, Language Technologies, another notion pillared in this volume, are understood in the context of these pages as a branch of artificial intelligence that deals with the processing and generation of natural language. Language Technologies enable us to interact with computers using natural languages, such as speech recognition, machine translation, and chatbots. Language Technologies also facilitate analyzing natural language data, such as natural language understanding, generation, information extraction, summa-

rization, and many more, which are well assessed and reflected in the pages of the present volume.

This book presents examples and applications of how these scientific areas can enrich our knowledge and appreciation of human culture and society. Moreover, this book will inspire new generations of scholars to explore the possibilities and challenges of Digital Humanities in their research and teaching practices.

Therefore, the research present in the chapters of this volume contributes to exploring new avenues regarding the cross-/inter-/multi-disciplinary intersections between the Digital Humanities, Computational Cultural and Literary Studies, and Computational Linguistics. From its very conception, this book results from a joint effort between the University of Antioquia and the University of Groningen and a firm belief in the cross-cutting domain nature of cultural and literary studies and how interdisciplinary approaches to everyday challenges, as recently brought up to the light by the UNESCO "Knowledge Driven Actions (2022), it an essential toolkit for the engineering of our future.

Every chapter has been rigorously evaluated by academic peers who are experts in one of the varied fields of knowledge in this volume. This book will be a valuable resource for researchers, students, and anyone interested in the broadly so-called "digital turn" and the Humanities. I thank the authors who contributed to this book and the academic peers who reviewed their work. I would also like to thank our colleagues at the University of Antioquia and the University of Groningen for their support in bringing this project to fruition. Digital Humanities, Corpus, and Language Technologies are rapidly growing fields that have the potential to revolutionize research across various disciplines. New technologies have opened up new perspectives for research, allowing scientists to analyze data in previously impossible ways.

The first part of this book is devoted to Digital Humanities. This section includes chapters on digital storytelling, data visualization, and text mining. These contributions demonstrate how Digital Humanities can enhance research in various fields, from literature to history to anthropology. For example, one chapter discusses how digital storytelling can be used to teach history. The authors argue that students can better understand historical events and their significance using multimedia elements such as images, videos, and audio recordings. Another chapter discusses how data visualization can be used to analyze literary texts. The authors demonstrate how visualizing patterns in language use can reveal insights into literary style and authorship.

The second part of this book focuses on linguistic corpora construction. A corpus is a collection of texts for linguistic analysis. Corpus-based research has become increasingly

popular in linguistics because it allows researchers to analyze large amounts of data. This section includes contributions to corpus annotation, corpus design, and corpus-based language teaching. Another chapter discusses how corpus-based research can study language change over time. The authors demonstrate how analyzing changes in word frequency over time can reveal insights into linguistic evolution. While another contribution discusses how corpus-based language teaching can improve second language acquisition. The authors argue that exposing learners to authentic language use through corpora can develop more naturalistic language skills.

This book's third part explores projects with corpus analysis and natural language processing as the main areas of interest. Computational linguistics studies how computers can process natural language data, while natural language processing is the application of computational techniques to analyze and understand human language. This section includes contributions to machine translation, named entity recognition, and text classification. For example, one of the chapter studies how machine learning can improve sentiment analysis. The authors demonstrate how training a machine learning algorithm on a large corpus of annotated data can improve its ability to classify sentiment accurately in new texts. Other scholars made substantial advancements in how named entity recognition can extract information. This book overviews current Digital Humanities, Corpus, and Language Technologies research. It demonstrates how these fields can enhance research across various disciplines. The conversation is now open. The data revolution has already changed everything. How would this inform the Humanities of tomorrow? This very question remains open, and yet its overwhelming and unattainable challenge is one of the most scientific quests that our generation must provide an answer to. The pages of this book are a modest but robust effort to create and find new paths.

Prof. dr. Pablo Valdivia
Academic Director Netherlands Research School for Literary Studies (OSL)
Chair-Full Professor European Culture and Literature – University of Groningen

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.

Part I
Digital Humanities

Chapter I

Understanding Outsider Art in the context of Digital Humanities

Entender el Arte Outsider en el contexto de las Humanidades Digitales

John Roberto & Brian Davis
Dublin City University – Ireland

Abstract: This chapter introduces the Outsider Art Project. "Outsiders" are highly innovative artists who have been aesthetically and socially marginalized because of their status as psychiatric patients, homeless, recluses, people with disabilities, migrants and ethnic minorities. Because of the need to characterize outsider art on a formal basis, this project is aimed at the automatic discovery of the semantic boundaries of outsider art in the context of digital humanities. From the methodological point of view, the Outsider Art Project is organized around three tasks: collecting a corpus of outsider art, generating a large dataset of digital images about outsider art and building the first ontology of outsider art.

Resumen: Este capítulo presenta el Proyecto de Arte Outsider. Los "outsiders" son artistas muy innovadores que han sido marginados estética y socialmente debido a su condición de pacientes psiquiátricos, personas sin hogar, reclusos, personas con discapacidad, migrantes y minorías étnicas. Debido a la necesidad de caracterizar el arte outsider de manera formal, este proyecto tiene como objetivo el descubrimiento automático de los límites semánticos del arte Outsider en el contexto de las humanidades digitales. Desde el punto de vista metodológico, el Proyecto de Arte Outsider se organiza en torno a tres tareas: recopilar un corpus sobre arte outsider, generar un gran conjunto de datos de imágenes digitales sobre arte outsider y construir la primera ontología del arte outsider.

1. Introduction

The world of art and culture can be divided into *mainstream art* and *outsider art*. Outsider artists are highly creative people who have been marginalized because they have broken, in some way, whether intentionally or not, rightly or wrongly, with the cultural conventions, rules and codes established by a community. Hence, we are referring to people with some form of physical, intellectual, or psychiatric disability, members of minority groups and social misfits involved in any artistic activity. Outsider artists often employ obsessive and repetitive patterns to represent disturbing themes such as sex and violence through the use of unconventional materials.

Outsider art is a concept that cannot be defined in absolute terms. The word was coined by Roger Cardinal in 1972 as an English equivalent for the term 'art brut', which was created around 1945 by the French artist Jean Philippe Arthur Dubuffet. Dubuffet stated that Art Brut was free from all social and cultural constraints because outsider artists are unfamiliar with the academic dogmas in which mainstream artists have been schooled. According to Professor Colin Rhodes, "as a category construction, 'art brut' was meant to highlight a creative tributary that was not so much different in kind from mainstream art, but rather in its lack of self-censorship or interest in following art world fashions" (C. Rhodes, personal communication, December 8, 2020). Throughout its history, the term outsider art has been associated with very closed terms that focused on a specific dimension of the notion. For example, the term 'naïve art' emphasizes the lack of formal training of some artists, 'neuve invention' is used to refer to subversive and inventive artists, and 'self-taught art' is a term which tries to avoid "the stigmas that some feel are attached to the Outsider Art definition" (Raw Vision magazine). Often, such definitions may end up in overlaps or even fall into circular reasoning: "Art Brut means 'Raw Art'" (Raw Vision magazine[1]) and "Outsider art is used to describe art that has a naïve quality" (the Tate website's glossary[2]).

In general, outsider art has always been the "other art". For many in of the mainstream art community, outsider art is considered an "anti-intellectual", "anti-professional" and "anti-academic" genre. Even, it is seen as "unsightly rubbish" by some art purists (Hernández, 2014). A significant part of the artistic mainstream despises outsider art, partly because its creators are seen to exist outside established culture and society, and partly because they are artists with a disability or untrained artists. A prototypical example of an outsider artist is Rodó[3]. Rodó is a Latin American artist diagnosed with paranoid schizoaffective

1 https://rawvision.com/
2 https://www.tate.org.uk/
3 The name "Rodó" is a pseudonym for the real identity of the artist. Outsiders wish to remain anonymous in

disorder. He emigrated to Barcelona (Spain) in the late 1990s, where he did not have an easy life: he slept on the streets and begged for money. When Rodó was a child, he enjoyed sculpting in clay and painting in oils. Nowadays, Rodó divides his time between his job as a cleaner and painting with watercolours. However, the truth is that despite his talent, Rodó has little hope of achieving fame.

From the analytical point of view, understanding outsider art is a considerable challenge, due to the large number of prejudices and misunderstandings surrounding the conceptualization of this artistic style. Although marginalization is a common trait of the artistic and cultural worlds, the marginalization of outsiders is the rule. For example, abstract expressionism was a mainstream movement defined by the machismo of its most representative figures, Jackson Pollock and Willem de Kooning. The New York School – which represented the abstract expressionists in America – rejected the painter Robert Rauschenberg for being gay and neglected the work of the American artist Lee Krasner for being a woman. Hans Hofmann once said, with regard to a painting by Krasner: "so good you would not know that it was done by a woman." Therefore, if gender inequality is predictable in mainstream art (Miller, 2016), then female outsider artists are discriminated against both because they are women and because they are outsiders. Indeed, there also seems to be a tendency towards the structural exclusion of women from the "canon" of outsider art. In a show organized by the Hayward Gallery featuring the most prolific outsiders of the last several decades, 91.3% were male and only 8.7% were female. However, what is particularly poignant for outsider artists is that some of them would not even consider themselves to be artists. An example is Barry Woo, who said the following when he was called an artist: "I thought I was just a 'schizophrenic'!".

In this chapter we present the Outsider Art Project, an innovative research project that applies digital technologies to the objective conceptualization of the artistic practices that lie outside the mainstream art world. Analysing outsider art by computational means is important for the characterization of a hermetic part of the world of creativity and, by extension, of society. From a scientific point of view, outsider art is an entry point for understanding a number of complex and interdisciplinary issues such as the *psychological* relationship between art and disability (Pettinari, 2019), how cultural (*sociological*) products are legitimated as art (Alexander & Bowler, 2021) and the *philosophical* role of artistic artefacts in the reproduction of power and domination in our society (Safina *et al.*, 2020),

a way that is similar to the street artist Bansky, for whom anonymity is vital because graffiti is illegal. In the case of outsider artists, anonymity protects them from social rejection.

among others. This project will provide a better understanding of an art often produced by people who are socially and culturally marginalized by assigning semantic meaning to huge amounts of textual and visual data.

This chapter is organized in five sections, in addition to this introduction. Section 2 discusses outsider art as a concept and describes its relationship to mainstream art. Section 3 deals with two main problems affecting the state of the art of scientific production in outsider art. Section 4 presents the methodological framework that we consider necessary to understand outsider art. Sections 5 and 6 briefly introduce the key resources with which we work: the corpus, ontology and dataset of images. Finally, Section 7, presents our conclusions and summarizes the most salient points made in this chapter.

2. Outsider Art, a Bargaining Chip for Contemporary Art

Outsider art must be considered an extremely complex phenomenon in which different "levels of reality" are present simultaneously. There have been many attempts to define outsider art across the disciplines, though most of them have limited themselves to presenting personal views and concerns about the concept without providing empirical evidence or having a formal basis. For example, the New York Times journalist Roberta Smith (1996) attempted to define the concept as "a somewhat vague, catchall term for self-taught artists of any kind". The critic, curator and writer Lyle Rexer (2005), in an attempt to characterize the confusing terminology around the term, defines outsider art as art "created under the conditions of a massively altered state of consciousness, product of an unquiet mind". Ramón Almela (2006), Ph.D. in Art, talks of "art created outside of conventional circumstances". David Davies (2009) proposed a theoretical characterization of the artistic status of outsider art on the basis of broader considerations regarding the philosophy of art. Jerry Saltz (2013) argues that outsider art does not exist at all, except as a discriminatory boundary preventing untrained artists from taking their rightful places in the canon. Linda Rainaldi (2015) later examined American and European perspectives on outsider art, focusing on biases, ideologies, and social factors, concluding that "I was no closer to articulating one comprehensive definition of outsider art". Rebecca Hoffman, director of the Outsider Art Fair, has her own, more general criteria: "I utilize the term 'outsider art' as an umbrella for a lot of different categories" (Acosta, 2015).

The point here is that outsider art is culturally marginalized by mainstream art. Thus, while mainstream artistic styles (e.g., cubism, realism, baroque or abstract) are usually described on the basis of artistic criteria such as the use of the colour, shapes, space or

techniques, outsider styles are most frequently described on the basis of negative non-artistic criteria such as the mental condition or the lack of training of the artist. In the cases in which aesthetic criteria were used, they tend to lead to a negative assessment of the works of art. Paradoxically, in spite of this, "outsiders" are considered to be highly innovative artists and the visibility of outsider art has increased dramatically in recent years. Even more paradoxical is the fact that mainstream artists have found inspiration in the work of their marginalized peers.

As a result, there is an unhealthy relationship between mainstream art and other forms of art. Experienced artists, such as Paul Klee, Wassily Kandinsky, Pablo Picasso, Jean Dubuffet, Max Ernst and André Breton, sought "inspiration" in the art of children, the art of "primitive" societies, the art of madness, mass culture and even in totally unintentional art such as that produced by animals. A well-documented story in this sense is that of the British zoologist Desmond Morris, who sold paintings by a chimpanzee named Congo to Salvador Dali, Pablo Picasso and Joan Miro. We also all know that Andy Warhol became a huge influence on popular culture by placing ordinary everyday items at the heart of his work. He said, "I don't think art should be only for the select few, I think it should be for the mass of the American people." With this in mind, Warhol turned art into a mass-produced commodity and the artist into a brand name. Max Ernst, who abandoned his studies in psychiatry at the University of Bonn for painting, was profoundly interested in the "art of the insane" as a way to access primal emotion. Ernst was probably responsible for bringing *art brut* into surrealism. Paul Klee wrote that "in our own time worlds have opened up which not everybody can see into, although they too are part of nature. Perhaps it's really true that only children, madmen and savages see into them" (MacGregor J., 1989). Joan Miró also turned to "extra-cultural art" for inspiration, including children's art and primitive and folk-art. Linda Ferrell (1983) states that "Miró has not only made use of a child's color scheme, but he has added the child's painting technique to the shapes and motifs he has chosen and to his use of space and line." Ferrell also argues that Jean Dubuffet's art shows a major influence from the art of children. Specifically, he used elements from the artwork of children in the dawning realism stage, which marks the transition between art as purely symbolic to art as a creative outlet. In the same vein, Heather Malin (2013) from Stanford University states that Wassily Kandinsky "gave special privilege to the lack of purpose in children's art making" and, in an article published by Sharla Ackles from Colorado State University, she stated that:

> *Most of the artists who have been influenced by the art of the primitive have included the art of children as an influence. One of the artists who used children's art as his main source of inspiration was Paul Klee. He had great respect and enthusiasm for the work of children (Ackles, 1988).*

The case of outsider art is paradigmatic in this regard because there are those who believe that outsider art has been used, reproduced and finally scrapped by mainstream art: "the mainstream appropriates artifacts as art but then insists that they occupy a marginal or degraded position" (Alexander & Bowler, 2021). As a result, there are mainstream artists who draw "inspiration" from outsider artists. For example, in Figure 1 we can see the similarities and coincidences between an illustration by the Spanish illustrator Ricardo Cavolo (Figure 1a) and a serigraphy by the outsider artist Antonio Roseno de Lima (Figure 1b). Therefore, the demarcation line between both artistic styles, outsider and insider, in terms of their mutual influence can be difficult to define. Consider, for example, the case of the self-taught artist Jean-Michel Basquiat, who has been directly classified by some art historians as an outsider because of his use of found materials and the obsessive and repetitive use of symbols in his work. Others, however, find this idea disturbing because Basquiat's work sells for millions. On the other hand, Jean Dubuffet, who was greatly inspired by the work of the outsider painter Adolf Wölfli, completely embraced this style. Along the same lines, but regarding the neural mechanisms regulating face and body perception in the work of the mainstream artist Francis Bacon, researchers on neuroaesthetics at University College London stated that "he [Bacon] subverted the normal neural representation of faces and bodies" (Zeki and Ishizu, 2013), leading to produce a "visual shock" in the spectator (see Figure 1c). We can observe a similar effect in the portraits of the outsider artist Jean-Marc Renault (see Figure 1d) who created "a dozen portraits of war victims who carry their physical deformation forever" (Chernetska, 2020).

Apart from mainstream art, it is very surprising – or perhaps not – the extent to which outsider art shares some common visual traits with the art of children. Figure 2 shows how both an outsider artist and a four-year old boy represent a human figure. Aside from the differences related to age, for instance the fact that the child has not introduced a baseline to organize objects in space, both subjects share a common vision of some parts of the body such as the feet, knees, waist (belt buckle), chest (right pocket), hands in pockets or arms that are drawn close to the body and big eyes. Typically, the drawings of children and outsiders are self-portraits and may be a realistic portrayal or an idealized image. In the case of children, it is known that egocentric thinking plays a crucial role in the self-defining process of four-year old boys and girls. In the case of outsiders, psychologists state that a

Figure 1. (a) Ricardo Cavolo's illustration (Cavolo, 2021). (b) *Bebado*, serigraphy by the outsider artist Antonio Roseno de Lima (Collection de l'Art Brut, undated). (c) Francis Bacon, *Self-Portrait* (Artnet'news, 1969). (d) Jean-Marc Renault, Portrait no. 9 (Renault, 2018).

"preoperational features such as egocentric thinking and perception-bound reasoning have been implicated in the association between schizotypy and creativity" (Winston *et al.*, 2014).

Figure 2. Left: Painting by the outsider artist Daniel Saracho (Marginarte, 2019). Right: drawing by a four-year old boy (Marginarte, 2019).

3. State-of-the-art in Outsider Art

Until now, outsider art has been analysed in the light of theoretical[4] but **not** computational models. According to the Scopus database, while 99% of the papers in computer science dealing with artistic styles are about mainstream art (e.g., pop, conceptual, abstract and street art), only 1% of papers are about outsider art. Thus, it is not uncommon to find papers on mainstream art describing a mathematical algorithm to produce abstract paintings (Spann, 2020), on applying optical techniques with the aim of identifying similarities and differences between the 17th century painting Madonna della Cesta by Rubens and a Piero Fevere tapestry (Dal Fovo, *et al*., 2020), on detecting the presence of graffiti art on building facades using Deep Learning models (Novack *et al*., 2020), or on generating pop art-like images from photographic images using binomial distribution methods (Hiraoka, 2020), among many others. However, this does not occur with outsider art, where we can refer to only two works in computer science: Roberto & Davis (2020) and Roberto *et al*. (2020). We call this problem the *computational gap*.

On the other hand, although there are no studies in this regard, there are reasons to think that less than 2% of the documents on outsider art are written in the first person. This is particularly strange considering that outsider artists are prone to expressing their feelings in writing. In contrast to outsider art, it is not uncommon to find papers on mainstream art written by artists in the first person. First-hand experience in fine art is a self-reflexive qualitative research method which foregrounds the artist's subjectivity. By probing the "artist's intent" it is possible to improve different tasks such as the conservation of works of art: "it seems that the conservation field is opening up towards the use of writing in first person in art research" (Quabeck, 2021). The value of first-person texts for fine art experts is based on the generation of reliable knowledge by co-constructing (with the artists) instead of reconstructing the experience of the artist. Unfortunately, the co-construction of knowledge based on artists' first-hand experience is not frequent in the research on outsider art, probably because researchers do not consider the artists a reliable source of information. We call this problem the *data imbalance problem*.

It is therefore necessary to develop methodologies for describing outsider art based on objective and formal knowledge, such as those provided by processes like digitization, computation and the quantification of linguistic and graphic data. Natural Language Pro-

4 For example, Baumann's general theory of artistic legitimation (Alexander & Bowler, 2021) or Bourdieu's conceptualization of disinterestedness (Ardery, 1997).

cessing and Machine Learning techniques play a significant role in this task. But first, it is necessary to define framework that support both approaches.

4. Methodological Framework for Understanding Outsider Art

The Outsider Art Project is being conducted within the framework of the digital humanities. However, there are two behaviours which, according to the critics, should be avoided in digital humanities projects. First, thinking that digital humanities is just "about introducing digital technologies where there were none before" (Brennan, 2017) and, secondly, believing that it is possible to "reveal the secrets of complex social and cultural processes" through algorithmic computation. Therefore, we are considering digital humanities as a methodological framework in order to place outsider artists at the centre of the research and to promote the development of digital infrastructures for the computational processing of outsider art. Other aims, different to those already proposed, should be evaluated on the basis of social and cultural criteria by attending to the voices of multiple stakeholders and considering the complexity of the subject matter. The latter leads us to talk about the transdisciplinary and multimodal nature of the Outsider Art Project.

According to different researchers such as Kemman (2019), "one of the defining characteristics of digital humanities is its emphasis on interdisciplinary collaboration" between disciplinary peers (research teams, faculties, laboratories and institutions). But describing digital humanities as interdisciplinary practices places limitations on our research. That is because of the possibility of collaborative work between scholars or "disciplinary peers" ruling out the voice of underprivileged and marginalized groups, including outsider artists (see "data imbalance problem" at Section 3). As Martin and Runyon (2016) recognise:

> *The digital humanities represent, for many researchers, the potential for extending their research in terms of audience, scope, methods, and opportunity for interdisciplinary collaboration. Ideally, this potential should also extend access to cultural engagement and preservation for marginalized groups.*

In order to overcome the limitations associated with interdisciplinary research, we considered it more appropriate to adopt a transdisciplinary approach. Adopting a transdisciplinary approach can influence scientific agendas and change the dynamics of research by promoting the participation of disadvantaged actors. Indeed, it is clear that social actors other than researchers play a crucial role in transdisciplinary research. Transdisciplinary research occurs when academics and non-academics contribute their different expertise to understanding a problem holistically by developing a common intellectual framework

that goes beyond particular perspectives. Seeking the collaboration of researchers and non-academic actors in order to develop a common definition of a problem is a way to deal with the complexity of real-world problems such as those referring to cultural marginalization. Therefore, in contrast to those who emphasise the interdisciplinary nature of the digital humanities, we prefer to state that the digital humanities is a transdisciplinary field. This assertion is supported by bibliometric analyses such as those obtained by Yang *et al*. (2020) and Isemonger (2018). At the same time, one ought not to forget that in order to resolve real world or complex problems, transdisciplinarity places the emphasis on humanities: "transdisciplinarity integrates the natural, social and health sciences in a humanities context, and transcends their traditional boundaries" (Choi, 2006). A transdisciplinary view of outsider art will enable us to make both societal and scientific advances by looking at a problem from many angles and by involving both academics and marginalized artists.

In addition to the need to establish a transdisciplinary framework for the project, we are aware of the fact that understanding outsider art depends on analysing both textual and pictorial information. It is therefore necessary to have a multimodal model of semantics that makes it possible to link textual information with its real-world counterpart, (digital) cultural objects, and, as we shall see below, with emotional information too. This is not a new approach, there are a number of voices arguing in favour of "visual digital humanities":

> *Since there are several overlaps in epistemic cultures of visually oriented and digitally supported research in art and architectural history studies, museology, and archaeology, as well as cultural heritage, we introduce 'visual digital humanities' as novel 'umbrella' term to cover research approaches in the digital humanities that are dependent on both consuming and producing pictorial, rather than textual, information to answer their humanities research questions (Münster and Terras, 2019).*

The multimodality of digital cultural information arises from external and internal factors from which outsider art is not exempt. First, this is due to the development of new Information and Communications Technologies (ICTs) for creating and linking textual and graphic information. There are many tools for creating digital exhibitions that allow experts to manage digital assets and create robust narratives and layouts for display online. For example, Contentdm and OmekaS are publishing platforms for institutions interested in connecting digital cultural heritage collections with other resources online. Digital technologies for cultural heritage have demonstrated their value by offering a virtual space in which to build ideas collectively. Currently, different museums around the world are using a number of digital technologies that allow the users to add digital content to cultural

items. This is the case of the GIFT Box[5], a set of apps that allow visitors to add new digital content to a physical exhibit and ArchAIDE[6], a software that automatically identifies archaeological ceramic fragments pieces thereby allowing experts to enter textual descriptions about them. Obviously, this enormous amount of cultural data (texts, images and audio) needs to be interpreted and contextualized in order to be useful.

The metaphorical meaning of cultural assets is the second reason to explain the multimodal digital humanities. This metaphorical meaning emerges from the symbolic nature of feelings and emotions for both creators and viewers. On the one hand, cultural artefacts are made by creators to be beautiful but also to express an important idea or feeling while, on the other hand, viewers use their own experiences, views, and preferences to "understand" cultural artefacts. As a result, heritage materials tend to be embedded in narratives and analogies that can be interpreted by expert curators and interested lay persons. That motivates us to think that the semantic enrichment of outsider art collections must be based on models that integrate visual and emotional information, in addition to linguistic information. Empirical work on semantic processing has shown that integrating both forms of information together with linguistic information plays an important role in understanding semantic data. Rotaru and Vigliocco (2020) found that including visual and emotional information performs better to capture affective information than purely linguistic models based on distributional models of semantics, such as Latent Semantic Analysis (Landauer & Dumais, 1997). They are even more specific: "we found that including visual information is particularly beneficial to more concrete concepts, whereas including emotional information is particularly beneficial to more abstract concepts" (p.16). Similar results have been shared by De Deyne *et al.* (2018) and Ponari *et al.* (2018), among others. Therefore, we assume that in order to understand outsider art it is necessary to combine linguistic information derived from objective text corpora (e.g., scientific papers), visual information derived from image collections (e.g., the textual descriptions that typically accompany objects in digital collections), and emotional information derived from first-person texts by outsider artists.

In this regard, it is important to emphasise that cultural artefacts are often enriched with and through linguistic information. Moreover, the way in which cultural heritage artefacts are represented and communicated has a significant impact on the way in which those artefacts are interpreted. A semiotic approach to the museum phenomenon consid-

5 https://gifting.digital/
6 http://www.archaide.eu/

ers museum objects as performing a social function, always enhanced by textual descriptions that contribute to the process of sign production and of sign interpretation. For example, museum catalogues are uniquely valuable sources because they encourage visitors to recover their freedom of decoding, while at the same time they function as a marketing tool that encourages people to come and buy cultural goods and may even confer additional value to a specific piece. Such publications must be capable of capturing the complexity of an exhibit in a written text. Additionally, cultural heritage artefacts need to be digitised and labelled with metadata standards in order to be shared across different environments and domains. In other words, metadata standards enable intra-collection searches and also support cross-boundary access to collections. This provides an opportunity for users to interconnect the cultural heritage objects to contextual information and vice-versa.

From our point of view, addressing the social, aesthetic and linguistic issues surrounding outsider art requires an analysis of both texts and images by computational methods. That is because, in the world of the arts, visual and textual languages are two sides of the same coin. Therefore, in our project we are applying Natural Language Processing to the interpretation of texts on outsider art while applying Machine Learning to the analysis of paintings by outsider artists.

5. Analysing Natural Language to Understand Outsider Art

This project draws on Natural Language Processing and Computational Linguistics to understand how society perceives outsider art or, more specifically, how outsider art is conceptualised in scientific and popular writing. According to the Stanford Encyclopedia of Philosophy[7] "Computational Linguistics (CL) is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artefacts that usefully process and produce language, either in bulk or in a dialogue setting." Similarly, Natural Language Processing (NLP) is broadly defined as the automatic manipulation of natural language by software. Natural Language Processing and Computational Linguistics are helping us to understand outsider art by automatically capturing/enriching data with metadata and by transforming textual content into a computer-reliable format. In the Outsider Art Project, the first of these tasks has been tackled through the compilation of the outsider art corpus and the second task is currently being carried out through the development and implementation of the outsider art ontology.

7 https://plato.stanford.edu/entries/computational-linguistics/

5.1. The Outsider Art Corpus

Collecting textual data about outsider art is the first step toward understanding this domain. Thanks to the explosion in the volume of machine-readable text and advances in available computing power, text corpora have become essential components of new developments in computational linguistics from 1980 until the present. Corpus linguistics provides a wealth of experience in dealing with language problems and also contribute to the understanding of specific domains. In both cases, the kind of data plays an important role in achieving research goals. In the case of the analysis of outsider art, we found it useful to make a distinction between primary and secondary data.

In general, raw text is classified as primary data, while annotations of these primary texts are considered secondary data. However, considering that "the term 'secondary' suggests that the data provide indirect access to the research domain" (Østergaard & Torst, 2017), we have adopted a broader vision of data types. So, primary data refers to those data that are collected directly from the source, in our case, first-person texts by outsider artists. In contrast, secondary data involve an existing document, which had previously been used by another researcher for a different research question. Secondary data often involve the interpretation of cultural artefacts and are distant from the time and place of the original artefact. An example of primary data with which we work is the illustrated novel entitled *The History of My Life*, the autobiographical narrative of the outsider artist Henry Darger. An example of secondary data is the book *Henry Darger, in the realms of the possibly real*, a biography of Darger by Jim Elledge.

For this project, we decided to compile a large text corpus of secondary data for two main reasons: the lack of primary data and the difficulties of anonymizing it. Although there is a long tradition in cultural heritage of capturing primary data, this type of data is scarce in the field of outsider art (see Section 3). Therefore, while there are many books, catalogues, magazines, webpages and articles on outsider art written by experts, there are not many artists' accounts of their own experiences captured through interviews or any other primary data collection method. Besides the problem of this lack of primary data, personal information on outsider artists should also be removed from primary data in order to reduce the risk of unnecessary information exposure to third parties. Encryption, pseudonymization and anonymization are methods for removing sensitive information from documents and are also known as de-identification methods. In Kacane (2021), anonymization is performed by the interviewees themselves who were asked about their habits in attending museums. Automatic de-identification methods, in turn, are typically limited to a few common named entity types (e.g., a person's name, hometown and work-

place) and "human supervision will still be needed for it to completely guarantee the anonymization of the messages" (Helbrink & Åkesson, 2020). However, the de-identification of fine-grained entities, such as the titles of artworks and nicknames, is of great importance for outsider artists. Therefore, it is necessary to seek ways to adjust sensitive personal data in such a way that it is no longer possible to identify the originating outsider artist before working with primary data. We assume that the fine-grained de-identification of personal information for research purposes involving marginalized groups is a pending task and this has a direct impact on corpus goals.

We compiled the outsider art corpus with the goal of describing how society understands outsider art by identifying the patterns of language use in the target textual domain. Specifically, we are interested in discovering *how* outsider art is conceptualised in writings about art. Therefore, the question that the outsider art corpus must be capable of responding to is: *what* are the terms/concepts and linguistic structures that characterise texts on outsider art? The outsider art corpus will be used as a silver standard for machine learning because it is (semi)automatically generated. Our aim is to use this corpus to train machine learning algorithms that are able to capture the main essentials of the outsider art knowledge domain: concepts and hierarchies.

The outsider art corpus currently contains 981,868 words extracted from 450 documents that have been collected by hand in order to ensure quality and relevance. The corpus includes English texts that talk about outsider art, *art brut*, folk art, naïve art and self-taught art. We include three main text types or genres: artist bios, scientific articles (e.g., books and papers) and op-ed articles (e.g., art criticism and art press releases). The texts in this corpus had been obtained from web pages and documents in PDF format. Additionally, there is a set of texts coming from printed books consisting of excerpts of text under copyright law[8]. Every text in the corpus is stored within a separate XML file (in UTF-8 text encoding). Two main types of XML annotations were added to the outsider art corpus: meta-information about the document (e.g., author, genre[9], if the text is an excerpt from a major work, theme/style[10], type of source and *url*) and information about the structure of the document (e.g., paragraphs, sentences, titles and subtitles).

In addition to the foregoing, a subset of 1,690 random sentences has been manually annotated with domain-specific terms belonging to three different semantic categories as shown in Table 1: (a) very typical outsider art terms, (b) terms that bear a relationship with

8 Only a minor part (10%) of the total document has to be scanned in order to obtain the raw text.
9 Artist bio, scientific article and op-ed article.
10 Outsider art, *art brut*, folk art, naïve art, self-taught art and autism.

the life and creative work of outsider artists, and (c) terms that include a wide range of specific entities not directly connected with outsider artists. We performed this task with CATMA [11] open-source software, which allowed us to define our own set of tag categories. Each annotation collection in CATMA is represented as one TEI XML file and terms can be retrieved by using a character offset (the position of the first letter and the last letter of the selected term). This subset of random sentences will be used as a gold standard domain model in order to establish a system for detecting outsider art terms automatically.

Table 1. Examples of domain-specific terms.

a.	Yet, for **outsider artists** , who are **self-taught** , **amateurish** and **reclusive** , the usual rules don't apply.
b.	Born in 1891, **Marino Auriti** was an **Italian-American self-taught artist** .
c.	**Roger Cardinal** published a book in 1972 with this title.

Finally, it is important to note that bias is an additional problem affecting secondary data related to outsider art. In our experience, language and gender are the most important factors influencing the process of the interpretation of outsider art. There is an overrepresentation of English-speaking articles and European and North American regions in the literature on the subject. This is not only because English is the dominant language (language-based bias) but also because most featured artists were born in the United States or Europe (geography-based bias). In the same way, gender is one of the most prevalent biases in this domain since the featured artists are mostly male. Gender imbalance in the art world (see Section 2) has been documented by Bocart *et al.*, 2017 and Cameron *et al.*, 2017, among many other researchers. Therefore, factors causing bias have been controlled for where possible by applying existing methods such as those described by Wang *et al.* (2020) and Sun *et al.* (2019).

5.2. The Outsider Art Ontology

Capturing and codifying knowledge related to outsider art is the second step towards understanding this domain. Therefore, an important task of the Outsider Art Project concerns encoding knowledge about outsider art in a machine-readable language or computational ontology. In computer science, an ontology is a linguistic/cognitive based representation of the concepts, relations, attributes and hierarchies that are present in a given domain of

11 https://catma.de/

knowledge. For example, in the expression "Adolf Wölfli was born in Bern" the term "Adolf Wölfli" is an instance of the category "outsider artist" and is linked to the word "Bern" (capital of Switzerland) by the relation "was born in". An ontology is filled with thousands of these relations, which makes it possible to draw complex inferences about the domain.

Ontologies for cultural heritage are interdisciplinary artefacts since they describe objective manifestations of the human mind, including customs, practices, places, objects, artistic expressions and values. There are a number of projects in Europe working to reduce the digital gap between the humanities and technology through the creation of ontologies and new metadata models for representing knowledge related to cultural heritage, including Europeana and POSTDATA (González-Blanco *et al.*, 2018). Europeana is an authoritative repository of more than 58 million cultural and scientific heritage objects represented in the Europeana Data Model (EDM[12]), a metadata framework for the interoperability and standardisation of cultural data. The EDM metadata standard contributes to the creation of new knowledge by incorporating semantic information from external resources located in different countries across Europe. The POSTDATA[13] (Poetry Standardization and Linked Open Data) project has as its main objective to provide a means to publish European poetry (EP) data as Linked Open Data (LOD) through the creation of a digital semantic web-based platform for poetry analysis and edition. Although there are several repertoires and databases that have the "poem" as object of study, they cannot communicate because they are not semantically interoperable. Therefore, POSTDATA applies a reverse engineering process by which the project team analyses the logical models of different databases in order to create a common conceptual model for all the existing ones.

To the best of our knowledge, there has been no attempt to formalize knowledge about outsider art via a computational ontology or any other tool for terminological standardization. Therefore, we are constructing the ontology of outsider art by assigning meaning to the large amount of relevant but scattered textual data stored in electronic form. Concretely, we are applying Natural Language Processing and Machine Learning techniques to the development of a machine-processable ontology in a semi-automatic fashion. It is important to point out that, when categorising aesthetic objects, the rule is to integrate several external resources. There are several examples of ontology integration in the cultural heritage field, including the Conservation Reasoning ontology (Moraitou *et al.*, 2018) and the Heritage Building ontology (Tibaut *et al.*, 2018).

12 The Europeana Data Model for Cultural Heritage.
13 https://postdata.linhd.uned.es/

However, due to the heterogeneity of the concepts potentially associated with the outsider art domain, we decided to build the ontology from scratch. Indeed, the outsider art ontology must deal with both the artistic/cultural and social issues associated with inequality, mental disorders, physical disabilities, racial and ethnic origins and geographical/geopolitical settings, among others. For example, as can be seen in Figure 3, Henry Joseph Darger is characterized by a set of artistic and non-artistic properties that depict him as an outsider artist (novelist, painter and draughtsman). Some of the artistic properties are "has exhibited in*: collection de l'art brut*", "creator of*: the story of the vivian girls*", "use of materials*: recovered paper*" and "deal theme sex*: nudity*". Some non-artistic properties associated with Darger are "worked as*: janitor*", "enrolled in*: mission of our lady of mercy*", "suffer mental condition*: tourette syndrome*", "born place*: chicago*" and "featured by*: john macgregor*". As can be seen in Figure 1, the central class in the ontology is the outsider artist, represented by the "Creator" category. This is one of the major differences with respect to other existing cultural heritage ontologies in which the collection or the artifact/object occupies a prominent position.

In a basic sense, the main goal of the outsider art ontology is to contribute to the transfer of knowledge between different sectors and disciplines by standardizing the terminology associated with this artistic phenomenon. Additionally, this resource will be used to preserve and disseminate outsider art collections and to develop high-level software tools (e.g., systems that recommend outsider art assets to tourists).

6. Analysing Images to Understand Outsider Art

Digital images play an essential role in cultural heritage. Encoding the image features of paintings for classifying art styles automatically is a typical task in the field of the computational analysis of visual aesthetics. A few datasets of fine-art images are commonly used to train automatic image classifiers but none of them are about outsider art. For example, Painting-91[14] (Khan *et al.*, 2014) is a dataset consisting of digital paintings from 91 different painters including Picasso, Rubens and Kandinsky; Art500K[15] (Mao *et al.*, 2017) is a large-scale dataset containing over 500,000 artworks annotated with detailed artist labels; the Sculptures 6k Dataset of images (Arandjelović & Zisserman, 2011) consists of 6,340 sculptures by Henry Moore and Auguste Rodin collected from Flickr; the Museum of Modern

14 http://www.cat.uab.cat/~joost/painting91.html.
15 https://deepart.ust.hk/ART500K/art500k.html.

Figure 3. A fragment of the outsider art ontology.

Art (MoMA) dataset[16] contains 15,236 records with basic metadata about all the artists who have work in the MoMA collection, although images must be requested separately via email; SemArt[17] is a collection with 21,384 digital paintings in which each image is associ-

16 https://github.com/MuseumofModernArt/collection.
17 http://noagarciad.com/SemArt/.

ated to a textual artistic comment; ErgSap[18] is a visual art gallery application that contains almost 60,000 images of art work grouped by artist; the WikiArt[19] dataset contains over 80,000 images of art work labeled across 27 varied art styles collected from WikiArt.org.

As with primary data, there is an important lack of datasets on outsider art painting which would allow for research to be carried out on visual aesthetics based on machine learning approaches. To resolve this problem, we are preparing a large dataset of outsider art paintings. A first version of this dataset with 3,616 images was used in Roberto *et al.* (2020) to establish an initial approach to the automatic classification of digital images related to outsider art. This limited version of the outsider art dataset merged 2,405 images labelled as Naïve Art from WikiArt, a category that is considered to be very close to the outsider art style (Van Heddeghem, 2016, p.13) and 1,232 outsider art images collected from different sources. In the referenced paper, we addressed the question of whether it is possible to classify different artistic styles by using Deep Learning methods. Preliminary results suggested that there are no significant differences between ten artistic styles, including outsider art. Additionally, we concluded that outsider art can be computationally modelled by objective means but it is necessary to dispose of a larger dataset in order to provide stronger and more robust assessments. For this reason, we are currently generating a large dataset with 10,000 images related to outsider art, folk art, naïve art and *art brut*. Generating a new dataset involves routine tasks such as collecting digital images via crawling and scanning, transforming images into digital format (if necessary), editing images and removing de-duplicates and noising images. These images are taken from social networks, non-governmental organization, museums, galleries, books and magazines, among other sources.

7. Conclusion

This chapter describes the main goals, the development status and the methodological details of the Outsider Art Project, which is being carried out at the ADAPT Centre of Dublin City University. We propose a transdisciplinary and multimodal framework for identifying and classifying the main concepts in the outsider art domain. We claim that, in order to properly understand this domain, it is necessary to analyse heterogeneous data including text and images, and to incorporate the voices of multiple stakeholder groups at different stages of the project. However, due to a lack of data for undertaking a computa-

18 https://art.ergsap.com/downloads.
19 https://github.com/lucasdavid/wikiart.

tional analysis of the domain, our efforts have mainly been aimed at collecting a corpus of texts about outsider art and a large dataset of digital images of outsider artworks. Additionally, we are developing the first ontology of outsider art to standardize the terminology of the domain in order to enable semantic interoperability between heterogeneous metadata and to examine the relationship between social exclusion and cultural artefacts. In general, *the Outsider Art Project posits outsider art as an object of study of digital humanities by entailing the existence of a research niche merging art, technology and society.*

References

Ackles, S. (1988). *The influence of primitive art on early modern European painters*. Colorado State University.

Acosta, A. (2015). A semantic analysis of the meaning of the word outsider art. *ArtsLife*. https://artslife.com/2015/07/23/a-semantic-analysis-of-the-meaning-of-the-word-outsider-art/

Alexander, V. D., & Bowler, A. E. (2021). Contestation in aesthetic fields: Legitimation and legitimacy struggles in outsider art. *Poetics*, *84*, 1-17. ISSN 0304-422X.

Almela, R. (2006). Outsider… deconstructing art from the outside. Epistemology of marginal art as an expressive visual practice. *Criticarte*. http://www.criticarte.com/Page/file/art2006/outsider_decons_ingles.pdf.

Arandjelović, R., & Zisserman, A. (2011). Smooth object retrieval using a bag of boundaries. *International Conference on Computer Vision*, 375-382.

Ardery, J. (1997). 'Loser wins': outsider art and the salvaging of disinterestedness. *Poetics*, *24*(5), 329-346.

Bocart, F., Gertsberg, M. & Pownall, R. A. J. (August 27, 2018). Glass Ceilings in the Art Market Available at SSRN: https://ssrn.com/abstract=3079017 or http://dx.doi.org/10.2139/ssrn.3079017.

Brennan, T. (2017). The Digital-Humanities Bust. *Chronicle of Higher Education, 64*(8). http://www.chronicle.com/article/The-Digital-Humanities-Bust/241424.

Cameron, L., Goetzmann, W. & Nozari, M. (2017). Art and Gender: Market Bias or Selection Bias? Available at SSRN: https://ssrn.com/abstract=3025923 or http://dx.doi.org/10.2139/ssrn.3025923.

Chernetska, A. (2020, August 5). Behind the mask. *Raw Vision Magazine, (106)*, 40-45.

Choi, B.C. & Pak, A.W. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clin Invest Med*. *29*(6): 351-64. PMID: 17330451.

Dal Fovo, A., Striová, J., Pampaloni, E., Fedele, A., Morita, M.M., Amaya, D., Grazzi, F., Cimó, M., Cirrincione, C., & Fontana, R. (2020). Rubens' painting as inspiration of a later tapestry: Non-invasive analyses provide insight into artworks' history. *Microchemical Journal*, *153*. 104472.

Davies, D. (2009). On the Very Idea of 'outsider art'. *The British Journal of Aesthetics*, *49*.

De Deyne, S., Navarro, D., Collell, G., & Perfors, A. (2018). Visual and affective grounding in language and mind. OSF.

Ferrell, L.L. (1983). *The influence of children's art on Joan Miró and Jean Dubuffet*. [Master thesis]. Mary Washington College of the University of Virginia.

González-Blanco, E., Ros, S., Ruíz, P. Díez, M. L., Bermúdez, H. *et al.* (2018). Poetry and Digital Humanities making interoperability possible in a divided world of digital poetry: POSTDATA

project. *EADH 2018: Data in Digital Humanities, European Association for Digital Humanities*, Dec 2018, Galway, Ireland.

Heather, M. (2013). Making Meaningful: Intention in Children's Art Making. *International Journal of Art & Design Education, 32*(1), 6-17.

Helbrink, J. & Åkesson, S. (2020). *Data Anonymization using Machine Learning and Natural Language Processing*. [Master Thesis]. Department of Computer Science. Lund University.

Hernández, J. F. (2014). Local Art, Global Issues: Tales of Survival and Demise Among Contemporary Art Environments. In L. Del Giudice (Ed.), *Sabato Rodia's Towers in Watts: Art, Migrations, Development* (pp. 29–68). Fordham University Press. https://doi.org/10.2307/j.ctt1c5cjcv.5.

Hiraoka T. (2020). Generation of pop art-like images using binomial distribution. *ICIC Express Letters*, *14*(3), 227-233.

Isemonger, I. (2018). Digital Humanities and Transdisciplinary Practice: Towards a Rigorous Conversation. *Transdisciplinary Journal of Engineering & Science*, *9*, 116-138.

Kacane, I. (2021). Heritage sites as means of bringing cultural awareness: intergenerational attitudes towards visiting museums. *Proceedings of INTED2021 Conference 8th-9th March 2021*. (pp. 8261-8266). Daugavpils University (LATVIA).

Kemman, M. (2019). Boundary Practices of Digital Humanities Collaborations. In W. Dillen, *et al.* (Eds.), *Integrating Digital Humanitites* (pp. 1-24). DH Benelux Journal.

Khan, F., & Beigpour, S, Weijer, J. & Felsberg, M. (2014). Painting-91: A large scale database for computational painting categorization. *Machine Vision and Applications*, *25*, 1385-1397.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

MacGregor J. (1989). *The discovery of the art of the insane. Princeton*: Princeton University Press.

Mao, M. & Cheung, M. & She, J. (2017). DeepArt: Learning Joint Representations of Visual Arts. *MM'17: Proceedings of the 25th ACM international conference on Multimedia* (pp. 1183–1191). https://doi.org/10.1145/3123266.3123405.

Martin, J., & Runyon, C. (2016). Digital humanities, digital hegemony: exploring funding practices and unequal access in the digital humanities. *SIGCAS Comput. Soc. 46*(1), 20-26.

Miller, D. (2016). Gender and the Artist Archetype: Understanding Gender Inequality in Artistic Careers. *Sociology Compass*, *10*(2), 119-131.

Moraitou, T., Aliprantis, J., & Caridakis, G. (2018). Semantic Preventive Conservation of Cultural Heritage Collections. *SW4CH@ ESWC.*

Münster, S. & Terras, M. (2020). The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures. *Digital Scholarship in the Humanities*, *35*(2), 366-389.

Novack, T., Vorbeck, L., Lorei, H., & Zipf, A. (2020). Towards Detecting Building Facades with Graffiti Artwork Based on Street View Images. *ISPRS International Journal of Geo-Information*, *9*(2), 98. http://dx.doi.org/10.3390/ijgi9020098.

Østergaard, S. & Torst, P. (2017). Research styles: data and perspectives in the human sciences. In C. Emmeche, D. Pedersen, & F. Stjernfelt (Eds.), *Mapping frontier research in the humanities*. Bloomsbury Academic.

Pettinari, G. (2019). The 'Art and Madness' debate in Italy and the life story of Antonio Tolomei. *Epidemiology and Psychiatric Sciences*, *28*(4), 369-370. doi:10.1017/S2045796019000258.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549.

Quabeck, N. (2021). Reframing the Notion of "The Artist's Intent:" A Study of Caring for Thomas Hirschhorn's Intensif-Station (2010), *Journal of the American Institute for Conservation*, DOI: 10.1080/01971360.2020.1826151.

Rainaldi, L. (2015). *outsider art: forty years out (T).* University of British Columbia. https://open.library.ubc.ca/collections/ubctheses/24/items/1.0221495.

Rexer, L. (2005). *How to Look at outsider art*. Harry N. Abrams, Inc. ISBN 10: 0810992027.

Roberto, J. & Davis, B. (2020). Towards the Ontologization of the outsider art Domain: Position Paper. *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation at LREC 2020*.

Roberto, J., Ortego, D. & Davis, B. (2020). Toward the Automatic Retrieval and Annotation of outsider art images: A Preliminary Statement. *Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access (AI4HI-2020)*. European Language Resources Association (ELRA), pp. 16-22.

Rotaru, A. S., & Vigliocco, G. (2020). Constructing Semantic Models From Words, Images, and Emojis. *Cognitive science*, *44*(4), e12830. https://doi.org/10.1111/cogs.12830.

Safina, A., Gaynullina, L., & Cherepanova, E. (2020). A work of art in the space of network culture: creativity as bricolage. *Creativity Studies*, *13*(2), 257-269. https://doi.org/10.3846/cs.2020.12264.

Saltz, J. (2013, February 1). Jerry Saltz on the outsider art Fair — and Why There's No Such Thing As 'Outsider' Art. *Vulture*. https://www.vulture.com/2013/02/jerry-saltz-on-the-outsider-art-fair.html.

Smith, R. (1996). The outsider art Fair' The Puck Building Lafayette and Houston Streets SoHo Through Sunday. *The New York Times*. https://www.nytimes.com/1994/01/28/arts/art-in-review-011215.html

Spann, R. (2020). An algorithm for abstract images. *Journal of Mathematics and the Arts*, *14*(1-2), 141-143. doi:10.1080/17513472.2020.1732804.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K., & Yang Wang, W. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, Florence, Italy.

Tibaut, A., Kaučič, B., Dvornik, P., Tiano, P., & Martins, J. (2018) Ontologizing the Heritage Building Domain. In: M. Ioannides, J. Martins, R. Žarnić, & V. Lim (Eds.), *Advances in Digital Cultural Heritage*. *Lecture Notes in Computer Science*, *vol 10754* (pp. 141-161). Springer, Cham.

Van Heddeghem, R. (2016). *Outsider art, In or Outside the World of Art? A study of the framing of the paradoxical position of outsider art*. [Master thesis]. Erasmus School of History, Culture and Communication, Erasmus University Rotterdam.

Wang A., Narayanan A. & Russakovsky O. (2020) REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In A. Vedaldi, H. Bischof, T. Brox, & J.M. Frahm. (Eds.) *Computer Vision – ECCV 2020. ECCV 2020*. *Lecture Notes in Computer Science*, *vol 12348*. Springer, Cham.

Winston, C. N., Tarkas, N. J., & Maher, H. (2014). Eccentric or egocentric? Preoperational features in schizotypic and creative adults. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 413-422.

Yang, M., Wang, M., Wang, H., Yang, G., & Liu, H. (2020). Exploring the Transdisciplinary Nature of Digital Humanities. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*.

Zeki S, Ishizu T. (2013). The "Visual Shock" of Francis Bacon: an essay in neuroesthetics. *Fronters in Human Neuroscience*, *7*(850).

Chapter II

La *Biblioteca Virtual de la Filología Española* (*BVFE*) y su acervo hispanoamericano[1]

The Biblioteca Virtual de la Filología Española (BVFE) and its Hispanic American heritage

Jaime Peña Arce & M.ª Ángeles García Aranda
Universidad Complutense de Madrid – España

A Manuel Alvar Ezquerra

Resumen: El objetivo de este capítulo es doble. Por un lado, se da a conocer la *Biblioteca Virtual de la Filología Española* (*BVFE*), un portal que recoge una gran cantidad de obras lingüísticas relacionadas con el español, a las que proporciona un acceso libre y gratuito. Por otro, se indaga en el componente hispanoamericano de su acervo, con el propósito de recapacitar sobre lo que ya se ha hecho y sobre lo que queda por hacer.

Abstract: The aim of this chapter is twofold. First, a presentation will be given of the *Biblioteca Virtual de la Filología Española* (*BVFE*), a portal that gathers numerous linguistic works on the Spanish language and provides free and open access to them. Secondly, it will examine the Latin American component of its heritage, with a view to reflecting on what has already been done and what remains to be done.

1 Este trabajo se enmarca en el Proyecto de Investigación "Biblioteca Virtual de la Filología Española. Fase lll: nuevas bibliotecas y nuevos registros. Información bibliográfica. Difusión de resultados" (FFl2017-82437-P), financiado por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España.

1. Introducción

Las páginas siguientes están dedicadas a mostrar la riqueza y utilidad de un recurso en línea a través de parte de sus materiales. Por un lado, nos sirven para presentar la *Biblioteca Virtual de la Filología Española* (a partir de ahora, *BVFE*), un portal que atesora un sinfín de títulos —diccionarios, gramáticas, diálogos, ortografías y otros textos de contenido lingüístico— relacionados con la lengua castellana,[2] muy reconocido ya entre la comunidad investigadora filológica a ambos lados del Atlántico. El *III Congreso Internacional de Lingüística Computacional y de Corpus (CILCC 2020) y V Workshop en Procesamiento Automatizado de Textos y Corpus (WoPATeC 2020)*, celebrado en la ciudad colombiana de Medellín entre el 21 y el 23 de octubre del 2020, nos dio la oportunidad de dar a conocer nuestra herramienta y sus recursos —aunque fuera de forma virtual, debido a las actuales condiciones de pandemia— a un amplio público, implicado en el estudio y en la descripción de la lengua de Cervantes, que aborda su trabajo desde las más variadas perspectivas que ofrece en la actualidad el panorama investigador.

Por otro lado, los miembros de este equipo de trabajo consideramos que dicho encuentro, organizado y amparado por la Universidad de Antioquia (en colaboración con la neerlandesa Rijksuniversiteit Groningen), podía ser un buen pretexto para examinar el corpus de obras y autores hispanoamericanos incluido dentro de la *BVFE*, un componente fundamental de nuestra herramienta. La construcción de este acervo es siempre una prioridad para nosotros, no en vano, el propio nombre de nuestro portal es un homenaje a una de las recopilaciones que más ha ayudado a los investigadores de Historiografía lingüística en el pasado, a saber, la *Biblioteca histórica de la filología castellana*, del Conde de la Viñaza (1978 [1893]), autor que también prestó una particular atención a la realidad lingüística del Nuevo Mundo en su *Bibliografía española de lenguas indígenas de América* (1892). Con estos antecedentes, el examen que contienen estas páginas resultaba más que obligado.

La metodología que hemos empleado para la elaboración de este trabajo, gracias a las variadas opciones de búsqueda que ofrece nuestro sitio web (http://www.bvfe.es), es bastante sencilla. A partir de los parámetros *autor*, *lugar de impresión*, *biblioteca* en la que se conservan los ejemplares físicos e *idioma*, mostraremos la importancia cuantitativa y cualitativa de la presencia hispanoamericana en la *BVFE*. Tanto el lugar de impresión como la biblioteca que atesora el ejemplar físico debían estar radicados en algún punto del continente hispanoamericano; los idiomas, además de los trasplantados desde Europa (español,

2 Los criterios seguidos para aceptar títulos dentro de nuestro portal son los siguientes: por un lado, se incluye toda obra de contenido lingüístico de cualquier autor español o natural de un país hispanohablante, con independencia del idioma que describa o estudie; por otro lado, se recoge todo texto que trate sobre la lengua española, al margen de la nacionalidad de su autor.

latín…), tenían ser los propios de las comunidades indígenas locales para ser tenidos en cuenta. Respecto a los autores, debían ser nacidos en algún rincón de los antiguos Virreinatos españoles y actuales estados soberanos o haber realizado en aquellas tierras la mayor parte de su actividad científica; esta última ponderación resultó imprescindible por la fuerte corriente migratoria que, con origen en la Península y destino a aquellas latitudes, ha existido en diferentes momentos de la historia.

La estructura de estas páginas está en consonancia con el doble propósito del que hablábamos más arriba. En primer lugar, vamos a realizar una presentación general de la *BVFE*, atendiendo a sus orígenes y trayectoria, para finalizar con la presentación de sus actuales datos de impacto. En segundo término, nos centraremos en el análisis de su componente hispanoamericano a partir de los parámetros ya señalados: lugar de impresión, biblioteca, idioma y autoría. Finalmente, se incluyen unas conclusiones que pretenden relacionar ambos bloques, con el objetivo de mostrar una perspectiva del acervo hispanoamericano de la *BVFE* lo más completa y contextualizada que sea posible y reflexionar sobre el camino a seguir en el futuro. El capítulo se cierra con el desarrollo de las referencias bibliográficas traídas a colación a lo largo de esta investigación.

2. La *Biblioteca Virtual de la Filología Española* (*BVFE*)

2.1. Orígenes

El origen de la *BVFE* hay que buscarlo en la idea que el profesor Manuel Alvar Ezquerra (1950-2020), de inolvidable memoria y uno de los investigadores más importantes que ha tenido la lengua española en fechas recientes, tuvo durante el segundo lustro del presente siglo. Su propósito inicial fue construir un catálogo que incluyera todos los repertorios lexicográficos del español[3] y, simultáneamente, crear una biblioteca virtual que ordenara los materiales disponibles en la red y garantizara su acceso de forma libre, gratuita y con las garantías de calidad de quien dedicó su vida al estudio de esta disciplina y de gran parte de sus títulos más importantes. Esa primera pretensión pronto se amplió y terminó dando cabida a cualquier obra de contenido lingüístico relacionada con nuestro idioma[4]. Así, tras varios años de esfuerzos, y

3 El trabajo que, a este respecto, se había realizado hasta aquel momento era bastante modesto. Además de obras de carácter más general (Esparza-Niederehe 1995, 1999 y 2005), solo existían una serie de aproximaciones sobre la dimensión de la producción lexicográfica de la lengua española realizadas en el ámbito académico italiano (Fabri 1979 y 2002; San Vicente 1995).

4 Si se quiere saber más sobre la historia de la *BVFE*, consúltese: Alvar Ezquerra y Miró Domínguez (2013), Calero Hernández, Fernández de Gobeo y Peña Arce (2018), Cazorla Vivas y García Aranda (2018) y García Aranda y Peña Arce (2019).

gracias al trabajo de los miembros del equipo y los colaboradores —junto a las ayudas públicas captadas—,[5] la *BVFE* se abrió al público como parte de la biblioteca de la Universidad Complutense de Madrid (en adelante, BUCM), https://webs.ucm.es/BUCM/nebrija/, en el año 2010. Desde entonces, la *BVFE* nos ha facilitado la investigación, pues los interesados en estas cuestiones tenemos acceso a numerosas obras sin tener que acudir a bibliotecas, sin tener que localizar ejemplares, sin tener que solicitar reproducciones y sin la necesidad de comparar catálogos, bibliografías y demás fuentes para comprobar si la información dada es fiable.

2.2. Desarrollo

El desarrollo, el crecimiento y la mejora que imponía la *BVFE* obligó a su cambio de ubicación, de manera que, desde el año 2014, nuestros materiales pueden consultarse en http://www.bvfe.es, página web que mantiene, desarrolla y edita la empresa especializada *Stílogo*.

Basta una comparativa cuantitativa para comprobar el trabajo llevado a cabo en este sentido en la *BVFE*:

Tabla 1. Comparativa del n.º de registros entre la *BUCM* y la *BVFE.*

BUCM (2010-2014)	www.bvfe.es (2014-2020)
• 2200 títulos lexicográficos	• 4638 obras lexicográficas. • 3641 gramáticas y tratados gramaticales. • 626 ortografías y prosodias. • 430 diálogos. **Total: 9335 registros**

El camino hasta llegar a la situación actual de la *BVFE* no ha sido fácil. Así, por ejemplo, los continuos cambios en las direcciones electrónicas de las obras digitalizadas obligan a una revisión permanente de los enlaces; el crecimiento exponencial de los libros digitalizados también supone, por las necesidades de actualización, un reto importante; los errores en las informaciones bibliográficas de los catálogos y las bibliotecas exigen una investigación concienzuda y la dificultad, por no decir la imposibilidad, de elaborar una lista completa y fiable de todas las obras lingüísticas del pasado nos obliga a replantearnos de forma constante nuestros objetivos y nuestra metodología de trabajo. A todos estos retos

5 Esta herramienta se ha beneficiado de tres planes de ayuda del Gobierno de España: "Creación y desarrollo de la BVFE" (FFI2011-24107), "Biblioteca Virtual de la Filología Española. Fase II. Consolidación, mejora y ampliación de los datos y de la web. Estudio de los materiales contenidos" (FFI2014-53851-P) y "Biblioteca Virtual de la Filología Española. Fase III: nuevas bibliotecas y nuevos registros. Información bibliográfica. Difusión de resultados" (FFI2017-82437-P).

y dificultades tratamos de buscar solución en nuestro quehacer cotidiano, en aras de la creación de un repositorio lo más completo que sea posible.

La *BVFE* facilita el acceso a obras lingüísticas seleccionadas a partir de una serie de criterios (en español, sobre el español, compuestas en otras lenguas por autores españoles, bilingües con el español, multilingües con el español) y que son integradas en un servidor diseñado para esta biblioteca virtual (autor, título, datos de edición/impresión, enlace, lenguas, notas, parte de otra obra…). La forma de trabajar es sencilla: se buscan las obras a partir de una serie de palabras clave en los catálogos de bibliotecas y repositorios para obtener los ejemplares de las obras lingüísticas digitalizados en ellos y se cargan en una base de datos específicamente diseñada para ello, donde se ponen todos esos datos, y un comentario o aclaración que puedan ser útiles al usuario. Cuando se han comprobado todas las informaciones (que son correctas, que no hay duplicaciones, etc.), los registros se depositan en el servidor. El usuario puede recuperar los datos de la *BVFE* a partir de una serie de búsquedas que realiza en la web a partir de una serie de parámetros:

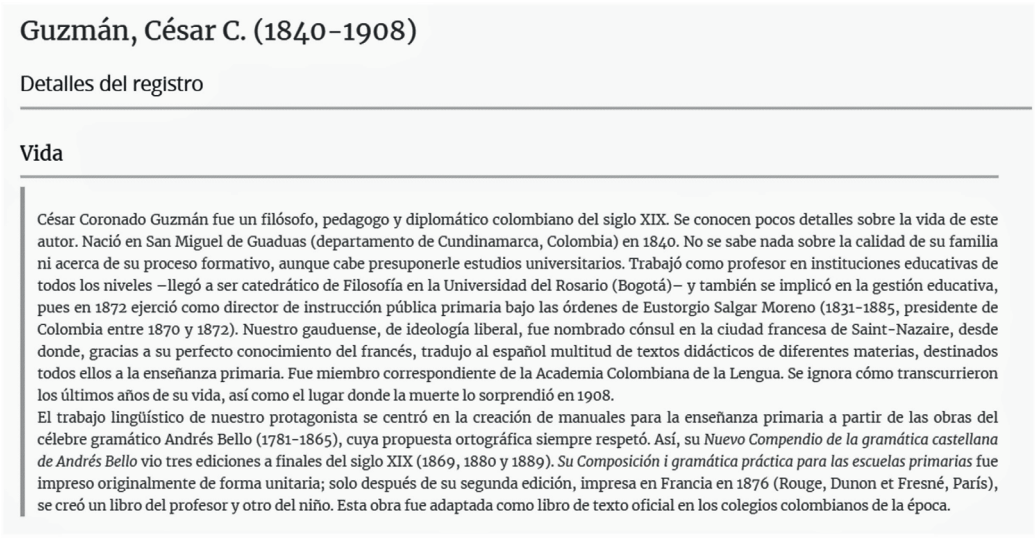
- En primer lugar, una búsqueda alfabética, seleccionando la letra inicial de la obra o tipo de texto que se desea localizar (gramática, tratado gramatical, ortografía, prosodia, nomenclatura, diccionario).
- En segundo lugar, una búsqueda sencilla en el buscador de la página principal, introduciendo el término de búsqueda.
- Y en tercer lugar, una búsqueda avanzada, en donde se pude filtrar por obra, fecha de publicación, impresor, lugar de impresión, lenguas de publicación, periodo cronológico, etc.

Y los resultados que arrojan estas búsquedas pueden, a su vez, ordenarse a partir de varios criterios, a saber: título ascendente/descendente, recientemente modificado, autor ascendente/descendente, fecha ascendente/descendente, impresor ascendente/descendente, lugar de impresión ascendente/descendente y biblioteca ascendente/descendente. Una vez finalizada la búsqueda y la ordenación, solo hay que pinchar en el título de la obra para acceder a los datos completos del registro (título, autor, ciudad y fecha de impresión, páginas que ocupa, procedencia del ejemplar digitalizado, signatura) y al ejemplar o a la ficha biobibliográfica del autor, de las que se habla en el párrafo siguiente.

La *BVFE* se sirve de discos de alta gama NVMe que mejoran considerablemente el rendimiento y la eficiencia de las conexiones gracias a la rapidez de lectura y al aumento de ancho de banda, lo que se aprecia en una navegación ligera y dinámica. Alexa, la aplicación sobre tráfico web, la sitúa en el *ranking* mundial (formado por más de 1800 millones de páginas web) en el puesto 2 421 083.

Para que la *BVFE* funcione correctamente son necesarios 1) un mantenimiento continuo del software, 2) actualizaciones periódicas, 3) controlar las defensas de los ataques de robots y mecanismos que desean acceder de forma ilícita a ella, 4) mejoras constantes de la interfaz (por ejemplo, con su traducción al inglés) y del motor de filtrado (parámetros incluyentes y excluyentes en las búsquedas avanzadas; filtros de ordenación "ascendente/descendente" de los resultados obtenidos en las búsquedas para todos los criterios utilizados…), auditorías de seguridad y optimización para evitar ralentizaciones y bloqueos.

Por otro lado, y desde la segunda fase o consolidación de la *BVFE*, esto es, desde finales del 2015, el corpus acopiado se ha enriquecido con la inclusión de las fichas biobibliográficas de los autores cuyas obras recogemos (actualmente, 1917). Estas fichas se estructuran así: 1.º) los datos biográficos del autor y una breve reseña de su producción, 2.º) la descripción de su obra lingüística, tanto de la incluida en la *BVFE* como de la que no se encuentra, 3.º) las principales referencias bibliográficas y 4.º) la firma del autor. En la actualidad, contamos con 911 fichas biobibliográficas, número que crece cada día gracias al trabajo de nuestros miembros y colaboradores. A continuación, se incluye un ejemplo del trabajo descrito en este párrafo:



Guzmán, César C. (1840-1908)

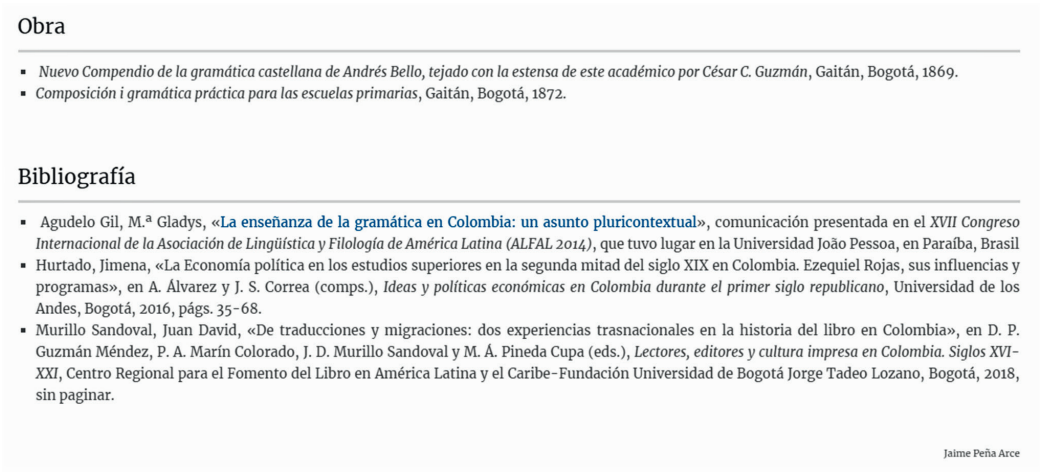
Detalles del registro

Vida

César Coronado Guzmán fue un filósofo, pedagogo y diplomático colombiano del siglo XIX. Se conocen pocos detalles sobre la vida de este autor. Nació en San Miguel de Guaduas (departamento de Cundinamarca, Colombia) en 1840. No se sabe nada sobre la calidad de su familia ni acerca de su proceso formativo, aunque cabe presuponerle estudios universitarios. Trabajó como profesor en instituciones educativas de todos los niveles –llegó a ser catedrático de Filosofía en la Universidad del Rosario (Bogotá)– y también se implicó en la gestión educativa, pues en 1872 ejerció como director de instrucción pública primaria bajo las órdenes de Eustorgio Salgar Moreno (1831-1885, presidente de Colombia entre 1870 y 1872). Nuestro gauduense, de ideología liberal, fue nombrado cónsul en la ciudad francesa de Saint-Nazaire, desde donde, gracias a su perfecto conocimiento del francés, tradujo al español multitud de textos didácticos de diferentes materias, destinados todos ellos a la enseñanza primaria. Fue miembro correspondiente de la Academia Colombiana de la Lengua. Se ignora cómo transcurrieron los últimos años de su vida, así como el lugar donde la muerte lo sorprendió en 1908.

El trabajo lingüístico de nuestro protagonista se centró en la creación de manuales para la enseñanza primaria a partir de las obras del célebre gramático Andrés Bello (1781-1865), cuya propuesta ortográfica siempre respetó. Así, su *Nuevo Compendio de la gramática castellana de Andrés Bello* vio tres ediciones a finales del siglo XIX (1869, 1880 y 1889). *Su Composición i gramática práctica para las escuelas primarias* fue impreso originalmente de forma unitaria; solo después de su segunda edición, impresa en Francia en 1876 (Rouge, Dunon et Fresné, París), se creó un libro del profesor y otro del niño. Esta obra fue adaptada como libro de texto oficial en los colegios colombianos de la época.

Figura 1. Ejemplo de una ficha biobibliográfica: vida de César Guzmán (Alvar, 2020).



Obra

- *Nuevo Compendio de la gramática castellana de Andrés Bello, tejado con la estensa de este académico por César C. Guzmán*, Gaitán, Bogotá, 1869.
- *Composición i gramática práctica para las escuelas primarias*, Gaitán, Bogotá, 1872.

Bibliografía

- Agudelo Gil, M.ª Gladys, «La enseñanza de la gramática en Colombia: un asunto pluricontextual», comunicación presentada en el *XVII Congreso Internacional de la Asociación de Lingüística y Filología de América Latina (ALFAL 2014)*, que tuvo lugar en la Universidad João Pessoa, en Paraíba, Brasil
- Hurtado, Jimena, «La Economía política en los estudios superiores en la segunda mitad del siglo XIX en Colombia. Ezequiel Rojas, sus influencias y programas», en A. Álvarez y J. S. Correa (comps.), *Ideas y políticas económicas en Colombia durante el primer siglo republicano*, Universidad de los Andes, Bogotá, 2016, págs. 35-68.
- Murillo Sandoval, Juan David, «De traducciones y migraciones: dos experiencias trasnacionales en la historia del libro en Colombia», en D. P. Guzmán Méndez, P. A. Marín Colorado, J. D. Murillo Sandoval y M. Á. Pineda Cupa (eds.), *Lectores, editores y cultura impresa en Colombia. Siglos XVI-XXI*, Centro Regional para el Fomento del Libro en América Latina y el Caribe-Fundación Universidad de Bogotá Jorge Tadeo Lozano, Bogotá, 2018, sin paginar.

Jaime Peña Arce

Figura 2. Ejemplo de una ficha biobibliográfica: obra de César Guzmán y referencias (Alvar, 2020).

2.3. Datos actuales

En la *BVFE* pueden consultarse registros digitalizados de la mayoría de las bibliotecas y repositorios españoles, europeos y extranjeros. Se han escrutado los catálogos de más de 200 instituciones. Dentro de nuestra colección priman los títulos atesorados en diferentes bibliotecas —físicas o virtuales— de España: ya pertenezcan a la administración general del estado (la Biblioteca Nacional de España, la *Biblioteca Virtual del Patrimonio Bibliográfico*, *Hispana*, la Universidad Nacional de Educación a Distancia, la Real Academia Española o las bibliotecas públicas estatales de las diferentes capitales provinciales), ya a las diferentes comunidades autónomas (Biblioteca de Catalunya, Biblioteca Valenciana, *Biblioteca Virtual de Andalucía*, *Biblioteca Virtual de Castilla y León*…) o a sus universidades (Complutense de Madrid, Salamanca, Zaragoza, Sevilla, Granada, Barcelona, Valencia, Santiago de Compostela…), ya a colecciones privadas (Fundación Sancho el Sabio, en Vitoria, o Fundación Sierra Pambley, en León) o municipales (Biblioteca Histórica Municipal, en Madrid).

También contamos con las aportaciones de las bibliotecas nacionales más importantes de Europa (la *Bibliothèque Nationale de France*, la *British Library*, *Bayerische Staatsbibliothek* de Múnich, la *Österreichische Nationalbibliothek* de Viena, la *Národní Knihovna České Republiky* de Praga o las bibliotecas nacionales italianas de Florencia, Roma y Nápoles) y del mundo (*Library of Congress*, en Washington, la Biblioteca Nacional de Colombia, la Biblioteca Nacional de Chile…). Asimismo, hemos incorporado los registros pertinentes de las principales bibliotecas universitarias de Europa (*Oxford University*, *Cambridge University*, *Universiteitsbibliotheek Gent*, *Université de Toulouse*, *Università degli Studi di Roma*

"La Sapienza"…), de los Estados Unidos (*Harvard University*, *University of Michigan*, *University of California*, *The John Carter Brown Library*, *Brown University*, *Columbia University*…), Canadá (*University of Toronto*), Hispanoamérica (Universidad Autónoma de Nuevo León, Universidad Nacional Autónoma de México…) o Australia (*La Trobe University*). Igualmente, recogemos las referencias depositadas en los más importantes repositorios virtuales, como *Google books* o *Archive*. En definitiva, estamos en condiciones de presumir de nuestro completo acervo, que recoge obras custodiadas por instituciones que van desde las más modestas, como el Instituto de Enseñanza Secundaria Alfonso X el Sabio, en Murcia, hasta las de primer nivel, como la *New York Public Library*.

Más interesante es, si cabe, el balance que arroja la comparativa, en cuanto a número de visitantes, de los últimos años. Cifras que evidencian el interés y la confianza de los usuarios por la *BVFE*[6]:

Tabla 2. Datos de impacto de la *BVFE.*

Año	2018	2019	2020
N.º total de visitas	126 872	210 548	397 681
Visitantes diarios distintos	69 004	81 255	197 025
Páginas vistas	1 043 598	7 815 384	8 388 692

España lidera la lista de países con mayor número de páginas vistas en estos años, seguida por los Estados Unidos, México, Francia, Alemania, Ecuador, Colombia, Argentina, Italia, Perú y Panamá. Cantidades que se convierten en un reto para seguir trabajando por la mejora y el crecimiento constantes de la *BVFE*. En cuanto a la posición de la *BVFE* en los resultados de búsquedas de Google Search, suele ocupar los primeros puestos al indagar sobre *diccionarios de metáforas, palabras en rifeño, diccionario mallorquín-castellano, diccionario menorquín, diccionario de andalucismos, vocabulario quirúrgico, gramática analítica* o *diálogos españoles* o al tratar de averiguar los datos biográficos de Ambrosio Calepino, Vicente Salvá, Esteban Pichardo, Carlos Felipe Beltrán, Pedro Marbán o Francisco de Paula Mellado.

De todo ello, tanto de los nuevos registros como de las biografías de los autores y de las novedades en la web, damos puntual cuenta cada final de mes con un boletín de novedades al que cualquiera puede suscribirse desde la página de la *BVFE*.

6 En los primeros meses de 2021, fecha en la que se escribe este trabajo, el número de visitas a páginas de la *BVFE* asciende a 38 011.

3. El componente hispanoamericano de la *BVFE*

En los siguientes epígrafes vamos a descomponer el acervo hispanoamericano contenido en nuestro portal. Tal como anunciamos al inicio del capítulo, el orden en el que se va a llevar a cabo el estudio es este: lugar de impresión, biblioteca, idioma y autoría.

3.1. Lugares de impresión

Más de 1000 ejemplares de los incluidos en la *BVFE* han sido impresos en imprentas hispanoamericanas (un 11.60 % del total). La llegada de la imprenta a los virreinatos de la Nueva España y del Perú en época temprana (después llegaría a la Nueva Granada y al Río de la Plata) y su desarrollo posterior en todo el continente explican esta cifra.

Tabla 3. Registros de la *BVFE* impresos en Hispanoamérica.

Totales	En Hispanoamérica
9335	1083 (11.60 %)

Los primeros textos impresos en estos talleres se deben a la labor de descripción realizada por los misioneros sobre las lenguas amerindias. El *Vocabulario en la lengua castellana y mexicana* de Alonso de Molina y el *Arte de la lengua de Michuacán* de Maturino Gilberti en el taller de Juan Pablos (1555 y 1558); el *Arte en lengua zapoteca* de Juan de Córdova, el *Arte en lengua mixteca* de Antonio de los Reyes y el *Vocabulario en lengua misteca* de Francisco de Alvarado en la imprenta de Pedro Balli (1578, 1593), o el *Vocabulario manual de las lenguas castellana y mexicana* de Pedro de Arenas en la imprenta de Henrico Martínez (1611) son buena muestra de la actividad en México. Por otro lado, el *Arte y vocabulario en la lengua general del Perú llamada quichua* de Alonso de Bárcena en el taller de Antonio Ricardo (1586), el *Arte y gramática general de la lengua que corre en todo el reyno de Chile* de Luis de Valdivia y la *Gramática y arte nueua de la lengua general de todo el Perú* de Diego González Holguín en la imprenta de Francisco del Canto (1606 y 1607) ilustran las producciones textuales limeñas.

Pero en estos primeros siglos no solo se publicaron obras misioneras, también hubo tiempo, dinero y dedicación para, entre otros, los *Discursos de la antigüedad de la lengua cántabra vascongada* de Balthasar Echave (México, Henrico Martínez, 1607) o para la *Ortografía castellana* de Mateo Alemán (México, Jerónimo Balli, 1609).

Ahora bien, el siglo que más resultados de impresiones hispanoamericanas proporciona es el XIX. Durante esta centuria se publicaron en México, Chile, Perú, Argentina y Colombia numerosas obras lingüísticas que testimonian la riqueza y el interés del periodo

para la Historia de la lingüística, pues entre ellas se pueden encontrar aportaciones a diferentes disciplinas lingüísticas (semántica, sociolingüística, dialectología, gramática, lexicografía, ortografía, traducción, enseñanza de la lengua o lingüística misionera) desde otras tantas perspectivas, metodologías y corrientes teóricas (tradicional, normativa, racionalista, general, lógica, historicista, didáctica…), lo que resulta una innegable contribución para la historia de la lengua española. Sirvan como muestra las que se citan a continuación:

- Diálogos (*Diálogos de Juan Luis Vives, traducidos en lengua castellana por el doctor Cristóbal Coret y Peris*, México, 1827).
- Ortografías y ortologías (*De la ortografía* México, 1847; *Ortografía española acomodada a la pronunciación megicana* México, 1851; *Principios de la ortología y métrica de la lengua castellana*, Santiago de Chile, 1835; *Acentuaciones viciosas*, Santiago de Chile, 1887; *Neógrafos kontemporáneos*, Santiago de Chile, 1896; *Ortografía fonética*, Santiago de Chile, 1897; *Ortografía castellana americana*, Buenos Aires, 1876; *Enseñanza de la lectura y la logografía. Instrucciones para los maestros*, Buenos Aires, 1887).
- Silabarios (*Silabario de idioma mexicano* México, 1849; *Silabario de idioma mexicano,* México, 1883).
- Repertorios lexicográficos (*Nuevo vocabulario filosófico-democrático*, México, 1834; *Diccionario de sinónimos castellanos* México, 1845; *Manual de voces equívocas sacadas del Diccionario de la lengua castellana* México, 1848; *Vocabulario del idioma comanche*, México, 1866; *Diccionario etimológico de la lengua castellana (ensayo)*, México, 1877; *Diccionario de dudas ortográficas formado con arreglo al último de la Real Academia*, México, 1881; *Diccionario de mejicanismos*, México, 1898; *Diccionario para el pueblo, republicano, democrático, moral, político y filosófico*, Lima, 1855; *Neologismos y americanismos*, Lima, 1896; *Diccionario hispano chileno*, Santiago de Chile, 1846; *Diccionario de chilenismos*, Santiago de Chile, 1875; *Diccionario filológico-comparado de la lengua castellana*, Buenos Aires, 1882; *El lenguaje gauchesco*, Buenos Aires, 1894; *Minucias lexicográficas. Tata, tambo, poncho, chiripá, etc*., Buenos Aires, 1896; *La religión en el idioma. Ensayo paremiológico*, Buenos Aires, 1899).
- Gramáticas (*Elementos de gramática castellana para el uso de las escuelas* México, 1843; *Arte del idioma othomí,* México, 1863; *Compendio de gramática de la lengua española, según se habla en Méjico* México, 1867; *Epítome de la gramática de la lengua castellana,* México, 1873; *Gramática de la lengua castellana, compuesta por la Real Academia Española,* México, 1877; *Estudios gramaticales sobre el "náhuatl",* México, 1887; *Compendio de la gramática castellana para el uso de las escuelas de primeras letras del Perú*, Lima, 1836; *Gramática de la lengua castellana*, Lima, 1872; Gramática latina, Santiago de Chile,

1831; *Gramática de la lengua chilena*, Santiago de Chile, 1846; *Gramática de la lengua castellana destinada al uso de los americanos*, Santiago de Chile, 1847; *Borrones gramaticales*, Santiago de Chile, 1894; *Gramera berria*, Buenos Aires, 1860; *Arte de la lengua lule y toconoté*, Buenos Aires, 1877).

- Métodos de enseñanza de segundas lenguas (*Novísima gramática francesa,* México, 1863; *La clave del francés*, México, 1886; *El maestro de inglés*, Lima, 1891; *Lecciones de gramática francesa*, Santiago de Chile, 1829).

En Colombia, país en el que se funda en 1894 la Imprenta Nacional en los talleres de los afamados Echavarría Hermanos,[7] se imprimieron, entre otros muchos, unos *Elementos de la gramática castellana y ortografía* (1825), la *Gramática y ortografía de la lengua castellana para uso de los niños en las escuelas de primeras letras del Departamento del Cauca* (1826), *La ortografía fijada en la Nueva Granada. Método perfeccionado de enseñanza para las primeras letras* (1833), *Nuevo epítome de gramática castellana* (1843), *Observaciones curiosas sobre lengua castellana* (1848), *Prontuario de ortografía de la lengua castellana* (1850), *Salvá reformado* (1850), *Diccionario ortográfico* (1867), *Apuntaciones críticas sobre el lenguaje bogotano* (1867-1872), *Gramática de la lengua latina para el uso de los que hablan castellano* (1869), *Análisis ideolójica de los tiempos de la conjugación castellana* (1872), *Gramática de la lengua castellana destinada al uso de los americanos* (1874) o *Ensayo de gramática hispano-goahiva* (1895).

Las razones expuestas explican que el país hispanoamericano que más textos suministra a la *BVFE* sea México, seguido de Chile, Perú, Argentina, Colombia y Costa Rica:[8]

7 Antes de esa fecha existían los talleres de Antonio Espinosa, de Salazar, de José A. Cuella, N. Gómez, de Francisco Torres Amaya, Arnulfo Guarín, Foción Mantilla, la Imprenta de El Día, la Imprenta del Neogranadino, Imprenta del Tradicionalista, entre otros.

8 Por ciudades, la distribución es la siguiente: Aguascalientes 1, Bogotá 71, Buenos Aires 92, Caracas 19, Cartagena de Indias 3, Chiapas 8, Concepción (Chile) 4, Córdoba (Argentina) 2, Cuenca (Ecuador) 1, Cuernavaca (México) 7, Cuzco 8, Guadalajara 13, Guanajuato 3, Guatemala 2, Habana/La Habana 28, Iquitos 1, La Paz 2, La Plata 8, La Victoria (Venezuela) 2, León (México) 1, Lima/Ciudad de los Reyes/Los Reyes 126, Matanzas 6, Medellín 3, México/Méjico/México D. F. 383, Mérida de Yucatán 26, Monterrey 2, Montevideo 9, Morelia 11, Oaxaca 6, Panamá 3, Ponce (Puerto Rico) 3, Puebla/Puebla de los Ángeles 23, Quito 3, Puerto Rico/San Juan de Puerto Rico 5, San Cristóbal de las Casas (México) 3, San José de Costa Rica 28, San Juan de los Lagos (México) 2, Santa Fe del Río (México) 1, Santiago/Santiago de Chile 124, Santiago de Cuba 1, Salta 1, Sucre/Chuquisaca 5, Tegucigalpa 4, Toluca (México) 1, Valdivia 2, Valparaíso 12, Veracruz 1, Zacatecas 2.

Tabla 4. Registros hispanoamericanos de la *BVFE* por países y ciudades (en %).

País y ciudad	Porcentajes %
México	45 %
Ciudad de México	35 %
Chile	13 %
Santiago de Chile	11.5 %
Perú	12.5 %
Lima	11.5 %
Argentina	8.7 %
Buenos Aires	8.4 %
Colombia	6.8 %
Bogotá	6.5 %
Costa Rica	2.3 %

3.2. Bibliotecas

En cuanto a las bibliotecas en que se localizan los ejemplares de la *BVFE* hay que destacar la Biblioteca Nacional de Colombia (https://bibliotecanacional.gov.co/es-co). Fundada en 1777 con una colección de los padres jesuitas expulsados de España, hoy constituye el fondo nacional hispanoamericano más importante para nuestro portal. Su página web, cómoda y sencilla; sus múltiples servicios para atender a todos los usuarios, y sus varias colecciones temáticas (corográfica, botánica, fondos especiales, bibliotecas digitales de autor, fondos gráficos, prensa del siglo XIX y hemeroteca digital) la convierten en un recurso útil y completo. Tras él se encuentran los fondos nacional y general de México (que están albergados en la Universidad Nacional Autónoma de México, https://www.bidi.unam.mx/), la Universidad Autónoma de Nuevo León, https://www.dgb.uanl.mx/?mod=bdigital, y la Biblioteca Nacional de Chile (https://www.bibliotecanacional.gob.cl/). Muy por detrás se encuentran los fondos de Costa Rica (http://www.sinabi.go.cr/bibliotecas/biblioteca_nacional.aspx), Argentina (https://www.bn.gov.ar), Perú (https://www.bnp.gob.pe) y Guatemala (http://mcd.gob.gt/biblioteca-nacional/).

Tabla 5. Registros de la *BVFE* en bibliotecas de Hispanoamérica.

Totales	En bibliotecas hispanoamericanas
9335	539 (5.77 %)[9]

9 Los ejemplares digitalizados en bibliotecas hispanoamericanas son algunos más, pero todavía no están cargados en la web de la bvfe , pues están a falta de un estudio detallado de sus contenidos.

En la actualidad, y esperamos que sea una realidad que se subsane lo antes posible, la cantidad de digitalizaciones de obras procedentes de bibliotecas de este hemisferio es notablemente inferior al de otros territorios, como Europa o América del Norte. Por este motivo, las cifras ofrecidas en este epígrafe son más un motivo de reflexión que algo realmente orientativo. Sea como fuere, los datos desglosados por bibliotecas son los que siguen:

Tabla 6. Registros de la *BVFE* en bibliotecas de Hispanoamérica (desglose).

Biblioteca y número de ejemplares	
Biblioteca Nacional de Colombia	192
Universidad Nacional Autónoma de México	118
Universidad Autónoma de Nuevo León	103
Biblioteca Nacional de Chile	60
Biblioteca Nacional Miguel Obregón Lizano, Costa Rica	19
Biblioteca Pública del Estado "Juan José Arreola", Guadalajara	15
Biblioteca Nacional Mariano Moreno de la República Argentina	10
Biblioteca Nacional de Maestros, Buenos Aires	7
Biblioteca Nacional del Perú	4
Biblioteca Palafoxiana, Puebla	4
Universidad de Chile	4
El Colegio de México	1
Universidad Francisco Marroquín, Guatemala	1
Universidad Nacional de Colombia	1

3.3. Lenguas amerindias

Una de las mayores riquezas de la *BVFE* es el número y variedad de lenguas que atesora. De las más de 230 lenguas que están presentes en la *BVFE*, 110 se hablan o se han hablado en territorio hispanoamericano, y con ellas se han compuesto 1007 obras, esto es, un 10.78 % del total de registros.

Tabla 7. Registros de la *BVFE* de lenguas amerindias.

Totales	En bibliotecas hispanoamericanas
9335	1007 (10.78 %)[10]

10 Los ejemplares digitalizados en bibliotecas hispanoamericanas son algunos más, pero todavía no están cargados en la web de la *BVFE* , pues están a falta de un estudio detallado de sus contenidos.

Destacan, en este sentido, los textos compuestos en náhuatl (135), quechua (79), mapuche (57), otomí (57), maya (35), tarasco (33), michoacano (28), cachi (28), purépecha (27), cachiquel (26), guaraní (26), quiché (24), zapoteco (24), cahíta (21) y cabécar (20)[11], que en su mayoría se utilizaron para componer textos correspondientes a la Lingüística misionera. En la *BVFE* no solo contamos con trabajos descriptivos sobre las lenguas amerindias mayoritarias, también atesoramos diccionarios y gramáticas sobre idiomas muy minoritarios, como, por ejemplo, el cuna (*Vocabulario castellano-cuna*, de A. L. Pinart, publicado en 1890), la lengua propia de un pueblo que habita entre Panamá y Colombia, o el ixil (*Arte y vocabulario de la lengua ixil*, anónimo, post 1935), empleada en el noroeste del altiplano guatemalteco y perteneciente al tronco mayense.

3.4. Autores y época

El último parámetro manejado para describir el componente hispanoamericano en la *BVFE* es el de autores o fichas biobibliográficas. En este apartado se ha incluido a) autores cuyas sus obras traten sobre lenguas amerindias; b) autores, con independencia de su lugar de nacimiento, cuyas obras fueran imprimidas en ese continente, y c) autores nacidos en América, con independencia de la temática de sus obras. La *BVFE* cuenta con un total de 1917 autores, de los que 911 cuenta actualmente con una ficha biobibliográfica; de ellos, 202 cumplen los criterios antes mencionados (un 22 % sobre el total de autores ya estudiados).

Tabla 8. Autores hispanoamericanos en la *BVFE* .

Registros totales (autores)	Registros con ficha	Fichas de autores hispanoamericanos
1917	911	202

11 Alfabéticamente, las lenguas que han aportado registros a las *BVFE* son: achagua 2, aimara 18, allentiac 12, arasairi 1, atacameño 5, ayook 1, baure 3, biceita/viceyta 2, boruca 1, bribri 1, cabécar 20, cachi 28, cachiquel 26, cahíta 21, campa 3, caviñeno 2, chaima 3, chanabal 3, chiapaneca 3, chibcha 19, chilote 1, chinanteco 1, chinchaisuyo 1, chiquito 5, chirripó 1, chol 6, choltí 5, coa 3, comanche 1, cora 3, cumanagoto 4, cuna 1, eudeve 1, guahibo 5, guaraní 26, guatuso 1, guaymíe 1, hegue 1, huasteco 14, ixil 1, kunza 1, lean 1, lenguas de México 23, lule 18, machiguenga 1, mam 12, mame 5, mapuche 57, matlatzinca 4, maya 35, mazahua 1, mazateco 1, michoacano 28, mixe 2, mixteco 10, mochica 1, mojo 19, morocosi 1, muisca 16, mulía 1, mutsun 7, nahua 135, névome 3, ópata 1, orosí 1, otomí 57, páez 2, pame 1, paria 2, pima 3, pocoman 8, pocomchí 4, popoluca 1, purépecha 27, quechua 79, quekchí 1, quiché 24, rusien (Canadá) 1, sáliba 1, setevo 1, siona 1, sipibo 1, subinha 1, tacana 2, talamanca 1, tarahumara 6, tarasco 33, tatché 2, telamé 2, tepehuán 1, tepeguano 1, térraba 2, timucua 1, toba 2, totonacalpa 1, totonaco 1, tucurrique 1, tupí 17, tzeltal 7, tzendal 7, tzotzil 4, tzutuhil 2, yaqui 1, yook 1, yunga 5, yupa 1, zapoteco 24, zend 9, zoque 8, zutunil 4.

En cuanto a la época en que estos autores desarrollaron su actividad, los datos reflejan, como era de esperar, un continuo crecimiento a medida que pasa el tiempo. De los ocho autores del siglo XVI se pasa a los 72 del siglo XIX.

Tabla 9. Autores y siglos.

Siglos	Número de autores
Siglo XVI	8
Siglos XVI-XVII	10
Siglo XVII	19
Siglos XVII-XVIII	4
Siglos XVIII	21
Siglos XVIII-XIX	7
Siglo XIX	72
Siglos XIX-XX	47
Siglo XX	14

En las primeras centurias destaca la presencia de misioneros de diferentes órdenes religiosas, mientras que en las últimas los protagonistas son prestigiosos lingüistas internacionales que desarrollaron su labor docente e investigadora o publicaron sus textos en sus países natales[12].

12 Algunos de los autores con fichas biobibliográficas y de los que, por tanto, tenemos noticia son alfabéticamente: Abeille, Luciano. XIX-XX; Ágreda, Antonio de. XVIII; Agüero, Cristóbal de, O. P. XVII; Aguilera Patiño, Luisita, XX; Alvarado, Francisco de, O. P. XVI-XVII; Amunátegui Aldunate, Miguel Luis. XIX; Anchorena, José Dionisio. XIX; Arenas, Pedro de. XVII; Arias de la Vega, Eusebio. XIX; Armentia, fr. Nicolás. XIX; Arroyo, Santiago. XVIII-XIX; Aza, José Pío, O. P. XIX-XX; Baralt, Luis A. XIX-XX; Bárcena, Alonso de, S. I. XVI; Basalenque, Diego, O. S. A. XVII; Bayo, Ciro. XIX-XX; Bello, Andrés. XIX; Belmar, Francisco. XIX-XX; Beltrán de Santa Rosa María, Pedro, O. F. M. XVIII; Bertonio, Ludovico, S. I. XVI XVII; Botello Movellán, José Ceferino. XVIII; C. F. B. XIX; Caballero, Darío Julio. XIX; Cáceres, José María. XIX; Caro, Miguel Antonio. XIX; Carochi, Horacio, S. I. XVII; Carricaburu, Alfredo. XIX; Chimalpopocatl Galicia, Faustino. XIX-XX; Chomé, Ignace, S. I. XVIII; Ciudad Real, Antonio de, O. F. M. XVI-XVII; Company Company, Concepción. XX; Conto, César. XIX; Córdova, Juan de, O. P. XVI; Cuervo, Rufino José, XIX; Dávila Garibi, José Ignacio Paulino. XX; Espinosa, Juan. XIX; Febrés, Andrés, S. I. XVIII; Fernández Garfias, Pedro. XIX; Flores, Ildefonso José, O. F. M. XVIII; Franco, José Félix. XIX; Frías, Heriberto. XIX; Fuentes, Ventura y Victor E. François. XIX-XX; Galván, Mariano. XIX; Gárate Arriola, Justo. XX; García del Río, Juan. XIX; Gilberti, Maturino, O. F. M. XVI; Gómez de la Maza, Manuel. XIX-XX; González del Valle, Manuel. XIX; González Holguín, Diego, S. I. XVI-XVII; Guerra, Juan, O. F. M. XVII; Gutiérrez, Rafael. XIX; Henríquez Ureña, Pedro. XX; Herranz y Quirós, Diego Narciso. XVIII-XIX; Huerta, Alonso de. XVI-XVII; Lemos Ramírez, Gustavo. XVIII-XIX; León, Nicolás. XIX-XX; Limardo, Ricardo Ovidio. XIX; López Yepes, Joaquín, O. F. M. XIX; Lugo, Bernardo de, O. P. XVII; Machoni de Cerdeña, Antonio, S. I. XVII-XVIII; Magdalena, Agustín de la, O. F. M. XVIII; Marroquín, José Manuel. XIX; Matto de Turner, Clorinda. XIX; Membreño, Alberto. XIX-XX; Meneses y Gómez, Sabas. XIX; Mesías, José Mercedes. XIX; Mossi, Miguel Ángel. XIX; Navarro, Manuel, O. F. M. XIX-XX; Neve y Molina, Luis de, O. F. M. XVIII; Obelar, Raimundo D. XIX-XX; Oroz, Rodolfo. XX; Pareja, Francisco, O. F. M. XVI-XVII; Peñafiel, Antonio. XIX-XX; Pichardo y Tapia, Esteban. XIX; Pinart, Alphonse Louis. XIX; Pinilla, Norberto. XX; Ponce de León, Néstor. XIX; Quesada, Ernesto. XIX-XX; Rabanales O., Ambrosio. XX; Restrepo, Félix, S. I. XX; Reyes, Antonio de los, O. P. XVI; Reyes, Rincón, Antonio del, S. I. XVI; Rivera, Gregorio. XVIII-XIX; Rivodó, Baldomero. XIX; Rojas, Arístides. XIX; Rojo Mejía y Ocón, Juan. XVII; Rosales, Carlos Joseph, O. F. M. XVIII; Ruz, Joaquín, O. F. M. XVIII-XIX; San Buenaventura, Gabriel de, O. F. M. XVII; Sarmiento, Domingo Faustino. XIX; Suárez, José Bernardo. XIX-XX; Suárez, José Bernardo. XIX; Tangol, Nicasio. XX; Tellechea, Miguel, O. F. M. XVIII-XIX; Thiel, Bernardo Augusto, C. M. XIX; Torres Rubio, Diego de, S. I. XVI-XVII; Torresano, fr. Estevan. XVIII; Uribe Uribe, Rafael. XIX-XX; Uricoechea, Ezequiel. XIX; Valdivia, Luis de, S. I. XVI-XVII; Velarde, Fernando. XIX; Vetancurt, Agustín de, O. F. M. XVII; Vico, Domingo de, O. P. XVI; Vicuña Cifuentes, Julio. XIX-XX; Villarreal,

Entre los primeros, ante la imposibilidad de nombrarlos a todos, queremos citar a Arenas, Ciudad Real, Córdova, Flores, Gilberti, González Holguín, Neve y Molina, Rosales o Vico; entre los segundos, a Bello, Cuervo, Caro, García del Río, Gómez de la Cortina, Marroquín, Obelar, Rojas, Sarmiento o a Lenz, quien cuenta con treinta registros en la *BVFE* actualizados los pasados meses (*La oración y sus partes*, estudios sobre el español de Chile, reflexiones sobre fonética y ortografía, *¿Para qué estudiamos gramática?* o el papiamento).

4. Conclusiones

La *BVFE* es un proyecto consolidado, al que avalan sus once años de trayectoria, y líder en su ámbito, como atestiguan los datos sobre el número total de visitas o sobre las páginas visitadas. Su aportación a la sociedad del conocimiento fue reconocida el pasado mes de septiembre con la concesión del primer premio de la V Edición de los Premios de Transferencia de Tecnología y de Conocimiento de la Universidad Complutense de Madrid (2020). Resulta justo decir que esta herramienta, nacida y desarrollada en España, no podría entenderse sin el componente hispanoamericano, al igual que le sucede al idioma que compartimos. Para corroborar esta afirmación, solo hace falta traer a colación unos cuantos datos que ya han sido apuntados más arriba:

En primer lugar, según el lugar de impresión de las obras, ese componente hispanoamericano alcanza al 11.60 % de nuestros registros. Dentro de los territorios de la América hispana destaca, respecto a la cuestión que nos ocupa, la zona septentrional de Mesoamérica, ocupada en el pasado por el Virreinato de la Nueva España y, desde comienzos del siglo XIX, por los Estados Unidos Mexicanos. Y en el seno del país azteca, brillan con luz propia las prensas de la Ciudad de México, antaño capital del más importante virreinato del Nuevo Mundo y hoy del país con el mayor número de hispanohablantes del orbe. A continuación, y justo por encima del otro gran reino de las Indias españolas, el del Perú con capital en Lima, ocupan un lugar destacado las prensas chilenas —y, en particular, las santiaguinas—; realidad que se justifica por el proceso de digitalización de documentos llevado a cabo por las instituciones culturales de ese país del cono de Sudamérica.

En segundo lugar, si hablamos del porcentaje de registros cuyo ejemplar físico correspondiente se custodia en una biblioteca de ese continente, el porcentaje asciende al 5.77 %. En este sentido y junto a la última alusión del párrafo anterior, cabe destacar en trabajo de la Biblioteca Nacional de Colombia y el de dos de las principales instituciones mexicanas

Federico. XIX-XX; Vingut, Francisco Javier. XIX; Vivero, Luis Fernando. XIX; Ybarra, Alejandro. XIX-XX; Zambrano Bonilla, José. XVIII.

de educación superior, la Universidad Nacional Autónoma de México y la Universidad Nacional Autónoma de Nuevo León.

En tercer lugar, un 11 % de nuestros diccionarios, gramáticas u ortografías profundizan en el estudio y la descripción de alguna lengua amerindia. Y, como ha quedado dicho, no solo de las más extendidas, sino también de algunas de las más desconocidas. El quehacer de los lingüistas misioneros fue especialmente fructífero en las áreas de los grandes virreinatos históricos: Nueva España —náhuatl (135), otomí (57), maya (35), tarasco (33) o michoacano— y Perú —quechua (79)—. El papel algo sobredimensionado de las lenguas amerindias chilenas —mapuche (57)— se debe a la ya mencionada (y muy completa) digitalización de las obras custodiadas en la Biblioteca Nacional de Chile.

En cuarto lugar y para terminar, del total de autores ya estudiados y que poseen su ficha biobibliográfica, un 22 % proceden de esta región del planeta. Temporalmente hablando, y tal como ocurre con el conjunto de registros de nuestro portal (García y Peña, 2019, 126-130), una mayoría de ellos pueden radicarse en el siglo XIX y durante el primer tercio del XX. En este sentido, ese porcentaje se debe, fundamentalmente, a los trabajos realizados por alguno de los miembros del equipo de investigación, como Jaime Peña Arce y Leticia González Corrales, o por alguno de nuestros colaboradores, como Darío Rojas, Susana Serra Sepúlveda, Érika Moreno o Viviana Ávila.

En definitiva, la *Biblioteca Virtual de la Filología Española* pretende dar soporte a cualquier investigador, con independencia del lado del Atlántico en el que viva, y acercarle aquellos materiales que, geográficamente, le queden más alejados. El objetivo final es seguir trabajando juntos por el estudio y el cuidado de la lengua española y de todos aquellos autores que han ayudado a engrandecerla.

Referencias

Alvar Ezquerra, M. (2020). *Biblioteca Virtual de la Filología Española (BVFE): directorio bibliográfico de gramáticas, diccionarios, obras de ortografía, ortología, prosodia, métrica, diálogos e historia de la lengua*. [Consulta: 10/10/2020]. https://www.bvfe.es/es/.

Alvar Ezquerra, M. y Miró Domínguez, A. (2013). Antecedentes y primeros pasos de la Biblioteca Virtual de la Filología Española. En P. Spinato, P. Bruschi, & J. J. Martínez (Eds.), *Cuando quiero hallar las voces, encuentro los afectos. Studi di Iberistica offerti a Giuseppe Bellini* (pp. 49-60). Consiglio Nazionale delle Ricerche.

Cazorla Vivas, M.ª C. y García Aranda, M.ª Á. (2018). Herramientas en red: la Biblioteca Virtual de la Filología Española. *E-Scripta Romanica*, *5*, 12-27.

Calero, E., Fernández, N. y Peña, J. (2018). La *Biblioteca Virtual de la Filología Española* (*BVFE*) y la digitalización de obras complutenses del siglo XVI. En A. Menéndez de la Cuesta González (Ed.), *Encuentros digitales: escrituras, colecciones, aprendizajes en español. Encontros digitais: escritas,*

colecçoes, aprendizagem em portugués (pp.150-176). Universidad Complutense de Madrid y Fundación BBVA.

Esparza Torres, M. Á. y Niederehe, H.-J. (1995). *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español (BICRES). Desde los comienzos hasta el año 1600*. John Benjamins.

Esparza Torres, M. Á. y Niederehe, H.-J. (1999). *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español (BICRES II). Desde el año 1601 hasta el año 1700*. John Benjamins.

Esparza Torres, M. Á. y Niederehe, H.-J. (2005). *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español (BICRES III). Desde el año 1701 hasta el año 1800*. John Benjamins.

Fabbri, M. (1979). *A Bibliography of Hispanic Dictionaries. Catalan, Galician, Spanish, Spanish in Latin America and the Philippines. Appendix: A Bibliography of Basque Dictionaries*. Galeati.

Fabbri, M. (2002). *A Bibliography of Hispanic Dictionaries. Catalan, Galician, Spanish, Spanish in Latin America and the Philippines. Supplement I*. Panozzo Editore.

García Aranda, M.ª Á. y Peña Arce, J. (2019). La Biblioteca Virtual de la Filología Española: de Antonio de Nebrija a Antonio de Nebrija. En J. M.ª Santos Rovira (Ed.), *Raíces y horizontes del español. Perspectivas dialectales, históricas y sociolingüísticas* (pp.119-135). Axac.

San Vicente, F. (1995). *Bibliografía de la lexicografía española del siglo XVIII*. Piovan editore.

Viñaza, Conde de la, (1892). *Bibliografía española de lenguas indígenas de América*. Sucesores de Rivadeneyra.

Viñaza, Conde de la, (1893). *Biblioteca histórica de la filología castellana*. Imprenta y Fundición de Manuel Tello.

Chapter III

De dos bases de datos relacionales a una base de datos XML. El proyecto COMREGLA

From two relational databases to an XML one. Project COMREGLA

Eveling Garzón Fontalvo [a], Berta González Saavedra [b], José Ignacio Hidalgo González [c], Iván López Martín [b], Alberto Pardal Padín [a], Guillermo Salas Jiménez [b] & Cristina Tur [a]
Universidad de Salamanca ([a]), *Universidad Complutense de Madrid* ([b]), *IES Sant Marçal* ([c]) – *España*

Resumen: Esta contribución tiene como objetivo presentar las modificaciones y adaptaciones que hemos hecho a dos bases relacionales del proyecto REGLA (REcción y complementación en Griego Antiguo y Latín) cuyo foco se encuentra en el estudio de predicaciones verbales. El fin de estos cambios –que se enmarcan en el proyecto COMREGLA– es que la información contenida en ellas sea compatible con otras herramientas de tratamiento automático del lenguaje y que el análisis no sea solo de predicaciones nucleares y básicas, sino de textos completos. Para ello, se ha creado un estándar de notación nuevo que permite reflejar la riqueza de la información morfológica, sintáctica, semántica y léxica de las bases de datos originales, dar cuenta de la propia recursividad del lenguaje (en términos de posibles relaciones de estructuras) y enriquecer el análisis con etiquetas para componentes que no se estudiaban antes (complementación no obligatoria de la predicación expandida).

Abstract: This paper aims to present the modifications made to two relational databases belonging to REGLA (Rección y complementación en Griego antiguo y Latín, Spanish acronym for *Government and complementation in Ancient Greek and Latin*), a research project centred on the study of verbal predications. This transformation, which is the main goal of the project COMREGLA, seeks to make the information stored in these databases compatible with other natural language processing tools,

as well as to expand their analysis beyond core and basic predications to cover the whole discourse. To do so, a new standard for linguistic annotation has been developed which not only enables the representation of the rich linguistic information on the source databases, but also allows for the recursive nature of language (understood as complex structures relations) and enriches the analysis with new data from elements not addressed hitherto, such as non-obligatory complementation within the expanded predication.

1. Introducción

El proyecto COMREGLA[1] tiene como objetivo hacer accesibles y compatibles con otros recursos digitales dos bases de datos relacionales que se concibieron para estudiar las estructuras predicativas de los verbos más frecuentes del griego antiguo y el latín. A raíz de la aparición de corpus anotados para estas dos lenguas a partir de los años 2000 y del nacimiento del proyecto Linking Latin (Passarotti *et al*., 2019), se ha hecho evidente la necesidad de abrir estas bases de datos y convertirlas en recursos accesibles y compatibles con otras herramientas disponibles de tratamiento automático del lenguaje.

Esta transformación ha supuesto una serie de dificultades que están directamente relacionadas no solo con el tipo de información almacenada en las bases de datos originales, sino también con la naturaleza de la información recogida en los otros recursos con los que se pretende hacer compatible nuestra herramienta.

En esta contribución, en primer lugar, presentaremos los datos contenidos en el recurso de partida (es decir, en las bases de datos relacionales) y explicaremos algunas de las dificultades que entraña su adaptación para, acto seguido, describir cómo otros recursos existentes abordan estas cuestiones (§ 2). A continuación, especificaremos el marco teórico en el que se encuadra nuestro proyecto (§ 3), así como los aspectos metodológicos de la transformación de las bases de datos relacionales (§ 4). Por último, profundizaremos en la descripción de algunos problemas relativos al análisis de las formas nominales del verbo y en las soluciones dadas a estos (§ 5). Para finalizar, plantearemos unas conclusiones (§ 6).

1 Financiado gracias a una *Ayuda a equipos de investigación científica en Humanidades Digitales* de la Fundación BBVA (convocatoria 2018).

2. Cuestiones preliminares. Presentación de los recursos de partida

COMREGLA ha supuesto toda una renovación de nuestros recursos que ha desembocado en la creación de una nueva base de datos. A continuación, describiremos nuestro proyecto de partida, REGLA, y otros proyectos similares que han servido de base teórico-técnica para el desarrollo de esta nueva herramienta.

2.1. Nuestro proyecto: REGLA

El grupo de investigación Rección y Complementación en Griego antiguo y Latín (REGLA), que es el inicio del actual proyecto COMREGLA, fue creado en 1992 por un grupo de investigadores de cuatro universidades españolas: U. Autónoma de Madrid, U. Complutense de Madrid, U. de Alcalá de Henares y la U. de Santiago de Compostela, al que se fueron incorporando otras como la U. de Salamanca y la U. de Oviedo.

En los últimos años, el equipo ha estado trabajando en el desarrollo de dos bases de datos relacionales, REGLA-Griego y REGLA-Latín, que tienen como objetivo último obtener un repertorio lo más completo posible de los marcos predicativos (MP), esto es, los esquemas de complementación obligatoria de los verbos más frecuentes en griego antiguo y latín. Así pues, estas bases de datos han sido diseñadas para recoger, organizar y recuperar las apariciones de cada verbo en un corpus seleccionado, con su correspondiente análisis sintáctico, semántico y léxico.

A pesar de sus diversas transformaciones (cambio en la nomenclatura de los distintos proyectos financiados y en la configuración del equipo de trabajo)[2], el objetivo del grupo ha sido siempre el estudio de la estructura oracional del griego antiguo y el latín y, en particular, de los aspectos relacionados con la sintaxis y semántica de los constituyentes que la integran.

Para ilustrar el tipo de análisis que recogen estas bases de datos, podemos observar la sección superior de una de las fichas del verbo *appello* 'nombrar, denominar' en latín:

2 En orden cronológico los proyectos concedidos son: *Corpus y base de datos sobre la complementación. Un estudio lingüístico sobre el griego y el latín* (CAM 06/0013/1999); *Sintaxis y semántica de la complementación II* (BFF2001-0135-C04); *Corpus de rección y complementación en griego y latín* (HUM2005-06622-C04); *Corpus de rección y complementación en griego y latín II* (FFI2009-13402-C04); *Problemas de complementación en griego y latín* (FFI2013-47357-C4); *Interacción del léxico y la sintaxis en griego y latín* (FFI2017-83310-C3). Como antecedentes de estos proyectos se pueden mencionar *Las funciones nominales en Griego y en Latín: Tucídides y Cicerón* (PS91-0014); *Las unidades funcionales en la oración en griego y en latín* (PB94-0197); *Sintaxis, semántica y pragmática de la complementación* (PB97-0005-C04), que desarrollaron las bases teóricas. En la actualidad, el proyecto vincula a más de una quincena de investigadores (entre profesores y alumnos de postgrado).

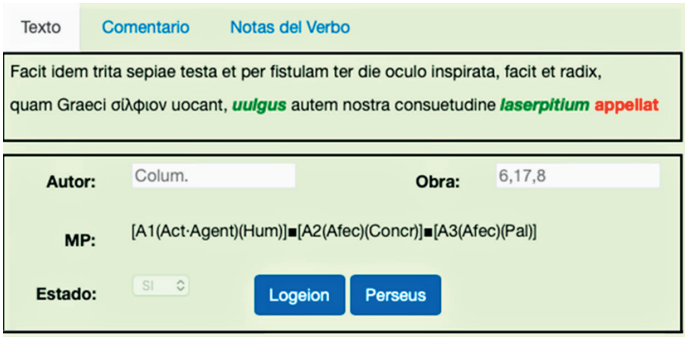


Figura 1. Ejemplo parcial de una ficha en la base de datos REGLA.

Aquí tenemos parte del texto recogido en la ficha de la Figura 1.

Ejemplo (1).

radix,	*quam*	*Graeci*	σίλφιον	*uocant,*	*uulgus*
raízNOM.SG	REL.AC.SG	griegosNOM.PL.	silfioAC.SG	llaman	vulgoNOM.SG
autem	*nostra*	*consuetudine*	*laserpitium*	*appellat*	
PART	nuestraABL.SG	tradiciónABL.SG	laserpicioAC.SG	denomina	

"la raíz que los griegos llaman *silfio*; el vulgo, en cambio, según nuestra tradición, la denomina *laserpicio*" (Colum. 6.17.8)

En concreto, en esta ficha se analiza la predicación *uulgus autem nostra consuetudine laserpitium appellat*, traducida como "el vulgo, en cambio, según nuestra tradición, la denomina laserpicio", y se recoge la estructura argumental del verbo *appello*. Así pues, los elementos destacados en verde, esto es, *uulgus* y *laserpitium*, se identifican con los elementos obligatorios –y, nótese bien, explícitos– de la predicación de este verbo. En la ficha se recoge también la formalización del análisis del verbo en este pasaje en la casilla MP, donde se nos indica que en esta construcción *appello* cuenta en realidad con los siguientes constituyentes obligatorios (dos explícitos y uno elíptico contextual): un Argumento 1 Actor-Agente tipificado como /+humano/ (*uulgus*); un Argumento 2 Afectado /+concreto/ (elíptico contextual) y un Argumento 3 Afectado con la caracterización léxica /+palabras/ (*laserpitium*).

A pesar de que los datos consignados en estas bases de datos son de bastante calidad, puesto que los análisis han sido llevados a cabo por miembros del equipo de investigación con formación en lingüística y en griego y latín, esta forma de organizar y almacenar los datos ha resultado no ser del todo efectiva, ya que plantea, sobre todo, dos dificultades:

i Incapacidad de dar cuenta del carácter recursivo del lenguaje. Cuando un constituyente de la oración forma, a su vez, una estructura predicativa propia (por ejemplo, otra oración), no se puede abordar el análisis de manera conjunta, sino que cada elemento predicativo ha de analizarse en una ficha diferente. En el ejemplo (1), el análisis de la oración de relativo (*quam Graeci* σίλφιον *uocant* 'que los griegos llaman silfio') no se puede poner en relación con el de la oración principal en la que se integra.

ii Limitación del análisis a constituyentes centrales de la predicación. Por esta razón, un sintagma como *nostra consuetudine* 'según nuestra tradición' (ejemplo 1), que funciona como un disjunto (esto es, un elemento que trasciende el ámbito de la predicación), queda fuera del ámbito de análisis en REGLA.

Identificar estas dos cuestiones problemáticas y darles una solución satisfactoria ha sido clave para cumplir con una parte crucial del proyecto COMREGLA, como es el hacer compatible los datos disponibles en REGLA con otras herramientas y recursos dedicados a las lenguas que nos ocupan.

2.2. Otros proyectos

De cara a resolver los tres problemas descritos, uno de los primeros pasos ha sido comprobar de qué manera se abordaban en otros treebanks con anotación semántica y sintáctica, especialmente los dedicados a las lenguas clásicas, como PROIEL (Haug & Jøhndal, 2008), el Index Thomisticus Treebank (ITTB; Passarotti, 2009) y el Ancient Greek and Latin Dependency Treebank (AGLDT; Bamman & Crane, 2011).

La primera de las tres herramientas se sirve del etiquetado morfológico de Universal Dependencies para el análisis sintáctico de textos con el objetivo de presentar de forma arbórea las distintas dependencias de un predicado; este sistema es aplicado a un pequeño corpus de obras latinas y griegas, entre otras lenguas.

El ITTB, por su parte, surge de uno de los proyectos pioneros en lingüística computacional, el Index Thomisticus. Su objetivo inicial era la anotación morfológica de las obras de Tomás de Aquino. Con todo, desde hace algunos años se ha ampliado el corpus con autores clásicos latinos, se ha comenzado a anotar también información sintáctica y semántica y se ha añadido un léxico de valencias basándose en el marco teórico desarrollado por el Prague Dependency Treebank, aunque con ciertas adaptaciones.

Por último, el AGLDT, de la Universidad de Leipzig, ofrece una recopilación de textos griegos y latinos de distintos géneros y épocas usando también el etiquetado de dependencias sintácticas del Prague Dependency Treebank[3].

Los *treebanks* citados ofrecen el análisis de obras completas, por lo que se han tenido que enfrentar a los problemas que planteábamos en el punto anterior: el análisis de estructuras complejas de subordinación y coordinación con sus propias funciones y la anotación de complementos no centrales. El análisis de estructuras complejas está resuelto por estos *treebanks*; sin embargo, no permiten un análisis tan pormenorizado como el que se ofrece en REGLA, que contempla más categorías y depura mucho más los datos[4]. La transformación directa al formato de uno de estos *treebanks* habría supuesto, por lo tanto, una pérdida de información de la base de datos de partida, razón por la que no se ha llevado a cabo. Con todo, sí resultó útil la observación y el conocimiento de los *treebanks* mencionados para comprobar cómo se anotaban los constituyentes no centrales de la predicación, que, en general, reciben etiquetas distintas para marcar su relación sintáctica y semántica menos estrecha con la predicación.

3. Marco teórico

Para explicar por qué el análisis preexistente en las bases de datos relacionales de REGLA es más preciso y no puede ser transformado directamente al formato usado por otros *treebanks* es necesario mencionar que nuestras bases de datos tienen como principal fundamento teórico la Gramática Funcional de S. Dik (1997). Este modelo se ha aplicado con notable éxito al estudio tanto del latín como del griego. Cabe destacar en esta línea el trabajo de Pinkster para el latín (2015; 2021) y los desarrollados por los miembros de REGLA tanto para el latín como para el griego (p. ej., Baños *et al*., 2003; Torrego *et al*., 2007; Baños, 2009; Jiménez López, 2020).

En concreto, es fundamental tener en cuenta el concepto de predicación y de MP (Dik, 1997, p.78ss; de la Villa, 2003) para comprender el desarrollo de la base de datos REGLA. El primero hace referencia a una estructura sintáctico-semántica formada por un verbo y los elementos que de él dependen, tanto si son obligatorios como si no. El segundo es el

3 Cabe mencionar además la existencia de algunos léxicos de valencias, herramientas que recogen bien la estructura sintáctica de los verbos, como el *Homeric Dependency Lexicon* para las obras homéricas (que anota según los parámetros teóricos sintácticos del *Prague Dependency Treebank*) o el *IT-VaLex* para la obra de Tomás de Aquino, bien su estructura semántica, como el *Latin Vallex* (desarrollado a partir de la anotación semántica del *Index Thomisticus Treebank*).

4 Algo similar ocurre con los léxicos de valencias de acceso abierto que, a pesar de la valiosa información que comparten, no aportan una tan detallada y completa como la que contiene REGLA.

esquema de complementación obligatoria de un verbo. Este estudio de los MP es, en última instancia, el responsable de que el interés de la base de datos previa se haya centrado sobre la complementación obligatoria y haya dejado de lado el análisis exhaustivo de todos los elementos de la predicación y la oración.

Esta perspectiva funcionalista se ha enriquecido a lo largo de los años con aportaciones de otros marcos teóricos afines como la Gramática Cognitiva (Langacker, 2008) o la Gramática de las Construcciones (Goldberg, 1995), así como con otras teorías funcionalistas posteriores a las de Dik, como la Gramática del Papel y la Referencia (Van Valin & LaPolla, 1997) y la Gramática Funcional del Discurso (Hengeveld & Mackenzie, 2008). Todas estas perspectivas comparten una visión de la lengua en la que priman la función comunicativa del lenguaje y el uso en contexto por encima de cuestiones puramente formales.

4. Aspectos metodológicos

Con el fin de hacer compatibles las bases de datos REGLA-Griego y REGLA-Latín con otras herramientas de procesamiento del lenguaje natural, era necesario hacer una migración de las dos bases de datos relacionales a una base de datos XML, COMREGLA, lo que supone un cambio estructural de gran calado, puesto que las formas de almacenamiento de la información son muy diferentes.

En un primer momento, tomamos como modelo un *standard* XML ya existente para el análisis sintáctico y semántico necesario en la creación de *treebanks*, el *Prague Markup Language* (PML), un sistema de marcado desarrollado para el *Prague Dependency Treebank* y que ya ha sido aplicado al latín en el ITTB, entre otros recursos (cf. § 2.2).

A grandes rasgos, el PML es un marcaje *stand-off* que se articula en cuatro capas o niveles de análisis: *tokens* o nivel *words*, morfología o nivel morfológico, análisis sintáctico o nivel analítico y análisis semántico-pragmático o nivel tectogramatical. No obstante, tal y como hemos mencionado (§ 2.2), no resultó ser del todo compatible con el tipo de información que se almacena en nuestras bases de datos relacionales. En efecto, si bien hasta el nivel morfológico el PML se adecúa correctamente al tipo de información de REGLA, en el nivel sintáctico y semántico, sigue preceptos teóricos diferentes a los que sustentan nuestro proyecto[5]. Por otra parte, PML resulta insuficiente para reflejar determinada información sintáctica y semántica que se tiene en cuenta en REGLA (como es el caso de las ca-

5 Por ejemplo, el PML distingue entre argumentos y adjuntos obligatorios, mientras que en COMREGLA los adjuntos son por definición constituyentes opcionales determinados por el predicado.

racterísticas semánticas de las predicaciones en su conjunto, cuando son componentes de una principal).

Así las cosas, decidimos que los elementos de la base de datos COMREGLA estarían anotados mediante un sistema propio de etiquetas XML que se ajustara lo más posible a los campos de las bases de datos relacionales de REGLA. Este sistema de etiquetas se basa en buena medida en el PML, pero también en otros sistemas de gramática de dependencias, como PROIEL.

Las bases de datos de REGLA contienen cuatro tipos de información lingüística: morfológica, sintáctica, semántica y léxica. Esta información se ha redistribuido, como se observa en la tabla 1, en dos niveles *stand-off*: WORDS, en el que se recoge la forma y el lema de cada palabra del texto, así como su información morfológica, y CLAUSES, que es de mayor complejidad, en el que se explicitan los rasgos léxicos de las unidades lingüísticas, las relaciones sintácticas y semánticas que se establecen entre ellas y las jerarquías de estructuras sintácticas en las que se insertan.

Tabla 1. Distribución de la información lingüística en los nuevos niveles.

	WORDS	CLAUSES
Morfología	Forma y lema Características morfológicas	-
Sintaxis	-	Palabras (WORDS) < Predicaciones (CLAUSES) < Oraciones (SENTENCES) Relaciones sintácticas (dependencias, funciones sintácticas, etc.) • entre las palabras de una oración, • entre las predicaciones que conforman una oración
Semántica	-	Características semánticas • de las relaciones (funciones semánticas, tipos de subordinación, etc.), • de las predicaciones (polaridad, diátesis, fuerza ilocutiva, control, aspecto léxico, etc.)
Léxico	-	Rasgos léxicos

Los aspectos sintácticos que se recogen en la capa CLAUSES parten de la división del texto en unidades. Todo texto se compone de palabras y otros *tokens* como la puntuación, números, etc., que constituyen la forma más básica (WORDS). Las unidades básicas comprendidas entre puntuación fuerte forman oraciones (SENTENCES). Entre ambas unidades se sitúa la unidad lingüística que para nosotros es central: las predicaciones (CLAUSES), que es, como se dijo en § 3, la unidad de análisis fundamental de las bases de datos relacionales de

REGLA. Una vez determinadas las unidades sintácticas, establecemos las relaciones entre estas unidades, tanto de las palabras entre sí, como de las predicaciones u oraciones.

Asimismo, las relaciones entre las unidades sintácticas tienen una dimensión semántica, para lo que se consignan, por ejemplo, las funciones semánticas, que definen el tipo de relación entre el verbo y sus elementos (Agente, Paciente, Beneficiario, etc.) o los tipos de subordinación (completiva, condicional, concesiva, etc.). Además, las propias predicaciones tienen ciertas características semánticas que les son propias, como pueden ser la polaridad, la diátesis, la fuerza ilocutiva o el aspecto léxico. Por último, se anota la información sobre el léxico de los elementos que funcionan como participantes en la oración.

Como se ha ilustrado anteriormente, en las bases de datos relacionales se analizan fragmentos sueltos sin conexión entre ellos, elegidos solamente con el fin de analizar los MP de ciertos verbos. En la nueva base de datos, en cambio, las oraciones se encontrarán en su contexto, ya que se analizan textos completos. Comparemos el análisis del ejemplo (1) en REGLA (Figura 1) con la forma que presenta el mismo ejemplo en la base de datos COMREGLA. En la capa WORDS, como se ha mencionado ya, aparece la información morfológica de cada palabra. Así, como se puede observar en la tabla 2, de la palabra *radix* 'raíz', por ejemplo, se recogerá el tipo de palabra (sustantivo), la declinación (3ª declinación), el caso, el número y el género. Para el verbo *uocant* 'llaman', se incluirán datos como la conjugación, el tiempo, el modo, la voz, la persona y el número.

Tabla 2. Análisis del ejemplo (1) en COMREGLA.

radix	**quam**	**Graeci**	**σίλφιον**	**uocant**	**uulgus**	**autem**	**nostra**	**consue-tudine**	**laserpi-tium**	**appellat**
Sust. 3ª decl. Nom. Sg. Fem.	Pron. Acus. Sg. Fem.	Sust. 2ª decl. Nom. Pl. Masc.	Sust. 2ª decl. Acus. Sg. Fem.	Verbo 1ª conj. Pres. Ind. Act. 3 pers. Pl.	Sust. 2ª decl. Nom. Sg. Neutr.	Indecl.	Det. Abl. Sg. Fem.	Sust. 3ª decl. Abl. Sg. Fem.	Sust. 2ª dec. Acus. Sg. Neutr.	Verbo 1ª conj. Pres. Ind. Act. 3 pers. Sg.

En la capa CLAUSES, por su parte, se muestran las relaciones entre las palabras dentro de las predicaciones y entre las predicaciones entre sí. Dado que, como se ha visto antes, el verbo es generalmente el núcleo de la predicación, en nuestro ejemplo, hay dos predicaciones, una cuyo núcleo es *uocant* 'llaman' y otra cuyo núcleo es *appellat* 'denomina'. Los demás elementos de la predicación se relacionan con ellos tanto sintáctica como semánticamente. De este modo, por ejemplo, *Graeci* 'los griegos' es el sujeto (función sintáctica)

Agente (función semántica) de *uocant* 'llaman', y *uulgus* 'el pueblo' cumple las mismas funciones (sujeto Agente) respecto a *appellat* 'denomina'.

Así mismo, los elementos de una predicación pueden remitir secundariamente a otros componentes. Por ejemplo, los nombres σίλφιον y *laserpitium*, que es como cada uno de los pueblos llama a la raíz en cuestión, cumplen una doble función: sintácticamente son complementos del objeto obligatorios de los verbos *llamar* y *denominar* ("a la raíz la llaman *laserpicio*"), semánticamente hacen referencia a *radix* 'la raíz'. Esta doble relación está marcada mediante una línea discontinua.

Por otro lado, las predicaciones en su conjunto también cuentan con sus propias características sintácticas (si se trata de una oración principal o subordinada; si es esto último, de qué tipo es y qué función sintáctica cumple) y semánticas (si es un evento controlado, su polaridad y diátesis).

Además de todo esto, se reflejan las características léxicas de los distintos elementos, sean palabras o predicaciones completas.

En la siguiente ilustración se muestra un modelo de representación gráfica del análisis en COMREGLA.

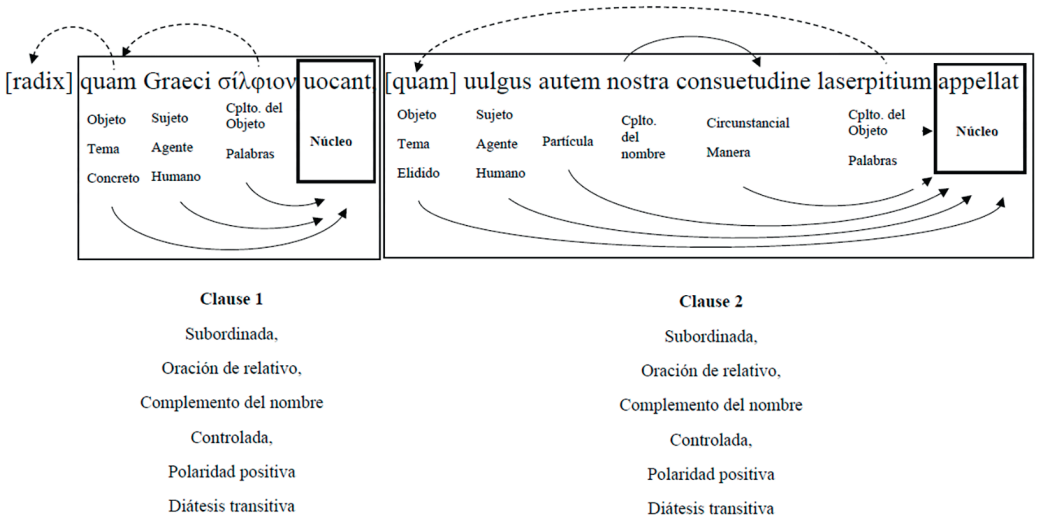


Figura 2. Modelo de representación de la capa CLAUSES para el ejemplo (1).

5. El problema de las formas nominales del verbo en latín y en griego antiguo

Una vez presentada la estructura general de la nueva base de datos XML, ahora profundizaremos en la descripción de algunos problemas relativos al análisis de las formas nominales del verbo –elementos altamente productivos en las lenguas estudiadas (§ 5.1)– y en las soluciones que se ofrecen desde el nuevo modelo COMREGLA (§ 5.2).

5.1. Descripción de los problemas

Las llamadas "formas nominales del verbo" tienen unas peculiaridades morfológicas que las hacen participar de una doble naturaleza nominal y verbal, pero la razón por la cual las hemos escogido para profundizar en los problemas que nos han surgido es que en el plano sintáctico y semántico se caracterizan, sobre todo, porque no suelen formar una oración independiente: no suelen constituir un mensaje completo, puesto que no tienen autonomía sintáctica ni comunicativa[6]. Tienen, pues, un carácter subordinado: están insertas en una oración y, a la vez, tienen su propio MP.

A través de los ejemplos que se analizan a continuación se ilustra la gran variedad de construcciones sintácticas a las que dan lugar estas formas nominales y se recoge de manera esquemática la información presente en las bases relacionales de REGLA.

Para comenzar, en el ejemplo (2) tenemos una construcción de infinitivo no concertado, donde el verbo en infinitivo (*facere*) se inserta en el MP del verbo principal (*uolo*), al tiempo que tiene su propia complementación: un sujeto (*te*) y un objeto (*hoc*). El infinitivo participa, en este sentido, en dos predicaciones al mismo tiempo[7].

6 Estas no son las únicas construcciones que forman oraciones subordinadas en griego y en latín, pues tenemos oraciones introducidas por conjunciones subordinantes, así como por pronombres relativos. Sin embargo, la elección de las formas nominales del verbo para este artículo es que son mucho más frecuentes y productivas en ambas lenguas.

7 En los modelos de representación de los ejemplos se han empleado las siguientes abreviaturas:
- ARG-SBJ: argumento-sujeto
- ARG-OBJ: argumento-objeto
- CN: complemento del nombre
- Coord: coordinación
- Disj: disjunto
- elip: elemento elíptico
- MP: marco predicativo
- Prep: preposición
- *: elemento sin correspondencia en el nivel WORDS

Ejemplo (2).

nunc	*ego*	*te*	*facere*	*hoc*	*uolo*
ADV	PRON.AC.SG	PRON.AC.SG	INF.PRES.ACT	PRON.AC.SG	QUERER1SG.PSTE.IND.ACT
"ahora yo quiero que tú hagas eso" (Plaut. *Bacch*. 93)					

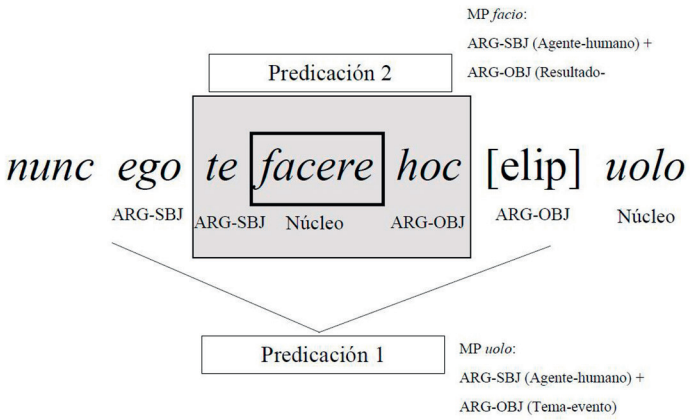


Figura 3. Modelo de representación del ejemplo (2).

Otra construcción típica de estas formas en las lenguas clásicas es la del participio sustantivado, ilustrado en (3). En ejemplos como este, a la participación de la forma nominal del verbo en dos predicaciones a la vez se añade el problema del marcaje del léxico. En efecto, debido a la sustantivación de οἱ οἰκοῦντες 'los que viven' nos encontramos con una dicotomía a la hora de establecer el léxico del participio: ¿es /+humano/ porque está sustantivado o es /+evento/ porque expresa un estado?

Ejemplo (3).

ἔμειναν	δὲ	καὶ	**οἱ**	παρὰ	τὴν	θάλατταν	**οἰκοῦντες**
permanecer 3PL.AOR.IND.ACT	PART	ADV	ART.NOM.PL	PREP	ART. AC.SG	marAC.SG	habitarPART. PRES.NOM.PL
ἐν	Σόλοις						
PREP	Solos DAT.PL						
"Y se quedaron también **los que viven** junto al mar, en Solos" (X. *An*. 1.2.24).							

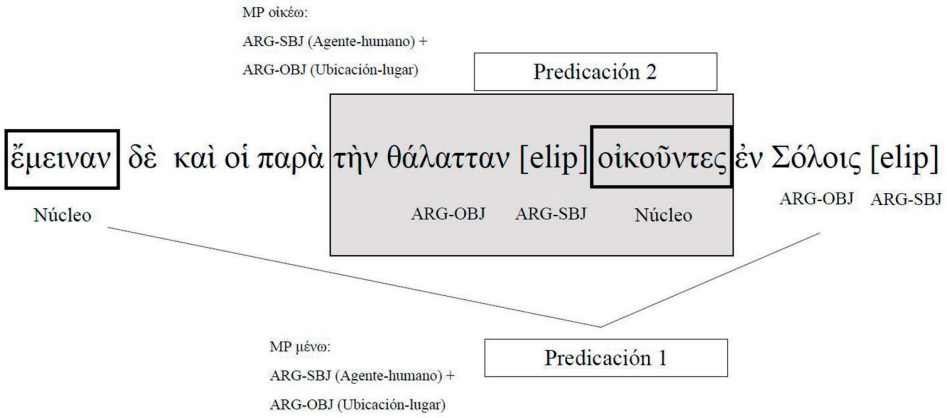


Figura 4. Modelo de representación del ejemplo (3).

El ejemplo (4) representa otra de las estructuras habituales a las que dan lugar estas formas: el participio atributivo. En este caso, el participio μείνας 'que permanece' funciona como un modificador de στρατὸς 'ejército'. Sin embargo, en su análisis se pierde información sobre su complementación, dado que se le asigna un sujeto elíptico contextual, a pesar de que tal sujeto sea el sustantivo στρατὸς.

Ejemplo (4).

ἀλλ'	οὐδ'	ὁ	**μείνας**	νῦν	ἐν	Ἑλλάδος	τόποις
CONJ	CONJ	ART.NOM.SG	permanecerPART. NOM.SG	ADV	PREP	GreciaGEN. SG	lugarDAT. PL.
στρατὸς	κυρήσει	νοστίμου	σωτηρίας				
ejércitoNOM. SG	conseguir3SG. FUT.IND.ACT	regresoGEN. SG	salvaciónAC.SG				

"pero ni siquiera el ejército **que permanece** ahora en territorio griego conseguirá la salvación del regreso" (A. *Pers.* 796-797).

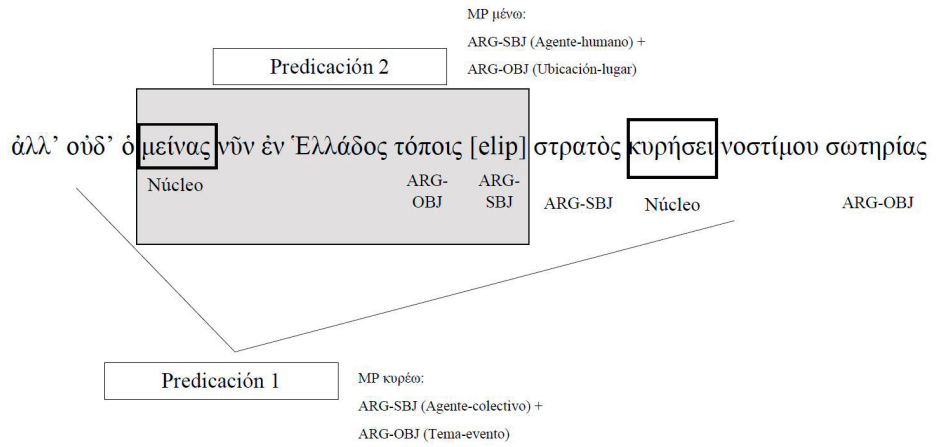


Figura 5. Modelo de representación de ejemplo (4).

En resumen, por la naturaleza de las lenguas clásicas, las formas no personales del verbo son uno de los escollos más frecuentes y que mejor ilustran este proceso de transformación de un sistema a otro, ya que obligan a condensar información que, hasta el momento, aparecía en dos (o más) fichas y a establecer cuál es la relación entre las predicaciones, sea esta de carácter obligatorio, tal como hemos visto en los ejemplos (2) y (3), o no, como en el ejemplo (4).

Por otra parte, vemos cómo hay otros elementos que están dentro de la predicación o que unen una oración con la anterior en el texto (en el ejemplo 4, ἀλλ᾽ y οὐδ᾽ cumplirían esta función) o que enlazan predicaciones y que quedarían sin etiqueta (al igual que la predicación segunda en 4) y tampoco aparecerían recogidos de ninguna manera.

5.2. Soluciones adoptadas en COMREGLA

El nivel CLAUSES del marcaje en XML de COMREGLA ofrece las herramientas necesarias para afrontar los problemas planteados por las formas no personales del verbo. Veamos cada uno de los ejemplos y comprobemos cuáles son las soluciones que proponemos en COMREGLA.

Figura 6: respecto a la integración de subordinadas en sus respectivas predicaciones principales, problema que se ilustró en el ejemplo (2), la nueva base de datos permite establecer la naturaleza morfológica y sintáctico-semántica del objeto de la predicación regida, a diferencia de REGLA; recordemos que en estas solo se recogía la información morfológica (un infinitivo) sin que se pudiese establecer la relación entre ambas predicaciones. Para relacionarlas, como se observa en la figura 6, la base de datos COMREGLA

se sirve de un elemento en la oración principal que no remite a ninguna realidad textual y que recoge la información de la predicación subordinada (*).

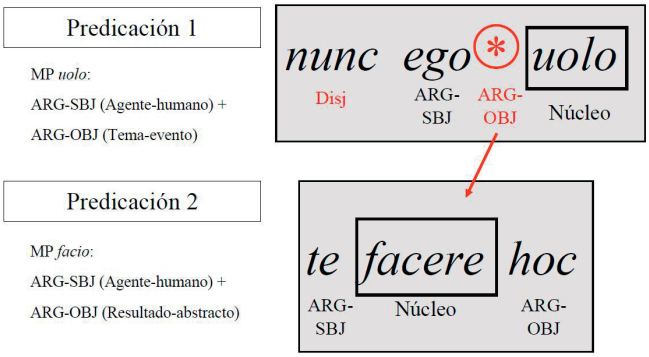


Figura 6. Solución de análisis en COMREGLA para el ejemplo (2).

Figura 7: el segundo de los problemas que plantean las formas nominales de los verbos y que se ha ejemplificado en (3) es la necesidad de recoger la información léxica de las predicaciones subordinadas cuando están sustantivadas. A este respecto, como se ilustra en la figura (7), la base de datos COMREGLA es capaz de almacenar esta información, añadiéndosela al elemento (*). Así, en la oración principal, el elemento que remite a la predicación 2 en su conjunto presenta el rasgo /+humano/ y la predicación en sí conserva su carácter de evento. Además, permite etiquetar elementos que no pertenecen a la predicación nuclear, como δὲ y καί.

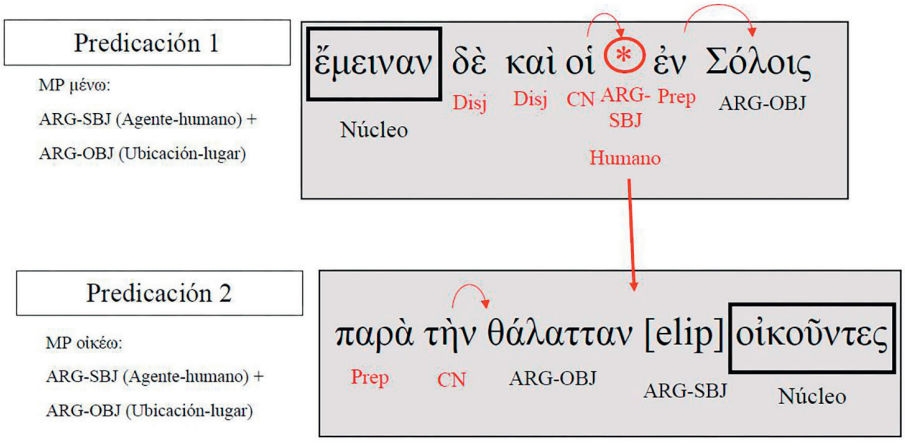


Figura 7. Solución de análisis en COMREGLA para el ejemplo (3).

Figura 8: en el análisis de la base de datos REGLA, no es posible establecer una relación entre el participio atributivo (μείνας) y el sustantivo al que complementa (στρατός). Por el contrario, la nueva base de datos, como se ve, permite relacionar ambos términos en dos sentidos: por un lado, mediante un elemento en la oración principal que no remite a ninguna palabra y que recoge la información de la predicación subordinada en su conjunto se marca la función de la predicación subordinada como complemento del nombre στρατός. Por el otro, en la predicación subordinada se considera un sujeto elíptico cuya información es coincidente con la de στρατός. Además, como ya sucedía en la figura 7, los elementos que no pertenecen estrictamente al ámbito de la predicación, sino al nominal (como los artículos) y al oracional (partículas discursivas y algunos adverbios), reciben sus etiquetas correspondientes.

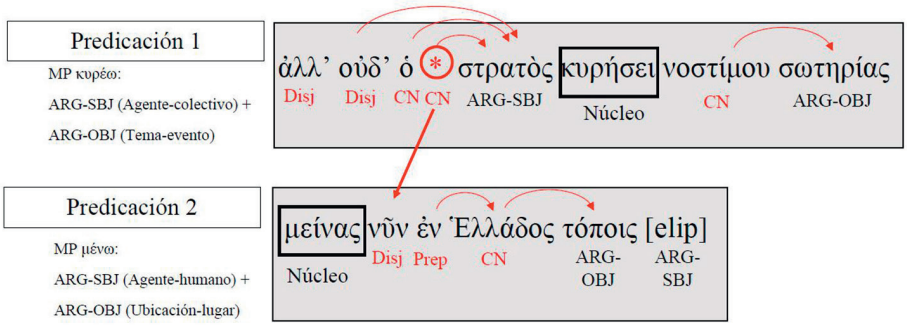


Figura 8. Solución de análisis en COMREGLA para el ejemplo (4).

A través de los anteriores ejemplos hemos podido mostrar cómo el nuevo análisis propuesto por COMREGLA permite solucionar los problemas principales que plantean las bases de datos relacionales REGLA: la relación entre predicaciones y el etiquetado de elementos que no pertenecen a la estructura obligatoria de la predicación.

De esta manera, la información recogida en COMREGLA mantiene el análisis refinado de las bases de datos predecesoras solventando sus carencias y consiguiendo, al mismo tiempo, ser compatible con otras herramientas de PLN.

6. Conclusiones

Como se ha podido comprobar, la nueva base de datos XML hereda de las antiguas bases de datos relacionales la capacidad de almacenar y gestionar un profundo análisis sintáctico-semántico que puede ser de enorme ayuda en la labor de investigación lingüística del griego antiguo y el latín, pero también supone algunas novedades respecto a sus predecesoras.

Como se recordará, las bases de datos relacionales de las que parte este trabajo se nutren de fragmentos no conectados entre sí, de los que solo podían analizarse el verbo y su complementación obligatoria. Frente a esto, la base de datos COMREGLA permite tanto etiquetar textos completos, estableciendo para ello las relaciones pertinentes entre distintas predicaciones, como analizar todos sus componentes, sean obligatorios o no.

Asimismo, posibilita unas búsquedas mucho más precisas y completas, al haber mucha más información analizada que poder recuperar: estructuras complejas como las formas nominales del verbo, adjetivos con función atributiva, construcciones no pertenecientes a la predicación, entre otras, sin perder la precisión que se había ganado con la anotación detallada de las estructuras predicativas.

Por otro lado, el hecho de emplear la misma tecnología que otros recursos similares, como, por ejemplo, LiLa, permite la compatibilidad con ellos y, aunque esté de momento centrado en el latín y el griego antiguo, es un modelo de etiquetado que podría aplicarse a otras lenguas.

Referencias

Bamman, D. & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebank. In C. Sporleder, A. van Den Bosch & K. Zervanou (Eds.), *Language Technology for Cultural Heritage, ser. Foundations of Human Language Processing and Technology* (pp. 79-89). Springer.

Baños, J.M. (coord.) (2009). *Sintaxis del latín clásico*. Liceus E-Excellence.

Baños, J.M., Cabrillana, C., Torrego, M.E. y de la Villa, J. (2003). *Praedicativa: complementación en griego y latín*. Universidade de Santiago de Compostela.

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

Goldberg, A.E. (1995). *Constructions: a Construction Grammar approach to argument structure*. The University of Chicago Press.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., … Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5*. Prague: Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID. (http://hdl.handle.net/11234/1-2621).

Haug D.T.T. & Jøhndal, M.L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In C. Sporleder & K. Ribarov (Eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (pp. 27-34). Marrakech.

Hengeveld, K. & Mackenzie, J.L. (2008). *Functional discourse grammar: a typologically-based theory of language structure*. Oxford University Press.

Jiménez López, M. D. (Coord. Ed.) (2020). *Sintaxis del griego antiguo*. 2 vols. CSIC.

Langacker, R.W. (2008). *Cognitive Grammar: an Introduction*. Oxford University Press.

Passarotti M. (2009). Theory and Practice of Corpus Annotation in the Index Thomisticus Treebank. *Lexis*, *27*, 5-23.

Passarotti M., Cecchini F.M., Litta E., Franzini G., Mambrini F. & Ruffolo P. (2019). LiLa: Linking Latin – A Knowledge Base of Linguistic Resources and NLP Tools. In T. Declerck, & J. P. McCrae (Eds.), *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK-PS 2019)*. University of Leipzig. DOI: 10.5281/zenodo.3358550

Pinkster, H. (2015). *The Oxford Latin Syntax. Volume 1: The Simple Clause*. Oxford University Press.

Pinkster, H. (2021). *The Oxford Latin Syntax. Volume II: The Complex Sentence and Discourse*. Oxford University Press.

Torrego, M.E., Baños, J.M., Cabrillana, C. y Méndez Dosuna, J.V. (2007). *Praedicativa II: esquemas de complementación verbal en griego antiguo y en latín*. Prensas de la Universidad de Zaragoza.

Van Valin, R. D. & LaPolla, R. J. (1997). *Syntax: Structure, Meaning, and Function*. Cambridge University Press.

Vendler, Z. (1967). Verbs and times. In Z. Vendler (Ed.), *Linguistics in philosophy* (pp. 97-121). Cornell University Press.

Villa, J. de la. (2003). Límites y alternancias en los marcos predicativos. In J. M. Baños, C. Cabrillana, M. E. Torrego, y J. de la Villa (Eds.), *Praedicativa. Complementación en griego y latín* (pp. 19-49). Universidad de Santiago de Compostela.

Chapter IV

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

Analysis of Colonel Anselmo Pineda's epistolary with Python: a glance to the collecting project from the study of the territory and social networks

Santiago Alejandro Ortiz Hernández
Red Humanidades Digitales – Colombia

Resumen: Este artículo analiza el coleccionismo del coronel Anselmo Pineda, quien fue el mayor coleccionista de documentos públicos del siglo XIX colombiano, a partir de su voluminoso epistolario conservado en la Biblioteca Nacional de Colombia. Se usa una metodología mixta que combina la tradicional lectura cercana y la lectura distante realizada por la máquina e implementada a través de técnicas propias de la ciencia de datos y los Sistemas de Información Geográfica implementados con Python. De manera que, a través de esa doble lectura, se propone alcanzar dos objetivos: I) plantear una aproximación basada en humanidades digitales e historia digital que permita descubrir el método de recopilación de documentos del coronel al examinar la composición de su red de colaboradores reconstruida exclusivamente mediante su correspondencia personal, y II) explorar el alcance espacial de esa red

de colaboradores de forma tal, que posibilite la evaluación de la dimensión espacial en la conformación de la biblioteca Pineda en el marco del proyecto civilizatorio de la naciente república en Nueva Granada.

Abstract: This article analyzes colonel Anselmo Pineda's collecting, who was the major documental collector of Colombian XIX century, taking as source his rich and abundant personal correspondence preserved at the National Library of Colombia. The previous through a mixed methodology that blend the traditional close reading of the letters and distant reading performed by the machine and implemented trough data science and GIS techniques with Python. Therefore, with that dual type of reading, this article proposes two goals: I) to pose both a theoretical and practical approximation, based on Digital Humanities and Digital History, that allows discover compilation method developed by colonel Pineda when examine the composition of his network of collaborators reconstructed exclusively through his personal letters, and II) to explore the spatial scope of that collaborators' network in a manner that allows the evaluation of the spatial dimension in the conformation of Pineda's library under the civilizatory project at the emergent republic of Nueva Granada.

1. Introducción

Con base en la abundante correspondencia personal del coronel Anselmo Pineda dispersa en varios archivos colombianos públicos y privados, en las pocas biografías juiciosas del coronel y en una investigación del autor del presente texto que contó con la financiación del Ministerio de Cultura de Colombia a través del programa de estímulos para la investigación en Humanidades Digitales, se reconstruyó tanto la trayectoria del militar, político e ilustrado, así como su estrategia coleccionista. En ese sentido, la trayectoria del coronel estuvo desde muy temprano marcada por la guerra y por un indiscutible patriotismo que se expresaba no solo en sus actos de lealtad a los ideales republicanos del siglo XIX, sino en sus consistentes esfuerzos por construir un monumento a la república, que en forma de colección documental, cumpliera el propósito de servir como archivo para el doble propósito de la conservación de la memoria y la identidad nacional, así como fuente de autoridad y legitimidad estatal desde un punto de vista jurídico y político.

Tras el proceso de independencia, la naciente república neogranadina resultó con un vacío simbólico y documental que requirió de la agencia de una extensa red de ilustrados, libreros, amigos, familiares y, en menor medida, autoridades estatales que colaboraron en

la consecución de un gran proyecto coleccionista materializado en dispersas colecciones privadas de diversa índole. Algunas de estas colecciones no solo fueron pensadas por sus propietarios en términos de su coherencia y orden interno, también fueron pensadas para enlazarse con otras y formar una sólida base documental que solventara la urgencia fundacional de un archivo de la historia y la ley de la república. Es el caso de la colección Pineda, la más grande del siglo, diseñada por su autor-coleccionista para eslabonarse con las colecciones de menor volumen que paralelamente organizaban el general Joaquín Acosta y Manuel Ancizar, y dotar así a la Biblioteca Nacional de un gran repositorio conocido como la Biblioteca de Obras Nacionales que desde la geografía, la historia, los soportes de documentos oficiales y la literatura legal y política, hiciera las veces de punto de partida para la historia del progreso, la formación nacional y la consolidación estatal. De manera que hubo un proyecto coleccionista consciente y colectivo que buscó hacer de algunas de las colecciones privadas un recurso indispensable para el Estado.

Este proyecto coleccionista fue consustancial al proyecto de colonización interna y a los esfuerzos civilizatorios de las élites criollas, en la medida de que los más importantes coleccionistas, como Anselmo Pineda, tuvieron la doble función de adelantar la colonización interna y exploración de los territorios, así como la de configurar una representación y narrativa nacional a través de la recolección, clasificación y disposición del universo documental condensado en esas colecciones privadas. Dicho esto, en la colección Pineda, a la luz de su biografía, se manifiesta la yuxtaposición del proyecto de colonización interna del territorio con el proyecto coleccionista de la élite ilustrada de la República de Nueva Granada, que tras la independencia comprendía a Ecuador, Venezuela, Panamá y Colombia, y pasó a llamarse la Gran Colombia.

Así emerge el carácter indudablemente político del coleccionismo, pues este no solo fue una práctica ilustrada con los fines ya mencionados, tampoco fue solo una manía compulsiva de algunos, sino que fue un instrumento político de promoción y defensa de una determinada visión nacionalista a la medida de su autor y de su red social. Es decir, el coleccionista, especialmente Anselmo Pineda, que recopila, ordena y cataloga su colección, también termina por manufacturar una poderosa arma de guerra oponible a otros proyectos nacionales en competencia y a la que debe defender mediante el debate público en periódicos, tertulias informales y discursos en el senado de la república, en búsqueda de suficiente legitimidad para elevar su colección privada al estatus de archivo de Estado, tal y como lo demostrará este artículo.

Siendo así, es vital señalarle al lector que el interés de este artículo versa más sobre el coleccionismo de Anselmo Pineda que sobre su colección propiamente, no obstante, la propuesta de investigación que se mostrará apunta a relacionar la colección con sus condiciones de

posibilidad mediante el abundante epistolario que el coronel premeditadamente decidió conservar para su estudio histórico. Las Humanidades Digitales y las Geohumanidades Digitales ofrecen una especial forma de análisis apropiada para el estudio de un especial y voluminoso corpus de epístolas compuesto por 3613 documentos personales que serán procesados con diversos algoritmos diseñados por el investigador e implementados en el lenguaje de programación Python. Se explicará en detalle el proceso en el apartado sobre la metodología.

2. Antecedentes

Sobre el coronel Anselmo Pineda se han escritos contadas investigaciones con diferentes niveles de profundidad historiográfica, por un lado, existen las biografías apologéticas mayormente publicadas a comienzos y hasta mitad del siglo XX, cuya característica es que dan al lector una imagen de Pineda coherente con los valores cívicos y republicanos. Entre estas encontramos *La Biografía de Anselmo Pineda* (León Gómez, 1907), y *Coronel Anselmo Pineda* (Giraldo, 1955). Por otro lado, existen las biografías con una narrativa histórica más rigurosa entre las que están *Anselmo Pineda* (Moreno de Ángel, 1981); *The Struggle for Power in Post-Independence Colombia and Venezuela* (Brown, 2012), y dos tesis de pregrado: *La Biblioteca de Obras Nacionales Formada por el Coronel Anselmo Pineda Como un Aporte a la Formación de la Nación Colombiana*, (Pardo, 2005) y finalmente la tesis *Vida y Obra del Coronel Anselmo Pineda. Un Estudio del Coleccionismo y las Redes Sociales en Nueva Granada Durante el Siglo XIX* (Ortiz, 2016).

Cabe resaltar que solo los últimos dos trabajos académicos emplean como fuentes primarias la correspondencia del coronel Pineda, pero únicamente el último trabajo comprende todo el epistolario encontrado hasta el momento en los repositorios de la Biblioteca Nacional de Colombia. El presente artículo introduce también la correspondencia del coronel, conservada en otros archivos colombianos como el Archivo Central del Cauca, Tomas Cipriano de Mosquera; el Archivo de la Universidad EAFIT; el Archivo Histórico Cipriano Rodríguez Santamaría - Universidad de la Sabana; el Archivo Histórico Universidad Nacional de Colombia y, de la sección de Libros Raros y Manuscritos, el Archivo Julio Arboleda de la Biblioteca Luis Ángel Arango. Por último, es necesario destacar que este artículo hace parte de los resultados de varios años de investigación y trabajo de archivo que, en adición, en 2019 recibió una beca de investigación del Ministerio de Cultura de Colombia. Con todo, la investigación aún se encuentra inacabada dadas las varias aristas y niveles de profundidad para el análisis del objeto de estudio y procesamiento de las numerosas fuentes.

3. Breve biografía del coronel

Anselmo Pineda nació en abril de 1805, en El Santuario, Antioquia, para entonces perteneciente a la jurisdicción de Marinilla, motivo por el cual ha existido confusión sobre su lugar de origen. Con 17 años, el joven Pineda fue remitido por su padre a estudiar jurisprudencia en el Colegio Mayor Seminario de San Bartolomé en Bogotá, pero como varios de sus contemporáneos abandona la academia en busca de un oficio que le permitiera iniciar una carrera en el Estado. Es así como por intermedio de su coterráneo y para el momento Secretario del Interior, José Manuel Restrepo, obtiene el cargo de ayudante archivero de la Secretaría del Interior para una año después ser promovido a oficial escribiente de la Secretaria de Hacienda. Ambos cargos son determinantes en la trayectoria del joven Pineda, pues al entrar en contacto con las desordenadas reservas documentales de la naciente república, termina por motivarse a iniciar el coleccionismo documental, dice Pineda en 1848: "adquirí el hábito importante del arreglo de papeles de un archivo, ya desde entonces el convencimiento íntimo, por el desorden en que se hallaba aquel y por el ímprobo trabajo que costaba dar con algún antecedente" (RM 630, 1848, folios 24-27)

Sin embargo, su carrera en los archivos estatales se vería brevemente interrumpida por un evento que obligaría a su escape rumbo a Antioquia en compañía de su entrañable amigo Mariano Ospina Rodríguez, quien se vio envuelto en la llamada conspiración septembrina de 1828, en contra de Simón Bolívar. En 1829, Pineda es nombrado por Manuel Antonio Jaramillo en el cargo de oficial archivero de la Secretaría de Gobierno de la provincia antioqueña, pero duraría poco en el cargo debido a su incorporación a las huestes del general José María Córdova conocidas como el Ejercito de la Libertad y que tenían como propósito enfrentarse al gobierno central de Bolívar (Pineda, 1831, págs. Pág. 1-2). El conflicto regional escaló hasta convocar a los dos ejércitos en el campo de batalla de El Santuario en 1829.

El resultado de la contienda dejó diezmado y acorralado al Ejército de la libertad, al general Córdova muerto por ejecución sumaria (Brown, 2012, cap. 4) y a nuestro personaje con graves heridas de bala que, de no ser por la ayuda del hermano menor del general Córdova, Salvador Córdova, hubiese tenido el mismo destino. Varios meses después de su recuperación y tras el indulto otorgado por Daniel O'Leary a los excombatientes en 1830, Pineda fue nombrado interventor de la Tesorería de Antioquia (Pineda, 1831, pág. 2), no obstante, las secuelas del conflicto de El Santuario estaban lejos de acabar y las relaciones de varios implicados en la contienda apenas comenzarían. Solo un año más tarde, en 1831, Pineda fue puesto en la cárcel acusado de conspirador e inepto en su cargo, pero tras fugarse se incorpora a las tropas de Salvador Córdova, esta vez para una nueva campaña militar en contra del gobierno central de Rafael Urdaneta (Pineda, 1831, págs. 7-8).

Una vez depuesto el presidente, inicia la persecución y exilio de los bolivarianos radicales (Brown, 2012, cap. 6), dando lugar a una reconfiguración de las redes de poder regionales en la que Pineda se beneficiaría. Con el patronazgo de José María Obando, ministro de guerra, Pineda fue restablecido en su puesto en la Tesorería de Antioquia y, en 1832, incorporado al ejército regular del gobierno central en donde le fueron reconocidos los rangos alcanzados en el Ejercito de la Libertad. Anselmo Pineda no solo se vio beneficiado en lo que respecta a su carrera militar, también comenzó a establecer importantes relaciones personales con la élite payanesa al contraer matrimonio con la viuda del prohombre de la independencia Pedro Acevedo Tejada. Esta nueva relación no solo le daría mejor estatus al antioqueño, también le daría los medios sociales para cimentar relaciones de cooperación con coleccionistas ilustrados del Cauca[1].

Pineda dedicó los siguientes 7 años al intercambio coleccionista con amigos como Antonio María Gutiérrez, quien le siguiere tener buenas relaciones con los correistas y "con este método para que lleves al cabo tus Colecciones" (RM 435, 1843, folio 150-160) y Tomás Cipriano de Mosquera, con quien compartía la afición botánica y naturalista (Carpeta 21, Pieza 106, folio 18133; RM 447, 1834, folio 86), además se concentró en la fundación de sociedades de instrucción, colegios e instituciones para la educación de niñas (RM 446 folio 92; RM 446 pág. 127; RM 445, folio 376). Sin embargo, la reconfiguración de las redes de poder del gobierno central, sumada a un ambiente político volátil y una tendencia a las armas devino en un nuevo conflicto bélico conocido como la Guerra de los Supremos. En este conflicto José María Obando, aprovechando la insurrección promovida por el cura Francisco Villota en Pasto por el cierre de ocho conventos, se levanta en armas en contra del presidente José Ignacio Márquez, por lo que fueron enviados el general Pedro Alcántara Herrán y el capitán Anselmo Pineda, que para entonces se ocupaba del arreglo del archivo general del ejército granadino (carpeta 35, Pieza 3, folio 10260), a pacificar la provincia del Cauca. En esta campaña la función de Pineda consistió en administrar las finanzas del ejército por lo que fue ascendido a tesorero de guerra (Carpeta 34, Pieza 25, folio 11346), y aunque no poseía conocimientos contables hizo una formidable labor en la organización y control de los recursos de campaña (Carpeta 34, Pieza 33, folio 11354), pero inconforme con las dificultades en su labor (Carpeta 34, Pieza 34, folio 11355; Carpeta 35, Pieza 11, folio 10268) solicitó un reemplazo y también ser colocado en primera línea de combate (Carpeta 35, Pieza 8, folio 10265). Una vez en el campo de batalla tuvo un destacado desempeño en la batalla de Chuaguabamba por lo que fue ascendido a sargento mayor.

1 Los principales colaboradores en Popayán fueron la familia Arroyo y Caicedo, pero también contó con el apoyo de los Arboleda y Mosquera.

Al levantamiento fueron sumándose caudillos de todas las provincias en oposición al gobierno central, incluyendo a Salvador Córdova en Antioquia (RM 439 Folio 74; RM 444, folio 87; RM 446, folio 64), motivo por el cual Pineda fue enviado por Márquez a solicitar apoyo al presidente de Ecuador, José María Flóres, así como también ordenó a Tomás Cipriano de Mosquera a unirse a Pedro Alcántara Herrán en el sur. Tras la victoria, Pineda y Mosquera fueron enviados a Antioquia para enfrentar a Córdova (Carpeta 53, pieza 45, folio 13471), quien al ser derrotado fue ejecutado por Mosquera, por su parte Pineda fue remitido de vuelta al Cauca con la misión de perseguir remanentes de guerrillas opositoras (Carpeta 84, Pieza 63, folio 14407; Carpeta 84, Pieza 64, folio 14408). El fin de esta guerra no solo cierra un ciclo de tensiones presentes desde la guerra de El Santuario, también marca el momento en que Pineda constituye nuevas lealtades e inicia una carrera política, coleccionista y militar en ascenso (Ortiz, 2015, pág. 47).

En el siglo XIX la esfera política, militar e intelectual suelen sobreponerse de modo que resulta imposible encasillar una figura de la época en alguna de esas categorías separadamente, por tal motivo, al mismo tiempo que Pineda mejora su posición social y asciende en el ejército también se va perfilando como un político de influencia. Es así como para dar por terminada la Guerra de los Supremos es comisionado a negociar una salida pacífica con el supremo de Panamá, Tomás Herrera, lo que consigue con éxito y es nombrado coronel de infantería por el presidente interino y pariente Domingo Caicedo, quien además habría facilitado el matrimonio de su sobrina María Josefa Valencia con Anselmo Pineda varios años atrás después de combatir hombro a hombro al gobierno del bolivariano Rafael Urdaneta. La carrera política de Pineda cobra forma con su elección como representante de Antioquia en 1843, pero es nombrado gobernador de Panamá poco tiempo después por el presidente Pedro Alcántara Herrán, motivo por el cual debe abandonar su curul en la Cámara de Representantes hasta su retorno en 1848.

En Panamá, Pineda puso en marcha proyectos de educación popular a través de escuelas-taller para el fortalecimiento del comercio de exportación; también mediante publicaciones periódicas como la Cartilla Popular, la que gozó del apoyo de la élite intelectual y política local y extranjera, es el caso del militar, intelectual y coleccionista Joaquín Acosta, quien al respecto comenta:

> *acabo de recibir el N. 2 de la Cartilla Popular [...]. Diríjase pues usted en mi nombre a Mr Hormes Secretario de la Sociedad de Educación del Liceo de Nueva York que él le procurará libros elementales escogidos por las escuelas por precios ínfimos y solo calculados para reembolsar una pequeña parte de los gastos de impresión y papel-- Hoy no tengo lugar de buscar el cuaderno que me pide, pero seguiré por el otro correo. He leído su carta al Sr Ordoñez en presencia de varios señores interesados en sus proyectos. Yo por mi parte nada puedo sino suscribirme a la Cartilla más como no he visto sino el N.2 ignoro el precio de la suscripción para remitirle (Acosta, RM 439, folio 313).*

Estos proyectos consistieron también en la fundación de la Sociedad Filantrópica de Panamá que contó con el respaldo de otras sociedades filantrópicas granadinas[2] y de influyentes amigos como el cura Antonio María Gutiérrez, quien le advirtió a Pineda sobre el rol político y la poderosa influencia de las sociedades, dice Gutiérrez:

> *El primero entraremos en los trabajos de Chagres, y ya te he dicho que no nos acompañas porque las filantrópicas, tienen ya y tendrán la parte influyente en las elecciones i como que he oído con disgusto que por allá trabajan bajo tus auspicios, por el B.M.O. [para referirse a Mosquera] hace para presidente pobre patria si tendrás en tus ultimas convulsiones un Maximiliano que te arranque las entrañas (RM 446, folio 143).*

Cabe señalar que el coronel Pineda no era un novato en este tipo de proyectos, dado que ya contaba con experiencia en la fundación de sociedades y a él le eran reportados con frecuencia los avances de sociedades filantrópicas en Antioquia en las que participó como fundador en años anteriores[3].

Expuesto así, es evidente la inseparabilidad anotada entre el ejercicio político, militar e intelectual de Pineda que se materializó en su Biblioteca de Obras Nacionales. En consecuencia, el coleccionismo respondía a intereses específicos de un nicho social ubicado en un determinado espectro político, pero también a un particular y singular proyecto civilizatorio que, en el caso específico de Pineda, consistía en desarrollar las bases para el progreso nacional que fundamentalmente buscaban educar a las masas en actividades prácticas para el comercio, la construcción de infraestructura y la exploración de las zonas de frontera inexploradas y alejadas del poder institucional del Estado como Panamá, Túquerres y Caquetá. En este sentido, Anselmo Pineda a pesar de ser uno de los padres fundadores del partido conservador, no tuvo como prioridad la enseñanza moral y si la educación práctica sin distinción de género, lo cual expresa el talante intelectual del coleccionista y su postura política

2 A modo de invitación Pineda recibe la siguiente comunicación de la sociedad filantrópica de Medellín: *"No creo demás indicar a U que en la actualidad tengo la dicha de pertenecer a la respetable y grande sociedad de instrucción primaria de esta capital, y también correspondo a su consejo administrativo que dignamente preside el muy ilustre señor Arzobispo y distinguido ciudadano José Manuel Mosquera, y yo desearía que la de esa provincia se pusiera en comunicación con la de esta capital y se estableciera entre todas las asociaciones de esta clase una marcha igual, acorde, constante y sostenida en la propagación de las escuelas de la enseñanza general".* (RM 441, folio 105)

3 Una comunicación de Elías Gonzáles a Pineda sobre los proyectos de la sociedad filantrópica en Salamina, Antioquía, dice: *"La sociedad filantrópica se reunió el día 4, i todas las noches se reúne a discutir varios proyectos que se han presentado cuales son la supresión de billares, la corrección de niños, una contribución para alumbrado, i gastos de escritorio, un reglamento interno que me mandó ud uno i últimamente estamos ensayando la ley que dispone se nombre un cabildo parroquial"* (RM 446, folio 109)

difícil de encasillar, muy semejante a la figura de Simón Rodríguez, a quien conoció durante su insospechado paso por Caquetá cuando Pineda fungía como prefecto[4].

De conformidad con esos presupuestos identificados en la visión de progreso de Pineda, durante su gobernación en Panamá, este convenció al presidente Herrán de la conveniencia de la construcción del canal en alianza tripartita de Nueva Granada, Francia e Inglaterra (Carpeta 21, Pieza 102, folio 18129; Carpeta 21, Pieza 103, folio 18130), pero tras el fracaso del proyecto este renunció al cargo y con su nombramiento como prefecto de Caquetá y luego como gobernador de Túquerres, emplea de nuevo esos instrumentos de colonización interna practicados en Panamá, esto es: construcción de infraestructura (Carpeta 47, Pieza 117, folio 19974), control del contrabando (Carpeta 41, Pieza 136, folio 21994), convocatoria de colonos con exención de impuestos y adjudicación de tierras baldías (Pineda, Pieza 469, 1845, folios 103-104), puesta en marcha de escuelas-taller sin distinción de género para el artesanado (RM 622, Pieza 29) y exploración de la geografía selvática. En este momento, Pineda conoce al maestro de Simón Bolívar, el célebre Simón Rodríguez (1990), con quien tuvo la oportunidad de desarrollar un proyecto civilizatorio único basado en la colonización del territorio efectuado por ciudadanos con habilidades manuales -agrícolas y artesanales – capaces de auto sustentarse y contribuir al desarrollo de la nación. En particular, se propusieron, en primer lugar, enseñar en las escuelas-taller varias técnicas de carpintería, agricultura y construcción, así como aritmética, civismo republicano, gramática y retórica, y en segundo lugar, moral y catecismo, tal y como lo propuso Rodríguez, pues se trataba de una educación a la medida de la realidad americana.[5]

4 La colaboración entre Rodríguez y Pineda al respecto del proyecto educativo y de la exploración de la geografía fronteriza, le cuenta el maestro a Pineda: *"No escribiré a usted largo, porque se me olvidó el día del correo, y la persona que lleva ésta a Pasto la está esperando para ponerse en talones. La casualidad ha traído aquí un médico naturalista suizo, que anda explorando, y me ha hecho el favor de dar algunos remedios a Manuelito. Pasó para Barbacoas y va al Puracé a analizar las aguas del río Vinagre. Hoy debe estar en cerro de Cumbal. No hay más noticias del País, y en las de Santa Fe corre que el General Mosquera es Presidente de la República y que su hermano es Arzobispo. Flores está en Norte América con un Ejército de mil demonios. Roca está haciendo confesión general. Los angloamericanos se han tragado a México como un pastelito. Yo estoy bueno. El doctor Orjuela ha pasado con su esposa de Gobernador de Barbacoas. Hasta el correo que viene."* (Rodríguez, "*Extracto sucinto de mi obra*, 1954, pág. 376). Y sobre los fondos solicitados por Pineda para la manutención de Rodríguez, Escribe Emeterio Gómez: "*Para el establecimiento del señor Rodríguez se ha adelantado cuanto ha sido posible*" (RM 446, folio 192)

5 Son varias las correspondencias entre Pineda y el presidente Mosquera sobre la llegada del educador y sobre la solicitud de fondos para financiarlo. Pineda anuncia la llegada de Rodríguez así: *"solo he regresado p[ar]*[a] *despachar la correspondencia, y asegurarle un alojamiento cómodo al ilustrado patriota* Simón Rodríguez *antiguo ayo y confidente del G[ene]ral* Bolívar *¡Ah! no le hablaré nada de esta respetable sujeto, porque recuerdo, que lo hice con vivo interés en el año pasado y V[uestra] E[xcelencia] no me contestó nada, enteram[en]*te *nada, le he pagado parte de su su viaje y en el proccimo d[iciem]bre, después que me deje bien establecido aquí la escuela normal seguirá conmigo a Bogotá voy a llevar a V[uestra] E[xcelencia] esta reliquia cuyo merito sobresaliente se conocerá tratándolo y viéndolo [Inserto: ocupado] en la grandiosa obra de dar luz al entend[imien]*to *embrutesido; desde q[u]*e *he tratado y conocido al s[eño]r Rod[rígue]z hasta he renunciado a la pación de vicitar la Europa, y el tiempo q[u]*e *había de consagrar en esto pienso ocuparlo recibiendo lecciones de este Rusó [!]. No crea q[u]*e

Si bien Pineda desarrolló un proyecto colonialista singular durante su ejercicio en cargos públicos en zonas de frontera, este no fue el único en emprender la colonización interna al explorar y documentar personalmente vastas selvas y ríos inexplorados, atraer nuevos pobladores y utilizar las sociedades filantrópicas para su educación a la luz de la ideología del progreso decimonónico, es el caso de su cercano amigo y dedicado colonizador interno Elías González con quien intercambia numerosas comunicaciones respecto a planes de fundación de poblaciones en Tolima y Huila[6], y también sobre las actividades de las sociedades filantrópicas en la comunidad (RM 446, Folio 109; RM 446, Folio 127), comenta González sobre el trabajo de Pineda: "complacido al ver que mi más querido amigo es quien marcha a la vanguardia en la noble e interesante empresa de ilustrar y de moralizar las masas populares de su patria" (RM 439, Folio 126).

Pineda también se apoyó constantemente en misioneros jesuitas como José Layner, quien emprendía viajes a través de las selvas del sur de Colombia para evangelizar indígenas, y cuyos reportes le servían a Pineda para conocer e incorporar a su colección diarios de viaje sobre la geografía todavía indocumentada (RM 444, Folio 201), lo que le mereció, según el propio Anselmo Pineda ante el congreso, el reconocimiento de autoridades en la materia como el geógrafo y militar Agustín Codazzi, quienes reconocen la valiosa información aportada por esos documentos, dice el coronel Pineda sobre carta de Codazzi: "en que manifiesta que la "colección Pineda" suministra conocimientos nuevos i mui importantes sobre la jeografia de territorios que nadie ha recorrido ni descrito tales son los Andaquies i Caqueta".[7]

No solo la agencia colonialista del coronel Pineda expandió sus alcances coleccionistas, también lo hizo para coleccionistas de raros artículos de historia natural como su colaborador, antes enemigo en el campo de batalla de El Santuario, Daniel O'Leary quien le solicita a Pineda: "Si en aquel distrito nuevo para la civilización encuentra V. algunos objetos de Historia natural que llamen la atención, suplico a V. los compre para mí, avisando de su clase y valor. Algunas muestras de fósiles y minerales serán muy apreciadas" (Moreno de Ángel, 1981, p.67)

A su regreso a Bogotá en 1848, el coronel Pineda ocupa su curul en la Camara de Representantes, y allí se opone a la expulsión de los jesuitas por considerarlos indispensables

le ecsajero, mi g[ene]ral, estoy encantado con el s[eño]r Rod[rígue]z y V[uestra] E[xcelencia], V[uestra] E[xcelencia] puede dejar monumentos perdurables. [...] No por esto mis afanes se han contraído únicamente a este punto, sino que mis atenciones se han dirijido a otros varios medios de adelantar estos pueblos moralisar y formar constumbres públicas y escuelas y caminos, he aquí programa" (Carpeta 41, Pieza 140, folio 21998)

6 Por la correspondencia de González con Pineda se puede establecer la cercanía del primero con el misionero jesuíta José Layner con el cual efectivamente colaboró en Antioquia en tareas civilizatorias. (RM 444, Folio 201); Además, González a su llegada a Neiva en 1842, le comenta a Pineda: "*Hace 20 días que llegué á esta con el objetivo de fundar un pueblo i ya tengo 200 vecinos cabezas de familia, voy a dar como 2 anegadas de monte, i como una legua de camino hecho*". (RM 446, Folio 100)

7 Memorial dirigido al congreso. No hay registro de la carta de Codazzi dirigida a Pineda. (RM 640, Pieza. 58)

para la causa civilizatoria. Durante los siguientes años se dedicó a asuntos personales[8], al intercambio de documentos, arreglo de la colección y al debate público mediante publicaciones sobre la importancia de la colección Pineda para la república (RM 640, Pieza 60). Gestión que procuró la legitimidad de la colección documental entre la élite intelectual y política con artículos de autoría propia o de terceros para convencer de la conveniencia de su compra por parte del congreso colombiano[9]. Resulta imprescindible señalar que esta fue una ardua tarea con encendidos debates sobre la relevancia de la colección, al respecto Pineda señala en comunicación al congreso:

> *[…] En cuanto a la importancia de la colección, apelo al testimonio de los que la han visto, la comisión nombrada por el cuerpo legislativo; y los que ni a estos, ni a los otros quieran creer, suspendan su juicio hasta la próxima reunión imparcial del congreso en que los señores Maldonado, Miranda y Paz habrán acabado su trabajo, a ellos me refiero al público imparcial, a los amigos que tan generosamente me han franqueado algunos documentos; y para decirlo de una vez, a los tres encargados de negocios de Francia i a la Gran Bretaña i al señor Bucconi encargado de la numeración Romana, que han hecho más aprecio de mi penoso trabajo que el recién venido que en un virulento articulo ha opacado mi colección basando su artículo sobre supuestos falsos unos, y equivocados otros (RM 640, Pieza 60)*

Este esfuerzo por llamar la atención hacia la colección documental y persuadir a la opinión pública de su relevancia, respondió también a otras circunstancias personales que obligaron a Pineda a publicar los catálogos y a buscar, incluso en Estados vecinos o europeos, el apoyo que con tanta dificultad obtendría en Colombia[10]. Esa contradicción entre ofrecer la colección al público o conservarla para sí, dado que a los ojos del coleccionista todavía permanecía inacabada, pone de manifiesto el nivel afectivo del coleccionismo y el coste personal de llevar a cabo esta empresa, por lo que el coleccionista manifiesta:

8 En carta con María Josefa Valencia, Pineda se refiere al proceso de divorcio iniciado por este y posible gracias a la Constitución liberal de 1853: *"Las diarias i multiplicadas ocupaciones de mi colección me habian impedido adelantar el juicio de divorcio provocado por ti i entablado por mi por exitacion tuya, según se vé de los documentos que reposan en mi poder. Aquellas dificultades han desaparecido i al presente me encuentro en estado de continuar esta lucha sin tregua, de 19 años que tal vez, i sin tal vez, vá a finalizar con mucha vida"*. (RM 622. Pieza. 126)

9 Dice el redactor del periódico oficial a Pineda: *"me excita para que le recomiende en el periódico oficial la importante empresa que Ud ha acometido de formar la colección estadística e histórica de los documentos celebres e importantes que se han publicado en la Nueva Granada desde una época remota Con mucho gusto haré la recomendación de su colección en nuestro periódico oficial"* (RM 444, folio 244). Son varias las cartas que demuestran el respaldo de amigos en tertulias informales y publicaciones periódicas a la colección (RM 439, folio 81; RM 445, folio 365; RM 445, folio 367; RM 437, folio 33)

10 Fueron varias las comunicaciones que demuestran pretendidas negociaciones con el gobierno británico, por intermedio del representante de la legación británica en Bogotá, Daniel O'Leary, para la adquisición de la colección Pineda, (Miscelánea 1440, Pieza 8. Biblioteca Luis Ángel Arango). Así como la respuesta negativa del gobierno venezolano a la propuesta del coronel para venderles la colección. (RM 444, folio 245)

> *Jamás había pensado desprenderme de la copiosa colección de documentos oficiales que poseo y de que voy a hablar, adquiridos a costa de mil privaciones desde 1825. Pero repentinamente sin casi sentirlo me encuentro al presente con enfermedades de cuidado, adquiridas en el servicio, que me están inhabilitando para ocuparme en una vida activa, y esta circunstancia fatal, agregada al deber de dar educación á 4 hijos me han determinado con harto pesimismo a publicar los índices de una parte de los documentos que tengo en mi poder y solicitar la aquiescencia de los hombres ilustrados de las 3 republicas en que se dividió la antigua Colombia para generalizar dichos documentos. Estoy persuadido que contando con las luces y la experiencia de los que tengan un mediano conocimiento de los consabidos documentos a la vez que se les puede dar el carácter de utilidad, que es lo que más me ha determinado a hacer la publicación podré desprenderme de ellos con un mediano provecho (RM 630, Pieza 24)*

Finalmente, se realizó la entrega de 1100 volúmenes y con esta la solicitud de baja del ejército por parte del coronel Pineda, pero unos meses más tarde es apresado por sospechas de su participación en la insurrección conservadora instigada por Pastor y Mariano Ospina Rodríguez. Una vez puesto en libertad, Pineda es nombrado custodio y curador de la Biblioteca Nacional por el vicepresidente de turno José de Obaldía.

Llegados a este punto, fueron dos los eventos trascendentales en la vida del coronel, por una parte, logra que su colección sea reconocida y aceptada oficialmente y, por otra parte, termina su matrimonio con la payanesa María Josefa Valencia, lo cual afecta poderosamente sus relaciones con ilustrados de Popayán, pero también abre nuevas posibilidades de relación con la élite costeña después de que contrajo matrimonio con Ana María Danies Kennedy a finales de la década del 50.

En lo que resta de los años 50, Anselmo Pineda se reincorpora al ejército para llevar a cabo el golpe de estado, en el que participaron mancomunadamente liberales y conservadores, en contra del presidente José María Melo y sus políticas económicas favorables hacia el artesanado (RM 447, folios 51-56). Pineda además contrae matrimonio por segunda vez e invierte buena parte de la contraprestación concedida por su colección en la producción de quina y caucho en el Huila y en continuar con su colección para una posterior entrega. Los esfuerzos coleccionistas de Pineda durante este último periodo se sirvieron del cargo que desempeñó en Magdalena como intendente de hacienda nombrado por el presidente Mariano Ospina Rodríguez, quien además le encargó al coronel civilizar, pacificar e insertar en los circuitos económicos a la Guajira (RM 441, Folio 65). Este cargo le permitió a Pineda expandir su poder político, fortalecer la sociedad de fomento a la industria que fundó (RM 440, folios 439, 445) y tener acceso privilegiado tanto a oportunidades de negocio con comerciantes extranjeros para su negocio de quinas (RM 445, folios 243, 245), como oportunidades de

negocio con agentes locales para su parentela (RM 445, Folio 272), así como también conectarse a fuentes documentales inéditas (RM 447, folios 175, 181, 182, 198).

Años después es encargado en el arreglo de los archivos de la Tesorería General del Estado (RM 640, Pieza 114). Al respecto recibe la siguiente comunicación que no solo demuestra el reconocimiento social alcanzado por Pineda en materia de organización y catalogación de archivos documentales, sino que también, en tanto que experto como ningún otro en materia de archivo[11], obtiene la confianza pública para ser encargado de tareas sensibles para el Estado, al respecto Pineda recibe la siguiente carta:

> *[..] en honor de la verdad debo decir a usted que a lo que se queria dar el nombre de archivo en la tesoreria jeneral, es un cuarto donde estaban amontonados en una confusion incomplicable, libros, legajos, documentos de deuda pública de la mayor importancia como se ha visto despues, restos de [ilegible], y en fin objetos de todas clases tan cubiertos de polvo tan revueltos que costaba trabajo creer que aquello hubiera podido ser algun tiempo el archivo de una de las oficinas mas importantes de la Republica. Fui testigo muchas veces, que necesitando el gobierno ó algun particular un dato, por importante que fuera, habia que renunciar a encontrarlo si se inferia que pudiera estar en el archivo, pues ni siquiera se pensaba en este, y decir, tal documento debe estar en el archivo, era lo mismo que decir, no existia. [...] Cuando sali de la tesoreria, ese cáos de papeles tomaba ya forma y usted habia clasificado muchos documentos importantes. Pero lo que no quiero dejar de consignar aqui es el importantisimo servicio que usted ha hecho a la nación desenterrando del polvo documentos de gran valor, tales como esqueletos firmados de vales de manumición, cupones de renta sobre el tesoro y muchos otros de un valor considerable que si hubieran caido en manos menos dignas, como desgraciadamente ha sucedido ya, habrían causado grandes perjuicios a nuestra hacienda [...] Me consta, así mismo, que cuando por falta de fondos en la tesoreria jeneral ó por cualquier otro motivo, no se pagavan sus ayudantes usted les daba adelantado de su bolsillo (RM 640, Pieza 114).*

Es importante cerrar esta condensada biografía del coronel Anselmo Pineda, haciendo hincapié en un aspecto clave de su estrategia coleccionista, consistente en el uso de su prestigio personal y la legitimidad de su colección, para solicitar formalmente a las administraciones regionales la remisión de cuanto documento fuera impreso por estos gobiernos. De manera que ya no dependía de intermediarios que reunieran y le enviaran documentos, pues ya gozaba de una relación directa con los gobiernos locales que destinaban algunos recursos para alimentar su colección como si se tratase de un depósito legal, dice Pineda:

11 Anselmo Pineda, por su larga experiencia en archivos públicos y actividad coleccionista, expresa la urgencia de profesionalizar al archivista: *"Este ramo merece tanta más profesión cuanto es mayor el deseo nacional que se advierte ya en algunos hombres ilustrados; deseo que en todos los países civilizados de la tierra ha llamado su atención"* (RM 630, Folios 24-27).

> *Desde 1865 y aun desde mucho antes que me propuse compaginar y arreglar la nueva Colección adicional que debo enlazar con la otra, dirigi circulares y comunicaciones oficiales a los ciudadanos presidentes de los estados sobre este asunto y lo relacionado con los impresos que se acompañan. Pero se me ha cobrado ultimamente por el oficio que original acompaño con la cubierta, y de seguro seguiran cobrandome por todas las notas oficiales y documentos importantes que para evitar su estrabio vengan con cubierta. Ultimamente han aparecido entre memorias, mensajes, proclamas del Libertador y del General Santander como mil otras piezas importantes de que no tan solamente no tenia noticia, sino que en publicaciones de 1829 habia asegurado y repetido despues no existian. […]. Por estas razones y otras que omito por ser cansado; en atencion al absoluto abandono que he hecho de mis negocios particulares desde hace tanto tiempo; por el desesperante anhelo de complementar este aservo publico en pro de mi patria (RM 640, Pieza 111).*

De la misma manera y no menos importante, el coronel le solicita a la oficina de correos que no se le cobre el envío de documentos pues se trata de un asunto de importancia oficial por las siguientes razones:

> *Primero: Poco más poco menos desde cuando han notado ustedes que con mucha mas frecuencia que antes los funcionarios de los Estados y aun los Presidentes de dichos Estados me remiten, memorias, codigos de leyes y toda una a una las publicaciones oficiales que se hacen en las Capitales. Segundo: Si han notado ustedes que viniendo comunicaciones oficiales con alguna frecuencia relativos a la segunda Colección de Obras Nacionales que hubiera ya compajinado si tuviera piezas que tengo que contestarle oficialmente y si a pesar de palpar que es sobre asuntos oficiales me han cargado el porte de los impresos que se remiten al Estado soberano del Ystmo (RM 640, pieza 112).*

Finalmente, Anselmo Pineda entrega una segunda parte de su biblioteca en 1868 y se retira a su casa en Fusagasugá, Cundinamarca. Muere en 1880 dejando las huellas de una vida de guerra, entrega a la república y a la actividad intelectual marcada por un pleno convencimiento patriótico cristalizado en su colección.

4. Metodología

En ese apartado se detallará la metodología empleada para analizar el epistolario del coronel Pineda ofreciendo una nueva perspectiva para leer y procesar un corpus documental voluminoso. Son varios los componentes que hacen parte del proceso y varias las relaciones entre estos, pues los distintos enfoques para el tratamiento de datos son capaces de generar nueva información que resulta provechosa para otros procesos de cómputo. Es el caso del modelo final de aprendizaje automático que emplea atributos generados en cada

uno de los procesos de exploración, georreferenciación, indicadores relacionales e indicadores de minería de textos aplicados al corpus.

Antes de explicar cada proceso, vale la pena comentar el procedimiento de captura de datos que se realizó de las 3613 cartas que hasta ahora componen el epistolario Pineda. Este proceso básicamente consolida en una base de datos la información de cada carta, tal como: remitente; destinatario; lugar y fecha de elaboración; descripción del contenido; transcripción de al menos 500 cartas y una columna con un código binario que servirá para identificar la relación de la carta con el coleccionismo y también como etiqueta de evaluación cuando se clasifiquen los colaboradores coleccionistas.

Una vez consolidada la base de datos, se exploró la distribución de los datos mediante estadísticas descriptivas básicas como frecuencia de remitentes y destinatarios, frecuencia de contactos epistolares relacionados y no relacionados con el coleccionismo, frecuencia de términos y su visualización sobre un eje temporal. Más tarde se llevó a cabo la exploración de las redes sociales del coronel mediante la generación de gráficos de red divisibles en duraciones temporales, pero que para el presente artículo se optó por un grafo de la red completa, aun así, se pueden distinguir interacciones interesantes. El análisis de interacciones permite también producir algunos indicadores de centralidad e intermediación útiles para identificar los nodos más importantes en la topología de la red, y además útiles para el modelo de aprendizaje automático posterior.

Simultáneamente, se procedió a georreferenciar mediante el geoetiquetado automático de la toponimia del lugar de elaboración de cada documento para producir mapas de distribución espacial del epistolario. Cada mapa comprende la ubicación de los lugares de producción de las cartas dentro de duraciones específicas de tiempo dadas por aquellos momentos de cambios abruptos en términos relacionales, identificables en el paso anterior y sustentados en la biografía de nuestro personaje. Por último, queda una de las fases más importantes y complejas en este estudio, conocido como Procesamiento de Lenguaje Natural (NLP), que busca producir nuevos atributos derivados de la minería de texto, además de servir para el reconocimiento de entidades (NER) como nombres de personas, lugares u organizaciones y para el cómputo de temas principales dentro de una colección documental. Cada uno de estos procesos permite el desarrollo de diferentes herramientas secundarias como un sistema de recomendación documental, basado en el cálculo de la semejanza (cosine similarity) de vectores numéricos que representan cada documento en tanto que conjunto de palabras vectorizadas según su identidad numérica, y además una interface con los temas principales basada en una colección de diccionarios conformados a partir de conjuntos de tres palabras, trigramas, y en un modelo de bolsa de palabras.

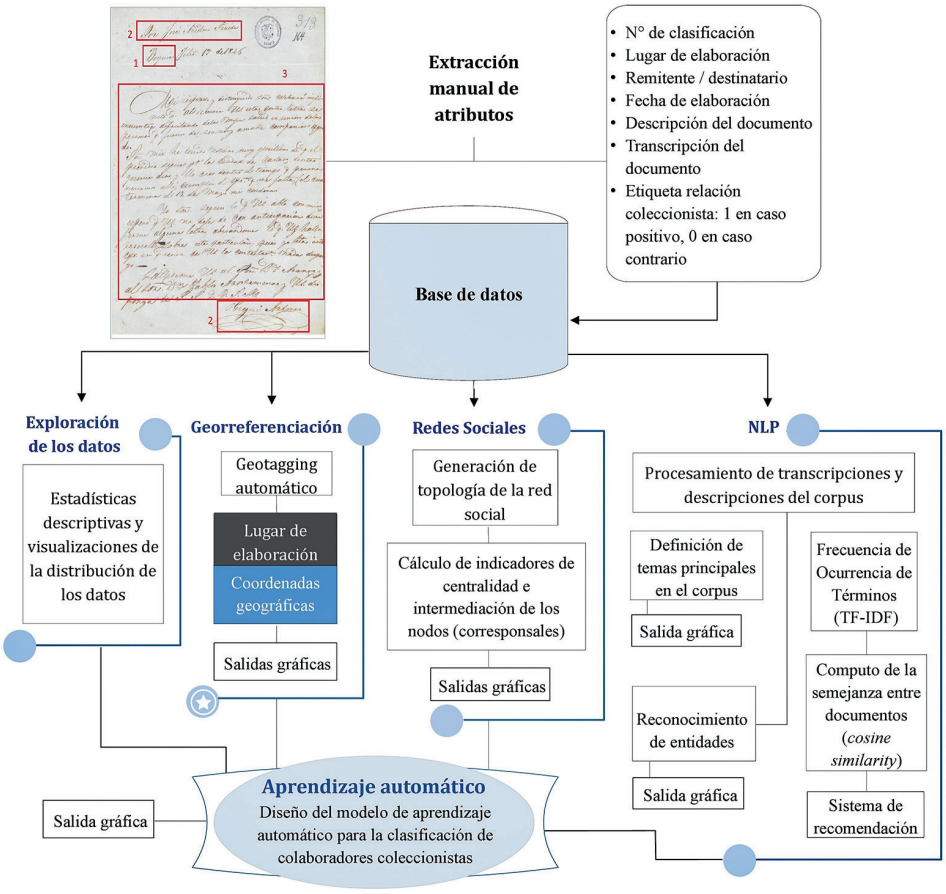


Figura 1. Esquema de la metodología sobre el levantamiento de datos en archivo.

Finalmente, todos los atributos numéricos generados en todas las fases descritas, además de algunos atributos cualitativos originales como la fecha de escritura de la carta, fueron el insumo para el algoritmo de aprendizaje automático que implementa el modelo Bosque Aleatorio (Random Forest) para clasificar cada registro con base en su probabilidad de pertenecer a un grupo u otro: colaborador y no colaborador. La clasificación usa una búsqueda informada de hiperparámetros para encontrar valores óptimos de clasificación, esta búsqueda se compone de una primera búsqueda aleatoria de hiperparámetros y luego de una búsqueda ordenada con los mejores hiperparámetros de la búsqueda aleatoria. Para evaluar la efectividad del algoritmo se usaron las etiquetas binarias insertadas por el equipo de investigación en la base de datos inicial y se computó una matriz de confusión que muestre los errores y aciertos del proceso de clasificación automático.

Las librerías empleadas para el análisis son: I) Pandas para la gestión de la base de datos; II) Seaborn y Matplotlib para las visualizaciones; III) Networkx y Holoviews para generar las redes de individuos; IV) Geopy y Folium para la georreferenciación y visualización web; V) NLTK, Gensim, Polyglot, pyLDAvis y Spacy para el procesamiento de lenguaje natural; VI) Scipy y Scikit-learn para implementar el modelo de aprendizaje automático.

5. Análisis del epistolario con Python

Las técnicas antes descritas permiten diseccionar con sumo detalle el corpus epistolar, de manera que, dada la extensión de un análisis que considere toda la vida de Pineda, en esta sección solo nos concentramos en la época más activa del coronel y relacionada con su coleccionismo, que como ya vimos en el apartado biográfico, tiende a coincidir con el apogeo de su carrera militar y política.

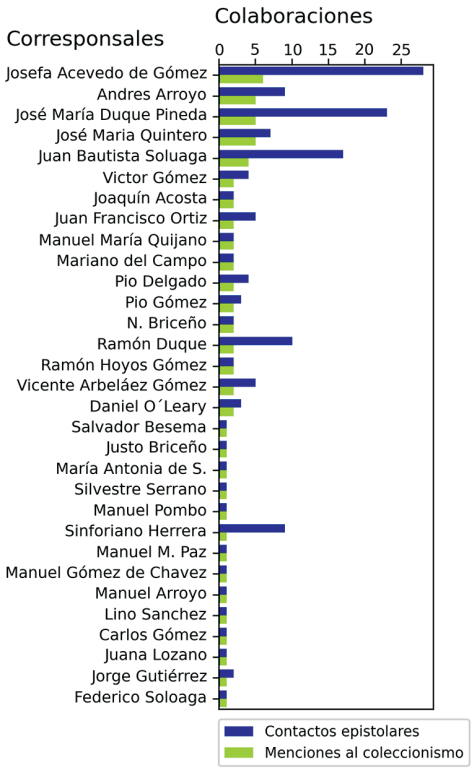


Figura 2. Número de epístolas y menciones al coleccionismo por remitente en 1848-1849.

En este sentido, conforme Pineda ganó mayor protagonismo como figura política y militar, mayor fue su capacidad de convocatoria para solicitar y recibir documentos para la colección, en especial durante los años que precedieron a la primera entrega. Es decir, a medida que la carrera política del coronel iba en ascenso, también lo hacía el número de cartas y, por ende, el número de contribuciones que las acompañaban, no obstante, el significativo esfuerzo del coleccionista no solo radicó en solicitar documentos, sino en persuadir de la importancia de la colección en tanto que archivo de Estado (RM 630, pieza 24).

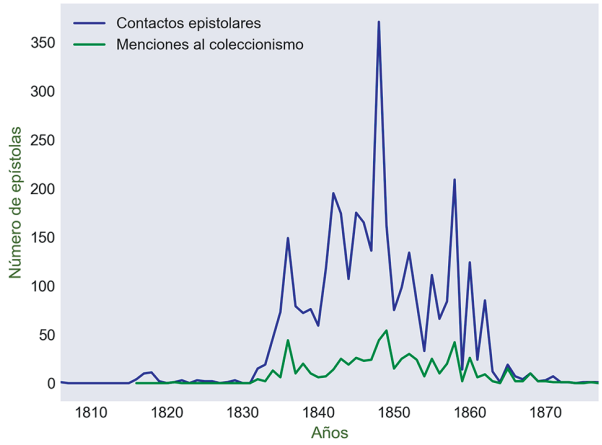


Figura 3. Actividad epistolar de Pineda, por número, y menciones al coleccionismo, por año.

Las gráficas anteriores, dedicadas a la actividad epistolar y coleccionista, muestran la tendencia de que, a mayor número de contactos epistolares, mayor la cantidad de contribuciones a la biblioteca Pineda. Por una parte, en la primera gráfica de barras aparece Josefa Acevedo de Gómez encabezando la lista, seguida de José María Duque Pineda, primo del coronel; Juan Nepomuceno Duque, primo; y otros corresponsales entre familiares y amigos de los cuales sobresalen Joaquín Acosta, Manuel María Quijano, Andrés Arroyo, Domingo Caicedo, Daniel O'Leary y Manuel María Paz. Cabe hacer la salvedad de que si bien durante el periodo entre 1848-1849, se da la mayor actividad coleccionista, antes existieron colaboradores muy importantes como la del cura dominico Antonio María Gutiérrez[12] quien, hasta meses previos a su muerte en 1846, aportó 80 epístolas de un total de 552 cartas que en el epistolario versan sobre el envío de documentos para la biblioteca Pineda.

Por otra parte, el segundo gráfico muestra la actividad epistolar durante toda la vida del coronel Pineda y las menciones al coleccionismo rastreadas con palabras clave como manuscrito, colección, gaceta, biblioteca, cuaderno, cartilla popular, libro, compilación o memorias, entre otros términos recurrentes en cartas que acusan envío adjunto de documentos. Esta gráfica también permite evaluar la asociación entre número de contactos y número de contribuciones, pero además posibilita la identificación del auge simultáneo de

12 Con toda certeza, Antonio María Gutiérrez fue uno de los amigos más cercanos de Anselmo Pineda. El sacerdote fue abogado, teólogo, orador, profesor y senador, pero además fue quien, a su regreso de Jamaica posterior al exilio a causa de su inclinación realista previa a la independencia, reclutó al joven Pineda para el Ejercito de la Libertad de José María Córdova en 1829 (Brown, 2012). Por otra parte, Gutiérrez participó en la fundación de la masonería en Nueva Granada junto a Francisco de Paula Santander entre 1820-1825 y respaldó a Pineda con sus buenas relaciones públicas e influencia política hasta 1846. (RM 446, folios 85-86).

actividad epistolar y coleccionista entre 1848-1849 previo a un abrupto descenso en 1850 y a la primera entrega en 1851.

El estudio de redes sociales aplicado a un corpus de correspondencia personal tiene como principal utilidad la visualización de las interacciones entre sujetos y la representación de su relevancia relacional a través de códigos visuales de color y tamaño. Semejante a un mapa geográfico, un mapa relacional permite ubicar nodos y trazar los caminos o vínculos que los interconectan, así como calcular el grado de centralidad o intermediación de cada uno de los individuos en consideración a los vínculos que posea.

En primer lugar, la centralidad, representada por color, es el coeficiente del número de contactos que un nodo particular tiene en la red, es decir, se basa en el hecho de que nodos importantes o populares tienen mayor número de contactos epistolares. En segundo lugar, la intermediación, representada por tamaño, mide el número de veces en que un nodo específico está presente en el camino más corto entre otros dos nodos en la red, es decir, los nodos con mayor grado de intermediación tienen un rol significativo en la comunicación y flujo de información. No menos importante es la configuración topológica de la red, pues resulta determinante en el acceso de los nodos a recursos e información que, al estar ubicados de manera desigual y asimétrica en la estructura social, poseen grados asimétricos de inserción y posibilidades de acceso a recursos sociales. La red que se presenta en la figura 4, es de tipo egocentrado, dado que el nodo central (ego) aglomera entorno a sí a la mayoría de los vínculos existentes en la red que abarca toda la duración comprendida entre la primera hasta la última carta del epistolario.

Esta red comprende el rango de 1806-1880 y ofrece gran cantidad de información visual, en ella se prefirió destacar con etiquetas los nodos de mayor centralidad. Se observa al ego principal, el coronel Anselmo Pineda, seguido por el general Joaquín Acosta, ambos compartían el proyecto coleccionista privado con propósito público, tal y como se describió en el apartado biográfico, pero también se muestran otros personajes relevantes en la historia del siglo XIX como Tomás Cipriano de Mosquera quien, como se comentó, fue un amigo coleccionista de Pineda en su faceta naturalista y botánica (RM 447, Folio 86, 90-91), se encuentran también Antonio María Gutiérrez, Domingo Caicedo, Pedro Alcántara Herrán y además se muestran otros nodos importantes en esta estructura social reconstruida desde el epistolario, por ejemplo, se observa la importancia relacional de la segunda esposa de Pineda, Ana María Danies Kennedy, quien fue para el coronel la vía de acceso a la élite costeña y la posibilidad de emprender los proyectos del gobierno central para la inserción de las zonas de frontera al circuito económico. Danies también posee el mayor indicador de intermediación observable en la gráfica de barras incluida, seguida por la primera es-

posa de Pineda, María Josefa Valencia, quien fue a su vez la vía de acceso a la élite payanesa décadas antes. Estos altos índices de intermediación, con los que cuentan ambas esposas, confirman la hipótesis sobre la importancia de los vínculos matrimoniales para conectar al coronel Pineda con las élites a las que pertenecieron cada una de estas mujeres, y que terminaron por impulsar tanto la carrera política y militar, como el proyecto coleccionista del coronel Pineda.

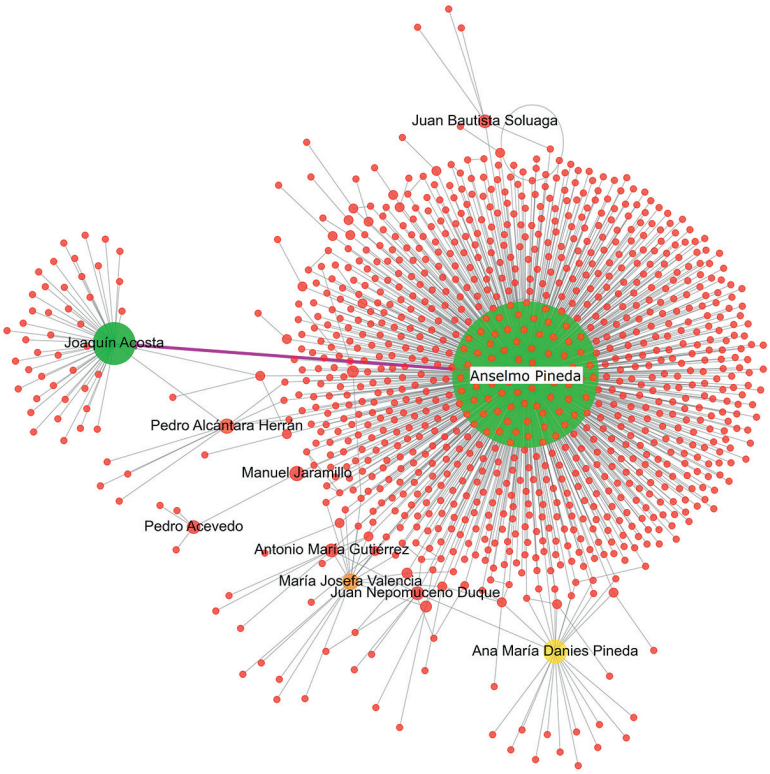


Figura 4. Red social de Anselmo Pineda: Red epistolar-coleccionistas 1806-1880.

La anterior figura, que representa la red epistolar centrada en Pineda, se expande y aclara en el siguiente indicador de intermediación de la figura 5:

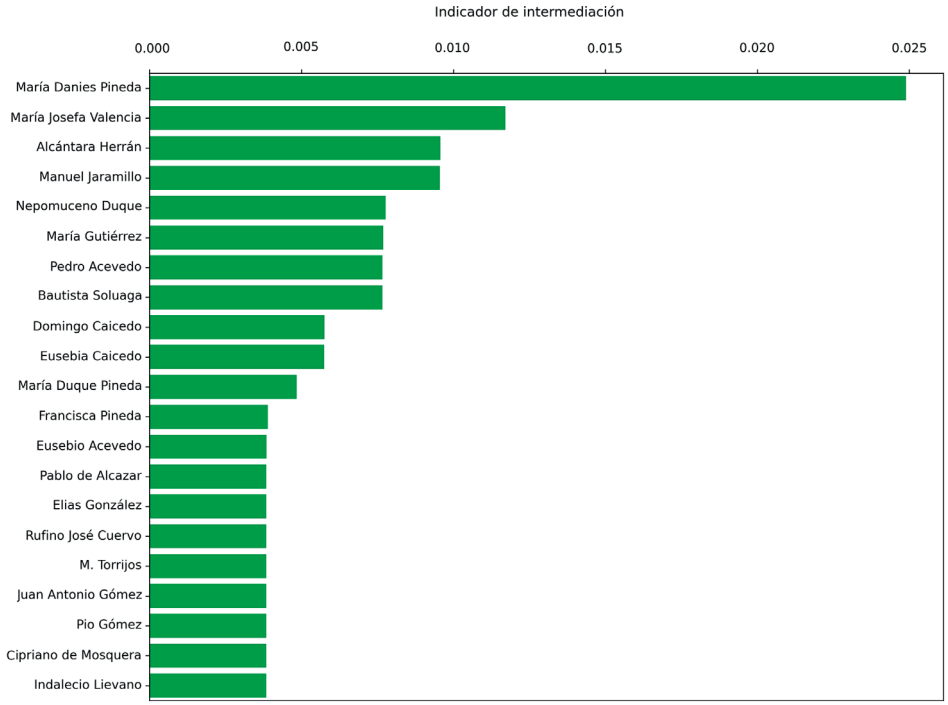


Figura 5. Complemento a la red epistolar a través de sus indicadores de intermediación.

Otro matiz interesante al que se puede acceder mediante esta aproximación en HD es el carácter espacial del epistolario que, a través del lugar de elaboración de las cartas georreferenciadas, permite estudiar la distribución espacial de las redes epistolares y el espacio de circulación de documentos puesto que, como ya se explicó, la correspondencia funcionaba como mecanismo para el tráfico de impresos y manuscritos. En este sentido, un mapa del epistolario hace posible dimensionar el alcance de las colaboraciones coleccionistas que Pineda sostenía con los viajeros a Europa y con proveedores locales.

Al respecto, en respuesta a las solicitudes del coleccionista, un remitente desconocido le cuenta a Pineda desde París:

> *No he olvidado las encarecidas recomendaciones de ud para solicitar las obras i escritos de todo género relativos a la historia de nuestra patria desde su descubrimiento hasta hoy [...] Aquí no es posible conseguir ninguno de los manuscritos u obras inéditas que especialmen-*

> *te me recomendó Ud, como la relación del mando del Virci, Montalvo, la de Quesada, i los demás documentos especiales antiguos i modernos de que Ud me halla en sus instrucciones. Esperaba hallar todo esto en el tiempo durante mi viaje a España, para tener copias auténticas de las interesantes piezas que Ud desea para su bella colección, i de todos los demás documentos que pudiera descubrir; […] Respecto de las obras de Mútis, Cáldas, Lozano. D'Eluyar i demas hombres ilustrados de nuestro pais que Emile trajo, de Bogotá a Madrid, procuraré descubrir su paradero, i formar, si es posible, copias de las menos voluminosas i más interesantes, pues de los escritos sobre botánica e historia natural no será esto fácil porque entiendo que [mutilado] descubrir su paradero (RM 447, folio 130).*

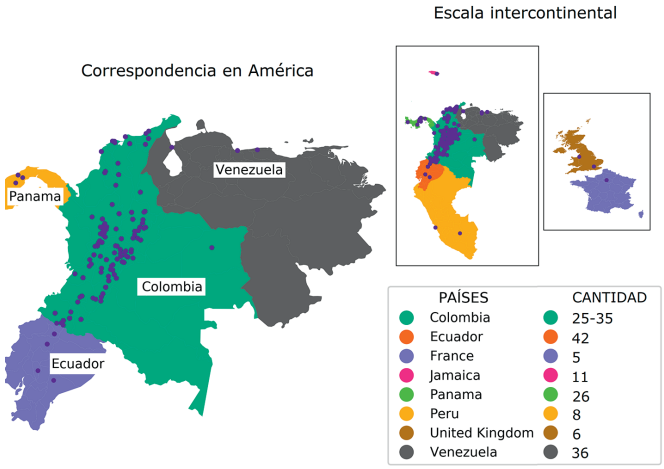


Figura 6. Mapa del alcance espacial en el epistolario de Pineda por países.

En el mismo sentido, el coleccionista comenta que:

> *Se han mandado sacar copias de documentos sumamente interesantes que deben existir en los archivos de Simancas; y otros de Europa; se han solicitado de las provincias documentos que pongan en claro, acontecimientos pasados que el tiempo i la indolencia han sepultado en el olvido; como son los pormenores de la guerra de Pasto desde 1813 hasta la época presente, y otros muchos que son de suma importancia (RM 640, folio 60).*

Una vez señalada la potencia de explotar la dimensión espacial del corpus, podemos focalizar la atención en el procesamiento de otro atributo de las cartas, a saber, su descripción y transcripción. Como se describió en la metodología, el objetivo es descubrir los temas principales en el corpus y discriminar todas aquellas entidades útiles para acceder a otra dimensión del epistolario, todavía en proceso, en lo relativo a personas o lugares referidos

en el contenido de cada carta. Las siguientes gráficas muestran, por un lado, la proporción de entidades reconocidas.

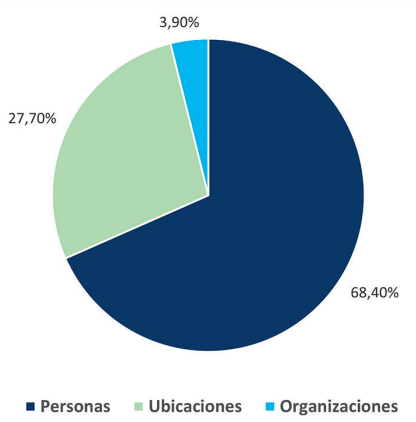


Figura 7. Proporción de entidades nombradas en el corpus.

Por otro lado, se incluye una gráfica del resultado del modelado de temas que muestra los términos más importantes extraídos mediante un popular algoritmo denominado Latent Dirichl*et al*location (LDA), que permite la organización y entendimiento, desde la lectura distante, de los temas subrepticios, pero significativos en una gran colección de textos (Jänicke, 2015). Empero, merece la pena decir que el modelado de temas no garantiza necesariamente que los términos sean fácilmente interpretables por el ser humano, sin embargo, existen métricas para determinar el grado de coherencia, en este caso, un indicador intrínseco basado en que la ocurrencia de un término sobresaliente debe ser precedida por otro término sobresaliente, en otras palabras, que la probabilidad de un término sobresaliente debe ser más alta en un documento si este ya contiene un término sobresaliente, esto es el cálculo de la probabilidad condicional de ocurrencia de un término siempre que ya exista otro término importante en el documento.

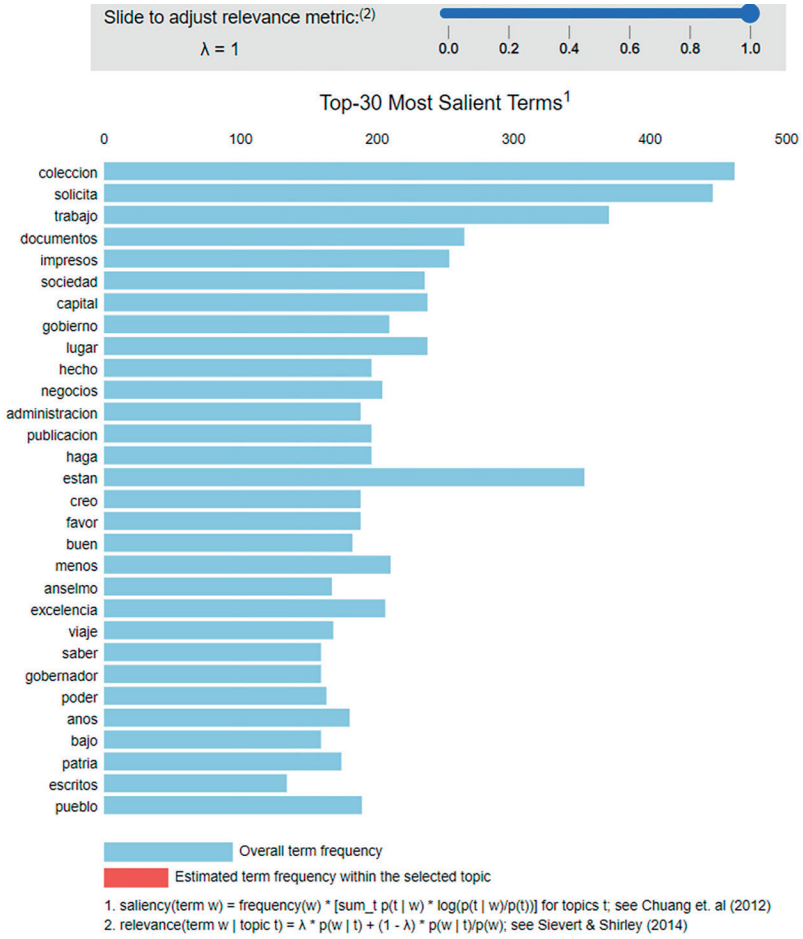


Figura 8. Identificación de términos sobresalientes.

La siguiente gráfica muestra el índice de coherencia *Umass* para todas las iteraciones del modelo, dando como resultado que la mejor coherencia esta alrededor de 30-35 temas por su cercanía al 0, coherencia perfecta.

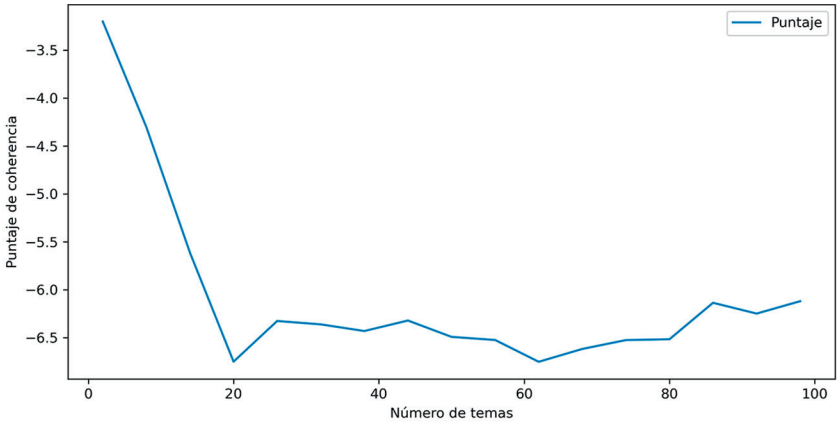


Figura 9. Gráfica del índice de coherencia en modelado de temas.

La lista de términos más sobresalientes incluye palabras como colección, documentos, impresos, gobierno, favor, publicación, gobernador, escritos, patria, viaje y pueblo, además de otros menos claros como trabajo y administración.

Hasta ahora se han mencionado 2 tipos de procesamiento de lenguaje natural para tratar el archivo epistolar de Anselmo Pineda, pero para el siguiente paso que consiste en introducir todas las entradas de la base de datos en un algoritmo de aprendizaje automático, es necesario darle una identidad numérica para hacerlo procesable. Existen varias maneras de surtir esa transformación, en esta investigación se usará la estadística TF-IDF para convertir cada palabra en el valor probabilístico dado por la frecuencia de un término en un solo texto dividida por el número de textos en el que aparece ese término, de manera que las palabras más frecuentes en un idioma y menos significativas, palabras vacías, son filtradas. Adicionalmente, se transforman los demás atributos cualitativos como nombres y lugares a su identidad numérica mediante *one hot encoding*, que busca codificar todas las categorías en una matriz binaria de ceros y unos.

Al modelo de aprendizaje automático supervisado *Random Forest*, elegido por obtener mejores resultados con este corpus que otros algoritmos, se le pasa como insumo la nueva base datos transformada desde la original con las coordenadas geográficas, fechas, contenidos de las cartas, nombres y demás datos para que tome como base de conocimiento el 80% de la muestra y realice la predicción sobre el 20% restante usando validación cruzada para evitar fuga de datos y, en consecuencia, sobreajuste del modelo. Por otra parte, se aplicó un modelo de aprendizaje no supervisado para identificar las agrupaciones geográficas presentes en el epistolario, de acuerdo con el valor de las distorsiones calculadas entre las distancias de los elementos de una agrupación a su centroide respectivo.

Los siguientes mapas muestran el resultado de aplicar aprendizaje no supervisado, junto a la gráfica de distorsiones para determinar el número óptimo de agrupaciones, y el resultado del aprendizaje supervisado.

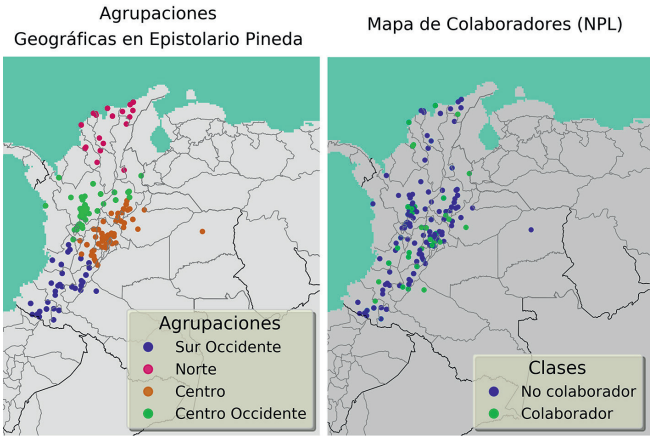


Figura 10. Mapas de agrupaciones geográficas y de clasificación automática.

Tras un examen más detallado del mapa de agrupaciones geográficas se puede determinar que estas coinciden, a grandes rangos, con cada ensanche o expansión de las redes sociales epistolares de Pineda al considerar el rango temporal de cada agrupación. Dicho esto, se calcula que la proporción de cartas en cada agrupación espacial es la siguiente: región sur occidente contiene el 19% de corresponsales en un rango temporal entre 1836-1870, coincidente con su primer matrimonio; la región norte contiene el 19% de corresponsales en un rango temporal entre 1852-1871, aproximadamente coincide con su segundo matrimonio; la región centro contiene el 39.7% comprendido en un rango temporal entre 1830-1876, que responde a las redes de parentesco y patronazgo tras la reconfiguración de las redes de poder; la región centro occidente conserva el 22.3% entre 1816-1877, coincidente con la mudanza de Pineda a la capital. Lo anterior solo corrobora la hipótesis de que a medida que Pineda expande sus horizontes relacionales, a través del matrimonio y las relaciones políticas, también expande su influencia en el territorio.

En cuanto a la evaluación del modelo predictivo, resulta muy útil el computo de una matriz de confusión para determinar que tantos aciertos o desaciertos tuvo el algoritmo. Este paso, a juicio del investigador, sirve más para probar qué tan útiles son los datos utilizados para la predicción, que para probar la utilidad del algoritmo. Los resultados son los siguientes:

Tabla 1. Matriz de confusión.

		Valores reales	
		Negativo	**Positivo**
Predicción	**Negativo**	Verdadero Negativo (608)	Falsos negativos (0)
	Positivo	Falsos positivos (36)	Verdadero Positivo (95)

De la muestra destinada a la predicción (20%), el algoritmo alcanzó una precisión de 0.95, una sensibilidad de 0.725 y una exactitud de 0.95. Esto quiere decir que el modelo tiene una excelente capacidad de predicción de positivos (precisión), así mismo una alta tendencia a producir falsos positivos (sensibilidad) y, finalmente, una buena capacidad de producir predicciones correctas (exactitud). Estos valores, al lado de la matriz de confusión permiten evaluar el comportamiento del modelo que, para este caso, se consideró menos riesgoso un falso positivo a un falso negativo, dados los costos temporales de verificación para los falsos negativos. En consecuencia, podríamos concluir que el modelo es aceptable al ponderar falsos positivos, falsos negativos y total de aciertos.

6. Conclusión

Como se evidenció en este artículo, las diversas y potentes metodologías de las humanidades digitales tienen la capacidad de colocar al investigador en una posición privilegiada al momento de enfrentarse a un complejo y voluminoso corpus documental que, en este caso, permanecía inexplorado, tanto como la figura histórica a la que perteneció y quien sin duda se descubre como un personaje clave para el estudio de la vida política e intelectual del siglo XIX. El archivo epistolar de Anselmo Pineda es el laboratorio perfecto para aplicar metodologías experimentales que sean capaces de asumir la retadora tarea de hacer historia, a la vez que un aporte metodológico poco convencional en el campo de las HD aplicado a la investigación social del siglo XXI en Colombia y a la historia digital. Aún son muchas las posibilidades abiertas para el estudio del epistolario con metodologías distintas a las presentadas o con metodologías semejantes, pero aplicadas a otros epistolarios del siglo XIX, en un esfuerzo por comprender las dinámicas sociales de uno de los periodos más interesantes en la historia americana.

Si bien la combinación de las diversas técnicas de análisis de datos expuestas resulta muy potente en el caso estudiado, cada una de ellas constituye un campo especializado que valdría la pena explorar y poner a prueba con otros archivos documentales semejantes y epistolarios del mismo periodo. En este sentido, una de las técnicas con mayor alcance es el análisis de redes, pues al incorporar no solo un epistolario, sino varios epistolarios de los

ilustrados de mediados de siglo, sería factible producir un mapa de topología relacional para la élite intelectual y política del momento y abrir la puerta a un estudio sin precedentes que en diferentes escalas pueda incorporar las demás técnicas de análisis digital y derivar en enfoques que podrían enmarcarse, bien sea, en la historia de la ciencia para el estudio del tráfico y difusión de saberes, textos y artículos científicos; en la historia cultural y política con el análisis de tendencias de agrupación y comportamientos sociales acorde al partido político, la parentela o lugar de nacimiento; o en la geografía histórica con la comprensión y visualización de la estructura social de este grupo ilustrado con un énfasis en su distribución espacial, entre otros posibles ángulos e intereses de estudio de la historia de Colombia.

Referencias

Benjamin, W. (2005). El coleccionista. *Libro de los Pasajes*. Akal.

Bourdieu, P. (2000). *Poder, derecho y clases sociales*. Desclée.

Brown, M. (2012). *The Struggle for Power in Post- independence Colombia and Venezuela*. Macmillan.

Castillo Gómez, A. (2002). Del tratado a la práctica. La escritura epistolar en los siglos XVI y XVII. En C. Sánchez, y C. Castillo (Coords.), *Actas del VI Congreso Internacional de Historia de la Cultura Escrita, Vol. 1, La correspondencia en la historia. Modelos y prácticas de escritura epístola* (pp. 79-108). Calambur.

Cerarols, R. y García, A. L. (2017). Geohumanidades. El papel de la cultura creativa en la intersección entre la geografía y las humanidades. *Treballs de la Societat Catalana de Geografia, 84,* 19-34.

Derrida, J. (1996) *Mal de Archivo. Una impresión freudiana*. Trotta.

González Stephan, B. (2000). Coleccionar y exhibir: la construcción de patrimonios culturales. *Revista de Literatura, 29*(86), 3-18.

Gutiérrez Lorenzo, M.P. (2002). Prácticas y modelos epistolares de un archivo decimonónico: la correspondencia del Hospicio Cabañas. En C. Sánchez. y C. Castillo (Coords.), *Actas del VI Congreso Internacional de Historia de la Cultura Escrita, Vol. 1, La correspondencia en la historia. Modelos y prácticas de escritura epístola* (pp. 305-328). Calambur.

Hernández de Alba, G. y Carrasquilla Botero, J. (1997). *Historia de la Biblioteca Nacional.* Instituto Caro y Cuervo.

Imízcoz, J. M. y Arroyo, L. (2011). Redes Sociales y Correspondencia Epistolar. Del Análisis Cualitativo de las Relaciones Personales a la Reconstrucción de Redes Egocentradas. *Redes. Revista Hispana para el Análisis de Redes Sociales*, *21*(4), 98-138.

Jänicke, S., Franzini, G., Cheema, M. F. & Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. *Procedimientos de EuroVis*. (pp. 1-21). STAR – State of The Art Report. http://dx.doi.org/10.2312/eurovisstar.20151113

König, H-J. (1994). *El Camino Hacia la Nación: nacionalismo en el proceso de formación del Estado y de la Nación de la Nueva Granada, 1750 a 1856*. Editorial Banco de la República.

Moreno de Ángel, P. (1981). *Anselmo Pineda. Colección Academia Antioqueña de Historia.* Editorial Vieco.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history.* Verso.

Ortiz, S. A. (2015). *Vida y Obra del Coronel Anselmo Pineda. Un Estudio del Coleccionismo y de la Redes Sociales en Nueva Granada Durante el Siglo XIX*. [Tesis de grado]. Pontifica Universidad Javeriana.

Pineda, A. (1844). Prospecto. En J. M. Bermúdez (Ed.), *La Cartilla Popular: periódico moral, industrial y noticioso.1843-1844*. Panamá

Rodríguez, S. (1990). Extracto sucinto de mi obra sobre la educación republicana. *Sociedades americanas. Biblioteca de Ayacucho, 2,* 278-306.

Silva, R. (2002). *Los Ilustrados de Nueva Granada, 1760-1808. Genealogía de una comunidad de interpretación*. Fondo Editorial Universidad EAFIT.

Wolf, E. (1980). Relaciones de Parentesco, de Amistad y de Patronazgo en las Sociedades Complejas. *Clásicos y Contemporáneos en Antropología*. Alianza.

Fuentes primarias

Correspondencia de Anselmo Pineda. Fondo Tomas Cipriano de Mosquera del Archivo Central del Cauca.

Correspondencia de Anselmo Pineda. Fondo Mariano Ospina Rodríguez del Archivo histórico de la Universidad EAFIT.

Correspondencia de Anselmo Pineda. Archivo Histórico Cipriano Rodríguez Santamaría de la Universidad de la Sabana.

Correspondencia de Anselmo Pineda. Fondo Manuel Ancizar Basterra en el Archivo Histórico Universidad Nacional de Colombia.

Correspondencia de Anselmo Pineda. Archivo Julio Arboleda de la sección de Libros Raros y Manuscritos en la Biblioteca Luis Ángel Arango.

Pineda, A. *Manifestación comprobada que José Anselmo Pineda oficial primero interventor de la tesorería departamental de Antioquia hace al público, de la injusta persecusión suscitada contra él en los días de la tiranía por el tesorero José Prieto.* Medellín: Impreso por Manuel Antonio Balcázar, 1831.

Pineda, A. Disposiciones del prefecto Anselmo Pineda para el aprovechamiento de Caquetá. Fondo Pineda Pieza 469, Folios 103-104

Tomo 435 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 437 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 439 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 438 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 441 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 444 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 445 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 447 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 622 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 640 en *Raros y Manuscritos de la Biblioteca Nacional de Colombia.*
Tomo 630 *en Raros y Manuscritos de la Biblioteca Nacional de Colombia.*

Part II
Corpus construction

Chapter V

Desarrollo de un corpus de atlas lingüísticos[1]

Development of a corpus of linguistic atlases

Carolina Julià Luna
Universidad Nacional de Educación a Distancia (UNED) – España

Resumen: El objetivo del presente capítulo es la presentación de algunas características y funcionalidades del *Corpus de los atlas lingüísticos* (*CORPAT*), una herramienta informática en la que se almacenan datos procedentes de los atlas lingüísticos regionales del español europeo con el fin de conservar el patrimonio lingüístico que contienen; de servir como fuente de divulgación de la variación y la riqueza lingüística; y de complementar los datos procedentes de corpus textuales y obras lexicográficas que permitan ampliar las investigaciones sobre el cambio lingüístico y la historia de la lengua española.

Abstract: The aim of this chapter is to present some characteristics and functionalities of the *Corpus of Linguistic Atlases* (*CORPAT*). This computer tool collects data from the different regional linguistic atlases of European Spanish to preserve the linguistic heritage; to serve as a linguistic resource to disseminate knowledge about variation; and to complement the data from textual corpora and dictionaries that allow further research on linguistic change and the Spanish language history.

1 El presente texto fue escrito a mediados de 2021, por lo que los datos que constan él (referencias al corpus y número de mapas y de registros que contiene) pertenecen a ese año. A lo largo de 2022 el corpus ha aumentado el número de registros y desde el 1 de diciembre de 2022, CORPAT se desarrolla en el marco del proyecto "CORPAT-PEPLEs: corpus digital para la preservación y el estudio del patrimonio lingüístico del español" (TED2021-130752A-I00), financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea "NextGenerationEU"/PRTR.

1. Introducción

Desde hace más de una década, en España se está trabajando en la digitalización del atlas lingüístico nacional (el *Atlas Lingüístico de la Península Ibérica* o *ALPI*, García Mouton 2010, 2017; Sousa, 2020); sin embargo, buena parte de los materiales de la geolingüística regional no pueden consultarse todavía hoy en formato electrónico. El volumen de información recogido en estos atlas regionales[2] impresos entre 1961 y 1999, y de acceso muy limitado (no se encuentran en cualquier biblioteca), es excepcional: casi 7000 mapas que se corresponden con más de 700 puntos de encuesta, lo que supone miles de formas lingüísticas que aportan datos fonético-fonológicos, léxico-semánticos, morfosintácticos y etnolingüísticos de las variedades dialectales de España.

Actualmente, el hecho de que buena parte de los atlas regionales del español dirigidos por Manuel Alvar solo puedan consultarse físicamente en algunas bibliotecas universitarias, y a veces de forma incompleta (pues no siempre disponen de todos los volúmenes), dificulta sus posibilidades de estudio y explotación tanto a investigadores como a cualquiera que esté interesado en conocer, por ejemplo, la historia lingüística de su pueblo natal o de otros lugares de España. Además, la forma en la que presentan los datos supone otra barrera para los interesados no expertos, ya que la información se recoge, en muchos de los mapas, en alfabeto fonético. Y, para mayor complejidad, en el alfabeto de la *Revista de Filología Española* (*ARFE*), un sistema de transcripción empleado en el ámbito hispánico que creó Tomás Navarro Tomás cuando se fundó la revista con el objetivo de servir para "los artículos que hubieran de requerirlo y para los estudios de dialectología, fonética y geografía lingüística que el Centro había emprendido" (Navarro Tomás, 1975, p.19).

Con el fin de cubrir esta parte de la geolingüística europea, se ha ideado y diseñado *CORPAT* (*Corpus de los atlas lingüísticos*). Se trata de un corpus en el que se pretende organizar y categorizar conjuntamente parte de la información de los atlas lingüísticos regionales (*ALEA, ALEANR, ALEICan, ALECant, ALCyL, ALeCMan, ADiM, CaLiEx*)[3] en una

2 *ALCyL* = Manuel Alvar (1999): *Atlas Lingüístico de Castilla y León*. Valladolid: Junta de Castilla y León/Consejería de Educación, 3 vols.; *ALEA* = Manuel Alvar con la colaboración de Antonio Llorente y Gregorio Salvador (1961-1973): *Atlas lingüístico y etnográfico de Andalucía.* Granada: Universidad de Granada/CSIC, 6 vols.; *ALEANR* = Manuel Alvar con la colaboración de Antonio Llorente, Tomás Buesa y Elena Alvar (1979-1983): *Atlas lingüístico y etnográfico de Aragón, Navarra y Rioja.* Madrid /Zaragoza: La Muralla / Institución Fernando el Católico de la Excma. Diputación provincial de Zaragoza / CSIC, 12 vols.; *ALECant* = Manuel Alvar con la colaboración de Carlos Alvar, José A. Mayoral, M.ª Pilar Nuño, M.ª del Carmen Caballero y Julia B. Corral (1995): *Atlas lingüístico y etnográfico de Cantabria*. Madrid: Arco/Libros, 2 vols. [Etnografía y láminas de Elena Alvar]; *ALEICan* = Manuel Alvar (1975-1978): *Atlas lingüístico y etnográfico de las Islas Canarias.* Las Palmas de Gran Canaria: Publicaciones del Excmo. Cabildo Insular, 3 vols.

3 Quiero hacer constar mi agradecimiento a los autores del *ALeCMan*, la Dr.ª Pilar García Mouton y el Dr. Francisco Moreno Fernández; del *ADiM*, la Dr.ª Pilar García Mouton y la Dr.ª Isabel Molina Martos; y de *CaLiEx*, el Dr. José González Salgado, por su apoyo en el inicio del desarrollo esta herramienta.

base de datos espacial consultable en línea. Antes de presentar la herramienta (epígrafes 3-4), se describe brevemente la historia de la relación que la geolingüística ha mantenido con la tecnología (epígrafe 2) y algunos de los resultados obtenidos de su aplicación.

2. La tecnología en la geografía lingüística

La geografía lingüística nace en Europa a finales del siglo XIX con el objetivo de representar la variación lingüística en mapas y dejar de lado la descripción intuitiva y fortuita de las áreas dialectales que se había realizado hasta la segunda mitad de esa centuria (Chambers y Trudgill, 1994, p.37). A principios del siglo XX, con la publicación del *Altas Linguistique de la France* (*ALF*), se consolida como método de investigación dialectal basado en la compilación de datos procedentes de testimonios orales. Desde entonces, el atlas lingüístico se convierte en una obra fundamental en el ámbito de los estudios variacionistas que irá perfeccionándose y modificándose con el paso del tiempo.

La historia de la renovación del método y de su producto principal, el *atlas lingüístico*[4], refleja cambios de diverso tipo; desde las innovaciones vinculadas con la organización de los datos (p. ej. el paso de la organización alfabética del *ALF* a la onomasiológica del *AIS*) hasta modificaciones relacionadas con el foco de interés lingüístico (p. ej. el surgimiento de atlas sintácticos como el SCOSYA o el *DynaSAND*, que atienden una parte de la gramática poco representada en los primeros atlas) y con el tipo de informante (p. ej. la ampliación de las encuestas a hablantes urbanos, más jóvenes y que incluyan tanto a hombres como a mujeres)[5], entre otras (Julià, 2020). Además de estas variaciones, asociadas a la evolución de la propia metodología y de las teorías lingüísticas, uno de los aspectos que ha supuesto un cambio mayor es la aplicación de los ordenadores a su creación, diseño y explotación. La aplicación de la tecnología a la geografía lingüística es sumamente importan-

4 La 23.ª ed. del *DLE* (2014) incorpora por primera vez en la historia del diccionario académico la definición de *atlas lingüístico* "Conjunto de mapas en que se presentan datos lingüísticos procedentes de encuestas" (*s. v. atlas*). Para más información sobre el concepto 'atlas lingüístico', véase Coseriu (1977).

5 Tradicionalmente, los cuestionarios tenían como objetivo recoger información procedente de hablantes (generalmente hombres) que conocieran el medio rural, sus tradiciones y sus costumbres; a este informante tipo, según Chambers y Trudgill (1994, p.57), se le denomina mediante el acrónimo *NORM* (*nonmobile, older, rural, males*). Sin embargo, "en las últimas décadas, las antiguas formas de vida y las tareas asociadas con ellas se han transformado hasta casi desaparecer" (García Mouton y Molina 2009, p.180) y también lo han hecho los informantes que son objeto de interés en la geografía lingüística. El *ALeCMan*, por ejemplo, incluye novedades respecto a sus antecesores (el *ALEA,* el *ALEANR* o el *ALEICan*) como, por ejemplo, la incorporación de dos informantes sistemáticamente por localidad, "un hombre y una mujer entre los que se reparten el contenido de un cuestionario muy extenso. Se hicieron sistemáticamente dos entrevistas por punto: una con un hombre y otra con una mujer" (Molina, 2018, p.4). Para una visión global de la representación del papel de la mujer como informante en la geografía lingüística de la península ibérica, véase García Mouton (1999a).

te para los estudios dialectales, para la investigación lingüística (Nerbonne *et al.*, 2021) e incluso para la historia de las humanidades digitales (Sousa, 2017).

Los primeros testimonios de la aplicación de los ordenadores al estudio geolingüístico se pueden fechar en la década de los sesenta (Ziamandanis, 1996, p.56). En 1966, Roger Shuy, en el capítulo titulado "An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada", explica cómo ideó un programa informático de tarjetas perforadas[6] —como hizo Busa en el proceso de lematización de la obra de Santo Tomás de Aquino en el *Index Thomisticus*— para trabajar en la automatización de los datos con el objetivo de que fueran más accesibles. Para demostrar la viabilidad de su idea, se centró en 78 informaciones gramaticales de una región de Estados Unidos (*The Linguistic Atlas of New England*) con el fin de trazar relaciones sociolingüísticas significativas. Los resultados de este primer acercamiento a la automatización de los atlas lingüísticos son, en opinión del autor, una demostración del potencial de los datos después de haber sido procesados electrónicamente:

> *This program, of course, is only suggestive of what can be done with the Atlas materials once the data are submitted to automation. These materials will be more accessible and reproducible than ever before. More significant, the dialectologist will be able to broaden his investigation of the sociological implications of American speech through improved handling of data. As indicated previously, one of the benefits of our program is in the area of distributions by occupation, sex, age, and type. (Shuy, 1966)*[7].

A la propuesta de Shuy (1966) empiezan a suceder otros estudios. Por un lado, investigaciones en las que con la incorporación de los ordenadores al análisis de los datos se pretendía extraer el máximo rendimiento a la información lingüística desde el punto de vista de la variación y de la delimitación de las áreas dialectales (Gordon, 1969, p.1). Entre ellas se sitúan, por ejemplo, los estudios en dialectometría[8]. Por otro lado, los primeros trabajos que emplean la informática para crear atlas se sitúan en la década de los setenta. Así, en la

6 Las tarjetas perforadas constituyen el primer medio de almacenamiento digital de información empleado para introducir y guardar datos en ordenadores. Este método fue muy empleado en la década de los setenta del siglo XX.

7 No se cita la página porque se ha consultado la edición electrónica del trabajo y en ella no constan las páginas.

8 García Mouton (1999b, p.335) define la *dialectometría* como "una disciplina clasificatoria, de carácter instrumental, que se apoya en la geografía lingüística y recurre a procedimientos objetivos —estadísticos y taxométricos—, para establecer relaciones de semejanza o diferenciación dialectales, en un intento de sintetizar los contenidos de un atlas lingüístico". Aunque en los primeros trabajos dialectométricos de Jean Séguy (1973) se prescindiera de la automatización, y los cálculos se hicieran manualmente (García Mouton, 1999, p.336; Aurrekoetxea 2019, p.23-24), el uso de la cuantificación informática es un pilar esencial de esta disciplina.

geolingüística estadounidense, Wood (1971 *apud* Ziamandanis, 1996, p.56) propone, en la línea de Shuy (1966), el uso de computadoras y tarjetas perforadas para editar atlas lingüísticos. Y es en la década de los setenta cuando se sitúa el inicio de la informatización de los atlas (Hoch y Hayes, 2010, p.25) que ofrecerá los primeros resultados en los años ochenta y noventa para la geografía lingüística europea: "The three projects which stand out as pioneers are *Computer Developed Linguistic Atlas of England* (Viereck y Ramisch, 1991-1997), *Atlas Linguarum Europae*[9] (Alinei *et al.*, 1983) and *Kleiner Deutscher Sprachatl*as (Veith *et al.*, 1984-1999)" (Sousa, 2017, p.22).

En las siguientes décadas, el acelerado progreso en el ámbito de la comunicación y la expansión del uso de la tecnología generó cambios en los estudios geolingüísticos y dialectales. Entre esos cambios, destacan las mejoras en los escáneres de imágenes, la proliferación de programas de bases de datos espaciales (BDE) y el surgimiento de numerosas aplicaciones y programas para crear mapas (Google My Maps, Gabmap, Diatech, OpenStreetMap, ArcMap, Carto, Mapbox o QGIS). El empleo de estas herramientas ha permitido, por ejemplo, digitalizar los primeros atlas lingüísticos y recogerlos en la web (a modo de facsímil) con el fin de preservarlos y ponerlos a disposición de cualquiera que quiera consultarlos. Entre otros, pueden mencionarse los proyectos de digitalización del *Sprachatlas des Deutschen Reichs* (*DSA*) de Georg Wenker, que actualmente se puede consultar en RegionalSprache.de (Herrgen 2010 y Limper, Pheiff y Williams 2020: 3744); el *Atlas Linguistique de la France* (*ALF*), disponible en CartoDialect (Davoine *et al.*, 2015); y el *Sprach und Sachatlas Italiens und der Südschweiz* (*AIS*) de Karl Jaberg y Jakob Jud, accesible en NavigAIS (Tisato, 2019). Algunos de ellos, además, incluyen la posibilidad de consultar bases de datos en las que la información está organizada y clasificada por categorías (formas y campos semánticos, por ejemplo).

En España es también en la década de los setenta cuando se empieza a pensar en la automatización de la geografía lingüística regional (Alvar, 1976[10]; Alvar y Verdejo, 1978 [1980]; Alvar y Nuño, 1981) y a partir de los ochenta se plantea el análisis automatizado de los datos (Enríquez, 1986). El proyecto del *ALES* (*Atlas Lingüístico de Santander*) —al que hoy se conoce como *Atlas Lingüístico y Etnográfico de Cantabria* (*ALECant*)— es la prime-

9 Sobre el *Atlas Linguarum Europae* (*ALE*) y la implementación de un proceso de cartografiado automático pueden leerse algunos de los primeros planteamientos en Putschke (1969 y 1972) a los que se van sucediendo otros trabajos y propuestas.

10 Esta referencia aparece citada en Alvar y Nuño (1981, p.359, nota 1). En la primera nota al pie se explica que es una publicación que deriva de una comunicación que Manuel Alvar había presentado en febrero de 1974 en el Simposio Ordenadores y Lingüística que organizó la Universidad Complutense. Según se indica en Alvar y Nuño (1981, p.359), el contenido del texto presenta resultados de los primeros contactos que Manuel Alvar mantuvo con W. Putschke para el *Atlas Linguarum Europae* (*ALE*).

ra muestra de aplicación de la tecnología a los atlas españoles. En el artículo de Alvar y Verdejo (1978), titulado "Automatización de atlas lingüísticos", se presentan las bases de los primeros pasos de la geolingüística española en el proceso de creación de atlas automatizados. Los autores toman como modelo el atlas de Andalucía (*ALEA*) para explicar la complejidad que supone el proceso manual de elaboración de cada uno de los mapas:

> *Cada cuaderno de formas es la base para que un cartógrafo dibuje un mapa por cada binomio —concepto, región— representando en él los testimonios —provincia, localidad, respuesta— pertenecientes al mencionado binomio. Más tarde se lleva a cabo la impresión. (Alvar y Verdejo, 1978, p.23)*

En palabras de los propios autores, se trata de un "complejo proceso manual" repleto de dificultades que "puede simplificarse en mucho con un proceso de automatización" (Alvar y Verdejo, 1978, p.26-27). Era evidente que la automatización del proceso se veía, principalmente, como una vía para reducir el tiempo dedicado a dibujar los mapas y para mitigar los errores que pudieran introducirse en el proceso de cartografiado manual de la información lingüística. Así, los autores describen con detalle en el artículo cuál tendría que ser el método de automatización que debería seguir un atlas; y, en la conclusión, explican que esta es la metodología que han empezado a aplicar para la publicación de los materiales del *ALECant*, cuya recopilación de datos terminó en julio de 1978. Sin embargo, el proceso de automatización descrito por Alvar y Verdejo (1978), que luego se complementa con el artículo de Alvar y Nuño (1981), fue más costoso de lo que parecía inicialmente. Tales fueron las dificultades del proyecto —asociadas a su proceso de informatización (como puede leerse en el epígrafe titulado "Lamento inicial" que precede a la nota preliminar del *ALECant*, 1995, p.7)— que el atlas no se publicó hasta casi veinte años más tarde. Después del atlas de Cantabria, se publican otros atlas de forma automatizada como el *ALCyL* y el ALeCMan. El primero, según Alvar, sigue los criterios del *ALECant* (*ALCyL*, Prólogo: 11); el segundo, en cuya informatización empezó a trabajarse desde 1996 en la Universidad de Alcalá de Henares, sigue un camino distinto: para su elaboración se creó un programa informático específico denominado *Atlante* que tenía por objetivo la automatización de las "labores que conducen a la confección de un atlas lingüístico, así como el aprovechamiento de toda la información lingüística que contiene una obra de estas características" (Moreno *et al*., 1997, p.202). Este atlas, que puede consultarse en internet actualmente, seguía la línea de trabajo iniciada en otros proyectos europeos y americanos en los que la informática permitía automatizar el proceso de cartografiado y gestionar las bases de datos espaciales.

Posteriormente, en la segunda década el siglo XXI, se inician los trabajos de edición digital del *Atlas Lingüístico de la Península Ibérica* (ALPI) parcialmente consultable en la red en la actualidad (García Mouton, 2017)[11]. Paralelamente a estos trabajos de digitalización de atlas tradicionales se ha consolidado el diseño y la producción digital de atlas, lo que ha generado que nos encontremos ante una nueva generación de contenidos geolingüísticos más sostenibles y accesibles que ya no se publican en papel; es el caso, por ejemplo, del *Atlas Dialectal de Madrid* (*ADiM*), que sigue la línea iniciada por el *ALeCMan*. Por otra parte, además de estos proyectos, cabe señalar que el empleo del mapa como medio de representación de datos lingüísticos se ha expandido más allá de la publicación de los atlas. Son diversos los trabajos en los que se (geo)localizan valiosas informaciones lingüísticas en mapas y que permiten realizar interesantes estudios de variación desde el eje diatópico (COSER).

En este proceso de digitalización e informatización de los atlas, la geografía regional del español (nos referimos a los atlas que dirigió Manuel Alvar desde la segunda mitad del siglo XX) cuenta con pocas iniciativas y, por el momento, son pocos los proyectos que trabajan en esta línea. Uno de ellos es el *Atlas Lingüístico y Etnográfico de la provincia de Zaragoza* (ALPEZ) cuyos datos proceden del *ALEANR* (*Atlas Lingüístico y Etnográfico de Aragón, Navarra y La Rioja*). Se trata de un atlas digital que recoge los materiales del cuarto volumen de este atlas. Se puede consultar en línea y ofrece los datos organizados e interpretados desde diferentes perspectivas, lo que permite realizar consultas de distinto tipo:

> *Este Atlas digital ofrece nuevas posibilidades de búsqueda (visual e interactiva), estudios con gráficos-estadísticos, multi-task, un mapa interactivo (actualizable), respuestas en transcripción ortográfica, un mapa-leyenda en colores que remite al del ALEANR y una base de datos informatizada. (Tranquilli, 2019, p.1)*

El acercamiento a los datos que ofrece este reciente recurso constituye una muestra de las posibilidades que brinda la aplicación de la tecnología a los datos de los atlas regionales. Asimismo, son interesantes otras investigaciones también recientes en las que se explotan los datos de los atlas regionales mediante la tecnología. En el proyecto VitaLex (desarrollado en la Universidad de Granada), que se centra en el análisis de la zona de la Alpujarra (Andalucía), el objetivo principal es analizar las respuestas léxicas de 10 puntos de encuesta del *ALEA* y contrastarlas con datos actuales obtenidos de nuevas entrevistas. Los resultados de este estudio permitirán ver los cambios que se han producido en cincuenta años

11 Sobre el español de América se inician también múltiples e interesantes proyectos de digitalización e informatización de atlas lingüísticos en la misma época; por ejemplo, sobre el *Atlas Lingüístico de Puerto Rico* - ALPR o el *Atlas lingüístico y etnográfico de Colombia* - ALEC Digital, entre otros.

en esta zona (Fernández Morell en prensa). En los capítulos 14 y 15 de Fradejas (2020), titulados "Mapas con R. Un poco de geografía lingüística", se muestra también algunos de los resultados de la aplicación de la tecnología a los datos que atesoran los mapas de la geografía lingüística regional. Es en este marco, en el de aprovechar las posibilidades que ofrecen las bases de datos espaciales y los sistemas de información geográfica (SIG), entre otros, en el que nace la idea de crear *CORPAT* (*Corpus de los atlas lingüísticos*), una base de datos cuyos objetivos, contribuciones y características se describen a continuación.

3. Objetivos y contribución

CORPAT se concibe como una herramienta digital que pretende, por un lado, preservar el patrimonio histórico-lingüístico y cultural de la lengua española y, por otro lado, aproximar la investigación de la variación lingüística a la sociedad. Para la consecución de estos objetivos, se parte de las posibilidades que ofrecen las nuevas tecnologías para la divulgación y la gestión de datos geolocalizados en el marco de las humanidades digitales. Mediante el traspaso de las formas de las cartas lingüísticas a bases de datos espaciales en transcripción ortográfica se favorece su difusión, además de permitir que los materiales permanezcan almacenados con el fin último de contribuir a su preservación y divulgación.

La creación y el diseño del corpus se justifica tanto desde la perspectiva histórica como actual para la geografía lingüística española y europea. El lento y desafortunado desarrollo de la geografía lingüística en España (Heap, 2002; García Mouton, 2009) impidió la publicación completa del Atlas Lingüístico de la Península Ibérica (ALPI). Para suplir este vacío, fueron publicándose sucesivamente, desde la década de los sesenta del siglo XX, un conjunto de atlas regionales que abarcan diferentes zonas: Andalucía (*ALEA*), Aragón, Navarra y La Rioja (*ALEANR*), las Islas Canarias (*ALEICan*), Cantabria (*ALECant*) y Castilla y León (*ALCyL*). Posteriormente, esta saga de atlas lingüísticos se ha completado con otros como el de Castilla-La Mancha (ALeCMan) y el de Madrid (ADiM) en formato digital y consultables en línea. A estos hay que añadir los atlas de las zonas bilingües que han ido publicándose de forma paralela a los del español, pero que abarcan solo el estudio de la lengua cooficial y que, en algunos casos, se encuentran en Internet (Galicia: ALGa, País Vasco: EEHHA y Cataluña, Valencia y Baleares: ALDC), y también los trabajos de González Salgado sobre el extremeño (Cartografía lingüística de Extremadura) que completan la cartografía por regiones. Así pues, a pesar de contar con datos geolingüísticos sobre el español europeo de una gran parte del territorio, lo cierto es que estos materiales no se han explotado ni estudiado de forma exhaustiva y contrastada y las comunidades lingüísticas de las que proceden frecuentemente ignoran su existencia. La cuantía de datos que incluyen y la gran cantidad de tiempo invertido en su

elaboración es uno de los principales motivos que ha generado que la última fase del método de la geografía lingüística —en la que se procede a su estudio— se haya desarrollado parcialmente (Del Barrio, 2018; Fernández Morell, en prensa).

Así pues, la contribución principal de *CORPAT* es la preservación del patrimonio lingüístico español; esto es, el almacenamiento y la gestión de los datos que actualmente se hallan distribuidos en bibliotecas y centros de investigación y que corren el riesgo de desaparecer por el formato en el que se conservan. Los mapas de los atlas son multidimensionales y permiten estudiar aspectos diversos desde perspectivas distintas, como la variación fonético-fonológica (Llorente, 1962), la caracterización y la delimitación de los campos semánticos (Salvador, 1965), los procedimientos de formación de palabras (Uritani y Berrueta, 1985), los procesos de creación léxica (Fuster, 1996), la historia de la lengua y la etimología (Prat, 2006; García Mouton, 2010, 2016; Fernández-Ordóñez, 2011); el cambio lingüístico (Molina 2006; Del Barrio 2018), etc.. Así, contar con un recurso informático que permita consultarlos de forma rápida y sistematizada aportará información muy valiosa para la investigación en lengua española desde múltiples perspectivas. Por ejemplo, se podrán estudiar los procesos de creación léxica más frecuentes en la lengua popular o contrastar la extensión y la vitalidad de los fenómenos fonético-fonológicos en la época en la que se recogieron los datos. Además, esto podría tomarse como punto de partida para entrevistar de nuevo los territorios y estudiar el cambio lingüístico en los últimos setenta años, de forma similar a lo que se está haciendo, por ejemplo, para otras lenguas como el inglés (http://tweetolectology.com/) o, a pequeña escala, con una parte del territorio andaluz (proyecto Vitalex). Los datos no serán solo útiles individualmente, también servirán como complemento a otros grandes bancos de datos digitales como son los corpus textuales, los diccionarios electrónicos y otros atlas lingüísticos digitales (en especial, el *ALPI*).

Asimismo, la divulgación digital de la información contenida en las cartas lingüísticas también contribuirá, por un lado, a educar en empatía lingüística (Ibarretxe-Antuñano, 2021), un aspecto con escasa presencia en el proceso de enseñanza-aprendizaje de lenguas; y, por otro, a conservar y a dar a conocer la memoria histórica de las comunidades lingüísticas de España. Por ejemplo, las localidades que fueron encuestadas a mediados del siglo XX podrán tener acceso a los datos sobre las herramientas y las técnicas de cultivo empleadas por sus antepasados, las creencias o las costumbres sobre juegos, tradiciones y fiestas populares, entre otros aspectos de carácter etnolingüístico. Conocer su pasado a través de los atlas lingüísticos, les permitirá entender su presente. A continuación, se describe brevemente la estructura y el contenido del corpus en el inicio de su configuración.

4. Estructura y contenido del corpus en la fase preliminar

El corpus, que se halla en una etapa preliminar (desarrollo en fase de pruebas en la que se han incorporado los datos de 50 mapas relativos a 15 conceptos, lo que supone, por el momento, más de 5500 registros), se recopila en una base de datos MySQL 5.6. Se trata de una base de datos relacional en código abierto compuesta por tablas (algunas formadas con catálogos y otras abiertas) en las que se relaciona la información lingüística con la geográfica. En la interfaz de introducción de datos, en la que se trabaja en línea —lo que permite que diferentes personas introduzcan datos a la vez— se pueden modificar, eliminar y crear registros de cada una de las tablas. Para cada una de las respuestas recogidas (formas) en un mapa se crea un registro en la base de datos que se categoriza y completa según los siguientes parámetros (que constituyen campos en la base de datos): concepto, punto de encuesta, lengua, información morfológica, información semántica, información sintáctica, información fonética, información etnolingüística, tipo de respuesta, otras informaciones. A continuación, se describen algunas de las funcionalidades básicas del corpus que atañen a una parte de la información que se incorpora en la base de datos para cada uno de los registros. Se trata de la parte que más se ha desarrollado hasta la actualidad (mayo de 2021) y que se refiere principalmente a las búsquedas de información léxico-semántica y geográfica.

El corpus se ha diseñado, igual que otras herramientas lingüísticas creadas mediante tablas relacionales (cfr. por ejemplo, la versión electrónica del *Diccionario Crítico Etimológico Castellano e Hispánico - DECH*, versión en CD ROM 2012), para que puedan realizarse búsquedas simples (por un solo criterio) o búsquedas múltiples (que combinan distintas opciones y permiten filtrar la información para obtener resultados más concretos). Al acceder a la interfaz, se llega a la consulta principal, que se divide en tres campos (figura 1):

Figura 1. Interfaz de consulta principal de *CORPAT*.

En la búsqueda por FORMA se recoge en transcripción ortográfica la palabra o secuencia de palabras que se corresponde con la respuesta de un punto de encuesta del atlas. Por ejemplo, si se introduce la palabra *jamón* en la caja de consulta, el desplegable ofrece la lista ordenada alfabéticamente de los diez registros que contienen esta cadena de caracteres,

bien sean palabras simples, derivadas o sintagmas que la contengan (*el jamón, hueso del jamón, jamón, jamoncete, jamoncillo*). El usuario puede elegir la forma que le interese del desplegable o verlas todas. Si se eligen todos los registros, se obtiene información sobre los conceptos, los atlas, los mapas y los puntos de encuesta en los que aparecen estas formas. La búsqueda arroja 10 registros relativos a los conceptos 'hueso de la cadera' y 'pulpejo' (figura 2):

FORMA	CONCEPTO	CAMPO SEMÁNTICO	ATLAS	MAPA	PUNTO DE ENCUESTA	PROVINCIA	LOCALIDAD	NÚMERO DE RESPUESTA
el jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Ma 101	Málaga	Teba	1.ª
hueso del jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Co 605	Córdoba	Castil de Campos	1.ª
hueso del jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Gr 301	Granada	Colomera	1.ª
hueso del jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Gr 302	Granada	Iznalloz	1.ª
hueso del jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Gr 500	Granada	Salar de Loja	1.ª
hueso del jamón	hueso de la cadera	El cuerpo humano	ALEA	1260	Ma 201	Málaga	Villanueva de Algaidas	1.ª
jamoncete	pulpejo	El cuerpo humano	ALECant	848	S 401	Cantabria	Villaverde de Trucios	1.ª
jamoncillo	pulpejo	El cuerpo humano	ALEANR	992	Sor 400	Soria	Ólvega	1.ª
jamón	pulpejo	El cuerpo humano	ALEA	1275	J 306	Jaén	Porcuna	1.ª
jamón	pulpejo	El cuerpo humano	ALEANR	992	Lo 502	La Rioja	Lumbreras	2.ª

Mostrando registros del 1 al 10 de un total de 10 registros

Figura 2. Resultados de la búsqueda por *forma* en *CORPAT.*

Los resultados obtenidos en esta búsqueda constituyen el reflejo de la necesidad de poder ver los datos de los atlas organizados de este modo para examinar qué relaciones lingüístico-conceptuales se establecen entre los diferentes conceptos y ámbitos semánticos que forman parte de los atlas (como las partes del cuerpo y los alimentos).

En la búsqueda por CONCEPTO se incluye el nombre identificativo del mapa que constituye la realidad que es objeto de investigación. Es el que suele aparecer en los índices de los atlas y habitualmente se ubica en la parte superior izquierda de las cartas geolingüísticas (véase la figura 4). Esta posibilidad de búsqueda está vinculada al orden onomasiológico en el que los atlas se conciben. Se parte, por tanto, del concepto (realidad) para llegar al lexema. El nombre del concepto se ha vinculado previamente a un subcampo semántico que, a su vez,

se relaciona con un campo semántico. Esta clasificación conceptual deriva de la organización de los índices de los atlas lingüísticos. Así, por ejemplo, en el campo semántico Agricultura (que en el *ALEA* ocupa del mapa 7 al mapa 287), se incluyen quince subcampos semánticos (aparejo para las bestias de carga, arado, carbonero, carro, el campo y sus cultivos, el corcho y su elaboración, molinos de harina y panificación, olivo y oleicultura, vid y vinificación, etc.) en cada uno de los cuales se clasifican los conceptos cartografiados. La jerarquía puede esquematizarse del siguiente modo con los mapas del *ALEA* referidos al subcampo semántico de la vid y la vinificación que se recoge en la figura 3:

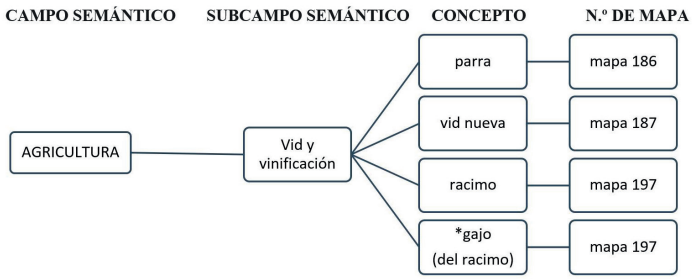


Figura 3. Ejemplo de jerarquía onomasiológica del corpus.

El corpus incorpora tanto los conceptos cartografiados como aquellos que no tienen mapa propio porque se consideró que presentaban poca variación para representarla en un mapa. En la mayor parte de los atlas, los conceptos no cartografiados suelen aparecer en otros mapas y señalados en el índice con un asterisco. El *ALCyL* es el único que incluye las respuestas a conceptos no cartografiados en una lista —titulada "Preguntas no cartografiadas" (pp. 921-937)— en lugar de incorporarlas en otros mapas. Véase, a modo de ejemplo, la información que sobre el concepto 'articulación' incluye el mapa 494 del *ALEICan* en el que las respuestas que aparecen cartografiadas son las del concepto 'hueso de la cadera' (figura 4):

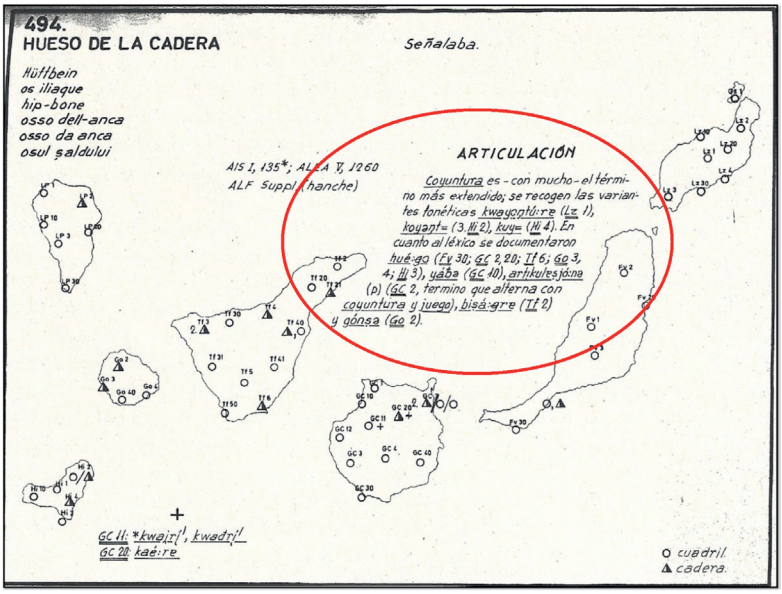


Figura 4. Ejemplo de mapa que incluye conceptos no cartografiados (*ALEICan*, mapa 494).

Esto es importante por cuanto amplía considerablemente el número de registros incorporados en el corpus. El *ALEA*, por ejemplo, en su primer volumen, incluye 86 conceptos no cartografiados en el interior de los mapas, lo que supone un incremento de un 30 % más de registros para este volumen.

Aunque los atlas lingüísticos regionales del español siguen una metodología homogénea —motivo por el cual sus datos pueden ser contrastados y analizados como una unidad—, existen pequeñas divergencias que han implicado un trabajo de unificación previo para sistematizar la búsqueda en este campo de la base de datos. Por ejemplo, algunos conceptos no se etiquetan con el mismo nombre, aunque se refieren a la misma realidad. Así sucede en el caso del concepto 'incisivo' que aparece identificado como 'incisivos superiores centrales' en el *ALEA* (mapa *1224), 'incisivos' en el *ALECant* (mapa 833), '(diente) incisivo' en el *ALEANR* (mapa 955) y 'dientes delanteros' en el *ALeCMan* (mapa 298).

En la búsqueda por CAMPO SEMÁNTICO el usuario puede seleccionar de una lista cerrada el ámbito de significación sobre el que desea realizar la consulta. Como se ha detallado anteriormente en la descripción de la búsqueda por CONCEPTO, los campos semánticos que aparecen en el corpus vienen determinados por las áreas de interés de los cuestionarios que aparecen organizados onomasiológicamente en los atlas: agricultura, animales domésticos, animales silvestres, apicultura, creencias populares y supersticiones, de la cuna a la sepultura, el cuerpo humano, el mar, el tiempo, etc. De igual modo que en el caso de la búsque-

da por concepto, la información de este apartado requiere de una unificación previa. Así, por ejemplo, mientras que el *ALEA*, el *ALEANR* y el *ALEICan* coinciden en dividir el campo semántico relativo a los vegetales en diversos apartados (plantas silvestres, flores, arbustos, hortalizas, árboles frutales, el bosque, etc.), el *ALECant* recoge la información bajo el epígrafe "Vegetales" sin establecer ninguna división. Por ello, los conceptos del *ALECant* que coinciden con los de los otros atlas, se han clasificado según estos. El concepto 'musgo' puede servir de ejemplo: aparece en el *ALEA*, el *ALEANR* y el *ALEICan* en el apartado "Plantas silvestres, flores y arbustos", por ello, en *CORPAT*, los registros del *ALECant* para este mapa se categorizan bajo este subcampo semántico que, por el momento, no se visualiza en la interfaz de consulta.

A las tres búsquedas principales que se han descrito (por FORMA, CONCEPTO y CAMPO SEMÁNTICO) se añaden otras opciones vinculadas a la fuente de obtención de datos. Se puede buscar por ATLAS, por NÚMERO DE MAPA, por PUNTO DE ENCUESTA, por NOMBRE DE LA LOCALIDAD y por PROVINCIA. Es posible, además, combinar estas búsquedas con las tres principales; así, el usuario puede obtener todas las formas que el corpus contiene, por ejemplo, para la provincia de Huelva en el campo semántico del cuerpo humano o todos los registros de una localidad (figura 5):

Figura 5. Interfaz de consulta secundaria de *CORPAT*.

En el campo PUNTO DE ENCUESTA se incluye el código que recibe el enclave geográfico en cada uno de los mapas según la metodología seguida por Alvar desde el *ALEA*:

> *cada lugar está representado por una sigla (que representa el nombre de la provincia, según la abreviatura oficial del Ministerio de Obras Públicas) y un número de tres cifras […] cada provincia está dividida idealmente en seis casillas de las cuales las que registras cen-*

> *tenas impares corresponden al oeste y las pares al este. Dentro de ellas, la localización (norte, centro, sur) se hace por orden creciente: 1 (noroeste), 3 (centro-oeste), 5 (sudoeste); 2 (nordeste), 4 (centro-este) y 6 (sudoeste)). (ALEA, Nota preliminar: 3)*

El corpus, por tanto, mantiene la codificación original de los atlas regionales. El punto de encuesta se recoge previamente en una tabla en la que se asocian con información sobre el atlas al que pertenece, el nombre de la localidad, la provincia y las coordenadas (la longitud y la latitud) que permiten la geolocalización. Cada registro se localiza en el mapa al pinchar en el nombre de la localidad (figura 6):

Figura 6. Localidad y punto de encuesta en CORPAT.

Además de permitir la consulta de formas por puntos de encuesta (figura 7), existe también la posibilidad de ver todos los puntos de encuesta, bien por atlas, bien en conjunto (figura 8):

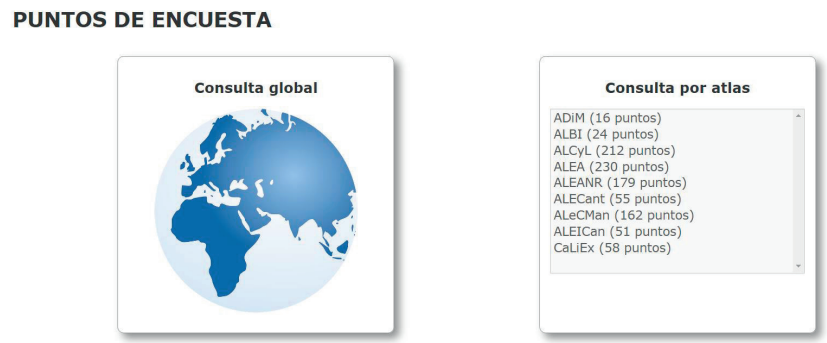


Figura 7. Consulta de formas por puntos de encuesta.

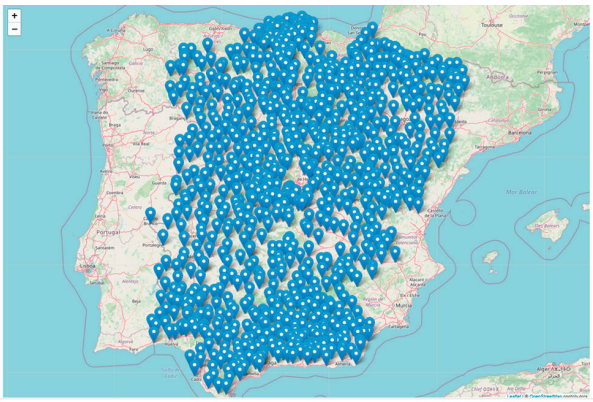


Figura 8. Puntos de encuesta de los atlas en *CORPAT*.

Igual que en algunos de los campos anteriores, se han tenido que ajustar y unificar algunas informaciones relativas a la codificación que generaban algunos problemas en el proceso de geolocalización. Por un lado, se han actualizado los nombres de algunas poblaciones bien por cambios ortográficos bien porque en la búsqueda actual del nombre aparecía información que no se halla en el atlas (tabla 1); se trata de un problema al que otros investigadores han hecho alusión con anterioridad (Pato, 2004, p.123-125).

Tabla 1. Algunos ejemplos los cambios de nombres de localidades.

Atlas	Punto de encuesta	Nombre en el atlas	Nombre en CORPAT
ALEANR	Na 103	Arcos	Los Arcos
	Na 303	Salinas	Salinas de Ibargoiti
	Lo 303	Tovía	Tobía
	Vi 600	La Guardia	Laguardia
ALEA	J 102	Isabela	La Isabela
	J 600	Pozo-Alcón	Pozo Alcón
ALEICan	L P 1	Garafía	Villa de Garafía
ALECant	S 202	Mortera de Piélagos	Mortera
ALCyL	Bu 602	Pinilla	Pinilla de los Moros
ALeCMan	GU 310	Abádanes	Abánades

Además de esta falta de coincidencia parcial con el nombre actual, en el *ALCyL* se han encontrado dos puntos de encuesta que tienen el mismo nombre: So 502 y So 602 se refieren a *Torrevicente*, aunque actualmente no se han podido identificar dos localidades con el mismo nombre. Siguiendo la ubicación del mapa del atlas, se ha identificado Torrevi-

cente en So 502. Además de estos casos, también se han tenido que modificar algunos de los códigos de los puntos de encuesta porque coincidían en más de un atlas y ello generaba un conflicto al etiquetar la localidad. Esto ha sucedido en los puntos de encuesta del *ALEANR* situados en Soria (So 400, So 402 y So 600) y Burgos (Bu 400), ya que el código empleado coincidía con el del *ALCyL*. Como se trata solo de cuatro casos, se han modificado ligeramente los nombres añadiendo una tercera letra a la abreviatura del nombre de la provincia. Así, los cuatro puntos del *ALEANR* mencionados se hallan en *CORPAT* etiquetados como Bur 400, Sor 400, Sor 402 y Sor 600, por lo que no existe posibilidad de confusión con los puntos del *ALCyL*. En el caso de los nombres de puntos de Cuenca y Guadalajara del *ALeCMan* que coinciden con algunos del *ALEANR*, no existe posibilidad de confusión porque en el atlas de Castilla-La Mancha las letras del código aparecen en mayúscula (CU 200, CU 400; CU 200 y CU 400) y en el *ALEANR* en minúscula (Cu 200, Cu 400; Gu 200, Gu 400).

5. Conclusión

El corpus, sobre el que se han descrito brevemente algunas de las funcionalidades (principalmente relativas al vocabulario dialectal) y características que presenta en esta primera etapa de su desarrollo (muy preliminar), se ha diseñado como herramienta complementaria a los corpus textuales y obras lexicográficas del español. No pretende, en ningún caso, sustituir ni al atlas ni a los mapas que lo conforman, pues constituyen documentos genuinos de un valor incalculable, sino que persigue la protección del patrimonio histórico, cultural y artístico. Consideramos, de acuerdo con Sousa (2017), que tanto los atlas como su contenido forman parte de los bienes materiales e inmateriales de la historia de la lengua española y que es necesario invertir tiempo en preservarlos antes de que se pierdan y el fruto de tanto esfuerzo económico y científico acabe olvidándose.

Referencias

Alvar, M. (1976). Ordenadores y geografia lingüística: el proyecto del Atlas plurilingüe de Europa (ALE). *Revista de la Universidad Complutense*, *XXV*, 78-85.

Alvar, M. y Nuño, M.ª P. (1981). Un ejemplo de atlas lingüístico automatizado: el *ALES*. *Lingüística Española Actual*, *3*(2), 359-370.

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

Aurrekoetxea, G. (2019). Sobre el valor de la dialectometría en la delimitación de las distancias lingüísticas. *GLOSEMA*. *Revista Asturiana de Llingüística*, *1*, 19-39.

Bonilla, J. E. y Bernal Chávez, J. A. (2020): Modelamiento de una base de datos espacial para el *Atlas Lingüístico-Etnográfico de Colombia*. *Revista Signos. Estudios de Lingüística*, *53*(103), 346-368. http://www.revistasignos.cl/index.php/signos/article/view/106/202

Chambers, J. K. & Trudgill, P. (1994). *La dialectología*. Visor Libros.

Coseriu, E. (1977). *El hombre y su lenguaje. Estudios de teoría y metodología lingüística*. Gredos.

Davoine, P.-A., Gally, S., Garat, P., Chauvin, C., Copi, O., & Cavalière, C. (2015, August): New approach to explore and to study cartographical heritage in dialectology: application to the Linguistic Atlas of France. *27th International Cartographic Conference (ICC 2015)*, Rio de Janeiro, Brazil. https://icaci.org/files/documents/ICC_proceedings/ICC2015/papers/18/fullpaper/T18-298_1430301727.pdf

Del Barrio de la Rosa, F. (2018). *Espacio variacional y cambio lingüístico en español*. Visor.

Enríquez, E. (1986). Análisis automático de la información fónica contenida en los Atlas lingüísticos. *Lingüística española actual*, *7*(1), 93-131.

Fernández Morell, M.ª L. (2019). Los nombres de animales y vegetales como patrimonio lingüístico alpujarreño a partir de los datos del proyecto VitaLex. *Proyecto Vitalex*. http://www.proyectovitalex.es/pdf/articulos/2019-05-publicaciones.pdf

Fernández-Ordóñez, I. (2011). *La lengua de Castilla y la formación del español*. Discurso leído el día 13 de febrero de 2011 en su recepción pública. Real Academia.

Fuster, M.ª T. (1996). Voces de creación metafórica sobre el maíz y el trigo en el *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja*. *Estudios de Lingüística de la Universidad de Alicante* (*ELUA*), *11*, 139-147.

Fradejas, J. A. (2020). *Cuentapalabras*. *Estilometría y análisis de texto con R para filólogos*. Universidad de Valladolid. http://www.aic.uva.es/cuentapalabras/

García Mouton, P. (1999a). *Cómo hablan las mujeres*. Arco/Libros.

García Mouton, P. (1999b). Dialectometría. En J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos* (pp. 335-356). Editorial Milenio.

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. (2016). Corominas tenía razón: *jamila* no *jámila*. En M. Quirós (Ed.), *Etimología e historia en el léxico del español. Estudios ofrecidos a José Antonio Pascual (Magister bonus et sapiens)* (pp. 293-302). Iberoamericana/Vervuert.

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

García Mouton, P. y Molina Martos, I. (2009). Trabajos sociodialectales en la comunidad de Madrid. *Revista de Filología Española*, *89*(1), 175-186.

Heap, D. (2002). Segunda noticia histórica del *ALPI*. *Revista de Filología Española*, *82*(1/2): 5-19.

Herrgen, J. (2010). The digital wenker atlas (www.diwa.info): An online research tool for modern dialectology. *Dialectologia: Revista electrònica*, *I* (Special Issue), 89-95.

Hoch, S. C. & Hayes, J. J. (2010). Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin*, *51*(1), 23-36.

Ibarretxe-Antuñano, I. (2021). Empatía lingüística. *Archiletras / Revista de Lengua y Letras*, *9*.

Julià Luna, C. (2020). Reseña a Alberto Manuel Arias García y Mercedes de la Torre García (2019): *Ictionimia andaluza. Nombres vernáculos de especies pesqueras del "Mar de Andalucía"*. Madrid: CSIC. *Dialectologia et Geolinguistica*, *28*, 163-172.

Lance, D. M. & Slemons, S. V. (1976). The use of the computer in plotting the geographical distribution of dialect items. *Computers and the Humanities*, *10*, 221-229.

Limper, J., Pheiff, J. & Williams, A. (2020). REDE SprachGIS: A Geographic Information System for Linguists. In S. Brunn & R. Kehrein (Eds.), *Handbook of the Changing World Language Map* (pp. 3743–3771). Springer. https://doi.org/10.1007/978-3-319-73400-2_145-1

Llorente, A. (1962). Fonética y fonología andaluzas. *Revista de Filología Española*, *45*(1/4), 227-240.

Molina, I. (2006). Innovación y difusión del cambio lingüístico en Madrid. *Revista de Filología Española*, *86*(1), 127-149.

Molina, I. (2018). Atlas lingüísticos castellanos: el *ALeCMan* y el *ADiM*. In *Coloquio Geolingüística Peninsular: investigaciones en curso* (pp. 1-9). Instituto de Lengua, Literatura y Antropología, CSIC, Madrid, 28 de septiembre de 2018.

Moreno, F., Moreno, J. E. y García de las Heras, A. (1997). Cartografiado automático y bases de datos. *Boletín de Filología de la Universidad de Chile*, *36*(1), 201-209.

Navarro Tomás, T. (1975). Noticia histórica del *Atlas Lingüístico de la Península Ibérica*. En *Capítulos de Geografía Lingüística de la Península Ibérica* (pp. 9-21). Instituto Caro y Cuervo.

Nerbonne, J., Heeringa, W., Prokić, J. & Wieling, M. (2021). Dialectology for computational linguists. In M. Zampieri & P. Nakov (Eds.), *Similar Languages, Varieties and Dialects. A Computational Perspective* (pp. 96-117). CUP.

Pato, E. (2004). La sustitución de *cantara / cantase* por *cantaría y cantaba*. Universidad Autónoma de Madrid. http://www.corpusrural.es/publicaciones/2004/2004_sustitucion.pdf

Putschke, W. (1969). Über ein Computerprogramm zur Herstellung von Sprachkarten, *Germanistische Linguistik*, *1*, 45-114.

Putschke, W. (1972). Planung einer Projektdurchführung: Automatische Kartierung des ATLAS LINGUARUM EUROPAE, *Germanistische Linguistik*, *4*, 547-577.

Prat Sabater, M. (2006). Reflejo espacial del cambio léxico: los atlas lingüísticos y el *DCECH*. *Actes del VII Congrés de Lingüística General (Barcelona, 18-21 de abril de 2006)*, 132-132.

Salvador, G. (1965). Estudio del campo semántico "Arar" en Andalucía, *Archivum: Revista de la Facultad de Filología*, *15*, 73-111.

Shuy, R. (1966). An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada. In P. L. Garvin (Ed.), *Computation in Linguistics: A Case Book* (pp. 60-75). Indiana University Press. https://publish.iupress.indiana.edu/read/74795c17-0e97-454a-926b-ad83db02cf76/section/6059e206-c137-403c-a001-96ddda48fdff#toc_8

Sousa, X. (2017). From field notebooks to automatic mapping: the *Atlas Lingüístico Galego* database. *Dialectologia et Geolingüistica*, *23*(1), 1-22.

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.

Tisato, G. (2019). Acquisizione Digitale dell'Intero *AIS*. Documento digital. https://www.aisv.it/aisv2019/abstracts/11.pdf

Tranquilli, R. (2019). *Atlas Lingüístico y Etnográfico de la provincia de Zaragoza* [Presentación]. Institución Fernando el Católico. https://ifc.dpz.es/index/alepz/Atlas_linguistico/Atlas_digital_provincia_de_Zaragoza/ALEPZ_DIGITAL

Uritani, N. y Berrueta de Uritani, B. (1985). Los diminutivos en los atlas lingüísticos españoles. *Lingüística Española Actual, 7*(2), 203-236.

Wood, G. (1969). Dialectology by computer. *International Conference on Computational Linguistics COLING 1969: Preprint* (19), 1-19. University Edwardsville.

Wood, G. (1971). Why Not a Computer as Editor? In L. H. Burghardt (Ed.), *Dialectology: Problems and Perspectives* (pp. 41-53). University of Tennessee.

Ziamandanis, C. M. (1996). Dialectología y ordenadores. En M. Alvar (Dir.), *Manual de dialectología. El español de España* (pp. 55-62). Ariel.

Fuentes primarias

ADiM = García Mouton, P. y Molina Martos, I. (2015): *Atlas Dialectal de Madrid*. CSIC. http://adim.cchs.csic.es/es

AIS = Jaberg, K. & Jud, J. (1928-1940): *Sprach-und Sachatlas Italiens und der Südschweiz*. Zofingen: Gedruck mit Unterstützung der Gesellschaft für Wissenschaftliche Forschung an der Universität Zurich und privater Freunde des Werkes von der Verlagsanstalt Ringier & Co., 8 vols.

ALEA = Alvar, M. con la colaboración de Llorente, A. y Salvador, G. (1961-1973). *Atlas lingüístico y etnográfico de Andalucía*. Universidad de Granada/CSIC, 6 vols.

ALEANR = Alvar, M. con la colaboración de Llorente, A., Buesa, T. & Alvar, E. (1979-1983). *Atlas lingüístico y etnográfico de Aragón, Navarra y Rioja.* La Muralla / Institución Fernando el Católico de la Excma. Diputación provincial de Zaragoza / CSIC, 12 vols.

ALECant = Alvar, M. con la colaboración de Alvar, Mayoral, J. A., Nuño, M.ª P., Caballero, M.ª del C. y Corral, J. B. (1995). *Atlas lingüístico y etnográfico de Cantabria*. Arco/Libros, 2 vols. [Etnografía y láminas de Elena Alvar].

ALEC = Instituto Caro y Cuervo (2018). *Atlas Lingüístico-Etnográfico de Colombia*. http://alec.caroycuervo.gov.co

ALeCMan = García Mouton, P. y Moreno Fernández, F. (2003). *Atlas lingüístico y etnográfico de Castilla-La Mancha*. Universidad de Alcalá de Henares. https://www.linguas.net/alecman/

ALEICan = Alvar, M. (1975-1978). *Atlas lingüístico y etnográfico de las Islas Canarias.* Publicaciones del Excmo. Cabildo Insular, 3 vols.

ALF = Gilliéron, J. & Edmont, E. (1902-1910). *Atlas Linguistique de la France*. Honoré Champion, 12 vols.

ALPI = García Mouton, P. (coord.), Fernández-Ordóñez, I., Heap, D., Perea, M.ª P., Saramago, J. y Sousa, X. (2016). ALPI-CSIC, edición digital de Navarro Tomás, T. (dir.): *Atlas Lingüístico de la Península Ibérica*. CSIC. http://www.alpi.csic.es/

ALPR = Navarro Tomás, T. (1948): Atlas Lingüístico de Puerto Rico. In *El español en Puerto Rico: Contribución a la geografía lingüística hispanoamericana*. Río Piedras. https://portfolio.umontreal.ca/view/view.php?id=255862

CaLiEx = González Salgado, J. A. (2005-2015). Cartografía Lingüística de Extremadura.

COSER = Fernández Ordóñez, I. (dir.) (2005-). *Corpus Oral y Sonoro del Español Rural*. www.corpusrural.es

DECH = Coromines, J. y Pascual, J. A. (1980-1991). *Diccionario Crítico Etimológico Castellano e Hispánico*. Gredos. Edición digital en CD-ROM (2012).

DLE = Real Academia Española (2014). *Diccionario de la lengua española*. Espasa Calpe. https://dle.rae.es/

DSA = Wenker, G. (1881). *Sprachatlas von Nord-und Mitteldeutschland.* Auf Grund von systematisch mit Hilfe der Volksschullehrer gesammeltem Material aus circa 30 000 Orten. Straßburg.

DynaSAND = Barbiers, S. *et al.* (2006). *Dynamische Syntactische Atlas van de Nederlandse Dialecten* (DynaSAND). Meertens Instituut. http://www.meertens.knaw.nl/sand/.

SCOSYA = Smith, J., Adger, D., Aitken, B., Heycock, C., Jamieson, E. & Thoms, G. (2019). *The Scots Syntax Atlas*. University of Glasgow. https://scotssyntaxatlas.ac.uk.

Chapter VI

The C-ORAL-BRASIL proposal for the treatment of multimodal corpora data: the BGEST corpus pilot project

La propuesta del C-ORAL-BRASIL para el tratamiento de datos multimodales en corpus: el proyecto piloto del corpus BGEST

Camila Barros & Heliana Mello
Federal University of Minas Gerais – Brazil

Abstract: Due to major technological advances, multimodal data treatment and compilation is a thriving possibility in Linguistics and provides new insights about the interplay of the sound signal and its corresponding gestuality in multimodal spontaneously produced data of how speech and gestures couple. This chapter discusses methodological issues associated with multimodal data compilation and treatment, especially regarding the crucial role of action. The main objective was to connect information structure organization, as it is treated through the Language into Act Theory – L-AcT (Cresti, 2000; Cresti & Moneglia, 2010; Moneglia & Raso, 2014), with the concept of spatio-motoric packaging as found in Kita & Özyürek (2003). The novelty of this methodological proposal stems from the crucial role prosody plays in the definitional categories found in L-AcT and its impact on the interpretation of gestures. The BGEST corpus, a pilot study within the C-ORAL-BRASIL research initiative, is presented as the basis of the discussion carried.

Resumen: Debido a los principales avances tecnológicos, la recopilación y el tratamiento de datos multimodales es una posibilidad animadora para brindar nuevas perspectivas sobre la interacción de la señal sonora con la gestualidad en datos multimodales producidos espontáneamente de cómo se acoplan el habla y los gestos. Este capítulo discute cuestiones metodológicas asociados con la recopilación y el tratamiento de datos multimodales, especialmente con respecto al papel crucial de la acción. El objetivo principal fue conectar la organización de la estructura de la información, tal como abordada a través de la Teoría de la lengua en Acto (Cresti, 2000; Cresti & Moneglia, 2010; Moneglia & Raso, 2014), con el concepto de empaquetado espacio-motor encontrado en Kita y Özyürek (2003). La novedad de esta propuesta sucede del papel crucial que la prosodia desempeña en las categorías informacionales de la L-AcT y su impacto en la interpretación de los gestos. El corpus BGEST, un estudio piloto dentro del grupo de investigación C-ORAL-BRASIL, es presentado como base para la discusión realizada.

1. Introduction

Technological advances have enabled researchers to study speech beyond its transcription. This has shown how much information is lost in the direct conversion of spoken texts to their written counterpart. Transcriptions can often be misleading and fail to provide a myriad of nuances that are crucial to the understanding of how speech is produced (Mello, 2014). Recently, the same conclusion could be drawn regarding multimodal data (Allwood, 2008). Considering that most daily human interactions happen in face-to-face contexts, what is lost if the study of these events is limited to their audio recordings?

The study of multimodal data may pose even bigger challenges than those found in speech data study when it comes to corpora compilation and treatment, because the process might demand even more planning and manual treatment. The use of high-quality equipment, such as wireless microphones, discreet cameras and powerful software is only part of the issue. Most of the work involved refers to pre-planning, in which the type of interaction, size, format, technical specifications, and usability of the corpus are established. These decisions affect directly the corpus and the kind of analysis that may be developed. Moreover, the theoretical path that led to the methodological decisions must be clear to enable a coherent analysis later on. The BGEST corpus will serve as a case-study, presenting methodological decisions designed to enable gesture-prosody interface studies, joining

the efforts made by both the study of speech and gesture in face-to-face interactions. These issues will be tackled in the next sections.

2. Background

The current state of the art of multimodal corpora resembles more closely an analytic heuristic for gesture studies than a set of guidelines towards the systematic collection of machine-readable linguistic data (Duncan, 2013). Most publications rely on *ad hoc* data collections that provide material for analysis but are not comparable to other data sets. Part of the problem is due to unclear legislation that fails to provide clear rules about how to guarantee participant anonymity while still making the data widely available. The other major problem is the amount of time required to collect, treat, and annotate the data.

While spoken corpora are growing in terms of length and automation, multimodal corpora fail to meet the criteria of variability, size and comparability that are common to spoken corpora. On multimodal corpora variability, Mello (2014) points out that the issues inherent to video recordings outnumber the available technical solutions. Alongside the additional costs, it is hard to predict how people will behave when video-recorded. A room filled with video cameras, as in a movie shooting, besides demanding enormous resources would impairs the intended spontaneity, even when the person is not camera-shy. The data treatment required also poses a constraint to multimodal data, regarding the time employed to select, edit, transcribe, and annotate the overwhelming amount of information that comes up in a recorded situation. When Loehr (2004, 2014) gave his first steps in this direction, he pointed out that annotation could take up to one hour per second of data: thus, only ten minutes of data could take 600h to be ready to be analysed.

Therefore, the corpus pre-planning phase should be guided initially by what can be feasibly accomplished (Mello, 2014). This means that, given the current possibilities of data compilation, it is better to have simple and well-structured data than to have many unreliable excerpts that cannot be directly compared. In comparison to spoken corpora, the size must be shrunken, to enable careful annotation and internal variability, given the previously mentioned compilation issues. A case study of the BGEST corpus, a multimodal corpus pilot project, stemming heavily from the C-ORAL spoken corpora family is presented in the following sections. Many practical considerations had to be made, as the following paragraphs show.

The protocol conducted in the BGEST corpus was intended mostly to allow studies on the interplay of gesture and prosody according to the Language into Act Theory (Cresti, 2000; Moneglia & Raso, 2014), resulting in a multimodal corpus comparable to the monologue

section of the C-ORAL spoken corpora family. The Language into Act Theory is a corpus-based theory about informational patterning in speech. The theory establishes that prosody is a necessary interface between the linguistic content and illocutions (speech acts) conveyed through speech (Cresti & Moneglia, 2010; Moneglia, 2011; Cavalcante, 2015). This means that speech is conducted by the actions performed in interaction, such as a question, assertion, among many others, technically referred to as *illocutions*. The prosody carries (most of) the illocutionary force. In terms of analysis, the basic is unit are utterances, perceived as pragmatically and prosodically autonomous units, which convey the illocution. An utterance can be internally divided in tonal units. The unit which carrying the illocution is called Commentary and appears without internal divisions in the utterance. In case the utterance has internal divisions, other units frame the illocution complementing it with textual informational or with discourse markers, regulating interaction (Moneglia & Raso, 2014).

The intention behind the use of L-AcT as a theoretical background to compile a multimodal corpus was grounded on the actional basis that underlies both prosody and gesture (Wagner, *et al*., 2014). In prosody, action is portrayed through an illocution, a highly conventionalized form that conveys a speech act (Cresti, 2000; Cresti & Moneglia, 2010). In gesture, action comes as a representation that is not entirely conventionalized, but it is packed as spatio-motoric information complementary to speech (Kita & Özyürek, 2003). As such, the research question that guided our research proposes a deep look into how action may frame multimodal information.

Cantalini (2018) dealt with this question, analysing excerpts of recited and spontaneous speech by three Italian actors. The author analysed up to ten minutes of data in both typologies and concluded that the internal divisions in gesture are temporally compatible to prosodic breaks, both terminal and non-terminal. Her research also showed that gestures align to speech at the lexical, informational and illocutionary levels. These findings may be seen as evidence that the informational patterning has a role in the organization of speech *and* gesture.

The BGEST corpus architecture was drawn from the C-ORAL corpora family, a multi-language corpora compilation project covering all major Romance languages (Cresti & Moneglia, 2005), including Brazilian Portuguese (Raso & Mello, 2012), Angolan Portuguese (Rocha, *et al*., 2018) in addition to English (Cavalcante & Ramos, 2016). The major difference between the C-ORAL corpora to other spoken corpora initiatives is the variability of situations portrayed, pre-planned to accurately encompass diaphasic variation. Diastratic variation resulted from the variation of recording situations and the diatopy was restricted to a metropolitan regional variety.

The C-ORAL family documents both formal and informal spontaneous registers, besides telephone conversations, television discourse, conferences, political debates, and teaching. Informal texts are normally not shorter than 1,500 words (around ten-minute recordings) and never longer than 3,000 words. This constraint warrants textual autonomy, but it does not overtly represent idiosyncratic characteristics (Mello, 2014). The two registers branch into public and private/family contexts. The division between public and private/family contexts takes into account the role the participants exercise in the interaction. Within the C-ORAL family, cultural differences moulded these definitions. Here, we only consider the C-ORAL-BRASIL.

Regarding the architecture above, some considerations must be made to make a multimodal corpus feasible. The first concern is the time required for data treatment, which will inevitably reduce the text's size. In gesture study tradition, texts are considerably small: Loehr (2004) analysed 164 seconds (summing up 147 gestures) in four dyatic interactions. Other authors worked with smaller time stretches: Condon and Ogston (1966) analysed five seconds of psychiatric consultations, Kendon (1972) worked on 90 seconds of data collected at a pub, and McClave (1994) analysed 125 gestures extracted from hours of filmed conversations. McNeill (1992) worked with 790 gestures in six different languages in elicited monologues. Cantalini (2018) analysed around 10 minutes of spontaneous speech and seven minutes of recited speech, which were used as a model for our research.

In an attempt to select texts that were not overtly long but still held their autonomy, it was decided that they should be no shorter than two minutes and no longer than three minutes. This provided around 400 words and 45 gestures per text. Regarding the type of register and its branches it was settled that a private/family informal context was more adequate to create a friendly environment that could compensate for the recording equipment embarrassment effect.

The C-ORAL family corpora have as a primary goal to be as representative as possible of the diaphasic variation in spontaneous speech. This motivation is based on the fact that the linguistic structure of a communicative event drastically changes from one situation to another, regulated by the ongoing activities. Monologues, interactions in which one speaker holds the floor to explain or tell a story, follow a semantic trajectory in which the main actionality is the speaking process itself. Dialogues and conversations have at least two participants that "perform co-dependent speech actions" (Mello, 2014, p.37). While monologues are more informative, with a richer elaboration of its content, dialogues and conversations are less informative but richer with respect to their illocutions. The C-ORAL family is divided in one third monologic and two thirds dialogical (dialogues and conversation) texts. The

justification for such division is grounded on the necessity to replicate what is found in authentic interactions and represent different degrees of interactivity, especially regarding different levels of actionality. In the BGEST corpus, adjustments had to be made to encompass the specificity of gesture capture, as gestures become more elaborate as the linguistic content complexifies. Thus, the BGEST corpus compilation was restricted to monologues, as their illocutionary monotony would be compensated by a richer gestural production.

Restricting the text typology to private/family monologues, the diaphasic variation was compromised. A greater diaphasic variation would require a whole set of cameras around one environment that allowed participants to move around freely, as the lapel-microphones do. For the moment, the amount and kind of data that monologues provided suffice for the analysis of the relation between gesture and prosody in this textual type.

The BGEST corpus followed the C-ORAL-BRASIL I guidelines (Raso & Mello, 2012) regarding the diatopic variety, capturing speakers aged 18 to 40, living in the metropolitan area of Belo Horizonte for at least two years, 50% of them originally are from that city. Ten participants are recorded in the almost 4,000 words comprising the BGEST corpus. Six of them are female and four are male, each one responsible for roughly 10% of the words uttered. All the participants were either enrolled in an undergraduate course or held college degrees. To avoid code-blending phenomena (Casey & Emmorey, 2009; Emmorey *et al.*, 2008), in which fluent sign language speakers gesture with signs while using an oral language, the participants who were fluent in Brazilian Sign Language were excluded (one participant). The dominant hand was controlled to guarantee that there was no side bias (eight were right-handed and two were left-handed). An analysis conducted after the data collection concluded that the gesture position and the dominant hand do not hold any correlation ($\chi^2 = 0.1(1)$, $p < 0.75$).

3. Data collection and treatment

3.1. Recordings

After the architecture was settled, recordings took place. The main concerns in this task were acoustic quality, video recording and gesture production. The participants provided their consent to the data collection beforehand, as well as their legal consent to image usage rights. There are still no clear guidelines in Brazilian legislation regarding how image can be distributed, which leads to the videos being only available to the research group members involved in the project. Participants' identities are not revealed, and they are only referred to by a codified sequence of letters.

For the BGEST corpus, the participants were recruited using the main researcher's personal network. The researcher would refer to the project without mentioning the specific interest in gestures, asking for an appointment at the participant's earliest convenience and, if given permission, beginning the recording. A comfortable situation was crucial to assure adequate data collection, especially considering that the recordings took place during the 2020's coronavirus pandemic.

The recording should enable high quality audio and video, in a way that allows phonetic studies and gesture analysis. The first constraint is easy to be overcome using high quality equipment, such as wireless lapel microphone system (Sennheiser EK100G3) preferably accompanied by a dedicated recording device (TASCAM DR-100MKII). This equipment has a friendly and non-invasive size that favours the recording session as it is easily forgotten by participants. The video recordings posed problems of a different nature: the image resolution for analysis does not need to be extremely high (e.g., 480p is sufficient when the facial expression is not relevant, according to ELAN's guidelines), but it should encompass different angles of the participant. This enables the participant to freely move while talking, not being constrained to a specific frame. Two or more cameras also give a three-dimensional sense to the footage, allowing fine-grained perception of gestures. The cameras should capture the participant's upper limbs to the extent of wide-open arms and should be placed as out of sight as possible. A simple, yet successful way to accomplish this is to place the researcher in between two cameras. By doing so, the participant tends to look more directly at the researcher than at the cameras. This also prevents the embarrassment that a recording session may cause to participants, because they usually forget about the equipment in a few minutes and carry the interaction naturally.

The distance in which the participants should be placed depends on the kind of lenses used. In the BGEST corpus, two kinds of lenses were used: 35mm (Panasonic HC-X900M) and 10mm (GoPro Hero 7). 35mm lenses are more common, accessible and distort less the image. Because the camera must be placed on a tripod at least 1.2 m from the participant, it draws some attention and has an inherent risk of something extraneous occurring in between the lenses and the participant (someone walking by, for instance). 10mm lenses have a smaller focal distance with a resulting broader angle of view, which causes a bigger distortion. Even so, the smaller design and higher stability (it does not require a tripod), makes it easier to be placed out of sight. Because it can be placed closer to the speaker, it diminishes the risk of something coming in between the lenses and the participant.

Following Mello's (2014) guidelines, some experience is required to find the equipment finest tune and recording of more time than what is intended to be transcribed should be done. This is important due to three main reasons (Mello, 2014, p.49):

a to allow for the possibility of choosing the best acoustic quality excerpt;
b to allow for the possibility of choosing the most interesting and actional excerpts;
c to allow for the possibility of choosing more than one excerpt from the same recording session.

The recording sessions were up to one hour long. This was more than enough for the participants to get acquainted with the situation, speak freely and (hopefully) move their hands. Excerpts up to three minutes long were collected from each recording, in which the participant was holding the floor for at least 30 seconds (Loehr, 2004). Each excerpt was then analysed concerning the informational units used and how comfortable the participant seemed. Out of fourteen recording sessions, one was excluded because the participant was fluent in Brazilian Sign Language (to avoid code-blending), three were excluded because the participants did not feel comfortable during the session or requested to be excluded. One was partially censored upon the participant's request. In the ten remaining recordings, three to five excerpts were analysed to meet the 30 second criteria. Out of each recording session, only one excerpt up to three minutes long was chosen.

The acoustic quality of the audios was measured by the script provided by Ferrari, Mello and Vieira (2020), also used on C-ORAL-BRASIL II (Raso *et al.*, in preparation). The criteria used for the analysis are fo, formants (F1 and F2) and signal-noise ratio. The method employed combines a series of Praat (Boersma & Weenink, 2020) measurements to a human evaluator's appraisal, which is crucial to double check all parameters. For audios from one to five minutes, five excerpts of two seconds long were analysed. Each parameter received a score and weighted average values with arbitrary weights were calculated. The tags are from A (best quality) to C (worst quality). The audios in the BGEST corpus received different tags: five were classified as (A), four were (AB) and one was classified as (B).

In the best-case scenario, all the recordings should be of (A) quality, to enable good prosodic analysis, as recorded by at least two cameras. However, because the recordings were carried during the coronavirus pandemic, attempting new recording sessions was not feasible.

3.2. Transcription, speech segmentation and informational tagging

The main points that must be taken into consideration in a transcription are the previous training of the team involved and decisions about which transcription criteria should be

adopted. The C-ORAL-BRASIL transcription guidelines were followed and are summarized in the following paragraphs (cf. Mello, 2014; Mello *et al.*, 2012). The overall architecture follows the CHAT guidelines (MacWhinney, 2000) used in the CHILDES project. This means that each speaker turn is represented on one line, started by a "*" followed a three-letter capitalized acronym for the participant. Each turn is delimited either by a non-terminal break "/" or by a terminal break "//". In the BGEST corpus, as in the C-ORAL family, terminal breaks signal pragmatic- and prosodically autonomous utterances, according to L-AcT (Cresti, 2000; Moneglia & Raso, 2014). Interruptions are delimited by a "+" and cancelled words are marked by the following convention: a "&" precedes the interrupted word and "[/n]" indicates how many words have been retracted. Other linguistic phenomena are represented by a symbolic convention: "hhh" indicates paralinguistic sounds such as laughter and coughs; "&he" indicates hesitation or taking time (regardless of the vowel enunciated); "<>" angular parentheses signal an overlap; "yyyy" indicates an incomprehensible sequence; and "xxx" indicates an incomprehensible word.

Example 1. Main criteria used in targeting – bgest_010[2-4]: [1]

*CLA: [2]	há eu não vou conseguir lembrar // <mas> +
	hhh I am not going to remember // <but> +
*CAM: [3]	<mas> cê era do / lado da promotoria ou do +
	<but> you was on / [the] prosecution or on +
*CLA: [4]	não / do juiz mesmo // &j [/1] então / a promotoria seria no / criminal //
	no / on the judge['s side] actually // &j [/1] so / the prosecution would be in criminal [law] //

In the example above (1), the speaker CLA laughs at the beginning of an utterance and produces an interrupted utterance, marked by "+". In [4], CLA stutters "j" at the beginning of the second utterance, abandoning the word. This is marked by "&j" with [/1] indicating that the previous word has been cancelled.

Orthographic conventions aim to guarantee readability, reliability and ease in the following computational treatment. Non-orthographic criteria tried to capture on-going phenomena of grammaticalization and lexicalization in Brazilian Portuguese, such as the apheresis of the verb *ser* (to be), as in *tá (>está), tar (> estar), tamos (> estamos)* forms. Phenomena, such as production and agreement errors are noted in the metadata that accompanies the transcription. Acronyms and abbreviations can be transcribed in two ways: only in capitaled words if uttered as a single word (e.g., *SUS*), or, when they are uttered as a sequence of letters, as syllables formed by a single letter (e.g., *uefeemegê* – UFMG/Federal University of Minas Gerais).

1 The icons indicate an associated audio or video that can be accessed in <>.

Example 2. Transcription incorporating orthographic and non-orthographic conventions: apheresis and cliticization – bgest_007[19]:

*CAR: [19] aí eu [/1] ele respondeu que e' tava bem / aí ele [/1] aí eu / cê sumiu né //
then I [/1] he replied that he was fine / then he [/1] then I / you've disappeared right //

In the example (2) above, other conventions are presented. In [19], the apheretic forms *tava* (>*estava*), as well as the cliticization of the subject pronouns *cê (>você)* and *e' (>ele)* are portrayed. The revision of the transcripts took place in two stages. The first, shortly after transcription, was performed by experienced reviewers from the C-ORAL-BRAZIL group. The second happened during the informational annotation also conducted by experient annotators from the C-ORAL-BRAZIL group.

The segmentation of recorded stretches of speech followed L-AcT in its assumption that utterances make up the basic pragmatic unit of study. Here, it will be argued that the pragmatic definition used by the Language into Act Theory is compatible with gesture studies for two reasons: it is grounded on the same actional principles that are believed to regulate and organize speech, and it is easily implemented. Furthermore, it will be argued that the segmentation of gesture and speech cannot be conducted separately.

As briefly said in section (2), the BGEST corpus is grounded on the L-AcT analytical categories. This theory holds as the basic unit of analysis the utterance, as it can be prosodically and pragmatically interpreted and conveys a speech-act. When an utterance only carries a single information unit, it necessarily corresponds to a Comment unit, i.e., the informational unit that conveys the illocution an utterance is simple if it only conveys one information unit and it is complex when it portrays two or more units. The informational units that frame the illocutionary one (Comment), can be either textual or dialogic units. Textual units make up the linguistic content in the utterance and can be: Topic, Appendix of Comment or Topic, Locutive Introducer and Parenthetical. Dialogic units can be roughly referred to as units that regulate the interaction (Raso, 2014; Raso & Vieira, 2016). Their specifications will not be explored in this paper (Moneglia & Raso, 2014). In some cases, the isomorphism of one illocution per utterance is not held, in which case there are textual units named Stanzas. This happens often in monologues, in which the textual content is divided in Bounded Commentaries, which indicate a sign of prosodic continuity, or in Multiple Commentaries, which form a prosodic pattern. Utterances can accommodate scanned units, which take place when the speaker must divide her/his uttering of speech for reasons other than to convey an information unit, e.g., breathing (Moneglia & Raso, 2014).

This approach differs from others focused on the syntactic or interactional segmentation of speech, based on complete predications or speech turns. By doing so, L-AcT is able

to describe more accurately verbless units and large differences in turn divisions caused by different text typologies (Cavalcante, 2015). Concerning gestures, this approach also differs from "apex-guided" approaches, such as Loehr (2004) that looked for the rhythm alignment of the apex of gestures and pitch accents following the ToBI model (Pierrehumbert, 1980). The L-AcT approach towards gestures is tightly bound to the coordination of prosodic breaks and the manual patterns that are associated with informational units. This has the practical benefit of being more easily implemented than approaches that adopt gesture-speech dissociated criteria.

Another layer of annotation that was implemented in the BGEST corpus is the informational one.

3.3. Gesture annotation

Gesture annotation followed the definitions proposed by Kendon (1972, 2004) organized in a hierarchy by Kita, van Gijn, and van der Hulst (1998), and systematized by Bressem, Ladewig, and Müller (2013). The gestural annotation was performed in the ELAN software (Wittenburg *et al.*, 2006)[2] a multimodal, free and open-source data, annotation tool. The annotation adopted in the BGEST corpus is simplified in relation to the protocol provided by Bressem, Ladewig, and Müller (2013), thus, it provides only crucial information about movement, direction, hand shape and spatial position.

The gesture is basically defined by its expressive phase, an energy peak that constitutes the semantic part of it. The stroke may be preceded by a preparation phase and followed by a retraction phase. The linear structure of (preparation), stroke (and retraction) is called a gesture phrase. They can be either isolated or compounded by sequences of phrases that are delimited by a rest position (when the hands and arms are relaxed). A sequence of gesture phrases is called a gesture unit. As an example of this first explanation, an excerpt of the bgest_001 file is shown. It is synchronized to the utterance "aí minha mãe conheceu meu pai lá //" (en. *and then my mom met my dad there //*).

2 http://www.tla.mpi.nl/tools/tla-tools/ela.

Figure 1. Gesture excursion (bgest_001, GU: 106, GP: 214).

The participant (JUL) initially has her hands on her lap in a rest position. Then, JUL raises her right hand in a flat form handshape towards the center. In the third frame, the retraction of the gesture is depicted. As there is only one movement peak, the gesture is a single phrase and unit.

It may happen that the stroke is composed of a series of repetitive movements, defined by Kita, van Gijn and van der Hulst (1998) as a repetitive phase, included in the attack label. When the stroke has a static peak of movement (McNeill, 2005), the stroke label is used and the hold marked in the movement tier. Figure 2 shows an excerpt from bgest_003, synchronous to the utterance "a ideia é tipo você quebrar isso em [/1] em / compreensão / né / &he / discussão / e reprodução / basicamente / né //" (en. *the idea is basically that you break it in [/1] in / comprehension / right / &he / discussion / and reproduction / basically / right //*).

Figure 2. Gesture excursion of a unit compound by three gesture phrases (bgest_003, GU: 103).

Each frame illustrates a different stroke with no rest position in between. This is a gesture unit compounded by three phrases.

At the level of fine-grained detailing, the annotation simplifies the protocols adopted. First, gesture types as indicated by McNeill's (1992) were not included. This decision was made, because this specific annotation would require an extra validation step that would not be feasible in the time available for the research. Another difference is that the annotation was done with sound support, as "(…) if the goal is to annotate the co-speech gesture then the removal of the information relating to speech, with respect to which the gesture finds relevance, does not seem justified as it eliminates perceptually relevant information for its identification." (Cantalini & Moneglia, 2020, p.11). This decision is supported by Loehr (2004) and Cantalini (2018).

As for the three levels of annotation for gestures predicted by Bressem, Ladewig, and Müller (2013), only some of the features were annotated. The annotation stage includes the three levels listed, all mandatory:

1 Determining units: **gestural unit** and **gesture phrase**;
2 Annotation of form: **hand shape, orientation, spatial position, movement type**, direction of movement, movement quality;
3 Motivation of form: mode of representation, action, motor pattern and image schema.

Only the bold items were noted, taking into account that i) this step was simplified so that the annotation was informative, but not excessive; ii) the motivation of form was not initially considered as relevant and, therefore, not considered in this annotation. Each of the annotated parameters is briefly explained below.

Handshape is annotated according to its form during the stroke. The fingers used were not annotated, for the hand shape was already informative enough for our purposes. The parameters are fist, flat hand, single fingers, and combination of fingers.

Figure 3. Hand configurations (Bressem, 2013, p.1085).

Orientation refers to the orientation of the palm in relation to the body, using McNeill's definition (1992, p.380). The features refer to the sagittal axis (considering a line perpen-

dicular to the body), which define if the gesture is *towards center* or *away from center.* When the gesture moves in relation to the torso, it can be *towards body* or *away from body.* The diagonal orientation of the hand was not noted.

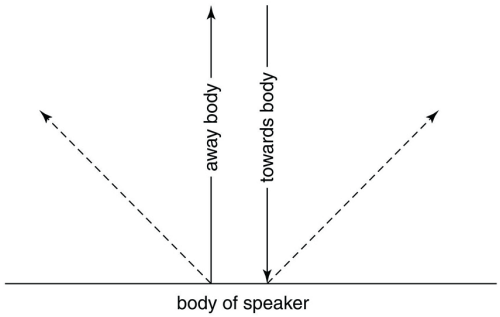


Figure 4. Orientation of movement (Bressem, 2013, p.1088).

There are six types of movement annotated in the corpus: *straight, arched, circle, spiral, zigzag,* and *S-line*.

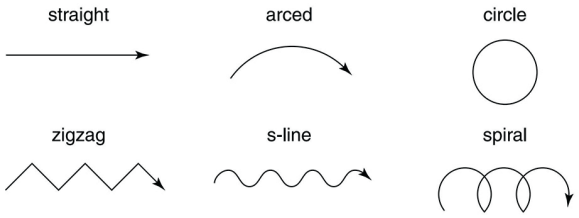


Figure 5. Movement types (Bressem, 2013, p.1088).

The spatial reference of the gesture is taken from McNeill (1992, p.86) and sets the parameters as *center-center*, *center*, *periphery*, and *extreme periphery.* They are arranged on a left-right and up-bottom axis, as shown below (Figure 6).[3]

3 To annotate all the 11 possibilities predicted in the amount of data available would only disperse the data. Thus, simplifying the annotation was a way to try to gain explanatory power.

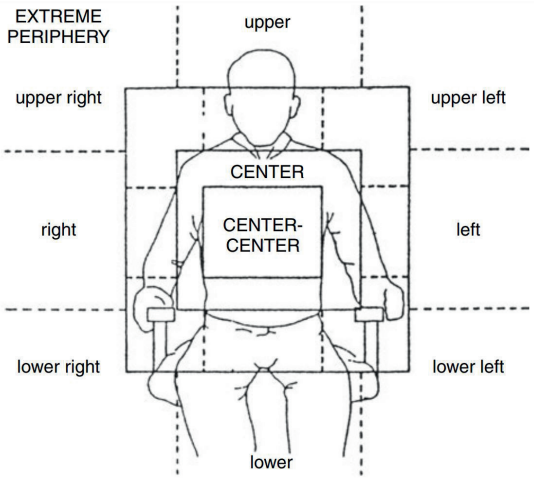


Figure 6. Gesture position (McNeill, 1992, p.89).

3.4. Usability

A multimodal, aligned corpus provides easy and ready access to sound, text and image of the excerpt under scrutiny, allowing fruitful exploitation of it. For BGEST, the text-to-sound alignment was done using Praat (Boersma & Weenink, 2020) and imported into ELAN (Wittenburg *et al*., 2006), where gesture and speech annotation were coupled. Both software were chosen because they are free, open source and the tiers can be imported from one to the other.

The annotation is hierarchically divided in tiers separated in speech and gesture. The speech tiers are annotated for terminal and no terminal breaks. The gesture tiers are annotated for gesture units, phrases, and phases. The phases are subdivided in orientation, movement, handshape, and position. This enables the user not only to watch the video and follow the transcription but also to export the alignment of the data in a data frame format, to be easily comparable. Ready access to the audio and video allows one to see how crucial the gesture and prosody interplay is to speech segmentation. This can be seen in the following example:

Example 3. Different segmentation possibilities based only on the transcription:

*GUI: e isso não só na cultura grega como a gente sabe que na cultura hebraica foi também a questão da procedência né judaico-cristã por muito tempo &he é tipo isso me diga com quem andas dirte-ei as manhas que tens sabe
and this not only in Greek culture as we know of the Hebrew culture was as well the matter of ancestry right Judeo-Christian [ancestry] for a long time &he it is like this tell me who do you walk with and I will tell you and I will tell who you are you know

The possible segmentations to this excerpt, without access to the corresponding audio, would be (almost exclusively) guided by a syntactic paradigm. Below are some possibilities for such a segmentation:

a [e isso não só na cultura grega como a gente sabe que na cultura hebraica foi também a questão da procedência né judaico-cristã por muito tempo] [&he] [é tipo isso me diga com quem andas dir-te-ei as manhas que tens sabe]

b [[e isso não só na cultura grega] [como a gente sabe que na cultura hebraica foi também a questão da procedência né judaico-cristã por muito tempo]] [[&he] é tipo isso me diga com quem andas dir-te-ei as manhas que tens sabe]]

c [[e isso não só na cultura grega] [como a gente sabe que na cultura hebraica foi também] [a questão da procedência [né judaico-cristã] por muito tempo]] [&he é tipo isso me diga com quem andas dir-te-ei as manhas que tens sabe]

d [[e isso não só na cultura grega] [como a gente sabe que na cultura hebraica foi também a questão da procedência né judaico-cristã por muito tempo]] [[&he] é tipo isso] [me diga com quem andas] [dir-te-ei as manhas que tens sabe]

In (a), we would have a complex clause followed by an assertion, without internal divisions. In (b), the first clause could be internally divided in two. In (c), the clause could be even more divided, with an insertion as "né judaico-cristã" (*right Judeo-Christian [ancestry]*). The last possibility envisioned without access to audio would be an internal division of the second clause.

Listening to the audio, the ambiguity concerning the syntactic organization of the utterance are restricted to two main possibilities, which would allow a corresponding accurate informational tagging.

e [[e isso] [não só na cultura grega] [como a gente sabe que na cultura hebraica] [foi também] [a questão da procedência] [né judaico-cristã] [por muito tempo] [&he] [é tipo isso]] [[me diga com quem andas] [dir-te-ei as manhas que tens] [sabe]]

f [[e isso] [não só na cultura grega] [como a gente sabe que na cultura hebraica] [foi também] [a questão da procedência] [né judaico-cristã] [por muito tempo]] [[&he] [é tipo isso] [me diga com quem andas] [dir-te-ei as manhas que tens] [sabe]]

Both possibilities sound plausible because they reflect the possible prosodic patterns. The doubt regards the placement of the terminal break that can follow *por muito tempo* or *é tipo isso.* The prosodic pattern supports both interpretations due to a sign of continuity in *por muito tempo*, weak enough to be a non-terminal break, but strong enough to not be

dismissed. The ambiguity is resolved by the video, which shows two gesture units aligned to each one of the utterances conveyed in the turn, thus leading to the segmentation in example 4.

Example 4. Final segmentation with audio and video (bgest_002[3-4]):

*GUI: [3] e isso / não só na cultura grega / como a gente sabe que na cultura hebraica / foi também / a questão da procedência / né / judaico-cristã / por muito tempo / &he / é tipo isso // [4] me diga com quem andas / dir-te-ei as manhas que tens / sabe //

Translation: *[3] and this / not only in Greek culture / as we know of the Hebrew culture / was as well / the matter of ancestry / right / Judeo-Christian [ancestry] / for a long time / &he / it is like this // [4] tell me who do you walk with / and I will tell you and I will tell who you are / you know //*

Figure 7. Different gesture patterns in the excerpt (bgest_002[3-4]).

In the first utterance, an iterated gesture with the right hand shaped in a combination of fingers moving in circles is made (frame 1). The second frame is synchronous to "por muito tempo" and is a straight movement. The third frame indicates how the participant used the rest position as a shifting device, implying it to indicate the termination of the last utterance. Another kind of pattern appears in "me diga com quem andas / dir-te-ei as manhas que tens / sabe //" with the right hand using the bench as support for a rhythmic gesture.

Without ready access to aligned transcription, audio *and* video, this discussion would not be possible, leading to misinterpretation of the data. Neither would it be possible to go through the audio, make measurements, and associate it with the gesture pattern.

This discussion indicates that despite the technological milestones that spoken corpora have reached, another stretch must be taken to include multimodal information in the analysis of human interaction. Despite the myriad of information in multimodal data, the gestures and facial expressions that appear in the data are of the utmost importance to accurately describe and understand ongoing communication processes.

4. Conclusion

The BGEST pilot project showed that the current state of art and technological devices at hand are not ideal but are sufficient to provide the means necessary for robust multimodal data compilation projects. It is crucial to understand the decisions that have to be made along the process and, by doing so, what is left behind, what is feasible and goals to be pursued in the future.

A set of those decisions was demonstrated in this paper, having action as its foundational point. The examples were drawn from the BGEST corpus to support the argument that the possibilities available for multimodal data compilation currently allow the development of multimodal corpora.

References

Allwood, J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 207-225). de Gruyter.

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

Bressem, J. (2013). 70. A linguistic perspective on the notation of form features in gestures. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/1* (pp. 1079-1098). de Gruyter. https://doi.org/10.1515/97831102613 18.1079

Bressem, J., Ladewig, S., & Müller, C. (2013). 71. Linguistic Annotation System for Gestures. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/1* (pp. 1098-1124). de Gruyter. https://doi.org/10.1515/97831102613 18.1098

Cantalini, G. (2018). *La gestualità co-verbale nel parlato spontaneo e nel recitato*. Università degli studi Roma Tre.

Cantalini, G., & Moneglia, M. (2020). The annotation of gesture and gesture/prosody synchronization in multimodal speech corpora. *Journal of Speech Sciences*, *9*, 7-30.

Casey, S., & Emmorey, K. (2009). Co-speech gesture in bimodal bilinguals. *Language and Cognitive Processes*, *24*(2), 290-312. https://doi.org/10.1080/01690960801916188

Cavalcante, F. A. (2015). *The topic unit in spontaneous American English* [Doctoral Dissertation]. Universidade Federal de Minas Gerais.

Cavalcante, F. A., & Ramos, A. C. (2016). The American English spontaneous speech minicorpus. *CHIMERA. Romance Corpora and Linguistic Studies*, *3*(2), 99-124.

Condon, W. S., & Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, *143*(4), 338-347. https://doi.org/10.1097/00005053-196610000-00005

Cresti, E. (2000). *Corpus del italiano parlato*. Accademia della Crusca.

Cresti, E., & Moneglia, M. (Eds.). (2005). *C-ORAL-ROM: Integrated reference corpora for spoken Romance languages*. J. Benjamins.

Cresti, E., & Moneglia, M. (2010). Informational Patterning Theory and the corpus-based Description of Spoken Language: The compositionality Issue in the Topic-Comment Pattern. In M. Moneglia & A. Panunzi (Eds.), *Bootstrapping information from corpora in a cross-linguistic perspective* (pp. 13-45). Firenze University Press.

Duncan, S. (2013). 65. Transcribing gesture with speech. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/1* (pp. 1007-1014). de Gruyter. https://doi.org/10.1515/9783110261318.1007

Emmorey, K., Borinstein, H., Thompson, R., & Gollan, T. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition*, *11*(1), 43-61. https://doi.org/10.1017/S1366728907003203

Ferrari, L., Mello, H., & Vieira, M. (2020). Reflexões sobre a classificação da qualidade acústica de dados de corpora orais. *Anais do Congresso Brasileiro de Prosódia*, *1*, 27-30.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177-210). Pergamon Press.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic variation. *Journal of Memory and Language*, *48*(1), 16-32.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer Interaction* (pp. 23–35). Springer. https://doi.org/10.1007/BFb0052986

Loehr, D. (2004). *Intonation and Gesture* [Doctoral dissertation, University of Georgetown]. University of Georgetown.

Loehr, D. (2014). 100. Gesture and prosody. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & J. Bressem (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/2* (pp. 1381-1391). de Gruyter. https://doi.org/10.1515/9783110302028.1381

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd Edition). Lawrence Erlbaum Associates. https://talkbank.org/manuals/CHAT.pdf

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45-66. https://doi.org/10.1007/BF02143175

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press.

Mello, H. (2014). Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL. In T. Raso & H. Mello (Eds.), *Studies in Corpus Linguistics* (pp. 27-68). John Benjamins Publishing Company. https://doi.org/10.1075/scl.61.01mel

Mello, H., Raso, T., Mittmann, M., Vale, H., & Côrtes, P. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In T. Raso & H. Mello, *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal* (pp. 125–176). Editora UFMG.

Moneglia, M., & Raso, T. (2014). Appendix: Notes on the Language into Act Theory. In T. Raso & H. Mello (Eds.), *Studies in Corpus Linguistics* (pp. 468-495). John Benjamins Publishing Company. https://doi.org/10.1075/scl.61.15mon

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* [Doctoral dissertation, Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy]. MIT repository.

Raso, T. (2014). Prosodic constraints for discourse markers. In T. Raso & H. Mello (Eds.), *Studies in Corpus Linguistics* (Vol. 61, pp. 411–467). John Benjamins Publishing Company. https://doi.org/10.1075/scl.61.14ras

Raso, T., & Vieira, M. A. (2016). A description of Dialogic Units/Discourse Markers in spontaneous speech corpora based on phonetic parame*ters*. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, *3*(2), 221–249. https://revistas.uam.es/chimera/article/view/6516.

Raso, T., & Mello, H. (Eds.). (2012). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Editora UFMG.

Raso, T., Mello, H., & Ferrari, L. (In preparation). *C-ORAL-BRASIL: corpus de referência do português brasileiro falado. II*.

Rocha, B., Mello, H., & Raso, T. (2018). Para a compilação do C-ORAL-ANGOLA. *Filologia e Linguística Portuguesa*, *20* (Especial): 139-157. https://doi.org/10.11606/issn.2176-9419.v20iEspecialp139-157

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209-232. https://doi.org/10.1016/j.specom.2013.09.008.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006*, 1556–1559. https://archive.mpi.nl/tla/elan

Chapter VII

Las tecnologías del lenguaje y las lenguas indígenas mexicanas: constitución de un corpus paralelo amuzgo-español

Human language technology and the indigenous languages in Mexico: the Amuzgo-Spanish parallel corpus

Antonio Reyes Pérez[a] & H. Antonio García Zúñiga[b]
Universidad Autónoma de Querétaro ([a]), *Instituto Nacional de Antropología e Historia* ([b]) – *México*

Resumen: En este artículo se describen las particularidades de la construcción del primer corpus paralelo amuzgo-español, el cual representa una fuente de datos reales para la investigación lingüística, particularmente, así como para el desarrollo de recursos y herramientas para lenguas escasamente representadas e, incluso, en peligro de extinción. Los procesos llevados a cabo durante la constitución del corpus se detallan de acuerdo con las siguientes fases: i) obtención de datos en la lengua mediante entrevistas realizadas en trabajo de campo, ii) transcripción de las entrevistas; iii) procesamiento de la señal sonora en PRAAT para realizar análisis espectrográficos; iv) creación de glosas y traducción al español; v) alineación semiautomática de traducciones a partir de la correspondencia lingüística entre lenguas. Finalmente, se muestra el resultado de la implementación del corpus en una plataforma web para la consulta pública.

Abstract. In this article, a collaborative project to build the first parallel corpus Amuzgo-Spanish is described. The goal of this project is to provide a source with data

collected from colloquial speech in Amuzgo (glossed and translated into Spanish) for research, as well as for the development of tools for scarce resources languages. The processes carried out to compile the corpus are described according to the following phases: i) data collection in Amuzgo by means of linguistic fieldwork; ii) data transcription; iii) acoustic data processing with Praat to carry out spectrographic analysis; iv) glossing and translating data into Spanish; v) semiautomatic alignment of translations. Finally, an open access tool is presented because of the corpus release.

1. Introducción

El lenguaje verbal es la vía más natural para que los seres humanos pueden manifestarse e interactuar entre sí. Las Tecnologías del Lenguaje Humano (TLH) buscan, desde una perspectiva que agrupa el conocimiento y las metodologías desarrolladas en diferentes campos y disciplinas, hacer que una computadora pueda analizar, interpretar, comprender y producir información que la faculte para la comunicación e interacción con cualquier ser humano a través del uso del lenguaje. Para lograrlo, además de un conjunto vasto de técnicas, métodos y algoritmos, es necesario que existan recursos que representen en un nivel micro el fenómeno lingüístico que sucede a nivel macro. En este sentido, una de las formas más comunes para representar el lenguaje verbal, sea en su vertiente oral o escrita, es la constitución de corpus lingüísticos. Con este tipo de recursos, todo sistema computacional podría tener estructurado el conocimiento lingüístico y así tener la posibilidad de determinar la estructura y significado de casi cualquier expresión lingüística (Manning y Shütze, 1999), desde la fonética y la fonología hasta el discurso, pasando por la morfología, la sintaxis y la semántica.

En este escenario de creación de recursos que sirvan como fuente de conocimiento, no solo para fines lingüísticos, antropológicos o sociales, sino incluso para cuestiones relacionadas con el desarrollo de tecnologías que permitan el tratamiento computacional del lenguaje, el trabajo realizado desde la segunda mitad del siglo pasado se ha centrado en un conjunto no muy amplio de lenguas, en donde el inglés es la lengua más representada; por citar un par de recursos muy conocidos, el Corpus Brown o el BNC. En este sentido, el español también ha sido una lengua que goza de una representación interesante en términos de corpus disponibles, baste mencionar tres de los más representativos: el Corpus de Referencia del Español Actual (CREA), el Corpus Diacrónico del Español (CORDE) y el Corpus del Español del Siglo XXI (CORPES). Asimismo, ha habido esfuerzos por representar algunas otras lenguas, muchas de ellas con una descripción lingüística muy completa, como es el caso del italiano, el árabe y el alemán, entre otras (Quasthoff *et al.*, 2006),

así como lenguas escasamente representadas, minoritarias o, incluso, en peligro de extinción (Prinsloo, 2015; Vinogradov, 2016; Midrigan *et al*., 2020). No obstante, hay muchas lenguas que en la actualidad carecen de representatividad, y no solo en términos de recursos, sino, en muchos casos, en términos de existencia de datos mínimos necesarios para realizar una descripción lingüística. Tal es el caso de varias lenguas indígenas mexicanas.

En México, además del español, coexisten más de 60 lenguas indígenas, con sus respectivas variantes, las cuales son, en algunos casos, ininteligibles entre sí. Esta enorme diversidad se describe en el *Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas* (INALI, 2008) en términos de 11 familias lingüísticas, 68 agrupaciones y 364 variantes. De las 68 agrupaciones identificadas, las más representativas en términos de hablantes son el náhuatl, el maya, el mixteco y el zapoteco. La primera con más de un millón de hablantes, la segunda con alrededor de 800,000 hablantes, mientras que las dos últimas con poco más de 400,000 hablantes cada una (INEGI, 2015). Del resto de lenguas, algunas no llegan a los 1,000 hablantes, mientras que algunas otras están en vías de desaparición[1]. Esta gran variedad de lenguas es, a todas luces, reflejo de una riqueza cultural y social, así como de una cosmovisión e identidad. No obstante, es evidente que desde la perspectiva de las TLH hay una insuficiencia de recursos, herramientas e, incluso, materiales lingüísticos para la gran mayoría de estas lenguas. Algunas de ellas, de forma sorprendente, a pesar de que han sido bien estudiadas y descritas.

Dado el contexto presentado, en este artículo se describe un trabajo interinstitucional (Universidad Autónoma de Baja California e Instituto Nacional de Antropología e Historia) relacionado con una lengua que, no obstante su estado de descripción y cantidad de hablantes, muestra ya un vínculo incipiente con las TLH: el amuzgo.

El amuzgo o *jnon3 nda3* se habla en algunas localidades de tres municipios de dos entidades federativas del sureste de México: Oaxaca y Guerrero. Cuenta con alrededor de 60,000 hablantes (INEGI, 2015). A pesar de que existen trabajos descriptivos importantes y notables (Buck, 2000 y 2018), la lengua no ha sido documentada ni descrita de forma exhaustiva. En términos gramaticales, el amuzgo se caracteriza por contar con un repertorio extenso de clases léxicas, lo que se manifiesta en una alta complejidad verbal (Smith y Tapia, 2002; Apóstol, 2014), un conjunto amplio de pronombres personales (Buck, 2015; Palancar y Feist, 2015), así como en el empleo de tonos fonológicos para la marcación de distintos significados morfológicos, tales como la posesión (Hernández *et al*., 2017; García *et al*., en prensa).

1 Algunos de los casos más extremos serían el ayapaneco, el oluteco, el tuzanteco, el *moocho'* y el kiliwa.

La constitución del corpus paralelo, que es el objetivo de este trabajo, se sustenta en la obtención de muestras reales de habla en amuzgo mediante entrevistas realizadas en campo con hablantes nativos de la lengua. Al respecto, es importante destacar que la creación de este recurso, además de ser un aporte para aumentar la atención a las lenguas escasamente representadas e, incluso, en peligro de extinción, permitirá el desarrollo de nuevos recursos que pueden aprovechar el conocimiento explícito e implícito de los materiales que integran el corpus. Por ejemplo, desde el ámbito de la traducción automática, para mejorar los procesos de alineación entre segmentos del texto origen y el texto meta o, por otro lado, para desarrollar sistemas de extracción de información sustentados en las características intrínsecas de la lengua.

A continuación se presenta la organización de los contenidos tratados en el artículo: en la Sección 2 se presentará el estado del arte de los trabajos de TLH relacionados con las lenguas indígenas mexicana. En la Sección 3 se detallarán algunas características lingüísticas representativas de la lengua amuzga. La Sección 4 describirá el proceso para la obtención de los datos orales, así como el procesamiento espectrográfico y textual de los mismos. En la Sección 5 se explicará el proceso de glosado y de traducción al español, así como el trabajo de alineación de las traducciones y la liberación de una primera versión del corpus en una plataforma web. Finalmente, en la Sección 6 se presentarán las conclusiones, centrando la atención en algunos resultados alcanzados, así como resaltando las líneas de trabajo futuro.

2. El tratamiento tecnológico de las lenguas indígenas mexicanas

De acuerdo con los datos presentados en el documento *Análisis del Sector de las Tecnologías del Lenguaje en México* (2018, p.49), la existencia de recursos en lenguas indígenas de América Latina es casi inexistente. Una de las principales causas, señalan, es la mínima presencia de datos, en el plano escrito, tanto en medios tradicionales, tales como textos impresos, así como en medios electrónicos, sean estos contenidos web o de redes sociales. A lo anterior, se puede añadir el hecho que se mencionó en la sección previa: hay varias lenguas indígenas que no cuentan con la descripción lingüística suficiente, ya sea porque no han sido atendidas en un sentido académico, o bien, porque su gramática es difícil o la consecución de datos es altamente complicada y, en ocasiones, riesgosa[2].

2 Al respecto hay que puntualizar que esta situación representa, además de una desventaja académica, una de las consecuencias inmediatas de lo que se conoce como brecha tecnológica o digital. Como se sabe, este es un problema de muy diversa índole (económica, educativa, informativa, política y social) que pone de manifiesto, por un lado, la marginación de las comunidades indígenas de México y otras latitudes del mundo y, por otro, la incapacidad para emplear, adquirir y generar recursos tecnológicos que, en un contexto generalizado de inequidad e injusticia, termina por excluir a estas comunidades (cf. Acosta & Aguilar, 2020; Arévalo, 2015).

A pesar de esta situación poco alentadora, en Mager, Gutiérrez, Sierra y Meza (2018), se listan algunos recursos digitales en estas lenguas. Entre ellos, destacan un par de corpus paralelos, así como herramientas para análisis morfológico para algunas lenguas de las familias otamangue y uto-azteca. De manera más específica, en tareas relacionadas con la constitución y explotación de corpus, se pueden citar los trabajos de Gutiérrez (2015) y Gutiérrez, Sierra y Hernández (2016) en los que presentan el trabajo realizado con un corpus paralelo náhuatl-español.

Por otra parte, en un artículo de 2016, Mager, Barrón y Meza describen un acercamiento a la traducción estadística automática entre dos lenguas que en términos tipológicos son muy diferentes: el wixarika y el español. Los autores detallan una aproximación basada en la descomposición morfológica para mejorar los procesos de alineamiento con las traducciones al español y paliar la ausencia de datos en wixarika (Mager *et al*., 2016, p.64-64). En otra línea de trabajo, en el proyecto Digging Early Colonial History[3] han utilizado técnicas y herramientas de PLN y aprendizaje automático para realizar tareas de anotación con documentos históricos, mayoritariamente en español, pero en los cuales también aparecen datos en lenguas como el náhuatl, el mixteco y el maya.

En trabajos más relacionados con la oralidad se puede citar la investigación publicada por Castellanos *et al*. (2019, p.21), en la que se detallan los resultados de una aproximación para evaluar la pronunciación de aprendices de lenguas indígenas, particularmente del mixe, aplicando técnicas de modelado y reconocimiento de voz. Asimismo, el trabajo desarrollado por Cruz y Waring (2019) acerca del uso de redes neuronales para el reconocimiento automático de voz en chatino o el de Adams *et al.* (2018), también para el chatino, en el que se focaliza la importancia y complejidad del proceso de transcripción y anotación de los datos orales, al igual que el tratamiento adecuado de la información tonal de esta lengua. Esto último es de suma importancia, puesto que el tono, como se verá más adelante en este trabajo, constituye un elemento de la lengua amuzga esencial para marcar (dotar de sentido) elementos gramaticales específicos, lo cual dista mucho de lo que ocurre en lenguas como el náhuatl, el wixarika o el mixe[4].

3 Información detallada del proyecto se puede consultar en https://www.lancaster.ac.uk/digging-ecm/es/inicio/

4 En efecto, la morfología y la sintaxis de estos dos tipos de lenguas, las tonales (el chatino), por un lado, y las no tonales (como el wixarika, uno de los ejemplos citados), por el otro, son diferentes. El primero de los casos es un ejemplo de lenguas no concatenativas (sus morfemas no están necesariamente representados por segmentos discretos, ya que el tono, o alguna derivación fonológica de este rasgo, es un recurso para la marcación; es decir, el tono no se ubica de forma exclusiva en un nivel léxico, sino que puede llegar a uno de contenido gramatical), en tanto que el segundo se trata de una lengua concatenativa discreta, esto es, siempre con morfemas segmentables.

Por último, desde una perspectiva más relacionada con la industria, se puede subrayar el trabajo realizado por algunas pequeñas empresas, así como grandes compañías como Google y Microsoft, que en conjunto con instituciones gubernamentales o académicas, han generado algunos recursos en lenguas indígenas mexicanas del tipo de repositorios de información, traductores o *apps* para su aprendizaje (cf. ASTLM, 2018:51-52).

3. Características lingüísticas del amuzgo

En esta sección se caracteriza la familia lingüística a la que pertenece el amuzgo con el propósito de facilitar la presentación de los rasgos lingüísticos esenciales de dicha lengua. Se verá que el término otomangue remite a un conjunto de sistemas complejos y diversificados.

3.1. Familia otomangue

La familia otomangue en su conjunto siempre ha sido objeto de interés debido, principalmente, a sus características lingüísticas, muy distintas a las de otras lenguas habladas en territorio mexicano, así como a la diversidad que existe en su interior. Pese a concentrarse en un espacio geográfico definido (la hipótesis que sustenta el origen y la integración de la familia considera al subtiaba y al mangue, hoy en día extintos, los cuales se hablaron en Nicaragua, lo cual rompería esta idea de continuum), cada una de las lenguas que componen la familia cuenta con un buen número de variantes, situación que obliga a pensar si se trata de una familia de lenguas o, más bien, de una macrofamilia de familias; esto es, algunas variantes, incluso, podrían llegar a considerarse lenguas diferenciadas de las otras variantes que componen a una agrupación, para emplear la terminología del Instituto Nacional de Lenguas Indígenas. Este es el caso de la llamada subfamilia amuzgo-mixtecana (Campbell, 1997), a la cual pertenece el amuzgo.

3.1.1. Subfamilia amuzgo-mixtecana

El conjunto de lenguas amuzgo-mixtecanas pertenecen al otomangue del este (Campbell, 1997: 158). En esta división también se encuentran el popoloca, el mazateco, el ixcateco, el chocho, el zapoteco y el chatino. Como se ha dicho, la variedad interna en estas lenguas es amplia. En el caso concreto del amuzgo se ha señalado que, en términos históricos, han existido tres variantes: Xochistlahuaca, San Pedro Amuzgos e Ipalapa (habría otra, Tlacoachistlahuaca, sobre la que no se conoce mucho). En la actualidad, se considera que solo en dos de estos municipios existen hablantes: Xochistlahuaca (Guerrero) y San Pedro Amuzgos (Oaxaca). Estas demarcaciones territoriales y administrativas conforman por sí mis-

mas dos variantes plenamente diferenciadas en casi todos los planos lingüísticos. No obstante, el INALI (2008) identifica cuatro variedades (amuzgo alto del este, amuzgo bajo del este, amuzgo del norte y amuzgo del sur). Por su parte, el resto de las lenguas amuzgo-mixtecanas tienen los siguientes números de variantes: mixteco (81), tacuate (1, la cual, en términos lingüísticos, parece haberse separado del mixteco), cuicateco (3) y triqui (4).

3.2. El amuzgo

Las características gramaticales del amuzgo se agrupan en torno de los niveles de análisis lingüísticos tradicionales[5]. De esta manera, en un sentido elemental, se reconocen aspectos fonético-fonológicos, morfológicos y sintácticos. Sin embargo, el amuzgo al ser una lengua en la que el tono (frecuencia acústica que se produce al interior de unidades fonológicas como la sílaba), además de las distinciones semánticas que produce en el léxico, (véase ejemplo 1), interactúa con la morfología (ejemplo 2) y la sintaxis (ejemplos 3 y 4)[6].

1. a. su^{2} ‘llano’
   b. su^{3} ‘copal’
2. a. ba’^{1} ‘su casa (de él/ella)’
   b. ba’^{4} ‘tu casa’
3. a. ki^{2}tsian^{3} ‘tigre’
   b. ki^{2}tsian^{3}an^{3} ‘el tigre’
4. a. ts’an^{2}jni^{2} ‘persona malvada’
   b. ts’an^{2}jni^{2}i^{1} ‘la persona malvada’

Como se puede ver en los ejemplos anteriores, un cambio en el tono (de medio a alto en 1a y 1b, así como de bajo a súper alto en 2a y 2b) comporta un cambio importante en el significado de la palabra. En los ejemplos de 3 y 4 lo que se muestra es la forma en la que se construye el sentido definido de una frase nominal, el cual también está asociado a un fenómeno tonal. Obsérvese que en 3b y 4b, ejemplos en los que las frases nominales se encuentran definidas, la última sílaba es una copia de la precedente. No obstante, en 3b el

5 Otro tipo de caracterizaciones de la lengua, como las de corte sociolingüístico, se delinean en varios sentidos. En las primeras secciones de este trabajo se incorporaron algunos de los datos más destacados en términos poblacionales. Al respecto se entiende que la descripción que se hace de una lengua en términos de las necesidades de las TLH debe ser lo más amplia posible o, por lo menos, tiene que estar apegada a los fines del trabajo. Por ejemplo, más allá de una descripción lingüística adecuada y profunda, en un ámbito donde la creación de recursos tenga que ver con lo judicial, sin duda, la pragmática, por un lado, y la entonación, por el otro, serían sumamente relevantes.

6 Los superíndices indican el tipo de tono: 1 bajo, 2 medio, 3 alto, 4 súper alto y 5 extra alto. Con estas posibilidades, se pueden formar ciertas combinaciones.

tono alto se mantiene en la sílaba que resulta de dicha copia, mientras que en 4b, esto no sucede. La explicación de esta circunstancia es que cuando el tono de la última sílaba de una palabra es medio, el llamado artículo definido no puede tener un tono medio, por lo que tiene que cambiar a uno bajo.

En concreto, el sistema fonológico del amuzgo se compone por 15 consonantes (entre las que se cuentan dos prenasales, tres que son producto del contacto con el español, la /p/, la /l/ y la /r/, así como una con baja frecuencia de uso, la /m/). Asimismo, existen 7 vocales, algunas de la cuales muestran oposiciones entre abiertas y cerradas, fundamentalmente en las medias (/e/, /o/), en tanto que otras tienen contrastes entre orales y nasales (de nueva cuenta, las medias, así como la baja, o sea, la /a/, y la anterior abierta, /ɛ/). Por otra parte, los tonos de la lengua son, en total, siete; cinco considerados de nivel (los explicados en nota 5: bajo, medio, alto, súper alto y extra alto) y 2 de contorno (medio-bajo, medio-alto)[7].

En cuanto a otros aspectos centrales de la lengua, esta es de marcación en el núcleo (salvo en las terceras personas), las relaciones sintácticas se dan por yuxtaposición, o sea, no se morfologizan y, como menciona Hernández (2019), el predicado no lleva de manera sistemática afijos para una referencia cruzada con el sujeto. Según Smith y Tapia (1984), el amuzgo presenta un orden de constituyentes Verbo-Sujeto-Objeto en las construcciones transitivas, mientras que para las intransitivas se mantiene el verbo en posición inicial. De igual manera, en palabras de estos autores, hay un sistema escindido en las intransitivas, de forma tal que la codificación es distinta entre las intransitivas agentivas, las intransitivas pacientivas y las intransitivas estativas.

El sistema de personas gramaticales se organiza en tres (primera, segunda y tercera) con sus distinciones respectivas entre singular y plural. En la tercera persona de plural se hace una diferencia entre inclusión del escucha y la exclusión de este. La complejidad morfológica ha obligado a proponer un peso fuerte de las clases léxicas.

4. Diseño del corpus: fase monolingüe en amuzgo

A continuación se describen las fases de trabajo para la construcción del corpus. En particular, las relativas a la obtención y procesamiento de los datos en amuzgo. Al respecto, es necesario remarcar que se trata de material recopilado en un ambiente natural, esto es, se planeó, registró y estructuró en campo. En consecuencia, el corpus se puede caracterizar como representativo de un habla natural, diverso y actual, en correspondencia con los

7 La complejidad fonológica de la lengua es amplia, por cuestiones de espacio no puede ser abordada aquí. Para mayores detalles, consúltese Hernández (2019).

grupos etarios que conforman la muestra. Estas características, sin duda, son las que, en determinado momento, resaltarán cuando la información se traduzca en aplicaciones relacionadas con las necesidades propias de la comunidad de habla, como aquellas relacionadas con la atención en servicios de salud y justicia.

4.1. Obtención de datos orales

Aunque en este trabajo se presenta una parte del corpus conformado, su totalidad engloba la participación de un grupo de personas adultas, jóvenes e infantes, tanto hombres como mujeres en cada subconjunto. En este sentido, hasta el momento se ha trabajado con dos personas en cada franja etaria (la cual no coincide necesariamente con la del sistema urbano debido a la forma de vida comunitaria en la que, desde la infancia, se adquieren responsabilidades familiares).

Asimismo, en relación con la información de corte social con la cual, tradicionalmente, se organiza e identifica un corpus, se consideraron las circunstancias de vida de cada participante con el propósito de observar su conocimiento, control y dominio de la lengua. Por ejemplo, se aplicó un pequeño y sencillo instrumento en el que se captó información referente a la frecuencia y los contextos de uso de la lengua. Esto contribuyó a catalogar a quienes colaboraron en la investigación en atención al bilingüismo o monolingüismo mostrados, o bien, a su actitud frente a la lengua (hablantes pasivos, por ejemplo). En las condiciones actuales del mundo, cada vez se hace más necesario abrir un espacio para hablar de la migración. En un estudio que parte de la configuración sistemática de un corpus, el estatus migratorio de las personas es relevante porque da una ilustración más precisa de su comportamiento lingüístico. En efecto, un(a) migrante reacciona, después de su experiencia como tal, de forma muy diversa a una interacción comunicativa. Al respecto, los extremos a considerar serían: desiste de hablar su lengua o se torna un(a) purista de ella. En el punto medio quedaría la facultad de introducir préstamos lingüísticos con mayor o menor resistencia. Lo anterior, no hay duda de ello, incide en el tipo de información que se recolecta y obliga a imaginar nuevas formas de documentación o, en todo caso, a la aceptación de la nueva realidad.

Debido a lo que se comenta (el panorama es mucho más complejo y amplio de que lo que aquí se presenta), a cada una de las personas que colaboró en la investigación se le solicitó una anécdota o historia de vida, propia o ajena, al igual que una narración tradicional; solamente en algunos casos se incluyeron diálogos y entrevistas. De esta manera, se procuró estructurar un corpus real, espontáneo, natural, diverso, con información suficiente, representativo y cuidado, en el que las diferentes fases de la vida cotidiana y formas de interacción

(con estructuras lingüísticas comunes y variadas) se encontraran representadas. Al final, se logró conformar un material cercano a las ocho horas de duración. En este trabajo se ejemplifica con la información concerniente a una narración (*La esposa del zorro*) en la que se relata el intento de rescate de la esposa del zorro, que emprenden, por separado, un tigre, una vaca y un conejo. Resulta llamativo que, en la cultura amuzga, se observe una divergencia de aquello que se ha mostrado en la tradición literaria conocida como occidental. En este caso no se trata de un animal astuto, inteligente, tramposo, malo, cizañero o sagaz, sino, más bien, de un ser pasivo que sufre y no actúa, no muestra ni coraje ni ánimo, lo que lo lleva a caer pronto en la desesperación. Por tal razón, el tigre, la vaca y el conejo, en diferentes oportunidades, le ofrecen su ayuda al mirar la impotencia con la que vive.

Todo el material que se obtuvo se registró en audios, los cuales posteriormente fueron utilizados para guiar el proceso de transcripción, así como el de análisis acústico en herramientas tales como Praat y ELAN. Los resultados del tratamiento de la señal sonora servirán como base para desarrollar una línea de trabajo futuro que contempla el diseño y construcción de un corpus oral en amuzgo, así como de herramientas que permitan sacar provecho al material ahí registrado.

4.2. Procesamiento de la señal acústica

Tal como se describió en la Sección 3.2, las características fonológicas del amuzgo son muy complejas, de ahí que el tratamiento de la señal acústica se convierta en un elemento relevante para el estudio de la lengua. En este sentido, aunque el objetivo del trabajo es crear un corpus paralelo de tipo textual, no se descarta que los datos recogidos para la construcción del corpus se utilicen para sentar las bases de un nuevo corpus de tipo oral a nivel monolingüe.

Ahora bien, independiente a esta línea de trabajo futuro, para la construcción del corpus paralelo amuzgo-español fue necesario procesar la señal acústica con el fin de tener un componente que sirviera de guía para el proceso posterior de transcripción. Para ello, se hizo uso de herramientas auxiliares para el análisis del habla que garantizaran la fidelidad de los datos. En principio, se utilizó Praat para estar en posibilidades de realizar análisis de habla, etiquetado y segmentación, síntesis y manipulación de habla, así como cuestiones relacionadas con representaciones gráficas y de experimentación. Hay que reconocer que, en general, esta herramienta no soporta de forma eficiente cadenas de habla largas. Por esto se utilizó, fundamentalmente, para analizar entradas léxicas en las que existiera alguna duda respecto al tono de la palabra fonológica. En la Figura 1 se presenta un ejemplo de una cadena en la que se contrastan palabras con la misma conformación silábica, pero con contrastes tonales.

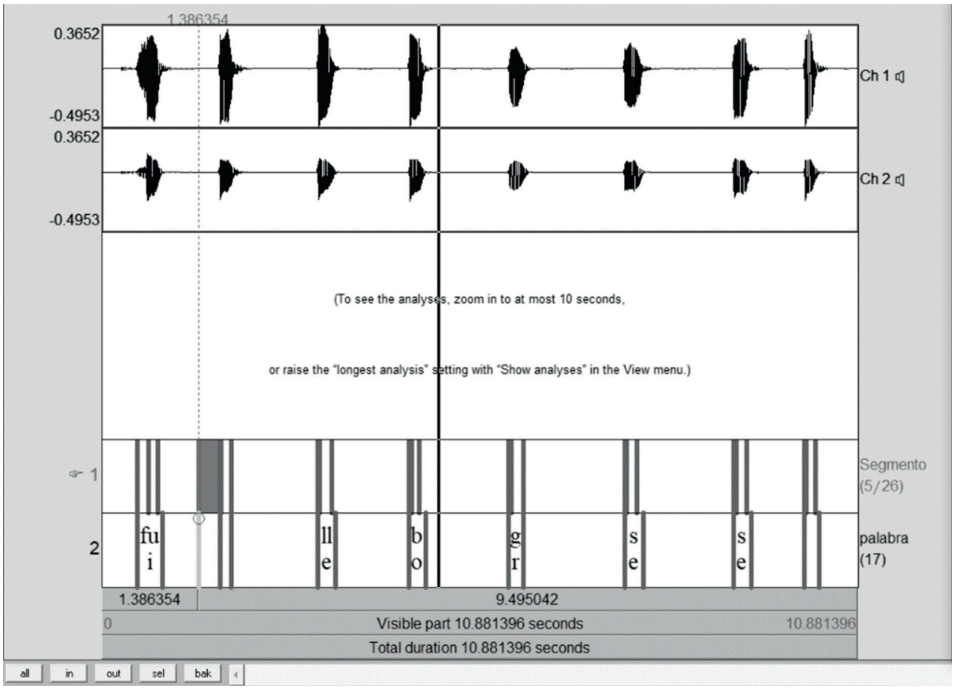


Figura 1. Diferenciación tonal del espectro acústico en amuzgo.

En contraparte, con ELAN se tuvo la posibilidad de analizar cadenas de habla más extensas; esto facilitará la incorporación de información multimodal a ese futuro corpus oral con el que se incrementarán las posibilidades de estudio de estos datos. Finalmente, como resultado del procesamiento acústico, se generó información relativa a la duración, el tiempo de emisión, el acento, así como a los formantes (pico de la intensidad o concentración de energía de una frecuencia) con los que, entre otras cuestiones, se distinguen las vocales.

4.3. Transcripción

El procedimiento específico de esta etapa se dio de la forma en la que se describe a continuación. En primer lugar, en lo que corresponde con la transcripción, se distinguieron los segmentos. A partir de este momento se planificó un cotejo entre lo hecho en las notas de campo y los espectrogramas que arroja Praat. En segundo lugar, se puntualizó en el registro de rasgos fonológicos específicos, tales como apertura vocálica, nasalidad y laringealización. En tercer lugar, como se muestra en el ejemplo 5, se hizo un primer acercamiento a los tonos de la lengua. Debido a que esta es una característica fundamental en amuzgo, se debía tener cuidado en una anotación precisa. Por tanto, se anticipaban y aceptaban modificaciones.

5. *1.* Twe'[3] nkwi[3]xue[12] m'an[3] kwi[3]ti'[12]tyo[3]ndye[35] ts'a[3] ti'[12],
   2. ts'ian[5] jndë[12], tyua'[12] ju'[35] sku'[35] ti'[12] k'a[35] ti'[12] jndë[12],
   3. Mo'[12] twe'[3] nkwi[3] xue[5] t-ja[3] ti'[12],
   4. të[1]ki[3]tsa[35] ti'[12] ts'ian[5],
   5. no'[1] ya[12] tje[35] ti'[12] tyua'[12]je[12],
   6. t'aa[3]'nna[3] sku'[12] ti'[12] k'on[3],
   7. No'[1] ma'[3]kje[35] tats'on[35] ti'[12] ng'e[3] të[1]non[35] tsan[34]nji[12],
   8. të[1]'yon[35] jon[35] sku'[12] ti'[12], xue'[35] jon[35] sku'[12] ti'[12]. No[1]jo[12],
   9. 'nni[12] 'nna[3] ntsa'[35] ti'[12]. Ma[3] të[1]kjo[35] ti'[12],
   10. t'eo[35] ti'[12],
   11. ndo'[1] hu[5]xjen[3] 'nein[3] tje[35] kwi[3] ki[5]tsian[35],
   12. tso'[35] ki[5]tsian[35]an':
   13. ndo'[1] u'[34] tyo[3]ndye[35]re[12], ndu[12] ma[3]tyo'[34],
   14. 'a[3]jo[1]je[12]. Ma[3]t'eo[5] ma[3]ng'e[3]
   15. hu[5]tsan[34] nji[12] tje[5] jon[35] b'a[5] no'[1] tsia[3]na[3] tja jon[35]
   16. të[1]'yon jon[35] sku'[5].

Para preparar las siguientes fases, cada uno de los textos resultantes se segmentó en cláusulas, las cuales se marcaron con número arábigos (en el ejemplo de arriba, la numeración aludida va del 1 al 16). Esta es una manera eficaz de organizar los textos y facilitar el análisis de glosado y traducción.

5. Construcción del corpus: fase paralela amuzgo-español

En esta sección se describen los procesos para la obtención de los datos en español con el propósito de conformar la estructura en paralelo amuzgo-español.

5.1. Glosado y traducción

Una vez que se realizaron los procesos previos, se tomó la decisión de trabajar con las transcripciones con el fin de expandir las posibilidades del corpus. Para ello, se realizó un proceso de glosado y de traducción de los datos. Las etapas relacionadas con la generación de glosas se esquematizan a continuación:

i Limpieza de las transcripciones para preparar el trabajo de glosado.
ii Empleo del sistema ortográfico más consistente (Tapia, 2006) y contraste con el propuesto por Hernández (2019).

iii Verificación de rasgos fonológicos relacionados con el acento para distinguir entre palabra fonológica (incluye los clíticos) y entrada léxica.
iv Marcación de los clíticos.
v Segmentación de los distintos tipos de frase: verbal, nominal, adverbial, etc.
vi Realización de la glosa de cada cláusula de acuerdo con las reglas de glosado de Leipzig (Comrie *et al.*, 2008).

El glosado, como se sabe, incluye tanto la segmentación como la identificación de la categoría gramatical (no funcional) de las unidades reconocidas. La última fase del análisis previo a la formalización de los datos en amuzgo fue la traducción de estos al español. Esta traducción se realizó en tres pasos. Primero, una interpretación general del texto en la lengua origen. Segundo, una alineación manual de las categorías gramaticales identificadas con sus respectivos significados (en este proceso se privilegió una traducción literal, manteniendo incluso el orden que se presentó en las oraciones en la lengua origen). Tercero, formalización de la traducción considerando el sentido oracional, la correspondencia entre categorías y la información producida mediante el proceso de glosado.

Este proceso de traducción fue realizado por un traductor humano, hablante nativo de amuzgo y español, con formación profesional en lingüística amerindia. Dadas estas características, se aseguró que la traducción fuera lo más fiel posible, tanto en términos de correspondencia lingüística como de función comunicativa, para poder realizar los procesos automatizados de alineación de segmentos. En la Figura 2 se ejemplifica el resultado del proceso general de traducción. En ella se observan algunos segmentos transcritos en la lengua origen (línea 1) con sus respectivas segmentaciones y glosas (líneas 2 y 3), así como una primera traducción basada en el tercer paso de la traducción (línea 4).

Twe'^{3} nkwi^{3}xue^{12} m'an^{3} kwi^{3}ti'^{12}tyo^{3}ndye^{35} ts'a^{3} ti'^{12}

T-we'^{3}	nkwi^{3}=xue^{12}	m'an^{3}	kwi^{3}	ti'^{12}=tyo^{3}ndye^{35}	ø-ts'a^{3}	ti'^{12}
CPL-haber.3SG	ART.INDEF.SG=día	HAB.estar.3SG	uno	compañero=zorro	PROG-hacer[3SG]	compañero

Hubo una vez un zorro

ts'ian^{5} jndë^{12}, tyua'^{12} ju'^{35} sku'^{35} ti'^{12} k'a^{35} ti'^{12} jndë^{12}.

ts'ian^{5}	jndë^{12}	tyua'^{12}	ø-ju'^{35}	sku'^{12}	ti'^{12}	ø-k'a^{35}	ti'^{12}	jndë^{12}
trabajo	monte	temprano	PROG-moler. 3SG	esposa[3SG]	compañero	HAB.ir[3SG]	compañero	monte

que trabajaba en el campo, temprano molía su esposa [y] él iba al monte.

Mo'^{12} twe'^{3} nkwi^{3} xue^{5} t-ja^{3} ti'^{12}

mo'^{12}	t-we'^{3}	nkwi^{3}	xue^{5}	t-ja^{3}	ti'^{12}
pero	CPL-haber.3SG	uno	día	CPL-ir[3SG]	compañero

Pero hubo un día

të^{1}ki^{3}tsa^{35} ti'^{12} ts'ian^{5}

të^{1}-ki^{3}-tsa^{35}	ti'^{12}	ts'ian^{5}
CPL-CAUS-hacer[3SG]	compañero	trabajo

que fue al trabajo

no'^{1} ya^{12} tje^{35} ti'^{12} tyua'^{12}je^{12}

no'^{1}	ya^{12}	t-je^{35}	ti'^{12}	tyua'^{12}=je^{12}
y	cuando	CPL-llegar[3SG]	compañero	temprano=INT

y cuando llegó más temprano a [su] casa

t'aa^{3}'nna^{3} sku'^{12} ti'^{12} k'on^{3}.

t'aa^{3}='nna^{3}	sku'^{12}	ti'^{12}	k'on^{3}
NEG=cosa	esposa[3SG]	compañero	HAB.estar[3SG]

ya no se encontraba su esposa.

Figura 2. Ejemplo de segmentos transcritos en amuzgo con sus respectivas glosas y traducción al español.

5.2. Alineación automática de segmentos

La siguiente fase de construcción consistió en realizar un proceso automático para alinear los textos transcritos en amuzgo con sus correspondientes traducciones al español. Esta fase es de suma importancia para poder concretar todo corpus que tenga como característica el ser paralelo. Para realizar este proceso se utilizó la herramienta de alineación que está implementada en el programa de Traducción Asistida por Computadora (TAC), OmegaT. Se decidió utilizar esta herramienta dado que el proceso de alineación se hace con base en el algoritmo de Gale-Church (1993), el cual ha sido utilizado en varios trabajos de lingüística computacional. Este algoritmo es independiente de la lengua, es decir, no es necesaria una gramática, en este caso del amuzgo, ni tampoco grandes volúmenes de datos para poder emparejar los segmentos. Pondera, en contraparte, la longitud de los segmentos para realizar la alineación con base en el supuesto de que las construcciones largas en la lengua origen deben corresponderse con construcciones de longitud similar en la lengua meta.

La alineación se hizo considerando los dos métodos de comparación de segmentos implementados en la herramienta: el método *parsewise* y el método *heapwise*. El primero privilegia el paralelismo sintáctico entre lenguas a partir de la alineación unitaria de segmentos, en tanto que el segundo privilegia una alineación global de los textos. Ambos métodos arrojaron resultados diferentes, cuya calidad fue evaluada con base en la información de las glosas y la traducción literal. En las figuras 3 y 4 se ejemplifican los resultados del proceso de alineación para un mismo fragmento. En la figura 3 se destaca el método *heapwise*, mientras que en la 4, el *parsewise*.

Alinear	
Original	Traducido
Twe' nkwixue m'an kwiti'tyondye ts'a ti', ts'ian jndë, tyua' ju' sku' ti' k'a ti' jndë, Mo' twe' nkwi xue t–ja ti', tëkitsa ti' ts'ian, no' ya tje ti' tyua'je, t'aa'nna sku' ti' k'on, No' ma'kje tats'on ti' ng'e tënon tsannji , të'yon jon sku' ti', xue' jon sku' ti'.	Hubo una vez un zorro que trabajaba en el campo, temprano molía su esposa [y] él iba al monte. Pero hubo un día que fue al trabajo y cuando llegó más temprano a [su] casa ya no se encontraba su esposa.
	De inmediato se enteró que pasó una persona malvada que se robó a su esposa.
Nojo, 'nni 'nna ntsa' ti'.	Entonces ¿qué haría el zorro?
Ma tëkjo ti', t'eo ti', ndo' huxjen 'nein tje kwikitsian, tso' kitsianan': ndo' u' tyondyere, ndu matyo', 'ajoje. Mat'eo mang'e hutsan nji tje jon b'a no' tsiana tja jon të'yon jon sku'.	Sólo se sentó y lloró, (y) en ese momento llegó un tigre, dijo el tigre: "y tú amigo zorro, ¿por qué lloras?" "Lloro porque la persona malvada llegó a mi casa y cuando se fue se llevó a mi esposa".
No'jo tso kitsianan': ti'ndyoto no' ti'nkon' tonyomatson' sku' nein jokionkwe, tso kitsianan' nnu ti'tyondyee'.	Entonces le dijo el tigre: "no llores y no te preocupes, tu esposa hoy la voy a ir a traer", dijo el tigre al zorro.
Jo jnon ti' na' t'ëo' ti'.	Entonces el zorro dejó de llorar.

Figura 3. Alineación mediante el método *heapwise*.

Alinear	
Original	Traducido
Twe' nkwixue m'an kwiti'tyondye ts'a ti', ts'ian jndë, tyua' ju' sku' ti' k'a ti' jndë, Mo' twe' nkwi xue t–ja ti', tëkitsa ti' ts'ian, no' ya tje ti' tyua'je, t'aa'nna sku' ti' k'on, No' ma'kje tats'on ti' ng'e tënon tsannji , të'yon jon sku' ti', xue' jon sku' ti'.	Hubo una vez un zorro que trabajaba en el campo, temprano molía su esposa [y] él iba al monte. Pero hubo un día que fue al trabajo y cuando llegó más temprano a [su] casa ya no se encontraba su esposa.
	De inmediato se enteró que pasó una persona malvada que se robó a su esposa.
Nojo, 'nni 'nna ntsa' ti'.	Entonces ¿qué haría el zorro?
Ma tëkjo ti', t'eo ti', ndo' huxjen 'nein tje kwikitsian, tso' kitsianan': ndo' u' tyondyere, ndu matyo', 'ajoje.	Sólo se sentó y lloró, (y) en ese momento llegó un tigre, dijo el tigre: "y tú amigo zorro, ¿por qué lloras?" "Lloro porque la persona malvada llegó a mi casa y cuando se fue se llevó a mi esposa".
Mat'eo mang'e hutsan nji tje jon b'a no' tsiana tja jon të'yon jon sku'.	
No'jo tso kitsianan': ti'ndyoto no' ti'nkon' tonyomatson' sku' nein jokionkwe, tso kitsianan' nnu ti'tyondyee'.	Entonces le dijo el tigre: "no llores y no te preocupes, tu esposa hoy la voy a ir a traer", dijo el tigre al zorro.
Jo jnon ti' na' t'ëo' ti'.	Entonces el zorro dejó de llorar.
Tja kitsian yana tje kitsianan' b'a tsan njii', yajo tsoo': nkiachjob'are, taa'nnat'a tsan njii' b'i jon.	Fue el tigre a la casa de la persona malvada, cuando llegó a la casa de la persona malvada, en ese momento dijo: "vengo a tu casa amigo", no contestó la persona malvada estaba enojada.
No' u're, 'nni 'nna mandue' manti' tsit'uantson're, janjokwaxye 'amayu u' njoyon sku ti' sku ti' tyondyee', t'io ti' m' a ti'.	"Y tú amigo, ¿qué cosa andas buscando?" "Disculpa amigo, vengo a preguntar si tú en verdad te robaste a la esposa del zorro, el amigo está llorando".
Ntjo jekindye ntjo jekindyenkwaxye nnö', tsannji b'i jon 'nni 'nna njaan, tyon tinkiandhe binaya'.	"Aquí no puedes preguntar nada", respondió enojada la persona malvada.

Figura 4. Alineación mediante el método *parsewise*.

5.3. Depuración manual y realineación de segmentos

Como se puede apreciar en las figuras anteriores, el resultado de alineación difiere bastante en los segmentos emparejados. Esta variación está en función del método de comparación. Así, cuando se hizo la alineación usando heapwise, los segmentos alineados no correspondían en buena medida con la información de la traducción. En cambio, cuando se hizo el proceso con el método parsewise, el resultado mejoró, por lo que se decidió utilizar este método para alinear los textos.

Cabe mencionar que, a pesar de la mejora que se observó con parsewise, la alineación de los segmentos aún distaba de ser totalmente paralela. Por tal motivo, se decidió hacer una depuración manual en la que se realinearon varios segmentos que no se correspondían.

Este proceso, si bien fue extenuante, en todo momento estuvo supeditado a la información que se obtuvo del proceso de glosado y de traducción. Ello, de alguna manera, garantiza que los segmentos emparejados exhiben de forma adecuada una correspondencia lingüística y comunicativa entre los datos en amuzgo y sus traducciones al español. Para finalizar esta sección, en la Figura 5 se evidencia el resultado de alineación después de realizar la depuración y su consecuente realineación.

Original	**Traducido**
Twe' nkwi xue m'an kwi ti' tyo ndye ts'a ti', ts'ian jndë, tyua' ju' sku' ti' k'a ti' jndë,	Hubo una vez un zorro que trabajaba en el campo, temprano molía su esposa [y] él iba al monte.
Mo' twe' nkwi xue t-ja ti', të ki tsa ti' ts'ian,	Pero hubo un día que fue al trabajo
no' ya tje ti' tyua' je,	y cuando llegó más temprano a [su] casa
t'aa 'nna sku ti' k'on,	ya no se encontraba su esposa.
No' ma' kje tats'on ti' ng'e të non tsan nji, të 'yon jon sku' ti', xue' jon sku' ti'.	De inmediato se enteró que pasó una persona malvada que se robó a su esposa.
No jo,	Entonces
'nni 'nna ntsa' ti'.	¿qué haría el zorro?
Ma të kjo ti',	Sólo se sentó
t'eo ti',	y lloró,
ndo' hu xjen 'nein tje kwi kit sian, tso' ki tsian an':	(y) en ese momento llegó un tigre, dijo el tigre:
ndo' u' tyo ndye re, ndu ma tyo', 'a jo je.	"y tú, amigo zorro, ¿por qué lloras?" "Lloro porque la persona malvada llegó a mi casa y cuando se fue se llevó a mi esposa".

Figura 5. Resultado de segmentos emparejados después de la depuración y la realineación.

5.4. Implementación y liberación del corpus

Una vez que se concluyó el proceso total de alineación, se buscó cómo implementar el material generado en un recurso que permitiera la consulta de los datos de una manera eficiente. Para ello, se utilizó la plataforma web GECO[8], la cual permite hacer una implementación de los datos en una interfaz sencilla para el usuario. Además de ello, ofrece algunas herramientas para explotar el contenido de los corpus, por ejemplo, la búsqueda de concordancias. Para ilustrar el resultado de la implementación en esta plataforma, en la Figura 6 se muestra una captura de pantalla del corpus en la que se focaliza la búsqueda de la palabra *sku* en amuzgo (base semántico-léxica de "esposa") y los contextos en los cuales aparece en ambas lenguas.

8 http://www.geco.unam.mx/.

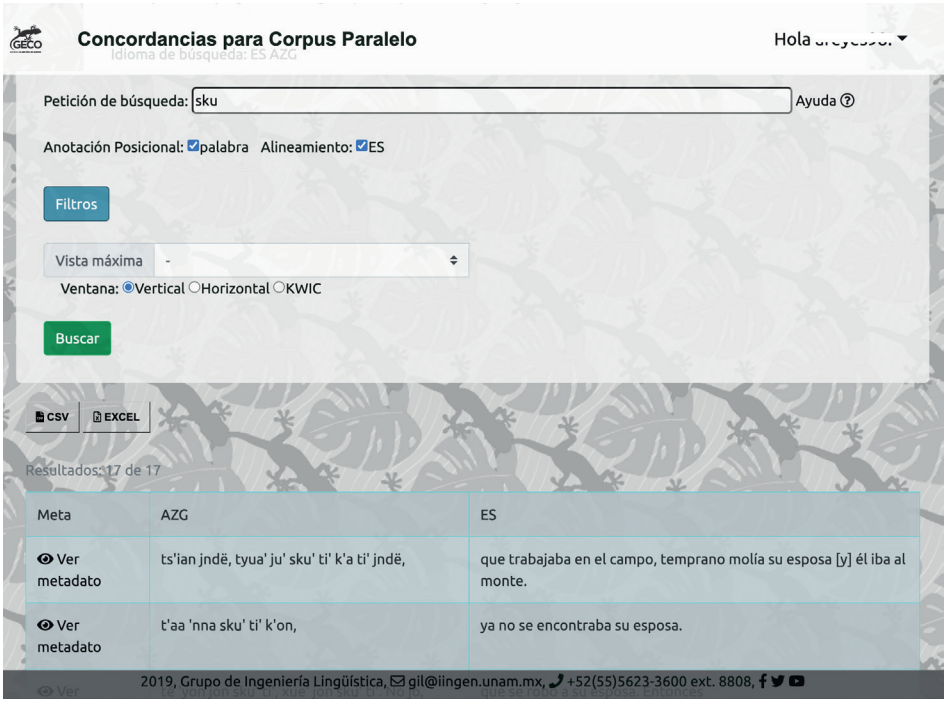


Figura 6. Concordancias amuzgo-español de la palabra *sku* (esposa) en el corpus.

Si bien en este momento la implementación del corpus aún no ha concluido, es importante destacar que el resultado de todo este conjunto de procesos es una primera versión que permite explotar, aunque sea de forma mínima, los datos paralelos del corpus. Es cierto que hay información pendiente de procesar e, incluso, de implementar (por ejemplo, en esta versión preliminar no se aportan estadísticas acerca de la relación types/tokens del mismo debido a que es poco representativo hablar en estos términos dadas las características morfológicas del amuzgo); no obstante, es importante recalcar que el corpus cuenta al momento con poco más de una hora de grabaciones procesadas conforme a las etapas descritas previamente. En este sentido, el corpus se está constituyendo con información que rebasa el espectro oral de los datos, es decir, se está incorporando información muy valiosa en las glosas y en las traducciones, la cual, una vez liberada la versión final del corpus, permitirá complementar y expandir la utilidad de este para estudiar y generar nuevo conocimiento, así como herramientas y recursos para esta lengua.

6. Conclusiones

En este artículo se ha descrito un trabajo para constituir un corpus paralelo amuzgo-español. Se ha enfatizado la problemática que implica la creación de recursos en lenguas indígenas. En específico, para lenguas cuya ausencia de datos dificulta, incluso, su descripción lingüística. De igual manera, se ha resaltado el trabajo realizado para la obtención de muestras reales de la lengua mediante trabajo de campo. Los datos aquí presentados corresponden a una primera fase de grabaciones, las cuales han sido procesadas considerando diferentes niveles que permitan generar un corpus de calidad: transcripción, procesamiento de la señal acústica y transcripción fonética; asimismo, se ha trabajado con las transcripciones para realizar el proceso de glosado y de traducción al español. Esta información, además de ser relevante para fines lingüísticos, puede ser provechosa para modelar sistemas sustentados en las características propias de la lengua. Cabe mencionar, por otra parte, que se espera que en próximas fases se presenten más transcripciones de las grabaciones, así como que el número de entrevistas aumente en el corto y mediano plazo, logrando así un corpus más amplio.

Ahora bien, los resultados que se han obtenido a la fecha permiten hacer una proyección de la utilidad del corpus por demás interesante. A saber, más de una hora de grabación de muestras reales de habla en amuzgo, es decir, a diferencia de algunos corpus que parten de documentos que reflejan muy poco el habla coloquial o son traducciones de documentos oficiales o religiosos, este corpus representará un habla lo más natural posible, tal como se da en la comunidad. Asimismo, este tipo de contenido permitirá, en el ámbito de las TLH, contar con un recurso con el cual se pueda experimentar, por ejemplo, con modelos de reconocimiento de voz, tomando en cuenta las características tonales de la lengua, así como con herramientas de traducción automática que consideren los rasgos tipológicos del amuzgo para segmentar las oraciones y alinearlas correctamente con los segmentos de la lengua de llegada. Aunado a lo anterior, es indudable que este tipo de recursos puede coadyuvar a los diferentes esfuerzos que desde diversos ámbitos intentan disminuir la brecha tecnológica entre comunidades y que, de forma ideal, como señala Crystal (2000), pueden aportar para evitar la potencial desaparición (o muerte) de lenguas.

Para concluir, se destaca una serie de líneas de trabajo que permitirán formalizar el trabajo realizado hasta ahora. La primera y más obvia es la consecución de nuevas muestras orales que permitan expandir la cantidad de datos que integrarán el corpus. Una segunda línea es la exploración de herramientas que permitan procesar la señal acústica de manera más rápida, de forma que el proceso de transcripción se vuelva, en cierto punto, más inmediato. Finalmente, se contempla una línea de trabajo más social en la que el corpus, en

tanto herramienta que refleja una forma de conceptualizar y verbalizar el mundo, permita poner de manifiesto las necesidades y oportunidades sociales de las comunidades indígenas, por ejemplo, en escenarios de interpretación social, médica o jurídica.

Referencias

Acosta, O., & Aguilar, C. (2020). A Critical Review of the Current State of Natural Language Processing in Mexico and Chile. In F. Pinarbaşi & M. Taşkiran (Eds.), *Natural Language Processing for Global and Local Business* (pp. 365-389). IGI Global.

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S. & Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds), *Proceedings of the 11th International Conference on Language Resources and Evaluation* (pp. 3356-3365) European Language Resources Association (ELRA).

Apóstol, J. (2014). *Clases flexivas verbales en el amuzgo de Xochistlahuaca (Guerrero)* [Tesis de Maestría, Centro de Investigaciones y Estudios Superiores en Antropología Social].

Arévalo, J. (2015). El problema de la brecha tecnológica: un asunto de cultura. *Revista Sinapsis*, *7*(7), 43-57.

ASTLM. (2018). Análisis del sector de las Tecnologías del lenguaje en México. *Plan del impulso de las tecnologías del lenguaje.* Gobierno de España.

Buck, M. (2000). *Gramática del amuzgo de San Pedro Amuzgos*. Instituto Lingüístico de Verano.

Buck, M. (2018). *Gramática del amuzgo de Xochistlahuaca*. Instituto Lingüístico de Verano.

Campbell, L. (1997). *American Indian languages: the historical linguistics of Native America*. Oxford University Press.

Castellanos, A., Estrada, E. y Domínguez, W. (2019). Implementación de algoritmos de procesamiento de lenguaje natural para la evaluación de la pronunciación efectiva en el aprendizaje de lenguas indígenas. *Revista Electrónica de Investigación e Innovación Educativa-REIIE*, *4*(2), 16-24.

Comrie, B., Haspelmath, M., Bickel, B. & Max Planck Institute for Evolutional Anthropology. (2008). *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses*. Max Planck Institute for Evolutionary Anthropology.

Crystal, D. (2000). *Language death*. Cambridge University Press.

Cruz, H. & Waring, J. (2019). Deploying Technology to Save Endangered Languages. *arXiv*.

Gale, W. & Church, K. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75-102.

García, H., Hernández, N. y Mora, A. (en dictamen). Posesión y otras relaciones semánticas en Amuzgo de San Pedro Amuzgos (otomangue). En Z. Estrada y M. Peregrina (Eds.), *Dependencias simétricas y asimétricas: Dominios semánticos y motivaciones*. Universidad de Sonora.

Gutiérrez, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. En D. Inkpen, S. Muresan, S. Lahiri, K. Mazidi, & A. Zhila (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 154-160). Association for Computational Linguistics.

Gutiérrez, X., Sierra, G., & Hernández, I. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. En N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard,

J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4210-4214). European Language Resources Association.

Hernández, N., Mora, A. y García, H. (2017). Estructura de la frase nominal posesiva en amuzgo (otomangue). UniverSOS. *Revista de Lenguas Indígenas y Universos Culturales*, *14*, 63-82.

Hernández, N. (2019). *El sistema tonal en el amuzgo de San Pedro Amuzgos: Interacción entre el tono de la base nominal y los clíticos* [Tesis de Maestría en Lingüística Indoamericana, Centro de Investigaciones y Estudios Superiores en Antropología Social].

INALI, (2008). Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas. En *Diario Oficial de la Federación*, 14 de enero de 2008.

INEGI. (2015). *Encuesta intercensal 2015. Lenguas indígenas y hablantes de 3 años y más*. http://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm.

Mager, M., Barrón, C. y Meza, I. (2016). Traductor estadístico wixarika-español usando descomposición morfológica. *COMTEL*, *6*, 63-68.

Mager, M., Gutiérrez, X., Sierra, G., & Meza, I. (2018). Challenges of language technologies for the Americas indigenous languages. In E. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27*th *International Conference on Computational Linguistics* (pp. 55-69). Association for Computational Linguistics.

Manning. C. & H. Schütze. (1999). *Foundations of statistical natural language processing*. MIT Press.

Midrigan, L., Boyd, V., Victoria, L., Sánchez, D., Malancea, D., Midrigan, D., & Corina, D. (2020). Resources in Underrepresented Languages: Building a Representative Romanian Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association: 3291-3296.

Palancar, E. y Feist, T. (2015). Agreeing with subjects in number: The rare Split of Amuzgo verbal inflection. *Linguistic Typology*, *93*(3), 337-383.

Prinsloo, D. (2015). Corpus-based Lexicography for Lesser-resourced Languages - Maximizing the Limited Corpus. *Lexikos, 25*(1), 285-300.

Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the fifth International Conference on Language Resources and Evaluation* (pp. 1799–1802). ELRA.

Smith, T. y Tapia, F. (2002). Amuzgo como lengua activa. En P. Levy (Ed.), *Del cora al maya yucateco. Estudios lingüísticos sobre algunas lenguas indígenas mexicanas* (pp. 81-129). Universidad Nacional Autónoma de México.

Tapia, F. (2006). *Diccionario amuzgo-español. El amuzgo de San Pedro Amuzgos*. CIESAS.

Vinogradov, I. (2016). Linguistic corpora of understudied languages: do they make sense? *Káñina*, *40*(1), 116-130.

Chapter VIII

Methodological bases: the construction of a corpus for the detection of deception and credibility assessment[1]

Bases metodológicas: la construcción de un corpus para la detección de mentiras y la evaluación de la credibilidad

Pedro Eduardo Hernández Fuentes
Universidad Nacional Autónoma de México –México

Abstract: Meta-analytic approaches reveal that, to identify lies or evaluate the credibility of a testimony, it is more reliable to perform a discursive or verbal material analysis in contrast to one based on non-verbal behavior. Hence, different research has been developed to make scientific contributions in this regard. These efforts make imperative the need to focus on the construction of a linguistic corpus that enables the study of the subject without ignoring the contributions made from cognitive psychology. Therefore, the methodological proposal for the construction of a corpus on the subject will be developed in this text. This is the result of a transdisciplinary work between linguistics and psychology integrated for a total of 54 cognitive interviews using a double-blind procedure.

Resumen: Los acercamientos metaanalíticos revelan que la información verbal es un indicador confiable para identificar mentiras o evaluar la credibilidad de un tes-

1 Translation from Spanish language by Leon Jacob Ortega Islas.

timonio. De aquí que actualmente se han desarrollado diversas investigaciones para realizar aportaciones científicas al respecto. Estos esfuerzos vuelven imperativa la necesidad de enfocarse en la construcción de un corpus lingüístico que posibilite el estudio del tema sin relegar las aportaciones realizadas desde la psicología (cognitiva). Por ello, en este texto, se desarrollará la propuesta metodológica para la construcción de un corpus en el tema. Ésta es el resultado de un trabajo transdisciplinario entre la lingüística y la psicología que consiste en la realización de 54 entrevistas cognitivas con el método de doble ciego.

1. Introduction

The study of detection of deception and credibility assessment has been of interest to many specialists and has been approached from different disciplines. Although scientific tools have been provided for its study, there is still a widespread false belief that there are key determinants, universal body signals or physiological indicators that are irrefutable proof that an individual is lying. Systematic review to analyze research results quantitatively (meta-analysis) reveals that most of the indicators that researchers typically examine in detection of deception are not related to deception at all (Vrij *et al.*, 2010).

Meta-analytic research also reveal that verbal information is a more reliable indicator to identify deception or assess the credibility of a testimony (DePaulo *et al.*, 2003; Vrij, 2018). Hence, research from forensic linguistics, sociolinguistics, psycholinguistics and mostly, cognitive psychology have currently been developed to make scientific contributions in this regard. These efforts make the need to focus on the construction of a linguistic corpus that allows the study of detection of deception and credibility assessment imperative.

Therefore, this chapter will develop the methodological proposal that allows the creation of a linguistic corpus to identify some characteristic features of the evaluation of truthfulness and lie detection in discourse. This project is the result of an inter- and transdisciplinary work between linguistics and psychology. The project proposal developed at the Language and Cognition Laboratory of the Cognitive Sciences Research Center (UAEM) will be presented, emphasizing the methodology followed for the construction of the sample; an in-depth explanation of the method and general description for the construction of the corpus is presented: type of study, type of participants, data collection procedure and ethical considerations We believe that, before making evaluations on truthfulness or falsehood in discourse, it would be necessary to explore theoretically and meth-

odologically the construction of the discursive corpus in order to begin to explore the still little-known map of deception and truthfulness. Beginning to establish methodological rigor in the construction of this type of samples is not an easy task, although it is necessary for the future experimental or quasi-experimental approach to a subject for which there are many questions and few answers.

We also aspire to introduce to the academic context a subject that has been little addressed in the scientific field, since there have not been enough studies that consider linguistic theory to address this phenomenon: most of the research has been conducted from the perspective of cognitive psychology. There is also a deficiency in the little research conducted on the Spanish language; although some recent proposals consider this language as a field of study there are still few efforts (Hwang *et al.*, 2016; Vrij *et al.*, 2020).

In short, although research has emphasized the preponderance of the analysis of verbal content in contrast to that of nonverbal behavior, there is a gap in this regard and not enough value has been given to the construction of the corpus so that, in the future, the main linguistic indicators that differentiate between a discourse that intends to deceive another and one that does not can be studied. This project will contribute to fill this gap.

2. Detection of Deception and Credibility Assessment

The subject of this paper has a long tradition within the scientific and non-scientific field. The approaches to this subject have been made mainly from philosophy and psychology, although there are also contributions from anthropology, behavioral economics, sociology, and linguistics, to mention a few examples. Possibly, the first major contribution that these works have given is the definition of the terms *lie* and *deception*, which have been used as synonyms, and are understood as an attempt to generate in someone else, from verbal or nonverbal means, a belief that the communicator assumes as false (Vrij, 2008; Masip, 2004) (§3.2). Other contributions concern the answer to questions such as: why do we lie or what are the reasons for lying? (Vrij, 2001, 2008), what are the characteristics of a good liar? (Vrij, 2008), what skills do people have to lie? (Salekin *et al.*, 2008), how often do we lie? (Feldman *et al.*, 2002), what are the basics of lie detection? (Vrij, 2008), and why are we bad lie detectors? (Vrij *et al.*, 2010).

The phenomenon acquired greater visibility from its association with the study of nonverbal behavior, whose most considered channels have been facial expression, physiology, paralanguage and oculesics. Although the study of nonverbal behavior also includes other channels, namely gestures, postures, orientation and movement, proxemics, haptics, and appearance (López *et al.*, 2016), these have been less regarded and studied. Within the study

of facial expression, one of the greatest proponents has been Paul Ekman, who has argued in various publications (Ekman, 2015, 2017; Ekman and Friesen, 1969, 1974; Ekman and O'Sullivan, 1991) that facial expressions of emotions are universal and have a biological, evolutionary, and adaptive origin, as Darwin (1872/2009) stated. Although Darwin's (1872/2009) and, therefore, Ekman's proposals were initially questioned, his findings have now been supported by more than a hundred research studies and different specialists; for example, Reissland *et al*. (2011) conducted a study on facial development based on 4-D ultrasound visualization of fetal facial movements.

These investigations have led to state that one of the most reliable ways to detect deception is the study of *microexpressions*, which are rapid facial movements lasting less than one-fifth of a second, which are important because they convey important information about what a subject is truly feeling or experiencing emotionally and is trying to hide (Ekman, 2017). Based on this, it has been suggested the idea that observable microexpressions on the face are more reliable indications of deception than other channels. This, moreover, is supported by the Filtering Hypothesis, which argues that, when a person lies, he or she experiences emotions that he or she tries to hide because they could reveal the truth; however, these are leaked through the subject's face for a brief moment (Ekman and Friesen, 1969).

However, the analysis of microexpressions as indicators of deception is still under discussion, since deception can generate positive or negative emotions, or even these may not be present and, therefore, the analysis of them is not the best way to determine when a person is hiding the truth (Burgoon, 2018; Vrij *et al*., 2010). In addition, it remains to delve into the relevance or not of other indicators of nonverbal behavior that have been less studied, such as those that DePaulo *et al*. (2003) registered: the movements of arms, hands, fingers, fingers, legs and feet and the use of illustrators. Therefore, Vrij *et al*. (2010) reviewed which is the most successful way for detection of deception when a subject tries to detect it without the help of technology: nonverbal behavioral analysis or discursive analysis, concluding that a promising way was discursive analysis.

Following the above, Vrij (2018) presented a literature review on the keys in detection of deception and pointed out that the projects that study the differences at the discursive level are the ones that are currently predominant, as there is scientific evidence on their level of reliability. This is also confirmed by the meta-analytical study of DePaulo *et al*. (2003), in which, from the review of the importance of 158 behaviors (verbal and nonverbal), it was concluded that the analysis of the verbal in contrast to the nonverbal is more relevant. So, is detection of deception a problem of linguistics?

2.1 Is detection of deception a problem of linguistics?

The fact that the subject has been widely approached from psychology does not imply that it is not a problem of linguistics. From this area, some research has been carried out, although it is not very abundant, since the study of detection of deception as a linguistic phenomenon has been relegated; hence there is a need to offer more specific contributions from this discipline that give a linguistic description of the phenomenon. It is likely that the limited existence of linguistic studies of lying is the result of the methodological difficulties involved in the design of experiments and the analysis of the information obtained (Infante, 2015). The still low number of contributions made from this area regarding the subject and some peripheral subjects could be listed more and more frequently; however, increasing interest in the construction of a corpus other than English – the language in which the experiments and samples have been mostly designed – may allow us to generate a more assertive approach to the matter.

Among the linguistic contributions, those developed from forensic linguistics stand out, for example, Picornell (2013) has studied the detection of deception in written witness statements and has proposed ways to look for signs of deception from the narrative characteristics of the witnesses. The author has criticized that one of the shortcomings that exist in several of the research studies is that they are conducted with university students because they are the closest participants, although they do not reflect the reality. For this reason, in the present study, the two variables to be controlled are not related to educational level, but to age and sex (§3.3). Also noteworthy are the contributions of Fitzpatrick (2009), who attempted to test the accuracy of some linguistic cues linked to deception.

From a more technological perspective, a number of tools have been developed, for example, the Linguistic Inquiry and Word Count (Pennebaker *et al.*, 2001), used to automate in a simple way the lexical analysis of deceptive text; the Voice Stress Analyzer (NITV Federal Services, 2020), whose hypothesis is that vocal stress indicators reveal deception; and the CSC Deceptive Speech (2013), a corpus developed to distinguish deceptive speech from non-deceptive speech based on machine learning techniques on features extracted from the corpus. These endeavors, which aim to identify and quantify linguistic indicators of deception, have generated several computational programs from different research areas and laboratories in the last fifteen years with the direct or indirect purpose of achieving a better identification of lies: Agent99 Analyzer, General Architecture for Text Engineering (GATE), iSkim or CueCal, Coh-Metrix, Automated Deception Analysis Machine (ADAM) (Hauch *et al.*, 2015).

3. Methodological proposal

The creation of this corpus responds to the interest and the need to create resources that generate research related to truthfulness and deception in discourse, since, as stated in the introduction, most of the current research indicates that the analysis of verbal content can provide more clues in the detection of deception and the evaluation of credibility. Thus, beginning to defragment and study how Spanish speakers lie in quasi-experimental conditions is a timely, though limited, approach for resource generation and future research purposes in this field.

The idea that there is no single totally reliable signal for deception detection is the most useful one because of the very difficulties of lie detection. In this sense, the set of several verbal and non-verbal indicators is the most accurate way to deal with this phenomenon; although the focus of this work is, in principle, linguistic, by obtaining recorded audiovisual material (§3.3), other types of approaches will be possible in the future. It should also be noted that, as mentioned, most of the research reviewed seeks to find patterns that help to determine whether there are indicators of deception, leaving aside the evaluation of truthfulness in discourse. This is also intended to be controlled in the present research.

Thus, the creation of this sample seeks to create a database with a general criterion specific to the Laboratory of Language and Cognition that: 1) favors projects related to the topic; 2) speeds up the necessary methodological processes of a research related to the topic; 3) allows the approach of inter- and transdisciplinary research from the same material whose methodological decisions have a justification; 4) allows finding characteristic patterns of truthful and fallacious discourse of a specific society and with a particular topic. In principle, the scope of the set of texts is limited to the collaborators of the Laboratory, i.e., only members will be able to consult it, since there is no platform on which it can be disseminated. Nevertheless, in the future, a greater transcendence is intended.

The first phase of this research involved the design of the interview and the selection of the participants (§3.3 and §3.4); the second phase involved sending more specific information through the informed consent form (§3.3 and Appendix 2. CI); the third phase involved conducting the cognitive interview divided into two sections (Appendix 3. GE): implementation of the double-blind method and conducting the interview; finally, the fourth phase involved the transcription and basic labeling that will allow for future analyses (Appendix 4. CT).

Upon completion, fifty-four narratives of experience were obtained from twenty-seven people who had some experience of the September 19, 2017, earthquake that occurred in Mexico; each participant provided one truthful narrative (twenty-seven total) and one

fallacious narrative (twenty-seven total). The testimonies were divided into three different groups (Table 2) to be able to perform comparative analyses.

3.1 Cognitive interviewing as a method for eliciting deceptive discourse

The lack of evidence that proves the usefulness of non-verbal parameters in lie detection and credibility assessment has generated the development of research that bets on the use of cognitive strategies. This has led to remarkable differences between those who express (verbally) a truth or a lie (Vrij, 2018) and, therefore, has prompted the design of experiments that assess these distinctions: telling a story backwards rather than in chronological order (Vrij *et al*., 2012; Vrij *et al*., 2008), looking at the direction of gaze (Vrij *et al*., 2010), asking unexpected questions of the participant (Lancaster *et al*., 2013), asking the subject to perform a secondary task (drawing, for example) during the interview (Lancaster *et al*., 2013), and providing a greater number of possible details in a story (Leal *et al*., 2015).

Throughout all of these approaches, the role of the interviewer is critical. For example, one could highlight the difference between the cognitive interview model and the Reid technique model of interviewing and interrogation, which is still used despite its proven ineffectiveness. So, it is important for the interviewer to take an active role and ask questions that generate distinctive reactions between the person who is lying and the person who is telling the truth (Masip and Herrero, 2015). This should be supported by protocols based on solid theoretical models, cognitively based, and supported by research, such as the Activation-Decision-Construction Model (ADCM) proposed by Walczyk and those previously discussed.

It is important to consider the limitations noted about the cognitive models currently developed, since specifying the reasons why lying is cognitively more complex is not the same as elaborating or contrasting models that specify the cognitive processes responsible for the distinctions between lying and telling the truth that clarify answers to questions such as what cognitive processes are activated when a person lies? (Blandón-Gitlin *et al*., 2017).

As part of the development of research studies that focus on the use of cognitive strategies, we can find the cognitive interview, designed by Geiselman *et al.* (1984) and Fisher and Geiselman (1992) with the purpose of obtaining quality information from the interviewee; in addition to developing an alternative interview method to the existing ones, focused on the mental processes of the witnesses instead of the events that occurred (Fisher and Geiselman, 2019). In its first version, channeled toward criminal investigation, the proposal contained four basic techniques: 1) context reinstatement, 2) telling everything, 3) change of perspective, and 4) change of order. In the second version of the interview (Fisher and Geiselman, 1992), called the enhanced cognitive interview, social and commu-

nicative factors were included, which were intended to improve the social interaction between the interviewer and the interviewee, improve the interviewee's memory and other cognitive processes, and achieve effective communication:

Table 1. Cognitive interview techniques (Fisher & Geiselman, 2019).

No.	Technique	Description	Improved psychological process
1	Rapport	It aims to create a good emotional climate and develop a good relationship between the interviewee and the interviewer.	Social interaction
2	Active participation of the interviewee	The interviewee actively generates information throughout the interview: he/she does not only answer the interviewer's questions.	Social interaction
3	Report everything	The interviewee includes all the memories that come to mind, as he/she is asked to report all the facts, whether he/she considers them important or not.	Memory and communication
4	Reset the context	The interview aims to re-establish the context of the original experience.	Memory
5	Describe in detail	Seeks a detailed account of events from the interviewee. It can sometimes be initiated from a model statement (Leal, Vrij, Warmelink, Vernham, & Fisher, 2015).	Communication
6	Close your eyes	The interviewee is asked to close his/her eyes. This instruction should be done after the relationship between the interviewee and the interviewer has been developed.	Cognition
7	No interruptions	The interviewee should not be interrupted during the interview.	Social interaction and cognition
8	Do not guess	It is made clear to the respondent that it is okay to say "I don't know" and not to guess the answer.	Cognition
9	Open questions	It calls for mainly open-ended questions; closed-ended questions will be asked only as a follow-up.	Social interaction and cognition
10	Multiple recovery	An attempt is made to encourage the interviewee to search through his or her memory more than once.	Memory
11	Varied recovery	It is intended to encourage the participant to search through his or her memory in different ways.	Memory
12	Questions compatible with the interviewee	It calls for questions that are compatible with the respondent's current accessibility.	Memory
13	Avoid suggesting questions	Avoid asking questions that suggest a specific answer.	Memory
14	Compatible output code	It allows respondents to produce their knowledge in the same form in which it is stored (often non-verbal).	Communication

Over the years, modifications have been made to the cognitive interview and a consensus has been reached on its effectiveness in contrast to other types of interviews such as structured interviews (Köhnken *et al*., 1999). It has also been successful in increasing the amount of correct information recalled by the interviewee (Fisher *et al*., 2011), it has proven to be effective in different contexts and in both criminal and non-criminal investigations (Fisher and Geiselman, 2019). Likewise, it has been widely used in the field of lie detection. Therefore, in this paper, we used this type of interview to obtain the required information.

3.2. Type of study

The type of study of this research is non-probabilistic quasi-experimental in which a corpus was obtained by convenience from the manipulation of two variables of interest: age and sex. For this, in each interview, a pre-post evaluation was conducted from which the baseline of the participants will be obtained according to the evaluation between the narration of the true story (experience of the earthquake of September 19, 2017, in Mexico) and the false version of the same story; both were conducted randomly, that is, in some cases it was decided that the interviewee first lied and then told the truth and vice versa to observe whether this has an effect on the discourse. It should be added that this work does not aim to evaluate the memory or recollection of the participants, but rather their intention to lie or tell the truth; this justifies the decision to use an event that occurred well in advance (see definition of lying, §2). Likewise, the participants were intended to be their own control.

The scientific method used to prevent the results of future research from being influenced by observer bias was the so-called double-blind method: in the collection of the corpus, the participants were unaware of the research topic (Appendix 1. D) while the interviewer and analyst are still unaware of the type of discourse they formulated first, true or false, as the information was determined by an instructor outside the interviewer.

3.3 Participants and interview

A non-probabilistic convenience sampling was carried out. To this end, 27 volunteers (Table 2) were invited to participate using a poster published on social networks, with the following requirements or inclusion criteria: internet access, time availability of approximately one hour, being of one of the requested ages, agreeing to sign an informed consent form (Appendix 2. CI) with the request to videotape their participation for strictly academic purposes, to have a camera and audio in the device to be connected and to have the video call program to conduct the meeting via this means. The exclusion criteria, in addition to non-compliance with any of the above, were neurological problems or language

pathology. Since these were self-declarations, the reliability of this information could not be controlled. Based on these requirements, men and women were selected from each of the three groups shown in Table 2. It should be noted that the initial intention was to obtain 30 volunteers, but only the number indicated was achieved and it was necessary to exclude some of the participants. In the future, we intend to complete the number of participants in order to have a fully gender-balanced sample.

Table 2. Participants.

Group	Age	Sex	No. of participants
1	20-25	5 women and 5 men	10
2	35-40	5 women and 5 men	10
3	50-55	5 women and 2 men	7

The project manager determined the eligibility of the participants according to the inclusion and exclusion criteria indicated, based on the answers provided by the volunteer. None of the three groups included vulnerable participants.

A virtual Zoom session was organized for each of the volunteers to conduct the interview. The first face-to-face (virtual) approach was by a person other than the interviewer, known as the "instructor", to give the participant the instructions developed in the interview guide, the instructor's guide (Appendix 3. GE). Once his/her participation was completed, the instructor informed the interviewer that he/she had finished so that he/she could enter the session via Zoom and continue with the meeting as detailed in the guide.

The two participant narrations (one true and one false) were both recorded on two different recordings. Each was labeled as follows: CMC0001vA. This label is comprised of basic information to systematize the use of the material, consisting of: 1) the letters *CMC* refers to the name of the corpus "Corpus mentiras y credibilidad"; 2) the sequence of four numbers corresponds to the number of the video and changes according to the number of testimony; 3) the letter *v* corresponds to the clarification that it is a video; 4) the capital letter corresponds to the letter assigned to each one of the participants.

Once the material was obtained, a Word transcription was made with the corresponding criteria (Appendix 4. CT). These files were labeled CMC0001tA, which is the same as the previous label, but with a change in the lowercase letter, which implies that it is a transcription. The transcription process involved two participants: the transcriber and the reviewer.

3.4 Data collection procedure

The participation of the volunteers was videotaped with the Zoom program. The instructor and the interviewer used the interview guide to help them (Appendix 3. GE). The cognitive interview proposals (§3.1) were considered in the elaboration of these materials; they were also reviewed and commented on by three experts.

As for the transcription criteria (Appendix 4. CT), great attention was paid to ensure that the use of marks was the minimum necessary to achieve the purposes of this project, while remaining rigorous. Thus, most of the elements linked to phonetic-phonological characteristics were omitted. Likewise, the participants were given the "Informed Consent" (Appendix 2. CI). All the forms are attached as annexes.

3.5 Ethical considerations

Regarding ethical considerations, this research had minimal risk for the participants, since only documentary research techniques were used (cognitive interview) in which sensitive aspects of behavior were not addressed. The research protocol was sent to the Centro de Investigación Transdisciplinar en Psicología, Universidad Autónoma de Morelos, on September 4, 2020, and was approved on November 30, 2020.

3.6 Current track and future projections

As mentioned at the beginning, this work is mainly of a methodological nature, as it is considered that, since this is a subject that has been little addressed in linguistic and corpus studies, the first approach to follow is to make a proposal that allows us to obtain the truthful and fallacious discourse. In spite of this, some of the results obtained have to do with the type of words present in the total narration, the number of total words, the lexical variety, the approximate duration of the narration and the number of words per minute (Tables 3 and 4).

Table 3. Group 1. Women aged 20 to 25 years.

	Type	Token	Lexical variety	Approximate duration	Words per minute
CMC0004-B	448	1826	4.07	11	166
CMC0007-D	303	1032	3.40	9	114.66
CMC0009-E	393	1452	3.69	8	181.50
CMC0014-G	715	3803	5.31	22	172.86
CMC0017-I	479	1947	4.06	11	177

Table 4. Group 2. Women aged 20 to 25 years.

	Type	Token	Lexical variety	Approximate duration	Words per minute
CMC0003-B	385	1526	3.96	10	152.60
CMC0008-D	533	2515	4.71	19	132.36
CMC0010-E	387	1255	3.24	7	179.28
CMC0013-G	525	2393	4.55	14	170.92
CMC0018-I	466	2072	4.44	10	207.20

The above tables show that we started from a general approach to proceed to a particular one in which potential linguistic indexes are codified to establish their quality. Some of them are part of the psychological, criminological and, to a lesser extent, linguistic literature that have been constantly mentioned and are currently considered as warning flags: full pauses, negation, adverbs, verb tenses, pronouns, number of syllables, number of sentences, number of big words, number of syllables per word, number of short sentences, number of long sentences, average number of words per sentence, conjunctions, simple sentences and adjectives (Burgoon *et al*., 2003; Fitzpatrick and Bachenko, 2009; Picornell, 2013; Villar and Castillo, 2016). Currently, the coding of filled pauses, pronouns, adverbs, reported memory, in addition to those previously mentioned, is part of the tasks of the coordinator of this research and the first results are expected to be available in March 2022.

4. Conclusions and discussion

Although research has emphasized the preponderance of the analysis of verbal content in contrast to that of nonverbal behavior, there is a lack of studies that delve into the considerations of linguistic theory and that focus, as a first step, on the construction of a corpus that allows the study of the main linguistic indicators that distinguish between a discourse that intends to deceive and one that does not.

It is necessary to consider that detection of deception is complex, so it could be easy to fall into the Othello error, a concept coined by Ekman (2015) to refer to the errors in which the evaluator may fall if he/she does not consider that a person who is telling the truth may "appear" to be a liar when only one level of analysis is considered. In this sense, it is easy for biases such as gaze direction or the different comfortable certainties mentioned in this work to induce error. Hence, this paper seeks to reduce this type of errors through a promising approach, which is a verbal and cognitive one in which the analyst's bias (with the double-blind method) is reduced.

Regarding detection of deception and truthfulness assessment, it is clear that there are currently different research studies that question the analysis of *microexpressions* or *paralinguistics* as viable channels of analysis. Although this paper does not go into this issue in depth, it is considered that the study of these channels can always provide valuable information if they are considered as part of a whole. This implies aiming at a constellation of evidence in which the analysis of linguistic behavior is as important as the analysis of non-linguistic behavior: an isolated analysis of non-verbal behavior would be just as dangerous as concentrating exclusively on a strictly linguistic analysis. For methodological reasons, however, in this work greater weight has been given to the construction of the corpus from a more linguistic angle, without disregarding the other channels. Hence the audiovisual recording of the participants who took part in this project.

In sum, this work makes different contributions. First, it establishes methodological rigor in the construction of a corpus for the identification of linguistic strategies linked to lies and truthfulness. This implied the careful selection of participants, the elaboration of instruments such as the interview guide with a solid theoretical basis, and the submission of the project itself to an ethics committee.

Moreover, by obtaining two types of discourse (one truthful and one false), it is intended that soon it will be possible to study both differences in the same subject, that is, to know the linguistic baseline of the participant when he/she tells the truth in order to recognize the relevant and significant differences when the same subject lies. It should be considered that, in the future, specialists in "detection of deception" should focus on assessing truthfulness in discourse rather than on identifying lies. This project thus emphasizes both fallacious and truthful discourse.

Finally, an advantage of the project is the transdisciplinary ethos that it aims to have so that, over time, more collaborative work with different disciplines can be carried out to understand a phenomenon that is present in our daily interactions.

Appendix

Below is a summary of each of the appendices attached to the research in Spanish.

1. D. Diffusion

This appendix corresponds to the poster used for the search of volunteers. It indicates the requirements, includes contact information and general information about the research. The poster was circulated by the Language and Cognition Laboratory of the Center for Research in Cognitive Sciences (UAEM).

2. **CI. Informed Consent Form**

This appendix contains the informed consent form. This appendix includes the consent of the volunteers to participate in the collection of interviews as part of the Language and Cognition Laboratory project. It specifies the risks, type of research technique, rights, benefits, and general structure of the interview.

3. **GE. Interview Guide**

This appendix is divided into two sections: Instructor's Guide and Interviewer's Guide. In the first section, the general instructions to be given by the instructor to the interviewer are detailed, that is, to welcome him/her and the instruction to lie or tell the truth in each of the narratives according to the order decided by the instructor himself/herself. In the second section, more specific information about the project is mentioned, the instruction given by the instructor is reinforced without discovering the double-blind, a model description of the type of narrative expected is made, the participant's acceptance is asked again, and the interviewee's narrative begins with the completion of the question in which the narrative of the experience of the earthquake of September 19, 2017, is requested.

4. **CT. Transcription Conventions**

This section details the transcription conventions used. The criteria used in terms of spelling and punctuation, phonic and lexical labeling, labeling of discursive dynamics, and format criteria are mentioned.

References

Burgoon, J. K. (2018). Microexpressions Are Not the Best Way to Catch a Liar. *Frontiers in Psychology*, *9*, 1-5.

Blandón-Gitlin, I., López, R. M., Masip, J. y Fenn, E. (2017). Cognición, emoción y mentira: implicaciones para detectar el engaño. *Anuario de Psicología Jurídica, 27*(1), 95-106.

Columbia University, SRI International, and University of Colorado Boulder. (2013). *CSC Deceptive Speech LDC2013S09*. Recurso electrónico. Linguistic Data Consortium. https://doi.org/10.35111/q500-9a28

Darwin, C. (1872/2009). *La expresión de las emociones*. Laetoli.

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

Ekman, P. & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, *32*, 88-106.

Ekman, P. & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, *29*(3), 288-298.

Ekman, P., & O'Sullivan M. (1991). Facial expression: methods, means, and moues. In R. S. Feldman, & B. Rimé, (Eds.), *Fundamentals of Nonverbal Behavior* (pp. 163-199). Cambridge University Press.

Ekman, P. (2015). *Cómo detectar mentiras. Una guía para utilizar en el trabajo, la política y la pareja*. Paidós.

Ekman, P. (2017). *El rostro de las emociones. Qué nos revelas las expresiones faciales*. RBA.

Feldman, R. S., Forrest, J. A., & Happ, B. R. (2002). Self-presentation and verbal deception: Do self-presenters lie more? *Basic and Applied Social Psychology*, *24*(2), 163-170.

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing Techniques for Investigative Interviewing: The Cognitive Interview*. Charles C. Thomas.

Fisher, R. P., & Geiselman, R. E. (2019). Expanding the Cognitive Interview to Non-Criminal Investigations. In J. Dickinson, N. Schreiber Compo, R. Carol, B. L. Schwartz, & M. McCauley (Eds.), *Evidence-based Investigative Interviewing Applying Cognitive Principles* (pp. 1-28). Routledge, Taylor & Francis Group.

Fisher, R. P., Milne, R., y Bull, R. (2011). Interviewing cooperative witnesses. *Current Directions in Psychological Science*, *20*, 16-19.

Fitzpatrick, E. & Bachenko, J. (2009). Building a forensic corpus to test language-based indicators of deception. In: S. T. Gries, S. Wulff & M. Davies (Eds.), *Corpus-linguistic applications. Current studies, new directions* (pp. 183-196). Rodopi.

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L., Sullivan, S. J., Avetissian, I. V., & Prosk, A. L. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police and Science Administration*, *12*, 74-80.

Hauch, V., Sporer, S. L., Michael, S. W. & Meissner, C. A. (2014). Does training improve detection of deception? A meta-analysis. *Communication Research*, *43*(3), 283-343.

Hwang, H. C., Matsumoto, D. & Sandoval, V. (2016). Linguistic Cues of Deception Across Multiple Language Groups in a Mock Crime Context. *Journal of Investigative Psychology and Offender Profiling*, *13*, 56-69.

Infante Arriagada, P. (2015). La mentira como fenómeno lingüístico: algunos aspectos centrales para su descripción. *LL Journal*, *10*(2), 1-20.

Köhnken, G., Milne, R. Memon, A., & Bull, R. (1999). A meta-analysis on the effects of the Cognitive Interview. *Psychology, Crime, & Law*, *5*, 3-27.

Lancaster, G. L., Vrij, A., Hope, L. & Waller, B. (2013). Sorting the liars from the truthtellers: The benefits of asking unanticipated questions on lie detection. *Applied Cognitive Psychology*, *27*, 107-114.

Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2015). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology*, *20*(1), 129-146.

López Pérez, R. M., F. Gordillo León y M. Gau Olivares (coords.). (2016). *Comportamiento no verbal. Más allá de la comunicación y el lenguaje*. Pirámide.

Masip, J., Garrido, E. y Herrero, C. (2004). La detección de la mentira mediante la medida de la tensión en la voz: una revisión crítica. *Estudios de Psicología*, *25*(I), 13-30.

Masip, J., y Herrero, C. (2015). Nuevas aproximaciones en detección de mentiras I. Antecedentes y marco teórico. *Papeles del Psicólogo*, *36*(2), 83-95.

NITV Federal Services (2020). *Voice Stress Analyzer*. https://www.cvsa1.com/.

Pennebaker, J. W., Francis, M. E. & Booth, R. J. (2001). *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates.

Picornell, I. (2013). Analysing Deception in Written Witness Statements. *Linguistic Evidence in Security, Law and Intelligence*, *1*(1), 41-50.

Reissland, N., Francis, B., Mason, J. & Lincoln, K. (2011). Do Facial Expressions Develop before Birth? *PlosOne*, *6*(8), 1-7.

Salekin, R. T., Kubak, F. A. & Lee, Z. (2008). Deception in children and adolescents. In R. Rogers, & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (p. 343-364). The Guilford Press.

Villar, G., & Castillo, P. (2016). The Presence of 'Um' as a Marker of Truthfulness in the Speech of TV Personalities. *Psychiatry, psychology, and law: an interdisciplinary journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *24*(4), 549-560.

Vrij, A. (2001). Detecting the liars. *Psychologist*, *14*, 596-598.

Vrij, A. (2008). *Wiley series in the psychology of crime, policing and law. Detecting lies and deceit: Pitfalls and opportunities* (2ª ed.). John Wiley & Sons Ltd.

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

Vrij, A., Fisher, R. P., Mann, S., Deeb, H., Jo, E., Castro Campos, C., & Hamzeh, S. (2020). The Efficacy of Using Countermeasures in a Model Statement Interview. *The European Journal of Psychology Applied to Legal Context*, *12*(1), 23-34.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection. *Psychological Science in the Public Interest*, *11*(3), 89-121.

Vrij, A., Leal, S., Mann, S. A. y Fisher, R. P. (2012). Imposing cognitive load to elicit cues to deceit: Inducing the reverse order technique naturally. *Psychology, Crime & Law*, *18*, 579-594.

Vrij, A., Mann, S. A., Leal, S. & Fisher, R. P. (2010). "Look into my eyes": Can an instruction to maintain eye contact facilitate lie detection? *Psychology, Crime & Law*, *16*, 327-348.

Chapter IX

Türkisch für Anfänger: propuesta de un corpus del alemán coloquial actual, ejemplificado a partir de las fórmulas rutinarias de saludo

Türkisch für Anfänger: proposal of a corpus of modern colloquial German, exemplified from routine phrases for greetings

Karen Lorena Baquero Castro
Universidad de Salamanca – España; *Universidad Ean – Colombia*

Resumen: En el contexto de la enseñanza del alemán como lengua extranjera, aprendices y docentes se enfrentan al vacío de una didáctica que optimice el proceso y los resultados de aprendizaje de unidades fraseológicas. En un sentido amplio de la fraseología, se encuentran las fórmulas rutinarias, expresiones cuya polifuncionalidad y complejidad pragmática evidencian la necesidad de crear materiales auténticos basados en el análisis de corpus lingüísticos que apelen al contexto. Para ello, presento en este artículo la metodología de creación de una base de datos compuesta por 12.911 líneas de diálogo de la serie alemana *Türkisch für Anfänger*[1], el análisis de un subgrupo de fórmulas de saludo presentes en la misma y la correspondiente implicación didáctica para el aprendizaje de dichas unidades basadas en el alemán actual cotidiano.

1 Serie de televisión alemana de comedia dramática, producida en los años 2006 a 2008.

Abstract: In the context of teaching German as a foreign language, learners and teachers encounter a lack of didactics that optimizes the process and results of a learner in phraseological units. In a broad sense of the phraseology, there are conversational routines, expressions whose polyfunctional nature and complex pragmatics show the need to create authentic materials based on the analysis of a linguistic corpus that applies to the context. The present article shows the methodology used to create a database made up of 12,911 lines of dialogue from the German television series *Türkisch für Anfänger*[2]. In addition, it shows an analysis of a subgroup of greeting routines available in the corresponding didactic proposal to learn such phrasemes based on quotidian German language used today.

1. Introducción

Igor Sosa Mayor (2006, p.62) expone que los fraseólogos incluyeron el estudio de las fórmulas rutinarias cuando investigaban y establecían las características de otras unidades como los fraseolexemas. A pesar de haber sido incorporadas por Burger desde 1973 en la investigación fraseológica del alemán, bajo la denominación de "pragmatische Phraseme"[3], los investigadores aún no han llegado a un consenso sobre las características que las definen. Paradójicamente, sí existe claridad suficiente para la consideración de sus múltiples funciones en la comunicación oral y escrita: estructuración de discursos, adecuada interacción situacional, descarga de tiempo y estrés, así como el fortalecimiento del contacto social a través de la precisión lingüística.

Dada su relevancia, en esta investigación[4] consideramos necesario crear una fuente lingüística auténtica[5] del alemán actual que permita entre otras, indagar sobre sus diferentes usos y a partir de ello desarrollar estrategias didácticas para su aprendizaje. Partimos de la creación de un corpus compuesto por las líneas de diálogo de la serie de televisión alemana *Türkisch für Anfänger*. Nos proponemos revisar qué tipo de datos recopilados

2 German television comedy-drama series, aired between 2006 and 2008.

3 A lo largo de la literatura se encuentran diferentes términos para referirse a dichas unidades: Pragmatische Idiome (Burger, 1973), Routineformeln (Coulmas, 1981; Burger, 1998; Stein, 1985; Gläser, 1986; Lüger, 1999; Sosa Mayor, 2006), Kommunikative Formeln (Fleischer, 1982), kommunikative Phraseologismen (Burger, 1998), Kommunikative Routineformeln (Hyvärinen, 2003).

4 Esta investigación hace parte del proyecto doctoral que desarrollo en la Universidad de Salamanca en el área de lenguas modernas y que tiene como enfoque la creación de un corpus lingüístico que permita la sistematización de datos sobre las fórmulas rutinarias del alemán coloquial actual.

5 Nos basamos en la propuesta de Lüger (2009, p.15), para quien la autenticidad es aquello que es "real", "verdadero", "fiel al original" o "no artificial".

pueden ser usados para que los aprendices desarrollen su sentido lingüístico[6] y sean capaces de comunicarse usando fórmulas actuales y propias de los contextos coloquiales del alemán.

1.1 Propiedades de las fórmulas rutinarias

Estas unidades, "[…] deben poseer las características comunes a todas ellas, la fijación y en ocasiones la idiomaticidad, […] pero además pueden presentar algún tipo de independencia como enunciados fraseológicos que son" (Alvarado, 2008, p.93). Dentro de estas propiedades se distingue la importancia de la fijación formal y psicolingüística, "referida a la convencionalización en la comunidad lingüística, es decir, a la estabilidad en su producción y a su frecuencia de uso" (Alvarado, 2008, p.93). Como advierte la autora, dichos rasgos pueden ocurrir de manera gradual.

En las fórmulas rutinarias, la independencia es una característica primordial. Alvarado (2008, p.116) distingue: la independencia entonativa, distribucional, semántica, sintáctica y textual. En la primera de estas, la entonativa, se tiene en cuenta que estas unidades "son actos de habla que presentan fuerza ilocutiva exclamativa de sorpresa, admiración, rechazo, susto, etc., por lo que tienen un esquema entonativo propio […] (Alvarado, 2008, p.124). La independencia distribucional, como describe la autora, se refiere a la libertad que tiene el hablante de usar dichas unidades cuantas veces lo requiera (Alvarado, 2008, p.125), "por lo tanto está estrechamente ligada con el concepto de dependencia situacional, ya que un gran número de fórmulas depende siempre de la situación que se esté produciendo" (Alvarado, 2008, p.126). La independencia semántica tiene que ver con que "el valor de la fórmula está fijado por el contexto habitual en el que se produce y significa por sí misma y no necesita de otros elementos" (Alvarado, 2008, p.127). La última de estas, la independencia textual, es aquella que el corpus permite ver con mayor claridad, "si la fórmula se puede dar tantas veces en el discurso como se quiera es porque no depende del contexto lingüístico, sino del situacional" (Alvarado, 2008, p.126).

Según Winzer-Kiontke (2016, p.34), las fórmulas rutinarias se definen a partir de su frecuencia, coherencia fonológica, uso y grado de independencia. Si bien se puede hablar de un relativo consenso alrededor de la mayoría de las propiedades definitorias de las fórmulas rutinarias, hay una, sobre la cual se generan discrepancias, a saber, la polilexicalidad. Alvarado (2008) no la postula como una característica necesaria. Para Winzer-Kiontke (2016,

6 El sentido lingüístico es definido en el diccionario Merriam-Webster como un sentido intuitivo de lo que es apropiado lingüísticamente.

p.22), estas unidades tienden a ser polilexicales, pero se incluyen las que no lo son, es decir, aquellas monolexicales. Por monolexicales, comprendemos aquellas fórmulas cuyo límite mínimo es la palabra. Como propone Sosa (2006, p.27), es justamente este aspecto el más problemático en la clasificación de las fórmulas rutinarias ya que si se aplica de manera categórica el criterio de polilexicalidad, se deben excluir unidades del campo de las fórmulas rutinarias que, según el autor y nuestro estudio, deben ser tenidas en cuenta. Añade Sosa (2006, p.33) que incluso se tienen en cuenta aquellas fórmulas que por su frecuencia de uso dejan de ser polilexicales y se convierten en molexicales debido a procesos lingüísticos de elisión. Así, la monolexicalidad debe encontrarse dentro de las propiedades definitorias de dichas unidades. Para ilustrar la relevancia de fórmulas rutinarias monolexicales, dentro del corpus de nuestra investigación se ha encontrado un número total de 238 fórmulas de saludo, en las que se incluyen fórmulas como *hallo!*, *Tag!*, *Morgen,* o *hey*.

2. Clasificación de fórmulas rutinarias

En la literatura de la fraseología, se encuentran diferentes propuestas clasificatorias. Winzer-Kiontke (2016) retoma en su sistema de clasificación los aportes de Coulmas (1981), Pilz (1981), Gläser (1986), Zenderowska-Korpus (2004) y Sosa Mayor (2006). Tipos de fórmulas que aparecen en cada una de estas publicaciones como las de saludo, despedida, pésame, agradecimiento, disculpas y deseos, se tienen en cuenta de manera directa en su clasificación. Según esta propuesta, la autora recopila los 33 tipos de fórmulas en sentido estricto que se muestran en la siguiente tabla (Winzer-Kiontke, 2016, p.84):

Tabla 1. Categorías de fórmulas rutinarias según Winzer-Kiontke (2016).

Base de datos-categorías		
1. Fórmula de rechazo	12. Fórmula de restricción	23. Fórmula de comentario
2. Fórmula de despedida	13. Fórmula emotiva	24. Fórmula de contacto
3. Fórmula de ocasión	14. Fórmula de disculpas	25. Fórmula de estornudo
4. Fórmula de tratamiento	15. Fórmula de información	26. Fórmula de reprimenda y grosería
5. Fórmula de exhortación	16. Fórmula de advertencia	27. Fórmula de lenguaje escrito
6. Fórmula de compasión	17. Fórmula de aliento	28. Fórmula de sorpresa
7. Fórmula de bienvenida	18. Fórmula de asombro	29. Fórmula de presentación
8. Fórmula de pésame	19. Fórmula de respuesta	30. Fórmula de advertencia (En sentido amplio: Fórmula de prohibición)
9. Fórmula de apaciguamiento	20. Fórmula de alimento y bebida	31. Fórmula de recibimiento
10. Fórmula de aseveración	21. Fórmula de saludo	32. Fórmula de deseo
11. Fórmula de agradecimiento	22. Fórmula institucional	33. Fórmula de consentimiento

Al igual que esta propuesta, se han planteado un sinnúmero de clasificaciones, que, en su mayoría, como la de Winzer-Kiontke (2016, p.84), apelan a la teoría de los actos de habla. Dentro de las más completas también se incluye la de Alvarado (2008, p.268), considerando que es otra clasificación precisa para este grupo de unidades fraseológicas, aunque pensada para las fórmulas rutinarias del español, y que permite de entrada incluir las fórmulas rutinarias discursivas.

En la propuesta de Alvarado, el hablante es el punto de partida "que codifica sus emociones en la fórmula rutinaria" (2008, p.268). Tiene en cuenta dos modalidades: "la modalidad lógica, que se relaciona con la verdad de lo que se dice, y la modalidad subjetiva, que muestra la valoración del hablante" (Alvarado, 2008, p.268). En el grupo de fórmulas rutinarias lógicas se distinguen las epistémicas, que "se vinculan con el ámbito de la posibilidad de que un enunciado sea cierto" (Alvarado, 2008, p.269) y las deónticas, que "expresan la obligatoriedad de que se cumpla lo que el hablante dice […]" (Alvarado, 2008, p.279), allí se incluyen las fórmulas declarativas, interrogativas, imperativas y exclamativas. En cuanto a las fórmulas rutinarias subjetivas, la autora plantea dos categorías, las afectivas, que expresan la emoción del hablante y las evaluativas, "que codifican la modalidad subjetiva, puesto que manifiestan la actitud del hablante frente al *dictum,* y evalúan dicho enunciado en términos valorativos" (Alvarado, 2008, p.315). En un último grupo, se encuentran las fórmulas rutinarias discursivas, cuya función consiste en darle orden al discurso a partir de tres distinciones: apertura, transición y cierre (Alvarado, 2008, p.318).

Consideramos la propuesta de clasificación de Winzer-Kiontke (2016) como la más adecuada para los fines propuestos en nuestro trabajo. Teniendo en cuenta el carácter del corpus, nos inclinamos por una clasificación pragmática basada en el uso de las unidades fraseológicas según su contexto o situación. Consideramos que el aporte de este corpus consiste justamente en la explotación de las unidades allí identificadas y clasificadas. Dicha clasificación permite que las reflexiones didácticas que de allí surjan sean más operativas y de este modo más sencillas de comprender para un aprendiz de la lengua.

3. El corpus

En el campo de la lingüística moderna, el uso de corpus se ha dado de manera extendida. De acuerdo con Villayandre, fue el uso de los computadores para "reunir, organizar, y procesar esos datos el que ha dotado de modernidad a esta tarea, hasta el punto de propiciar el despegue de una forma de hacer lingüística, la llamada 'lingüística de corpus' (2008, p.330).

El concepto de corpus previo al desarrollo de los computadores se definía a partir de la recopilación de textos con el fin de analizar fenómenos de lenguas muertas y tenía como objeto indagar sobre la adquisición del lenguaje a temprana edad, precisar reglas de ortografía, hacer listas de vocabulario, comparar lenguas y crear gramáticas (Villayandre 2008, p.330). Aunque durante el siglo XIX se vive en esta disciplina un acelerado desarrollo, es solo a partir del siglo XX cuando esta se convierte en metodología con la lingüística americana estructuralista.

Con la postura de Chomsky que cuestionaba la metodología del empirismo, la disciplina pierde auge y desarrollo. La postura del autor se centraba esencialmente en tres puntos: carencia del uso de la intuición a la que debe recurrir el lingüista, el carácter incompleto de los datos que contienen los corpus y la metodología dispendiosa que implica el análisis de datos (Villayandre, 2008, p.333). Sin embargo, dichas críticas se superaron a partir de argumentos sobre la gramaticalidad de los elementos del corpus, los datos cuantitativos y su representatividad y el uso de computadores. Así, el mayor desarrollo de la lingüística de corpus se aprecia desde la década de 1980 (Villayandre, 2008, p.337).

Algunos de los corpus más representativos creados en dicha década son: el 'Bank of English' el CREA (Corpus de Referencia del Español Actual) y CORDE (Corpus Diacrónico del español). En dichos corpus se debe cumplir con características primordiales como tener un formato digitalizado, criterios que permitan la selección de información bien sea lingüística o extralingüística, representatividad estadística y tamaño por lo general finito (Villayandre, 2008, p.341). Nuestro corpus, al conformarse a partir de las líneas de dialogo de la serie alemana mencionada previamente, se define como corpus del alemán oral actual. En la creación de este, se llevó a cabo el proceso de transcripción de 52 capítulos que constituyen las 3 temporadas de la serie[7]. Para el alemán existe un gran número de corpus[8], sin embargo, estos no están recopilados propiamente para un uso didáctico, como lo expone Wallner (2014). Así, el tamaño de nuestro corpus permite un manejo adecuado de información enriquecida para el desarrollo de materiales didácticos de aprendices del alemán coloquial actual.

7 El proceso de transcripción utilizado se realizó manualmente con el fin de garantizar la precisión de las transcripciones, usando como fuente de apoyo los subtítulos descriptivos, junto con el uso de programas como oTranscribe y Amberscript. Las líneas de dialogo fueron revisadas durante y después de la transcripción. El corpus tuvo un proceso de revisión extenso a cargo de un ingeniero de bases de datos y una doctoranda especializada en la enseñanza del alemán con nivel de alemán C1 y C2, respectivamente. Una vez definida la base de datos para el uso del material lingüístico, el corpus pasará por una tercera revisión de un hablante lingüista y nativo del alemán.

8 Dentro de los corpus del alemán escrito se cuentan, entre otros: das Deutsche Referenzkorpus - DeReKo (Instituts für Deutsche Sprache, 2018), Digitales Wörterbuch der deutschen Sprache – DWDS (Berlin-Brandenburgischen Akademie der Wissenschaften, s.f.), das Projekt deutscher Wortschatz (Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig, 1998) y el corpus Südtirol (Team Korpus Südtirol, s. f.). Para el alemán oral existe el Datenbank gesprochenes Deutsch - DGD2 (Deppermann & Schmidt, 2014) y el GeWiss (Herder-Institut - Universität Leipzig, s. f.).

En el contexto de esta investigación, entendemos por corpus el conjunto de textos orales que han sido digitalizados a partir de la recopilación y estructuración de las líneas de diálogo de la serie alemana *Türkisch für Anfänger*. Como describe Jens (2015, p.16), desde mediados de los noventa, diferentes cómicos como Mundstuhl, Kaya Yanar Spaß-Duo Erkan y Stefan potenciaron el lugar de los llamados etno-formatos en la radio y con estos, los diferentes estilos de la lengua. En este contexto, surge la etno-comedia *Türkisch für Anfänger*. Entre los años 2006 y 2008, ya se habían creado tres temporadas de esta serie que llamaba la atención sobre la relación de la familia turco-alemana Schneider-Öz-Türk, conformada por una madre alemana de Berlín-Neukölln y un padre turco, ambos con sus dos hijos de tradición turco y alemana, respectivamente. A través del humor en la sobre-actuación de los clichés de ambas culturas, se logran plasmar aspectos de la actualidad alemana como la inmigración, la interculturalidad y la búsqueda de identidad de los inmigrantes. En este proceso, la lengua usada comienza a proponer reflexiones de índole social, como advierte Jens (2015) sobre uno de los personajes de la serie:

> *con su elección lingüística, de la prosodia, como también de su lenguaje corporal y su ropa, remite Cem a ambientes sociales característicos en los que se desenvuelve. En esta forma extrema estilizada de hablar unifica elementos de la cultura Hip-Hop, como de anglicismos adaptados (…) con elementos típicos juveniles (p.16).*

Justamente esta riqueza semántica, física y visual es la que nos interesa para proponer la construcción de un corpus lingüístico de este idioma que contenga variedades diatópicas, diafásicas y diastráticas del alemán oral actual y que den cuenta de las diferentes estrategias sintácticas, fonológicas, gestuales, corporales y lexicales que dan lugar a lo que la autora denomina la "realización de una categoría identitaria" (Jens, 2015, p.18).

Consideramos que justamente es este proceso de construcción de identidad el que experimentan los aprendices de lengua y, por tanto, la finalidad de este corpus consiste en conducir a los aprendices a la lengua auténtica, entendida esta como aquella que es cercana a la lengua en uso, y a un proceso de identificación con su propia construcción de identidad que se da en la lengua meta.

Es importante aclarar que no desconocemos que la lengua usada en una serie de televisión corresponde a lo que diversos autores han denominado "la oralidad fingida" (concepto introducido por Goetsch (1985, p.202) para describir la oralidad de textos literarios y que se refiere a la "ilusión de autenticidad" que existe en lo escrito que ha sido creado para lo oral, como lo es un guion de televisión), de modo que el corpus que creamos a partir de una lengua con estas características se permea de ellas.

Consideramos, por tanto, que este tipo de oralidad creada "puede contribuir a crear la ilusión de verosimilitud, ayudar a situar la acción en una determinada época y región, contrastar el lenguaje de los personajes según la pertenencia a cierta clase social o según la educación, y cotejar la incorporación de elementos procedentes de la tradición y el saber orales" (Goetsch, 1985, p.217). A pesar del reconocimiento de dicha ilusión consideramos que este corpus representa un material cercano a lo auténtico y real en el uso oral del alemán actual.

Para la comprensión de los resultados cuantitativos, en términos de usos de las fórmulas rutinarias de acuerdo con el interlocutor, proponemos el siguiente cuadro descriptivo de personajes:

Tabla 2. Personajes de la serie Türkisch für Anfänger.

Personaje	Actor/Actriz	Rol	Descripción
Lena Schneider (personaje principal)	Josefine Preuß	Hija de Doris y Markus; hermana de Nils; hermanastra de Yagmur y Cem	Es una adolescente de 16 años, estudiante de Secundaria de origen alemán.
Doris Schneider	Anna Stieblich	Madre de Lena y Nils; hija de Hermi, hermana de Diana; madrastra de Yagmur de Cem; esposa de Metin	Mujer adulta y psicoterapeuta de origen alemán.
Metin Öztürk	Adnan Maral	Padre de Cem y Yagmur; padrastro de Nils y Lena; esposo de Doris	Es un adulto comisario de origen turco.
Cem Öztürk	Elyas M'Barek	Hijo de Metin; hermano de Yagmur; hermanastro de Lena y Nils; Ex novio de Ching y Ulla	Es un joven estudiante que al terminar sin éxito el examen de secundaria estudia para formarse como policía. Su origen es turco.
Yagmur Öztürk	Pegah Ferydoni	Hermana de Cem, hija de Metin; hermanastra de Nils y Lena	Joven estudiante de secundaria de origen turco. Posteriormente se dedica a la traducción de textos turcos al alemán y trabaja para el Parlamento Alemán. Su origen es turco.
Costa Papavassilou	Arnel Taci	Mejor amigo de Cem; prometido de Yagmur	Joven estudiante de secundaria que al terminar la secundaria crea su propio negocio de moda. Su origen es griego.

4. Análisis cuantitativo del corpus

Con el fin de analizar el uso de las fórmulas rutinarias en el contexto de saludo, es importante tener en cuenta la representación que cada personaje tiene en la serie en términos de

su participación como interlocutor, esto es, en términos del tiempo de intervención. A continuación, se observa que en consecuencia con el rol que asume Lena, la protagonista, es quien más participa; asimismo, Doris, su hermanastro Cem y su padrastro Metin.

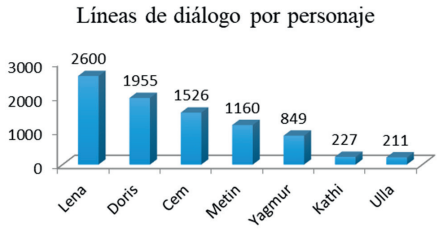


Figura 1. Líneas de diálogo por personaje.

Dentro de las fórmulas rutinarias de saludo informales encontradas en el corpus, la más utilizada es *hey*. De las 200 ocurrencias de esta fórmula rutinaria, 107 tienen la función de saludo, las demás 93 se utilizan en contextos en los que los interlocutores llaman la atención y se categorizan como fórmula rutinaria de contacto. La segunda más usada es *hallo!* y le sigue *hi!* De estas, las más frecuentes en los textos de aprendizaje suelen ser *hallo!* y *hi!*; esporádicamente se incluye *hey*, contrario a lo que muestra el corpus. Este fenómeno también se presenta en las fórmulas *Guten Morgen* y *Morgen,* la segunda de estas es más usada en el corpus y no necesariamente en los libros de enseñanza como Studio d (2010), Berliner Platz neu (2017) o incluso más recientes como Linie 1 (2017). Se explica esto teniendo en cuenta que la mayoría de los manuales se suelen regir por la norma escrita y no por la norma hablada.

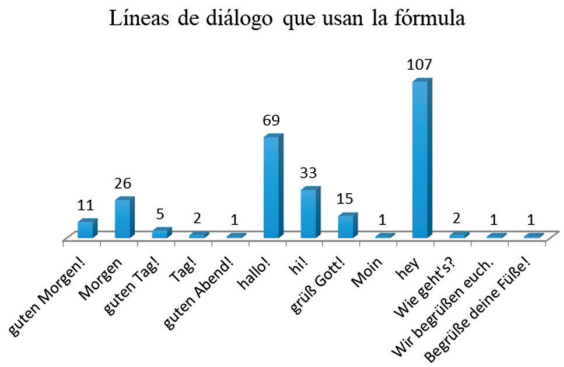


Figura 2. Líneas de diálogo que usan fórmulas rutinarias de saludo o recibimiento.

La fórmula rutinaria más frecuente, *hey*, es usada incluso por personajes que tienen pocas intervenciones. En cuanto a fórmulas de recibimiento explícitas, únicamente se identifican dos a lo largo de todo el corpus: *Wir begrüßen euch* y *Begrüße deine Füße!* Se destaca que esta fórmula es mayoritariamente usada por interlocutores jóvenes, resaltados en negrilla en la siguiente gráfica. Precisamente, Doris, siendo un personaje principal y que tiene en el corpus una participación comparable a la de Lena, no hace uso de la fórmula *hey*. Por tanto, habría que tener en cuenta esta variable relativa a la edad.

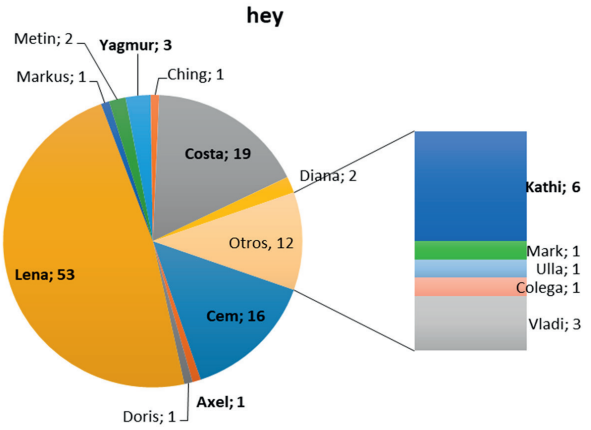


Figura 3. Uso de la fórmula rutinaria *hey*, por personaje.

Contrario a esta particularidad, la fórmula rutinaria *hallo* es usada tanto por jóvenes como adultos, como se observa a continuación:

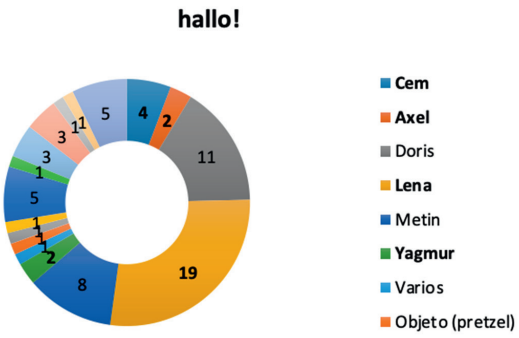


Figura 4. Uso de la fórmula rutinaria *hallo!*, por personaje.

Si bien la fórmula rutinaria *hi* tiene un uso algo frecuente, es usada solo una vez por el personaje Doris (madre) y no es usada por Metin (padre), dos de los personajes de mayor edad en la serie.

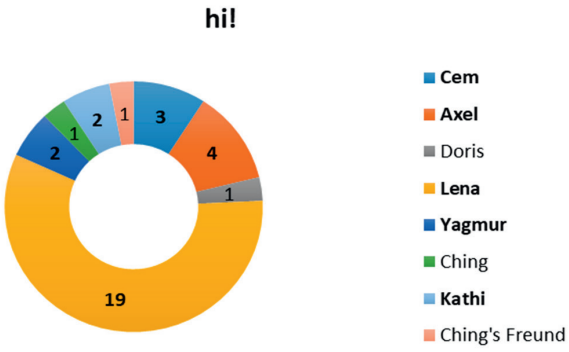


Figura 5. Uso de la fórmula rutinaria *hi!*, por personaje.

Teniendo en cuenta los personajes de mayor participación, se puede inferir que los personajes más jóvenes tienden a utilizar en igual medida tanto *hallo* como *hi* y en menor proporción usan *hey*. Por otra parte, vemos que los personajes adultos utilizan más *hallo* y no *hi* ni *hey*.

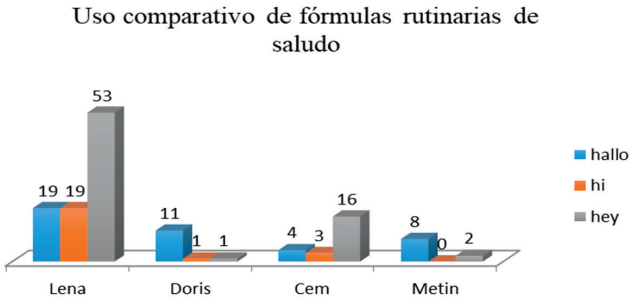


Figura 6. Uso comparativo de las fórmulas rutinarias de saludo hallo, *hi* y *hey*, por personaje.

Morgen aparece como una fórmula destacada dentro de las unidades fraseológicas de saludo (26 veces), incluso con más del doble de las ocurrencias de *guten Morgen* (11 veces). En ambas se observa una frecuencia de uso independiente de la edad del interlocutor.

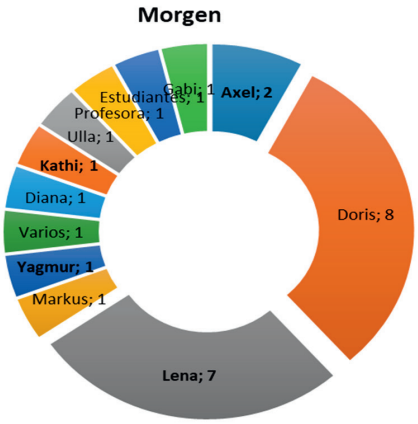


Figura 7. Uso de la fórmula rutinaria *Morgen*, por personaje.

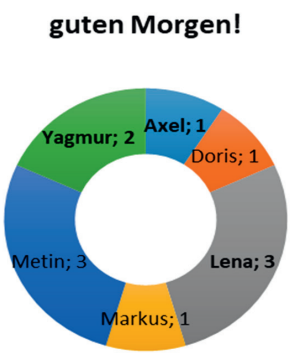


Figura 8. Uso de la fórmula rutinaria guten *Morgen!*, por personaje.

La gráfica a continuación nos muestra la posibilidad de usar la mayoría de fórmulas rutinarias de saludo agregando un nombre a su estructura, como por ejemplo, *hey Kathi*, *Morgen Cem, hallo Metin, hi Axel, Wie geht's Yagmur?, grüß Gott Cem!* A excepción de *guten Abend*, *Mahlzeit! y Moin,* en todas las demás fórmulas aparece un nombre dentro de su estructura, al menos una vez.

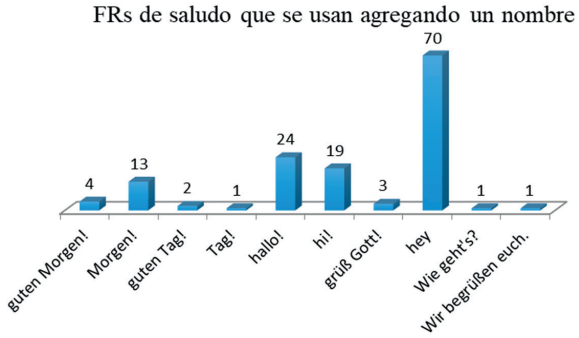


Figura 9. Fórmulas rutinarias de saludo que se usan en combinación con un nombre.

5. Reflexiones didácticas

Con los hallazgos descritos, es consecuente plantear ejercicios didácticos en los que el profesor de lengua entregue a sus aprendices un *input* de fórmulas rutinarias que los sensibilice frente a la posibilidad de crear conversaciones que consten únicamente de dichas unidades, como lo muestra el siguiente ejemplo de nuestro corpus: *Hallo, hier ist die Doris. Macht euch 'nen ganz schönen Abend und kommt auf keinen Fall vor vier nach Hause, ist das klar*? A su vez, la fuente lingüística de los aprendices puede proponerse a manera de ejercicio en el que se deban ordenar las líneas del diálogo y que de este modo el aprendiz se enfrente a la toma de decisiones frente a los espacios del diálogo más adecuados en el uso de las fórmulas rutinarias. Con ello, notará el aprendiz que para algunas de estas unidades la dependencia del contexto será más o menos rigurosa.

Además, en concordancia con los resultados de los datos cuantitativos, podemos plantear didácticas de aquellas unidades que hayan sido recurrentes en su uso como la fórmula *hey, hallo o hi*, pero a la vez sobre aquellas no tan representativas como *grüß Gott*, de las que se obtenga información que pueda ser revisada a la par de aquello propuesto en los manuales de enseñanza. Así, algunas de las actividades lingüísticas pueden partir de la asignación de tareas por parte del profesor que permitan la exploración del uso de fórmulas rutinarias con particularidades de tipo regional. Allí por ejemplo el aprendiz puede indagar, desde un punto de vista analítico y a través de la observación del contexto, qué interlocutor hace uso de la fórmula *grüß Gott*[9], sus características como hablante y las

9 *grüß Gott* es utilizado únicamente por Ulla, quien se caracteriza por ser muy religiosa. Algunas situaciones en las que usa la fórmula son: en el saludo del buzón de su teléfono ("*Grüß Gott hier spricht Ulla!*"), al saludar en persona ("*Grüß Gott Mr. Rimp.*") y al presentarse ("*Grüß Gott ich bin Ulla*"). Se puede tener en cuenta que esta

condiciones de uso de esta unidad frente a los contextos. El docente, por su parte, puede integrar en el aula guías didácticas que aprovechen el potencial de todo lo que un medio visual ofrece: imagen, sonido, texto. La imagen podrá ser revisada en términos de la gestualidad que conlleva el gesto de una fórmula; el sonido permitirá reflexionar sobre la fonética o entonación y el texto se convertirá en una fuente para el desarrollo de ejercicios de tipo lingüístico o cultural. Estos ejercicios permitirán que el aprendiz se acerque a lo que Lavid (2005, p.142) denomina el conocimiento pragmático que implica el saber del contexto *lingüístico-discursivo*, así como del *extra-lingüístico*.

Dentro de nuestras propuestas también sugerimos abordar el corpus desde la perspectiva de la fraseodidáctica contrastiva. En esta, los aprendices recurren a sus conocimientos de lengua materna y de su mundo conocido con el fin de crear traducciones en la forma de subtitulación o doblajes de la serie de la que cuentan con un texto recopilado en la forma de corpus y que puede ser llevado a la comprensión de los significados de las fórmulas rutinarias en el contexto auténtico y real de su uso. Así, la reflexión desde la lengua materna les permitirá hacer deducciones sobre fenómenos que caractericen dichas unidades como su gestualidad o entonación y con ello fortalecer las competencias comunicativas orales de la lengua en fase de aprendizaje, en este caso, del alemán.

6. Conclusiones

Este análisis de tipo cuantitativo nos permitió clasificar los datos observados y describir aspectos de la lengua que a continuación pueden ser tenidos en cuenta en la reflexión didáctica. Hemos detectado, a partir de resultados representativos, que *hey*, incluso aunque no sea una fórmula usualmente incluida en los textos de enseñanza, sí cuenta con un uso extendido por parte de interlocutores jóvenes. Por otro lado, al comparar otras fórmulas de saludo, para los adultos de la serie fue más frecuente el uso de *hallo*. Notamos también que otras fórmulas pueden ser utilizadas en la comunicación oral en combinación con un nombre propio o un pronombre. Algunos casos muestran también el uso de dos fórmulas rutinarias como *Hallo Metin, schön dich zu sehn; Vorzimmer Dr. Schneider, guten Tag, was kann ich für Sie tun?; Hi Cem! Na, was geht so; Hi Kathi! Tschuldige, dass ich mich jetzt erst melde...*[10].

Ahora bien, recurriendo a los datos, a su clasificación, a su análisis y uso, proponemos algunas reflexiones didácticas que permitirán además la recepción y producción de dis-

fórmula tiene una marca regional del sur de Alemania y de Austria.

10 Dentro del corpus se encontró un total de 52 fórmulas de saludo combinadas con otra fórmula.

cursos que articulen un lenguaje cercano a lo auténtico del alemán como lengua extranjera. Partimos de la hipótesis según la cual "en una palabra, los análisis cuantitativos permiten explorar y llevar a cabo descubrimientos sobre los patrones de uso de la lengua de forma rigurosa y fiable, ya que permiten comprobar empíricamente las hipótesis sobre el uso de la lengua" (Lavid, 2005, p.325).

El carácter representativo del alemán coloquial actual que muestra el corpus descrito al inicio del artículo, así como su fácil manipulación[11], permitirá que este sea explotado tanto por aprendices como por profesores de la lengua alemana, facilitando la creación de aplicaciones didácticas. Al respecto, de acuerdo con la propuesta de Lavid (2005, p.139), si bien la ventaja en el uso de corpus se basa en la posibilidad de indagar los significados de determinados términos de acuerdo con su aparición y distribución, son también relevantes en el análisis los términos que no estén representados de manera significativa. Las fórmulas rutinarias son complejas por su componente social y contextual y deben ser puestas en conocimiento del aprendiz desde el principio del proceso de aprendizaje. Tal es el caso del subgrupo de fórmulas rutinarias de saludo, para las que hemos concluido que, aunque ciertos hablantes adultos no utilizan, es común en el contexto del alemán de los jóvenes. Ejemplos de ellos son *hey* o *hi*, dos fórmulas que se descuidan en los manuales y por ende muchas veces en el aula de la enseñanza del alemán como lengua extranjera.

Referencias

Alvarado, M. (2018). *Las fórmulas rutinarias en español actual* [Tesis doctoral, Universidad de Alicante]. Repositorio Institucional de la Universidad de Alicante. http://rua.ua.es/dspace/handle/10045/7726

Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig. (15 de enero de 2019). *Deutscher Wortschatz / Leipzig Corpora Collection, 1998*. http://wortschatz.uni-leipzig.de/de

Berlin-Brandenburgischen Akademie der Wissenschaften. (11 de diciembre). *DWDS – Digitales Wörterbuch der deutschen Sprache*. https://www.dwds.de

Burger, H. (1973). *Idiomatik des Deutschen (Germanistische Arbeitshefte)*. Niemeyer.

Burger, H. (1998). *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Erich Schmidt Verlag.

Coulmas, F. (1981). *Routine im Gespräch: zur pragmatischen Fundierung der Idiomatik.* Athenaion.

Deppermann, A. & Schmidt, T. (2014). Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik: Eine exemplarische Untersuchung auf Basis des Korpus FOLK in

11 Con el corpus creado, se está desarrollando una herramienta que posibilita una consulta rápida por contenido, permitiendo la depuración de este y observación de estadísticas. Inicialmente, la herramienta que se ha usado para realizar estas tareas es Mongo DB y el objetivo final es desarrollar una herramienta sencilla que ofrezca una funcionalidad de consulta a través de un navegador web y facilite interactuar con el corpus de una forma más intuitiva. Dicha funcionalidad permitirá observar el contexto de uso de las fórmulas rutinarias dentro del corpus.

der Datenbank für Gesprochenes Deutsch (DGD2). *Mitteilungen des Deutschen Germanistenverbandes, 61*, 4-17. https://doi.org/10.14220/mdge.2014.61.1.4

Fleischer, W. (1982). *Phraseologie der deutschen Gegenwartssprache.* VEB Biblio- graphisches Institut.

Gläser, R. (1986). *Phraseologie der englischen Sprache.* Niemeyer.

Goetsch, P. (1985). Fingierte Mündlichkeit in der Erzählkunst entwickelter Schriftkulturen. *Poetica,* (17), 202-218.

Herder-Institut - Universität Leipzig. (25 de febrero de 2019). *Gesprochene Wissenschaftssprache*. https://gewiss.uni-leipzig.de

Hyvärinen, I. (2003). Kommunikative Routineformeln im finnischen DaF-Unterricht. *Info DaF: Informationen Deutsch als Fremdsprache,* (4), 335-351. https://www.academia.edu/1025371/Ein_Terrain_des_Fremdsprachenunterrichts_Deutsch_Interkulturelle_Kompetenz_in_der_Tourismusausbildung

Hyvärinen, I. (2011). *Beiträge zur pragmatischen Phraseologie.* Peter Lang.

Institut für Deutsche Sprache, Ausbau und Pflege der Korpora geschriebener Gegenwartssprache. (26 de enero de 2019). http://www1.ids-mannheim.de/kl/projekte/korpora

Jens, M. (2015). Mehrsprachigkeit: flexibles Repertoire statt Defizit. Die deutsche Ethno-Comedy Türkisch für Anfänger. *Sprachreport: Informationen und Meinungen zur deutschen Sprache,* (28), 16-19.

Lemnitzer, L. & Zinsmeister, H. (2015). *Korpuslinguistik: Eine Einführung.* Narr Francke Attempto.

Lüger, H. (1999). *Satzwertige Phraseologismen: Eine Pragmalingustische Untersuchung.* Präsens Verlag.

Lüger, H. (2009). Authentische Mündlichkeit im fremdsprachlichen Unterricht?, *Beiträge zur Fremdsprachenvermittlung. Sonderheft*, (15), 15-37.

Merriam-Webster. (2019). *Merriam-Webster Dictionary*. https://www.merriam-webster.com/dictionary/sprachgefühl

Pilz, K. (1981). *Phraseologie: Redensartenforschung.* Metzler.

Sosa, I. (2006). *Routineformeln im Spanischen und im Deutschen. Eine pragmalinguistische Analyse.* Präsens Verlag.

Stein, S. (1995). *Formelhafte Sprache. Untersuchungen zu ihren prag-matischen und kognitiven Funktionen im gegenwärtigen Deutsch*. Lang.

Team Korpus Südtirol. (26 de febrero de 2019). *Korpus Südtirol*. http://www.korpus-suedtirol.it/

Villayandre, M. (2008). Lingüística con corpus (I). *Estudios Humanísticos. Filología,* (30), 329-349. https://doi.org/10.18002/ehf.v0i30.2847

Wallner, F. (2014). Lehren und lernen mit Korpora im DaF-Unterricht. *Magazin Sprache vom Goethe-Institut München*.

Winzer-Kiontke, B. (2016). *"Gäbe es das Lehrwerk, würden wir es Ihnen empfehlen." Routineformeln als Lehr-/Lerngegenstand*. IUDICIUM Verlag.

Zenderowska-Korpus, G. (2004), *Sprachliche Schematismen des Deutschen und ihre Vermittlung im Unterricht DaF.* Peter Lang.

Chapter X

CLEC - Colombian Learner English Corpus: first learner corpus of written production in English online in Colombia

CLEC - Corpus Colombiano de Aprendices de Inglés: primer corpus de producción escrita de aprendices de inglés en Colombia disponible en línea

María Victoria Pardo Rodríguez[a] & Antonio Jesús Tamayo Herrera[b]
Universidad de Antioquia ([a]) *–Colombia*; *Instituto Politécnico Nacional* ([b]) *– México*

Abstract: This article aims to introduce CLEC's web application (Colombian Learner English Corpus) to the research community. This application was created to search for information within a learner corpus labeled with error tags to add, modify and eliminate data. After having the corpus collected and tagged, it was necessary to create a tool that systematically searches for information within the labeled data. The compilation of the learner corpus followed the guidelines of the Computational Corpus Linguistics (McEnery & Hardie, 2011) and the parameters of learner corpus Granger (2002), Gilquin (2015). The result is a web app designed to seek error tags within a context that can be easily revised and expanded through the system administrator. This corpus is available online, and it is open to any researcher who wants to consult it or contribute with data to enhance the corpus.

Resumen: Este artículo tiene como objetivo presentar la aplicación web de CLEC (Colombian Learner English Corpus) a la comunidad investigadora. Esta aplicación

fue creada para buscar información dentro de un corpus de aprendices etiquetado con etiquetas de error para agregar, modificar y eliminar datos. Luego de haber recolectado y etiquetado el corpus, fue necesario crear una herramienta que hiciera búsquedas sistemáticas de información dentro de los datos etiquetados. La compilación del corpus de aprendices siguió las pautas de la Lingüística de Corpus Computacional (McEnery & Hardie, 2011) y los parámetros de los corpus de los aprendices Granger (2002), Gilquin (2015). El resultado es una aplicación web diseñada para buscar etiquetas de error dentro de un contexto que se puede revisar y expandir fácilmente a través del administrador del sistema. Este corpus está disponible en línea y está abierto a cualquier investigador que quiera consultarlo o que quiera aportar nuevos datos para aumentar el corpus.

1. Introduction

Learner corpora (LC) emerged in the late 1980s (Granger *et al*., 2015) as a valid scientific way to analyze learners' output and has the same characteristics attributed to other corpora with the difference that the source of data is the output of language learners. Defined as "electronic collections of natural or almost natural data produced by foreign or second-language students (L2) and gathered according to explicit design criteria" by Granger (2002, p.7) and Gilquin (2015, p.1). LC has gained significance in the analysis of students' production. Regarding the authenticity of the data produced in a classroom, it is important to remember that the environment is not completely natural because the activities to obtain that input involve some kind of "artificiality" (Granger, 2002, p.8). Also, special attention must be paid to the criteria to build the corpus. The learner corpus' metadata, such as students' characteristics and the task they develop, are important factors for data collection.

The growth of LC in the late 1980s was in part to its potential to investigate authentic output from students. This methodology gives researchers access to outstanding amounts of data samples to do searches for collocations, patterns, and statistics. In the field of research on second and foreign language acquisition and teaching, learner corpora give access to learners' errors when they have been previously tagged, facilitating the analysis of such errors.

Error Analysis (EA) appeared in the early 1970s, and Corder (1967) was the first author to propose the idea that second language learners generated an autonomous linguistic system that he called "*transitional competence*". The author argued that learners gradually modify their native language rules towards target language rules, probably using a univer-

sal grammar or what he called a "*built-in syllabus*". Later, Selinker (1972) called the built-in syllabus *interlanguage*, and this is the term that has prevailed in time. It refers to the version of language produced by a learner. The analysis of the interlanguage of learners can be performed through the analysis of errors. Error analysis is "the investigation of the language of second language learners" (Corder, 1971, p.14). These analyses can be done using electronic learner corpora to obtain statistics and patterns and analyze what learners lack or need in their learning process. A learner corpus can be very useful when it has error labels to facilitate extensive studies.

Although the usefulness of a corpus of learners' language with error labeling is undeniable, it does not, on itself, facilitate extensive studies that could be carried out on it. For that reason, taking advantage of the fact that this corpus has a marking of errors in a set of texts, a collection of documents was generated and later uploaded into a database. After having the corpus collected in electronic format, there was a need for a tool that allowed researchers access to the corpus and provided the possibility of making queries with different filters.

The present paper starts with a brief description of the previous related work in learner corpora. Then, it describes the theoretical framework that supports this work along with the process followed during the compilation of the present corpus and the error tagging process. Afterwards, it narrates how the CLEC[1] app was designed and how it works to obtain its best performance. This project was developed with the research group Translation and New Technologies (TNT) of the School of Languages at Universidad de Antioquia and makes part of the products of a doctoral thesis.

2. Previous work

There are numerous corpora of English learners that contain samples of learners who have Spanish as their mother tongue, UC Louvain, (2018). Some of them are the Written Corpus of Learner English (WRICLE) Mendikoetxea *et al*., (2009); the Santiago University Learner of English Corpus (SULEC) Santiago University, (2002): the Gachon Learner Corpus (GACHON) Carlstrom and Price, (2012); the NOn-native Spanish corpus of English (NOSE) Díaz-Negrillo, (2012); the International Corpus of Learner English (ICLE) Granger, (2003). The ICLE and the NOSE can be highlighted as corpora of English language with samples of learners who have Spanish as their mother tongue. The ICLE is considered a pioneer in the field of learner language corpus. It has a relatively large collection (approx-

1 CLEC can be accessed via this URL: https://grupotnt.udea.edu.co/clec

imately 3.7 million words) of learners' written output from 16 different mother tongues, including Spanish. A CD containing the collection of texts must be purchased along with a desktop software to carry out searches and analysis on them to have access to this corpus. On the other hand, the NOSE (The NOn-native Spanish Corpus of English) has a collection of approximately 1000 argumentative and descriptive texts from students at the University of Granada and University of Jaen. It has labeling of errors under the EARS system Diaz-Negrillo, (2009). Apparently, this corpus had a web interface for its consultation allowing filtering by subject, text type, and parameters of the student's profile, but it is currently not accessible. Most of these corpora lack error labeling, and none of them currently has an accessible interface for researchers or the public to allow searches on them.

The corpus of the present analysis has a collection of documents labeled with error tags. It lets researchers, students, and teachers carry out searches systematically and with the possibility of filtering errors on different categories and types. Also, with this app, it is possible to obtain examples of these errors and their corrections. For the case of errors that represent more than one error category, a new functionality was developed to change error tags when necessary. This development results from a long process of trial and error, plus tests to achieve an app that allows adding, modifying, or eliminating errors or documents. These functionalities are carried out with a corpus management system that is powerful, versatile, and friendly. Initially, the development of this app was carried out in a technology called Django, which makes use of the Python language, but it was determined that the app should allow not only to consult but also to comply with all the initials of the CRUD concept (James, 1980) (Create, Read, Update, Delete). Therefore, to carry out this scalability process, an architecture and a technology analysis exercise were developed to enable the web application to perform these functions.

3. Corpus collection process

There are several options to collect a learner corpus. It can be collected as part of an academic activity in which all students participate, e.g., as an exam with its corresponding permission for data use. Another option is to ask students to volunteer their work if they are willing to participate. In this second option, attention must be paid not to introduce a bias considering that the most successful students would be more willing to participate than those with a low performance, which would compromise the balance and representativeness of the data.

Regardless of how a corpus is collected, texts in a learner corpus do not occur strictly in a natural way because they are produced in a classroom context and are the result of

activities designed to improve the learners' skills in the target language. In the present research, the output collected results from elicitation techniques that searched for the most natural output from students. The output resulted from questions that elicited students' information or opinions from current situations that affect their daily lives. Participating students were able to choose their own words to express their opinions in their compositions. The present research was based on the analysis of a written corpus from a cross-sectional study.

A written corpus can start with handwritten or typed texts. In the case of handwritten texts, the researcher must make sure the transcription is accurate; therefore, in typing, it is essential to trace the texts for any involuntary addition or loss of data. When all texts are collected, they should be coded, indicating a reference and information that make them traceable. Attention must be paid to quotations that do not belong to the learners' production. Guilquin (2015, p.19) recommends to "remove quotations (which do not represent the learner's own use of language and may therefore have to be excluded from the analysis of the corpus)." In the present work, quotations were not removed to keep the entire context from errors. In some cases, removing quotations would mean losing fundamental parts of the text indispensable to understand the context. On the contrary, they were kept, but close attention was paid to not analyze those parts. On the other hand, in the case of direct computerized versions of learners' texts, they can be kept in files as TXT texts to make sure they can be uploaded in the most appropriate software to conduct the tagging process.

The principles of learner corpora guided the collection of the present corpus (Pardo, 2020). These are some of the guidelines that should be taken into account when designing a corpus of learners, according to Granger, (2002), see Table 1.

Tabla 1. Guidelines for designing a learner corpus (Granger, 2002, p.9).

Learner	Task settings
Learning context	Time limit
Mother tongue	Use of reference tools
Other foreign languages learned	Type of test
Level of performance of English as a Foreign Language (EFL)	Audience / speaker
(The researcher could add other information that consider relevant)	(The researcher could add other information that consider relevant)

After having the institution's permission to carry out the research, several stages were needed to accomplish the collection process. Students did a placement test consisting of an online test supplied by Oxford University Press (Oxford University Press, 2017) and

available at www.oxfordenglishtesting.com. After a brief registration and the introduction of a password, the student starts a one-hour test of about 100 questions that the system sorts out with different degrees of difficulty to determine the student's language level. This test type guarantees that students are classified according to their performance following the Common European Framework of Reference for Languages (Europe, 2001).

In Table 2 it can be observed how the population of the present study was distributed. Participating students in this study were registered in different semesters from several BA programs offered by the university: Architecture, Basic Sciences, Health Sciences, Law, Politic Sciences, International Affairs, Business School, Humanities and Social Sciences, Engineering, Education Studies, and Mathematics. All participants share the same mother tongue: Spanish and their average age is 23.

Table 2. University classification according to CEFR (Pardo, 2019).

	Introductory Level	Level							
U. Norte Levels		1	2	3	4	5	6	7	8
CEFR	A1	A2	A2	B1	B1	B1	B2	B2	B2
Number of Students	110	496	439	409	325	356	377	335	286
				Pre-Intermediate	Interme-diate		Intermediate II	Upper-Intermediate	

After the files were collected, they were processed in different ways because they were submitted in different formats. For instance, and because their final work was handwritten, for level B1 the process started with the scanning followed by the texts' typing. External assistants did the typing of texts in their final year of their BA in languages at Universidad de Antioquia. They were given clear instructions regarding neither adding nor subtracting any words from the original handwritten compositions. After all texts were transcribed, they were thoroughly checked for mistakes and to make sure they were exactly as the original. Next, they were converted into TXT texts to do error annotation. Students from level B2 directly did the digital version; therefore, those texts were immediately converted into TXT format for the error tagger. The handwritten files were in total 373, and the process of typing lasted approximately seven months. After all the previous preparation, all files were ready to start annotation.

3.1. Error annotation process

As any other kind of corpora, learner corpora start as raw texts of electronic versions or transcribed texts from spoken learner output. Van Rooy (2015, p.79) mentions three advantages of using learner corpora to do research in language teaching: size, variability, and automation. *Size* refers to the amount of data that can be processed (computerized corpus allows analyses of great amounts of data). *Variability* refers to the possibility of having more individuals and more text types to include in a corpus. This advantage is also linked to the possibility of having a computerized corpus. Finally, *automation* refers to some automatic aspects of data analyses possible thanks to information technologies (IT).

Corpus annotation is "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Wynne, 2005, p.25). The added information comes in the form of tags, which can be defined as single entities added to one part or parts of the speech. Tags are unique and can identify features of the analyzed learner corpus. There are different types of annotation, and they require different tags depending on the goal of the researcher. For instance, descriptive linguistic uses Part of Speech (POS) tags to obtain grammatical annotation in a corpus. Another example is semantic annotation that requires assigning each word a semantic field used to do refined searches and classifications according to the research purpose. For error analysis, the annotation process is done to identify errors according to various categories and types.

To annotate errors, it is necessary to interpret learners' choices and decide in what category the error best fits. This entails the construction of one or several target hypotheses that the researcher must test. It is impossible not to interpret data. Only through interpretation, the researcher will find ways to unhide possible hypothesis to do an essential analysis. Assigning a tag to an error means that it was the researcher's interpretation, and that interpretation is publicly available for the reader. For that reason, when an error-tag is assigned, there could be other interpretations, but the most important is to keep uniformity in the way the tags are used. "The usefulness of error annotated corpora depends on the consistency on the annotation" (Ludeling & Hirschmann, 2015, p.148). Once the present learner corpus was annotated, it was easier to identify and extract data to analyse because the data was organized and ready to be used with software that permits further analyses.

For the present work, the learner corpus was tagged with a standardized error taxonomy that permitted the search and counting of errors analyzing within their context. The software used to extract error tags was WordSmith (Scott, 2005) and LancsBox. (Brezina *et al.*, 2015). WordSmith was used to obtaining the total statistics of errors, the dispersion,

and patterns that most affect the learner's production. LancsBox was used to obtain a more detailed profile of each error type and the corresponding graphics.

Regarding the annotation types in error analysis, there are two different types of annotation: *emendation* and *categorization* (Rosen *et al*., 2012). In the first case, the researcher establishes one or more target hypotheses and does the correction according to the author's intention. On the other hand, the categorization is done following a previous established list of errors, because error annotation relies on error taxonomies and their categories for error classification. In the present work, after choosing a target hypothesis the researcher did an error categorization, adding predefined tags according to the Manual of Error Tagging from Louvain University version 1.2 (Dagneaux *et al*., 2005). The corpus contained in the CLEC is a digital collection of 515 written files from English as a Foreign Language (EFL) university students registered in different careers. After the corpus was collected, the files were labeled. When an error was detected, the label was placed just before the error, and the correction followed the error between two-dollar signs: $ correction $ as the manual indicates:

Example:

Nowadays, we have seen (GADJN) differents $different$ (This error corresponds to the Grammar category and refers to the pluralization of an adjective (ADJN) in English).

The errors labeled and corrected in the CLEC are classified in the following eight categories that grouped a total of 56 error types. Please refer to appendix 1 of the present article to see the error types in detail.

- Form (F): groups the words used that do not exist in English and other errors of a formal type.
- Grammar (G): groups the errors that violate the general rules of English grammar.
- Lexical-grammar (X): errors where the morphosyntactic properties of a word are violated.
- Lexis (L): errors related to the semantic properties of words or sentences.
- Words (W): redundant words, missing words, or wrong word order.
- Punctuation (Q): errors related to punctuation marks.
- Style (S): incomplete sentences and unclear sentences.
- Infelicities (Z): registration problems (related to the field, the mode and the tenor of the speech) and issues of political correctness.

The next step after doing the error labeling was the extraction and alignment of the corpus. This process was carried out using an extraction software that searched for the labels and grouped them according to each error type. Tags were extracted within a context that granted proper analysis. The corpus's alignment was done using WordSmith, Scott, (2005) and LancsBox software, Brezina *et al.* (2015), which permitted the identification of language patterns obtaining statistics of the data with their respective graphs. After this process, the analysis of the findings took place.

3.2. Corpus metadata summary

The following are the main features of the corpus.

- Medium: written production
- Students belong to different university majors
- The EFL courses are 64 hours with an intensity of 4 hours per week for 16 weeks
- Native language of learners: Spanish
- Target language: English
- Genre of texts: there is a combination of genres between opinion paragraphs on different topics for level B1 and argumentative essays for level B2
- Tokens per text: at level B1 a maximum of 200, at level B2 up to 700
- Type: local corpus that seeks to identify needs and failures of learners
- Data compilation: it is a synchronous corpus with data collected in the second semester of 2015
- The incidence analysis was done by calculating the percentage of errors per 100 tokens to guarantee the proportionality of the analysis
- Corpus characteristics 149,325 tokens, 12,164 types and 12,337 lemmas

4. Methodology in the designing of the web application CLEC

After having the corpus collected and labeled with error tags, it was necessary to develop an application that systematically allowed the search of errors with the possibility to filter them according to different categories and types. It was also required that the app could allow changes in the error tags when they overlap among error categories. Therefore, a web application was developed with a frontend and a backend layer. After several tests, the functions of adding, modifying, or eliminating unnecessary data in the corpus were defined to be implemented. The development was possible thanks to a new technology where the frontend and backend responsibilities could be separated, and they were not

codependent. The alternative was a backend developed in Node.js (Dahl, 2009) together with Express.js (a web application framework for Node.js) for its construction as a REST API (Fielding, 2000) and a frontend in a JavaScript-based technology in which the options were React (Walke, 2013). It was decided to develop these technologies as they have excellent documentation and constant updates. Likewise, it was considered that the Node.js and React technologies have better support and a much broader community to guarantee a better response to the problems that arise throughout the development.

During the process, it was decided to use the persistence layer MongoDB (Merriman *et al*., 2007) database management system (DBMS), which is document oriented because it is consistent with the data of the corpus in the present study. This DBMS allows efficient access when making inquiries. The structure shown in Figure 1, allows to store the contexts after being processed. In this structure, it can be observed how the data is organized by level, name of file, context, error type, and its correction.

```
{
        level:,
        name:,
        context:,
        errors: [{type:,
                  error:,
                  correction:,
                  pos:
                 }
                ]
}
```

Figure 1. Document structure in MongoDB.

After defining the technologies to use, the development of the backend started by developing the methods for the search of errors. The additional services were defined and developed to enable the functions to create, read, modify, and delete contexts and create, read, and delete errors.

In this case, the method for modifying errors was left out as this meant an unnecessarily large load for processing due to the data's nature. Instead, it was decided to leave this functionality implicit as a combination of elimination and addition of errors. The database of contexts was populated with the help of preprocessing Python scripts that allowed structuring the data in the way it was previously defined. The new method of creating contexts included all this preprocessing that was required for new contexts.

In Figure 2, it is shown the architecture of the system described above.

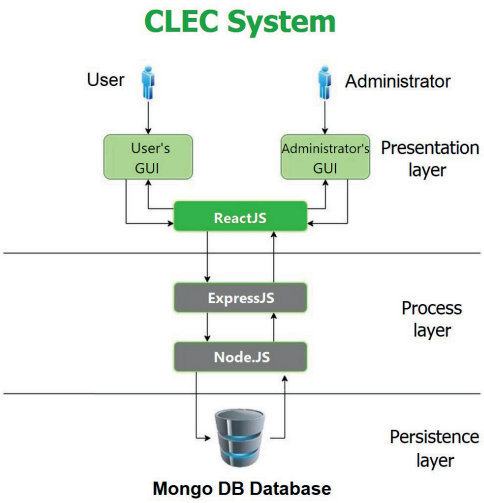


Figure 2. CLEC System Architecture.

As may be observed in Figure 2, the proposed system has two roles: administrator and user. The administrator can modify the application's data, whereas the user can only use the application. The most important use cases for both administrator and users are shown below in figure 3 and 4, respectively.

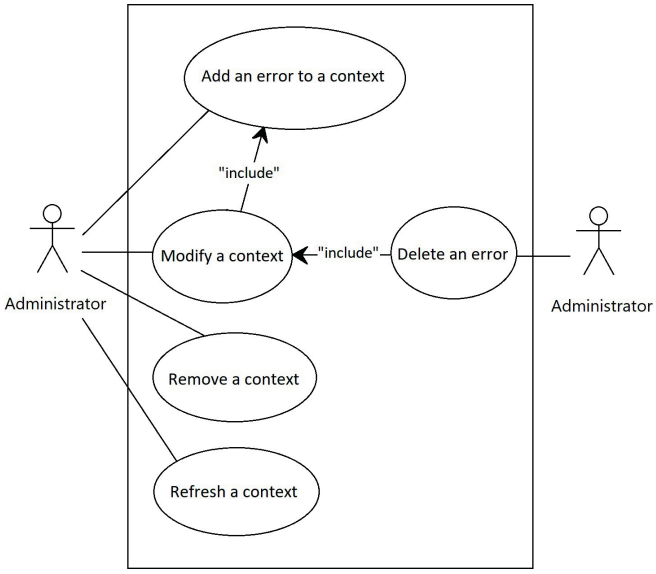


Figure 3. The administrator's use cases.

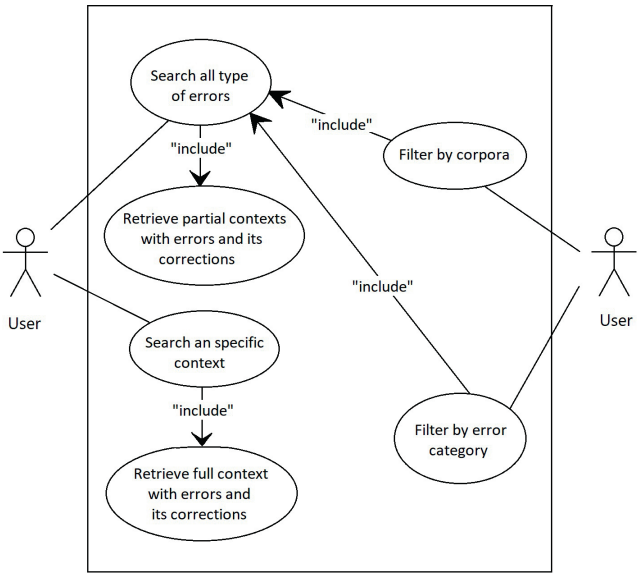


Figure 4. Use cases available for all users.

Each of the use cases depicted above will be illustrated below.

There were two ways to obtain the text contexts, one that displayed all the texts for a general view of different errors within their contexts, and one that obtained a specific text for a detailed view of each error within its context. Let us see the general view of different errors in Figure 5.

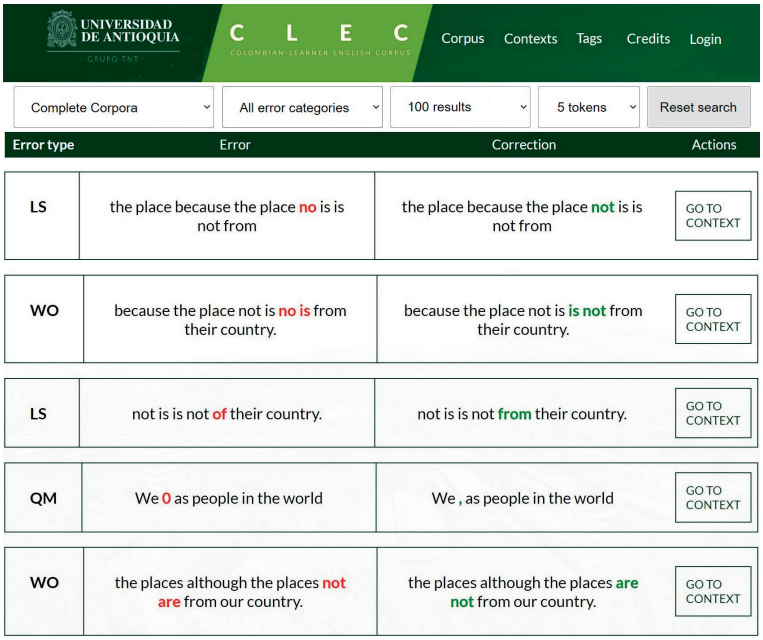


Figure 5. General view of different error types with their corrections (Pardo *et al*., 2018)

In Figure 5, for every sentence, it can be observed at the right side of the menu a button link that redirects the search to see each error's whole context. Clicking that button implies seeing the text's whole context that contains the error mentioned at the left side of the sentence. When you hit the button "go to context," you will see what is shown in Figure 6, the same error within the full context, and the correction in green.

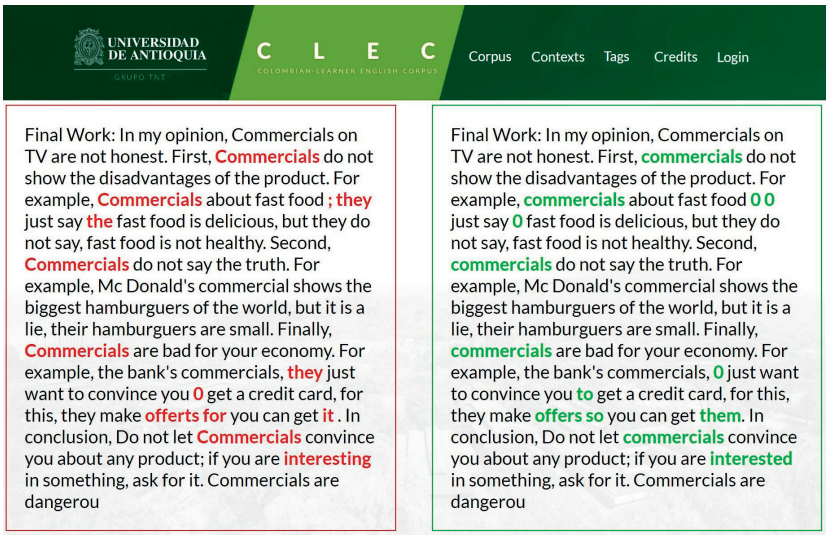


Figure 6. View of errors with full context and corrections (Pardo *et al.*, 2018).

Considering the nature of the data and these functionalities, the possibility of modifying contexts only to the parts of each text that did not contain errors was added. This was done in case the researcher wants to focus only on the text with errors. There were two methods to achieve this goal, one that creates lists of both context parts that contained and did not contain errors, and a second method that receives similar lists with the modifications made.

Similarly, the services corresponding to creating, reading, and eliminating errors were developed. All of them included verifications so that the rest of the errors did not enter conflict for their positions and/or for their content. For this part of the process, the service to modify errors was left out because it resulted in multiple cases in which some verifications of the data required excessive processing. This was replaced by a new possibility to modify errors by eliminating a previous error and adding a new one. It was an easier function, both for the development process and for the end-user.

Down, on the right side of Figure 7, 4 buttons allow changes in the corpus: add error, modify context, remove context, and refresh context.

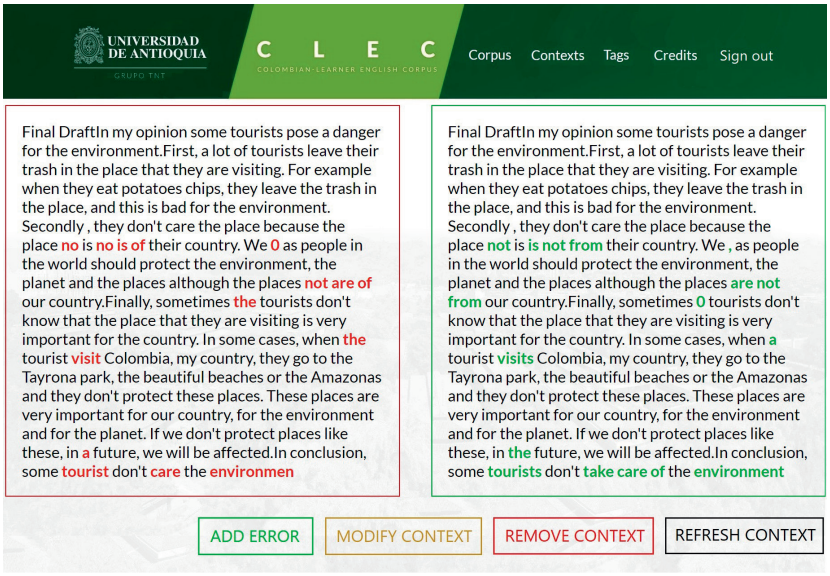


Figure 7. View of buttons to make modifications in the corpus.

These new functionalities are a plus in case there is need for a more detailed work in the corpus or to focus on specific parts of the texts.

A view of the search filters can be viewed in Figure 8. These filters were grouped by level: the corpus was divided into 4 levels of English A1, A2, B1, B2. They were arranged in an element of type selected:

- Basic (A1)
- Pre-intermediate (A2)
- Intermediate (B1)
- Advanced (B2)

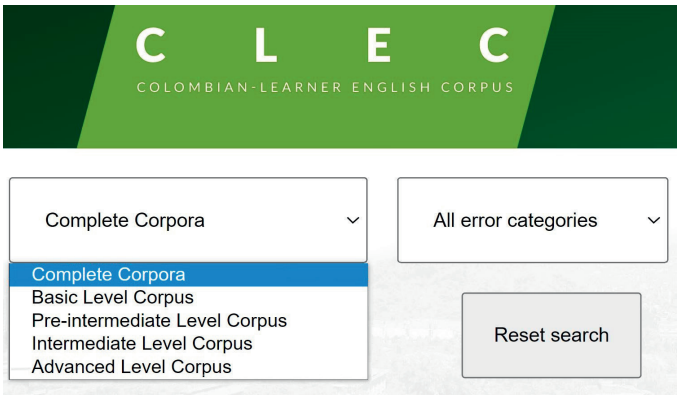


Figure 8. View of levels in the corpus.

In Figure 9, it can be noticed how the error types explained in the corpus collection section of this article were arranged as an element of type select.

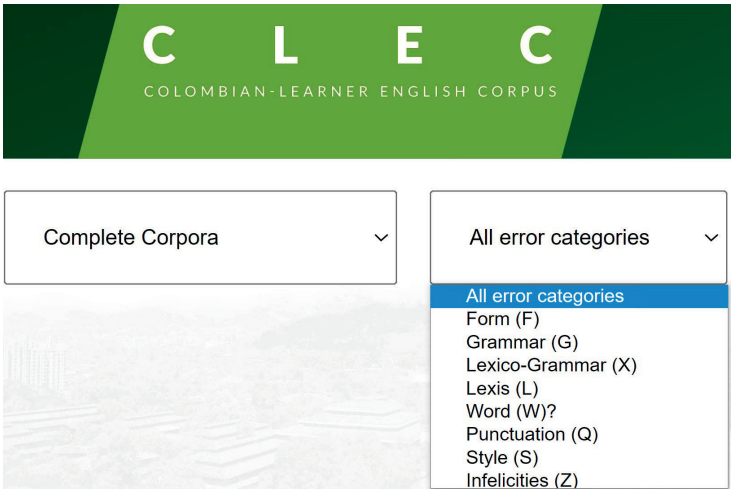


Figure 9. View of error categories (Pardo *et al.*, 2018).

In Figure 10, it may be noted how a condition was created so that check boxes with the corresponding class error types would be displayed when the selection was changed. In all this process, it can be noted how the system's graphic design was created, selecting the university's institutional colors (dark and light green).

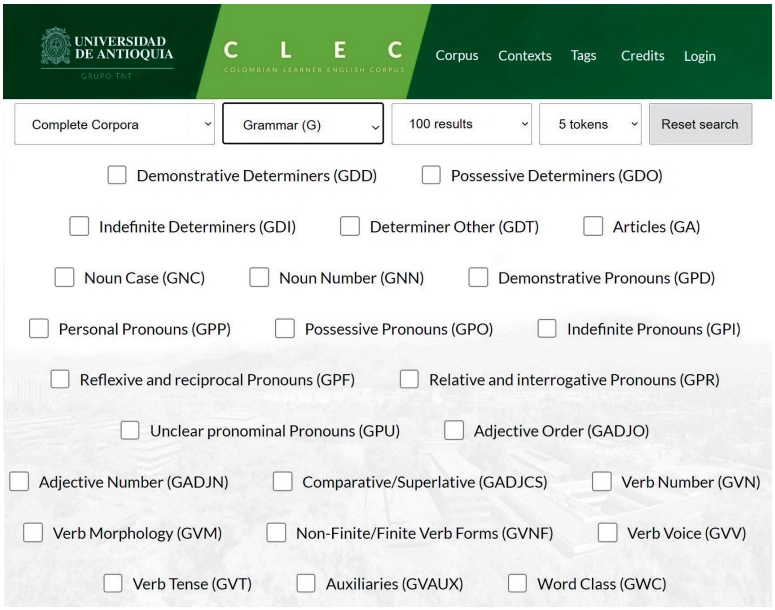


Figure 10. Check boxes to choose error types to analyze in the Grammar category.

In this case, Figure 10 shows error types from the grammar category, but if the category changes, the error types will correspond to the chosen category.

In Figure 11, it is possible to observe errors within the context of one sentence. The errors are in red and in front of the whole text with the corrections in green.

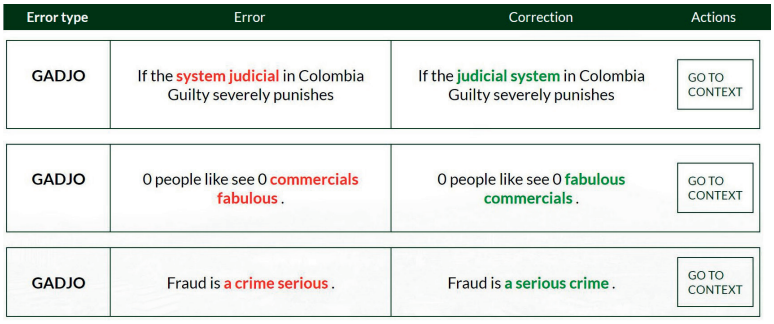


Figure 11. View of errors within a small context.

The same errors can be viewed in the whole context when hitting the button "go to context." In Figure 12, we may note the view of the whole context for one of the errors.

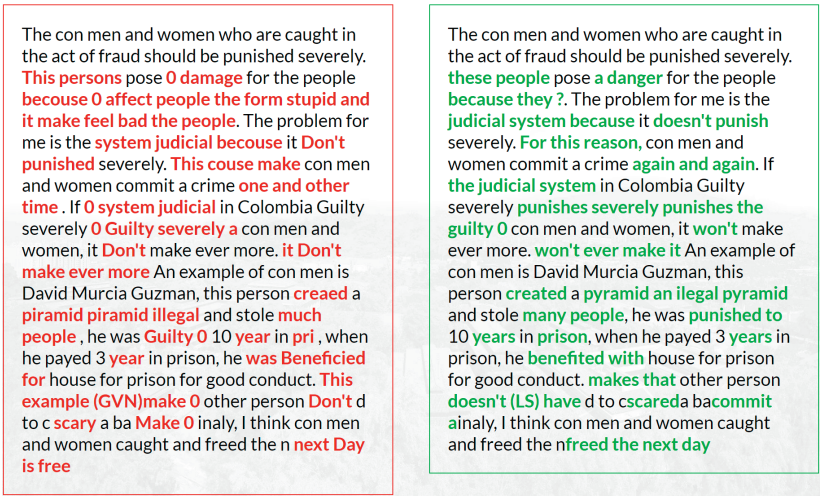


Figure 12. View of errors in one file.

It is necessary to clarify that the view of errors in Figure 12 shows all the different errors the student made in his composition, for that reason, there are several categories and types of errors.

All the previous functionalities were oriented for the use of all users, including unauthenticated ones. For authenticated users (administrator role), additional components were made available for the other functionalities, including a button, in the context view, for each error that would allow the possibility to eliminate them if necessary. Let us see the detail in Figure 13.

Figure 13. View of the button to delete errors (Pardo *et al.*, 2018).

Besides, a set of buttons were included at the bottom of the whole contexts, and the buttons are: Add, Modify, Remove and Refresh. By displaying a pop-up window, the user selects

the context section on which he/she wants to introduce a modification. The same process is followed for each case. There is another button to remove the context and the last button to refresh the context with the changes made. Let us see Figure 14.

Figure 14. View of full contexts and buttons to add, modify and remove data (Pardo *et al.*, 2018)

5. Results

From the previous process, the result was a web responsive application that completely performs searches and does analysis on the tagged corpus of errors. This app contains a learner corpus of English as a Foreign Language (EFL) learners that has the potential of being easily revised and expanded through the role of the system administrator. This new functionality will be very useful to enrich the system that can be used by linguists, teachers, and students who may consider it to do research. This corpus is available in the given URL

and is open to any researcher if you want to consult it or if you want to contribute with learner corpora[2].

The development of the backend as a REST API allowed the tests to be carried out independently of the frontend, allowing future developers to use this API for new versions or refactoring of the frontend.

Regarding the front end, it was also possible to deliver a design that is very aesthetic and friendly. This will allow that existing method and those that would be open to the public were simplified and more understandable for use.

Finally, the web application was deployed on the Translation and New Technologies (TNT) research groups of Universidad de Antioquia server. The Colombian Learner English Corpus (CLEC) is available online at: https://grupotnt.udea.edu.co/clec.

5.1. Graphical view of errors

The findings of errors in the corpus were grouped by category and type. Figure 15 shows a view of errors by category.

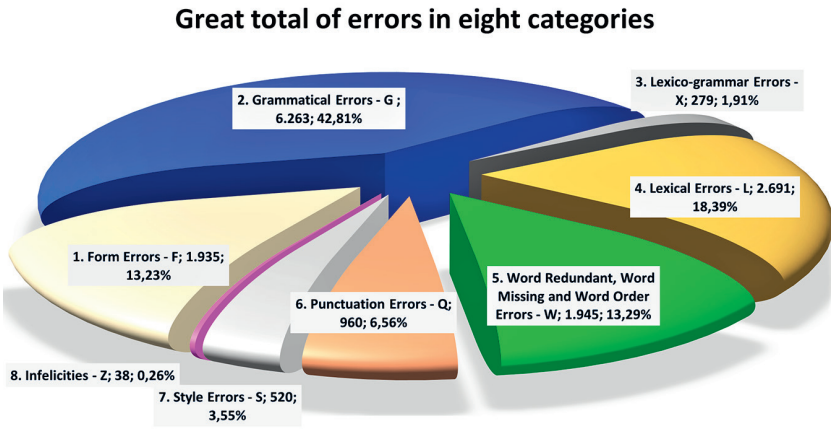


Figure 15. Incidence of errors by category (Pardo, 2019).

It is clear in figure 15 that the category of errors with most frequency in the corpus was Grammar. A more detailed view of errors is displayed by type in Figure 16.

2 If you want to contribute with data to this project, please contact the authors.

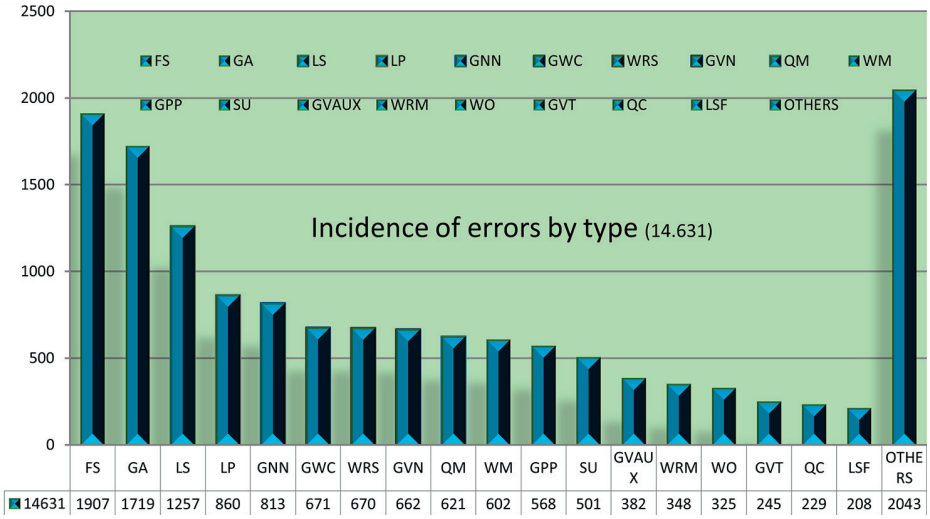


Figure 16. Incidence of errors by type (Pardo, 2019).

In this case, the frequency by type can give us an idea of the frequency of each type of error. All this information can be easily retrieved for its analysis using the CLEC app.

6. Conclusions

This work presented the CLEC app, the first corpus of written production of Colombian students learning English as a Foreign Language available online for the research community. CLEC works with a modern technology that offers agile maintenance options and allows a user interface design that is friendly and allows a satisfying interaction with the app.

Similarly, it was possible to achieve the construction of a complete, friendly, and safe administration system to manage the data of the treated corpus allowing its scalability and maintenance to create, read, edit, and eliminate contexts. These functions give the application an invaluable utility for didactic and research matters.

There were several advantages brought with the technologies used in this project. Using React, future development teams will be able to take over the project and add new functionalities.

Despite the complexity of the structure in which the contexts and errors were handled, it was possible to reduce the complexity of the entire process for the end-user through the correct planning of the development and the views. Now it is an interface that allows the use of its features in a practical way.

Finally, this work gives the academic community an invaluable free access web application, which facilitates the teaching-learning process of English as a foreign language through an efficient and friendly error analysis.

Acknowledgements

Thanks to Universidad del Norte for allowing the collection of the data.

We would like to acknowledge Manuel Gómez and Nicolás Henao for their participation in the design of the CLEC app.

The research reported here was supported by a COLCIENCIAS scholarship.

Appendix

1. Error categories and types according to the manual of Louvain University

FM	**F**orm, **M**orphology
FS	**F**orm, **S**pelling
FSR	**F**orm, **S**pelling, **R**egional
GDD	**G**rammar, **D**eterminer, **D**emonstrative
GDO	**G**rammar, **D**eterminer, P**O**ssessive
GDI	**G**rammar, **D**eterminer, **I**ndefinite
GDT	**G**rammar, **D**eterminer, O**T**her
GA	**G**rammar, **A**rticles
GADJCS	**G**rammar, **Adj**ectives, **C**omparative / **S**uperlative
GADJN	**G**rammar, **Adj**ectives, **N**umber
GADJO	**G**rammar, **Adj**ectives, **O**rder
GADVO	**G**rammar, **Adj**erbs, **O**rder
GNC	**G**rammar, **N**ouns, **C**ase
GNN	**G**rammar, **N**ouns, **N**umber
GPD	**G**rammar, **P**ronouns, **D**emonstrative
GPP	**G**rammar, **P**ronoun, **P**ersonal
GPO	**G**rammar, **P**ronoun, P**O**ssessive
GPI	**G**rammar, **P**ronoun, **I**ndefinite
GPF	**G**rammar, **P**ronoun, Re**F**lexive/Reciprocal
GPR	**G**rammar, **P**ronoun, **R**elative/ Interrogative
GPU	**G**rammar, **P**ronoun, **U**nclear reference
GVAUX	**G**rammar, **V**erbs, **Aux**iliaries
GVM	**G**rammar, **V**erbs, **M**orphology
GVN	**G**rammar, **V**erbs, **N**umber
GVNF	**G**rammar, **V**erbs, **N**on-Finite / **F**inite
GVT	**G**rammar, **V**erbs, **T**ense
GVV	**G**rammar, **V**erbs, **V**oice
GWC	**G**rammar, **W**ord **C**lass

LCC	**L**exis, **C**onjunctions, **C**oordinating
LCLC	**L**exis, **C**onnectors, **L**ogical, **C**omplex
LCLS	**L**exis, **C**onnectors, **L**ogical, **S**ingle
LCS	**L**exis, **C**onjunctions, **S**ubordinating
LP	**L**exical **P**hrase
LPF	**L**exical **P**hrase, **F**alse friends
LS	**L**exical **S**ingle
LSF	**L**exical **S**ingle, **F**alse friends
QC	Punctuation, **C**onfusion
QL	Punctuation, **L**exical
QM	Punctuation, **M**issing
QR	Punctuation, **R**edundant
SI	**S**entence, **I**ncomplete
SU	**S**entence, **U**nclear
WM	**W**ord **M**issing
WO	**W**ord **O**rder
WRS	**W**ord **R**edundant **S**ingle
WRM	**W**ord **R**edudant **M**ultiple
XADJCO	Le**X**ico-Grammar, **Adj**ectives, **Co**mplementation
XADJPR	Le**X**ico-Grammar, **Adj**ectives, Dependent **Pr**eposition
XCONJCO	Le**X**ico-Grammar, **Conj**unctions, **Co**mplementation
XNCO	Le**X**ico-Grammar, **N**ouns, **Co**mplementation
XNPR	Le**X**ico-Grammar, **N**ouns, Dependent **Pr**eposition
XNUC	Le**X**ico-Grammar, **N**ouns, **U**ncountable / **C**ountable
XPRCO	Le**X**ico-Grammar, **PR**epositions, **Co**mplementation
XVCO	Le**X**ico-Grammar, **V**erbs, **Co**mplementation
XVPR	Le**X**ico-Grammar, **V**erbs, Dependent **Pr**eposition
Z	Infelicities

References

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. https://doi.org/10.1075/ijcl.20.2.01bre

Carlstrom, B., & Price, N. (2012). *The Gachon Learner Corpus*. Retrieved from https://app.box.com/s/erq3w1d7v711fq5ze76kzt56lomk3c06

Corder, S. (1967). The significance of learner's errors. *IRAL - International Review of Applied Linguistics in Language Teaching*, *5*(1-4), 161-170. https://doi.org/10.1515/iral.1967.5.1-4.161

Corder, S. (1981). Error Analysis and Interlanguage. *Oxford University Press*.

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.

Dahl, R. (2009). *NODE.JS.* Open JS Foundation. https://nodejs.org/es/docs/

Diaz-Negrillo, A. (2009). *EARS: a User's Manual.* Lincom Academic Reference.

Díaz-Negrillo, A. (2012). Learner corpora: the case of the NOSE corpus. *Journal of Systemics, Cybernetics and Informatics*, *10*(1), 42-47. https://www.iiisci.org/journal/pdv/sci/pdfs/HEB467AV.pdf

Europe, C. of. (2001). The Common European Framework of Reference for Languages: Learning, teaching, assessment. *Common European Framework*. https://doi.org/10.1093/elt/cci105

Fielding, R. (2000). Architectural Styles and the Design of Network-based Software Architectures [Doctoral dissertation, University of California, Irvine]. Donald Bren School of Information and Computer Sciences. https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

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. https://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 Foreign Language Teaching* (pp. 3-33). John Benjamins Publishing Company.

Granger, S. (2003). The International Corpus of Learner English : A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, *37*(3), 538-546.

James, M. (1980). *Managing the database environment*. Savant Research.

Ludeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135-157). Cambridge University Press.

McEnery, A., & Hardie, A. (2011). *Corpus Linguistics: Method, theory and practice*. Cambridge University Press.

Mendikoetxea, A., O'Donnell, M., & Rollinson, P. (2009). *WriCLE: A learner corpus for Second Language Acquisition Research*. 2010. http://ucrel.lancs.ac.uk/publications/cl2009/351_FullPaper.doc

Merriman, D., Horowit, E., & Ryan, K. (2007). *MongoDB Documentation*. https://docs.mongodb.com/

Pardo, M. (2019). *Error Analysis in a Written Corpus of Spanish Speakers EFL Learners. A Corpus-based Study.* Universidad de Antioquia.

Pardo, M., Quiroz, G., Tamayo, A., Henao, N., Ortega, M., & . (2018). *CLEC Colombian Learner English Corpus*. https://grupotnt.udea.edu.co/clec/corpu

Rosen, A., Jirka, H., Stindlová, B., Feldman, A., & Svatava, S. (2012). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, *48*(1), 65-92. https://doi.org/10.1007/s10579-013-9226-3

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

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(1–4), 209-232.

UC Louvain. (2018). *Centre for English Corpus Linguistics*. Learner Corpora Around the World. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

University, T. S. (2002). *The Santiago University Learner of English Corpus (SULEC)*. https://sulec.cesga.es/

Van Rooy, B. (2015). Annotating learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 79105). Cambridge University Press.

Walke, J. (2013). *React. Una biblioteca de JavaScript para construir interfaces de usuario.* React. https://es.reactjs.org/

Wynne, M. (Ed.) (2005). *Developing linguistic corpora: a guide to good practice*. Oxbow Books.

Part III
Corpus analysis and Natural Language Processing

Chapter XI

Pronunciation of consonant clusters in Spanish speakers based on the Czech read speech corpora

La pronunciación de los grupos de consonantes en hispanohablantes basándose en el corpus oral leído checo

Kateřina Pugachova & Jitka Veroňková
Faculty of Arts, Charles University – Czech Republic

Abstract: The purpose of this study was to determine which Czech consonant clusters are difficult to pronounce for Spanish speakers, and identify the sound changes that are more frequent due to the difference in syllable structure between these two languages. A set of 26 consonant clusters in initial, medial and final word positions was selected. The 75 words containing the target clusters were included in a coherent text written in Czech (838 words long). Then, the speech of 13 Spanish speakers reading this text was recorded. Based on perceptual analysis, 27% of clusters were pronounced incorrectly. The number of correct items among the cluster types and within the types varied considerably. Substitution, elision and prothesis represented almost 90% of all the sound changes. Substitution, being the most dominant, affected all studied consonant cluster types.

Resumen: El propósito de este estudio fue determinar qué grupo de consonantes checas son difíciles de pronunciar para los hispanohablantes e identificar los cambios de sonido que son más frecuentes, debido a la diferencia en la estructura de las sílabas entre estos dos idiomas. Se seleccionó un conjunto de 26 grupos de consonantes en las posiciones inicial, media y final de la palabra. Las 75 palabras que contenían

los grupos de consonantes estudiados se incluyeron en un texto coherente escrito en checo (con 838 palabras en total). Luego, se grabó el discurso de 13 hispanohablantes que leían este texto y se realizó un análisis perceptivo. El 27% de los grupos de consonantes se han pronunciado incorrectamente. El número de grupos de consonantes pronunciados correctamente varió mucho entre los tipos de agrupaciones e incluso dentro de las propias agrupaciones. La sustitución, elisión y prótesis representaron casi el 90% de todos los cambios de sonido. La sustitución, que fue la más dominante, afectó a todos los tipos de grupos de consonantes estudiados.

1. Introduction

In recent years, the Czech Republic has been hosted to an increasing number of Latin American and Spanish people who usually work or carry out their studies at universities. Smaller Spanish-speaking groups regularly take part in summer Czech language schools for foreigners or attend preparatory courses aimed at potential international students at Czech universities.

However, there are a limited number of textbooks for Spanish speakers on the market. Available materials are often a translated version or an older edition. Regarding the Czech language of Spanish speakers, rather informal observations of teachers are available, but systematic data-based research has not been carried out.

Our experiment aims to contribute to the research of sound aspects of Czech in Spanish speakers. It focuses on one of the difficult areas, i.e., the pronunciation of consonant clusters. Perception analysis is based on recordings of Czech read speech in speakers with Spanish as a first language.

The difficulties of Spanish learners with the pronunciation of consonant clusters or consonants in the positions restricted in Spanish have been mainly evidenced by studies on the acquisition of English. Based on the review of literature, Moore and Marzano (1979) presented a list of possible errors of Spanish students learning English, including consonants and their clusters. Based on Helman (2004), some of these are possible adaptations of unfamiliar English consonant endings, the simplifying of a consonant cluster by deleting a consonant, substituting to create an ending permissible in Spanish or a change leading to a vowel ending. According to Magen (1998), initial schwa inserted by Spanish speakers in English syllable onsets formed by fricative + stop clusters and deleting of final /s/ belonged among factors listeners were sensitive to when they rated the extent of foreign accent. The application of Spanish phonological and orthographic rules on English is recog-

nized from spelling in written texts as well. (Fashola *et al.*, 1996; Sun-Alperin and Wang, 2008; Hevia-Tuero *et al.*, 2021)

The difficulties L2 learners may encounter are not only due to the influence of the L1 features on the target language. Piske (2001) provides an overview of factors that may affect the acquisition of L2 including pronunciation, e.g., the length of stay in the target country and the use of language, gender or existing or lacking formal instructions; the existence of the so-called critical period is widely discussed (comp. also Singleton, 2005; Rothman, 2008). Individual differences among learners might be caused, for example, by the cognitive and learning styles, language aptitude, motivation and personality (Ellis, 1985, pp. 639-723; Hummel, 2014, pp. 193-222). Regarding our speakers, we were mainly interested in circumstances related to staying in the Czech Republic, studying Czech and using Czech in daily communication; however, our research is not focused on examining the influence of any certain factor.

2. Theoretical framework

Sound characteristics distinguishing Czech and Spanish include syllabic structure and consonant clusters. The primary difference lies in the number of consonants within a single syllable, their frequency, and phoneme combinatory aspects including constraints in specified positions. In Czech, for example, some sonorants (mainly /l/ and /r/) may form a syllabic nucleus, unlike in Spanish.

Czech and Spanish syllables tend to be open. In both languages, the predominant syllable type is the CV type, which occurs in 59.76% in Czech (Těšitelová *et al.*, 1985, p.149) and in 55.81% in Spanish (Guerra, 1983, as cited in Quilis, 1993, p.370). However, a significant difference is the number of consonants within one syllable. In Spanish, onset and coda are usually formed by one, rarely two consonants, and thus the CCCV syllable type, for instance, containing three consonants in onset, is not present in Spanish, unlike in Czech in which it has a frequency of occurrence of 0.72% (Těšitelová *et al.*, 1985, p.149). CCVCC is the longest Spanish syllable type – occurrence of 0.01% (Guerra, 1983, as cited in Quilis, 1993, p.370), the same syllable type in Czech occurs with the higher frequency of 0.26% (Těšitelová *et al.*, 1985, p.149). Based on the analysed texts, the longest Czech syllable type is CCCVCC (ibid; Kučera & Monroe, 1968, p.47) with frequency of 0.08% (Těšitelová *et al.*, 1985, p.149); however, it is possible to find samples even for types with longer consonant sequences (Bičan, 2013, p.122) and the number of consonants in the onset may increase by including a non-syllable preposition.

In Czech, there are no such restrictions for one-segment or multi-segment onset and coda, as in Spanish. (Ludvíková & Kraus, 1966; Kučera & Monroe, 1968; Bičan, 2013) In the Spanish CC-onset in the initial word position there can be only combinations of obstruent and sonorant, namely 12 clusters /pr, br, fr, tr, dr, kr, gr, pl, bl, fl, kl, gl/ (Saporta & Olson, 1958, p.263; Quilis, 1993, p.381; Ríos Mestre, 1999, section 6.2.2.2.) and /tl/ in words of Náhuatl origin (Quilis, 1993, p.381; RAE, 2011, p.302-303). The loanwords containing initial /s/ followed by another consonant are adapted by a prothetic vowel, e.g., *escena* (RAE, 2011, p.305). In loanwoards, e.g., from Latin or Greek, other consonant groups such as *cn-, gn-, mn-, pt-* and *ps-* may occur in the initial position of the word. However, in Spanish, the groups remain preserved only in written form, the pronunciation is simplified (the first consonant is elided). Simplified forms appear even in written form as parallel variants, e.g., *gnomo – nomo, psíquico – síquico, ptolemaico – tolemaico* (RAE, 2011, p.304-305; RAE, 2021).

For the Spanish coda -C at the end of a word, studies present a limited set of phonemes as well. It is the loanwords that are the source of new codas including -CC in the word final position, otherwise unusual in Spanish (Saporta & Olson, 1958, p.266), e.g., *golf* or *vals* (RAE, 2011, p.315). However, there is a tendency towards simplification in pronunciation too. Parallel variants may occur, e.g., *cinc*/*zinc* is pronounced both with a full coda or without a final consonant, or only simplified pronunciation is used, e.g., *robots* with elision of /*t*/. (RAE, 2011, p.315-317).

The sequence of consonants may be increased by the contact of a coda and an onset in the medial position of a word. In Spanish, changes occur in those cases as well. For example, in the combination *bs* + consonant, /b/ is usually weakened or skipped. According to RAE (2011, p.320-321), nowadays it is possible to omit *b* not only in pronunciation but even in writing and the simplified spelling is primary; comp. e.g., *oscuro – obscuro, sustantivo – substantivo*, *sustituir – substituir* (RAE, 2021). The cause is mainly the syllable boundary. Unlike in Czech, where the position of the syllable boundary may vary to some extent (Palková, 1997; Šturm, 2018), in Spanish there are precise rules governing this process; the main rule is the permission or restriction of a fixed combination of sounds within a syllable. (Quilis, 1993, pp. 368-370; Ríos Mestre, 1999, section 6.2.3.) For example, the 12 clusters defined for the initial position of a word (see above) cannot be split within a word (Quilis & Fernández, 1979, p.140).

3. Methodological framework

3.1. Target consonant cluster set

In the first step, we determined a set of target consonant clusters. Since the aim was not to test the pronunciation of individual segments, but consonant groups as a whole, the condition was determined that consonants absent in the Spanish language would not be included in the consonant cluster set used for this research. Otherwise, any potential difficulties of speakers might be primarily related to the pronunciation of that segment, not to the combination of the given cluster as a whole. For example, clusters with a specific Czech vibrant fricative /r̝/ or with a laryngeal consonant [ɦ] (in Czech, unlike most languages, voiced), none of which have equivalent in Spanish, were not tested.

The starting point was a set of consonant clusters occurring in Spanish. Based on Quilis (1993), RAE (2011), and Čermák (2015), those consonant clusters were selected, whose pronunciation may differ between Czech and Spanish or those that may present difficulties for L2 Czech speakers with Spanish as L1 because of position restriction etc. Due to a large number of such clusters, another selection procedure followed. The set was limited to two-component clusters with an initial consonant [s], with an initial consonant [p], namely [pt], [ps], [pn], and the cluster [gn]. Three-component clusters [pst] and [psk] were also included. Those clusters were then systematically supplemented based on Czech language, e.g., by combinations containing voiced/voiceless counterparts.

In the S + consonant type, we tested all two-member combinations existing in Czech, the first member of which is the consonant [s] (with the exception of less common or problematic combinations such as [sf] or [st͡ʃ]). Those items were [s] + voiceless stops [p], [t], [c], [k], fricative [v], nasals [m], [n], [ɲ] and oral sonorants [l], [r], [j].

Due to the use of the nasal palatal [ɲ] in conjunction with [s], we decided to test the combination of the nasal [ɲ] with other initial consonants already used, i.e., the cluster [pɲ] and [gɲ] were added.

Due to the fact that in Czech the voicing opposition plays an important role, four more clusters [bn], [bɲ], and [kn], [kɲ] were added as voiced and unvoiced equivalents to the existing clusters [pn], [pɲ], and [gn], [gɲ]. In these nasal clusters, the voicing property of obstruents should be preserved.

Altogether, 23 clusters divided in 6 types were included in the experiment (see Table 1).

Table 1. Set of consonant cluster types.

2-consonant clusters	• consonant [s] combined with defined unvoiced obstruents, sonorants and [v] (S+cons) • [ps] • [pt] • obstruent bilabials [p], [b] and velars [k], [g], each combined with nasals [n] and [ɲ] (O+nas)
3-consonant clusters	• [pst] • [psk]

Note: In the following text, capital letters, i.e. [ps] PS are used, and palatals [ɲ] and [c] are written as Ň and Ť.

3.2. Target words set

A set of words containing the observed consonant clusters was created. For each consonant cluster, the position in the word selected for the test was established: initial – I, medial – M and final – F. The purpose of the experiment and the ideal number of tested units were taken into account.

In the S+cons type, we focused on the initial position, because that is where Spanish native speakers use a prothetic vowel, which is a significant difference compared to Czech. The originally determined nasal clusters PN and GN were tested in I and M positions. The groups with voicing counterparts and palatal [ɲ] were tested only in M position. For other types PS, PT and PST, PSK, an attempt to find a representative for all three positions was made.

The Index Database (Databáze heslářů) was used for searching suitable words. It contains over 900,000 entries from 14 Czech written sources with items from both older dictionaries and new vocabulary occurring in newspapers or magazines. In the process of creating the word sets, it was found that we could not always fill a defined I / M / F position. The PST, for example, appeared only in positions M and F. For some clusters, although lexemes were available, their occurrence was either restricted to scientific terminology, or very limited in general frequency. For that reason, the GŇ cluster was eventually excluded from the test. Regarding the type and position, 31 subgroups were defined.

To ensure that any errors would be a matter of personal pronunciation and not a case of ignorance of orthoepic rules, in S+cons, only words in which the graphic form and pronunciation of the target cluster did not differ due to voicing assimilation, as in the word *zkoušky* [sko͡uʃkɪ] (En. exams, Sp. exámenes), were tested eventually. The need to perform voicing assimilation occurs in our set in less frequent groups: a) in all five representatives of PST, in

graphic form of *bst*, e.g., *obstarávat* (En. to procure, Sp. procurar), b) once in PT in the M position (*drobty* (En. crumbs, Sp. pizcas) vs. *poptávka* (En. demand, Sp. demanda).

Table 2. Set of consonant clusters regarding word position with examples.
CC – consonant cluster, IMF – position in a word: I – initial, M – medial, F – final, N – number of words per cluster, Ť – [c], Ň – [ɲ], fem. – femininum, n. – noun, sust. – sustantivo. B – underlined – a graphic form of a consonant cluster does not correspond to the pronunciation, + the form of a Czech example is not a nominative case.

CC	IMF	N	Example	Pronunciation	In English	In Spanish
SP	I	2	**sp**ekulace	[**sp**ɛkulat͡sɛ]	speculation	especulación
ST	I	2	**st**udentka	[**st**udɛntka]	student (fem.)	la estudiante
SŤ	I	2	**st**ěží	[**sc**ɛʒiː]	hardly	apenas
SK	I	2	**sk**upina	[**sk**upɪna]	group	grupo
SV	I	2	**sv**obodu	[**sv**obodu]	liberty+	libertad+
SM	I	2	**sm**utná	[**sm**utnaː]	sad (fem.)	triste (fem.)
SN	I	2	**sn**ad	[**sn**at]	perhaps	quizas
SŇ	I	2	**sň**atek	[**sɲ**atɛk]	marriage	matrimonio
SL	I	2	**sl**unce	[**sl**unt͡sɛ]	sun	sol
SR	I	2	**sr**azila	[**sr**azɪla]	(she) crashed	chocó (fem.)
SJ	I	2	**sj**ezdu	[**sj**ɛzdu]	exit (n.)+	salida (sust.)+
PS	I	6	**ps**ala	[**ps**ala]	(she) wrote	escribió (fem.)
			psychologie	[**ps**ɪxologɪjɛ]	psychology	psicología
PS	M	7	na**ps**aly	[na**ps**alɪ]	(they) wrote (fem.)	escribieron (fem.)
			ka**ps**ičky	[ka**ps**ɪt͡ʃkɪ]	pockets	bolsillos
PS	F	2	kola**ps**	[kola**ps**]	collapse	colapso
PSK	M	3	Li**psk**a	[lɪ**psk**a]	Leipzig+	Leipzig+
PST	M	4	su**bst**anci	[su**pst**antsɪ͡]	substance+	sustancia+
	F	1	zá**bst**	[zaː**pst**]	to freeze	tener frío
PT	I	3	**pt**át	[**pt**aːt]	to ask	preguntar
	M	3	konce**pt**y	[kont͡sɛ**pt**ɪ]	concepts	conceptos
	F	3	rece**pt**	[rɛt͡sɛ**pt**]	recipe	receta
PN	I	3	**pn**ula	[**pn**ula]	twined (fem.)	se enroscó (fem.)
	M	3	oslepne	[oslɛ**pn**ɛ]	(it) will go blind	se quedará ciego
PŇ	M	3	tra**pně**	[tra**pɲ**ɛ]	embarrassingly	embazarosamente
BN	M	2	dro**bn**é	[dro**bn**ɛː]	change (n.)	cambio (sust.)
BŇ	M	2	bezchy**bně**	[besxɪ**bɲ**ɛ]	flawless	sin falta
GN	I	1	**gn**ómon	[**gn**oːmon]	gnomon	gnomon
	M	3	i**gn**orovat	[ɪ**gn**orovat]	to ignore	ignorar
KN	M	2	pě**kn**ou	[pjɛ**kn**oʊ̯͡]	beautiful (fem.)+	bella+
KŇ	M	2	baro**kn**í	[baro**kɲ**iː]	baroque	barocco

A list of words containing the selected clusters in defined positions was created. We assumed that a coherent text would be a better disguise for the target phenomenon and that a story would be easier to read than, say, single sentences without wider context. In order to examine as many items as possible while avoiding excessive text length, the following numbers of words were used: a) two words for each S+cons cluster, b) regarding PS, six clusters in I and 7 in M (and two in F) to obtain more items for comparison, c) for remaining clusters, an average of 2–3 words per cluster and position. The set of words examined also depended on the number of suitable candidates. In cases where the number of words of a certain type of cluster was insufficient in any of the I, M, F positions, we tried to increase the representation of the cluster in another position, e.g., the PST cluster was represented only by one word in F, but 4x in M. Where possible, a loanword was used for the given cluster and the position. Each word contained just one target consonant cluster, with the exception of two words – *skeptiků* (En. sceptics, gen., Sp. escépticos, gen.), *skepse* (En. scepticism, Sp. escepticismo) containing two examined consonant clusters. Table 2 presents the set of defined clusters according to their position and the samples of target words. A total of 73 different words (containing 75 target consonant clusters) were selected: 47 % words in I, 45 % words in M and 8 % in F. The most numerous were disyllabic (40.0%) and trisyllabic words (30.7%), then 4-syllabic (12, 16.0%). Monosyllables were represented by seven words and 5- and 6-syllabic items were attested in three cases altogether. A text – story (838 words long) was created. In order to prevent the spread of a consonant cluster across a word boundary, the I-cluster was preceded by a vowel, and a vowel followed the F-cluster, or it was assumed that a pause would be realized.

3.3. Speakers

The group of participants consisted of 13 speakers with Spanish as L1 who were either from the first author's circle of acquaintances or responded to requests on social media, through which the community of foreigners living in Prague was addressed. Women showed significantly less interest, which resulted in groups not being balanced by sex: 10 males and 3 females were eventually available for the experiment. There were 9 Latin Americans from six different countries and 3 Spanish, each coming from different cities in Spain. The length of stay of speakers in the Czech Republic (CR) ranged from 1.5 years to 9.5 years, for most speakers it was a continuous stay. Five speakers completed a one-year preparatory course in Czech, then they studied in the CR at technical universities. One speaker stated the study of Czech lasted 1.3 years. For other speakers, the study of Czech was shorter – from two weeks to six months, with the characteristic that those studies took place several years

ago, and in two cases it was self-study; the speaker declaring two-week study had lived in the CR for 1.5 year. Speakers also differed in the degree of use of Czech or the intensity of contact with the Czech environment – some speakers used Czech at work or in communication with their family or friends, while others did not use Czech in their daily life at all. With some exceptions, however, all indicated English as their primary language for communication. There was one more speaker, who might be considered bilingual. His father was from Peru and his mother was Czech. This speaker had a Czech and Spanish high school diploma and at the time of recording he was currently studying at a Czech university. According to his words, however, he started speaking Czech at a preschool age and he had not always felt confident in Czech in some respects. Throughout his life, he had been alternating between both Czech-dominated and Spanish-dominated environments. All speakers interested in participating were recorded including the bilingual one as his speech showed similar features to the rest of the speakers (see Table 3).

Table 3. Information about speakers.
F – female, M – male, es – Spanish, pt – Portuguese, cz – Czech, CR – Czech Republic, y./m./w. – year, month, week.

Speaker	F/M	Country	L1	Stay in CR (in years)	Study Czech (+University study)	Primary language used in daily life
S1	F	Paraguay	es, pt	8.5	1 y. (+6 y.)	es, cz
S2	F	Honduras	es	9.5	1 y. (+5 y.)	en
S3	M	Bolivia	es	8.5	1 y. (+5 y.)	en, cz
S4	M	Peru	es	8.5	1 y. (+5 y.)	en
S5	M	Colombia	es	8.5	1 y. (+4 y.)	en
S6	M	Peru	es	2	10 m.	en
S7	M	Spain	es	7	6 m.	en
S8	M	Spain	es	3	6 m.	en
S9	M	Honduras	es	2.5	3 m.	en
S10	M	Spain	es	1.5	3 m.	en
S11	F	Colombia	es	4.5	1 m.	en
S12	M	Ecuador	es	2.5	2 w.	en
S13	M	Peru/CR	es, cz	–	–	cz, es

3.4. Recording procedure

Reading of the Czech story by the 13 Spanish speakers were recorded individually in a sound-treated and sound-proofed room (AKG C 4500 B-BC microphone, sample rate 32 kHz, 16-bit depth). Their main task was to read the text. In a short introductory dialogue, relevant information regarding speakers' personal data and exposure to Czech lan-

guage was gathered. The form of a dialogue was preferred to a questionnaire in order to capture the circumstances of each individual speaker.

Before recording, each speaker had been given time to get accustomed to the text. All, but one speaker, were ready in less than 10 minutes. Only 4 speakers asked for a translation of some less frequent words. No speaker asked for guidance in pronunciation. During the recording, one of the authors was present in a soundproof room to reduce stress of speakers due to the unknown environment. Before reading the actual text, speakers introduced themselves shortly. This was done in order to ensure that the speaker started reading the text in their standard voice and got accustomed to being recorded. Based on an informal discussion following the recording, none of the speakers were able to identify the topic of the experiment.

3.5. Perception analysis

Perception analysis supported by acoustic representation was performed using Praat software (Boersma & Weenink, 2019). Target words were transcribed, and the following procedure was executed:

1. Presence or absence of intonation juncture between the target word and adjacent words was examined.
2. The fluency of the target word as a whole was assessed on the 4-point scale: 0 meant fluent pronunciation with 1–3 signalling degrees of dysfluency. Only words with 0 rating were processed further.
3. Intelligibility of words thus determined was assessed (5-point scale).
4. Further analysis concerned the target consonant clusters was performed in multiple steps.
   a. It was determined whether the cluster was pronounced correctly or incorrectly. During the analysis, cases emerged in which the decision-making was uncertain. Since this group was not large, we opted for the following solution: based on repeated listening, a consonant cluster with little inaccuracy was rated as correct, while clusters with greater inaccuracy were rated as incorrect.
   b. This rough categorization disregarded the fact that some pronunciation variants were less intelligible than others; therefore, we proceeded to the subsequent evaluation of that aspect (5-point scale).

   In case of incorrect realization,

   c. the type of sound changed and d) affected segments were determined.

The following sound changes were studied: substitution, elision, prothesis, epenthesis, metathesis, lengthening of the consonant, weakening. Based on the analysis, another type was added, namely splitting, i.e., the splitting of a word cluster into two parts. In some consonant clusters, multiple sound changes co-occurred. In cases where sound changes affected different segments, these changes were accounted for separately, e.g. [barokɲi:] → [baro(k)ni:] as weakening of [k] and substitution [ɲ] → [n]. Another typical example was the addition of a prothetic sound to a cluster and affecting a consonant simultaneously. The category of accumulation was newly introduced for cases where a consonant was affected by several sound changes [prokopskɛ:ɦo] → [prokops:(ʃ)skɛɦo], or when it was not possible to clearly determine the type of sound change, e.g. [sɲatɛk] → [stək].

In the following analysis we use the data obtained in step 2 and present the results of phase 4a, 4c and partially 4d.

4. Data analysis

4.1. Correctness rate: overview

The resulting set of 975 target clusters was analyzed (75 words x 13 speakers): 7.0% of target words were affected by slips of tongue, dysfluency (see step 2 above) or repetition and those items were excluded from further analysis, 65.7% of consonant clusters were pronounced correctly, 27.3% of them incorrectly.

Concerning the position within a word, the I, M, F positions did not differ in the number of excluded cases, ranging from 6.4% to 7.2%. The correctness rate in M and F was similar (M: 70.1%, F: 69.2%), in I it was a little bit lower (60.9%).

In the following sections 4.2 and 4.3, the results presented have already all the above-mentioned exclusions.

4.2. Correctness rate: consonant clusters

In this part, the results regarding consonant clusters are presented. Fig. 1 shows the number of correct variants of each cluster type (for types see section 3.1). Each type achieved at least 60% of correct realizations. The S+cons and O+nas types narrowly crossed this line. The greatest correctness rate was indicated in the PS and PSK types (about 85%). The PT and PST types were situated roughly in the middle of the range.

Nevertheless, these summarizing results may disguise differences within cluster types according to their phonetic composition or within the same consonant cluster according

to the positions I / M / F. Fig. 2 provides the comparison of correctness rate for consonant clusters in which different positions in the word were tested.

For the PS and PT types, all three positions were tested. The PS type achieved a very high correctness rate in M and F (slightly above 90%); the correctness rate was lower in I, but still very high (almost 80%). For PT, the correctness rate differed for all positions, decreasing in the direction I – M – F, the difference between I and F is about 20% (I: 86.1%, F: 64.9%).

In the other three consonant cluster types, only two positions were tested. The biggest difference between the positions was seen in the PN type, where the realization in M was very successful (86.8%). On the contrary, in I, incorrect realizations prevailed (the number of correct variants was only 34.3%). In another type with nasal GN, the M position was as successful as in PN (86.1%). In I, the correctness rate was slightly lower compared to M, however, unlike in PN, the correctness rate of M in GN was still relatively high (75.0%).

The three-segment cluster PST, similarly to PT, indicated a lower correctness rate in F compared to M. For PT, the difference between these positions was about 10%; for PST, it was even about 20% (M: 81.4%, F: 60.0%). The number of correct realizations of PSK, which was tested only in M, was similar to PST in this position (86.5%).

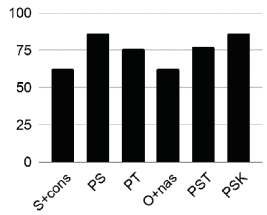


Figure 1. Correctness rate of consonant cluster types (in %).

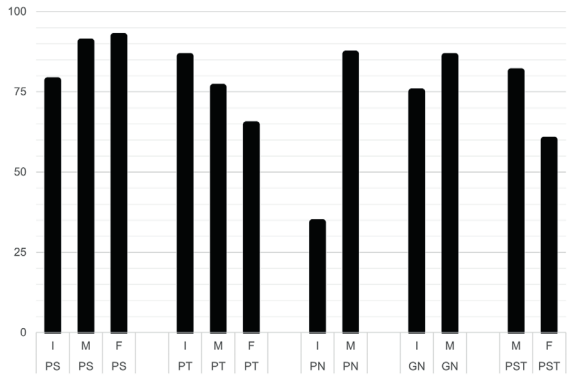


Figure 2. Correctness rate of consonant clusters tested in two positions in the word as minimum (in %).

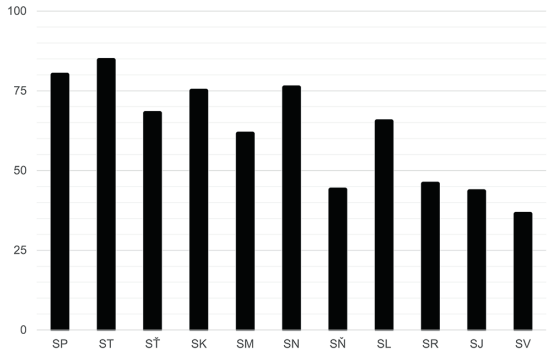


Figure 3. Correctness rate of S+cons clusters tested in the initial (I) position in the word (in %).

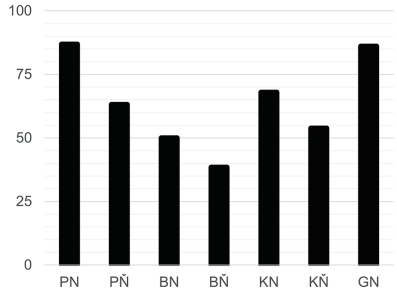


Figure 4. Correctness rate of O+nas cluster types tested in the medial (M) position in the word (in %).

Fig. 3 shows all two-segment clusters belonging to the S+cons type which was tested in the I position. The correctness rate of pronunciation was very high for clusters ST and SP (between 80% and 90%). In the next band (between 70% and 80%) there were SN and SK clusters. The limit of 60% was exceeded by three more clusters of the S+cons type – SŤ, SL and SM. The S+cons clusters can therefore be divided into two groups. There were seven clusters with the correctness rate of over 60%, representing four tested combinations of two obstruents (SP, ST, SŤ, SK), two combinations with nasals SN and SM and a combination with lateral SL. The remaining four clusters did not reach even 50% of correct variants – these were the remaining combinations with sonorants SR, SŇ, SJ and the cluster SV with fricative [v].

Fig. 4 compares the correctness rate of O+nas clusters in the M position, i.e., the combinations with palatal [ɲ] (occurred only in this position in our set) and the combinations with alveolar [n] (tested in the I and M positions, see above). As we have already shown in the previous explanation, the correctness rate of pronunciation was very high for clusters PN and GN in the M position (between 80% and 90%). Unlike them, the correctness rate

of BN cluster was very low (50%) and the rate of KN is situated roughly in the middle of the range (70%).

The PŇ type was the only combination with a nasal palatal in which the number of correct realizations exceeded 60%, for KŇ the number of correct realizations was around half of the cases, for BŇ it did not even reach 40% (GŇ was not eventually included in the set, see section 3.2). For all pairs of clusters N / Ň, the number of correct realizations was higher for the cluster with alveolar [n] than for the cluster with palatal [ɲ]; the highest difference was in the pair PN – PŇ (24%). The same observation was made for clusters SN – SŇ (32%) belonging to S+cons type.

4.3. Sound changes

4.3.1. Sound changes: overview

In this section, we provide an overview of sound changes that occurred in the set of incorrect pronunciation (step 4c, see 3.5).

Table 4. Sound changes according to their frequency. + the form of a Czech example is not a lemma.

Type of sound changes	Frequency (in %)	Example, correct pronunciation	Example, real pronunciation	In English	In Spanish
substitution	44.3	[ɪ**gn**orovat] →	[ɪ**xn**orovat]	to ignore	ignorar
		[pro**gn**o:zu] →	[pro**kn**o:zu]	prediction+	pronóstico+
		[**sl**ɛt͡ʃnu] →	[**ʃl**ɛt͡ʃnu]	young lady+	señorita+
elision	22.0	[su**pst**ants͡ɪ] †	[su**st**ants͡ɪ]	substance+	substancia+
		[**ps**ɪsko] →	[**s**ɪsko]	dog	perro
prothesis	20.2	[**st**at͡ʃɪlo] →	[ɛ**st**at͡ʃɪlo]	to be enough+	ser suficiente+
		[**sr**ovnala] →	[ɛ**sr**ovnala]	to compare+	comparar+
weakening	2.8	[krɛ**ps**ɪlonɛm] →	[krɛ**(p)s**ɪlonɛm]	crepe+	crepé+
epenthesis	2.1	[**pn**ɛumatɪka] →	[**psn**ɛumatɪka]	tyre	neumático
lengthening	1.7	[**ps**ɛm] →	[**ps**::ɛm]	dog+	perro+
metathesis	0.7	[**sj**ɛzdu] →	[**sɛ̆j**ʒdu]	exit+	salida+
accumulation	3.5	[proko**pske**:ĥo] →	[proko**ps:(ʃ)ske**ĥo]	Prokop+ (adj.)	Prokop (adj.)
splitting	2.8	[po**pt**a:fka] →	[po**p t**avka]	demand	demanda

Within the whole set, a multiple occurrence of incorrect realisations within the consonant group occurred in 19 cases. There was a co-occurrence of two changes, with the exception

of one case with three changes. The total number of sound changes was thus 20 higher than the number of incorrect implementations.

Among the types of changes, substitution was the most frequently represented (44.4%). The second most numerous were elision (22.0%), and prothesis (20.3%); their frequency was therefore about half that of substitution. The frequency of other types (weakening, epenthesis, lengthening, metathesis, and accumulation and splitting into two stress groups) did not reach 5% (see Table 4 for more details); their total share in the number of sound changes was 13.3%.

4.3.2. Sound changes in types of consonant clusters

In this section, the distribution of sound changes in consonant cluster types is presented. Based on previous findings, three most common types of changes, i.e., substitution, elision and prothesis, have been distinguished; the remaining changes are included in the group "others".

Fig. 5 shows two types of values for each type of consonant clusters. The first value represents the number of incorrect variants. Other values indicate the distribution of sound changes for a given cluster type.

It is obvious that the types of clusters differed in the types and the amount of sound changes they evoked. The most visible finding was that prothesis occurred only in S+cons. For this type, prothesis covered the entire half of all sound changes (51.8%). Another relatively common sound change in this type was substitution. However, the distribution of sound changes varied among single clusters of this type (see below).

Substitution was the most common sound change for O+nas, where it applied to ⅔ of all sound changes (67.6%). One-fifth of the sound changes in this type was elision. However, almost all the instances of elision appeared only in the I position of PN, which also contained a lot of incorrect realizations overall (the position I of GN was rather successful). In M, nearly all incorrect realizations were the matter of substitutions, regardless of the number of incorrect forms, or whether the cluster contained N or Ň.

Elision covered more than half of the sound changes for PS and PT (53.8%, 57.7%). However, in the case of PS it was elision in I, and in the case of PT the cluster in F was simplified. PS and PT types, compared to other cluster types, had relatively more sound changes included in the group "others" (for PS about 30%). These changes occurred mainly in M.

The figure does not include the types PST and PSK, for which there were only 12 and 5 sound changes respectively; in both cases, it was mainly a substitution, in M of PST elision as well.

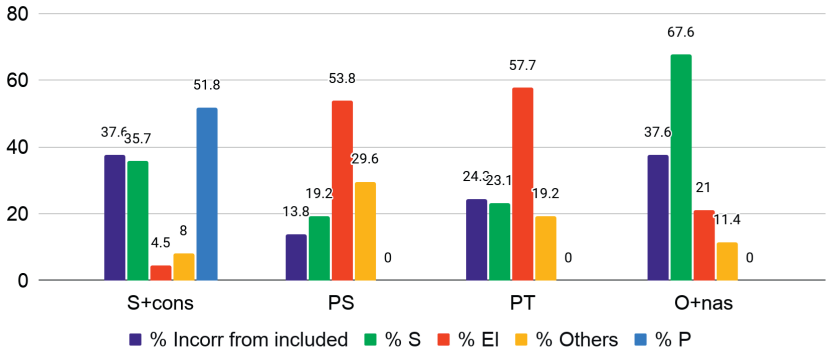


Figure 5. Distribution of sound changes within consonant cluster types (in %). Incorr – incorrect realizations, S – substitution, El – elision, P – prothesis.

Fig. 6 shows the number of incorrect realizations and the distribution of sound changes in consonant clusters of the S+cons type (the absolute values of). In two of the four least successful clusters SŇ and SV, there was a considerable number of substitutions; prothesis reached about half of the cases there. On the other hand, in the four most successful clusters, which were three obstruent clusters SP, ST, SK and SN, substitution did not occur at all (except for one occurrence in SN). For the remaining clusters, the number of instances of prothesis and substitution were either comparable or the number of substitutions was lower. Elision occurred only individually; changes included in the "others" were also limited and occurred in the least successful clusters with a sonorant.

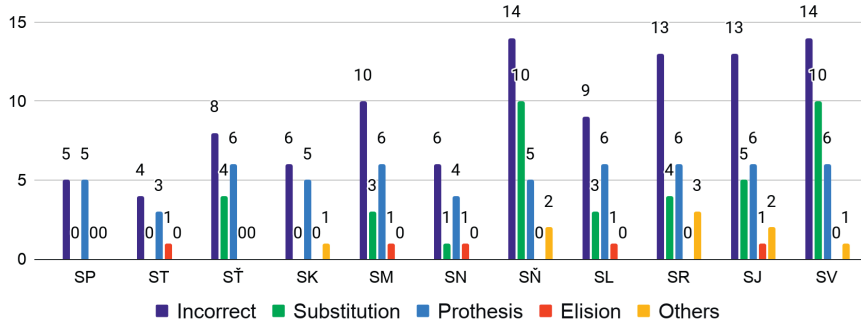


Figure 6. Distribution of sound changes in the S+cons type (absolute values).

4.4 Success rate and sound changes in individual speakers

Concerning individual speakers (see Table 5a), the number of correct forms ranged from 44.0% to 84.0%, while the number of incorrect forms ranged from 12.0% to 52.0%. Speakers also differed in the number of excluded cases that ranged from 1.3% to 13.3%. The number of excluded items did not correlate with the number of correct realizations (r=0.13, using Spearman's coefficient).

Table 5b indicates the number of incorrect realizations for each speaker and the distribution of sound changes. In the speech of speakers with fewer than 25 incorrect items (⅓ of all target clusters), it was substitution that prevailed, except S13, who tended to elision.

Table 5. a) Number of correct and incorrect realizations and excluded clusters regarding speakers (in %). b) Number of incorrect realizations and number and type of sound changes regarding speakers. Corr / Incorr – correct / incorrect realizations, Ex – excluded items, S – substitution, El – elision, P – prothesis.

	a)				b)					
Speaker	**Corr**	**Incorr**	**Ex**		**Incorr**	**S**	**El**	**P**	**Others**	**Total**
S1	80.0	12.0	8.0		9	6	1	0	2	9
S2	73.3	13.3	13.3		10	7	2	1	1	11
S3	77.3	13.3	9.3		10	5	3	2	0	10
S4	57.3	41.3	1.3		31	8	11	12	1	32
S5	53.3	34.7	12.0		26	19	3	0	7	29
S6	69.3	25.3	5.3		19	10	4	1	4	19
S7	62.7	33.3	4.0		25	16	7	1	2	26
S8	53.3	41.3	5.3		31	17	4	8	6	35
S9	69.3	20.0	10.7		15	8	4	1	4	17
S10	84.0	14.7	1.3		11	6	2	0	4	12
S11	50.7	40.0	9.3		30	11	8	14	1	34
S12	44.0	53.3	2.7		40	13	8	18	4	43
S13	78.7	13.3	8.0		10	1	6	0	3	10
				Sum	**267**	**127**	**63**	**58**	**39**	**287**
				%		44.3	22.0	20.2	13.6	100

A more detailed analysis was applied to speakers with at least 25 incorrect variants. These were six out of 13 analysed speakers (marked in grey in the Table 5a). The ratio between correct, incorrect and excluded cases in these speakers is clearly shown in Fig. 7. In one of these speakers, the number of incorrect realizations prevailed over the correct ones (S12 53.3% of incorrect variants). There were speakers with both the low number of excluded items (S4 1.3%) and the higher number of excluded items (S5 12.0%). The distribution of sound changes was to a large extent variable (see Fig. 8). Speaker S12 and S11 manifested

the largest number of prothesis (more than 40%). Unlike them, S5 had no prothesis, but dominated in the number of substitutions (65.5%); similar number of substitutions and almost no instance of prothesis were observed by S7. Speaker S4 applied elision to a larger extent than most of the others (34.4%). Speaker S5 had a noticeably higher number of "others" types of sound changes compared to most other speakers (24.1%). Possible influence of the factors we obtained (duration of stay in the Czech Republic, studying of Czech, etc.) on the correctness rate are discussed in the next section 5.

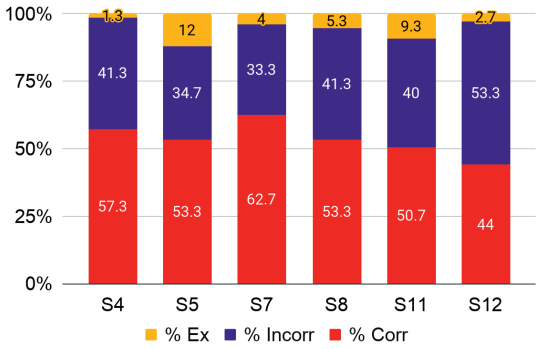


Figure 7. Number of correct (Corr) and incorrect (Incorr) realizations and excluded (Ex) items (in %) regarding six mostly unsuccessful speakers.

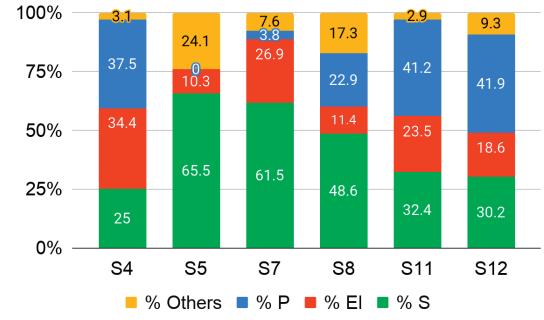


Figure 8. Distribution of sound changes regarding six mostly unsuccessful speakers (in %). P – prothesis, El – elision, S – substitution.

5. Discussion

Pronunciation of defined consonant clusters was proved to present difficulties for Spanish speakers, including the advanced ones. On average, ⅔ of realizations were correct, ⅓ contained errors, slips of tongue or dysfluency. It seems that the initial position was slightly

more difficult, however compared to M and F positions, the difference was not so remarkable. Nevertheless, we need to bear in mind that the clusters were not tested in a balanced way in I, M, F.

The correctness rate among the cluster types and within the types varied considerably. There was a tendency for clusters containing only obstruents to be more successful. This was evidenced by the number of correct realizations of both disyllabic clusters PS and PT, clusters of /s/ + stop – SP, ST, SŤ, SK, as well as three-syllable clusters PST and PSK. Even the least successful obstruent clusters achieved a correctness rate of over 60% (with the exception of the SV cluster, see below). Numerical values also indicated the tendency: clusters consisting only of obstruents had a correctness rate of 81.0%, clusters with nasals (O+nas and relevant clusters of S+cons type – SM, SN and SŇ) 62.0% and clusters containing oral sonorants SJ, SL, SR achieved the lowest correctness rate of 52.1%.

The SV cluster, indicating the lowest correctness rate of all the clusters tested – only 36.4%, was not included in the calculations above. In Spanish, [v] can be heard for example in the word *afgano* as the voiced variant of /f/ (RAE, 2011, p.186). In Czech, it functions as phoneme /v/, and phonetically, it is classified as a fricative, however, due to historical development, it behaves like a sonorant in certain positions. For example, it does not cause voicing assimilation of the previous unvoiced obstruent. So, in the SV cluster, [s] remains voiceless in Czech. Both analysed words containing SV, *sváteční* and *svobodu*, achieved the high number of incorrect forms (14/22). Substitution, namely sonorization [s] → [z], was very frequent (9/14). Prothesis was also relatively common (6/14), with one speaker combining both of these sound changes within a word. The incorrect realization of the SV words was caused by the application of the incorrect orthoepic rule and the sound change typical of the S+cons clusters following the structure of the Spanish syllable.

It was indicated that correctness rate may be influenced by the position of the cluster in the word. In I, M and F, two clusters PS and PT were tested. The correctness rate of PS was very high in all positions, in M and F of about 90%, in I slightly lower. In PT, the tendency was reversed and the difference between I and F was more evident: the I position was the most successful – 86%, F the least successful – 65%. Elision, namely that of [p], obviously prevailed among the incorrect realizations of PT and PS.

In the PT type, three words were tested in F. Two words *manuskript* and *položept* contained a greater number of incorrect realizations (14/24). This may be because these are trisyllabic words, less frequent, and the Spanish equivalent of *manuscrito* no longer contains the consonant cluster *pt*. The word *recept*, on the contrary, was relatively successful (incorrectness 3/13). It is a quite common disyllabic word; in Spanish, in addition to the

word *receta*, there is also *recepta*, which might encourage the preservation of the consonant cluster in pronunciation. This parallel could also be seen in tested PS words in F *biceps* and *kolaps* with a large number of correct realizations. Both words are loanwords and in Spanish spelling *bíceps*, *colapso* they have retained the consonant cluster.

For PS, a potential difference may be found between the pronunciation of native and loanwords in I. For the latter, the tendency towards elision seems stronger. In the words *psychologie* and *pseudogotický*, where it is possible to omit *p* in Spanish equivalents in writing as well, 8/22 incorrect realizations occurred. For native vocabulary, e.g., *psi*, *psala*, there were only 7/48 incorrect realizations. However, the word length might have affected pronunciation as well.

In I of PN, with a considerable number of incorrect realizations (23/39), this difference was not detected. The speakers pronounced both loanwords *pneumatika*, *pneumatiky*, whose Spanish counterpart is spelled only without *p* – *neumático*, and the native word *pnula* incorrectly. PN was also another example of a cluster with a significant difference between positions – unlike in I, the speakers were more successful in M (only 5/39 incorrect forms). In addition, substitution applied mostly in M, opposite to I where elision prevailed in both PS and PT.

An interesting tendency was noted regarding nasals – for the respective pairs PN – PŇ, BN – BŇ and KN – KŇ tested in M, the cluster containing an alveolar was always more successful than the one with a palatal. This applied not only to stop + nasal clusters, but also to SN – SŇ, for which the difference within the pair was most considerable. However, a more detailed word-level analysis will be required to account for possible factors. For instance, in the words *snubní* and *barokní*, substitutions [ɲ] → [n] was applied frequently. The impact of spelling on pronunciation cannot be excluded as a factor: In these words, the grapheme *n* is the part of the digram *ní*, which is pronounced as [ɲi:], not [ni:].

Regarding sound changes, substitution, elision and prothesis represented almost 90% of them. Substitution, which affected all analysed clusters, was the most frequent. This may have been caused by the fact that the category of substitution is very extensive and may include different types of processes (voicing assimilation, articulatory assimilation both in place and manner, etc.). In BN/BŇ, KN/KŇ and GN in M, substitution was obviously the dominant sound change, as it occurred at least in ¾ of realizations. Examination of the substitution types may help explain the low correctness rate of clusters containing /b/. In accordance with Spanish rules, Spanish L1 speakers often weakened the closure and pronounced the sound as an approximant or a fricative. The occurrence of substitution was also significant for PN in M (see above) and S+cons (about ⅓ of sound changes). In the

latter, the type of substitution may contribute to explaining the lower correctness of some clusters as well. For example, [s] followed by a sonorant was quite often assimilated to [z], similar as in SV (see above).

Elision appeared in both disyllabic and trisyllabic clusters, beginning with [p]; it was this consonant that was mostly elided. See the discussion on PS, PT and PN above. Unlike most of the other sound changes, prothesis was present only in S+cons, and it accounted for more than half of all changes in this type. This may be due to the /s/ + consonant group being widely spread in Spanish but not appearing as an onset at the beginning of a word. In this position, it is standardly divided into two syllables adding a vowel prior to the /s/ + consonant group.

The range of correctness rate in terms of speakers was relatively wide, which was not so surprising, given the composition of the speakers group and the interview data. Based on the correctness rate, the speakers were divided into two groups. Although the research did not focus on the possible influence of extralinguistic factors, we wondered if there were some common features within the groups. The obtained data did not allow for greater generalization; however, some findings may be presented.

Of the 13 speakers, only four regularly used Czech on a daily basis (S1, S3, S6, S13) with two of them working in Czech environment (S1, S3); a total of three mentioned Czech as one of the two languages they speak mostly (S1, S3, S13). All four speakers belonged to the group with higher correctness rates. However, as the example of the S13 speaker showed, active use, supported here by partial school attendance in Czech, was not a guarantee of mastering pronunciation at the highest level. Although this speaker mentioned Czech besides Spanish as his mother tongue, he did not deviate from other speakers with low frequency of incorrect forms.

Three speakers from a more successful group shared the experience of a one-year Czech preparatory course and subsequent study at a university in Czech (S1, S2, S3). However, even studying in Czech is not in itself a guarantee of a correct pronunciation, unless supported by other factors. Namely, speakers S4 and S5 also went through the same type of course and university, but practically didn't use Czech afterwards and, based on the analyses, they belonged to a less successful group. The same may be said about the period of stay in the Czech Republic – out of the whole group of respondents, all five named above stayed in the Czech Republic the longest (if S13 is omitted), around 9 years, but the correctness rate was different.

Speaker S10 is a very interesting case. He made a comparable number of errors as respondents who had graduated from a Czech university and used Czech regularly. However, S10 moved to the Czech Republic only a year and a half before recording and had only

three months of self-study. He mentioned that he loves literature, writes stories himself, and although he did not have particularly intense contacts with the Czech environment, he tried to listen to Czech as much as possible on the street and in the media.

Thus, it seems that the active use of Czech or an active approach and probably motivation are likely to be beneficial. Speakers in the less successful group mentioned English as the language of communication, some barely associated with Czechs and did not use Czech. When they did use it, it was a less frequent use in the city, listening to TV / radio or in meetings with Czech extended family.

6. Conclusion and perspectives

The presented experiment brought useful findings that can be followed up. Within the already analysed material, it would be useful to compare in more detail the realization and sound changes of individual words. Due to the length of the recordings, the already carried out analysis of 975 units could be expanded up to double in the framework of the current set of consonant clusters; however, because of unintentional occurrences, the balance of all clusters and positions is not guaranteed. Undoubtedly, it will be useful to expand the set of analysed consonant clusters, both in terms of segment combinations and their number. It will be appropriate to verify the identified tendencies on a larger number of respondents and to obtain a more balanced group of males and females. The analysis was performed on the read text, which posed both advantages (controlled occurrence of target clusters, by speakers no need to formulate themselves) and disadvantages (potential influence of the graphic form on pronunciation, more difficult vocabulary), so it will be appropriate to expand the research material with recordings of spontaneous speech. The rating of intelligibility processed by authors was for information only; perception tests focusing on the impact on a native speaker in terms of foreign accent, intelligibility and comprehensibility would also be beneficial. Recordings of Czech native speakers started to be gathered to compare native and non-native speech. In addition, it would be useful to analyse the production of consonant clusters in speakers of other L1s, which could not only enhance our theoretical knowledge, but also be beneficial for improving methods in teaching pronunciation of Czech as L2.

Acknowledgements

This research was supported by the Czech Science Foundation Project No. 18-18300S "Phonetic properties of Czech in non-native and native speakers' communication".

We would like to thank anonymous reviewers for their constructive comments and recommendations.

Appendix

1. A sample of a Czech text that was read and recorded (the target clusters are indicated)

Sára, původem Švédka, začala spolu s rodiči žít v Praze krátce po sametové revoluci. Stěhovat se nejdřív nechtěla. Svoje priority si nicméně po**st**upně **sr**ovnala a později nelitovala. Odjakživa ji lákala **ps**ychologie, po maturitě proto skládala přijímací zkoušky na Filozofickou fakultu, bohužel neúspěšně. Nepochybně byla zklamaná, ale nerezignovala. Další rok se na vytoužené studium dostala. Byla nadšená, že si konečně plní své sny a jako **st**udentka poprvé v životě pocítila opravdovou **sv**obodu.

Diplomovou práci **ps**ala na téma psychologie skeptiků na území Evro**psk**é unie. V průběhu studia ji totiž zaujaly **sp**ekulace, které se týkaly vnímání **sk**epse a její různé kon**cept**y. Včera složila státnice. Byla nesmírně šťastná a ačkoli byla a**bst**inentka [pst], měla sraz s kamarády a šla slavit. Ti se jí smáli, když okolo ho**ps**ala a radovala se jako malá holka. I**gn**orovat ji nemohla ani skupina lidí stojících opodál. Blonďatá „Sněhurka" s modrýma očima, štíhlé sportovní postavy snadno přitahovala pozornost. Měla na sobě velice pě**kn**ou baro**kn**í [kɲ] sukni skořicové barvy a jemnou stylovou blůzu. Dokonalý **sv**áteční vzhled doplňovala bílá ma**gn**ólie, která se Sáře **pn**ula ve vlasech.

Kolem se šouralo nějaké **ps**isko s ježatými chlupy. Tohoto psa, u něhož lékařka vyslovila pro**gn**ózu, že brzo osle**pn**e, a který **st**ěží [sc] slyšel na jedno ucho, k sobě zavolala starší, **sm**utná paní. Dávala si v kavárně pozdní **sn**ídani [sɲ] – popíjela svou oblíbenou vídeňskou kávu s čerstvým meruňkovým koláčkem a četla další román Milana Kundery. Jakmile zahlédla Láru, začala ji pozorovat a bezchy**bn**ě [bɲ] odhalovat všechny dro**bn**é detaily její trochu extravagantní sukně. Například, že svrchní látka byla zhotovena z dvojvlákna, a spodní, která pomáhala sukni na**pn**out a udržet její tvar, byla jistě bavlna s krajkovou ozdobou dole a kre**ps**ilonem. Sukně byla tak dlouhá a splývavá, že v ní člověka **sn**ad ani nemohlo zá**bst** [pst].

2. English translation of the Czech text sample

Sarah, originally from Sweden, started living with her parents in Prague shortly after the Velvet Revolution. At first, she didn't want to move, however, she gradually put her priorities straight and later did not regret it. She has always been attracted to psychology, so after graduating from high school she attended the entrance exams to the Faculty of Arts, but unfortunately was not accepted. No doubt she was disappointed, but she did not give up.

The next year she got into the university. She was excited that she was finally fulfilling her dreams and, as a student, for the first time in her life she felt real freedom.

She wrote her diploma thesis on the topic of psychology of skeptics in the European Union. During her studies, she became interested in speculations concerning the perception of skepticism and its various concepts. Yesterday she passed the state exam. She was extremely happy and although she didn't drink, she met her friends and went to celebrate. They laughed at her as she jumped around and rejoiced like a little girl. Even a group of people standing nearby could not ignore her. A blond "Snow White" with blue eyes and slender athletic figure would easily attract attention. She was wearing a very nice baroque cinnamon color skirt and a delicate stylish blouse. The perfect festive look was complemented by a white magnolia, which decorated Sarah's hair.

An older, sad-looking lady called a rough-looking dog that was running around to come close to her. It could barely hear in one ear and a doctor warned that it would go blind soon too. The lady was having a brunch in the café; she was sipping her favorite Viennese coffee with a fresh apricot pie and reading another novel by Milan Kundera. As soon as she spotted Sarah, she began to observe her, precisely revealing all the small details of her somewhat extravagant skirt. For example, the top fabric was made of double fiber, and the bottom fabric, which helped tighten the skirt and maintain its shape, was certainly cotton and crepe with a lace ornament at the bottom. The skirt was so long and flowing that you definitely wouldn't feel cold in it.

3. Spanish translation of the Czech text sample

Sarah, nacida en Suecia, comenzó a vivir con sus padres en Praga poco después de la Revolución de Terciopelo. Al principio no quería mudarse, sin embargo, gradualmente puso sus prioridades en orden y no se arrepintió. Siempre le atraía la psicología, por lo que después de realizar el bachillerato asistió a los exámenes de ingreso a la facultad, pero lamentablemente no fue aceptada. Sin duda, estaba decepcionada pero no renunció y al año siguiente ingresó a la universidad. Estaba emocionada de que finalmente estaba cumpliendo sus sueños y, como estudiante, por primera vez en su vida sintió verdadera libertad.

Escribió su trabajo fin de grado sobre el tema de "La psicología de los escépticos dento la Unión Europea". Durante sus estudios, se interesó por las variantes de la percepción del escepticismo y sus diversos conceptos. Ayer aprobó el examen estatal y estaba extremadamente feliz. Aunque no bebía alcohol se fue a celebrar con sus amigos. Se rieron de ella mientras saltaba y se regocijaba como una niña, incluso un grupo de personas que estaban

cerca no podían ignorarla. Una rubia "Blancanieves" con ojos azules y una figura atlética esbelta fácilmente llamaba la atención. Llevaba una falda estilo barroco muy bonita de color canela y una blusa elegante y delicada. El look festivo perfecto se complementó con magnolia blanca, que decoraba el cabello de Sarah.

Una señora mayor y con aspecto triste llamó al perro con pelo de punta que se movía de un lado a otro para que se acercara a ella. El perro apenas oía por un oído y el veterinario advirtió que pronto tambien se quedará ciego. La señora estaba tomando un brunch en el café, bebía su café vienés favorito con una tarta de albaricoque recién hecho y leía otra novela de Milan Kundera. Tan pronto como vió a Sarah, comenzó a mirarla, observando con precisión todos los pequeños detalles de su falda tan extravagante. Notó que la tela superior estaba hecha de doble fibra y la tela inferior, que tensaba la falda y mantenía su forma, era de algodón y crepé con un adorno de encaje en la parte inferior. La falda era tan larga y fluida que una seguramente no tendría frío con ella puesta.

References

Bičan, A. (2013). *Phonotactics of Czech.* Peter Lang Verlag. https://doi.org/10.3726/978-3-653-03482-0

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.25. http://www.praat.org.

Čermák, P. (2015). *Fonetika a fonologie současné španělštiny*. Karolinum.

Ellis, R. (1985). *Understanding second language acquisition* (2nd Ed.). Oxford University Press.

Fashola, O. S., Drum, P. A., Mayer, R.E., & Kang, S. J. (1996). A Cognitive theory of orthographic transitioning: Predictable errors in how Spanish-speaking children spell English words. *American Educational Research Journal*, *33*(4), 825-843. https://doi.org/10.2307/1163417.

Helman, L. A. (2004). Building on the sound system of Spanish: Insights from the alphabetic spellings of English-language learners. *The Reading Teacher*, 57(5), 452-460. http://www.jstor.org/stable/20205383.

Hevia-Tuero, C., Incera, S. & Suárez-Coalla, P. (2021). Does English orthography influence bilingual Spanish readers? The effect of grapheme crosslinguistic congruency and complexity on letter detection. *Cognitive Development*, 59, 101074. https://doi.org/10.1016/j.cogdev.2021.101074.

Hummel, K. M. (2014). *Introducing second language acquisition: Perspectives and practices*. John Wiley & Sons.

Kučera, H. & Monroe, G. K. (1968). *A comparative quantitative phonology of Russian, Czech and German*. Elsevier.

Ludvíková, M. & Kraus, J. (1966). Kvantitativní vlastnosti soustavy českých fonémů. Slovo a slovesnost, *27*(4), 334-344.

Magen, H. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*(4), 381-400. https://doi.org/10.1006/jpho.1998.0081.

Moore, F. B., & Marzano, R. J. (1979). Common errors of Spanish speakers learning English. *Research in the Teaching of English*, *13*(2), 161-167. http://www.jstor.org/stable/40170752.

Palková, Z. (1997). *Fonetika a fonologie češtiny – s obecným úvodem do problematiky oboru* (2nd ed.). Karolinum.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, *29*, 191-215. https://doi.org/10.1006/jpho.2001.0134.

Quilis, A. (1993). *Tratado de fonología y fonética españolas*. Gredos (Biblioteca románica hispánica III, 74).

Quilis, A., & Fernández, J. (1979). *Curso de fonética y fonología españolas para estudiantes angloamericanos* (9th ed.). C. S. I. C.

RAE. (2011). *Nueva gramática de la lengua española. Fonética y fonología*. Espasa Libros.

RAE. (2021). *DLE (Diccionario de la lengua española)*. https://dle.rae.es.

Ríos Mestre, A. (1999). *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico*. Estudios de Lingüística del Español, *4*. ISSN: 1139-8736. http://elies.rediris.es/elies4/

Rothman, J. (2008). Why all counter-evidence to the critical period hypothesis in second language acquisition is not equal or problematic. *Language and Linguistics Compass 2*(6), 1063-1088. https://doi.org/10.1111/j.1749-818X.2008.00098.x.

Saporta, S., & Olson, D. (1958). Classification of Intervocalic Clusters. *Language*, *34*(2), 261-266. https://doi.org/10.2307/410830.

Singleton, D. (2005). The Critical Period Hypothesis: A coat of many colours. International *Review of Applied Linguistics in Language Teaching*, *43*(4), 269-285. https://doi.org/10.1515/iral.2005.43.4.269.

Sun-Alperin, M. Kendra & Min Wang (2008). Spanish-speaking children's spelling errors with English vowel sounds that are represented by different graphemes in English and Spanish words. *Contemporary Educational Psychology*, *33*(4), 932-948

Šturm, P. (2018). Experimental evidence on the syllabification of two-consonant clusters in Czech. *Journal of Phonetics*, *71*, 126-146. https://doi.org/10.1016/j.wocn.2018.08.002.

Těšitelová, M., Confortiová, H., Králík, J., Ludvíková, M., Nebeská, I., & Uhlířová, L. (1985). *Kvantitativní charakteristiky současné češtiny*. Studie a práce lingvistické, sv. 19. Academia.

Chapter XII

Relacionando los análisis cualitativo y cuantitativo. Una propuesta de modelo estadístico predictivo para completar la descripción compleja de los verbos cognitivos[1] [2]

Relating qualitative and quantitative analysis. A predictive statistical model proposal to complete the complex description of cognitive verbs

M. Amparo Soler Bonafont
Universidad Complutense de Madrid – España

Resumen: El objetivo del presente capítulo es realizar una propuesta de descripción de los usos semántico-pragmáticos de unas formas verbales complejas como son las formas performativas de los verbos cognitivos, concretamente, de su forma paradigmática *creo*, en la interacción oral. Para ello se lleva a cabo una aproximación cognitiva con base en una propuesta estadística predictiva, creada a partir de un sistema de regresiones multinomiales (con la herramienta STATA). Se persigue que el modelo diseñado permita reconocer con un elevado grado de explicatividad ante qué

1 Este capítulo profundiza en algunos de los resultados parciales que son fruto de la tesis doctoral de la autora (Soler, 2019), así como de la ampliación que de ellos se realiza en Soler (2021b).

2 La investigación se enmarca en el proyecto *Los procesos de gestión de la imagen y la descortesía: perspectivas históricas, lingüísticas y discursivas*, concretamente, en la subdivisión de análisis de procesos discursivos (ref. PID2019-107668GB-100, Ministerio de Ciencia e Innovación, Gobierno de España).

significados y funciones pragmáticas de unidades polisémicas y polifuncionales como *creo* nos encontramos, una vez sistematizadas las principales circunstancias de aparición cualitativas que las rodean. El estudio de formas se da en un corpus compilado de conversaciones coloquiales y de discursos de debate parlamentario. Como resultado, se obtiene un modelo de análisis predictivo replicable en otros textos y géneros en los que pueden aparecer unidades epistémicas similares.

Abstract: The goal of this chapter is to bring a description proposal for the semantic and pragmatic uses of some complex verbal forms like the performative forms of cognitive verbs, specifically *creo*, in oral interaction. For this purpose, it is carried out a cognitive approach based on a predictive statistical pattern created with a multinomial regressions system (through STATA tool). It is intended that the designed model allows the researcher to recognize what senses and pragmatic functions is dealing with in so polysemic and polyfunctional units as *creo*, with a high degree of explanation, once the main circumstances of its qualitative appearances are systematized. The study of forms is done in a compiled corpus of colloquial conversations and parliamentary debates. As a result, it is obtained a predictive model of analysis which is replicable in other texts and genres in which some similar epistemic units can appear.

1. Introducción

1.1. Dificultades descriptivas en el grupo de las formas performativas de los verbos cognitivos

La explicación funcional del uso semántico-pragmático de algunas unidades epistémicas en los textos resulta aún hoy, y tras siglos de disquisiciones, compleja. Tal es el caso, reveladamente frecuente en la oralidad, de algunas formas verbales de primera persona del singular del presente de indicativo: *creo*, *pienso*…, también conocidas como formas performativas de los verbos cognitivos o de opinión (Fetzer y Johansson, 2010; Fetzer, 2014; González Ruiz, 2015; Soler, 2018). Estas formas verbales son subjetivas y, en algunas ocasiones, pueden manifestarse de manera integrada (*creo que* + verbo) o parentética (uso de *creo* con movilidad posicional), desde el punto de vista morfosintáctico. No obstante, estas características que las identifican no son tan llamativas como otros de sus rasgos definitorios, los cuales dificultan su reconocimiento: estos son su polisemia y su polifuncionalidad anunciadas (Hartwell *et al.*, 2017; Jansegers, 2017; Soler, 2019).

Los diferentes significados y funciones que pueden manifestar unidades como *creo*, la forma paradigmática de este conjunto por tratarse de la más compleja y la más polisémica y polifuncional de su clase (Soler, 2019), han sido estudiadas en diferentes géneros (tanto en español como en otras lenguas), entre los que destacan los de interacción oral, especialmente, la conversación y el debate parlamentario. Así bien, incluso en este tipo de géneros, *creo* y formas verbales semejantes a esta manifiestan desde funciones atenuantes hasta intensificadoras (Cutting, 2007; Fuentes Rodríguez, 2010, 2016; De Hoop *et al*., 2018), a la vez que despliegan una gran variedad de valores semánticos, desde la creencia hasta el juicio (Soler, 2019, 2021). Distinguir la multiplicidad de sus posibilidades semántico-pragmáticas no es tarea sencilla para el lingüista, que se encuentra, desde hace más de un siglo con un escollo adicional en estos verbos: la limitación de las herramientas lingüísticas tradicionales para el estudio de fenómenos complejos como el citado. Los pragmatistas se preguntan cómo definir los significados y significados en uso de unidades subjetivas como las que son objeto de este trabajo, para los que no son suficientemente explicativas las pruebas veritativo-condicionales ni las de la pragmática clásica. Por estos motivos, son cada vez más numerosos los estudios que realizan una aproximación cognitiva a estas formas, gracias a su concepción de la semántica y de la pragmática como un mero *continuum* (Achard, 1998; Buceta, 2014; Jansegers, 2017; Jansegers y Gries, 2017; Boas y Ziem, 2018), lo que ayuda a superar algunos obstáculos definitorios.

No obstante, y de acuerdo con diferentes estudios pragmáticos y sociolingüísticos recientes (Díaz-Campos y Gradoville, 2011; González *et al*., 2014), la explicación cualitativa cognitiva queda incompleta si no se realiza un análisis riguroso de corpus, de tipo cuantitativo (Roldán 2005; Abdulrahim, 2014; Milin *et al*., 2016). Dicha incompletud se observa en la falta de diseños metodológicos cuantitativos capaces de dar una respuesta procedimental adecuada a la hora de operar ante estos casos, así como en la falta de homogeneidad ante la descripción tanto de unos valores semánticos cerrados de *creo*, como de las funciones concretas que puede desarrollar.

1.2. Planteamiento de este trabajo

El objetivo de esta investigación, una vez enunciadas algunas de las fallas metodológicas en el estudio de fenómenos lingüísticos semántica y pragmáticamente complejos, es tratar de llevar a cabo una descripción sistemática del funcionamiento de las formas performativas de verbos cognitivos como *creo* en la interacción oral. Para ello, este trabajo combina el análisis cualitativo de corte cognitivo y el análisis cuantitativo, en lo que se propone como una propuesta predictiva de reconocimiento de valores de *creo*. Se expone un modelo diseñado

mediante distintas regresiones multinomiales de variables cualitativas de análisis (elaboradas con una base cognitiva), las cuales se realizan a través de herramientas como STATA. Con este modelo se consigue reconocer con un elevado grado de explicatividad ante qué significados y funciones pragmáticas de la unidad objeto de estudio podemos encontrarnos, una vez sistematizadas las principales circunstancias de aparición que las rodean.

Las formas analizadas como *tokens* se han extraído de un corpus constituido por dos géneros discursivos de tipo interactivo, en el español de España, de los últimos 20 años: conversaciones coloquiales (de los corpus disponibles *COGILA*, *COJEM*, *Val.Es.Co. 2002* y *Val.Es.Co. 2.0*) y discursos de debate parlamentario (pertenecientes al archivo del *Congreso de los Diputados* del Gobierno de España y de *Les Corts Valencianes* y accesible en línea). Sobre los datos obtenidos, se han aplicado las bases de la estadística descriptiva y predictiva, como también se ha realizado en otros trabajos de corte lingüístico previos (Abbhul y Mackey 2013, James *et al*., 2013). En definitiva, se obtiene un diseño predictivo propio, el cual es replicable en otro tipo de textos y géneros textuales susceptibles de contener unidades epistémicas de funcionamiento similar al de las formas performativas de los verbos cognitivos.

2. Acercamiento teórico a *creo* y otras formas performativas de los verbos cognitivos

Una de las grandes preocupaciones de los filósofos del lenguaje desde el siglo XIX (desde Frege o Russell, hasta Kripke o Richard), y que ha perdurado en la lingüística aún hasta nuestros días, es la de la descripción de aquellas unidades subjetivas cuyo valor de significado no puede ser suficientemente explicado desde la semántica, pero para las que la pragmática tampoco puede ofrecer una solución aislada. Tenemos un claro ejemplo en las formas performativas, esto es, aquellos verbos en primera persona del singular del presente de indicativo, y que son de carácter cognitivos. Son casos como *considero*, *creo*, *opino*, *pienso*, *supongo*…, con los que no solo se hace evidente el *origo*, la presencia del hablante en la escena en que se produce lo dicho, sino que se observa que el significado del referente viene enriquecido con aspectos intencionales que le superpone el hablante gracias a una doble posibilidad de lectura: proposicional y también extraproposicional. Esta naturaleza hace ver, pues, que unidades como las comentadas se encuentran en el límite mismo entre la semántica y la pragmática.

Disquisiciones aparte, en este trabajo abordamos la problemática concreta que ofrece una forma verbal paradigmática: *creo*, por ser considerada esta la más compleja del conjunto de las formas performativas de los verbos cognitivos. Con su estudio pueden verse resumidas cuestiones que atañen al resto de unidades de primera persona del singular de

estos verbos, que funcionan de modo semejante, y cuyas dificultades definitorias (si no todas, sí muchas de ellas) pueden verse subsumidas en las que aquí planteamos para *creo*.

2.1. Polisemia de *creo*

Creo es considerada una forma verbal con un valor altamente subjetivizador de lo dicho (Soler, 2019). Es la forma performativa del verbo *creer*, el cual se caracteriza por ser polisémico, si bien esta polisemia no había sido aclarada hasta los últimos años. El reciente interés investigador por esta polisemia ha cristalizado en el reconocimiento de una alta complejidad cifrada en el conjunto de varios aspectos: su polimorfismo construccional (*creer en*, *creer que*, *no creer*…) (Buceta, 2014; Soler, 2019), la multiplicidad de contextos de aparición (conversación coloquial, entrevistas políticas, debates, etc.) (Fetzer, 2014; Fetzer & Johansson, 2010; González Ruiz, 2015; Soler, 2018), la frecuencia de un fuerte componente argumentativo en su cotexto próximo (Fuentes Rodríguez, 2010, 2016), y la diferente variedad funcional, incluso complementaria (desde la atenuación a la intensificación, pasando por la neutralidad), que puede manifestar (González Ruiz, 2015; Soler, 2019).

De todo ello se desprende que *creo*, la forma más peculiar de su paradigma morfológico, supone un escollo para la investigación, que si bien ha observado las causas de su complejidad, no había conseguido dar hasta la fecha con una descripción consistente de sus usos. Y es que la bibliografía se ha tratado de acercar repetidamente a sus significados, los cuales fluctuaban entre dos y seis valores, sin que pudiera haber acuerdo, sino solo un resumen tradicional de los valores primordialmente en dos: el epistémico o débil y el de opinión o fuerte (Fetzer, 2014; Fetzer & Johansson, 2010; González Ruiz, 2015). Estos dos significados polares se resumen en los siguientes ejemplos:

B: pero ¿qué es/¿que ya lo has dejado oo?
A: **creo que** ya lo he dejado un poco por imposible (*valor epistémico o débil*)
B: yo **creo que** tienes que insistir (*valor de opinión o fuerte*)

Puede observarse que el valor débil de *creo* presenta a modo de duda y no de una convicción lo dicho por A, y expresa que el hablante puede no disponer de pruebas para manifestar lo dicho con mayor grado de seguridad. Por su lado, el valor de opinión se corresponde con la expresión de un juicio personal, independientemente de las pruebas de las que se disponga sobre lo aseverado. En ambos casos está presente la subjetividad, pero esta pone su foco en diferentes aspectos (bien en las pruebas de las que se dispone sobre ello, bien en la confianza de que lo dicho sea de tal o cual modo), incluso con el uso de una misma construcción formal.

La distinción básica revisada puede resultar viable en un primer momento, pero no lo es si nos encontramos ante casos como los que siguen: creo que *tu papi va a jugar con el barquito más que tú* (en que además del grado de seguridad, también podríamos hablar de opinión); *hospital de la Vega Baja, hospital* –creo recordar– *de Elda*… (en que tenemos construcciones de doble acusativo, muy características, en las que tampoco es fácil discernir ante qué valor nos encontramos); *eso es* lo que creo (en que una nueva construcción encapsulada en función de atributo parece estar acercándose más al valor de certeza que al de duda o al de opinión), etc. Con ello, vemos que la polisemia debe abordarse desde un criterio efectivo, que ordene los semas de cada valor de manera rigurosa para poder reconocer límites entre ellos, que supere la diversidad de descripciones bibliográficas y que, de acuerdo con lo visto, evidencie los puntos de conexión con las diferentes construcciones formales del verbo. Asimismo, y como persigue este capítulo, se espera que la categorización obtenida se acompañe de un criterio de reconocimiento sencillo y viable para el analista.

2.2. Polifuncionalidad de *creo*

La polifuncionalidad, no solo de *creo*, sino también de otras unidades de su mismo conjunto de formas performativas, subjetivas y cognitivas, viene de la mano de su reconocida polisemia. Como hemos avanzado, en usos como los de *creo* se han reconocido tradicionalmente funciones de atenuación (creo que *ya lo he dejado un poco por imposible*, Val.Es.Co. 2002), neutralidad (*hospital de la Vega Baja, hospital* –creo recordar– *de Elda*, Les Corts Valencianes), e incluso intensificación (A: es que los mayores↑ además a mí seguro que se me comen (RISAS)/ tienes que tener un SEXTO= // B: NO↓ **yo creo que** exige más↑, Valesco 2.0, C. 1, 68-69).

Diversos estudios monográficos previos que han versado sobre el objeto de estudio de este capítulo se han preguntado si existe una correlación entre los significados reconocidos y las funciones pragmáticas de *creo*. La bibliografía ha llegado a establecer una correlación casi directa entre el valor débil y la atenuación, por un lado, y el valor de opinión, y la intensificación (Fuentes Rodríguez, 2016; González Ruiz, 2015), por otro lado*o*. Estas correlaciones establecidas de forma automática y asumidas por la comunidad científica llevan, no obstante, a arrastrar varios errores conceptuales básicos. Así, por ejemplo, cabe destacar que los estudios de corpus realizados hasta la fecha no aportan una amplitud suficiente de datos basados en corpus de lengua real ni cotejan las observaciones con pruebas objetivas y replicables a partir de las que puedan ofrecerse resultados concluyentes, con lo cuales pudiera confirmarse dicha automaticidad de relaciones semántico-pragmáticas de *creo*. Asimismo, en los estudios se observa una ausencia de criterio para la detección de otros posibles significados, o funciones, distintos a los básicos, ya comentados. Prueba de ello es

que la neutralidad suele quedar fuera de los análisis, pese a que algunos investigadores han llegado a reconocer esta función en casos aislados, o incluso que la atención a las diferentes manifestaciones formales de *creo* y sus repercusiones a nivel semántico y pragmático suelen estar ausentes en las investigaciones. Serán estos aspectos los que tratará de solventar este capítulo con la propuesta de un modelo de análisis concreto, que se presenta como replicable también para otros análisis de unidades doxásticas complejas.

3. Exploración de un análisis cognitivo experimental

La búsqueda de metodologías de análisis, si no alternativas, sí complementarias a las explicaciones cualitativas de la semántica tradicional, ha llevado a la comunidad científica a explorar enfoques integradores, como es el caso de la lingüística cognitiva. Abdulrahim 2014; Fetzer y Johansson 2010; Jansegers 2017; Jansegers y Gries 2017; Milin *et al.*, 2016; o Roldán 2005 son algunos de los casos de análisis semántico-pragmáticos de tipo cognitivo combinados con estadística. Este marco teórico entiende la semántica y la pragmática como un continuo, lo cual ha facilitado la comprensión de formas como *creo* desde este paradigma, como prueban dichos estudios. Gracias a este enfoque, la observación cualitativa del analista no se ve anulada, sino que es, además de reconocida, apoyada en datos reales y comprobables. Se trata, por tanto, del motivo por el que el enfoque cognitivo está tomando cada vez más auge en los últimos años. Asimismo, el acercamiento estadístico predictivo y experimental también se ha visto incrementado recientemente en distintos trabajos lingüísticos, cognitivos, e incluso funcionales y sociolingüísticos (Boas & Ziem, 2018; Díaz-Campos & Gradoville, 2011), en los cuales, como planteamos en este trabajo, un sistema de análisis cuantitativo riguroso completa adecuada y necesariamente la aproximación cualitativa.

4. Metodología del estudio

De acuerdo con lo expuesto, la hipótesis de partida que planteamos es que debe de existir la posibilidad de realizar un cálculo aproximado de los valores semánticos y pragmáticos que manifiestan formas performativas como *creo*, si el inventario de categorías (significados y funciones) que se les reconoce es cerrado[3]. Por esta razón, la pregunta de investigación a la

3 Desde el punto de vista de la Semántica Cognoscitiva, los significados de *creo*, así como los de otras palabras polisémicas, pueden concebirse como continuos y ordenables a partir de la ganancia o pérdida de algunos semas. Ahora bien, lo que esta concepción del significado conlleva es que existan valores básicos, prototípicos, desde los que derivan extensiones significativas. Luego los significados nucleares sí que componen inventarios cerrados y, por consiguiente, pueden ser estudiados de una manera más sistemática que si el investigador se enfrentara a toda la polisemia de elaboraciones y extensiones semánticas en su conjunto que puede generar una palabra.

que se pretende responder es qué método, complementario al análisis cualitativo, puede permitir una descripción más amplia y certera de los usos de unidades complejas como *creo*.

Este capítulo se propone, por consiguiente, aplicar un análisis de *creo*, como forma paradigmática del conjunto de unidades performativas complejas de los verbos cognitivos, desde el paradigma del cognitivismo, el cual ha resultado eficaz para la descripción de otras formas lingüísticas (adverbiales y verbales) de funcionamiento semejante a la que es objeto de estudio (Abdulrahim, 2014; Fetzer y Johansson, 2010; Jansegers, 2017; Jansegers y Gries, 2017; Milin *et al*., 2016; Roldán, 2005), y probar su operatividad. Asimismo, se quiere determinar qué parámetros afectan en el proceso de detección de la semántica y la pragmática de la forma verbal para establecer un protocolo jerárquico de las características observables y que, a partir de estas, pueda certificarse un alto grado de reconocimiento del significado y de la función pragmática de *creo*.

Con este fin, planteamos una metodología de análisis de corpus. Se compila un conjunto de textos disponibles de interacción oral de diferentes géneros discursivos: conversación coloquial y debate parlamentario. Son estos dos los formatos en los que más se ha estudiado hasta la fecha el comportamiento de los verbos cognitivos, tanto en el caso del español como en otras lenguas. Asimismo, se trata de géneros que suponen puntos opuestos de un continuo tanto de formalidad como de otros rasgos como dialogicidad, grado de planificación y determinación en el reparto de los turnos de los participantes, lo que permite obtener un espectro ancho de circunstancias de la oralidad adecuadas para realizar un estudio general de tendencias de uso de *creo* lo más amplio posible. La compilación la conforman textos de conversaciones coloquiales de los corpus *COGILA*, *COJEM*, *Val.Es.Co. 2002* y *Val.Es.Co. 2.0*; y sesiones de debate parlamentario del *Congreso de los Diputados* (del Gobierno de España) y de *Les Corts Valencianes* (del gobierno autonómico de la Comunitat Valenciana), en una proporción equitativa. En el caso de los corpus conversacionales, se analizan en su totalidad el COGILA (36 000 palabras); el COJEM (100 000 palabras); Val.Es.Co. 2002 (91 366 palabras); y Val.Es.Co. 2.0 (128 394 palabras). De los corpus parlamentarios se obtiene, de manera aleatoria, una muestra de una cantidad similar de palabras, repartida esta entre las dos fuentes: *Congreso de los Diputados*, 177 522 palabras; *Les Corts Valencianes*, 174 366. La siguiente tabla resume esta base de la muestra:

Tabla 1. Datos de la muestra, base para el análisis.

Género	N. º palabras	N. º casos *creo*
conversación coloquial	355 760	427
debate parlamentario	351 888	303
TOTAL	**707 648**	**730**

Como se observa en esta Tabla 1 ilustrativa, de los corpus se extraen manualmente los ejemplos de *creo* (bien con buscadores de los archivos de PDF manejados para el caso de los debates parlamentarios, bien a través de la escucha de las conversaciones coloquiales grabadas). Estos suponen un total de 730 casos, los cuales se analizan desde el punto de vista cualitativo, mediante la observación de 30 variables de análisis determinadas en análisis previos (Soler, 2019), bajo un criterio de aproximación cognitiva, sobre todo, aquellos que realizan una aproximación semántica y funcional a *creo*. Se trata de las siguientes variables:

I. Parámetros formales

MORFOSINTÁCTICOS

1. Construcción de *creo*,
2. Integración parenticidad de *creo* en la cláusula,
3. Sujetosintácticode *creo*,
4. Pronominalización del objetodirectode *creo*,
5. Pronominalización del objetoindirectode *creo*,
6. Negación de *creo*,
7. Posición sintáctica de *creo*,
8. Negación del verboregidopor *creo*,
9. Persona y númerodel verboregidopor *creo*,
10. Tiempoverbal del verboregidopor *creo*,
11. Modoverbal del verboregidopor *creo*;

DE COAPARICIÓN

12. Coaparición *creo* + marcadores del discurso,
13. Coaparición *creo* + formas y estructuras lingüísticas relevantes en el reconocimientode su semántica/pragmática, norepetidas,
14. Coaparición *creo* + formas y estructuras lingüísticas relevantes en el reconocimientode su semántica/pragmática, repetidas en el cotexto;

II. Parámetros semánticos

ARGUMENTATIVOS

15. Tipode argumentoen el que se sitúa *creo*,
16. Polifonía de *creo*;

DE SIGNIFICADO

17. Naturaleza factual del predicadode *creo*,
18. Compartición de las pruebas odatos para avalan lodichocon *creo*,
19. Gradode subjetividad,
20. Tipode intervención en la que aparece *creo*,
21. Gradode convencimientodel hablante sobre loexpresado,
22. Valor semánticobásicomanifestadopor *creo*;

III Parámetros discursivo-funcionales

PRAGMÁTICOS

23. Tipos de actos de habla de *creo*,
24. Posición discursiva de *creo*,
25. Gradode asertividad,
26. Funciones pragmáticas;

SOCIOPRAGMÁTICOS

27. Actividades de imagen;

PARALINGÜÍSTICOS

28. Otros aspectos relevantes;

IV. Parámetros textuales

29. Tipología textual de la secuencia de *creo* ,
30. Género discursivo.

Realizado el análisis cualitativo con la observación de los aspectos cifrados en las variables previas sobre el total de los 730 casos obtenidos, pasamos a realizar el análisis cuantitativo principal que este trabajo presenta. Este consiste en la aplicación de una estadística exploratoria (mediante tablas de contingencia comunes) que permite discriminar algunos datos básicos (ej. la determinación de algunos resultados semánticos, a partir de algunos aspectos formales de las manifestaciones del verbo). Tras ello, se propone un modelo de análisis de estadística descriptivo-predictiva basado en un protocolo de tres pasos: 1. regresiones logísticas, 2. obtención de valores de verosimilitud de cruces de las variables en la determinación del grado de explicación sobre la semántica y sobre la pragmática de *creo*, y 3. cálculo de errores. Todos estos cálculos se realizan en una programación experimental de 1 000 iteraciones, mediante el programa STATA. Ahora bien, para poder aplicar las pruebas estadísticas, se crea un corpus ampliado en el que se aumentan los datos hasta llegar a un mínimo de 5 casos por cada variante de las contenidas por variable aplicada (ya que se trata del número mínimo de casos para que los que las pruebas estadísticas pueden arrojare resultados significativos). Estos ejemplos se obtienen de los corpus *COLAm* y *CORPES XXI*, para el caso de la conversación coloquial, y de otras sesiones no consultadas de las mismas fuentes parlamentarias, para el caso del debate. La Tabla 2 resume los datos de *creo* extraídos del corpus ampliado (un total de 865 casos), sobre los que se aplica el protocolo de análisis, frente a los del corpus base (730 ejemplos).

Tabla 2. Datos de los corpus *base* y *ampliado* , para el análisis estadístico predictivo significativo

	Corpus base	**Corpus ampliado**
Ocurrencias de *creo*	730	865

Las regresiones logísticas que planteamos para este análisis son de tipo multinomial. Las regresiones son un cálculo predictor sobre la incidencia de una variable dependiente (Y) sobre una independiente (X). El valor de la regresión ($Y \approx \beta o + \beta 1X$) permite obtener un coeficiente que cifra la estimación de los valores, el cual se denomina *R*2. Ahora bien, cabe destacar que esta prueba estadística presupone linealidad entre las variables. Dado que esta no se da entre aspectos cualitativos de análisis lingüísticos como el que presentamos y, por consiguiente, el cálculo obtenido en el primer paso no es exacto, en un segundo paso o instancia calculamos complementariamente un número de verosimilitud de la relación entre las variables cotejadas. Lo hacemos a partir del modelo de McFadden, el cual permite obtener, frente a la estimación de valor de R^2, un valor probabilístico de pseudo-R^2. Este valor permitirá ordenar jerárquicamente las variables preestablecidas de mayor a menor grado de expli-

cación sobre el valor semántico de *creo*, por un lado, y sobre el valor pragmático, por otro. Finalmente, como este cálculo no es exacto y se realiza sobre 1 000 repeticiones del experimento, se calculan posibles errores a partir de la creación de dos variables: la máxima probabilidad de acierto del resultado y la mínima probabilidad de esta. Estas también se entrecruzan con las previas para obtener las diferencias y el margen de error.

En lo que sigue, se verán los resultados obtenidos de la aplicación de este modelo de análisis. Asimismo, se comprobará su viabilidad como metodología replicable.

5. Análisis y discusión de los resultados

El análisis efectuado sobre la semántica y la pragmática de *creo* ha ofrecido resultados en diferentes planos. En lo que sigue, presentamos los obtenidos en cada fase del estudio, y un resumen del modelo metodológico aplicado, el cual puede considerarse también como un resultado de la investigación.

5.1. Fases del análisis

El primer resultado que ofrece el acercamiento cognitivo a los usos discursivos de *creo* ha permitido reconocer cinco valores semánticos básicos: *creencia*, *certeza*, *conjetura*, *predicción* y *juicio*, de acuerdo con lo apuntado en estudios previos (Soler, 2018; 2019)[4]. Estos valores se ordenan en un continuo de subjetividad, según el grado de implicación del hablante en la escena que proyecta. Nuestro estudio estadístico descriptivo del corpus base, así como del corpus ampliado, permite ver que la construcción formal de *creo* determina en el 100 % de los casos alguno de estos cinco valores. Véanse las tablas de contingencia 3 y 4:

4 Si bien la descripción de los significados de *creo* excede los objetivos de este capítulo (véase, para ello, Soler 2021), describimos mínimamente los semas básicos de cada uno de ellos para aclarar su lectura. El valor de *creencia* describe la adhesión completa del hablante a lo dicho, con independencia de las pruebas que se tengan para ello (ej. *creo en dios*). El valor de *certeza* describe verdades que son absolutas únicamente para el propio hablante, el cual también las presenta como independientes de su comprobación (ej. *me lo creo*). El valor de *conjetura* hace referencia a un cálculo realizado por el hablante cuando este dispone de algunas pruebas sobre lo dicho (ej. *creo que fue ayer*). La *predicción*, como la *conjetura*, se basa en algunas pruebas, pero se proyecta sobre hechos futuros (ej. *creo que viene mañana*). Por último, el *juicio* manifiesta una opinión personal, basada en la comprobación de lo dicho, que ahora no es factual, sino que se basa en la única escala de valores que son los personales del hablante (ej. *creo que eso no está bien*).

Tabla 3. Cruce de datos obtenidos entre la construcción y el valor semántico de *creo* , con prevalencia del valor semántico (Soler, 2019).

Variable formal (1): Construcción						
24. SIGNIFICADO	**creencia**	**certeza**	**conocimiento**	**posibilidad**	**juicio**	**intr. impreciso**
no creo en	**100 %**	0 %	0 %	0 %	0 %	0 %
(no) me (lo) creo (X)	0 %	38,46%	0 %	0 %	0 %	0 %
ya lo creo (X)	0 %	38,46%	0 %	0 %	0 %	0 %
(no) lo creo	0 %	21,15%	0 %	7,89%	0 %	0 %
creo	0 %	0 %	11,67%	0 %	4,79%	0 %
no creo	0 %	0 %	0 %	17,54%	0 %	0 %
creo que	0 %	0 %	59,17%	34,21%	75,80%	**100 %**
no/tampoco creo que	0 %	0 %	0,00%	35,09%	0,23%	0 %
sí/también creo que	0 %	0 %	1,25%	0 %	3,88%	0 %
creo que no/Æ o verbo	0 %	0 %	7,92%	5,26%	4,57%	0 %
creo que sí/también + Æ o verbo	0 %	1,92%	9,58%	0 %	2,74%	0 %
X + creo + PVO del OD/ pron. + creo + CC	0 %	0 %	0 %	0 %	4,57%	0 %
creo + infinitivo	0 %	0 %	8,33%	0 %	0 %	0 %
lo que creo + Æ o verbo	0 %	0 %	2,08%	0 %	3,42%	0 %

Tabla 4. Cruce de datos obtenidos entre la construcción y el valor semántico de *creo* , con prevalencia de la construcción (Soler, 2019).

(1) CONSTRUCCIÓN						
24. SIGNIFICADO	**creencia**	**certeza**	**conocimiento**	**posibilidad**	**juicio**	**intr. impreciso**
(no) creo en	**100 %**	0,00%	0 %	0 %	0 %	0 %
(no) me (lo) creo (X)	0 %	**100 %**	0 %	0 %	0 %	0 %
ya lo creo (X)	0 %	**100 %**	0 %	0 %	0 %	0 %
(no) lo creo	0 %	55 %	0 %	45 %	0 %	0 %
creo	0 %	0 %	57,14%	0 %	42,86%	0 %
no creo	0 %	0 %	0 %	**100 %**	0 %	0 %
creo que	0 %	0 %	27,63%	7,59%	64,59%	0,19%
no/tampoco creo que	0 %	0 %	0 %	97,56%	2,44%	0 %
sí/también creo que	0 %	0 %	15 %	0 %	85 %	0 %
creo que no/Æ o verbo	0 %	0 %	42,22%	13,33%	44,44%	0 %
creo que sí/también + Æ o verbo	0 %	2,78%	63,89%	0 %	33,33%	0 %
X + creo + PVO del OD/ pron. + creo + CC	0 %	0 %	0 %	0 %	**100 %**	0 %
creo + infinitivo	0 %	0 %	**100 %**	0 %	0 %	0 %
lo que creo + Æ o verbo	0 %	0 %	25 %	0 %	75 %	0 %

Como puede observarse, tanto el valor de *creencia* (*(no) creo en*) como el de *certeza* (*(no) me (lo) creo (X)* o *ya lo creo (X)*) vienen determinados en el 100 % de los casos por una construcción concreta de *creo*. Asimismo, en la totalidad de los casos analizados en los que aparece una construcción concreta de *creo*, el valor semántico reconocido es el mismo, si bien esta relación no se da ahora siempre en el sentido inverso. Se trata de *creo* + *infinitivo*, que conlleva el valor de *conjetura*; y *no creo*, que expresa predicción; y *X* + *creo* + *PVO del OD/ pron.* + *creo* + CC, asociada al *juicio*. De ello se desprende que la aproximación cognitiva es eficaz, y que la estadística descriptiva ofrece una prueba patente de ello, pues certifica la viabilidad de las pruebas para discernir algunas de las relaciones de variables determinantes en el reconocimiento, en este caso, del valor semántico de *creo*. No obstante, no es determinante para el reconocimiento de su pragmática, ni explica todos los valores semánticos que ha distinguido el enfoque cognitivo aplicado. Por consiguiente, en una segunda fase del estudio, se aplica la estadística predictiva al corpus ampliado, con el fin de alcanzar resultados más concretos.

Implementamos la metodología diseñada a partir de sucesivas pruebas de regresiones logísticas previas al corpus base ampliado. El método de ensayo y error nos permite obtener un protocolo de actuación ordenado y aplicado, finalmente, para 1 000 iteraciones, mediante STATA. Este experimento lo realizamos dos veces ya que, al no tratarse de un cálculo exacto (porque las variables cotejadas son cualitativas) las pruebas son de realización extensa y apenas puede llegarse a un valor de verosimilitud, y no a un 100 % de exactitud, aunque sí lo más cerca posible de este porcentaje. Así, en una primera instancia, se aplica una regresión logística multinomial tomando como variable dependiente la relativa al valor semántico de *creo*, lo cual se lleva a cabo para 1 000 iteraciones o repeticiones. Tras ello, se repite el proceso, esta vez partiendo de la variable de la función pragmática como dependiente, con el mismo número de repeticiones. En el siguiente apartado aportamos los resultados obtenidos en ambas repeticiones del protocolo diseñado, siguiendo los pasos concretos y ordenados del diseño.

5.2. Resumen del diseño de un modelo predictivo de los valores de *creo* en tres fases

En el modelo diseñado para el análisis predictivo de los valores de significado de *creo*, en primer lugar, y de sus funciones pragmáticas, en segundo lugar, determinamos para comenzar *(1) la capacidad explicativa de las variables cotejadas*. Obtenemos una tabla como la que sigue con los valores de R^2 de McFadden por cada uno de los cruces de variables:

Tabla 5. R^2 de McFadden para las variables independientes en la determinación del significado de *creo* (Soler, 2019).

Modelos de regresión multinomial	Log Likelihood	*pseudo-R2* de McFadden (1.ª INSTANCIA)
SIGNIFICADO (sin variables)	-753.04064	
SIGNIFICADO - CONSTRUCCIÓN	-570.74281	0,242082326
SIGNIFICADO - INTEGRACIÓN	-744,02822 (el modelo converge)	0,011968039
SIGNIFICADO - OD	-718,37835 (el modelo converge)	0,046029773
SIGNIFICADO - OI	-751,82063 (el modelo converge)	0,001620112
SIGNIFICADO - NEGACIÓN V.	-647,25325 (el modelo converge)	0,14048032
SIGNIFICADO - NEGACIÓN V. SUB.	-751,94636 (el modelo converge)	0,001453149
SIGNIFICADO - PERS. Y NÚM. V. SUB.	-683,25781 (el modelo converge)	0,09266808
SIGNIFICADO - TIEMPO V. SUB.	-639,50454 (el modelo converge)	0,150770216
SIGNIFICADO - MODO V. SUB.	-639,71697 (el modelo converge)	0,15048812
SIGNIFICADO - SUJETO	-730,20534 (el modelo converge)	0,030324127
SIGNIFICADO - POSICIÓN SINT.	-704,22015 (el modelo converge)	0,064831149
SIGNIFICADO - REPETICIONES	-706,214 (el modelo converge)	0,062183417
SIGNIFICADO - MMDD	-727,2134 (el modelo converge)	0,034297272
SIGNIFICADO - OTROS ELEMENTOS	-642,77243 (el modelo converge)	0,146430623
SIGNIFICADO - GÉNERO	-667,35646 (el modelo converge)	0,113784271
SIGNIFICADO - TIP. TEXTUAL	-598,01081 (el modelo converge)	0,205871797

Seguidamente, a partir de estos datos, se calcula *(2) la jerarquía de las variables en el aumento paulatino de explicación* que proporcionan sobre el significado de *creo*. En la ordenación de esta jerarquía, nos fijamos en el valor de verosimilitud proporcionado por R^2 de McFadden, si bien también se tienen en cuenta cuestiones cualitativas de aplicación de las variables al análisis. Así, por ejemplo, se observa cualitativamente que las características de tipo formal son más rápidamente reconocibles por parte del analista (las cuales subimos en la escala de jerarquía), y que otras de tipo semántico presentan una detección más compleja (razón por la que, en algunos casos, las relegamos a puestos inferiores de la jerarquía de aplicación). Así mostramos los resultados de la segunda instancia en dos tablas. Primeramente, observamos que en la Tabla 6 aparecen todos los resultados de verosimilitud obtenidos. Seguidamente, en la Tabla 7 reordenamos los parámetros de análisis de mayor a menor grado de explicación sobre el valor semántico del verbo y añadimos el porcentaje de error que este pueda estar generando.

Tabla 6. R^2 de McFadden ordenados por valor, en la determinación del significado de *creo* (Soler, 2019).

	Variables jerarquizadas	*Pseudo-R*2 de McFadden porcentual (2.ª instancia)
1	TIPOLOGÍA TEXTUAL	25,44 %
2	CONSTRUCCIÓN	24,21 %
3	TIEMPO VERBO SUBORDINADO	16,43 %
4	OTROS ELEMENTOS	15,58 %
5	GÉNERO	12,43 %
6	PERSONA Y NÚMERO VERBO SUB.	6,44 %
7	POSICIÓN SINTÁCTICA	6,42 %
8	SUJETO	5,36 %
9	NEGACIÓN *CREO*	2,94 %

Tabla 7. R^2 de McFadden reordenados por jerarquía de aplicación, en la determinación del significado de *creo* (Soler, 2019).

Variables ordenadas	Error común estándar
CONSTRUCCIÓN	43,82 %
SUJETO	32,80 %
NEGACIÓN *CREO*	44,45 %
TIEMPO VERBO SUBORDINADO	30,11 %
PERSONA Y NÚMERO VERBO SUB.	32,80 %
POSICIÓN SINTÁCTICA	45,70 %
OTROS ELEMENTOS	27,82 %
TIPOLOGÍA TEXTUAL	20,70 %
GÉNERO	31,72 %

En la tabla 7 vemos cómo, en el último paso de nuestro protocolo **(3) se obtiene un error ajustado de los cálculos realizados.** Este permite ver que no ha habido desfases entre la extracción de los valores de verosimilitud de las tablas previas y los de la probabilidad total de que se reconozcan los datos de cada variable. Dado que, en este caso, para la semántica de *creo*, todos los valores obtenidos son menores al 50 % y no presentan diferencias relevantes respecto a los datos de verosimilitud de las regresiones llevadas a cabo, no se plantea una nueva reorganización en la jerarquía de aplicación de las variables, respecto a la ya propuesta.

Para el caso de la determinación de las funciones pragmáticas de *creo*, que se han establecido en las tres categorías reconocidas por la bibliografía previa (a saber, *atenuación*, *neutralidad* e *intensificación*), dado que el análisis cualitativo cognitivo aplicado las reco-

noce, efectivamente, en los mismos términos, se repite el experimento de tres fases diseñado, pero ahora, sobre la base del significado, ya reconocido gracias a la aplicación de las fases de análisis explicadas. Véase la tabla final obtenida:

Tabla 8. R^2 de McFadden reordenados por jerarquía de aplicación, en la determinación de la función pragmática de *creo* (Soler, 2019).

Regresión multinomial	Log Likelihood	*Pseudo-R2* de McFadden (1.ª instancia)
FUNCIÓN-**SDO**	-1227,0094	21,37 %
FUNCIÓN-SDO-**INTERSUBJLOC.**	-1117,7345	8,91%
FUNCIÓN-SDO-**POSIC. SINT.**	-1.125	8,34%
FUNCIÓN-SDO-**GEN.**	-1132,3951	7,71%
FUNCIÓN-SDO-**IMAGEN**	-1.134	7,62%
FUNCIÓN-SDO-**POLIF.**	-1142,6586	6,87%
FUNCIÓN-SDO-**ASERTIVIDAD**	-1.144	6,78%
FUNCIÓN-SDO-**OTROSELS.**	-1147,1233	6,51%
FUNCIÓN-SDO-**TXT.**	-1148,9373	6,36%
FUNCIÓN-SDO-**CONVENC.**	-1164,7245	5,08%

En este segundo experimento, se parte de que el significado de *creo* ya ha sido establecido con la primera aplicación del protocolo. De este modo, se reduce el número de variables en el cálculo de la función pragmática. En la obtención de errores, se estima que estos, de nuevo, no alteran los datos de verosimilitud de R^2 de McFadden y, por consiguiente, el orden y jerarquía de aplicación de las variables para la determinación de la función pragmática de *creo* se mantiene como muestra, más arriba, la Tabla 8.

El análisis demuestra, pues, que del total de variables cognitivamente descritas para el posible análisis semántico-pragmático de creo, solo algunas de ellas son eficaces con más de un 20 % de explicación y hasta más de un 50 %, mientras que otras, pueden descartarse, al menos, en un estudio genérico para detectar lo más automáticamente posible ante qué tipo de *creo* nos encontramos.

6. Conclusiones

El análisis de este capítulo confirma que es posible diseñar un modelo de análisis cuantitativo que, siempre como complemento del análisis cualitativo de fenómenos lingüísticos como el del funcionamiento de las formas performativas de los verbos cognitivos, permite determinar más del 60 % de sus valores semánticos, así como entre el 80 y el 100 % de sus funciones pragmáticas (si sumamos el valor de verosimilitud de la aplicación de las varia-

bles jerarquizadas en el protocolo). Asimismo, el modelo planteado para el caso de *creo* es replicable en otro tipo de textos y géneros discursivos. En contraposición con las carencias metodológicas de la bibliografía previa, el modelo de análisis creado mejora y perfecciona la aplicación de pruebas estadísticas que han resultado insuficientes en otros estudios. Este hecho confirma la hipótesis de partida de este trabajo, ya que es posible completar el análisis cualitativo de *creo* con el acercamiento cuantitativo riguroso no solo descriptivo, sino también predictivo. Este modelo de análisis puede describirse a partir de tres fases: (1) determinación de la capacidad explicativa de las variables seleccionadas con criterios cognitivos, (2) jerarquización de las variables para la descripción semántica y pragmática de *creo* (o la forma verbal considerada), y (3) cálculo de errores cometidos en el proceso, las cuales dan respuesta a la pregunta de investigación del trabajo, la cual se cuestionaba si era posible llegar a una sistematización de análisis para el reconocimiento semántico-funcional de unidades lingüísticas complejas como la que nos atañe.

En conclusión, cabe decir que este capítulo ha pretendido ofrecer un paradigma de estudio que es compatible con los ya conocidos, pero que viene a completar los puntos que no habían sido solventados hasta ahora por la investigación lingüística más tradicional. Queda para el futuro próximo replicar este patrón propuesto y perfilar el modelo de análisis y las fases de su consecución, así como también cotejar los resultados específicos que pueda dar su aplicación a otros formatos textuales y fenómenos lingüísticos.

Referencias

Abbhul, R. & Mackey, M. (2013). Experimental research design. In R. Abbuhl, S. Gass & M. Mackey, *Research Methods in Linguistics* (pp. 116-134). Cambridge University Press.

Abdulrahim, D. (2014). Annotating corpus data for a quantitative, constructional analysis of motion verbs in Modern Standard Arabic. In N. Habash, & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing* (*ANLP*) (pp. 28-38). Association for Computational Linguistics.

Achard, M. (1998). Representation of cognitive structures. *Cognitive Linguistics*, *15*(4), 588-594.

Boas, H. & Ziem, A. (2018). Constructing a constructicon for German. Empirical, theoretical, and methodological issues. In B. Lyngfelt, L. Borin, K. Ohara, & T. Timponi (Eds.), *Constructicography: Constructicon development across languages* (pp. 183-228). John Benjamins. https://doi.org/10.1075/cal.22.07boa.

Buceta, O. (2014). Construcciones del verbo 'creer'. *Factótum*, *12*, 74-90.

Cutting, J. (Ed.). (2007). *Vague Language Explored*. Palgrave MacMillan.

De Hoop, H., Foolen, A., Mulder, G. & Van Mulken, V. (2018). *I think* and *I believe*: Evidential expressions in Dutch. In A. Foolen, H. de Hoop & G. Mulder (Eds.), *Evidence for Evidentiality* (pp. 77-97). John Benjamins. https://doi.org/10.1075/hcp.61.04hoo.

Díaz-Campos, M. & Gradoville, M. (2011). An Analysis of Frequency as a Factor Contributing to the Diffusion of Variable Phenomena: Evidence from Spanish Data. In L. Ortiz (Ed.), *Selected Proceedings of the 13th Hispanic Linguistics Symposium*, (pp. 224-238). Cascadilla Proceedings Project.

Fetzer, A. (2014). *I think*, *I mean* and *I believe* in political discourse. Collocates, functions and distribution. *Functions of Language*, *21*(1), 67-94.

Fetzer, A. & Johansson, M. (2010). Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse. *International Journal of Corpus Linguistics*, *15*(2), 240-266.

Fuentes Rodríguez, C. (2010). La aserción parlamentaria: de la modalidad al metadiscurso. *Oralia*, *13*, 97-125.

Fuentes Rodríguez, C. (2016). Atenuación e intensificación estratégicas. In C. Fuentes Rodríguez (Ed.), *Estrategias argumentativas y discurso político* (pp. 163-222). Arco/Libros.

González Ruiz, R. (2015). Los verbos de opinión entre los verbos parentéticos y los verbos de rección débil: aspectos sintácticos y semántico-pragmáticos. *Círculo de Lingüística Aplicada a la Comunicación*, *62*, 148-173.

González, J., Boeck, P. & Tuerlinchx, F. (2014). Linear mixed modelling for data from a double mixed factorial design with covariates: a case-study on semantic categorization response times. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(2), 289-302.

Hartwell, L. M., Esperança-rodier, E. & Tutin, A. (2017). *I think we need…*: Verbal expressions of opinion in conference presentations in English and in French. *Romance Corpora and Linguistic Studies*, *4*(1), 35-60.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Jansegers, M. (2017). *Hacia un enfoque múltiple de la polisemia. Un estudio empírico del verbo multimodal "sentir" desde una perspectiva sincrónica y diacrónica*. Mouton de Gruyter.

Jansegers, M. & Gries, S. (2017). Towards a dynamic behavioral profile: a diachronic study of polysemous 'sentir' in Spanish". *Corpus Linguistics and Linguistic Theory*, *16*(1), 145-187.

Milin, P., Divjak, D., Dimitrijević, S. & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, *27*(4), 507-526.

Roldán, A. (2005). Applications of cognitive linguistics (CI) to languages for specific purposes (LSP). In M. L. Carrió (Coord.), *Perspectivas interdisciplinares de la lingüística aplicada*, Vol. 2 (pp. 325-332). Universitat de València.

Soler, M. A. (2018). Algunos apuntes bibliográficos en torno a los verbos de opinión. In C. J. Álvarez López & M. R. Martínez Navarro (Coords.), *En busca de nuevos horizontes. Algunas líneas actuales en los estudios hispánicos* (pp. 59-70). Ediçoes Húmus,

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

Soler, M. A. (2021a). Análisis cognitivo de la semántica de *creo* en el español occidental hablado. En L. E. Aguilera, E. de los Santos, M. E. Flores & J. Haidar (Eds.), *Enfoques alternativos en los estudios del discurso* (pp. 92-107). Universidad Autónoma de Nuevo León.

Soler, M. A. (2021b). *Semántica de* creo. *Análisis cognitivo de la polisemia de una forma verbal doxástica en la interacción oral en español*. Peter Lang.

Chapter XIII

Use of Bayesian networks for the analysis of corpus of local problems related to the Sustainable Development Goals

Uso de redes Bayesianas para el análisis de corpus de problemas locales relacionados con los Objetivos de Desarrollo Sostenible

Manuel Caro Piñeres & Ernesto Llerena García
Universidad de Córdoba – Colombia

Abstract: Bayesian networks are a widely used formalism for data analysis, modelling, and decision support in various domains. Currently, there is a need for techniques and tools that automatically build Bayesian networks from massive text or literature data. Collecting people's perception of the problems they face in their daily lives generates a great deal of textual information. Textual descriptions increase as new data collections are made. Due to the lexical differences between different regions of a country, it is necessary to constantly update the new modelled data.

Resumen: Las redes bayesianas son un formalismo ampliamente utilizado para el análisis de datos, el modelado y el apoyo a la toma de decisiones en varios dominios. Actualmente, existe la necesidad de técnicas y herramientas que construyan automáticamente redes bayesianas a partir de textos masivos o datos bibliográficos. La recopilación de la percepción de las personas sobre los problemas que enfrentan en su vida diaria genera una gran cantidad de información textual. Las descripciones textuales aumentan a medida que se realizan nuevas recopilaciones de datos. Debi-

do a las diferencias léxicas entre las diferentes regiones de un país, es necesario actualizar constantemente los nuevos datos modelados.

1. Introduction

The 17 Sustainable Development Goals (SDGs) are a plan of the United Nations to achieve a better and more sustainable future for people and the planet by 2030. In these goals there are aspects related to poverty, hunger, good health and well-being, quality education, clean water, clean energy among others. With just under ten years left to achieve the Sustainable Development Goals, world leaders at the SDG Summit in September 2019 called for a Decade of Actionand delivery for sustainable development, and pledged to mobilize financing, enhance national implementation and strengthen institutions to achieve the Goals by the target date of 2030, leaving no one behind. Thus, it was necessary to use reliable technology for understanding people's needs all around the world, and during this decade achieve the 17 Sustainable Development Goals (SDGs) lead by the United Nations. In that way, Bayesian network was used for collecting data through a software created by EduTLan group which helps to gather and analyze all the information needed to reach these goals. Bayesian networks are used for modelling knowledge in computational biology and bioinformatics, learning, medicine, biomonitoring, document classification, information retrieval, semantic search, image processing, data fusion, decision support systems, engineering, games and law. For decision-making at the governance level, it is necessary to know how non-compliance with the SDGs affects the well-being of the population. However, the SDGs are little known by the general population, so it is necessary to have techniques that can relate people's speech in relation to the language of the SDGs. To fulfil this purpose, it is necessary to collect many descriptions of problems related to the SDGs in the communities.

The main goal of this study is to describe the process of collecting, organizing, tagging and validating a corpus of more than 3,000 descriptions of problems related to compliance of the SDGs in three regions in Colombia. The main result of this study was a large digital corpus of descriptions of problems related to compliance of the SDGs in three regions in Colombia. The potential of the corpus was verified by evaluating the results of a Bayesian network algorithm. In the evaluation, the standard processing of the text by the algorithm produces a high rate of correct answers.

The rest of the paper is organized as follows. Section 2 describes the theoretical framework that supports this research. Section 3 summarizes the methodological framework

based on Design Science Research (DSR) used to design the machine learning approach based on Bayesian networks for the analysis of corpus of local problems related to the Sustainable Development Goals. In Section 5 the results are describes. Finally, the conclusions are presented.

2. Theoretical framework

For this research work, theoretical elements on structural semantics and digital lexicography were used. Lexicon organization of the corpus from selected words were done based on the structural semantics proposed for the semantic fields as well as the other levels of the linguistic structure that have a structural nature and functioning. For this reason, this position was welcomed on this research, and it is applied for the collection of information. According to this structural organization, the entire lexicon must be organized into semantic fields. A semantic field, in linguistics, is one that makes up a group of words that share one or more features in their meaning. This semantic field is organized through hypernyms and hyponyms (In this investigation the term holonym is related to hypernym and the word meronym is related to hyponym. Theoretically, the difference between hypernyms-hyponyms and hollonyms-meronys is that the former has conceptual inclusion and the latter have material inclusion -i.e., part of-). A hypernym is a general term that can be used to refer to the reality named by a more specific term.

For this research, each field is equivalent to the following development objectives, which functioned as hypernyms: no poverty, zero hunger, good health and well-being, quality education, gender equality, clean water and sanitation, affordable and clean energy, industry, innovation and infrastructure, reduced inequalities, sustainable cities and communities, *responsible consumption and production, climate action, life below water, life on land, peace, justice and strong institutions, partnerships for the goals*. Each one of these referential fields presents, in turn, relations of hyponymy. The hyponyms are words that have all the semantic features, or semes, of a more general one – its hypernym – but that in its definition adds other semantic characteristics that differentiate it from others. The hyponyms of each hyperonym were determined, so when the words that the interviewee was saying were extracted from the recordings, they were distributed according to each hypernym and the default hyponyms for each one. For example, the hypernym **no poverty** has the following hyponyms: *displaced women, social security, extreme poverty, poverty line, multidimensional poverty, multidimensional poverty index. For a more related relation.*

This form to extract semantic relations of related words was based primarily from the digital lexicography; the basic approaches of semantic organization were led by the way

Wordnet was elaborated. WordNet is an electronic lexical reference system, developed in the form of a lexical database, created by the psycholinguist George A. Miller which is in line with psycholinguistic theories regarding the organization of lexical information in the mind of the speaker (Baars, 1986). WordNet is a project that was supported from the beginning by various US government and private institutions: The Department of Naval Research, the James S. McDonnell Foundation and Princeton University. Apart from being an example of government and public cooperation, it is also a project whose results have been made public and can be freely distributed for academic purposes. WordNet is available to any user who wishes to consult its resources through the internet and the system can be used in online mode (See http://wordnet.princeton.edu/). The primary objectives of WordNet, and that The following are fundamental bases in the elaboration of this software: a) The validation of psycholinguistic theories on lexical organization; b) Its foreseeable use in various applications that require access to lexical information The basic difference between this and other projects for the implementation of computational lexicons is that it is the only relatively large-scale project in which the organization of the Lexis in semantic fields can handle information for the purpose of gathering semantic approaches. In fact, the main motivation for its realization has been the idea of testing, through its direct implementation in a digital computer, psycholinguistic and lexicological theories regarding the structure of the mental lexicon. Following a model of semantic networks for organizing the mental lexicon, the group of researchers that made up WordNet set out in 1985 to create a tool that would allow moving through the structure of a dictionary conceptually and not just alphabetically. The differences from a traditional dictionary are obvious: WordNet divides the lexicon into five categories: nouns, verbs, adjectives, adverbs, and functional elements. However, Wordnet presents a considerable amount of redundant information that would not appear in a traditional dictionary, in those cases where a word belongs to more than one category.

On the other hand, this type of organization greatly facilitates the analysis of the semantic organization differences that exist between these five syntactic categories, and it is also important to note that, by not having to force the different categories into the same representational scheme, it is possible to search the most suitable way for each one of them separately. WordNet is an attempt to reflect the lexical memory model based on semantic networks proposed by Collins and Quillian (1969) in a lexicographic model of lexical organization. One of the first examples of a semantic memory network model is the TLC (Teachable Language Comprehender) (Collins & Quilliam, 1969). According to this model, each node is a word that represents a concept (such as "bird"). With each node, a series

of properties is stored (such as "can fly" or "has wings"), as well as directions (for example, links) to other related nodes (for example, "dove"). A node is directly linked to those others that are a subclass or a superclass (for example, "bird" would be related to both the "pigeon" subclass and the "animal" superclass). Thus, the TLC model assumes a hierarchical representation of knowledge, in which high-level nodes representing broad categories are connected (either directly or indirectly—Through the nodes of lower classes—) to a multitude of elements belonging to those categories. The nodes that represent concrete examples of these supracategories would be at a lower level, only connected to the immediately higher categories. Also, properties are stored at the highest level of categorization to which they can be applied. For example, "is yellow" could be stored with "canary"; "Has wings" could be stored with "bird" (one level up); and "can move" could be stored with "animal" (another level up).

Nodes can also store the negation of the properties of their superordinate nodes (for example, "can't fly" could be stored with "penguin"). This provides an economy of representations, in which properties are only stored at the level of categorization for which they are essential, that is, at the point where they become critical characteristics. According to the TLC, processing is a form of activation propagation, that is, when a node is activated, the activation spreads to other nodes through the links that join them. In that case, the response time to the question "Is the pigeon a bird?" it depends on the distance that mediates between the nodes "dove" and "bird" (for example, the number of intermediate nodes that may exist).

3. Methodological framework

This section describes the Design Science Research Methodology (DSRM) (Hevner *et al.*, 2007) used in the present study to address the use of Bayesian networks in the analysis of corpus of local problems related to the Sustainable Development Goals (SDGs).

This study adopted the DSRM due to it seeks to enhance human knowledge with the creation of innovative artifacts and the generation of design knowledge (DK) via innovative solutions to real-world problems. The DSRM approach, followed in this study, has been used before in the development of knowledge-based systems and Natural Language Processing (NLP) Systems. As an example, we could refer to the work of Pereira, Ferreira, & Lopes (2020) in knowledge representation and NLP case study in innovation processes (O'Riain, Curry & Buitelaar, 2012). This study includes the following five steps for the development of a software artifact according to DSRM.

Step 1. Problem identification and motivation. In this stage the objectives for a solution are described. Resources required for this activity include the state of the problem and the importance of its solution.

Step 2. Objectives for a solution. Infer the objectives of a solution from the problem definition and knowledge of what is possible and feasible. Resources required for this include knowledge of the state of problems and current solutions.

Step 3. Design and development. Create the artifact. Such artifacts are potentially constructing, models, methods, or instantiations (each defined broadly) (Hevner *et al*., 2007) or "new properties of technical, social, and/or informational resources (Jarvinen, 2007)".

Step 4. Demonstration. Demonstrate the use of the artifact to solve one or more instances of the problem. This could involve its use in experimentation, simulation, case study, proof, or other appropriate activity.

Step 5. Evaluation. Observe and measure how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from use of the artifact in the demonstration. It requires knowledge of relevant metrics and analysis techniques.

4. Results

This section describes the results obtained from the follow-up of each of the steps of the DSR methodology.

4.1. Problem identification and motivation

In this step the problem formulation for the proposed research approach is stated. The problem is described in the form of functional requirements (Eekels & Roozenburg, 1991; Baskerville, *et al*., 2018). Listed below are some of the functional requirements that are necessary to address the development of a system for translating community problem descriptions into language of the SDGs.

- Collect many descriptions of problems related to the SDGs that affect the communities of different regions of Colombia.
- Relate the documents of the corpus with the language of the SDGs, considering the lexicon of regionalisms.
- Provide graphical reports about the problems that each population describes.
- Develop a model that translates natural language into the language of the SDGs.
- Develop an App that allows the collection, storage and translation of the problems expressed by the communities.

4.2. Define the objectives for a solution

Create a corpus with the descriptions of the problems and actions that are carried out in the communities, which have some relationship with the SDGs. Design a system that translates the problem descriptions of different communities into the language of the SDGs. The system must allow:

- Log in through an account.
- Record an interview by voice and convert it to text.
- Enter the data related to the interviewed user.
- Record the priority topics for the interviewed user.
- Record by voice the three main problems in your community and the system converts it to text.
- Record by voice the three actions that have been taken
- implemented in your community for each of the three problems and the system converts it to text.
- The system, through Artificial Intelligence, reports on the SDGs related to each problem of the interviewee.
- The system, through Artificial Intelligence, reports on the goals of each SDG related to each problem of the interviewee.
- The system reports the percentage that relates each SDG to the problem reported by the interviewee.
- The system incorporates new vocabulary related to the SDGs using machine learning.

4.3. Design and development

The classification method used in this study is Naïve Bayes Classifier, to classify online testimonial data from leading e-traveling sites. The current Naïve Bayes Classifier method has been developed to calculate the probabilistic size of each word and provide an assessment for each class. One of them is the Multinomial Naïve Bayes model developed by Schütze *et al.*, (2008). This method estimates the conditional probability of a token that has a class, as the relative frequency of the word t in the document belonging to the class c. In NBC, the probability of a document *d* (e.g., problem description) being in class *c*, *P(c|d)*, is computed as shown in this equation formula:

$$P(c|d) \propto P(c) \prod_{k=1}^{m} P(t_k|c) \quad (1)$$

The Naïve Bayes Multinomial Method takes into account the number of occurrences of the word t in class c training documents, as well as several existing events.

$$P(t|c) = \frac{t_{ct}}{\sum_{t' \in V} T_{ct'}} \qquad (2)$$

The data collection processing mechanism for training and prediction to be used by the ECHO application has the following phases:

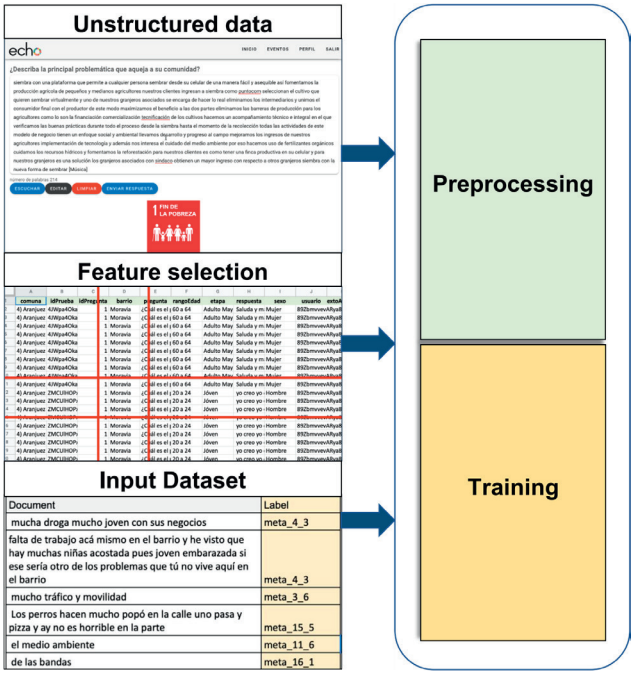


Figure 1. Input data processing and training protocol.

Phase I: Preparation of the initial data matrix (Pre-processing). This matrix can be created as follows:

1 From the cleaning of the matrix extracted from the events carried out or a particular subset of them.

Phase II: Creation of the training dataset. This Dataset can be created in the following way:

2 From experts tagging directed speeches captured by ECHO (SDG translation App).

3 From the review of the application output of an Event

The Training Dataset has two columns, as shown in Figure 2.

Document	Label
mucha droga mucho joven con sus negocios	meta_4_3
falta de trabajo acá mismo en el barrio y he visto que hay muchas niñas acostada pues joven embarazada si ese sería otro de los problemas que tú no vive aquí en el barrio	meta_4_3
mucho tráfico y movilidad	meta_3_6
Los perros hacen mucho popó en la calle uno pasa y pizza y ay no es horrible en la parte	meta_15_5
el medio ambiente	meta_11_6
de las bandas	meta_16_1
la inseguridad y violencia padres que maltratan los niños y en la calle	meta_4_7
mi barrio el principal problema Es la falta de seguridad	meta_11_2
la delincuencia que hay en el barrio el cobro de extorsiones a los tenderos y a toda la gente que tiene su negocio	meta_11_a
Está más suave ya no hay casi venta así como los otros barrios porque sólo como ya hicieron la paz 2010	meta_16_1
la violencia mujeres también hay mucho maltrato Con todo	meta_4_7

Figure 2. Training dataset with two columns: Document and Label.

The theoretical foundation of the system has its origin in the idea of the "vocabulary matrix" (Miller *et al.*,1993) (vocabulary matrix). Miller uses the term lexical form (word form) to refer to the physical expression that is written or pronounced and meaning.

Also, by using this methodology of "nodes" words from interviews were linked to words related and their goals (17 goals in total) as we can see on Table 1.

Table 1. SDG and related words.

SDGS	Related words
Goal 1 No poverty	Displaced women, social security, extreme poverty, poverty line, multidimensional poverty, multidimensional poverty index.
Goal 2 Zero hunger	Agricultural product, agricultural production, agricultural productivity, environment, agricultural sector, safe food.
Goal 3 Good health and well-being	Health centers, environmental sanitation, public health, family planning, reproductive health, sexual health, work accident, work accident.
Goal 4 Quality education	Educational infrastructure, early childhood, vocational training, preschool education, university education, higher education, secondary education, drinking water, educational infrastructure, scholarships available, qualified teachers, teacher training, high enrollment, labor law, adults, literacies, literacy, literate, high enrollment, high school fees, illiterate, illiterate, quality learning, good learning, good school, good teacher, good education, good teaching, good school.

SDGS	Related words
Goal 5 Gender equality	Reproductive health, sexual health, physical violence, sexual violence, psychological violence, forms of violence, sexual exploitation, labor law, reproductive rights, domestic work.
Goal 6 Clean water and sanitation	Open defecation, street poop, environmental sanitation, water resources, drinking water.
Goal 7 Affordable and clean energy	Electric power, street lighting, power service.
Goal 8 Decent work and economy growth	Tax incentives, environment, labor law, child labor, child soldiers, human trafficking, forced labor, labor rights, labor law, safe work, precarious employment, migrant workers, jobs.
Goal 9 industry, innovation and infrastructure	Economic development, rural area, sustainable city, bad connection, slow connection.
Goal 10 Reduced inequalities	Sexual harassment, human rights.
Goal 11 Sustainable cities and communities	Private sector, urban area, public roads, housing project, sports venues, sustainable city, sexual harassment, harsh winter, housing construction.
Goal 12 Responsible consumption and production	Organic waste, environment, solid waste, teacher training, teacher training, material consumption, responsible consumption, sustainable consumption.
Goal 13 Climate action	Secondary education, mitigation activities, improve education, extreme weather conditions, climate change, early warning.
Goal 14 Life below water	Marine biodiversity, marine technology, scientific knowledge, research activities.
Goal 15 Life on land	Drinking water, global warming, climate change, stop deforestation, ecosystems in planning, terrestrial ecosystems.
Goal 16 Peace, justice and strong institutions	Micro-trafficking, armed groups, armed conflict, armed group, extorted, sexual abuse, sexual harassment, criminal gangs, illicit weapons, arms trafficking, workplace harassment, armed conflict, right to vote, human rights.
Goal 17 Partnerships for the goals	Economic development, internet, internet of things.

```
1.  Algorithm_1. Training document by multinomial naïve bayes
2.  Input: Document D, Class C
3.  Output: Vocabulary V, Prior Knowledge, Likelihood condprob
4.  a) Extract vocabulary V from document D
5.  b) Calculate the number of N documents D
6.  c) For every c ∈ C
7.    Calculate Nc as number of D documents that have class c
8.     Calculate prior [c] = Nc / N
9.     Combine all text in document D that has class c into textc
10.    for every t V
11.     Calculate Tct as the number of tokens appearing from textc which has class c
12.      for every t V

13.       Calculate Likelihood condprob [t] [c]= formulae (2)
```

The Naïve Bayes Classifier performance can be improved by using corpus data that has been created and developed in the previous stage. The use of corpus aims to give more weight to the parameters of the probability value, for each token listed in the corpus. The corpus used is the corpus that deals with the topic of hotel parameters, namely comfort, cleanliness, location of the hotel, food, and friendly service.

Corpus value weights are obtained from probabilistic values. The occurrence of the term t on the existing topic, the goal is to normalize the weight. In this study using the proportionality of token numbers for each class c, positive classes p + = 0.65 (for inclusion into a class) and negative p- = 0.35 (for not inclusion into a class) in the data sequence. So that *condprob* can be calculated by a formula such as,

$$score[c] = \sum_{t' \in V} \log(condprob[t][c] \times (1 + (\sum_{t' \in K} wk_{t'} \times p_c))) \qquad (3)$$

To get a score for each class [c] can use the following formulae.

$$condprob[t][c] = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \times (1 + (\sum_{t' \in K} wk_{t'} \times p_c)) \qquad (4)$$

With the knowledge base generated, the algorithm can make inferences and reasoning based on the input from the new interviews to generate predictions regarding the SDGs and targets that are related to the inputs.

```
Belief{
  id: problem_space_ODS
  ISA:[Problem]
  HAS:{
   dataset: learning_dataset_RBHMCM_ODS
  }
}

Belief{
  id: learning_dataset_RBHMCM_ODS
  ISA:[Dataset]
  HAS:{
   cols:[
     has_col_meta_1_x,
     has_col_meta_3_x,
     has_col_meta_4_x,
   ],
   cells:[
     b_a_droga_meta_1_x,
     b_a_droga_meta_3_x,
     b_a_droga_meta_4_x
   ]
  }
}

Belief{
  id: has_col_meta_3_x
  ISA:[ColDataset]
  HAS:{
   head: meta_3_x
   probability: 0,9
  }
}
```

```
Belief{
  id: meta_3_x
  ISA:[TargetODS]
  HAS:{
   holonym: ODS_x
   meronym:[
    droga,
    vicio
   ]
  }
}

Belief{
  id: b_a_droga_meta_3_x
  ISA:[CellDataset]
  HAS:{
   field: droga
   head: meta_3_x
   probability: 0,06
  }
}
```

Figure 3. Semantic network created from the terms processed by the algorithm. Part I.

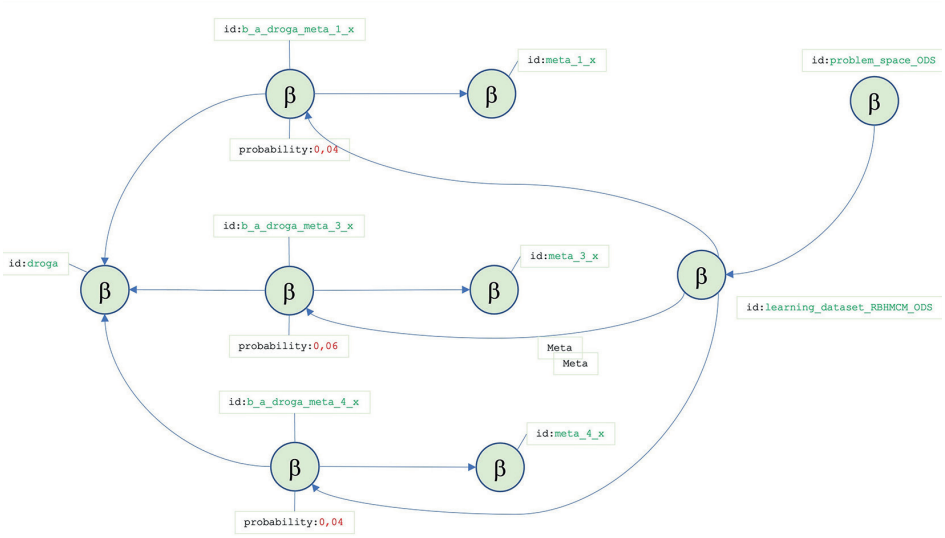


Figure 4. Semantic network created from the terms processed by the algorithm. Part II.

For prediction, the algorithm deployed into the ECHO App captures the information using Speech Recognition. The testing phase based on the results of training data can be used Algorithm_2.

1. Algorithm_2. Testing document by multinomial naïve bayes
2. Input: Class C, Vocabulary V, Prior Knowledge, Likelihood condprob, Test document d
3. Output: arg $max_{c \in C}\, score[c]$
4. Extract token W from test document d based on Vocabulary v b).
5. For each $c \in C$
   Calculate score [c] = $\log prior[c]$
   For every t W
   Calculate score [c] + = $\log condporb[t][c]$
6. Count arg $max_{c \in C}\, score[c]$

The backend of the application and the main algorithm were developed using the framework Nodejs in JavaScript. The front-end was developed with the Vue.js framework, while semantic and procedural memory data were stored in MongoDB. Below is an example of the prediction output for the algorithm in the ECHO App.

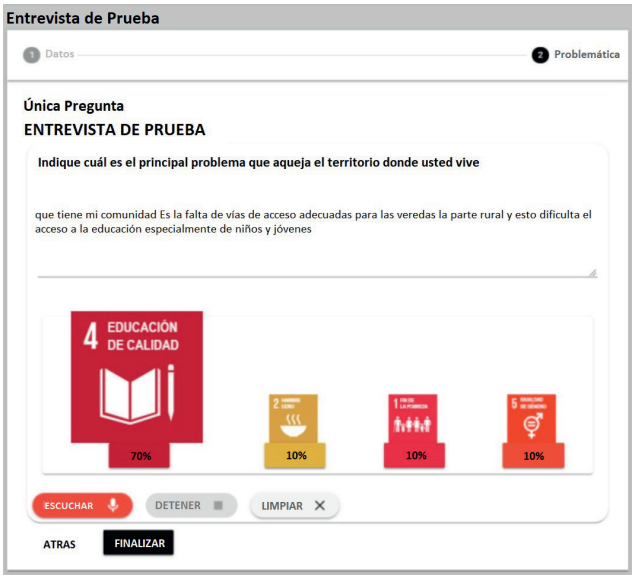


Figure 5. Window for information input.

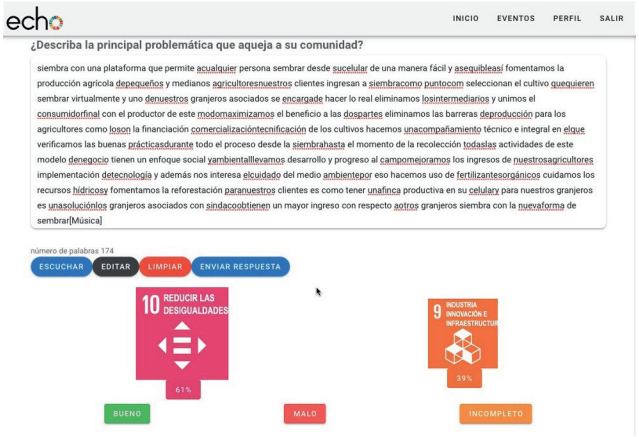


Figure 6. Prediction of the algorithm.

4.4. Demonstration

The descriptions were collected verbally for three years and contain regionalisms related to the SDGs from the Caribbean region, Antioquia and Bogotá. The tool was tested in the cities of Cartagena and Medellín, where the application processed 3456, 5249 and 2345 descriptions of community problems. To facilitate the gathering of testimonies through the ECHO tool, a 5-day information gathering session was held within the framework of the

project "Testing ECHO amplifying the citizen's voices for the SDG's". More than 30 university students participated in the sessions, who were volunteers to collect the problems of the communities and were trained in the use of the ECHO tool. Cell phones with Android operating system, microphone and internet connection were used to collect information.

Figure 7. Event registration window for collecting problems in the community.

4.5. Evaluation

The corpora were taken through oral interviews with people (men and women) from diverse social levels (mainly 1, 2, 3 social levels). The interviewer recorded the interview with a cell phone and instantly or when a WIFI connection was able, all the information was gathered and analyzed. Thus, the system shows how people think about their necessities related to the United Nations's goals. This information will be used to promote prosperity while protecting the planet. Initially, the algorithm presented a level of precision of 84% in the translation of the corpus into the language of the SDGs.

Precision refers to the proportion of concepts that is accurately detected relative to all the concept elements that are represented in the corpus (Brewster *et al.*, 2004). The numerator of Eqs. (1) describe that knowledge that is accurately detected and corresponds to the intersection of the relevant entities and the retrieved entities.

$$precision = \frac{|\{relevantentities\} \cap \{retrievedentities\}|}{|\{retrievedentities\}|} \quad (1)$$

The erroneous results were analyzed by a team of OSDGsDS experts, linguists, and data engineers to determine the causes of the failures. In this process it was found that regionalisms were the main cause, in this sense the application training was refined with a corpus that contained the regionalisms expressed in the problem descriptions. Thus, on the last day of testing in both cities, a precision of 90.1% was obtained.

5. Conclusion

The main result of this study is a large digital corpus of descriptions of problems related to compliance of the SDGs in three regions in Colombia. The potential of the corpus was verified by evaluating the results of a Bayesian network algorithm. In the evaluation, the standard processing of the text by the algorithm produces a high rate of correct answers. The use of semantic methodology for the organization of information in semantic fields was very efficient. Semantic field was organized through hyperonyms and hyponyms which allow to organize all the information in key words related for each goal. The system took every word in discourse and classify it according to a specific sustainable development goal. Starting from oral discourse, organizing it and taking it to quantitative data, it verifies that words can be used to be able to analyze a discourse with practical uses. This type of methodology allows quantifying large amounts of oral information that are extracted from interviews to find out what people think about a specific topic, for this research, about the 17 sustainable development goals.

References

Baars, B. (1986). Interview with George Miller. In B. Baars (Ed.), *The cognitive revolution in psychology* (pp. 200-203). Guildford Press.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: finding a balance between artifact and theory. *Journal of the Association for Information Systems*, *19*(5), 358-376. https://dx.doi.org/10.17705/1jais.00495

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004* (pp. 641-644). European Language Resources Association (ELRA). https://aclanthology.org/volumes/L04-1/

Collins, A. & Quilian, R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior 8*(2), 240-247.

Devi, S., Priya, M.V., Akhila, P., & Vasundhara, N. (2018). Analysis and prediction of student placement for improving the education standards. *International Journal of Engineering & Technology*, *7*(2.8), 303-306. https://doi.org/10.14419/ijet.v7i2.8.10429

Eekels, J., & Roozenburg, N. F. (1991). A methodological comparison of the structures of scientific research and engineering design: their similarities and differences. *Design studies*, *12*(4), 197-203.

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

Järvinen, P. (2007). Action research is similar to design science. *Quality & Quantity*, *41*(1), 37-54.

Kutela, B., and Teng, H. (2019). Prediction of drivers and pedestrians' behaviours at signalized mid-block Danish offset crosswalks using Bayesian networks. *Journal of Safety Research 69*, 75-83. https://doi.org/10.1016/j.jsr.2019.02.008

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., & Tengi, R. (1993). Five papers on WordNet (TM). *International Journal of Lexicography*, *3*(4), 235-244.

Naciones Unidas (2019). *Informe de los objetivos de desarrollo sostenible*. Naciones Unidas.

O'Riain, S., Curry, E., & Buitelaar, P. (2012). Engaging Practitioners within Design Science Research: A Natural Language Processing Case Study. In M. Helfert, & B. Donnellan (Eds.), *Design Science: Perspectives from Europe. EDSS 2012. Communications in Computer and Information Science, vol 388* (pp. 155-169). Springer, Cham. https://doi.org/10.1007/978-3-319-04090-5_14

Pereira, A. R., Ferreira, J. J. P., & Lopes, A. (2020). A knowledge representation of the beginning of the innovation process: The Front End of Innovation Integrative Ontology (FEI2O). *Data & Knowledge Engineering*, *125*, 101760. https://doi.org/10.1016/j.datak.2019.101760

Sandri, M.; Berchialla, P.; Baldi, I.; Gregori, D.; & De Blasi, R., A. (2014). Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. *Journal of biomedical informatics*, *48*, 106-13.

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

Chapter XIV

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[1]

Correlation between the orientational metaphor GOOD IS UP / BAD IS DOWN and positive/negative polarity in Spanish verbs: a study with corpus statistics

Benjamín López Hidalgo, Irene Renau & Rogelio Nazar
Pontificia Universidad Católica de Valparaíso –Chile

Resumen: La metáfora conceptual se ha estudiado ampliamente mediante lingüística de corpus, pero es necesario seguir proponiendo métodos estadísticos que permitan hallar evidencia cuantitativamente significativa sobre su uso en el discurso. Además, la metáfora orientacional en particular ha sido poco abordada en la investigación sobre metáfora conceptual. Esta investigación tiene como objetivo comprobar la relación entre la orientación vertical (ARRIBA/ABAJO) y la polaridad (POSITIVA/NEGATIVA, respectivamente) que existe en las metáforas orientacionales del tipo BUENO ES ARRIBA / MALO ES ABAJO halladas en corpus. 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' (resp.), etiquetadas mediante

1 Agradecemos al Proyecto Fondecyt Regular n.º 1231594 (ANID, gobierno de Chile).

un lexicón de polaridad. Los resultados indican que existe tal asociación en el 80% de los casos analizados.

Abstract: Conceptual metaphors have been extensively studied by means of corpus linguistics, but there is a need to continue proposing statistical methods that allow us to find quantitatively meaningful evidence on its use in discourse. Moreover, orientational metaphors in particular are yet to be sufficiently addressed in conceptual metaphor research. The present research aims to test the relationship between vertical orientation (UP/DOWN) and polarity (POSITIVE/NEGATIVE, respectively) that exists in orientational metaphors of the type GOOD IS UP / BAD IS DOWN found in corpora. Ten Spanish verbs with meaning 'up' / 'down' were selected and their association was measured in corpus concordances with lexical units with 'positive' / 'negative' value (resp.), labeled by means of a polarity lexicon. The results indicate that such an association exists in 80% of the analyzed cases.

1. Introducción

La metáfora orientacional (Lakoff & Johnson, 1980, 1999; Lakoff, 1993; Langacker, 1986; Kövecses, 2002, 2008; Soriano, 2012) es un tipo de metáfora que organiza un sistema conceptual en términos de una orientación espacial. Tal es el caso de los conceptos FELIZ, BUENO, OPTIMISTA... y TRISTE, MALO, PESIMISTA..., que son considerados de forma universal como positivos y negativos, respectivamente. En estos casos, la metáfora orientacional FELIZ, BUENO, OPTIMISTA... ES ARRIBA / TRISTE, MALO, PESIMISTA... ES ABAJO funciona como un dispositivo conceptual que permite organizar, expresar, comprender y reforzar cognitivamente estos conceptos abstractos. Por ejemplo, en expresiones como "Mi moral está *por los suelos*" se hace explícita la relación entre 'estar pesimista' y la posición 'abajo' a través de la locución verbal *por los suelos;* al contrario, en "Mi moral está *por las nubes*" se muestra una relación entre 'arriba' y 'optimista'. Esta relación entre la orientación espacial ARRIBA/ABAJO y la consideración de algo como POSITIVO/NEGATIVO se ha evidenciado empíricamente sobre todo a partir de la psicología experimental y también de algunos estudios de corpus (véase el apartado 2). Sin embargo, la evidencia es escasa y, en particular, faltan propuestas que permitan observar este fenómeno cognitivo a través de expresiones en el discurso, de forma cuantitativamente significativa y con métodos que permitan replicar los estudios en distintos tipos de textos y lenguas.

En vista de lo anterior, esta investigación se propuso comprobar si la metáfora orientacional BUENO ES ARRIBA / MALO ES ABAJO puede evidenciarse empíricamente a través del análisis estadístico de corpus. Para ello, se seleccionaron verbos del español que, en su acepción prototípica, tuvieran el significado de 'subir' o 'bajar', y se analizó su coocurrencia con unidades léxicas con sentido de 'bueno' o 'malo', respectivamente. Se etiquetaron estas unidades como BUENO o MALO mediante un lexicón de polaridad, que tiene ya previamente etiquetadas las unidades léxicas como 'positivas' o 'negativas'.

La hipótesis que se planteó es que los verbos con significado 'subir' (como *ascender, elevar, levantar,* etc.) coocurren más a menudo con unidades léxicas (sustantivos, adjetivos, verbos y adverbios, locuciones incluidas) con significado 'bueno' (ej., *maravilloso*, *alegrar*, *felizmente*, *bondad*, *en las nubes*), y los verbos con significado 'bajar' (como *caer*, *descender*, *tumbar*, etc.) coocurren más a menudo con unidades con significado 'malo' (ej., *horrible*, *entristecer*, *desgraciadamente*, *maldad*, *a duras penas*, etc.). El trabajo, como ya se indicó, tiene interés al proponer un método puramente estadístico y, por tanto, fácil de aplicar a otras lenguas y a distintos corpus, y la única herramienta externa utilizada (el lexicón de polaridad) es muy común en muchos idiomas debido a su uso extendido en el área de la minería de opinión (Alm *et al.*, 2005; Baccianella *et al.*, 2010). Desde un punto de vista más amplio, este trabajo es un aporte a los estudios de metáfora en corpus, y en particular, a la evidencia empírica sobre la teoría de la metáfora conceptual en el discurso.

2. Antecedentes y marco teórico

La teoría de la metáfora conceptual (Lakoff & Johnson, 1980, 1999; Lakoff, 1993; Langacker, 1986; Kövecses, 2002, 2008; Soriano, 2012) postula que la metáfora es un mecanismo cognitivo utilizado por el ser humano para comprender el mundo o expresar su concepción de la realidad; una metáfora conceptual toma como dominio de origen una realidad conocida y generalmente concreta y material, y la utiliza para categorizar el dominio de destino, correspondiente a una realidad más desconocida y abstracta. Las metáforas conceptuales pueden expresarse mediante dibujo, fotografía, danza, música, etc., pero es muy común su uso en expresiones lingüísticas, no solo en literatura, sino en cualquier discurso de la vida cotidiana. Así, a través de expresiones como *dejamos la vida en la cancha, el equipo atacó con fuerza, salimos derrotados en la final del campeonato,* etc., el FÚTBOL (dominio de destino) es caracterizado como una GUERRA (dominio de origen) a través de la metáfora conceptual EL FÚTBOL ES UNA GUERRA.

El tipo de metáfora mencionado se denomina *estructural* porque organiza el conocimiento del dominio meta mediante la estructura conceptual importada del dominio fuen-

te. Las metáforas ontológicas, por su lado, sirven para caracterizar elementos abstractos (como eventos, emociones, experiencias, ideas, etc.) mediante entidades materiales. Por ejemplo, LA MENTE ES UNA MÁQUINA es una metáfora ontológica que permite comprender la mente como un artefacto complejo; esta metáfora se observa en múltiples expresiones lingüísticas, como *mi cerebro está un poco oxidado hoy, tengo el disco duro demasiado lleno de distracciones,* etc.

La metáfora orientacional, que centra nuestra investigación, fue definida por Lakoff & Johnson (1980, 14) como "another kind of metaphorical concept, one that does not structure one concept in terms of another but instead organizes a whole system of concepts with respect to one another". En otras palabras, da coherencia a un conjunto de conceptos, debido a que estos comparten el mismo dominio de origen (Langacker, 1986). Lakoff y Johnson (1980, 14) las llamaron metáforas orientacionales "since most of them have to do with spatial orientation: up-down, in-out, front-back, on-off, deep-shallow, central-peripheral". Por ejemplo, los conceptos FELIZ / BUENO / SALUD / PODER se unifican bajo el concepto ARRIBA, mientras que TRISTE / MALO / ENFERMEDAD / AUSENCIA DE PODER se unifican en ABAJO: estas dos estructuras conceptuales, a su vez, se unifican bajo una de las metáforas orientacionales más universales: BUENO ES ARRIBA / MALO ES ABAJO.

Existe una línea ya extensa de trabajos que han abordado la teoría de la metáfora conceptual, sobre todo la metáfora estructural, desde el análisis de corpus (Charteris-Black, 2000; Semino *et al.*, 2004; Deignan, 2008; Semino *et al.*, 2016; Potts & Semino, 2019 ; Liu & Mo, 2020). Este enfoque ha permitido comprobar cómo las metáforas, empleadas en discursos de diversos tipos (prensa, textos especializados, escritura académica, etc.) contribuyen a configurar y transmitir determinados marcos cognitivos y culturales. En el caso de la metáfora orientacional en concreto, las evidencias parten más bien de los estudios experimentales, con algunos pocos estudios de corpus. El enlace entre orientación espacial ARRIBA / ABAJO y la connotación POSITIVA / NEGATIVA, respectivamente, se ha evidenciado en el área de la psicología experimental (Meier & Robinson, 2004, 2006; Crawford *et al.*, 2006; Cassanto & Dijkstra, 2010; Santana & De Vega, 2011). En estos trabajos se confirma empíricamente que el recuerdo de experiencias positivas facilita el realizar actividades motrices ascendentes, pero entorpece la actividad motriz cuando es descendente (Casasanto & Dijkstra, 2010). Asimismo, se compueba que colocar tarjetas de vocabulario en ubicaciones particulares después de estudiarlas ayudan a los estudiantes a aprender las definiciones de palabras con valencia emocional positiva (colocación arriba) o negativa (colocación abajo) (Casasanto & De Bruin, 2019). Según estos estudios, pues, existe una correlación positiva entre el concepto ARRIBA y BUENO, y ABAJO y MALO.

Las metáforas orientacionales de diversos tipos se han estudiado también en el discurso económico, político y del marketing. Por ejemplo, Fernández Rodríguez (2020) compara corpus de textos de economía en español y en francés y estudia las expresiones metafóricas orientacionales. En sus datos, el 75% de estas metáforas corresponden a la orientación ARRIBA / ABAJO (ej., "la inflación china *baja*", "la *caída* de los precios de los alimentos", *cf.* Fernández Rodríguez, 2020, p.121), y en otros casos a la orientación ENTRAR / SALIR, CENTRO / PERIFERIA, etc. Estas metáforas, como indican Graupe y Steffestun (2018), sirven para facilitar la comprensión de conceptos abstractos de la economía mediante conceptos más intuitivos y cercanos, como ocurre con las metáforas conceptuales en general (Lakoff & Johnson, 1980). En determinados textos, no obstante, pueden dificultar también el pensamiento crítico en tanto que proponen marcos conceptuales que no se discuten: por ejemplo, el mercado visto como un CONTENEDOR que se conceptualiza con la oposición DENTRO / FUERA(*cf.* Graupe & Steffenstun, 2020). Luque (2020) también encuentra la metáfora orientacional de tipo BUENO ES ARRIBA / MALO ES ABAJO en un corpus de discursos políticos euroescépticos (por ejemplo, "esta Unión Europea ha *caído* en una serie de errores de los que será difícil recuperarse", Luque, 2020, p.358). Feng Dezheng (2011), desde una perspectiva multimodal, analiza el sistema de orientaciones espaciales en el marketing, en específico en anuncios publicitarios de automóviles, donde identifica metáforas orientacionales como IDEAL / ABSTRACT IS UP – REAL / CONCRETE IS DOWN, entre otras del mismo tipo. Finalmente, el uso de metáforas orientacionales se ha analizado también en la literatura. Así pues, Zhao, Han y Zhao (2019) realizaron un análisis de corpus de las metáforas conceptuales en *Pavilion of Women,* de Pearl S. Buck, y en su estudio hallan que las metáforas orientacionales son las menos frecuentes, aunque de ellas, la más frecuente es UP IS GOOD / DOWN IS BAD (por ejemplo, "She let her heart *down*", *cf.* Zhao, Han & Zhao, 2019, p.107).

Las mencionadas aportaciones contribuyen al desarrollo de la propuesta seminal de Lakoff y Johnson (1980), aunque, como se ha podido comprobar, las investigaciones son escasas. Además de ello, las propuestas de corpus que han estudiado este tipo de metáfora han empleado en ocasiones software de gestión de corpus, como AntConc o Wordsmith, pero el análisis en sí ha sido manual y restringido a corpus de pequeñas dimensiones. Ello, como se indicó en la introducción, motiva la presente propuesta, que plantea un método de explotación de grandes cantidades de datos, lo que supone un nuevo avance hacia el estudio de este tipo de metáfora conceptual en el discurso.

3. Marco metodólogico

3.1. Materiales

Para llevar a cabo esta investigación, se utilizó un listado de verbos con significado 'arriba' y 'abajo', un corpus de trabajo y un lexicón de polaridad que permitiese etiquetar como 'positivas' o 'negativas' las unidades léxicas (sustantivos, adjetivos, verbos y adverbios, incluidas expresiones pluriverbales) que coocurrieran con los verbos. Naturalmente, en algunos casos los adverbios de negación pueden modificar la polaridad positiva o negativa de las palabras, pero ello representa una variable aleatoria y, como tal, no puede afectar los resultados.

En cuanto al listado de verbos empleado, se seleccionaron unidades que prototípicamente tuvieran significado 'arriba' y 'abajo'. Para ello, se buscaron verbos definidos, en su primera acepción, mediante los hiperónimos *subir* o *bajar* en dos diccionarios electrónicos (Battaner, 2003; RAE, 2014). Para el primer diccionario, se utilizó la búsqueda compleja del CD-ROM, y para el segundo se empleó la búsqueda avanzada de la plataforma EnclaveRAE. Del listado que se obtuvo, se seleccionaron los 5 de cada uno más frecuentes, menos ambiguos y comunes a las distintas variedades del castellano: *ascender, elevar, escalar, levantar* y *trepar* como hipónimos de *subir*, y *agachar, caer, derribar, descender* y *tumbar* en el caso de *bajar*.

Como corpus de trabajo, se utilizó el EsTenTen (Kilgarriff & Renau 2013), en concreto, la versión Spanish Web 2011 (esTenTen11, Eu + Am), que consta de, aproximadamente, 10.000 millones de palabras, divididas entre el español peninsular y el español de Latinoamérica.

Finalmente, se utilizó el lexicón de polaridad de Martínez (2018) para etiquetar los adjetivos, verbos, sustantivos y locuciones con carga positiva o negativa que coocurrieron con los verbos seleccionados. Un lexicón de polaridad es un conjunto de unidades léxicas que presentan una carga subjetiva que dirige hacia lo negativo o lo positivo, como *aburrirse* (–), *admirable* (+), etc. (Fauconnier, 1975; Giannakidou, 2001). Los lexicones de polaridad se utilizan en minería de opinión para, por ejemplo, el análisis de la expresión del texto a la voz (Alm *et al.*, 2005), la búsqueda de contenido emocional en foros o noticias (Lloyd *et al.*, 2005; Balog *et al.*, 2006) o el análisis de debates políticos y las respuestas a las preguntas (Yu & Hatzivassiloglou, 2003). Actualmente, el análisis de sentimiento ha tenido un gran desarrollo (Bosco *et al.*, 2013; Cambria *et al.*, 2014; Mäntylä *et al.*, 2018; Nassif *et al.*, 2020) y sus herramientas, recursos y métodos se han ido ampliando más allá de la minería de opinión; la presente investigación es un ejemplo de ello.

El lexicón de polaridad utilizado en esta investigación cuenta con aproximadamente 5.000 unidades léxicas, cada una en una línea del fichero seguidas de [N] en caso de ser negativa o de [P] en caso de ser positiva (véase un fragmento en la tabla 1 a modo de ejemplo), mientras que las unidades neutras (del tipo *mesa, estar, ahí,* etc.) se encuentran ausentes del lexicón. Algunos de los 10 verbos seleccionados estaban recogidos en el lexicón de polaridad empleado, por lo que, naturalmente, fueron deshabilitados del listado para que no alteraran el análisis del algoritmo.

Tabla 1. Fragmento del lexicón de polaridad utilizado. P = positivo; N = negativo.

Afable P
Afectado N
Afectar N
Afecto P
Afectuoso P

3.2. Métodos

En primer lugar, se preparó la muestra y se creó la herramienta de medición, que consistió en un script desarrollado en el lenguaje de programación Perl. Este script registra la frecuencia de coocurrencia en el corpus entre los verbos y las unidades del vocabulario de polaridad. En segundo lugar, se establecieron los criterios de análisis que nos permitieron controlar mejor las variables. En tercer lugar, se aplicaron pruebas preliminares en otros grupos de verbos que sirvieron para probar la validez del método, con el objetivo de, en la última etapa, aplicarlo una vez validado por dichas pruebas.

Para preparar la muestra se extrajo, con la herramienta virtual Jaguar (Nazar *et al.*, 2008; http://www.tecling.com/jaguar), una muestra aleatoria de 5.000 concordancias por cada uno de los 10 verbos (*ascender, elevar, escalar, levantar, trepar, agachar, caer, derribar, descender* y *tumbar*), cada una con una ventana de contexto de máximo 10 palabras a la izquierda y 10 palabras a la derecha (el total de la muestra, pues, fue de 50.000 concordancias). El corpus EsTenTen tiene etiquetado morfosintáctico con TreeTagger (Schmid, 1994), que durante décadas se consideró el sistema más avanzado para ello, tanto en castellano como en otras lenguas, lo que permitió obtener las concordancias con las unidades léxicas lematizadas. Esto facilitó el cruce con las unidades del lexicón de polaridad, que se encuentran también lematizadas. Como último paso de preparación de este material, cada muestra de 5.000 concordancias de cada verbo se trasladó a un archivo distinto.

Con el objetivo de medir la polaridad de las unidades léxicas que coocurren con alguno de los 10 verbos en cuestión, el script en lenguaje Perl que desarrollamos permite buscar, evaluar, agrupar y contar las unidades léxicas del lexicón en nuestra muestra. Este código, en concreto, se separa en tres acciones que se describen a continuación:

1 *Lectura e instrumentalización del lexicón de polaridad.* Se asignó un valor a cada unidad léxica del lexicón de polaridad para luego reconocer y contabilizar dichas unidades en las concordancias. El objetivo fue hacer que tanto las unidades léxicas negativas como las positivas del lexicón sumaran 1 por cada vez que aparecieran en una concordancia (a menos que la unidad léxica tuviese 3 o menos letras: esto se hizo para evitar ruido de adverbios de negación, entre otros problemas).
2 *Clasificación de concordancias.* Luego, se realizó un conteo de las unidades léxicas positivas y de las negativas que se encontraron en cada concordancia. Como output, se obtuvo la polaridad de cada concordancia. Si la concordancia presentaba más casos de unidades léxicas positivas que negativas, la concordancia se clasificó como positiva, y viceversa. Si se contaba el mismo número de unidades léxicas positivas que negativas, la concordancia se clasificó como neutra. Por último, si no había unidades léxicas del lexicón de polaridad en la concordancia, esta también se clasificó como neutra.
3 *Clasificación de verbos.* Finalmente, se sumó el resultado de la clasificación anterior a nivel de concordancias por cada verbo, con el fin de determinar la tendencia del verbo hacia 'positivo' o 'negativo'. La mayor cantidad de concordancias etiquetadas como positivas por cada verbo daba como resultado que el verbo se clasificaba como 'positivo', y viceversa.

4. Análisis de datos

4.1. Criterios de análisis

Una vez conformados los materiales y establecidos los métodos se tomó la decisión de fijar un umbral de comportamiento neutro de los verbos. En concreto, se postuló que si un verbo poseía un 51% o más del total de concordancias que no resultaran ni positivas ni negativas, ese verbo se consideraría neutro, ya sea por una igualdad entre los resultados locales (+) y (−) en el verbo en cuestión o porque fueron más las concordancias en las que el algoritmo no encontró unidades léxicas del lexicón de polaridad, debido a la extensión de este último. Con esto se controló que la cantidad de concordancias con polaridad fuera significativa respecto con el total de concordancias por cada verbo. Para determinar la

significación estadística de los resultados se empleó el nivel alfa de 0.05, tal como es habitual en ciencias sociales.

4.2. Pruebas preliminares

Antes de analizar el grupo de verbos que eran objeto de estudio, se realizaron pruebas con dos grupos de verbos para evaluar la efectividad del método. La prueba 1 se realizó para medir la confiabilidad del instrumento, y consistió en aplicar el algoritmo a 5 verbos con sentido positivo y 5 verbos con sentido negativo, en ambos casos no vinculados a las metáforas orientacionales que son objeto de estudio y con sentidos positivo o negativo muy evidentes: *agradecer, bendecir, felicitar, festejar, sonreír, destruir, empeorar, entristecer, lamentar, llorar*. La prueba 2 consistió en observar el resultado del algoritmo con 10 verbos a los que no se podría asociar a priori un sentido positivo ni negativo, es decir, verbos considerados neutros: *pensar, decir, estar, dibujar, escribir, tomar, traducir, consistir, leer, vestir*. Ambas pruebas fueron realizadas con el mismo corpus empleado para los verbos en estudio. Los resultados de estas dos pruebas preliminares se muestran en la tabla 2.

Tabla 2. Resultados de las pruebas preliminares.

Prueba 1					
Verbos	**Total +**	**Total –**	**% concordancias con polaridad del verbo**	**Polaridad resultante +/–**	**Valor *p***
agradecer	3260	366	73	+	< 2.2e-16
bendecir	3143	449	72	+	< 2.2e-16
felicitar	3194	366	71	+	< 2.2e-16
festejar	2545	685	65	+	< 2.2e-16
sonreír	2504	1045	71	+	< 2.2e-16
destruir	1339	1998	67	–	< 2.2e-16
empeorar	1242	2131	67	–	< 2.2e-16
entristecer	963	1278	66	–	= 2.85e-11
lamentar	1357	1916	65	–	< 2.2e-16
llorar	1569	1909	70	–	= 8.156e-09
Prueba 2					

Verbos	Total +	Total –	% concordancias con polaridad del verbo	Polaridad resultante +/–	Valor *p*
pensar	1873	1336	64	+	< 2.2e-16
decir	1904	1215	62	+	< 2.2e-16
estar	1987	1233	64	+	< 2.2e-16
dibujar	2096	969	61	+	< 2.2e-16
escribir	1958	1013	59	+	< 2.2e-16
tomar	1865	1211	62	+	< 2.2e-16
traducir	2177	1064	65	+	< 2.2e-16
consistir	2126	921	61	+	< 2.2e-16
leer	2006	982	60	+	< 2.2e-16
vestir	1985	1034	60	+	< 2.2e-16

La tabla 2 indica que, con respecto a la prueba 1, ninguno de los 10 verbos superó el umbral de comportamiento neutro que se estableció (51% o más), lo que implica que la cantidad de concordancias evaluadas como positivas o como negativas es significativa en consideración al total de concordancias por cada verbo. En segundo lugar, se observa que el algoritmo reconoció en el grupo de verbos de la prueba 1 los 5 verbos de polaridad positiva como positivos y los 5 verbos de polaridad negativa como negativos, tal como se esperaba. Por otra parte, se puede observar que ningún verbo presenta un valor *p* mayor a 0.05, por tanto, ninguno de estos resultados puede atribuirse al azar, lo que demuestra que hay una dependencia estadística entre estos 10 verbos y la polaridad que obtuvieron como resultado.

En el caso del grupo de verbos de la prueba 2, los 10 verbos presentaron polaridad positiva (+), lo que constituye un hallazgo imprevisto. Igual que en la prueba 1, en este caso el valor *p* también fue siempre menor a 0.05, lo que significa que la probabilidad de que estos resultados hayan sido producto del azar es remota (0.001). Este resultado indica probablemente que ciertos verbos, aunque no tengan una polaridad aparente, generalmente presentan una tendencia hacia la polaridad positiva (+); por ejemplo, se identifican actividades como *pensar, leer* o *escribir* como positivas en la mayoría de los casos. La profundización en el estudio de este hallazgo, que no se encuentra entre los objetivos de la investigación, se deja para trabajo futuro.

5. Resultados y discusión

Como ya se mencionó en el apartado 3.1, se analizaron 5 verbos con orientación arriba (*ascender, elevar, escalar, levantar* y *trepar*) y 5 verbos con orientación abajo (*agachar, caer,*

derribar, descender y *tumbar*), que sirvieron para reflejar el binomio orientacional ARRIBA / ABAJO. Para analizar su relación con aquellas unidades léxicas que reflejan los conceptos BUENO / MALO se aplicó el método descrito en el apartado 3, una vez ya realizadas las evaluaciones que permitieron asegurar la confiabilidad (prueba 1) y flexiblidad (prueba 2) del instrumento de medición. Los resultados del estudio se presentan en la tabla 3.

Tabla 3. Resultados del análisis del grupo de verbos en estudio.

Verbos	Total +	Total –	% concordancias con polaridad del verbo	Dif. total + y total –	Polaridad resultante +/–	Chi cuadrado	Valor *p*
ascender	1688	1078	55	610	+	1.345.264	< 2.2e-16
elevar	2055	1131	64	924	+	2.679.774	< 2.2e-16
escalar	1803	1474	66	329	+	330.305	= 9.072e-09
levantar	1665	1472	63	193	+	118.741	= 0.0005692
trepar	1811	1044	57	767	+	2.060.557	< 2.2e-16
agachar	1508	1670	64	162	–	8.258	= 0.004057
caer	1304	1890	64	586	–	1.075.128	< 2.2e-16
derribar	1425	1846	65	421	–	541.856	= 1.824e-13
descender	1410	1459	57	49	–	0.8369	= 0.3603
tumbar	1535	1620	63	85	–	2.29	= 0.1302

En la tabla 3 se muestra, en primer lugar, que ninguno de los 10 verbos superó el umbral de comportamiento neutro que se estableció (51% o más). El mayor porcentaje analizado se presenta en el verbo *escalar* con 66% y el menor porcentaje analizado se presenta en el verbo *ascender* con 55%), por lo que, como se explicó anteriormente, la polaridad fue estadísticamente significativa en consideración al total de concordancias por cada verbo. En segundo lugar, los resultados arrojaron que los verbos de orientación ARRIBA se vinculan con el sentido positivo, mientras que los verbos de orientación ABAJO se vinculan con el sentido negativo. La probabilidad de que este resultado fuera por azar es de 0.001 y, por tanto, prácticamente nula.

Para comprobar en cuántos casos existe o no dependencia estadística entre las dos variables, se aplicó el test del chi cuadrado, que arrojó que *descender* (0.3603) y *tumbar*

(0.1302) presentan un valor *p* mayor a 0.05 y, por tanto, los resultados no son estadísticamente significativos. Los otros 8 verbos presentan, sin embargo, un valor menor al alfa 0.05, lo que muestra que hay una dependencia estadística entre estos verbos y el sentido positivo o negativo que se obtuvo como resultado de la aplicación del método. Es decir, en cuanto a la formulación de nuestra hipótesis, en el 80% de los casos esta se confirmó.

Estos resultados, en consideración con lo estipulado en los criterios de análisis, permiten comprobar que existe una relación entre la variable orientación vertical y la variable polaridad positiva o negativa en contextos reales de uso de las unidades de análisis. Ello permite comprobar empíricamente y mediante métodos de estadística de corpus la metáfora orientacional BUENO/FELIZ ES ARRIBA y MALO/TRISTE ES ABAJO 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'. Así, por ejemplo, véase la concordancia 45 de *agachar:*

> ...lo he visto– dijo finalmente al tiempo que ***agachaba*** la mirada con *tristeza*...

En este contexto se observa una polaridad negativa que es reconocida por el script al detectar una unidad negativa presente en el lexicón de polaridad utilizado (*tristeza*) y ninguna positiva; el resto de unidades (*ver, finalmente, tiempo* y *mirada*) son neutras. Un caso opuesto se muestra en la concordancia 32 de *elevar:*

> ...cambios estructurales han *permitido avanzar* significativamente hacia la *estabilidad*, ***elevar*** la *eficiencia* de la economía...

En este contexto, el script reconoció cuatro unidades positivas (*permitir*, *avanzar*, *estabilidad* y *eficiencia*) y ninguna negativa (pues el resto son neutras: *cambio, estructural, significativamente, economía*). (Se recuerda que tanto *agachar* como *elevar,* igual que el resto de verbos en estudio, se excluyeron del lexicón para no interferir en los resultados y, por tanto, no fueron contabilizados como positivos ni negativos).

Finalmente, el siguiente ejemplo (concordancia 14 de *agachar*) muestra que las categorías POSITIVO y NEGATIVO pueden ser controvertidas, lo que mueve a considerar que sería difícil obtener un 100% de precisión con este método, como es habitual en semántica:

> ...ahora nos *falta* [–] a nosotros **agachar** la cabeza de una vez y *reconocer* [+] nuestros *errores* [–]...

En este caso, el algoritmo identifica las unidades *faltar* y *error* como unidades negativas y *reconocer* como positiva y, por tanto, adjudica un resultado de polaridad negativa a esta concordancia. Si bien la expresión *agachar la cabeza* es claramente negativa, podría considerarse que *reconocer nuestros errores*, y especialmente el conjunto del contexto, es una secuencia positiva. Esto ocurre también con adjetivos como *gran(de)* (+) o *poco* (–), que pueden generar secuencias de polaridad contraria a la del adjetivo aislado: *gran pena* (–), *pocas* críticas (+). Estos casos, si bien producen cierto porcentaje de error, se compensan con la gran cantidad de datos analizados (5.000 concordancias por cada verbo), lo que reduce el impacto de este tipo de secuencias en la muestra.

Además, cabe destacar que los resultados arrojaron una mayor circulación de unidades léxicas positivas a nivel general de los verbos analizados, con independencia de la polaridad con la que fueron evaluados. Este fue un resultado sorprendente, sobre todo por la diferencia reflejada en el total de concordancias analizadas como positivas y en el total de concordancias analizadas como negativas (60% + frente a 40% total) de los verbos analizados (prueba 1, prueba 2 y grupo en estudio). Además, la prueba 2 dio como resultado la polaridad positiva en 10 de 10 verbos sin una polaridad aparente, lo que es otra prueba de esta tendencia. Asimismo, la mayor diferencia entre total + y total – se dio en los verbos evaluados como positivos, lo que habla de que, por lo general, tienen una polaridad más marcada que los negativos (véase la figura 1 para ampliar el panorama de los datos).

En último lugar, el total de concordancias con polaridad en el total de verbos fue de 95.102, es decir, un promedio de 64,1% del total analizado (148.389 concordancias) (véase la figura 1). Este resultado, si bien es estadísticamente suficiente, puede mejorar conforme se emplee un lexicón de polaridad más amplio o se amplíe el utilizado, y el instrumento de medición se vaya complejizando.

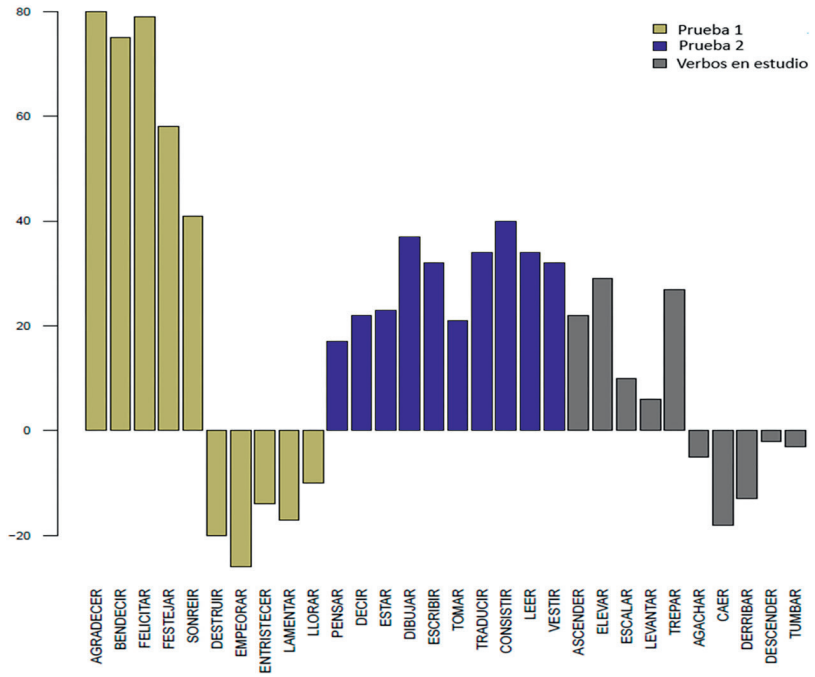


Figura 1. Porcentaje de diferencia entre total de polaridad + o – en cada verbo analizado (prueba 1, 2 y grupo de verbos en estudio).

6. Conclusión y perspectivas

Esta investigación se situó en la problemática de la metáfora conceptual y su estudio desde el análisis de corpus. En concreto, su enfoque radica en el análisis estadístico de un tipo de metáfora orientacional y su materialización lingüística en contextos reales de uso. Para observar el binomio ARRIBA/ABAJO se buscaron verbos que presentaran en su definición el verbo *subir* o *bajar,* mientras que para observar los dominios BUENO y MALO se empleó el recurso del lexicón de polaridad con el fin de observar el comportamiento discursivo de estos dominios conceptuales que física, cultural y socialmente son entendidos a nivel general como positivos y negativos, respectivamente.

A partir de los resultados mostrados en el apartado anterior, se puede confirmar que la relación entre verbo con orientación ya sea ARRIBA o ABAJO y la polaridad 'positiva' y 'negativa', respectivamente, se manifiesta a nivel lingüístico y es coherente con los postulados de la metáfora orientacional (Lakoff & Johnson 1980, 1999b; Lakoff, 1993). Es decir, un verbo con significado 'arriba' tiende a aparecer combinado con unidades léxicas son sentido positivo, y un verbo con significado 'abajo' tiende a aparecer combinado con unidades léxicas con sentido negativo.

Como trabajo futuro, el algoritmo confeccionado se puede aplicar empleando otros lexicones que permitan analizar el uso de otras expresiones metafóricas, como puede ser, por ejemplo, el caso de un lexicón de términos bélicos que aporte en el análisis de la metáfora estructural LA DISCUSIÓN ES UNA GUERRA en su dimensión lingüística. Para ello, se podrían, eventualmente, extraer expresiones de foros o situaciones comunicativas en las que personas debatan con respecto a un tema y hacer la búsqueda de las unidades del lexicón de términos bélicos en estas expresiones de situaciones comunicativas de debate o discusión. Este es uno de los tantos ejemplos en los que el algoritmo puede contribuir en los estudios de las metáforas conceptuales con métodos de estadística de corpus. Asimismo, el léxico trabajado en cuanto a verbos con polaridad ARRIBA/ABAJO se puede ampliar mediante otras técnicas, como por ejemplo utilizando algoritmos de aprendizaje automático. Alternativamente, también se podría intentar la expansión del lexicón de polaridad utilizando los mismos métodos de esta investigación. Por ejemplo, 8 de 10 los verbos estudiados tienen una dependencia estadística con la polaridad asignada, lo que implica que se pueden agregar al lexicón de polaridad *escalar* y *trepar* como unidades léxicas positivas y *agachar* como una unidad léxica negativa, entre otros verbos que actualmente no se encuentran en dicho recurso.

Referencias

Alm, C., Roth, D. & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. En R. Mooney, C. Brew, L.-F. Chien & K. Kirchhoff (Eds.), *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579-586). Association for Computational Linguistics.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, (pp. 2200-2204). European Language Resources Association.

Balog, K., Mishne, G. & De Rijke, M. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. En D. McCarthy & S. Wintner (Eds.), *11th Conference of the European Chapter of the Association for Computational Linguistics. Proceedings of the conference* (pp. 207-210). Association for Computational Linguistics.

Battaner, P. (2003). *Diccionario de uso del español de América y España*. Spes. Versión CD- ROM.

Bosco, D., Patti, V. & Bolioli, A. (2013). Developing corpora for sentiment analysis and opinion mining: a survey and the Senti-TUT case study. *IEEE Intelligent Systems*, *28*(2), 55-63.

Cambria, E., Gelbukh, A., Poria, S. & Kwok, K. (2014). Sentic API: a common-sense based API for concept-level sentiment analysis. En M. Rowe, M. Stankovic & A.-S. Dadzie (Eds.), *Proceedings of the the 4th Workshop on Making Sense of Microposts* (pp. 19-24).

Casasanto, D. & Dijkstra, K. (2010). Motor action and emotional memory. *Cognition*, *115*(1), 179-185.

Casasanto, D. & De Bruin, A. (2019). Metaphors we learn by: directed motor action improves word learning. *Cognition*, *182*, 177-183.

Charteris-Black, J. (2000). Metaphor and vocabulary teaching in ESP economics. *English for Specific Purposes*, *19*(2), 149-165.

Crawford, E., Margolies, S., Drake, J. & Murphy, M. (2006). Affect biases memory of location: evidence for the spatial representation of affect. *Cognition and Emotion*, *20*(8), 1153-1169.

Deignan, A. (2008). Corpus linguistics and metaphor. En R. Gibbs (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 280-294). Cambridge University Press.

Dezheng, F. (2011). Visual space and ideology. A critical cognitive analysis of spatial orientations in advertising. En K. O'Halloran & B. Smith (Eds.), *Multimodal studies. Exploring issues and domains* (pp. 55-75). Routledge.

Fauconnier, G. (1975). Polarity and the scale principle. *Chicago Linguistic Society*, *11*, 188-199.

Fernández Rodríguez, Á. (2020). La metáfora orientacional en traducción económica (fr-es-fr). *Çédille. Revista de Estudios Franceses*, *17*, 115-139.

Giannakidou, A. (2001). The meaning of free choice. *Linguistics and Philosophy*, *24*(6), 659-735.

Gibbs Jr, R. W., Gibbs, R. W., & Gibbs, J. (1994). *The poetics of mind: figurative thought, language, and understanding*. Cambridge University Press.

Graupe, S. & Steffestun, T. (2018). 'The market deals out profits and losses' – How standard economic textbooks promote uncritical thinking in metaphors. *Journal of Social Science Education*, *17*(3, 5-18.

Hatzivassiloglou, V. & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. En M. Kay (Ed.), *COLING '00: Proceedings of the 18th Conference on Computational Linguistics*, (pp. 299-305). Association for Computational Linguistics.

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.

Liu, D. & Mo, Q. (2020). Conceptual metaphors and image schemas: a corpus analysis of the development of the *on track/off track* idiom pair. *Journal of English Linguistics*, *48*(2), 137-165.

Lloyd, D. K. & Skiena, S. (2005). Lydia: a system for large-scale news analysis. En M. Consens & G. Navarro (Eds.), *String Processing and Information Retrieval. 12th International Conference, SPIRE* 2005 (pp. 161-166). Springer.

Luque, F. (2020). La metáfora conceptual en el discurso político euroescéptico (francés-español). *Logos: Revista de Lingüística, Filosofía y Literatura*, *30*(2), 349-364.

Mäntylä, M. V., Graziotin, D. & Kuutila, M. (2018). The evolution of sentiment analysis: a review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32.

Martínez, R. (2018). *La incidencia de las interacciones verbales en la configuración de la red social twitter: un análisis desde la polaridad, la novedad y el prestigio* [Tesis doctoral]. Pontificia Universidad Católica de Valparaíso.

Meier, B. & Robinson, M. (2004). Why the sunny side is up: associations between affect and vertical position. *Psychological Science*, *15*(4), 243-247.

Meier, B. & Robinson, M. (2006). Does "feeling down" mean seeing down? Depressive symptoms and vertical selective attention. *Journal of Research in Personality*, *40*(4), 451-461.

Nassif, A., Elnagar, A., Shahin, I. & Henno, S. (2020). Deep learning for Arabic subjective sentiment analysis: challenges and research opportunities. *Applied Soft Computing Journal*, *98*, 106836, 1-27.

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.

Potts, A. & Semino, E. (2019). Cancer as a metaphor. *Metaphor and Symbol*, *34*(2), 81-95.

Real Academia Española. (2014). *Diccionario de la lengua española* (23.ª ed.). Espasa.

Santana, E. & De Vega, M. (2011). Metaphors are embodied, and so are their literal counterparts. *Frontiers in Psychology*, *2*, 1-12.

Semino, E., Demjén, Z. & Demmen, J. (2016). An integrated approach to metaphor and framing in cognition, discourse, and practice, with an application to metaphors for cancer. *Applied Linguistics*, *39*(5), 625-645.

Semino, E., Heywood, J. & Short, M. (2004). Methodological problems in the analysis of metaphors in a corpus of conversations about cancer. *Journal of Pragmatics*, *36*(7), 1271-1294.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*.

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

Yu, H. & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. En *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 129-136). Association for Computational Linguistics.

Zhao, X., Han, Y., & Zhao, X. (2019). A corpus-based study of metaphor in *Pavilion of women*. *Chinese Semiotic Studies*, *15*(1), 95-117.

Chapter XV

UnderRL Tagger[1]: a free software for Under-Resourced Languages POS tagging

UnderRL Tagger: un software libre para etiquetar POS en Under-Resourced Languages

José Luis Pemberty Tamayo & Jorge Mauricio Molina Mejía
Universidad de Antioquia – Colombia

Abstract: This chapter presents a free software program that can be used for POS tagging in a multiplicity of languages that do not have automatic taggers. The program aims to facilitate the work with corpora in these languages through Natural Language Processing. Its operation allows the manual tagging process to be gradually automated thanks to a system that makes it possible to recall and reuse tags, as well as to handle large amounts of text and to generate output files in XML format with tags based on the EAGLES system.

Resumen: En este capítulo se presenta un software libre que puede utilizarse para el etiquetado de POS en una multiplicidad de lenguas que no cuentan con etiquetadores automáticos. El programa busca facilitar el trabajo con corpus en estas lenguas a través de la lingüística computacional. Su funcionamiento permite que el proceso manual de etiquetado se convierta poco a poco en automático gracias a un sistema que permite recordar y reutilizar las etiquetas, de la misma manera en que permite manejar grandes cantidades de textos y generar archivos de salida en formato XML con etiquetas basadas en el sistema EAGLES.

1 UnderRL Tagger is a free software for semi-automatic POS tagging of languages without many linguistic resources, which has been created within the framework of the college work of J. L. Pemberty Tamayo (2020), within the research team Corpus Ex Machina (Facultad de Comunicaciones y Filología, Universidad de Antioquia). The computer program has been patented in 2020 by J. L. Pemberty Tamayo, J. M. Molina Mejía and M. I. Marín Morales (2020).

1. Introduction

One of the most notorious aspects in the research and study of current Linguistics is the use of textual corpora for various purposes, for example: grammatical analysis (Parodi, 2010; Biber & Finegan, 2014; Jones & Waller, 2015), anaphora resolution (Mitkov, 2014; Poesio, Stuckardt & Versley, 2016; Grajales Ramírez & Molina Mejía, 2019), statistical analysis by means of corpora (Beaudouin, 2016; Brezina, 2018; Wallis, 2021), etc. On the other hand, it is possible to observe the way in which a strong relationship has been established with Computational Linguistics (Mitkov, 2004; Wilks, 2010; Molina Mejía, 2021), precisely for the processing, handling, and interpretation of required amounts of data (Zeroual & Lakhouaja, 2018). Within this scenario, written texts play a prominent role, since they lend themselves to computational processes more easily than other forms of language use (Baquero Velásquez, 2010; Parodi, 2010). Such ease has made it possible to standardize different levels of annotation or tagging, which are ways of enriching the information in the text, making the linguistic notions underlying their use patent (McEnery & Hardie, 2011). An example of this is the POS (Part-of-Speech) level, the simplest and most necessary as a first step in the annotation of texts with linguistic information (Parodi, 2010; Straka & Straková, 2017).

The aforementioned process acquires importance when considering the purposes pursued by Corpus Linguistics, because it permits computers to process information to which they would not otherwise have access. In this sense, software products have also been built that, based on different systems of rules or artificial intelligence, can automatically perform, with a high degree of success, common forms of tagging in different languages, generally the most widely spoken ones such as Spanish, English, French, German, among others (Molina Mejía, 2021).

Automation in the case of corpus tagging is of great importance, since the manual work that would be required to annotate a robust corpus of texts is quite expensive in time, effort and human resources, not to say that it can often seem impossible. This situation places languages that do not have the computerized means to be processed efficiently, at a disadvantage; since the need for manual work limits the information that can be taken for an investigation, as well as it can dissuade potential scholars from dedicating themselves to taking them as an object of work. This group is known as Under-Resourced Languages (henceforth URLa) (Krauwer, 2003).

Considering all of the above, this chapter presents "UnderRL Tagger" (Pemberty Tamayo, Molina Mejía & Marín Morales, 2020), a software that aims to help researchers in the process of tagging textual corpora in URLa, based on a system that permits to recall the tags associated with certain words and automating their annotation as much as possi-

ble (Pemberty Tamayo, 2020). It should be noted that the aim of the work is not to achieve fully automatic tagging, but to assist the manual process, as will be seen in the following pages. This program is the result of work done at the level of conception and elaboration of semi-automatic POS tagging systems for Under-Resourced Languages (Pemberty Tamayo, 2020; Pemberty Tamayo & Molina Mejía, 2020; Pemberty Tamayo *et al*., 2023).

2. State of the Art

As mentioned in the previous section, a clear antecedent of the works whose subject is corpus annotation are the computer platforms and computational tools that currently fulfill the task of automatically tagging large amounts of texts in different languages. Some well-known free access tools are TreeTagger[2] (Schmid, 1994) and TagAnt[3] (Anthony, 2015), which could help with the tagging of some different languages at the Part of Speech -POS- level (Weisser, 2018).

Other prominent names are FreeLing[4] (Padró, Collado, Reese, Lloberes & Castellón, 2010) and Stanford Parser[5] (Schuster & Manning, 2016), which allow annotation at different levels of analysis such as parsing (generation of syntactic trees from dependency grammar and immediate constituents, alternatively), recognition of coreferential chains (anaphora and cataphora), elaboration of semantic graphs, analysis of named entities, etc. Regarding FreeLing, it is important to note that this program uses the EAGLES system as a standard for the annotation of the different human languages.

The EAGLES are a series of conventions adopted by different groups in the work with corpora; they were proposed by the "Expert Advisory Group on Language Engineering Standards" (Leech & Wilson, 1996) and consist of a series of regulations in the use of certain codes for the different possible values in the tagging of POS notions. Bearing this in mind, the work presented here also embraces this standardization, its existence being an important antecedent in the definition of the algorithms described later in this chapter.

Within the framework of the creation of a computer system destined to under-resourced languages and minority languages, it is important to start from a standardized morphosyntactic tagging system. In this way, both researchers and specialists in this type

2 TreeTagger is a tool for annotating text with POS and lemma information. More information can be found at the following link: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

3 TagAnt is a freeware POS tagger built on TreeTagger tool. You can download the tool and find more information at the following link: https://www.laurenceanthony.net/software/tagant/

4 Information regarding FreeLing and the possibility of downloading the tool can be found at the following link: http://nlp.lsi.upc.edu/freeling/node/1

5 The Stanford Parser can be viewed and downloaded at the following site: https://nlp.stanford.edu/software/lex-parser.shtml

of language will be able to understand each other. Starting from this premise, it was decided to aim to have the tags proposed by the EAGLES project. This should permit the program to be used by specialists in minority and under-resourced languages in different geographical and linguistic contexts, and the data obtained from research in different languages to be shared globally. It is also worth mentioning different academic works that focus on the computational treatment of URLa; These works are based on approaches as varied as the annotating of specific languages, such as Arabic and Vietnamese (El-Haj, Kruschwitz & Fox, 2015; Le & Besacier, 2009); speech recognition (Besacier, Barnard, Karpov & Schultz, 2014) or corpus collection by obtaining texts from the web (Scannell, 2007). These works share with "UnderRL Tagger" their concern for this group of languages, but they also have the difference that they do not properly deal with automated assistance in manual corpus tagging and their approaches are, in most cases, monolingual.

Unlike these studies, two remarkable computer programs have also been found, since, although they do not mention the concept of URLa in their documentation, they mark more notable antecedents in relation to the objective of this work. These are "FieldWorks Language Explorer" (Moe, 2008) and "Field Linguist's ToolBox" (Buseman & Buseman, 2013), both designed to manage corpora in different languages, mainly with the intention of processing them at the lexicographic level and in order to finally produce a dictionary of the languages worked by each of them (Rogers, 2010).

However, these software programs, given the breadth of their field of application, could hinder the simplest task of obtaining an annotated corpus in each language, in addition to the fact that they also lack a standardization in the field of Corpus Linguistics such as those mentioned in EAGLES. In this sense, they are established as antecedents of this work, but their functionalities are not the same as those of "UnderRL Tagger" (Pemberty Tamayo, 2020).

3. Theoretical Framework

3.1. Computational Linguistics and Natural Language Processing

Computational Linguistics is usually defined as a discipline whose purpose is the construction of computer systems that process linguistic structures and simulate human linguistic capabilities (Moreno Sandoval, 1998, pp. 29-30). This discipline is framed within Applied Linguistics (Moreno Sandoval, 1998; Tordera Yllescas, 2011, Molina Mejía, 2021) and, following the opinion of several authors (Sáiz Noeda, 2002; Tordera Yllescas, 2011), it will be considered in this chapter as a synonym of NLP (Natural Language Processing).

Although many authors agree on this general definition, there are different ways of delimiting the scope of Computational Linguistics. From practical approaches that include all types of computer language processing (Mitkov, 2004, p.15), to more theoretical points of view, which focus on how the simulation of linguistic capacity helps to understand linguistic behaviour of natural languages (Tordera Yllescas, 2011). Considering, in addition, the use or creation of computational models or tools that allow the computational processing of natural languages, which should permit, a fortiori, that the language itself can serve as an input for scientific research and/or formulation of programs that can be applied in life, in society in general, thanks to the analysis of linguistic corpora in context (Molina Mejía, 2021).

In this difference of opinions, intermediate approaches have been found, such as that of Moreno Sandoval (1998), who proposes the following applications: a) systems that try to emulate the human capacity to process natural languages; b) programs to aid writing and textual composition; and c) computer-assisted teaching and linguistic task support systems (pp. 27-29). This last group includes tools for managing and annotating linguistic corpora, i.e., the work presented here. This list of applications can be extended with more current functionalities, following Nerbonne (2007) and Molina Mejía (2021): a) speech recognition; b) speech synthesis; c) data mining; d) automatic completion systems in smartphones; e) management of academic documents and databases; f) conversational systems; g) automatic topic detection; h) automatic summarization; i) automatic document classification, among others.

It is also common to find that Computational Linguistics is understood from its division into theoretical and applied. Theoretical Computational Linguistics deals with the construction of linguistic abstractions that encompass both computer and natural language phenomena, as well as the construction of algorithms that help model and test these abstractions (Nerbonne, 2007, p.3). Applied Computational Linguistics is dedicated to the construction of computer tools to manipulate language for different purposes (Nerbonne, 2007). The delimitation of these applications, as mentioned above, varies depending on the authors, however some may be mentioned: a) automatic translation; b) information retrieval; c) human-machine interfaces; d) text analysis tools; e) lexicographic databases; f) spelling, syntax, and style checkers; and g) educational programs for language teaching (Moreno Sandoval, 1998, pp. 27-29).

3.2. Corpus Linguistics

Corpus Linguistics is defined as a "methodology for languages and language research, which allows empirical **investigations** to be carried out in authentic contexts" (Parodi, 2010, p.15). Considering the empirical and authentic character indicated by this definition, this methodology can be related to the functionalist model of linguistics, which seeks to understand linguistic phenomena in real situations. This model is opposed to the generativist model, which is dedicated to theorizing about phenomena through linguistic intuition (Baquero Velásquez, 2010, p.25; McEnery & Hardie, 2013).

s tasks that fit within Corpus Linguistics, we can include the collection, processing and analysis of large amounts of data representative of the use of the language or languages that are assumed as object of study (Baquero Velásquez, 2010; Bernal Chávez & Hincapié Moreno, 2018; McEnery & Hardie, 2011). There is, moreover, a marked interdisciplinarity in this methodology, as it works both for the investigation of phenomena at any level of the language and to help in meeting the objectives of different fields of Applied Linguistics (Parodi, 2010, p.15).

Given that authenticity, representativeness and interdisciplinarity have been such important aspects in working with corpora; the relationship that can be established between Computational Linguistics and Corpus Linguistics becomes evident, since the former has provided the necessary mechanisms for handling large amounts of data information and its processing by various means (Baquero Velásquez, 2010; Bernal Chávez & Hincapié Moreno, 2018; Parodi, 2010) and, on the other hand, the need for corpora that possess a high level of quality and variety in discourses and textual typologies (Molina Mejía, 2021).

This relationship is even taken for granted nowadays, through authors who go so far as to define a corpus as a series of texts that can be processed by computers (McEnery & Hardie, 2011, p.1). However, this relationship has not always been present, and in previous times, such as the mid-twentieth century (Bernal Chávez & Hincapié Moreno, 2018, p.12) and even the nineteenth century (Baquero Velásquez, 2010), it has been necessary to carry out work with corpora manually. This implied enormous complications, since the more the amount of data with which one works grows, the greater sums of time, money, effort, and human capital are necessary, making some tasks unfeasible (Mitkov, 2001, p.110).

The help of computational means has therefore come to reduce the resources required in these jobs and also the risk of human errors and loss of information. However, not all languages have the appropriate tools to make use of these technologies, which places them at a considerable disadvantage, insofar as it is not possible to carry out work of the same

magnitude with them as with languages that are more accessible to computer processing (Baquero Velásquez, 2010, p.28).

3.2.1. What is a corpus?

The term corpus has already been used in the previous sections and, before continuing, it is necessary to dedicate a few paragraphs to clarify its definition. We will start from the proposal of Bernal Chávez and Hincapié Moreno (2018), for whom a corpus is a set of digital texts that are collected and systemized following linguistic criteria. Note in this definition the importance of computational means with respect to the need for texts to be digital; in addition to this, it is also fundamental the fact that the collection and systematic organization of the corpus is done with respect to these linguistic criteria; this is the main characteristic that distinguishes a corpus from any other collection of texts.

For its part, Parodi (2010) proposes a more specific list of characteristics that can guide us in understanding what a perfect corpus is:

1 Collection of texts in natural environments.
2 Explicitly of the defining features shared by the constituent texts.
3 Final plain digital type format (*.txt) for each text or document.
4 Size, preferably large.
5 Respect for ecological principles.
6 Semi-automatic computational tagging or annotation of a morphosyntactic or other nature for each text.
7 Availability through computational means.
8 Access to complete visualization of the texts that compose it in plain format.
9 Search for principles of proportionality or representativeness (possibly statistical).
10 Livelihood or initial provenance specified.
11 Identification of an organization around themes, types of texts, registers, genres, etc.
12 Record of quantitative data that allows the comparison and possible normalization of figures (p.26).
13 And to comply with all these elements at the same time, but that the importance of each one can vary depending on the specific objectives of each collection of texts (p.27).

In these characteristics, the need for computational processing is also evident, as well as the need to make explicit the features shared by the texts; this may or may not be part of a tagging or annotation, which is also part of the above list. With this in mind, an important part of corpus work is usually the enrichment of textual information with other types of

information that provides clarity about the underlying linguistic notions. This process is known as tagging, and it will be the object to be dealt with in the next section.

3.3. Corpus Annotation

The construction of a corpus is a process that goes through different phases, which include its design, data capture, storage system planning and text processing (Bernal Chávez & Hincapié Moreno, 2018, p.53). Within this last step is a process called annotation.

A clearer definition of corpus annotation can be found in the work of McEnery and Hardie (2011): "[…] is largely the process of providing —in a systematic and accessible form— those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they worked with" (p.13). It is very important to take into account, from this definition, the fact that the data included in the tagging are those that a linguist could extract from the collected texts, that is, the linguistic information that is implicit within the use of language and that it must be made visible in a systematic way so that it can be recognized and processed by computer programs.

To achieve this systematic way of describing the information, specialized languages are used in tagging, which help to assign different types of values to each of the elements of the text, depending on what is to be said about them. Some of these languages are XML (Extensible Markup Language), HTML (HyperText Markup Language) and GML (Generalized Markup Language), as Bernal Chávez and Hincapié Moreno (2018, p.57) explain. JSON (JavaScript Object Notation) language and some standardized formats such as TEI (Text Encoding Initiative) are also used very frequently, according to Molina Mejía (2021). Thus, the result of a tagging process is usually a text in a format different from the original, in which part of its implicit information is made visible.

The information that could be included in corpus annotation can be as wide as the elements that play a role in communication are different and as varied as the objectives that each researcher has when planning the construction of the corpus. In this sense, there is great freedom in choosing what will be explicit in the tags of a corpus. However, in current work it is possible to note that some forms of tagging have become standardized.

Two common types of annotations are the syntactic parsing, which focuses on analysis of the functions that each word fulfils in the syntax of the sentence (Parodi, 2010, p.40) and the POS (Part-of-Speech) tagging, also known, following Mitkov (2004), as morphological or lexical annotation. Although the term part-of-speech refers to something specific, this type of tagging usually presents, in addition to this data, information on gender, number, case, tense, mood, aspect and person (p.225).

There are different approaches to perform this task. For McEnery and Hardie (2011, p.49), a corpus can be tagged manually, automatically or an automatic process followed by a manual review. The application of these methods may vary in their margin of error and in the time and effort to be devoted to tagging, but as will be seen below, their choice depends on how easy it is for a researcher to access automatic tagging methods in a given language.

3.4. Under-Resourced Languages

Considering the aforementioned concepts, the importance of having properly compiled and annotated corpora is evident, as well as the availability of tools for automatic language processing in the studies that can be carried out in a given language (Pemberty Tamayo, 2020). Thus arises the concept of Under-Resourced Languages, which can be defined as the set of languages that do not have the computer resources for their automatic processing, as well as the lexicographic and corpus inputs that would serve as the basis for the construction of these tools (Krauwer, 2003).

A definition can also be found in a series of criteria proposed in the works of Krauwer (2003) and Berment (2004), which propose the tools that a language must have in order to be considered as having a basic level of access to computational linguistics technologies. Languages that lack several of these elements are thus considered to be Under-Resourced Languages:

a Lack of a single writing system or a stable spelling.
b Limited presence on the web.
c Lack of experts in Linguistics.
d Lack of electronic resources for speech and language processing.
e Lack of monolingual corpus.
f Lack of electronic bilingual dictionaries.
g Lack of transcribed oral corpus.
h Lack of pronunciation dictionaries and vocabularies.

As Maxwell & Hughes (2006, p.29) mention, the availability of such tools in a language, coupled with other extralinguistic factors, can greatly influence a researcher's decision to work with it. This means that the lack of tools makes research in some languages less frequent and, therefore, the creation of the same tools could be slow and difficult. The availability of these elements, at the same time, makes different applications of information and communication technologies, such as machine translation or digital dictionaries, available

to speakers of the language. That is why filling the gap in terms of tools for computational processing in these languages is not only an academic interest, but also benefits the communities in which the language is spoken (Pemberty Tamayo, 2020).

Based on all the topics explored in this section, the need for tools for corpus tagging in Under-Resourced Languages is evident. The UnderRL Tagger tool (Pemberty Tamayo *et al.*, 2020) proposes, through Computational Linguistics, a system that allows manual tagging of large amounts of texts in different languages, with the help of the computer, which provides the facility to speed up the process by a significant proportion. This process can also produce content that can be reused to annotate other corpora in the same language and serve as a basis for the creation of applications that allow the fully automatic tagging of texts (Pemberty Tamayo, 2020; Pemberty Tamayo *et al.*, 2023).

4. Methodological Framework

Before describing the methodology through which this software is built, it is necessary to explain some elements that have served to frame it in a standard that facilitates its use in the current environment.

Taking into account that the main objective of the application has been selected as the POS level in tagging, the use of the EAGLES tag system (Leech & Wilson, 1996) was accepted for this purpose, which allows coding information such as grammatical category, gender, number, etc., in a brief way, through different numbers and letters. An example is shown below:

Table 1. Example of EAGLES tags for a Spanish sentence.

I	*BUY*	*BREAD*
PP1CSN0	VMIP1S0	NCMS000

The table above shows how EAGLES tags are used to specify the information for each of the words. However, these series of letters and numbers must be converted into a markup language that can be computationally processed and parsed. To achieve this goal, the program uses the XML language, which allows assigning individual elements within a series of defining characteristics. Thus, in this language the corresponding tag can be assigned to each of the text components. Both the EAGLES tags and the XML language correspond to standards widely used in the corpus tagging environment, so their use guarantees understanding by a wide variety of researchers in the field, as well as easy integration with previous projects or work that may have been carried out.

4.1 Description of the program structure

The UnderRL Tagger software interface consists mainly of a window that can be interacted with to navigate between corpus files, set tags and save or retrieve previous sessions. This window constantly interacts with other files and folders that record everything necessary to make the tagging process as efficient and correct as possible.

One of the folders is used by the system to store the data of the different dictionaries that are created. The dictionary is a file in which the tags that can be reused in a given corpus are stored, so that it is not necessary to re-enter them manually.

Another important location is the folder where the XML files containing the already tagged texts are stored; this folder is automatically created in the same directory as the original corpus texts. In addition, there is also a set of files that record at all times which annotation projects are running and what their progress is; so, it is easy to interrupt the tagging task at any time and come back to it later.

From here, the program can enter all the texts that make up the corpus, which must be in plain text format (*.txt) and UTF-8 encoding, in which the computer will recognize a wide variety of characters. All of them must be stored in a single folder, the address of which will be entered in the application.

Once the texts are available, the software will proceed to go through each of them, as selected by the user, and perform a process that consists of separating the text by words. Once the words have been separated, the main window shows the user each one of them, allowing the user to select more than one when necessary. For each word, the user can select, through several controls, the characteristics of the word to be tagged and the program takes care of representing them according to the EAGLES model. In addition, a space in the interface permits the creation of new tags or the editing of the default ones; in this way it is possible to expand the tagging possibilities according to the needs outside the POS. Finally, once a tag has been established, the user can save it in the final XML file, where it will be arranged with the rest of the text, with its corresponding tag and a unique identifier.

In addition to simply tagging the word, the user can choose to save that tag in the dictionary, so that each time the same word appears in the corpus, it will be automatically tagged without user intervention. This is how this software helps to greatly automate annotation, as it allows human intervention to be reduced to the points where it is really necessary. Each time the tagger encounters a new word, it looks it up in the dictionary before displaying it on the screen, so the same text can go through considerable chunks before requiring human attention.

As a consequence of this procedure, the dictionary can be strengthened as the tagging progresses, permitting for greater automation and also providing a file that can be used to tag other texts in the same language or as a basis for other programs that require knowledge of these notions for language processing.

When a user perceives that the tagging of a word cannot be automated because it may present variations in its tags throughout the corpus, he can simply choose not to save it in the dictionary, so that each time it appears he will be presented in the main window of the interface and will be allowed to choose the tag he considers appropriate for each occasion, as mentioned in Pemberty Tamayo (2020).

5. Analysis of the algorithms

UnderRL Tagger is a software written in Python language that can be used for semi-automatic tagging of POS in Under-Resourced Languages, putting the methods of Natural Language Processing at the service of Corpus Linguistics, and allowing the tagging process to be significantly speeded up by automating several of its stages (Pemberty Tamayo, 2020; Pemberty Tamayo *et al.*, 2023).

When a user correctly enters the address of a folder containing the texts of a corpus, the first actions performed by the program are to verify the existence of the texts and to create the files and folders necessary to store the records involved in the process (Figure 1), as described in the methodological framework.

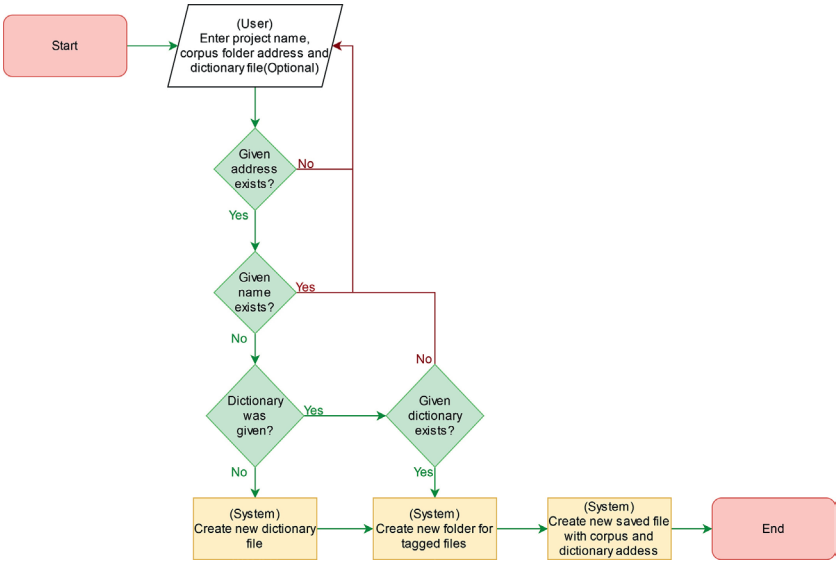


Figure 1. Flowchart: Starting a New Project. Adapted from Pemberty (2020).

All the information that the System stores in addition to the XML tagged texts is in folders that must be in the same directory in which the program is running, and for this purpose files are used that are also in plain text format, so that they can be easily read and modified in case a mistake has been made, for example, by creating an erroneous tag in the dictionary.

Once these files have been prepared, the tool goes on to tag the texts. To exemplify what will happen in each of the steps, we will take here the same sentence that is proposed in the work from which this program arises. This fragment is an example of the Creole language of the islands of San Andres (Colombia) and is shown below along with a brief analysis (Table 2):

Table 2. Description of the "Sentence A" (Pemberty, 2020, p.31).

Sentence A							
Word	*Di*	*bwai*	*gwain*	*da*	*di*	*niu*	*house*
POS	Article	Name	Verb	Preposition	Article	Adjective	Name
Translation	The	boy	goes	to	the	new	house

Before showing the user the texts to be tagged and the diverse options, it is necessary that the text is processed in a specific way. In previous sections it has been said that the text is divided into words and categories are assigned to each of them. In this sense, it is important to specify that the appropriate concept is not that of a word, but that of a token. According to Mitkov (2004), a token is a minimal linguistic unit that can correspond to a word, a number, or a punctuation mark. An important difference between a token and a word is that the latter remains a single element regardless of whether it appears several times in one or in many texts, whereas the former corresponds to a single occurrence, so each of them must be differentiated in relation to the others. The process of dividing a text into its component tokens is called tokenization.

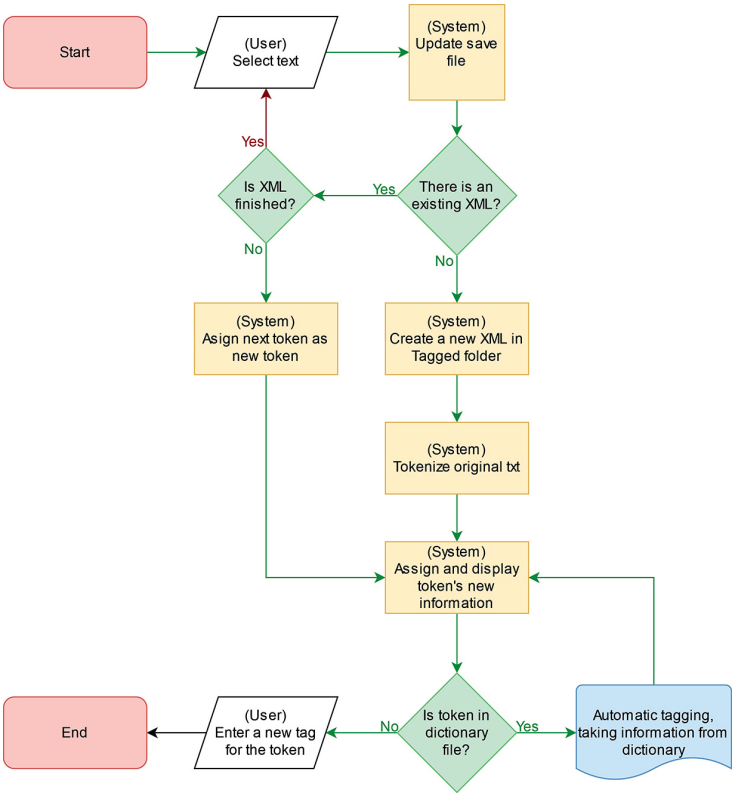


Figure 2. Flowchart: Pre-processing of a selected text. Adapted from Pemberty (2020).

The software checks the file system to see if there is previous information on the same text so that it can be retrieved and continue where the work left off, as well as checking from the first token of the text if there is a set of tags for it in the dictionary, as can be seen in the diagram above. Assuming that this is a new project that has no tags in its dictionary, the result of this process will simply be the tokenized text.

It is also important to note that tokens are usually identified through the blank space between two words; however, there are also many units that are made up of two or more words separated by spaces that would be erroneous to tagged as distinct or non-consecutive tokens. These units are called multi-token words and examples of them can be phrases or some ways of referring to numbers (Mitkov, 2004). To annotate these units, the system offers the possibility of chaining some tokens with others, being able to create a composite unit between one element and the one that follows it.

All the checks seen in Figure 2 are performed automatically by the system, so for the user only a moment passes between selecting a text to tag and the first tokens and controls to set the tags are displayed in the window.

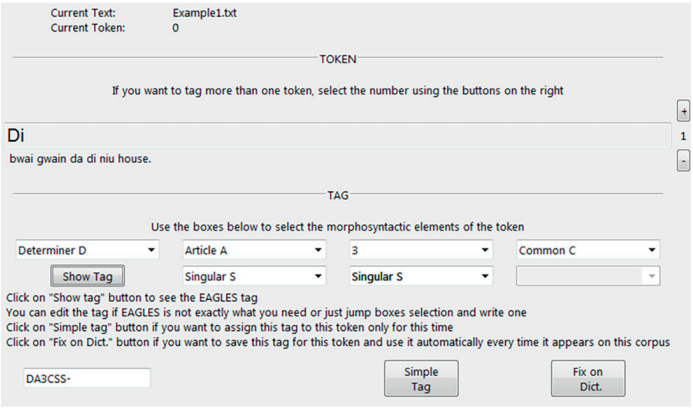


Figure 3. Example of the program window with a tagged unit (Pemberty Tamayo, 2020, p.33).

The program presents the user with the first token of "Sentence A" as well as others that are useful for understanding the context in which each one appears, as shown in Figure 3. Likewise, a series of drop-down lists are enabled for the user that will permit him to choose between distinct categories that could be assigned to the token that is selected. From the various selections, the tag will be created.

The diverse possibilities available to the user vary depending on the first selection to be made, that of the part of speech to be attributed to the token, from which the others are derived. Thus, the amount of information required and its type change when one of these categories is selected.

Once you have selected the appropriate items in the drop-down lists, click on the "Show tag" button, which permits the user to visualize, in the text bar at the bottom, the tag that has been created from the information entered and following the EAGLES system. In the drop-down lists the options are expressed with words commonly used in the field of Linguistics, while the tag only shows its equivalent in the annotation system, as shown in the previous image; in this way, it is not necessary for the user to be perfectly familiar with the EAGLES tags to be able to use them, since the program takes care of establishing which characters are necessary.

The user can already set that tag for that token; however, he be able also to edit it, in case he needs to add additional information of interest for his work. Thus, the tagger per-

mits researchers to create their own tags based on EAGLES or completely new ones, so it could be used not only for URLs, but also in other languages to tag phenomena outside the POS level. This flexibility let the user to work according to the theory or linguistic approach he prefers or needs.

There are also two options to fix the tag and bring it definitively to the output XML file. The first is "Simple Tag", which takes whatever is on the bar where the tag appears and fixes it in the output file associated with that particular token and its ID number.

On the other hand, there is a button called "Fix on Dict". It permits to fix what is written in the tag bar in the dictionary file associated to the selected token; besides that, it performs the procedure of fixing that occurrence of the token in the XML file.

This second option should only be applied when there is certainty that the same tag could be used on all occasions when the same word or combination of words occurs in the token. This can easily be applied to articles, punctuation marks, prepositions, or adverbs, and even to most nouns, adjectives and verbs. This feeds the dictionary, which will be used to automatically tag tokens that match the information it contains. For cases where the tag may vary, the first option will be used, as the absence of that tag in the dictionary will always prompt the user to manually select the appropriate categories. An example dictionary file is shown below:

```
entry_ . **** Fp
entry_ bwai ***** NCMS ---
entry_ di ***** DA-CNS-
entry_ house ***** NCFS---
entry_ niu ***** AQ-CS--
```

Figure 4. Tokens and dictionary entries (Pemberty Tamayo, 2020, p.38).

As shown in Figure 4, this file consists of several lines of text that associate each token with the tag that has been assigned to it. The characters found at the beginning and in the middle of each line are used by the system to differentiate these two elements. The dictionary lookup consists of going through this set of alphabetically ordered lines and taking from them the tag if a match is found, and then taking it to the output file.

By constantly repeating the process of feeding the dictionary with new tokens and tags and allowing the tagger to automatically find and fix as many word occurrences as possible, a significant reduction in the effort required to have a fully XML tagged corpus is achieved.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<text name="Ejemplo1.txt">

<token form="Di" tag="DA3CNS" id="t.0.1"/>
<token form="bwain" tag="NCMS--" id="t.1.1"/>
<token form="gwain" tag="VMIP3SC" id="t.2.1"/>
<token form="da" tag="SP----" id="t.3.1"/>
<token form="di" tag="DA3CNS" id="t.4.1"/>
<token form="niu" tag="AQ-FS-S" id="t.5.1"/>
<token form="house" tag="NCFS--" id="t.6.1"/>
<token form="." tag="Fp" id="t.7.1"/>
</text>
```

Figure 5. Final XML example.

Finally, Figure 5 illustrates what "Sentence A" tagged with the UnderRL Tagger system would look like in your output file. The XML file has an identification of the text in question and all the tokens that make it up. For each of these tokens, the form information is available, which is the exact way it appears in the text; tag, which is the annotation that was established for it and an ID, which is a number that identifies it and differentiates it from all other tokens in the text. This ID is composed of the letter "t", an integer that refers to the position of the token in the text and another integer that refers to the number of words that make up the token, which varies in the case of multi-token words.

6. Conclusions and Perspectives

During this chapter we have seen how it is possible to use Natural Language Processing applications in corpus tagging in languages that do not yet have access to automatic annotation tools, making it possible that, through diverse processes, to achieve a part of what would be enormously expensive if executed completely manually.

The UnderRL Tagger software (Pemberty Tamayo *et al.*, 2020), the tool described in the previous pages, aims to bring URLa closer to information and communication technologies, as well as to facilitate to have them as an object of investigation. For all these reasons, as we have seen in the theoretical framework of this chapter, the existence of computer tools capable of processing and tagging corpora in these languages is of utmost importance.

Thus, through a window-based interface and simple controls, UnderRL Tagger enables a highly computer-assisted and automated manual handling tagging process, offering users the possibility to adhere to international standards in the field of Corpus Linguistics, choose their own tagging system and even annotate outside the POS with any other desired phenomena. Similarly, it allows the management of dictionary files that can be used in the future to further tag texts in the same language or share them with other researchers. Finally, it is important

to note that this software is freely available and can be found in the repository of the main author of this work: https://github.com/jluispemberty/UnderRlTagger.

References

Anthony, L. (2015). *TagAnt (Version 1.2. 0)[Computer Software]*. Waseda University. http://www.laurenceanthony.net/software/tagant/

Baquero, J. M. (2010). *Lingüística computacional aplicada*. Universidad Nacional de Colombia.

Beaudouin, V. (2016). Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis. *Glottometrics*, 33.

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues " peu dotées "* [PhD Thesis, Université Joseph-Fourier - Grenoble I]. https://tel.archives-ouvertes.fr/tel-00006313

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

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100.

Biber, D., & Finegan, E. (2014). On the Exploitation of Computerized Corpora in Variation Studies. In *English Corpus Linguistics* (pp. 216-232). Routledge.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Buseman, K., & Buseman, A. (2013). *Field Linguist's ToolBox (Version 1.6.1)*. SIL International. https://software.sil.org/toolbox/

El-Haj, M., Kruschwitz, U., & Fox, C. (2015). Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation*, *49*(3), 549-580.

Grajales Ramírez, A. & Molina Mejía, J. (2019). Problemática actual del procesamiento computacional anafórico: el caso de FreeLing 4.1. *Lenguaje*, *47*(2S), 537-568.

Jones, C. & Waller, D. (2015). *Corpus Linguistics for Grammar: A Guide for Research*. Routledge.

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM 2003* (pp. 8-15).

Le, V.-B., & Besacier, L. (2009). Automatic speech recognition for under-resourced languages: Application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(8), 1471-1482.

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

Maxwell, M., & Hughes, B. (2006). Frontiers in linguistic annotation for lower-density languages. In T. Baldwin, F. Bond, A. Meyers, & S. Nariyama (Eds.), *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006* (pp. 29-37). Association for Computational Linguistics. https://aclanthology.org/W06-06

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics*. Edinburgh University Press.

McEnery, T., & Hardie, A. (2013). The history of corpus linguistics. *The Oxford handbook of the history of linguistics*, *727*, 745.

Mitkov, R. (2001). Outstanding Issues in Anaphora Resolution. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 110-125). Springer.

Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics*. OUP Oxford.

Mitkov, R. (2013). *Anaphora Resolution*. Routledge.

Moe, R. (2008). FieldWorks Language Explorer 1.0. *SIL Forum for Language Fieldwork 2008-011*. SIL Forum for Language. https://www.sil.org/resources/publications/entry/7793

Molina Mejía, J. M. (2021). *Lingüística computacional y de corpus: teorías, métodos y aplicaciones*. Editorial Universidad de Antioquia.

Moreno Sandoval, A. (1998). *Lingüística computacional: Introducción a los modelos simbólicos, estadísticos y biológicos*. Editorial Síntesis.

Nerbonne, J. (2007). Linguistic Challenges for Computationalists. In N. Nicolov, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005* (pp. 1-16). John Benjamins Publishing.

Padró, L., Collado, M., Reese, S., Lloberes, M., & Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & Daniel Tapias (Eds.), *7th International Conference on Language Resources and Evaluation* (pp. 931-936). European Language Resources Association (ELRA).

Parodi, G. (2010). *Lingüística de corpus: De la teoría a la empiria*. Iberoamericana.

Pemberty Tamayo, J. L. (2020). *Concepción y elaboración de un sistema de etiquetado semiautomático para under-resourced languages* [trabajo de grado, Universidad de Antioquia]. Grupo de Estudios Sociolingüísticos]. Repositorio Institucional Universidad de Antioquia. https://bibliotecadigital.udea.edu.co/handle/10495/16570

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.

Pemberty Tamayo, J. L.; Molina Mejía, J. M. & Vallejo Zapata, V. J. (2023). UnderRL Tagger: un etiquetador gramatical para lenguas infrasoportadas tecnológicamente y lenguas minoritarias. *Forma y Función*, *36*(2). https://doi.org/10.15446/fyf.v36n2.101984

Poesio, M., Stuckardt, R., & Versley, Y. (2016). *Anaphora Resolution*. Springer.

Rogers, C. (2010). Review of fieldworks language explorer (flex) 3.0. *Language Documentation & Conservation*, *4*, 78-84.

Sáiz Noeda, M. (2002). Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español. *Procesamiento del lenguaje natural*, *28*, 113-114.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, *4*, 5-15.

Schmid, H. (1994). *TreeTagger-a language independent part-of-speech tagger*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Schuster, S. & Manning, C. D. (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In *LREC 2016*.

Straka, M. & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In J. Hajič, D. Zeman (Eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88-99). Association for Computational Linguistics. https://aclanthology.org/K17-3

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing.

Tordera Yllescas, J. C. (2011). *Lingüística computacional: Tecnologías del habla*. Publicacions de la Universitat de València.

Wallis, S. (2020). *Statistics in Corpus Linguistics Research: A New Approach*. Routledge.

Wilks, Y. (2010). Corpus Linguistics and Computational Linguistics. *International Journal of Corpus Linguistics*, *15*(3), 408-411.

Zeroual, I. & Lakhouaja, A. (2018). Data Science in Light of Natural Language Processing: An Overview. In J. Boumhidi, P. Érdi, Y. Ghanou, E. H. Nfaoui, & Y. Oubenaalla (Eds.), *Procedia Computer Science 127* (pp. 82-91). https://doi.org/10.1016/j.procs.2018.01.101

DATA SCIENCE, CULTURE & SOCIAL CHANGE

"*Digital Humanities, Corpus and Language Technology: A look from diverse case studies is* an outstanding collection of research contributions that explores the intersection of technology and the humanities. The authors provide a comprehensive overview of how these technologies can enhance research across various disciplines, from literature to history to anthropology. This book is a must-read for anyone interested in future research in the humanities. Digital Humanities, Corpus, and Language Technologies are rapidly growing fields that have the potential to revolutionize research across various disciplines. New technologies have opened up new perspectives for research, allowing scientists to analyze data in previously impossible ways. The interdisciplinary approach and practical applications make it an invaluable resource for researchers, students, and anyone interested in the intersection of technology and the humanities."

Andrés Grajales Ramírez is a Hispanic philologist from the University of Antioquia (Colombia) and holds a Master's degree in Cinematografía from the University of Córdoba (Spain).
Jorge Molina Mejía is an associate professor in the area of linguistics at the University of Antioquia, professor of computational linguistics and Spanish as a foreign language, coordinator of the research group Corpus Ex Machina, he is part of the Committee of the Doctorate in Linguistics of the Faculty of Communications and Philology (University of Antioquia).
Pablo Valdivia Martin is Chair-Full Professor of European Culture and Literature (University of Groningen), Accreditated Full Professor [Catedrático Universidad] of Arts and Humanities (ANECA, Spain), Associate in Applied Physics at Harvard Paulson School of Engineering and Applied Sciences (Harvard University), Academic Director of the Netherlands Research School for Literary Studies (OSL), Scientific Advisor of the Netherlands Institute of Advanced Studies in Social Sciences and Humanities and the Netherlands Royal Academy of Arts and Sciences (NIAS-KNAW), Coordinator Research Theme Group Data Science, Culture & Social Change at Research Centre for the Study of Democratic Cultures and Politics (DemCP, RUG), Co-Editor of the Routledge Companions to Hispanic and Latin American Studies and Research Fellow "Corpus Ex Machina" Research Group Incubator (UdeA).