Teira, Celia y Polo, Nuria (2021). Digitalización y recursos para la investigación en lingüística. *Revista española de lingüística,* 51(1), 157-176.
DOI: http://dx.doi.org/10.31810/RSel.51.1.9

Resumen:
La investigación lingüística se ha beneficiado del desarrollo de las tecnologías, y a su vez, ha contribuido al mismo de manera significativa. El valor interdisciplinar de la lingüística y sus diferentes ramas requieren nuevos materiales, herramientas y métodos, cuyo acceso y manejo ha facilitado la tecnología. Dentro de ella, se distinguen las tecnologías que permiten un mayor entendimiento del sistema lingüístico de los productos tecnológicos que lo utilizan para su funcionamiento.
Se presenta a continuación un panorama general de recursos digitales que podrían contribuir en las diferentes etapas de una investigación lingüística, desde la revisión bibliográfica hasta la publicación final de resultados.

Palabras clave: digitalización; tecnologías; investigación; lingüística.

DIGITAL RESOURCES IN LINGUISTIC RESEARCH

Abstract:
Linguistic research has not only gained from technologies improvements, but also has contributed significantly to its development. The interdisciplinary nature of Linguistics and its different fields require new materials, tools, and methods. Technologies are also facilitating our access to all these options and how we understand them.
In our review, we are going to distinguish technologies that improve our knowledge of the linguistic systems from those other technologies that presuppose the linguistic system to operate. Our goal is to present an overview of digital resources that might help at different research stages in Linguistics, from the initial literature review to the publication of the research output.

Key words: digitization; technologies; research; Linguistics.

1.INTRODUCCIÓN

El estudio del lenguaje hoy es impensable sin la asistencia de la tecnología y la digitalización. De hecho, tal es así que gracias a este desarrollo la pandemia de covid-19 que azota el mundo desde 2020, ha visto reducido su impacto en el mundo académico. Sirvan como ejemplo algunos de los congresos desarrollados en línea durante estos años 2020 y 2021, como *Abralin ao vivo* con participación de lingüistas de diferentes países desde mayo de 2020, entre ellos España, cuyo cursos han quedado recogidos aquí; el 36 Congreso Internacional de Informática del 22 al 24 de septiembre de 2020; o *Language and aging: The International Online Workshop on Language in Healthy and Pathological Aging*, organizado por las Universidades de Sevilla y Salamanca, en abril de 2021.

No obstante, este desarrollo sin el que ahora no se puede concebir el trabajo académico, y la lingüística no es una excepción, tiene apenas 70 años. Su origen se encuentra en los primeros intentos de traducción automática por la década de 1950 y, en España concretamente, en la fundación de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN) en 1983. A pesar de que ya han pasado bastantes años, y el nivel de desarrollo en España de la calidad y cantidad de digitalización y el uso de las tecnologías para la investigación en lingüística es significativo, es difícil equipararlo con el de otras lenguas en otros países (como ya señalaban Martí, 1999; Llisterri, 2007; Melero et al., 2012).

Antes de continuar, conviene esclarecer la relación entre la lingüística y la tecnología, ya que hay que buscar esta en dos vertientes. Por un lado, en un afán por ampliar el conocimiento del lenguaje, la tecnología no es más que una herramienta para conseguirlo. Por el otro, la tecnología trata de simular o reproducir la capacidad humana del leguaje. Por eso es importante no confundir los primeros, los instrumentos o fuentes de datos para su creación, con los segundos, los productos tecnológicos que requieren del lenguaje (Fernández Pérez, 2005). Entre los primeros se encuentran las herramientas tecnológicas específicas para la *codificación, análisis, etiquetaje y explotación* de datos (Cassany, 2016). En este sentido, la lingüística se ha convertido en una disciplina interdisciplinar que necesita nuevos materiales, herramientas y métodos porque la tecnología ha permitido acceder a una gran cantidad de información que necesita formas nuevas de manejarla.

Entre los segundos, la actividad simuladora de las máquinas, que se suele denominar *procesamiento del lenguaje natural*, se divide a su vez en procesamiento de lenguaje oral y escrito. Al primer grupo pertenecen el reconocimiento automático del habla (que los ordenadores entiendan), síntesis de habla (que los ordenadores hablen), los sistemas de diálogo (como Alexa, Siri, etc.); al segundo la creación y comprensión de textos escritos. Esto permite el dictado automático o conversión de habla en texto, la traducción automática, la transcripción automática, la identificación o reconocimiento automático del locutor, la redacción automática, etc. Sirva para ilustrar la situación de España que entre los años 2016-2020 el gobierno puso en marcha un «Plan Nacional de Impulso a las Tecnologías del Lenguaje», dentro del Marco de la Agenda Digital para España. Este plan se presentaba como una iniciativa para «fomentar el desarrollo del procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales en España, y especialmente en lengua española y lenguas cooficiales» (Bel y Rigau, 2015; MINECO,

2015; MINECO, 2018). Sus resultados se resumen en una serie de estudios, de infraestructuras lingüísticas (recursos de datos y recursos de software —SW—), desarrollos SW, plataformas donde se presentan las herramientas finales desarrolladas en el Plan, y demostradores, que posibilitan de forma interactiva entender su funcionamiento. Actualmente, numerosas empresas se dedican a estas tecnologías del habla, por ej. Nuance, por citar una extranjera, y Verbio, por mencionar una española, con diferentes proyectos para el desarrollo de la investigación, la enseñanza y aprendizaje de lenguas, la accesibilidad, etc.

Un compendio relativamente actual de estas tecnologías del lenguaje y los grupos de investigación existentes se encuentra en la web de la SEPLN y en publicaciones como la de Lahoz Bengoechea y Pérez Ramón (2019), Gonzalo (2016) o Rubio Ayuso y Hernández Rioja (2005). Las aplicaciones lingüísticas de estas tecnologías pueden ser muy diversas: desde la identificación de hablantes en lingüística forense (se puede consultar para ello la Asociación Internacional de Fonética Forense y Acústica, IAFPA), a su utilización en lingüística clínica, para proporcionar sistemas alternativos y aumentativos de comunicación a personas que presentan dificultades en este ámbito (véase la Sociedad para el desarrollo de los Sistemas de Comunicación Aumentativos y Alternativos, ESAAC). Para ampliar el conocimiento sobre las bondades del procesamiento del lenguaje natural se puede consultar Llisterri (2003, 2007) o Moreno Sandoval (2019), entre otros.

En ese sentido, el objetivo de este breve resumen es presentar un panorama no tanto exhaustivo, pero sí ilustrativo, de los recursos y herramientas digitales que un investigador en cualquier rama de la lingüística puede necesitar. Si se repasan las herramientas y recursos digitales disponibles, se descubre que estos benefician a las diferentes ramas de la lingüística de maneras muy diversas. La lingüística histórica ha podido servirse de «los corpus electrónicos, los diccionarios informatizados y las bases de datos bibliográficas» (Juliá-Luna y Paz-Afonso, 2011, p.320). La lingüística teórica a su vez, además de contar con corpus y diccionarios, recurre a instrumentos para el reconocimiento, la segmentación, el etiquetado... en los distintos niveles del lenguaje (fonético, fonológico, morfológico, sintáctico, léxico y semántico). Por último, en lingüística aplicada, no solo es innegable la entidad y relevancia de estas herramientas, sino que el crecimiento de todas sus áreas ha sido exponencial con el uso de las tecnologías: enseñanza de lenguas, traducción, lingüística clínica, lingüística forense o judicial, etc.

Seguidamente se ofrece una recopilación no sistemática de estos materiales, herramientas y métodos que pueden servir para la búsqueda y manejo del conocimiento lingüístico.

2. Recursos para la investigación

Como se expone a continuación, los recursos influyen directa o indirectamente en la elección del tema de investigación; facilitan la revisión bibliográfica (catálogos en línea, revistas digitales, etc. que evitan tener que visitar físicamente todas las bibliotecas de su comunidad), y permiten desarrollar una u otra metodología (materiales, instrumentos, procedimientos y análisis de resultados). De igual modo, contribuyen a la redacción (ej.

traducciones, correctores ortográficos y gramaticales, etc.) y al modo de difusión de dicha investigación (mediante redes sociales, buscadores académicos, etc.). Se trata tan solo de ofrecer una muestra de todo lo existente en este mundo virtual que pretende dar cuenta del panorama general y sus posibilidades.

2.1. *La revisión bibliográfica*

El acceso a la información en línea es uno de los grandes avances para el investigador: hay mucha información, en diferentes lenguas (con posibilidad de traducción automática). Son recursos bibliográficos los libros, revistas y artículos científico-técnicos, las publicaciones de congresos, las tesis y trabajos académicos… que podemos consultar en buscadores generales y especializados, en bases de datos, catálogos y repositorios.

No nos vamos a detener especialmente en los *buscadores generales* (*Google*, *Bing*, *Yahoo Search*, etc.) y vamos a nombrar a continuación algunos de los buscadores especializados en los que se pueden encontrar materiales para la investigación lingüística: Google Académico, Microsoft Academic, Research Gate, ScienceResearch, Academia.edu, Springer Link, JURN, BASE o ERIC. En el caso de nuestro país, resulta interesante dentro de los repositorios (espacios centralizados de contenidos digitales, propios de las instituciones), el del Consejo Superior de Investigaciones Científicas (Digital.CSIC), la posibilidad de acceder a varios repositorios españoles a través del buscador conjunto de la FECYT (Recolecta) y el Repositorio español Científico y Tecnológico de esta misma institución (Recyt).

Los buscadores acuden a las *bases de datos* bibliográficas para recuperar información. Estas no hacen una búsqueda general en la red, como los buscadores, tienen una estructura concreta, de registros y campos, y contienen información específica y contrastada. Ejemplos de algunas de las bases de datos internacionales especializadas son *Scientific Electronic Library Online* (SciELO); *Modern Language Association* (MLA), sobre lengua y literatura inglesa de EE.UU.; y una base de datos de *Proquest*, especializada en lingüística y ciencias del lenguaje, es *Linguistics and Language Behavior Abstracts* (LLBA). En España, contamos con DIALNET (portal dedicado a las publicaciones científicas de España, Portugal y Latinoamérica); TESEO (base de datos de las tesis españolas defendidas desde 1976; mientras que OATD permite acceder a las tesis internacionales); ISOC, base de datos multidisciplinar sobre Ciencias Humanas y Sociales en español del CSIC; y también es de interés saber que la Fundación Española para la Ciencia y Tecnología (FECYT) gestiona las licencias de dos bases de datos científicas internacionales, *Web of Science* y *Scopus*; para más información se puede consultar los capítulos correspondientes a búsquedas bibliográficas en Chacón Beltrán (2021).

Al mismo tiempo, cada vez es mayor la digitalización de las bibliotecas. Sus *catálogos y repositorios digitales* posibilitan el acceso a fuentes diversas. Dentro de los catálogos, cabe mencionar el de la Biblioteca Nacional de España (BNE) que permite consultar las referencias bibliográficas de todos sus documentos (algunos de sus manuscritos originales se están digitalizando —ver la Biblioteca Digital Hispánica—); y el de la Biblioteca Virtual Miguel de Cervantes (Cervantes Virtual) con materiales bibliográficos en lengua española, centrándose en las bibliotecas virtuales y las ubicadas en países

hispanohablantes. De interés igualmente para la consulta de fondos antiguos, es el Archivo Digital de Textos y Manuscritos Españoles (ADMYTE) con transcripciones íntegras de obras en español o cualquiera de sus dialectos, durante el Medievo. Existe un catálogo de Bibliotecas y Hemerotecas Españolas (BIDI) que posibilita consultar bibliotecas y centros de documentación españoles en diferentes países. Por último, mencionaremos la Red de Bibliotecas Universitarias (REBIUN), cuyo catálogo unifica en una búsqueda tanto catálogos de las Universidades españolas como los de centros de investigación y el WorldCat, un catálogo que rastrea entre las bibliotecas de todo el mundo. En este caso solo tendremos acceso a los contenidos que tengan publicados en línea, pues el préstamo interbibliotecario solo existe entre bibliotecas universitarias españolas. En relación con estos, podemos consultar el directorio de bibliotecas españolas y de todo el mundo de la Universidad de Salamanca.

2.2. *La metodología*

Los materiales para la investigación en lingüística son principalmente corpus orales y escritos, que pueden apoyarse en imágenes (imprescindibles para trabajar con lenguas de signos) o ser multimodales. A continuación, se presentan los recursos que permiten explotar estos corpus, instrumentos y procedimientos para realizar experimentos y analizar los resultados.

2.2.1. Materiales

2.2.1.1 *Los corpus*

La digitalización ha permitido acceder a grandes fuentes de datos (lingüísticos) que hasta ahora era imposible localizar, almacenar y manejar. Tanto es así que esto ha supuesto una nueva rama de la lingüística denominada *lingüística de corpus* (McEnery y Wilson, 1996; Kennedy, 1998; por citar los primeros trabajos; véase también la Sociedad española de Lingüística de Corpus, AELINCO). En palabras de Rojo (2016: 285) un *corpus* es un conjunto de textos naturales, almacenados en formato electrónico, representativos de una variedad lingüística. Con textos se refiere a textos orales y escritos. Veamos algunos ejemplos.

Entre los *corpus orales* se encuentran los de habla conversacionales y monólogos; de nativos o hablantes de una segunda lengua, de niños o adultos, etc., tanto del español como de otras lenguas.

El corpus PRESEEA dentro del Proyecto para el Estudio Sociolingüístico del español de España y de América, recoge muestras de habla representativas de la diversidad sociolingüística hispanohablante. De similares características es el Atlas Oral del proyecto *Wikilengua del español* de la FundéuRAE. Cuenta con un corpus oral en abierto construido por una comunidad de hablantes de variedades del español. La Dialectoteca del español y el catálogo Voces Hispánicas del Centro Virtual Cervantes son también bibliotecas audiovisuales con muestras de diferentes hablantes nativos de variedades del español. CORLEC es el Corpus Oral de Referencia de la lengua española contemporánea. Contiene textos de diferentes tipos: administrativos, científicos, familiares, etc. El corpus del español oral ESLORA, elaborado por los miembros del Grupo de Gramática del

Español de la Universidad de Santiago de Compostela. El Corpus Oral Didáctico Anotado Lingüísticamente (C-Or-DiAL) del Laboratorio *Linguistico del Dipartimento di Italianistica dell'Università di Firenze* es un corpus oral de la lengua española, que recoge la transcripción ortográfica, con etiquetado prosódico y anotación de las funciones comunicativas. El Corpus integrado de referencia en lenguas romances (C-ORAL-ROM) es un corpus multilingüe de habla espontánea, formado por grabaciones de francés, español, italiano y portugués, perteneciente al Laboratorio de Lingüística Informática de la UAM. Asimismo, el corpus del español coloquial realizado por el grupo Val.Es.Co de la Universidad de Valencia. Finalmente, COSER es un Corpus Oral y Sonoro del Español Rural y el corpus oral de las hablas en la frontera entre España y Portugal (FRONTESPO). Se puede consultar Rojo (2016) para otros corpus pasados y actuales del español.

Entre los *corpus escritos*, destacan en español los corpus de la Real Academia de la Lengua Española (Corpus de Referencia del Español Actual, CREA; Corpus Diacrónico del Español, CORDE; Corpus del español del siglo XXI, CORPES. El corpus del español de Mark Davies es también uno de los corpus de referencia; el Corpus del Español Actual (CEA), lematizado y etiquetado con información morfológica y categorial, coordinado por Carlos Subirats y Marc Ortega, y un proyecto que aúna distintos corpus de textos, el proyecto Letras y Números en Análisis Lingüísticos (LYNEAL); el Corpus del Español Mexicano Contemporáneo perteneciente a El Colegio de México. Fue pionero en la confección de corpus el *Survey of English Usage* sobre la lengua inglesa.

Dentro de los *corpus diacrónicos* podemos encontrar, el ya mencionado CREA, los tres corpus de documentos, literatura y prensa, que reúne el Corpus diacrónico y diatópico del español de América (CORDIAM). A su vez, el corpus Hispánico y Americano en la red: Textos Antiguos (CHARTA) incluye archivos de los siglos XII a XIX, en español. Siguiendo sus criterios de etiquetaje, encontramos el Corpus de Documentos Españoles Anteriores a 1800 (CODEA). Además, cabe mencionar el portal de Corpus Históricos Iberorrománicos (CORHIBER), un recopilatorio de los Atlas Lingüísticos tradicionales españoles (CORPAT) y el corpus de Ediciones de Clásicos Latinos en el Renacimiento publicados en España. Esta es la propuesta de Francisco Gago Jover y Javier Pueyo Mena, con más de 32 millones de palabras en español medieval.

Algunos *corpus de hablantes aprendices* de español se pueden encontrar a través del Indexador del Corpus de Aprendices de Español del Departamento de Lengua Española de la UCM de Madrid. A estos se puede añadir el corpus de grabaciones en vídeo de conversaciones naturales para enseñar español (Columbia Corpus de Conversaciones para E/LE). Asimismo, el proyecto L de lengua con un corpus oral sobre la fonética de aprendices de español (Fonoele) y el Corpus escrito de español L2 con textos escritos por aprendices y nativos de español (CEDEL2). Para corpus de aprendices de cualquier lengua se puede consultar el *Learner Corpora around the World* del *Centre for English Corpus Linguistics* de la *Université Catholique* de Lovaina.

Para el estudio de la *prosodia*, destacan el corpus de habla anotado para los estudios prosódicos en español y catalán (GLISSANDO); o los diferentes atlas de entonación, como el Atlas interactivo de la entonación del español, el Atlas interactivo de la entonación romance, el Atlas Multimedia de la prosodia del español románico (AMPER) o Atlas interactivo del portugués y el Atlas sonoro de distintas lenguas.

De gran interés para la lingüística aplicada, como corpus de habla de poblaciones con trastornos del lenguaje en nuestro país destaca el Corpus de Percepción, Lenguaje y Afasia (PerLA) y el corpus de trastornos del lenguaje de la Universidad de Cádiz, que no está publicado en abierto (Paredes Duarte y Martín-Sánchez, 2018).

Para concluir este apartado, mencionamos el Centro de Normalización de la Lengua de Signos Española (CNLSE), con recursos lingüísticos en relación a la LSE, tales como una biblioteca virtual, un diccionario de signos (DILSE) o un corpus signado (corpus de la lengua de signos española). Además, la base de datos en abierto LSE-Sign y los proyectos de la universidad de Vigo (http://isignos.uvigo.es/ y http://griles.webs.uvigo.es/coralse.html). En cuanto a la lengua de signos catalana (LSC), se está trabajando sobre un corpus, pero todavía no está en abierto. Y para otras lenguas de signos existen corpus como: lengua de signos británica (BSL), lengua de signos americana (ASL), lengua de signos sueca, lengua de signos de neerlandesa (NGT), lengua de signos alemana (DGS), lengua de signos australiana (Auslan) y varias lenguas de signos de Asia. Además de los sistemas específicos de cada lengua, existe un Sistema de Signos Internacional.

2.2.1.2 *Las bases de datos*

Entre los *bancos de datos*, encontramos el Portal del Léxico Hispánico, que aúna datos bibliográficos, lingüísticos y documentales sobre el léxico de las lenguas y de los dialectos iberorrománicos. La Asociación Española de Terminología (AETER) recoge en su página web un listado de bancos de datos terminológicos de España y del extranjero. En cuanto al *léxico*, la base de datos sobre asociaciones léxicas entre miles de variedades de lenguas *Cross-Linguistic Colexifications* (CLICS), del *Max Planck Institute for the Science of Human History*, permite representaciones interactivas de las asociaciones semánticas. Otras bases de datos léxicas similares para el inglés son *Wordnet* o *Framenet*,

La base de datos HESPERIA para la recopilación, ordenación y tratamiento de materiales lingüísticos antiguos relativos a la Península Ibérica y el sur de Francia, reúne varias bases de datos (epigrafía, numismática, onomástica, lexicográfica y bibliográfica). La base de datos sobre Marcas de impresores permite enlazar los registros de impresor con los registros bibliográficos del Catálogo de la Biblioteca de la Universitat de Barcelona. Más allá del español, *el Diachronic Atlas of Comparative Linguistics* (DiACL), una base de datos para comparar lenguas genéticamente, el CALC (*Computer-Assisted Language Comparison*), ofrece diferentes recursos para comparar lenguas con interés filogenético, especialmente del sudeste asiático, del Max Planck Institute for the Science of Human History, y la base de datos *Causal Hypotheses in Evolutionary Linguistics* (CHIELD), una herramienta para evaluar hipótesis sobre la evolución del lenguaje.

La base de datos fonológica comparativa L1-L2*map*, desarrollada en la *Norwegian University of Science and Technology*, permite realizar análisis contrastivos entre la fonología de dos lenguas. También sobre *información fonológica* podemos consultar las bases de datos *Phonetics Information Base and Lexicon* (PHOIBLE), *UCLA*

Phonological Segment Inventory Database) (UPSID) y *Lyon-Albuquerque Phonological Systems Database* (LAPSyD), que reúnen inventarios fonológicos de diferentes lenguas del mundo. Sobre información fonética con medidas acústicas de vocales y sibilantes, la base de datos VoxClamantis, o proyectos colaborativos sobre cómo suenan las lenguas del mundo. También pueden ser fuentes de interés el banco de datos Lengua y prensa, de la Universidad de Málaga, sobre información que aparece en la prensa española acerca de lingüística, las lenguas de España y sus variedades; o *Verba volant*, base de datos sobre los telediarios de La 1 de Televisión Española.

Como bases de datos generales que recogen información de las lenguas del mundo son referentes Ethnologue (Eberhard, et al., 2020), Glottolog (Hammarströ et al., 2019), *The World Atlas of Language Structure* (WALS) (Dryer y Haspelmath, 2013) y *The Universal Archive*, sobre universales lingüísticos. Y sobre las lenguas en peligro de extinción se puede consultar la sección específica de la *Linguistic Society of America* o el Proyecto de Lenguas en Peligro.

De gran interés para la lingüística aplicada, es el proyecto TALKBANK, en el que podemos encontrar muestras de habla conversacionales, muestras de habla infantil, de población clínica, de hablantes multilingües…, en diferentes lenguas (entre ellas, el español). Dentro de este proyecto, destaca la base de datos CHILDES, que recoge muestras de habla durante el desarrollo infantil, con su nueva interfaz CHILDES-DB, que permite procesar estos datos en R, y Wordbank, la base de datos sobre el desarrollo de vocabulario infantil en diferentes lenguas.

2.2.2. *Instrumentos*

En este apartado se recogen instrumentos con los que tratar los datos recopilados anteriormente. No hablaremos del registro de los datos, ya que en cada tipo de corpus se especifican los requisitos de la muestra (cuantitativos y cualitativos). Presentamos brevemente algunas de las herramientas y recursos existentes para la codificación, análisis, etiquetaje y explotación de los materiales.

A menudo el desarrollo de recursos digitales para la investigación lingüística proviene desde lo que se ha llamado Humanidades Digitales (véase como ejemplo la actividad de la Asociación Humanidades Digitales Hispánicas, o la página web del Portal de Lingüística Hispánica). Cabe destacar CLARIN (*Common Language Resources and Technology Infrastructure*), un espacio virtual para alojar proyectos y recursos lingüísticos digitales de humanidades y ciencias sociales. También ofrece herramientas para explotar y analizar estos proyectos. En España está liderado por la Universitat Pompeu Fabra.

Existen distintos repositorios que recogen recursos generales de diversa índole, como Portal de Lingüística Hispánica sobre el español, el Linguistic Dada Consistorium sobre lenguas en peligro o la sección de recursos de LinguistLit. Algunas guías para el uso de datos estandarizados a la hora de comparar lenguas y para la anotación, como las propuestas por TEI (*Text Encoding Iniciative).*

A la hora de confeccionar un *corpus oral* hay que elegir un tipo de soporte que permita alinear el audio (o video) con la transcripción ortográfica (o fonética) del texto, ya que disponer del audio en un archivo y en otro distinto la transcripción escrita no es operativo.

Para ello se pueden usar herramientas como ELAN, creada en el *Max Planck Institute for Pyscholinguistics*, o EXMARaLDA, creada dentro del proyecto *Computer assisted methods for the creation and análisis of multilingual data* del Centro de Multilingüismo (*Sonderforschungsbereich Mehrsprachigkeit* – SFB 538) de la Universidad de Hamburgo. También es destacable Phon, creado por Yvan Rose (*Memorial* University,Canadá), que segmenta archivos de vídeo o audio y estos se pueden alinear con las correspondientes transcripciones. Actualmente es compatible con *Praat* para realizar análisis acústicos. Además de estas, para anotar video y audio se usan otros softwares como *ANVIL* o *Transana*, y específicos para lenguas de signos como *SignStream* o iLex y técnicas de captura de vídeo como WATSMART (Schembri, 2010).

En el ámbito de la *psicolingüística* se pueden encontrar páginas con una recopilación de los test estandarizados generales o específicos del desarrollo, que evalúan el lenguaje, recursos para experimentación psicolingüística, o textos estandarizados que se usan en los experimentos. Para el estudio del desarrollo del lenguaje, el Inventario de Desarrollo Comunicativo MacArthur tiene ahora una versión web y para realizar estudios sobre la adquisición de vocabulario destacamos el programa gratuito lognostics.

2.2.2.1 *Instrumentos para la investigación en fonética y fonología*

En el ámbito de la fonética y fonología, son ampliamente reconocidos programas para el análisis de habla como *PRAAT*, *Wavesurfer*, o el *Vocal Profile Analyses*. Estos programas permiten visualizar el espectro vocal, analizar parámetros acústicos, editar y manipular la señal, etc. También destacamos los recursos desarrollados por el prof. Granqvist, del KTH, para medir y simular sistemas acústicos tanto hablados como cantados (http://www.tolvan.com). Como simuladores de sonidos destacan los sonidos del habla de la Universidad de Iowa o *Seeing Speech* y como simuladores del tracto vocal, Pink trombone o Interactive Sagittal Section, simuladores de ondas sonoras, y, finalmente, un software para analizar las vocalizaciones animales (https://koe.io.ac.nz/#).

Dentro de los recursos del Laboratorio de Fonética Antonio Quilis (UNED) se encuentran algunas herramientas disponibles para este estudio fonético y fonológico (herramientas basada en PRAAT para llevar a cabo la notación prosódica, FonetiToBI, o su segmentación, SegProso; para la estilización y etiquetado de la frecuencia fundamental (f_0), MelAn; para modificar la f_0, ModProso; herramienta para la generación de prosodia en sistema de síntesis de habla, GenProso; un creador de diccionarios fonéticos, TransDic y un generador de transcripciones fonéticas, en español y catalán, TransText; así como guiones para la realización de análisis de diferentes estímulos, ej. *Vowel analyses* y *Fricatives*).

En la página web del Laboratori de Fonètica de la universidad de Barcelona se encuentra el acceso a un transcriptor fonológico multidialectal del español (TRAFO), y a un programa léxico que devuelve automáticamente la evolución fonética del latín al castellano del mismo. Su página web también se alcanza al Proyecto AMPER-CAT, que mencionábamos anteriormente por su corpus (dialectometrización de datos prosódicos, *ProDis* y *Calcu-dista*; notación prosódica mediante el *FonetiToBI*, y su aplicación a la

enseñanza de la entonación de una lengua extranjera, *AMPER_Dídac* y AMPER_Forensic).

Y, finalmente, también son interesantes los recursos que ofrece el departamento de *Speech, Hearing and Phonetics* de la UCL, como tutoriales para enseñar fonética, o un software online para análisis de la f_o.

Como herramientas de análisis fonológico destacan las derivaciones tradicionales y el software para análisis optimales. En cuanto al español, el buscador de estructuras silábicas Bufon o un diccionario multilingüe (Dicofon).

2.2.2.2 *Instrumentos para la investigación en morfología y sintaxis*

En morfología y sintaxis, el Grupo de Estructuras de Datos y Lingüística Computacional del Departamento de Informática y Sistemas de la Universidad de Las Palmas de Gran Canaria ha desarrollado aplicaciones de morfología computacional, sintaxis automatizada, análisis de textos y lexicografía (en su página web se puede encontrar un desambiguador morfosintáctico y un Flexionador y Lematizador Automático de Palabras en Español, FLAPE). El grupo ProLNat de la Universidad de Santiago de Compostela cuenta también con aplicaciones interesantes en este terreno. Entre ellas, un conjugador verbal para el gallego (CONSHUGA) o una analizador sintáctico para varias lenguas (DepPattern). Algunos de sus miembros son también autores *Linguakit* una propuesta multilingüe para el tratamiento de textos escritos que incluye aplicaciones diversas: para la conjugación, etiquetadores morfosintácticos y analizadores sintácticos, etc. El *Centre de Llenguatje i Computació* (CLiC) entre sus demos de morfología y sintaxis presenta varias aplicaciones (flexionadora, lematizadora y etiquetadora, junto a un analizador sintáctico). Es interesante revisar la página del Grupo de Procesamiento del Lenguaje Natural (IXA) de la Universidad del País Vasco. En ella, se muestran recursos lingüísticos relacionados con diferentes áreas (entre ellas, la morfosintáctica: como la base de datos léxica para el vasco EDBL; el analizador sintáctico para el vasco EDGK, etc.). GRAMPAL es un analizador morfológico de acceso libre desarrollado por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid (LLI-UAM). Dentro del grupo de investigación de Recursos en Tecnologías del Lenguaje (TRL) del Instituto de Lingüística Aplicada de la Universitat Pompeu Fabra se presentan distintas herramientas de análisis, entre las que se encuentra IULA Spanish LSP Treebank; el analizador de dependencias MaltParser; etc. También el Archivo Gramatical de la Lengua Española (AGLE) de Salvador Fernández Ramírez sobre la gramática del español. Un recurso de la *Association for Linguistic Typology*, que remite a descripciones gramaticales en diferentes lenguas, es Grammar Watch.

2.2.2.3 *Instrumentos para la investigación del léxico y la semántica*

En el terreno del léxico y la semántica trabajan también muchos de los grupos mencionados. Así los lexicones de CliC construidos a partir del corpus Ancora (AncoraNet, Ancora-Verb, Ancora-Nom). Las bases de datos léxicas y redes que se presentan entre los recursos de la página web del Centro de Tecnologías y Aplicaciones del Lenguaje y Habla (TALP) de la Universitat Politècnica de Catalunya: un generador de mapas entre diferentes redes léxicas (WN-Map); un repositorio central multilingüe

(MCR); etc. El grupo de Procesamiento de Lenguaje Natural y Sistemas de Información presenta entre sus recursos corpus y ontologías, como *Semantic Package* desarrollada para extraer rasgos semánticos predeterminados en textos escritos u OntoLegoLanguage, propuesta no solo para la extracción de información semántica sino para la generación de información nueva a partir de la misma. Finalmente, la plataforma web para el estudio morfogenético del léxico. En este ámbito también nos podemos encontrar bases de datos léxicas que permiten buscar palabas en función de determinadas características, como Palabras Tip o EsPAL. Dos últimas herramientas de gestión de corpus y análisis textual que queríamos mencionar son *wordSmith* o *sketchEngine*.

Entre los *diccionarios* de español, destacan el diccionario de la lengua española (DLE); Diccionario panhispánico de dudas (Dpd); Nuevo Tesoro Lexicográfico de la lengua española (NTLLE); Nuevo diccionario histórico del español (DH); entre otros), el Nuevo Diccionario Histórico del Español, CNDH); junto con el Mapa de Diccionarios, que permite contrastar entre las diferentes ediciones de los diccionarios. También es necesario incluir otros diccionarios digitales como el Diccionario del español de México, Diccionario Clave. Diccionario de uso del español actual, Diccionario de neologismos del español actual, diccionarios de Disponibilidad Léxica (DispoLex) para las diversas zonas del mundo hispánico y base de datos y herramientas para realizar los cálculos más habituales en disponibilidad léxica; y los diccionarios inversos como IEDRA y diccionarios de ideas afines. Para las lenguas de signos cabe destacar el diccionario multilingüe *Spreadthesign*.

2.2.3. *Procedimiento y análisis de resultados*

El *procedimiento* se refiere a diferentes aspectos que engloban la selección de personas participantes (tratamiento de la muestra poblacional); la creación de estímulos y su administración en el caso de estudios experimentales; las instrucciones proporcionadas; el registro de las respuestas, etc.

Actualmente, la posibilidad de crear cuestionarios en línea y su distribución es relativamente sencilla (ej. a través de Google Forms, Typeform, SurveyMonkey, Limesurvey o Qualtrics). Asimismo, en este enlace del Max Planck se tiene acceso a diversos tipos de cuestionarios (para extraer información fonética, morfosintáctica o léxica) que se utilizan para investigar lenguas de muy diversas tipologías.

De igual modo, la creación de experimentos y su administración virtual está en pleno auge, apareciendo propuestas como PsychoPy para experimentos de tipo psicolingüístico, con la versión Pavlovia para su realización online, o PsyToolkit (Lumsden, 2019), y plataformas como Ibex, que permite realizar experimentos a partir de tareas de lectura y de juicios de gramaticalidad. Una herramienta virtual para la experimentación perceptiva desarrollada en España se encuentra entre los proyectos de Instituto da Lingua Galega: FOLErPA. Esta plataforma permite crear test perceptivos y distribuirlos. Con funcionalidades similares el programa TP disponible en portugués y español.

El *análisis estadístico de datos* es imprescindible para organizar, analizar e interpretar correctamente los resultados que hemos obtenido. Asimismo, nos va a permitir la presentación posterior de los resultados de forma gráfica. Podemos obtener respuestas categóricas (datos cualitativos) o numéricas (datos cuantitativos) de nuestra experimentación. Normalmente, nos va a interesar conocer la media aritmética de los

valores y la desviación estándar, para conocer cómo se distribuyen los datos; y querremos visualizar gráficamente los datos con diagramas de barras, gráficos circulares, histogramas, etc. Al mismo tiempo, podemos querer conocer si existe relación entre variables (correlación), o si la evolución de una de ellas afecta al resto (regresiones). Quizá queramos predecir lo que va a pasar a partir de datos pasados (series temporales), etc.

Para calcular el tamaño de la muestra que se necesita en un estudio se puede usar una prueba de potencia (*power test*). Existen softwares en internet que pueden calcularlo automáticamente, como G*Power, https://www.surveysystem.com/sscalc.htm o https://homepage.divms.uiowa.edu/~rlenth/Power/.

Excel es el programa de análisis más sencillo, pero vamos a presentar ahora programas con amplias funcionalidades. Entre los investigadores son ampliamente conocidos los programas de análisis de datos *SPSS*, *STATA*, *Stats iQ*, *MatLab* o *MINITAB* o simuladores online que permite hacer estos análisis. Existen también procesadores que facilitan el análisis cualitativo de grandes bases de datos como *Atlas.ti*, *NVivo*, *The Etnograph*, AQUAD o MAXQDA. Los lenguajes de programación no deberían ser ajenos a nuestra profesión (ej. R o Python son algunos de ellos y su conocimiento permite aprovechar al máximo las posibilidades de análisis de grandes conjuntos de datos). Los lingüistas necesitamos dominar cada vez más estos programas para entender los resultados y poder replicar estudios.

Finalmente, para hacer gráficos puede resultar interesante esta herramienta. Para más información sobre la construcción de gráficos en Excel (versión libre, OpenOffice), R y SPSS (versión libre, JASP) se puede consultar el tutorial online de Bross (2019).

2.3 *La redacción y publicación de la investigación*

Hay una serie de herramientas informáticas para la redacción de textos científicos a las que nos hemos habituado, aunque sus posibilidades están en actualización constante. Entre los procesadores de texto más conocidos se encuentran *OpenOffice* o *Microsoft Office*. Estos incluyen correctores ortográficos y gramaticales, que, por supuesto, también podemos encontrar en línea (ej. *Stylus*, Corrector, etc.) Incluyen además diccionarios multilingües y de sinónimos que evitan en ocasiones, tener que consultar en la red. Resultan interesantes las herramientas de ayuda para la redacción de textos académicos como Harta, el portal de escritura de la UCM o la propuesta de ArText que se describe como un «sistema automático de ayuda a la redacción de textos en español de ámbitos especializados». Entre tales ámbitos se encuentran la administración pública, la medicina, el turismo y textos académicos como el Trabajo de Fin de Grado. También podemos querer disponer de generadores de resúmenes o palabras clave (Resumidor, *Resoomer*, *AutoSummarizer*, etc.); y de traductores automáticos (ej. *TTS*, *LinguaVox*, *DeepL*, etc.).

De igual modo, los gestores de referencias bibliográficas son especialmente relevantes para la investigación. Poder almacenar la cita, con su autor, año, título… como antes se hacía en fichas de papel, ahorra mucho tiempo y esfuerzo. Entre estos se encuentran Zotero, *Refworks*, *EndNote*, *Bibme*, *Mendeley*, *Citavi*, etc. La elección de uno u otro depende de nuestros gustos personales, de que su acceso sea gratuito o de que la biblioteca nos forme en su uso. Una comparación de los distintos gestores

bibliográficos se puede consultar en esta página de Wikipedia donde se ofrecen las características técnicas y de uso de cada uno de ellos.

El último paso consiste en decidir *dónde publicar* nuestros trabajos (lo que suele ser un tema que abordan la mayoría de las bibliotecas universitarias en su apartado de «Apoyo a la investigación»). Para dar a conocer cualquier trabajo, además de las bases de datos que se mencionaron al inicio (en el apartado de revisión bibliográfica), las tecnologías han permitido contar con listas de distribución, páginas personales y blogs de los investigadores, así como con las redes sociales. En 1996, se crea la lista de distribución *Infoling* con el fin de difundir información científica y técnica referente a distintos ámbitos de la lingüística hispánica (tenía como referente la australiana *Linguist List*, creada en 1990). En ella podemos encontrar información varia sobre publicaciones, peticiones de contribuciones, congresos, ofertas de trabajo, etc. En el Portal del Hispanismo del Instituto Cervantes se puede encontrar información similar.

El perfil público del investigador, que incluirá como mínimo el nombre, contacto, el centro en que trabaja, su currículo, proyectos, publicaciones, etc., es importante para dar a conocer a uno mismo y a su trabajo. *ORCID* es una iniciativa que promueve tener un código de identificación personal en la investigación, lo que evita las coincidencias en los nombres de los investigadores. Hay varias redes sociales específicas donde presentarse profesionalmente (*LinkedIn*, *ResearcherId*, *Research Gate*…), así como redes sociales generales, donde las interacciones no siempre van a tener el rigor científico esperado, pero favorecen el contacto y la difusión (*Facebook*, *Twitter*, *Instagram*, etc.) Muchos departamentos universitarios y grupos de investigación participan en las mismas.

Además, numerosos lingüistas e instituciones publican entradas en sus blogs (ej. Blog de la Fundeu, Blog de la RAE, Lenguaje Administrativo, blog de Cálamo&Cran, Sottovoce, etc.). Hypotheses es la plataforma para blogs académicos. En relación con esta temática de tecnologías y Lingüística, existía el blog Informática para las lenguas (cuya última entrada el 18 de febrero de 2018 tiene que ver con el tema de la publicación de la investigación: «¿En qué revista publico mi artículo»), si bien también en distintos blogs, se pueden consultar las entradas bajo la etiqueta «recursos lingüísticos» para obtener información sobre estos recursos digitales (como en el blog Morforetem).

3. Conclusiones

Los recursos digitales se caracterizan por su rápida evolución. Esto obliga también a la actualización constante de los investigadores. Cierto es que la lingüística ha acompañado la investigación tecnológica desde sus inicios, pero las posibilidades que puede proporcionar la digitalización a la investigación lingüística aún están por explorar. En las distintas fases de la investigación podemos contar con recursos de todo tipo: nadie es ajeno a los dispositivos de registro, transcripción, almacenamiento, análisis y presentación de datos, o a las propuestas que te permiten intercambiarlos, traducirlos, difundirlos, etc. Al mismo tiempo, la investigación lingüística proporciona información que no puede ser ignorada por estos avances digitales, respecto a sus unidades, características y posibilidades combinatorias.

Especialmente de interés para los lingüistas en formación, si bien no exhaustiva, esperamos que esta recopilación les pueda servir en su trabajo, ya que a menudo la gran

cantidad de información disponible hace no saber discernir sobre si un recurso o herramienta merece la pena o simplemente queda perdido entre todos los existentes.

4. BIBLIOGRAFÍA

Bel, N., Rigau, G. (eds.) (2015). *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*. España: Ministerio de Industria, Energía y Turismo. Recuperado de: https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf

Bross, F. (2019). *Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis*. Version 1.0. Mimeo.

Cassany, D. (2016). Recursos lingüísticos en línea: Contextos, prácticas y retos. *Revista Signos* 49(1). DOI: 10.4067/S0718-09342016000400002

Chacón Beltrán. R. (coord.) (2021). *La elaboración del TFM en filología. Guía práctica para estudiantes*. Editorial UNED.

Dryer, M. S. y Haspelmath, M. (eds.) (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Eberhard, D. M., Simons, G.F. y Fennig, C.D. (eds.) (2020). *Ethnologue: Languages of the World.* 23 edición. Dallas, Texas: SIL International. Versión online: http://www.ethnologue.com.

Fernández Pérez, M. (2005). Aplicaciones de la lingüística y Nuevas Tecnologías: De hecho, pareja, 29-44. En Cal Varela, M. (coord.), *Nuevas tecnologías en lingüística, traducción y enseñanza de lenguas.* Santiago de Compostela: Universidad de Santiago de Compostela.

Garrido Almiñana, J.M. (2015). Fonética experimental y tecnologías del habla. *NORMAS*, 5, 67-79. DOI: 10.7203/Normas.5.6822

Gonzalo, A.L. (coord.) (2016). *Tecnologías del lenguaje en España. Comunicación inteligente entre personas y máquinas*. Madrid: Fundación Telefónica. Recuperado de: https://www.fundaciontelefonica.com/cultura-digital/publicaciones/565/

Hammarströ, H., Forkel, R., Haspelmath, M. (2019). *Glottolog 4.0*. Jena, Alemania: Max Planck Institute for the Science of Human History.

Juliá-Luna, C., Paz-Afonso, A. (2011). Aplicación de las nuevas tecnologías a la historia de la lengua española: el estudio de la evolución del léxico. En Cerezo, M. (dir.), *III Jornada nacional sobre estudios universitarios. El presente de los nuevos títulos*. Castelló de la Plana: Universitat Jaume I, 319-326. DOI: 10.13140/RG.2.1.1733.1688

Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics*. Londres: Longman.

Lahoz Bengoechea, J.M., Pérez Ramón, R. (2019). *Subsidia: Tools and resources for speech science.* Málaga: Universidad de Málaga. Recuperado de: https://riuma.uma.es/xmlui/handle/10630/18177

Llisterri, J. (2003). Lingüística y tecnologías del lenguaje. *Lynx*, 2, 9-71. Recuperado de: http://liceu.uab.cat/~joaquim/publicacions/Llisterri_03_Linguistica_Tecnologias_Lenguaje.pdf

Llisterri, J. (2007). El español y las nuevas tecnologías. En M. Lacorte (Ed.), *Lingüística aplicada del español*, 483-520. Madrid: Arco/Libros. Recuperado de: http://liceu.uab.cat/~joaquim/publicacions/Llisterri_07_Tecnologias_Linguisticas_Espanol.pdf

Lumsden, J. (2019). So, you want to run an online experiment? [Blog Post]. Recuperado de https://ocean.sagepub.com/blog/how-to-run-an-online-experiment

Martí Antonin, M.A. (1999). Panorama de la Lingüística Computacional en Europa. Revista Española de Lingüística Aplicada, 1, 11-24. Recuperado de: https://dialnet.unirioja.es/descarga/articulo/227023.pdf

McEnery, T. y Wilson, A. (1996). *Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Melero, M., Badia, T., Moreno, A. (2012). *La lengua española en la era digital: The Spanish language in the digital age.* Berlin: Springer.

Ministerio de Economía y Empresa. (2015). *Plan de Impulso de las Tecnologías del Lenguaje*. Recuperado de: https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Impulso-Tecnologias-Lenguaje.pdf

Ministerio de Economía y Empresa. (2018). *Estudio de caracterización del sector de tecnologías del lenguaje en España*. Recuperado de: https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/Estudios%20del%20sector/Sector%20de%20tecnolog%C3%ADas%20del%20lenguaje%20en%20Espa%C3%B1a/estudio-caracterizacion-sector-TL.pdf

Molina Mejía, M., Valdivia Martín, P., Venegas Velásquez, R. (eds.) (2020). *Actas III Congreso Internacional de Lingüística Computacional y de Corpus-CILCC 2020 y V Workshop en Procesamiento Automatizado de Textos y Corpus-WoPATeC 2020. Libro de Resúmenes.* Medellín: Universidad de Antioquia. Recuperado de: https://cilcc20.files.wordpress.com/2020/11/libro-de-resumenes-actas-iii-cilcc-2020-y-v-wopatec-2020-virtual.pdf

Moreno Sandoval (2019). *Lenguas y computación*. Madrid: Síntesis.

Paredes-Duarte, M.J., Martín-Sánchez, V.M. (2018). *Corpus de Trastornos del Lenguaje*. Cádiz: Universidad de Cádiz.

Rojo, G. (2016). Los corpus textuales del español. En Gutiérrez-Rexach, J. (ed.), *Enciclopedia de lingüística hispánica*. Oxon: Routledge, 285-296.

Rubio Ayuso, A., Hernández Rioja, I. (2005). *Libro Blanco de las Tecnologías del Habla*. Recuperado de: http://lorien.die.upm.es/~lapiz/rtth/docs/LibroBlancoTecnologiasDelHabla.pdf

Schembri, A. (2010). Documenting sign languages. En P. K. Austin (ed.), *Language Documentation and Description 7: Lectures in Language Documentation and Description*, 105-143.London: School of Oriental and African Studies.