Dialectología digital del español

Edición a cargo de
Ángel J. Gallego
Francesc Roca Urgell

Verba
Anexo 80

2020
Servizo de Publicacións e Intercambio Científico • UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DIALECTOLOGÍA DIGITAL DEL ESPAÑOL

Edición a cargo de
Ángel J. Gallego
Francesc Roca Urgell

Verba
Anexo 80

2020
Universidade de Santiago de Compostela

Edita
Servizo de Publicacións
e Intercambio Científico
Campus Vida
15782 Santiago de Compostela
www.usc.es/publicacions

DOI https://dx.doi.org/10.15304/9788418445316
ISSN 2341-1198
ISBN 978-84-18445-31-6

ÍNDICE

Alcalá), Ana María Cestero Mancera (Coordinadora técnica de PRESEEA, Universidad de Alcalá)

139 Humanidades digitales y geografía lingüística: la edición digital del *Atlas Lingüístico de la Península Ibérica*

Xulio Sousa (Instituto da Lingua Galega, Universidade de Santiago de Compostela)

159 *Variación Gramatical del Español en el Mundo* (VARIGRAMA): una visión panorámica de los rasgos sintácticos del español

Toshihiro Takagaki (Universidad de Estudios Extranjeros de Tokio), Hiroto Ueda (Universidad de Tokio), Antonio Ruiz Tinoco (Universidad Sofía)

189 Codificación y anotación del habla en un contexto bilingüe: el corpus ESLORA de español de Galicia

Victoria Vázquez Rozas (Universidade de Santiago de Compostela), Mario Barcala (NLPgo Technologies, S.L.), Eva Domínguez Noya (Centro Ramón Piñeiro para a Investigación en Humanidades, Universidade de Santiago de Compostela), Alba Fernández Sanmartín (Universidade de Santiago de Compostela), Guillermo Rojo (Universidade de Santiago de Compostela), María Paula Santalla (Universidade de Santiago de Compostela)

INTRODUCCIÓN[*]

Introduction

ÁNGEL J. GALLEGO
CLT, Universitat Autònoma de Barcelona

FRANCESC ROCA URGELL
GLG, Universitat de Girona

Este volumen monográfico ofrece una visión panorámica de una serie de iniciativas actuales centradas en el estudio de la variación lingüística del español. Tales iniciativas tienen como denominador común la aplicación de tecnologías digitales, la ampliación de los objetivos y áreas de estudio de la investigación filológica tradicional y la voluntad de proporcionar nuevos datos y herramientas tanto al investigador de diferentes aspectos de la lengua española como a toda aquella persona interesada en adentrarse en el conocimiento de esta lengua y de sus variedades. Los artículos recogidos giran en torno a recursos que han dado un nuevo impulso a los estudios de variación y/o han planteado nuevas perspectivas o retos, ya sea porque se trata de proyectos innovadores o porque, siendo proyectos con una trayectoria larga, recientemente han actualizado sus planteamientos y han conseguido mejorar tanto sus prestaciones hacia el usuario como la recogida, tratamiento, clasificación y presentación de los datos.

Todas las iniciativas que recogemos son significativas de la relevancia que ha adquirido la dialectología digital, así como de los caminos que puede seguir en su evolución futura, y constituyen una muestra de cómo el progreso científico y tecnológico ha servido para revitalizar los estudios de dialectología superando limitaciones anteriores y profundizando en el análisis de los datos; es decir, proporcionándonos, en definitiva, un mejor y más sólido conocimiento de la lengua.

LA VARIACIÓN EN LOS ESTUDIOS LINGÜÍSTICOS

El estudio de la variación lingüística no es nuevo. La concepción de la lingüística como disciplina con presupuestos científicos incorporó la comparación entre lenguas (o sistemas lingüísticos) desde sus inicios en el siglo XIX, como bien reflejan los trabajos en el marco de la lingüística comparada e histórica (primera mitad de siglo) y del programa neogramático (segunda mitad). Estos trabajos reconstruían estadios previos de una lengua y establecían relaciones entre lenguas a partir de correspondencias formales y semánticas basadas en la idea de la regularidad de los cambios fonéticos. En el ámbito particular de la dialectología del español, destaca, desde principios del siglo pasado, una obra que se erigió en referencia de trabajos posteriores: el *Atlas Lingüístico de la*

[*] Esta investigación ha sido parcialmente subvencionada por las ayudas concedidas por el Ministerio de Economía y Competitividad (FFI2017-87140-C4-1-P [Gallego] y FFI2017-87140-C4-2-P [Roca Urgell]), la Generalitat de Catalunya (2017SGR634 [Gallego] y 2017 SGR 01194 [Roca Urgell]), y la Institució Catalana de Recerca i Estudis Avançats (ICREA Acadèmia 2015 [Gallego]). Asimismo, agradecemos muy especialmente a Laura Arias y a Cristina Ruiz su colaboración en el proceso de edición de este volumen.

Península Ibérica – ALPI, iniciado bajo la dirección de T. Navarro Tomás en el primer tercio del siglo XX. Este atlas seguía la línea de atlas lingüísticos de lenguas europeas cercanas como el *Sprachatlas des Deutschen Reichs* de Georg Wenke (finales del siglo XIX), el *Atlas linguistique de la France* de Jules Gilliéron (primera década del siglo XX) o el *Atlante Linguistico Italiano* de Matteo G. Bartoli (desde 1924).

La obra de Navarro Tomás se ha alargado en el tiempo (v. Sousa, en este volumen) y fue fuente de inspiración, hasta bien entrada la segunda mitad del siglo XX, para los atlas lingüísticos y etnográficos que han constituido la base de los estudios dialectológicos tradicionales sobre el español tanto en Europa (*Atlas Lingüístico y Etnográfico de Andalucía*, *Atlas Lingüístico y Etnográfico de las Islas Canarias*, *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja*, etc.) como en América (*Atlas Lingüístico-Etnográfico de Colombia, Atlas Lingüístico y Etnográfico de Chile*, *Atlas Lingüístico de México*, etc.). Durante todo este período, los trabajos sobre dialectología se interesaban, esencialmente, por la variación detectada en los componentes léxico y fónico y solían partir de información recogida mediante trabajo de campo basado en cuestionarios aplicados a un tipo muy concreto de informante: rural, de edad avanzada y de sexo masculino.

Dos revoluciones cambiaron el panorama tradicional de la dialectología en Occidente durante el siglo XX. La primera de ellas es de carácter científico y consistió en la aparición y desarrollo de corrientes lingüísticas como el estructuralismo, el generativismo y la sociolingüística. Estas líneas de investigación rompieron con los objetivos, enfoques y métodos de la filología tradicional ampliando el campo de estudio y abriéndolo a nuevas perspectivas y a la interacción con otras disciplinas científicas. La relevancia de la sociolingüística de base laboviana (véanse Labov 1972, 1994, 2001) para avanzar en la investigación sobre variación es evidente: bajo el presupuesto general de analizar las estructuras lingüísticas en el habla de la comunidad (atendiendo a cómo evolucionan y sustentándose en una recogida amplia de datos), incluyó, como objeto de estudio, las variantes lingüísticas de carácter urbano y potenció el conocimiento sobre las condiciones de usos lingüísticos por parte de colectivos diferentes, las actitudes lingüísticas de los diversos miembros de la comunidad y la relación entre variante lingüística y clase, grupo o nación (es decir, entre lengua e identidad, en términos más generales). Por su parte, los planteamientos estructuralistas y generativistas, centrados en el análisis de la lengua como sistema y en la discusión de la naturaleza del lenguaje humano, resultaron muy eficaces para descubrir las propiedades de los componentes gramaticales, así como las limitaciones de los mecanismos de cambio que caracterizan la evolución del sistema lingüístico. La investigación sobre el cambio lingüístico, una noción fundamental en los estudios comparativos entre lenguas y entre estadios sucesivos de una lengua, se ha beneficiado en gran medida de los principios teóricos y metodológicos proporcionados por la sociolingüística y, más recientemente e incluyendo el plano diacrónico, por el generativismo (véanse Batllori *et alii* 2005, Lightfoot 1979, 1999, 2006, Kroch 1994, Roberts y Roussou 2003, Roberts 2007). El enfoque generativista actual ofrece, gracias a la noción de parámetro y a su desarrollo en términos de variación macro y microparamétrica, un marco formal idóneo para el estudio tanto de variantes lingüísticas tipológicamente alejadas (lenguas y familias de lenguas) como cercanas (lenguas próximas o variantes de una misma lengua), respectivamente (véanse Chomsky 1981, Kayne y Cinque 2005, Baker 2008, Roberts 2019).

La segunda revolución es de corte tecnológico (y, por tanto, con una incidencia más directa y más transversal en la sociedad): la irrupción, desarrollo y difusión de la tecnología digital, especialmente, en lo relativo a las denominadas nuevas tecnologías de la información y la comunicación. Estas innovaciones técnicas han dotado (y están dotando) a los lingüistas con un conjunto de herramientas que, primero, permitieron almacenar y clasificar una cantidad enorme de datos lingüísticos (bases de datos, corpus), posteriormente, dieron pie a diseñar procesos específicos para consultar, identificar y

relacionar estos datos (etiquetaje, codificación, frecuencia, localización geográfica), y, actualmente, permiten obtener nuevos datos de manera rápida y a gran escala (internet, *crowdsourcing*, redes sociales, …). Las consecuencias más inmediatas de la incorporación de este tipo de herramientas a la investigación lingüística han sido el crecimiento sin precedentes de la base empírica (el lingüista tiene fácil acceso a grandes cantidades de datos de diferentes estadios diacrónicos, variedades geográficas y registros sociales) y la posibilidad de efectuar análisis más rigurosos, más precisos y de mayor calado.

En combinación con las concepciones teóricas sobre el lenguaje y con los métodos de análisis desarrollados en el seno de las corrientes lingüísticas mencionadas, las innovaciones de carácter técnico han ampliado el abanico de variantes lingüísticas, de fenómenos gramaticales y de condicionantes de uso de la lengua susceptibles de ser analizados y, por tanto, de ser incluidos en la agenda del investigador. A los tradicionales estudios de cuestiones léxicas y fónicas[1], se incorporó, con un protagonismo notorio, el análisis de aspectos morfológicos, sintácticos y pragmáticos. Así pues, la convergencia de los dos factores, el científico y el tecnológico, ha configurado un marco de trabajo del cual los estudios dialectológicos se han beneficiado especialmente. Por un lado, se han superado las limitaciones del enfoque tradicional, pues la metodología clásica basada en cuestionarios cerrados aplicados a sujetos de un entorno rural ha podido ser sustituida por métodos más abiertos de obtención de datos que incluyen técnicas de elicitación, juicios de gramaticalidad, etc. y que son aplicables a diferentes entornos y registros; por otro lado, se ha dado un nuevo aire a los estudios sobre variación llevando los proyectos enmarcados en esta disciplina hacia el centro de la investigación lingüística actual y futura. En consecuencia, actualmente la descripción y análisis de fenómenos de variación está muy presente en las principales obras gramaticales de referencia del español como Bosque y Demonte (1999), RAE-ASALE (2009, 2011) o Gutiérrez-Rexach (2016), y los grandes proyectos de investigación sobre variación dialectal y social se extienden por todo el dominio lingüístico hispánico. Los que se presentan en este volumen son una buena muestra de ello.

CONTENIDO DEL VOLUMEN

En los artículos que siguen a esta introducción el lector comprobará que la digitalización permite no solo dar mayor difusión a los contenidos de los diferentes recursos, sino también aumentar su alcance incorporando nuevos aspectos gramaticales y sociolingüísticos de variación, y ofreciendo posibilidades de complementación y colaboración entre recursos, proyectos y obras de referencia. Algunas de las aportaciones detallan los pasos que se han seguido para tener digitalizados y disponibles en la red una serie de atlas y de corpus, tanto de nuevo cuño como procedentes de obras de referencia tradicionales. En este caso, es interesante advertir el tipo de problemas con que se han enfrentado y cómo ha sido posible introducir diversas mejoras. La información que aportan estos trabajos puede servir de inspiración a iniciativas presentes y futuras en la misma línea. En otros artículos, por el contrario, los autores se concentran en presentar las características de los recursos mostrando mediante casos de estudio concretos cómo puede el investigador obtener el máximo rendimiento y cómo estos recursos se complementan con otras obras de referencia o con otras herramientas también digitalizadas.

Hemos elegido recursos de naturaleza diversa (atlas, bases de datos, habla rural/urbana, dominio lingüístico general/específico, etc.) que responden a proyectos

[1] Los estudios sobre variación léxica y fonética también se han beneficiado considerablemente de los avances tecnológicos. Además de los diversos diccionarios, vocabularios y glosarios disponibles en la red actualmente se dispone de los atlas de entonación como, por ejemplo, el *Atlas Interactivo de la Entonación del Español* (Grup d'Estudis de Prosòdia, Universitat Pompeu Fabra).

con características y con historias diferentes. Así, los hay que son específicos de ciertas zonas del dominio lingüístico como el ALPI, el ALEC, el COSER y el ALCu, que se distinguen entre sí en que, mientras los dos primeros parten de obras tradicionales ya existentes en versión impresa y se han reconvertido (y reinventado) mediante la digitalización, los otros dos ya fueron concebidos en soporte digital (portal web y/o versión en DVD). También se explican las posibilidades que ofrecen dos recursos muy recientes: ASinEs y CORPES. Ambos parten de una conexión muy estrecha con las obras gramaticales de referencia (especialmente las académicas) y ofrecen diferentes procedimientos de consulta, clasificación de datos y obtención de información relacionada que permiten profundizar en la caracterización de los fenómenos. Finalmente, contamos con trabajos cuyo objetivo es detallar las características de proyectos de investigación que amplían muy considerablemente el tipo y la cantidad de datos objeto de estudio atendiendo a factores geográficos y sociolingüísticos, ya sea en un dominio lingüístico amplio (PRESEEA, VARIGRAMA) o acotado a una zona especialmente interesante por su carácter blingüe (ESLORA). En suma, este volumen brinda la oportunidad de conocer bien una serie de recursos digitalizados dedicados específicamente al estudio de la variación lingüística del español y construidos con objetivos y metodologías diferentes.

En suma, los trabajos recogidos en este volumen nos proporcionan información relevante y actualizada para el manejo de atlas, corpus y proyectos de investigación que mejoran el conocimiento de los fenómenos de variación del español en diversos contextos sociolingüísticos y desde perspectivas diferentes. A continuación, ofrecemos una breve visión de cada uno de los artículos.

En el primer artículo Bonilla *et alii* describen las diversas etapas de elaboración por las que ha pasado el *Atlas Lingüístico-Etnográfico de Colombia* (ALEC) desde sus inicios en 1956 hasta la actualidad. Se detalla cómo del proyecto inicial, basado en los tradicionales cuestionarios sobre aspectos léxicos, fonéticos y morfosintácticos, se ha pasado, en el marco del proceso de digitalización en que está inmerso, a un corpus de carácter geolingüístico y oral. El corpus actual incluye las grabaciones originales llevadas a cabo por los encuestadores en la fase inicial, así como la información de los mapas que formaban la versión impresa, que es enriquecida mediante las posibilidades que ofrecen los actuales sistemas de información geográfica. El artículo nos muestra, asimismo, los sistemas de búsqueda y de manejo de mapas interactivos y las líneas de trabajo que se seguirán en el futuro inmediato, que incluyen el etiquetado morfosintáctico de los textos. Este trabajo es un excelente ejemplo de cómo los procesos de digitalización consiguen transformar un recurso de base tradicional en un recurso de futuro incrementando significativamente sus posibilidades.

El siguiente trabajo presenta las propiedades del *Corpus del Español del siglo XXI* (CORPES XXI) poniéndolo en relación con otra obra reciente de referencia: la *Nueva Gramática de la Lengua Española*. Buenafuentes y Sánchez Lancis advierten que uno de los rasgos característicos de la última gramática académica es el alto grado de reconocimiento de la variación morfológica y sintáctica existente en español, y apuestan por el CORPES XXI como la herramienta ideal para detectar su extensión y, en consecuencia, completar la información diatópica y diastrática de la gramática. Los autores apoyan esta conclusión con el estudio de dos fenómenos concretos de variación del español: la combinación de comparativos sintéticos y analíticos, y la discordancia del clítico acusativo. A través de estos estudios nos muestran, por un lado, cómo las diferentes opciones de búsqueda habilitadas en el corpus permiten identificar datos relativos al fenómeno en cuestión y fijan la distribución por país y por tipo de texto y, por otro lado, cómo se pueden extraer las conclusiones pertinentes. Gracias a este trabajo podemos comprobar que CORPES XXI es una herramienta que delimita de una manera fiable la vitalidad de los fenómenos gramaticales gracias a la identificación de la extensión geográfica, la frecuencia y el tipo de texto.

El tercer artículo es el de Castillo, Colomina y Fernández, que analizan las características del *Atlas Sintáctico del Español* (ASinEs), una herramienta de origen muy reciente. Las autoras destacan que este recurso, orientado hacia el estudio de la variación sintáctica en español, consiste en una base de datos construida a partir de datos extraídos de obras gramaticales de referencia y de redes sociales (Twitter). A diferencia de las herramientas de índole más tradicional, ASinEs incluye información gramatical descriptiva y teórica, hecho que resulta especialmente útil para reforzar los análisis que puedan realizar los usuarios. Como muestra de ello, acompañan la explicación de las características y la organización de ASinEs con el estudio de una construcción en particular: la pasivización de dativos en español. La presentación de este estudio, construido a partir de las posibilidades que ofrece la herramienta, es interesante porque ejemplifica cómo el usuario puede abordar un fenómeno concreto y cómo el análisis se enriquece delimitando mejor su distribución geográfica, confirmando sus propiedades gramaticales (advertidas o no en trabajos anteriores) y, eventualmente, aportando una mejor base para descripciones y análisis más refinados.

Fernández-Ordóñez y Pato subrayan el carácter pionero del *Corpus Oral y Sonoro del Español Rural* (COSER) como un corpus que, iniciado a finales del siglo pasado, supera estadios anteriores de la dialectología tradicional hispánica. El corpus está dedicado específicamente al habla rural, añade a los datos de los dialectos históricos datos de zonas como las islas Canarias o las Baleares y presta atención a fenómenos morfosintácticos, superando así las obras de corte tradicional basadas mayoritariamente en la variación fonética y léxica. Se destaca la utilidad del COSER como herramienta que permite comparar con otros registros lingüísticos (por ejemplo, la variante estándar o la norma culta), con variantes urbanas y con la evolución diacrónica. Tras introducir las características generales del corpus y contextualizarlo adecuadamente, los autores muestran sus prestaciones como herramienta consultable *online* enfatizando la posibilidad de acceder a las grabaciones de audio y de realizar búsquedas a partir del etiquetado gramatical. Lo ponen de manifiesto comentando una serie de construcciones de carácter morfosintáctico que han sacado a la luz propiedades relevantes tanto de fenómenos ya recogidos en estudios tradicionales (leísmo, laísmo, formas analógicas en la flexión verbal) como de otros que habían pasado más desapercibidos (neutro de materia, concordancia en verbos existenciales y en infinitivos y gerundios). El lector encontrará aquí una excelente introducción a las características, uso y utilidad de este recurso consultable en línea.

Se comenta a continuación el *Atlas Lingüístico de Cuba* (ALCu). El ALCu es una obra que sigue la tradición de atlas lingüísticos como, por citar dos presentes en este volumen, el ALPI o el ALEC, pero se distingue de estos en que es mucho más reciente, ya que se inició a finales del siglo pasado (en 1983) y fue publicado la pasada década (en 2013). Figueroa describe las diferentes etapas en la breve historia del atlas, que es una aplicación en la web, y el proceso de elaboración que se ha seguido. Parcialmente en línea con la tradición de los atlas lingüísticos del siglo pasado, el ALCu parte de cuestionarios de carácter eminentemente léxico (la presencia de la fonética y de la morfosintaxis es menor), pero los aplican a áreas tanto de carácter rural como de carácter urbano. Se combinan así presupuestos tradicionales como los que guiaron el ALPI o el COSER con otros más actuales como los que caracterizan, por ejemplo, a PRESEEA. La autora explica el tipo de datos que contiene la web del ALCu (grabaciones y transcripciones fonéticas) y su proyección en mapas, y subraya cómo la herramienta ya ha dado lugar a la creación de otros recursos y obras paralelas como el *Atlas lingüístico de Cuba dinámico e interactivo,* el *Atlas lingüístico de Cuba* en soporte digital o el *Diccionario Geolectal de Cuba*. Este trabajo muestra cómo la creación de una nueva herramienta que parte de presupuestos tradicionales puede proporcionar, gracias a las posibilidades de la digitalización, resultados inmediatos de diverso tipo como, por ejemplo, una mejor

identificación de las isoglosas y de las zonas dialectales o la creación de diccionarios geolectales.

El *Proyecto para el Estudio Sociolingüístico del Español de España y América* (PRESEEA) es analizado en este volumen por Moreno y Cestero, quienes destacan su carácter sociodialectológico. Se trata de un corpus que recoge muestras de lengua hablada en grandes ciudades de todo el ámbito hispanófono y, en consecuencia, proporciona información sobre la variación del español tanto en lo referente al registro social como a la zona geográfica. El artículo nos presenta una breve historia del proyecto, de su organización a partir de equipos de trabajo en cada una de las 44 ciudades en que se está desarrollando y de la información que se puede encontrar en el portal abierto al usuario. Además de indicar las características del corpus y de los mecanismos de búsqueda, se ofrece una visión general de los tipos de materiales y trabajos generados, donde, si bien destacan estudios de carácter pragmático y fonológico (variables de fonemas y patrones de entonación), también se ha prestado atención a aspectos léxicos y morfosintácticos. El artículo ofrece, pues, una visión del amplio abanico de cuestiones gramaticales y sociolingüísticas tratadas en el seno del proyecto. Asimismo, es muy destacable la conexión entre PRESEEA y el sistema LYNEAL (*Letras y Números en Análisis Lingüísticos*), que facilita el análisis cualitativo y cuantitativo. Este es un buen ejemplo de cómo herramientas creadas independientemente se pueden unir para facilitar al investigador una mejor comprensión de los datos y, consecuentemente, reforzar la base de los análisis.

Sousa nos presenta la situación actual de la primera gran obra dialectológica del español: el *Atlas Lingüístico de la Península Ibérica* (ALPI). Este artículo es especialmente interesante en la medida en que proporciona al lector una visión actualizada de las características del ALPI, de cómo está diseñado y de cómo se puede manejar. Se relata la historia de la elaboración del atlas y de cómo se ha llevado a cabo su proceso de digitalización (desde la iniciativa de Heap hasta el momento actual, con García Mouton). A lo largo de los diferentes apartados se describe cómo es el material abierto a consulta, cómo se han introducido los datos, cuál es la información que mantiene, cuál es la novedad respecto a los cuadernos originales del atlas y cómo está estructurada la información. Se ofrece, pues, tanto la visión del investigador-creador de la herramienta como la del investigador-usuario, por lo que el artículo resulta útil no solo para situar el ALPI en el contexto actual, sino también para advertir las dificultades que entraña el proceso de digitalización de una obra clásica de esta envergadura y poder influir en el planteamiento, diseño y mejora de otras herramientas similares (presentes y futuras).

Takagaki, Ueda y Ruiz-Tinoco forman parte de los impulsores del proyecto *Variación Gramatical del Español en el Mundo* (VARIGRAMA). Este proyecto recoge datos del habla urbana de varias ciudades de España y de Latinoamérica, de manera que, al igual que PRESEEA, abarca zonas representativas de toda la extensión geográfica del español. En su aportación, estos autores nos describen la herramienta, que contiene datos obtenidos a partir de cuestionarios dirigidos específicamente a aspectos de morfología y de sintaxis, y ejemplifican su uso mediante cuatro fenómenos diferentes: el régimen preposicional, la alternancia entre indicativo y subjuntivo, el leísmo, y el queísmo y el dequeísmo. Los porcentajes de respuestas indican la vitalidad de un determinado tipo de construcción en una zona concreta o en el dominio general y, como indican los autores, se pueden asimilar a índices de aceptabilidad. El artículo también muestra cómo se pueden obtener mapas que reflejen la extensión del fenómeno e incluye en el apéndice la lista de frases empleadas en los cuestionarios.

El volumen se ciera con el artículo de Vázquez *et alii*, donde se exponen las características del *Corpus para el estudio del español oral* (ESLORA) y se describe el proceso de elaboración que se ha seguido. Se destaca que este corpus sobre el español hablado en Galicia se ha construido no solo mediante grabaciones, recurso habitual en

la mayoría de trabajos dedicados a la variación dialectal, sino también mediante un cuestionario que atiende a la situación de bilingüismo en la zona prestando especial atención a las actitudes hacia cada una de las lenguas y a la inseguridad lingüística. Los autores explican cómo se deben hacer las consultas gramaticales en el corpus, muestran el tipo de información facilitada (que incluye transcripción ortográfica y audio) y proporcionan una serie de ejemplos que ponen de manifiesto la variación morfológica respecto a la variante peninsular estándar. Los ejemplos sirven también como muestra de los problemas que surgen y del tratamiento de que son objeto en ESLORA. Una característica especialmente relevante de ESLORA es que la lematización morfológica y la identificación morfosintáctica facilitan análisis de frecuencia que superan los estudios léxicos. Este trabajo nos da la situación actual del corpus, que está en evolución continua, y subraya la relevancia y la utilidad de los corpus orales en relación con factores sociolingüísticos, especialmente en un contexto de blingüismo.

REFERENCIAS BIBLIOGRÁFICAS

BAKER, M. C. (2008): "The macroparameter in a microparametric world", in T. Biberauer (ed.): *The limits of syntactic variation*. Amsterdam: John Benjamins, pp. 351-373.

BATLLORI, M., M.L. HERNANZ, C. PICALLO & F. ROCA (eds.) (2005): *Grammaticalization and parametric theory*. Oxford: Oxford University Press.

BOSQUE, I. & V. DEMONTE (dirs.) (1999): *Gramática descriptiva de la lengua española*. Madrid: Espasa.

CHOMSKY, N. (1981): *Lectures on Government and Binding*. Dordrecht: Foris.

CINQUE, G. & R. S. KAYNE (eds.). (2005): *The Oxford handbook of comparative syntax*. Oxford: Oxford University Press.

GUTIÉRREZ-REXACH, J. (ed.) (2016): *Enciclopedia de lingüística hispánica*. London: Routledge.

KROCH, A. (1994): "Morphosyntactic variation", in K. Beals (ed.): *Papers from the 30th Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, pp. 180-201.

LABOV, W. (1972): *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

LABOV, W. (1994): *Principles of linguistic change. Volume 1: Internal Factors*. Oxford: Blackwell.

LABOV, W. (2001): *Principles of linguistics change. Volume 2: Social Factors*. Oxford: Blackwell.

LIGHTFOOT, D. (1979): *Principles of diachronic syntax*. Cambridge: Cambridge University Press.

LIGHTFOOT, D. (1999): *The development of language. Acquisition, change, and evolution*. Oxford/Malden: Wiley-Blackwell.

LIGHTFOOT, D. (2006): *How new languages emerge*. Cambridge: Cambridge University Press.

RAE - ASALE (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.

RAE - ASALE (2011): *Fonética y fonología. Nueva gramática de la lengua española*. Madrid: Espasa.

ROBERTS, I. (2007): *Diachronic syntax*. Oxford: Oxford University Press.

ROBERTS, I. (2019): *Parameter hierarchies and Universal Grammar*. Oxford: Oxford University Press.

ROBERTS, I. & A. ROUSSOU (2003): *Syntactic change: A minimalist approach to grammaticalization*. Cambridge: Cambridge University Press.

PROYECTO DE DIGITALIZACIÓN Y NUEVAS PERSPECTIVAS TECNOLÓGICAS DEL *ATLAS LINGÜÍSTICO-ETNOGRÁFICO DE COLOMBIA*

Digitization project and new technological perspectives of the Linguistic-Ethnographic Atlas of Colombia

JOHNATAN ESTIVEN BONILLA
Universidad de Gante, Instituto Caro y Cuervo

RUTH YANIRA RUBIO LÓPEZ
Instituto Caro y Cuervo

ANDREA LIZETH LLANOS CHÁVEZ
Instituto Caro y Cuervo

DANIEL EDUARDO BEJARANO BEJARANO
Instituto Caro y Cuervo

JULIO ALEXANDER BERNAL CHÁVEZ
Instituto Caro y Cuervo

Resumen
El presente trabajo describe el corpus geolingüístico del *Atlas Lingüístico-Etnográfico de Colombia* y su proceso de digitalización y puesta en línea. El corpus se compone de 1523 mapas lingüísticos, etnográficos y mixtos, más de 16000 fotografías y 765 sesiones de audio. La información fue recolectada en 262 localidades de Colombia entre 1956 y 1978. Desde el 2015 la línea de investigación en lingüística de corpus y computacional del Instituto Caro y Cuervo ha trabajado en la digitalización de los materiales mediante la construcción de bases de datos y aplicativos en línea que permiten la conservación del material y su consulta y análisis en la web. En total se presentan tres aplicativos que contienen la información del ALEC: el Sistema de Información Geográfica del ALEC (SIG ALEC), ALEC Digital y el Corpus Oral del ALEC (CLICC).

Palabras clave: geolingüística, atlas web, sistema de información geográfica, corpus oral, español de Colombia

Abstract
This paper describes the geolinguistic corpus of the *Linguistic-Ethnographic Atlas of Colombia* and its process of digitization and online publication. The corpus is made up of 1,523 linguistic, ethnographic and mixed maps, and 765 audio sessions. The information was collected in 262

localities in Colombia between 1956 and 1978 with the participation of 2234 informants. The research line in corpus and computational linguistics of the Caro y Cuervo Institute has worked on the digitization of materials since 2015, through the construction of databases and applications in line that allows the conservation of the material and its consultation and analysis on the web. We present three applications in total: the ALEC Geographic Information System (SIG ALEC), ALEC Digital, and the Oral Corpus of ALEC (CLICC).

Keywords: geolinguistics, web atlas, geographic information system, spoken corpus, Colombian Spanish

1. INTRODUCCIÓN

El objetivo del proyecto de digitalización del *Atlas Lingüístico-Etnográfico de Colombia* (ALEC) es el diseño y desarrollo de un corpus geolingüístico y oral en línea con los materiales de la investigación. La construcción del corpus consiste en la sistematización, divulgación y explotación de los materiales a través de herramientas y plataformas virtuales. Este proyecto es liderado por la línea de investigación en Lingüística de Corpus y Computacional (LICC) del Instituto Caro y Cuervo (ICC) con el apoyo del Núcleo de Investigación en Datos Espaciales (NIDE) de la Universidad Distrital Francisco José de Caldas (UDFJC).

Los atlas lingüísticos son producto de largos procesos de investigación que permiten recoger grandes cantidades de datos en distintos puntos de un territorio. Tradicionalmente, una vez finalizada la publicación de los mapas la labor del investigador se daba por terminada, por lo cual, varios autores mencionan la falta de aprovechamiento de los materiales producto de la geografía lingüística (Fernández 2010). De esta manera, uno de los motivos principales del proyecto es facilitar el uso y divulgación de los datos para nuevas investigaciones y el diseño de estrategias pedagógicas para la enseñanza y apropiación del español de Colombia.

Adicionalmente, el uso de herramientas tecnológicas y la posibilidad de almacenamiento de grandes de cantidades de información posibilitan la publicación de materiales que con anterioridad se veía limitada en los atlas en papel. García Mouton explica que la plataforma diseñada para el *Atlas Lingüístico de la Península Ibérica* (ALPI) "permite conservar y ofrecer todo lo que en un atlas tradicional quedaría sepultado en nota y ni siquiera pasaría a los márgenes de un mapa" (2015: 107).

Los datos representados en los atlas suelen ser homogéneos, ya que se obtienen a través de un cuestionario idéntico; con investigadores capacitados en la técnica de la encuesta; informantes con perfiles concretos y coincidentes en características sociolingüísticas; y puntos que cubren un territorio específico (Fernández 2010; García Mouton 2015). Estas condiciones consolidan a los materiales de los atlas en corpus. Los corpus son "un conjunto de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, y seleccionados y ordenados, de acuerdo con criterios explícitos, para ser utilizados como modelo de un estudio o nivel de lengua determinado" (Santalla 2005: 45-46). Los materiales del ALEC son una colección de mapas, audios, fotografías e imágenes recogidos a través de encuestas recopiladas en una red de puntos del territorio colombiano que pueden servir como modelo del español de Colombia de 1958 a 1978. Cabe resaltar que los mapas, fotografías e imágenes del ALEC se encontraban en formato de papel y los audios en cintas de carrete abierto, por ende, ha sido necesario realizar varios procesos para el paso de estos formatos al electrónico. Estos procedimientos son fundamentales para la consolidación de los materiales en

corpus, puesto que en la actualidad una de las principales características de estos es la posibilidad de manejar los datos de forma electrónica.

Ahora bien, la información recolectada en los atlas lingüísticos tiene como eje central la relación de las muestras y los datos con su ubicación espacial. De ahí que cualquier dato que haga parte del corpus estará relacionado con un punto del territorio establecido. De esta manera, podemos decir que los materiales de los atlas sistematizados e informatizados se consolidan como corpus geolingüísticos. Por lo cual, llamamos a nuestro corpus "el corpus geolingüístico del ALEC".

En primer lugar, describiremos los aspectos más importantes de la investigación del ALEC. En segundo lugar, presentaremos el corpus geolingüístico compuesto por tres sistemas principales: la base de datos espacial, el Sistemas de Información Geográfica (SIG) y el Atlas Web. Para continuar, expondremos el corpus oral del ALEC y su relación con el corpus geolingüístico. En último lugar, hablaremos de las perspectivas futuras y conclusiones.

2. EL *ATLAS LINGÜÍSTICO ETNOGRÁFICO DE COLOMBIA* (ALEC)

El objetivo de la investigación ALEC fue conocer de primera mano las principales características del español de Colombia y, a partir de esto, poder establecer diferencias y afinidades con otras variedades del español como el peninsular y su relación con las lenguas precolombinas (Buesa Oliver & Flórez 1954). La investigación se inició con una prueba piloto en 1956 y durante veintidós años se recorrieron un total de 264 localidades pertenecientes a veintiocho de los treinta y dos departamentos del país. En total participaron 2234 informantes y 23 encuestadores. Los resultados de la investigación se publicaron en seis tomos, cada uno de 50 x 35 cm, que reúnen 1696 láminas, de las cuales 1523 son mapas lingüísticos, etnográficos o mixtos. Según Flórez (1983), en los mapas lingüísticos (figura 1) se registran los nombres que los informantes dieron a los conceptos por los cuales se preguntó en las encuestas y los mapas etnográficos muestran las áreas o la difusión geográfica de cosas u objetos de la vida material popular. Como material complementario fueron publicados un índice alfabético, un manual, con información adicional de las localidades, informantes y encuestadores, y un suplemento del Tomo III que incluye muestras de habla espontánea y dos discos de vinilo con grabaciones de juegos y cantos de velorio de las costas caribe y pacífica colombianas.

Para la recolección de información se realizaron encuestas directas a los informantes con base en un cuestionario lingüístico de 1600 preguntas en los niveles léxico, fonético y morfosintáctico. El cuestionario, al igual que los tomos producto de la investigación, fue dividido en dieciséis campos semánticos relacionados con la vida en el campo y cotidianidad de los informantes: el cuerpo humano, el vestido, la vivienda, la alimentación, la familia y el ciclo de vida, las instituciones y vida religiosa, las festividades y distracciones, el tiempo y el espacio, el campo --los cultivos y otros vegetales, industrias relacionadas con la agricultura, ganadería, animales domésticos, animales silvestres, oficios y empleos, transporte y embarcaciones y pesca. Sumado a esto, encontramos preguntas concernientes a aspectos fonéticos y gramaticales, específicamente, onomatopeyas en el Tomo II, variación fonética de vocales y consonantes en diversos contextos de palabra y fenómenos gramaticales en el Tomo VI (Bonilla *et al.* 2017). Gracias al interés etnográfico de la investigación, de manera no sistemática, también se recolectaron registros de habla, fotografías y objetos. Algunos de estos, como los juegos y cantos de velorio, fueron publicados en el suplemento o como material adicional de los mapas, sin embargo, las fotografías y sesiones de grabación quedaron archivadas.

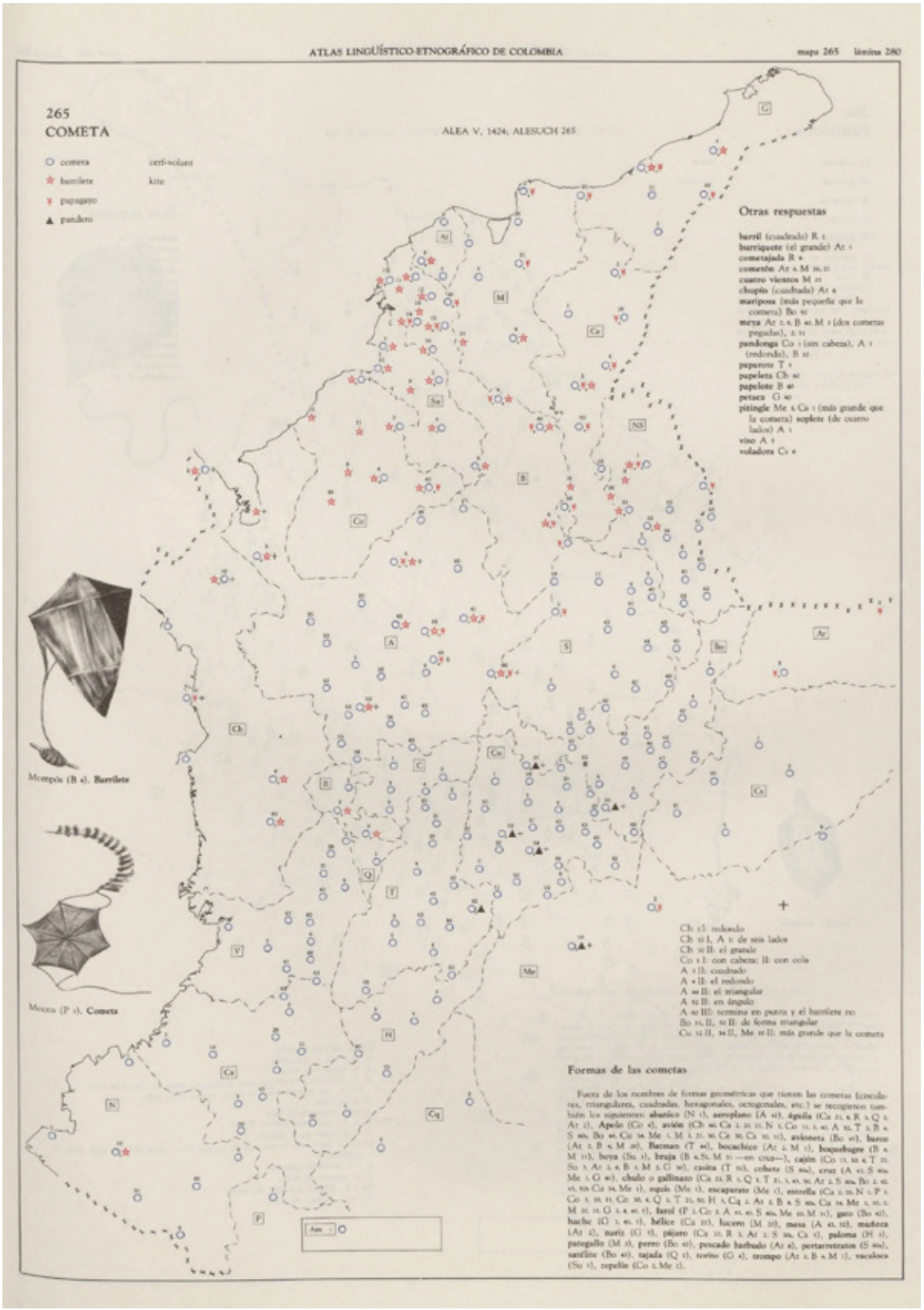


FIGURA 1. Mapa digitalizado del ALEC - Tomo III, Lámina 280, Mapa 265, Cometa

De acuerdo con Flórez (1983) los informantes seleccionados eran nativos de la localidad o habían vivido en ella la mayor parte de su vida. También se tuvo como criterio que los informantes contaran con poca instrucción escolar y fueran de origen campesino. En cuanto a la edad había preferencia por adultos entre los 40 y 60 años. Al sistematizar la información contenida en el *Manual del Atlas Lingüístico-Etnográfico de Colombia* (*Manual*) (cf. Flórez 1983) encontramos que el 32.9% de los informantes fueron mujeres y el 67.1% hombres. Los informantes se encontraban en edades entre los 16 a 100 años distribuidos de la siguiente manera:

Edades	Porcentajes de informantes	Edades	Porcentajes de informantes
16 a 20 años	0.45	61 a 70 años	14.80
21 a 30 años	5.42	71 a 80 años	6.95
31 a 40 años	18.64	81 a 90 años	0.99
41 a 50 años	28.02	91 a 100 años	0.50
51 a 60 años	24.10	Más de 100 años	0.09

TABLA 1. Rangos de edad de informantes del ALEC

En cuanto a la formación académica, el 21,5% eran analfabetas y más del 70% contaban con estudios incompletos de escuela primaria. Las ocupaciones más destacadas eran agricultura (32.7%), oficios varios o domésticos (27.9%), ganadería (3.9%), comercio (3.8%), carpintería (2.6%) y pesca (2.3%).

En cuanto a las localidades el criterio principal era de tipo geométrico, es decir, que se encontraran a distancias relativamente iguales unas de otras y se cubriera la mayor proporción posible del territorio (Flórez 1983), la geografía colombiana obligó a los encuestadores a ir seleccionando las localidades a medida que se realizaba el trabajo de campo, en ocasiones teniendo en cuenta criterios cronológicos, con relación a la época de fundación de la localidad, demográficos, por la cantidad de habitantes, topológicos, climáticos y socioeconómicos, buscando siempre la representación de la biodiversidad geográfica y cultural del país.

3. EL CORPUS GEOLINGÜÍSTICO

El diseño y desarrollo de una herramienta en línea que permitiera el tránsito de los datos impresos en el ALEC para su posterior consulta y análisis se ha venido realizando en tres etapas. La primera etapa (2016) consistió en el modelamiento de una base de datos espacial con base en las particularidades de los materiales del ALEC y las necesidades a futuro del proyecto. Sumado a esto, a partir del diálogo de disciplinas como la lingüística y la geomática se llevó a cabo el levantamiento de los requerimientos técnicos del software esperado mediante la definición de los objetivos del sistema, de los usuarios y las especificaciones de software, de lo que se destaca la disposición hacia la

implementación de software libre. En el mismo orden también se dio inicio a la digitación de la información contenida en el *Manual*, el Tomo III y el suplemento con el fin de contar con datos suficientes para pruebas e implementación de los sistemas y se llevó a cabo la digitalización en alta resolución de los materiales impresos con el apoyo de la Biblioteca Nacional de Colombia para extraer símbolos, imágenes, ilustraciones, partituras y textos.

3.1. Base de datos espacial del ALEC

Con base en la información del *Manual*, la base de datos espacial del ALEC fue diseñada para contener información de las localidades en cuanto a nombre, departamento, latitud, longitud, año de fundación, altura sobre el nivel del mar, población, actividades económicas, vías de acceso, fecha de encuesta, comentarios y página web; de los informantes relaciona el nombre, apellidos, edad, sexo, ocupación, nivel de escolaridad, proveniencia o lugar de nacimiento, viajes y la proveniencia de sus padres y cónyuge; sobre los encuestadores encontramos nombres, apellidos, años de vinculación con el proyecto, localidades donde realizó encuestas y número total de localidades.

Si bien actualmente es posible cargar la información de los mapas directamente a la base de datos espacial mediante las herramientas de administración del Sistema de Información Geográfica (SIG), en principio fueron diseñadas hojas de cálculo que sirvieron también para el modelamiento. En la primera hoja, denominada mapa, fue digitada la información concerniente a título del mapa, nombre científico, número del mapa, número de la lámina, referencia a otros atlas lingüísticos, campo semántico y el nombre de su imagen digitalizada para posterior ubicación en el servidor. En la segunda hoja se asignaba un número identificador a cada variante y se relaciona con un código asignado al símbolo que la representa en el mapa. Es importante distinguir que en el ALEC existen variantes cartografiadas y no cartografiadas. Las variantes cartografiadas cuentan con un símbolo que se despliega sobre el mapa para indicar el lugar de realización, mientras que las variantes no cartografiadas son variantes de menor frecuencia y con el fin de no sobrecargar el mapa impreso fueron georreferenciadas mediante la asignación de un código único de la localidad conformado por las primeras letras del nombre del departamento y un número y puestas a la izquierda del mapa (figura 1). La tercera hoja y cuarta hoja relacionaban el número identificador de la variante cartografiada o no cartografiada con el identificador de la localidad. Por, último, la quinta hoja incluía la información adicional del mapa (audio, texto, fotografía, ilustración, partitura,) asignando un identificador que luego era georreferenciado en la sexta hoja relacionándolo con la localidad.

Con base en lo anterior se llevó a cabo el modelamiento de una Base de Datos Espacial (figura 2) que se reproduce en un servidor de base de datos *PostgreSQL* y funciona para los dos desarrollos realizados durante la segunda etapa: un SIG[1] que cuenta con herramientas de administración de la base de datos en términos de consulta, edición y adición de información, consultas avanzadas y cruces de información y herramientas de análisis geoespacial y un Atlas Web[2] que permite la consulta simple de los mapas del ALEC y la información contenida en el manual.

[1] http://atlasweb.caroycuervo.gov.co/
[2] http://alec.caroycuervo.gov.co/alec/

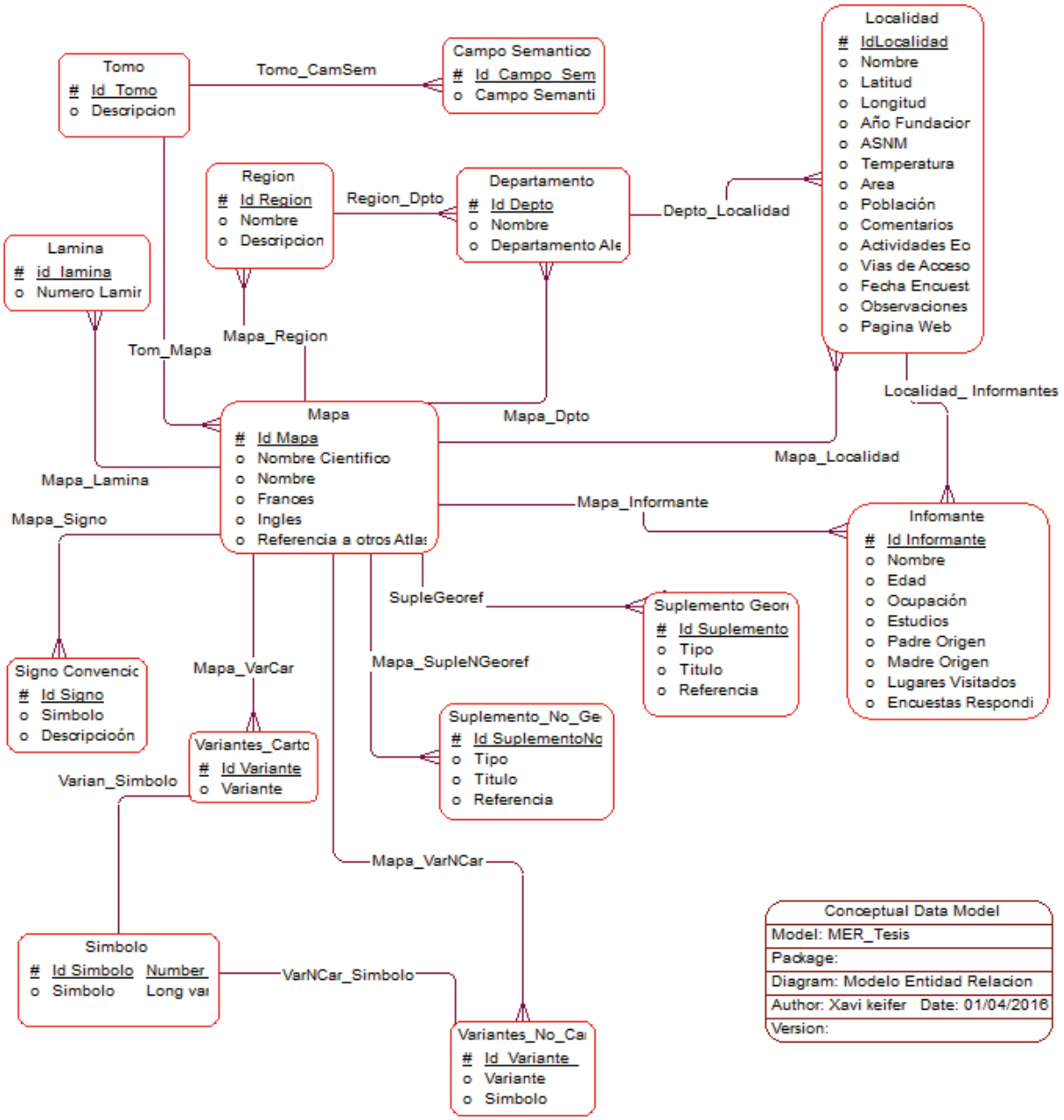


FIGURA 2. Modelo de Base de Datos Espacial del ALEC

3.2. Sistema de Información Geográfica y Atlas Web

Durante la segunda fase del proyecto (2017), además de la implementación de la estructura de base de datos y conexión con el esquema de base de datos del SIG, se llevaron a cabo las tareas de diseño y programación de las interfaces de administración y usuario del SIG y del Atlas Web mediante la implementación del servidor de mapas Geoserver. Geoserver sirve de puente entre la información geográfica la BDE y la interfaz de usuario ya que permite la creación de servicios geográficos para luego utilizarlos a través de la aplicación web acercando al usuario final con la información del SIG y el Atlas Web. Además, es importante resaltar que los servicios geográficos cumplen con los

estándares de la *Open Geographic Consortium*. El servidor de mapas *GeoServer* utiliza *PostGIS* como su base de datos espacial conectada como se mencionó anteriormente a *PostgreSQL*.

La biblioteca para mapas interactivos implementada tanto en el SIG como en el Atlas Web es *Leaflet* ya que, aunque es ligera, tiene todos los elementos que se pueden necesitar para los mapas online. Está diseñado teniendo en mente la simplicidad, el rendimiento y la usabilidad. Trabaja eficientemente con la mayoría de las plataformas de escritorio y móviles. Se puede ampliar con muchos complementos. Tiene una API fácil de usar, bien documentada y un código fuente simple y legible.

A grandes rasgos el SIG del ALEC cuenta con distintas capas de acceso y ha sido diseñado específicamente para tareas avanzadas. El SIG, en su interfaz de usuario (figura 3), cuenta con seis funciones. La primera denominada "seleccionar" permite llamar un mapa del ALEC siguiendo las rutas de Tomo, Campo Semántico, Mapa o escribiendo directamente una variante y acceder a toda la información relacionada como audios, imágenes o texto. En el mismo menú es posible consultar la información relacionada con encuestadores, informantes y localidades.

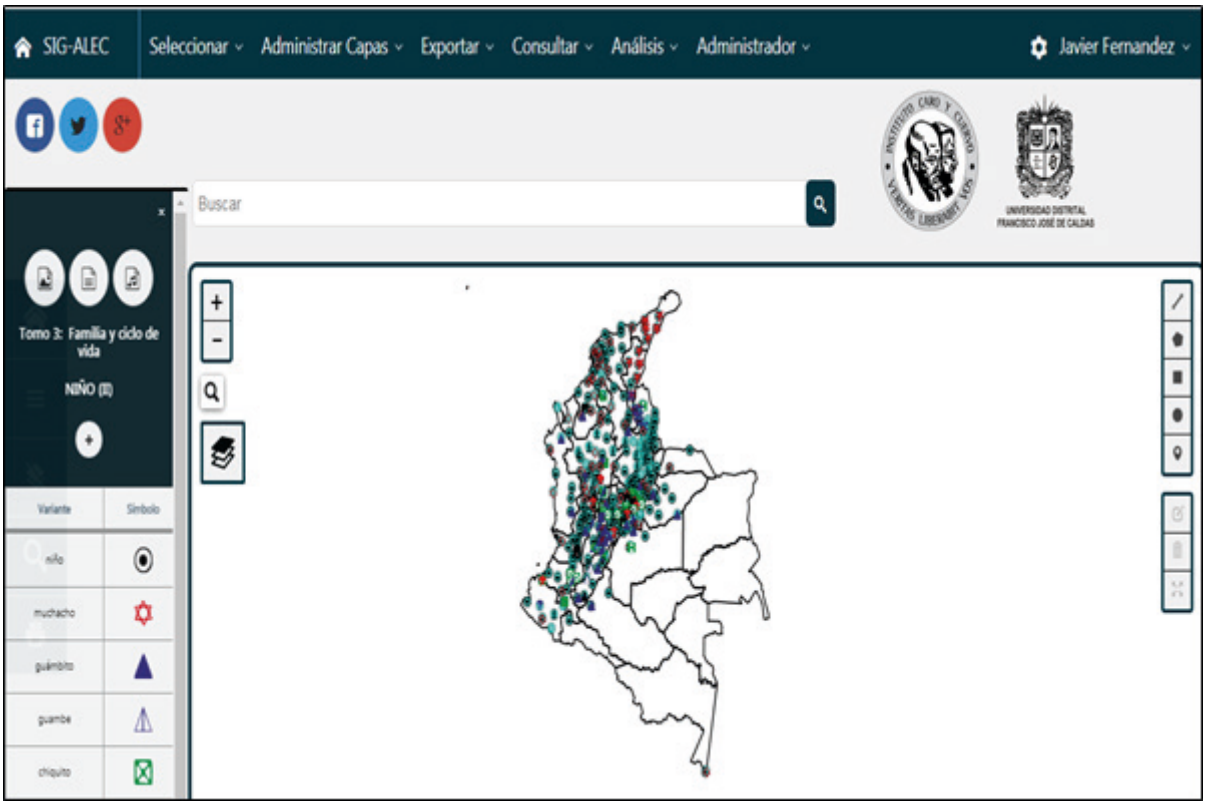


FIGURA 3. Interfaz de usuario del SIG ALEC - Consulta del mapa *Niño (II)*

La segunda funcionalidad se denomina "administrar capas", allí es posible tomar múltiples mapas, variantes, informantes, encuestadores y/o localidades e irlos agregando al mapa de acuerdo con las necesidades del investigador (figura 4). La tercera funcionalidad denominada "exportar" permite acceder a los datos en formatos compatibles (shapefile, kml, CSV) con cualquier SIG de escritorio u otro software de manejo de datos espaciales para su implementación y análisis.

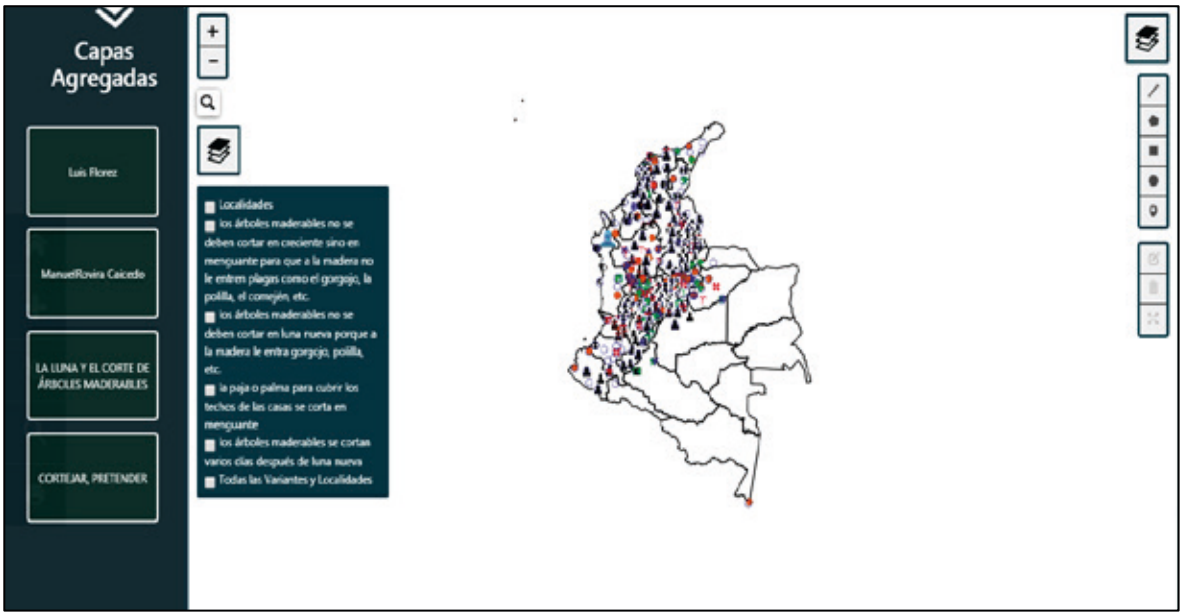


FIGURA 4. Consulta de múltiples capas. Localidades donde realizó encuesta *Luis Flórez*. Localidad del informante *Manuel Rovira Caicedo*. Variantes de los mapas *La luna y el corte de árboles maderables* y *Cortejar, Pretender*

En cuarto lugar, contamos con la herramienta "consultar" que permite realizar cruces en la información de la base de datos, es decir, se supera la mera representación sobre el mapa y posibilita utilizar los metadatos concernientes a informantes, localidades, encuestadores y mapas con el fin de delimitar las búsquedas de manera específica (figura 5). En cuanto a las herramientas de análisis, la quinta funcionalidad, a la fecha se han implementado las herramientas de intersección, *buffer* y *disjoint* (figura 6).

FIGURA 5. Consulta estadística, media de edad de los informantes por departamento

FIGURA 6. Consulta intersección municipios donde se dice *chiras* al demonio

Por último, contamos con la funcionalidad de administrador que solo aparece cuando el usuario cuenta con los permisos necesarios y permite la gestión de la base de datos espacial del ALEC, es decir, agregar, editar o eliminar mapas, encuestadores, informantes, localidades y campos semánticos (figura 7).

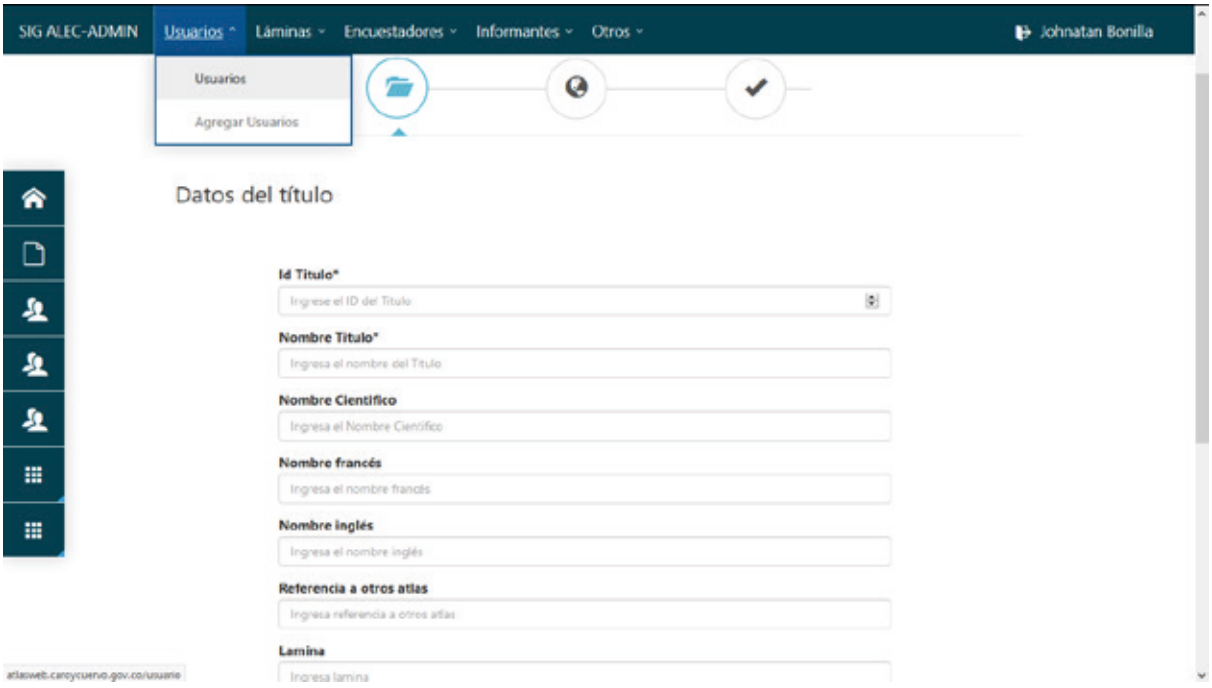


FIGURA 7. Visualización del administrador para ingreso de nuevo mapa

Por su parte, el Atlas Web denominado ALEC Digital (figura 8), se presenta como una herramienta más sencilla completamente abierta al público. Entre sus funcionalidades el ALEC Digital permite una consulta especializada desde Tomo, Campo Semántico y Mapa y búsquedas sencillas por mapa y palabra en lo que refiere a la información lingüística y etnográfica. Asimismo, es posible acceder mediante la consulta del mapa a la información adicional ya sea imagen, audio o texto. La información concerniente a encuestadores, localidades e informantes se consulta por separado, cada una con un enlace desde el panel de menú lateral y, por último, es posible acceder a los fondos no publicados del ALEC, es decir, más de 16000 fotografías desde Registro Fotográfico y a una muestra del Corpus Oral con búsqueda sencilla por localidad desde Registro Sonoro.

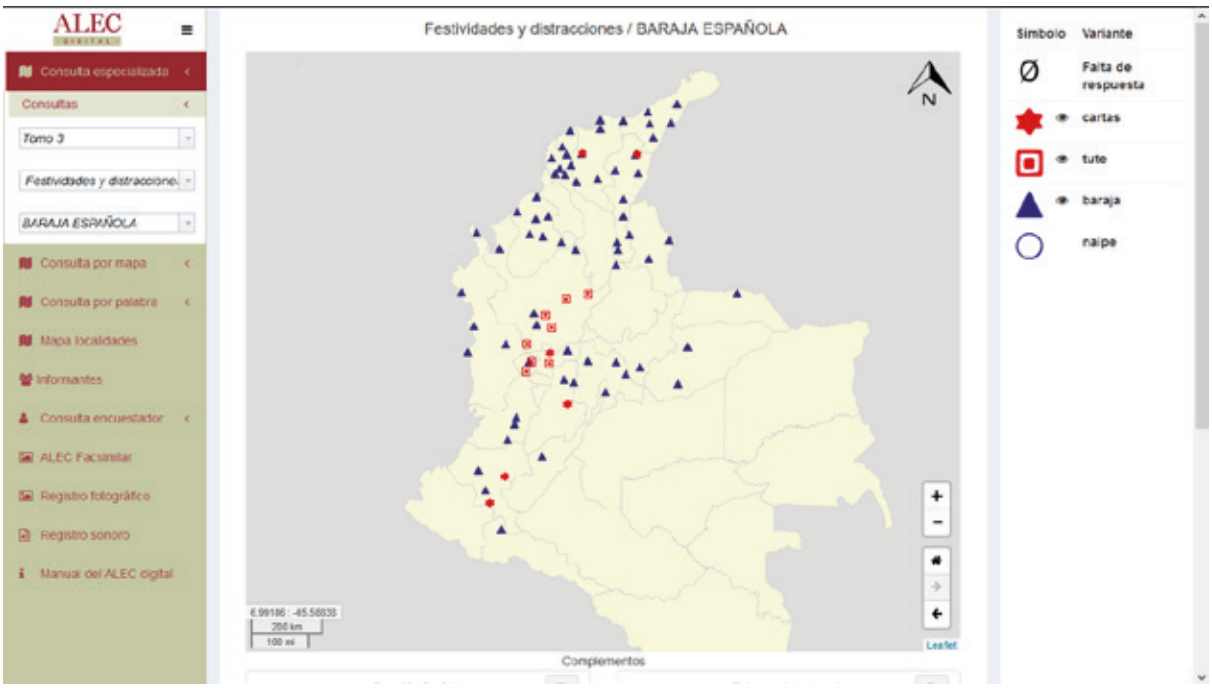


FIGURA 8. Interfaz de Atlas Web del ALEC. Consulta del mapa Baraja Española

La última fase del proyecto que se está desarrollando actualmente consta del mantenimiento y mejoramiento de los desarrollos de software. Además, se realizan pruebas de usabilidad con usuarios externos para optimización de los sistemas. En el mismo orden se está realizando el ingreso de datos de los demás tomos del ALEC (I, II, IV, V, VI) en la base de datos espacial ya que todas las pruebas se realizaron sobre la información del tomo III.

4. EL CORPUS ORAL

El corpus oral del ALEC está compuesto por 765 sesiones de audio. Cada sesión puede tener una duración de entre 2 a 60 minutos. Los investigadores se valían de magnetófonos para grabar además de romances "canciones de cuna, villancicos y otras canciones tradicionales, voces de llamada y ahuyentadores de animales, aspectos concretos del léxico (como la descripción del cultivo del café), conversaciones, cuentos cortos, historietas y otros breves relatos locales" (Buesa y Flórez 1954: 166). De esta

manera, el corpus cuenta con audios que contienen encuestas fonéticas, canciones, relatos, muestras folclóricas y entrevistas relacionadas con los campos semánticos del ALEC. Los informantes son principalmente hombres y mujeres de 36 a 55 años de edad con formación educativa en los primeros años de primaria o sin escolarización. Sin embargo, se pueden encontrar hablantes de los 5 a los 100 años de edad y con distintos niveles educativos que conforman un total de 1176 informantes.

La construcción del corpus oral del ALEC inició en 2013 junto con el corpus oral del *Español Hablado en Bogotá* (EHB) y el corpus oral del *Habla Culta de Bogotá* (HCB). Para su constitución se han llevado a cabo varios procedimientos que describiremos a continuación:

1. El primer paso (2013) fue la digitalización de los audios que se encontraban en cintas de carrete abierto. Este proceso fue realizado por la Fundación Patrimonio Fílmico Colombiano quienes entregaron los audios del ALEC en formato WAV en un disco duro y en CDs con la organización que mantenían antes de su digitalización. Es decir, almacenados por cajas y marcados de acuerdo con el número de la caja y la localidad en la que fue tomada la grabación. Asimismo, se realizaron copias de seguridad para garantizar la conservación y facilitar el trabajo con los audios.
2. El segundo paso (2014) fue la definición de metadatos generales sobre las muestras y los informantes. Los metadatos se subdividieron en cuatro áreas centrales: datos de la muestra (nombre del archivo o id, ubicación, calidad del audio y comentarios sobre la calidad); datos del informante (nombres, edad, fecha y lugar de nacimiento, nivel educativo, etc.); datos de la sesión (temas, descripción, fecha de la muestra, encuestados, entre otros.): y datos de los fragmentos para el corte de los audios (tiempo de inicio, ubicación, tiempo de finalización, primera parte, etc.).
3. El tercer paso (2015) consistió en escuchar los audios, completar la tabla de metadatos ya definida y verificar que la información y el marcado coincidiera con el contenido del audio. Vale decir, que una sesión podía estar ubicado en varias cintas y no coincidir con la localidad con la que estaba marcada. Por esto, actualmente se está realizando un proceso en el que se cortan y pegan los fragmentos de audios que conforman una sola sesión con base en la tabla de metadatos completada por los investigadores. Este proceso facilitará la consulta y la transcripción de las grabaciones.
4. El cuarto paso (2016) fue la revisión de las tablas de metadatos.
5. El quinto paso (2017) consistió en la transcripción ortográfica y la alineación de muestras de audios. Para iniciar este proceso se realizaron 20 transcripciones ortográficas de sesiones cortas alineadas con el programa ELAN[3]. Para la realización de la transcripción se estableció un protocolo como herramienta para normalizar el proceso de transcripción ortográfica de los corpus orales de la institución. Este documento funciona como manual de orientación para transcribir nuevos corpus o normalizar los ya existentes bajo una serie de pautas concretas y homogéneas que facilitan el ingreso, consulta y presentación uniforme de la información en la plataforma CLICC (Bejarano *et al.* 2018).

[3] https://tla.mpi.nl/tools/tla-tools/elan/

Paralelo a los procesos mencionados y por la necesidad de contar con una plataforma para la consulta de los corpus, la línea de investigación de lingüística de corpus y computacional inició en 2017 el proyecto "Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC)"[4]. El proyecto ha ido desarrollando un Sistema Gestor de Contenidos (SGC) para el almacenamiento, organización, consulta y explotación de los corpus del Instituto. La plataforma está compuesta por la base de datos, la interfaz administrativa y la interfaz de usuario final. Los primeros corpus que están siendo ingresados son los corpus orales ALEC, EHB y HCB. La plataforma permite la consulta de los audios y las transcripciones del ALEC a partir de los metadatos definidos para su consulta. Actualmente, CLICC cuenta con dos tipos de búsqueda: la búsqueda sencilla que se puede hacer por palabra, si el audio tiene transcripción, o por 4 metadatos principales (edad, género, fecha y lugar de la encuesta) (Ver, figura 4); y la búsqueda avanzada que es para usuarios registrados permite hacer búsquedas por diversos metadatos (temas, localidad, encuestador, edad y género del informante, etc.) y por palabra (Ver, figuras 5 y 6).

FIGURA 9. Visualización de resultados de búsqueda rápida por palabra

FIGURA 10. Visualización de consulta avanzada para usuarios registrados

[4] http://corpus.caroycuervo.gov.co

La búsqueda avanzada permite realizar mayor cantidad de tareas: posibilita la consulta con varios metadatos y por palabra; permite la visualización de la alineación del audio con la transcripción; autoriza la descarga de los audios y las transcripciones en varios formatos; y permite la realización de análisis morfosintáctico automático con el programa *TreeTagger*[5] cuando se cuenta con la transcripción de los audios.

FIGURA 11. Visualización de reproducción de un audio del ALEC alineado con su respectiva transcripción

5. PROYECCIÓN Y CONCLUSIONES

La proyección del Corpus Geolingüístico del ALEC a cinco años (2018-2023) está pensada en tres áreas principales que son afines a los objetivos del ICC y que buscan promover investigaciones sobre el español de Colombia y el reconocimiento del patrimonio lingüístico y cultural de la nación: lingüística de corpus y computacional, dialectología y nuevos métodos de análisis estadístico y geoespacial, y actualización de los datos del ALEC.

Dentro del área de la lingüística de corpus y computacional actualmente se trabaja en el etiquetado morfosintáctico de las variantes y textos del ALEC con el fin de implementar en el SIG ALEC y Atlas Web la posibilidad de realizar consultas mediante expresiones regulares. En cuanto al Corpus Oral se continuará con la transcripción de los audios para su posterior etiquetado de manera automática gracias a la implementación de la herramienta TreeTagger. Adicionalmente, se están vinculando los datos del atlas con los metadatos del corpus oral para relacionar la base de datos de las dos plataformas. Estos procedimientos facilitarán la relación de los audios con su ubicación geográfica.

Respecto a la dialectología y los nuevos métodos de análisis estadístico y geoespacial, a la fecha se han hecho análisis dialectométricos con la información de la base de datos espacial del ALEC utilizando herramientas disponibles en la web como Gabmap[6] y Diatech[7], pero se espera que a futuro este tipo de herramientas sean integradas en el SIG y se pueda brindar a la comunidad académica diversas posibilidades de análisis y retroalimentación de datos geolingüísticos en línea.

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[6] http://www.gabmap.nl/
[7] http://eudia.ehu.es/diatech/index

Y en lo referente a la actualización de los datos del ALEC, debido a que fueron recogidos entre 1956 y 1978 se han dado discusiones acerca de la vigencia de los materiales y sobre la falta de representación de algunas regiones que no fueron visitadas en la época (Guaviare, Guainía, Vichada, Vaupés y vastas zonas del Amazonas) donde en la actualidad se encuentran hablantes de español. Frente a esto el ICC, en cabeza de la línea de Lingüística de Corpus y Computacional, está llevando a cabo los primeros acercamientos teóricos y metodológicos para fundamentar el proyecto de actualización de los materiales del ALEC bajo el proyecto Nuevo ALEC por regiones que inició en el 2019. Por ejemplo, buscando implementar nuevas técnicas de recolección de la información con base en Tics, actualmente se realizan estudios sobre dialectos de Colombia con base en información de un corpus de Twitter que puede ser base para la creación de mapas de tipo léxico y morfosintáctico (Rodriguez- Díaz et al. 2018)

Agregado a lo anterior, es importante mencionar que los datos, las muestras y las plataformas pueden servir como herramienta pedagógica para trabajar sobre el español de Colombia y su diversidad. Desde el 2018 se han realizado varios talleres en universidades y bibliotecas para el manejo de estas herramientas y explorar sus posibilidades pedagógicas. También, se visualizan las posibilidades para la creación de nuevas aplicaciones multimedia de educación sobre el español de Colombia o para la enseñanza aprendizaje de español como lengua extranjera.

Adicionalmente, se deben mencionar las facilidades que proporcionan los datos para la creación y apoyo a la construcción de diccionarios. Al respecto, cabe decir que uno de los estudiantes de la maestría en lingüística del ICC desarrolló su proyecto de grado sobre la creación de la plantilla para la construcción de un diccionario del ALEC. Dicho proyecto espera consolidarse desde el 2019 como insumo para LEXICC, una plataforma para la administración de diccionarios propia.

Para finalizar, resaltamos la importancia que tienen los materiales del ALEC para los estudios del español de Colombia y como patrimonio lingüístico y cultural de la nación. El cambio de formatos (análogo a digital), la sistematización de los materiales, y el diseño y construcción de las plataformas ha sido una tarea compleja de varios años de desarrollo, que estamos seguros servirá para fomentar la investigación lingüística y para el aprovechamiento de los materiales del ALEC en varias áreas del conocimiento. Consideramos que las herramientas tecnológicas permiten el desarrollo de investigaciones desde nuevas perspectivas y facilitan el análisis automático de grandes cantidades de datos. Asimismo, la capacidad de almacenamiento, las herramientas de búsqueda y las interfaces de usuario facilitan la visualización y búsqueda de datos y materiales de una manera ágil y amigable para el usuario, lo que fomenta su uso y explotación. Varios materiales del ALEC que con anterioridad estaban ocultos y olvidados están ahora disponibles para diferentes tipos de usuarios y desde cualquier lugar con conexión a internet. Para cerrar, resaltamos la importancia del cambio de formatos y la creación de copias de seguridad para garantizar la conservación de los materiales.

REFERENCIAS BIBLIOGRÁFICAS

BEJARANO, D., A. LLANOS, R. RUBIO, J. & J. E. BONILLA (2018): *Protocolo de transcripción ortográfica CLICC*. Bogotá: Instituto Caro y Cuervo.

BONILLA, J. E., D. BEJARANO, J. BERNAL CHÁVEZ, R. RUBIO & A. LLANOS (2017): "Procesamiento informático de los materiales del Atlas Lingüístico-Etnográfico de

Colombia: Sistema de Información Geográfica", in *Estudios Lingüísticos*: La Habana: Instituto de Literatura y Lingüística.

BUESA OLIVER, T. & L. FLÓREZ (1954): "El Atlas Lingüístico-Etnográfico de Colombia (ALEC) Cuestionario Preliminar", *THESAURUS Boletín del Instituto Caro y Cuervo, X (1,2,3)*, pp. 147-315.

FLÓREZ, L. (1983): *Manual del Atlas Lingüístico-Etnográfico de Colombia*. Bogotá: Instituto Caro y Cuervo.

FERNÁNDEZ, X. S. (2010): "Entre el atlas lingüístico y el diccionario. Un diccionario de léxico tradicional a partir de los materiales del ALPI", in I. Ahumada (ed.): *Metalexicografía variacional: diccionarios de regionalismos y diccionarios de especialidad*. Málaga: Universidad de Málaga, pp. 237-256.

GARCÍA MOUTON, P. (2015): "Lengua y espacio. Revisión metodológica" in E. Hernández & P. M. Butragueño (eds.): *Variación y diversidad lingüística: Hacía una teoría convergente (Vol. 61)*. México: El Colegio de México AC, pp. 99-115.

RODRIGUES-DIAZ, C. A., S. JIMÉNEZ, G. DUEÑAS, J. E. BONILLA & A. GELBUKH (2018): "Dialectones: Finding statistically significant dialectal boundaries using twitter data", *Computación y Sistemas, 22(4),* pp. 1213-122.

SANTALLA, Mª P. (2005). "El proceso de elaboración de corpus lingüísticos", in M. Cal Varela & I. M. Palacios Martínez (eds.): *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*. Santiago de Compostela: Universidad de Santiago de Compostela, pp.46-62.

LA VARIACIÓN SINTÁCTICA DEL ESPAÑOL A LA LUZ DEL *CORPUS DEL ESPAÑOL DEL SIGLO XXI* (CORPES XXI)*

Syntactic Variation of Spanish in the light of the Corpus del español del siglo XXI *(CORPES XXI)*

CRISTINA BUENAFUENTES DE LA MATA
Universitat Autònoma de Barcelona

CARLOS SÁNCHEZ LANCIS
Universitat Autònoma de Barcelona

Resumen
Esta investigación muestra las posibilidades que ofrece el *Corpus del español del siglo XXI* (CORPES XXI) de la Real Academia Española para la descripción de la variación sintáctica del español. Para ello, se analizan dos fenómenos presentes en determinadas zonas del español americano: la combinación de comparativos sintéticos y analíticos (*Lo más mejor de todo*) y la discordancia de género y número en el pronombre acusativo (*Aquello se las dije bien claro a tus hermana*s). Las informaciones que pueden extraerse del CORPES en relación a ambas construcciones contribuyen a su mejor caracterización y permiten corroborar o complementar los datos proporcionados por otras fuentes de referencia, como, por ejemplo, la NGLE (2009) o los atlas lingüísticos

Palabras clave: lingüística de corpus, variación sintáctica, comparativo sintético y analítico, discordancia de género y número, lengua española

Abstract
This research shows the possibilities which the *Corpus del español del siglo XXI* (CORPES XXI) of the Royal Academy of Spanish offers to account for the syntactic variation in the Spanish language. For this purpose, we will analyse two phenomena existing in certain areas of Latin American Spanish: combination of synthetic and analytic comparatives (*Lo más mejor de todo*) and gender and number disagreement in accusative pronouns (*Aquello se las dije bien claro a tus hermanas*). The information available in CORPES regarding both constructions not only helps performing a better characterization of these phenomena, but also allows corroborating or complementing the data provided by other notable sources, such as the NGLE (2009) and the linguistic atlases.

* El presente trabajo ha sido parcialmente financiado con una ayuda del MICINN y FEDER (F FFI2017-87140-C4-1-P y PGC2018-094768-B-100) y de la CIRIT del Comissionat per Universitats i Recerca de la Generalitat de Catalunya (2017 SGR 634 y 2017 SGR 1251).

Keywords: corpus linguistics, syntactic variation, synthetic and analytic comparative, gender and number disagreement, Spanish language

1. INTRODUCCIÓN

Con la aparición de la *Nueva gramática de la lengua española* (NGLE) (*cfr.* RAE - ASALE 2009), contamos en estos momentos con una nueva obra que recoge una gran cantidad de información dialectal, la cual hasta ahora se encontraba dispersa en estudios de diferentes tipos. En esta gramática se atiende de forma especial a la variación diatópica, ya que son continuas las referencias a los usos dialectales de determinadas construcciones en los dialectos del español, sobre todo en referencia a las variantes americanas, lo que la convierte en una valiosa fuente de información para el análisis de la variación sintáctica en esta lengua (*vid.* Buenafuentes & Sánchez Lancis 2012 y Buenafuentes 2015). Al mismo tiempo, desde el año 2007, la Academia ha venido desarrollando un nuevo corpus, el *Corpus del español del siglo XXI* (CORPES XXI), en el que el 70% de las formas pertenecen al español de América. Ambos proyectos reflejan un cambio muy significativo en una concepción de la lengua española no atomista, sino como un todo panhispánico. Asimismo, aparte de poder obtener datos sobre variación diatópica, el CORPES XXI refleja claramente los avances experimentados en el ámbito de la lingüística de corpus en los últimos años[1], que van más allá de su aprovechamiento desde el punto de vista estrictamente léxico y lo convierten en una herramienta viable también para la extracción y tratamiento de datos de índole morfológica y sintáctica.

En la actualidad, el CORPES XXI (versión beta 0.91) está constituido por un total de 286 millones de formas (de los 300 millones de formas previstas) e incluye textos escritos y orales comprendidos entre 2001 y 2016, procedentes de España, América, Filipinas y Guinea Ecuatorial[2]. Su aprovechamiento desde la perspectiva de la variación sintáctica es posible gracias a que se trata de un banco de datos textual cuyas formas están lematizadas y categorizadas. Además, presenta dos posibilidades de consulta de gran interés desde el punto de vista investigador: en primer lugar, la búsqueda por proximidad, que permite analizar el corpus a partir de la coaparición de formas (según su categoría gramatical o especificando la forma exacta o el lema); y, en segundo lugar, la consulta por subcorpus, que posibilita acotar el análisis según criterios diatópicos, cronológicos o de tipología textual y temática. Por este motivo, si se conocen bien las posibilidades del corpus y se combinan entre sí, el CORPES XXI se convierte en una herramienta que puede arrojar datos muy interesantes sobre determinados fenómenos de variación sintáctica (*vid.* Buenafuentes & Sánchez Lancis en prensa).

Así pues, el presente trabajo quiere evidenciar la utilidad de los corpus, en general, como herramienta para el análisis de la variación sintáctica y, en concreto, poner de manifiesto el aprovechamiento que se puede obtener de un corpus como el CORPES XXI desde esta perspectiva. Para llevar a cabo este objetivo, se analizan dos fenómenos de variación sintáctica presentes en determinadas zonas del español americano: en primer lugar, la combinación de comparativos sintéticos y analíticos (*Esto es más mejor que aquello*) y, en segundo lugar, la discordancia de género y número en el pronombre acusativo (*Aquello se las dije bien claro a tus hermanas*). A partir de las posibilidades

[1] De hecho, estos avances se aprecian claramente si se contrasta el CORPES XXI con su precedente, el CREA, en su versión originaria. Véase Rojo (2016) para el contraste entre los dos corpus académicos.

[2] Puede hallarse más información sobre este corpus en http://www.rae.es/sites/default/files/CORPES_Sistema_de_codificacion_12_2015.pdf y en http://www.rae.es/sites/default/files/CORPES.pdf.

que permite el diseño del CORPES XXI, se quiere demostrar su idoneidad en la descripción de la variación sintáctica del español, ya que los datos obtenidos contribuyen a corroborar y, sobre todo, a complementar las informaciones proporcionadas por otras obras que atienden a la variación sintáctica, como la NGLE (2009) o los atlas lingüísticos[3].

2. COMBINACIÓN DEL COMPARATIVO SINTÉTICO Y ANALÍTICO

El grado comparativo actual refleja la confluencia de los dos procedimientos que existían en latín para expresar dicho grado. El primero de ellos era de carácter analítico y consistía en el uso de las siguientes perífrasis en función del tipo de comparativo: *minus* + adjetivo + *quam*, para el de inferioridad; *tam* + adjetivo + *quam*, para el de igualdad; y *magis / plus*[4] + adjetivo + *quam,* para el de superioridad. Además, el adjetivo se graduaba también a partir de la adjunción de las desinencias *-ior* (masculino/femenino), *-ius* (neutro) (*fortior*, *fortius*). Este sistema contaba, además, con los adjetivos comparativos sintéticos *melior*, *peior*, *minor* y *maior*, que se constituían como formas irregulares en este paradigma al producirse en ellos el cambio de la base derivativa de los adjetivos *bonus*, *malus*, *parvus* y *magnus*, respectivamente, de ahí que hayan recibido la denominación de comparativos supletivos.

Las estructuras analíticas solo eran obligatorias en un principio para aquellos adjetivos con un hiato en posición final, del tipo *arduus*, *idoneus*, etc. (Penny 2014 [1993]: 129). Sin embargo, las tres perífrasis comparativas se extendieron de tal modo que desencadenaron la desaparición del sistema sintético, por lo que se convirtieron en el único procedimiento en español para dar cuenta del grado comparativo. No obstante, las formas irregulares del comparativo sintético de superioridad, gracias a su alta frecuencia, resistieron a esta evolución, de modo que *mejor*, *peor*, *menor* y *mayor* (procedentes de *meliorem, peiorem*, *minorem* y *maiorem,* respectivamente) se han mantenido en español como restos de este comparativo latino.

La evolución descrita da como resultado en español un paradigma formado por un procedimiento general para crear el grado comparativo de tipo analítico (*más bueno que*) y cuatro formas irregulares y residuales[5] que incorporan el procedimiento sintético en la comparación de superioridad (*mejor que*). Sin embargo, como señala Martinell (1989: 1255), «Todos somos conscientes de que el uso general y frecuente de una forma provoca la pérdida de su fuerza expresiva […]. Por esa causa los medios de gradación se duplican». En esta línea, tal y como indica la NGLE (2009: §13.3d), «en la lengua rural de muchos países hispanohablantes se documentan comparativos léxicos combinados con los

[3] Sin embargo, Massanell (2017: 274) señala que «Lo primero que debemos tener presente es que los atlas lingüísticos tradicionales no están pensados específicamente para recoger datos de interés sintáctico. La geolingüística nació ligada a la gramática histórica, que en sus orígenes se preocupaba esencialmente por el componente fonético-fonológico de la lengua». En los últimos años, no obstante, se han desarrollado atlas lingüísticos con ricas informaciones sintácticas, como el COSER (http://www.corpusrural.es), e incluso otros que recogen exclusivamente fenómenos sintácticos, como el ASinEs (http://asines.org).

[4] La forma *plus* se empezó a utilizar también para el comparativo de superioridad en época imperial (*vid.* Urrutia Cárdenas & Álvarez Álvarez 1983: 103). Lapesa (1981: § 21.3) apunta la preferencia de los romances peninsulares y el rumano por la forma *magis* frente a los romances centrales, como el italiano y el francés, que optaron por la partícula comparativa *plus*, pese a la existencia de algunos testimonios de este último elemento en las *Glosas Emilianenses* y en Berceo.

[5] Las formas sintéticas irregulares del comparativo de superioridad conviven en español actual con los adjetivos etimológicamente comparativos *anterior, posterior, superior, interior, exterior* y *ulterior*. Estos adjetivos entraron como préstamos cultos del latín en época medieval o postmedieval (Penny 2014 [1993]: 131) y se emplean, en la actualidad, como adjetivos en grado positivo, si bien «una consecuencia de su significado comparativo es el que no puedan ser precedidos por adverbios comparativos» (Sánchez López 2006: 40) y, en este sentido, se aproximarían a los comparativos sintéticos *mejor, peor*, etc. Para las diferencias entre estos adjetivos y los comparativos sintéticos m*a*yor, *menor, peor* y *mejor*, véanse Urrutia Cárdenas & Álvarez Álvarez (1983: 103) y Sánchez López (2006: 39-40).

sintácticos» como, por ejemplo, *Esto es más mejor que aquello*. Aparentemente esta construcción parece partir de un reanálisis de la forma sintética como adjetivo en grado positivo, que, por tanto, si se quiere graduar, se integra en una estructura analítica. Como se desprende de la cita de la NGLE (2009), la información proporcionada en relación a la extensión geográfica de este fenómeno no es muy precisa. En cambio, Kany (1969 [1963]: 71) no solo señala su empleo también en España, sino que refiere ejemplos en textos americanos de Chile, Argentina, Paraguay, Ecuador, Colombia, Venezuela y México. Este autor, al igual que la NGLE (2009), relega el uso de esta construcción «a los iletrados en el uso popular y rústico general, tanto en España como en América». Es por este motivo que en la mayoría de las gramáticas se censura su uso (incluida la misma NGLE 2009). Pese a ello, este fenómeno se documenta ya en latín (*vid*. Iordan y Manoliu 1972) y «ha tenido continuidad en todas las épocas del idioma» (Romero Cambrón 1998: 33)[6].

Las informaciones proporcionadas por otras fuentes pueden (y deben) ser complementadas por las que se pueden extraer de corpus como el CORPES XXI. Así, desde el punto de vista diatópico, este corpus ofrece un reflejo de la extensión de esta construcción, lo que permite al investigador comprobar, por un lado, si se trata de un fenómeno general en América o si solo tiene uso en determinados países, y, por otro, si se registra también en España. A partir de ahí, se pueden analizar las diferencias que se pueden dar en su uso en función de la zona geográfica. Desde el punto de vista diastrático-diafásico, los resultados que ofrece el corpus pueden mostrar si, a pesar de la censura normativa, este fenómeno va más allá de la lengua rural y se localiza en textos de formalidad alta.

El CORPES XXI, como ya se ha señalado, ofrece al investigador una desarrollada herramienta, ya que, gracias a su diseño, hace que sea posible extraer datos de índole morfológica y sintáctica. La búsqueda por lema permite tener en cuenta cualesquiera de las formas flexivas de una palabra, lo que trasciende la mera búsqueda de tipo léxico y amplía el rango de resultados obtenidos en una sola consulta (si el corpus no estuviera lematizado, sería necesario buscar cada una de las formas flexivas de manera independiente). En el caso que nos ocupa, lo más relevante del corpus para la obtención de datos sobre una estructura como la combinación de los comparativos sintéticos y analíticos (y, en general, para cualquier búsqueda de tipo sintáctico en este corpus) es, sin duda, que se pueda indicar la categoría gramatical del elemento. Y no solo eso: CORPES XXI va más allá y permite que todavía pueda afinarse más esta concreción por categoría gramatical, ya que dentro de cada clase de palabra se pueden precisar distintos tipos de informaciones (en el sustantivo, por ejemplo, se puede señalar si es propio o común, su género y su número; en el verbo se puede indicar el modo, el tiempo, el número o la persona). En el caso del adjetivo, aparte de poder indicar, si se considera pertinente, su género, número[7] o tipo, también se puede concretar su grado, lo que permite extraer datos sobre las formas sintéticas del grado comparativo que son objeto de nuestra investigación.

La categorización del CORPES XXI es una de las mayores ventajas para su explotación desde el punto de vista sintáctico, pero no es la única, ya que el corpus permite, mediante la búsqueda por proximidad, acotar las consultas teniendo en cuenta los elementos que se combinan con el lema o con la forma. Esta función es lo que hace posible interrogar al corpus por estructuras sintácticas como la combinación de los comparativos sintéticos y analíticos. Además, las posibilidades de extraer estructuras

[6] Véanse ejemplos en la lengua latina y en español medieval y clásico en Vigara (2010: 134-136).

[7] Esto implica que la búsqueda por categoría incorpora, por defecto, todas las posibilidades flexivas de la categoría gramatical seleccionada.

sintácticas más complejas se amplían al poder realizar consultas teniendo en cuenta varios elementos en la combinación, e indicando su posición o el intervalo o distancia que se desea que haya entre ellos.

Por tanto, para obtener datos sobre la combinación de los comparativos sintéticos y analíticos en CORPES XXI, la búsqueda debe partir de los adverbios comparativos analíticos de superioridad, inferioridad e igualdad, es decir, *más*[8], *menos* y *tan*, que deben indicarse en el campo *forma* de la página principal de consulta. A continuación, hay que emplear la búsqueda por proximidad, que nos permite señalar la categoría gramatical «adjetivo», así como el grado comparativo y, además, precisar en qué posición interesa que aparezca ese comparativo sintético (a la derecha) y a cuánta distancia del comparativo analítico (justo al lado), tal y como se ilustra en la siguiente figura:

FIGURA 1. Consulta para la combinación del comparativo analítico de superioridad con el comparativo sintético

A partir de esta búsqueda, se obtiene que las estructuras comparativas analíticas de superioridad son las que arrojan más casos de combinación con los comparativos sintéticos (217 casos en 190 documentos), seguidas de las de igualdad (106 en 91 documentos) y, a mucha distancia, de las de inferioridad (33 casos en 33 documentos).

Obviamente estos resultados esconden muchas variables que cabe tener en cuenta en la investigación del uso de estas estructuras. En este sentido, la herramienta también permite al investigador matizar el valor de los datos obtenidos, gracias a las distintas posibilidades para su clasificación. Por ejemplo, no todas las estructuras que se atestiguan en el corpus pertenecen a la combinación de comparativos sintéticos y analíticos. Si bien el corpus no permite distinguir casos que, pese a tener formas y categorías idénticas, reflejan estructuras con diferentes funciones, el CORPES XXI ofrece múltiples posibles ordenaciones de los ejemplos que, además, pueden combinarse, lo cual hace posible agrupar, en algunos casos, aquellas construcciones que no corresponden a la que interesa al investigador. En el caso que nos ocupa, esta herramienta del corpus permite discriminar la combinación de los comparativos analíticos y sintéticos de aquellos casos en los que los adverbios *más* y *menos* forman parte de otro tipo de estructuras. Para ello, realizamos la misma búsqueda que hasta el momento, y ordenamos, en primer lugar, los ejemplos por el primer lema[9] a la derecha y, en segundo lugar, indicamos que, a su vez, estos resultados se ordenen por el primer lema a la izquierda, tal y como se ilustra en la figura 2:

[8] Para evitar la confusión con la conjunción adversativa, en la búsqueda se puede seleccionar o bien que se respete la grafía original, de modo que al escribir la forma con tilde, el corpus desechará los casos sin ella y, por tanto, los que corresponden a la conjunción; o bien que, independientemente de su escritura, la forma solo corresponda a la categoría *adverbio*.

[9] Esta ordenación también puede realizarse por forma o por categoría gramatical.

FIGURA 2. Ordenación de los ejemplos

Esta ordenación permite agrupar casos como los siguientes, que deberán ser desechados, pues no son realmente reflejo de la combinación de los comparativos sintéticos y analíticos[10]:

- *al menos* + adjetivo sintético comparativo (7 casos en 7 documentos)
- *echar de menos* + adjetivo sintético comparativo (1 caso en 1 documento)
- *más o menos* + adjetivo sintético comparativo (2 casos en 2 documentos)
- *ni mucho menos* + adjetivo sintético comparativo (1 caso en 1 documento)
- *cuanto menos* + *mejor* (2 casos en 2 documentos)
- *cuanto más* + *mejor* (21 casos en 21 documentos)
- *mientras más+mejor* (1 caso en 1 documento)
- *contra más mejor* (1 caso en 1 documento)
- *cada vez más*+ adjetivo sintético comparativo (1 caso en 1 documento)

Por tanto, una vez eliminadas estas estructuras, el corpus permite atestiguar la presencia en español de la combinación del comparativo analítico con el sintético en 307 casos: 185 para el comparativo de superioridad, 18 para el de inferioridad y 104 para el de igualdad[11]. Una desventaja del corpus es que no permite seleccionar los ejemplos que interesan al investigador y desechar aquellos que no responden a sus intereses, lo cual sería un valor añadido que facilitaría más, si cabe, el análisis de los resultados[12].

Además de separar los casos que no responden a la combinación de los comparativos sintéticos y analíticos, esta ordenación aporta datos interesantes teniendo en cuenta la forma del comparativo sintético que se combina. Se puede apreciar, a partir de la agrupación de los ejemplos, que en las estructuras analíticas más frecuentes según

[10] Sin embargo, hay que discriminar otros casos de forma manual, como, por ejemplo, cuando el adjetivo comparativo sintético se ha lexicalizado en sustantivo o cuando el adjetivo comparativo sintético forma parte de un nombre propio, como en el siguiente ejemplo:

(i) No. Fíjese que yo no jugué beisbol juvenil y mucho **menos Mayor A** porque era muy delgadito, muy débil (…). (Rodríguez Centeno, Edgard: «"Después de dios y mi familia, el beisbol"». *La Prensa.com.ni*, 2007).

[11] No se indica el número de documentos, solo los casos hallados, debido a que en un mismo texto pueden aparecer ejemplos de dos o más usos de combinación de los distintos comparativos.

[12] Por ejemplo, el CORDIAM permite seleccionar los ejemplos y borrar aquellos que no interesan al investigador.

los datos anteriores, es decir, las de superioridad e igualdad, el comparativo sintético mayoritario es *mayor* (84,3% para las primeras y 82,7% para las segundas), mientras que en las comparativas analíticas de inferioridad prácticamente el único adjetivo comparativo sintético registrado es *peor*. La siguiente tabla resume estos resultados:

	mayor	*menor*	*mejor*	*peor*	TOTAL
más	156	4	16	9	185
menos	0	0	1	17	18
tan	86	14	3	1	104

TABLA 1. Datos según el tipo de comparativa analítica y el comparativo sintético

Por tanto, en cuanto a la combinación de comparativos analíticos y sintéticos se refiere, las comparativas de desigualdad, pese a ser tratadas habitualmente como un conjunto homogéneo[13] que no presenta diferencias en su comportamiento sintáctico, en la selección del comparativo sintético muestran tendencias diferentes, a tenor de los datos que refleja el corpus. Asimismo, mediante el uso de las distintas ordenaciones, se pueden extraer datos interesantes sobre los distintos elementos constitutivos de la estructura comparativa: observar la presencia de cuantificadores o expresiones diferenciales (término empleado por la NGLE 2009) (*mucho más mayor que*, *bastante menos peor que*), ya que se puede realizar la ordenación por el primer lema a la izquierda, así como analizar la elisión o aparición del complemento comparativo y, en este último caso, examinar cuál es el elemento que lo introduce (*que, de, como*, etc.), porque se pueden clasificar los ejemplos por el segundo lema a la derecha.

Hasta el momento, se ha podido observar el aprovechamiento del CORPES XXI para la obtención de informaciones sobre una estructura sintáctica como la combinación de los comparativos sintéticos y analíticos. Estas informaciones pueden someterse a la luz de la variación diatópica a partir de otro de los recursos que también ofrece este corpus. Los resultados pueden ser filtrados a partir de la búsqueda por subcorpus según la procedencia geográfica del texto en tres niveles de concreción: (i) España/América/Guinea/Filipinas, (ii) área dialectal y (iii) país.

FIGURA 3. Consulta para la combinación del comparativo analítico de superioridad con el comparativo sintético en América

Así pues, si tomamos los datos obtenidos de la búsqueda efectuada para la comparación de los comparativos sintéticos y analíticos y limitamos en el corpus según los textos sean americanos o españoles[14], se observa que la representación de esta

[13] De hecho, las comparativas de superioridad e inferioridad se engloban dentro de las comparativas de desigualdad en oposición a las comparativas de igualdad (*vid.* NGLE 2009).

[14] No se han tomando en consideración las variedades de Guinea y Filipinas.

estructura en España es mayor que en América, ya que el total de casos en textos americanos es de 98 frente a los 209 que se registran en documentos españoles.

	mayor		*menor*		*mejor*		*peor*		TOTAL	
	Esp.	Am.	Esp.	Am.	Esp.	Am.	Esp.	Am.	Esp.	Am.
más	139	17	1	3	2	14	1	8	143	42
menos	0	0	0	0	0	1	0	17	1	17
tan	62	24	3	11	0	3	0	1	65	39
TOTAL combinación comparativo analítico+sintético									209	98

TABLA 2. Datos según el tipo de comparativa analítica y el comparativo sintético en España o América

Aparte de realizar una selección de los casos que el corpus ofrece, sus informaciones, obviamente, deben ser sometidas al análisis del investigador, pero el solo hecho de poder obtener datos de índole sintáctica teniendo en cuenta la perspectiva diatópica es una gran ventaja. De este modo, los resultados reunidos en la tabla 2, si bien indicarían que en España la combinación de comparativos analíticos y sintéticos está más extendida que en América, en contra de lo indicado por la NGLE (2009), un análisis más minucioso de los resultados permite matizar esta generalización en dos sentidos. En primer lugar, en España las combinaciones *más mayor* y *tan mayor* suponen el 96,2% del total de los casos de esta estructura, mientras que en textos americanos ambas construcciones representan un 41,8%. Conviene, por tanto, acudir al análisis más concreto de los ejemplos que ofrece el corpus para comprobar si se puede hallar lo que motiva esta diferenciación entre el español peninsular y el americano. Así, si se efectúa este examen, se comprueba que lo que acrecienta la cifra de casos de combinación de los comparativos sintéticos de superioridad e igualdad junto al adjetivo *mayor* en España es que en prácticamente[15] todos ellos el contexto se refiere a la edad, rentabilidad ya puesta de manifiesto en el estudio de Vigara (2010 y 2011). De hecho, tal y como señala esta autora,

> parece evidente que ni *mayor* ni *grande* están generalizados con valor temporal en todo el ámbito español. *Mayor* sí lo está en España, aunque *grande* y su opuesto, *chico*, se conocen también y se usan sobre todo en Andalucía. *Mayor* es también (re)conocido en toda Hispanoamérica, incluso en los países en que no se usa. (Vigara 2011: 120).

El corpus refleja claramente la extensión del uso de *más mayor* referido a la edad en América, ya que en el conjunto de ejemplos de *más mayor* y *tan mayor*, menos numeroso que en la península, como se ha señalado, solo en 1 caso se emplea fuera de ese contexto (véase ejemplo 3).

En segundo lugar y derivado de lo anterior, mientras que el español de España ofrece poca variedad en el uso de las diferentes combinaciones de comparativos analíticos y sintéticos (como se ha señalado, es muy mayoritario el empleo de *más mayor* y *tan mayor*), el americano registra ejemplos de todas ellas, excepto de la combinación de *menos* con los adjetivos sintéticos *mayor* y *menor* (que tampoco se atestiguan en España).

Aparte de poder contrastar estructuras sintácticas entre el español americano y el peninsular, también es posible observar las diferencias entre las distintas áreas

[15] El único ejemplo que se ha localizado en el corpus en el que el contexto no se refiere a la edad se reproduce en (11).

dialectales americanas o entre países, ya que el corpus permite acotar por estos dos criterios. Esto permite constatar y precisar, por ejemplo, la indicación geográfica que proporciona la NGLE (2009) o actualizar los datos que ofrecía Kany (1969 [1963]) en relación a este fenómeno.

Según se aprecia en el corpus, la estructura que combina los comparativos analíticos y sintéticos se documenta prácticamente[16] en toda América, aunque de forma bastante escasa, como se ha visto anteriormente[17]. Sin embargo, esto podría indicar que se trata de un fenómeno no restringido a un área o país concretos, sino general en el español americano. Los países en los que se han detectado más casos y que reflejaban distintas combinaciones de comparativos son, en el área de México y Centroamérica, Panamá, Nicaragua, México, Guatemala y El Salvador; en la zona del Caribe Continental, Colombia; en la zona del Río de la Plata, Argentina y Uruguay; en la zona andina, Perú; en las Antillas, República Dominicana, y en Chile y EE.UU (véanse ejemplos de (1) a (4), así como los recogidos de (5) a (8)).

(1) — Luego será año nuevo, es **menos peor** que la Pascua, aunque todos los años son peores. (Giusti, Isabel Vera: *CRIS o la plenitud del vacío*, 2003, Chile).

(2) Por fortuna, al oír eso de que estoy **tan mejor**, Myriam decide llevar la contraria (…). (Araújo, Helena: *Las cuitas de Carlota*, 2003, Colombia).

(3) F. JUAN DE TALABAN: He contado en vuestro registro en lo menos… ¡cuatrocientos de naturales convertidos! ¡Eso es un ejército **más mayor** que las tropas! (Galindo Moriega, Edeberto: *La furia de los Mansos*, 2008, México).

(4) Yo hago los deberes en la mesa. La tabla del cuatro es **más mejor** que la del tres. (Rosencof, Mauricio: *Sala 8*, 2011, Uruguay).

Finalmente, el corpus también permite clasificar y filtrar los textos en función de su tipología y temática a partir de la búsqueda por subcorpus, lo que proporciona otra óptica más desde la que analizar los datos: la perspectiva textual y su posible incidencia en la descripción lingüística. En el caso de la combinación de los comparativos analíticos y sintéticos este factor es especialmente relevante, ya que la NGLE (2009) relega su empleo al habla rústica y, por este motivo, censura su uso.

El primer nivel de tipología textual que permite el CORPES XXI es indicar si el texto es oral o escrito. En el caso que nos ocupa, todos los testimonios obtenidos son escritos, lo que demuestra un dato interesante: el fenómeno, a pesar de considerarse incorrecto, ha permeado en la lengua escrita. En los textos escritos, el corpus posibilita señalar si se trata de ficción o de no ficción. Los textos de ficción, *grosso modo*, se encuentran más cercanos a la lengua literaria (novelas, obras de teatro, relatos), mientras que los textos de no ficción son de carácter divulgativo (prensa y libros de temáticas diversas, documentos de internet). Por tanto, la presencia en documentos de ficción indicaría que la combinación de comparativos analíticos y sintéticos es un fenómeno que también está presente en la lengua literaria, es decir, en la lengua culta.

Sin embargo, de nuevo, el corpus proporciona los datos, pero el investigador debe analizarlos con detalle para no errar en su interpretación. En este sentido, el CORPES

[16] Salvo en Costa Rica y Puerto Rico.

[17] La censura normativa es uno de los factores que seguramente ha incidido en la difusión y extensión de esta estructura.

XXI no clasifica los textos según el grado de formalidad de cada contexto específico[18] y, por lo tanto, debe ser el investigador quien interprete la situación discursiva. Así, cabe precisar que la mayoría de los ejemplos hallados, pese a pertenecer a textos de ficción, se caracteriza por ser fragmentos cercanos a la oralidad, como, por ejemplo, parlamentos de personajes en las novelas y en los relatos; u opiniones personales en el caso de la prensa (véanse ejemplos (5)-(8)).

(5) — Mi tío Generoso –dijo Merceditas–, que es el que **más mejor** de todos sabe escribir. (Ferrini, Ernesto: *La tristeza de los burros*. Lima: Planeta, 2006, Novela, Perú).

(6) Es unánime que *Sin tetas no hay paraíso* es el peor filme del año. **Tan peor** como el libro, la miniserie y la novela. (Morales, Nicolás: «Los sopores del 2010. Sopor y piropos». Revista Arcadia.com, 2010-12-15. Colombia).

(7) La definición general era: "Kerry no es una alternativa, pero es **menos peor** que Bush" […]. (Rascón, Marco: «El modelo Kerry de legitimación». La Jornada, 2005-05-10, México).

(8) Sea el establecimiento pecuario que fuere y sobre todo si éste cuenta con gran cantidad de animales y se basa en un sistema del tipo intensivo (**más peor**, aún). (Montiel, Eduardo F.: «Medio Ambiente: Manejo de desperdicios, qué hay que hacer… lo que hay que cuidar». *Producción Agroindustrial del Noa*, 2001-08, Argentina).

Los datos obtenidos permiten localizar casos de esta estructura y corroborar su pertenencia al habla informal, pues los fragmentos en los que aparece documentada presentan una gran cercanía a la oralidad. Además, los pocos ejemplos hallados para España de esta construcción no referida a la edad se hallan en unos contextos muy semejantes a los observados en los textos americanos, es decir, en la reproducción de citas textuales o en la expresión de la opinión personal (véanse ejemplos (9)-(11)).

(9) (…), *porque si la gente está sin trabajo, está parada, lo que vamos a tener es eso, delincuencia, y cosas* ***más peores***[19]. (Elejabeitia Tavera, Carmen: *Mujeres inmigrantes en la educación de personas adultas*, 2006, España).

(10) Presentador: Hola, amigos, ¿qué tal? Bienvenidos al concurso Llama y Gana. Soy Alfonso Séptico, y empieza aquí, ahora, en este estudio, ya, aquí, en este instante, el concurso **más mejor** de la radio. (*Especialistas secundarios. Podría ser peor*, 2010, España).

(11) (…) pero enseguida me sentí arrastrado por el irresistible caudal narrativo de uno de los **más mayores** novelistas del gran siglo de la novela. (Rodríguez Rivero, Manuel: «Dickens también es para el verano». *El País.com. Babelia*, 2011-08-13, España).

En definitiva, el CORPES XXI ha permitido no solo dar cuenta de la existencia y vigencia de esta estructura en español, a pesar de la censura normativa, sino también

[18] Otros corpus como el PRESEEA (http://preseea.linguas.net/Inicio.aspx) sí que ofrecen esta clasificación de los textos atendiendo a la situación comunicativa.

[19] La cursiva es del original, pues es la reproducción de una cita textual.

precisar su extensión geográfica y complementar su restricción diafásica. Asimismo, se han podido extraer diferentes tendencias en su uso entre el español peninsular y el americano, ya que en el primer caso la combinación se limita de forma prácticamente exclusiva a *más mayor* y *tan mayor* en contextos que refieren a la edad, mientras que en el segundo caso existe más variedad de estructuras en la combinación de los comparativos analíticos y sintéticos.

3. LA DISCORDANCIA DE NÚMERO EN EL PRONOMBRE ACUSATIVO

Un aspecto muy interesante de variación sintáctica es la discordancia de número en el pronombre acusativo en oraciones del tipo *Eso se los dije ayer a tus hermanos*, en donde este clítico personal realiza su concordancia con los rasgos morfológicos del complemento indirecto (*a tus hermanos*) y no del complemento directo (*eso*). Este fenómeno, recogido por la NGLE (2009: §35.2h), se sitúa en el marco de las oraciones con complemento directo e indirecto representados por pronombres personales clíticos de tercera persona. En este tipo de construcciones, como la forma pronominal de dativo *se* solo indica persona (ni género ni número), el pronombre acusativo, único elemento que posee flexión nominal, varía morfológicamente concordando con el referente del dativo[20], pero no con el suyo propio. En principio, esta clase de estructura solo se produce cuando el complemento indirecto mantiene una relación de referencia con un elemento plural y el complemento directo con un elemento singular, por lo que, en definitiva, no se trata más que de un simple fenómeno de hipercaracterización del número. Si nos atenemos a la caracterización tanto diatópica como diastrático-diafásica que realiza la Academia en su gramática, esta explica que se trata de una construcción «frecuente en la lengua oral y coloquial de amplias zonas de América, así como del español canario», pero seguidamente realiza una distinción, ya que aunque en ciertas áreas lingüísticas esta estructura se extiende progresivamente «a los registros cultos (México, el Caribe continental y parte de las áreas centroamericana, rioplatense y andina), en otras (Chile, España y parte de las áreas andina y antillana) no se considera propia de estos registros» (NGLE 2009: §35.2h)[21].

En este caso, el CORPES XXI se nos presenta como un corpus idóneo para poder comprobar la existencia de esta variante sintáctica en el español actual, tanto europeo como americano, así como su desarrollo tanto a nivel diatópico como diastrático. Con esta finalidad, y debido a la peculiaridad, por su alta frecuencia de aparición, de los elementos gramaticales que componen esta construcción[22], se ha realizado una primera consulta, de tipo general, a partir de la secuencia pronominal *se los* y el verbo de lengua *decir* conjugado[23], con el fin de poder hallar un primer conjunto de posibles casos, pues no todos los ejemplos resultantes pertenecerán a esta clase de estructura, ya que no es posible encontrar únicamente casos de esta variante sintáctica mediante los diferentes

[20] Algunos autores como Rivarola (1985), Mello (1992) y Fernández Soriano (1999: 1257-1258) interpretan la secuencia de los dos clíticos como una unidad morfológica, lo que explicaría la inserción del morfema -s de número al final. Por su parte, Company (1992 y 1998) considera que se trata de una construcción reciente, de principios del siglo XX, que se gramaticaliza, fruto de una subjetivización, a partir de un reanálisis del morfema *-s*, que añade a su valor de pluralidad el rasgo [+ animado, + humano] característico del dativo.

[21] Kany (1969 [1963]: 140-143) lo considera un ejemplo de «interferencia asociativa», un solecismo, muy arraigado en la lengua y muy difícil de eliminar; raro en España, pero de progresiva difusión en numerosas zonas del español de América, y aunque en algunas es de uso popular, en otras es de empleo general incluso entre la gente culta y en el estilo literario, lo cual indica su alto grado de aceptación.

[22] Una búsqueda en la que se combinaran simplemente ambos clíticos proporcionaría un resultado totalmente inmanejable por parte del investigador (17.458 casos en 9.088 documentos), por lo que es necesario establecer *a priori* una estrategia basada en la combinatoria más factible de cara a la obtención de ejemplos válidos.

[23] De no estar lematizado el CORPES XXI, el análisis de esta construcción sería una tarea muy costosa.

etiquetadores del corpus[24]. De este modo, a partir de la forma *los*, sin especificar la categoría gramatical de ninguno de estos elementos, se ha buscado la forma *se* antepuesta inmediatamente a su izquierda y el lema *decir* (o sea, todas las formas conjugadas de este verbo) inmediatamente a su derecha[25]. El resultado de la búsqueda es el siguiente:

FIGURA 4. Consulta para la combinación de los clíticos *se los* antepuestos al verbo *decir* conjugado

Se han encontrado 212 casos en 182 documentos de esta construcción[26]. Algunos ejemplos de esta estructura son los siguientes:

(12) Algunos se van gritando y pateando, otros lo hacen con una aguja –real o metafórica– en el brazo. Otros ya se han ido y simplemente no lo saben –alguien tiene que **decírselos**. Supongo que yo soy uno de los que **se los dice**. (Solórzano, Fernanda: «Ruta a la ciudad de los fantasmas». *Letras Libres*. Coyoacán: letraslibres.com, 2004-02-29, México).

(13) — Bah –dijo–, lo del dinero es una excusa de mujeres, ya **se los dijo** Ovidio: «Mayor sin duda es vuestro placer que el que les corresponde a los hombres». (Franco, Jorge: *Paraíso Travel*, 2001, Colombia).

(14) **Se los dije**, que esto podía ser muy divertido. Para qué esperar a morirse para oír lo que dicen tus amigos de ti. Mejor oírlo en vida. De todas maneras es mentira. También **se los dije**. Todo lo que se dice en un velorio es mentira: nadie tiene el valor de decir la verdad sobre el muerto. (González-

[24] Pero sí se trata de la combinación que nos permite acotar el mayor conjunto de ejemplos válidos.

[25] Se emplea el verbo *decir* en la búsqueda por tratarse de un verbo que posee dos complementos, uno directo y otro indirecto, y además por ser habitualmente el primero de tipo oracional, por lo que su pronominalización se realiza mediante un acusativo neutro.

[26] De estos 212 casos, solo 14 no se corresponden con la construcción analizada (4 de México, 4 de Colombia, 3 de España, 1 de Chile, 1 de Argentina y 1 de Honduras), lo que demuestra que se trata de una combinación con un valor muy marcado.

Llorente, José Miguel: «Con la muerte no se juega (Monólogo desde el ataúd)». *La confesión del Comandante y otras historias*, 2006, Cuba).

(15) Acabo de decir que me sentí aliviado por el hecho, la certeza, de que no se conocían entre ellos y, antes, les había dicho que me tranquilizaba la posibilidad de que llegasen a conocerse: no me contradigo, evoluciono en mis ideas. Ya lo escribí, es decir, **ya se los dije**, les dije que al llegar a una conclusión, ésta me prueba la opuesta. (Mujica, Hugo: «El testamento inconcluso». *Bajo toda la lluvia del mundo*, 2008, Argentina).

(16) Pero miren cómo **se los digo**: *El cojo y el loco*: puta obra maestra. *El cojo y el loco*: una de las mejores novelas del 2010. (Juan Mal-herido: «El cojo y el loco, de Jaime Bayly». *Lector Mal-herido Inc.* www.lectormalherido.wordpress.com, 2011-01-16, España).

(17) Estas evidencias tienen a los defensores de la teoría del peak oil preparándose para gritar a los cuatro vientos "¡**se los dije**!". («Petróleo negro oscuro». *Capital.cl*, 2008-05-30, Chile).

(18) Me preguntaron mi nombre, y cuando **se los dije** me vieron con lástima. (Sandoval, Ana María: *¿Quién mató a Molly Blum?*, 2010, Guatemala).

En todos estos ejemplos, el clítico personal *los* está en plural, a pesar de tener un antecedente singular oracional (*mayor sin duda…*; *que esto podía ser muy divertido*; *que al llegar a una conclusión…*; *El cojo y el loco: puta obra maestra*; *mi nombre*; etc.), mientras que el clítico *se* remite a un referente plural explícito o no.

Si bien se ha dicho antes que esta construcción repercute básicamente en el número del clítico de acusativo, es posible hallar también ejemplos en los que además se altera su género. La Academia considera estas variantes sintácticas más marcadas que las anteriores, por lo que recomienda explícitamente en su gramática evitar esta discordancia de género (NGLE 2009: §35.2i), contrariamente a la indicación realizada sobre la discordancia de número, en la que no se señala tal prescripción. Para encontrar casos de esta variación estructural, se ha seguido el mismo criterio de búsqueda que en el caso anterior, pero sustituyendo la forma *los* por *las*, lo que ha permitido hallar 49 casos en 46 documentos, de los cuales solo uno corresponde a 1 caso de discordancia morfológica[27]:

Figura 5. Consulta para la combinación de los clíticos *se las* antepuestos

[27] Al ser esta construcción más propia de la lengua oral, es mucho más difícil encontrar ejemplos en el CORPES XXI debido a que este corpus posee todavía un porcentaje muy bajo de textos orales.

al verbo *decir* conjugado.

(19) Pero yo creo que Teletón ha sido una gran solución para toda esta problemática, y desde aquí les invito. Y una modalidad que tiene Teletón en este año, y **se las digo**, a partir de una semana antes de lo que será el Teletón del 12 y 13, se podrá hacer la recaudación de el de su aportación. (*Primera emisión: grabación en directo, 11/11/03, Imagen* Informativa, 2003, México).

En este caso, el clítico de acusativo concuerda en número y también en género con el referente del clítico de dativo. Así, en el ejemplo (19), el clítico *las* parece que hace referencia al público femenino asistente al programa de televisión, pues el clítico de acusativo tiene como referente un elemento oracional (*a partir de una semana antes* (…) *se podrá hacer la recaudación de* (…)). Por otra parte, este ejemplo pertenece a una zona y país (México) en donde la discordancia de número está mucho más extendida, pues ha llegado incluso al registro culto, como señala la NGLE (2009), si bien cabe advertir que este ejemplo reproduce la lengua oral, ya que se trata de la grabación directa de un programa. El hecho de que los casos de discordancia de género sean muchísimo menos frecuentes en el corpus que los de discordancia de número demuestra la poca vitalidad de esa estructura, al mismo tiempo que nos permite establecer una periodización de ambas: de este modo, la discordancia de número es necesariamente la más antigua, la más frecuente, la más extendida diastráticamente y, por consiguiente, la menos marcada, frente a la discordancia de género, que se caracterizaría por ser la más reciente, la menos extendida, tanto diatópica como diastráticamente, y, en consecuencia, la más marcada.

Por otro lado, es muy significativa la mayor o menor frecuencia de combinación de estos elementos (*se* + *los* + el verbo *decir*) en las diferentes zonas lingüísticas del corpus, distinción que se puede visualizar gracias a las estadísticas que ofrece el CORPES XXI. Estas estadísticas no solo proporcionan los datos absolutos de la consulta efectuada, sino la frecuencia normalizada, es decir, el número de casos por millón de palabras, lo que permite obtener una visión más real de la representatividad en el corpus del fenómeno objeto de estudio. Según el CORPES XXI, la frecuencia normalizada de esta construcción es de 0,75 casos por millón. Resulta muy significativo constatar que las zonas y los países (México, el Caribe continental y parte de las áreas centroamericana, rioplatense y andina) en los que la discordancia está más extendida en los distintos registros lingüísticos, como señala la NGLE (2009), son los que presentan una mayor frecuencia normalizada, como se puede ver en la siguiente figura:

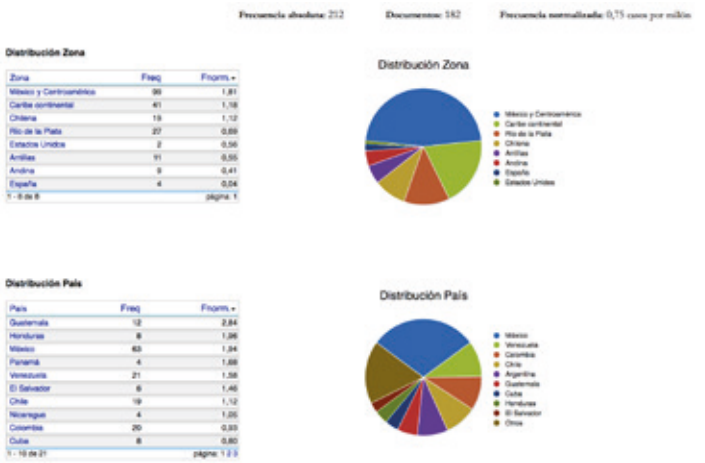


FIGURA 6. Estadística de la frecuencia absoluta y la frecuencia normalizada por zonas y países de la combinación de los clíticos *se los* antepuestos al verbo *decir* conjugado

De los datos anteriores del corpus, se constata la poca vitalidad de esta construcción en español actual (apenas 212 casos en 182 documentos, de los cuales 198 (93,4%) se corresponden con casos de discordancia de número del clítico acusativo)[28]. Además, se puede comprobar su escasa presencia en el español europeo (apenas 0,04 casos por millón), frente a su alta frecuencia en distintas zonas del español de América (hasta 1,81 casos por millón). Y dentro de estas últimas, destacan, por un lado, las zonas de México y Centroamérica (1,81fn), Caribe continental (1,18fn), zona chilena (1,12fn) y Río de la Plata (0,69fn), como era esperable, por lo que las zonas en las que esta construcción ha pasado a la lengua culta presentan una mayor frecuencia normalizada; y por países, Guatemala (2,84fn), Honduras (1,96fn), México (1,94fn), Panamá (1,68fn), Venezuela (1,58fn), El Salvador (1,46fn), Chile (1,12fn), Nicaragua (1,05fn), Colombia (0,93fn), Cuba (0,80) y Argentina (0,71fn), también como era esperable por las mismas razones. Además, la estadística también ofrece datos sobre la tipología textual de los documentos que constituyen el corpus y en los que se atestigua el fenómeno buscado. De este modo, en el caso de esta construcción, se confirma que la tipología textual es un factor que favorece una mayor presencia de discordancia morfológica en una determinada zona o país, ya que la ficción presenta 1,98fn y la entrevista 0,85fn; y por temática, el teatro, claro ejemplo de oralidad, presenta 4,62fn, seguido del guion (3,13fn), el relato (2,08fn) y la novela (1,51fn).

4. CONCLUSIÓN

El presente estudio demuestra la validez de la explotación de un corpus para realizar un estudio de variación sintáctica del español actual. A partir de su análisis, se ha podido comprobar y completar una pequeña parte de la gran cantidad de información diatópica y diastrática que aparece en la NGLE (2009); y también se ha demostrado las grandes posibilidades del CORPES XXI como una herramienta de gran utilidad para la investigación de esta variación. Así, aparte de la obtención de una gran cantidad y variedad de ejemplos de los diferentes casos de variación sintáctica del español actual, gracias a la lematización y la categorización gramatical de las palabras que componen el corpus, este permite sobre todo precisar con mayor fiabilidad las diferentes zonas lingüísticas, así como los distintos países, en los que aparecen estas variantes sintácticas. El análisis en su conjunto de dichas variaciones hace posible además evaluar su presencia real en cada dialecto a partir de la obtención de su frecuencia normalizada. Y este mismo dato nos permite también estudiar su mayor o menor presencia en las distintas zonas y países en función de la tipología textual o la temática del texto, ya sea tanto escrito como oral.

Por consiguiente, el CORPES XXI se presenta como una herramienta muy útil, ya que su diseño permite la explotación, desde el punto de vista de la variación sintáctica. Su empleo resulta imprescindible para complementar, corroborar o precisar las informaciones sobre variación sintáctica que aparecen dispersas en otro tipo de obras, como las gramáticas o los atlas lingüísticos.

[28] A pesar de que el CORPES XXI no nos permite separar los ejemplos que presentan discordancia de los que no la poseen, los datos obtenidos conjuntamente de ambas estructuras, siendo conscientes del peso de cada una de ellas, como así se indica, pueden ser tratados estadísticamente con total validez para poder constatar las zonas que poseen una mayor discordancia a partir de su tendencia a una mayor frecuencia de empleo de esta combinación sintáctica en general.

RECURSOS ELECTRÓNICOS MENCIONADOS

ASINES: Á. J. Gallego (dir.) (2014-2020): *Atlas Sintáctico del Español*. Barcelona: Universitat Autònoma de Barcelona. http://www.asines.org. ISSN: 2462-3431.

CORPES XXI: Real Academia Española: *Corpus del español del siglo XXI*. http://www.rae.es.

COSER: I. FERNÁNDEZ-ORDÓÑEZ (dir.) (2005-): *Corpus Oral y Sonoro del Español Rural*. www.corpusrural.es. ISBN 978-84-616-4937-2.

CREA: Real Academia Española: *Corpus de referencia del español actual*. http://www.rae.es.

PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* Alcalá de Henares: Universidad de Alcalá. http://preseea.linguas.net

REFERENCIAS BIBLIOGRÁFICAS

BUENAFUENTES, C. (2015): "Variación morfosintáctica en los dialectos del español: un análisis desde la diacronía", *Zeitschrift für romanische Philologie* 131 (2), pp. 383-407. http://doi.org/10.1515/zrp-2015-0024

BUENAFUENTES, C. & C. SÁNCHEZ LANCIS (2012): "Procesos de gramaticalización y lexicalización a la luz de los corpus académicos", in T. Jiménez *et alii* (coords.): *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Universidad de Santiago de Compostela: Servizio Publicacións, pp. 153-165.

BUENAFUENTES, C. & C. SÁNCHEZ LANCIS (en prensa): "The *Spanish Twenty-first Century Corpus* (CORPES XXI): A Tool for the Study of Syntactical Variation in Spanish", in A. Cerrudo, Á. J. Gallego & F. Roca (eds.): *Syntactic geolectal variation: Traditional approaches, current challenges and new tools*. Amsterdam: John Benjamins.

COMPANY COMPANY, C. (1992): "Un cambio en proceso: «el libro ¿quién *se los* prestó?», in E. Traill (ed.): *Scripta philologica in honorem Juan M. Lope Blanch*. México: UNAM, pp. 349-363.

COMPANY COMPANY, C. (1998): "The interplay between form and meaning in language change. Grammaticalization of cannibalistic datives in Spanish", *Studies in Language* 22 (3), pp. 529-565. https://doi.org/10.1075/sl.22.3.02com

FERNÁNDEZ SORIANO, O. (1999): "El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos", in I. Bosque & V. Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, vol. 1, pp. 1209-1273.

IORDAN, I. & M. MANOLIU (1972): *Manual de lingüística románica*. Madrid: Gredos.

KANY, Ch. E. (1969 [1963]): *Sintaxis hispanoamericana*. Madrid: Gredos.

LAPESA, R. (1981): *Historia de la lengua española*. 9ª ed. corregida y aumentada. Madrid: Gredos.

MARTINELL, E. (1992): "Estilística en la gradación de los adjetivos", in A. Vilanova (ed.): *Actas del X Congreso de la Asociación Internacional de Hispanistas*. Barcelona: PPU, pp. 1253-1262.

MASSANELL I MESSALLES, M. (2017): "Los atlas lingüísticos como fuente para estudios de sintaxis dialectal: el caso del *Altes Lingüístic del Domini Català (ALDC)*", *Verba. Anuario Galego de Filoloxía* 44, pp. 271-315. http://dx.doi.org/10.15304/verba.44.2949

MELLO, G. de (1992): "*Se los* for *se lo* in the spoken cultured Spanish of eleven cities", *Hispanic Journal* 13 (1), pp. 165-179.

PENNY, R. (2014 [1993]): *Gramática histórica del español*. Barcelona: Ariel.

RAE - ASALE (2009): *Nueva gramática de la lengua española. Morfología y sintaxis*. Madrid: Espasa Libros.

RIVAROLA, J. L. (1985): "*Se los* por *se lo*", *Lexis* 9 (2), pp. 239-242.

ROJO, G. (2016): "*Citius, maius, melius*: del CREA al CORPES XXI", in J. Kabatek (ed.): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlin-Boston: Walter de Gruyter, pp. 197-212.

ROMERO CAMBRÓN, Á. (1998): *Historia sintáctica de las comparativas de desigualdad*. Cuenca: Ediciones de la Universidad de Castilla-La Mancha.

SÁNCHEZ LÓPEZ, C. (2006): *El grado de adjetivos y adverbios*. Madrid: Arco Libros.

URRUTIA CÁRDENAS, H. & M. ÁLVAREZ ÁLVAREZ (1983): *Esquema de morfosintaxis histórica del español*. Bilbao: Publicaciones de la Universidad de Deusto.

VIGARA TAUSTE, A. M.ª (2010): "Gramática, «excepción», norma y uso: a propósito de la construcción *más mayor*. Aspectos sincrónicos y diacrónicos (I)", *Revista de la Sociedad Española de Lingüística* 40 (2), pp. 123-140.

VIGARA TAUSTE, A. M.ª (2011): "Gramática, «excepción», norma y uso: a propósito de la construcción *más mayor*. Aspectos sincrónicos y diacrónicos (II)", *Revista de la Sociedad Española de Lingüística* 41 (1), pp. 113-127.

EL *ATLAS SINTÁCTICO DEL ESPAÑOL* (ASINES): UNA HERRAMIENTA PARA CODIFICAR LA VARIACIÓN*

The Syntactic Atlas of Spanish *(ASinEs): A tool to encode variation*

LORENA CASTILLO
Universitat Autònoma de Barcelona

M. PILAR COLOMINA
Universitat Autònoma de Barcelona

IRENE FERNÁNDEZ
Universitat Autònoma de Barcelona

Resumen
El objetivo de este artículo es presentar el *Atlas Sintáctico del Español* (ASinEs), una herramienta en línea para codificar la variación sintáctica en español. La base de datos del ASinEs contiene tanto datos empíricos, procedentes de las gramáticas de referencia en combinación con datos extraídos de redes sociales, como información teórica. Nos basaremos en un fenómeno concreto, la pasivización de dativos (Montalbetti 1999), para mostrar el funcionamiento del atlas, cómo se introducen y tratan los datos, incluyendo los extraídos de redes sociales, y cómo se realiza la búsqueda por parte del usuario.

Palabras clave: variación del español, atlas sintáctico, redes sociales, pasivización de dativos

Abstract
This paper introduces the online atlas for Spanish syntactic variation (ASinEs) as a tool for gathering and understanding syntactic phenomena of Spanish dialects. To this end, we focus on a particular dialectal phenomenon: "dative passivization" (Montalbetti 1999). We show how the phenomenon is encoded by researchers in the Atlas database, which combines data and theoretical information from the reference grammars of Spanish and social networks, and how users can find the relevant information.

* Nos gustaría agradecer a los asistentes de los congresos *European Dialect Syntax* (Edisyn), Workshop IX, y 49th *Linguistic Symposium on Romance Languages* (LSRL) los comentarios recibidos a una versión previa de este trabajo. También queremos dar las gracias a los revisores anónimos, cuyos comentarios fueron útiles para la versión final del artículo. Este trabajo se ha beneficiado de las ayudas concedidas 2017SGR634 Grup de Lingüística Teòrica, ICREA (ICREA Acadèmia 2015, Ángel J. Gallego) y Ministerio de Economía y Competitividad (FFI2017-87140-C4-1-P; PIF2016, FPU2015 y FPU2016).

Keywords: Spanish variation, syntactic atlas, social networks, dative passivization

1. INTRODUCCIÓN

El objetivo de este artículo es presentar el *Atlas Sintáctico del Español* (ASinEs), una herramienta en línea para codificar la variación sintáctica en español. La base de datos del ASinEs contiene tanto datos empíricos, procedentes de las gramáticas de referencia en combinación con datos extraídos de redes sociales, como información teórica. A lo largo del artículo, se describirá la manera en la que los datos y la información están codificados y se mencionarán algunos de los problemas y cuestiones que estos plantean.

Para mostrar el funcionamiento del atlas, nos basaremos en un fenómeno concreto, la pasivización de dativos (Montalbetti 1999), PD a partir de ahora. En primer lugar, se presentará dicho fenómeno y, a continuación, explicaremos cómo se introducen y tratan los datos en el ASinEs, incluyendo los extraídos de redes sociales.

El artículo está organizado de la siguiente forma: en el siguiente apartado se presentará el proyecto ASinES. Explicaremos en primer lugar los antecedentes y los objetivos del proyecto además de su estructura y funcionamiento. En el apartado 3, nos centraremos en el fenómeno que será objeto de estudio: la pasivización de dativos. Se describirá el fenómeno en cuestión y se ofrecerá un breve resumen de los análisis que se encuentran en la bibliografía. A continuación, se explicará cómo se ha introducido la PD en el ASinES y las conclusiones a las que podemos llegar a través de las búsquedas en la red social Twitter.

2. EL PROYECTO ASINES

2.1 Antecedentes y objetivos

La elaboración de atlas lingüísticos y etnográficos ha constituido uno de los objetivos clave para el desarrollo de la dialectología tradicional (*cfr*. Alvar 1996, García Mouton 1994, Vaquero Ramírez 1996). En ellos se encuentran representados los rasgos lingüísticos característicos de determinadas zonas geográficas. Tradicionalmente, las variantes que aparecen en los diferentes enclaves geográficos corresponden a la información extraída de un cuestionario elaborado previamente por un lingüista. Asimismo, los hablantes son preseleccionados de acuerdo a una serie de criterios que garantizan la asiduidad de sus respuestas.

En el ámbito europeo, Mouton (2016) distingue dos fases en las que podemos dividir la creación de atlas: en primer lugar, una fase de desarrollo de atlas de tipo nacional (macroatlas), que presentan un alcance más amplio y, en segundo lugar, una fase de elaboración de atlas regionales (microatlas). En el ámbito 'macro' destaca la aparición de proyectos como el *Atlas Lingüístic del Domini Català* (ALDC), impulsado por Germà Colón, o el *Atlas Lingüístico de la Península Ibérica* (ALPI), iniciado por Navarro Tomás, entre otros. En relación con los atlas de alcance más reducido, son ampliamente conocidos el *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja* (ALEANR), el *Atlas Lingüístico y Etnográfico de Castilla-La Mancha* (ALeCMan) y el *Atlas Lingüístico y Etnográfico de Andalucía* (ALEA).

Uno de los rasgos que comparten los atlas mencionados es que centran su atención en fenómenos de tipo fonético, morfológico y léxico. Esto se debe en gran medida a que el alcance de estos trabajos no es meramente lingüístico, sino que cubre también aspectos etnográficos. Por este motivo, es habitual que el léxico reciba especial atención y pueda de esta forma representar aspectos sociales y culturales. A lo largo de los diferentes atlas

los campos semánticos como la ganadería, la medicina, la casa y la familia aparecen de forma reiterada. Los siguientes mapas extraídos del *Atlas Lingüístico de la Península Ibérica* (ALPI) constituyen un ejemplo de ello:

Figura 1. Uso de la palabra *aguijón* (y sus variantes)

FIGURA 2. Verbo usado para expresar *mecer la cuna*
[tomado de Fernández-Ordóñez 2011: 51-52]

La información que aparece representada en menor medida es, sin duda, la variación sintáctica. Desde el punto de vista dialectológico, algunos lingüistas llegan a afirmar que este nivel es el que presenta diferencias menos acusadas entre los dialectos del español (Sánchez Lobato 1994: 560). En realidad, las diferencias sí existen, pero la dialectología ha tendido a obviar este tipo de información en sus trabajos (Camus & Gutiérrez, en prensa). Como apunta Morillo-Velarde (1992: 219-220), algunas de las razones pueden ser de tipo metodológico: las encuestas y entrevistas estabas diseñadas para recoger información fonética, morfológica y léxica, siendo de poca utilidad para documentar la variación sintáctica. Además, la ausencia de modelos teóricos que permitan analizar adecuadamente este tipo de variación ha contribuido a su invisibilidad.

El proyecto ASinEs tiene como principal objetivo cubrir este vacío existente y ofrecer una herramienta útil para el estudio de la variación sintáctica del español (ver también Cerrudo *et alii* 2015). Concretamente, el ASinEs es una herramienta interactiva

en línea con formato de atlas que ofrece un estudio sistemático de los fenómenos de las variedades geográficas del español peninsular y de América.

Como se detallará a lo largo del artículo, en el presente proyecto se incorporan nuevas herramientas *en línea* (basadas en Sistemas de Información Geográfica) que permiten una descripción más detallada de los diferentes fenómenos y facilitan la creación de un sistema más sofisticado de mapificación. Además, el ASinEs parte de una visión ampliada del concepto de "atlas", puesto que incorpora información procedente de diversas fuentes que incluyen gramáticas, redes sociales, estudios monográficos, trabajo de campo, etc. La multiplicidad de sus fuentes posibilita que también sea una amplia base de datos de referencia.

Tal y como se desarrollará en la siguiente sección, la ficha, que es el formato mediante el cual se articulan los diferentes fenómenos, permite incorporar información gramatical que suele estar ausente en los atlas tradicionales. En este sentido, el ASinEs sigue la línea de otros atlas interactivos, muchos de los cuales están listados en la web del proyecto Edisyn[1].

Debido a la diversidad de información y de fuentes, el ASinEs es una herramienta que puede servir tanto a investigadores interesados en la sintaxis del español como a profesores y estudiantes de diferentes niveles. En la siguiente sección, se explicará con más detenimiento el funcionamiento del atlas y la codificación de los diferentes datos y fenómenos.

2.2. Estructura y funcionamiento del ASinEs

En esta sección introduciremos brevemente el funcionamiento general del ASinEs, antes de tratar el caso de estudio seleccionado. El atlas está organizado en torno a tres ejes fundamentales: **las fichas**, **el glosario** y **la base de referencias bibliográficas**. Estos tres elementos están relacionados entre sí mediante hipervínculos y se estructuran con un sistema de etiquetado (**ontología**). A continuación describiremos con más detalle cada uno de estos tres ejes: en primer lugar, hablaremos de la base de referencias y el glosario y, a continuación, del sistema de fichas. Haremos más hincapié en este último punto, puesto que las fichas contienen los datos y mapas que codifican la información gramatical y dialectal.

El usuario tiene a su disposición la pestaña "bibliografía" para consultar **la lista de referencias bibliográficas** (*vid.* figura 3). Esta lista contiene actualmente alrededor de 4.000 referencias sobre variación, entre las que se incluyen diccionarios y gramáticas, volúmenes, monografías y artículos de revista. Esta base de datos pretende, no solo informar al usuario de las fuentes consultadas para la elaboración del atlas, sino también ser en sí misma una herramienta útil para el investigador, que puede realizar búsquedas según distintos criterios, como la variedad del español sobre la que trata la obra o la temática específica.

[1] http://www.meertens.knaw.nl/edisyn/searchengine/

FIGURA 3. Lista de referencias bibliográficas

Otra herramienta que ofrece el ASinEs es un **glosario** de términos lingüísticos (http://asines.org/glosario) que aclara y amplía los conceptos gramaticales usados en el atlas. El glosario presenta una triple función: además de ser un recurso para la consulta y resolución de dudas, actúa como eje vertebrador de la variación documentada, ya que recoge las diferentes fichas que tratan el fenómeno que se define. Finalmente, es otro método de búsqueda para acceder a las fichas mediante hipervínculos.

Los términos escogidos para su definición en el glosario provienen del sistema de etiquetado u ontología del atlas. Cada entrada del glosario está compuesta por cinco apartados: *ejemplos, definición, características, bibliografía* y *fichas relacionadas* a los que se suma un apartado en el que se indica la autoría de la entrada (*vid.* figura 4).

FIGURA 4. Ejemplo de entrada del glosario (http://asines.org/loismo)

En *ejemplos* y *definición* se ofrece una descripción más general del fenómeno y en el apartado *características* se amplía el contenido anterior. Para esta sección partimos de la información presente en las gramáticas de referencia (*Gramática Descriptiva de la Lengua Española* 1999, *Nueva Gramática de la Lengua Española* 2009[2]) y la *Enciclopedia de Lingüística Hispánica* (2016), que se completa mediante la consulta de otras fuentes especializadas. Todas estas referencias se listan en el apartado *bibliografía*, en el que no solo se citan las obras consultadas para elaboración de la entrada, sino también otras obras relacionadas que pueden ser de interés para el usuario. Finalmente, en el apartado *fichas relacionadas*, aparecen las fichas relevantes para el fenómeno, listadas mediante hipervínculos.

Vistas las herramientas de la base de referencias y el glosario, pasemos a ver cómo se estructuran las **fichas**, en las que aparece la información fundamental del atlas. En primer lugar, es necesario señalar qué información constituye una ficha. Aunque en ocasiones los "fenómenos" codificados en el atlas tienen relación con la semántica, la morfología o la fonética, solamente consideramos parte del atlas aquellos que reflejan variación sintáctica y con distribución diatópica (aunque también se recogen algunos fenómenos sujetos a variación diastrática y diafásica). Por ello, la distribución es el criterio para determinar qué constituye una ficha o, en otras palabras, si fenómenos

[2] A partir de ahora nos referiremos a estas obras con las siglas GDLE y NGLE respectivamente.

similares deben aparecer en una misma ficha o tener sus propias entradas. Esta jerarquía de criterios se especifica en (1):

(1) Distribución geográfica > distribución de otro tipo > características sintácticas/gramaticales

Consideremos el ejemplo de la ficha de la figura 5:

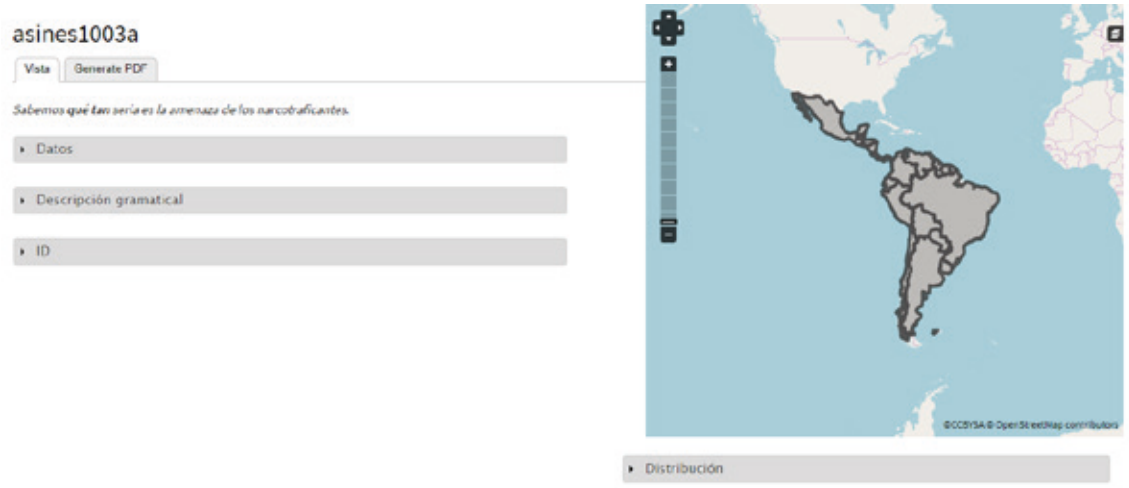


FIGURA 5. Ejemplo de ficha

Según la NGLE la secuencia "qué tan" como locución interrogativa cuantificativa se extiende a las variedades de la mayoría de países hispanoamericanos. Por tanto, no es necesario crear dos fichas independientes para *qué tan + adjetivo* y *qué tan + adverbio*, ya que la distribución es la misma. Por otro lado, si se documentase variación diatópica en cuanto a las características sintácticas, esto se reflejaría en dos fichas separadas que se conectarían mediante la entrada del glosario relacionada con este fenómeno. Es decir, en el ejemplo, si la variante con adverbio solo fuera característica de un dialecto X y con adjetivo, de un dialecto Y, las dos estructuras estarían recogidas en dos fichas independientes.

El ASinEs no pretende ser un proyecto que refleje un trabajo de campo específico (como sí lo son otros proyectos como el COSER; Fernández-Ordóñez (2005-)), sino una herramienta donde poder volcar otros trabajos. Para empezar a estructurar el atlas, se pactó partir de la variación documentada en la NGLE, aunque no sea una obra especializada en dialectología. A partir de aquí, se irá precisando la información mediante otras obras más especializadas, como artículos, tesis o monografías, como los recogidos en la lista de referencias comentada anteriormente. Como consecuencia, el ASinEs estará en constante actualización y no será extraño que el número de fichas disponibles varíe en el tiempo.

Además de la información de gramáticas y monografías, se está llevando a cabo la inclusión de datos de otras fuentes, en concreto, de la red social Twitter. Este tipo de fuentes ha despertado un interés creciente en los investigadores en lingüística, no solo para explorar la variación léxica o discursiva, sino también sintáctica (ver Fernández-Serrano 2017 y referencias allí citadas para un resumen). Entre otras ventajas[3], destacan

[3] Como es obvio, también existen desventajas en cuanto a la utilización de Twitter como fuente que no discutiremos

el rápido acceso a los datos que, en general, son más cercanos a la lengua espontánea y que, por lo tanto, suelen reflejar más fielmente el idiolecto del hablante y, por otro lado, también la posibilidad de obtener datos de prácticamente todas las variedades del español. Aunque el atlas no pretende tener una función de corpus, creemos que la inclusión de ejemplos extraídos de Twitter en las fichas puede ampliar el conocimiento de un fenómeno, especialmente en cuanto a su distribución geográfica, e incluso ayudar a detectar nuevas líneas de investigación.

Una vez descritos los constituyentes y las fuentes de una ficha, explicaremos su estructura. Cuando accedemos a una ficha, podemos ver dos columnas (*vid.* figura 5)[4]: a la izquierda aparece la información gramatical y a la derecha, su distribución geográfica. Esta es la caracterización básica del "fenómeno"[5] recogido en la ficha, cuya información se amplía en el resto de apartados, presentados como un menú desplegable para comodidad del usuario.

Como vemos en la figura 6, la ficha dispone de una caracterización básica de la variación que se está documentando. En este apartado el objetivo es ofrecer una descripción superficial que simplemente informe al usuario de qué debe observar en el ejemplo (normalmente señalado en negrita). Debe tenerse en cuenta que el ejemplo en concreto no es crucial al documentar la variación, como se ha mencionado en el apartado anterior, y que por ello se ofrecen otros ejemplos en la ficha.

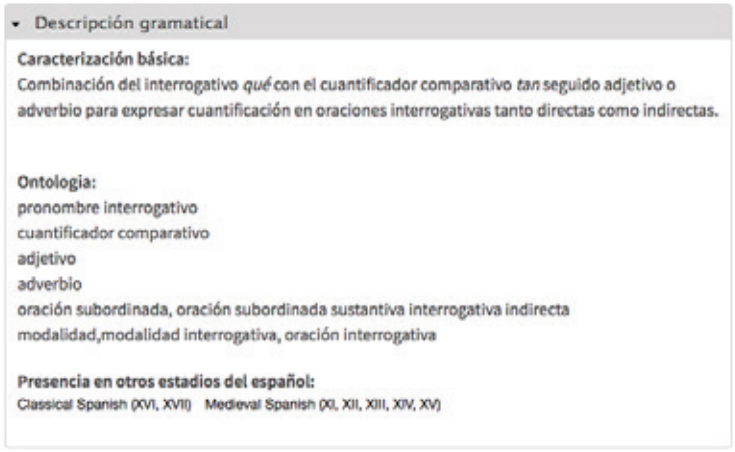


FIGURA 6. Apartado de "descripción gramatical" de una ficha

Por el contrario, en el apartado "análisis" sí que se incluye una breve discusión, que implica un acercamiento teórico al fenómeno. Este doble nivel, uno más descriptivo y otro más analítico/teórico, responde a la voluntad de que el ASinEs sea una herramienta útil para diferentes tipos de usuarios: que sea provechosa para investigadores y estudiosos de la lengua y que a la vez no resulte inaccesible para el público general.

Además de los campos que hemos comentado, es posible incluir información adicional en la ficha si se dispone de ella y se considera pertinente, como información

aquí por motivos de espacio (véase Fernández-Serrano 2017).

[4] Aunque la concepción general de la ficha se mantiene desde el inicio del proyecto, la apariencia y algunas secciones de las fichas han ido variando. La ficha que presentamos aquí puede compararse con una versión previa descrita en Cerrudo *et alii* (2015).

[5] Utilizamos aquí "fenómeno" como término amplio para referirnos a cualquier tipo de variación sintáctica documentada, ya que no hay una etiqueta que englobe de forma general la relación entre variantes respecto a un elemento o estructura concretos. Sin embargo, en el ASinEs solo aparece la palabra "fenómeno" referida a aquellos casos en los que existe una etiqueta más o menos aceptada tradicionalmente como "leísmo" o "dequeísmo".

diacrónica, diastrática o diafásica. También puede indicarse si el fenómeno está presente en otras lenguas.

Finalmente, se pueden consultar las etiquetas utilizadas para clasificar la ficha y hacerla accesible para su búsqueda. El hecho de que las etiquetas sean visibles favorece la navegación del usuario a través del atlas, ya que puede hacer clic en dichas etiquetas y consultar otras fichas que también estén relacionadas con esa palabra clave. En el caso del ejemplo anterior, vemos que podemos acceder a esta ficha mediante la búsqueda de términos muy generales como sería "adverbio", pero también otros más específicos como "cuantificador comparativo". Cabe mencionar que no se etiquetan todos los términos del ejemplo, sino solo aquellos que tienen relación con la variación específica que se está documentando, ya que, como ya hemos comentado, las palabras léxicas seleccionadas para ejemplificar el fenómeno no son relevantes para la variación. En otras palabras, aunque en los ejemplos anteriores aparecen sustantivos o determinantes, no están etiquetados porque no son relevantes para la variación. Lo mismo ocurre con otras etiquetas, como las de funciones sintácticas o estructuras oracionales y no oracionales. Consideremos, por ejemplo, la ficha asines083:

(2) Ejemplo: "**Yo** no pude estar allí, **yo** oí la gritería, pero **yo** estaba en mi oficina y los oí gritando."

Distribución: en el español antillano, en el de otras partes del Caribe, así como en el hablado en las islas Canarias y en el occidente de Andalucía (España)

Ontología: sujeto, tácito/pronombre, pronombre personal, pronombre personal tónico

En este ejemplo se etiqueta la función "sujeto" y la categoría gramatical de "pronombre tónico", pero no se especifica ni tipos de oración, ni el tiempo del verbo, ni otras características del ejemplo concreto porque no son relevantes para documentar la variación en este caso.

En la sección §4 discutiremos cómo se crea una ficha, basándonos en el fenómeno de la pasivización de dativos, y veremos qué problemas presenta codificar la variación sintáctica y a qué soluciones o "compromisos" se puede llegar para ser lo más fieles posible a la realidad lingüística y, a la vez, cumplir los objetivos de nuestra herramienta.

3. PASIVIZACIÓN DE DATIVOS

Con el fin de ilustrar cómo funciona la elaboración de una ficha, hemos escogido un fenómeno concreto que se encuentra registrado en el atlas: la pasivización de dativos, ejemplificado en (3b), en contraste con la oración activa de (3a):

(3) a. Prohibí *a María* leer un libro.
b. *María* fue prohibida de leer un libro.

[Datos extraídos de Montalbetti 1999: 134]

No obstante, antes de mostrar cómo se presenta en el atlas la información, es necesario, por un lado, describir brevemente los contextos sintácticos en los que puede

aparecer este fenómeno y, por otro, explicar los diferentes análisis propuestos en la bibliografía para dar cuenta de su distribución.

En primer lugar, la PD se considera un fenómeno del español, dado que no es gramatical en la mayoría de dialectos, ni peninsulares ni hispanoamericanos. Por otro lado, en las pocas variedades que se permite la PD, el fenómeno está sujeto a muchas restricciones sintácticas, que se comentarán en esta sección. En cambio, todos los dialectos hispanos aceptan la pasiva del objeto acusativo:

(4)	a.	(Le) di un regalo a Juan.	voz activa
	b.	Un regalo le fue dado a Juan.	objeto directo pasivizado
	c.	*Juan fue dado un regalo.	objeto indirecto pasivizado

(4b) ejemplifica la pasiva de acusativo, que de ahora en adelante llamaremos "pasiva no marcada". Por su parte, (4c) muestra un caso de PD, agramatical en todos los dialectos del español[6]. Sin embargo, la GDLE (1999: 1622) señala que se encuentran documentados algunos datos de PD, aunque muy excepcionalmente y restringidos a géneros periodísticos.

(5) Cuando fue llamado la atención por circular por dirección prohibida.

[Datos extraídos de Marcos Marín 1980]

Este tipo de casos se han tratado en la bibliografía (Mendikoetxea 1999, Marcos Marín 1980) como errores gramaticales, por ser considerados meros anglicismos. Es plausible atribuir a este tipo de pasivizaciones el estatus de anglicismos, debido a su escasa aparición en textos tanto orales como escritos, limitada a titulares de prensa. De hecho, las variedades del español que permiten la pasivización de dativos no la aceptan en configuraciones sintácticas como las de (5), donde el OD es un SD, y no una oración subordinada.

A diferencia del nominativo y el acusativo (o ergativo y absolutivo si se trata de una lengua ergativa), el dativo presenta menos homogeneidad entre las lenguas en cuanto a su distribución sintáctica. En efecto, nominativo y acusativo exhiben un comportamiento sintáctico similar en todas las lenguas. Ambos pertenecen a la clase de casos llamada "estructural" (Chomsky 1986)[7], que se caracterizan por ser sensibles a la configuración sintáctica y estar asociados a la concordancia de rasgos de número y persona. Así, si se manipula la estructura argumental, como en una pasiva perifrástica o refleja, el caso se ve alterado, ya que las condiciones estructurales son diferentes (*vid* (6) y (7)). En cambio, los casos no estructurales son aquellos que no dependen de la configuración sintáctica. Se mantienen pese a que la derivación sintáctica cambie (*vid* (8)).

[6] La pasivización de dativos en los dialectos del español que lo permiten solo es posible bajo ciertas condiciones sintácticas (explicadas en esta sección). En contextos como (4c) es agramatical en todas las variedades.

[7] Asumiendo la distinción clásica entre Caso estructural y Caso no estructural (Chomsky 1986), el primero se asigna como resultado de una relación de concordancia de número y persona entre un núcleo funcional y un argumento sintáctico. Este tipo de Caso no es interpretable en el componente semántico, dado que está motivado únicamente por requisitos morfológicos. La única restricción que impone la operación de concordancia es la localidad. El argumento sintáctico más cercano al núcleo funcional activo se llevará el Caso estructural independientemente de si ambos constituyentes se han generado en la misma cláusula. Por lo que respecta al Caso no estructural, se obtienen a partir del ensamble externo entre un núcleo léxico y un argumento. Solo el primero es sensible a las configuraciones sintácticas, participando en alternancias.

(6) a. La vi (a María). María= acusativo
b. María fue vista (por mí). María= nominativo

(7) a. Vi que María/ella corría. María= nominativo
b. La vi correr. María= acusativo

(8) a. Escribí a María. María= dativo
b. *María fue escrita. María= dativo

Mientras que nominativo y acusativo son casos estructurales en todas las lenguas, el dativo suele presentar un comportamiento híbrido, exhibiendo rasgos de ambos tipos de caso. En función de la lengua, estos rasgos pueden acercarle más al caso estructural, como ocurre en inglés, o al no estructural, como en español.

Por lo que respecta al español (y al resto de lenguas románicas), el dativo[8] no puede participar en alternancias de pasivas perifrásticas o reflejas, pero sí lo puede hacer en otro tipo de construcciones en las que el OD está encabezado por un SD que contiene a su vez un genitivo de posesión (*vid* (9)):

(9) a. Pisé las gafas de Juan. Juan= genitivo
b. Le pisé las gafas a Juan. Juan= dativo

Este tipo de alternancia es la única que se documenta en las variedades del español que no permiten la PD. Lo relevante aquí es que no puede ocupar nunca la posición de sujeto, a diferencia del inglés. Como se puede ver en (10), esta lengua acepta tanto la pasiva del acusativo, como la del dativo:

(10) a. John gave Mary a book. / John gave a book to Mary.
John dio María un libro John dio un libro a Mary
'John le dio a Mary un libro/John le dio un libro a Mary'
b. A book was given to Mary.

[8] Las características de los dativos en español son las que listamos a continuación: (i) van precedidos siempre por la preposición *a*, (ii) pueden pronominalizarse por *le/les*, (iii) pueden aparecer doblados por el pronombre *le/les* en cualquier contexto e independientemente del tipo de SD, (iv) no pueden someterse a pasivización, (v) pueden no concordar en número con el pronombre *le*, (vi) pueden participar en alternancias con el genitivo y (vii) no van asociados a un rol temático concreto.

(i) Di un regalo *(a) Juan.
(ii) Envié un paquete a los niños.
(iii) Les envié un paquete (a los niños).
(iv) a. Fue enviado un paquete a los niños.
b. *Fueron enviados los niños un paquete.
(v) El chocolate le(s) gusta a todas las niñas.
(vi) a. Le lavé el pelo a María.
b. Lavé el pelo de María.
(vii) a. Le di un regalo. (destinatario)
b. Le gusta el chocolate. (experimentante)
c. Se le rompió el vaso. (afectado)

un libro fue dado a Mary

'Un libro le fue dado a Mary'

c. Mary was given a book.

Mary fue dada un libro

'A María le fue dado un libro'

Como puede observarse en (10), en inglés el dativo exhibe un comportamiento sintáctico más parecido al acusativo y nominativo que en las lenguas románicas. De hecho, la pasivización del dativo en esta lengua está poco restringida. Es posible en muchos tipos de configuraciones sintácticas.

Sin embargo, tal y como indica Montalbetti (1999), en algunas variantes hispanoamericanas es posible encontrar ejemplos como los siguientes:

(11) a. Prohibí *a María* leer un libro.

b. *María* fue prohibida de leer un libro.

[Datos extraídos de Montalbetti 1999: 134]

Según Montalbetti (1999), la variedad peruana permite dos tipos de pasiva perifrástica en función del objeto interno que sube a la posición de sujeto: directo (12b) o indirecto (12c).

(12) a. Prohibí *a María* leer un libro.

b. Leer un libro le fue prohibido *a María*.

c. *María* fue prohibida de leer un libro.

Este fenómeno solo puede tener lugar en los contextos siguientes: (i) el verbo que permite la PD tiene que ser ditransitivo (*cfr*. (13)), (ii) debe contener un OD oracional (*cfr*. (14)), (iii) la oración subordinada debe estar encabezada por un infinitivo (*cfr*. (15), (iv) el sujeto del infinitivo tiene que ser correferente con el OI (*cfr*. (16)):

(13) a. Escribí a María.

b. *María fue escrita.

(14) a. Di un regalo a María.

b. *Fue dada María (de) un regalo.

(15) a. Prohibí a María que leyera un libro.

b. *María fue prohibida (de) que leyera un libro.

(16) a. Prometí a María leer un libro.

b. *María fue prometida (de) leer un libro.

En cuanto a la oración pasiva, el OD debe ir precedido de *de* (*cfr*. (17)), el OI en la pasiva debe preceder al verbo principal (*cfr*. (18)):

(17) a. Prohibí a María leer un libro.
b. María fue prohibida *(de) leer un libro.

(18) a. Prohibí a María leer un libro.
b. *Fue prohibida María de leer un libro.

3.1. PD: Propuestas de análisis

Montalbetti (1999), además de proporcionar la descripción de los datos de la variedad peruana, propone un análisis para explicar este fenómeno en términos generativistas. Considera la opción de que el OI sea la cláusula de infinitivo (*leer un libro* en (19)) y el dativo de la pasiva pase a ser el OD (*María* en (19)). Sin embargo, descarta esta vía de análisis, ya que no explica por qué la cláusula de infinitivo necesita la preposición "de" y no "a" en la oración pasiva (19b) y que en la activa la "a" nunca sea posible:

(19) a. Juan le prohibió a María (*a) leer un libro.
b. María fue prohibida (*a/de) leer un libro.

Por el contrario, sigue un razonamiento basado en los casos del español que no presentan una relación directa entre Caso y función sintáctica. Según el autor, la estructura que contiene la pasivización del OI no es de control, a diferencia de la pasiva del OD (*cfr*. (20)), sino de Marcado Excepcional de Caso (MEC)[9] (*cfr*. (21)):

(20) a. Juan le prohibió [a María] [leer el libro]
b. [Leer el libro]$_i$ le fue prohibido h$_i$ [a María] pasivización de OD (control)

(21) a. Juan le prohibió [a María leer el libro]
b. María$_i$ fue prohibida de [h$_i$ leer el libro] pasivización de OI (MEC)

De acuerdo con esta hipótesis, en el español de Perú las estructuras de MEC no se limitan a verbos de percepción (como *ver*, *oír*, etc.) y causatividad (*dejar* y *hacer*) (*cfr*. (22a), sino que también se extienden a los verbos de "influencia"[10] (*cfr*. (22b)):

[9] En Gramática Generativa, Marcado Excepcional de Caso (*Exceptional Case Marking*) alude a una estructura en la que un argumento sintáctico no puede recibir Caso de un núcleo que pertenece a su misma cláusula, por tratarse de una oración reducida (no verbal) o no finita:

(i) a. John believes Mary/her/*she to be nice.
b. Mary/she/*her is believed to be nice.

'Mary' en (ia) recibe Caso de la oración principal, y no de la subordinada en infinitivo, como lo demuestra el hecho de que en la correspondiente pasiva, (ib), este SD debe ocupar la posición de sujeto. Además, el SD recibe Caso acusativo independientemente del tipo de verbo subordinado (transitivo/intransitivo).

[10] Montalbetti (1999) denomina "performance verbs" a aquellos verbos ditransitivos de control de objeto indirecto que expresan algún tipo de mandato o prohibición. Pertenecen a esta lista verbos como *prohibir, permitir, ordenar,*

(22) a. Juan vio/hizo [a María leer el libro]
b. Juan le prohibió [a María leer el libro]

Sin embargo, es importante indicar que, pese a que para Montalbetti (1999) la estructura que subyace a oraciones de la PD es la misma que la de los verbos de percepción y causatividad (*cfr*. (22)), el sujeto del infinitivo no recibe Caso acusativo como en un MEC canónico, sino dativo, como lo demuestra el hecho de que solo con los verbos de influencia sea posible el doblado de clítico:

(23) a.	Juan (*le) vio/hizo [*a María* leer el libro]	Acusativo
b.	Juan (le) prohibió [*a María* leer el libro]	Dativo

Por otro lado, defender que los SPs pasivizados no son argumentos del verbo principal (caso de control), sino sujetos de la cláusula de infinitivo (configuración de MEC) evita tener que formular más condiciones *ad hoc* para explicar los contrastes entre (24), por un lado, y (25), (26), (27), por otro. Efectivamente, si se atribuyera a las estructuras de PD (*cfr*. (24)) la estructura subyacente de control de objeto, no habría ninguna razón sintáctica para suponer que el mismo SP pudiera pasivizarse en (24), pero no en el resto de casos, (25)-(27):

(24) a.	Juan le prohibió a María leer el libro.	
b.	María fue prohibida de leer el libro.	PD con infinitivo
(25) a.	Juan le prohibió a María que leyera el libro.	
b.	*María fue prohibida (de) que leyera el libro.	PD con sub. finita
(26) a.	Juan le prometió a María leer el libro.	
b.	*María fue prometida de leer el libro.	PD con control de sujeto
(27) a.	Juan le dio un regalo a María.	
b.	*María fue dada un regalo.	PD con SD

En efecto, el análisis de Montalbetti (1999) predice correctamente la imposibilidad de pasivizar el dativo en casos como los de (25)-(27), dado que este no puede analizarse en ninguno de ellos, por motivos diferentes, como sujeto. En (25) el OD está encabezado por una oración subordinada finita, por lo que el SP no puede ocupar la posición de sujeto[11]. Por su parte, (27) contiene un SD como OD. Por lo que respecta a (26), la explicación también se sigue correctamente de la hipótesis de Montalbetti, ya que no hay ambigüedad. La única estructura posible es la de control, a diferencia de (24), porque el

impedir, mandar. Aquí lo traducimos como "verbos de influencia".

[11] En las configuraciones de MEC el verbo subordinado tiene que estar en forma no personal. Eso explica que el sujeto subordinado no pueda legitimarse dentro de su dominio oracional (no puede recibir Caso nominativo) y establezca una relación de Caso con el verbo de la oración principal. Si la oración subordinada es finita, como en (25), el sujeto estará en nominativo y, por tanto, un SP no podrá ocupar esa posición.

SD que es correferencial con el sujeto del infinitivo es el sujeto de la oración principal, no el dativo.

No obstante, este análisis no está exento de problemas; desde un punto de vista teórico no es plausible que un SD pueda recibir casos diferentes ocupando la misma posición sintáctica en una configuración *x*. Se debería estipular alguna condición que diera cuenta de las divergencias que presenta una misma estructura en cuanto a la asignación de Caso.

Por otro lado, Montalbetti defiende que los OIs en todas las variedades del español se comportan igual por lo que respecta a la imposibilidad de pasivizarse. Sin embargo, no explica por qué un dativo, aunque no desempeñe la función de OI, puede hacerlo, ya que en español los únicos objetos que pueden someterse a tal operación transformacional son los ODs.

Otra propuesta que pretende explicar este fenómeno se encuentra en Castillo (2017). Esta autora, a diferencia de Montalbetti (1999), aboga por una estructura de control en ambas oraciones, (20) y (21), repetidas aquí como (28) y (29):

(28) a. Juan le prohibió [a María] [leer el libro]

b. [Leer el libro]$_i$ le fue prohibido h$_i$ [a María] pasivización de OD (control)

(29) a. Juan le prohibió [a María] [leer el libro]

b. [María$_i$] fue prohibida de [h$_i$] [leer el libro] pasivización de OI (control)

De acuerdo con este análisis, el caso nominativo, por ser estructural, está sujeto a restricciones de localidad. Es decir, se asigna al argumento sintáctico más cercano. En una configuración como la de (30a), en la que dos argumentos compiten por el mismo caso estructural (acusativo o nominativo), el argumento SZ es el que lo recibe. Este es el caso de la pasivización de OD (*cfr*. (28a)). Sin embargo, cuando uno de los dos objetos recibe caso no estructural, se vuelve invisible para la asignación de dicho caso, de manera que no hay competición entre los dos argumentos (*cfr*. (30b)). Este es caso de la PD (*cfr*. (29b)), donde la cláusula de infinitivo se desmarca de la competición por el caso nominativo, dado que recibe caso no estructural de la preposición "de", a diferencia de la respectiva oración activa. Así, el OD no bloquea la asignación de caso nominativo al OI.

(30) a. x [[SZ] … SY] SZ= necesita caso

b. x [[P SZ] … SY] SZ[12]= ya lleva caso (no estructural)

Según Castillo (2017), el infinitivo en la estructura de la PD se comporta en algún nivel sintáctico como una oración de gerundio o participio en el sentido de que no necesita caso. La pregunta entonces es: ¿por qué el español de Perú permite insertar preposiciones delante de cláusulas de infinitivo argumentales?, ¿tienen infinitivos "preposicionales" como el inglés (*to see*)? Estas preguntas quedan todavía por responder

[12] "SZ, SY, SP" son siglas que indican, respectivamente, "sintagma Z", "sintagma Y", "sintagma preposicional". "Y" y "Z" son variables.

en los análisis que se han comentado, pero no suponen un impedimento para codificar el fenómeno en el atlas. Dedicaremos la siguiente sección a este propósito.

4. El fenómeno en el ASinEs

A continuación, detallaremos cómo se ha introducido el fenómeno de la pasivización de dativos en el atlas. Se ha dedicado una ficha específica a este fenómeno, ya que no es propio del español general, sino un fenómeno de variación. Es interesante destacar que, aunque se trata de una alternancia con la pasiva "estándar" o "no marcada", el único dato que aparece recogido en el ASinEs la variante con el dativo pasivizado, es decir (31a) no tiene ninguna ficha.

(31) a. Leer el libro le fue prohibido a María. pasiva no marcada
b. María fue prohibida de leer el libro. pasivización de dativos

Esto responde a la necesidad de documentar exclusivamente la variación, ya que el ASinEs no es una herramienta que describa toda la sintaxis del español. Sin embargo, sí que se menciona en la descripción (*cfr*. figura 7) y un usuario que quiera tener esta información más detallada la puede encontrar recogida en el glosario en la entrada dedicada a las "pasivas" en español.

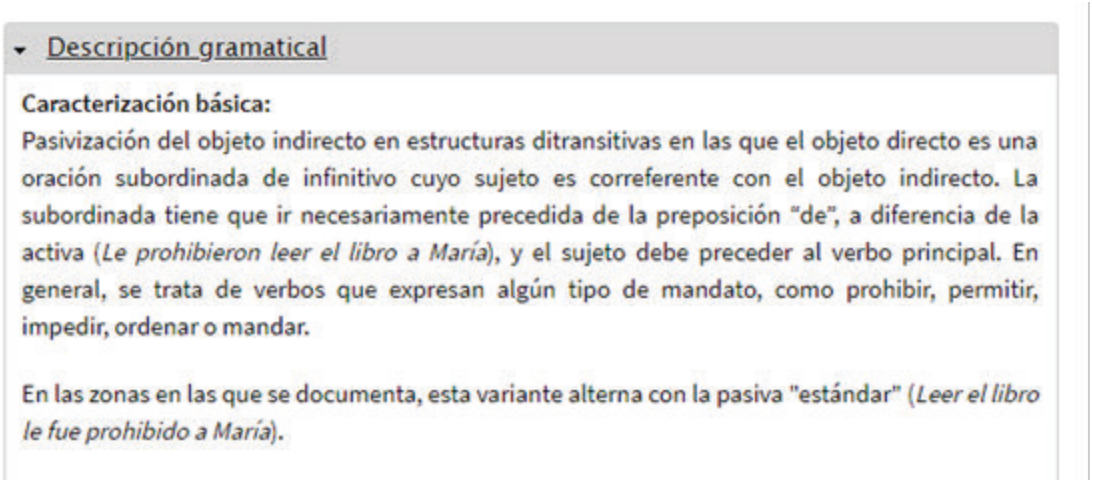


Figura 7. Sección de descripción gramatical de la ficha asines2000

Por otro lado, el etiquetaje de esta ficha es bastante complejo, ya que la variación implica muchos niveles dentro de la estructura. Estos se reflejan de más a menos generales en el orden de las etiquetas:

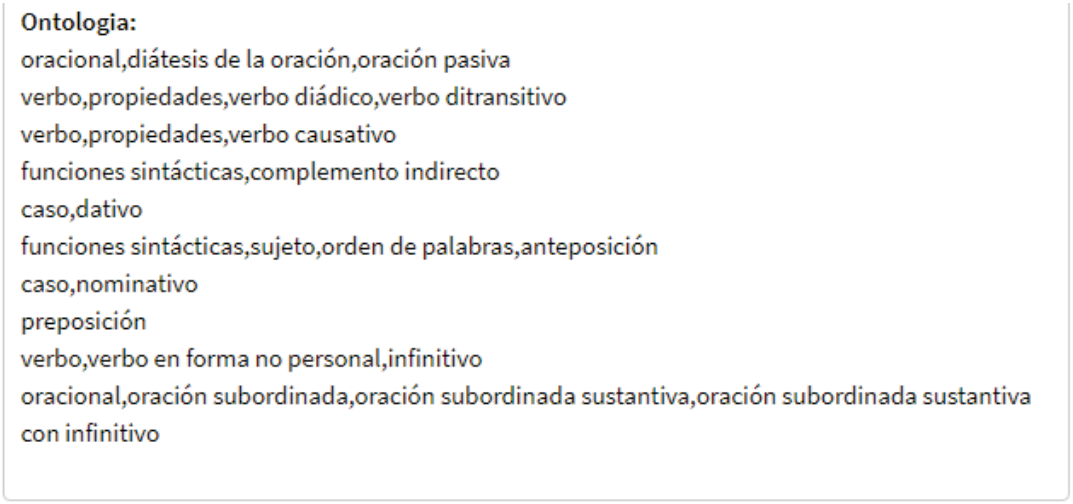


FIGURA 8. Ontología en la ficha asines2000

La etiqueta más general es la que clasifica la diátesis de la oración (pasiva), acto seguido se caracteriza el tipo de verbo que permite este tipo de estructuras (ditransitivo, causativo), a continuación el tipo de argumentos implicados (su función -complemento indirecto-, sus rasgos relevantes -caso dativo y caso nominativo- y el orden de palabras -anteposición-). Aparte de esto también se especifica que es necesaria una preposición y que la cláusula es subordinada de infinitivo.

Todas estas etiquetas permiten al usuario llegar a esta ficha tanto si están interesados en conocer variación sobre pasivas, caso dativo, infinitivos o preposiciones (y todo el resto de etiquetas especificadas). Otra forma de acceder a esta ficha es buscar información sobre el español del Perú en general en el apartado “mapas”:

FIGURA 9. Búsqueda por mapa

Como ya se ha comentado, las fichas también están conectadas al glosario, en este caso se puede consultar la entrada "activa-pasiva". En esta entrada se puede contrastar el fenómeno descrito en la ficha con la formación de las pasivas en el español general, ya que, como se comentó, no hay fichas dedicadas a las variantes "no marcadas".

4.1. Datos de Twitter

Como se ha mencionado en el apartado §2.2., Twitter es una fuente de datos que resulta muy útil para documentar la variación. En el caso de la PD parece especialmente relevante pues, como se indica en la bibliografía, la distribución geográfica de los datos no está estudiada en profundidad. Montalbetti (1999) señala que el fenómeno se produce en diferentes variedades hispanoamericanas, pero él se centra en describir y analizar los datos del Perú.

Las búsquedas en Twitter de verbos de actuación pasivizados ("fue prohibido/a de", "fue permitido/a de", etc.) dan como resultado dos tendencias: por un lado, parece que efectivamente se emplea el PD en lenguaje periodístico, especialmente en titulares y, por otro lado, también hay hablantes que lo producen de manera espontánea y que en este caso la distribución está más restringida al área andina.

En el primer caso, parece que dentro del lenguaje de prensa, estos dativos son más frecuentes en las zonas hispanoamericanas, como se muestra en (32) más abajo. Además, en el ejemplo de (33) se puede comprobar como los titulares de la misma noticia varían en Perú y España:

(32) a. Sindicato policial denunció que otra policía fue impedida de trabajar por tener cáncer de mama. (Uruguay)[13]

b. Marcha opositora fue impedida de llegar a su destino final. (Venezuela)

(33) a. Niña de 10 años fue prohibida de abortar. (Perú)

b. Prohíben abortar a una niña de diez años. (España)

Se documentan también otros casos de 'falsa PD' con verbos como *preguntar*, que parecen regir un OI en estructuras como (34), pero en realidad se está utilizando con la estructura propia del verbo "interrogar" de OI y argumento preposicional[14]:

(34) a. Los griegos *fueron preguntados por* el futuro de su dinero y de su país. (España)

b. Barata SÍ *fue preguntado* por la contratación de periodistas. (Perú)

Por otro lado, aparece documentada en Twitter una PD con la locución "llamar la atención". La extensión de este dato comprende todas las variedades del español. De hecho, la GDLE (Mendikoetxea 1999) aporta un dato con esta construcción (anterior (5)), pero lo considera agramatical por tratarse de un dato excepcional (Marcos Marín 1980 solo lo documento en un caso de lenguaje periodístico). Las búsquedas en Twitter demuestran que no se trata de un caso tan excepcional, puesto que se documenta en numerosos tuits. Además, su uso no parece restringirse al ámbito periodístico:

(35) a. Aclaro q luego *fue llamado la atención* x subir video a You Tube. (Argentina)

b. No le parece que los periodistas *deben ser llamados la atención*. (Ecuador)

Respecto a la distribución geográfica del fenómeno, los datos de Twitter parecen demostrar que el fenómeno es común al área andina, en Perú (*cfr*. (36a)), Bolivia (*cfr*. (36b)) y Ecuador (*cfr*. (36c)). Aunque se documentan casos esporádicos en otras zonas.

(36) a. Odebrecht fue prohibida de estar en Perú desde 1993.

b. Expresidente de Bolivia denuncia que fue prohibido de ingresar a Cuba.

c. Fue prohibida de viajar. Le retiraron el pasaporte.

En cuanto a los contextos sintácticos en los que se documenta este fenómeno, los datos de Twitter apoyan la caracterización que ofrece Montalbetti (1999). No se documentan casos de PD con subordinadas flexionadas (ver (14) y (15) anterior) ni

[13] Los tuits no se han modificado, así que mantienen faltas de ortografía y/o abreviaturas propias del lenguaje de las redes sociales. El país de origen señalado entre paréntesis se corresponde al país del medio (revista o periódico), en el caso de los titulares de prensa, o al que el autor o autora del tuit muestra en su perfil de usuario como su lugar de origen.

[14] Este uso aparece recogido en el *Diccionario Panhispánico de dudas* (RAE-ASALE 2005).

tampoco con OD nominales. Además, en todos los casos encontrados aparece el infinitivo precedido por la preposición 'de' en la pasiva. Respecto al tipo de verbo que permite PD, los datos de Twitter coinciden con la descripción de Montalbetti (1999): todos los datos encontrados se producen con verbos que seleccionan infinitivos de control de OI, como se puede ver en los siguientes ejemplos:

(37) a. la prensa no fue *permitida* de cubrir el evento debidamente (Ecuador)

b. Esta sra debe de ser *impedida de salir* del país [...] (Perú)

Los resultados obtenidos en Twitter permiten mejorar la información que aparece en la ficha del ASinEs. En primer lugar, permiten extender los casos documentados en Perú a toda el área andina (*vid.* figura 10). Además, también se incluye la observación de que en el lenguaje periodístico aparece también en otras áreas hispanohablantes.

FIGURA 10. Sección "distribución" en la ficha asines2000

Los datos que se recogen en Twitter se añaden en el campo "ejemplos" y permiten complementar el dato de referencia tomado de Montalbetti (1999).

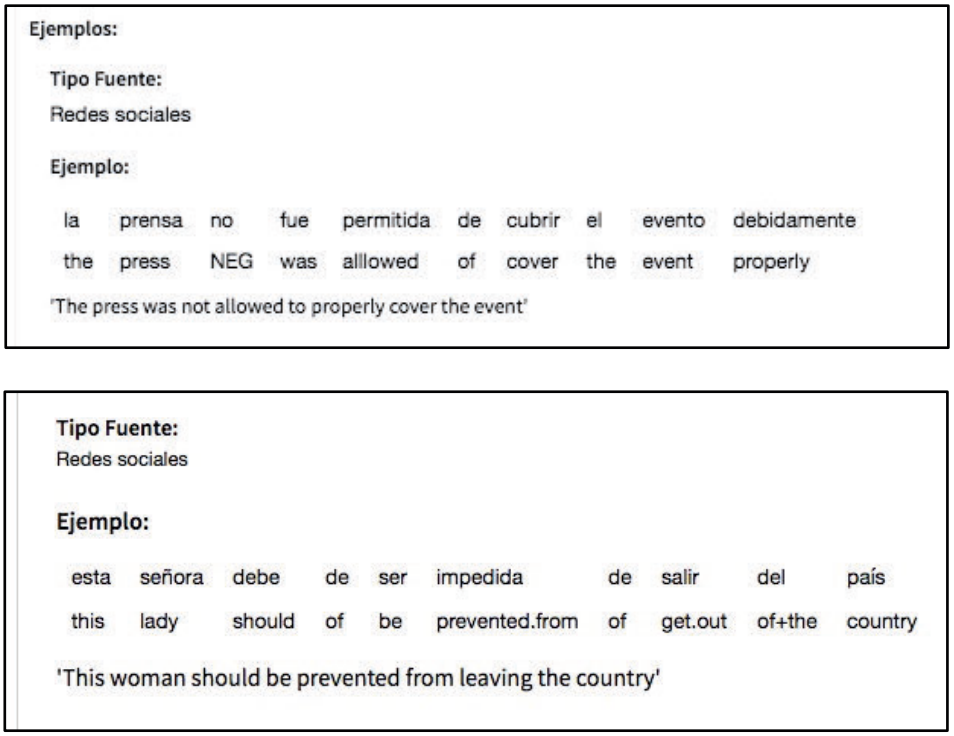


FIGURA 11. Ejemplos de Twitter en la ficha asines2000

Como se ha visto, la información de Twitter permite completar las descripciones que aparecen en las monografías, como en este caso sucede con la de Montalbetti. Aunque es evidente que es necesario un estudio más pormenorizado para acabar de delimitar la distribución geográfica, las redes sociales son una fuente de información que puede dar cuenta de determinadas tendencias que con otros métodos más tradicionales serían difíciles de observar.

5. CONCLUSIONES

En este artículo hemos discutido cómo un fenómeno concreto de variación puede codificarse en una herramienta online como es el Atlas Sintáctico de Español (ASinEs). hemos visto cómo tal recurso combina tres niveles de información. Visualmente, muestra la distribución a través de mapas; descriptivamente, presenta ejemplos de modo más próximo a un corpus; gramaticalmente, proporciona aspectos teóricos más avanzados. Además, hemos recogido los datos que proporciona Twitter y, de esta forma, hemos podido mejorar la información inicial que se ofrecía en la ficha. Las búsquedas de Twitter han corroborado a grandes rasgos la caracterización que señalaba Montalbetti (1999) y han permitido ampliar la información relativa a la distribución del PD.

Como hemos visto, las redes sociales posibilitan realizar búsquedas en un área geográfica muy amplia de una forma rápida. Una de las grandes ventajas que presentan las redes sociales es que permiten observar la lengua coloquial, más cercana a la oralidad, sin intervención directa de un lingüista (por medio de entrevistas). Aunque las conclusiones que los datos de Twitter pueden ofrecer deben tratarse con cautela y, a poder ser, corroborarse mediante otras fuentes o métodos, está claro que pueden observarse tendencias que pueden ser valiosas para el investigador e interacciones entre

diferentes variantes que no deberían ser ignoradas por los investigadores de la lengua, ya que forman parte de un contexto de globalización.

RECURSOS ELECTRÓNICOS MENCIONADOS

ASINES: Á. J. Gallego (dir.) (2014-2020): *Atlas Sintáctico del Español*. Barcelona: Universitat Autònoma de Barcelona. http://www.asines.org. ISSN: 2462-3431.

COSER: I. FERNÁNDEZ-ORDÓÑEZ (dir.) (2005-): *Corpus Oral y Sonoro del Español Rural*. www.corpusrural.es. ISBN 978-84-616-4937-2.

REFERENCIAS BIBLIOGRÁFICAS

ALVAR, M. (1996): *Manual de dialectología hispánica. El español de América*. Barcelona: Ariel.

CAMUS, B. & E. GUTIÉRREZ (en prensa): "Syntactic features and dialect areas in Peninsular Spanish", in A. Cerrudo, Á. J. Gallego & F. Roca (eds.): *Syntactic geolectal variation: traditional approaches, current challenges and new tools*. Amsterdam: John Benjamins.

CASTILLO, L. (2017): "A note on an ECM asymmetry in Spanish", *Isogloss* 3(1), pp. 69-76.

CERRUDO, A., Á. J. GALLEGO, A. PINEDA & F. ROCA (2015): "ASinEs: Prolegómenos de un atlas de la variación sintáctica del español", *Linguamática* 7(2), pp. 59-69. https://linguamatica.com/index.php/linguamatica/article/view/V7N2.5

FERNÁNDEZ-ORDÓÑEZ, I. (2011): "La lengua de Castilla y la formación del español". Discurso leído el 13 de febrero de 2011 en su recepción pública por la Excma. Sra. D.ª Inés Fernández Ordóñez y contestación del Excmo. Sr. D. José Antonio Pascual, Madrid.

FERNÁNDEZ SERRANO, I. (2017): *Long distance agreement in Spanish*. Tesis de máster, Universitat Autònoma de Barcelona. http://hdl.handle.net/2445/116923

GARCÍA MOUTON, P. (1994): *Lenguas y dialectos de España* (Vol. 20). Madrid: Arco Libros.

GUTIÉRREZ, S. (1999): "Los dativos", in I. Bosque & V. Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 1855-1930.

MARCOS MARÍN, F. (1980): *Curso de gramática española*. Madrid: Cincel-Kapelusz.

MENDIKOETXEA, A. (1999): "Construcciones inacusativas y pasivas", in I. Bosque & V. Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 1575-1630.

MONTALBETTI, M. (1999): "Spanish passivized datives: The relevance of misanalysis", in K. Johnson & I. Roberts (eds.): *Beyond Principles and Parameters*. Kluwer: Dordrecht, pp. 133-144.

MORILLO-VELARDE, R. (1992): "Un modelo de variación sintáctica dialectal: el demostrativo de realce en andaluz", in M. Ariza Viguera, R. cano Aguilar, J. Mendoza & A. Narbona (eds.): *Actas del II Congreso Internacional de Historia de la Lengua Española*, vol II. Madrid: Pabellón de España, pp. 219-227.

RAE - ASALE (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.

SÁNCHEZ LOBATO, J. (1994): "El Español en América", in J. Sánchez Lobato & I. Santos Gargallo (coords.): *Problemas y métodos en la enseñanza del español como lengua extranjera. Actas del IV Congreso Internacional de ASELE*. Madrid: SGEL, pp. 553-570.

VAQUERO RAMÍREZ, M. (1996): *El español de América*. Madrid: Arco Libros.

EL *CORPUS ORAL Y SONORO DEL ESPAÑOL RURAL* (COSER) Y SU CONTRIBUCIÓN AL ESTUDIO DE LA VARIACIÓN GRAMATICAL DEL ESPAÑOL*

Audible Corpus of Spoken Rural Spanish *(COSER) and its contribution to the study of grammatical variation in Spanish*

INÉS FERNÁNDEZ-ORDÓÑEZ
Universidad Autónoma de Madrid

ENRIQUE PATO
Université de Montréal

Resumen
El objetivo principal de este trabajo es dar a conocer la contribución pionera del COSER al estudio de la variación gramatical del español europeo y actualizar los informes previos. Su estructura es la siguiente: tras una breve introducción, se lleva a cabo una descripción general del corpus, así como de la metodología empleada en su elaboración. Después, se da a conocer el estado del COSER en línea (enero de 2019). Posteriormente, en el apartado principal del estudio, se resumen y sintetizan las aportaciones concretas que se han realizado hasta la fecha sobre numerosos aspectos de variación gramatical del español gracias a los datos del COSER (morfología verbal, concordancia de caso en los clíticos de objeto, concordancias de género y número, construcciones reflexivas, medias e impersonales con *se*, selección del modo, expresión del complementante, entre otros).

Palabras clave: español europeo, dialectología, variación gramatical, corpus lingüísticos, COSER

Abstract
The main objective of this work is to show COSER's pioneering contribution to the study of grammatical variation in European Spanish and to update previous reports. Its structure is as follows: after a brief introduction, a general description of this corpus is carried out, as well as the

* Este trabajo se enmarca dentro del proyecto "Cambios en el habla rural: del siglo XX (Atlas Lingüístico de la Península Ibérica, ALPI) al siglo XXI (Corpus Oral y Sonoro del Español Rural, COSER)", financiado por el Ministerio de Ciencia, Innovación y Universidades (PGC2018-095077-B-C2). El *Principio del deber contraído* (Labov) y el *Principio de gratitud lingüística* (Wolfram) establecen que todo investigador que disponga de datos lingüísticos de alguna comunidad de habla está en la obligación de ponerlos en conocimiento de dicha comunidad, y debe además buscar los modos prácticos de devolver el favor lingüístico proporcionado. Este trabajo es uno de esos modos para agradecer a todos los informantes su participación en las encuestas del COSER.

methodology used in its elaboration. Then, the status of COSER online is released so far (January 2019). Subsequently, in the main section of the paper, we summarize and synthesize all the contributions that have been made on numerous issues of grammatical variation in Spanish thanks to COSER data (verb morphology, case selection in object clitics, gender and number agreement, *se* constructions -reflexive, middle, and impersonal- mood selection, and complementizer expression, among others).

Keywords: European Spanish, dialectology, grammatical variation, linguistic corpora, COSER

1. INTRODUCCIÓN

Como es bien sabido, el estudio de la variación dialectal del español hablado en España (tanto en la península como en las islas) se ha basado principalmente en los datos de los diversos atlas lingüísticos regionales publicados y en algunas monografías de distinto valor e interés hasta hace relativamente poco tiempo (*cfr*. Fernández-Ordóñez 2007c y ss.). En ambos tipos de trabajos la variación fonética y la variación léxica copaban casi toda la atención. Al hecho de que las monografías dialectales apenas estudiaban la variación gramatical debe sumarse el problema de que se habían centrado sobre todo en los dialectos llamados "históricos" (navarro, aragonés y asturleonés), sin atender en la misma medida al castellano rural. Esta situación deficitaria empezó a cambiar radicalmente con la creación –hace ya treinta años– del *Corpus Oral y Sonoro del Español Rural* (COSER), corpus de grabaciones del habla rural que empezó a compilarse en el año 1990 para complementar, precisamente, las fuentes tradicionales de información y los datos dialectales recopilados mediante cuestionarios no grabados, en especial, en lo relativo a la variación gramatical. Desde entonces, el COSER ha ido creciendo cada año, con sucesivas campañas de encuesta en la península, en Canarias y en Baleares, tal y como veremos más adelante y se puede comprobar en su página web (corpusrural.es).

Por otro lado, las fuentes escritas objeto de estudio en los trabajos sincrónicos solían reflejar solo la lengua estándar castellana, o el estándar regional de cada provincia, en el mejor de los casos. Por ese motivo, había toda una serie de fenómenos de variación gramatical que no se habían documentado en los corpus de referencia anteriores; y en el caso de que se encontraran, podía suceder que su localización geográfica fuera insegura, o que no se documentaran todos los contextos pertinentes para su completa descripción y análisis. No menos importante es el hecho de que la lengua oral –la de todos los sociolectos– tiene todavía poca presencia en las fuentes habituales (CREA, CORPES XXI) que los investigadores emplean en sus trabajos.

Las primeras campañas de encuesta realizadas en el marco del COSER siguieron de cerca los modelos de trabajo de Diego Catalán para el estudio de los romances tradicionales (cuyo fruto sería el *Archivo Sonoro del Romancero 'Débora Catalán'*) y de Flora Klein-Andreu para el estudio de los clíticos *le/s*, *la/s*, *lo/s* en Castilla (la Vieja). Las grabaciones semidirigidas del COSER pronto mostraron que el trabajo de campo ofrece numerosísimas ventajas para la descripción y la explicación de los fenómenos gramaticales, al identificar por primera vez el trazado de la isoglosa de un rasgo determinado o al modificar y precisar una distribución geográfica ya conocida previamente. También sirven para desarrollar y evaluar nuevas hipótesis y explicaciones sobre aspectos de la variación gramatical, pues la cantidad y la calidad de los datos recopilados es mucho mayor de lo que se disponía previamente y, por tanto, su descripción lingüística más precisa. Las encuestas incluso proporcionan información gramatical desconocida, o muy poco estudiada hasta la fecha, por lo que se han podido

descubrir y describir nuevos fenómenos de variación morfosintáctica. De este modo, se han podido entender mejor numerosas estructuras dialectales del español europeo en sus distintas modalidades, septentrional, meridional, occidental y oriental, así como del español en contacto con otras lenguas peninsulares (§ 4).

Por otro lado, y como también tendremos ocasión de mostrar, los datos del COSER permiten comparar mejor los fenómenos registrados en el habla rural con aquellos propios del habla urbana, recopilados en los varios proyectos efectuados sobre la norma culta de las diversas ciudades españolas e hispanoamericanas, como el estudio sociolingüístico del español de España y de América (PRESEEA), o sobre el habla coloquial de algunas ciudades españolas (*Corpus Val.Es.Co* en Valencia, *Vernáculo Urbano Malagueño* en Málaga, *Corpus Oral de Conversa Col·loquial* en Barcelona, entre otros muchos). En efecto, la uniformidad en la metodología empleada en el COSER lo hace útil para medir tanto la distancia lingüística que separa diversas áreas (distancia geográfica) como la distancia lingüística que separa el grupo social rural de mayor edad de otros como el de los hablantes de mayor nivel sociocultural o el de los hablantes más jóvenes (distancia sociológica)[1].

Por último, los datos del COSER también permiten valorar en su justa medida la evolución diacrónica de ciertos fenómenos gramaticales al ser comparables con otras fuentes previas de la dialectología, como los atlas y las monografías dialectales.

El objetivo principal de este trabajo, como indica su título, es dar a conocer la contribución pionera del COSER al estudio de la variación gramatical del español y actualizar los informes previos presentados sobre este corpus rural. Su estructura es la siguiente: tras esta breve introducción (§ 1), se lleva a cabo una descripción general del corpus, así como de la metodología empleada en su elaboración (§ 2). Después, se da a conocer el estado del COSER en línea, a día de hoy (§ 3). Posteriormente, en el apartado principal del estudio, se resumen y sintetizan las aportaciones concretas que se han realizado hasta la fecha sobre fenómenos gramaticales de variación del español gracias a los datos del COSER (§ 4). Las conclusiones finales cierran el trabajo (§ 5).

2. DESCRIPCIÓN GENERAL DEL COSER Y METODOLOGÍA

Varias de las cuestiones generales y metodológicas sobre el COSER han sido previamente tratadas en otros trabajos (*cfr*. Fernández-Ordóñez 2011, 2010a, 2010b, 2009a, 2007c y corpusrural.es), por lo que nos limitaremos en esta ocasión a ofrecer únicamente la información más relevante para su descripción general y la metodología empleada.

Lo primero que hay que recordar es que el COSER es un corpus de entrevistas semidirigidas restringido al habla rural de informantes mayores de escasa escolarización, que han nacido y que todavía viven en la ruralia, los pueblos de España, es decir, está dedicado a las personas que fueron objeto de interés en la dialectología tradicional. Sin embargo, aunque el COSER se nutre del mismo tipo de informantes que los atlas lingüísticos y que muchas de las monografías dialectales publicadas, la metodología empleada y los objetivos son muy diferentes. La diferencia más importante es, quizá, que el desarrollo de la entrevista semidirigida grabada permite investigar el empleo de cualquier fenómeno gramatical en su contexto de uso. En efecto, en lugar de las oraciones aisladas y descontextualizadas propias de un cuestionario lingüístico, la entrevista recoge oraciones emitidas dentro de un discurso real, en el que se pueden investigar, además,

[1] Las conclusiones obtenidas del contraste entre grupos lingüísticos resultan metodológicamente adecuadas si tenemos en cuenta el tipo conversacional en que se obtuvieron los datos, tanto en el COSER como en otros corpus lingüísticos: la entrevista, tipo de conversación sometido al intercambio pregunta-respuesta.

los valores contrastivos, las motivaciones afectivas y las inferencias pragmáticas asociadas a una determinada estructura sintáctica[2].

Las distintas campañas de encuesta (desde 1990 hasta noviembre de 2018) han permitido obtener 1656 entrevistas en 1325 enclaves rurales de 52 provincias e islas de España, lo que suma un total de 1752 horas de grabación. Todo ello hace que la densidad de la red de puntos encuestados y de los datos recogidos sea muy superior a la del único atlas lingüístico nacional o de gran dominio (ALPI) y a la de los distintos atlas lingüísticos (y etnográficos) regionales (ALCyL, ALEA, ALEANR, ALECant, ALECMan y ALEICan). El número de informantes también es mucho mayor, pues el COSER cuenta a día de hoy (enero de 2019) con 2574 informantes registrados: 1351 mujeres (52,4 %) y 1223 hombres (47,5 %). La edad media es de 73 años, siendo ligeramente menor en las mujeres (72,2) que en los hombres (73,7). El número de hablantes de cada sexo es estadísticamente representativo, por lo que permite investigar también diferencias lingüísticas asociadas a este factor.

La recogida de los datos se ha efectuado a ritmo anual desde el verano de 1990 (con una prueba piloto en León en 1988). En septiembre de 2012 y de 2013 se llevaron a cabo dos grandes campañas de encuesta para terminar el sur y el este peninsular, y desde 2016 se vienen realizando las encuestas en las islas Canarias y en Baleares. En total, se han organizado hasta el momento 48 campañas (*cfr*. corpusrural.es/campanas.php).

Por lo que respecta a la metodología empleada, en el COSER se prioriza la entrevista oral semidirigida por parte del equipo de encuestadores, durante la cual uno de ellos suele guiar la conversación, primero hacia ciertos módulos temáticos de la vida rural tradicional como la matanza del cerdo y los cultivos propios del lugar. No obstante, pasado un tiempo y ganada la confianza del informante, la conversación se desplaza hacia otros temas como la vida en el pueblo, la educación recibida, las experiencias personales y la familia, dependiendo siempre del grado de comodidad y de espontaneidad que muestre el informante. La entrevista se centra en estos temas para facilitar que el informante potencial acepte ser entrevistado, al reconocer que está en posesión de ciertos conocimientos sobre un sistema de vida en decadencia, producto tanto de su experiencia personal como de su edad. Todo ello le confiere una autoridad informativa ante el entrevistador que viene de la capital. Los informantes aceptan la entrevista al comprender el interés por el testimonio de su sistema de vida, del que pocos guardan recuerdo y del que es experto.

Por lo general, el primer contacto con los informantes se realiza sin ninguna gestión previa, nada más llegar al enclave seleccionado entre los habitantes del lugar que reúnan los requisitos mencionados (persona mayor de escasa escolarización, que haya nacido y vivido en el pueblo y a la que le guste hablar). Todo investigador que haya realizado trabajo de campo sabe que el éxito de una entrevista no está asegurado, y que una entrevista puede ser mejor o peor bajo las mismas condiciones de partida. A este respecto cabe señalar que no todas las entrevistas son igualmente informativas, sino que su calidad depende de la disposición del informante, de la habilidad de los

[2] Las carencias de los atlas lingüísticos y de las monografías dialectales en lo concerniente al estudio de la gramática no deben, sin embargo, conducir a descartar su utilidad, ya que cuando algunas de estas obras se concibieron y realizaron ni la sintaxis ni la sociolingüistica habían alcanzado el desarrollo teórico que han experimentado en los últimos cincuenta años. Aunque la metodología del cuestionario geolingüístico empleada para reflejar el habla en el ALPI y los atlas regionales posteriores es muy diferente de la metodología de la entrevista sociolingüística del COSER, debe reconocerse que ambas son hijas del estado teórico de la dialectología de su tiempo. El desarrollo de la sociolingüística ha puesto de manifiesto las múltiples limitaciones de la metodología de los atlas lingüísticos, pero hay que tener en cuenta también que, puesto que no existen grabaciones del habla de épocas pasadas equivalentes a las actuales, sus datos siguen constituyendo un testimonio válido, por imperfecto que sea, para estudiar la variación, incluida la gramática (*cfr*. Pato 2003, Heap 2006, entre otros trabajos).

entrevistadores y de la interacción entre ambos. En algunos casos extremos solo es posible identificar *a posteriori* a los informantes que, por circunstancias varias –como haber pasado largas temporadas fuera del pueblo o haber disfrutado de una formación escolar mayor que el resto–, no encajan lingüísticamente ni en su área ni en su perfil sociolingüístico.

Respecto al número de informantes encuestado en cada enclave, para el COSER se suele entrevistar a un solo hablante en profundidad, ya sea hombre o mujer. Esto no impide que, en ocasiones, la entrevista pueda incluir esporádicamente a otros individuos, generalmente miembros de la familia y conocidos del entrevistado, que, atraídos por la circunstancia extraordinaria de la entrevista, se sienten tentados a intervenir y dar también su testimonio.

La duración media de las grabaciones realizadas es de aproximadamente una hora y tres minutos por enclave. Sin embargo, algunas de ellas cuentan con media hora y otras con hasta más de dos horas y media. Como todo investigador de campo sabe, la calidad de los datos no es directamente proporcional a la duración de una entrevista, ya que existen grabaciones excelentes y altamente informativas de media hora, cuyos resultados son equiparables a los obtenidos en una entrevista de mayor duración, y viceversa.

En cuanto al procesamiento de los datos, en 2009 se digitalizaron todos los materiales anteriores que estaban en formato analógico (desde 1990). Desde 2012 las entrevistas se graban tanto en audio como en vídeo, siempre con el consentimiento oral del informante.

La edición de las transcripciones ha seguido los avances de la tecnología, desde la transcripción a mano o a máquina hasta la informatización total de las mismas, como parte de diversos proyectos de investigación y de los trabajos finales de curso para las asignaturas que, en sucesivos planes de estudio en la Universidad Autónoma de Madrid, han tratado la dialectología del español desde 1990 hasta 2018. Las primeras transcripciones, de carácter bastante libre, se convirtieron en transcripciones pautadas desde 2005, con indicación de distintas marcas de conversación, temas tratados y fenómenos dialectales documentados. Las transcripciones disponibles están ahora todas homogeneizadas y se ha empleado para ello un sistema de ortografía convencional en el que se reflejan algunas características propias de la oralidad, como la omisión de algunos segmentos fonológicos (*comprao* por *comprado*) y la adición de otros (*muncho* por *mucho*), así como algunas acentuaciones dialectales (*pajáro* por *pájaro*). Todo ello permite proporcionar una imagen algo más cercana del discurso de los informantes rurales y de su habla, pero manteniendo globalmente las convenciones generales de la ortografía y favoreciendo la legibilidad de los textos (*cfr.* http://www.corpusrural.es/transcripcion.php)[3].

Se reseñan, además, con marcas específicas las hablas simultáneas, las conversaciones cruzadas, otros elementos como risas, toses, chasquidos, elementos de comunicación no verbal –como los gestos–, así como la mayor o menor inteligibilidad de lo transcrito. También se indican las interrupciones (que pueden servir al hablante de herramienta para planificar el discurso) y las autocorrecciones (como indicadores de conciencia sociolingüística), entre otros muchos aspectos (*cfr.* http://www.corpusrural.es/marcas.php).

La decisión de transcribir en ortografía convencional algunos aspectos de la pronunciación dialectal obligó a crear un sistema de desambiguación de algunas formas

[3] Aspectos como la pérdida de [d] intervocálica, la pérdida de algunas consonantes finales (-r, -l, -n, -d), la asimilación de [rl]/ [rs] > [l]/ [s] en los infinitivos con clítico o el desplazamiento acentual se aceptan transgrediendo la ortografía convencional. No obstante, la mayor parte de los cambios fonéticos o fonológicos no se reflejan: seseo, ceceo, yeísmo, neutralización, glotalización y pérdida de -s, -z, etc.

para su posterior etiquetado morfosintáctico. Por ejemplo, la transcripción de la pérdida de la *-d-* intervocálica y de la *-r* final condujo a desambiguar los participios en *-ada* y los infinitivos de la primera conjugación *-ar* para que el lematizador no pudiera confundirlos[4].

Editar el COSER ha implicado tomar decisiones importantes sobre ciertas convenciones necesarias. En efecto, ya que desafortunadamente no contamos con un sistema universalmente aceptado para todos los corpus orales, y que el proceso de transcribir es una tarea hermenéutica sujeta a errores de audición y de comprensión, la uniformidad deseable supone siempre un reto. A esa dificultad debe añadirse que la transcripción y revisión es una tarea colectiva en la que han participado muchos estudiantes y colaboradores a lo largo de los años.

Desde 2016 las labores de gestión del proyecto se realizan en una base de datos en línea, en la que, con distintos permisos de acceso, pueden editarse y transcribirse las grabaciones y someterlas a revisiones sucesivas de forma centralizada[5].

3. El COSER en línea, hoy

Por lo que respecta a la difusión del corpus, desde 2005 a 2015 se alojó en la página web www.uam.es/coser. En 2015 migró a www.corpusrural.es. En esta página se incluye toda la información general sobre el corpus (descripción y mapa de localidades encuestadas), la recogida de los datos (metodología y campañas), la transcripción (marcas de la conversación, transcripción ortográfica, temas tratados), los recursos disponibles (archivos de audio y texto así como muestras dialectales breves de aspectos sujetos a variación) y todo lo relativo al proyecto del COSER (su historia, los participantes, la financiación, las publicaciones, cómo colaborar con el proyecto, y los agradecimientos). Además, se habilitó una consulta básica de las grabaciones entonces transcritas. Desde el año 2017 esa web incorporó una consulta avanzada en pruebas que se ha actualizado, con diversas mejoras, en enero de 2019. En 2018 se completó la información de cada uno de los 1325 enclaves con la ficha técnica relativa a las entrevistas realizadas (código, encuestadores, informantes, duración, fecha de la encuesta), con independencia de que estén o no disponibles como archivo de texto y audio y de que sean accesibles a las consultas básica y avanzada.

Desde 2019 se ofrecen en la web, en texto, audio e interrogables mediante la consulta, 175 entrevistas correspondientes a 169 enclaves, que suman más de 229 horas de grabación y 4 771 943 palabras de texto transcrito, un 13 % del total del corpus[6]. Componen una muestra representativa de las siguientes provincias: Valladolid (6 enclaves), Burgos, Cáceres, Cantabria, León, Palencia, Toledo, Vizcaya, Murcia (5 enclaves), Álava, Albacete, Alicante, Almería, Asturias, Ávila, Badajoz, Cádiz, Castellón, Ciudad Real, Córdoba, Cuenca, Guipúzcoa, Granada, Guadalajara, Huelva, Huesca, Jaén, La Rioja, Madrid, Málaga, Navarra, Salamanca, Segovia, Sevilla, Soria, Teruel, Valencia, Zamora, Zaragoza (4 enclaves) y Lérida (1 enclave)[7].

[4] Un ejemplo concreto de ello sería el caso de *amarrá*, por el adjetivo/participio *amarrada* y por el verbo en infinitivo *amarrar*, por lo que tendríamos dos formas posibles: *amarr(á=ada)* y *amarr(á=ar)*.

[5] https://corpusrural.fe.uam.es/coser-editor/login.php, de uso interno para los colaboradores del proyecto. Esta base de datos permite asimismo controlar las estadísticas.

[6] Aparte de esas 229 horas transcritas y revisadas (13 % del corpus), contamos con otras 322:45 horas transcritas, sin revisar, que corresponden a la transcripción completa de 344 entrevistas, y otras 177:58 horas transcritas, que corresponden a la transcripción parcial de 304 entrevistas (cuya duración suma 311:30 horas). En total, 450:43 horas pendientes de revisión, un 25,6 % del corpus, que, sumado al 13 % ya revisado, alcanza la cifra de 38,6 % del total.

[7] Las transcripciones que ofrezcan una muestra equivalente de Barcelona, Gran Canaria, Fuerteventura, La Gomera, Lanzarote, Tenerife, Tarragona, Mallorca, La Coruña, La Palma, Lugo, Orense y Pontevedra están en proceso. De Gerona,

En la consulta básica el buscador permite interrogar los datos delimitándolos por «Opciones de búsqueda» (exacta, aproximada y por expresión regular) y ordenar los resultados por provincia, frecuencia por provincia y frecuencia por enclave. Entre los filtros de búsqueda están el contexto de la búsqueda (entrevistas completas, informantes, hombres o mujeres), la provincia y la década de nacimiento del informante, las marcas (asentimiento, pausa, silencio, conversación cruzada, habla simultánea, literatura popular, nombre propio, gesto, risa, pronunciación, onomatopeya, ruido, chasquido, otras emisiones) y los temas (por orden alfabético: agricultura; alimentación; animales domésticos; bodas y noviazgos; caza y pesca; construcción y casas; costumbres y tradiciones; economía; educación y escuela; ejército y servicio militar; familia; fiestas populares; ganadería; industria básica; labores del hogar; matanza del cerdo; oficios; sanidad y salud; vida religiosa). También es posible el cartografiado dinámico de los resultados obtenidos tras la consulta.

En la consulta avanzada (*cfr*. Imagen 1) la búsqueda (opcionalmente sensible a las mayúsculas) puede realizarse por forma ortográfica, lema y etiqueta gramatical (de acuerdo con el etiquetado proporcionado por FreeLing, http://nlp.lsi.upc.edu/freeling/node/1). Los filtros de búsqueda de la consulta básica se mantienen igual. En cuanto al uso de FreeLing, para los efectos prácticos del COSER se ha debido modificar el tokenizador (con etiquetas XML), se ha ampliado considerablemente su diccionario léxico y se han incluido las anotaciones de sufijos, terminaciones verbales y clíticos de las variedades dialectales (*cfr*. De Benito, Pueyo & Fernández-Ordóñez 2016 para los problemas de transcripción y lematización).

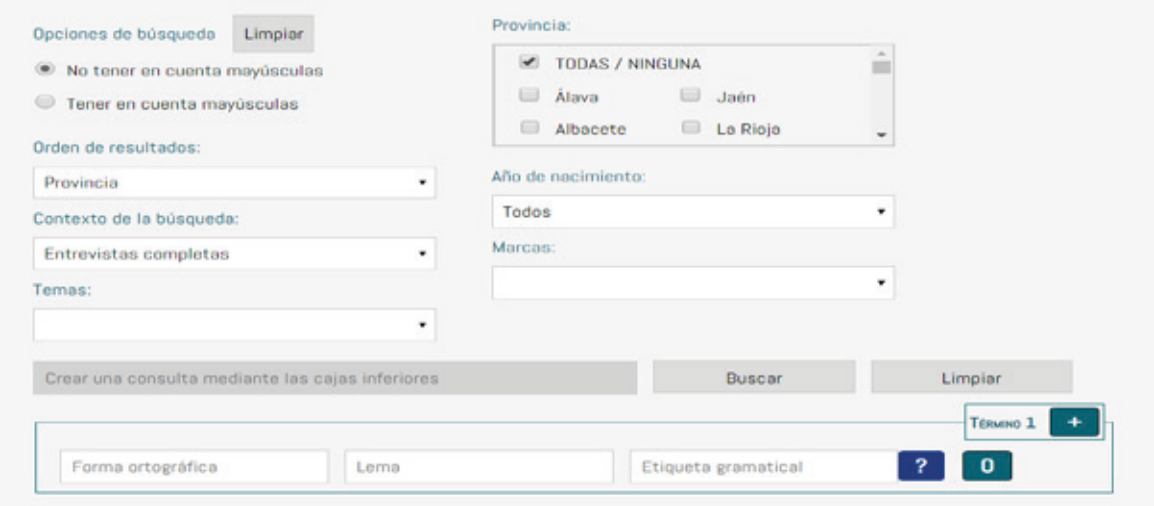


FIGURA 1. Consulta avanzada del COSER

Desde 2019 todos los resultados se pueden descargar en formato Excel, son imprimibles y el cartografiado se realiza de forma dinámica y automática según el interés y las necesidades del investigador (con posibilidad de eliminar del mapa los enclaves que no se correspondan con lo buscado). Las ventajas que ofrece el COSER son varias, ya que permite un acceso rápido y fácil a la información (además de gratuito) y la posibilidad de acceder a todo el soporte original (el archivo audio de cada entrevista, *cfr*. http://www.corpusrural.es/archivos.php). En un futuro se espera completar la alineación de texto y audio de todas las entrevistas disponibles, la mejora de la

Ibiza, Menorca y El Hierro todavía (enero de 2019) no hay materiales recolectados.

lematización y el etiquetado actual, así como la implementación de un etiquetado sintáctico basado en constituyentes o en dependencias[8]. La muestra de entrevistas disponibles, que hoy cubre toda la península ibérica, con la salvedad de Galicia y Cataluña, seguirá aumentando con el objeto de tener representación de esas áreas, así como de las islas Baleares y Canarias.

4. APORTACIONES BASADAS EN EL COSER

Como quedó señalado en (§ 1), el COSER es un corpus creado para el estudio de la variación que pueda encontrarse en el habla de los grupos socioculturales de menor educación del ámbito rural. El hecho de que la mayor parte de los corpus orales del español se nutran de fuentes como los medios de comunicación proporciona cierta singularidad al COSER y hace su aportación especialmente útil, pues facilita el estudio de las soluciones gramaticales no estándares, las cuales suelen ser sistemáticamente evitadas en la lengua escrita y en el habla de los grupos socioculturales de mayor educación. No son pocos los aspectos de variación gramatical que han sido estudiados a partir de los datos del COSER. Nos centraremos en los siguientes.

4.1. Morfología verbal dialectal

La morfología de la flexión verbal es un aspecto altamente sujeto a la variación dialectal que, sin embargo, cuenta con pocos estudios geográficos y teóricos, quizá porque generalmente no penetra en la lengua estándar y que, por ello, suele ser recesiva y difícil de documentar. Los atlas lingüísticos permiten hacerse una idea parcial de esta variación (*cfr*. Mondéjar (1970) a partir del ALEA, o Llorente (1965) y Buesa & Castañer (1994) sobre el ALEANR), amén de las monografías dialectales. Sin embargo, hay pocos estudios recientes dedicados a ella en castellano (frente al interés y publicaciones que suscita la morfología dialectal del verbo en catalán, por ejemplo). Los datos del COSER, comparados con otras fuentes, han permitido definir o confirmar las áreas de uso, atestiguar la vitalidad y explorar las raíces históricas de algunos aspectos.

4.1.1. Los pretéritos perfectos fuertes

La 3ª persona del plural con desinencia analógica *-n* en los pretéritos perfectos fuertes del tipo *dijon*, *puson*, *estuvon* (Pato 2004a, 2010b), propia del leonés centrooriental, el castellano occidental y el extremeño, se conserva hoy de manera restringida, especialmente en León, Zamora, Salamanca, Palencia y Cáceres, y en puntos aislados y limítrofes de Valladolid, Ávila y Badajoz. La comparación con los datos del ALPI y otras fuentes anteriores a la Guerra Civil permite observar que en la segunda mitad del siglo XX su uso ha retrocedido en esas últimas provincias, así como en Burgos y Segovia. En ninguna fuente se registran las formas analógicas al norte de la cordillera Cantábrica, sino en los territorios repoblados al sur de la misma, es decir, el área originaria del posterior reino de León desconoce estas formas de perfecto. Ello unido a la ausencia en la documentación medieval leonesa del siglo XIII de este tipo de perfectos induce a pensar que se trata de un aspecto potenciado en época posterior, quizá gracias a la nivelación lingüística que tuvo lugar como efecto de la repoblación. Las primeras documentaciones de *hizon* o *pudon* en la zona occidental proceden de la literatura pastoril de Lucas Fernández, de finales del siglo XV.

[8] Para ello contamos con el proyecto, dirigido por Miriam Bouzuita, «A Respeaking and Collaborative Game-Based Approach to Building a Parsed Corpus of European Spanish Dialects», financiado por el FWO (Medium-scale research infrastructure project: I000418N).

4.1.2. Los gerundios

Otra variante de flexión verbal estudiada gracias a los datos del COSER han sido los gerundios "analógicos" (del tipo *supiendo*, *hiciendo, tuviendo*), formas que siguen vivas en algunas de las variedades peninsulares (Pato & O'Neill 2013). En el ALPI las formas *tuviendo* y *quisiendo* se registran especialmente en el área oriental, de norte a sur, desde Cantabria hasta Alicante, pasando por Burgos, La Rioja, Soria, Guadalajara y Cuenca, así como en Álava, sur de Navarra, Huesca, Zaragoza y Teruel. En el área occidental, se localizan de manera mucho más esporádica en diversos enclaves del sureste de León, centro de Palencia, Salamanca y norte de Cáceres. El COSER corrobora los datos del ALPI al mostrar que el área de conservación de estos gerundios es fundamentalmente oriental (registrados en las provincias de Álava, Soria, Guadalajara, Zaragoza y Teruel, con ejemplos también de Alicante y Castellón), siempre en informantes mayores. No obstante, el COSER también registra ejemplos en Cantabria y Palencia. Pato & O'Neill (2013) explican estas formas, en las que el tema de perfecto se extiende al gerundio, por condicionamiento fonológico del segmento temático [ˈje], común tanto al perfecto (y tiempos afines) como al gerundio, en lugar de a través de la analogía proporcional con el gerundio del verbo *venir (viniendo)*, tal y como se venía explicando anteriormente.

4.1.3. La acentuación

La variación acentual en formas del pretérito imperfecto de indicativo, subjuntivo y condicional (como *cantabámos* por *cantábamos*) (Pato 2012a) es otro de los rasgos estudiados. Estas formas que mantienen el acento latino en la 1ª y 2ª persona del plural han sido documentadas en el COSER para el pretérito imperfecto de indicativo en algunos enclaves de Burgos, La Rioja, Aragón (Huesca, Zaragoza y Teruel), Navarra, Álava, Guipúzcoa, Soria, Albacete, Almería, Castellón y Valencia, con algún ejemplo en Madrid, Toledo y León. En realidad, estas formas con acentuación etimológica se comportan como las del tema de presente (presente de indicativo y de subjuntivo), tiempos que no tienen acento columnar y que no mantienen la misma categoría morfológica acentuada en la 1ª y 2ª persona del plural. La conservación de la acentuación originaria, especialmente viva en Navarra y Aragón, puede deberse a que el castellano aragonés tiende a cambiar la acentuación esdrújula hacia formas paroxítonas (*sabádo* por *sábado*).

4.1.4. El pretérito imperfecto de indicativo

Además, los imperfectos de la 2ª y 3ª conjugación que mantienen o extienden analógicamente *-ba-* en su formación (*caíba, traíba*) también se documentan en el COSER en el área oriental del castellano (Pato 2018b). Tanto en la Península, Canarias como en América son más frecuentes con verbos cuya raíz finaliza en vocal y pueden por ello entenderse como una estrategia analógica, basada en el morfema *-ba* de la 1ª conjugación y apoyada por el imperfecto del verbo *ir* (*iba*), con la que se persigue deshacer el hiato originario. Su aparición independiente, en áreas no necesariamente vinculadas desde un punto de vista histórico, apunta a considerar este rasgo como un universal vernáculo (Chambers 2004).

4.1.5. La desinencia verbal -éis > -ís

Por último, la evolución de las desinencias de 2ª persona plural en *-éis* > *-ís* que documentan los atlas lingüísticos desde los años 30 del siglo XX se atestigua también en

los datos del COSER en toda el área del castellano oriental (Del Barrio 2018) y, además, ha hecho posible su cuantificación. De acuerdo con ella el foco de esta innovación o de su pervivencia parece localizarse en la provincia de Zaragoza, desde donde se extiende a las vecinas de Huesca, Teruel, Guadalajara, Soria y sur de Navarra, con presencia más reducida en Cuenca, Albacete, Ciudad Real, Toledo, Jaén y Almería. El análisis de los datos permite constatar que la reducción se da tanto en los verbos auxiliares (*habís*) y modales (*querís, podís*) como en los léxicos (*sabís, hacís, vis*). Asimismo, el tiempo verbal en que tuvo lugar la reducción primero es el presente de indicativo, común a todas las áreas. Desde ese paradigma la desinencia se extendió al presente de subjuntivo (*cantís*) y al futuro de indicativo (*cantarís*), extensión exclusiva de las zonas de mayor frecuencia.

Aunque todas estas formas de la morfología del verbo son recesivas en el español peninsular, los datos del COSER dan fe de su pervivencia en el ámbito rural a finales del siglo XX y principios del siglo XXI.

4.2. Concordancia de caso en los clíticos de objeto

4.2.1. Leísmo, laísmo y loísmo

La escasa tolerancia que la lengua estándar muestra hacia la variación gramatical hace que numerosos fenómenos dialectales se hayan visto sometidos con cierta frecuencia a un filtrado sociolingüístico, hecho que puede distorsionar los principios lingüísticos que explican su funcionamiento originario. Este ha sido el caso de los usos "anómalos" de los pronombres átonos, esto es del leísmo, el laísmo y el loísmo. Gracias a las entrevistas sociolingüísticas de Klein-Andreu (2000) y los datos del COSER (*cfr*. Fernández-Ordóñez 1994, 1999) ha podido demostrarse que lo que los gramáticos percibían como usos desviados del empleo pronominal general son, en realidad, muestras parciales de paradigmas pronominales alternativos que se manifiestan plenamente en el habla de los grupos socioculturales bajos y en donde la selección del pronombre se realiza de acuerdo con principios lingüísticos diversos a los vigentes para el español estándar. Es decir, las hipótesis explicativas se habían creado únicamente sobre los datos que ofrecía la lengua escrita culta, hecho que dificultaba su correcta interpretación.

Si bien es cierto que en las descripciones previas se había percibido la conexión entre el leísmo, el laísmo y el loísmo (*cfr*., ente otros, Lapesa 1968), estos usos no se sometían fácilmente a una única explicación. Por un lado, la tendencia a distinguir los objetos directos personales (con *le* y leísmo personal) de los no personales (con *lo*, *la*, sin leísmo) explicaba el leísmo personal, pero dejaba sin aclarar los motivos por los que esta confusión afecta fundamentalmente a los objetos masculinos ni las razones por las que puede acompañarse de leísmo no personal (o de cosa), laísmo y loísmo. Por otro lado, la tendencia a asignar los pronombres recurriendo al género de su antecedente, sin considerar su función sintáctica, explicaba el leísmo masculino (personal y no personal) y el laísmo, pero no por qué el leísmo no acababa de establecerse plenamente para todo tipo de objetos masculinos, siendo siempre más frecuente cuando el antecedente era un objeto personal. Tampoco explicaba que el leísmo no se estableciera plenamente en el plural, donde además contendía con el loísmo.

A este respecto, la aportación del COSER ha sido fundamental para aclarar este aspecto mal comprendido de la sintaxis del español (*cfr*. Fernández-Ordóñez 1993, 1994 y 1999, Fernández-Ordóñez 2001 desde el punto de vista histórico, y Fernández-Ordóñez 2012 para la gestación de las áreas dialectales). Se han acotado geográficamente las áreas que presentan leísmo, laísmo y loísmo, al tiempo que se ha demostrado que la aparente

falta de coherencia en los usos documentados responde a la existencia de varios paradigmas pronominales dialectales, alternativos al general del español[9], empleos que se mezclaban en los estudios previos distorsionando su interpretación[10]. Con todo, además de soluciones de transición, existen tres grandes paradigmas (de los que solo se reseñan abajo los empleos discrepantes del general):

a) El del romance hablado en contacto con el vasco, caracterizado por presentar leísmo asociado a objetos animados, masculinos y femeninos (*A Jon/ A Miren le veo*), y pronombres nulos para los objetos inanimados (*La bicicleta Ø tengo ahí*).

b) El cántabro, caracterizado por presentar leísmo con los objetos contables masculinos en singular (*A Juan le veo*; *El libro le tengo*) y reservar *lo* para los objetos no contables, masculinos y femeninos (*El vino/ La leche lo bebo*).

c) El castellano, también llamado paradigma referencial, caracterizado por haber eliminado las distinciones de caso entre objetos directos e indirectos. Los pronombres se asignan de acuerdo a su género, número e interpretación ±contable. Así, hay leísmo para contables masculinos en singular (*A Juan le veo*; *El libro le tengo*) y, en la mitad norte del área, también en plural (*Los niños les veo*; *Los libros les tengo*); *lo* para objetos no contables, masculinos y femeninos (*El vino/ La leche lo bebo*), también cuando son indirectos (*A la leche lo echan de todo*); laísmo para los objetos indirectos femeninos (*La doy un libro*; *Las gusta el cine*); y loísmo para los objetos indirectos plurales en la mitad sur del área (*A los niños los gusta jugar*).

Mientras que el paradigma vasco representa la materialización de la hipótesis del leísmo como extensión del dativo a los objetos personales, tanto masculinos como femeninos, los paradigmas cántabro y castellano se caracterizan por basar la selección pronominal en la categorización semántica del antecedente como contable o no contable, categoría lingüística que no se había considerado anteriormente y que explica el hecho de que el leísmo fuera universal con los antecedentes personales masculinos (siempre contables y referidos por *le*) pero no se generalizara con los no personales, ya que estos pueden ser contables (referidos por *le*) o no contables (referidos por *lo*). El paradigma castellano, a su vez, se distingue del cántabro por haber eliminado la categoría de caso, generalizando *la(s)* y *lo* como pronombres de dativo. Además, en el plural masculino presenta dos soluciones diferentes según el pronombre preferido: a) *les* al norte (en el noroccidente de Burgos, Palencia y Valladolid); y b) *los* al sur (en el este de Salamanca y Cáceres, Ávila, oeste de Toledo y norte de Madrid). Los territorios en que se emplea este

[9] El paradigma del español estándar (y el hablado en la mayor parte de variedades americanas) emplea *lo(s)* para OODD masculinos (*A Juan lo veo*; *El Libro lo tengo*) y neutro (*Eso no lo sé*), *la(s)* para OODD femeninos (*A María la veo*; *La caja la tengo*) y *le(s)* para OOII (*A María / A Juan le regalé un libro*; *A eso no le doy importancia*).

[10] Los diferentes tipos de leísmo tradicionalmente descritos son: a) OD masculino de persona singular *(¿Conoces a Juan? Sí, le conozco hace tiempo*); b) masculino de cosa singular *(¿Sabes dónde está mi libro? No, no le he visto por aquí*); c) de persona plural (*Esta tarde voy a recoger a los niños del colegio y les llevaré al parque*) y de cosa plural (*Fui a buscar los discos que querías y les encontré en la tienda de abajo*); d) femenino de persona singular (*A María hace tiempo que no le veo*) y plural (*Aquí no hay monjas. En la guerra les mataron a todas*). El laísmo de persona puede ser singular y plural (*Cuando vi a Pepa, la di su regalo; A las niñas de hoy ya no las gusta coser*) y de cosa (*Coges la sartén, la das la vuelta y ya tienes lista la tortilla*). Y el loísmo de persona plural (*Cuando recojo a los niños del colegio, los llevo la merienda*) y de cosa singular y plural (*Cuando el arroz está cocido, lo echas la sal*; *Para arreglar esos trajes, hay que sacarlos el bajo*), y de objeto neutro (*Yo no lo doy ninguna importancia a eso*).

sistema son, pues, los del centro y occidente de Castilla, desde el sur de la cordillera cantábrica hasta La Mancha.

Esta perspectiva general ha sido algo matizada en un análisis detallado de los pronombres átonos en la provincia de Álava (Camus Bergareche y Gómez Seibane 2015). Los datos del COSER han demostrado que el sistema propiamente vasco se da sobre todo al norte de la provincia, en contacto con el euskera, mientras que el área occidental se aproxima más al uso castellano y la oriental al uso navarro-riojano.

4.2.2. El doblado de objetos directos

Otro de los fenómenos relacionados con los pronombres átonos ha sido el estudio de la duplicación de objetos directos posverbales en el español en contacto con el vasco (*Le veo a Juan*) (Gómez Seibane 2017), teniendo en cuenta para ello variables sintácticas, semánticas y cognitivas. Sobre los datos del COSER de zona vasca, con un corpus de control del castellano central, se ha podido probar que el doblado de objetos directos posverbales es mucho más frecuente en la primera área que en la segunda (71, 2 % vs. 51,5 %). Además, el tipo de objeto doblado por el clítico es mayoritariamente pronominal (52,9 %) e inanimado (67,6 %) en zona castellana (*Lo teníamos que hacer todo*) mientras que es nominal (76,1 %) y humano (64,3 %) en zona vasca (*Tenía que ir a ayudarle al marido*). La mayor extensión del doblado al norte se revela también en que los antecedentes del clítico son, desde un punto de vista cognitivo, objetos semiactivos e inactivos, mientras que los del corpus de control son mayoritariamente activos y semiactivos[11].

4.2.3. La pronominalización con el verbo haber

Por último, los datos del COSER han sido la fuente para estudiar la posible pronominalización del argumento que acompaña al verbo *haber* existencial, siempre de carácter indefinido e inespecífico (*Antes había médico pero ahora no lo hay*) (De Benito 2016). En primer lugar, se ha precisado el área geográfica que acepta el pronombre, no restringida al noroeste peninsular, tal y como se creía, sino que se extiende desde Asturias y Cantabria por todo el occidente y centro de la Península hacia el sur, dejando fuera la Castilla oriental y el este peninsular. Mientras que en las variedades centrooccidentales se acepta el pronombre en el 43,4 % de los casos, en las orientales ello solo sucede en el 6,7 %. Dos de los factores que influyen en esta pronominalización son la animación del argumento y el tipo de construcción existencial. De hecho, los objetos animados aparecen referidos por un clítico con mayor frecuencia que los inanimados (40,3 % vs. 31,7 %) y la aparición de los pronombres átonos es más común en las construcciones adscriptivas (en las que el argumento se acompaña de una predicación secundaria) (*Sillas las hay pequeñas y grandes*) (50 %) que en las de locación (*Sillas las hay*) (26,5 %). Por otro lado, de gran importancia teórica es la documentación esporádica de ejemplos de *haber* concordado en plural con el argumento pronominalizado (*También las habían otras*), que vienen a sostener la idea de que el argumento único de *haber* presenta características mixtas de objeto y de sujeto (*cfr. infra*, 4.4.1).

[11] Los activos han sido mencionados de forma inmediata en la interacción, los semiactivos han sido mencionados con cierta distancia, pero son identificables a través de alguna vinculación textual o inferencial, y los inactivos son completamente nuevos y no son accesibles textual e inferencialmente.

4.3. Concordancia de género

4.3.1. El neutro de materia

En algunos de los casos que hemos revisado anteriormente los datos del COSER han hecho posible estudiar y entender mejor fenómenos dialectales que ya eran conocidos. El interés del COSER resulta mayor si atendemos al hecho de que ha permitido registrar fenómenos completamente ignorados en las gramáticas y monografías dialectales hasta ahora, como el *neutro de materia*, fenómeno que se suponía exclusivo del asturiano centrooriental y de algunos puntos de Cantabria. Sin embargo, esta concordancia se registra también en todo el centro y occidente de Castilla hasta Toledo (Fernández-Ordóñez 2006, 2007a, 2007b, 2009b, 2019).

Por *neutro de materia* se entiende tradicionalmente la expresión morfológica, en los pronombres y adjetivos concordantes, de la categorización del nombre como discontinuo (contable) o como continuo (no contable)[12]. La concordancia está restringida a pronombres y adjetivos, y no se da en el nombre, los determinantes, cuantificadores o adjetivos prenominales (*el/ este/ mucho/ buen pelo*; *la/ esta/ mucha/ buena lana*, pero **lo/ esto/ bueno/ mucho pelo/ lana*, a pesar del carácter no contable del nombre). Cuando el nombre masculino o femenino se interpreta como no contable, los pronombres personales y demostrativos, el artículo con núcleo elíptico y los adjetivos que siguen al nombre, sean modificadores adyacentes o de carácter predicativo, emplean una morfología diversa que los masculinos y femeninos contables. Por ejemplo, *La lana blanco está sucio vs. La manta blanca está sucia*; *Lo blanco está sucio vs. La blanca está sucia*; *La lana blanco lo lavo vs. La manta blanca la lavo*[13].

En los adjetivos, la expresión de la concordancia de materia presenta menor vitalidad que en los pronombres. Desde el punto de vista sintáctico, está restringida a los adjetivos postnominales y a aquellos situados en posiciones predicativas. La variable distribución geográfica y la cuantificación de la concordancia en los adjetivos referidos a nombres no contables femeninos ha permitido concluir que se manifiestan con intensidad diversa según la clase de palabra y la posición sintáctica: pronombre clítico de OD (81,4 %) > predicativo (59,2 %) > atributo con *estar* (54,9 %) > atributo con *ser* (31,3 %) > adjetivo modificador (13 %)[14].

Así, la concordancia es regular en los pronombres de objeto y bastante elevada en los predicados secundarios (*Esa lana de las ovejas se lava bien lavao*) y en los atributos con la cópula *estar* (*Cuando estaba la leña bien quemadito ya, pues lo dejabas, lo esparcías bien por todo el horno*). Estos dos tipos de predicados tienen en común que suelen denotar estados acotados o propiedades de estadio o episódicas. En cambio, la concordancia neutra desciende drásticamente con los atributos introducidos por la cópula *ser* (*La leche de cabra es muy bueno, mejor que lo de vaca*) y con los adjetivos que modifican directamente al nombre (*Si lo coge el agua, si lo siegas y te lo coge el agua, ye una hierba blanquino que…*), quizá porque suelen denotar propiedades individuales.

[12] Este conjunto de concordancias no debe confundirse con la existencia de un tercer género neutro, ya que todos los nombres que reciben interpretación continua son, desde el punto de vista léxico, masculinos o femeninos. Se trata de una concordancia basada en rasgos de carácter semántico, no léxico, y equiparable a otras concordancias *ad sensum*.

[13] Las únicas diferencias morfológicas en el área son las siguientes: en los clíticos de OD se emplea *lu* en Asturias y *le* en Cantabria y Castilla para los contables masculinos; en los adjetivos masculinos los contables acaban en *-u* y los no contables en *-o* en Asturias (*el coche rotu* vs. *el vino frío*) mientras que en Cantabria y Castilla todos terminan en *-o* (*el coche roto*, *el vino frío*).

[14] Únicamente en Asturias la concordancia ha llegado a afectar al adjetivo postnominal y en alguna de las variedades asturianas puede manifestarse, si bien de forma irregular, en la morfología del nombre (Lena, Quirós).

La concordancia de materia propia de los dialectos peninsulares refleja el patrón de la jerarquía universal de la concordancia propuesta por Corbett (2006): *modificador* < *predicado* < *pronombre relativo* < *pronombre personal*. De hecho, el examen detallado de los datos dialectales de la concordancia de materia permite refinar la jerarquía propuesta, distinguiendo dentro de la posición *predicado* entre predicados de propiedades individuales y de propiedades episódicas, siendo este segundo grupo el que expresa antes la concordancia de base semántica: *adjetivo modificador* < *predicados individuales* < *predicados episódicos* < *relativo* < *pronombre*.

4.3.2. El sincretismo en los cuantificadores evaluativos

El COSER también ha servido para documentar formas de difícil registro y aumentar la casuística de los ejemplos. Es el caso del sincretismo femenino en los cuantificadores evaluativos del oriente cántabro, del tipo *mucha trabajo*, que suele aparecer con nombres de materia (*mucha vientu*) y con nombres abstractos e infinitivos (*poca talentu*; *tanta trabajar*) (Fernández-Ordóñez 2015). Esta extraña concordancia se documenta, según los datos del COSER, en un área muy reducida del este cántabro (Vega de Pas, Arredondo, Rubalcaba, Quintana de los Prados (Cantabria) y Orbaneja del Castillo (Burgos)). El sincretismo de género –y número– se registra en cuantificadores universales en no pocas variedades románicas y, aunque más raramente, también en los evaluativos, pero siempre a favor del masculino. La existencia de estos cuantificadores cántabros plantea la cuestión de si podrían representar restos del neutro plural latino como los que se documentan en napolitano antiguo (*tanta angustie* 'tanta angustias'; *quanta homini* 'cuanta hombres').

4.4. Concordancia de número

Uno de los aspectos de variación gramatical más frecuentes en las lenguas es la variable concordancia de número vinculada al reanálisis sintáctico de algunos constituyentes. Los datos del COSER han hecho posible estudiar algunos ejemplos, ya conocidos, pero ni cuya distribución geográfica en la Península ni sus factores condicionantes estaban bien definidos.

4.4.1. La concordancia de haber *existencial en 3ª persona*

La construcción impersonal existencial del verbo *haber* con un único argumento como OD tiende a reanalizarse en español de forma que ese argumento se interpreta como sujeto e induce la concordancia plural en el verbo (*Había casas* > *Habían casas*). Esta concordancia es muy frecuente en América, Canarias y en el español en contacto con el catalán, mientras que en el resto de la península ibérica se considera propia del habla popular o poco prestigiosa, sin haber sido objeto de un estudio de conjunto. Los datos del COSER han permitido demostrar la amplia difusión geográfica peninsular de las formas pluralizadas, que había escapado, por su baja frecuencia, a los atlas (Pato 2016, Bouzouita & Pato 2019). Si bien la concordancia alcanza su mayor frecuencia en el área valenciano-catalana (65 %), también existe en Murcia (46 %) y, esporádicamente, en porcentajes inferiores al 3 %, en otras comunidades (Cantabria, Extremadura, Castilla-La Mancha, Andalucía, Aragón, Castilla y León y el País Vasco). El contacto lingüístico con el valenciano favorece las formas plurales. También es frecuente en las construcciones perifrásticas (*Pueden haber problemas*). Aunque se ha observado para América que el rasgo semántico [+ humano] del argumento propicia la concordancia (*Han habido chicos que…*), tal diferencia no es tan clara en el español europeo. La importancia de la semántica se observa en que la concordancia puede darse incluso con

relativos (*Habían quien iban*) o nombres colectivos (*Habían gente*) [+ humanos] de semántica plural y número singular.

4.4.2. La concordancia de haber *existencial en 1ª persona plural*

Muy relacionada con la concordancia en 3ª persona plural de las construcciones existenciales figura aquella en que el verbo *haber* concuerda en 1ª personal plural con su argumento con lectura inclusiva del hablante (*Los que habemos no cabemos en el mundo*) (Castillo Lluch & Octavio de Toledo 2016). Mientras que es muy frecuente en América, en España se daba como típica del habla popular, sin más precisiones. Los atlas lingüísticos apenas documentan esta concordancia, por lo que los datos del COSER han hecho posible reconstruir por vez primera su distribución geográfica peninsular y atestiguar su empleo en distintos tiempos (*habíamos, habremos, habríamos, hemos habido*). La concordancia aparece sobre todo en las áreas periféricas de la Castilla central, desde León a Andalucía occidental, pasando por Extremadura, y desde Huesca a Andalucía oriental y Murcia, sin olvidar las zonas castellanohablantes valencianas. El rasgo semántico [+ humano] del argumento es obligado en la construcción que comenzó a ser plenamente existencial solo una vez que *avemos* perdió del todo su empleo como verbo posesivo, desde mediados del siglo XVII. La distribución geográfica de *habemos* existencial parece tener que ver con la pervivencia, en los mismos territorios y hasta el siglo XIX, de *habemos* auxiliar, mientras que en Castilla ya en el siglo XVI *habemos* había sido reemplazado por *hemos*.

4.4.3. La concordancia en los infinitivos

Otro ejemplo de expresión variable de la concordancia de número lo proporcionan los infinitivos (o gerundios) de verbos o construcciones reflexivas de sujeto plural (*Tenían que irsen para | porque no estuvieran las chicas por los pisos metiéndosen con hombres*) (Heap & Pato 2012, con datos no completos). Esta concordancia en las formas no personales de los verbos reflexivos es un rasgo propio de la zona oriental (navarro-aragonesa y castellana), según los datos del ALPI que corrobora el COSER; no es, por tanto, un rasgo general al español de las clases bajas, contra lo que se había creído. Aunque actualmente predominan las formas en *-se*, los datos del COSER han permitido demostrar que la concordancia afecta a todo tipo de verbos y construcciones reflexivas (*cfr. infra*, 4.5): los reflexivos directos, los únicos acreditados por el ALPI (*No quieren abrigarsen*), los reflexivos indirectos (*Hacersen unas mangas*), los recíprocos (*Casarsen no se casaron*), los anticausativos (*Los chorizos pa secarsen*), los conversivos (*Tienen que venir a preocuparsen un poco*) e incluso los verbos intransitivos (*Salirsen de casa*) y transitivos (*Empiezan a comprarsen cosas*) sin cambio valencial, además de los no reversibles (*No vinieron más que a pitorrearsen de nosotros*). Además, la concordancia se extiende a la pasiva refleja, aunque el argumento no sea específico (*Ya echaron a formarsen sindicatos*).

4.5. Construcciones reflexivas, medias e impersonales con *se*

Las grabaciones del COSER son la base de uno de los estudios más completos y detallados efectuados hasta el momento sobre las construcciones reflexivas, mediopasivas e impersonales con *se* en español actual (De Benito 2012, 2013, 2015a, 2015b). La novedad reside en que, por vez primera, hay una medición cuantitativa de la presencia o ausencia del reflexivo en los distintos tipos de construcciones que pueden aceptarla. La cuantificación permite hipotetizar sobre las vías de extensión del reflexivo, desde su locus original, las construcciones reflexivas, a los demás tipos y medir, además,

si esa extensión fue condicionada por parámetros sintácticos o más bien semánticos (De Benito 2015b).

4.5.1. Las construcciones reflexivas y recíprocas

El estudio ha permitido determinar el comportamiento sintáctico diferenciado de las construcciones reflexivas y recíprocas en atención a su contenido semántico. La presencia del reflexivo y de su refuerzo enfático *a sí mismo*, aunque posible en toda construcción definida como reflexiva, no es idéntica en términos porcentuales. El refuerzo se da más con aquellos verbos que pueden alternar entre una construcción reflexiva o no (*Juan se mira vs. Juan mira la tele*) o reflexivos 'puros'. En cambio, los reflexivos 'naturales', aquellos en que la acción se dirige prototípicamente a uno mismo, como los de cuidado personal, aceptan con mucha menor frecuencia el refuerzo reflexivo *a sí mismo* (*Juan se peina / se lava / se ducha*). Esta diferencia semántica con consecuencias sintácticas tiene su reflejo en la distribución geográfica: en el área noroccidental el reflexivo puede faltar en los reflexivos naturales (*Juan peina*) pero siempre está presente en los puros (*Juan se mira*).

También la cuantificación del uso ha permitido captar sutiles diferencias sintácticas entre los distintos papeles semánticos del reflexivo indirecto. El reflexivo indirecto puede tener distintos roles: destinatario (*Me envié el paquete*), benefactivo (*Me he conseguido una entrada*) y posesivo (*Me he roto la pierna*; *Me pongo el abrigo en los hombros*). Mientras que el destinatario puede ser el sujeto u otro argumento (*Le envié el paquete*), los reflexivos posesivos indican prototípicamente que el OD o el complemento locativo es parte del sujeto. Por ello, es más fácil su ausencia, tal como sucede en las variedades noroccidentales (*Rompí la pierna*; *Pongo el sombrero*). En cambio, los reflexivos destinatarios no solo son comunes a todas las áreas sino que son los que aceptan más el refuerzo enfático *a sí mismo* (*Me envié el paquete a mí mismo* vs. *?Me he roto la pierna a mí mismo*).

Las construcciones recíprocas también son sensibles a la interpretación semántica del verbo. Los verbos naturalmente recíprocos son generalmente simétricos (*Juan y María se encontraron*) y suelen presentar con menor frecuencia el refuerzo *uno al / con el otro, entre ellos / sí*, en contraste con lo que ocurre con los recíprocos 'puros', generalmente no simétricos (*Juan y María se miran*), que, por ello, presentan en mayor porcentaje el refuerzo (*Juan y María se miran el uno al otro*). Otra diferencia sintáctica entre estos dos tipos de verbos recíprocos es el hecho de que la construcción llamada "discontinua" (que presenta el sujeto en singular y codifica el otro miembro de la acción recíproca con un sintagma preposicional: *Juan se encontró con María*) es propia de los verbos naturalmente recíprocos. Tal como sucedía con los reflexivos indirectos, los recíprocos indirectos, generalmente dativos poesivos, son mucho menos frecuentes en el área noroccidental (*Juan y María dieron la mano*).

4.5.2. Las construcciones 'medias'

La puesta a prueba con un análisis estadístico de las hipótesis manejadas para explicar los distintos tipos de *se* ha arrojado importantes conclusiones. Hay acuerdo en que el reflexivo sigue un camino de reanálisis sintáctico desde el punto de vista tipológico que obedece a este patrón: *reflexivo* > *recíproco* > *medio* > *pasivo* > *impersonal*. Sin embargo, dentro de los usos medios caben muchas configuraciones sintácticas y es lícito preguntarse por los mecanismos precisos y construcciones concretas en que tuvo lugar esa reinterpretación. Los usos llamados 'medios' se documentan en los siguientes tipos de alternancias diatéticas:

a) La diátesis anticausativa, en que se elimina el agente y el paciente se promueve a la posición de sujeto (*Juan rompe la mesa* > *La mesa se rompe*).

b) La diátesis conversiva, en que la causa expresada por el sujeto se degrada a complemento de régimen y el experimentante objeto se promueve a sujeto (*La noticia alegró a María* > *María se alegró de la noticia*).

c) La diátesis deobjetiva, que implica la eliminación del objeto, generalmente abstracto y que adquiere una lectura genérica (*Juan explicó (algo)* > *Juan se explicó*).

d) La diátesis antipasiva, que consiste en la supresión del objeto, el cual, a veces, puede expresarse como un complemento de régimen (*Juan aprovechó tu experiencia* > *Juan se aprovechó (de tu experiencia)*).

Además, el reflexivo se ha extendido a otros tipos de verbos que no manifiestan cambio valencial, tanto intransitivos (*ir* > *irse*) como transitivos (*comer algo* > *comerse algo*). Las diferencias porcentuales de expresión del reflexivo en cada tipo y en la geografía han permitido concluir que el camino de difusión fue el siguiente. La extensión del reflexivo comenzó con la diátesis conversiva, con uso casi categórico del reflexivo en todas las zonas y sujetos siempre animados (*María se asusta*), tal como las construcciones reflexivas. De ahí se extendió a la deobjetiva (*María se aguanta*) y a la diátesis anticausativa con sujeto animado (*María se levanta*). La difusión alcanza más tarde a la antipasiva (*Juan se olvidó*) y a la diátesis anticausativa cuando presenta un sujeto inanimado (*La ropa (se) seca*). Este itinerario se deduce de las frecuencias globales del empleo del reflexivo por tipo diatético: *conversiva* (99 %) > *deobjetiva* (87,2 %) > *anticausativa con sujeto animado* (80 %) > *antipasiva* (76,9 %) > *anticausativa con sujeto inanimado* (71,6 %).

Todas las configuraciones diatéticas con frecuencias más altas del reflexivo coinciden en la animación del sujeto, lo que las conecta con la diátesis originaria, la reflexiva. Por otro lado, el análisis de los datos permite configurar dos áreas dialectales: la noroccidental, en que la diátesis anticausativa presenta los índices de frecuencia más bajos, tanto para sujetos animados (56,33 %) como inanimados (42,3 %), y el resto del territorio, en que esos porcentajes se elevan notablemente (97,1 % y 75,5 %). Así, resulta evidente que el reflexivo es originariamente un procedimiento sintáctico intransitivizador del verbo que indica el cambio diatético.

4.5.3. El reflexivo sin cambio valencial

La difusión del reflexivo a los verbos sin cambio valencial es necesariamente posterior. A diferencia de la gran homogeneidad en el comportamiento sintáctico de los verbos anteriores, la dispersión léxica en el empleo del reflexivo es enorme. Incluso en los verbos que aceptan el reflexivo, su uso no supera el 25,1 % en los intransitivos (*salirse, irse, venirse, estarse*) y el 11 % en los transitivos (*comerse, saberse algo*). De nuevo, los sujetos animados favorecen la presencia del reflexivo (28 %) frente a los inanimados (15,8 %), al menos en los intransitivos, y el área noroccidental se resiste a aceptar estos usos más que el resto del territorio peninsular.

La extensión tardía del reflexivo a los verbos intransitivos sin cambio valencial puede explicarse como un proceso analógico basado en la semejanza de estos verbos con el contenido prototípico de los que experimentan esos cambios diatéticos (movimiento,

irse < *levantarse,* o emoción, *reírse* < *alegrarse*). Por lo general, el reflexivo indica un aumento de la agentividad o de la afectación del sujeto.

Necesariamente posterior debe haber sido su extensión a los verbos transitivos sin cambio valencial, en que el reflexivo parece dar cuerpo a un dativo benefactivo (*comerse algo, saberse algo, temerse algo*) que funciona como adjunto. Estos dativos también enfatizan la agentividad o la afectación del sujeto en el evento denotado. En realidad, apenas se distinguen del reflexivo que aparece en las construcciones reflexivas o recíprocas indirectas con un dativo benefactivo no argumental (*hacerse algo*), el cual, como sabemos, apenas aparece en el noroeste y que, por ello, se conforma como la fuente potencial de la analogía.

Además de reconstruir el camino por el que avanzó la reinterpretación del reflexivo mediante la cuantificación de su presencia en la estructura lingüística y en la geografía, otra importante conclusión teórica de este estudio es poner en tela de juicio la función supuestamente aspectual del reflexivo en los verbos sin cambio valencial. Para discutir esa interpretación, son pertinentes dos aspectos: por un lado, la ausencia de conexión diacrónica entre los empleos del reflexivo como indicador de un cambio diatético y su supuesta función de marcador aspectual de eventos culminados o logros; por otro, la gran cantidad de ejemplos en que el reflexivo no se relaciona con ningún cambio tal en la interpretación aspectual del evento (por ejemplo, *El agua se está saliendo*; *Me estoy terminando el café*; *Ayer me leía el periódico en un bar*).

4.5.4. Áreas dialectales vinculadas a los verbos reflexivos

Desde el punto de vista areal, este estudio identifica por vez primera una zona noroccidental en que los empleos del reflexivo están menos extendidos en general y, en concreto, en la diátesis anticausativa, en los reflexivos indirectos posesivos y benefactivos así como en los verbos intransitivos sin cambio valencial. Además, también se descubre y describe un área oriental, desde Castellón hasta Murcia, en que tiene lugar un sincretismo en el paradigma morfológico de los reflexivos en el plural a favor de *se* (*se casamos, se vais*) (De Benito, 2015a). Los datos muestran que el sincretismo se produce de forma más regular y geográficamente extendida en la 2ª persona de plural (*se queréis*) y que se ve favorecido por el uso del infinitivo en lugar del imperativo plural (*¡Sentarse de una vez!*). Aunque el fenómeno se da también en catalán, incluido el dialecto valenciano, las causas del cambio en esa lengua parecen fonéticas. En el castellano el contacto puede haber sido una motivación concomitante, pero también se ajusta al sincretismo derivado del empleo del infinitivo, extendido por toda Andalucía y claramente independiente del contacto bilingüe con el catalán. A estas áreas vinculadas a los reflexivos debe sumarse la oriental que expresa la concordancia plural del sujeto de infinitivos y gerundios mediante el reflexivo *-sen* (*cfr. supra*, 4.4.3).

4.5.5. Las construcciones impersonales reflejas

El *se* impersonal parece el último desarrollo de la extensión del reflexivo (*Aquí se vive bien*). Tradicionalmente se estima como prueba de la configuración impersonal su empleo con verbos intransitivos o, si el verbo es transitivo, que acepte la pronominalización del paciente siempre que sea humano y definido, esto es, vinculada al MDO (Marcado diferencial de Objeto) (*Se ve a Juan* > *Se le ve*). Ciertos dialectos aceptan, sin embargo, una pronominalización ampliada (De Benito 2012, 2013) en la que es posible referir con clítico a pacientes no humanos. Sobre los datos del COSER se demuestra que esa pronominalización se da exclusivamente en aquellas áreas en que los clíticos de 3ª persona tienden a eliminar las diferencias entre acusativo y dativo, la

referencial castellana, la cántabra y en el castellano hablado en contacto con el euskera. Por otro lado, gracias a la cuantificación de los datos se prueba que la pronominalización, nunca obligatoria, se extiende gradualmente siguiendo la Jerarquía de la animación: *humanos* (52, 5%) > *animados* (38, 8 %) > *inanimados contables* (12, 5 %) > *inanimados no contables* (6 %).

Otra importante aportación de este estudio ha sido descartar la hipótesis de que la pronominalización ampliada esté relacionada con un uso impersonal, esto es, no concordado del verbo con los objetos inanimados (*Se vende patatas ≠ Se las vende*). La desconexión entre los dos aspectos se prueba por la ausencia de coincidencia geográfica entre ambos. La falta de concordancia se da en áreas (como el Alto Aragón) en que la pronominalización ampliada es inexistente. Por otro lado, la topicalización del paciente definido no fuerza la pronominalización ampliada (solo el 4,2 % de los casos presentan ejemplos con copia pronominal como *Esa tela se la descose*), quizá porque en la mayoría de los casos se interpreta como sujeto de una pasiva refleja (*Esa tela se descose*).

El pronombre que figura en la impersonal refleja era originariamente un dativo, tanto para pacientes humanos masculinos y femeninos como para inanimados, y así continúa siendo en la mitad septentrional de América (*A María se le ve*; *La montaña se (le) ve*). En aquellas variedades peninsulares en que se han perdido total o parcialmente las distinciones de caso, se emplean, sin embargo, los pronombres de acuerdo con el paradigma pronominal en uso en cada variedad. En el País Vasco, solo aparece con *le*, y en el área referencial, los pronombres *le(s), la(s), lo(s)* seleccionados de acuerdo con los parámetros de género, número y ±contable que determinan su uso general en ese dialecto, si bien siempre pervive el *le* originario, compitiendo en minoría. Ello abre la posibilidad de que esté teniendo lugar una transitivización plena de la construcción en esos dialectos, aunque fuera originalmente intransitiva.

4.6. Variación modal

El COSER ha sido la base documental principal de estudios dedicados al desplazamiento del subjuntivo por el indicativo en la Península, en concreto, el empleo del condicional (o el imperfecto de indicativo) en lugar del pretérito imperfecto de subjuntivo en cualquier tipo de oraciones subordinadas (*Las costillas y todas esas cosas se metían en ollas para que se conservarían / conservaban*) (Pato 2003, 2004b, 2012c). El contraste entre los datos de los atlas y los datos del COSER confirma, en términos generales, la coincidencia en el área geográfica en que tiene lugar el fenómeno, pero resalta la incapacidad de los atlas en reflejar la variante minoritaria, el imperfecto de indicativo, y en detectar la existencia del mismo desplazamiento en las correspondientes formas compuestas del verbo. La ventaja del COSER es el elevado número de casos obtenidos (3588), lo que permite detectar la existencia de variantes minoritarias y cuantificar los datos para matizar su distribución. Así se ha podido identificar el área focal de este empleo (en la Castilla Vieja, el sur de Cantabria, el occidente de Vizcaya, el norte de Palencia, la Navarra media y las provincias de Álava y La Rioja) y las áreas de transición (en el norte de Cantabria, el oriente de Vizcaya, Guipúzcoa, la Navarra septentrional y meridional, el sur de Palencia, el este de León y el norte de Valladolid, Segovia y Soria). En las áreas en que tiene lugar este desplazamiento el condicional es la forma preferida (96 %) frente al imperfecto de indicativo (4 %) y es mucho más frecuente en los tiempos simples (61, 9 %) que en los compuestos (21, 6 %).

Por otro lado, la cuantificación ha hecho posible aplicar pruebas estadísticas como la regresión logística para valorar la influencia simultánea de varias variables sobre la manifestación del fenómeno (*cfr*. Pato 2012c para una visión de conjunto). La presencia del condicional en la prótasis de las oraciones condicionales (*Si tuviera / tendría dinero,*

lo compraría) se había reconocido siempre como un uso dialectal estereotipado del habla de los vascos. Pero, en realidad, en el castellano septentrional la prótasis condicional no constituye el contexto originario de este fenómeno, sino el más visible a los gramáticos y dialectólogos. Gracias al análisis estadístico, se ha podido demostrar que la extensión de las formas indicativas en *-ría* y *-ba* a costa de las formas subjuntivas en *-ra* y *-se* tuvo lugar primero en las oraciones completivas, extendiéndose luego a las relativas y adverbiales relacionadas (modo, lugar y tiempo), y más tarde a las condicionales, para alcanzar finalmente al resto de contextos sintácticos (Pato 2003, 2004b): *completivas* (72,1 %) > *relativas* (61,7 %) > *condicionales / finales* (57, 5 %).

Sin duda alguna, el orden *completivas* > *relativas* > *condicionales* pasó desapercibido porque en las oraciones completivas y en las relativas es posible encontrar alternancia de modo en el español sin que, a veces, sean nítidas las diferencias en la interpretación. En las primeras, las completivas, la alternancia de contenido modal suele explicarse por la presuposición del valor de verdad de aquello expresado en la subordinada, esto es si se presupone cierto o se afirma (*María sabía que su novio vendría a visitarla*, selección del indicativo) o no (*María deseaba que su novio viniera a visitarla*, selección del subjuntivo). Sin embargo, hay contornos que en los dos modos alternan de forma sutil sin que se identifiquen claramente esas diferencias de interpretación (*María esperaba / no creía que su novio vendría / viniera a visitarla*) y que se perfilan como el locus originario de la extensión del indicativo a costa del subjuntivo.

En cuanto al segundo tipo oracional, las relativas y adverbiales relacionadas, sucede algo semejante. La alternancia de modos se explica por el carácter más o menos específico del antecedente. Si es específico (existente o identificable) se emplea el indicativo (*El hombre, que sabría aquel oficio, había fallecido*), y si es inespecífico (no se afirma su existencia), el subjuntivo (*No hubo nadie que supiera aquel oficio*). Pero de nuevo hay lecturas en que los dos modos alternan con significado muy similar (*El hombre que sabría / supiera aquel misterio había desaparecido*). En consonancia con estos hechos, en el castellano septentrional la presencia del indicativo en lugar del subjuntivo es tanto más frecuente cuanto más específico es el antecedente. Un criterio formal indirecto permite medir la especificidad: la escala de la definitud. Aunque la interpretación de la especificidad de los sintagmas nominales definidos e indefinidos es contextual, los datos revelan una mayor presencia del desplazamiento acorde con ella: *antecedentes definidos* (64,5 %) > *antecedentes indefinidos* (58,3 %) > *nadie* (0 %).

Desde un punto de vista tipológico, la pérdida del subjuntivo según la escala *completivas* > *relativas* > *condicionales / resto* probablemente sea de aplicación a las variedades del español en América y a otras lenguas romances, ya que es un cambio predecible tanto en lo relativo al modo perdido (subjuntivo) como a los tiempos afectados (pasado), pues los elementos marcados tienden a perder la expresión flexiva de diferencias que se mantienen en los no marcados.

4.7. Rección

El estudio de la variación en la expresión del complementante *de* que rige un infinitivo subordinado, fenómeno conocido como *deísmo*, ha avanzado asimismo a partir del análisis de los datos del COSER (*Me han hecho mis padres de correr*) (De Benito & Pato 2015). En primer lugar, porque sobre el análisis del COSER se ha ofrecido el primer mapa que cartografía la extensión geográfica del deísmo, ya que los atlas apenas consiguen documentar este fenómeno. El mapa revela que hay ejemplos aislados de deísmo en la mitad septentrional peninsular, pero la expresión del complementante es más intensa y constante en la mitad sur (con ejemplos regulares en Andalucía, Murcia,

Extremadura, Castilla-La Mancha, Comunidad Valenciana y Madrid). Además, el número de ejemplos aumenta según nos acercamos hacia el sur, por lo que se configura como un rasgo sintáctico característico del español meridional.

Además, se han podido caracterizar los tipos de verbos con los que aparece: a) verbos de afección (*Me costó de adaptarme*; *Me da pena de verla*), a veces con lectura genérica; b) verbos de percepción (*Le oía de decir*); c) verbos de manipulación (*De bebérmelo no me dejó*); d) verbos de volición e intención (*Están deseandito de venir*), e incluso; e) verbos modales en perífrasis verbales (*Suele de pasar*).

La categoría a la que pertenece de, preposición o complementante, ha sido debatida tanto en el español como en otras lenguas que presentan infinitivos regidos por la preposición, como el catalán, el francés o el italiano (De Benito & Pato 2015). Después de constatar la distribución geográfica desigual y paulatina de los tipos de verbos que aceptan el deísmo, estos autores deducen que los primeros en aceptarlo fueron los de afección, en los que el valor preposicional de *de* es perceptible en que a veces aceptan un complemento nominal (*Me da miedo del perro*; *Me da pena de mamá*), y los más antiguos documentados. A partir de esos empleos parece haberse gramaticalizado como complementante, ya que no introduce nombres (*Estoy deseando del regalo; *Le oía de sus pasos). En una etapa posterior, se habría convertido en partícula introductora de infinitivos, similar a *to* en inglés, fase solo presente en los verbos modales, mucho menos frecuente y más restringida en la geografía. Por tanto, esta hipótesis predice que hay diferentes tipos de deísmo según las áreas y las fases de evolución.

4.8. Fonética y gramática

La aparición de variantes fonéticas en palabras gramaticales o la incidencia de factores gramaticales en la evolución fonética también ha sido objeto de investigación con los datos del COSER.

4.8.1. El cuantificador muncho

El contraste entre los datos del ALPI y el COSER ha permitido conocer la extensión geográfica de la forma *muncho* (Pato 2013), fenómeno antiguo pero hoy de uso frecuente en Andalucía centrooriental, Murcia, Alicante y Albacete (con ejemplos esporádicos aragoneses y en las zonas castellanohablantes de Castellón y Valencia). Un área independiente se configura en Asturias. Esta *-n-* epentética se documenta también en otras lenguas y dialectos romances, como el asturiano y los dialectos septentrionales italianos, así como históricamente en francés. Documentado desde el español antiguo, su frecuencia de uso fue notable, sobre todo, en el siglo XVI.

4.8.2. La pérdida de -d- *intervocálica*

La pérdida de la *-d-* intervocálica, uno de los fenómenos más característicos de la fonética oral del español, ha sido estudiada en profundidad sobre los datos del ALPI y del COSER (Estrada 2017, 2018). Por primera vez, se ha podido analizar este cambio desde una perspectiva geográfica sobre mallas uniformes en la recolección de los datos. Como resultado, se ha delimitado el área geográfica afectada y cuantificado el alcance de la pérdida en distintos contextos fonotácticos, morfológicos y sintácticos.

La pérdida de la *-d-* es característica de la mitad sur de la Península. A partir del sistema central hacia el norte disminuye mucho, salvo en los participios en *-ado*, y solo se detecta con intensidad en un área aislada, la Asturias centrooriental y Cantabria, en la que influye el contacto con el asturiano. La isoglosa delimitadora es casi idéntica en el

ALPI y el COSER, lo que indica que, pese al tiempo transcurrido entre ambas muestras, se trata de una situación estable.

En cuanto a los factores que condicionan la pérdida, el estudio demuestra que: a) la pérdida apenas se da en contacto con una *glide* mientras que la apertura de las vocales en contacto con la *-d-* podría favorecerla (*a, o, e* vs. *i, u*); b) el acento previo a la *-d-* favorece su pérdida, así como la acentuación paroxítona; c) las palabras de tres o más sílabas también propician más la pérdida que las bisílabas; d) desde el punto de vista gramatical, la pérdida ordena su frecuencia de acuerdo a esta escala: *morfemas flexivos* (62,5 %) > *lexemas gramaticales* (37,9 %) > *morfemas derivativos* (30,3 %) > *lexemas* (19,7 %); e) es teóricamente importante constatar que los diminutivos eliminan la *-d-* siguiendo el modelo de su base (*delgado> delgao> delgaúcho*, pero no *delgaducho*); y f) la preposición *de* apenas se pierde, aunque desde el punto de vista fonotáctico se encuentre en las mismas condiciones intervocálicas que dentro de una palabra. Ello parece depender de si la preposición es parte de una locución lexicalizada (*ojo de gallo*; *lucero del alba*) o si enlaza dos nombres en un sintagma (*una caja de madera*), en cuyo caso la pérdida es casi inexistente. g) Finalmente, la frecuencia del elemento en que la *-d-* se incluye parece ser determinante. Por ese motivo, los nombres propios son especialmente renuentes a perderla mientras que los participios en *-ado*, los cuantificadores *todo* > *to*, *nada* > *na* o el verbo *poder* dan cuenta de la inmensa mayoría de los casos.

4.9. Léxico y gramática

En las entrevistas del COSER se registran adjetivos o adverbios poco o nada comunes en el habla común, tanto por conservación de formas arcaicas como por desarrollo de nuevos significados y posibilidades sintácticas. Algunos ejemplos estudiados son los siguientes.

4.9.1. Mucho *como cuantificador de adjetivos y adverbios*

La distribución geográfica del uso del superlativo absoluto con la forma plena *mucho* como cuantificador de adjetivos y adverbios (*mucho bien*, *mucho guapa*) ha sido también confeccionada gracias al COSER (Pato & Viejo Fernández 2017). Aparte de su empleo en Asturias occidental, es uso propio del español oriental del norte (con ejemplos desde el oriente de Burgos, La Rioja, Soria, Guadalajara, Navarra, Aragón –sobre todo Teruel– y Murcia). Como uso vigente en hablantes mayores de 70 años, es un fenómeno en retroceso y siempre en alternancia con *muy*. No obstante, *mucho* parece conservarse mejor cuando modifica a un adverbio que a un adjetivo.

4.9.2. El adjetivo semejante

En el habla rural se documenta también un uso nuevo del adjetivo *semejante* (Pato 2018a). A los valores usuales de *semejante* como adjetivo calificativo 'similar' (*Un libro semejante al cuadro*), cuantificador ponderativo 'tal' (*No es lícito valerse de semejantes medios*) y determinante (*No he visto a semejante hombre*), se suma, en algunos enclaves de Madrid y Ciudad Real, un uso adverbial. Esta recategorización de adjetivo a adverbio parece derivar del segundo significado ('de manera similar, de la misma manera, igualmente'). En los ejemplos registrados *semejante* modifica a adverbios deícticos de lugar (*aquí*) y modo (*así*) (*Tenías que tumbar la oveja, la oveja tripa arriba, el cuello aquí, la cabeza semejante aquí, y las patas pa abajo*; *El hombre o la mujer, cogía así un cacho, y lo metía semejante así, en el, en el, en el centro*). En estos ejemplos *semejante*

presenta un valor anafórico que hace referencia a una situación inmediatamente anterior, de modo que la paráfrasis de *semejante así* sería 'de esta manera, por ejemplo'.

4.9.3. Los adverbios relativos e interrogativos

La distribución geográfica y sintáctica de los adverbios relativos e interrogativos *dónde, ónde, adónde, aónde* y *ánde* también ha sido estudiada sobre los datos del COSER y de los atlas lingüísticos (Del Barrio 2018). El ALPI permite descubrir una nueva isoglosa entre dos áreas dialectales ya conocidas por otros aspectos: la occidental, en que se emplea *onde* para la ubicación y *a(o)nde* para el destino, y la oriental, en que *a(o)nde* expresa ambos contenidos. El sincretismo en sentido contrario, a favor de *onde*, solo aparece en algunos puntos andaluces, extremeños y leoneses. Los datos del COSER revelan la desdialectalización acaecida a lo largo del siglo XX, pues *ónde, a(ó)nde* son ahora minoritarios en comparación con las formas estándar *dónde* y *adónde*. Aun así, *ánde* mantiene mayor vigencia en el área oriental y *ónde* en Asturias y Cantabria. Para expresar la ubicación *dónde* es la forma más frecuente (79,9 %) frente a *ánde, a(d)ónde* (20 %). Para la dirección, en cambio, *ánde, a(d)ónde* (58,8 %) superan a *(d)ónde* (43,1 %).

4.9.4. Otros cuantificadores y adverbios

Además, en las entrevistas del COSER se registran cuantificadores y adverbios supuestamente desaparecidos de la Península desde los siglos XVI y XVII. En Toledo aflora el cuantificador *algotro* (Octavio de Toledo 2016) y en varios enclaves de las provincias de Teruel y de Zaragoza aún está vivo el adverbio *agora* 'ahora', que ha perdurado sin interrupción en el ámbito aragonés (Pato 2010a).

5. CONCLUSIONES

La dialectología moderna sabe que la gramática dialectal es una fuente de información preciosa, hasta hace poco tiempo insuficientemente valorada, no solo para la caracterización de un dominio lingüístico concreto sino también para el estudio teórico y tipológico de las lenguas (*cfr*., entre otros, Chambers 1995, Kortmann 1999, 2004a, 2004b, Bariers, Corsnip & van der Kleij 2002, Biberauer 2008). Como hemos visto a lo largo de este trabajo, los datos del COSER sirven para alcanzar interpretaciones más correctas de los principios lingüísticos operantes en las variedades orales.

El COSER constituye, además, un complemento tanto de los atlas lingüísticos como de los diversos corpus de habla culta y urbana que se han compilado, o se están compilando, en el mundo hispanohablante, con la posibilidad de estudiar cambios lingüísticos en tiempo real. Puesto que, como vimos, los informantes del COSER pertenecen al mismo grupo social de los informantes seleccionados para los atlas lingüísticos, resulta posible la comparación entre el habla de varias generaciones sucesivas de hablantes de las mismas áreas geográficas. Pese a las diferencias existentes entre una y otra metodología (cuestionario *vs*. entrevista oral), el contraste entre los datos permite investigar cambios en el tiempo transcurrido desde comienzos del siglo XX hasta comienzos del siglo XXI.

Esta metodología comparativa entre el COSER y los atlas y otros corpus para obtener conclusiones sobre la evolución diacrónica y la implantación sociolingüística de los usos sometidos a variación se ha puesto a prueba en distintos trabajos como los de Pato (2003, 2004b, 2010b, 2012c, 2013), Pato & Heap (2012), Pato & O'Neill (2013), Pato & Viejo (2017), Heap (2006), De Benito (2015b, 2016), Castillo Lluch & Octavio de Toledo (2016), León Zurdo (2017), Estrada (2017), Lara (2018) o Del Barrio (2018), con

conclusiones relevantes sobre el mantenimiento o la desdialectalización de los aspectos estudiados (*cfr*. §§ 4.1.1, 4.1.2, 4.1.5, 4.4.3, 4.6, 4.8.1, 4.8.2, 4.9.3).

De este modo, se ha podido constatar la desdialectalización de algunos aspectos propios de la gramática rural. Además de los citados *supra*, conviene mencionar el declive de las secuencias *me se* y *te se*, la de 2ª persona más frecuente que la de 1ª en extensión geográfica (Heap 2006), o del uso de *yo* y *tú* como pronombres regidos por preposición en el castellano de Aragón, rasgo común al aragonés, catalán y provenzal (Pato 2012b). Estas formas se documentan hoy día en el COSER solo en muy pocos enclaves de las provincias de Huesca (*De tú no me despido, pimpollo de albahaca fina*) y de Teruel (*Espérate allí hasta que venía el amo, a por tú*), mientras que tenían difusión más amplia en el ALEANR. En cambio, se ha podido constatar que otros aspectos dialectales mantienen incólume su área de uso, como es el caso del sistema referencial de los pronombres átonos de 3ª persona (León Zurdo 2017), el desplazamiento modal del imperfecto de subjuntivo (Pato 2003) o la pérdida de la *-d-* intervocálica (Estrada 2017).

Como hemos visto, no son pocos los rasgos morfosintácticos que han sido estudiados sobre datos del COSER. El resultado ha sido una mejor descripción y análisis de aspectos de variación que solo se describían de forma parcial o impresionista, tanto en lo relativo a sus límites geográficos como a sus factores lingüísticos condicionantes[15]. Esta investigación ha modificado notablemente la idea tradicional de que la variación gramatical en el español rural era casi inexistente, por lo que se ha obtenido un panorama mucho más complejo y también más ajustado a la realidad sobre la dialectología moderna del español, con no pocas repercusiones también para la historia de la lengua.

Así, la visión tradicional de las áreas dialectales del español europeo se ha visto enriquecida y complementada con aspectos relativos a la gramática. En algunas ocasiones, se corroboran áreas ya conocidas por la fonética –como el español meridional o el andaluz occidental– pero otras veces se postulan áreas nuevas que no se habían identificado claramente: el español oriental, el español occidental, el español noroccidental –y en contacto con el gallego y el asturiano–, el español septentrional, el español en contacto con el valenciano y de Murcia o el español en contacto con el vasco, entre otras (Fernández-Ordóñez 2016).

Finalmente, y es un aspecto que no suele señalarse, el COSER y los numerosos proyectos a él asociados han servido para formar a toda una serie de investigadores, tanto en España como fuera de ella, especialistas en dialectología, variación gramatical y sintaxis dialectal. El COSER ha servido de modelo, asimismo, en la creación de otros corpus orales del español, como el COLEM (*Corpus oral de la lengua española en Montreal*) y comienza poco a poco su extensión a tierras americanas (COSER-Cuba, COSER-Uruguay). Además, se emplea también como material en la enseñanza del español como lengua extranjera a alumnos de niveles avanzados (Sanmiguel Mariño 2011).

En el momento actual, solo un 13 % de sus materiales se ofrece en línea, y se espera seguir desarrollando el corpus hasta completar todas las provincias e islas, así como corregir los posibles errores en las transcripciones existentes. Por ello, el COSER tiene ante sí notables desafíos desde el punto de vista técnico para poder ser una herramienta aún más útil para la investigación gramatical. Retos pendientes son la mejora de la anotación morfosintáctica existente y su parseo o análisis sintáctico (*parsing*) –por constituyentes y dependencias–, así como la transcripción automática mediante

[15] No podemos citar todos aquí. Otra dimensión en vías de estudio es la sintaxis de los posesivos, por el ejemplo, el posesivo tónico (Pato 2015), con valor habitualmente focal, o el adjetivo pospuesto a un adverbio, en lugar de un sintagma preposicional con núcleo pronominal (*detrás de ti* > *tuyo, tuya*) (Salgado & Bouzouita 2017).

reconocimiento del habla (*respeaking*) para, al menos, adelantar la ardua labor de transcripción de las encuestas que todavía no han sido abordadas.

REFERENCIAS BIBLIOGRÁFICAS

ALCyL: Alvar, M. (1999): *Atlas lingüistico de Castilla y León*. Valladolid: Junta de Castilla y León, 3 vols.

ALEA: Alvar, M. (1961-1973): *Atlas lingüístico y etnográfico de Andalucía*. Granada: Universidad de Granada/ CSIC, 6 vols.

ALEANR: Alvar, M. (1979-83): *Atlas lingüístico y etnográfico de Aragón, Navarra y La Rioja*. Zaragoza: Institución Fernando el Católico/ CSIC, 12 vols.

ALECant: Alvar, M. (1995): *Atlas lingüístico y etnográfico de Cantabria*. Madrid: Fundación Marcelo Botín, 2. vols.

ALECMan*:* García Mouton, P. & F. Moreno Fernández, *Atlas lingüístico (y etnográfico) de Castilla-La Mancha*. Alcalá de Henares: Universidad de Alcalá.

ALEICan: Alvar, M. (1975): *Atlas lingüístico y etnográfico de las islas Canarias*. Madrid: Arco/ Libros, 3 vols.

ALPI: Navarro Tomás, T. (dir.) (1962): *Atlas Lingüístico de la Península Ibérica*, vol. 1, *Fonética*. Madrid: CSIC. Véase también Pilar García Mouton (coord.), Inés Fernández-Ordóñez, David Heap, Maria Pilar Perea, João Saramago, Xulio Sousa, 2016, ALPI-CSIC [www.alpi.csic.es], edición digital de Navarro Tomás, Tomás (dir.), Atlas Lingüístico de la Península Ibérica, Madrid, CSIC.

Barbiers, S., L. Cornips & S. van der Kleij (eds.). (2002): *Syntactic microvariation*. Amsterdam: Meertens-Instituut.

Biberauer, T. (ed.) (2008): *The limits of syntactic variation*. Amsterdam/ Philadelphia: John Benjamins.

Bouzouita, M. & E. Pato (2019): "*Antes había(n) pozos en el pueblo*. La pluralización del verbo *haber* existencial en español rural europeo", *Revue de linguistique romane* 83, pp. 137-165.

Buesa, T. & R. M. Castañer (1994): "El pretérito perfecto simple en las hablas pirenaicas de Aragón y Navarra", *Archivo de Filología Aragonesa* L, pp. 65-132.

Camus Bergareche, B. & S. Gómez Seibane (2015): "La diversidad del español en Álava: Sistemas pronominales a partir de las encuestas del COSER", *Revista de Filología Española* XCV, pp. 279-206.

Castillo Lluch, M. & Á. Octavio de Toledo y Huerta (2016): "*Habemos muchos que hablamos español*: distribución e historia de la concordancia existencial en primera persona del plural", in C. de Benito & Á. Octavio de Toledo y Huerta (eds.): *En torno*

a haber*: construcciones, usos y variación desde el latín hasta la actualidad*. Bern: Peter Lang, pp. 111-168.

CHAMBERS, J. K. (1995): *Sociolinguistic theory*. Oxford: Blackwell.

CHAMBERS, J. K. (2004): "Dynamic typology and vernacular universals", in B. Kortmann (ed.): *Dialectology meets typology*. Berlin/New York: Mouton de Gruyter, pp. 127-145.

CORBETT, G. C. (2006): *Agreement*. Cambridge: Cambridge University Press.

COLEM: Pato, E. (dir.) (2013-2020): *Corpus oral de la lengua española en Montreal*. Montreal: Université de Montréal <esp-montreal.jimdo.com>.

CORPES XXI*: Corpus del español del siglo XXI*. Madrid: RAE <http://www.rae.es>.

COSER: Fernández-Ordóñez I. (dir.) (2005-2019): *Corpus Oral y Sonoro del Español Rural*. Madrid: Universidad Autónoma de Madrid <corpusrural.es>.

CREA: *Banco de datos. Corpus de referencia del español actual*. Madrid: RAE <http://www.rae.es>.

DE BENITO MORENO, C. (2012): "The pronominal coding of the patient in reflexive indefinite agent constructions in Peninsular Spanish", *Journal of Portuguese Linguistics* 11(1), pp. 45-60.

DE BENITO MORENO, C. (2013): "*(Esa tela) se la descose*: la pronominalización del paciente en las impersonales reflejas del español peninsular", *Borealis* 2(2), pp. 129-157.

DE BENITO MORENO, C. (2015a): "*Pero se escondíamos como las ratas*: syncretism in the reflexive paradigm in Spanish and Catalan", *Isogloss* 1, pp. 95-127.

DE BENITO MORENO, C. (2015b): *Las construcciones con "se" desde una perspectiva variacionista y dialectal*. Tesis doctoral. Universidad Autónoma de Madrid.

DE BENITO MORENO, C. (2016): "La pronominalización en las construcciones existenciales con *haber*: ¿hay restricciones o no las hay?", in C. de Benito & Á. Octavio de Toledo y Huerta (eds.): *En torno a* haber*: construcciones, usos y variación desde el latín hasta la actualidad*. Bern: Peter Lang, pp. 209-237.

DE BENITO MORENO, C. & E. PATO (2015): "On the *de* + infinitive construction (*deísmo*) in Spanish", *Dialectologia* (Special issue V), pp. 29-51.

DE BENITO MORENO, C., J. PUEYO & I. FERNÁNDEZ-ORDÓÑEZ (2016): "Creating and designing a corpus of rural Spanish", *Proceedings of the 13th Conference on Natural Language Processing* (*KONVENS 2016*). Bochum: Rühr-Universität Bochum, pp. 78-83.

DEL BARRIO DE LA ROSA, F. (2018): *Espacio variacional y cambio lingüístico en español*. Madrid: Visor.

ESTRADA ARRÁEZ, A. (2017): *The loss of intervocalic /d/ in the Iberian Peninsula*. Tesis doctoral. Universidad Autónoma de Madrid/Universidad de Friburgo.

ESTRADA ARRÁEZ, A. (2018): "Factors involved in the evolution of intervocalic /d/ in the Iberian Peninsula: the case of frequency", in F. Sánchez Miret & D. Recasens (eds.): *Production and Perception Mechanisms of Sound Change*. Múnich: Lincom Europa, pp. 145-156.

FERNÁNDEZ-ORDÓÑEZ, I. (1993): "Leísmo, laísmo y loísmo: estado de la cuestión", in O. Fernández Soriano (ed.): *Los pronombres átonos*. Madrid: Taurus, pp. 63-96.

FERNÁNDEZ-ORDÓÑEZ, I. (1994): "Isoglosas internas del castellano. El sistema referencial del pronombre átono de tercera persona", *Revista de Filología Española* LXXIV, pp. 71-25.

FERNÁNDEZ-ORDÓÑEZ, I. (1999): "Leísmo, laísmo y loísmo", in I. Bosque & V. Demonte (dirs.): *Gramática descriptiva de la lengua española*. Madrid: Espasa, vol. I, cap. 21, pp. 1317-1397.

FERNÁNDEZ-ORDÓÑEZ, I. (2001): "Hacia una dialectología histórica. Reflexiones sobre la historia del leísmo, el laísmo y el loísmo", *Boletín de la Real Academia Española* LXXXI, pp. 389-464.

FERNÁNDEZ-ORDÓÑEZ, I. (2006): "Del Cantábrico a Toledo: "El "neutro de materia" hispánico en un contexto románico y tipológico" (primera parte), *Revista de Historia de la Lengua Española* 1, pp. 67-118.

FERNÁNDEZ-ORDÓÑEZ, I. (2007a): "Del Cantábrico a Toledo. El "neutro de materia" hispánico en un contexto románico y tipológico", *Revista de Historia de la Lengua Española* 2, pp. 29-81.

FERNÁNDEZ-ORDÓÑEZ, I. (2007b): "El neutro de materia en Asturias y Cantabria. Análisis gramatical y nuevos datos", in I. Delgado Cobos & A. Puigvert Ocal (eds.): *Ex admiratione et amicitia. Homenaje a Ramón Santiago*. Madrid: Ediciones del Orto, pp. 395-434.

FERNÁNDEZ-ORDÓÑEZ, I. (2007c): "Nuevas perspectivas en el estudio de la variación dialectal del español: El *Corpus Oral y Sonoro del Español Rural* (*COSER*)", in D. Trotter (ed.): *Actes du XXIV Congrès de Linguistique et Philologie Romanes*. Tübingen: Niemeyer, vol. 3, pp. 29-44.

FERNÁNDEZ-ORDÓÑEZ, I. (2009a): "Dialect grammar of Spanish from the perspective of the Audible Corpus of Spoken Rural Spanish (or Corpus Oral y Sonoro del Español Rural, COSER)", *Dialectologia* 3, pp. 23-51.

FERNÁNDEZ-ORDÓÑEZ, I. (2009b): "The development of mass/count distinctions in Indo-European languages", in V. Bubenik, J. Hewson & S. Rose (eds.): *Grammatical change in Indo-European languages*. Amsterdam/Philadelphia: John Benjamins, pp. 55-68.

FERNÁNDEZ-ORDÓÑEZ, I. (2010a): "La grammaire dialectale de l'espagnol à travers le Corpus oral et sonore de l'espagnol rural (COSER, *Corpus Oral y Sonoro del Español Rural*)", *Corpus: La syntaxe de corpus/Corpus syntax* 9, pp. 81-114.

FERNÁNDEZ-ORDÓÑEZ, I. (2010b): "New methods for the study of grammatical variation and the Audible Corpus of Spoken Rural Spanish", in G. Aurrekoetxea & J. L. Ormaetxea (eds.): *Tools for Linguistic Variation*. Bilbao: Universidad del País Vasco, pp. 119-130.

FERNÁNDEZ-ORDÓÑEZ, I. (2011): "Nuevos horizontes en el estudio de la variación gramatical del español: el *Corpus Oral y Sonoro del Español Rural*", in G. Colón i Domènech & L. Gimeno Betí (eds.): *Noves tendències en la dialectologia contemporània*. Castellón de la Plana: Universitat Jaume I, pp. 173-203.

FERNÁNDEZ-ORDÓÑEZ, I. (2012): "Dialect areas and linguistic change: Pronominal paradigms in Ibero-Romance dialects from a cross-linguistic and social typology perspective", in G. de Vogelaer & G. Seiler (eds.): *The Dialect Laboratory. Dialects as a testing ground for theories of language change*. Amsterdam/Philadelphia: John Benjamins, pp. 73-106.

FERNÁNDEZ-ORDÓÑEZ, I. (2015): "*Mucha trabajo*: sincretismo femenino en los cuantificadores evaluativos de Cantabria", in *Studium Grammaticae. Homenaje al profesor José Antonio Martínez*. Oviedo: EdiUNo, pp. 337-349.

FERNÁNDEZ-ORDÓÑEZ, I. (2016): "Dialectos del español peninsular", in J. Gutiérrez Rexach (ed.): *Enciclopedia lingüística hispánica*. London/New York: Routledge, vol. 2, pp. 387-404.

FERNÁNDEZ-ORDÓÑEZ, I. (2019): "Mass/count distinctions in Ibero-Romance dialects", in Á. Gallego (ed.): *The syntactic variation of Spanish dialects*. Oxford: Oxford University Press, pp. 60-108.

GÓMEZ SEIBANE, S. (2017): "Español en contacto con la lengua vasca: datos sobre la duplicación de objetos directos posverbales", in A. Palacios (ed.): *Variación y cambio lingüístico en situaciones de contacto*. Madrid/Frankfurt: Iberoamericana/Vervuert, pp. 143-159.

HEAP, D. (2006): "Secuencias «invertidas» de clíticos: un cambio (¿?) en tiempo real", in J. J. de Bustos Tovar & J. L. Girón Alconchel (eds.): *Actas del VI Congreso Internacional de Historia de la Lengua Española*. Madrid: Arco/Libros, vol. I, pp. 785-798.

HEAP, D. & E. PATO (2012): "Plurales anómalos en los dialectos y en la historia del español", in E. Montero Cartelle & C. Manzano Rovira (eds.): *Actas del VIII Congreso Internacional de Historia de la Lengua Española*. Santiago de Compostela: AHLE/ Meubook, vol. 1, pp. 829-840.

KLEIN-ANDREU, F. (2000): *Variación actual y evolución histórica: los clíticos le/s, la/s, lo/s*. Múnich: Lincom Europa.

KORTMANN, B. (1999): "Typology and dialectology", in B. Caron (ed.): *Proceedings of the 16th International Congress of Linguists*. Amsterdam: Elsevier Science, Multimedia CD.

KORTMANN, B. (2004a): "Why dialect grammar matters", *The European English Messenger* XIII, pp. 24-29.

KORTMANN, B. (ed.) (2004b): *Dialectology meets typology. Dialect grammar from a cross-linguistic perspective*. Berlin/New York: Mouton de Gruyter.

LARA BERMEJO, V. (2018): "El superlativo absoluto en el español peninsular del siglo XX", *RILCE* 34 (1), pp. 225-239.

LAPESA, R. (1968): "Sobre los orígenes y evolución del leísmo, laísmo y loísmo", in K. Baldinger (ed.): *Festschrift Walther von Wartburg*. Tübingen: Max Niemeyer, pp. 523-551.

LEÓN ZURDO, O. (2017): "Case variation in unstressed third person pronouns in the Linguistic Atlas of the Iberian Peninsula", *Dialectologia* 18, pp. 43-72.

LLORENTE MALDONADO DE GUEVARA, A. (1965): "Algunas características lingüísticas de La Rioja en el marco de las hablas del valle del Ebro y de las comarcas vecinas de Castilla y Vasconia", *Revista de Filología Española* 48(3/4), pp. 321-350.

MONDÉJAR, J. (1970): *El verbo andaluz. Formas y estructuras*. Madrid: CSIC.

OCTAVIO DE TOLEDO Y HUERTA, Á. (2016): "Sin CORDE pero con red: *algotras* fuentes de datos", *Revista Internacional de Lingüística Iberoamericana* 28, pp. 19-47.

PATO, E. (2003): *La sustitución del imperfecto de subjuntivo por el condicional simple y el imperfecto de indicativo en el castellano septentrional peninsular. (Estudio de variación dialectal)*. Tesis doctoral. Universidad Autónoma de Madrid.

PATO, E. (2004a): "Los perfectos fuertes analógicos en español", in M. Trinidad (ed.): *Actas del Congreso Internacional "APLEx 2004". Patrimonio Lingüístico Extremeño* (Cáceres, 4-6 noviembre de 2004). Cáceres: Editora Regional (CD Rom).

PATO, E. (2004b): *La sustitución de* cantara/cantase *por* cantaría/cantaba *(en el castellano septentrional peninsular)*. Madrid: Universidad Autónoma de Madrid.

PATO, E. (2010a): "Los adverbios *agora* y *ahora*: Dos orígenes, un mismo resultado", *Revista de Historia de la Lengua Española* 5, pp. 167-173.

PATO, E. (2010b): "Linguistic levelling in Spanish: The analogical strong preterites", *Canadian Journal of Linguistics* 55(2), pp. 209-225.

PATO, E. (2012a): "*Cantabámos* por *cantábamos*: forma 'etimológica' del español rural", *Archivo de Filología Aragonesa* 68, pp. 219-236.

PATO, E. (2012b): "Nivelación lingüística y simplificación: El uso de preposición + *tú* en la historia de la lengua", in E. Montero Cartelle & C. Manzano Rovira (eds.): *Actas*

del VIII Congreso Internacional de Historia de la Lengua Española. Santiago de Compostela: AHLE/Meubook, vol. 1, pp. 1029-1039.

PATO, E. (2012c): "Variación dialectal y análisis estadístico: Formas indicativas por subjuntivas en español rural", in E. Pato & J. Rodríguez Molina (eds.): *Estudios de filología y lingüística españolas. Nuevas voces en la disciplina*. Bern: Peter Lang, pp. 93-133.

PATO, E. (2013): "Sobre la forma *muncho*", *Estudios de Lingüística. Universidad de Alicante* 27, pp. 329-342.

PATO, E. (2015): "El posesivo (antepuesto) tónico en español: ¿fenómeno de foco y contraste?", *Dialectologia* 14, pp. 47-73.

PATO, E. (2016): "La pluralización de *haber* en español peninsular", in C. de Benito & Á. Octavio de Toledo y Huerta (eds.): *En torno a* haber*: construcciones, usos y variación desde el latín hasta la actualidad*. Bern: Peter Lang, pp. 357-391.

PATO, E. (2018a): "Semejante adjetivo es *semejante*. Sus valores en español actual", *Estudios filológicos* 61, pp. 59-74.

PATO, E. (2018b). "*Queriba una cosa y traíba otra*. Los pretéritos imperfectos 'analógicos' en español", *Philologica Jassyensia* XIV(2), pp. 83-100.

PATO, E. & P. O'NEILL (2013): "Los gerundios 'analógicos' en la historia del español (e iberorromance)", *Nueva Revista de Filología Hispánica* LXI(1), pp. 1-27.

PATO, E. & X. VIEJO FERNÁNDEZ (2017): "*Mu(n)cho guapa* y *mu(n)cho bien*: *mu(n)cho* como cuantificador 'pleno' de adjetivos y adverbios en castellano y en asturiano", *Bulletin of Spanish Studies* 94(1), pp. 1-23.

SALGADO, H. & M. BOUZOUITA (2017): "El uso de las construcciones de adverbio locativo con pronombre posesivo en el español peninsular: un primer acercamiento diatópico", *Zeitschrift für romanische Philologie* 133(3), pp. 766-794.

SANMIGUEL MARIÑO, A. (2011): "Reseña: Corpus Oral y Sonoro del Español Rural (COSER)", *Boletín de ASELE* 44, pp. 58.

EL *ATLAS LINGÜÍSTICO DE CUBA* (ALCU): NOVEDAD Y ORIGINALIDAD EN LOS ESTUDIOS DE GEOGRAFÍA LINGÜÍSTICA CONTEMPORÁNEOS

The Linguistic Atlas of Cuba *(ALCu): Novelty and originality in contemporaneous linguistic geography studies*

AILYN FIGUEROA GONZÁLEZ

Universidad Autónoma Metropolitana, Unidad Iztapalapa

Resumen

El *Atlas Lingüístico de Cuba* (ALCu) constituye la obra dialectológica y de geografía lingüística de mayor envergadura realizada en Cuba. Con su publicación se llena un vacío en los estudios sobre estas disciplinas lingüísticas en el país. En el trabajo se presentan los presupuestos metodológicos que se tuvieron en cuenta en su realización, el diseño de la aplicación web con sus múltiples utilidades investigativas y los principales resultados alcanzados. El ALCu es un referente para los proyectos actuales y futuros de corte dialectal por el tratamiento de la información recopilada y por el cartografiado automático mediante el empleo de los presupuestos metodológicos de la geografía lingüística pluridimensional.

Palabras clave: dialectología, geografía lingüística pluridimensional, metodología lingüística, cartografiado automático

Abstract

The *Linguistic Atlas of Cuba* (ALCu) is the largest dialectological and linguistic geography work carried out in Cuba. Publication of this work fills a gap in studies on these linguistic disciplines in the country. In this article I put forward the methodological assumptions that were taken into account in its realization, the design of the web application with its multiple investigative utilities, and the main results accomplished. The ALCu is a reference for current and future dialect projects because of both the information treatment collected and the automatic mapping using the methodological assumptions of multidimensional linguistic geography.

Keywords: dialectology, multidimensional linguistic geography, linguistic methodology, automatic mapping

1. INTRODUCCIÓN

La publicación del *Atlas Lingüístico de Cuba* (ALCu) en 2013 significó la culminación de un largo y arduo período de trabajo dialectológico y cartográfico, a la vez de suponer la continuación de las tareas investigativas para efectuar el análisis del enorme acervo lingüístico recopilado en la obra. El proyecto desarrollado por investigadores del

Instituto de Literatura y Lingüística "José Antonio Portuondo Valdor" (ILL)[1], constituye el estudio de la modalidad cubana del español más amplio, ambicioso y completo que se haya realizado en el país hasta la fecha.

La obra se llevó a cabo en dos etapas (1986-1996 y 2009-2013). El ALCu, concebido en un principio para ser en formato papel, deriva en un atlas en formato digital que aprovecha a la vez los adelantos metodológicos de la geografía lingüística y los beneficios de los procesos automatizados. En la versión digital del ALCu se pueden generar mapas según los requerimientos de los usuarios, entre otras utilidades.

El objetivo de este trabajo es presentar la obra en sus diferentes etapas de realización, así como el resultado final: la aplicación web. En la actualidad el ALCu es una referencia para nuevas obras dialectales y de geografía lingüística por su propuesta metodológica en el tratamiento de los datos lingüísticos recogidos en el terreno, la preparación de la aplicación informática y los aportes que tiene para la investigación lingüística.

2. Primera etapa de trabajo (1986-1996)

Las labores del ALCu se iniciaron en 1986 con la aprobación en el ILL del proyecto "Caracterización geolingüística del español de Cuba" diseñado por Raquel García Riverón[2] y ejecutado por un gran equipo multidisciplinario.

El equipo de investigadores del ILL estuvo integrado por Lourdes E. Montero Bernal, Lidia Santana González, América J. Menéndez Pryce, Marcia Morón García y F. Vladimir Pérez Casal. De 1986 a 1991 la dirección y coordinación de las tareas del proyecto estuvo a cargo de García Riverón. Luego, la sucedieron Luis R. Choy López (de 1991 a 1993), y Sergio Valdés Bernal (de 1994 hasta 1996), ambos investigadores de la misma institución.

2.1. La concepción de la obra

En la concepción del ALCu se partió de la hipótesis de que dentro de la unidad lingüística del español de Cuba existe una variabilidad que permite caracterizar subregiones lingüísticas (García 1989). En correspondencia, el objetivo del proyecto fue identificar y caracterizar las regiones lingüísticas del país mediante el estudio geoespacial y multilateral del léxico, la morfosintaxis y la fonética de la modalidad cubana del español.

Para el diseño del proyecto y el entrenamiento del equipo se contó con la valiosa asesoría de Manuel Alvar López y Antonio Quilis, quienes se encontraban en el país a propósito de la aplicación del cuestionario del *Atlas Lingüístico de Hispanoamérica* (ALH)[3].

2.2. El cuestionario

Para la conformación del cuestionario del ALCu, de tipo onomasiológico, se realizó una ardua labor de cotejo con los cuestionarios de los atlas lingüísticos hispánicos publicados hasta la fecha, para incluir las preguntas recogidas de forma constante en

[1] El Instituto de Literatura y Lingüística "José Antonio Portuondo Valdor" (ILL) pertenece al Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) de Cuba.

[2] Desde la fundación del ILL en 1965, su director José Antonio Portuondo Valdor promovió la realización del atlas lingüístico. Para ello le encargó al lingüista rumano Marius Salas la elaboración de un cuestionario dialectológico que fue aplicado en 1967 en localidades de la provincia Pinar del Río. De esta etapa solo se conservan cinco cuestionarios manuscritos en el ILL.

[3] Hasta la fecha no se tiene noticia de la edición del volumen del ALH que incluye al español hablado en Cuba.

estos. Ello permitiría comparar los datos recopilados en Cuba con otras regiones hispanohablantes.

Además, el cuestionario se enriqueció con los ítems más representativos de la realidad cubana recogidos en los repertorios lexicográficos, en especial, aquellos identificados como localismos o regionalismos, referidos sobre todo a la fauna y a la flora, endémicas o no, con gran variación léxica según las obras consultadas. Sirvan de ejemplo las preguntas 957. tomeguín de la tierra, y 1020. cocuyo.

El cuestionario quedó conformado por un total de 2747 preguntas: 1980 de léxico, 230 de fonética y 370 de morfología y sintaxis. También se incluyeron 62 preguntas relacionadas con entonación y 105 de lenguaje gestual.

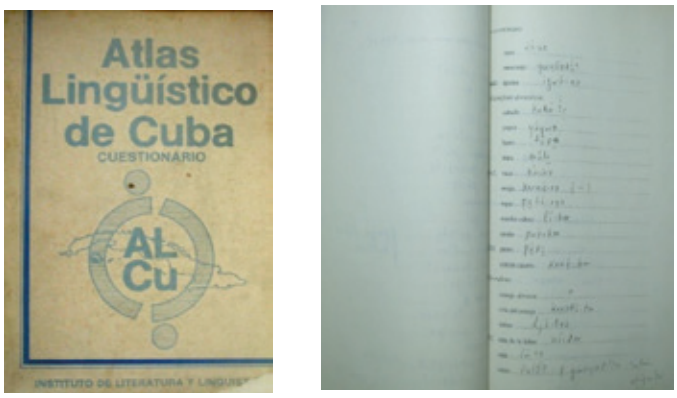


FIGURA 1. Portada y hoja interior del Atlas Lingüístico de Cuba. Cuestionario[4]

Las preguntas sobre léxico se organizaron en 21 campos semánticos relacionados con el hombre y su ambiente social, la flora, la fauna, la geografía y el tiempo cronológico. En los apartados de morfología y sintaxis y de fonética se tuvieron en cuenta los fenómenos de variación más frecuentes en Hispanoamérica y Cuba.

La aplicación del cuestionario se realizó de 1989 a 1995. Los investigadores se auxiliaron de láminas, imágenes y dibujos para facilitar la comprensión de la realidad encuestada.

2.3. Localidades

La mayoría de las localidades encuestadas fueron rurales, tal como se prescribe en los presupuestos metodológicos de la dialectología tradicional. Además, se aplicaron encuestas en zonas urbanas, tomando en consideración los adelantos en los estudios dialectales de ese momento, con el fin de tener una visión más completa de la realidad lingüística del país y estar en sintonía con los estudios de geografía urbana.

La red de puntos poblados se seleccionó de acuerdo con los datos del *Censo de población y viviendas de la República de Cuba de 1980*. El diseño de la red de localidades se proyectó lo más densa posible para que pudiera obtenerse un trazado de isoglosas mejor definidas y un mapa sintético más preciso.

En total, se seleccionaron 88 puntos poblados: 69 rurales y 19 urbanos, repartidos por toda la geografía nacional. Además, se atendieron a otros factores como la información sociocultural e histórica de la zona y la ubicación geográfica: región costera,

[4] Las imágenes presentadas en este artículo fueron tomadas de la obra.

montañosa o llana. No se incorporó la capital del país, por ser una zona cosmopolita cuyas características lingüísticas requieren un estudio especial[5].

FIGURA 2. Mapa con la ubicación de las localidades rurales encuestadas en el ALCu

FIGURA 3. Mapa con la ubicación de las localidades urbanas encuestadas en el ALCu

2.4. Selección de los informantes

Tal como se contemplaba en la metodología dialectológica tradicional, los informantes seleccionados debían reunir los requisitos de oriundez, haber permanecido en el lugar la mayor parte de su vida y tener disposición para responder el cuestionario.

En cada localidad se entrevistaron de 3 a 6 informantes teniendo en cuenta el sexo (masculino/femenino), la edad (tres grupos generacionales) y nivel instruccional (dos niveles). Para la estratificación de la muestra según las variables sociales mencionadas, ver tabla 1.

[5] En ese momento se estaba desarrollando en la capital el «Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades del mundo hispánico», dirigido por J. M. Lope Blanch.

Niveles instruccionales	Grupos generacionales		
	I (18-29 años)	II (30-49 años)	III (50 o más)
Primario (De 0 a 6 grados)	M/F	M/F	M/F
Secundario (De 7 a 9 grados)	M/F	M/F	M/F

TABLA 1. Estratificación de la muestra del ALCu

2.5. Resultados de la primera etapa del ALCu

Entre los resultados de la primera etapa de trabajo está la redacción de la versión final del *Atlas Lingüístico de Cuba*. *Cuestionario* a partir de su aplicación, a modo de pilotaje, en 14 puntos poblados de todo el país. Ello permitió que se incluyeran nuevos datos y que se desecharan otros por su irrelevancia. El pilotaje además sirvió para adiestrar a los investigadores con el trabajo *in situ*.

A partir de entonces fue posible la realización de la encuesta dialectal del ALCu en la red de puntos poblados establecida. Todas las respuestas fueron escritas en los cuadernos en transcripción fonética, para ello se utilizó el Alfabeto fonético de la *Revista de Filología Española* (ARFE), empleado en estudios hispánicos de esta índole desde su publicación en 1915.

El procesamiento de parte de la información recopilada permitió la elaboración de diferentes monografías, las cuales fueron publicadas en la obra *Visión geolectal de Cuba* (Valdés, Montero, Morón *et alii* 2007) y la confección del «Glosario de regionalismos y ruralismos de Cuba» (Cárdenas, Menéndez y Santana 1999) [inédito].

También, el estudio de los datos del ALCu para las zonas rurales pudieron corroborar la delimitación de las zonas lingüísticas de Cuba esbozadas por Choy (1989) a partir de sus estudios sobre el consonantismo en el español de Cuba. Vale aclarar que el trazado de isoglosas del ALCu, resultó más preciso dada la densidad de la red de localidades y la abundancia del material dialectológico recopilado de índole fonética y léxica fundamentalmente.

FIGURA 4. Representación de las cinco zonas geolectales de Cuba propuestas para el habla rural (Valdés, Montero, Morón *et alii* 2007)

Sin embargo, no se pudo concretar en esta etapa el tercer objetivo del proyecto: el cartografiado de los datos del ALCu.

3. Segunda etapa del proyecto (2009-2013)

El año 2009 fue propicio para el reinicio de los trabajos referidos al ALCu, pues se contaba en el ILL con un grupo de investigadores del Departamento de Lingüística, especialistas en informática y con las condiciones para elaborar la obra desde otra perspectiva: su publicación en formato digital.

El equipo de trabajo en esta etapa estaba conformado por los investigadores Roxana Sobrino Triana, Elisa García González, Adianys Collazo Allen, Lidia Santana González, Ailyn Figueroa González, Yanelys Abreu Babi y Sixto Gómez Echemendia; y los informáticos Rogelio Caballero Justo (programador principal) y Raudel Rodríguez Rodríguez, todos dirigidos por Lourdes Montero Bernal.

En un inicio fue crucial el intercambio profesional realizado por Montero Bernal con especialistas en el tratamiento automatizado de los datos dialectológicos: Pilar García Moutón, codirectora del *Atlas Lingüístico (y etnográfico) de Castilla la Mancha* (ALeCMan), del Centro de Ciencias Humanas y Sociales (CCHS) del Consejo Superior de Investigaciones Científicas (CSIC), España; y Miguel Ángel Quesada Pacheco, director del *Atlas Lingüístico-Etnográfico de Costa Rica* (ALECORI) y del *Atlas Lingüístico de América Central* (ALAC), del Departamento de Lenguas Extranjeras, de la Universidad de Bergen, Noruega. Con la ayuda de ambos investigadores y de Isabel Molina Martos codirectora del *Atlas Dialectal de Madrid* (ADiM), profesora titular de la Universidad de Alcalá e investigadora del CSIC, quedó esbozado el primer mapa para el ALCu.

3.1. Diseño de la aplicación web

La tarea inicial desarrollada por el equipo de trabajo fue el diseño de la aplicación web para el vaciado de los datos dialectológicos recolectados en la primera fase del ALCu. La versión definitiva de la aplicación quedó en dos módulos fundamentales: el de entrada de datos y el de consulta de mapas. También se previó que se pudiera exportar a diferentes formatos la información de la base de datos y los mapas.

En el módulo de entrada de datos fue necesario primero el diseño de apartados para almacenar la información general referida a las localidades encuestadas, a la ficha de los informantes y a las preguntas del ALCu, agrupadas por niveles y campos semánticos. Cumplida esta exigencia, fue posible el vaciado de la información dialectológica de los cuestionarios en ortografía regular y en transcripción fonética. Para esto se adapta el ARFE al sistema de notación del Alfabeto Fonético Internacional (AFI), publicado por la *Asociación de Fonética Internacional* (*International Phonetic Association*, IPA), versión del 2005, ya que su uso está mucho más generalizado. En caso de que el informante hubiera ofrecido una segunda respuesta también se colocaba siguiendo las mismas pautas que en la respuesta principal.

FIGURA 5. Módulo de entrada de las respuestas léxicas en la base de datos geolectal

Las grabaciones recogidas para el nivel fonético primero se convirtieron a formato wav. La edición de los archivos sonoros quedó dividida en dos partes: cuestionario y conversación libre. Para el cartografiado del ALCu solo se emplearon las grabaciones del cuestionario, las cuales fueron segmentadas, luego editadas con el programa PRAAT (versión 2005), para su publicación en el ALCu digital.

Los datos de entonación también fueron procesados y cartografiados. Los referidos a gestos no ameritaron su cartografiado, pero se encuentran disponibles para investigaciones relacionadas con este tópico.

Toda la información hasta aquí mencionada constituye la gran Base de Datos Dialectal (BDG) confeccionada con el motor de bases de datos relacionales MySQL, la cual junto con la base cartográfica en formato SVG permite generar los mapas de forma dinámica e interactiva y exportarlos a PDF, GIF, JPG y PNG.

3.2. Edición y cartografiado de las respuestas

En la conformación de los mapas se tuvieron en cuenta varios aspectos: los parámetros sociales incluidos en la obra, el tipo de respuesta de que se trataba y las características geográficas del terreno donde se iban a cartografiar los datos. Por este motivo se recurrió tanto a los métodos de representación geográfica lingüística tradicional como a los de la pluridimensional. Esta última había sido aplicada por primera vez en el *Atlas Lingüístico Diatópico y Diastrático del Uruguay* (ADDU 2000) de H. Thun y A. Elizaincín, el *Atlas Lingüístico pluridimensional de Nicaragua, nivel fonético* (ALN 2008) de M. A. Rosales Solís, el *Atlas Lingüístico pluridimensional de Costa Rica, nivel fonético* (2010) de L. Vargas, el *Atlas Lingüístico de pluridimensional de Costa Rica, nivel morfosintáctico* (2010) de M. Castillo; los últimos bajo la dirección de M. Á. Quesada Pacheco, entre otros (citado por Montero 2013). Resulta una metodología innovadora porque permite interrelacionar y reflejar en el mapa información diversa: edad, sexo, nivel sociocultural, etc.

En el ALCu se presentan dos tipos de mapas: el sociolingüístico y el lingüístico. En el primero, con la metodología de la geografía lingüística pluridimensional, se reflejan en el mapa la información lingüística según los indicadores sociales considerados, para ello se empleó la siguiente simbología:

- Variable sexual: se emplearon símbolos para distinguir cada sexo: ○ para el femenino y □ para el masculino
- Variable nivel educacional: se representó mediante colores dentro de las figuras que simbolizan la variable sexual: amarillo para el nivel primario y rosa para el nivel secundario.
- Información lingüística: las respuestas fueron identificadas mediante letras, las cuales se colocan dentro del símbolo que corresponde según el tipo de informante que dio la respuesta.

FIGURA 6. Mapa sociolingüístico correspondiente al nivel léxico, pregunta 339 *gemelos*

También aparece en el mapa:

- *En el extremo superior izquierdo*: el referente encuestado, la pregunta realizada al informante, la lámina en caso de que haya sido empleada y el cotejo con otros atlas.
- *En el extremo superior derecho*: aparece la leyenda con el tipo de localidad: rural o urbana, el grupo etario, los símbolos del sexo y el nivel de escolaridad. Debajo, en la misma columna aparecen las respuestas dadas a la pregunta con la letra del alfabeto que la identifica y ordenadas por la frecuencia de uso.
- *En el extremo inferior izquierdo*: la información adicional: segunda respuesta (+) y respuesta de otro informante (~), la cual viene acompañada del número de la localidad donde fue registrado el dato y el número del informante que lo ofreció.

Dadas las características físico-geográficas del país y el volumen de información recogido, el cartografiado del ALCu tuvo la limitación de que no se pudo colocar en un mismo mapa toda la información referida a los tres grupos etarios y las zonas rurales y urbanas. De este modo, la información concerniente para cada pregunta léxica quedó registrada en 6 mapas sociolingüísticos: 3 mapas generacionales rurales y 3 mapas generacionales urbanos.

Ante esta circunstancia, el equipo ideó otro tipo de mapa que permitía una consulta rápida de todos los datos lingüísticos cartografiados, en el que solo se representa el vocabulario recolectado mediante el uso de colores, al cual se le denominó mapa lingüístico.

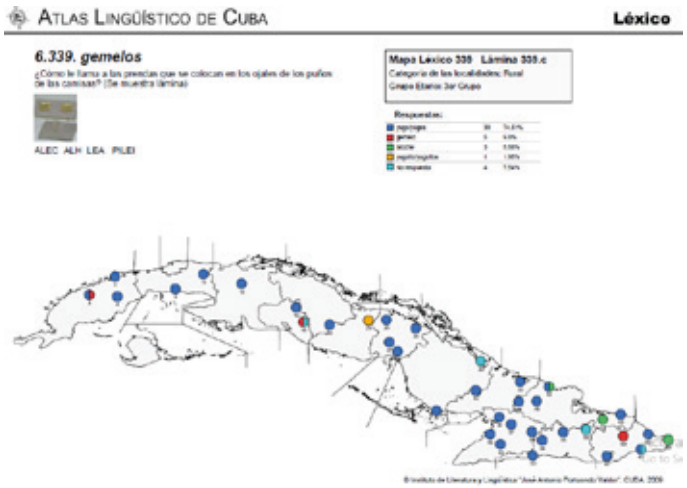


FIGURA 7. Mapa lingüístico correspondiente al nivel léxico, pregunta 339 *gemelos*

3.2.1. Tratamiento de los datos léxicos y gramaticales para el cartografiado

En la transcripción ortográfica de las respuestas dadas por los informantes se diseñó una metodología[6]. De manera general, siempre que las respuestas recogidas aparecieran en diccionarios se respetó la grafía con que fueron asentadas; además, se registraron todas las respuestas de los informantes: simples, compuestas y complejas; así como, todas las respuestas con valor apreciativo, hipocorísticos, acortamientos y préstamos adaptados a la ortografía del español.

Por su parte, las formas verbales conjugadas, participios y gerundios se pasaron a su correspondiente en infinitivo, salvo en casos de interés léxico-semántico o gramatical.

También se consideraron algunos cambios fonéticos en el inventario léxico y gramatical: sustituciones vocálicas y consonánticas, adición de sonidos: epéntesis vocálica y consonántica, aféresis vocálica y vocálica-consonántica, síncopa vocálica, consonántica (excepto /s/ morfemática, /r/ y /l/ distensivas por tratarse de fenómenos generalizados en el español de Cuba) y silábica.

Una vez introducidas todas las respuestas para un determinado mapa, este era revisado por el equipo de trabajo. Luego de ser editado y aprobado el mapa quedaba listo para su consulta y exportación en el módulo de salida de datos.

[6] Para ampliar sobre este particular, ver en la obra en el menú Mapas el apartado El cartografiado de los datos en el *Atlas Lingüístico de Cuba* (ALCu). Metodología.

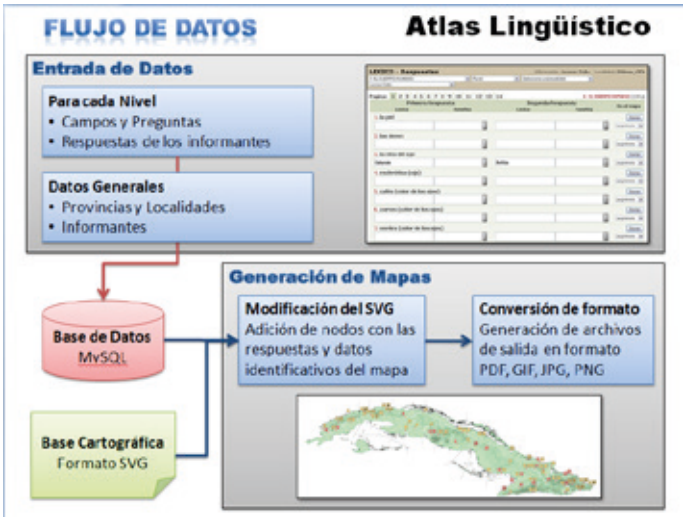


FIGURA 8. Esquema del flujo de datos en el *Atlas Lingüístico de Cuba*

Del total de entradas del cuestionario procesadas, se cartografiaron 1314 respuestas léxicas, 345 gramaticales y 100 fonéticas, de ellas 20 sobre entonación. Es preciso aclarar que no toda la información procesada fue cartografiada, porque en algunos casos esta no resultó interesante desde el punto de vista geosociolingüístico por su escasa o nula variación, y en otros porque los registros estaban incompletos o confusos. Tampoco se procesó el léxico especializado, cuyos datos recolectados fueron exiguos. (Montero 2013)

Entre las versiones de salida de datos se encuentran el *Atlas lingüístico de Cuba* dinámico e *interactivo* (en la intranet del ILL), el *Atlas lingüístico de Cuba* (publicado en soporte digital, DVD−ROM), y la primera versión del Diccionario geolectal de Cuba (DiGeCu) actualmente en proceso de elaboración.

3.3. El *Atlas Lingüístico de Cuba* (ALCu) en su versión de DVD-ROM

El *Atlas Lingüístico de Cuba* (ALCu) en su versión de DVD-ROM está dividido en cinco volúmenes: Atlas Lingüístico de Cuba. Vol. 1. Léxico El hombre y su ambiente social (I), Atlas Lingüístico de Cuba. Vol. 2. Léxico El hombre y su ambiente social (II), Atlas Lingüístico de Cuba. Vol. 3. Léxico Animales silvestres, domésticos y de corral, Atlas Lingüístico de Cuba. Vol. 4. Léxico La agricultura y el tiempo y Atlas Lingüístico de Cuba. Vol. 5. Gramática. Fonética.

La estructura de la obra es similar en cada uno de los volúmenes y está compuesta por un menú con los siguientes apartados: Inicio, El ALCu, Datos, Mapas, Índice, Ayuda y Acerca de.

En el Inicio aparece una introducción a la obra y la referencia del volumen en el que se encuentra el usuario.

FIGURA 9. Inicio del ALCu. Introducción a la obra

En el apartado El ALCu se presenta la historia del proyecto en sus diferentes etapas de desarrollo. Se describe minuciosamente la ejecución del proyecto con el apoyo de imágenes y fotos. Además, a lo largo de los textos aparecen hipervínculos a obras, bibliografías y resultados alcanzados durante la elaboración del proyecto. En este último caso, se presentan los estudios parciales realizados a partir del análisis de los datos: conferencias, cursos, tesis de diploma y maestría, artículos científicos, por solo mencionar algunos. Estos trabajos muestran los primeros pasos en lo que debiera ser la tercera etapa del ALCu, que consiste en el análisis e interpretación de los datos cartografiados.

FIGURA 10. El ALCu. Historia del proyecto

En Datos aparece la información general de la obra, así como el cuestionario del volumen en que se encuentre el usuario.

En el listado de Campos y Preguntas al desplegar el campo seleccionado aparecen el número que le corresponde a la pregunta en el cuestionario general, el referente o entrada seguida de la pregunta realizada al informante, el cotejo con otros atlas, un vínculo a los mapas lingüístico y sociolingüísticos correspondientes y la ilustración empleada en la entrevista en caso de que se haya empleado.

FIGURA 11. Datos. Cuestionario

En este mismo apartado de Datos aparece el listado de localidades organizadas por provincias y punto encuestado. Cada localidad aparece referenciada con el número que le corresponde en el mapa y su tipo: rural o urbana. En ese submenú aparecen los mapas con la distribución de las localidades por todo el territorio nacional.

FIGURA 12. Datos. Localidades

En Informantes aparece la ficha de las personas que participaron como informantes en la obra identificados por un número y con los datos personales: sexo, edad y nivel de escolaridad. El listado de individuos se organiza por provincia y localidad de procedencia.

FIGURA 13. Datos. Informantes

Tanto los listados de campos y preguntas como los de los informantes y las localidades pueden ser exportados a formato PDF.

El menú de Mapas se dedica al cartografiado de las respuestas recogidas en la base de datos geolectal. Por defecto al entrar en este apartado aparece la metodología con que fueron tratados los datos léxicos y gramaticales con sus respectivos ejemplos. La explicación aparece separada en tres submenús: "La ortografía de los datos léxicos en el ALCu", "El tratamiento de los datos léxicos y gramaticales en el ALCu" y "La información fonética en los datos léxicos y gramaticales del ALCu", de manera que los usuarios puedan acceder a la explicación detallada de cada una de las decisiones tomadas por el equipo de trabajo durante el procesamiento de los datos para su posterior cartografiado.

FIGURA 14. Mapas. Metodología para el tratamiento de los datos léxicos y gramaticales del ALCu

Al desplegar tanto el apartado de Mapa sociolingüístico como Mapa lingüístico el usuario tiene acceso a esta información, la cual por defecto comienza en la primera pregunta del primer campo semántico del volumen consultado.

La metodología y ambos tipos de mapas también pueden ser exportados a formato PDF.

Una de las posibilidades que ofrece la aplicación del ALCu es que los usuarios pueden confeccionar sus mapas según sus intereses, de modo que se puede visualizar la información por grupos de edades, tipo de localidad, pregunta o campo semántico. Cada mapa refleja en la parte inferior la leyenda y las localidades con su respectivo número.

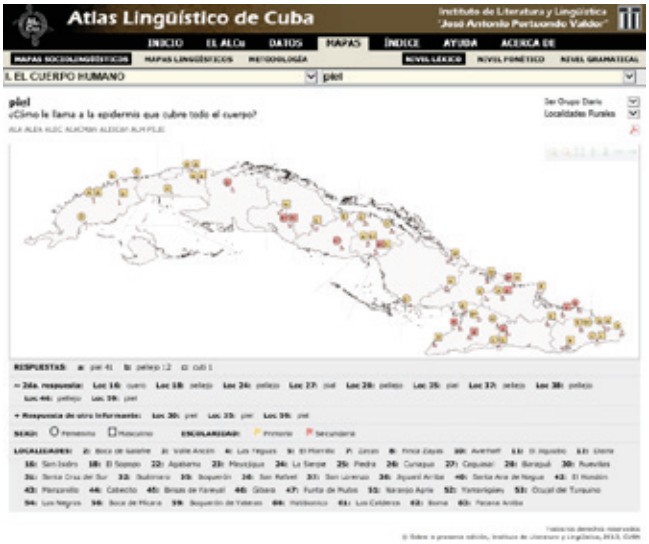


FIGURA 15. Mapa sociolingüístico del ALCu. I Cuerpo humano, pregunta 1 *piel*

En el mapa sociolingüístico aparecen todas las respuestas según el grupo etario y la localidad seleccionados, a diferencia del mapa lingüístico en el que además de todas las respuestas, se puede visualizar una en particular, lo cual permite realzar o descartar la información que desee el usuario. En ambos mapas las respuestas vienen acompañadas con su frecuencia de uso.

FIGURA 16. Mapa lingüístico del ALCu. I Cuerpo humano, pregunta 1 *piel*

Uno de los apartados más útiles de la obra es el Índice, en el que se recogen todas las respuestas del volumen ordenadas alfabéticamente. De manera que los usuarios tienen acceso al vocabulario general de la obra, información útil cuando se busca un fenómeno determinado. Por ejemplo: alteraciones fonéticas con incidencia en el léxico, presencia de extranjerismos en el español de Cuba, vitalidad de vocablos de procedencia indígena, respuestas complejas, entre otros. En este apartado al hacer clic en una de las respuestas se muestra un recuadro con la información del número de la pregunta del Cuestionario en donde aparece y los informantes que la dieron. Todos estos datos tienen un hipervínculo al mapa sociolingüístico respectivo. Para cada uno de los listados según la letra que corresponda, se registra el total de respuestas dadas.

FIGURA 17. Listado de respuestas que comienzan con la letra A, de la sección Índice del ALCu

A continuación, en Ayuda, se presenta el Manual de Usuario, con la explicación detallada de cómo se opera la obra. En Acerca de aparecen los créditos de la obra, así como la nómina de autores y colaboradores que intervinieron en sus dos etapas de realización.

Manual de usuario del ALCu

El Atlas Lingüístico de Cuba (ALCu) es una aplicación web diseñada para el cartografiado automático de los mapas léxicos, gramaticales y fonéticos reunidos en el ALCu, tanto lingüísticos como sociolingüísticos. Debe ejecutarse en el navegador Mozilla Firefox bajo los sistemas operativos 32 y 64 bits, aunque se pudieran utilizar otros navegadores.1Para la ejecución de dicho programa es aconsejable disponer, como mínimo, de un procesador Intel Celeron 1.7, o su equivalente AMD, con 256 MB de RAM (aunque se recomienda contar con 512 MB).

Es necesario un lector de DVD pues la obra ha sido concebida para su publicación en soporte digital DVD-ROM. Los cinco volúmenes son: *Atlas Lingüístico de Cuba. Vol. 1. Léxico El hombre y su ambiente social (I), Atlas Lingüístico de Cuba. Vol. 2. Léxico El hombre y su ambiente social (II), Atlas Lingüístico de Cuba. Vol. 3. Léxico Animales silvestres, domésticos y de corral, Atlas Lingüístico de Cuba. Vol. 4. Léxico La agricultura y el tiempo, y Atlas Lingüístico de Cuba. Vol. 5. Gramática. Fonética.*

La aplicación dispone de un **Menú principal** en la parte superior con acceso a siete secciones principales: Inicio, El ALCu, Datos, Mapas, Índice, Ayuda y Acerca de (Fig.1)

Figura 1. Menú principal

En la sección **Inicio** se hace una breve presentación sobre la información contenida en el *Atlas Lingüístico de Cuba*, además de explicar la estructura de la aplicación (Fig. 2).

FIGURA 18. Manual de usuario

FIGURA 19. Créditos de la obra

4. REFLEXIÓN FINAL

El análisis completo del ALCu en sus etapas de desarrollo permite identificar que en la obra existe una dualidad irrefutable: por una parte, mantiene aspectos de la dialectología tradicional y por la otra, incorpora los avances más novedosos en los parámetros para efectuar el trabajo de campo, el tratamiento de los datos dialectales y su cartografiado automático.

En cuanto al mantenimiento de los aspectos tradicionales se encuentran: la observancia de los requerimientos para la selección de los informantes: oriundez y permanencia en la localidad encuestada, así como sus escasos o básicos estudios. También están el empleo de los lineamientos para seleccionar las localidades: historia, tradiciones, ubicación geográfica; y la confección de un cuestionario dialectal que se ajusta a la realidad y al contexto social del territorio y sobre todo a la vida rural.

Dentro de las novedades de la obra se encuentran la ampliación de la red de poblados, tanto en la densidad de localidades como en la selección de puntos rurales y urbanos; la estratificación de los informantes según sexo, edad y nivel de escolaridad.

Todos estos aspectos novedosos enriquecen la información recopilada pues permiten obtener una visión más completa y real de las características de la modalidad cubana del español.

Por su parte, la aplicación informática que da vida al ALCu constituye un salto cualitativo en los estudios de este tipo y envergadura porque incorpora los avances más recientes en el cartografiado automático de datos lingüísticos. Si bien en la actualidad no existe consenso para la representación en un mismo mapa de información diversa: lingüística, social y de localización, el ALCu ofrece soluciones a muchas interrogantes y problemas surgidos sobre este tópico. La experiencia del trabajo de los especialistas cubanos constituye un referente por las decisiones tomadas y por la adaptación a la realidad geográfica del territorio estudiado.

El tratamiento de la información dialectal para su cartografiado definitivo pudiera recibir cuestionamientos por la manipulación de los datos “puros” recogidos en el terreno.

Sin embargo, el equipo del ALCu lo tomó como una alternativa para lograr reunificar información diversa que no tenía un valor especial para el tipo de dato que se quería representar, por ejemplo: las variantes flexivas en verbos recogidos en el apartado léxico. Además, mediante el tratamiento de los datos léxicos y gramaticales se logró representar mejor en el mapa este tipo de respuestas al no verse tan cargado de información superflua.

En cuanto a la elaboración de mapas según los intereses de los usuarios, se aprecia que la obra aporta diferentes posibilidades al permitir crear mapas atendiendo a las variables sociales empleadas en la obra, así como, al tipo de localidad. Además, el sistema simbólico de representación permite una visualización de los datos dialectales más fácil y cómoda para la investigación.

Por último, la posibilidad de contar con el Índice posibilita a los usuarios localizar voces concretas y fenómenos, a la vez de permitir el vínculo con el mapa correspondiente, de manera que se logra interrelacionar la información buscada con el resto de la obra.

Son múltiples las utilidades investigativas que han surgido y surgirán a partir del ALCu. Uno de los ejemplos concretos es que en estos momentos en el ILL se lleva a cabo el Diccionario geolectal de Cuba, a partir de las respuestas para el nivel léxico recogidas en el ALCu. Gracias a que ya se tiene toda la información dialectal digitalizada, con su respectiva referencia y localización, el trabajo de los investigadores ha sido más fácil y se han logrado avances más rápidos en la culminación por etapas de las tareas del trabajo.

Ver materializado y concluido el Atlas Lingüístico de Cuba (ALCu) significa la continuación y apertura de nuevos emprendimientos en la investigación de corte dialectal en Cuba. La obra es el resultado de la intensa labor de los especialistas involucrados que, en la mayoría de los casos, como sucede en obras de tal envergadura, trabajan más por amor, que por beneficios económicos. De eso se trata la investigación dialectal: una obra de amor.

REFERENCIAS BIBLIOGRÁFICAS

CÁRDENAS MOLINA, G., A. J. MENÉNDEZ PRYCE & L. SANTANA GONZÁLEZ (1999): *Glosario de regionalismos y ruralismos de Cuba*. [inédito]

CHOY LÓPEZ, L. R. (1989): "Zonas dialectales en Cuba", *Anuario L/L* 20, pp. 83-100.

ILL (Instituto de Literatura y Lingüística "José Antonio Portuondo Valdor) (2013): *Atlas Lingüístico de Cuba*. La Habana: Instituto de Literatura y Lingüística "José Antonio Portuondo Valdor" (ILL). DVD-ROM.

ILL (2009-2013): *Atlas Lingüístico de Cuba-Dinámico (ALCu-Din)*. http://illwebserver/atlas-dinamico/ .

GARCÍA RIVERÓN, R. (1989): "Caracterización geolingüística del español de Cuba", *El español en Cuba. Anuario*, pp. 69-92.

MONTERO BERNAL, L. E. (2013): "El *Atlas Lingüístico de Cuba* (ALCu). Historia del proyecto", in *Memorias VIII Conferencia Internacional Lingüística 2013*. La Habana: Instituto de Literatura y Lingüística "José Antonio Portuondo Valdor" (ILL), CD-ROM.

VALDÉS BERNAL, S., L. E. MONTERO BERNAL, M. MORÓN GARCÍA, L. SANTANA GONZÁLEZ & A. MENÉNDEZ PRYCE (2007): *Visión geolectal de Cuba*. Frankfurt: Peter Lang.

EL PROYECTO PRESEEA: DESARROLLOS ANALÍTICOS

The PRESEEA Project: Analytical developments

FRANCISCO MORENO FERNÁNDEZ
Coordinador general de PRESEEA, Universidad de Alcalá

ANA MARÍA CESTERO MANCERA
Coordinadora técnica de PRESEEA, Universidad de Alcalá

Resumen
El proyecto PRESEEA reúne y ofrece materiales procedentes de comunidades urbanas hispanohablantes de muy distintos espacios geográficos. Se trata de un proyecto de naturaleza sociodialectológica cuyo principal fundamento metodológico es la comparación. Estas páginas tienen como fin exponer cuáles son los principales desarrollos analíticos que ha conocido el proyecto PRESEEA a lo largo de su trayectoria. Ello implica el desarrollo de herramientas de almacenado y recuperación de los materiales reunidos por los equipos de investigación y la propuesta de análisis sociodialectales comparados sobre algunos de los numerosos aspectos lingüísticos que ofrecen las muestras de lengua reunidas, así como la implementación de recursos analíticos de base informática. Este trabajo también ofrece información sobre diversos elementos cuantitativos relativos a la dinámica y el alcance del proyecto PRESEEA.

Palabras clave: sociolingüística, PRESEEA, variación lingüística, lingüística comparada

Abstract
The PRESEEA project brings together and offers materials from Spanish-speaking urban communities in very different geographical spaces. PRESEEA is a socio-dialectological project whose methodological basis is comparison. The purpose of these pages is to explain the main analytical developments that the PRESEEA project has known throughout its trajectory. This implies the development of tools for storing and retrieving the materials gathered by the research teams and the proposal of comparative socio-dialectal analysis on some of the numerous linguistic aspects offered by the language samples gathered, as well as the implementation of computer-based analytical resources. This paper also provides information on various quantitative elements related to the dynamics and scope of the PRESEEA project.

Keywords: sociolinguistics, PRESEEA, linguistic variation, comparative linguistics

1. INTRODUCCIÓN

El "Proyecto para el Estudio Sociolingüístico del Español de España y América", conocido por sus siglas (PRESEEA), tiene una trayectoria de largo recorrido que lo ha convertido

en una iniciativa de referencia para el estudio de la lengua española hablada contemporánea. Como es sabido (Moreno Fernández 1996), el proyecto nació en respuesta a una demanda planteada por Carmen Silva-Corvalán en 1992, aunque ha ido enriqueciendo sus objetivos conforme los equipos participantes se han coordinado y se han ido publicando materiales y trabajos de investigación.

Claramente situado en el ámbito de la sociolingüística, el proyecto PRESEEA se distingue de otros proyectos de este tipo por su capacidad para reunir y ofrecer materiales procedentes de comunidades urbanas hispanohablantes de muy distintos espacios geográficos. De este modo, una iniciativa de base sociolingüística se convierte en un proyecto de naturaleza sociodialectológica cuyo principal fundamento metodológico es la comparación. La variación de la lengua se aborda, por un lado, en su complejidad sociodialectal y, por otro, desde el análisis de sus semejanzas y desemejanzas. Todo ello implica la adopción de una perspectiva social y comparada. En 2002, Sali Tagliamonte recuperó para la sociolingüística el concepto de «comparación» o «comparatismo», que durante 200 años había demostrado una gran capacidad explicativa. Desde una perspectiva actual, la sociolingüística exige una metodología capaz, por un lado, de desenmarañar las múltiples influencias que los hechos lingüísticos reciben y, por otro lado, de explicar la realidad lingüística a través de la comparación de sus manifestaciones (Tagliamonte 2002; Claes 2014).

Estas páginas tienen como fin exponer cuáles son los principales desarrollos analíticos que ha conocido el proyecto PRESEEA a lo largo de su trayectoria. Por desarrollos analíticos entendemos aquellos objetivos, trabajos e instrumentos que han surgido como consecuencia de las investigaciones sociolingüísticas de PRESEEA a lo ancho de la geografía hispanohablante. Ello implica el desarrollo de herramientas de almacenado y recuperación de los materiales reunidos y la propuesta de análisis sociodialectales comparados, cualitativos y cuantitativos, sobre algunos de los numerosos aspectos lingüísticos accesibles en las muestras, así como la implementación de recursos analíticos de base informática que facilitan el estudio de los datos. Complementariamente, este trabajo ofrece información sobre diversos elementos cuantitativos que permitirán conocer mejor la dinámica y el alcance del proyecto PRESEEA.

2. Los objetivos de PRESEEA

El objetivo prioritario de la red de equipos PRESEEA es la constitución de un corpus de lengua hablada que incluya muestras representativas de las comunidades hispanohablantes investigadas, teniendo en cuenta su diversidad geográfica y las variables sociales más claramente correlacionadas con la conducta lingüística de los hablantes. De este modo, PRESEEA aspira a poner en manos de los investigadores unos materiales representativos de una amplia geografía, sociolingüísticamente definidos, comparables, de origen contrastado, bien transcritos, en las mejores condiciones técnicas y de la forma más eficaz posible. Las bases metodológicas del proyecto fueron publicadas en 1996 (Moreno Fernández 1996) y revisadas en 2003 (PRESEEA 2003). En ellas se establece que la finalidad del proyecto es coordinar investigaciones sociolingüísticas de los países americanos y de España para facilitar la comparabilidad de los estudios y el intercambio de información básica.

Desde el comienzo del proyecto, en 1996, hasta la actualidad, los trabajos se han distribuido en diferentes etapas. En un primer periodo, concluido en 2010, la finalidad prioritaria de PRESEEA fue la creación de una amplia red de equipos de investigación sociolingüística. La adscripción a la red internacional PRESEEA es de carácter voluntario

y simplemente implica el seguimiento de los criterios metodológicos y técnicos comunes, así como la puesta a disposición para toda la red de los materiales reunidos en cada comunidad. Al tiempo que la red iba ampliando su número de miembros, los trabajos de los equipos propiciaron la consecución de su principal fin: la recogida y almacenamiento de materiales de lengua hablada procedentes de conversaciones semidirigidas y reunidos a partir de criterios sociolingüísticos.

Hasta el año 2010 la red PRESEEA aspiraba a formar un corpus de español hablado que permitiera abordar los análisis comparados que los investigadores, pertenecientes o ajenos a la red, consideraran de mayor interés. Sin embargo, los resultados obtenidos en esas fechas y la positiva dinámica de investigación que PRESEEA había generado, al estrechar las relaciones entre equipos procedentes de una decena de países hispanohablantes, condujeron a un replanteamiento del calendario de investigación y de una parte de los objetivos generales del proyecto. Efectivamente, a partir de 2010 se hicieron patentes tres necesidades:

1. La ampliación de la red de equipos PRESEEA.
2. La continuidad de las tareas de recogida de materiales de lengua hablada sin un límite temporal fijo.
3. El aumento de la capacidad de análisis de los propios equipos para aprovechar en lo posible el importante volumen de materiales disponibles.

En relación con los dos primeros puntos, hay que tener en cuenta que, en la etapa inicial del proyecto, se pensó en la conveniencia de circunscribir los materiales reunidos a un periodo temporal cerrado y no excesivamente amplio. De ahí que se pensara en 2010 como límite, dado que los primeros datos se reunieron en la década de los noventa. Esta decisión fue revocada por varias razones, entre las que llegaron a pesar más las siguientes: en primer lugar, la propia dinámica internacional de PRESEEA aún mostraba un gran potencial para propiciar la recogida de materiales por parte de equipos que no habían podido afrontarla previamente; en segundo lugar, en los análisis comparados, la mayor complicación surge del desconocimiento del origen de los materiales; ahora bien, si la fecha de los materiales es bien conocida, sea cual sea, sería responsabilidad de los propios estudiosos valorar la importancia que pueda tener la variable temporal y el modo de tratarla. No habría razones objetivas, pues, para limitar el proceso de recogida de materiales a un periodo concreto.

En lo que se refiere al desarrollo analítico en PRESEEA, el mismo documento de bases metodológicas publicado en 1996 aludía a la necesidad del análisis e interpretación de los materiales lingüísticos reunidos dentro del proyecto y señalaba la libertad de la que disfrutaban los investigadores vinculados a PRESEEA a la hora de afrontar esas tareas. Esa libertad se ejerció desde el inicio del proyecto por parte de numerosos investigadores, que planificaron sus estudios de los materiales de PRESEEA según sus particulares intereses. Sin embargo, las posibilidades que ofrece el trabajo en red permitían ir más allá en la forma de proceder a los análisis, buscando la formación de grupos especializados en determinadas líneas de investigación e interesados por objetos de estudio específicos. Así comenzaron a formarse grupos de trabajo en torno a problemas lingüísticos concretos, que han ido uniendo sus esfuerzos para conseguir financiación y que ya han confeccionado varios volúmenes monográficos, dedicados a diversas variables lingüísticas, así como a otros aspectos, como la atenuación o las actitudes (San Martín Núñez y Guerrero González 2016; Cestero Mancera y Moreno Fernández 2017; Cestero Mancera y Paredes García 2018; Molina Martos, Paredes García

y Cestero Mancera 2020; Villena Ponsoda en preparación; Albelda Marco *et alii* en preparación).

FIGURA 1. Puntos encuestados para PRESEEA (febrero 2019)

3. RECOGIDA, ALMACENAMIENTO Y RECUPERACIÓN DE MATERIALES: ASPECTOS CUANTITATIVOS

La red de investigación PRESEEA cuenta con una metodología propia para la recogida, almacenamiento y recuperación de materiales de lengua española hablada. La recogida de materiales es realizada por cada uno de los equipos miembros de la red. Con fecha enero de 2019, como se detalla en la sección de los equipos de la página PRESEEA, los equipos que han trabajado o están trabajando en tareas de recolección de materiales sociolingüísticos son 44, procedentes de 14 países: Argentina, Chile, Colombia, Cuba, Ecuador, España, Estados Unidos, Guatemala, México, Paraguay, Perú, Puerto Rico, Uruguay, Venezuela.

Los materiales recolectados reciben un tratamiento de transcripción, etiquetado, marcado y revisión antes de ser finalmente almacenados (PRESEEA 2008). Los equipos con corpus recolectados y etiquetados, ya publicados o en fase de publicación son los siguientes, por orden alfabético: Alcalá de Henares (España), Barranquilla (Colombia), Caracas (Venezuela), Cartagena de Indias (Colombia), Granada (España), Guatemala (Guatemala), La Habana (Cuba), Las Palmas de Gran Canaria (España), Lima (Perú), Madrid (España), Málaga (España), Medellín (Colombia), Mérida (Venezuela), México (México), Monterrey (México), Montevideo (Uruguay), Santiago de Chile (Chile), Santiago de Compostela (España) y Valencia (España).

Por otro lado, los equipos cuyos corpus se encuentran en distintas fases de elaboración y revisión son los siguientes: Asunción (Paraguay), Barcelona (España), Bogotá (Colombia), Buenos Aires (Argentina), Cádiz (España), Cipolletti (Argentina), Culiacán (México), Gijón (España), Guadalajara (México), Lérida (España), Mérida (México), Mérida (México), Mexicali (México), Nueva York (Estados Unidos), Oviedo (España), Palma de Mallorca (España), Pereira (Colombia), Puebla (México), Quito

(Ecuador), San Juan de Puerto Rico (Puerto Rico), Santander (España), Sevilla (España), Tunja (Colombia), Valledupar (Colombia), Valparaíso (Chile) y Zaragoza (España).

Cada uno de los equipos de PRESEEA es responsable del tratamiento dado a sus materiales, si bien la coordinación general y técnica del proyecto desarrollan una labor de asesoramiento de los trabajos y revisión final de los materiales que contribuye a que unos y otros cumplan con las condiciones previstas en la metodología general. Asimismo, buena parte de los equipos cuenta con su propia plataforma informática o página web en la que los materiales reunidos se ponen a disposición de la comunidad investigadora y a la que se remite desde la pestaña correspondiente a cada equipo de la página de PRESEEA.

Hasta el momento, el trabajo coordinado de esta amplia red de investigación ha proporcionado materiales cuyo recuento ofrece números dignos de tenerse en cuenta. Los equipos que ya han concluido las tareas de recolección, transcripción, revisión y almacenamiento (un total de 17), han permitido construir un corpus de muestras cuyo tamaño arroja las cifras que aparecen en la tabla 1.

Ciudades	**número de informantes**	**número de palabras**	**número de minutos**
Alcalá de Henares (España)	54	484623	2798
Caracas (Venezuela)	108	1094353	6169
Granada (España)	54	406339	2217
Guadalajara (México)	72	662076	4002
Guatemala (Guatemala)	72	512302	3984
La Habana (Cuba)	108	751486	5220
Lima (Perú)	54	396657	2602
Madrid (España)	108	1027242	6184
Málaga (España)	72	689508	3962
Medellín (Colombia)	72	485338	3510
Mexicali (México)	54	368204	2396
Ciudad de México (México)	108	1095001	6692
Monterrey (México)	108	1113286	7463
Montevideo (Uruguay)	54	410967	2369
Santiago de Chile (Chile)	108	887328	6366
Santiago de Compostela (España)	54	587045	3411
Valencia (España)	72	484560	2623
TOTALES (17 ciudades)	**1332**	**11456315**	**71968 (1200 h.)**

TABLA 1. Números totales de informantes, palabras y minutos del corpus PRESEEA de 17 ciudades

La tabla 1 nos permite destacar los siguientes aspectos relativos al tamaño y las características del corpus PRESEEA disponible:

a) Por el número de informantes, se trata del mayor corpus sociodialectal reunido hasta el momento.

b) Las horas de grabación disponibles constituyen una fuente de primer orden para el análisis de cuestiones fonéticas y prosódicas, a pesar de la natural irregular calidad de las grabaciones.

c) Por número de palabras, el tamaño del corpus PRESEEA supera ampliamente al componente oral de cualquier corpus de referencia.

Aunque se trata de solo una parte del corpus PRESEEA, el tamaño del material recogido en esas 17 ciudades nos lleva a considerar que estamos ante un corpus con número de informantes y un total de palabras y minutos suficiente para llevar a cabo investigaciones sobre distintos fenómenos y obtener resultados relevantes y significativos.

4. DESARROLLO DEL PORTAL PRESEEA

El portal informático de PRESEEA es uno de los instrumentos más representativos y valiosos del proyecto, aunque no el único. En su forma actual, inició su presencia en la red en el año 2014. Las primeras y principales funciones que cumple el portal son la de informar y la de visibilizar a los equipos integrados en la red (figura 1). La información proporcionada sobre cada equipo permite acceder directamente al contacto de sus responsables, así como conocer las páginas o portales donde se ofrecen otras informaciones particulares, el estado del corpus y las publicaciones derivadas de estudios enmarcados en PRESEEA. Además de la información, la visibilidad proporcionada por el portal PRESEEA es importante con vistas a la conciencia de pertenencia a la red y a la solicitud de ayudas financieras de naturaleza nacional o internacional, dada la creciente importancia que se concede a los proyectos de investigación desarrollados desde redes internacionales. El proyecto en su conjunto cuenta también con la evaluación positiva del Comité de Ética de la Investigación de la Universidad de Alcalá, entidad desde la que se coordina.

FIGURA 2. Página de información de equipos integrantes de la red. Portal PRESEEA (http://preseea.linguas.net/Equipos.aspx)

Por otro lado, como instrumento de coordinación e información, el portal PRESEEA ofrece de forma abierta toda la documentación de investigación generada por el proyecto, tanto para la fase de adscripción de equipos, como para la aplicación de la metodología común y para el desarrollo de análisis coordinados, entre otros fines. Los documentos de trabajo no solamente resultan de utilidad para los equipos que los han generado y aplicado, sino que se presentan abierta y gratuitamente para que cualquier investigador ajeno al proyecto pueda hacer un seguimiento del desarrollo de los trabajos o aplicar las técnicas propuestas desde PRESEEA sobre sus propios materiales. Sirvan como ejemplo, aparte del documento de metodología general o el documento de marcas y etiquetas mínimas para la transcripción, las guías para el estudio de la prosodia, del presente narrativo o de la atenuación, o las propuestas de codificación de variables para el estudio de la expresión del sujeto, de los usos del verbo *haber* o de /d/ intervocálica, como se puede ver en la figura 3.

FIGURA 3. Página de metodología. Portal PRESEEA
(http://preseea.linguas.net/Metodología.aspx)

Pero, sin duda, uno de los principales atractivos del portal informático está en el acceso al corpus PRESEEA a través de una herramienta de búsqueda y recuperación de los materiales de lengua hablada recolectados por los equipos de la red (figura 3). La herramienta permite buscar cualquier cadena de caracteres seleccionando el perfil de los hablantes-informantes por su sexo, edad, nivel de estudios y ciudad de procedencia. Los materiales transcritos se recuperan de forma gratuita e inmediata a través de archivos descargables TXT que permiten posteriores búsquedas y una contextualización de los ejemplos según las necesidades de los investigadores. Los documentos con las transcripciones se acompañan de muestras de audio de aproximadamente diez minutos del comienzo de cada entrevista, que permiten oír un fragmento de las grabaciones.

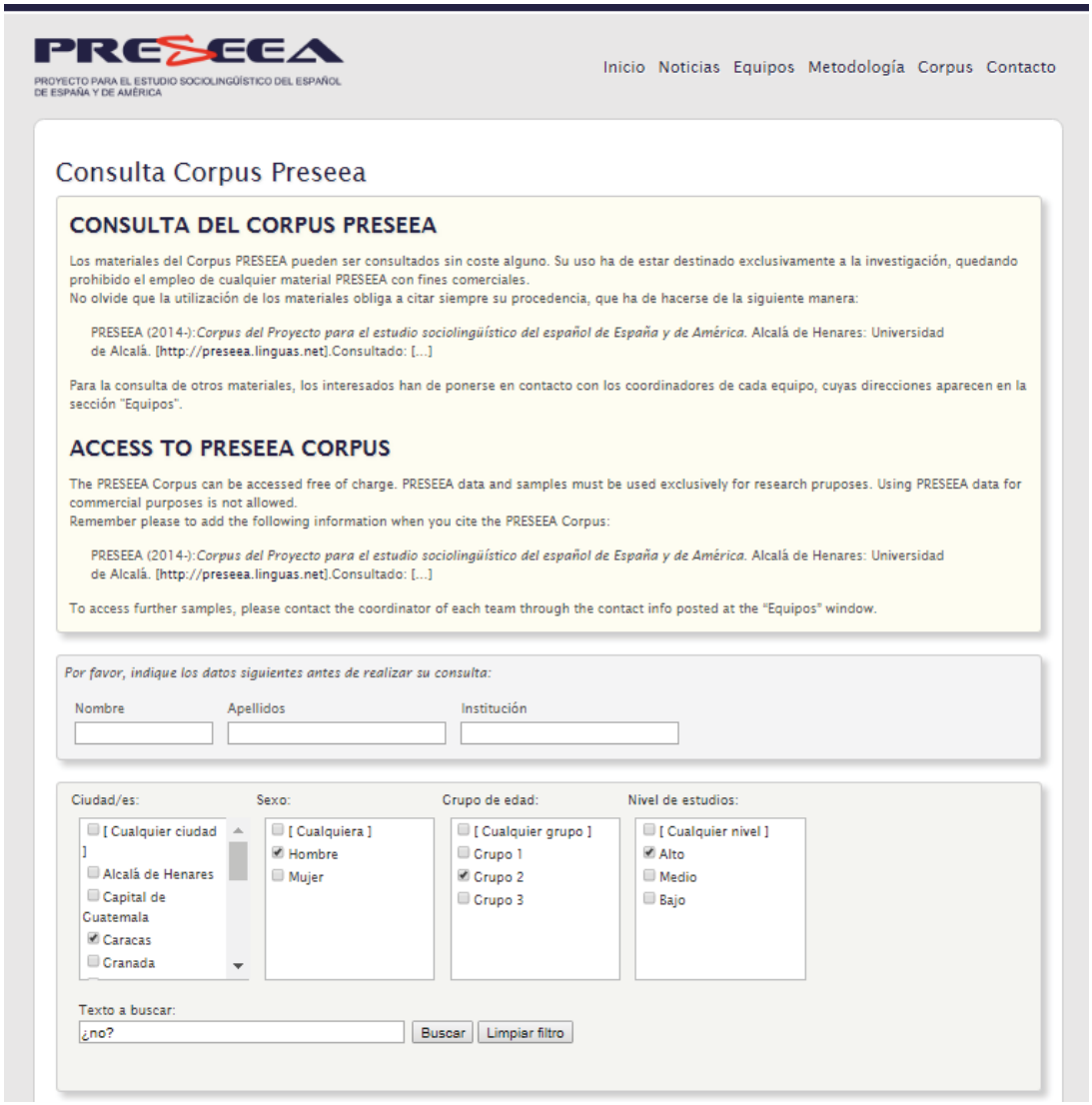


FIGURA 4. Corpus PRESEEA en abierto. Portal PRESEEA (http://preseea.linguas.net/Corpus.aspx)

Los materiales del corpus PRESEEA ofrecidos a través del portal no constituyen el corpus completo, sino una muestra de 18 informantes de cada una de las comunidades investigadas: 1 por cada cuota de estratificación manejada en la muestra, según sexo, edad y nivel de estudios. Asimismo, se ofrecen 18 audios de cada una de las comunidades, que pueden ser utilizados con fines de enseñanza y, eventualmente, de investigación. Estos archivos tienen formato MP3 y una duración aproximada de 10 minutos y un peso de unos 15Mb. En la actualidad (febrero 2019), el portal PRESEEA ofrece muestras en transcripción y audio de 17 comunidades hispanohablantes, cuyo tamaño puede verse en la tabla 2.

Ciudades	MUESTRAS		
	número de informantes	número de palabras	número de minutos
Alcalá de Henares (España)	18	161541	916
Caracas (Venezuela)	18	182392	1028
Granada (España)	18	135446	739
Guadalajara (México)	18	165519	975
Guatemala (Guatemala)	18	128075	996
La Habana (Cuba)	18	125248	870
Lima (Perú)	18	132219	867
Madrid (España)	18	171207	1117
Málaga (España)	18	172377	990
Medellín (Colombia)	18	121334	877
Méxicali (México)	18	122735	799
Ciudad de México (México)	18	182500	1115
Monterrey (México)	18	185548	1100
Montevideo (Uruguay)	18	136989	790
Santiago de Chile (Chile)	18	147888	1061
Santiago de Compostela (España)	18	195682	1137
Valencia (España)	18	121140	656
TOTALES (17 ciudades)	306	2587840	16033 (267 h.)

TABLA 2. Tamaño del subcorpus de 17 muestras ofrecido a través del portal PRESEEA (febrero 2019)

En lo que se refiere a la proyección y visibilidad del portal PRESEEA, el analizador SISTRIX nos permite conocer datos como los siguientes (febrero 2019):

a) Las palabras clave de mayor peso en las búsquedas son *linguas*, *corpus* y *equipos*.

b) El portal aparece vinculado a 478 enlaces, 71 dominios y 63 redes.

c) En relación con las señales sociales, el portal se vincula a 228 señales de *Facebook*.

Estos son indicadores estandarizados. Las señales de *Facebook* tienen un valor relativo porque son pocas, en comparación con el número de páginas de otros sectores, esas apariciones en un medio no especializado significan que el proyecto va algo más allá

de la investigación sociolingüística. En cuanto a los países desde los que se accede, su número y ubicacion son relevantes porque ofrecen una imagen del alcance geográfico del proyecto. La evolución de la visibilidad del portal PRESEEA, solamente para búsquedas efectuadas desde España, es la que se refleja en la figura 5.

FIGURA 5. Evolución de la visibilidad del portal PRESEEA desde España.
Fuente: SISTRIX (9 de febrero de 2019)

Desde una perspectiva internacional, los enlaces vinculados al portal PRESEEA están radicados en los siguientes países, por número de enlaces más significativos: España, Estados Unidos, Alemania, México, Colombia, Perú, Chile, Suiza, Holanda, Bélgica, Francia, Reino Unido.

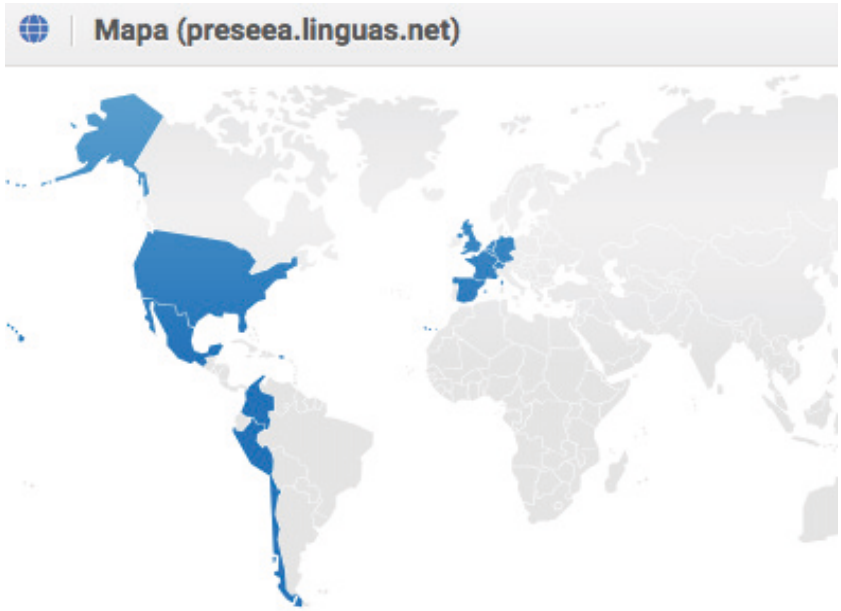


FIGURA 6. Mapa de visibilidad mediante enlaces del portal PRESEEA (http://preseea.linguas.net/).
Fuente: Informe SISTRIX (febrero 2019)

Finalmente, es importante destacar que, por acuerdo con la Real Academia Española, los materiales de PRESEEA posteriores al año 2000 se integran en el *Corpus del Español del Siglo XXI (CORPES XXI)* (Real Academia Española s.f.), en la parte oral del corpus de referencia. Dentro de este corpus académico, los materiales de PRESEEA

son etiquetados gramaticalmente, de modo que, a través del portal del CORPES XXI, es posible proceder a búsquedas mediante criterios lingüísticos más refinados.

5. Análisis de materiales. Los análisis coordinados

Como se ha mencionado previamente, PRESEEA tiene como finalidad realizar una serie de investigaciones sociolingüísticas con iguales procedimientos analíticos, de manera que se pueda comparar datos, así como identificar y explicar patrones sociolingüísticos y geolectales, además de procesos de variación, convergencia, divergencia y cambio (Silva-Corvalán 1994; Moreno Fernández 1996). El estudio coordinado del español hablado comenzó en 2010, fecha en la que se completó la primera fase del corpus, y partió de fenómenos priorizados de los distintos niveles lingüísticos. Como consecuencia de esa coordinación, varios equipos de investigación han concluido ya investigaciones sociolingüísticas que han revelado resultados de gran interés sobre fenómenos como la elisión de /d/ intervocálica, la variación de la /s/ implosiva, el uso de *haber* impersonal, la expresión del sujeto nominal y pronominal y, de manera novedosa por tratarse de una estrategia discursiva, la atenuación. Además, la nómina de fenómenos priorizados se ha ido ampliando progresivamente; en estos momentos, incluye prosodia basada en el uso, yeísmo, leísmo, laísmo y loísmo, dequeísmo y queísmo, disponibilidad léxica, formas de tratamiento, diminutivo, presente narrativo, expresión del tabú y apéndices interrogativos de control de contacto.

Por otra parte, se han afrontado un gran número de trabajos sobre temas diversos a partir de los subcorpus del proyecto, en atención a los intereses personales de los miembros de los equipos. Gracias a ellos, tenemos datos importantes sobre el funcionamiento, en puntos geolingüísticos determinados, de diversas variables fónicas (variación de /-r/, /f-/ y /ĉ/, variación acentual, realizaciones de /x/, distinción de sibilantes, ataques y codas silábicos; acústica de las vocales), morfosintácticas (perífrasis verbales de infinitivo, cláusulas relativas-adjetivas de preposición más (artículo) más *que*, pronombres posesivos, tiempos verbales del pasado, oraciones subordinadas de finalidad, discurso directo e indirecto, usos preposicionales, construcción *sino* más (*que*) más verbo, variación proposicional en estructuras biactanciales estativas, orden de palabras, estructuras hipotácticas adverbiales, futuros morfológico y perifrástico, oraciones hendidas, oraciones pasivas, variación de ser y estar con adjetivos de edad o perífrasis deber y deber de), léxico-semánticas (anglicismos, regionalismos, neologismos, fraseología, léxico de los colores, transferencias y cambios de código) y pragmático-discursivas (fenómenos estructurales tales como el intercambio de turnos de habla, la interrupción y la producción de apoyos, estructura de la información, recursos de función fática, signos paralingüísticos como la risa, marcadores discursivos, reformulación, intensificación, ironía, recursos para expresar desacuerdo y actos disentivos, expresión de la impersonalidad, expresión de la evidencialidad, estructuras narrativas, argumentación, modalización, secuencias descriptivas, cortesía y descortesía). Todas las investigaciones realizadas en el marco de PRESEEA combinan diversos procedimientos analíticos considerados de utilidad en la sociolingüística de la variación. Incluso, en algunas de ellas, se han desarrollado herramientas específicas de transformación, marcación y etiquetado (Villena Ponsoda *et alii* 2010), codificación y cuantificación, que ha supuesto un avance metodológico en la disciplina.

Cada estudio coordinado comienza con la elaboración de una ficha de análisis, diseñada, testada y validada por miembros de diferentes equipos, que constituye la base de la *Guía de investigación* del fenómeno, y que se encuentra disponible, cuando se considera terminada, en la página electrónica de PRESEEA. La ficha permite codificar

cada ocurrencia a partir de variables y variantes lingüísticas y extralingüísticas, y efectuar análisis cuantitativos de distinto tipo, que dependen, en cada caso, de la naturaleza del fenómeno. Así, los estudios sociopragmáticos, como el de la atenuación, no precisan necesariamente de estadística inferencial, de manera que la estadística descriptiva y las pruebas de significación, junto con tablas de contingencia y un análisis cualitativo de base, permiten obtener datos relevantes y significativos acerca del funcionamiento de la estrategia pragmática. Los estudios de fonética y morfosintaxis sí precisan del uso combinado de estadística descriptiva e inferencial y, en algunos casos, del empleo de nuevos programas o de la utilización de pruebas y aplicaciones que no son habituales en lingüística y que permiten avanzar y profundizar en el conocimiento. Para los trabajos sobre léxico, por su parte, ha sido necesario el diseño de programas específicos o la aplicación de técnicas de uso poco común en el análisis de corpus como el que nos ocupa (Ávila Muñoz 2014).

Sin duda, uno de los estudios coordinados más originales y novedosos abordados desde PRESEEA es el de la *prosodia basada en uso*. Esta investigación toma el habla como material básico de investigación. El objetivo fundamental es dar cuenta de patrones prosódicos del español. Ello se hace con un enfoque que considera indispensables el estudio de muestras reales, que permite el reconocimiento y análisis de procesos de variación y cambio sociolingüísticos y geolectales. Como apuntan los coordinadores del grupo de investigación, Martín Butragueño y Velásquez Upegui (2014), se parte de la base teórica de que el uso de la prosodia se presenta, esencialmente, en el habla y de que existen parámetros prosódicos generales en el español, pero también parámetros específicos que llevan a establecer distinciones entre comunidades de habla, grupos sociales o tipos discursivos. Para la descripción tonal, se utiliza el modelo Métrico Autosegmental; las mediciones se realizan con Praat y el etiquetaje, con el sistema Sp-ToBI (*Spanish-Tones and Break Indices*). Se combinan análisis cualitativos y cuantitativos y se aplican modelos estadísticos inferenciales. En este momento, se han propuesto dos objetos de estudio coordinado: los enunciados representativos de foco informativo amplio y el fraseo entre sujeto y predicado. Pueden consultarse ya algunos resultados relevantes, como los aportados por Martín Butragueño (2015, 2016) sobre prosodia de actos de habla aseverativos, directivos y expresivos en el español mexicano y los ofrecidos por Velásquez Upegui (2019) sobre enunciados declarativos de foco amplio, estudiado, en todos los casos, con un enfoque multivariable.

Asimismo, el estudio coordinado de variables fónicas, especialmente de la /s/ implosiva y la /d/ intervocálica, ha llegado a un nivel de profundización considerable, lo que ha proporcionado un conocimiento importante sobre patrones sociolingüísticos y geolectales y, más allá de ello, sobre procesos de divergencia y convergencia entre variedades del español. A este respecto, y a modo de ejemplo, podemos mencionar las investigaciones realizadas sobre el español hablado en Málaga y, concretamente, sobre la producción de /s/ en coda silábica. Los análisis cuantitativos (análisis probabilístico multivariante del mantenimiento, la elisión y la aspiración de la unidad) han permitido confirmar el carácter innovador de la variedad con respecto al debilitamiento y la pérdida de la sibilante, pues se ha documentado escasa o nula influencia de factores gramaticales para frenar el proceso de elisión, reflejo de un cambio en marcha (Vida Castro 2012).

Con una metodología rigurosa, que concede un lugar destacado a los análisis cuantitativos, se han estudiado, también, otras variables fónicas características del andaluz en los corpus de Málaga, Granada y Sevilla. Por los resultados obtenidos y el avance que supone el empleo de programas poco utilizados en la investigación lingüística, son destacables los últimos trabajos sobre ceceo y seseo en Málaga y Granada, en los que se da cuenta de cómo las mujeres jóvenes con estudios universitarios lideran un proceso

de convergencia hacia el estándar, esto es, de distinción entre /θ/ y /s/ (Villena Ponsoda y Vida Castro 2017; Moya Corral y Sosinsky 2015). La importancia de estos datos ha llevado al empleo de procedimientos analíticos que permiten ir más allá de lo que indica la percepción, calculando la distancia acústica de las realizaciones y teniendo en cuenta la variación paramétrica individual en el proceso de escisión; para ello, se ha propuesto un conjunto de análisis estadísticos que hace posible medir objetiva y numéricamente la distancia acústica y reflejar la variación paramétrica individual mediante un modelo gráfico basado en el concepto de coordenadas polares: test *Mann-Whitney*, *RWizard* (interfaz de R que permite realizar análisis complejos bi y multivariantes), gráfico de coordinadas polares (posibilita la representación multidimensional de la variación). La investigación llevada a cabo ya ha confirmado el proceso de convergencia protagonizado por jóvenes universitarios malagueños y ha permitido cuantificar el progreso de la escisión (Molina García 2020, en prensa).

No han faltado los estudios coordinados sobre fenómenos morfosintácticos en el seno de PRESEEA, siempre abordados con una metodología que prima los análisis cuantitativos. Pueden servir como muestra los análisis bivariantes y multivariantes y el cálculo del índice de seguridad lingüística realizado por Gómez Molina sobre el corpus de Valencia, complementado con cuestionarios de aceptabilidad, para conocer los procesos de variación de *haber* impersonal (Gómez Molina 2013a) y que lo llevan a comprobar que la concordancia del verbo con su argumento puede considerarse un fenómeno en expansión, propiciado por distintos grupos sociales, concretamente, en Valencia, jóvenes y sujetos de nivel de instrucción media. Los mismos tipos de análisis estadísticos (descriptivos e inferenciales) se han efectuado para el estudio de otras variables sociolingüísticas como la expresión del sujeto (Guerrero González 2019; Martín Butragueño 2020) o el dequeísmo y queísmo (Gómez Molina 2013b).

El léxico no ha sido desatendido en el marco de PRESEEA, si bien no se han propuesto temas prioritarios para estudios coordinados, lo que no resta relevancia a la investigación realizada ni a los resultados obtenidos. Por su originalidad, podemos destacar, en este nivel, los estudios sobre riqueza léxica llevados a cabo por Ávila Muñoz (Ávila Muñoz 2014) sobre el corpus de Málaga, en los que se ofrece un modelo de análisis que estima la densidad léxica virtual de los hablantes y permite hallar patrones sociolingüísticos en relación con la variación cuantitativa del léxico. Para realizar los análisis, se emplea un algoritmo de optimización aleatoria, llamado *Simulated Annealing*, que ofrece información, además, sobre el fluido léxico de los intercambios lingüísticos y sobre momentos de mayor o menor densidad léxica en la interacción. Ávila ha podido concluir que los sujetos de diferentes estratos sociales emplean un léxico distinto, cualitativa y cuantitativamente, lo que parece relacionarse con el nivel de instrucción y la ocupación, si bien ello no condiciona su eficiencia comunicativa. Por último, pueden servir también de ejemplo las investigaciones de Terrádez Gurrea sobre el léxico de los valencianos, en las que se emplean diversas técnicas cuantitativas: análisis de frecuencias léxicas, extracción de palabras clave (con el programa *AntConc*) (Anthony 2018) y pruebas estadísticas para detección de redes o categorizaciones léxicas (análisis de conglomerados, de correspondencia y discriminante, y escalamiento multidimensional), que le permiten concluir que, en general, los sociolectos pueden caracterizarse por una serie de preferencias léxicas (Terrádez Gurrea 2013a, 2013b).

6. DESARROLLOS ESTADÍSTICOS

Los materiales sociolingüísticos, como es lógico, pueden ser sometidos a análisis cualitativos y cuantitativos mediante el empleo de las herramientas y los recursos que la

investigación pone a disposición de cualquier interesado. Sin embargo, gracias a una colaboración con el investigador japonés Hiroto Ueda, los materiales de PRESEEA son, no solamente accesibles a través de una página electrónica denominada "PRESEEA en LYNEAL", sino que dentro de esta misma página pueden recuperarse para recibir la aplicación de pruebas estadísticas de muy variado tipo. El acceso a "PRESEEA en LYNEAL" puede hacerse a través de los portales de la Universidad de Tokio y de la Universidad Autónoma de Madrid, cuyos responsables son el propio Hiroto Ueda (http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/preseea.htm) y Antonio Moreno Sandoval (http://shimoda.lllf.uam.es/ueda/lyneal/preseea.htm).

El sistema LYNEAL (Letras y Números en Análisis Lingüísticos) es un desarrollo informático que permite la búsqueda de formas por medio de expresiones regulares simplificadas, el reemplazo de textos en las búsquedas (por ejemplo, para la eliminación de etiquetas), el filtrado de materiales por atributos de las fuentes (por ejemplo, sexo, edad, nivel de estudios, relación entre informante y encuestador…) o la combinación de atributos para la realización de búsquedas avanzadas.

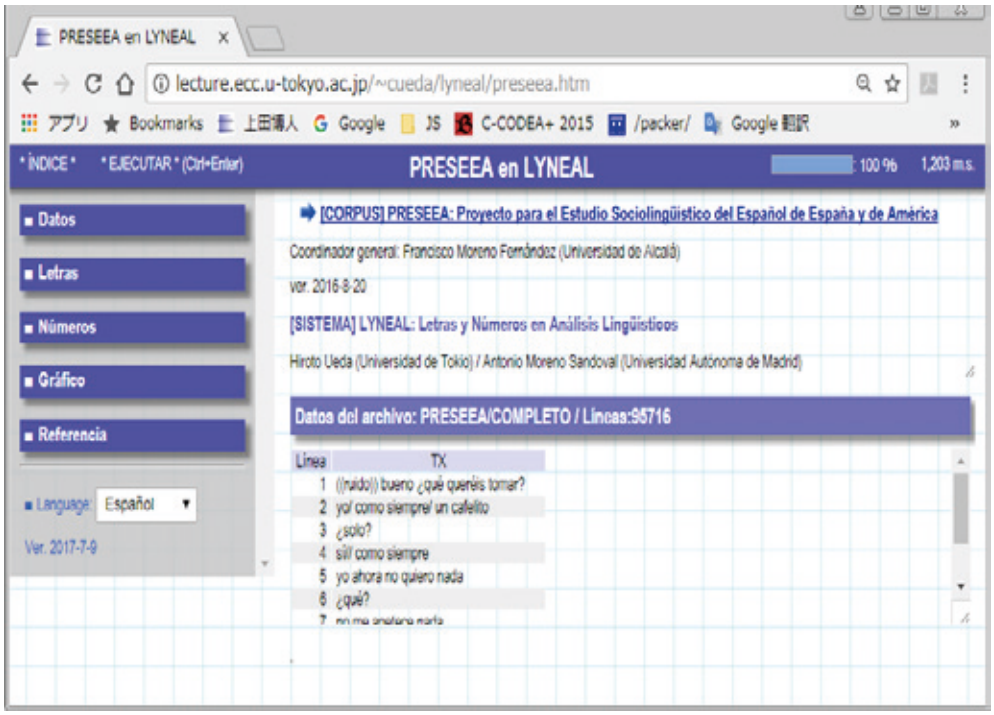


FIGURA 7. Pantalla de "PRESEEA en LYNEAL"

Una vez que los materiales han recibido el tratamiento y la forma adecuados, y una vez seleccionados los componentes deseados, los datos pueden ser sometidos directamente a una diversidad de pruebas estadísticas, que incluyen el cálculo de distintos tipos de frecuencias (absoluta, relativa, normalizada, probabilística, tipificada…), el cálculo de valores estadísticos básicos (media, varianza, desviación típica, mediana, cuartiles, rango, coocurrencias…), la realización de análisis multivariantes, incluidas correlaciones, cálculo de proximidades, análisis de Pareto…). Como componente esencial, se ofrece la posibilidad de obtener directamente una amplia variedad de cuadros y gráficos (barras, curvas, histogramas, gráficos de dispersión, gráficos de cajas, dendrogramas, diagramas de Pareto…). Asimismo, se está

desarrollando una nueva función de cartografiado automático, que permitirá producir mapas interactivos a partir de los resultados de diferentes pruebas estadísticas.

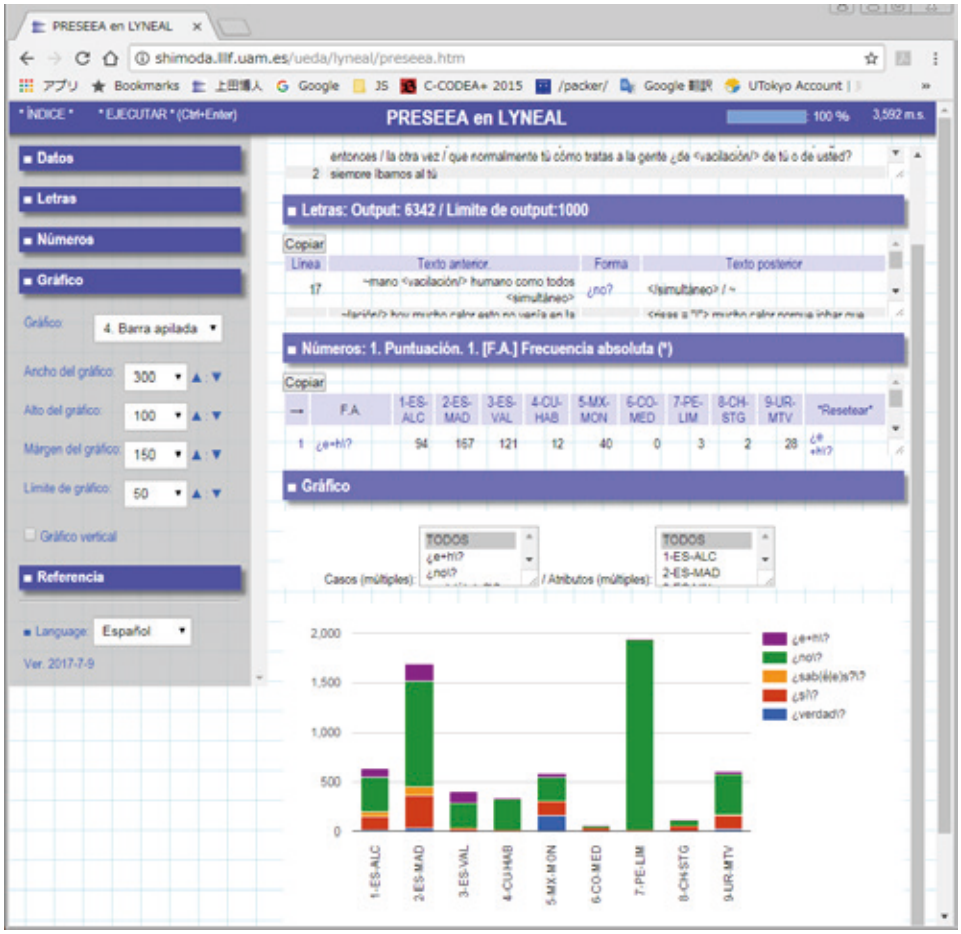


FIGURA 8. Pantalla de "PRESEEA en LYNEAL" con datos y gráfico sobre uso de elementos conversacionales de confirmación en español

Los materiales de PRESEEA se van incorporando a "PRESEEA en LYNEAL" conforme se integran también en el portal PRESEEA. La posibilidad de acceder a esta herramienta integrada proporciona a los materiales del proyecto un desarrollo analítico cuantitativo del que no disfrutan otras colecciones de materiales de naturaleza similar.

7. DESARROLLOS FUTUROS

Las investigaciones sobre el corpus PRESEEA están permitiendo caracterizar numerosas variables lingüísticas y describirlas como fenómenos en variación continua o estable. Además, ofrecen la información necesaria para descubrir los factores internos y externos a la lengua que determinan o inciden en los diferentes procesos de variación y cambio. Estamos en un excelente momento para seguir, poco a poco, cumpliendo metas y profundizando en el conocimiento de la variación sociolingüística.

La coordinación ofrece un marco ideal para investigaciones de gran envergadura, por ello, el desarrollo presente y futuro del proyecto debe apoyarse en la ampliación de la nómina de fenómenos priorizados y en el desarrollo de herramientas que ofrezcan más

información en los propios materiales (etiquetados específicos para anotación fónica, gramatical, léxico-semántica o pragmática de las muestras; alineamiento de transcripción y grabación en los subcorpus de acceso libre, etc.). Cada estudio debe partir, como hasta ahora, de la realización de una ficha de codificación y una *Guía de estudio* que permitan atender a todas las dimensiones incidentes y efectuar análisis cualitativos y, especialmente, cuantitativos (multivariantes).

Por otro lado, los resultados obtenidos ya hacen posible abordar la investigación de procesos de acomodación por contacto de variedades (especialmente en situaciones de migración), lo que llevará, en un futuro no muy lejano, a caracterizar sociedades multiculturales. Desde una dimensión aplicada, se ha empezado con la transferencia de conocimientos al ámbito de la enseñanza de lenguas adicionales y de la dinámica social, presentando patrones característicos de hombres y mujeres, que van más allá del descubrimiento de desigualdades sociales.

Dado que el macrocorpus PRESEEA tiene ya 20 años, se hace necesario comenzar la recogida de muestras de lengua hablada en formato digital de alta definición, incluso con grabación audiovisual, de modo que se pueda comprobar el estado de procesos de variación y cambio, y realizar análisis que requieran sonido de calidad y visionado de los modos de producción. Por otro lado, la profundización en el estudio de algunos fenómenos exige el diseño y la aplicación de herramientas complementarias de recogida y análisis de materiales, que permitan atender a diferentes dimensiones, desde la fonética, al discurso y la percepción.

Por último, el proyecto PRESEEA está en condiciones de abordar, en un futuro inmediato, estudios, teóricos y aplicados, sobre individuos, grupos, redes y comunidades de diferentes tamaños, estratificación y contacto, inmigración y discriminación, convergencia y divergencia, indización y significado social, así como sobre el cambio lingüístico en tiempo aparente y tiempo real. Confiamos en ir dando cuenta de ello en un futuro no muy lejano y en despertar el interés de lingüistas, sociolingüistas y dialectólogos por el proyecto, el corpus y su investigación.

REFERENCIAS BIBLIOGRÁFICAS

ALBELDA MARCO, M., A. M. CESTERO MANCERA, S. GUERRERO GONZÁLEZ & M. SAMPER HERNÁNDEZ (en preparación): *Variación sociopragmática y geolectal en el uso de atenuación*.

ANTHONY, L. (2018): *AntConc*. Waseda University: Center for English Language Education in Science and Engineering, School of Science and Engineering. http://www.laurenceanthony.net/software.html. [última consulta: 18/02/2019]

ÁVILA MUÑOZ, A. M. (2014): "Patrones sociolingüísticos de la riqueza léxica. Estudio basado en una propuesta original para el cálculo del índice de la densidad léxica virtual de los hablantes", *LEA* XXXVI (2), pp. 171-194.

BRIZ GÓMEZ, A. (2005): "Los corpus de español hablado", *Oralia* 8, pp. 7-12.

BRIZ GÓMEZ, A. & M. ALBELDA (2009): "Estado actual de los corpus de lengua española hablada y escrita: I+D", in *El español en el mundo. Anuario de del Instituto Cervantes*. Madrid: Instituto Cervantes, pp. 165 - 225.

CESTERO MANCERA, A. M. & F. MORENO FERNÁNDEZ (eds.) (2017): *Procesos de variación y cambio en el español de España. Estudios sobre el corpus PRESEEA, Lingüística en la Red*, Monográfico XV. http://www.linred.es/numero15_monografico.html. [última consulta: 15/02/2019]

CESTERO MANCERA, A. M. & F. PAREDES GARCÍA (eds.) (2018): *Percepción de las variedades cultas del español: creencias y actitudes de jóvenes universitarios hispanohablantes. Boletín de Filología* 53 (2), Monográfico. https://boletinfilologia.uchile.cl/index.php/BDF/issue/view/5038. [última consulta: 15/02/2019]

CLAES, J. (2014): "Sociolingüística comparada y gramática de construcciones: Un acercamiento a la pluralización de *haber* presentacional en las capitales antillanas", *Revista Española de Lingüística Aplicada* 27, pp. 338-364.

GÓMEZ MOLINA, J. R. (2013a): "Usos del verbo *haber* impersonal", in J. R. Gómez Molina (ed.): *El español de Valencia. Estudio sociolingüístico*. Bern: Peter Lang, pp. 109-143.

GÓMEZ MOLINA, J. R. (2013b): "Las construcciones '*de*/Ø + *que* + verbo en forma personal'", in J. R. Gómez Molina (ed.): *El español de Valencia. Estudio sociolingüístico*. Bern: Peter Lang, pp. 183-224.

GUERRERO GONZÁLEZ, S. (2019): "Expresión de sujeto pronominal en el corpus PRESEEA de Santiago de Chile". Comunicación presentada en el Seminario internacional *Las investigaciones lingüísticas en el mundo hispánico*, celebrado en La Habana (Cuba), del 28 de enero al 1 de febrero de 2019.

MARTÍN BUTRAGUEÑO, P. (2015): "Hacia una prosodia basada en el uso: actos de habla en el español mexicano", *Normas* 5, pp. 97-115.

MARTÍN BUTRAGUEÑO, P. (2016): "A veces lloro mis lágrimas. Acercamiento multivariable a la prosodia de los actos de habla expresivos en el español de México", *Estudios de Lingüística Aplicada* 63, pp. 59-102.

MARTÍN BUTRAGUEÑO, P. (2020): "An approach to subject pronoun expression patterns in data from the 'Project for the Sociolinguistic Study of Spanish in Spain and America', *Spanish in Context* 17 (2), pp. 294-316.

MARTÍN BUTRAGUEÑO, P. & E. P. VELÁSQUEZ UPEGUI. (2014): *PRESEEA_PROSODIA. Guía de estudios de prosodia basada en el uso en los corpus PRESEEA*. http://preseea.linguas.net/. [última consulta: 15/02/2019]

MOLINA GARCÍA, Á. (2019): "Percepción y distancia acústica: la variación paramétrica individual en la escisión fonemática de /θ/ en el español andaluz. Datos de la ciudad de Málaga", *ELUA* 33, pp. 111-140.

MOLINA GARCÍA, Á. (en prensa): "Percepción comunitaria de la distinción fonemática fonemática de la /s/ y la /θ/ en hablantes andaluces. Estudio en la ciudad de Málaga".

MOLINA MARTOS, I., F. PAREDES GARCÍA & A. M. CESTERO MANCERA (eds.) (2020): *Sociolinguistic patterns and processes of convergence and divergence in Spanish*, *Spanish in Context* 17 (2), Monográfico.

MORENO FERNÁNDEZ, F. (1996): "Metodología del 'Proyecto para el estudio sociolingüístico del español de España y de América' (PRESEEA)", *Lingüística* 8, pp. 257-287.

MOYA CORRAL, J. A. & M. SOSINSKY (2015): "La inserción social del cambio. La distinción s/θ en Granada. Análisis en tiempo aparente y en tiempo real", *Lingüística Española Actual* 37 (1), pp. 33-72.

PRESEEA (2003): "Metodología del 'Proyecto para el estudio sociolingüístico del español de España y América (PRESEEA)'". Versión revisada 31-10-2003. http://preseea.linguas.net. [última consulta: 15/02/2019]

PRESEEA (2008): "Marcas y etiquetas mínimas obligatorias". Versión 1.0. 31-01-2008. http://preseea.linguas.net. [última consulta: 15/02/2019]

RAE: *Banco de datos (CORPES XXI). Corpus del español del siglo XXI (CORPES)*. http://www.rae.es. [última consulta: 15/02/2019]

SAN MARTÍN NÚÑEZ, A. & S. GUERRERO GONZÁLEZ (eds.) (2016): *Estudios sobre la lengua española hablada en el mundo hispánico en su variedad geográfica y social con materiales del PRESEEA*. *Boletín de Filología de la Universidad de Chile,* 51 (2), Monográfico. https://boletinfilologia.uchile.cl/index.php/BDF/issue/view/4510. [última consulta: 15/02/2019]

SILVA-CORVALÁN, C. (1994): "Direcciones en los estudios sociolingüísticos de la lengua española", in *Actas del Congreso de La Lengua Española, Sevilla, 7-10 de octubre de 1992*. Madrid: Instituto Cervantes, pp. 399-415.

TAGLIAMONTE, S. (2002): "Comparative sociolinguistics", in J. K. Chambers, P. Trudgill & N. Shilling (eds.): *The handbook of language variation and change*. Malden: Blackwell, pp. 729-763.

TERRÁDEZ GURREA, M. (2013a): "El corpus PRESEVAL. Estudio cuantitativo del léxico", in J. R. Gómez Molina (ed.): *El español de Valencia. Estudio sociolingüístico*. Bern: Peter Lang, pp. 225-245.

TERRÁDEZ GURREA, M. (2013b): "Estratificación y frecuencias léxicas", in J. R. Gómez Molina (ed.): *El español de Valencia. Estudio sociolingüístico*. Bern: Peter Lang, pp. 247-274.

VELÁSQUEZ UPEGUI, E. P. (2019): **"**Enunciados declarativos de foco amplio en el corpus PRESEEA: Exploración descriptiva desde la Prosodia Basada en el Uso", Comunicación presentada en el Seminario internacional *Las investigaciones lingüísticas en el mundo hispánico*, celebrado en La Habana (Cuba), del 28 de enero al 1 de febrero de 2019.

VIDA CASTRO, M. Á. (2012): "Las consonantes en la coda silábica. Las causas universales y las razones sociales del mantenimiento y elisión de /s/ en el español hablado en Málaga", in J. A. Villena & A. M. Ávila (eds.): *Estudios sobre el español de Málaga. Pronunciación, vocabulario y sintaxis*. Málaga: Editorial Sarriá (colección Universidad), pp. 191-208.

VILLENA PONSODA, J. A. (coord.) (en preparación): *Estudio coordinado de patrones sociolingüísticos del español de España (Resultados del proyecto ECOPASOS).*

VILLENA PONSODA, J. A. & M. Á. VIDA CASTRO (2017): "Variación, identidad y coherencia en el español meridional. Sobre la indexicalidad de las variables convergentes del español de Málaga", *Lingüística en la Red*, Monográfico XV. http://www.linred.es/numero15_monografico_Art1.html. [última consulta: 15/02/2019]

VILLENA PONSODA, J. A., A. M. ÁVILA MUÑOZ, J. M. SÁNCHEZ SÁEZ & Mª C. LASARTE CERVANTES (2010): "Problemas de anotación e intercambio en los corpus orales. Estrategias para la transformación de textos etiquetados en documentos XML. El caso de los corpus PRESEEA", *Oralia* 13, pp. 261-323

HUMANIDADES DIGITALES Y GEOGRAFÍA LINGÜÍSTICA: LA EDICIÓN DIGITAL DEL *ATLAS LINGÜÍSTICO DE LA PENÍNSULA IBÉRICA*[*]

Digital Humanities and Linguistic Geography: The digital edition of the Atlas Lingüístico de la Península Ibérica

XULIO SOUSA

Instituto da Lingua Galega, Universidade de Santiago de Compostela

Resumen

El artículo presenta los procesos de diseño y desarrollo de la edición digital del *Atlas Lingüístico de la Península Ibérica*. El ALPI fue un atlas concebido en el marco de la dialectología tradicional occidental y siguiendo el modelo de las obras pioneras de geografía lingüística publicadas en Europa a principios del siglo XX. El proyecto de edición digital de los materiales editados e inéditos de esta obra, promovido por el CSIC, se fundamenta en los principios metodológicos de las humanidades digitales y tiene como objetivo ofrecer a los usuarios un acceso a la información en distintos formatos (imágenes de los cuadernos, transcripciones, visualizaciones, etc.). En la contribución se dará cuenta de las características generales del proyecto y de la utilidad de la herramienta de consulta.

Palabras clave: humanidades digitales, dialectología, geolingüística, lenguas iberorrománicas

Abstract

The paper describes the planning and development processes of the digital edition of the *Atlas Lingüístico de la Península Ibérica* (ALPI). The ALPI was an atlas conceived within the framework of traditional Western dialectology and following the example of the pioneering works of linguistic geography published in Europe at the beginning of the 20th century. The project of digital edition of the published and unpublished materials promoted by the CSIC is based on the methodological guidelines of the digital humanities and is designed to provide users access to information in diverse presentations (images of the notebooks, transcriptions, visualizations, etc.). The article shows an account of the general characteristics of the project and the usefulness of the query tool.

Keywords: digital humanities, dialectology, geolinguistics, linguistic atlas, Ibero-Romance languages

[*] Este trabajo se realizó en parte gracias a una ayuda de investigación para proyectos de I+D (referencia PGC2018-095077-B-C44, financiado por MCIU/AEI/FEDER, UE).

El *Atlas Lingüístico de la Península Ibérica* fue un proyecto de geografía lingüística iniciado a principios del siglo XIX en el seno del Centro de Estudios Históricos por Ramón Menéndez Pidal y Tomás Navarro Tomás. Los trabajos del proyecto fueron interrumpidos por la guerra civil y hasta 1962 no apareció el primer y único volumen publicado. En este artículo se presentan las características generales del proyecto de edición digital de los materiales inéditos del *Atlas Lingüístico de la Península Ibérica* (ALPI). El ALPI fue un atlas concebido en el marco de la dialectología tradicional y siguiendo el modelo de las obras pioneras de geografía lingüística publicadas en Europa a principios del siglo XX. El proyecto de edición digital de los materiales editados e inéditos de esta obra promovido por el CSIC se fundamenta en los principios metodológicos de las humanidades digitales y tiene como objetivo ofrecer a los usuarios el acceso a la información en distintos formatos (imágenes de los cuadernos, transcripciones, visualizaciones cartográficas, etc.). Esta contribución se centra en las tareas vinculadas con la digitalización e informatización de la información lingüística del ALPI.

1. INTRODUCCIÓN

El cartógrafo Ferjan Ormeling al abordar el uso de los mapas lingüísticos en la historia de la cartografía distingue entre dos categorías de representaciones: los mapas etnolingüísticos y los mapas dialectales (Ormeling 2015). Los primeros muestran las comunidades lingüísticas en un área geográfica y se comenzaron a emplear a partir del siglo XVI. Los segundos ayudan al usuario a visualizar la variación lingüística en un territorio y no aparecieron hasta la segunda mitad del siglo XIX, cuando los dialectólogos empezaron a desarrollar trabajos sistemáticos de documentación lingüística de las variedades lingüísticas regionales. Este tipo de mapas se reúnen para constituir los atlas lingüísticos, un tipo de obra cartográfica que en occidente nace al mismo tiempo que comienza a desarrollarse la dialectología, en sus inicios disciplina complementaria de los estudios de lingüística histórica (Hickey 2018), y es coetáneo de las primeras visualizaciones temáticas de datos realizadas sobre la base de mapas nacionales y regionales (Friendly 2008).

A pesar de que se reconocen algunos precedentes de trabajos sobre variedades regionales que utilizan mapas para mostrar la distribución de variantes lingüísticas, se considera que este segundo tipo de visualización de la documentación lingüística se inauguró en Europa en el siglo XIX con el *Sprachatlas des Deutschen Reichs* de Georg Wenker, bibliotecario, filólogo y fundador de la Escuela de dialectología de Marburgo. En esta época, los atlas lingüísticos se concebían como una forma ventajosa de presentación de información dialectal recolectada en varias localidades de un territorio. La representación de los datos en un mapa era el fin fundamental de los proyectos de investigación dialectal. La dialectología tradicional había nacido con propósitos descriptivos e históricos y las investigaciones vinculadas a los atlas lingüísticos eran más cartográficas que geográficas, en tanto que el proyecto se consideraba finalizado una vez que la colección de mapas se publicaba (Scheuringer 2010; Britain 2014).

Los mapas que componían el atlas de Wenker daban cuenta de los resultados obtenidos a partir de un cuestionario de 40 enunciados enviado a maestros de todo el imperio alemán. El proyecto, comenzado como iniciativa particular, acabó convirtiéndose en una empresa de estado y Wenker pudo ocuparse de forma exclusiva de elaborar manualmente los mapas del atlas a partir de los aproximadamente 45.000 cuestionarios recibidos por correo postal. El primer mapa compuesto se finalizó en 1889

y mostraba la variación vinculada con el numeral *sechs* 'seis'. Las variantes se representan sobre el mapa combinando isoglosas de colores, etiquetas y símbolos. Las isoglosas y las etiquetas ayudan a identificar las grandes áreas de las formas más extendidas; los símbolos señalan las variantes especiales y menos frecuentes. El valor de estos elementos se explica en una leyenda que acompaña a cada una de las tres hojas que componen un mapa. A pesar de las limitaciones de la época, la visualización cartográfica fue utilizada de forma inteligente por Wenker para confeccionar los mapas de su atlas. Todavía hoy el estudioso interesado en conocer la variación regional del alemán es capaz de identificar de forma cómoda y rápida la distribución de las variantes más frecuentes, gracias al uso acertado de isoglosas de colores y etiquetas. El *Sprachatlas des Deutschen Reichs* supuso un avance en la presentación cartográfica de la información lingüística que no todos los proyectos de geografía lingüística iniciados en las décadas sucesivas supieron comprender y seguir con acierto.

El sucesor de Wenker en el proyecto fue su asistente y también profesor de la universidad de Marburgo Ferdinand Wrede. En junio del año 1913 Wrede recibe en Marburgo la visita de Tomás Navarro Tomás, que por esa altura disfrutaba de una beca de la Junta para Ampliación de Estudios para conocer proyectos de fonética experimental y geografía lingüística que se desarrollaban en universidades del centro de Europa (Pérez Pascual 2016). El viaje de estudios de Navarro Tomás había sido iniciativa de Ramón Menéndez Pidal y tenía como motivos principales averiguar «qué aparatos necesitarían para montar un Laboratorio de Fonética en Madrid» y conocer «qué metodología resultaría adecuada para hacer el atlas previsto» (García Mouton 2015). A principios del siglo XIX Menéndez Pidal había ideado un proyecto de «mapa lingüístico» de las hablas castellanas y aragonesas para el que contaría con la asistencia de Navarro Tomás. La idea inicial fue el origen del atlas lingüístico de los romances peninsulares que el discípulo de Menéndez Pidal dirigiría y al que daría el nombre de *Atlas Lingüístico de la Península Ibérica*. La historia de este proyecto la conocemos hoy en detalle gracias a las contribuciones de varios investigadores (Cortés & García Perales 2009; Alonso Montero 2011; García Mouton 2011; García Mouton 2015; Pérez Pascual 2016).

Cuando Navarro Tomás visita el taller en el que trabajaban Wrede y sus colaboradores tiene oportunidad de conocer en profundidad los detalles de la obra: la metodología utilizada en la recolección de datos y en el diseño de mapas, los primeros borradores e incluso algunas hojas de mapas terminados (Cortés & García Perales 2009). Navarro Tomás comprobó las diferencias con respecto al atlas de Gilliéron, identificó sus defectos y también algunas «ventajas que pueden aprovecharse y que seguramente no es difícil llevarlas a la práctica» en el proyecto que estaba diseñando (Cortés & García Perales 2009). A pesar de que no es muy explícito en la carta dirigida a Menéndez Pidal desde Marburgo, es posible suponer que los inconvenientes a los que se refiere Navarro Tomás tenían que ver con el tipo de cuestionario (traducción de enunciados desde el alemán estándar), el método usado en la obtención de datos (colaboración de maestros y envío de cuestionarios por correo) y el poco interés prestado en el atlas a la fonética. El diseño final del proyecto del ALPI debe más a los atlas de Gilliéron (*Atlas linguistique de la France*, ALF) y Jaberg y Jud (*Sprach- und Sachatlas Italiens und der Südschweiz*, AIS) que a la obra iniciada por Wenker. Navarro Tomás y Menéndez Pidal fueron capaces de reconocer que las características de los proyectos románicos suponían un gran avance metodológico que permitiría investigar con mayor profundidad y rigor la variación lingüística en las hablas rurales románicas de la península ibérica.

Más de un siglo después de comenzados estos proyectos, el desarrollo de las humanidades digitales y la regeneración de los estudios dialectales, ahora incorporados en el ámbito de la investigación en variación lingüística, han vuelto a ligar el destino de

estas cuatro obras de geografía lingüística europeas. Desde principios de este siglo los materiales del atlas de Wenker pueden consultarse a través de una página web, en sus inicios bajo el nombre de *Digitaler Wenker-Atlas* (Schmidt & Herrgen 2001) y desde hace unos pocos años en el sitio del proyecto REDE (Regionalsprache.de; Limper, Pheiff & Williams 2019), que hace posible tanto la visualización de las 1619 hojas de mapas como la búsqueda y análisis de todas las variantes recogidas (Schmidt, Herrgen & Kehrein 2008; Rabanus, Kehrein y Lameli 2010). A la información del AIS de Jaberg y Jud puede accederse gracias al programa *NavigAIS,* desarrollado bajo la dirección de Graziano G. Tisato en el Consiglio Nazionale delle Ricerche de Padua. Los mapas están digitalizados y la información ha sido introducida en una base de datos que permite búsquedas según distintos criterios (Tisato 2009). El programa de consulta también puede descargarse y utilizarse en modo local. Desde hace menos de una década es posible consultar y descargar los mapas del ALF de Gilliéron gracias a la aplicación *CartoDialect*, desarrollada en colaboración por varias universidades y centros de investigación franceses. Esta aplicación, todavía en desarrollo, permite la explotación de la información semántica y geográfica de los datos recogidos por Jules Gilliéron y Edmond Edmont (Davoine *et alii* 2015). Las 1920 cartas del atlas pueden ser visualizadas en una imagen de calidad que permite comprobar sin dificultad las respuestas asociadas a cada uno de los 639 puntos de encuesta.

Los materiales del ALPI todavía no pueden ser consultados del mismo modo que en estos tres proyectos hermanos. A diferencia de ellos, solo una pequeña parte de la información recogida en los cuadernos de campo llegó a editarse como atlas en papel. El único volumen publicado en 1962 contiene 75 mapas, 5 de ellos con información básica del proyecto (lugares estudiados, nombre de las localidades, denominación de los habitantes, denominación del habla local y encuestadores) y otros 70 elaborados a partir de otras tantas preguntas de la sección de fonética del cuaderno I del cuestionario. El resto de los datos sobre la variación dialectal en las hablas rurales romances peninsulares tardó muchas décadas en estar a disposición de los investigadores (Heap 2012). En el año 2009, el Centro Superior de Investigaciones Científicas puso en marcha un proyecto intramural en colaboración con investigadores de diferentes universidades para desarrollar una edición digital completa de los materiales inéditos del ALPI. El proyecto, todavía en desarrollo, está dirigido por Pilar García Mouton (CSIC) y cuenta con la participación de Inés Fernández-Ordóñez (Universidad Autónoma de Madrid), Maria Pilar Perea (Universitat de Barcelona), João Saramago (Universidade de Lisboa), Xulio Sousa (Universidade de Santiago de Compostela) y David Heap (University of Western Ontario).

En esta contribución se presentan las principales características del proyecto de informatización del ALPI, fundamentalmente en los aspectos referidos al tratamiento de los materiales originales, al diseño de la base de datos y a la configuración de las aplicaciones de edición y consulta de los materiales. El objetivo de este artículo es mostrar el desarrollo y resultados obtenidos hasta el momento en el proyecto de edición de los materiales del *Atlas Lingüístico de la Península Ibérica*. Este atlas de las variedades románicas peninsulares, iniciado a comienzos del siglo XX bajo la iniciativa de Ramón Menéndez Pidal y Tomás Navarro Tomás, es todavía hoy una obra desconocida, ya que la mayor parte de sus materiales no ha sido publicada ni puesta a disposición de los investigadores. Este trabajo es un complemento a las detalladas presentaciones de las distintas fases del proyecto publicadas en los últimos años por Pilar García Mouton, coordinadora de esta nueva empresa científica (García Mouton 2010; García Mouton 2011; García Mouton 2012a; García Mouton 2012b; García Mouton 2015;

García Mouton & Pedrazuela Fuentes 2015; García Mouton *et alii* 2016; García Mouton 2017).

2. Los atlas lingüísticos: la geolingüística y las humanidades digitales

La geografía lingüística, y de forma general los estudios dialectales, no ha merecido mucha atención en las monografías de humanidades digitales publicadas en los últimos años. En los capítulos dedicados a las bases de datos de documentación lingüística y a la utilización de los sistemas de información geográfica, pueden encontrarse las escasas noticias sobre el uso de las herramientas digitales en la investigación en geografía lingüística y dialectología. En la última edición del *Companion to Digital Humanities* (Schreibman, Siemens & Unsworth 2016), Todd Presner y David Sheppard dedican un capítulo a abordar la relevancia que para los estudios humanísticos tiene el empleo de sistemas de información geográfica (visualización, interpretación y análisis de datos lingüísticos). En estas páginas se hace referencia a cómo la geografía lingüística se benefició de estas nuevas aplicaciones informáticas. Los autores fijan en 1983, con la impresión informatizada de algunos mapas del *Atlas Linguarum Europae* (ALE), bajo la dirección de Mario Alinei (1983), el primer uso de tecnologías digitales en un proyecto de geografía lingüística (Presner & Shepard 2016). En un trabajo anterior de revisión de la metodología de los proyectos de geografía lingüística tradicionales y modernos publicado en 2010, Alfred Lameli sitúa los primeros proyectos de geolingüística informatizada en la década de 1970 (Lameli 2010). Los tres proyectos destacados como precursores son el *Computer Developed Linguistic Atlas of England* (Viereck & Ramisch 1991-1997), el *Atlas Linguarum Europae* (Alinei *et alii* 1983) y el *Kleiner Deutscher Sprachatlas* (Veith, Putschke & Hummel 1984-1999).

De acuerdo con estas referencias, resulta evidente que la dialectología incorporó el empleo de técnicas informáticas al tratamiento, análisis y explotación de datos no mucho después que otras disciplinas que forman parte del área de estudios que hoy es común denominar humanidades digitales. Desde la década de los años setenta hasta la actualidad, los avances en la aplicación de sistemas informatizados a la geografía lingüística modificaron profundamente los métodos de interpretación y visualización de la información dialectal (Hoch & Hayes 2010). Las mudanzas ocurridas en el ámbito de los estudios lingüísticos con la incorporación de innovaciones informáticas y metodológicas obligaron también a modificar las técnicas de preservación y explotación de la información geolingüística (Brun-Trigaud 2016). Los atlas lingüísticos son proyectos que requieren el almacenamiento y administración de datos de naturaleza heterogénea (documentos de texto, dibujos, fotos, archivos de sonido, archivos de vídeo, etc.) y en consecuencia el uso combinado de diferentes herramientas informáticas. Este tipo de información difusa y dispersa debe estructurarse de manera que se asegure su conservación y se facilite la consulta y explotación (Jessop 2007). Además, en la actualidad resulta imprescindible que los datos se pongan a disposición de la comunidad científica en formatos que agilicen su tratamiento y análisis.

Los proyectos de geografía lingüística emprendidos en los últimos años se diseñan ya desde su inicio como proyectos digitales. Las herramientas informáticas son utilizadas para capturar la información (audio, vídeo, texto, fotos, etc.), para almacenarla (bases de datos, archivos digitales, etc.), para explotarla y para difundir los resultados (programas SIG, programas de análisis cualitativo y cuantitativo, programas de visualización gráfica, sitios web, etc.). El destinatario de estos proyectos es la comunidad académica que puede tener intereses muy diversos, desde buscar ejemplos que ilustren una teoría y encontrar

contraargumentos para rebatir propuestas hasta analizar datos brutos para descubrir patrones de distribución espacial que puedan relacionarse con informaciones históricas, demográficas, genéticas o de otro tipo. Igualmente, los proyectos actuales siguen estándares en su diseño que garantizan la reutilización de los datos y la interoperabilidad de las aplicaciones de análisis (Di Buccio, Di Nunzio & Silvello 2014). Los proyectos geolingüísticos de edición de materiales inéditos o de reedición de atlas ya publicados deben procurar adaptarse a estos estándares, ya que es la forma de asegurar su utilidad y garantizar la reutilización de la información (Embleton, Uritescu & Wheeler 2004).

3. Del cuaderno a la pantalla: la informatización del ALPI

El primer intento de poner a disposición de la comunidad científica los materiales inéditos del ALPI lo emprendió en el año 2000 David Heap de la Universidad de Western Ontario. Después de localizar los cuadernos, realizar una copia en papel y digitalizarla, creó un sitio web desde el que era posible consultar la imagen escaneada de las páginas del primer cuaderno de 529 localidades (Heap 2002; 2012). Desde esta página, actualmente inactiva, era posible acceder a los materiales a partir de la lista de localidades y de la relación de preguntas del cuestionario. A pesar de la peregrinación de los cuestionarios del ALPI por distintos países, despachos, archivos y viviendas particulares, los cuadernos se conservaron en buen estado y tanto los textos impresos como las respuestas anotadas a lápiz o con pluma podían leerse sin dificultad en estas copias. Gracias a esta forma de difusión del proyecto, varios investigadores pudieron realizar trabajos de análisis a partir de estos datos (Heap 2012).

Casi una década después de la puesta en marcha de esta iniciativa se emprende el proyecto de elaboración y edición completa de los materiales del ALPI. El objetivo del nuevo proyecto es editar y tratar los materiales del ALPI «utilizando las posibilidades que la tecnología informática proporciona actualmente a los trabajos geolingüísticos» (García Mouton 2010: 169). Los avances en tecnologías digitales y el conocimiento de experiencias similares hacen posible a esta altura el diseño de un proyecto complejo que aborde el tratamiento y conservación de los materiales originales, la transcripción de las respuestas en alfabeto fonético estandarizado, la creación de una base de datos y la disposición de toda la información en formatos compatibles y fácilmente accesibles. El propósito de este nuevo proyecto es que el investigador interesado en conocer las variedades rurales peninsulares de la primera mitad del siglo XX tenga acceso no solo a una imagen digital de los cuadernos originales, sino también que pueda realizar búsquedas sobre los datos (textuales y gráficos), visualizar la distribución de las variantes de forma simple y descargar los resultados de las búsquedas en formatos que permitan su posterior tratamiento y explotación, según los estándares aplicados en proyectos similares (Kretzschmar & Gray Potter 2010).

Desde los inicios del nuevo siglo se habían puesto en marcha distintos proyectos de informatización de atlas lingüísticos y por lo tanto se contaba con modelos para el diseño de bases de datos dialectales y para la constitución de un sistema de consulta que combinase la información textual con una visualización cartográfica de los resultados (Bauer & Goebl 2000; Aurrekoetxea 2008; Kretzschmar, Bounds & Palosaari 2011; Olariu & Olariu 2014; Clarke 2016; Kumagai 2016). En primer lugar, era necesario trasladar los materiales del ALPI a un formato que garantizase la conservación y permitiese su utilización mediante distintas aplicaciones (sistemas de información geográfica, servicios web, aplicaciones de análisis cuantitativo, etc.). Los materiales de los proyectos de geografía lingüística acostumbran a tener un carácter heterogéneo y requieren un tratamiento diferenciado que no siempre se adapta a los estándares

utilizados en otros ámbitos de las humanidades digitales. Los materiales inéditos del ALPI, a diferencia de otros proyectos posteriores (Sousa 2017), son de dos tipos, textuales y gráficos, y requieren un tratamiento de menor complejidad. La información fundamental del proyecto se recoge en los cuadernos de campo en los que los investigadores fueron anotando las respuestas que proporcionaban los informadores de cada localidad. Además de esta información textual, los entrevistadores también recogieron información gráfica de carácter etnográfico, unas veces como dibujos simples en las páginas de los cuadernos y otras como fotografías[1]. En las secciones siguientes se exponen las tareas de tratamiento de los materiales textuales y de diseño de las aplicaciones de edición y consulta de la información contenida en los cuadernos de campo.

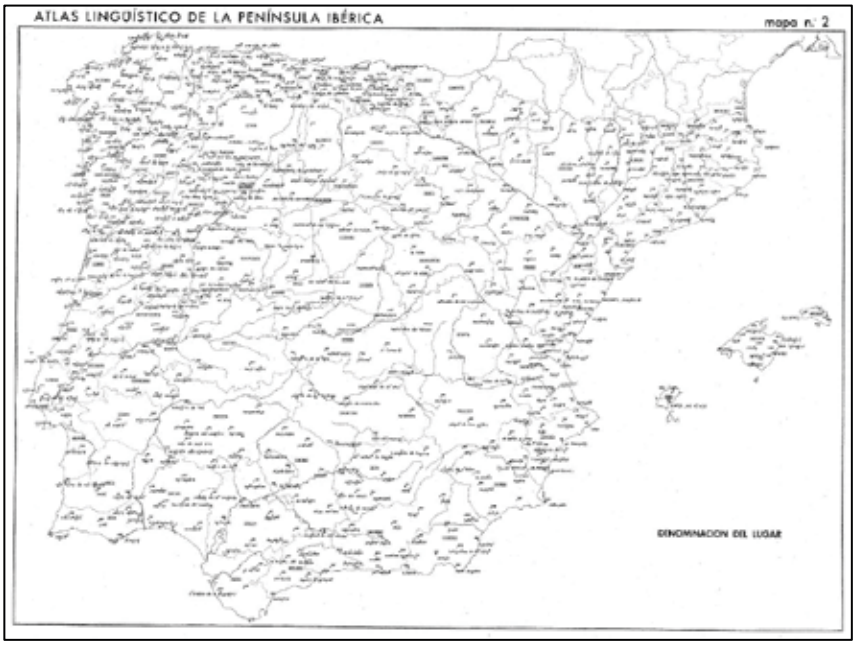


FIGURA 1. Mapa de localidades del ALPI del volumen publicado (Navarro Tomás 1962, mapa 1). Los puntos de encuesta se identifican con un número y el nombre de la localidad

3.1. Los cuadernos de campo

Los cuadernos de campo son los cuestionarios en papel en los que los investigadores anotaron las respuestas de los informadores en cada una de las localidades investigadas. La red del ALPI consta de 529 localidades, en su mayoría poblaciones rurales menores, y de cada una de ellas existen dos cuadernos. El cuaderno I reúne las preguntas de fonética y morfosintaxis (411 preguntas), y el cuaderno II las de léxico y etnografía. Del segundo cuaderno se editaron dos versiones, una extensa (IIE, 50 páginas) y otra reducida (IIG, 27 páginas)[2]. En la mayoría de los puntos investigados se utilizó el cuaderno más amplio, de 833 preguntas.

[1] En la página del proyecto puede consultarse la galería de fotos correspondientes al Fondo Rodríguez-Castellano. Estas son las únicas fotografías que se conservan del proyecto inicial.

[2] En la página web del proyecto pueden consultarse copias digitalizadas de los cuadernos e información detallada sobre el proyecto original (García Mouton *et alii* 2016).

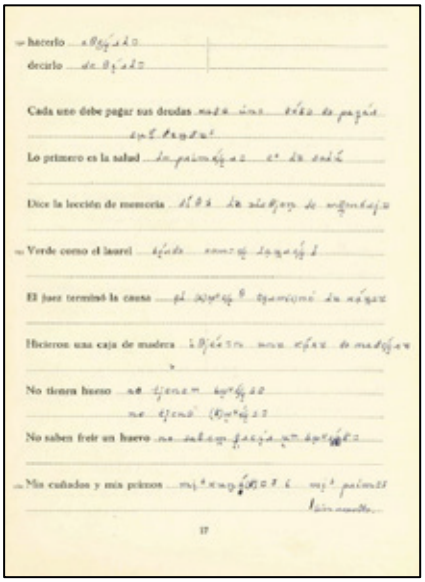


FIGURA 2. Página 17 del cuaderno I de Cadalso de los Vidrios, Madrid

En la cubierta de los cuadernos se registraron los datos básicos de identificación de la localidad: código numérico, nombre del lugar y provincia. Las páginas iniciales recogen información detallada sobre el informador (nombre y apellidos, edad, profesión, residencia, viajes, etc.), la investigación (fecha de la entrevista y nombre del investigador) y la localidad (parroquia, diócesis, nombre en el habla local, nombre dialectal de los habitantes, número de habitantes, modo de vida, etc.). A partir de la página 4 del cuaderno I se encuentran las preguntas sobre las variables lingüísticas. Las diez primeras son frases simples agrupadas en una sección denominada *Notas de orientación fonética* y a continuación figuran las preguntas de la sección *Cuestionario*, que continúa en el cuaderno II.

Como en la mayoría de los atlas tradicionales, el estímulo con el que se pretendía obtener la información es una palabra o una frase simple que el investigador utilizaba como referencia para enunciar la pregunta (por ejemplo: 625 *Sordo*, 658 *Hijos del hermano* (*sobrinos*), 674 *Nombre y empleo de las habitaciones superiores*, 695a *Nombre y substancias de las comidas más comunes*). Este texto aparece seguido de unas líneas punteadas en las que se anotaban las respuestas en transcripción fonética y también las comentarios y dibujos que se considerasen de interés. Como puede observase en las figuras 2 y 3, la complejidad de las anotaciones no es la misma en las distintas partes del cuestionario. En el cuaderno I (fonética y morfosintaxis) las anotaciones adicionales son muy escasas, pues el interés estaba en las secuencias de sonidos y en la ordenación de las formas (figura 2). En el cuaderno II los comentarios auxiliares, tanto en transcripción fonética como en ortografía convencional, son muy habituales, dado que los encuestadores consideraban de utilidad añadir observaciones a las respuestas (matices de significado, palabras relacionadas, referencias al conocimiento de los conceptos, etc.). En algunas ocasiones el investigador acompañó el texto explicativo de un dibujo simple que aclara el valor de las observaciones, muestra la forma de un objeto o da cuenta del nombre de sus partes. En la pregunta 783a del cuestionario de Cadalso de los Vidrios, Madrid, puede apreciarse el dibujo con que Aurelio M. Espinosa Jr. ilustra

las diferencias entre el *surco* y el *caño* que se forman al arar la tierra para ser sembrada (figura 3).

Además de los datos referidos a preguntas que forman parte del cuestionario, los entrevistadores apuntaron en los márgenes de las páginas otro tipo de informaciones que suponían de interés y que surgían durante la encuesta o en otros momentos de la estancia en el lugar investigado. En el margen inferior de la página 15 del cuaderno de Santiago del Monte, Asturias, Rodríguez Castellanos anota la palabra *cencerrada*, seguida de su transcripción fonética y de la correspondiente a *pandorgada*, que identifica como sinónimo (con el símbolo =). En estas páginas figuran preguntas referidas a las relaciones entre los novios y a la familia, pero ninguna que corresponda al significado de las formas anotadas al margen. En una de las últimas páginas en blanco de este mismo cuestionario, el entrevistador apunta una descripción muy precisa del proceso de elaboración de la manteca, acompañada de un dibujo de la *olla* y otros instrumentos que se usan para batir la leche. Este tipo de anotaciones no son exclusivas del cuaderno de léxico. En el cuaderno I de la misma localidad, en el grupo de preguntas de morfosintaxis (407 *Sirvió sólo dos semanas*), se registra en transcripción fonética *sirvió dos semanas solas* y debajo se añade, también con caracteres fonéticos, *tengo numás unu solu*.

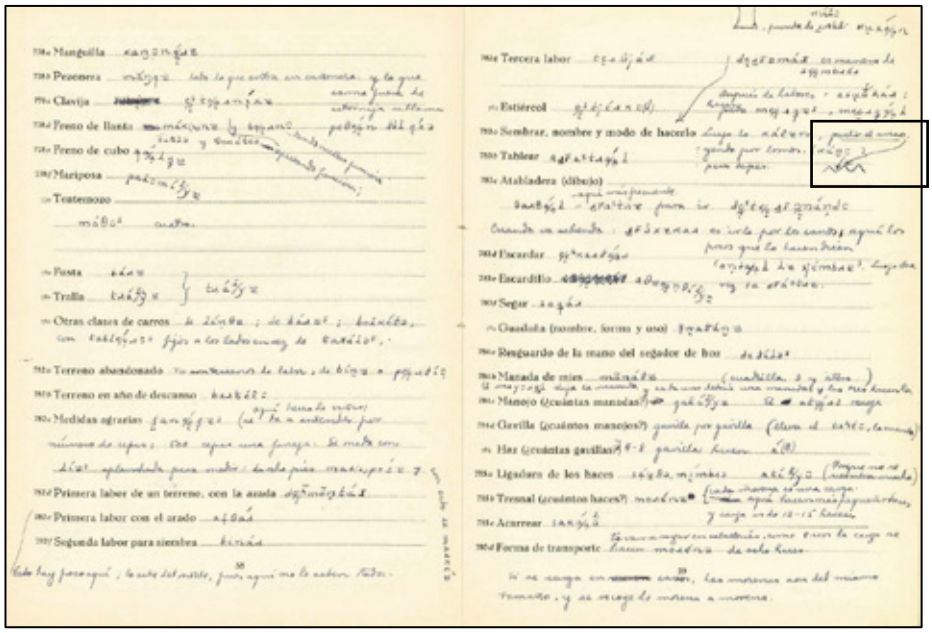


FIGURA 3. Páginas 38 y 39 del cuaderno I de Cadalso de los Vidrios, Madrid

Los ejemplos anteriores muestran que la complejidad de la información contenida en los cuadernos debe ser tratada de un modo especial al transferirla a una base de datos. A esta dificultad para el tratamiento informático de los datos se le añade la de la singularidad del sistema de notación fonética empleado por los encuestadores. Sobre la base del Alfabeto Fonético Internacional y de trabajos dialectales precedentes, Navarro Tomás había diseñado un sistema de símbolos extremadamente complejo con el que pretendía registrar con detalle todos los matices en la pronunciación de las variedades peninsulares. En la introducción del volumen publicado del ALPI se explica el origen de este alfabeto y se justifica su uso.

> En el sistema de transcripción fonética del ALPI se ha pretendido que las notaciones tengan rigor y detalle. Nuestro alfabeto fonético se caracteriza, pues, por su matización, generalmente más minuciosa que en otros Atlas lingüísticos. Es un simple desarrollo del alfabeto definido por nuestro director, el Prof. Tomás Navarro, en *RFE*, 1915, II, 374, y en el *Manual de Pronunciación española*, § 31. El Prof. Navarro adoptó como base, el modo de transcripción usado en las principales revistas de estudios lingüísticos, en el ALF y en el AIS, al cual añadió varios signos diacríticos para la representación detallada de las variedades dialectales. (Navarro Tomás 1962: 6)

A pesar de que existen en la actualidad aplicaciones que permitirían crear una fuente tipográfica que recogiese todos los símbolos del alfabeto original, los investigadores del proyecto consideramos que de esta forma se dificultaría el acceso a los datos. Las transcripciones debían ser volcadas a una base de datos, pero utilizando un alfabeto fonético convencionalizado, el Alfabeto Fonético Internacional, que facilitase a los usuarios la consulta, visualización y explotación de la información de la base (García Mouton 2017). Quien consulte la edición del atlas debe tener la posibilidad de acceder a la información original de los cuadernos, pero también debe poder realizar búsquedas textuales sobre estos datos de forma simple y sin necesidad de conocer en detalle el sistema complejo utilizado por Navarro Tomás y sus colaboradores.

3.2. Digitalización y tratamiento de los materiales

El proyecto de edición de los materiales del ALPI comparte características con otros proyectos de digitalización de atlas lingüísticos que se emprendieron en los últimos años, pero es también una empresa con ciertas singularidades. En primer lugar, el ALPI fue concebido desde el inicio como un atlas lingüístico tradicional para ser publicado en papel. Además, el ALPI no es un proyecto finalizado y del que contemos con una versión completa editada, como sucede con el ALF o el AIS, para la que se busque una nueva forma de difusión y explotación (Embleton, Uritescu & Wheeler 2009). Los materiales de la investigación son los cuadernos en papel del cuestionario y solo una pequeña parte había sido elaborada y publicada en un volumen. En consecuencia, primero era necesario poner todos los materiales del proyecto a disposición de la comunidad científica de forma simple y manejable. El formato de esta nueva edición debía ser digital y tendría que combinar el acceso a las imágenes de los cuadernos originales, la disposición de toda la información textual en un base de datos consultable y la posibilidad de obtener visualizaciones cartográficas de los resultados de las consultas.

La primera tarea abordada fue la digitalización de los cuestionarios de campo correspondientes a las 529 localidades estudiadas. Se obtuvo una imagen de calidad a doble página de los cuadernos que asegura su conservación y permite la vinculación con la base de datos. Cada una de las páginas fue marcada y etiquetada de forma que fuese posible acceder directamente desde la base al recorte de imagen correspondiente a cada pregunta. De esta forma tanto el editor y transcriptor de los materiales como el usuario final tendrían acceso inmediato a las anotaciones originales, una posibilidad que se ofrece en muy pocos de los proyectos de atlas digitalizados y en línea. La digitalización de los cuestionarios era también la forma de preservar los materiales originales del proyecto.

Al mismo tiempo que se digitalizaron los materiales, se comenzó el diseño de las herramientas de edición de la información lingüística. La base de datos que albergaría la información de los cuadernos debía adecuarse a las características complejas de los materiales. Su diseño tenía que facilitar la labor de edición de los datos y permitir la explotación posterior a través de distintos formatos y aplicaciones (búsquedas complejas,

análisis cuantitativos, visualizaciones con SIG, etc.). La información lingüística registrada en los atlas tiene carácter pluridimensional, tanto los atlas tradicionales como en los propiamente denominados multidimensionales. Cada variante está a asociada a un conjunto de rasgos de tipo distinto: localización geográfica, nivel de análisis lingüístico, hablante (sexo, edad, profesión, etc.), fecha de registro, etc. El diseño de la base de datos debe hacer posible la recuperación de toda esta información de forma simple.

3.3. La base de datos del ALPI

La utilización compartida de innovaciones informáticas en el ámbito de los estudios geolingüísticos ha hecho posible que en las dos últimas décadas el diseño de los proyectos tienda a la utilización de ciertos estándares. Con todo, los intentos realizados para establecer un catálogo de clases y propiedades de datos de los proyectos de geografía lingüística no han dado todavía el fruto deseado. La propuesta más completa y sólida de una ontología estandarizada de datos es la sugerida por Buccio para el *Atlante Sintattico d'Italia* (Di Buccio, Di Nunzio & Silvello 2013). Esta falta de consenso y la heterogeneidad de los datos de los atlas lingüísticos han sido utilizadas para justificar la dificultad de reutilización de los datos y la escasa presencia de este tipo de proyectos en repositorios abiertos de datos lingüísticos (LLOD).

La organización de la base de datos del ALPI toma como base la experiencia de otros atlas similares, fundamentalmente el *Atlas Lingüístico Galego* (Sousa 2017) y el *Atlas Dialectal de Madrid* (García Mouton & Molina 2015), y se adapta a las características propias del proyecto. La organización básica parte de la distinción de tres bloques principales que agrupan las distintas tablas de contenido: a) localidades; b) informadores; y c) documentación lingüística.

A. Localidades
   En esta categoría se registra la información referida a cada una de las localidades que constituyen la red del ALPI: código de la localidad, nombre, parroquia, municipio, diócesis, provincia, comunidad autónoma, datos socioeconómicos y coordenadas geográficas correspondientes. Estos datos identifican cada uno de los cuestionarios, permiten localizar geográficamente las repuestas y se pueden utilizar para procesar y visualizar la información con cualquier aplicación que funcione como sistema de información geográfica.

B. Informadores
   Bajo esta clase se agrupa la información referida a los sujetos entrevistados. En la tabla se recogen los datos que fueron anotados en los cuestionarios: nombre, edad, lugar de nacimiento, lugar de residencia habitual, ocupación, alfabetización, etc. Cada informante está asociado con una localidad y en consecuencia con un cuestionario.

C. Documentación lingüística
   Este conjunto de tablas reúne toda la información lingüística de la base de datos: la relación de preguntas de cuestionario, las transcripciones de las respuestas, los comentarios y los dibujos añadidos al margen y también la imagen digitalizada de las páginas de los cuadernos. Estas tablas están vinculadas con otras que contienen datos añadidos en la edición de los

materiales (clasificadores semánticos, etiquetas morfosintácticas, traducciones a otros idiomas, etc.).

Las relaciones entre las tablas de datos de estos tres grupos de contenidos están trazadas para facilitar la labor de los editores y permitir que los usuarios finales pueden acceder, de distintas formas, a toda la información lingüística registrada. Por ejemplo, la vinculación entre las imágenes de los cuadernos y la tabla de respuestas permite que tanto el editor como el usuario final puedan consultar los cuadernos originales; la conexión entre las localidades, los dibujos y las repuestas hace posible, por ejemplo, la elaboración de un mapa ilustrado de los distintos tipos de cántaros.

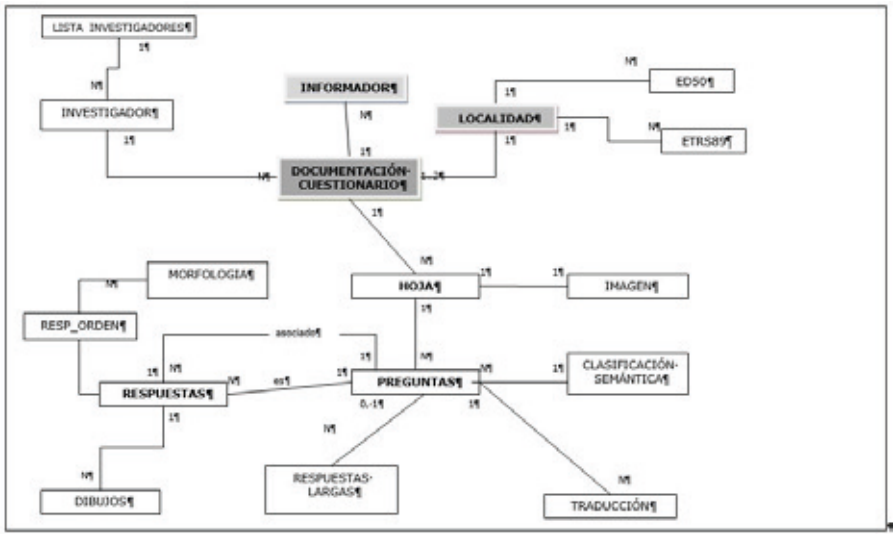


FIGURA 4. Esquema de las tablas y relaciones que constituyen la base de datos del ALPI. Con distintos sombreados se destacan los tres tipos de componentes básicos

3.4. La aplicación de edición

A partir de la estructura y componentes de la base de datos se diseñó una herramienta de edición para facilitar la tarea de transcripción de los materiales. Como en el proyecto participan equipos de distintas universidades y centros de investigación, se trata de una aplicación web accesible desde un navegador y conectada con la base de datos general del proyecto. El editor comienza eligiendo la localidad y el cuaderno sobre los que desea trabajar. En respuesta la aplicación muestra en pantalla la imagen del cuaderno y los campos en que deben introducirse las transcripciones. La herramienta permite navegar por las páginas del cuestionario para seleccionar las preguntas y también permite ampliar la imagen (figura 5). En la mitad inferior de la pantalla aparecen los campos en los que deben introducirse la información textual. En el campo *Respuesta IPA* el editor tiene que registrar la transcripción fonética de la respuesta con los símbolos del alfabeto fonético internacional. Para esta tarea se dispone de una guía que muestra las correspondencias entre los símbolos originales y los símbolos del AFI (García Mouton 2010). Los símbolos fonéticos se insertan a través del teclado convencional o bien usando un teclado virtual, muy útil para incorporar diacríticos. En el campo *Respuesta ORTO* se introduce el texto de la respuesta en la grafía convencional de la lengua correspondiente[3]. Los campos restantes se utilizan para introducir

[3] Los iconos de las banderas dan acceso a diccionarios electrónicos de español, gallego, asturiano, portugués y

informaciones complementarias de distinto tipo (anotaciones de los encuestadores, anotaciones del editor, etiquetas morfosintácticas, clasificadores semánticos, etc.). La herramienta permite también registrar en la base los apuntes que aparecen en los márgenes de las páginas y en las páginas finales de los cuadernos. Toda esta información es siempre recuperable desde la aplicación de consulta.

La aplicación de edición incorpora además una herramienta de revisión y otra de captura y clasificación de dibujos. La primera permite obtener un listado de todas las repuestas a una misma pregunta para identificar incoherencias en las transcripciones y realizar correcciones. La segunda ayuda a recortar los dibujos y a vincularlos con la localidad y la respuesta correspondientes.

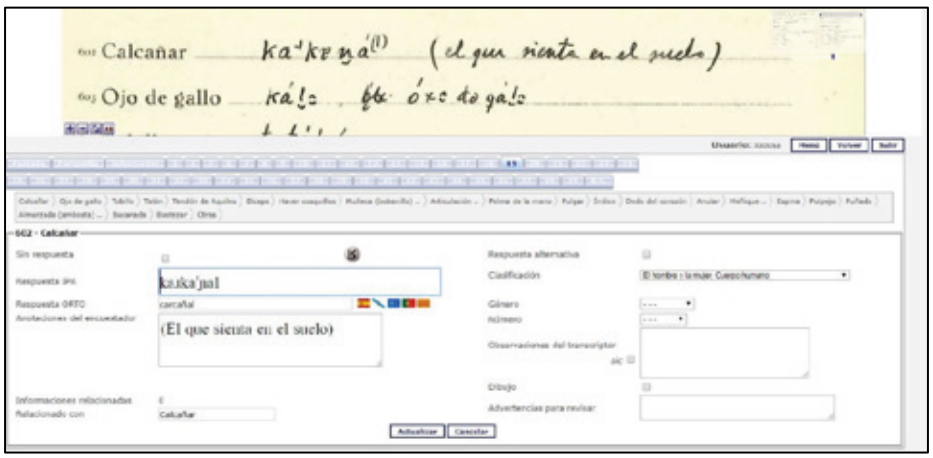


FIGURA 5. Captura de pantalla de la herramienta de edición del ALPI

3.5. La aplicación de consulta

Desde la página web principal del proyecto de edición y elaboración de los materiales del ALPI se puede acceder a una versión de muestra de la herramienta de consulta de los materiales[4]. Seleccionando la pestaña *Consulta de mapas* se abre una nueva página que contiene la relación de las 10 preguntas consultables y una descripción de la información que se ofrece. La consulta actual (figura 6) está limitada a los campos *Pregunta* (texto de la pregunta en el cuestionario original), *Respuesta IPA* (secuencia de caracteres fonéticos) y *Respuesta ortográfica* (secuencia de caracteres en ortografía convencional). La versión final ofrecerá más opciones de búsqueda (número de la pregunta, segmentos de texto, textos marginales, dibujos, campos semánticos, clases de palabras, unidades morfológicas, etc.), que además podrán combinarse de distintas formas. La consulta puede realizarse sobre todos los puntos de la red o bien sobre los puntos de una zona geográfica, una provincia o sobre una localidad. En la figura 7 se muestra la tabla de resultados para la pregunta 458 *Guisantes* en las localidades de Cantabria. En las columnas figura información sobre la localidad, el número de cuaderno, la provincia, el texto de la pregunta en el cuestionario original, la transcripción de la respuesta en ortografía convencional y en alfabeto fonético y los iconos de las opciones de visualización (ficha completa de la transcripción, imagen de cuaderno original y eliminación de la fila de datos). En la versión final la tabla de datos podrá descargarse en formatos que permitan el tratamiento de la información en distintos programas de

catalán para consultar las formas gráficas recomendadas.

[4] En la misma página es posible descargar un documento de ayuda en el que se ofrece información detallada sobre el funcionamiento de la herramienta (http://alpi.csic.es/es/consulta).

análisis de información lingüística y datos georreferenciados (texto, csv, tab, xml, etc.). Los datos obtenidos de las consultas pueden visualizarse sobre un mapa utilizando símbolos o polígonos. En la figura 8 se ofrece un ejemplo de la distribución de la respuesta *arveja(s)* para la pregunta *Guisantes*. Los círculos rojos identifican los puntos en que se registró esta respuesta. El mapa se obtiene gracias a una herramienta GIS conectada con la base de datos[5].

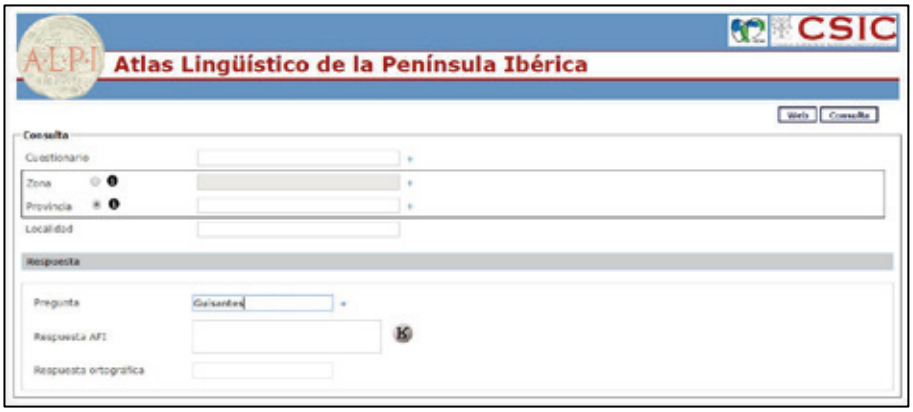


FIGURA 6. Captura de pantalla de la aplicación provisional de consulta del ALPI

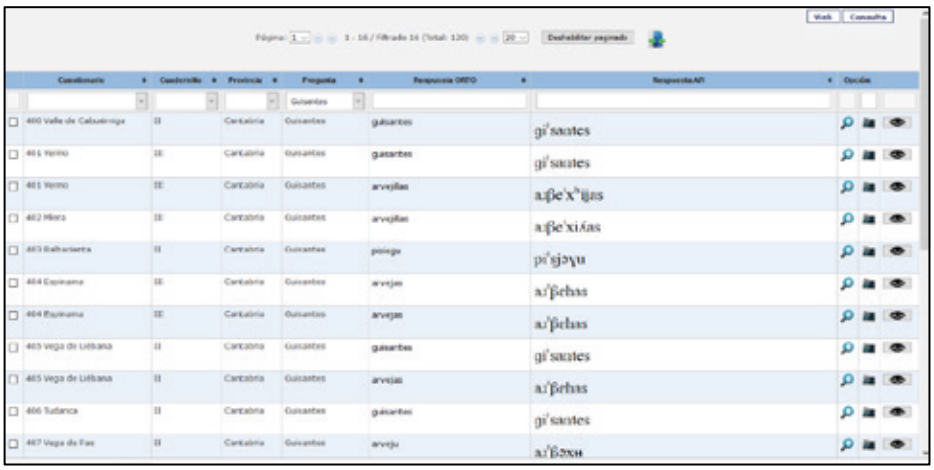


FIGURA 7. Tabla con las repuestas en localidades de Cantabria a la pregunta *Guisantes*

[5] La aplicaciones informáticas vinculadas con el proyecto fueron diseñadas y desarrolladas en el Centro de Ciencias Humanas y Social del CSIC con la colaboración de Juan Carlos Martínez Torres, Ángel Díaz del Castillo, Pilar García Mouton, responsable científica, y el departamento de SIG del centro (García Mouton 2017).

FIGURA 8. Mapa con la distribución de la variante léxica *arveja(s)* como respuesta a la pregunta *Guisantes*

4. CONCLUSIÓN Y TRABAJO FUTURO

En las últimas tres décadas los estudios de dialectología han gozado de un desarrollo teórico y metodológico que ha contribuido a la renovación de esta vieja disciplina, a la que en los años sesenta algunos autores habían sentenciado de muerte. La situación actual en que se encuentra la geografía lingüística, subdisciplina de la dialectología, se debe en buena medida a las consecuencias positivas que ha tenido la extensión de las humanidades digitales al ámbito de los estudios dialectales. Los actuales proyectos de geografía lingüística son un paradigma de aprovechamiento de los recursos que la tecnología informática pone a disposición de la investigación en ciencias humanas.

Los atlas lingüísticos son obras resultado de proyectos complejos y costosos que en pocas ocasiones ven reconocido su auténtico valor como fuentes de documentación lingüística (García Mouton 2012b). Los atlas editados en papel son con frecuencia obras de formato singular, de manejo difícil y de consulta complicada. La edición moderna de un atlas lingüístico debe aprovechar todos los medios que la informática pone a su disposición con el propósito de facilitar el acceso a la información (Kretzschmar 2006; 2013) y satisfacer las demandas actuales de la investigación en lingüística: i) los datos deben ser accesibles de forma libre y de acuerdo con los principios de la ciencia abierta; ii) la información debe ofrecerse en formatos estándares que permitan su explotación de forma interdisciplinar; y iii) las herramientas de acceso deben diseñarse para distintos tipos de usuarios, tanto expertos como no expertos.

Los materiales del *Atlas Lingüístico de la Península Ibérica* son un tesoro documental de las hablas romances de la península ibérica del siglo XX que debe ser difundido y puesto a disposición de la comunidad científica. Su valor documental e histórico era reconocido ya por Navarro Tomás poco antes de su fallecimiento: «el ALPI es como una especie de acta documental del carácter y fisonomía del habla popular de la Península en los años inmediatamente anteriores a la guerra civil. La honda conmoción producida por esta guerra en todo el país, y el movimiento de población ocasionado después por motivos económicos y sociales, habrán modificado sin duda alguna las líneas

del ALPI, lo cual acentúa su interés como testimonio de valor histórico» (Navarro Tomás 1975: 14). El propósito de los colaboradores de este proyecto es convertir al ALPI en un repositorio abierto que estimule la investigación lingüística sobre las hablas rurales peninsulares (García Mouton 2017).

REFERENCIAS BIBLIOGRÁFICAS

ALINEI, M., J. ALLIÈRES, R. I. AVANESOV, T. ITKONENE, W. VIERECK & A. A. WEIJNEN (coords.) (1983): *Atlas linguarum Europae (ALE)*. Assen/Maastricht/Rome: van Gorcum/Istituto Poligrafico e Zecca dello Stato.

ALONSO MONTERO, X. (2011): *Aníbal Otero: Lingüística e política en España na Guerra Civil e no franquismo*. Vigo: Xerais.

AURREKOETXEA, G. (2008): "Basque linguistic atlas-EHHA: From speech to automatic maps", *Dialectologia* 1, pp. 107–119.

BAUER, R. & H. GOEBL (2000): "Utilisation nouvelle de l'informatique dans les atlas linguistiques en Europe (1980-2000)", *Verbum* 22, pp. 169-185.

BRITAIN, D. (2014): "Geographical dialectology", in J. Holmes (coords.): *Research methods in sociolinguistics: A practical guide*. Chichester West Sussex UK, Malden MA: Wiley Blackwell, pp. 246–261.

BRUN-TRIGAUD, G. (2016): "Vers un renouveau des atlas linguistiques régionaux?", *Géolinguistique* 16, pp. 7–20.

CLARKE, S. (2016): "From legacy regional language materials to public engagement: The interactive online Dialect Atlas of Newfoundland and Labrador", in J. C. Beal, K. P. Corrigan & H. Moisl (coords.): *Creating and digitizing language corpora*. London: Palgrave, pp. 99–132. https://doi.org/10.1057/978-1-137-38645-8_4

CORTÉS, S. & V. GARCÍA PERALES (2009): *La historia interna del "Atlas Lingüístico de la Península Ibérica" (ALPI): Correspondencia (1910-1976)*. Valencia: Universitat de València.

DAVOINE, P.-A., S. GALLY, P. GARAT, C. CHAUVIN, O. COPI & C. CAVALLIERE (2015): "New approach to explore and to study cartographical heritage in dialectology: application to the Linguistic Atlas of France", in *27th International Cartographic Conference*. Rio de Janeiro: ICA.

DI BUCCIO, E., G. M. DI NUNZIO & G. SILVELLO (2013): "A geolinguistic web application based on linked open data", in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*. New York: ACM, 1101-1102. http://dx.doi.org/10.1145/2484028.2484219

DI BUCCIO, E., G. M. DI NUNZIO & G. SILVELLO (2014): "A linked open data approach for geolinguistics applications", *International Journal of Metadata, Semantics and Ontologies* 9, p. 29. http://dx.doi.org/10.1504/IJMSO.2014.059125.

EMBLETON, S., D. URITESCU & E. S. WHEELER (2004): "An exploration into the management of high volumes of complex knowledge in the social sciences and humanities", *Journal of Quantitative Linguistics* 11, pp. 183–192. http://dx.doi.org/10.1080/09296170420003149 30.

EMBLETON, S., D. URITESCU & E. S. WHEELER (2009): "Lessons from digitizing a linguistic atlas", in L. Botoşineanu, E. Dănilă, C. Holban & O. Ichim (coords.): *Distorsionări în comunicarea lingvistică, literară şi etnofolclorică românească şi contextul european*. Iaşi: Editura ALFA, pp. 137–145.

FRIENDLY, M. (2008): "A brief history of data visualization", in W. Chen, W. Hardle & A. Unwin (coords.): *Handbook of data visualization*. Berlín: Springer, pp. 15–56. https://doi.org/10.1007/978-3-540-33037-0_2

GARCÍA MOUTON, P. (2010): "El procesamiento informático de los materiales del Atlas Lingüístico de la Península Ibérica de Tomás Navarro Tomás", in G. Aurrekoetxea & J. L. Ormaetxea (coords.): *Tools for linguistic variation*. Bilbao: Universidad del País Vasco, Servicio Editorial = Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua, pp. 167–174.

GARCÍA MOUTON, P. (2011): "Tomás Navarro Tomás y la metodología del Atlas Lingüístico de la Península Ibérica (ALPI)", in M. Nevaci (coord.): *Studia linguistica et philologica: Omagiu profesorului Nicolae Saramandu la 70 de ani*. Bucureşti: Universitaţii din Bucureşti, pp. 377–383.

GARCÍA MOUTON, P. (2012a): "Editar el 'Atlas lingüístico de la Península Ibérica' en el siglo XXI", in C. Corrales Zumbado & D. Corbella Díaz (coords.): *Lexicografía hispánica del siglo XXI: Nuevos proyectos y perspectivas homenaje al profesor Cristóbal Corrales Zumbado*. Madrid: Arco Libros, pp. 323–330.

GARCÍA MOUTON, P. (2012b): "El Atlas Lingüístico de la Península Ibérica (ALPI) como fuente de documentación", in E. Montero Cartelle & C. Manzano Rovira (coords.): *Actas del VIII Congreso Internacional de Historia de la Lengua Española: Santiago de Compostela, 14-18 de septiembre de 2009*. Santiago de Compostela: Meubook, pp. 273–276.

GARCÍA MOUTON, P. (2015): "Los trabajos del Atlas Lingüístico de la Península Ibérica (ALPI) y la Revista de Filología Española", in P. García Mouton & M. Pedrazuela Fuentes (coords.): *La ciencia de la palabra: Cien años de la Revista de filología española*. Madrid: Consejo Superior de Investigaciones Científicas, pp. 175–208.

GARCÍA MOUTON, P. (2017): "El Atlas Lingüístico de la Península Ibérica (ALPI) en línea. Geolingüística a la carta", *Estudis romànics* 39, pp. 335–343.

GARCÍA MOUTON, P., I. FERNÁNDEZ-ORDÓÑEZ, D. HEAP, M. P. PEREA, J. SARAMAGO & X. SOUSA (2016): *ALPI-CSIC. Edición digital de Navarro Tomás, Tomás (dir.), Atlas Lingüístico de la Península Ibérica, from* www.alpi.csic.es.

GARCÍA MOUTON, P. & I. MOLINA (2015): *Atlas Dialectal de Madrid* (ADiM). Madrid: CSIC. http://adim.cchs.csic.es

GARCÍA MOUTON, P. & M. PEDRAZUELA FUENTES (coords.) (2015): *La ciencia de la palabra: Cien años de la Revista de filología española*. Madrid: Consejo Superior de Investigaciones Científicas.

HEAP, D. (2002): "Segunda noticia histórica del ALPI (a los cuarenta años de la publicación de su primer tomo)", *Revista de Filología Española* 82, pp. 5–19. http://dx.doi.org/10.3989/rfe.2002.v82.i1/2.141.

HEAP, D. (2012): "Ten years of the online ALPI (Atlas Lingüístico de la Península Ibérica)", *Dialectologia Special Issue III: Linguistic Atlas of the Iberian Peninsula (ALPI): Progress and Perspectives*, pp. 43–56.

HICKEY, R. (2018): "Dialectology, philology, and historical linguistics", in C. Boberg, J. A. Nerbonne & D. J. L. Watt (coords.): *The handbook of dialectology*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 23–38. https://doi.org/10.1002/9781118827628.ch1

HOCH, S. & J. J. HAYES (2010): "Geolinguistics: The incorporation of geographic information systems and science", *Geographical Bulletin - Gamma Theta Upsilon* 51, pp. 23–36.

JESSOP, M. (2007): "The Inhibition of geographical information in digital humanities scholarship", *Literary and Linguistic Computing* 23, pp. 39–50. http://dx.doi.org/10.1093/llc/fqm041.

KRETZSCHMAR, W. A. (2006): "Art and science in computational dialectology", *Literary and Linguistic Computing* 21, pp. 399–410. http://dx.doi.org/10.1093/llc/fql033.

KRETZSCHMAR, W. A., P. BOUNDS & N. PALOSAARI (2011): "Issues in using legacy data", in M. Di Paolo & M. Yaeger-Dror (coords.): *Sociophonetics: A student's guide*. London, New York: Routledge, pp. 46-57.

KRETZSCHMAR, W. A. & W. GRAY POTTER (2010): "Library collaboration with large digital humanities projects", *Literary and Linguistic Computing* 25, pp. 439–445. http://dx.doi.org/10.1093/llc/fqq022

KRETZSCHMAR, W. A., J. (2013): "Making sociolinguistic data accessible", in C. Mallinson, B. Childs & G. van Herk (coords.): *Data collection in sociolinguistics: Methods and applications*. New York: Routledge Taylor & Francis Group, pp. 206–209.

KUMAGAI, Y. (2016): "Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical dis-tributions of dialects", in M.-H. Côté, R. Knooihuizen & J. Nerbonne (coords.): *The future of dialects*. Berlin: Language Science Press, pp. 333–362.

LAMELI, A. (2010): "Linguistic atlases. Traditional and modern", in P. Auer & J. E. Schmidt (coords.): *Language and space: An international handbook of linguistic variation. Theories and methods*. Berlin, New York: De Gruyter Mouton, pp. 567–592. https://doi.org/10.1515/9783110220278.567

LIMPER, J., J. PHEIFF & A. WILLIAMS (2019): "REDE SprachGIS: A geographic information system for linguists", in S. Brunn & R. Kehrein (eds.): *Handbook of the*

changing world language map. Cham: Springer. https://doi.org/10.1007/978-3-319-73400-2_145-1

NAVARRO TOMÁS, T. (coord.) (1975): *Capítulos de geografía lingüística de la Península Ibérica*. Bogotá: Instituto Caro y Cuervo.

NAVARRO TOMÁS, T. (dir.) (1962): *Atlas Lingüístico de la Península Ibérica, I, Fonética*. Madrid: CSIC.

OLARIU, F.-T. & V. OLARIU (2014): "The Romanian linguistic cartography in the digitizing era: the electronic atlases", *Dialectologia et geolinguistica* 22, pp. 75-90. http://dx.doi.org/10.1515/dialect-2014-0005.

ORMELING, F. (2015): "Linguistic map", in M. Mononier (coord.): *Cartography in the Twentieth Century*. Chicago: University of Chicago Press, pp. 776–782.

PÉREZ PASCUAL, J. I. (2016): *Los primeros pasos de un largo caminar: Los comienzos del Atlas Lingüístico de la Península Ibérica*. San Millán de la Cogolla: Cilengua - Centro Internacional de Investigación de la Lengua Española.

PRESNER, T. & D. SHEPARD (2016): "Mapping the geospatial turn", in S. Schreibman, R. G. Siemens & J. Unsworth (coords.): *A new companion to digital humanities*. Chichester West Sussex UK: John Wiley & Sons, Inc, pp. 201–212. https://doi.org/10.1002/9781118680605.ch14

RABANUS, S., R. KEHREIN & A. LAMELI (2010): "Creating digital editions of historical maps", in A. Lameli, R. Kehrein & S. Rabanus (coords.): *Language and space: An international handbook of linguistic variation. Language mapping*. Berlin, New York: De Gruyter Mouton, pp. 375–385. https://doi.org/10.1515/9783110219166.1.375

SCHEURINGER, H. (2010): "Mapping the German language", in A. Lameli, R. Kehrein & S. Rabanus (coords.): *Language and space: An international handbook of linguistic variation. Language mapping*. Berlin, New York: De Gruyter Mouton, pp. 158–179. https://doi.org/10.1515/9783110219166.1.158

SCHMIDT, J. E. & J. HERRGEN (2001): *Digitaler Wenker-Atlas (DiWA): Erste vollständige Ausgabe von Georg Wenkers "Sprachatlas des Deutschen Reichs". 1888–1923 handgezeichnet von Emil Maurmann, Georg Wenker und Ferdinand Wrede*. http://www.diwa.info

SCHMIDT, J. E., J. HERRGEN & R. KEHREIN (2008): *Regionalsprache.de (REDE): Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. https://regionalsprache.de/

SCHREIBMAN, S., R. G. SIEMENS & J. UNSWORTH (coords.) (2016): *A new companion to digital humanities*. Chichester West Sussex UK: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118680605

SOUSA, X. (2017): "From field notebooks to automatic mapping: the 'Atlas Lingüístico Galego' database", *Dialectologia et geolinguistica* 25, pp. 1–22. http://dx.doi.org/10.1515/dialect-2017-0001.

TISATO, G. G. (2009): *NavigAIS: AIS Digital Atlas and Navigation Software. [Digital version of Jaberg & Jud] 1928–1940]*. http://www3.pd.istc.cnr.it/navigais/

VEITH, W. H., W. PUTSCHKE & L. HUMMEL (1984-1999): *Kleiner Deutscher Sprachatlas (KDSA)*. Tübingen: Niemeyer.

VIERECK, W. & H. RAMISCH (1991-1997): *The Computer Developed Linguistic Atlas of England (CLAE)*. Tübingen: Niemeyer.

VARIACIÓN GRAMATICAL DEL ESPAÑOL EN EL MUNDO (VARIGRAMA): UNA VISIÓN PANORÁMICA DE LOS RASGOS SINTÁCTICOS DEL ESPAÑOL

Variación Gramatical del Español en el Mundo *(VARIGRAMA): A panoramic view of some syntactic features of Spanish*

TOSHIHIRO TAKAGAKI
Universidad de Estudios Extranjeros de Tokio

HIROTO UEDA
Universidad de Tokio

ANTONIO RUIZ TINOCO
Universidad Sofía

Resumen
En este trabajo se describen el objeto y el planeamiento del proyecto, la encuesta y la recogida de datos. A continuación, presentaremos varios casos concretos de análisis de temas gramaticales del español como los de 'entrar a / en', del subjuntivo, del *leísmo* y *queísmo* / *dequeísmo*, acompañados de mapas elaborados para esta ocasión. Por último, se menciona la aplicación del sistema LYNEAL (elaborado por H. Ueda) que nos permite ver cómo funcionan nuestros datos en la web.

Palabras clave: gramática española, variación sintáctica, geografía lingüística, dialectología

Abstract
This study is intended to describe the object and approach of our project VARIGRAMA, and the results obtained in the questionnaires carried out in situ. Several cases of syntactic variation will be discussed, such uses of 'entrar a / en', of subjunctive mood, *leísmo* and *queísmo* / *dequeísmo* illustrated with distributional maps. Special mention should be made of the web-based system LYNEAL, which has made the comparison of the data collected in the successive stages of our questionnaire possible.

Keywords: Spanish grammar, syntactic variation, linguistic geography, dialectology

1. EL PROYECTO VARIGRAMA

El proyecto de investigación VARIGRAMA (*Variación Gramatical del Español en el Mundo*)[1], iniciado en 2001, tiene por objetivo conocer la situación actual de ciertos temas gramaticales del español utilizado en las principales ciudades (capitales en su mayor parte) de todo el territorio de habla española. El proyecto consiste en realizar *in situ* una encuesta dirigida a una media de 20 informantes nativos. El cuestionario consta de unas 110 oraciones en español provistas de algunos rasgos sintácticos que consideramos de mayor relevancia. Lógicamente el cuestionario no puede modificarse en las diferentes etapas de la encuesta, y tras haber pasado cerca de quince años desde las primeras encuestas en 2004 en España, el contenido no ha sido sujeto a cambios sustanciales.

La misma encuesta se ha llevado a cabo en diferentes etapas[2] en 10 ciudades españolas [Oviedo, Pamplona, Alcalá de Henares, Madrid, Barcelona, Salamanca, Sevilla, Huelva, Tenerife, Las Palmas], por un lado y en 18 ciudades latinoamericanas, por otro. [Ciudad de México (México), La Habana (Cuba), Santo Domingo (República Dominicana), San Juan (Puerto Rico), Ciudad de Guatemala (Guatemala), San Salvador (El Salvador), San José (Costa Rica), Ciudad de Panamá (Panamá), Caracas (Venezuela), Bogotá (Colombia), Quito (Ecuador), Lima (Perú), La Paz (Bolivia), Asunción (Paraguay), Santiago (Chile), Montevideo (Uruguay) y Buenos Aires (Argentina)].

［Ejemplo 1］ Yo la dije la verdad. (la = María)

✓(A) Yo lo diría así.

(B) Yo no lo diría, pero lo he oído decir.

(C) Yo no lo diría ni lo he oído decir.

(D) Comentarios: _____

［Ejemplo 2］ No estoy seguro(a) de que tenían dinero.

(A)Yo lo diría así.

✓(B) Yo no lo diría, pero lo he oído decir.

(C) Yo no lo diría ni lo he oído decir

(D) Comentarios: *Diría: tuvieran*

FIGURA 1. Ejemplos del cuestionario

El formato del cuestionario es muy sencillo. Ante las oraciones preguntadas los encuestados tienen que contestar eligiendo solo entre las tres opciones: (A) Yo lo diría así, (B) Yo no lo diría, pero lo he oído decir, y (C) Yo no lo diría ni lo he oído decir, como se ve en los ejemplos de la figura 1. Pueden añadir comentarios en el espacio (D) para cualquier clarificación.

Los resultados de las encuestas se clasifican en tablas como la siguiente, (tabla 1), donde figuran los datos relativos a ciudad, sexo y edad del encuestado. A la derecha aparecen recogidas tales opciones abreviadas como DIGO, OIGO y NO, de las cuales

[1] Los miembros del proyecto son Noritaka Fukushima y Masami Miyamoto, junto con los autores de este trabajo.

[2] Las encuestas han tenido lugar en los años 2004, 2008, 2010, 2014, 2016 y 2018. Le estamos muy agradecidos a Luis Ortiz la valiosa colaboración en las últimas dos encuestas de El Salvador y Guatemala, cuyos resultados no están reflejados en este trabajo.

identificamos las respuestas positivas en la columna DIGO como índice de aceptabilidad de la oración en cuestión[3].

TABLA 1. Resultados de las encuestas

País	Sexo	Edad	1: DIGO	2: OIGO	3: NO	TOTAL
1:México	Hombre	10-19	2			2
		20-29	9			9
	Hombre Total		11			11
	Mujer	20-29	8	1		9
		40-49		1		1
	Mujer Total		8	2		10
1:México Total			19	2		21
2:Colombia	Hombre	10-19	7			7
		20-29	6			6
	Hombre Total		13			13
	Mujer	10-19	10			10
		20-29	1	1		2
	Mujer Total		11	1		12
2:Colombia Total			24	1		25
3:Chile	Hombre	20-29	6			6
	Hombre Total		6			6
	Mujer	20-29	19			19
	Mujer Total		19			19
3:Chile Total			25			25
4:Paraguay	Hombre	10-19	2			2
		20-29	5	2		7
	Hombre Total		7	2		9
	Mujer	10-19	3			3
		20-29	8			8
	Mujer Total		11			11
4:Paraguay Total			18	2		20
5:Argentina	Hombre	10-19	1			1
		20-29	7			7
	Hombre Total		8			8
	Mujer	20-29	5			5
		30-39	2			2
		40-49	1			1
		50-59	4			4
	Mujer Total		12			12
5:Argentina Total			20			20
Total			106	5		111

[3] Los resultados se pueden consultar en la página web del proyecto VARIGRAMA. (http://lecture.ecc.utokyo.ac.jp/~cueda/varigrama/index.html). Este proyecto ha sido subvencionado por KAKENHI, de Japan Society for the Promotion of Science.

2. ANÁLISIS SINTÁCTICOS DE LOS DATOS—A MODO DE EJEMPLO

A continuación presentaremos algunos ejemplos de análisis sintáctico, basándonos en los resultados de las encuestas. Los temas abordados en este trabajo se centran en *entrar en/a*, los modos, *leísmo* y *queísmo-dequeísmo*, que resultan ser de carácter variado.

Las 110 frases del cuestionario (Aparecen en el APÉNDICE [2]) fueron seleccionadas para cubrir los más variados temas posibles que nos parecían relevantes para la investigación de la gramática española. Nuestra intención inicial era proporcionar al alcance de los interesados los resultados obtenidos en las encuestas que facilitaran un estudio más profundizado sobre cualquier tema contenido en el cuestionario. Lo que se expone en estos apartados no son estudios monográficos pormenorizados, sino que se limita a demostrar posibles aplicaciones de investigación en función de los datos cosechados en las encuestas de las principales ciudades del mundo hispánico.

2.1. *Entrar en/a*

En muchos estudios (Lope Blanch 1964, De Bruyne 1999, Butt y Benjamin 2000, RAE 2005) se generaliza que en Latinoamérica se prefiere la preposición *a* en combinación con el verbo *entrar*, mientras que en el español peninsular se utiliza principalmente *en*, aunque se escucha también *entrar a*.

Veamos al respecto los resultados de la oración (1) incluida en nuestro cuestionario.

Entrar a /en

(1) Ellos entraron *al edificio. (N. 14[4])

En la tabla 2 se muestran los resultados correspondientes a las respuestas DIGO, OIGO, NO, conseguidas en los lugares encuestados de España[5]:

Frec.	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	18	16	8	11	10	16	10	4	17	19
2:Oigo.	9	4	11	7	1	5	13	12	3	3
3:No.		1	1				1	4		2

TABLA 2. Respuestas (España) *Ellos entraron *al edificio.* (N. 14)

[4] La numeración corresponde a la del cuestionario que aparece en el APÉNDICE [2]. En el presente trabajo, hemos utilizado el nuevo sistema LYNEAL de nuestra elaboración para realizar tanto los cálculos de frecuencia y porcentaje como la visualización cartográfica de distribución en España y América. El mismo sistema se ha implementado en Madrid y en Tokio:

http://shimoda.lllf.uam.es/ueda/lyneal/

http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/

[5] Las abreviaturas de las ciudades españolas que se mencionan en el estudio son las siguientes: Ovi E01:Oviedo; Pam: E02:Pamplona; Bar E03:Barcelona; Sal E04:Salamanca; Alc E05:Alcalá; Mad E06:Madrid; Sev E07:Sevilla; Hue E08:Huelva; Ten E09:Tenerife; Pal E10:Las Palmas.

%	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	67	76	40	61	91	76	42	20	85	79
2:Oigo.	33	19	55	39	9	24	54	60	15	13
3:No.		5	5				4	20		8

TABLA 3. Porcentajes (España) *Ellos entraron *al edificio*. (N. 14)

Los números exhibidos en la tabla 3 son los porcentajes de cada respuesta respecto al total de los encuestados para cada lugar. Estas proporciones de respuestas se podrían considerar como índices de "aceptabilidad" de la oración por parte de los participantes locales en la encuesta, como acabamos de mencionar.

Por otro lado, tenemos los resultados que corresponden a los países latinoamericanos donde hemos realizado las mismas encuestas. Podemos ver las respuestas en la tabla 4 y sus porcentajes en la tabla 5[6]:

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	19	21	32	18	29	21	24	24	20	22	33	25	18	19	20
2:O	2	2	3		2	1	1	1	3	1	5		2	1	
3:N			2								1				

TABLA 4. Respuestas (Latinoamérica) *Ellos entraron *al edificio*. (N. 14)

%	ME	JO	PA	HA	DO	JU	CA	BO	QU	LI	LP	STG	AS	MO	BU
D	90	91	86	100	94	95	96	96	87	96	85	100	90	95	100
O	10	9	8		6	5	4	4	13	4	13		10	5	
N			5								3				

TABLA 5. Porcentajes (Latinoamérica) *Ellos entraron *al edificio*. (N. 14)

La supremacía numérica general de la preposición *a* se observa en los mapas lingüísticos (figura 3: España y figura 4: Latinoamérica)[7]. Se nota cierta diferencia entre España y América en el sentido de que, en la primera, sobre todo en tres ciudades, Barcelona, Sevilla y Huelva, se prefiere el uso de *en*, en lugar de *a*. En general, sin embargo, no se puede hablar de una marcada tendencia a usar *en*.

En cambio, la extensión de *a* en Latinoamérica es definitiva, sin excepción. Es interesante relacionar esta distribución contrastiva entre España y América con los cambios históricos. Según el diccionario etimológico de Corominas y Pascual (s.v. *entrar*), "en la España de Edad Media y en los clásicos, era *entrar en* o *entrar a* indiferentemente", lo cual confirma la continuidad cronológica de *entrar a* a lo largo de la historia. La misma historia se remonta al latín (Lewis, 1975: 439), donde el verbo INTRARE regía tanto la preposición IN como AD. La tendencia actual parece dirigirse al

[6] FA representa la frecuencia de aceptabilidad. Asimismo, las abreviaturas de las ciudades americanas corresponden a las siguientes: ME L01:Ciudad de México; JO L02:San José; PA L03:Panamá; HA L04:La Habana; SD L05:Santo Domingo; JU L06:San Juan; CA L07:Caracas; BO L08:Bogotá; QU L09:Quito; LI L10:Lima; LP L11:La Paz; STG L12:Santiago; AS L13:Asunción; MT L14:Montevideo; BU L15:Buenos Aires. Las iniciales de los porcentajes pertenecen a "Digo" (D), "Oigo" (O) y "No" (N).

[7] Los mapas geográficos (Figuras 3 y 4) aparecen en Apéndices.

uso general de la preposición *a*, a costa de *en*, tendencia atestada en la amplia geografía de la lengua[8].

2.2. Los modos

De las trece frases relativas a los modos (Véanse las frases N.1 a N.13 del cuestionario, en APÉNDICE [2]), comparamos la aceptabilidad de las siguientes tres del llamado uso de "emoción y evaluación", que requiere, en principio, que los verbos que aparecen en la cláusula sustantiva se expresen en modo subjuntivo. Los verbos utilizados en estas frases están en presente, pretérito perfecto compuesto, y pretérito perfecto simple, respectivamente. Nuestro interés, pues, se centra en la posibilidad de encontrar alguna correlación entre los tiempos y la frecuencia del uso de subjuntivo[9].

Indicativo / Subjuntivo

(2) Es interesante que usted *sigue con el mismo trabajo. (N.4)

(3) Me alegro (de) que Ud. *se ha mejorado. (N.7)

(4) Es una lástima que no *hizo [hubo] sol ayer. (N.2)

Los verbos de estas frases están en tres tiempos diferentes: presente, pretérito perfecto compuesto y pretérito perfecto simple, respectivamente. Las tablas 6, 8 y 10 reúnen las respuestas para las frases (2)-(4), respectivamente. Y en las tablas 7, 9 y 11, se muestran sus porcentajes:

FA	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.		2	1				1	2		2
2:Oigo.	7	7	6	7	4	6	5	2	8	8
3:No.	20	12	13	12	7	14	18	15	12	14

TABLA 6. Respuestas (España) *Es interesante que usted *sigue con el mismo trabajo.* (N.4)

%	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.		10	5				4	11		8
2:Oigo.	26	33	30	37	36	30	21	11	40	33
3:No.	74	57	65	63	64	70	75	79	60	58

TABLA 7. Porcentajes (España) *Es interesante que usted *sigue con el mismo trabajo.* (N.4)

[8] Por otra parte, somos conscientes de que otros métodos de recogida de datos puedan llevar a una distribución geográfica diferente. Para una primera aproximación hemos hecho una cala en un corpus georreferenciado de tuits para observar otros contextos en diferentes lugares geográficos. El corpus se ha obtenido a través del Streaming API de *Twitter*, y se han reunido los ejemplos en una base de datos (entre marzo de 2018 y febrero de 2019) con 10.715.724 tuits, algo más de 100 millones de palabras. En este corpus solo el promedio del 31% del total prefiere el uso de *a*, en lugar de *en*, contrariamente a lo que suponemos. Creemos que un estudio en más profundidad ampliará nuestra comprensión del fenómeno y nos aclarará algunas diferencias con los resultados obtenidos a través de las encuestas.

[9] La idea inspirada en García y Terrell (1977), al que volveremos más adelante.

FA	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.										1
2:Oigo.	5	8	9	7	5	9	9	3	7	9
3:No.	22	11	11	12	6	11	15	16	13	14

TABLA 8. Respuestas (España) *Me alegro (de) que *Ud. se ha mejorado*. (N.7)

%	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.										4
2:Oigo.	19	42	45	37	45	45	38	16	35	38
3:No.	81	58	55	63	55	55	63	84	65	58

TABLA 9. Porcentajes (España) *Me alegro (de) que *Ud. se ha mejorado*. (N.7)

FA	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	5	5		2	1	3	6	3	6	5
2:Oigo.	12	10	14	10	7	10	13	10	14	16
3:No.	10	6	6	7	3	7	5	5		3

TABLA 10. Respuestas (España) *Es una lástima que no *hizo [hubo] sol ayer*. (N.2)

%	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	19	24		11	9	15	25	17	30	21
2:Oigo.	44	48	70	53	64	50	54	56	70	67
3:No.	37	29	30	37	27	35	21	28		13

TABLA 11. Porcentajes (España) *Es una lástima que no *hizo [hubo] sol ayer*. (N.2)

La aceptabilidad media para cada oración que se observa en estas tablas parecen demostrar una marcada diferencia en las frecuencias del indicativo entre las oraciones (2) y (3), por una parte y en (4), por otra, en España, como se observa en la tabla 12.

	Porcentajes de aceptabilidad[10]	Tiempo
(2)	3,8 %	Presente
(3)	0,4 %	Pretérito perfecto compuesto
(4)	17,1 %	Pretérito perfecto simple

TABLA 12. Porcentajes de aceptabilidad del indicativo (España)

[10] Estas cifras son el promedio de los porcentajes de las respuestas DIGO de las tablas 7, 9,11.

Los verbos en indicativo son casi inadmisibles en presente (2) y pretérito perfecto compuesto (3) en España. En contraste, en pretérito perfecto simple sube notablemente la aceptabilidad (17,1%).

Ahora fijémonos en los resultados de Latinoamérica:

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	7	3	7	4	10	11	1	10	4	3	10	2	2	1	2
2:O	10	15	17	10	15	10	12	12	7	14	22	5	10	9	11
3:N	3	5	11	5	3	1	12	3	12	6	7	18	7	10	6

TABLA 13. Respuestas (Latinoamérica) *Es interesante que usted *sigue con el mismo trabajo*. (N.4)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	35	13	20	21	36	50	4	40	17	13	26	8	11	5	11
2:O	50	65	49	53	54	45	48	48	30	61	56	20	53	45	58
3:N	15	22	31	26	11	5	48	12	52	26	18	72	37	50	32

TABLA 14. Porcentajes (Latinoamérica) *Es interesante que usted *sigue con el mismo trabajo*. (N.4)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	5	3	11	5	11	3	1	5	5	6	11	1	4	2	2
2:O	6	5	17	7	16	15	15	16	11	12	23	10	12	10	16
3:N	8	15	5	5	3	4	9	4	6	4	6	13	2	7	2

TABLA 15. Respuestas (Latinoamérica) *Me alegro (de) que *Ud. se ha mejorado*. (N.7)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	26	13	33	29	37	14	4	20	23	27	28	4	22	11	10
2:O	32	22	52	41	53	68	60	64	50	55	57	42	67	53	80
3:N	42	65	15	29	10	18	36	16	27	18	15	54	11	37	10

TABLA 16. Porcentajes (Latinoamérica) *Me alegro (de) que *Ud. se ha mejorado*. (N.7)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	9	22	26	8	13	11	19	21	12	12	12	1	5	9	11
2:O	9	1	9	8	10	10	5	4	8	8	21	15	10	9	7
3:N	1			2	5	1	1		3	3	5	9	5	2	2

TABLA 17. Respuestas (Latinoamérica) *Es una lástima que no *hizo [hubo] sol ayer*. (N.2)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	47	96	74	44	46	50	76	84	52	52	32	4	25	45	55
2:O	47	4	26	44	36	45	20	16	35	35	55	60	50	45	35
3:N	5			11	18	5	4		13	13	13	36	25	10	10

TABLA 18. Porcentajes (Latinoamérica) *Es una lástima que no *hizo [hubo] sol ayer*. (N.2)

Al igual que en España, las respuestas recogidas en Latinoamérica reflejan una distribución similar, pero con porcentajes mucho más altos en todas las frases. En contraste con las frases (2) y (3), (4) marca un porcentaje considerablemente alto de aceptabilidad del indicativo (52,1%), como vemos en la tabla 19.

	Porcentajes de aceptabilidad[11]	Tiempo
(2)	20,7 %	Presente
(3)	20,1%	Pretérito perfecto compuesto
(4)	52,1%	Pretérito perfecto simple

TABLA 19. Porcentajes de aceptabilidad del indicativo (Latinoamérica)

Los usos de "emoción y evaluación" se caracterizan por implicar la veracidad del evento presupuesto en la cláusula subordinada. Al expresar "Me alegro de que te hayas mejorado", el hablante debe de haber confirmado que el interlocutor se ha mejorado en realidad. Para manifestar su sentimiento de lástima en la oración (4), el hablante debe de saber que de hecho ayer no hizo sol. Este carácter "factivo" inherente en estas subordinadas permite explicar por qué el subjuntivo en estos usos es propenso a sustituirse por el indicativo con mayor facilidad que en los otros usos del subjuntivo, según argumentan Terrell y Hooper (1974) y García y Terrell (1977). Por la comparación de las tablas 12 y 19, sabemos que el proceso de sustitución del subjuntivo por el indicativo está relativamente más avanzado en Latinoamérica que en España.

Según se sugiere en García y Terrell (1977) con sus datos recogidos en México y la zona fronteriza México-EE.UU., hay desequilibrios en las correspondencias de tiempos y modos del español, como vemos en la tabla 20. Se encuentra vacía la casilla del pretérito perfecto simple en subjuntivo. Se podría conjeturar que el vacío por el lado del subjuntivo es capaz de dar origen a la atracción de la forma correspondiente en indicativo (*hizo*).

	INDICATIVO	SUBJUNTIVO
Presente	hace	haga
Pretérito perfecto compuesto	ha hecho	haya hecho
Pretérito imperfecto	hacía	hiciera
Pretérito perfecto simple	hizo	Ø

TABLA 20. Las correlaciones de los tiempos y los modos

[11] Estos porcentajes son el promedio de las respuestas en DIGO de las tablas 14,16,18.

Se podría conjeturar que el índice relativamente alto del pretérito perfecto simple de nuestros resultados (17,1% en España; 52,1% en Latinoamérica) podría atribuirse a este desajuste estructural de las formas verbales entre los modos indicativo y subjuntivo. Obviamente, sin embargo, faltarían muchos más datos para justificar este supuesto. Podemos confirmar este desequilibrio entre presente y pretérito perfecto compuesto, por un lado y pretérito perfecto simple, por otro, en la distribución geográfica de las frecuencias relativas de DIGO (del uso de indicativo), ilustradas en los mapas lingüísticos (figuras 5: España y figura 6: Latinoamérica).

En España (figura 5), efectivamente existe cierta tendencia a admitir el uso del indicativo en la oración (4) especialmente en Pamplona (24%), Sevilla (25%), Tenerife (30%) y Las Palmas (21%). En América Latina, por su parte, se comprueba visualmente un mayor uso del indicativo en todos los lugares encuestados, como se observa en la figura 6.

Curiosamente, la distribución geográfica del indicativo en España (figura 5) parece presentar una gradación continua, es decir, las ciudades cercanas, por ejemplo, Oviedo y Pamplona, Alcalá y Madrid, Sevilla y Huelva, Tenerife y Las Palmas, muestran proporciones más parecidas que las ciudades distantes. En América (figura 6) también la tendencia de adyacencia se reconoce en casi todos los grupos: México, Centroamérica y el Caribe (La Habana, Santo Domingo, San Juan); Caracas (Venezuela), Bogotá (Colombia) y Quito (Ecuador); Lima (Perú) y La Paz (Bolivia); Buenos Aires (Argentina) y Montevideo (Uruguay). Por otra parte, Santiago de Chile es peculiar en no admitir el uso del indicativo en ninguna de las tres oraciones, lo que, sin embargo, no constituye un contraejemplo de la tendencia de distribución adyacente, puesto que geográficamente Chile se encuentra separado de los países colindantes por el desierto de Atacama y la cordillera de los Andes.

2.3. Leísmo

El *leísmo*, la sustitución de *le* por el pronombre átono para el complemento directo *lo* (o *la*), se considera un fenómeno generalizado en el castellano septentrional de España, mientras que en Latinoamérica se documenta en mucha menor escala. En nuestro cuestionario se registran tres ejemplos de *leísmo* (N. 47, N. 44, N.45). Es interesante comprobar cómo en los resultados de la encuesta se refleja esta distribución de la propagación *leísta*. En este apartado, sin embargo, veremos solo el caso de *leísmo* en singular representado en (5).

Leísmo

(5) Un coche *le atropelló y *le mató. [le = mi hijo] (N. 47)

FA.	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	12	14	8	12	9	16	9	12	11	8
2:Oigo.	12	5	11	7	2	4	15	6	4	13
3:No.		1	1					1	5	2

TABLA 21. Leísmo (España) *Un coche *le atropelló y *le mató*. [le = mi hijo] (N. 47)

%	OVI	PAM	BAR	SAL	ALC	MAD	SEV	HUE	TEN	PAL
1:Digo.	50	70	40	63	82	80	38	63	55	35

2:Oigo.	50	25	55	37	18	20	63	32	20	57
3:No.		5	5					5	25	9

TABLA 22. Leísmo (España) *Un coche *le atropelló y *le mató*. [le = mi hijo] (N. 47)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	1		4	1	6		1		7	1	10	2	12	1	
2:O	15	2	13	4	11	5	8	6	12	15	24	3	7	1	13
3:N	3	21	17	14	13	17	16	19	4	6	5	20		18	7

TABLA 23. Leísmo (Latinoamérica) *Un coche *le atropelló y *le mató*. [le = mi hijo] (N. 47)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	5		12	5	20		4		30	5	26	8	63	5	
2:O	79	9	38	21	37	23	32	24	52	68	62	12	37	5	65
3:N	16	91	50	74	43	77	64	76	17	27	13	80		90	35

TABLA 24. Leísmo (Latinoamérica) *Un coche *le atropelló y *le mató*. [le = mi hijo] (N. 47)

Como era de esperar, el promedio de aceptación sobrepasa el 50% en España (tabla 25). Alcalá de Henares y Madrid son las dos ciudades más representativas del *leísmo* que marcan índices superiores al 80%. Entendemos que solo en Barcelona, Sevilla y Las Palmas se prefiere la forma *lo*.

	Porcentajes de aceptabilidad
España	57,6%
Latinoamérica	12,2%

TABLA 25. Porcentajes de aceptabilidad del leísmo (España y Latinoamérica)

Según el mapa lingüístico (figura 7, en Apéndices) vemos que el *leísmo* se atestigua principalmente en Pamplona (70%), Alcalá de Henares (82%) y Madrid (80%), lugares que se podrían asociar con las dos Castillas, la Vieja y la Nueva. Por otra parte, el *leísmo* se observa en mucha menor escala en Oviedo (50%), Barcelona (40%), Sevilla (38%) y Las Palmas (35%). Con estos datos es posible suponer un contraste entre la zona central propensa al leísmo y la zona periférica reacia al mismo proceso.

En cambio, en Latinoamérica el porcentaje de aceptación se reduce a una media de 12,2%. Nos sorprenden, sin embargo, los índices de aceptabilidad inesperadamente altos que demuestran Asunción (63%), Quito (30%), La Paz (26%) y Santo Domingo (20%). Véase el mapa lingüístico en la figura 8 (en Apéndices) con especial atención a Asunción.

De acuerdo con Fernández-Ordóñez (1999: 1341), se puede pensar en "la posibilidad de la interferencia de lenguas no indoeuropeas como el quechua, el aimara, el guaraní, que están en contacto con el español". Efectivamente en Asunción, por ejemplo, donde se emplea el guaraní diariamente, es común el uso de l*e* "como único clítico de tercera persona para los objetos animados"[12]. También Haboud *et al.* (2008:170) testimonian la convivencia del sistema pronominal etimológico y el

[12] Véase también Palacios (2000:137-9).

simplificado *leísta*, basado en el uso de *le* para objeto directo e indirecto sin diferenciación de género[13].

Nuestros datos pueden corroborar dicha hipótesis puesto que los lugares donde se registran altos índices de *leísmo*, parecen coincidir con las regiones en las que el español se encuentra en contacto con los idiomas autóctonos.

2.4. Queísmo-dequeísmo

En este apartado comparamos dos fenómenos gramaticales, la ausencia no normativa de la preposición 'de', denominada 'queísmo' y la presencia superflua de la misma ('dequeísmo') delante de la cláusula completiva de 'que'. Nos parece cuestionable el tratamiento simétrico que se ha venido realizando en los estudios anteriores (Seco 1961: 114, RAE 1973: 522, Rabanales 1974: 25, Arjona 1978: 75, Schwenter 1999: 74, entre otros). Por otra parte, los estudios recientes han contribuido a precisar las características no paralelas, es decir asimétricas, de la ausencia - presencia de la preposición en sus aspectos gramaticales (Gómez Torrego 1999), geolingüísticos entre España y América (RAE y AALE 2009: 3248-57) e históricos (Ueda 2017). Echamos de menos un estudio sobre el mismo tema basado en las encuestas *in situ*.

Nuestro proyecto VARIGRAMA, para explorar las variaciones lingüísticas diatópicas, ha tratado las cinco oraciones (6) a (10) de *queísmo* y *dequeísmo*, en Latinoamérica a partir de la encuesta del 2008[14].

Queísmo-dequeísmo

(6)	Estoy seguro *que vendrá. (N. 63)	[queísmo]
(7)	Mi hermana está contenta *que hayas aceptado la invitación. (N. 64)	[queísmo]
(8)	Sospecho *de que me mintió. (N. 65)	[dequeísmo]
(9)	Ella dijo *de que no sabía nada. (N. 66)	[dequeísmo]
(10)	Supongo *de que es verdad. (N. 67)	[dequeísmo]

Nuestro interés se centra en las correlaciones entre los dos fenómenos aparentemente opuestos de *queísmo* y *dequeísmo*, *i.e.*, la elisión de la preposición *de* con función de conector entre predicado y cláusula subordinada, y, por el contrario, la inserción de dicho elemento innecesario en la misma posición. Las respuestas afirmativas en las ciudades latinoamericanas muestran las siguientes proporciones de aceptabilidad.

Los resultados y porcentajes de las primeras oraciones de *queísmo* (N. 63, N. 64) se muestran en las tablas 26, 27 (*Estoy seguro que…*) y 28, 29 (*Mi hermana está contenta que…*), respectivamente:

[13] Queda por averiguar la situación del contacto de lenguas en los otros puntos de alta aceptación de *le*, como El Alto, La Paz, y Santo Domingo.

[14] En la encuesta de España no contábamos con las preguntas N. 66 y N. 67, que añadimos a las de Latinoamérica a partir del año 2008. En este trabajo se analizan solo los resultados de Latinoamérica, ya que nos parece más significativo comparar el mayor número de preguntas posibles sobre el mismo tema aunque se limiten solo a un territorio.

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	16	9	34	14	24	20	17	22	15	19	32	18	15	15	14
2:O	1	13	2	3	6	2	6	3	7	4	7	5	4	4	5
3:N	3	1	2	1			2		1		1	2	1	1	1

TABLA 26. Queísmo (Latinoamérica) *Estoy *seguro que vendrá* (N. 63)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	80	39	89	78	80	91	68	88	65	83	80	72	75	75	70
2:O	5	57	5	17	20	9	24	12	30	17	18	20	20	20	25
3:N	15	4	5	6			8		4		3	8	5	5	5

TABLA 27. Queísmo (Latinoamérica) *Estoy *seguro que vendrá* (N. 63)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	11	5	28	8	18	15	10	16	12	15	22	11	12	11	12
2:O	7	14	5	9	8	5	10	9	7	7	16	11	7	6	5
3:N	3	4	4	1	3	2	5		4		1	3	1	3	3

TABLA 28. Queísmo (Latinoamérica) *Mi hermana está *contenta que hayas aceptado la invitación.* (N. 64)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	52	22	76	44	62	68	40	64	52	68	56	44	60	55	60
2:O	33	61	14	50	28	23	40	36	30	32	41	44	35	30	25
3:N	14	17	11	6	10	9	20		17		3	12	5	15	15

TABLA 29. Queísmo (Latinoamérica) *Mi hermana está *contenta que hayas aceptado la invitación.* (N. 64)

Los datos de *dequeísmo* se señalan en las tablas 30, 31 (*Sospecho de que…*); las 32, 33 (*Ella dijo de que…*), y las 34, 35 (*Supongo de que…*):

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	4	3	13	1	14	6	3	2	4	7	17	5	9	1	2
2:O	9	14	18	14	14	13	16	21	16	14	21	16	11	14	16
3:N	6	6	5	3	2	3	6	2	3	1	1	4		5	2

TABLA 30. Dequeísmo (Latinoamérica) *Sospecho *de que me mintió.* (N. 65)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	21	13	36	6	47	27	12	8	17	32	44	20	45	5	10
2:O	47	61	50	78	47	59	64	84	70	64	54	64	55	70	80
3:N	32	26	14	17	7	14	24	8	13	5	3	16		25	10

TABLA 31. Dequeísmo (Latinoamérica) *Sospecho *de que me mintió*. (N. 65)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	2	2	15	2	6	5	2	10	8	4	16	3	5		3
2:O	4	12	13	10	15	8	15	11	10	19	23	16	15	11	13
3:N	13	9	7	6	9	9	8	4	5		1	6		9	4

TABLA 32. Dequeísmo (Latinoamérica) *Ella dijo *de que no sabía nada*. (N. 66)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	11	9	43	11	20	23	8	40	35	17	40	12	25		15
2:O	21	52	37	56	50	36	60	44	43	83	57	64	75	55	65
3:N	68	39	20	33	30	41	32	16	22		3	24		45	20

TABLA 33. Dequeísmo (Latinoamérica) *Ella dijo *de que no sabía nada*. (N. 66)

FA	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	1	1	10		4	1	2	2	2	5	10	1	4		3
2:O.	5	14	18	10	17	15	14	19	14	15	26	19	16	8	12
3:N	13	8	6	8	10	6	9	4	7	3	4	5		12	5

TABLA 34. Dequeísmo (Latinoamérica) *Supongo *de que es verdad*. (N.67)

%	ME	JO	PA	HA	SD	JU	CA	BO	QU	LI	LP	ST	AS	MT	BU
1:D	5	4	29		13	5	8	8	9	22	25	4	20		15
2:O	26	61	53	56	55	68	56	76	61	65	65	76	80	40	60
3:N	68	35	18	44	32	27	36	16	30	13	10	20		60	25

TABLA 35. Dequeísmo (Latinoamérica) *Supongo *de que es verdad*. (N.67)

En la tabla 36 se muestra el promedio de los porcentajes de aceptabilidad para cada una de las cinco frases preguntadas. Se entiende que las primeras dos de *queísmo* (*Estoy seguro que…, Mi hermana está contenta que…*) cuentan con una mayor aceptación relativa. En cambio, los tres casos de *dequeísmo* (*Sospecho de que…, Ella dijo de que…, Supongo de que…*) no parecen tan aceptados como los de *queísmo*.

Frases encuestadas		Porcentajes de aceptabilidad
Queísmo	(N. 63) Estoy seguro que vendrá.	75,5%
	(N. 64) Está contenta que hayas aceptado….	54,9%
Dequeísmo	(N. 65) Sospecho de que me mintió.	22,9%
	(N. 66) Dijo de que no sabía nada.	20,6%
	(N. 67) Supongo de que es verdad.	11,1%

TABLA 36. Promedio de porcentajes de aceptabilidad (Latinoamérica)

Los índices de aceptabilidad, representados gráficamente (figura 2), ilustran, aparte de las diferencias entre las ciudades, una fuerte correlación entre el grupo de las dos oraciones de *queísmo* (*Estoy seguro de que*…y *Está contenta que*…), por un lado, y, por otro, el de las tres de *dequeísmo* (*Sospecho de que*…, *Dijo de que*… y *Supongo de que*…), respectivamente.

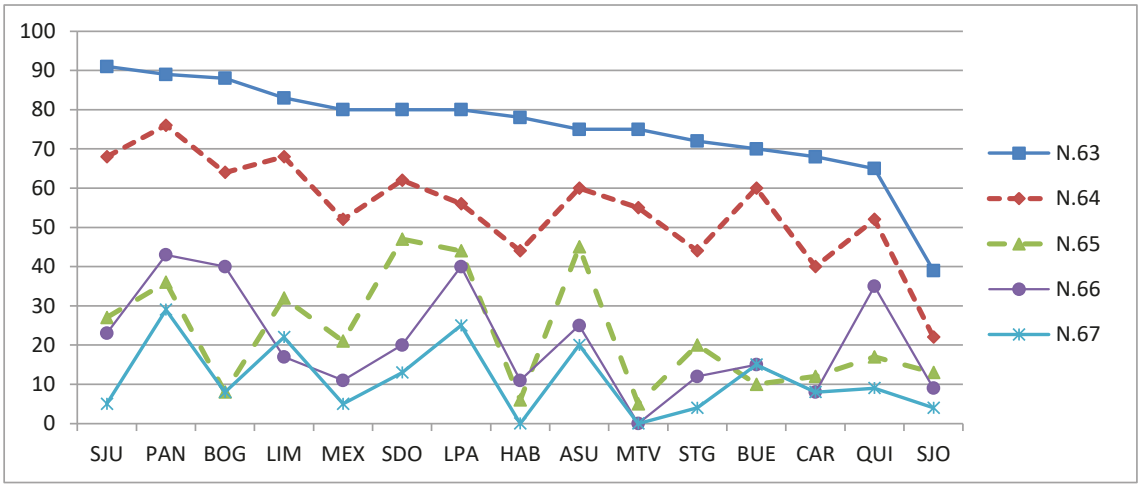


FIGURA 2. Índices de aceptabilidad de *queísmo* y *dequeísmo* (Latinoamérica)

De hecho, la siguiente tabla de correlaciones (tabla 37) entre estas cinco frases nos permite confirmar esta distribución desnivelada.

%	N.63	N.64	N.65	N.66	N.67
N.63	1.000	.859	.329	.425	.317
N.64	.859	1.000	.461	.554	.584
N.65	.329	.461	1.000	.466	.727
N.66	.425	.554	.466	1.000	.652
N.67	.317	.584	.727	.652	1.000

TABLA 37. Correlaciones entre las cinco frases

Esta distribución de *queísmo* y *dequeísmo* también se comprueba visualmente en el mapa lingüístico (figura 9, en Apéndices).

En síntesis, debemos pensar, aunque sea solo por estos pocos ejemplos, que *queísmo* y *dequeísmo* son dos procesos considerablemente independientes entre sí, lo cual nos hace cuestionar si es adecuado el tratamiento simétrico de la ausencia y presencia de la preposición *de* delante de la cláusula precedida de *que*[15].

3. CONCLUSIONES

En los apartados anteriores hemos presentado brevemente nuestro proyecto de variación gramatical del español en el mundo, denominado VARIGRAMA, junto con los ejemplos de análisis de los cuatro temas de gramática española de gran envergadura. En el cuestionario se incluyen otros temas gramaticales como los de relativos, comparación, pasiva refleja, pronombres átonos, gerundio, concordancia, conjunciones, preposiciones, etc. como se observa en APÉNDICE [2].

El proyecto VARIGRAMA, aunque lleva más de una docena de años, es un estudio preliminar que deberá irse ampliando sucesivamente y los resultados actuales son orientativos para poco más de cien oraciones concretas investigadas en contextos muy restringidos. Por ejemplo, disponemos de solo una oración *Ellos entraron al edificio* (N.14), mencionada arriba para conocer la realidad lingüística de la selección entre las preposiciones *a* y *en*.

El mayor mérito de nuestro proyecto, sin embargo, consiste en el hecho de que las mismas frases hayan sido sometidas, de modo uniforme, a juicio de grupos del mismo número de encuestados de la misma franja de edad y nivel cultural en distintas ciudades del extenso territorio del habla española.

En definitiva, a pesar de que nuestro acercamiento a las realidades lingüísticas tratadas en este trabajo no deja de ser hipotético, esperamos que los datos obtenidos mediante una metodología sistemática puedan proporcionar una nueva perspectiva a la investigación lingüística del español actual.

REFERENCIAS BIBLIOGRÁFICAS

ARJONA, M. (1978): Anomalías en el uso de la preposición de en el español de México, in *Anuario de Letras*, UNAM 16. pp. 67-90. https://revistas-filologicas.unam.mx/anuario-letras/index.php/al/ [Acceso: 24-11-2017]

BOSQUE, I. & V. DEMONTE (dirs.) (1999): *Gramática descriptiva de la lengua española*. Madrid: Espasa.

BUTT, J. & C. BENJAMIN (2000): *A New Reference Grammar of Modern Spanish*. 3rd edition, Chicago: McGraw-Hill.

COROMINAS, J. & J. A. PASCUAL (1980): *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos.

DE BRUYNE, J. (1999): "Las preposiciones", in I. Bosque & V. Demonte (dirs.): *Gramática descriptiva de la lengua española*. Madrid: Espasa. pp. 657-703.

[15] Véase Ueda (2017).

FERNÁNDEZ-ORDÓÑEZ, I. (1999): "Leísmo, laísmo y loísmo", in I. Bosque & V. Demonte (dirs.): *Gramática descriptiva de la lengua española*. Madrid: Espasa, pp. 1219-1397.

GARCÍA, M. E. & T. TERRELL (1977): "Is the use of mood in Spanish subject to variable constraint?", in M. P. Hagiwara (ed.): *Studies in Romance Linguistics*. University of Michigan, pp. 214-226.

GÓMEZ TORREGO, L. (1999): "La variación en las subordinadas sustantivas. Dequeísmo y queísmo", in I. Bosque & V. Demonte (dirs.): *Gramática descriptiva de la lengua española*. Madrid: Espasa. pp. 2105-2148.

HABOUD, M. & E. DE LA VEGA. (2008): "Ecuador", in A. Palacios (coord.): *El español en América*. Barcelona: Ariel, pp.161-187.

LEWIS, C. T. (1975): *An Elementary Latin Dictionary*. Oxford: Oxford University Press.

LOPE BLANCH, J. M. (1964): "Estado actual del español en México", in *Presente y futuro de la lengua española 1*. Madrid: Ediciones Cultura Hispánica, pp. 79-91.

PALACIOS, A. (2008): "Paraguay", in A. Palacios (coord.): *El español en América*. Barcelona: Ariel, pp. 279-300.

PALACIOS, A. (2000): "El sistema pronominal del español paraguayo: Un caso de contacto de lenguas", in J. Calvo Pérez (ed.): *Teoría y práctica del contacto: el español de América en el candelero*. Frankfurt-Madrid: Vervuert-Iberoamericana. pp. 122-143.

PALACIOS, A. (ed.) (2008): *El español de América. Contactos lingüísticos en Hispanoamérica*. Barcelona: Ariel.

RABANALES, A. (1974): "Queísmo y dequeísmo en el español de Chile", in L. Quiroga Torrealba, M. Torrealba Lossi & P. Díaz Seijas (dirs.) *Estudios filológicos y lingüísticos. Homenaje a Ángel Rosenblat en sus setenta años. Onomázein*. Caracas: Instituto Pedagógico Pontificia Universidad Católica de Chile, pp. 413-444. http://www.redalyc.org/pdf/1345/134516558002.pdf [acceso: 22 de nov. 2107]

RAE (1973): *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa.

RAE (2005): *Diccionario panhispánico de dudas*. Madrid: Santillana.

RAE - ASALE (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.

SCHWENTER, S. (1999): "Evidentiality in Spanish morphosyntax: a reanalysis of (de)queísmo", in Mª J. Serrano (ed.): *Estudios de variación sintáctica*. Frankfurt-Madrid: Iberoamericana Vervuert, pp. 65-87.

SECO, M. (1961): *Diccionario de dudas y dificultades de la lengua española*. Madrid Aguilar.

TAKAGAKI, T. (2014): "Variación gramatical del español: Algunos resultados del Proyecto Varigrama", *Actas del Congreso Internacional sobre el español y la cultura hispánica*. Tokio: Instituto Cervantes. pp. 248-254. http://cvc.cervantes.es/ensenanza/biblioteca_ele/publicaciones_centros/PDF/tokio_2013/28_takagaki.pdf

TERELL, T. & J. HOOPER (1974): "A semantically based analysis of mood in Spanish", *Hispania* 57, pp. 484-495.

UEDA, H. (2017): "Asimetría unidireccional de queísmo y dequeísmo—Aproximación a la realidad histórica y actual de la variación a través de estudios anteriores, corpus y encuestas", *Actas del Congreso Spanish Dialect Syntax,* Universidad Complutense de Madrid (21 de abril de 2017). https://lecture.ecc.u-tokyo.ac.jp/~cueda/kenkyu/rekisi/de-queismo/de-queismo.pdf

APÉNDICES

[1] MAPAS LINGÜÍSTICOS

1) *Entrar a/en* (Apartado 2.1)

*Ellos entraron *al edificio*.* (N. 14)

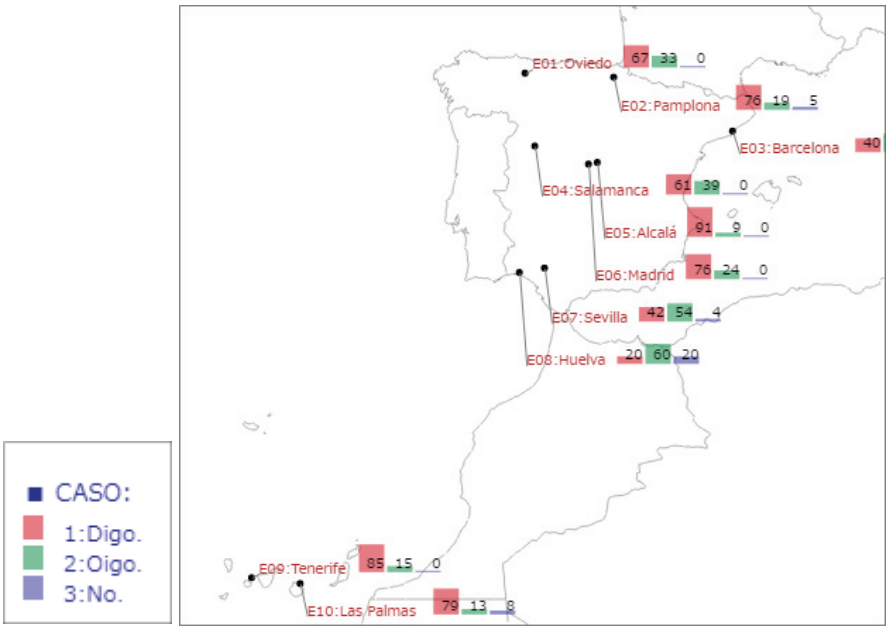


FIGURA 3. *entrar a/en* (España)

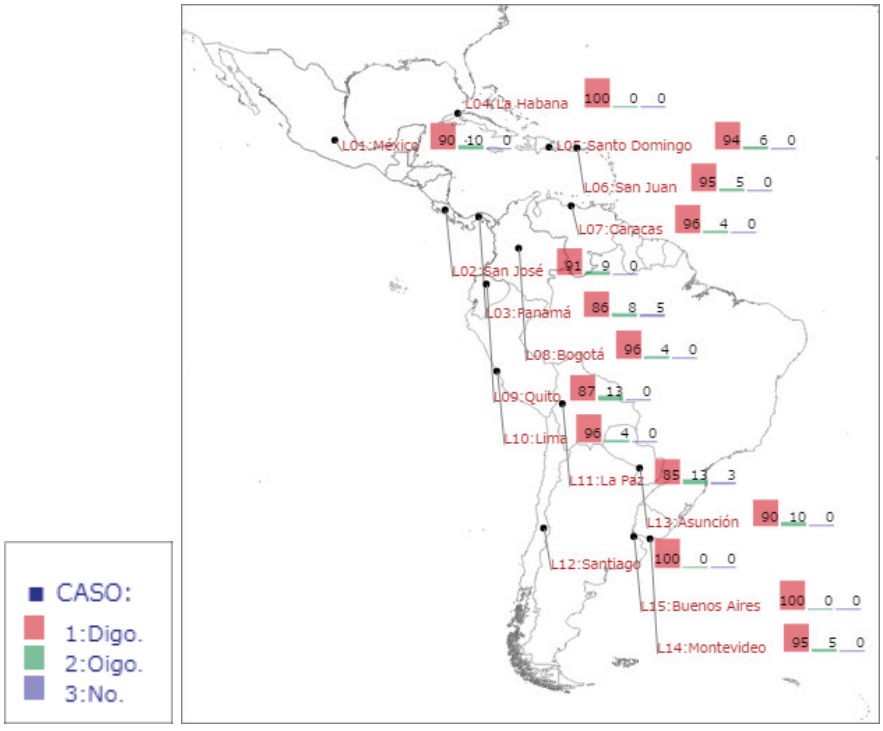


FIGURA 4. *entrar a/en* (Latinoamérica)

2) Indicativo/Subjuntivo (Apartado 2.2)

*Es interesante que usted *sigue con el mismo trabajo*. (N.4)

*Me alegro (de) que Ud. *se ha mejorado*. (N.7)

*Es una lástima que no *hizo [hubo] sol ayer*. (N.2)

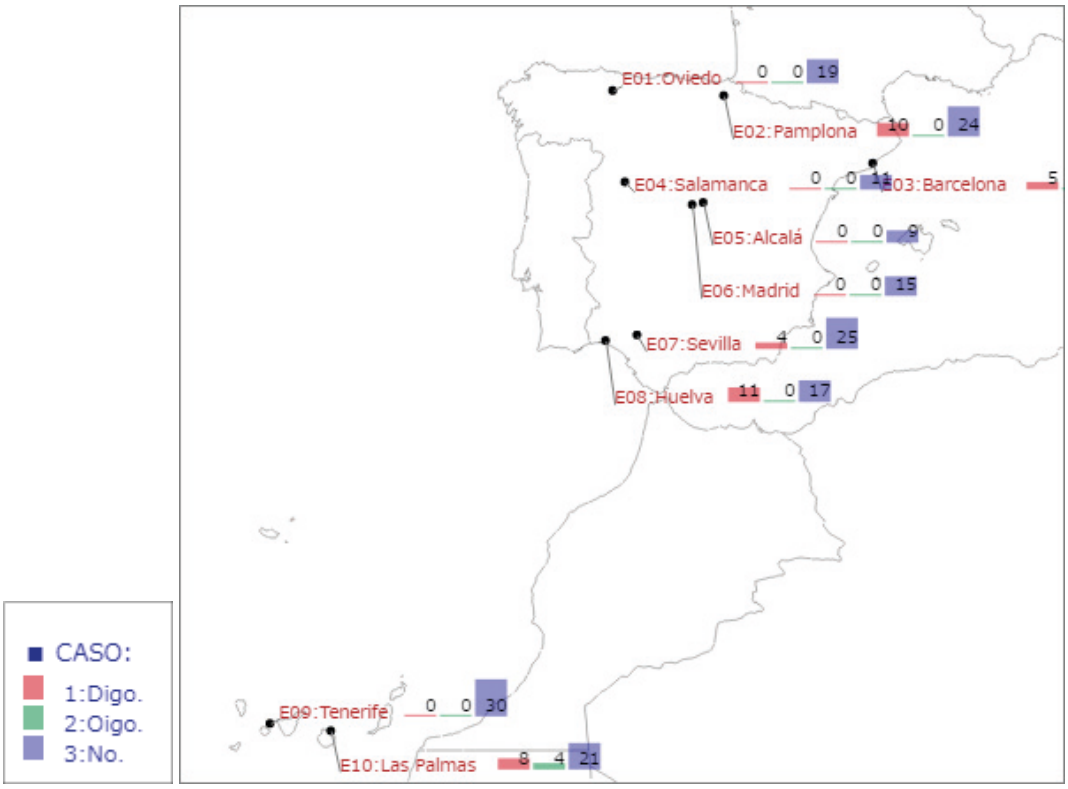


FIGURA 5. Indicativo/Subjuntivo (España)

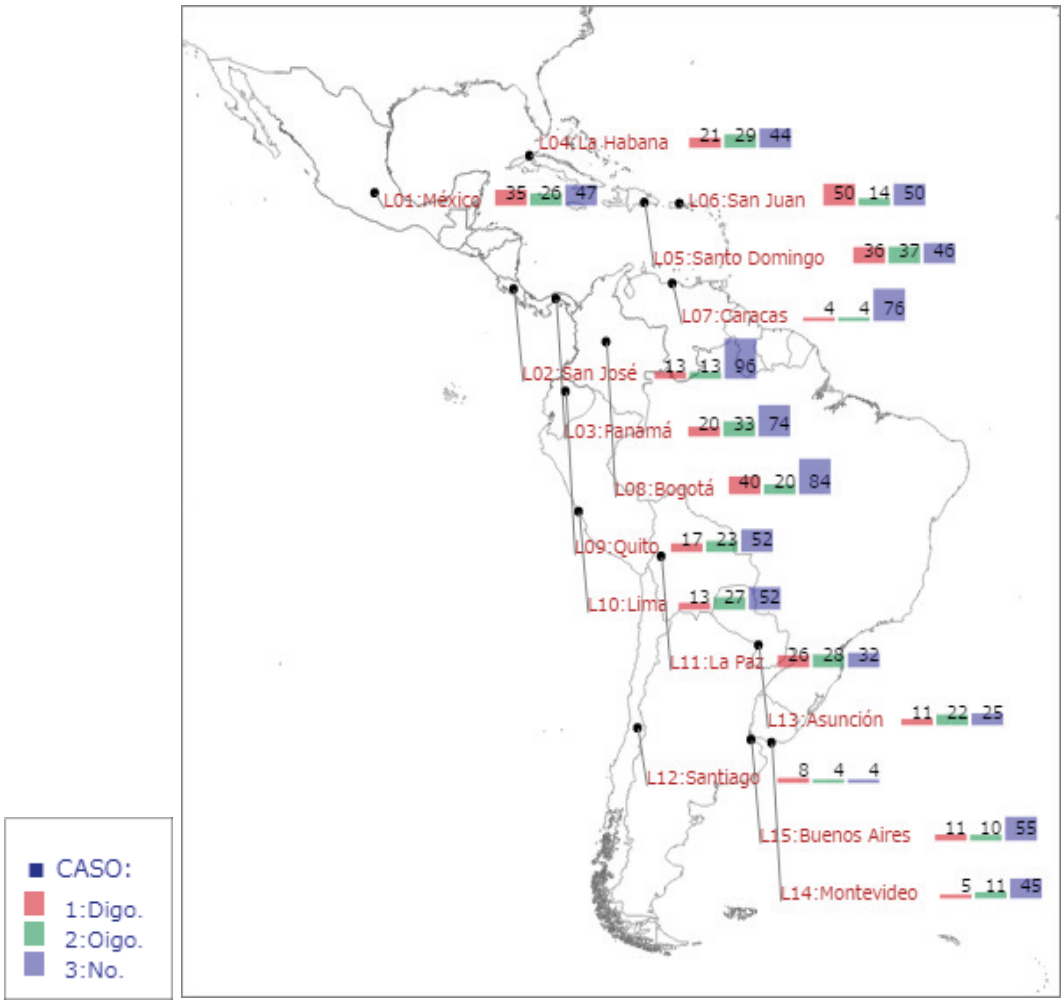


FIGURA 6. Indicativo/Subjuntivo (Latinoamérica)

3) Leísmo (Apartado 2.3)

*Un coche *le atropelló y *le mató.* [le = mi hijo] (N. 47)

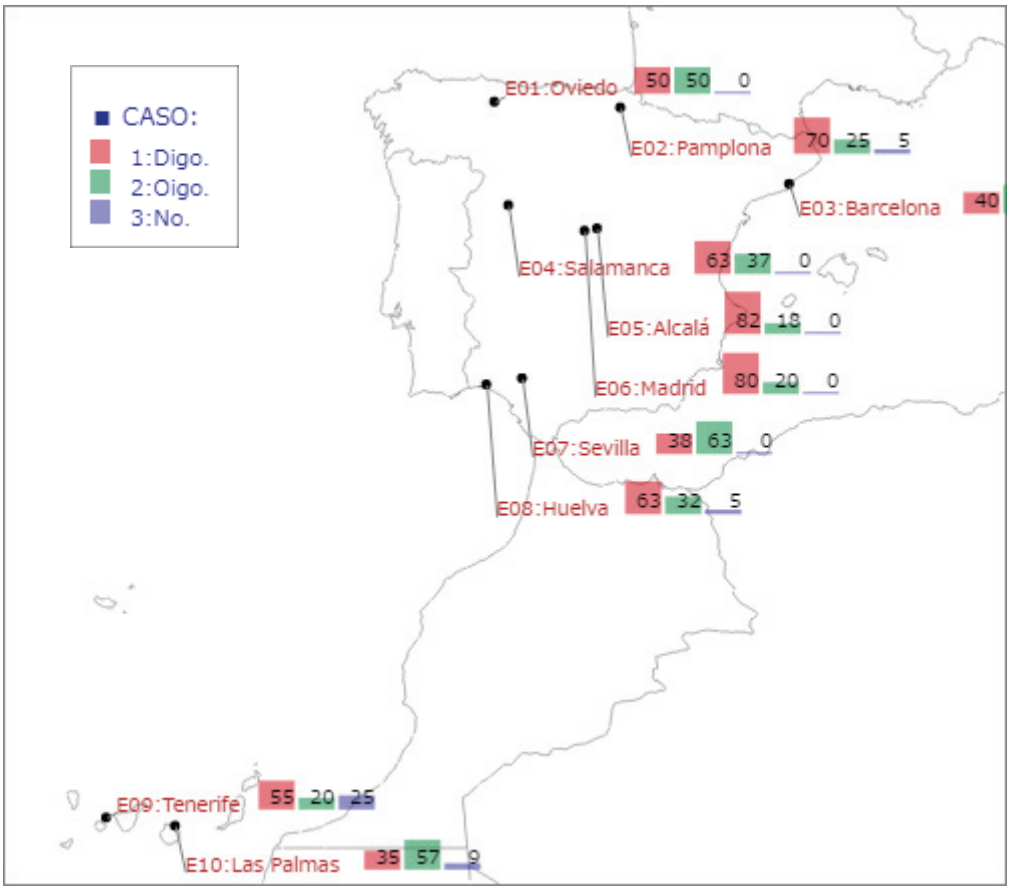


FIGURA 7. Leísmo (España)

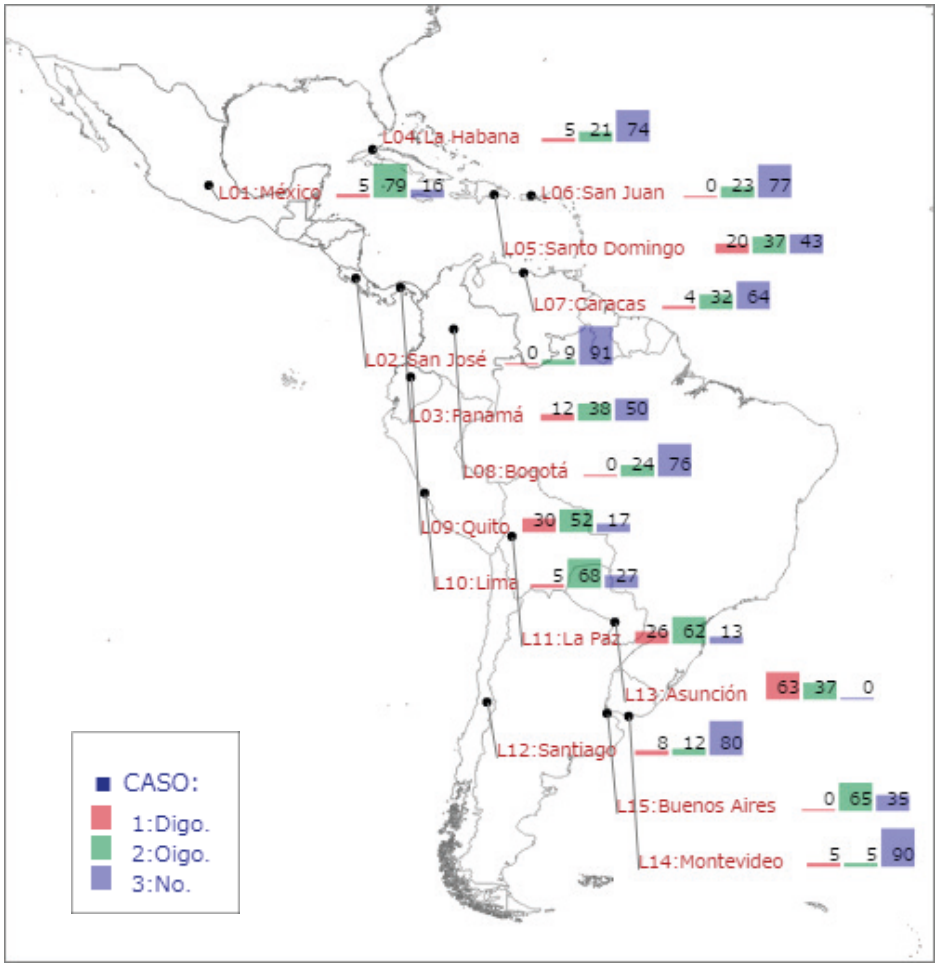


FIGURA 8. Leísmo (Latinoamérica)

4) Queísmo/Dequeísmo (Apartado 2.4)

*Estoy seguro *que vendrá*. (N. 63) [queísmo]
*Mi hermana está contenta *que hayas aceptado la invitación*. (N. 64) [queísmo]
*Sospecho *de que me mintió*. (N. 65) [dequeísmo]
*Ella dijo *de que no sabía nada*. (N. 66) [dequeísmo]
*Supongo *de que es verdad*. (N. 67) [dequeísmo]

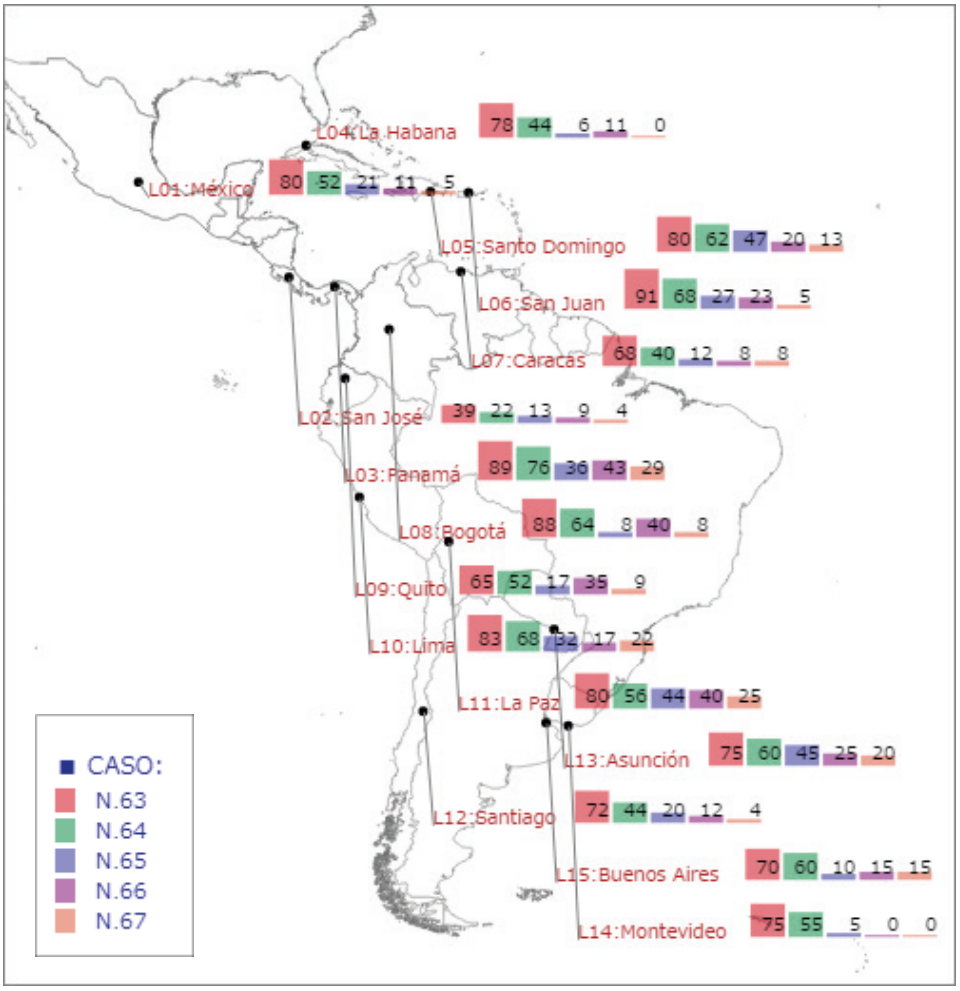


FIGURA 9. Queísmo / Dequeísmo (Latinoamérica)

[2] FRASES DEL CUESTIONARIO (VERSIÓN DE LA ENCUESTA EN SAN SALVADOR, 2018)

Esta es la versión más reciente utilizada en San Salvador, El Salvador, en 2018. Las frases están agrupadas bajo los temas gramaticales. Para ajustarse a la realidad lingüística y cultural del lugar donde se realiza la encuesta se sustituyen ciertos términos, como se ejemplifican particularmente en el cuestionario salvadoreño los nombres propios: *salvadoreño* (Frases 12, 13, 83), *la Avenida Independencia* (47), *San Salvador* (51), *Santa Ana* (51), *Álvaro Torres* (83) y objetos: *pupusas* (12,13), *carro* (47), *micro* (52), *diario* (40, 55,56).

Modos

1. Es una pena que (él) no *puede* venir.
2. Es una lástima que no *hizo* sol ayer.
3. Es una lástima que no *le ha gustado*.
4. Es interesante que usted *sigue* con el mismo trabajo.
5. ¡Qué bueno que *tenemos* tiempo para visitar a Juan también!
6. ¡Qué bueno que *tuvimos* tiempo para visitar a Juan también!
7. Me alegro (de) que *Ud. se ha mejorado*.
8. Me alegro (de) que *Ud. está* bien de salud.
9. El hecho de que el tabaco o el alcohol *están legalizados,* no da derecho a promover el consumo y legalización de otras drogas.
10. Quisiera enfatizar el hecho de que en Finlandia la libertad de prensa *está garantizada* en la legislación.
11. No hay duda (de) que lo *ha logrado*.
12. Aunque *soy salvadoreño*, no me gustan las pupusas.
13. Aunque *sea salvadoreño*, no me gustan las pupusas.

Preposiciones

14. Ellos *entraron a*l edificio.
15. *Metió* todos los libros *a* la mochila.
16. Voy *a por* los vasos.
17. Voy *por* los vasos.

Relativos

18. Ésta es la muchacha *que su* padre era profesor de música.
19. Hay muchos estudiantes *que* no *les* gusta leer libros.
20. Hay muchos estudiantes *a los que* no *les* gusta leer libros.
21. Los gramáticos aconsejan muchas cosas *que* nadie *las* dice.

22. No se acuerda del hotel que estuvimos el año pasado.
23. Éste es el martillo que yo hice el mueble con él.
24. Es una palabra que cualquier hablante puede identificar su significado.
25. A menudo paseo por el bosque, en cuyo lugar encuentro paz y serenidad.

Expresiones comparativas

26. La calidad del agua en las ciudades es inferior que en el campo.
27. Su hijo es menor al mío.
28. Yo me he tropezado con problemas más peores que ése.
29. Los intereses públicos son muchos más importantes que los privados, señaló el alcalde.
30. La literatura es muy bonita: contra más lee uno, más le gusta.
31. La mejor que lo dibujó fue Isabel.

Construcciones con *se*

32. Se *debe pedir* informes.
33. Se *puede aceptar* los defectos de los demás cuando se tiene clara conciencia de los propios.
34. Se *acepta* cheques de viajero [o de viaje].
35. *Léase* las páginas siete y ocho. [indicación en un manual]
36. Uno de los factores que se *ha de tener* en cuenta en la oxidación de metales es la humedad relativa.
37. A estas cosas se les *llaman* con diferentes palabras.
38. ¡Quédate*la*! (la = la foto)
39. ¡Quédate *con ella*! (ella = la foto)
40. En esta tienda [o acá] no se *vende* diarios.
41. Ahí no se *incluyó* sus firmas.

Clíticos

42. No la haga esperar más.
43. No la haga repetir la misma tontería.
44. Tenía muchas ganas de verles. (les = el padre y la madre)
45. El tractor hace tiempo que le vendimos para comprar uno nuevo.
46. Le digo la verdad a mis padres.
47. Un carro que venía por la Avenida Independencia le atropelló y le mató. (le = mi hijo)
48. Yo la he dicho que se habían quedado en la escuela. (la = mi madre)

49. ¿Les dijiste que no venía? —Ya se *los* dije.

Usos de gerundio

50. Hotel de tres estrellas necesita recepcionista *sabiendo* inglés y francés.
51. Llegaron a San Salvador el día quince, *dirigiéndose* a Santa Ana al día siguiente.
52. Un micro chocó con el camión *llevando* vacas.

Concordancia de *haber*

53. *Pueden* haber problemas.
54. *Hubieron* días que no salíamos a trabajar.

Topicalización

55. *El diario* compró Pedro.
56. *El diario lo* compró Pedro.
57. *Agua*, conseguimos.
58. Agua, *la* conseguimos.

Queísmo y dequeísmo

59. Me da igual *con tal que* no esté roto.
60. Lo arregló *antes que* se dieran cuenta.
61. El partido comenzó poco *después que* empezara a llover.
62. Le he dicho [o le dije] mil veces que *debe de estar* aquí a las ocho.
63. Estoy *seguro que* vendrá.
64. Mi hermana está *contenta que* hayas aceptado la invitación.
64-1. Mi hermana está *contenta que* aceptaras la invitación.
65. Sospecho *de que* me mintió.
66. Ella *dijo de* que no sabía nada.
67. Supongo *de que* es verdad.
68. Le gusta *de* comer en casa.

Ser* y *estar

69. Alfredo *es* feliz de vacaciones en Londres.
70. La situación *era* mucho más grave de lo que nos imaginábamos.
71. La vida *es* llena de sorpresas.
72. No *estoy* responsable de lo que hagan.
73. *Estamos* capaces de ganar a cualquiera.

74. Todo *está* inútil.

Adverbialización de adjetivos

75. María duerme *tranquila*.
76. Antiguamente lo hacíamos *manual*.
77. Ellos cantan muy *suave*.
78. Es que piensa *distinto*.
79. Eso lo hago yo *fácil*.

Adverbio + posesivo

80. *Cerca mío* había una mujer con una bolsa en la mano.
81. Siempre que me dice algo, habla *en contra tuya*.
81-1. Siempre que me dice algo, habla *en contra tuyo*.
82. *Detrás mío* venía un carro siguiéndome.

Concordancia de número

83. La mayoría de los salvadoreños *conocen* a Álvaro Torres.
84. Esto *son* mis opiniones.
85. *Estas* son mis opiniones.
86. Por favor, tengan *presente* mis palabras.
87. ¿Me da un vaso de agua? — Sí, un momento. Ahora se *lo* doy.
88. ¿Me da un vaso de agua? — Sí, un momento. Ahora se *la* doy.

Otros

89. En este punto de mi vida, *me* es importante comprender qué es lo que quiero hacer.
90. En este punto de mi vida, es importante *para mí* comprender qué es lo que quiero hacer.
91. Por favor, *salir* afuera todos.
92. Te quiero *hace* muchos años, tú lo sabes.
93. Nos tranquilizamos cuando por fin volviste en *sí*.
94. Los periodistas *subieron* la montaña en helicóptero.
95. Los periodistas *subieron a* la montaña en helicóptero.
96. Me duele *mi* cabeza.
97. Tómate *tu* leche.
98. Yo *me le* acerqué.
99. Pensé que *me les* escaparía.

100. *Me les* incorporo mañana.
101. Ya no quiero *más nada*.
102. En la familia, estábamos seguras de que *más nunca* regresarías.
103. *Más nunca* se la vio por aquí.
104. Esta roca es imposible de *ser movida*.
105. Esta roca es posible de *mover*.
106. Este libro es fácil de *leerse*.
107. Su casa está segura mientras *que* viaja.
108. *Me se* perdió el paraguas.
109. *Te se* acabaron las vacaciones.
110. Algún día tendrás que contárme*lo todo* sobre mi padre.
111. Algún día tendrás que contarme *todo* sobre mi padre.

CODIFICACIÓN Y ANOTACIÓN DEL HABLA EN UN CONTEXTO BILINGÜE: EL CORPUS ESLORA DE ESPAÑOL DE GALICIA*

Speech codification and annotation in a bilingual context: The ESLORA corpus of Galician Spanish

VICTORIA VÁZQUEZ ROZAS
Universidade de Santiago de Compostela

MARIO BARCALA
NLPgo Technologies, S.L.

EVA DOMÍNGUEZ NOYA
Centro Ramón Piñeiro para a Investigación en Humanidades, Universidade de Santiago de Compostela

ALBA FERNÁNDEZ SANMARTÍN
Universidade de Santiago de Compostela

GUILLERMO ROJO
Universidade de Santiago de Compostela

MARÍA PAULA SANTALLA
Universidade de Santiago de Compostela

* El corpus ESLORA fue financiado por el Ministerio de Economía y Competitividad a través de los proyectos de investigación ESLORA (FFI2010-17417) y ESLORA2 (FFI2014-52287-P), y actualmente por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto ESLORA+ (FFI2017-86379-P). El equipo del proyecto forma parte del grupo de investigación Gramática del español de la Universidad de Santiago de Compostela, beneficiario de una ayuda para «Consolidación e estruturación de Grupos con Potencial de Crecemento 2017» de la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (ED431B 2017/39). El equipo de ESLORA está asimismo integrado en la Red Temática en Estudios de Análisis del Discurso, financiada por el Ministerio de Ciencia, Innovación e Universidades a través de la AEI (FFI2017-90738-REDT).

Resumen
El artículo ofrece en primer lugar una caracterización general del diseño y composición del corpus ESLORA y muestra su utilidad para el análisis de la variación social y situacional. El corpus supone asimismo una aportación al estudio de los procesos de cambio ligados a la variación geográfica del español, puesto que registra su uso en un territorio con lengua propia, lo que facilita el reconocimiento del español hablado en Galicia como objeto de investigación de la dialectología hispánica y permite su comparación en pie de igualdad con otras variedades geográficas. En el núcleo del trabajo se describen algunas de las dificultades que surgen en la transcripción, codificación y anotación de registros de habla en un contexto de contacto lingüístico y se exponen los argumentos que sustentan las soluciones adoptadas.

***Palabras clave*:** corpus oral, español, variación, transcripción, anotación

Abstract
This article provides an overview of the design and composition of the corpus ESLORA and shows its usefulness in analysing social and situational variation. The corpus also contributes to the study of the processes of change related to the geographical variation of Spanish, since it records its use in a region with its own distinctive language. This aspect facilitates the recognition of the Spanish spoken in Galicia as a research object in the field of Hispanic dialectology, allowing for its comparison with other geographical varieties on an equal footing. The main focus of the paper is on some difficulties encountered during transcription, codification, and annotation of spoken recordings as well as with the arguments that justify the solutions taken by the research team.

***Keywords*:** spoken corpus, Spanish, variation, transcription, annotation

1. INTRODUCCIÓN

El desarrollo de corpus orales en español se inició hace más de medio siglo con el *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades de Iberoamérica y de la Península Ibérica* (*cfr.* Lope Blanch 1986), una empresa colectiva que supuso un cambio radical en la consideración de los registros de habla como base empírica de la investigación lingüística. Superando limitaciones técnicas, materiales y humanas, el proyecto logró con creces sus objetivos, pues no solo amplió el conocimiento de la realidad del español en América y en España sino que abrió nuevas perspectivas teóricas y metodológicas para el estudio de la variación en los diferentes niveles de análisis.

La propuesta de la *Norma culta* recogía algunos parámetros de variación que se han incorporado en buena parte de los corpus orales del español desarrollados desde entonces. Además de la variable geográfica, que permitió abordar estudios comparativos, los materiales documentan de forma equilibrada el habla de hombres y mujeres de diferentes franjas etarias mediante muestras de discurso formal e informal. Por otra parte, frente a la investigación dialectal tradicional, basada en el uso lingüístico de hablantes rurales de edad avanzada y con una escolarización limitada, la *Norma culta* se centró en los hablantes urbanos «cultos». No se contemplaba, por tanto, la variable estrática, común en la investigación variacionista, que sí se incluyó en el diseño de otros corpus orales elaborados posteriormente.

Frente al interés creciente por el registro del español hablado, sea con propósitos sociolingüísticos (PRESEEA, por ejemplo) o dialectológicos (COSER)[1], se echa en falta

[1] Las referencias completas a los corpus y otros recursos electrónicos mencionados en el texto se encuentran al final

una mayor atención al uso oral del español en comunidades bilingües. Ciñéndonos a la situación en España, el retraso en documentar adecuadamente el español hablado en Galicia, Cataluña, Valencia o el País Vasco es, en parte, el resultado del sesgo normativo que ha dominado en los estudios sobre estas variedades y, en parte, la consecuencia de su exclusión como objetos de estudio con entidad propia en la dialectología tradicional. No obstante, en los últimos años algunas iniciativas han empezado a llenar ese vacío (Vila Pujol 2001, Sinner 2001, Gómez Molina 2001, Briz y grupo Val.Es.Co 2002, Vann 2009, Paasch-Kaiser 2015, además de los equipos correspondientes del proyecto PRESEEA, así como diversos estudios basados en materiales procedentes de COSER, entre otros, Fernández Ordóñez 2007, De Benito 2015, Camus & Gómez Seibane 2015, Gómez Seibane 2015).

En este contexto se hizo patente la necesidad de documentar el uso oral del español en Galicia, ya que el estudio de esta variedad se reducía prácticamente a la identificación y caracterización de sus peculiaridades («galleguismos»), casi siempre a partir de su empleo en obras literarias o textos periodísticos de autores gallegos y observaciones aisladas de la lengua hablada (*cfr*. Rabanal 1967, García 1986, Acín 1996, Rojo 2004, Fernández-Ordóñez 2016)[2]. La creación de un corpus de habla permitiría ofrecer una base empírica adecuada para fundamentar las descripciones de una variedad apenas investigada y facilitar su comparación con otras variedades, al tiempo que reivindicaba su legitimidad como objeto de estudio lingüístico más allá de consideraciones prescriptivas.

El diseño de un corpus oral y los procesos de registro, codificación y tratamiento de los materiales para construir un recurso útil para el análisis lingüístico traen consigo múltiples decisiones teóricas y metodológicas, que en el corpus ESLORA están asimismo condicionadas por el objetivo de documentar el uso del español en un contexto bilingüe. A las dificultades inherentes al trabajo con muestras orales se añade, pues, el hecho diferencial de elaborar un corpus del español hablado en Galicia, un territorio con lengua propia.

En el presente trabajo se exponen las principales características de ESLORA, desde su diseño hasta las múltiples posibilidades de consulta y explotación que ofrece (*cfr*. también Barcala *et alii* 2018). En sus diferentes apartados se describen y justifican los distintos aspectos y fases de construcción del corpus, pero se desarrollan y fundamentan con más detalle aquellos fenómenos que presentan un interés especial en el ámbito de la variación lingüística. Se destaca así la aportación imprescindible de los corpus de variedades distintas de la estándar a la investigación lingüística en general y a la lingüística de corpus en particular.

El contenido del capítulo está estructurado de la siguiente manera. En la sección 2 se describen la composición y características generales del corpus. En la sección 3 se discuten los aspectos más relevantes de la construcción, codificación, anotación y consulta del recurso. El apartado 3.1 está dedicado al proceso de construcción (obtención del consentimiento informado, grabación, herramientas de transcripción y alineación, y anonimización); en el 3.2 se exponen y justifican las particularidades del sistema de codificación ortográfica adoptado; el 3.3 se centra en el sistema de anotación, con especial atención a la etiquetación morfosintáctica; y en el 3.4 se detallan las características y opciones que ofrece la aplicación de consulta en línea. Por último, la sección 4 aborda aspectos cuantitativos del corpus y ofrece un ejemplo ilustrativo de la

del trabajo.

[2] Los trabajos de Celia Pollán (2001, 2002) constituyen una excepción a esta tendencia, puesto que están basados en el análisis del contenido de un corpus de lengua oral, el de la lengua hablada en A Coruña (*cfr*. Fernández Rodríguez no publicado).

rentabilidad del corpus para el estudio de las frecuencias léxicas. El capítulo se cierra con un apartado final de síntesis y proyección hacia el trabajo futuro.

2. EL CORPUS ESLORA

El objetivo general del proyecto ESLORA es poner a disposición de investigadores y personas interesadas un corpus de español hablado en Galicia en las condiciones más adecuadas para su uso y explotación en la investigación lingüística y especialmente en el estudio de la variación. La finalidad del corpus ha guiado su diseño y características, que se manifiestan en tres aspectos nucleares: (i) la composición del corpus, (ii) su condición de recurso de acceso abierto, (iii) su enriquecimiento mediante herramientas de PLN.

(i) El corpus está formado por 54 entrevistas semidirigidas, de una hora aproximadamente, y 20 horas de conversación espontánea. Las entrevistas corresponden a 27 mujeres y 27 hombres divididos en tres grupos de edad (de 19 a 34 años, de 35 a 54 y de 55 o más años) y tres niveles de estudios (primarios, medios y universitarios). En la versión accesible actualmente en internet (1.2.2, de noviembre de 2018) están disponibles 53 entrevistas y 3 conversaciones, que suman un total de 647 758 formas ortográficas (medida poco adecuada para un corpus de este tipo) y 776 260 elementos gramaticales[3].

Los informantes de la muestra de entrevistas residen en Santiago de Compostela y su entorno, y en su mayoría son originarios de la ciudad y su área de influencia. Los participantes en las conversaciones proceden de diferentes puntos de Galicia y, en casos contados, de fuera de la comunidad. De todos ellos se recopila la información sociológica básica (mujer/hombre, edad, nivel educativo), que constituye el núcleo de los metadatos. Se registran además el papel comunicativo (entrevistador, informante o audiencia en las entrevistas), la relación o conocimiento previo entre los participantes y las circunstancias, localización y tiempo del encuentro.

Pero para documentar adecuadamente el uso del español por parte de hablantes gallegos no basta con registrar entrevistas y conversaciones y anotar los metadatos habituales para llevar a cabo estudios de corte sociolingüístico. Para alcanzar una imagen más completa de la realidad lingüística reflejada en el corpus, los participantes entrevistados respondieron a un cuestionario detallado sobre sus usos y actitudes lingüísticas en relación al español y al gallego y realizaron un test de inseguridad lingüística. Tanto el cuestionario como el test fueron grabados con el fin de analizar sus ventajas y desventajas metodológicas como instrumentos para el estudio de ideologías y actitudes lingüísticas (*cfr*. Recalde 2012).

(ii) Muchas de las características del corpus ESLORA están relacionadas con el objetivo prioritario de su puesta a disposición pública en condiciones apropiadas para ser realmente útil a las personas interesadas. Para poder ofrecer el acceso abierto a los materiales, los participantes firmaron su consentimiento informado con la condición de preservar su privacidad mediante la anonimización de las transcripciones y los audios, un proceso que se llevó a cabo de forma meticulosa (*cfr*. *infra* apdo. 3.1.4).

La accesibilidad y legibilidad del corpus se plasma también en el tipo de transcripción elegida, basada en la convención ortográfica y con un reducido componente interpretativo. La alineación de los textos transcritos con el sonido permite

[3] Llamamos elementos gramaticales a los que resultan de llevar a cabo los procesos de lematización y análisis morfosintáctico habituales en el procesamiento de corpus. Así, en una forma como *diciéndomelo* se reconoce la existencia de tres elementos (el gerundio *diciendo*, el pronombre *me* y el pronombre lo), en *al* hay dos elementos, mientras que *Instituto Nacional de Estadística* es considerado como una unidad.

la consulta inmediata del fragmento de audio correspondiente a cada búsqueda. Está disponible asimismo la descarga del corpus completo en formato textual y, bajo petición, se pueden obtener las grabaciones completas, el texto etiquetado y los datos sociolingüísticos recogidos en los cuestionarios mencionados. Para facilitar la recuperación de la información contenida en el corpus se diseñó una completa aplicación de consulta en línea con su correspondiente guía de uso, en la que se detallan las amplias posibilidades de búsqueda que ofrece el sistema (*cfr. infra* apdo. 3.4).

(iii) Dado que el corpus fue concebido como un instrumento para el análisis del español oral, el diseño general del recurso y las decisiones específicas de transcripción y codificación estuvieron dirigidas a facilitar el posterior enriquecimiento de los materiales mediante herramientas de PLN. Así, se optó por un sistema ortográfico de representación textual que permitiera la anotación morfosintáctica de las transcripciones con los programas de lematización y etiquetación disponibles, pensados para textos escritos estandarizados. Se evitó, por tanto, el empleo de elementos gráficos convencionales, como subrayados, cursivas, comillas, guiones, etc., y se optó por signos y etiquetas unívocas y procesables por medios automáticos. Otro elemento central en el diseño del corpus fue el empleo de programas de alineación de la transcripción con el audio, que permite el tratamiento conjunto de texto y habla y abre nuevas vías de aprovechamiento y explotación de los materiales.

3. Construcción, codificación y anotación de ESLORA

3.1. El proceso de construcción

3.1.1. Grabaciones

Todas las entrevistas se registraron en archivos WMA con grabadoras digitales Olympus DS-40 con micrófono integrado y sonido estéreo de extra alta calidad (ST XQ). En la mayor parte de los casos, la grabadora fue accionada antes del encuentro con el informante, de manera que quedaron registrados los saludos, presentaciones y, en general, charlas espontáneas previas a la entrevista propiamente dicha. Durante la entrevista la grabadora estaba a la vista, sin obstáculos físicos que pudiesen entorpecer la grabación. En general, se obtuvieron audios de buena calidad y libres de ruido de ambiente, pues las entrevistas se realizaron en casas particulares y, ocasionalmente, en los lugares de trabajo de los participantes.

En el caso de las conversaciones, la gran mayoría se grabó en formato WAV y, más raramente, en MP3, mediante aplicaciones de grabación para dispositivos electrónicos, casi siempre teléfonos móviles. En casi todos los casos se empleó la aplicación *Tape-a-talk Voice Recorder*, disponible para dispositivos Android de manera gratuita. Dicha aplicación permite efectuar grabaciones de buena calidad en formatos mp3, wav o 3gp. En estos casos, la disminución de calidad con respecto a la grabadora profesional se ve compensada por el hecho de que los móviles, al formar parte de la vida cotidiana de todos los participantes, podían situarse en el lugar más conveniente para la grabación sin levantar ningún tipo de sospecha, mientras que la grabadora profesional tenía que permanecer oculta en algún lugar no visible, lo cual podía obstaculizar la grabación.

En cualquier caso, todos los archivos que forman finalmente parte del corpus tienen una calidad de grabación media-alta, ya que fueron descartados todos aquellos que presentaban cualquier tipo de problema que dificultase la transcripción y la posterior escucha por parte del usuario del corpus (ruido de fondo, interferencias…).

3.1.2. Transcripción y alineación

Todas las entrevistas fueron transcritas y alineadas de forma manual empleando el programa Transcriber[4], que proporcionaba un entorno cómodo y rápido para codificar interacciones de dos participantes con un esquema de pregunta-respuesta. Transcriber cuenta además con un sistema de introducción de etiquetas fácilmente personalizable y de manejo sencillo.

Sin embargo, Transcriber presenta serias limitaciones a la hora de transcribir interacciones espontáneas con más de dos participantes. Por poner un ejemplo, la interfaz no permite introducir solapamientos en los turnos de más de dos hablantes. A ello se suma su falta de mantenimiento y actualización en los últimos años, que da lugar a continuos problemas técnicos. Por estos motivos, en la segunda parte del proyecto se decidió transcribir las conversaciones utilizando el programa ELAN[5], que, con un sistema de transcripción por niveles o *tiers* (el denominado sistema «en pentagrama»), y una gran cantidad de posibilidades de personalización y exportación, permite reflejar más fielmente la estructura de la conversación espontánea.

3.1.3. Permisos

El consentimiento informado de los participantes es un requisito legal y ético para el registro de los datos, que adquiere especial relevancia en el caso de las conversaciones, ya que son registradas sin el conocimiento de los participantes en ese momento concreto. Así, mientras que en las entrevistas se consideró suficiente la obtención de un consentimiento verbal anterior, y de uno escrito posterior, en el caso de las conversaciones se obtuvo de los participantes un doble permiso por escrito: uno previo, en el que aceptaban participar en el proyecto y ser grabados en cualquier momento, y otro posterior, en el que daban su autorización para emplear las grabaciones realizadas. En ambos casos, el equipo investigador se comprometía a restringir el uso de las grabaciones y transcripciones a los ámbitos de la investigación y la docencia, así como a la preservación del anonimato de los participantes.

3.1.4. Anonimización

La anonimización de los materiales, además de responder a un compromiso adquirido a través del formulario de consentimiento, constituye una obligación ética del investigador, entre cuyos deberes está evitar a los participantes cualquier tipo de perjuicio derivado de su participación en el proyecto.

El proceso no estuvo exento de complicaciones, ya que los elementos que pueden conducir a la identificación de una persona son múltiples y complejos. De este modo, preservar el derecho a la intimidad de los participantes y su entorno sin comprometer la validez y coherencia de los datos necesitó de toda la atención, buen juicio y en algunos casos ingenio de los transcriptores y revisores.

La anonimización se llevó a cabo sobre lo que la página web del *UK Data Archive*[6] denomina *identificadores directos*, que se clasifican en tres categorías: nombres, apellidos y apodos de personas; nombres de lugares (ciudades, pueblos, calles, etc.), y nombres de instituciones, agrupaciones o asociaciones (colegios, partidos políticos, etc.). Puesto que el corpus pone a disposición del investigador tanto la transcripción de las grabaciones como el audio, la anonimización se efectuó también en ese doble plano.

[4] http://trans.sourceforge.net
[5] https://tla.mpi.nl/tools/tla-tools/elan
[6] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative

Las opciones a la hora de anonimizar la transcripción son muy variadas. Algunas de las más frecuentes consisten en sustituir los nombres propios por iniciales –opción empleada, por ejemplo, en el *Corpus de habla culta de Salamanca* (CHCS)–, por nombres genéricos, del tipo *topónimo* o *nombre propio* –o simplemente *name*, como en el *British National Corpus* (BNC)–, por símbolos (como en el *Corpus de lengua hablada de la ciudad de A Coruña*, *cfr*. Vázquez Veiga 2003: 9) o incluso por códigos alfanuméricos. En el caso de ESLORA la opción elegida fue la empleada, entre otros, en el *Santa Barbara Corpus of Spoken American English*, consistente en sustituir aquellos nombres que se deseaba anonimizar por otros nombres del mismo tipo. Sin duda, este es el sistema cuya aplicación resulta más compleja, pero presenta una serie de ventajas para el usuario del corpus. Entre otras, permite captar con más facilidad la coherencia interna de las referencias, así como recibir gran parte de la información que los nombres propios originales proporcionan acerca de sus referentes.

En primer lugar, los nombres, apellidos y apodos de personas fueron sustituidos por otros nombres reales, tratando, no solo de que fuesen métricamente equivalentes, sino que también lo fuesen social y culturalmente. Como señala Sampson (2000), en muchos casos los nombres pueden estar asociados a un determinado grupo de edad, clase social o procedencia. En la misma línea, Agha (2007: 65-66) sostiene que los nombres propios pueden proporcionar una gran cantidad de información acerca de sus referentes, concerniente a su género, lugar o circunstancias de nacimiento, afiliación, religión o pertenencia a un grupo. Toda esa información se pierde irremediablemente si los nombres se sustituyen por símbolos, códigos o iniciales.

Por este motivo, en ESLORA se llevó a cabo un proceso de sustitución que no solo respeta la oposición de género, las diferencias entre nombres y apellidos o los diminutivos, sino que también refleja otros aspectos, como por ejemplo, las pistas acerca de las características demográficas de los referentes. Para empezar, dado que se trata de un corpus recogido en Galicia, aparecen en él multitud tanto de nombres (Breixo,[7] Xoán, Catuxa) como de apellidos (Ferreiro, Barreiro, Calviño) típicamente gallegos. En todos los casos se sustituyeron por otros nombres (Antón, Anxo, Sabela) o apellidos (Piñeiro, Cunqueiro, Mariño) que conservan la referencia al origen de los referentes. Del mismo modo se actuó con nombres extranjeros como Richard, sustituidos por equivalentes como por ejemplo Peter.

Además de la información geográfica, algunos nombres también proporcionaban pistas sobre otros aspectos, como la edad de la persona aludida o su grado de cercanía con el emisor. Así, en las grabaciones aparecen nombres muy frecuentes en el pasado pero que no lo son tanto en la actualidad, y que por tanto suelen corresponder a personas de edad avanzada. Es el caso de nombres como Brígida, Casilda, Socorro o Celestino que fueron sustituidos por otros con esa misma característica como Amelia, Luciana, Angustias o Edelmiro. En el caso de los nombres abreviados, como Tito, Chema, Cuca, Concha o Paqui, se reemplazaron por otros que conservan su carácter cercano y familiar, como Fito, Berto, Charo, Espe o Puri. Este proceso de anonimización supuso un laborioso trabajo que en muchos casos sobrepasó los límites de la lingüística y se introdujo de lleno en la sociología. En este sentido, resultó de gran utilidad la página del Instituto Nacional de Estadística, que, en una de sus secciones, ofrece la posibilidad de averiguar la edad

[7] Los ejemplos de nombres sustituidos no corresponden en ningún caso con los nombres reales que aparecen en las grabaciones originales, ya que esto constituiría una infracción del acuerdo de confidencialidad. Se trata de nombres inventados similares a los eliminados. Los nombres que se presentan para ejemplificar las sustituciones sí que forman parte de los que se emplean realmente en el corpus.

media de las personas que tienen un determinado nombre o apellido, además de su procedencia geográfica[8].

Cabe señalar, para terminar con los nombres de persona, que aquellos pertenecientes a personajes de la esfera pública, como actores, deportistas, políticos, escritores, etc., se excluyeron del proceso de anonimización, siempre que las referencias no atañesen a su vida privada y no resultaran lesivas para su imagen ni atentasen contra su honor (Sampson 2000). En estos casos la conservación de las referencias puede tener interés documental, más allá del ámbito estrictamente lingüístico, ya que reflejan la visión que los distintos informantes tenían de la sociedad en la que vivían en un momento histórico concreto.

Sí se cambiaron algunos nombres de personajes públicos cuando la referencia podía llevar a la identificación del hablante, como en un caso en el que el informante menciona durante la entrevista que ganó un certamen literario cuyo premio consistía en unas clases con un famoso escritor. Su nombre hubo de ser cambiado por el de otra figura literaria, ya que a través del nombre del escritor se podría identificar el premio y por tanto a su ganador.

Otro tipo de identificadores directos son los nombres de ciudades, pueblos, aldeas, etc., muy frecuentes en la mayor parte de las grabaciones. En estos casos, la sustitución no se llevó a cabo de manera sistemática, sino valorando en cada caso cuáles eran las ventajas y los riesgos de conservar la referencia original. En primer lugar, se conservaron las abundantísimas alusiones a la ciudad de Santiago, ya que todos los receptores del corpus saben de antemano que los informantes son naturales de dicha ciudad o residentes en ella. Tampoco se modificaron sistemáticamente los nombres de otras ciudades cuando aparecían en alusiones a viajes, opiniones, descripciones, etc. En síntesis: solamente se cambiaron las referencias que pudiesen conducir a la identificación del informante o de alguna persona de su entorno. Se operó de modo similar con los nombres de barrios, calles, plazas, etc., que solo se modificaron en caso de hacer referencia a la localización exacta del domicilio del hablante o de otra persona aludida en la interacción, así como a su lugar de trabajo o cualquier otro elemento identificador. Y otro tanto con los nombres de colegios, institutos, instituciones, comercios o asociaciones.

En los casos en los que se consideró oportuno modificar los topónimos, se llevó a cabo el mismo procedimiento que con los nombres de personas, tratando de conservar, en la medida de lo posible, la información que cada denominación proporcionaba acerca de su referente. Así, por supuesto, los nombres de barrios se cambiaron por nombres de barrios, los pueblos por pueblos, ciudades por ciudades y así sucesivamente. Pero además, se trató de escoger nombres que compartiesen con los originales sus características más definitorias. Por ejemplo, si el informante mencionaba sus vacaciones en un pueblo típico de veraneo de la costa gallega como Sanxenxo, la sustitución se hizo con otro de características similares como por ejemplo Baiona.

En todos los casos se trató de mantener la coherencia interna que permitiese el seguimiento de la continuidad referencial del discurso, tratando además de resultar verosímil en lo que se refiere a la información geográfica y demás indicaciones contextualizadoras.

Los elementos modificados aparecen en la aplicación de consulta resaltados en color amarillo, lo cual indica al lector que el nombre que figura en la transcripción no es el que aparece en el audio original.

[8] https://www.ine.es/widgets/nombApell/index.shtml.

Por su parte, la anonimización del audio consistió en introducir un ruido en todos los fragmentos anonimizados en la transcripción, empleando para ello el programa de tratamiento de sonido Audacity[9].

3.2. El sistema de codificación

La creación de una transcripción a partir de una grabación de habla implica la plasmación textual selectiva y parcial de un registro sonoro del evento comunicativo original. De la misma manera, tanto los archivos de audio como las transcripciones de las entrevistas y conversaciones que integran el corpus ESLORA constituyen una selección de datos relevante para el análisis del español de Galicia, pero no dejan de ser representaciones indirectas y fragmentarias del objeto de estudio como tal. Como apuntó Elinor Ochs (1979), las decisiones que el analista adopta en el proceso de transcripción no son neutrales, sino que están condicionadas por sus presupuestos teóricos e ideológicos y por el propósito de su investigación. Duranti (2006: 308) afirma que «transcription is a cultural activity used for creating and sustaining a science of the particular slice of the universe that interests us (…)». Son por tanto esos intereses los que orientan el proceso de registro y representación del habla.

El objetivo de documentar y analizar el uso oral del español en Galicia ha guiado el diseño y construcción del corpus ESLORA, tanto en la composición de la muestra –selección de hablantes y géneros discursivos–, como en la configuración de las transcripciones y de los metadatos recogidos a través de cuestionarios durante su elaboración.

Para la transcripción de las grabaciones se optó por un único nivel de representación que sigue la ortografía convencional excepto en dos aspectos. Uno de ellos es la puntuación, limitada a los signos de interrogación y admiración. Se prescinde, por tanto, de puntos, comas, etc., dado que su uso tiene una función convencional de estructuración de enunciados y organización textual propia de la escritura, mientras que en la codificación del habla interesa dejar constancia de las pausas, que se clasifican dependiendo de su duración en pausa breve, pausa larga y silencio.

El segundo aspecto diferencial es el empleo de mayúsculas, que se reduce a los nombres propios puesto que no hay mayúsculas dependientes de la puntuación. Como la distinción entre nombre común y nombre propio es problemática en ciertos casos, para garantizar un tratamiento homogéneo de las mayúsculas se adoptó como criterio básico de transcripción la norma del *Diccionario panhispánico de dudas* (DPD), que era el texto de carácter normativo más reciente de la Real Academia Española en el momento en que arrancó el proyecto. El DPD no disipa sin embargo todas las dudas, pues admite ambas posibilidades, por ejemplo, para los nombres de marcas comerciales, dependiendo de si la referencia es más o menos específica, una distinción no siempre determinable en las muestras del corpus.[10] En todo caso, para solventar la dificultad planteada por alguna discordancia que subsista en el corpus, en el sistema de consulta se ofrece también la opción de recuperar las formas con independencia del uso de mayúsculas o minúsculas (*cfr. infra* apdo. 3.4)

En relación con las mayúsculas, cabe indicar que se han usado en el término gallego *Rúa* 'calle', presente en denominaciones como *Rúa Nueva* (gal. *Rúa Nova*), *Rúa del*

[9] https://www.audacityteam.org/.

[10] «Las marcas comerciales son nombres propios, de forma que, utilizados específicamente para referirse a un producto de la marca, han de escribirse con mayúscula: *Me gusta tanto el Cinzano como el Martini*; *Me he comprado un Seat*; pero cuando estos nombres pasan a referirse no exclusivamente a un objeto de la marca en cuestión, sino a cualquier otro con características similares, se escriben con minúscula: *Me aficioné al martini seco en mis años de estudiante* (al vermú seco, de cualquier marca)» (DPD, *s.v. mayúsculas*, apdo. 4.22.).

Villar (gal. *Rúa do Vilar*), porque se considera parte del nombre de la calle, a diferencia de los sustantivos españoles *calle* y *avenida*, que se han tratado como comunes siguiendo el DPD. Por otra parte, en la escritura de los topónimos se reproduce la elección de cada participante, sea la forma original gallega (*Ourense, Vilagarcía, Conxo*, etc.), sea la adaptación al español (*Orense, Villagarcía, Conjo*).

Frente a una posible representación fonética, el uso de un sistema ortográfico aligera el trabajo de transcripción, permite abordar la construcción de un corpus más amplio y tiene ventajas de legibilidad y comparabilidad; además, la escritura convencional facilita la búsqueda y recuperación de información y posibilita la aplicación de herramientas estándar de PLN para la anotación textual (*vid.* Poplack 1989: 430; 1993: 265-266; Edwards 2001: 324; Torres Cacoullos & Travis 2018: 46-47). No obstante, la codificación ortográfica puede verse también como un obstáculo para la documentación de la variación, especialmente en un corpus que tiene entre sus objetivos visibilizar usos que se alejan de los considerados normativos.

La cuestión de la representación escrita de las variantes lingüísticas no canónicas ha sido objeto de debate al menos desde que Dennis Preston llamó la atención sobre el efecto de los *respellings* utilizados en la tradición folclorista. Preston (1982, 1985, 2000) muestra cómo la «escritura dialectal» (*eye dialect*) estigmatiza a los hablantes representados, sobre todo si, como suele ser el caso, la codificación de las variantes menos prestigiosas no afecta por igual a todos los participantes (*cfr.* también Bucholtz 2000; Jaffe & Walton 2000; Beal 2005). La alternativa, si se requiere la codificación detallada de las particularidades del habla, es recurrir a un sistema de transcripción fonética como el *Alfabeto Fonético Internacional*. En ESLORA, al igual que en otros corpus de construcción reciente, el acceso inmediato al audio alineado deja abierta la posibilidad de añadir a la representación ortográfica una línea o nivel de representación fonética (un nuevo *tier* en el sistema de ELAN). No obstante, la puesta a disposición libre de los datos sonoros puede no ser la opción más adecuada en otros casos. Torres-Cacoullos & Travis (2018: 48) advierten de las interpretaciones erróneas a que pueden dar lugar los registros de audio si los investigadores no están familiarizados previamente con el uso lingüístico representado, además de las consecuencias no deseables que derivan del reforzamiento de estereotipos negativos hacia una comunidad que habla una variedad no estándar.

Menos problemática para la imagen de los hablantes, pero tampoco exenta de dificultades, es la representación textual de otros fenómenos característicos del habla para los que no disponemos una convención gráfica fija, como los alargamientos vocálicos y consonánticos, las palabras truncadas o la risa. El fragmento de (1) muestra varias etiquetas XML de alargamientos, palabras cortadas, énfasis, pausa corta y pausa larga, marcas que si bien enriquecen la transcripción y permiten refinar las opciones de recuperación automática de información, presentan el inconveniente de impedir la lectura fluida del texto. En la aplicación de consulta del corpus ESLORA se ha resuelto este problema de visualización mediante indicaciones que se activan al pasar el cursor por los elementos afectados (*cfr. infra*, apdo. 3.4).

(1) <alargamiento>pues</alargamiento><pausa/> eeh mi madre había <alargamiento>sido</alargamiento> <palabra_cortada>me</palabra_cortada> eh mecanógrafa <pausa/> y yo tenía nociones <alargamiento>de</alargamiento> <pausa_larga/> <palabra_cortada>taquin</palabra_cortada> de <énfasis_inicio/>mecanografía<énfasis_fin/> (SCOM_M22_034)

El fragmento incluye también la representación de las vocalizaciones *ehh* y *eh*, que manifiestan vacilación y funcionan como prolongadores. En esta categoría se incluyen elementos interjectivos, unidades denominadas «cuasi-léxicas», «pausas sonoras», «apoyos vocálicos», que en muchos casos no poseen una forma escrita estandarizada, aunque sí se representan de forma convencional con más o menos variantes en algunos tipos de textos, como en las viñetas de *cómics* y novelas gráficas. Algunos protocolos de transcripción restringen las opciones de representación ortográfica a unas pocas formas preestablecidas. Por ejemplo, Tagliamonte (2004: 5-6) admite solo tres «hesitation words» (*um, ah, oh*) y establece una lista cerrada de «legal fillers», cada uno con su correspondiente significado.

En el ámbito del español, el macroproyecto PRESEEA también propone codificaciones convencionales para unidades como *ah, ay, aha, mmm, eeh, pff, bah*, y para la representación de onomatopeyas (*zas, bum, plas*), aunque no limita explícitamente el repertorio de variantes posibles[11]. Prueba de la convencionalidad de algunas de esas expresiones es su reconocimiento como entradas de diccionario. El DLE, por ejemplo, incluye *ah, ay, bah, zas, bum, plas*, pero no *mmm, ehh, pff y aha* (aunque sí *ajá*).

Una de las tareas actualmente en curso en el proyecto ESLORA tiene como objetivo ofrecer una codificación consistente y homogénea de tal tipo de unidades, que por una parte represente la variedad y particularidades de su forma y uso en las muestras del corpus y que, por otra parte, introduzca un cierto nivel de abstracción y generalización que permita contar con un conjunto limitado de opciones. Teniendo en cuenta que el acceso al audio alineado es inmediato y está a disposición de los estudiosos para análisis más delicados, el nivel básico de transcripción ortográfica debe acotar la dispersión que contiene la versión actual, con variantes como *eh, ee, eeh, ehh, eeeh*, o como *buf, buff, bufff, buuf*, entre otras.

Sin embargo, la unificación de la forma escrita de interjecciones y vocalizaciones requiere especial cautela, ya que la estandarización ortográfica tiende a favorecer la representación y el reconocimiento de las formas propias de variedades asociadas con el estándar y en cambio puede ignorar las formas y usos propios del habla considerada «dialectal». En ESLORA se documentan formas propias, y quizá exclusivas, del español de Galicia (y del gallego), que deben ser identificadas ortográficamente. Véase, por ejemplo el uso de *ho* en (2), que no debe confundirse con la interjección *oh* recogida en (3):

(2) lo malo que hay eh pue pue <pausa/> nos pues bueno nosotros con el repertorio que tenemos ya nos llega *ho* <pausa/> (SCOM_H31_046)

(3) en esa excursión fue como muy divertido ¿ sabes ? en plan <pausa/> españoles se encuentran a un español <pausa/> famoso y vas en plan *oh* no sé qué ¿ sabes ? <pausa/> fue como muy <risa/> <pausa/> muy gracioso sí <pausa/> (SCOM_M13_008)

En (4) se observa la forma *boh*, que hay que distinguir de *bah*, ejemplificada en (5):

[11] http://preseea.linguas.net/Portals/0/Metodologia/Marcas_etiquetas_minimas_obligatorias_1_2.pdf.

(4) otro de mis hermanos <pausa/> que hizo Químicas <pausa/> salió de casa <pausa/> eeh todo convencido de que iba a hacer Medicina <pausa/> y iba ya a matricularse en Medicina <pausa/> dio una vuelta por la Alameda <pausa/> porque era temprano y no sé qué <pausa/> y se encontró a unos mmm amigos de él <pausa/> que <pausa/> que habían hecho el bachillerato <pausa/> iban para Químicas <pausa/> y dijo ¡ *boh* ! pues mira <pausa/> voy para Químicas <pausa/> y se marchó a Químicas <pausa/> (SCOM_M33_005)

(5) (…) cuando veo que son las ocho de la tarde <pausa/> o las nueve que ya llevo doce horas conduciendo <pausa/> digo *bah* pues ahora el próximo hotel que vea <pausa_larga/> paro (SCOM_M23_018)

Otra interjección registrada es *mima* (6), probablemente una abreviación de *mi madriña*, expresión ponderativa también presente en el corpus con un valor similar a *madre mía* o *mi madre* (7).

(6) (…) buff la Navidad ¡*mima* ! yo pongo siempre ya pues el árbol el belén <pausa/> pongo ahí todo el misterio <pausa/> que lo hice yo de manualidades (SCOM_M23_004)

(7) no hay peor cosa que te idealicen una cosa <pausa/> después llegas allí <pausa/> mira <pausa_larga/> (…) llevé una decepción tan grande tan grande tan grande que digo yo *mi madriña* ¿quién me mandaría a mí venir aquí? (SCOM_M31_045)

Se ha hecho referencia a las ventajas de la ortografía estandarizada que prescinde de variantes fónicas al representar los registros orales. Con ella se facilita la búsqueda y recuperación de las unidades lingüísticas y el enriquecimiento de los materiales al aplicarles los mismos recursos de lematización y anotación morfosintáctica disponibles para textos escritos. Sin embargo, de forma contradictoria, el objetivo de documentar adecuadamente los hechos de variación léxica y morfológica justifican la opción de reflejar las opciones de los hablantes que no coinciden con las convenciones normativas (*cfr*. Preston 1982: 323; Tagliamonte 2004; Nagy & Sharma 2013: 240; Torres Cacoullos & Travis 2018: 46-47). La decisión no está exenta de problemas, ya que, como señala Tagliamonte (2004: 2), los límites entre el componente fónico y el morfológico no siempre están claros: «interpreting what is phonological and what is morphological variation is often not clear-cut!».

En la línea de los autores citados, en ESLORA, por una parte, se unifica la transcripción de las variantes de pronunciación manteniendo la ortografía establecida, lo cual implica, por ejemplo, que se escriba *entonces*, *es decir*, *para* o *para adelante*, independientemente de que estas y otras expresiones se reduzcan con frecuencia en la pronunciación. Un caso especial es el de los lapsus de dicción, que se representan tal como se han pronunciado, pero se marcan con la etiqueta <sic> para evitar que se interpreten como errores de transcripción:

(8) mmm tampoco fff siento una gran <sic>actración</sic> <pausa/> atracción por caer en eso ¡eh! <pausa/> realmente <pausa_larga/> (SCOM_H11_052)

Pero, por otra parte, se adopta el criterio de no estandarizar las variantes morfológicas empleadas por los hablantes. La consecuencia es que en el corpus se representen formas no canónicas como *estes* (9) y *eses* (10) para los demostrativos de plural.

(9) Al Aquasella por ejemplo llevo yendo todos los años desde que tengo diecisiete <pausa_larga/> y solo tengo fotos de *estes* dos últimos años <pausa/> (SCOM_H11_047)

(10) es diferente <pausa/> en Cuba no tanto pero bueno <pausa/> Venezuela <pausa/> Colombia ya no te cuento y todos *eses* países sí sí <pausa/> (SCOM_M12_020)

También se registran formas verbales no estándares para la segunda persona de singular del pretérito de indicativo (*dijistes, estuvistes, hicistes,* etc.) y para el presente de subjuntivo del verbo estar (*estea*, *esteamos*). La identificación de estos casos permite su incorporación al sistema de formas reconocibles por el etiquetador morfosintáctico. Retomaremos esta cuestión en el apdo. 3.3.

Requiere asimismo un tratamiento específico la sufijación apreciativa en -iño, que se da con sustantivos y adjetivos tanto de lema español (*hijiño*) como gallego (*filliño*). En el apdo. 3.3 se detalla la casuística del fenómeno y se explican las soluciones de anotación adoptadas.

Como es de esperar, se documentan en el corpus palabras originalmente gallegas integradas en secuencias en español sin que quepa hablar propiamente de cambio de código sino más bien de un repertorio léxico compartido por ambas variedades. Así ocurre con formas como *colo* 'regazo o disposición adoptada por los brazos para sostener a un niño' (11), *cacharela* 'hoguera' (12), *latar* 'hacer novillos' (13), entre otras.

(11) si tengo que ir a algún sitio <pausa/> hombre a lo mejor un día me apetece ir de compras <pausa/> y no me la llevo <pausa/> porque sé que a los diez minutos me dice mamá *colo* <pausa/> mamá <pausa/> no entres no pruebes más <risa/> me dice no ahí a esa tienda no <pausa/> más comercios no <pausa_larga/> (SCOM_M12_030)

(12) hacemos ahí la *cacharela* y montamos una cantina <pausa_larga/> (…) eeh regalamos las sardinas y el pan <pausa_larga/> (SCOM_M23_001)

(13) yo pues sí eh con los diecisiete años en el instituto pues sí las hacías <pausa/> *latabas* <pausa/> te ibas para el bar <pausa/> (SCOM_H12_027)

Lo mismo ocurre con palabras de otras lenguas, sobre todo del inglés, formas como *hall, freelance, realities, jazz, funky*, etc., que están integradas en el uso habitual de diferentes variedades del español (y del gallego) y para las que tampoco se ha considerado conveniente utilizar una etiqueta de lengua en la transcripción.

Por el contrario, sí se marca el cambio de código cuando la alternancia entre español y gallego (u otra lengua) va más allá de una sola palabra, tanto si representa la sucesión entre dos hablantes, con frecuencia en discurso referido (14), como si el cambio se produce en el interior de un enunciado de un hablante (15).

(14) *<gl> e logo <pausa_larga/> non vamos dar una voltiña?</gl>* <pausa/> no mamá porque Sonia no está <pausa/> y podemos caer <pausa_larga/> (SCOM_M31_045)

(15) tienes que estar así todo el día <pausa/> o sea no <pausa/> pues yo cojo y yo me largo para casa vamos *<gl>e non traballo</gl>* (SCOM_H21_039)

3.3. El sistema de anotación

El corpus ESLORA ha sido morfosintácticamente anotado y lematizado al modo habitual en este tipo de recursos, de modo que cada elemento del corpus es adscrito a un lema, es asignado a una clase de palabras y se indica el valor de cada una de las subcategorías gramaticales que le son de aplicación. Las dos últimas informaciones se expresan en lo que se conoce habitualmente como «etiqueta», en la que, de modo sintético, se sitúa en primer lugar la clave que corresponde a la clase de palabras[12] y luego se van añadiendo los símbolos que se refieren a las subcategorías. Así, por ejemplo, en el caso de una forma verbal, la etiqueta VIP1S indica que se trata de una forma de un verbo (V) perteneciente al modo indicativo (I), el tiempo presente (P), primera persona (1) del singular (S). Es decir, el número y el carácter de las clases de palabras, subclases y categorías gramaticales identificados en el sistema de anotación de ESLORA está plenamente en línea con las recomendaciones presentes en los estándares de anotación morfosintáctica (EAGLES en primer lugar) y en la práctica más común de anotación de corpus en español.

El sistema es peculiar, sin embargo, en un aspecto de su desarrollo: para las palabras que, en una determinada subcategoría, puedan presentar más de un valor *x* o *y* en razón del contexto, hay una etiqueta que les atribuye el valor *x*, otra que les asigna el valor *y,* y una tercera que las identifica como con uno de los dos valores *x* o *y*. Cuando el contexto considerado permite identificarlo, a la palabra en cuestión se le asignará en el corpus una de las etiquetas con valores *x* o *y*. Cuando el contexto considerado no lo permite, a la palabra en cuestión se le asignará la etiqueta que se refiere a la indeterminación entre *x* o *y* para la categoría implicada. Este planteamiento, obviamente, extiende el número de etiquetas. Así, un adjetivo como *interesante* tendrá, por causa del género, al menos 5 etiquetas posibles, puesto que, dependiendo del contexto, puede ser masculino [*un libro interesante*], femenino [una exposición interesante], neutro [*Eso es interesante*], masculino o femenino [*Cualquier artista interesante*] y masculino, femenino o neutro [*Resulta interesante*]. El coste que supone el aumento de etiquetas se compensa, sin embargo, con las ventajas que esta aproximación produce en el proceso automático de etiquetación.

Como resultado de todo lo anterior, el sistema de anotación aplicado a ESLORA consta de 455 etiquetas morfosintácticas distintas: 198 de pronombres, 136 de determinantes, 78 de verbos, 15 de sustantivos, 13 de adjetivos, 1 de adverbio, 1 de preposición, 1 de conjunción, 1 de interjección y 1 de puntuación. Las 8 etiquetas restantes sirven para elementos tales como fechas, símbolos o cifras, elementos que, en realidad, no se esperan en el corpus, bien por su carácter oral (es el caso de los símbolos), bien por haber sido transcritos ortográficamente (caso de las fechas o las cifras).

Para anotar el corpus se ha utilizado el etiquetador XIADA, el cual, aunque en principio diseñado para el gallego en el *Centro Ramón Piñeiro para a investigación en humanidades*, se ha podido adaptar con facilidad para etiquetar español en el marco del proyecto ESLORA. Si bien se trata de un etiquetador fundamentalmente estadístico,

[12] En nuestro caso, adjetivo, adverbio, conjunción, determinante, interjección, preposición, pronombre, sustantivo y verbo.

permite también la introducción de reglas lingüísticas que contribuyen a incrementar sensiblemente su nivel de acierto. Aparte de por sus buenos resultados, comprobados para el gallego (*cfr*. Domínguez *et alii* 2009), el etiquetador XIADA fue el elegido para la anotación de ESLORA porque era un etiquetador que se podía adaptar de manera mucho más simple que el resto de los considerados (Freeling en primer lugar) para la gestión simultánea de la etiquetación morfosintáctica y las marcas de oralidad empleadas en el corpus. Todo ello ha tenido como resultado que el producto final, la aplicación web de explotación del corpus, pueda manejar y proporcionar ambos tipos de información, etiquetación y marcas de oralidad, de manera mucho más eficaz y rentable desde el punto de vista del usuario.

Con la herramienta XIADA, la anotación del corpus se ha hecho entonces de manera automática, aunque la parte del mismo que luego sirvió para entrenar el sistema automático implicado (50 000 formas aproximadamente), ha sido también revisada a mano. Para desarrollar ese corpus de entrenamiento, creamos en primer lugar un minicorpus con el tamaño estrictamente necesario para que cada etiqueta del sistema de anotación estuviera representada al menos una vez. Ese minicorpus incluía secuencias reales tomadas del corpus siempre que era posible, y secuencias inventadas en caso contrario. Con ese recurso se etiquetó automáticamente un corpus de 25 000 palabras de entrevistas semidirigidas. Este corpus se revisó manualmente y se utilizó después para etiquetar automáticamente otro corpus de 25 000 palabras de conversaciones, corpus que también fue después revisado manualmente. Para la revisión manual, en cada una de sus fases, los revisores trabajaban con un editor XML personalizado[13] en el que se veían todas las etiquetas posibles de cada elemento (no solo las seleccionadas por el etiquetador), de manera que se podía, de forma cómoda, aprobar la etiqueta propuesta por el sistema, cambiarla por alguna de las rechazadas o incluso proponer una nueva.

Veamos ahora algunos de los problemas surgidos en este proceso, tanto manual como automático, de aplicación de las etiquetas al corpus, en concreto los que tienen que ver con la situación de contacto de lenguas en que se encuentran español y gallego en Galicia. En comparación con los que aparecen en cualquier proceso de anotación automática, se entrecruzan aquí dos tipos de problemas adicionales. Por una parte, están todos aquellos que son propios de la lengua oral y también los que, aunque puedan estar presentes en la lengua escrita, se presentan en la oralidad de un modo diferente y, por tanto, requieren un tratamiento distinto. En este punto, conviene tener en cuenta que la tradición descriptiva suele dar soluciones para los textos escritos, pero no para los orales. Por otra parte, aparecen aquí los problemas causados por las características específicas de una variedad del español, el español de Galicia, surgida, directa o indirectamente, como consecuencia del contacto entre español y gallego[14]. Los rasgos propios de esta variedad pueden, en efecto, requerir actuaciones sobre los recursos utilizados para la etiquetación (el lexicón o módulos de reconocimiento específicos, por ejemplo) o bien exigir la toma de decisiones que afectan a la relación de las etiquetas con las formas y comportamientos gramaticales a los que se refieren (casos como *medio* o formas verbales en *-ra*, *vid. infra*).

En el primer grupo, se encuentran, por ejemplo, los problemas que plantea la frecuente formación de los diminutivos a partir de la forma que es propia del gallego, el sufijo *-iño*. En su forma más simple, se trata del empleo de este sufijo sobre una base del español estándar:

[13] XMLmind XML Editor, URL: http://www.xmlmind.com/xmleditor/.

[14] Las características propias del español de Galicia no son un tema suficientemente tenido en cuenta en el estudio de las variedades del español, tal como señala Rojo (2004: 1091 y ss.). De esta presentación nos servimos en este apartado para la organización de buena parte de los problemas de etiquetación aquí tratados.

(16) eh <pausa/> y comemos <pausa/> a la vuelta comimos en el <pausa/> en el *barquiño* que que íbamos <pausa_larga/> (SCOM_H31_042)

Complicando un poco el problema, la forma en diminutivo es construida ya no a partir de una palabra del español, sino del gallego (*grolo* 'sorbo'), una palabra cuyo uso puede estar más o menos normalizado en el español de Galicia:

(17) si no tomo ahora un *groliño* de café estoy medio (SCOM_M13_008)

A veces, incluso, encontramos en el corpus las formas diminutivas en *-iño* asociadas tanto a una palabra en español como a su equivalente de traducción en gallego:

(18) y yo *filliño* <pausa/> además lo comprara en Portugal por nada por <pausa/> por nada (SCOM_M31_045)

(19) y yo le dije a mi marido dije oy *hijiño* no mejor es no ir porque pft (SCOM_M33_009)

Y, en ocasiones, la palabra en cuestión es también propia del español común, pero su diminutivo solo es posible en la variedad usada en Galicia y tiene además con esa forma un significado muy específico ('amable, entrañable' en el caso de *riquiño*):

(20) ¿sabes? <pausa/> y estás en el reservado hasta las cuatro de la mañana <pausa/> que eso es lo bueno que tienen <pausa/> en el Carretas que no tienen pro ¿sabes? <pausa/> tienen mm <pausa/> son muy *riquiños* en ese sentido (SCOM_M13_008)

A veces también sucede que la forma en diminutivo se constituye en (parte de) una unidad discursiva lexicalizada:

(21) y mi *madriña* <pausa/> las fiestas de Ribeira colega pero pfff (SCOM_H12_027)

En nuestro conjunto de etiquetas no está previsto que las formas apreciativas, ya sean diminutivos u otras modalidades, dejen ningún rastro específico: un sustantivo diminutivo es etiquetado exactamente igual que la forma no diminutiva de la que procede. Pero, naturalmente, el proceso de anotación requiere que ese diminutivo sea reconocido como tal y reciba la etiqueta que le corresponde. Ese objetivo puede ser alcanzado mediante dos caminos distintos. Algunos diminutivos se reconocen como tales a partir del diccionario al que acude la herramienta de etiquetación y otros por medio de un módulo de asignación de etiquetas a formas desconocidas sobre la base del contexto (este módulo se está desarrollando actualmente para la reconstrucción del lema original y su contraste con el lexicón). Ambas estrategias pueden utilizarse para la anotación de los diminutivos formados a partir del sufijo gallego *-iño*. Por supuesto, el problema de que el lema reconstruido no se halle en el diccionario siempre existe[15], tanto si la forma base

[15] Hay diminutivos previsibles, por ser de uso recurrente en español de Galicia, pero hay otros de aparición aislada. En la versión actual de ESLORA hemos hallado, en diferentes géneros y números, los siguientes: *airiño, barquiño, besiño, bueniño, cabeciña, casiña, cojiño, escueliña, filliño, gentiña, groliño, grupiño, hambriña, hijiño, hombriño, madriña, maliño, mujeriña, neniño, normaliño, pachuchiño, pesadiño, pitufiño, pobriño, poquiño, puertiña, riquiño, santiño, tranquiliño, viejiño*.

es española como si es gallega, pero, con una herramienta diseñada para el español estándar, es obvio que el problema se incrementa enormemente en el segundo caso. De hecho, consideramos perfectamente razonable la posibilidad de enriquecer las herramientas de etiquetación para diferentes variedades de una lengua (en situación de contacto con otras o no) con cuantos más y más detallados recursos para cada una de ellas. Con independencia de ello, resulta además muy conveniente la identificación inequívoca en los sistemas de etiquetación de los recursos específicos utilizados para el análisis de características propias de una variedad, de modo que sea posible activarlos o desactivarlos según sea aconsejable en cada caso. Entre estos recursos se encontrarán, sin duda, los que puedan servir para la identificación y anotación correcta de las formas diminutivas en *-iño* en el español hablado en Galicia.

Sin que podamos entrar en detalles al respecto de cada una de ellas, de naturaleza similar son los problemas que para la etiquetación plantea la presencia en español de Galicia de palabras o unidades fraseológicas de distinto tipo procedentes del gallego. Dado que es preciso, por supuesto, asignarles la etiqueta morfosintáctica que les corresponde, es necesario habilitar los recursos para ello. Hay que decidir qué palabras y unidades fraseológicas deben entrar en los recursos generales (el diccionario), identificadas como formas gallegas propias del español de Galicia, y qué otras formas, de uso esporádico o fortuito, han de ser reconocidas bien por el módulo de asignación de etiquetas a palabras desconocidas, bien mediante un diccionario de la lengua en contacto. Con algunos ejemplos de ESLORA, parece claro que el uso de la forma *estes* como demostrativo plural masculino en lugar de *estos* se encuentra en el primer grupo (para empezar es un determinante, con una función como palabra más gramatical que léxica), así como el uso de las formas *dea*, *estea* para los presentes de subjuntivos de los verbos *dar* y *estar*:

(22) ¿sabes? <pausa/> y cosas así ¿sabes? son <pausa/> no en plan <pausa/> como a los <pausa/> tipos *estes* que <pausa/> acaban locos y ven y escuchan cosas en plan <pausa/> ¿qué te decían las voces? ¿sabes? porque yo tuve un colega que le decían unas voces que matase a su madre (SCOM_H11_047)

(23) ya esta semana ya la estás preparando a las nueve para que *estea* <pausa_larga/> durmiendo para irla acostumbrando ya (SCOM_M12_030)

Sin embargo, las cosas dejan de estar tan claras si pensamos en integrar en el diccionario palabras como *home, neno, alá, alí, bico, pola, mallados*, etc., que también encontramos en ESLORA. Con más facilidad, creemos, se tiende a la integración de las unidades fraseológicas que, siendo calco de las gallegas correspondientes o incorporándolas tal cual, se utilizan en español de Galicia, y ello porque las unidades fraseológicas, al fin y al cabo, lo son, en buena medida, por el uso recurrente que se hace de ellas: *de aquella* (gallego *daquela* 'por aquel entonces')[16]*, si cuadra* (gallego *se cadra* 'tal vez', *cfr*. Rodríguez Espiñeira 2019)*, y más* (gallego *e mais* 'y, y sin embargo')*, de carallada* ('de juerga, de broma'), etc. son algunas de las que hemos encontrado en nuestro corpus.

(24) bueno fui al cuartel y tal <pausa/> hice la mili que *de aquella* había que hacer la mili (SCOM_H21_039)

[16] Esta forma no es exclusiva del español de Galicia, pues se encuentra también en las variedades de influencia asturleonesa. *Cfr*., entre otros, Le Men (2002-2012: *s.v. aquel*).

(25) supuestamente hay un santo <pausa/> enterrado <pausa/> para quien crea *si cuadra* es un cerdo no lo sé (SCOM_H13_012)

(26) crisis evidentemente hay <pausa/> yo la noto <pausa/> *y más* no me quitaron <pausa/> a mi marido no le quitaron <pausa/> nada de la nómina (SCOM_M11_040)

(27) dijo no no no te lo tomes *de de carallada* (SCOM_H21_053)

En el segundo tipo de problemas, los de más calado y concernientes a la relación entre etiquetas, formas y comportamientos gramaticales para los que estas están previstas, podemos mencionar, en primer lugar, la utilización de *medio/a/os/as* concordado con un adjetivo a continuación, allí donde en español estándar se utilizaría la forma invariable *medio*[17]. Así ocurre en (28) y (29):

(28) pero después así que eran ya *medios* adolescentes (SCOM_M33_005)

(29) aunque sean así <pausa/> *medias* así <pausa/> *medias* raras (SCOM_M22_019)

En español estándar la etiqueta que correspondería, a nuestro modo de ver, a *medio* como forma invariable en esa circunstancia sería la de adverbio (W). Pero, obviamente, en español de Galicia la forma que aparece, que es variable, no puede ser etiquetada de esa manera. Aun teniendo a nuestra disposición las etiquetas previstas en el sistema para *medio* como determinante o pronombre partitivo (DP?? o PP??)[18], no creímos, por diferentes razones, que su aplicación fuera apropiada en el contexto de los ejemplos (28) y (29), por lo que decidimos la introducción de una nueva etiqueta, la de adjetivo (A??), para este ítem en tales circunstancias. Respondimos, pues, en este caso, al problema planteado en español de Galicia por una asociación no prevista por nuestro sistema de anotación para español estándar entre una palabra y un comportamiento gramatical concreto creando esa asociación y etiquetándola, y eso lo hicimos porque tal asociación solo afectaba a un ítem léxico (era, por lo tanto, económico crearla) y porque la etiqueta correspondiente y su aplicación no se prestaban a discusión o duda.

Frente a lo decidido respecto a *medio*, se optó por la solución contraria en la etiquetación de las formas verbales en *-ra* (*amara*), que en la variedad que nos ocupa asumen valores modotemporales que en español estándar están ligados a otras formas, pero que en gallego son propios de las formas en *-ra* (aparte de los valores de anterioridad al origen en subjuntivo, los de anterioridad al origen o a una referencia anterior al origen en indicativo, Rojo & Vázquez Rozas 2014). Así ocurre con la forma *comprara* en (18), o con otras formas en *-ra* en los ejemplos (30) y (31).

(30) y estábamos Laura y yo y <pausa/> además *fuera* <pausa/> *fuera* curioso porque <pausa/> <ruido tipo="chasquido boca"/> <pausa/> eran eeh mmm mogollón de ellos (SCOM_H21_039)

[17] Esta utilización de *medio* concordado con los adjetivos a los que acompaña no es, de todos modos, exclusiva del español de Galicia. Se da también en otras variedades del español (*vid.* NGLE, § 19.4k y ss.). Cabe, pues, considerar que no sea un rasgo del español de Galicia debido al contacto de lenguas o que al menos solo resulte un rasgo reforzado por ese hecho.

[18] En la posición de los signos de interrogación aparecen los caracteres referentes al género y número correspondiente.

(31) y el otro <pausa/> eh este *estudiara* <pausa_larga/> este mmm <pausa/> ¿ cómo se llama ? <pausa_larga/> de Empresariales (SCOM_H31_042)

Se decidió en estos casos que se etiquetaría siempre de acuerdo estrictamente con la forma, asociada al lugar que se asigna habitualmente en el paradigma verbal a las formas en *-ra* en español estándar europeo: pretérito de subjuntivo [19]. Es decir, adoptamos aquí una solución contraria a la que hemos descrito para *medio* y no añadimos una etiqueta adicional para una asociación de forma y comportamiento gramatical que no es propia del español estándar y que, por tanto, no teníamos prevista. Hay que tener en cuenta que estaríamos hablando en este caso de una etiqueta que habría de ser añadida para todos los verbos. Y, sobre todo, que, aunque procedente de la situación de contacto con el gallego, no dejaría de ser esta en español de Galicia una etiquetación no de formas, sino de valores de las formas verbales, tarea que podría, con toda probabilidad, dar lugar a discusión, duda, discrepancia e inconsistencia ya en la anotación manual y, desde luego, tarea muy difícil de llevar a cabo en una anotación automática.

Finalmente, hay que aludir a aspectos que solo suponen una cierta diferencia entre la consideración que merece un fenómeno en español de Galicia y la que recibe en otras variedades del español, como sucede con los casos de segunda persona de singular del pretérito de indicativo terminadas en *-s*. Estas formas, que aparecen con cierta frecuencia en los registros orales de otras variedades del español[20], surgen también en el español de Galicia quizá aquí reforzadas por la influencia del gallego, donde terminan regularmente en *-s*[21].

(32) eso nunca lo *vistes* jugar ¿no? (SCOM_H21_053)

(33) *estuvistes* pendiente de alguien <pausa/> bueno yo y todos <pausa/> más o menos toda mi familia pero <pausa/> a cada uno le afecta a su manera (SCOM_M11_040)

Con más urgencia que para la etiquetación de español, se hace entonces necesario en el español de Galicia habilitar algún recurso o estrategia para la etiquetación automática de estas formas. De momento, la etiquetación que se les atribuye en ESLORA no incluye ninguna marca adicional con respecto a las etiquetas que reciben las formas estándares sin *-s* final.

3.4. La aplicación de consulta

Para facilitar la recuperación de información que ofrece ESLORA se ha diseñado una aplicación de consulta que se organiza en varias pestañas, de las cuales las más destacables son *Información, Guía, Descargas* y *Búsquedas*, esta última el núcleo de la aplicación.

En *Información* se proporcionan tanto los datos globales del corpus (número de versión, fecha de esta, número de documentos y número de palabras ortográficas o

[19] Que no es el único posible: los verbos modales conservan usos indicativos de las formas en *-ra* en español estándar peninsular (*debiera* equivalente a *debería*, *quisiera* a *querría* y semejantes). También son indicativos los usos del tipo *el libro que publicara en 1998* (*cfr.* Rojo 2011b).

[20] Como en el caso de *medio* (*vid*. nota 17), podría dudarse de que sea un fenómeno atribuible o siquiera reforzado por la situación de contacto.

[21] En ESLORA hasta ahora se han identificado en total 28 casos de 17 formas distintas: *abaratastes, anduvistes, dijistes, empezastes, enterastes, estuvistes, fuistes, hicistes, llamastes, nacistes, oístes, pasastes, pensastes, pusistes, salistes, tomastes, vivistes.*

elementos gramaticales que constituyen el corpus), como el número de palabras ortográficas y elementos gramaticales que posee el corpus en función de cada uno de los parámetros de clasificación establecidos (grupo de edad, papel del hablante, sexo, nivel de estudios y tipo de interacción).

Por su parte, *Guía* ofrece una descripción detallada del sistema de consultas, así como la relación de marcas aplicadas en la transcripción, que se reflejan en los resultados mostrando las palabras afectadas por la marca en cuestión sobre fondo amarillo. Las marcas empleadas son las siguientes:

Alargamiento: aumento de cantidad que afecta a algún sonido de la palabra marcada.

Cita: el fragmento resaltado reproduce estilo directo.

Énfasis: señala casos de pronunciación especialmente acentuada.

Ficticio: el nombre ha sido cambiado para preservar el anonimato de los hablantes.

Lengua [nombre=xx]: el fragmento está en una lengua diferente al español; xx puede ser *gl*: gallego, *en*: inglés, *pt*: portugués, *it*: italiano, *fr*: francés, *el*: griego.

Palabra_cortada: el segmento marcado representa un fragmento de una palabra.

Risa: marca un segmento en que se ríe un hablante.

Sic: señala errores de dicción para que no se interpreten como errores de transcripción.

Sigla: indica que una cierta forma es una sigla.

Dado que el corpus se enriqueció con su etiquetado morfosintáctico automático, como ya hemos visto en el apdo. 3.3, en *Guía* se ofrece asimismo el etiquetario que se emplea en una descripción que, organizada por clases de palabra, muestra cada etiqueta, su significado y un ejemplo de uso.

A su vez, la sección *Descargas* facilita la descarga del corpus en formato textual y proporciona un formulario para, previa justificación, disponer del corpus en formato ya etiquetado, los audios correspondientes o la información sociolingüística de los hablantes.

El sistema de consultas, al que se accede a través de la pestaña *Búsquedas,* constituye el núcleo de la aplicación y presenta el aspecto que se observa en la figura 1:

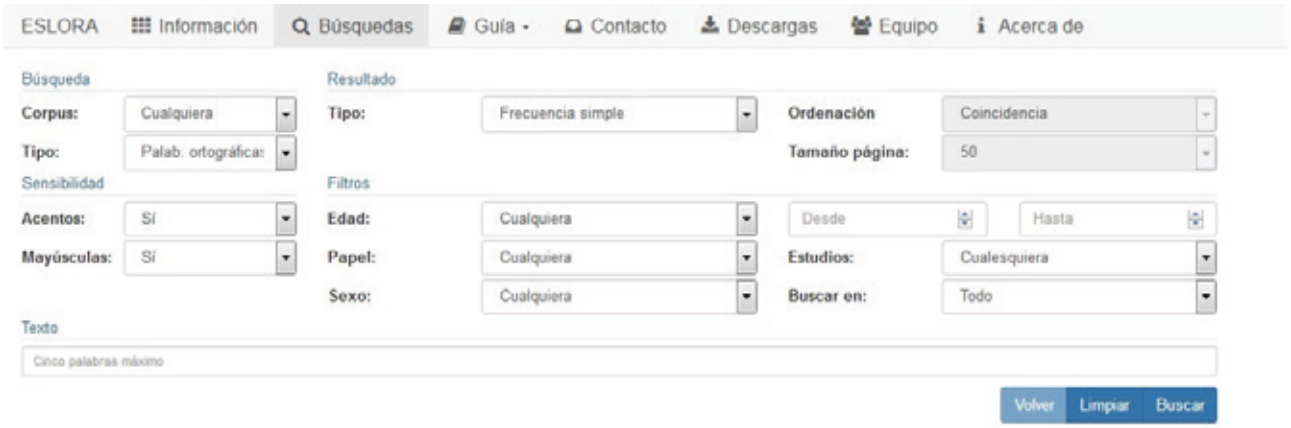


FIGURA 1. Pantalla inicial de captación de datos

Grosso modo, el sistema de consultas permite:

i. realizar consultas combinando variables sociales con variables léxicas y gramaticales;

ii. acceder a los fragmentos de audio que corresponden al resultado de la consulta;

iii. descargar el resultado de la consulta en formato TSV (*Tab Separated Values*).

Las consultas pueden realizarse sobre la totalidad del corpus, opción por defecto que muestra la figura 1, o bien sobre uno de los tipos específicos de interacción incluidos bajo el ítem *Corpus*: entrevistas o conversaciones. Además, es posible referir las búsquedas a la totalidad del texto, que es la opción habitual, o bien, mediante la selección de una de las alternativas presentes en el bloque *Buscar en*, centrarlas en fragmentos que han recibido cierta etiquetación (entre risas, con marca de énfasis o alargamiento, reproducción de una cita o identificación de siglas).

Asimismo, el sistema ofrece la posibilidad de combinar libremente valores de las diferentes variables sociales empleadas en la clasificación del corpus, de modo que el usuario crea subcorpus virtuales a la medida. Los parámetros clasificatorios de los que dispone y sus valores son los siguientes:

i. Edad del hablante: Cualquiera, 19-34, 35-54, >54, Desconocida.

ii. Papel o rol en la comunicación: Cualquiera, Audiencia, Entrevistador, Informante.

iii. Sexo: Cualquiera, Hombre, Mujer.

iv. Nivel de estudios: Cualesquiera, Universitarios, Medios, Primarios, Desconocidos.

A su vez, la búsqueda textual se organiza en el ítem *Tipo* del bloque *Búsqueda* por palabras ortográficas (*ir, del, yéndome*) o por elementos gramaticales (*voy, de* –incluyendo los casos de la contracción *del*–, *yendo* –incluyendo los casos del tipo *yéndome*, *yéndonos*–, etc.). La primera modalidad restringe las consultas a una cadena

formalmente coincidente con la grafía, lo que resulta de utilidad en consultas léxicas pero limita en grado sumo las opciones de búsqueda cuando interesa trabajar con información gramatical. La segunda modalidad permite dar un salto cualitativo en la recuperación de información y posterior análisis gramatical al poder introducir en los parámetros de consulta información de carácter gramatical, no solo ya la forma léxica, sino también lemas (todas las formas del verbo *ir*), clases de palabras (verbo, por ejemplo) o valores de las subcategorías gramaticales aplicables en cada caso (persona, por ejemplo, en el caso de los posesivos y verbos).

La importancia que adquieren las búsquedas gramaticales en el análisis de la lengua queda patente en el modo de obtener los datos para estudiar el influjo del gallego en la perífrasis *ir + infinitivo*. La consulta por palabras ortográficas es inútil en este caso debido, por un lado, a la variación e irregularidad del verbo auxiliar (*iba*, *íbamos*, *ir*, *vayamos*, *yendo*, *yéndome*, *váyase*, *irte*, etc.) y, por otro, debido a la ingente relación de infinitivos que pueden aparecer como auxiliado (*hacer, jugar, trabajar, ver*...). Sin embargo, el etiquetado morfológico automático del corpus y la implementación en la aplicación de consultas de las variables relativas a elemento gramatical, etiqueta y lema, combinables entre sí, y combinables asimismo con las variables sociales, no lo olvidemos, junto con la posibilidad de trabajar con hasta cinco elementos sucesivos, hacen que baste realizar la búsqueda que muestra la figura siguiente: lema *ir* en la línea correspondiente al primer elemento y etiqueta *VN*,* equivalente a verbo en infinitivo para el segundo elemento concurrente.

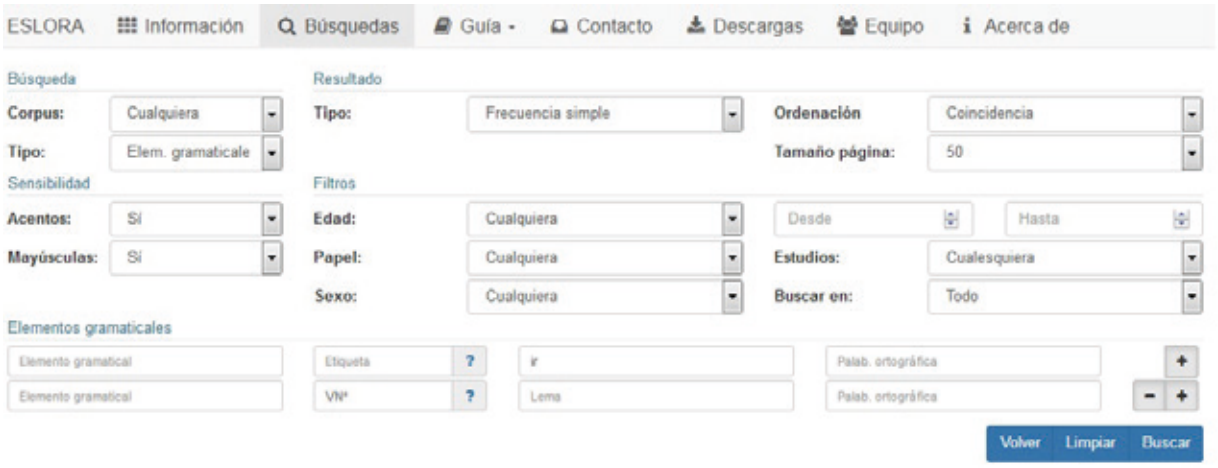


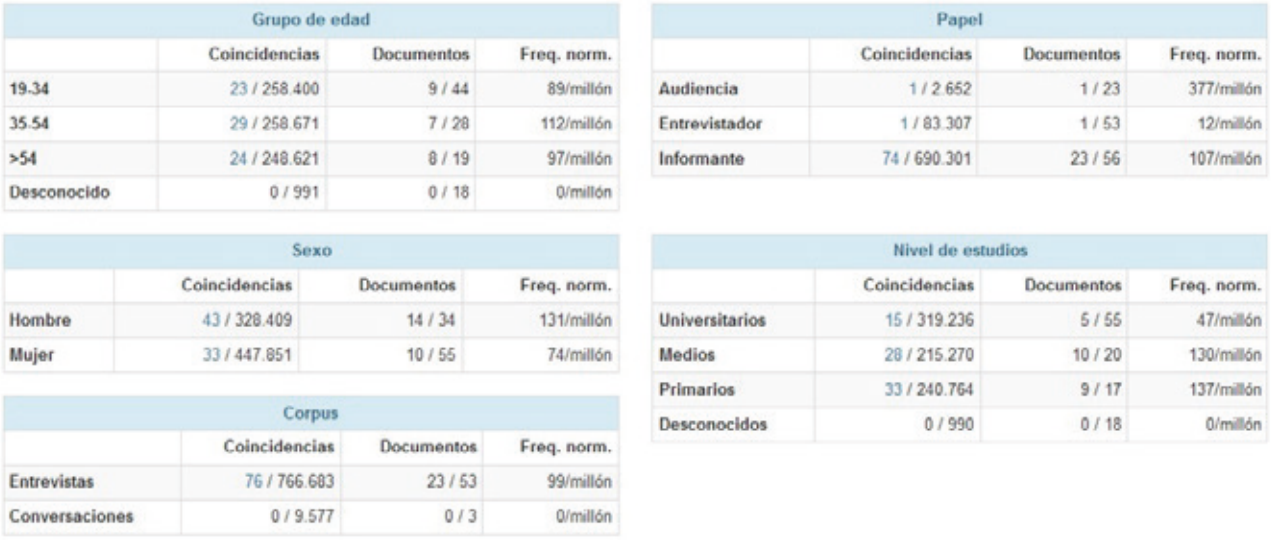
FIGURA 2. Pantalla de captación de datos gramaticales

En cuestión de segundos, modificando la opción *Tipo* del bloque *Resultado*, se obtienen:

1) los datos relativos a su frecuencia simple (figura 3):

FIGURA 3. Información sobre la frecuencia simple

2) las frecuencias completas según los distintos parámetros de clasificación del corpus (figura 4):



Grupo de edad	Coincidencias	Documentos	Freq. norm.
19.34	23 / 258.400	9 / 44	89/millón
35.54	29 / 258.671	7 / 28	112/millón
>54	24 / 248.621	8 / 19	97/millón
Desconocido	0 / 991	0 / 18	0/millón

Papel	Coincidencias	Documentos	Freq. norm.
Audiencia	1 / 2.652	1 / 23	377/millón
Entrevistador	1 / 83.307	1 / 53	12/millón
Informante	74 / 690.301	23 / 56	107/millón

Sexo	Coincidencias	Documentos	Freq. norm.
Hombre	43 / 328.409	14 / 34	131/millón
Mujer	33 / 447.851	10 / 55	74/millón

Nivel de estudios	Coincidencias	Documentos	Freq. norm.
Universitarios	15 / 319.236	5 / 55	47/millón
Medios	28 / 215.270	10 / 20	130/millón
Primarios	33 / 240.764	9 / 17	137/millón
Desconocidos	0 / 990	0 / 18	0/millón

Corpus	Coincidencias	Documentos	Freq. norm.
Entrevistas	76 / 766.683	23 / 53	99/millón
Conversaciones	0 / 9.577	0 / 3	0/millón

FIGURA 4. Datos de las frecuencias completas

3) las muestras con todos los casos de la perífrasis (figura 5) en formato KWIC (*Key Word in Context*):

FIGURA 5. Pantalla con algunas de las muestras finales

Además, y no menos importante dado que se trata de un corpus oral, el usuario puede acceder al audio del segmento que se corresponde con el objeto de la búsqueda y comprobar que no está presente la preposición *a*.

Los resultados de las búsquedas incluyen, también, la posibilidad de descarga de las concordancias o la ampliación del contexto con la representación de las etiquetas morfosintácticas, como muestra la figura 6:

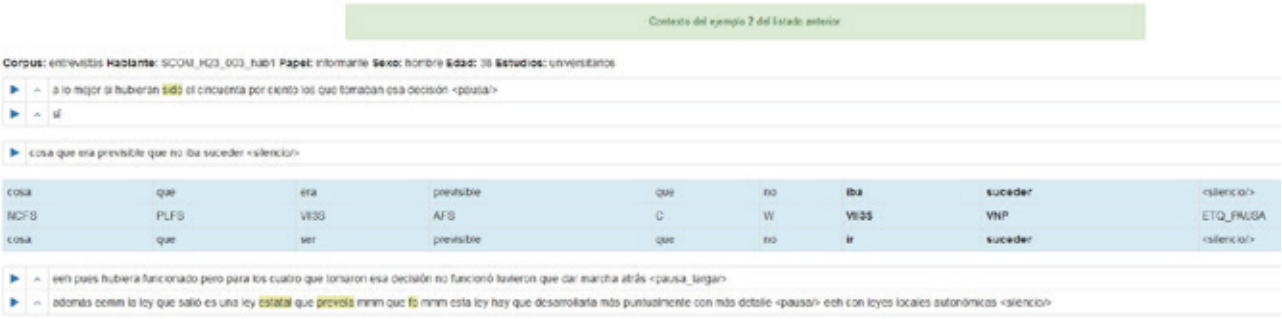


FIGURA 6. Visualización de anotación morfosintáctica

Conscientes de la dificultad que entraña conocer el complejo sistema de etiquetación, en la herramienta de consulta de ESLORA hemos habilitado la introducción de la etiqueta a través de un menú intuitivo que va guiando al usuario en la formulación de la búsqueda gramatical. De este modo, como se muestra en la imagen siguiente, el usuario formula su consulta apoyándose en un menú desplegable que a modo de cascada va mostrando los valores posibles, organizados en un primer nivel en clases de palabras y posteriormente en los valores disponibles para las diferentes categorías gramaticales asociadas a cada clase de palabra, siempre en función de las elecciones previamente realizadas:

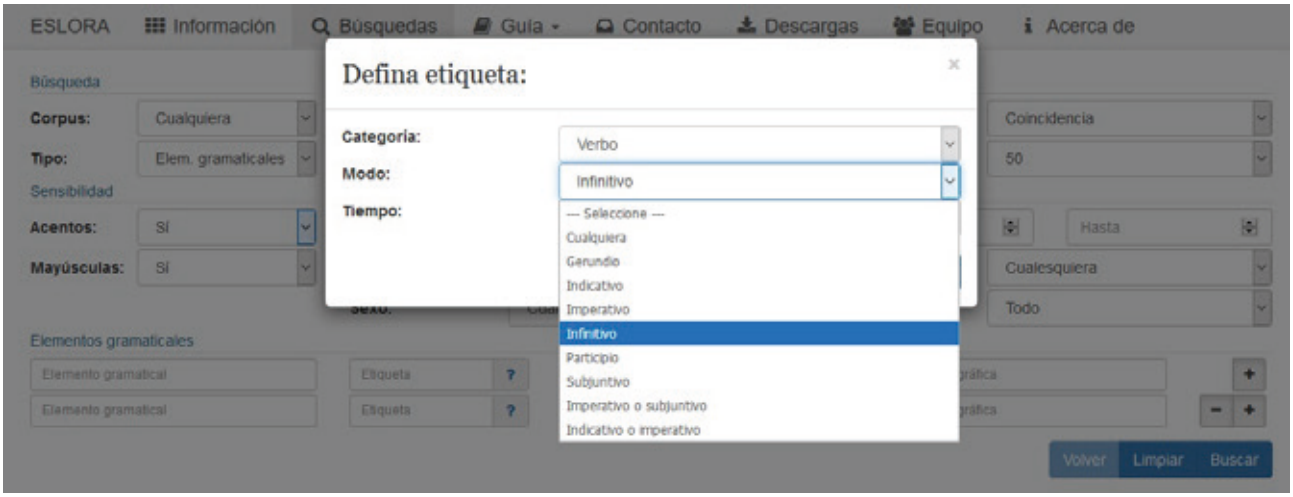


FIGURA 7. Ejemplo de consulta por etiqueta

Asimismo, en la observación de los resultados, ya sea en las muestras ya sea a través del contexto de estas, siempre que esté explícita una etiqueta, la herramienta desarrolla dicha etiqueta al situar encima el puntero del ratón pues, al igual que sucede con las marcas de codificación, emerge un cuadro de texto que la desglosa proporcionando la caracterización morfosintáctica de la unidad en cuestión en un lenguaje comprensible para cualquier usuario.

Esta doble actuación, menú desplegable para formular las consultas y cuadro de texto emergente con el desarrollo de la etiqueta en las concordancias o en el análisis completo de la secuencia visible desde el contexto, permite que un usuario no

familiarizado con el sistema de etiquetación aplicado pueda realizar consultas a través de la modalidad elementos gramaticales y comprender fácilmente los resultados.

La posibilidad de utilizar directamente rasgos gramaticales, independientes de las formas que los soportan, ilustra la potencia de la recuperación de información y su utilidad para el trabajo gramatical, pero el sistema de explotación de ESLORA cuenta además con otras opciones que permiten refinar todavía más las búsquedas. Existe la opción de realizar consultas acerca de la copresencia tanto de palabras ortográficas como de elementos gramaticales que no están limitadas a la aparición de una de ellas inmediatamente antes o después de la otra, sino que admiten la existencia de varios elementos (hasta diez) entre ambos. Gracias a esta posibilidad, es posible recuperar, por ejemplo, todos los casos del lema *olvidar* –información presente en el Elemento 1– que concurran con la preposición *de* –información aportada para el Elemento 2– a una distancia máxima de 4 o menos elementos –hay que indicar un valor de 1 a 10, pudiendo ser ese valor distancia exacta o inferior a la expresada–, con independencia de que existan casos de enclisis pronominal o interpolación de elementos varios.

De otra parte, el sistema cuenta con la posibilidad de emplear en los campos de búsqueda los comodines habituales: el cierre de interrogación (?) sustituye a un carácter en la posición que ocupe en la cadena y el asterisco (*) sustituye a ningún carácter, uno o varios en la posición que ocupe en la cadena, de forma que *p?so* devolverá los casos que contengan la palabra *paso*, *peso*, *piso*, *poso y puso*, mientras que **miento* devolverá todas las secuencias que incluyan *-miento*, incluida la primera persona del singular del presente de indicativo de *mentir*.

Asimismo, el usuario puede recurrir en la formulación de la consulta a dos operadores booleanos: el cierre de la admiración (!) equivalente a NO, y la barra vertical (|) equivalente a O, de modo que, como muestra la figura 8, es posible recuperar todas las formas verbales –etiqueta *V** en el campo *Etiqueta*– pertenecientes a la segunda y tercera conjugaciones salvo las de los verbos *ser* e *ir* –en el campo *Lema *er|*ir!ser!ir*–.

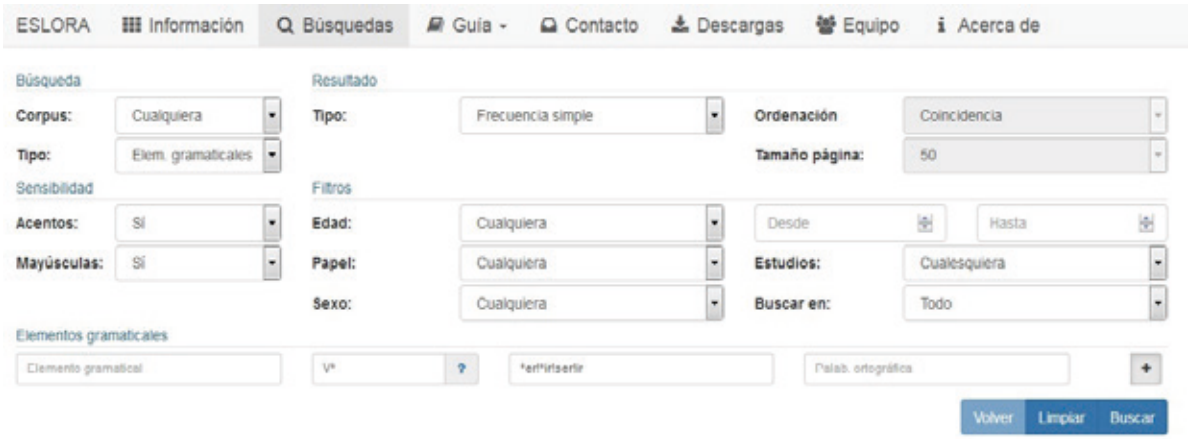


FIGURA 8. Pantalla de captación de datos con comodines y operadores booleanos

Naturalmente, existe la posibilidad de combinar en una misma consulta comodines, operadores booleanos, sensibilidad a acentos o mayúsculas y variables sociales con los distintos tipos de modalidad de búsqueda, por palabras ortográficas o elementos gramaticales, bien sucesivos bien discontinuos, lo que convierte al corpus ESLORA en una herramienta muy útil para obtener datos del español de Galicia.

4. ASPECTOS CUANTITATIVOS

Se ha aludido en numerosas ocasiones al escaso interés que los estudios acerca de la frecuencia de elementos o fenómenos lingüísticos ha despertado en la lingüística española. Si bien es cierto que las frecuencias léxicas han sido las menos desatendidas, también lo es que son muy pocos los estudios referidos a la lengua oral[22]. El más destacado de ellos es el que Ávila Muñoz (1999) dedicó al análisis de la lengua hablada en Málaga. Está planteado en el estilo tradicional de los diccionarios de frecuencias, de modo que resulta sencillo examinar lo referente a un lema y a las formas integradas en él, pero no está pensado para estudios que manejen elementos más abstractos como, por ejemplo, las clases de palabras o las subcategorías verbales.

El hecho de que ESLORA haya sido concebido desde el principio como un corpus de lengua oral y las transcripciones se hayan lematizado y analizado morfosintácticamente permite enfocar el análisis de las frecuencias, no solo las léxicas, desde una óptica más amplia y ambiciosa. Las 53 entrevistas que hemos podido procesar arrojan un total de 669 502 elementos lingüísticos. Nótese que aquí no se habla de palabras ortográficas (que, aunque comprensible, resultaría un tanto chocante cuando se trata de transcripciones de lengua oral). En esa cifra se incluyen también los escasos signos ortográficos que se han utilizado en la transcripción (*cfr*. *supra* apdo. 3.2), pero no entran en cambio, indicaciones de pausa, silencio, ruido, etc.[23]. Si se eliminan los signos ortográficos, operación precisa para poder hacer comparaciones con textos escritos, la cifra anterior se reduce a 647 574 elementos lingüísticos. En un segundo recorte, podemos eliminar las 9699 apariciones de nombres propios (el 1,5 % del total de elementos sin signos ortográficos). Por tanto, deducidos estos dos bloques, que no resultan de interés para el análisis de las frecuencias léxicas, queda un total de 637 875 elementos. Ciertamente, no son muchos para los volúmenes a los que estamos acostumbrados en los corpus, pero superan a los que han servido para la confección de obras de gran utilidad como el FDSW y también al construido sobre el español hablado en Málaga, que tiene 523 639 palabras (Ávila Muñoz 1999: 93).

Las cifras anteriores se refieren al total de apariciones de los elementos lingüísticos, esto es, a los *tokens* contenidos en las entrevistas. Los elementos distintos, es decir, los *types*, son 25 891, que se reducen a 23 074 una vez descontados los correspondientes a signos ortográficos y nombres propios. Para posibles comparaciones con los datos procedentes de otros corpus, es preciso tener en cuenta que lo que aquí consideramos elemento único (*type*) es más complejo y abstracto que lo que se maneja cuando se diferencia entre *types* y *tokens* trabajando con formas ortográficas. En efecto, el *type* en nuestro caso implica también una cierta etiqueta morfosintáctica y la pertenencia a un lema determinado. Por tanto, la palabra ortográfica *vacío* corresponde a tres elementos distintos (la forma singular del sustantivo *vacío*, la forma masculina singular del adjetivo *vacío* y la forma de primera persona del singular del presente de indicativo del verbo *vaciar*). Por otro lado, las contracciones *al* y *del* son sistemáticamente analizadas en dos elementos cada una de ellas y, en sentido contrario, las expresiones multipalabra del tipo *sin embargo*, *a pesar de*, lo mismo que los nombres propios formados por varias palabras ortográficas, se agrupan en un elemento único.

[22] Como es previsible dada la época en que fue confeccionado, no hay análisis de textos orales en el FDSW (Juilland & Chang 1964). En el diccionario de frecuencias de Davies (2006) aparecen indicaciones acerca de si una palabra muestra una predisposición especial a aparecer en textos orales, pero hay que tener en cuenta que el corpus sobre el que se hacen los cálculos es el primero de los elaborados por Davies para el español (el formado por cien millones de formas), donde la consideración de texto oral se aplica a, por ejemplo, entrevistas publicadas en periódicos (*cfr*. Rojo 2010).

[23] Eso explica la discrepancia entre estas cifras y las que se dan en la página de la aplicación o el apdo. 2.

La relación entre el número total de elementos (*tokens*) y el número de elementos distintos (*types*) permite calcular la TTR (*type-token ratio*), que ha sido utilizada con cierta frecuencia como un índice de la riqueza léxica de un texto o un conjunto de textos. Aunque la fórmula más sencilla admite bastantes mejoras (*cfr*. Torruella & Capsada 2013, 2017), es siempre una medida poco refinada y, lo que es más grave, muy sensible al tamaño del corpus, de modo que desciende de forma muy marcada con el aumento del volumen. El núcleo básico de los elementos (esto es, sin signos ortográficos ni nombres propios) del corpus ESLORA arroja una TTR de 0,036. Para poder situar estas cifras con relación a las que pueden corresponder a textos de otro tipo hemos extraído del *Corpus del español del siglo XXI* (CORPES) dos muestras de tamaño similar al que tiene ESLORA en la actualidad: una de ellas está constituida por noticias de prensa publicadas en España en 2002 y la otra está formada por libros (tanto de ficción como de otras clases) del mismo país y año. Los datos procedentes de los tres subconjuntos aparecen en la tabla 1:

	ESLORA		Muestra CORPES prensa		Muestra CORPES libros	
	Total elementos (tokens)	Total elementos distintos (types)	Total elementos (tokens)	Total elementos distintos (types)	Total elementos (tokens)	Total elementos distintos (types)
Elementos identificados	669 502	25 891	667 805	58 420	807 174	61 237
Signos Ortográficos	21 928	7	77 804	35	102 920	23
Nombres propios	9699	2810	24 323	10 020	20 520	56
Elementos sin signos ortográficos ni nombres propios	637 875	23 074	565 678	48 365	683 684	55 606
TTR	0,036		0,085		0,081	
1 elemento nuevo cada	27,645		11,696		12,295	

TABLA 1. Frecuencias de elementos y elementos distintos en tres subcorpus diferentes. Fuentes: ESLORA y CORPES. Elaboración propia

Puede observarse que la «riqueza léxica» en los textos orales es considerablemente menor que la que se aprecia en los textos escritos. Cabe pensar que una agrupación de noticias periodísticas, de escasa extensión, con los cambios de tema que suponen, trae consigo una variedad temática que forzosamente tiene que reflejarse en variedad léxica. Sin embargo, el resultado del análisis de textos procedentes de libros, no solo de ficción, muestra que las cifras son similares a las obtenidas en textos de prensa y ambas muy superiores a las que se observan en textos orales: con una ilustración clara, en corpus de tamaño similar, en los textos orales aparece uno nuevo cada 27,6 elementos, mientras que en los textos escritos es suficiente con unos 12. Es una diferencia muy evidente que, por otro lado, confirma lo esperable.

Una medida mucho más adecuada de la riqueza léxica pasa por utilizar las ventajas que proporciona la anotación de ESLORA. Dado que la anotación se ha hecho de forma

automática, hay que tener en cuenta siempre la posibilidad de que la información manejable contenga un cierto número de errores, además de elementos no identificados. Incluyendo en la caracterización del lema la pertenencia a una determinada clase de palabras, la versión actual de ESLORA (la 1.2.2) está formada por un total de 11 147 «lemas» distintos, que se reducen a 9368 si eliminamos los signos ortográficos, los nombres propios y los no identificados[24]. Esos 9368 lemas suman en total 639 160 elementos, lo cual arroja una frecuencia media de 67,7 elementos por lema, cifra bastante elevada. Para poder valorar adecuadamente estas cifras, hemos calculado también las correspondientes a las dos muestras de tamaño similar extraídas del CORPES. Los resultados figuran en la tabla 2.

	ESLORA		Muestra CORPES prensa		Muestra CORPES libros	
	Número de lemas distintos	Frecuencia total	Número de lemas distintos	Frecuencia total	Número de lemas distintos	Frecuencia total
Total «lemas»	11 747	669 502	31 142	667 805	29 913	807 174
Signos ortográficos	7	21 928	25	77 804	28	102 970
Nombres propios	2366	9699	9842	24 323	5623	20 520
No identificados	6	3715				
Total lemas	9368	634 160	21 275	565 678	24 262	683 684
Frec./lemas	67,7		26,6		28,2	

TABLA 2. Número de lemas y frecuencia total en ESLORA y dos subcorpus del CORPES.
Fuentes: ESLORA y CORPES. Elaboración propia

Como era de esperar, los textos escritos contienen un número de lemas considerablemente superior al que encontramos en ESLORA. La muestra de prensa, de menor tamaño que el corpus oral, tiene, a pesar de ello, más del doble de lemas diferentes. La divergencia entre los tres conjuntos textuales se hace muy evidente en la frecuencia media: 67,7 apariciones por lema en ESLORA y 26,6 y 28,2 en los dos subcorpus escritos.

La configuración resumida en la ley de Zipf se cumple también en los textos orales: los 25 elementos más frecuentes suponen el 38,97 % del total, mientras que los 25 lemas más frecuentes alcanzan el 44,53 %. Por otra parte, el porcentaje de hápax (es decir, elementos con frecuencia igual a 1) es del 53,43 % del total de elementos distintos y el 30,70 % de lemas. La tabla 3 muestra estos datos en comparación con los obtenidos de las dos muestras escritas que estamos usando para la comparación[25].

[24] Entrecomillamos «lemas» para señalar la peculiaridad de aplicar este término a signos ortográficos y nombres propios. Por otra parte, dado que aquí se usa lema para la unión de un término y una clase de palabras, se comprenderá que, como se muestra en la tabla 2, los lemas no identificados sean solo seis, tantos como clases de palabras distintas son asignadas a lemas no identificados. Los elementos distintos (elemento y etiqueta morfosintáctica) no identificados en la versión actual son 1694, con una frecuencia total conjunta de 3715, indicada en la tabla 2.

[25] Los porcentajes han sido obtenidos sobre el número de elementos o lemas ya sin signos ortográficos ni nombres propios ni unidades no identificadas.

	Totales y % sobre el total de ESLORA	Totales y % sobre el total del subconjunto de CORPES (prensa)	Totales y % sobre el total del subconjunto de CORPES (libros)
25 elementos lingüísticos más frecuentes	248 588 (38,97 %)	217 829 (38,48 %)	252 110 (36,84 %)
25 lemas más frecuentes	301 214 (44,99 %)	252 099 (44,53 %)	301 908 (44,02 %)
Elementos con frecuencia = 1 (hápax)	12 330 (53,43 %)	29 323 (60,54 %)	27 829 (49,96 %)
Lemas con frecuencia = 1	9440 (40,91)	14 872 (30,70 %)	9398 (31,41 %)

TABLA 3. Totales y porcentajes de elementos lingüísticos y lemas (sin signos de puntuación ni nombres propios) en ESLORA y en una muestra del CORPES escrito. Elaboración propia

Los porcentajes que suponen sobre los totales respectivos los 25 lemas o elementos más frecuentes son muy próximos entre sí y también a los que se obtienen en corpus de tamaño mucho mayor. En cuanto a los elementos con frecuencia igual a 1 (hápax), los tres porcentajes difieren, pero no demasiado, y el hecho de que resulten más altos que los obtenidos para muestras de mayor tamaño puede explicarse por el escaso volumen de los corpus que estamos examinando aquí. Por fin, el porcentaje de lemas con frecuencia igual a 1 es congruente con lo que se observa en corpus más grandes en las muestras escritas, mientras que resulta más alto en el corpus oral (*cfr*. Rojo 2008 y 2017 para datos de este tipo correspondientes a CREA y CORPES).

Más allá de estos resultados, todavía bastante toscos, relacionados con las frecuencias generales, las cuestiones realmente importantes se refieren a la posible existencia de diferencias en la configuración gramatical de los textos orales del tipo al que corresponden las entrevistas semidirigidas en comparación con (ciertos tipos de) textos escritos. Como primera aproximación, forzosamente muy superficial y tentativa, a un problema que requiere análisis mucho más profundos, hemos examinado el peso de las clases de palabras de contenido más léxico en ESLORA y en dos muestras procedentes del CORPES, tanto en el inventario como en los textos[26]. Los porcentajes del número de adjetivos, adverbios, sustantivos comunes y verbos sobre el total de los lemas (de nuevo sin signos ortográficos ni nombres propios ni lemas no identificados) es el que muestra la tabla 4. Llama la atención el hecho de que los textos orales y los escritos procedentes de libros están muy próximos en cuanto a lo que la suma de estas cuatro clases de palabras supone con respecto al total del inventario de lemas, mientras que los textos de prensa se quedan ocho puntos porcentuales por debajo de los otros dos. Con respecto a la distribución de las cuatro clases de palabras, lo más destacable es, sin duda, que los

[26] Se trata de la distinción establecida en Rojo (2011a) entre frecuencia de inventario y frecuencia de uso, que es reelaboración de la propuesta por Bybee (2007), entre otros, entre *type frequency* y *token frequency*, pero en una formulación bastante más general. La frecuencia de inventario es el número de elementos distintos de un cierto tipo (en nuestro caso, por ejemplo, adjetivos o verbos) que se localizan en un determinado corpus, es decir, en su lemario. La frecuencia de uso, en cambio, es la que arrojan *todas* las apariciones de un determinado tipo de elemento en un corpus (en nuestro caso, por ejemplo, el número total de adjetivos o de verbos, contando todas y cada una de las apariciones de elementos de la clase correspondiente en los textos que integran el corpus). Los artículos, las preposiciones o las conjunciones, por ejemplo, pesan muy poco en el inventario de los elementos de una lengua, pero, en cambio, tienen una frecuencia muy alta si nos referimos al total de sus apariciones en los textos.

textos orales muestran un menor porcentaje de adjetivos, compensado con una diferencia de signo contrario en los verbos.

	Inventario				Textos		
	ESLORA	Muestra CORPES prensa	Muestra CORPES libros		ESLORA	Muestra CORPES prensa	Muestra CORPES libros
Adjetivos	16,35	21,02	22,04	Adjetivos	1,78	7,65	6,87
Adverbios	4,23	3,25	3,74	Adverbios	14,48	4,63	5,35
Sustantivos comunes	55,63	52,23	56,65	Sust. com.	12,06	25,10	22,61
Verbos	19,81	12,08	14,26	Verbos	20,64	13,53	15,22
Totales	96,02	88,58	96,70	Totales	48,95	50,91	50,05

TABLA 4. Porcentajes de cuatro clases de palabras sobre el total de los lemas y el total de los textos en ESLORA y dos subcorpus del CORPES. Fuentes: ESLORA y CORPES. Elaboración propia

Más diferencias muestra la frecuencia en los textos. Aquí las dos muestras escritas resultan mucho más próximas entre sí y las entrevistas de ESLORA se diferencian de ambas con bastante claridad. Todas las clases examinadas presentan diferencias importantes en el corpus oral con respecto a los textos escritos: los porcentajes de adjetivos y sustantivos comunes son muy inferiores en los textos orales, mientras que los correspondientes a adverbios y verbos resultan muy superiores. No se puede descartar la posibilidad de que una parte de estas diferencias procedan de los diferentes programas de anotación automática que se han aplicado (uno en el caso de ESLORA y otro en las dos muestras del CORPES), pero no son estas clases de lemas las que pueden variar en mayor medida y, por otro lado, las diferencias son de una entidad considerable. Todo indica, por consiguiente, que aquí hay una diferencia importante que será necesario investigar con mayor profundidad.

Veamos, por último, algunos datos procedentes del cruce del número de lemas distintos con los parámetros utilizados habitualmente en los estudios sociolingüísticos. Con las eliminaciones habituales y sin tener en cuenta tampoco las intervenciones de los encuestadores, en las entrevistas con hombres como informantes documentamos 6371 lemas diferentes (sobre 266 035 elementos totales) y 5879 lemas diferentes en las protagonizadas por mujeres (sobre un total de 300 239 elementos). Con estas cifras, la «riqueza léxica» parece mayor en los hombres que en las mujeres. En la tabla 5 pueden verse estos datos y también los correspondientes a los otros dos parámetros.

Sexo	Lemas distintos	Nivel educativo	Lemas distintos	Edad	Lemas distintos
Hombre	6371	Bajo	4709	19-34	4688
Mujer	5879	Medio	5032	35-54	5042
		Alto	5300	>54	5300

TABLA 5. Número de lemas distintos en las entrevistas de ESLORA en relación con diferentes parámetros sociolingüísticos. Fuente: ESLORA. Elaboración propia

El número de lemas diferentes documentados en las entrevistas aumenta con el nivel educativo y también con la edad, pero las diferencias no parecen realmente significativas.

5. CONCLUSIONES

En este trabajo se ha presentado el corpus ESLORA, un corpus de lengua oral del español de Galicia. En la introducción hemos reflexionado sobre la necesidad y la utilidad de un corpus de estas características en especial para los estudios de dialectología. ESLORA viene, en efecto, a sumarse a otros corpus que documentan las variedades del español propias de las zonas en las que hay lenguas en contacto. El artículo revisa en primer lugar la composición del corpus desde el punto de vista de los tipos de interacciones incluidas y de las características sociolingüísticas de los informantes. Se explica a continuación el proceso de recogida (grabación) y preparación (transcripción, alineación, anonimización y codificación) de los datos, así como el sistema de anotación morfosintáctica utilizado en ellos. Todo ello desde la perspectiva de las peculiaridades o problemas específicos que presenta el desarrollo de un corpus oral para documentar una variedad de lengua en situación de contacto de lenguas. Se ha descrito también la aplicación de consulta en línea que pone el corpus a disposición de los investigadores y se ha hecho un estudio cuantitativo de frecuencias léxicas (*types* en relación con *tokens*, lemas y clases de palabras) en ESLORA en comparación con corpus de lengua escrita.

Actualmente continuamos trabajando en la introducción de conversaciones, así como en la mejora de algunos aspectos de la codificación y la anotación morfosintáctica. Próximamente iniciaremos también la anotación sintáctica del corpus, que pretendemos facilitar en dos versiones: constitutiva y dependencial, de acuerdo con las directrices del proyecto Universal Dependencies.[27] Esto por lo que respecta al enriquecimiento del corpus. Por lo que toca a su explotación, aparte de los estudios para los que esperamos que sea útil a otros usuarios, es nuestra intención trabajar en distintos aspectos, entre ellos, la elaboración de diccionarios de frecuencias léxicas y gramaticales de lengua oral, el estudio de unidades, construcciones y formas verbales en lengua oral y la alternancia de código español-gallego. Por último, también está abierta una línea de trabajo para utilizar el corpus en ELE, conectando con la elaboración de diccionarios de frecuencias, para la investigación en secuenciación de contenidos de aprendizaje de destrezas orales y también desde el punto de vista de la elaboración de materiales para ese aprendizaje.

RECURSOS ELECTRÓNICOS MENCIONADOS

BNC: *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/>

CORPES: *Corpus del español del siglo XXI*. Real Academia Española <http://rae.es/recursos/banco-de-datos/corpes-xxi>

COSER: *Corpus Oral y Sonoro del Español Rural*. <http://www.corpusrural.es/>

ESLORA: *Corpus para el estudio del español oral* <http://eslora.usc.es>, versión 1.2.2 de noviembre de 2018, ISSN: 2444-1430.

[27] URL: https://universaldependencies.org/.

PRESEEA: *Proyecto para el Estudio Sociolingüístico del Español de España y América*. <http://preseea.linguas.net/>

XIADA: Etiquetador/Lematizador do Galego Actual. Centro Ramón Piñeiro para a investigación en humanidades. <http://corpus.cirp.gal/xiada>

REFERENCIAS BIBLIOGRÁFICAS

ACÍN VILLA, E. (1996): "Galleguismos en la prensa gallega escrita en castellano", *in* M. Casado Velarde *et alii* (eds.): *Scripta Philologica in memoriam Manuel Taboada Cid*. A Coruña: Universidad de A Coruña, vol. 1, pp. 267-277.

AGHA, A. (2007): *Language and social relations*. New York: Cambridge University Press. Disponible en https://epdf.tips/language-and-social-relations-studies-in-the-social-and-cultural-foundations-of-.html

ÁVILA MUÑOZ, A. (1999): *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

BARCALA, M, E. DOMÍNGUEZ, A. FERNÁNDEZ, R. RIVAS, Mª P. SANTALLA, V. VÁZQUEZ & R. VILLAPOL (2018): "El corpus ESLORA de español oral: diseño, desarrollo y explotación", *CHIMERA. Romance Corpora and Linguistic Studies* 5/2, pp. 217-237. https://doi.org/10.15366/chimera2018.5.2.003

BEAL, J. (2005): "Dialect representation in texts", in *The Encyclopedia of Language and Linguistics*. Amsterdam / London: Elsevier, 2.ª ed., pp. 531-538. https://doi.org/10.1016/B0-08-044854-2/00504-6

BRIZ, A. & Grupo Val.Es.Co (2002): *Corpus de conversaciones coloquiales*. Madrid: Arco Libros.

BUCHOLTZ, M. (2000): "The politics of transcription", *Journal of Pragmatics* 32, pp. 1439-65. https://doi.org/10.1016/S0378-2166(99)00094-6

BYBEE, J. (2007): *Frequency of use and the organization of language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

CAMUS, B. & S. GÓMEZ SEIBANE (2015): "La diversidad del español en Álava: Sistemas pronominales a partir de las encuestas del COSER", *Revista de Filología Española* XCV, pp. 279-206. https://doi.org/10.3989/rfe.2015.11

DAVIES, M. (2006): *A frequency dictionary of Spanish. Core vocabulary for learners*. Nueva York: Routledge.

DE BENITO, C. (2015): *Las construcciones con «se» desde una perspectiva variacionista y dialectal*. Tesis doctoral. Universidad Autónoma de Madrid.

DOMÍNGUEZ NOYA., E. Mª, F. Mario BARCALA RODRÍGUEZ & M. Á. MOLINERO (2009): "Avaliación dun etiquetador automático estatístico para o galego actual: Xiada", *Cadernos de Lingua* 30-31, pp. 151-193.

DU BOIS, J. W., W. L. CHAFE, C. MEYER, S. A. THOMPSON, R. ENGLEBRETSON & N. MARTEY (2000-2005): *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

DURANTI, A. (2006): "Transcripts, like shadows on a wall", *Mind, Culture, and Activity* 13/4, pp. 301-310. https://doi.org/10.1207/s15327884mca1304_3

EAGLES (1996): *Recommendations for the morphosyntactic annotation of corpora*. EAGLES Document EAG-TCWG-MAC/R. http://www.ilc.cnr.it/EAGLES96/browse.htlm [última consulta: 20/6/2018].

EDWARDS, J. A. (2001): "The transcription of discourse", in D. Tannen, D. Schiffrin, & H. E. Hamilton (eds.): *Handbook of discourse analysis*. Oxford: Blackwell, pp. 321-348.

FERNÁNDEZ JUNCAL, Mª C. (2005): *Corpus de habla culta de Salamanca* (CHCS). Burgos: Fundación Instituto Castellano y Leonés de la Lengua.

FERNÁNDEZ RODRÍGUEZ, M. (dir.) (no publicado): "Formación de un corpus de lengua hablada en la ciudad de A Coruña". Proyecto financiado por la Universidad de A Coruña, la Xunta de Galicia (XUGA10402A90) y la DGICYT (PB90-0324).

FERNÁNDEZ-ORDÓÑEZ, I. (2007): "El 'neutro de materia' en Asturias y Cantabria. Análisis gramatical y nuevos datos", *in* I. Delgados Cobos & A. Puigvert Ocal (eds.), *Ex admiratione et amicitia. Homenaje a Ramón Santiago*. Madrid: Ediciones del Orto, pp. 395-434.

FERNÁNDEZ-ORDÓÑEZ, I. (2016): "Dialectos del español peninsular", *in* J. Gutiérrez Rexach (ed.): *Enciclopedia lingüística hispánica*. Londres & New York: Routledge, vol. 2, pp. 387-404.

GARCÍA, C. (1986): "El castellano en Galicia", in V. García de la Concha *et alii*: *El castellano actual en las comunidades bilingües de España*. Salamanca: Junta de Castilla y León, pp. 49-64

GÓMEZ MOLINA, J. R. (2001): *El español hablado en Valencia: Materiales para su estudio*. Valencia: Universidad de Valencia.

GÓMEZ SEIBANE, S. (2017): "Español en contacto con la lengua vasca: datos sobre la duplicación de objetos directos posverbales", *in* A. Palacios (ed.), *Variación y cambio lingüístico en situaciones de contacto*. Madrid: Iberoamericana/Vervuert, pp. 143-159. https://doi.org/10.31819/9783954876648-008

JAFFE, A. & S. WALTON (2000): "The voices people read: Orthography and the representation of non-standard speech", *Journal of Sociolinguistics* 4/4, pp. 561-587. https://doi.org/10.1111/1467-9481.00130

JUILLAND, A. & E. CHANG-RODRÍGUEZ (1964). *Frequency dictionary of Spanish words*. La Haya: Mouton.

LE MEN, J. (2002): *Léxico del leonés actual*. León: Centro de estudios e investigación *San Isidoro*, 2002-2012, 6 vols.

LOPE BLANCH, J. M. (1986): *El estudio del español hablado culto. Historia de un proyecto*. México: UNAM.

NAGY, N. & D. SHARMA (2013): "Transcription", *in* R. J. Podesva & D. Sharma, (eds.): *Research methods in linguistics*. Cambridge: Cambridge University Press, pp. 235-256. https://doi.org/10.1017/CBO9781139013734.014

OCHS, E. (1979): "Transcription as theory", in E. Ochs & B. Schieffelin (eds.): *Developmental pragmatics*. New York: Academic Press, pp. 43-72.

PAASCH KAISER, C. (2015): *El castellano de Getxo: estudio empírico de aspectos morfológicos, sintácticos y semánticos de una variedad del castellano hablado en el País Vasco*. Berlin: De Gruyter. https://doi.org/10.1515/9783110366518

POLLÁN, C. (2001): "The expression of pragmatic values by means of verbal morphology: A variationist study", *Language Variation and Change* 13, pp. 59-89. https://doi.org/10.1017/S0954394501131030

POLLÁN, C. (2002): "The morphological expression of pragmatic values in oral and written Galician", in M. Fernández Ferreiro & F. Ramallo (eds.): *Sociolinguistics in Galicia: Views on diversity, a diversity of views* [= *Estudios de Sociolingüística*, 3/2 (2002) & 4/1 (2003)], pp. 113-138.

POPLACK, S. (1989): "The care and handling of a mega-corpus: The Ottawa-Hull French project", in R. Fasold & D. Schiffrin (eds.): *Language change and variation*. Amsterdam: John Benjamins, pp. 411-451. https://doi.org/10.1075/cilt.52.25pop

POPLACK, S. (1993): "Variation theory and language contact", in D. R. Preston (ed.): *American dialect re-search: An anthology celebrating the 100th anniversary of the American Dialect Society*. Amsterdam: John Benjamins, pp. 251-286. https://doi.org/10.1075/z.68.13pop

PRESTON, D. R. (1982): "'Ritin' Fowklower Daun 'Rong: Folklorists' failures in phonology", in *Journal of American Folklore* 95, pp. 304-326. https://doi.org/10.2307/539912

PRESTON, D. R. (1985): "The Li'l Abner syndrome: Written representations of speech", *American Speech* 60/4, pp. 328-336. https://doi.org/10.2307/454910

PRESTON, D. R. (2000): "Mowr and mowr bayud spellin': Confessions of a sociolinguist", *Journal of Sociolinguistics* 4/4, pp. 614-621. https://doi.org/10.1111/1467-9481.00132m

RABANAL, M. (1967): "Gramática breve del castellano hablado en Galicia y otros temas", in M. Rabanal: *Hablas hispánicas. Temas gallegos y leoneses*. Madrid: Ed. Alcalá, pp. 11-69.

RAE - ASALE (2009): *Nueva gramática de la lengua española*. Madrid, Espasa-Calpe.

RECALDE FERNÁNDEZ, M. (2012): "Aproximación a las representaciones sociales del español de Galicia", in T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas & A.

Veiga (eds.): *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, Universidad de Santiago de Compostela, pp. 667-680.

RODRÍGUEZ ESPIÑEIRA, M. J. (2019): "La expresión epistémica *si cuadra* en español de Galicia", *Estudos de Lingüística Galega* 11, pp. 197-231. http://ojs3usc.devxercode.es/index.php/elg/article/view/5343. https://doi.org/10.15304/elg.11.5343

ROJO, G. (2004): "El español de Galicia", in R. Cano Aguilar (coord.): *Historia de la lengua española*. Barcelona: Ariel, pp. 1087-1101. 2.ª ed., 2005.

ROJO, G. (2008): "Lingüística de corpus y lingüística del español", ponencia plenaria en el XV congreso de la *Asociación de Lingüística y Filología de América Latina* (Montevideo, 18-21 de agosto de 2008). Montevideo. Edición en CD. [ISBN 978-9974-8002-6-7]

ROJO, G. (2010): "Sobre codificación y explotación de corpus textuales: Otra comparación del *Corpus del español* con el CORDE y el CREA", *Lingüística* 24, pp. 11-50.

ROJO, G. (2011a): "Frecuencia de inventario y frecuencia de uso", *Revista española de lingüística* 41/1, pp. 5-43.

ROJO, G. (2011b): "Me pidieron que reseñara~reseñase el libro que ?publicara / *publicase Bosque en 1980", in Mª V. Escandell Vidal, M. Leonetti & C. Sánchez López (eds.): *60* problemas *de gramática dedicados a Ignacio Bosque*. Akal: Madrid, pp. 213-219.

ROJO, G. (2017): "Sobre la configuración estadística de los corpus textuales", *Lingüística* 33/1, pp. 121-134. https://doi.org/10.5935/2079-312X.20170008

ROJO, G. & V. VÁZQUEZ Rozas (2014): "Sobre las formas en *-ra* en el español de Galicia", in A. Enrique-Arias, M. J. Gutiérrez, A. Landa & F. Ocampo (eds.): *Perspectives in the study of Spanish language variation. Papers in honor of Carmen Silva-Corvalán*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 237-270. DOI: dx.doi.org/10.15304/va.2014.701

SAMPSON, G. (2000): "CHRISTINE Corpus: Documentation". http://www.grsampson.net/ChrisDoc.html.

SINNER, C. (2001): *Corpus oral de profesionales de la lengua castellana en Barcelona*. Accesible en http://www.carstensinner.de/castellano/corpusorales/index.html

TAGLIAMONTE, S. A. 2004. *Transcription protocol*. https://www.cambridge.org/gb/files/3713/6689/9690/2847_APPENDIX_C.pdf. [Última consulta: 19/01/2019].

TORRES CACOULLOS, R. & C. TRAVIS (2018): *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235259

TORRUELLA, J. & R. CAPSADA (2013): "Lexical statistics and typological structures: A measure of lexical richness", *Procedia. Social and Behavioral Sciences* 95, pp. 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668

TORRUELLA, J. & R. CAPSADA (2017): "Métodos para medir la riqueza léxica de un texto. Revisión y propuesta. Aplicación en el Corpus Informatizado del Catalán Antiguo", *Verba* 44, pp. 347-408. https://doi.org/10.15304/verba.44.3155

VANN, R. E. (2009): *Materials for the sociolinguistic description and corpus-based study of Spanish in Barcelona. Toward a documentation of colloquial Spanish in naturally occurring groups*. Lewiston, NY: The Edwin Mellen Press.

VÁZQUEZ VEIGA, N. (2003): *Marcadores discursivos de recepción*. Santiago de Compostela: Universidade de Santiago de Compostela.

VILA PUJOL, Mª R. (2001): *Corpus del español conversacional de Barcelona y su área metropolitana*. Barcelona: Universitat de Barcelona.


USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA