TFM Clara Yaling Colás Rodríguez

por Clara Yaling Colás Rodríguez

Fecha de entrega: 04-sep-2024 09:48p.m. (UTC+0200)

Identificador de la entrega: 2445066595

Nombre del archivo:
5290_Clara_Yaling_Colas_Rodriguez_TFM_Clara_Yaling_Colas_Rodriguez_564278_1658662077.pdf (1.51M)

Total de palabras: 19870

Total de caracteres: 110641

MÁSTER EN LETRAS DIGITALES: ESTUDIOS AVANZADOS EN TEXTUALIDADES ELECTRÓNICAS

TRABAJO FIN DE MÁSTER

Curso 2023-2024

Problemas, desafíos y soluciones en la creación de un corpus diacrónico en línea: el Corpus PROLEGRAMES (análisis DAFO: debilidades, amenazas, fortalezas y oportunidades)

ESPECIALIDAD: Organización de la información y metadatos

APELLIDOS Y NOMBRE: Colás Rodríguez, Clara Yaling

DNI: 51485579X

CONVOCATORIA: SEPTIEMBRE 2024

TUTORES: Dr. Fr. Javier Herrero Ruiz de Loizaga y Dr. Daniel M. Sáez Rivera

ANEXO I: DECLARACIÓN DE NO PLAGIO

Dña. Clara Yaling Colás Rodríguez con NIF 51485579X, estudiante de Máster en la Facultad de Filología de la Universidad Complutense de Madrid en el curso 2023 - 2024, como autor/a del trabajo de fin de máster titulado Problemas, desafíos y soluciones en la creación de un corpus diacrónico en línea: el Corpus PROLEGRAMES (análisis DAFO: debilidades, amenazas, fortalezas y oportunidades) y presentado para la obtención del título correspondiente, cuyos tutores son: Fr. Javier Herrero Ruiz de Loizaga y Daniel M. Sáez Rivera

---

DECLARO QUE:

El trabajo de fin de máster que presento está elaborado por mí y es original. No copio, ni utilizo ideas, formulaciones, citas integrales e ilustraciones de cualquier obra, artículo, memoria, o documento (en versión impresa o electrónica), sin mencionar de forma clara y estricta su origen, tanto en el cuerpo del texto como en la bibliografía. Así mismo declaro que los datos son veraces y que no he hecho uso de información no autorizada de cualquier fuente escrita de otra persona o de cualquier otra fuente.

De igual manera, soy plenamente consciente de que el hecho de no respetar estos extremos es objeto de sanciones universitarias y/o de otro orden.

En Madrid, a 03 de septiembre de 2024

Fdo.: Clara Yaling Colás Rodríguez

Agradecimientos

Mis más sinceros agradecimientos a mis tutores (Daniel Sáez Rivera y a Francisco Javier Herrero) a mi familia y a mis amigos, en especial a Jaime, a Mika, a Soul y a Lalo por estar siempre ahí. Sin vosotros nunca lo hubiera logrado.

Resumen

El objetivo de este Trabajo de Fin de Máster es ofrecer una perspectiva general del corpus PROLEGRAMES, un corpus diacrónico digital que estudia la estructura del español desde sus orígenes hasta su gramaticalización y lexicalización y su nueva herramienta en línea que está siendo desarrollada por el Proyecto de Investigación PROGRAMES del Departamento de Lengua Española y Teoría de la Literatura de la Universidad Complutense de Madrid y la empresa de IT, Avantopy. La aplicación de esta perspectiva general se realizará a partir del análisis DAFO del proyecto de investigación PROGRAMES, los documentos PROGRAMES y sus características y su nueva herramienta en línea. Se propondrán soluciones a los resultados del análisis DAFO centrándonos sobre todo en las posibilidades de explotación de la nueva herramienta.

Palabras clave: análisis DAFO, lingüística de corpus, heurística de corpus, corpus diacrónico digital.

Abstract

The aim of this Master's Thesis is to provide an overview of the PROLEGRAMES corpus, a digital diachronic corpus that studies the structure of Spanish from its origins to its grammaticalization and lexicalitation, and its new online tool that is being developed by the PROGRAMES Research Project of the Department of Spanish Language and Literature Theory of the Universidad Complutense de Madrid and the IT company, Avantopy. The application of this overview will be based on the SWOT analysis of the PROGRAMES research project, the PROGRAMES documents and their characteristics, and their new online tool. Solutions to the results of the SWOT analysis will be proposed focusing mainly on the exploitation possibilities of the new tool.

Keywords: SWOT analysis, corpus linguistics, corpus heuristics, digital diacronic corpus.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

PROLEGRAMES, procesos de lexicalización y gramaticalización: cambio, variación y pervivencia en la historia discursiva del español, es un proyecto heredero de los proyectos PROGRAMES que arrancan en 2001 en la Universidad Complutense de Madrid, y se especializa en el estudio de los procesos de lexicalización y gramaticalización del español. Su principal objeto de estudio son los textos no literarios pertenecientes a tradiciones discursivas que se acercan a la lengua hablada de las distintas épocas y que han sido recopilados en forma de corpus. El corpus PROLEGRAMES es un corpus diacrónico del español que tiene la peculiaridad de que todos sus textos han sido etiquetados, en la forma y en los fenómenos lingüísticos, de forma manual.

El corpus PROLEGRAMES es de acceso libre y puede consultarse libremente por cualquiera en la página institucional del proyecto alineándose con la afirmación de Calderón Campos de que "la investigación lingüística histórica no se concibe en la actualidad sin el recurso a los corpus digitales en línea" ("Los corpus" 16). Además, desde el proyecto, se ofrece la posibilidad de descargar los textos para su análisis en plataformas externas al corpus. Con el fin de garantizar la continuidad y mejora del proyecto, PROLEGRAMES ha empezado a desarrollar su propia herramienta de análisis de corpus que pueda explotar su rico etiquetado. Esta interfaz de usuario pretende agrupar en un mismo sitio tanto los textos como las herramientas de análisis dando respuesta a la necesidad de un análisis más accesible y directo del corpus, que anteriormente dependía de herramientas externas que a menudo no podían explotar todos los matices que ofrece la anotación de estos documentos.

El trabajo se estructurará en una serie de apartados que presentaremos a continuación. En primer lugar, se expondrán los objetivos del trabajo. Siguiendo con el segundo apartado, se llevará a cabo un estado de la cuestión, incidiendo particularmente en las herramientas de análisis de corpus. Además, se proporcionarán ejemplos que permitan observar la funcionalidades más empleadas y comunes entre las herramientas con el fin de, más tarde, poder utilizarlas como modelo con el que comparar la nueva herramienta en línea del corpus PROLEGRAMES. En el tercer apartado, se describirá detalladamente la metodología empleada para realizar los correspondientes análisis del proyecto PROLEGRAMES, su corpus y su nueva herramienta en línea. Seguidamente, en el cuarto apartado se presentará el resultado del análisis DAFO, así como la descripción

y el análisis del corpus PROLEGRAMES y se analizará de la herramienta de consulta su interfaz y sus funcionalidades. En el quinto apartado se propondrán estrategias de optimización del corpus PROLEGRAMES y nuevas formas de explotación de la herramienta en línea en base a los análisis del apartado anterior. Por último, el trabajo terminará con unas conclusiones.

2. Objetivos del trabajo

El objetivo principal de este Trabajo de Fin de Máster es la realización de un análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) del proyecto de investigación PROLEGRAMES y su corpus con el fin de obtener datos objetivos sobre el estado actual del proyecto y así poder proponer estrategias que contribuyan a su optimización y mejora. Además, para lograr este propósito, se han seguido los siguientes objetivos específicos que han sido: realizar una presentación detallada del corpus PROLEGRAMES basándonos en los criterios de clasificación de la lingüística de corpus, describir la interfaz y las funcionalidades de la nueva herramienta en línea del corpus PROLEGRAMES, plantear y proponer nuevas formas de explotación del corpus y su herramienta en línea y desarrollar un análisis detallado que pueda servir como referencia para proyectos y trabajos futuros del proyecto de investigación PROLEGRAMES.

3. Estado de la Cuestión

Es imprescindible para este trabajo contextualizar el corpus PROLEGRAMES y su necesidad de crear y desarrollar una herramienta para analizar el corpus. Por tanto, en este apartado mostraremos qué son los corpus lingüísticos, las herramientas más frecuentes para el análisis de los corpus y ejemplos que lo ilustren.

3.1. Corpus lingüísticos informatizados

Un corpus se define como "conjunto lo más extenso posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación" (RAE, *s. v.*). Esta acepción resulta ambigua e insuficiente para disciplinas más específicas y, en consecuencia, en el ámbito de la lingüística se ha desarrollado un concepto más preciso

que se conoce como "corpus textual" o "corpus lingüístico" (Instituto Cervantes s.p; Torruella 16).

Un corpus lingüístico se concibe como "una recopilación extensa de textos (escrito, orales o de ambos tipos)" (Instituto Cervantes s.p) "e informatizados" (Torruella 36) "producidos en situaciones reales que han sido seleccionados siguiendo unos criterios específicos" (Pérez s.p) "para que reflejen el conjunto del estado o del nivel de lengua representado" (Torruella 36). De esta definición destacamos el hecho de que los textos "estén informatizados" y "que hayan sido seleccionados siguiendo unos criterios específicos".

En primer lugar, es fundamental recalcar que la informatización de los corpus, impulsada la irrupción de la informática al ámbito académico, ha significado una revolución en el ámbito de la investigación gracias a que ha permitido la introducción de nuevas formas metodologías para el análisis de corpus que de forma manual no hubieran sido posible hacerlas (Torruella 32).

En segundo lugar, la selección de textos y, por consiguiente, la construcción de los corpus lingüísticos permite clasificar los corpus atendiendo a su representación de la lengua, lo cual resulta conveniente para los investigadores que buscan muestras concretas en periodos concretos de una lengua (Pérez s.p). Además, estos criterios de selección permiten diferenciar los corpus lingüísticos informatizados de los archivos informatizados y de las bibliotecas de textos electrónicos que no están pensados para hacer estudios lingüísticos (Torruella 34). La disciplina que aborda el estudio de los corpus lingüísticos se denomina Lingüística de corpus y se detallará en mayor profundidad en el apartado 4.2.

3.2. Análisis de los corpus

Con la llegada de los corpus digitales que produjo una revolución en el ámbito de la filología, pues al informatizar los procesos se logró que las investigaciones se pudieran realizar de forma más rápida, sencilla y sistemática, a la vez que se podían consultar una gran cantidad de datos que de otra forma no podría haberse hecho (Pérez s.p; Torruella 36). Esta innovación en el campo creó nuevas metodologías para el análisis para estos corpus y entre ellos se destacan los análisis cuantitativos y cualitativos, es decir, análisis estadísticos (Pérez s.p).

Los análisis cualitativos consisten en hacer "una descripción detallada y completa de un fenómeno lingüístico o del comportamiento de una palabra o grupo de palabras" (Pérez s.p), mientras que el cualitativo consiste en crear "índices de frecuencia a los fenómenos lingüísticos observados en el corpus y éstos pueden servir para construir modelos estadísticos más complejos, que expliquen la evidencia hallada en el texto." (Pérez s.p; Toruella 247). Como apuntan Pérez y Torruella, estas metodologías son complementarias y no excluyentes, pues su combinación puede proporcionar resultados más profundos y ricos (s.p; 247).

Dado que los corpus lingüísticos aspiran a ser muestras representativas de la lengua, es fundamental que las investigaciones se apoyen en datos reales y cuantificables que muestren una base sostenible y empírica. Torruella añade que estos análisis estadísticos son esenciales para avanzar en las investigaciones, describir las características de un grupo, determinar las frecuencias de un fenómeno o comprobar hipótesis de relación casual entre variables (249).

3.2.1. Herramientas para el análisis de textos de un corpus

Como mencionamos anteriormente, con el aumento y desarrollo de las tecnologías relacionadas con el análisis de corpus, han surgido nuevas metodologías para llevar a cabo análisis estadísticos en el ámbito de la filología que en el pasado habrían resultado inviables por haber tenido que hacerlas de forma manual. (Torruella 32) Estas herramientas se han creado con el fin de analizar corpus digitales y procesar grandes cantidades de datos mediante el uso de XML-TEI (Pérez s.p). Actualmente existen interfaces de análisis de licencia de uso libre como *Voyant Tools*, de uso comercial como *Sketch Engine* y aquellas especializadas en un corpus como es la interfaz de CORDE.

Las herramientas más habituales en los corpus lingüísticos se fundamentan en la anotación lingüística a partir de TEI y los procesos de lematización y *tokenización*. La lematización consiste "en la asignación del lema correspondiente a cada *token*, esto es, la forma que representa al conjunto de variantes morfológicas de una palabra y encabeza la entrada de un diccionario" (Vaamonde 57). Por su parte, la tokenización "se refiere al proceso de segmentar un texto para identificar y delimitar las unidades del corpus que van a ser anotadas —los *tokens*—, habitualmente palabras, números y signos de puntuación" (Vaamonde 57). Gracias a estos procesos, es posible el funcionamiento de las herramientas de análisis de corpus, así como la interoperabilidad entre interfaces

debido a la codificación en TEI. Las funcionalidades más frecuentes de las herramientas para los análisis de corpus "permiten a los usuarios generar y manipular la frecuencia de palabras, listas, concordancias y colocaciones" (Menéndez-Barzanallana s.p)

3.2.1.1. Lista de frecuencias

Las listas de frecuencias son una herramienta que permite a los usuarios comprobar la cantidad de palabras diferentes en un corpus y su frecuencia de aparición. Estas listas no tienen una visualización preestablecida y dependiendo de la herramienta los criterios de ordenación variarán. Además, pueden ayudar a establecer el grado de representatividad de un corpus (Pérez s.p; Menéndez-Barzanallana s.p).

Orden	Frec. absoluta	Frec. normalizada
1.	de	13249078 56694.391
2.	que	8893512 38056.402
3.	y	8120570 34748.891
4.	la	7436390 31821.203
5.	el	5807480 24850.902
6.	en	5701768 24398.548
7.	a	4164527 17820.510
8.	los	3941664 16866.852
9.	se	2959002 12661.924
10.	por	2803887 11998.168
11.	no	2416777 10341.678
12.	las	2412584 10323.736
13.	con	2372493 10152.182
14.	del	2177430 9317.484
15.	su	1947547 8333.787
16.	e	1731843 7410.764
17.	lo	1651577 7067.296
18.	es	1522056 6513.060
19.	al	1345928 5759.387
20.	un	1309651 5604.153

Figura 1: ejemplo de lista de frecuencia de CORDE. Elaboración propia. [1]

3.2.1.2. Listas lematizadas

Las listas lematizadas clasifican las palabras de un corpus por lemas y, además, se pueden combinar con las listas de frecuencias. Una de las dificultades que pueden surgir a la hora de la lematización automática lo produce la existencia de los homógrafos, palabras idénticas, que a menudo pueden generar confusión a los programas a la hora de realizar este proceso. Una forma de combatir este problema reside en el uso de procedimientos de desambiguación o la lematización y/o revisión manual (Menéndez-Barzanallana s.p y Torruella 228).

[1] Datos recuperados de: https://corpus.rae.es/frecCORDE/5000_formas.TXT

Group	Count	WPM	Percent
Juan	1,257	1,312.17	63.32
Jua	295	307.95	14.86
Ju	182	189.99	9.17
Joan	148	154.5	7.46
Juo	37	38.62	1.86
Jun	26	27.14	1.31
Jhoan	11	11.48	0.55
Jno	9	9.39	0.45
Jon	7	7.31	0.35
Guan	3	3.13	0.15
Jvan	3	3.13	0.15
Jn	3	3.13	0.15
J	2	2.09	0.1
Ja‖Juan	1	1.04	0.05
Jan	1	1.04	0.05

Figura 2: ejemplo de lista lematizada en ODE (Juan). Elaboración propia[2].

3.2.1.3. Listas de parada

Las listas de parada contienen aquellos elementos que el usuario desea que el programa informático ignore. (Pérez s.p y Menéndez-Barzanallana s.p)

```
▼<stopwords>
   <word>aquel</word>
   <word>aquellas</word>
   <word>aquellos</word>
   <word>esta</word>
   <word>estas</word>
   <word>este</word>
   <word>estos</word>
   <word>esto</word>
   <word>ese</word>
   <word>esa</word>
   <word>eso</word>
   <word>esos</word>
   <word>esas</word>
 </stopwords>
```

Figura 3: lista de parada en formato XML (demostrativos). Elaboración propia.

[2] Datos recuperados de: http://corpora.ugr.es/ode/index.php?action=home

3.2.1.4. Concordancia

La concordancia es una funcionalidad que permite recuperar todas las ocurrencias de un patrón de búsqueda en particular en sus contextos inmediatos y los muestra en un formato fácil de leer (Menéndez-Barzanallana s.p). A menudo se representa en una estructura de KWIC (*Key Word in Context*) que recupera todas las apariciones de una palabra en un texto o conjunto de textos para facilitar el estudio de fenómenos morfosintácticos, junto con un número determinado de caracteres de contexto anterior y posterior mientras que la palabra clave buscada se encuentra en medio, a menudo resaltada por un formato o color diferente (Pérez s.p). No obstante, la forma de la visualización de la concordancia a menudo cambia dependiendo de la herramienta. Además, como señala Pérez, "en la mayoría de las ocasiones se facilita al usuario los llamados caracteres comodín con los que se puede buscar diferentes formas de una misma palabra o realizar búsquedas difusas, múltiples y de frases idiomáticas con un cierto grado de variación" (s.p). Así mismo, además de los caracteres contextuales, los programas ofrecen funcionalidades que amplían el contexto enseñando el párrafo donde se inserta una muestra y/o permiten al usuario a acceder al texto completo donde aparece (Pérez s.p).

Por otra parte, las muestras recopiladas en la mayoría de los programas de análisis de corpus pueden ser almacenadas de forma local por el usuario. Los formatos más frecuentes de exportación son el texto plano .txt, hojas de Excel .xlsx, o formato XML .xml. Como bien apunta Pérez, la concordancia es una de las herramientas fundamentales para el estudio lingüístico, ya que recupera muestras lingüísticas que "revelan el contexto en el que se encuentran las ocurrencias individuales de palabras y junto a sus opciones para ordenar y mostrar los datos pueden facilitar el proceso de observar y distinguir patrones de comportamiento lingüístico" (s.p).

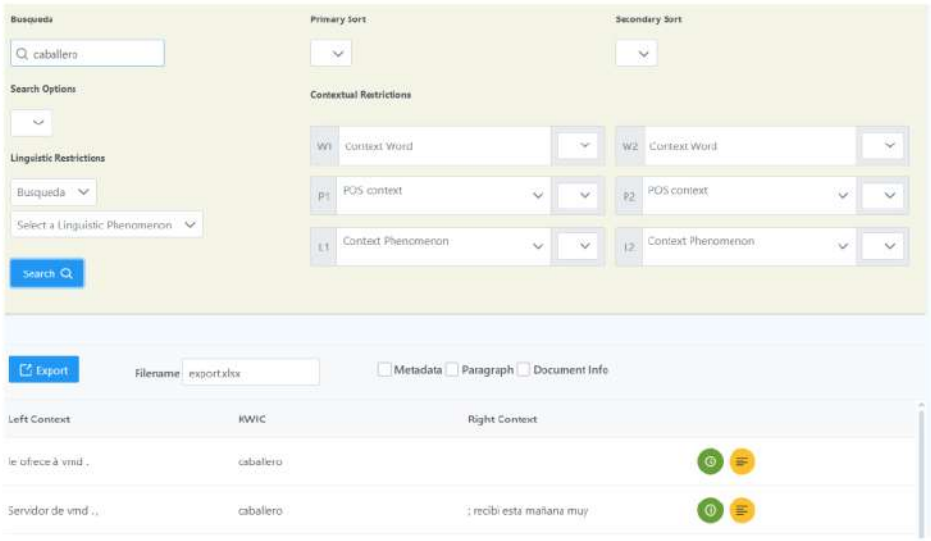

Figura 4: ejemplo de concordancia en PROLEGRAMES (caballero). Elaboración propia. [3]

3.2.1.5. Colocaciones

Las colocaciones constituyen una funcionalidad que permite "calcular las colocaciones, entendidas como patrones característicos de coocurrencias de las palabras" (Menéndez-Barzanallana s.p), es decir, "calcula la probabilidad de que las dos palabras (x y z) aparezcan juntas, calculando la probabilidad de que x y z aparezcan de forma independiente y después compara los dos valores" (Pérez s.p). Estas colocaciones se visualizan mediante índices ordenados por orden alfabético o a partir de una lista de frecuencia ordenada de mayor a menor frecuencias. Por último, "la frecuencia de asociación denominado *índice de información mutua* (MI Score), en el que se mide la fuerza de asociación entre dos palabras" (Menéndez-Barzanallana s.p).

[3] Datos recuperados de: https://prolegrames-iump.ucm.es

Coapariciones Clase de palabra por la ...

Lema	Cat.	Frec.	MI	T-Score	LL
jurel	sustantivo	47	17,82	6,85	470,35
Don Ernesto	sustantivo	10	17,02	3,16	94,4
arenque	sustantivo	18	16,27	4,24	161,64
bonito	sustantivo	20	16,2	4,47	178,79
sardina	sustantivo	74	15,76	8,6	646,26
boquerón	sustantivo	11	15,07	3,31	90,56
atún	sustantivo	48	14,6	6,92	383,66
salmón	sustantivo	32	13,99	5,65	243,18
merluzo	sustantivo	11	13,65	3,31	81,05
trucho	sustantivo	12	13,22	3,46	85,3

Figura 5: ejemplo de colocaciones en CORPES (caballa). Elaboración propia.[4]

3.2.1.6. Búsqueda por anotación lingüística

La búsqueda por anotación lingüística es relativa a las consultas que se pueden realizar en los corpus cuyos fenómenos lingüísticos hayan sido anotados (Menéndez-Barzanallana s.p). Esta anotación puede ser un proceso de lematización o un proceso realizado a partir del etiquetado con TEI de fenómenos lingüísticos. Además, puede realizarse de forma manual, como en el corpus PROLEGRAMES o de forma automática por medio de "programas de ordenador denominados *taggers* que pueden efectuar el proceso automáticamente" (Menéndez-Barzanallana s.p) como en el corpus GITHE con TEITOK. La ventaja de los corpus anotados sobre el resto es que permiten hacer búsquedas dentro de los textos más focalizadas e implementar nuevas funcionalidades como las consultas CQL o la generación de mapas geográficos (Menéndez-Barzanallana s.p).

[4] Datos recuperados de: https://www.rae.es/corpes/

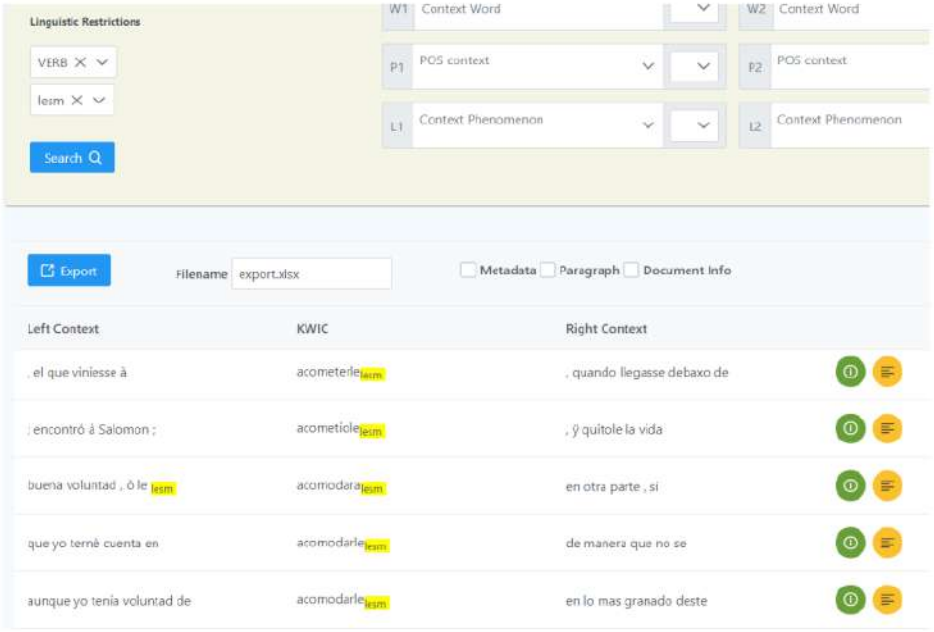

Figura 6: ejemplo de búsqueda por anotación lingüística en PROLEGRAMES (leismo). Elaboración propia. [5]

3.2.1.7. CQL (Corpus Query Language)

Corpus Query Language es un lenguaje de búsqueda creado por *Sketch Engine* que permite buscar dentro del corpus patrones complejos tanto gramaticales como léxicos (Jakubíček, et al. s.p). Este lenguaje surgió en un principio para uso propio en la plataforma de *Sketch Engine*, pero –actualmente– CQL se está implementando en otras interfaces de análisis de corpus como TEITOK o *Corpus Tool* entre otros.

[5] Datos recuperados de: https://prolegrames-iump.ucm.es

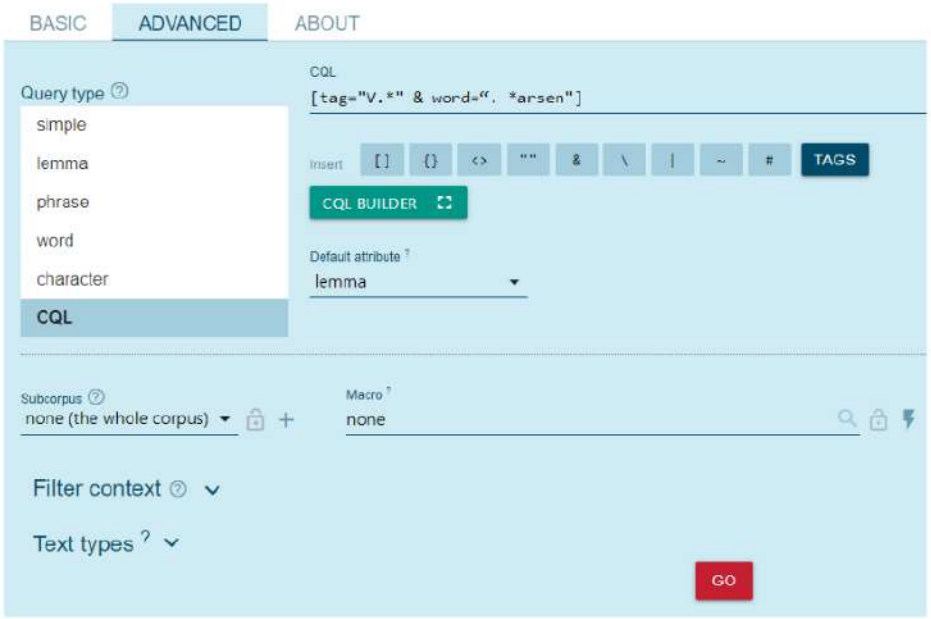

Figura 7: ejemplo de búsqueda con CQL en Sketch Engine (infinitivos en -arsen). Elaboración propia. [6]

3.2.2. *Ejemplos de herramientas de análisis de corpus*

En los siguientes apartados presentaremos brevemente tres ejemplos de herramientas para el análisis de corpus, dos que se relacionan con el corpus PROLEGRAMES, *WordSmith Tools* y *LYNEAL,* y una que actualmente está en auge entre los corpus históricos, *TEI:TOK*. Además de presentar unos ejemplos de corpus diacrónicos que tienen sus propias plataformas de análisis de corpus.

3.2.2.1. WordSmith Tools

WordSmith Tools es un *software* de análisis de corpus para encontrar patrones de palabras en los textos. Esta plataforma ofrece tres herramientas que son *Concord* para buscar casos de palabras o construcciones, *KeyWords* para buscar palabras en un texto o en varios y *WordList* que sirve para crear listas de palabras ordenadas en orden alfabético o por frecuencia. Además, *WordSmith Tools* hace la lematización de forma automática, lo cual resulta muy conveniente a la hora de realizar corpus *ad hoc* (Scott s.p).

[6] Datos recuperados de: https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Festenten18_fl6

Cabe destacar que *WordSmith Tools* ha sido el software preferente del proyecto PROLEGRAMES para la realización de las consultas, pues el etiquetado se pensó para ser reconocido por esta plataforma.

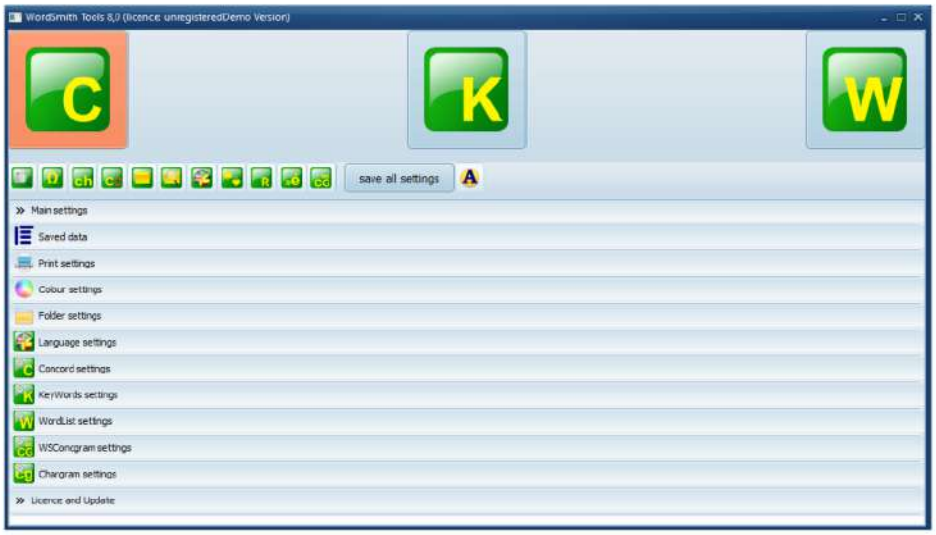

Figura 8: interfaz de WordSmith Tools. Elaboración propia.

3.2.2.2. LYNEAL

LYNEAL (Letras y Números en Análisis Lingüísticos) "es un sistema en línea de análisis de textos que facilita a los procesamientos de datos textuales tanto de los archivos almacenados en el servidor como los propios del usuario" (Ueda s.p). Esta plataforma ofrece "una amplia gama de herramientas de análisis, especializándose en el análisis gráfico de documentos" (Sánchez "Repercusión" 36).

Además, cabe señalar que el corpus PROLEGRAMES se puede consultar bajo el nombre de PROGRAMES en LYNEAL. No obstante, como esta plataforma está pensada para el análisis gráfico, aprovecha bien la transcripción paleográfica de los Documentos PROGRAMES, pero no sirve para aprovechar el etiquetado morfosintáctico. una de las razonas por las que el proyecto necesitaba de su propia interfaz de consultas (Sáez Rivera "Algunas posibilidades" 147).

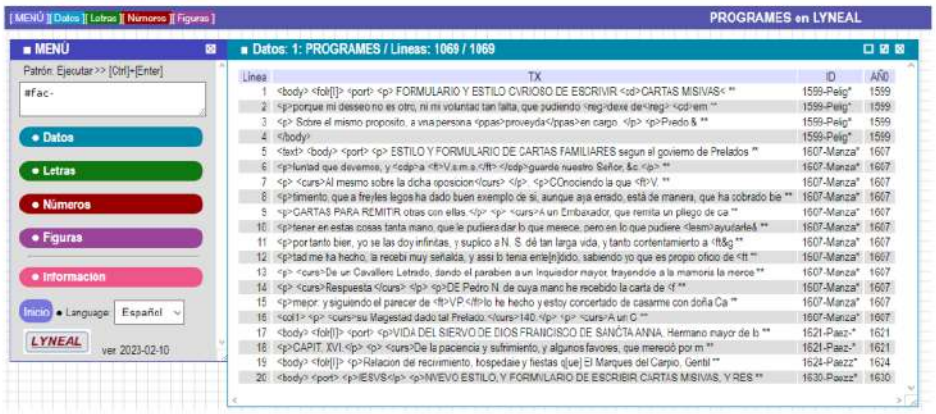

Figura 9: corpus PROGRAMES en LYNEAL. Elaboración propia.

3.2.2.3. TEITOK

TEITOK, *The Tokenized TEI Environment,* es una plataforma web diseñada para visualizar, crear y editar corpus a partir de la anotación textual y el etiquetado lingüístico mediante el uso de XML-TEI. En esta plataforma, se ofrecen dos opciones: una para los usuarios, que es una interfaz gráfica que permite ver los textos de diferentes formas; y otra para los administradores del corpus con la misma interfaz gráfica, pero que además permite editar los datos en XML para que puedan corregir el corpus mientras lo consultan. Actualmente más de una treintena de proyectos han implementado TEITOK en sus corpus (Janssen s.p).

La popularidad de TEITOK entre los corpus históricos radica en su sistema automático de *tokenización* y lematización, en su versatilidad a la hora de poder modificar y compatibilizar los archivos XML-TEI con otras herramientas y plataformas y, además, permite hacer una edición textual de una transcripción a la vez que la edición lingüística (Janssen s.p). Cabe destacar de este sistema que, aunque TEITOK sigue un estándar TEI, emplea su propio sistema de marcado de palabras con <tok>, en vez de usar <w>. En cuanto a TEITOK, la plataforma ofrece las siguientes funcionalidades: la creación de árboles sintácticos tanto en formato SVG como en formato de corchete; una herramienta de reconocimiento automático de la lengua en que está escrita el texto; y, por último, ofrece la posibilidad de poder hacer consultas con CQL (Corpus Query Language) (Janssen s.p).

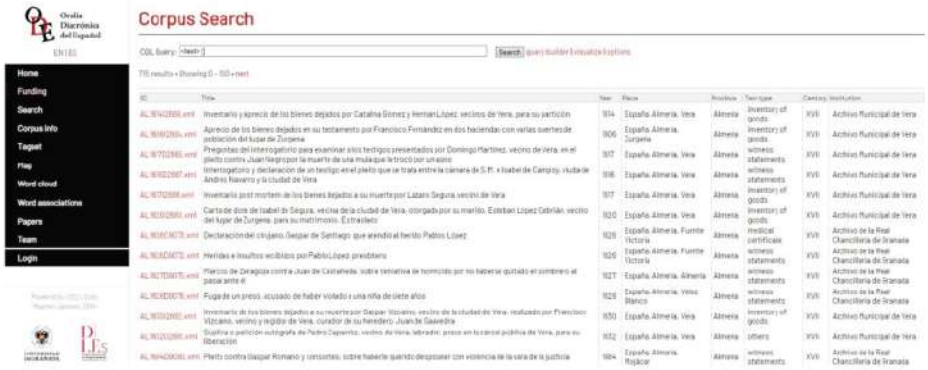

Figura 10: ejemplo de interfaz realizada con TEITOK (ODE). Elaboración propia.

3.2.3. Ejemplos de corpus diacrónicos con herramientas de análisis de corpus propias

Los corpus diacrónicos son corpus históricos que recogen textos y los organizan en diferentes etapas temporales sucesivas con el fin de poder observar las evoluciones que se producen en la lengua (Torruella 45). El corpus PROLEGRAMES es un corpus diacrónico que actualmente está desarrollando su propia interfaz de consulta para consultar los Documentos PROGRAMES. Por este motivo, hemos considerado pertinente presentar algunos corpus diacrónicos que poseen sus propias herramientas de análisis de corpus y así mostrar las funcionalidades más utilizadas en las interfaces de consulta de un corpus diacrónico. A continuación, se presentarán brevemente cuatro corpus diacrónicos y sus respectivas herramientas de análisis de corpus.

3.2.3.1. CORDE

CORDE, acrónimo de Corpus Diacrónico del Español, es un corpus diacrónico textual creado por la Real Academia Española (RAE) que recoge textos desde los primeros inicios de la lengua hasta el año 1974 de todas las variedades de geográficas, históricas y genéricas y múltiples tipos textuales. Actualmente cuenta con más de 250 millones de textos (RAE s.p). El sitio web del corpus posee las siguientes funciones para el análisis del corpus: consultas simples y consultas avanzadas por medio de filtros y operadores lógicos, clasificación de los resultados a partir de los autores y obras y visualización de los datos de las concordancias a partir de la clasificación por contexto y agrupaciones (RAE s.p).

3.2.3.2. CORDIAM

CORDIAM, acrónimo de Corpus Diacrónico y Diatópico del Español de América, es un corpus textual especializado en textos escritos en América desde el siglo XV hasta

el siglo XVIII en que se divide en tres subcorpus: CORDIAM-documentos, CORDIAM-Literatura y CORDIAM-Prensa. La interfaz de CORDIAM posee las siguientes características para el análisis del corpus: consultas simples o avanzadas, filtrado y ordenamiento de búsquedas a partir de los metadatos de los documentos, información cuantitativa del universo de palabras sobre el cual se ha realizado una búsqueda, visualización y guardado del documento completo y posibilidad de guardar automáticamente las concordancias seleccionadas en una base de datos (AML s.p).

3.2.3.3. DIACOM-es y OCCOR-es

DIACOM-es, acrónimo de Diacronía y Comercio, es un corpus textual del español especializado en el comercio internacional desde 1850 a 2018 e incorpora textos españoles tanto de América como de Europa (De Beni et al. s.p). Por otra parte, OCCOR-es, acrónimo de Occidente y Oriente, es un corpus textual del español de textos especializados en temas relacionados con Asia Oriental, especialmente China desde 1850 a 1939 (De Beni et al. s.p). Estos dos corpus pueden consultarse mediante la plataforma *Kontext*. Esta interfaz permite buscar a partir de consultas simples o consultas avanzadas que pueden filtrarse en función de parámetros lingüísticos y/o extralingüísticos. La visualización de los datos es en formato KWIC y, además, se ofrecen estadísticas relativas a su distribución y contextos ampliados (De Beni et al. s.p).

4. Metodología

El análisis del Corpus PROLEGRAMES se realizará a partir de la combinación de tres metodologías que son el análisis DAFO (4.1) que servirá para evaluar de forma estratégica el proyecto de investigación PROLEGRAMES con el objetivo de identificar de forma objetiva sus fortalezas, debilidades, oportunidades y amenazas; la lingüística de corpus (4.2) se empleará para la realización de un análisis detallado del corpus PROLEGRAMES; y finalmente, se empleará el análisis heurístico (4.3) para observar y analizar la interfaz y la usabilidad de la nueva herramienta en línea del proyecto.

4.1. Análisis DAFO

El análisis DAFO, acrónimo de Debilidades, Amenazas, Fortalezas y Oportunidades[7], es una herramienta que permite identificar y evaluar todos los aspectos

[7] En inglés se conoce como A*nalysis SWOT (Strengths, Weaknesses, Oportunities, Threats)*.

positivos y negativos de una organización. Esta metodología consiste en el desarrollo de un plan estratégico, un plan de negocios o un estudio de mercado a partir de la realización de un estudio de las condiciones reales de la organización y esta información servirá para mejorar la toma de decisiones. El análisis DAFO, además, permite observar las condiciones reales en que se encuentra la organización mediante dos análisis diferentes: el análisis externo y el análisis interno (Infoautónomos s.p, Pérez Capdevila 2, Santamaría y Alcalde 287).

El análisis externo según Santamaría y Alcalde "permite analizar las amenazas y oportunidades del sector donde se sitúa la organización para así anticiparse a ellas y poder superarlas o aprovecharlas según las circunstancias que se desarrollen" (288).

- **Amenazas**. Se relaciona con las limitaciones, bloqueos o riesgos que no son inmediatos, pero sí potenciales y que repercuten negativamente en la organización a largo plazo. Es crucial localizar las amenazas de una organización en un estado temprano y establecer un plan de contingencia y poder así anticiparse a ellas y reducir los riesgos que conllevan (Infoautónomos s.p).
- **Oportunidades**. Se refieren a los recursos, competencias, capacidades o situaciones favorables que, al igual que las amenazas, no son inmediatas, pero sí potenciales. Por tanto, es importante identificar las oportunidades y asegurar su incorporación en el futuro (Infoautónomos s.p).

Por su parte, el análisis interno localiza las fortalezas y debilidades, es decir, permite reconocer los factores internos de la organización con el objetivo de fortalecer y mejorar aquellos que no son eficientes (Santamaría y Alcalde 288).

- **Fortalezas**. Son los recursos que ya se disponen dentro de una organización y que permiten alcanzar los objetivos de forma más eficiente. La identificación y utilización de estas fortalezas facilitan el progreso y, además, contribuyen a mitigar el daño que puedan causar las amenazas y las debilidades (Infoautónomos s.p).
- **Debilidades**. Alude a las limitaciones o carencias internas que dificultan el progreso de la organización. Es importante identificar las debilidades de la organización para que a la hora de planificar se puedan desarrollar estrategias que minimicen su impacto o, en el mejor de los casos, las eliminen (Infoautónomos s.p).

Por último, hay que mencionar que el análisis DAFO a menudo se presenta de forma visual en forma de matrices DAFO, que mejoran la visualización y reconocimiento de los distintos aspectos tratados.

Figura 11: ejemplo de matriz DAFO. Elaboración propia.

Por consiguiente, la aplicación del análisis DAFO (5.1) al proyecto de investigación PROLEGRAMES podrá ofrecer una visión del estado actual del proyecto a la vez que sus resultados servirán para crear planes de futuro que lo fortalezcan y que contribuyan a mejorarlo.

4.2. Lingüística de corpus

Por tanto, para la presentación y el análisis del corpus PROLEGRAMES, utilizaremos la lingüística de corpus tomando de referencia la clasificación propuesta por Joan Torruella Casañas de su manual *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación lingüística.*

La lingüística de corpus es una rama de la lingüística computacional constituida por un conjunto de metodologías que se relacionan con el diseño, compilación y explotación de corpus para el estudio de la lengua a partir de grandes colecciones de textos que recogen muestras reales de uso de la lengua (Torruella 25; García-Miguel 11). Los orígenes de esta disciplina se remontan al año 1964 con la aparición del primer corpus, el corpus Brown que contaba con alrededor de un millón de palabras y a partir de él, empezaron a surgir múltiples corpus de referencia (García-Miguel 12). El objetivo

principal de esta disciplina es "encontrar en la colección de textos patrones regulares de uso que nos ayuden a entender la estructura y funcionamiento de un sistema lingüístico o a entender prácticas sociales, actitudes e ideologías que se reflejan en los usos lingüísticos" (García-Miguel 15). Con el aumento de la accesibilidad a textos y vestigios de la lengua y gracias a la evolución de las herramientas informáticas para el procesado y análisis masivo de estos datos, es fundamental la creación y clasificación de los corpus a partir de unos criterios de selección de textos que nos permitan conocer la representación del área de la lengua que pretende mostrar así los sesgos que pueda tener (García-Miguel 16). A continuación, procederemos a presentar los criterios de clasificación según Torruella (41).

Los parámetros clasificatorios más generales son los siguientes:

- **Modalidad**: orales, escritos o mixtos.
- **Temática**: general o especializado.
- **Época**: contemporáneos o históricos.
- **Temporalidad**: sincrónicos, diacrónicos.
- **Magnitud**: grande, restringido o pequeño.
- **Evolución**: abierto, cerrado o monitor.
- **Distribución**: proporcional o equivalente.
- **Número de ediciones**: monoedición o pluriedición, que se subdividen en comparables o paralelos.
- **Número de lenguas**: monolingüe o plurilingüe.
- **Tipo edición**: facsímil, diplomática o paleográfica, diplomática-interpretativa, normalizada, crítica o multiedición.
- **Muestras**: textual, de referencia o léxico.
- **Marcaje**: simples o etiquetados, dentro de este último pueden ser codificados o anotados. También dentro de los anotados se pueden subdividir en corpus anotados morfológicamente (*tagged*), lematizados, parentizados (*shallow* o *partial parsing*) o analizados (*full parsing*).

A continuación, presentaremos los ejes principales:

- **Eje temporal**. Permite facilitar la descripción de periodos desde un punto de vista lingüístico y de los cambios que se producen en cada periodo, así como poder constatar un estado de lengua con otro. Para establecer este eje, en primer lugar,

se han de establecer los límites temporales y establecer cuando empieza y termina. Para los corpus diacrónicos, además, es importante establecer una periodización, que puede ser interna o externa, dentro de esos límites temporales que hemos mencionado anteriormente (Torruella 67).

- **Eje diatópico**. Hace referencia a la geografía lingüística del corpus y a cómo se distribuyen los fenómenos lingüísticos en el territorio. Este eje es complicado de clasificar dado que los autores, entendidos como las personas que dictaban a los escribanos los mensajes que querían transmitir, en ocasiones no eran los escritores de los textos. Esto hace que no se vea reflejado o que resulte ambigua la variedad dialectal que se emplea en la escritura, pues un dialectismo podría ser del autor o del propio escriba (Torruella 85).
- **Eje tipológico**. Se refiere a la perspectiva comunicativa en la que se ha realizado el texto de acuerdo con los registros de la lengua que se recojan en los textos, pues no todos los tipos textuales son iguales, ya que no son fijos. Esta clasificación ayuda a observar si los cambios y las variaciones ocurren en unos tipos en concreto o si se han extendido a otros. Además de que permiten observar si estos cambios ocurren de forma generalizada en la lengua común o solo se mantienen en ciertos círculos reservados (Torruella 100).

Los criterios de selección de documentos para un corpus habitualmente son los siguientes, aunque varían dependiendo qué tipo de corpus se esté formando:

- **Autoridad**. Se refiere a que de entre todos los textos, el que se debe priorizar es el manuscrito original o el más cercano a él en caso de que no hubiera un original, ya que refleja mejor la intencionalidad del autor, así como su estilo y forma de escritura (Torruella 154).
- **Integridad**. Consiste en elegir siempre el documento más completo, aunque esto depende del tipo de corpus, ya que en los corpus de referencia solo se necesita un fragmento, mientras que en el corpus textual el texto debe estar completo (Torruella 154).
- **Comprensibilidad**: elegir el documento más claro o fácil de comprender (Torruella 154).
- **Accesibilidad**. Se debe escoger el documento que sea más fácil de acceder y se priman aquellos de libre acceso o que no tienen derechos de autor (Torruella 154).

- **Edición**. Consiste en adaptar la edición a las necesidades del proyecto y por tanto es imprescindible que cada proyecto presente sus normas de edición, no solo para que los lectores lo tengan en cuenta, sino también los programas puedan reconocer los documentos para su futuro análisis (Torruella 154).
- **Soporte.** Es el medio o soporte en el que está recogido el documento. Actualmente el formato más preferido y utilizado es el electrónico (Torruella 154).
- **Economía**. Se debe elegir las opciones más asequibles desde el punto de vista financiero, aunque esto depende enteramente de la financiación del proyecto (Torruella 155).
- ***Copyright***. El documento libre de derechos de autor o en caso de tenerlos respetarlos (Torruella 155).

La filiación de los documentos, también conocida como metadatos, es imprescindible para brindarle al lector información sobre la obra. Lo que, es más, permite, dependiendo del etiquetado, a los programas y herramientas filtrar y clasificar los documentos en subgrupos o realizar búsquedas con diferentes criterios a partir de estos metadatos (Torruella 156; García-Miguel 17). Los metadatos imprescindibles en un texto son los siguientes:

- **Título**. Para el título de la obra se prefiere elegir la forma más estándar de los nombres y si empieza por un artículo, preposición, etc., se prefiere que se quiten para agrupar todos los documentos que empiecen por la misma palabra (Torruella 157).
- **Autor.** Es imprescindible que solo haya una variante de nombre para los autores y filiar cada documento con ese nombre de autor para que no haya confusiones. Los criterios para los nombres de los autores son: nombre completo, el orden es preferible que sea 1º apellido + 2º apellido + nombre(s), no utilizar "varios autores" o similar e indicar a todos los autores. Se debe utilizar "anónimo" en caso de que no haya un autor concreto. Los nombres compuestos tienen que desarrollarse y se deben mantener los guiones. Las partículas quedan detrás de los nombres de pila y no se tendrán en cuenta los antenombres en la alfabetización, por lo que habría que colocarlos delante de los nombres mientras que los sobrenombres se tendrán que colocar detrás de los nombres (Torruella 158).

- **Fecha**. Lo óptimo es utilizar la fecha original de la obra original o del testimonio que hayamos utilizado. En caso de que no hubiera una fecha concreta, se puede utilizar una fecha aproximada (Torruella 159).
- **Tipo textual**. Se refiere a la clasificación tipológica de los textos (Torruella 162).
- **Dialecto**. Consiste en separar los documentos a partir de la variedad dialectal que se utilice (Torruella 162).

Tras la selección y creación del corpus viene la preparación de los textos que consiste en hacer una unificación de los documentos, primero para que sean homogéneos desde el punto de vista filológico del proyecto (edición textual) y segundo para que sean aptos para que un programa informático pueda interpretarlos adecuadamente (edición digital). La edición textual se refiere a editar textos según los criterios y normas de transcripción y de edición de los proyectos (Torruella 179). No obstante, aunque es habitual que los corpus sigan los criterios de estandarización de Red CHARTA o CORDE, existen corpus que han establecido sus propios criterios de edición como es el caso del corpus PROLEGRAMES. Por otro lado, la edición digital consiste en hacer una edición informática por la cual los editores pueden ofrecer al lector la mayor cantidad de información relevante a partir del etiquetado. En filología, la llamada edición filológica digital incluye habitualmente una edición digital o no de un texto, la codificación de los elementos estructurales y una anotación de los elementos lingüísticos. Dependiendo de la profundidad y la exhaustividad del etiquetado, el texto dispondrá de mayor o menor información (Torruella 190). Los principios para hacer un buen etiquetado serían estos:

- No escribir para software o hardwares específicos y en un lenguaje de marcado no estándar (Torruella 194).
- Mantener la diferencia de contenido entre los datos y la presentación (Torruella 194).
- Evitar innovaciones técnicas innecesarias (Torruella 194).

El cumplimento de estas normas es fundamental, pues como nos indica Torruella, con ellas se pretende hacer que los textos sean reutilizables, es decir, que sean compatibles con cualquier herramienta; y para que los textos etiquetados sean atemporales, es decir, que sirvan independientemente del paso del tiempo y que nunca estén desactualizados (195).

Las marcas de etiquetado más frecuentes y reconocidas entre la edición lingüística de textos se basan en TEI (*Text Encoding Initiative*). "TEI es un consorcio que desarrolla y mantiene colectivamente un estándar de representación de textos en formato digital[8]". (Scott s.p). Actualmente más de 200 proyectos utilizan el etiquetado TEI[9] además de que ha servido de inspiración para otros proyectos, entre los que incluimos el corpus PROLEGRAMES, para la creación de estándares propios basados en TEI.

Por último, dentro de la edición digital filológica encontramos la edición lingüística, que como su nombre nos indica, consiste en anotar en el texto los fenómenos lingüísticos que previamente se han delimitado e indicado para estudiar en el proyecto, pues no siempre es viable estudiar todos los fenómenos de un texto. Este tipo de anotación se hace a partir de la lematización, que consiste en asignar un lema a cada unidad léxica y la categorización, que consiste en asignarle una categoría gramatical (Torruella 224).

4.3. Heurística de interfaces web

Como hemos mencionado anteriormente, la nueva herramienta de análisis del corpus de PROLEGRAMES es una plataforma en línea, lo que implica que deba ser considerada y analizada como una interfaz web. Para el análisis, nos hemos basado en los nueve principios de la heurística que propusieron Rolf Molich y Jakobs Nielsen en 1990. Actualmente consisten en diez principios, tras haber sido revisados y replanteados por Nielsen años más tarde.

Antes de presentar los diez principios, cabe mencionar que la heurística según Molich y Nielsen "es un método informal de análisis de la usabilidad"[10] (249) que consiste en "observar una interfaz para intentar formar una opinión sobre lo bueno o malo que tiene una interfaz"[11] (249). A continuación, se mostrarán los diez principios en los que se basa la heurística aplicada a interfaces web:

- **Visibilidad del estado del sistema**[12]. La aplicación o web debe mostrar en todo momento al usuario el estado del sistema y en el punto en el que se encuentra dentro de él. Por ello es imprescindible que haya una buena comunicación entre el sistema y

[8] Versión original: *The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.*
[9] Datos extraídos de https://tei-c.org/Activities/Projects/
[10] Versión original: *is an informal method of usability analysis.*
[11] Versión original: *looking at an interface and trying to come up with an opinión about what is good and bad about an interface.*
[12] Versión original: *Visibility of System Status.*

el usuario, proveyéndole con mapas interactivos que adviertan su posición o mostrando el flujo de pasos que ha seguido y que tendrá que seguir (Nielsen s.p).

- **Relación entre el sistema y el mundo real**[13]. Se refiere a que el lenguaje que se muestra en la plataforma debe ser el mismo que utiliza el usuario, por lo que se deben evitar conceptos técnicos y emplear un lenguaje familiar para el usuario. También se recomienda utilizar iconos que faciliten el entendimiento de forma visual (Nielsen s.p)
- **Dejar control y libertad al usuario**[14]. Es importante dejar total libertad al usuario, aunque este se equivoque al realizar acciones. Por tanto, es necesario crear opciones de salida como botones de "Cancelar" o "Deshacer y Rehacer" que ayuden a los usuarios a deshacer y volver al estado anterior de cometer el error (Nielsen s.p).
- **Consistencia y estándares**[15]. Los usuarios no necesitan aprender acciones diferentes para entender cada plataforma que utilizan. Así pues, es necesario que se utilicen los mismos estándares y repetir patrones, como colores o iconos, para no confundir al usuario y establecer una cohesión dentro de la aplicación (Nielsen s.p).
- **Prevención de errores**[16]**.** Es importante realizar detecciones de los errores de forma anticipada para que se puedan resolver antes de causar un problema real. Algunas formas para prevenir errores son, por ejemplo, los mensajes de validación de los formularios, los mensajes de fallo de envío por un fallo de red o los mensajes de mantenimiento de las aplicaciones (Nielsen s.p).
- **Reconocer antes que recordar**[17]**.** Este principio se refiere a que hay que facilitar la información del uso de la interfaz a los usuarios de tal forma que no tengan que recordar acciones anteriores (Nielsen s.p).
- **Flexibilidad y eficiencia de uso**[18]. Los atajos de teclado o *shortcuts* son herramientas que facilitan el uso y mejoran la eficiencia de la plataforma favoreciendo las experiencias de los usuarios más experimentados sin perjudicar al resto (Nielsen s.p).
- **Estética y diseño minimalista**[19]. El diseño de la interfaz debe ser claro, sin espacios para elementos irrelevantes que pueda invisibilizar la información relevante (Nielsen s.p).

---

[13] Versión original: *Match between System and the Real World.*
[14] Versión original: *User Control and Freedom.*
[15] Versión original: *Consistency and Standards.*
[16] Versión original: *Error Prevention.*
[17] Versión original: *Recognition Rather Than Recall.*
[18] Versión original: *Flexibility and Efficiency of Use.*
[19] Versión original: *Aesthetic and Minimalist Design.*

- **Ayudar a los usuarios a reconocer, diagnosticar y solucionar los errores**[20]. Los mensajes de error deben expresarse, como se aclaró en segundo principio, de una forma que el usuario pueda entenderlo y sea fácil para él identificar el error y solucionarlo con o sin ayuda. Es preferible evitar los mensajes de error con códigos (Nielsen s.p)
- **Ayuda y documentación**[21]. A pesar de que el diseño en sí mismo debería ser suficiente para entender el uso de la interfaz, es crucial documentar y crear un espacio de ayuda al usuario para ayudarle en caso de que lo necesite (Nielsen s.p).

El análisis de la interfaz web de la herramienta en línea (5.4) es fundamental para mejorar y optimizar la usabilidad y la experiencia de los futuros usuarios.

5. Presentación de la investigación

En este apartado se procederá a mostrar el análisis DAFO del corpus PROLEGRAMES, realizado a partir de la evaluación de sus debilidades, amenazas, fortalezas y oportunidades. Este análisis nos permitirá indicar tanto los aspectos que necesitan mejorar como aquellos que pueden ser explotados y optimizados en beneficio del corpus. Después de este análisis se presentará el corpus PROLEGRAMES, así como la posibilidad de explotación del mismo a partir de su herramienta en línea.

5.1. Análisis DAFO

El análisis DAFO del proyecto que observaremos a continuación se basa en parte en las ideas ya planteadas en los trabajos "Algunas posibilidades de explotación" de Daniel Sáez Rivera (2018) y "Repercusión de TEI (*Text Encoding Initiative*) en España: evolución, situación actual y caso de corpus en español (Documentos PROGRAMES)" de Nayra Sánchez Vera (2020), pues en ellos ya se expusieron y plantearon mejoras necesarias para optimizar el corpus.

Por un lado, Sáez Rivera propuso una serie de mejoras que fueron "actualizar y completar la bibliografía sobre gramaticalización y conceptos conexos; crear una interfaz propia de consulta que permita buscar por fenómenos de manera usable o amigable; agilizar con herramientas automáticas la transcripción y etiquetado lingüístico; e incluir

[20] Versión original: Help Users Recognize, Diagnose, and Recover from Errors.
[21] Versión original: *Help and Documentation.*

reproducción facsímil en forma de edición múltiple" ("Algunas posibilidades" 157) Sánchez por su parte, tomando de base las propuestas anteriores, sugirió cambiar la jerarquía de etiquetado que no funcionaban bien con otros programas de análisis que no fuera *Wordsmith Tools*; la creación de una interfaz propia de consulta y análisis; la creación de un programa de automatización de etiquetado de los textos para agilizar y homogeneizar el etiquetado con TEI. Además, manifestó las necesidades de atraer colaboradores para el proyecto; la necesidad de una uniformidad en el etiquetado de los textos; y la necesidad del proyecto de recibir una mayor financiación ("Repercusión" 38). Todas estas propuestas han sido consideradas y adaptadas en nuestro análisis DAFO que de forma visual se representaría de la siguiente forma:

Figura 12: matriz del análisis DAFO del proyecto PROLEGRAMES. Elaboración propia. [22]

[22] Matriz realizada con: https://dafo.ipyme.org/Home

Antes de presentar el análisis de los resultados, hay que señalar que la matriz DAFO anterior se realizó en colaboración con los miembros del proyecto PROLEGRAMES. A continuación, presentaremos las debilidades, fortalezas, amenazas y oportunidades.

5.1.1. Debilidades

Las debilidades se refieren a las limitaciones técnicas y metodológicas que existen dentro de la organización (4.1). Por ello tras un análisis exhaustivo, hemos considerado las siguientes debilidades:

5.1.1.1. Corpus sin normalizar gráficamente

Esta cuestión se refiere a que los documentos únicamente presentan una transcripción paleográfica manteniendo fielmente las grafías originales. Esto dificulta la lematización gráfica y gramatical y, por consiguiente, los documentos tienen que ser etiquetados manualmente reduciendo el riesgo de error que podría conllevar el uso de un etiquetador automático debido a la cantidad de formas léxicas diferentes (Sáez Rivera "Algunas posibilidades" 152). La falta de normalización gráfica también afecta en la lectura, pues si no se está familiarizado con el español no contemporáneo puede dificultarla y más teniendo en cuenta que el corpus PROLEGRAMES ofrece una versión para lectura. No obstante, a falta de una versión facsimilar del texto original, esta carencia permite observar y apreciar las formas gráficas originales.

Por último, hay que mencionar que además de afectarle al texto en sí, los nombres de los autores y de las obras tampoco están normalizados, pues no hay un criterio para su normalización gráfica. Esto hace que tanto las búsquedas como las agrupaciones por autores o títulos se dificulten, pues un mismo autor puede aparecer duplicado porque aparecen varias variantes de su nombre o se dificulte la búsqueda de una obra porque ha mantenido su grafía original.

5.1.1.2. Equipo pequeño y cambiante

Como ya anunciaba Sánchez la falta de personal activo dificulta el progreso continuo del proyecto ("Repercusión" 38). Esto no solo repercute en la distribución de la carga de trabajo entre los colaboradores, sino que también afecta a la homogeneización de los criterios del corpus, pues es necesario reemplazar a los antiguos etiquetadores e invertir tiempo en formar nuevos. También es importante enseñarles bien, pues así se evita tener que invertir mucho tiempo en revisiones y correcciones.

5.1.1.3. Dificultad para fidelizar a los transcriptores/etiquetadores

Esta debilidad va directamente enlazada con la anterior, pues el principal motivo por la que el equipo es pequeño y cambiante es por la falta de fidelización. El corpus PROLEGRAMES desde sus inicios, como indica Sáez Rivera, se ha valido de colaboradores externos, en su mayoría alumnos de Grado y de Máster, doctorandos y jóvenes doctores, para hacer las transcripciones y el etiquetado ("Algunas posibilidades" 146). Esta falta de fidelización afecta sobre todo a los alumnos de Grado, al contrario que ocurre con los alumnos de Máster y doctorandos (Sánchez "Repercusión" 38). Por último, hay que mencionar que otro motivo por el que no se fidelizan es debido a la escasa financiación del proyecto que no puede permitirse contratar a mucho personal.

5.1.1.4. Falta de versiones facsimilares

El proyecto no ofrece una versión facsimilar con la que poder comparar las versiones de análisis y lectura (Sáez Rivera "Algunas posibilidades" 152), haciendo que el lector o bien busque por su cuenta las versiones originales o bien confíe firmemente en la correcta transcripción paleográfica que se haya hecho. Lo que sí es seguro es que esta carencia puede entorpecer las investigaciones al no ofrecer una versión facsimilar con la que se pueda hacer un cotejo, aunque los investigadores internos al proyecto sí disponen de esas versiones facsimilares, que se encargaron en su debido tiempo a diferentes bibliotecas, o últimamente se extraen de versiones facsimilares en línea, fácilmente consultables.

5.1.1.5. Los textos no están en formato .txt

Esta cuestión afecta a la hora de facilitar el trabajo a los investigadores que tienen que utilizar herramientas externas para el análisis, pues en el caso de PROLEGRAMES, los archivos están en formato .pdf o .doc haciendo que los investigadores tengan que cambiar el formato de forma manual. El formato .txt es imprescindible a la hora de manejar cualquier programa informático pues necesita de un texto puro (.txt) que no tenga las codificaciones internas de cada procesador de textos (Torruella 166). El único inconveniente de este formato, que no afecta a los documentos PROLEGRAMES porque cuenta con etiquetas de formato de texto, es que no permite mantener los formatos como las cursivas o caracteres volados. No obstante, esta debilidad no es responsabilidad directa del proyecto, ya que en la página web oficial donde están cargados los documentos no se permite compartir el formato .txt (sí lo permitía en la primera versión en línea del proyecto

como http://portal.ucm.es/web/programes) , y en todo caso la conversión de .doc a .txt resulta fácil.

5.1.2. Fortalezas

Las fortalezas se refieren a las ventajas y aspectos positivos que posee la organización y hemos identificado las siguientes:

5.1.2.1. Etiquetado de fenómenos lingüísticos

El etiquetado metalingüístico, referido a la información del documento y al formato y el etiquetado de fenómenos lingüísticos con TEI permiten "la preservación, la reutilización o posible integración en repositorios digitales" (Vaamonde 48) de cualquier texto etiquetado además de facilitar "enormemente la interoperabilidad del corpus para ser analizado con distintas herramientas y programas de análisis de corpus de lengua basados en los estándares" (Sánchez "Repercusión" 38). En definitiva, una de las mayores fortalezas de los documentos PROGRAMES es su rico etiquetado que permite un estudio en profundidad de cada texto.

5.1.2.2. Autores y textos poco conocidos

Los textos que se decide transcribir y etiquetar en PROLEGRAMES son textos no canónicos o no literarios, "que no se han incluido en ningún corpus lingüístico y que a menudo no tienen una edición moderna o que no cuentan con una transcripción válida para un estudio diacrónico" (Sáez Rivera "Algunas posibilidades" 149). Esto hace que desde el proyecto se dé acceso y visibilidad a textos y a autores a los que se ha prestado poca atención en las investigaciones pero que "son especialmente valiosos para calibrar el grado de acercamiento de la lengua escrita a lo que podría ser la lengua hablada de las distintas épocas" (Sáez Rivera "Algunas posibilidades" 149).

5.1.2.3. Varios textos de un mismo autor y varias ediciones de un mismo texto

Desde el punto de vista lingüístico e histórico es interesante estudiar el uso de la lengua de un autor durante las etapas de su producción, no solo para estudiar las estructuras gramaticales y el léxico que utilizaba, sino que también es interesante observar los cambios e influencias que pudo tener derivado de su contexto social o de otros autores de la época. Del mismo modo pasa con el estudio de varias ediciones de un mismo texto. Adscribiéndonos a la teoría de Octavio de Toledo "las variantes de una tradición textual, […] contienen información potencialmente muy útil a efectos para el esclarecimiento de

las evoluciones lingüísticas" (196). En sus estudios de cotejos de variantes textuales, Octavio Toledo advierte las siguientes generalizaciones:

- "Los testimonios más recientes modernizan los rasgos lingüísticos que en los más antiguos parecen mostrar más apego a la lengua del original (197).
- "La adscripción a una determinada tradición discursiva con creciente elaboración intensiva o la manipulación deliberada de las estructuras discursivas mismas inciden en la clase de información que manifiestan las variantes" (212).
- "[Las variantes] ofrecen también datos muy útiles para comprobar la coexistencia sincrónica de soluciones en competencia e incluso para rastrear las últimas fases de la extinción de un fenómeno" (212).

Sumado a esto, Sáez Rivera, añade que además de poder observar los cambios lingüísticos, también se puede apreciar los usos gráficos y su variación ("Algunas posibilidades" 155).

5.1.2.4. Formación de personal especializado

A pesar de que existe una clara dificultad para fidelizar a los colaboradores, como hemos visto en las debilidades (5.1.1), hay que puntuar positivamente y puntualizar que desde el proyecto PROLEGRAMES se capacitan profesionales especializados en transcripciones y etiquetado con TEI. Como se puede observar en la sección "Miembros" de la página web oficial, actualmente muchos de los antiguos colaboradores han continuado su formación académica y profesional como investigadores en diversas universidades.

Como ejemplos de egresados podemos nombrar a Nayra Sánchez Vera, colaboradora desde 2020, doctoranda actual en la UCM, bajo la dirección de Francisco Javier Herrero Ruiz de Loizaga y Daniel M. Sáez Rivera, de una tesis con el título actual de *La evolución de las formas de tratamiento del español desde el siglo XVI al XIX en el corpus etiquetado PRO(LE)GRAMES;* Mario Casado Mancebo, antiguo colaborador externo de 2013-2014 e investigador becado de lingüística en el departamento de Lengua española y Lingüística de la Universidad Nacional de Educación a Distancia (UNED); Alexandra Duttenhofer, antigua becaria del proyecto durante PROGRAMES5 (2016-2019) doctora en Lengua Española y sus Literaturas por la UCM; Sergio Montalvo Marcca también doctor en Lengua Española y sus Literaturas por la UCM y actual profesor ayudante en esta misma universidad; Manuel Peralta Céspedes doctor en Lengua

Española y sus Literaturas por la UCM y actual profesor del curso de "Sintaxis y Semántica" de la Universidad de Murcia (UM). O Patricia Fernández Martín, doctora en Lenga Española y sus Literaturas por la UCM y Profesora Titular en el Departamento de Filologías y su Didáctica de la Universidad Autónoma de Madrid (UAM), entre otros.

Por esta razón podemos afirmar que PROLEGRAMES contribuye a la formación de expertos en el ámbito de la lingüística.

5.1.2.5. Documentación y divulgación

Sáez Rivera propuso la mejora del proyecto a partir de la inclusión de más bibliografía y conceptos conexos ("Algunas posibilidades" 157). Esta ampliación de la bibliografía se realizó en 2020 bajo el nombre de *Bibliografía PROLEGRAMES* donde no solo se documentan los trabajos citados, sino que también se incluyen las contribuciones de los miembros del equipo. Además de la documentación, desde el sitio web del proyecto, en la sección "Enlaces", se proporcionan una serie de recursos en línea de interés para la investigación de la lengua española como enlaces a asociaciones, atlas lingüísticos, bases de datos, bibliotecas virtuales, catálogos bibliográficos y corpus entre otros recursos que pretenden fomentar, visibilizar y apoyar a la comunidad académica.

5.1.2.6. Accesibilidad

El corpus PROLEGRAMES no es simplemente un corpus lingüístico, sino que también funciona en parte como un archivo de textos. Esto es porque PROLEGRAMES ofrece a todos una versión de lectura de todos sus documentos en formato .pdf donde se eliminan todas las etiquetas de fenómenos lingüísticos manteniendo únicamente las de formato y de este modo facilitar la lectura de los textos para aquellos usuarios que estén interesados en el contenido textual (Sáez Rivera "PROLEGRAMES" s.p). Asimismo, el corpus es de acceso libre y se encuentra subido en línea para lectura y descarga en la página institucional del proyecto.

5.1.2.7. Longevidad

El proyecto de investigación PROLEGRAMES lleva activo desde 2001 y ha pasado por múltiples fases y áreas de estudio:

- "PROGRAMES1: "Procesos de gramaticalización en la historia del español" (MEC, REF: BFF2001-1340, 2001-2004, Investigador principal (IP) José Luis Girón Alconchel)" (Sáez Rivera "Algunas posibilidades" 143)

- "PROGRAMES2: "Procesos de gramaticalización en la historia del español (II): Formación de Variedades, Tipología, Periodización, Criollización" (MEC, REF: HUM04-3610, 2004-2007, IP José Luis Girón Alconchel)" (Sáez Rivera "Algunas posibilidades" 143)
- "PROGRAMES3: "Procesos de gramaticalización en la historia del español (III): gramaticalización, lexicalización y tradiciones discursivas" (MEC, REF: FFI02828/FILO, 2008-2011, IP José Luis Girón Alconchel)" (Sáez Rivera "Algunas posibilidades" 143)
- "PROGRAMES4: "Procesos de gramaticalización en la historia del español (IV): gramaticalización y textualización" (MinEco, REF: FFI2012-31427, 2012-2016, IP José Luis Girón Alconchel)" (Sáez Rivera "Algunas posibilidades" 143)
- "PROGRAMES5: "Procesos de gramaticalización en la historia del español (V): gramaticalización, lexicalización y análisis del discurso desde una perspectiva histórica" (MinEco, REF: FFI2015-64080-P, 2016-2019, IP Francisco Javier Herrero Ruiz de Loizaga)" (Sáez Rivera "Algunas posibilidades" 143)
- "PROGRAMES6: "Procesos de gramaticalización, lexicalización, pragmática y discurso en la historia del español" (Proyecto Santander-UCM, REF: PR108/20-11, 2020-2021, IP, Francisco Javier Herrero Ruiz de Loizaga)" (Sáez Rivera "PROLEGRAMES" s.p)
- "PROLEGAMES: "Procesos de lexicalización y gramaticalización en la historia del español: cambio, variación y pervivencia en la historia discursiva del español" (MinEco, REF: PID2020-112605GB-I00, 2021-actualidad, IP Francisco Javier Herrero Ruiz de Loizaga)" (Sáez Rivera "PROLEGRAMES" s.p)

Esta longevidad no solo confiere prestigio al proyecto, ya que muestra su capacidad de adaptación a los cambios de su entorno y de innovación, sino que también han realizado contribuciones significativas a su área de investigación ininterrumpidamente.

5.1.3. Amenazas

En cuanto a las amenazas (5.1.3), estas se refieren a los riesgos que podrían afectar de forma crítica al proyecto PROLEGRAMES. Antes de presentarlas, cabe apuntar que, la segunda amenaza, la fragilidad de los servidores, afecta actualmente al proyecto, pero no de forma crítica, pues la herramienta en línea aún sigue en fase de desarrollo y no está en funcionamiento.

5.1.3.1. Pérdida de financiación

Esta es la amenaza más crítica para el proyecto, pues al ser un proyecto universitario depende de las becas y ayudas para subsistir. Si llegara a terminar la financiación sería imposible continuar con el proyecto, pues no solo se tendría que prescindir de los colaboradores, sino que no se podría mantener y seguir desarrollando la nueva herramienta en línea que está siendo desarrollada por la empresa Avantopy.

5.1.3.2. Fragilidad de los servidores

Esta amenaza va ligada a la anterior de manera directa, pues se ve afectado directamente por la pérdida de financiación. Primero porque al tratarse de un corpus completamente digitalizado y al contar con una herramienta en línea resultaría catastrófico no poder acceder a los servidores. Por otro lado, la pérdida de recursos financieros afecta a la seguridad de la página web y a la navegación de los usuarios, pues al no poder pagar un certificado de seguridad, los navegadores, que advierten a los usuarios cuando un sitio web no es seguro mostrando alertas del tipo "la conexión no es privada", desalientan el acceso al usuario y provoca una disminución del acceso al corpus y a la herramienta en línea de PROLEGRAMES.

5.1.4. Oportunidades

Las oportunidades, referidas a las posibilidades de expansión, desarrollo y de explotación del corpus (4.1), son las siguientes:

5.1.4.1. Creación de una herramienta en línea propia y explotación de la nueva herramienta en línea de Avantopy.

Tanto Sáez Rivera como Sánchez veían necesario la creación de una interfaz propia de consulta que "permitiera localizar y buscar fenómenos lingüísticos en los textos del corpus y aprovechar así su rico etiquetado morfosintáctico" (Sánchez "Repercusión" 38) sin recurrir a las herramientas externas LYNEAL o *Wordsmith Tools*. Actualmente, como hemos mencionado, la herramienta en línea que presentaremos más adelante en el apartado 5.3 está en vías de desarrollo.

5.1.4.2. Automatización del etiquetado

La automatización del etiquetado a partir de un programa de etiquetado automático presentaría una gran ayuda al proyecto si tenemos en cuenta que está formado por un equipo pequeño y cambiante, ya que "agilizaría la transcripción y el etiquetado lingüístico" (Sáez Rivera "Algunas posibilidades" 157) y se tendría que depender tanto

de los etiquetadores. Siguiendo esta línea, Sánchez propone que, al menos, la futura aplicación pudiera hacer las etiquetas de formato dado que supondría menos dificultad, pues se rigen por criterios más rígidos y no dependen tanto de la lingüística ("Repercusión" 38). Sin embargo, gracias al avance de la inteligencia artificial sería posible la creación de una IA y a partir del *machine learning* entrenarla para que reconozca los fenómenos lingüísticos y poder así etiquetarlos a la vez que se podría entrenar para que también pudiera realizar el etiquetado de formato y la transcripción. Además, durante el entrenamiento de esta futura herramienta también se podría entrenar para que etiquetara aquellos fenómenos lingüísticos que en 2003 y 2005 tuvieron que recortarse por falta de tiempo y fallos de los transcriptores durante el etiquetado (Sáez Rivera "Algunas posibilidades" 13).

En definitiva, la creación de una IA para el proyecto PROLEGRAMES no solo agilizaría los procesos de transcripción y etiquetado, sino que, además, se podría ampliar el etiquetado de los fenómenos lingüísticos del proyecto.

5.2. El corpus PROLEGRAMES

En este apartado procederemos a realizar el análisis del corpus PROLEGRAMES siguiendo los criterios de la lingüística de corpus (4.2).

PROLEGRAMES: "Procesos de Lexicalización y Gramaticalización en la Historia del Español" es un proyecto de investigación heredero de los proyectos PROGRAMES que empiezan en 2001, y que pertenece al Departamento de Lengua Española y Teoría de la Literatura de la Universidad Complutense de Madrid, además de formar parte de la red INTELE (Infraestructuras de Tecnologías del Lenguaje), mientras que su herramienta de consulta en línea se inserta en el marco institucional del IUMP (Instituto Universitario Menéndez Pidal). Actualmente está dirigido por Francisco Herrero Ruiz de Loizaga (que heredó la batuta de José Luis Girón Alconchel) y sus miembros se conforman por investigadores de diversas procedencias y colaboradores externos. Dada su longevidad, el proyecto PROLEGRAMES ha pasado por varias fases donde se han ido estableciendo y concretando los parámetros de estudio, por el momento, nos encontramos en la fase: PROLEGRAMES, "Procesos de lexicalización y gramaticalización en la historia del español: cambio, variación y pervivencia en la historia discursiva del español" (ref. PID2020-112605GB-I00) (Sáez Rivera "PROLEGRAMES" s.p). El proyecto está colaborando con la empresa de IT, Avantopy, para la creación de

una herramienta de consulta en línea propia que facilite el análisis y la búsqueda de los textos del corpus.

5.2.1. Propósito general y delimitación del dominio de especialidad

El propósito general del proyecto de investigación PROLEGRAMES es el estudio del español desde sus orígenes hasta su gramaticalización y lexicalización, centrándose en las etapas menos estudiadas de la lengua, concretamente en "las etapas donde la creación de sistema a partir del discurso es más determinante como como la segunda mitad del siglo XIV y las transiciones del castellano medieval al español clásico y de este al moderno" (Sáez Rivera "PROLEGRAMES" s.p; Sáez Rivera "Algunas posibilidades" 148). Se escogió esta etapa de la lengua porque, a pesar de que es de las menos estudiadas en el campo de la investigación por su falta de prestigio cultural y político venida por supuesta decadencia histórica de la época según cierta historiografía (Sáez Rivera y Octavio de Toledo 19), se considera fundamental para entender el cambio morfosintáctico y las modificaciones en las técnicas de construcción del discurso (Sáez Rivera y Octavio de Toledo 24).

La delimitación del dominio de estudio de los periodos temporales del corpus PROLEGRAMES atienden a la propuesta de José Luis Girón Alconchel, antiguo investigador y director del proyecto, en "Cambios gramaticales en los siglos del oro" (2004):

1) *"1492-1555 ó 1556*: desde el final de los Reyes Católicos a la abdicación de Carlos V; desde la Gramática de Nebrija a la de Villalón y los Anónimos de Lovaina, así como desde la Celestina al Lazarillo" (*apud* Sáez Rivera "Tesis" 91).
2) *"1556-1648*: desde el comienzo del reinado de Felipe II a la Paz de Westfalia, desde las Gramáticas de Lovaina a la del Padre Villar, así como desde el Lazarillo a Gracián (este subperíodo central constituye el verdadero Siglo de Oro de nuestra literatura)" (*apud* Sáez Rivera "Tesis" 91).
3) *"1648-1726*: desde los últimos años del reinado de Felipe V a la mitad del reinado de Felipe V, el primer Borbón; desde el Padre Villar a la Real Academia Española, así como desde Calderón y los epígonos de la literatura barroca a Feijoo" (*apud* Sáez Rivera "Tesis" 91).

Más tarde, se amplió el rango temporal y se empezaron a incluir textos de finales del siglo XVI al siglo XIX (Sáez Rivera "PROLEGRAMES" s.p).

5.2.2. Parámetros clasificatorios generales

5.2.2.1. Muestras, magnitud, evolución y distribución

El corpus PROLEGRAMES está compuesto por 42 textos y 1.690.004 palabras. Según la clasificación de Torruella, el corpus se considera pequeño dado que no supera los 5.000.000 de palabras (48). No obstante, el corpus es abierto y dinámico, es decir, que con cada edición el número de textos y, por consiguiente, palabras, va en aumento.

Como hemos indicado, el corpus PROLEGRAMES es un corpus de textos escritos, es textual, y por eso, todos los textos que lo componen tienen que estar completos. Esto es algo fundamental para los corpus especializados que estudian la lengua ya que "la información lingüística, conceptual y pragmática de las unidades terminológicas o fraseológicas puede aparecer en cualquier parte de un documento" (Beni y Hourani-Martín "El discurso" 90). Cabe señalar, además, que los textos recogidos no se restringen por el número de palabras, sino que cada uno cuenta con una extensión diferente.

5.2.2.2. Lengua

Al ser un corpus que estudia la lengua española, solo encontramos textos escritos en español, es un corpus monolingüe. Sin embargo, podemos encontrar palabras o expresiones extranjeras, normalmente en francés, latín o catalán que se encuentran debidamente marcadas con la etiqueta <foreign></foreign> (Sáez Rivera "Algunas posibilidades" 148).

5.2.2.3. Autores y ediciones

Los textos que se recogen en este corpus como menciona Sáez Rivera "no han sido incluidos en ningún corpus lingüístico del español en red, […] carecen […] de edición moderna, o esta no sigue criterios de transcripción válidos para un estudio diacrónico" ("PROLEGRAMES" s.p). Además, suelen ser de autores "que han despertado poca atención en la historia de nuestra lengua" (Sáez Rivera "Algunas posibilidades" 149) y de "tradiciones discursivas en la que apenas se ha reparado hasta ahora" (Sáez Rivera "Algunas posibilidades" 149). Actualmente, el corpus ha transcrito y etiquetado obras de 23 autores diferentes y 13 obras escritas por autores anónimos. En la siguiente tabla, mostraremos los nombres de los autores:

Autor	Nº de obras
Jacques de Liaño	1
Juan Vicente Peliger	1
Gerónimo Paulo de Manzanares	1
Juan Páez de Valenzuela y Castillejo	4
Diego Ossorio y Basurto	1
Alejandro Domingo de Ros	1
Feliu de la Peña	1
Joseph Barzia y Zambrana	1
Francisco Garau	2
Bartolomé Guerrero Ludeña	1
Carlos Antonio Puertas	1
Gil de la Cotera	1
Juan Muñoz y Peralta	1
Sebastián Fernández de Medrano	1
Antonio López del Águila	1
Luis Salazar y Castro	1
Melchor de Alcázar y Zuñiga	1
Castellví	1
José del Campillo y Cosío	3
J, A. Santiago Rial	1
Francisco Mariano Nifo y Cagigal	2
M. Fernández	1
Carlos Pellicer.	1

Tabla 1: autores y obras. Elaboración propia.

De entre estos autores, hemos de resaltar a Juan Páez de Valenzuela y Castillejo, a Francisco Garau, a José del Campillo y Cosío y a Francisco Mariano Nifo y Cagigal, pues son los únicos autores que cuentan con más de una obra en el corpus. También , cabe mencionar la existencia de dos textos escritos por Francisco Garau que son dos ediciones diferentes de un mismo texto, *La Fee triunfante*…, de 1691 y de 1755. Como hemos mencionado en las fortalezas, en el apartado 5.1.2, tener varias ediciones de un mismo texto y/o varias versiones de un mismo autor, es relevante "para el estudio de la variación y la evolución lingüística" (Octavio Toledo 196) Para este tipo de estudios de cotejos, Octavio Toledo propone el concepto de "mapa variacional" que consiste en examinar las

variantes que siguen patrones tanto regulares como irregulares para así identificar la variación morfosintáctica de los textos (214).

5.2.2.4. Tipología textual

La tipología textual del corpus PROLEGRAMES ha variado a lo largo de sus investigaciones para adaptarse a las necesidades y criterios de cada etiquetador del proyecto, pues no existía una tipología normalizada. El criterio de selección indicaba que se priorizaban los textos no literarios ya que son considerados como "especialmente valiosos para calibrar el grado de acercamiento de la lengua escrita a lo que podría ser la lengua hablada de las distintas épocas" (Sáez Rivera "Algunas posibilidades" 148).

Los tipos de texto que se utilizaban anteriormente eran los siguientes: biografía, carta, diálogo escolar, manual de cartas, narración histórica, prosa jurídico-administrativa, prosa periodística, prosa técnica, relación de auto de fe, relación de sucesos, sermón, tratado y vida de santos (Sánchez "Guía" 11). No obstante, como hemos mencionado, se ha empezado la normalización de la tipología textual para la nueva herramienta en línea y evitar así que los etiquetadores elijan otras diferentes. Los tipos de textos que se manejan actualmente en el proyecto son, en cambio, estos: biografía, carta, consulta, crónica, diálogo, informe, manual de cartas, periodismo, prosa miscelánea, relación, sermón y tratado.

Los tipos de textos se registran en los documentos en la cabecera TEI dentro de <profileDesc></profileDesc>, que contiene "los elementos contextuales del texto original" (Sánchez "Guía" 10), enmarcados por la etiqueta <textClass></textClass>.

5.2.2.5. Producción, edición y marcado de los textos

Los textos del corpus PROLEGRAMES están transcritos manualmente por los etiquetados del proyecto de forma paleográfica manteniendo las grafías, la acentuación y la puntuación originales. El corpus no sigue un criterio estandarizado como podría ser el de red CHARTA, sino que tiene uno personalizado y acorde con la finalidad del estudio del proyecto de investigación.

Esta diferencia se aprecia realmente el desarrollo de las abreviaturas, pues las generales se completan entre corchetes, mientras que las que se refieren a fórmulas de tratamiento no lo hacen. Esto es porque la teoría que plantea Sáez Rivera es que en la primera mitad del siglo XVIII "una abreviatura de tratamientos como *v.m* podría tener un desarrollo oral espontáneo como *usted* pero también como *vosa merced* (o más

probablemente *vuesa merced*)" (Sáez Rivera y Octavio Toledo 29) además de que a ello se le suma la idea de que "el desarrollo de las abreviaturas de cortesía en ediciones modernas suponen un falso problema, porque estas deberíamos valorarlas como una variante más de tratamiento" (Sáez Rivera y Octavio Toledo 29).

Tras la transcripción paleográfica, se crean las dos versiones (lectura y análisis) de los textos que se ofrecen en el corpus. La versión para lectura constituye un texto en formato .pdf pensado para su lectura a partir de un etiquetado únicamente de formato que refleja fielmente la forma del texto original. Por su parte, la versión para análisis se ofrece en un documento .doc, ya que la página web oficial del proyecto no permite compartir los documentos en texto plano .txt, con un encabezado TEI con todos los metadatos y con todas las etiquetas de codificación de fenómenos.

5.2.2.6. Marcaje de los textos

El corpus PROLEGRAMES es un corpus etiquetado cuyo marcaje textual se hace siguiendo un estándar TEI adaptado a las necesidades del proyecto y que se puede consultar como *Guía de etiquetado: Documentos Programes* por Nayra Sánchez Vera (2020). Debido a que están etiquetados con un estándar TEI adaptado, que permite que los documentos puedan ser leídos por cualquier herramienta que soporte el lenguaje XML-TEI como, por ejemplo, *WordSmith Tools* o *AntConc*.

El marcaje de los textos de PROLEGRAMES cumple enteramente con dos de los tres principios de marcado que expone Torruella (190) vistos en el apartado 4.2. La anotación distingue entre metadatos, formato y fenómenos lingüísticos y no añade innovaciones a las etiquetas, solo adapta los nombres de las mismas para que encajen con las investigaciones como por ejemplo hacen con las cursivas <curs></curs>, en vez de utilizar las etiquetas de HI REND en CREA o CORDE. No obstante, respecto al último principio, las etiquetas de PROLEGRAMES fueron adaptadas al uso inicial de *WordSmith Tools*, y, por eso, como veremos a continuación, fol. además indicar si es recto o vuelto aparece con un número detrás, pues esa herramienta solo detecta números al final de la etiqueta (Sánchez "Guía" 14). A pesar de todo ello, como las etiquetas utilizan el estándar de XML-TEI, pueden ser reconocidas por cualquier programa que reconozca este lenguaje. A continuación, mostraremos el proceso de etiquetado de formato y lingüístico.

En un primer momento, los textos ya transcritos, son marcados con las etiquetas de formato que permiten señalar la forma original de un texto:

Etiqueta	Descripción
<pag#></pag#>	Página
<folr#></folr#>	Folio (recto)
<folv#></folv#>	Folio (vuelto o reverso)
<folr[I]></folr[I]>	Hoja (recto) con número romano añadido
<folr[II]></folr[II]>	Hoja (recto) con número romano añadido
<folr[III]></folr[III]>	Hoja (recto) con número romano añadido
<folr[IV]></folv[IV]>	Hoja (recto) con número romano añadido
<folv[I]></folv[I]>	Hoja (vuelto) con número romano añadido
<folrvII]></folv[II]>	Hoja (vuelto) con número romano añadido
<folv[III]></folrvIII]>	Hoja (vuelto) con número romano añadido
<folv[IV]></folv[IV]>	Hoja (vuelto) con número romano añadido
<col#></col#>	Columna
	Párrafo
<foreign></foreign>	Lengua extranjera
<curs></curs>	Cursiva
<sic></sic>	Error
<port></port>	Portada
<prelim></prelim>	Preliminar
<lim></lim>	Liminar
<colf></colf>	Colofón
<marg></marg>	Nota al margen
<ms></ms>	Nota manuscrita
<pie></pie>	Pie de página
<verso></verso>	Fragmento en verso
<line></line>	Línea o verso
<tach></tach>	Tachado
<superp></superp>	Superposición

Tabla 2: etiquetas de formato. Elaboración propia.

Por su parte, las etiquetas de fenómenos lingüísticos son las siguientes:

Etiqueta	Descripción
<cdp></cdp>	Complemento directo preposicional
<conda></ conda>	Condicional analítico
<dcl></dcl>	Duplicación clítica
<ft></ft>	Fórmulas de tratamiento

<futa></futa>	Futuro analítico
<lasm></lasm>	Laísmo
<lesm></lesm>	Leísmo
<losm></losm>	Loísmo
<marc></marc>	Marcadores de discurso (con que/conque, a saber, a pesar de (que), no obstante, de hecho, últimamente).

Tabla 3: etiquetas de fenómenos lingüísticos. Elaboración propia.

Cabe mencionar que el corpus no cuenta con una lematización gráfica ni gramatical y, que, además, solo se realiza en los documentos el etiquetado de estos fenómenos lingüísticos. Entre los años 2002-2003 se etiquetaban más fenómenos lingüísticos[23], pero se tuvieron que dejar de lado esas etiquetas porque se invertía demasiado tiempo en ellas (dado que incluía etiquetar los complementos directos o indirectos) y generaban confusiones cuya revisión también exigía un exceso de inversión de tiempo, más difícil aún con el magro equipo con el que siempre se ha contado Sáez Rivera "Algunas posibilidades" 153).

5.2.3. Ejes principales

5.2.3.1. Eje temporal

El corpus PROLEGRAMES es un corpus histórico y diacrónico del español que se centra en las épocas poco estudiadas de las etapas donde la creación de sistema a partir del discurso es más determinante como como la segunda mitad del siglo XIV y las transiciones del castellano medieval al español clásico y de este al moderno, así como del moderno al contemporáneo" (Sáez Rivera "PROLEGRAMES" s.p). Los textos que lo componen abarcan un periodo de tiempo de cuatro siglos, siendo el más antiguo del siglo XVI, *Vocabvlario de los vocablos que mas comunmente se suelen usar [...]* de Jacques de Liaño (1565), y el más reciente del siglo XIX, *El Secretario español, ó nuevo manual de Cartas y sus respuestas*: *según el gusto del día* de Carlos Pellicer (1861) (Sáez Rivera "Algunas posibilidades" 148).

En la siguiente tabla, hemos clasificado los textos a partir del siglo al que pertenecen, el número de palabras y el número de obras:

Siglo	Palabras	Obras
XVI	52.877	2

[23] Se pueden consultar estas etiquetas en la siguiente tabla: 5.3.2.4

XVII	765.742	18
XVII-XVIII	28.250	1
XVIII	779.699	19
XIX	63.436	2
TOTAL	1.690.004	42

Tabla 4: eje temporal. Elaboración propia.

5.2.3.2. Eje diatópico

En cuanto al eje diatópico o geográfico, el corpus PROLEGRAMES es un corpus español conformado por textos escritos en español europeo. Nos suscribimos a Virginia Bertolotti en su distinción de las variedades del "español europeo" frente al "español americano", pues uno de los objetos de estudio del proyecto es, precisamente, las fórmulas de tratamiento del español europeo en la primera mitad del siglo XVIII. Esta distinción de Bertolotti viene propiciada por la diferencia en la evolución de las formas plurales de segunda persona durante los siglos XVII, XVIII y XIX, que también afectó a sus respectivas formas verbales (Azevedo 3).

Por otra parte, PROLEGRAMES también documenta en el encabezado TEI, en <sourceDesc></sourceDesc> "los datos relativos a la fuente de origen del manuscrito o del impreso original" (Sánchez "Guía" 9), pero no se indica el lugar de creación del texto original ni el lugar de procedencia del autor.

5.2.3.3. Eje tipológico

Como presentamos en el apartado 5.2.2.4, la tipología textual se ha replanteado para adaptarse a las nuevas necesidades del corpus PROLEGRAMES.

Esta tipología se ha realizado de acuerdo con la distinción entre tradición discursiva y tradicionalidad discursiva. La tradición discursiva, entendida también como clases de textos, se define como "cualquier relación que se puede establecer entre dos elementos de tradición […] que evocan una determinada forma textual o determinados elementos lingüísticos empleados" (Sáez Rivera y Octavio Toledo 25) y "pueden asociarse a elementos o construcciones gramaticales o léxicos repetidos, pero ni las rutinas gramaticales ni las unidades del lexicón son en sí mismas tradiciones discursivas" (Sáez Rivera y Octavio Toledo 26). Por su parte, la tradicionalidad discursiva "viene, pues, determinada por los itinerarios textuales que este traza a lo largo del tiempo, difundiéndose de componente textual en componente textual […] que resulten recurrentes

en diferentes textos o clases textuales" (Sáez Rivera y Octavio Toledo 26) y además, "permiten describir, sin necesidad de recurrir a una taxonomía textual precisa, las relaciones entre textos, partes de textos y grupos de textos que cabe establecer a partir de la observación de fenómenos lingüísticos que comparten" (Sáez Rivera y Octavio Toledo 27).

En la siguiente figura se ilustrará el universo discursivo que se propone desde PROLEGRAMES:

CUADRO 2. *Universos del discurso, formas discursivas y temas discursivos de la selección de textos de la primera mitad del siglo XVIII*

Literatura	*Fe*	*Ciencia y técnica*	*Cotidianidad*
I. Prosa narrativa a. extensa [1-2, 26-30] b. breve [31-34]	III. Prosa de asunto religioso [7-13, 37-39]	IV. Prosa de asunto lingüístico y literario [14-17, 40-42]	IX. Prosa epistolar [25, 55-56]
II. Prosa historiográfica [3-6, 35-36]		V. Prosa especulativa, filosófica y moral [18-19, 43-46]	X. Prosa circunstancial, periódica y de divulgación [57-60]
		VI. Prosa de asunto político y socio-económico [20, 47-48]	
		VII. Prosa científica y técnica [21-23, 49-52]	
		VIII. Prosa jurídica y administrativa [24, 53-54]	

Figura 13: eje tipológico

Además, en la siguiente tabla mostraremos la distribución del corpus en cuanto a su tipología:

Tipo textual	Palabras	Obras
Biografía	105.043	2
Carta	7.905	2
Consulta	4.605	1
Crónica	193.686	1
Diálogo	32.107	3
Informe	43.872	3
Manual de cartas	397.068	4
Periodismo	229.628	2
Prosa miscelánea	78.916	3

Relación	75.368	9
Sermón	26.773	1
Tratado	495.033	11
TOTAL	1.690.004	42

Tabla 5: eje textual. Elaboración propia.

5.2.4. Filiación de los documentos (metadatos)

Los documentos del corpus PROLEGRAMES contienen metadatos descriptivos, es decir, contienen información que describen el documento y proporcionan información contextual sobre ellos. Además, estos metadatos pueden servir para filtrar y ordenar los documentos de un corpus cuando en los programas de análisis de corpus (Torruella 156). Esta información podemos encontrarla en el encabezado TEI descrito en "*Guía de etiquetado: Documentos Programes*" de Nayra Sánchez Vera (2020).

5.2.4.1. Título de la obra

Los títulos de las obras de PROLEGRAMES se pueden encontrar como nombres de las versiones para lectura y análisis, y, además, se indican en el encabezado TEI dentro de las etiquetas <titleStmt></titleStmt>, que contiene "la información relativa al titlo de la obra, a la edición de la misma y al responsable de transcripción y codificación" (Sánchez "Guía" 7), en <title></title>.

Empezando por el nombre de los documentos del corpus para descargar (que se corresponden con las versiones de lectura y análisis), observamos que se forman a partir de la estructura: código (821) + fecha de creación del documento (AAAA-MM-DD) + el nombre de la obra o el nombre del autor. De ejemplo tendríamos: "821-2018-05-30-Dialogo Fenrandez 1838 para analizar (1)" o "821-2018-05-30-carta gil de la cotera para analizar". Cabe destacar de los títulos que no están normalizados y se nombran de forma diferente dependiendo del etiquetador. No obstante, el título de la obra que es relevante realmente para las herramientas de análisis de corpus es el título del encabezado TEI. Este como hemos mencionado se encuentra en la etiqueta <title></title> y mantiene las grafías originales, no está normalizado. Además, no siguen el criterio de Toruella de eliminar las preposiciones o artículos (157), sino que las conservan.

5.2.4.2. Autor de la obra

El autor de la obra se indica en el nombre de los documentos de descarga, aunque esto depende del criterio que haya seguido el etiquetador. Un ejemplo de ello es: "821-

2014-10-02-Pellicer análisis". El nombre del autor de la obra también se puede encontrar en el encabezado TEI y se encuentra en este caso en <sourceDesc></sourceDesc>, que contiene los "datos relativos a la fuente de origen […] del manuscrito o impreso original (Sánchez "Guía" 9), en <resp></resp>, una etiqueta versátil en la que se indica el cargo de la persona y en <name></name> el nombre. Al igual que el título de las obras, los nombres de los autores no están normalizados dentro de los encabezados TEI. Además de no estar normalizados gráficamente, tampoco lo están en estructura, apareciendo como nombre + 1º apellido + 2º apellido en vez de utilizar la forma estándar 1º apellido + 2º apellido + nombre(s). Por otro lado, señalamos también el uso diferente de "Anónimo" en los documentos cuyos autores se desconocían llegando a omitirse o cambiándolo por "Desconocido". Por último, hay que mencionar que el encabezado PROLEGRAMES permite indicar varios autores con la inclusión de otra etiqueta de <resp></resp>.

A pesar de esta falta de normalización de los autores, cabe destacar que recientemente se están normalizando los nombres de los autores de las obras para mejorar la búsqueda por nombre de autor para mejorar el funcionamiento de la herramienta en línea.

5.2.4.3. Fecha

La fecha que se indica en los documentos del corpus PROLEGRAMES es triple dado que se indica tanto la fecha de creación de la obra original o testimonio como las fechas de creación de la transcripción y la fecha de revisión de este documento.

La fecha de creación del texto original se puede encontrar en <profileDesc></profileDesc>, que indica "los elementos contextuales del texto original" (Sánchez "Guía" 10) que son la fecha de creación y la tipología textual, en la etiqueta <date></date> en formato Año (+ mes + día, si procediera). Actualmente para la herramienta en línea, esta etiqueta se ha reemplazado recientemente por <startDate></startDate> y <endDate></endDate> a pedido del desarrollador de Avantopy.

Las fechas de creación de la transcripción se pueden comprobar en el título de los documentos al descargarlos y en <editionStmt></editionStmt> que "contiene los datos relativos a la fecha de la primera versión de la transcripción y el etiquetado del texto" (Sánchez "Guía" 9). La fecha de revisión se encuentra en

<revisionDesc></revisionDesc> que se actualiza con cada cambio o revisión del archivo. En este caso ambas fechas se muestran en formato DD/MM/AAAA.

5.2.4.4. Tipología textual

La tipología textual se indica en <profileDesc></profileDesc> dentro de la etiqueta <textClass></textClass>. Se puede consultar la tipología textual en los apartados 5.2.4.4 y 5.2.3.3.

5.2.4.5. Dialecto

En PROLEGRAMES no se indica el dialecto y la lengua se marca en el <sourceDesc></sourceDesc> en la etiqueta <idno type="origin">E</idno>. Siendo "E" quien indica que el texto es español.

5.2.4.6. Más metadatos del corpus PROLEGRAMES

Además de los metadatos imprescindibles que señala Torruella en el TEI HEADER de PROLEGRAMES se indican más que son: <publicationStmt></publicationStmt> que "contiene la información relativa a la publicación y distribución del texto" (Sánchez 8a), entendiendo texto como la versión transcrita; <sourceDesc></sourceDesc>, que además de indicar el autor y la lengua, también indica el impresor y el paradero del documento original; como hemos mencionado también se indica con <editionStmt></editionStmt> los datos de la edición; y por último, <extent></extent> que "incluye información relativa al tamaño del texto, cuya información está expresada en los atributos de medida "*bytes*" y "*words*"" (Sánchez "Guía" 9).

Gracias a la inclusión de estos metadatos, no solo podemos tener más información relevante respecto a los textos, sino que es posible que en un futuro la herramienta en línea pueda utilizarlos como criterios de filtrado o para realizar análisis estadísticos.

5.3. Presentación y explotación de la herramienta en línea

Según la afirmación de ODE de que "los dos aspectos que permiten mayores posibilidades de explotación de un corpus es la anotación lingüística y la interfaz de usuario" (s.p), podemos afirmar que el corpus PROLEGRAMES ya contaba con uno de los aspectos fundamentales para su explotación efectiva (anotación lingüística), y, que, gracias al desarrollo de la nueva herramienta en línea, el proyecto de investigación PROLEGRAMES ha incrementado significativamente sus oportunidades de explotación

del corpus. La interfaz de usuario es fundamental para la realización de consultas, ya que permite recuperar la información de forma eficaz y rentabilizar el rico etiquetado de PROLEGRAMES.

A lo largo de este trabajo, se ha mencionado el desarrollo de una interfaz de usuario, una herramienta en línea en colaboración del proyecto PROLEGRAMES con la empresa Avantopy para mejorar el análisis del corpus. Anteriormente, el análisis del corpus se recomendaba realizarlo con *WordSmith Tools* y LYNEAL, era a partir de un programa externo, obligando al investigador a crear un corpus ad hoc con los documentos que se tenía que descargar y convertir a texto plano individualmente para más tarde realizar el análisis. Con esta nueva interfaz, se facilitan todos esos procesos que se debían hacer manualmente.

Esta nueva interfaz en línea se llama *Complutense PROLEGRAMES*[24] y a continuación en los siguientes apartados vamos a explicar sus tres niveles de organización y funcionalidades: cabecera, motor de búsqueda y visualización de los resultados.

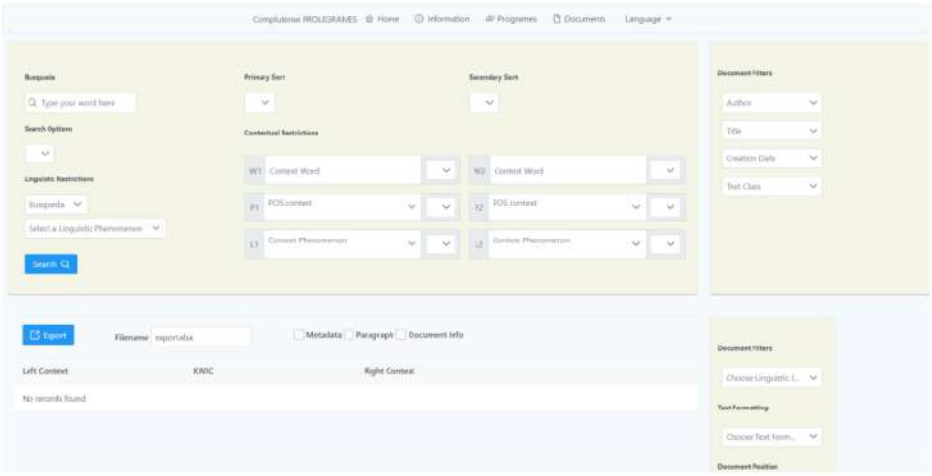

Figura 14: interfaz de Complutense PROLEGRAMES. Elaboración propia

5.3.1. Cabecera

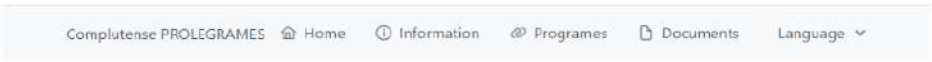

Figura 15: cabecera. Elaboración propia

[24] Se puede acceder a la herramienta mediante el siguiente enlace: https://prolegrames-iump.ucm.es

En el primer nivel, se encuentra la cabecera de la herramienta que se compone a su vez por cinco secciones.

- **Home**: es un botón que permite regresar a la página principal del proyecto.
- **Information:** es el lugar donde encontraremos la información sobre el proyecto, aunque actualmente se encuentra vacía.
- **Programes**: redirige a la página institucional del proyecto PROLEGRAMES.
- **Documents**: conduce a la sección de "Documentos PROGRAMES" de la página web institucional.
- **Language:** ofrece la posibilidad de cambiar la lengua de la interfaz, aunque por el momento la herramienta solo está disponible en inglés.

5.3.2. Motor de búsqueda

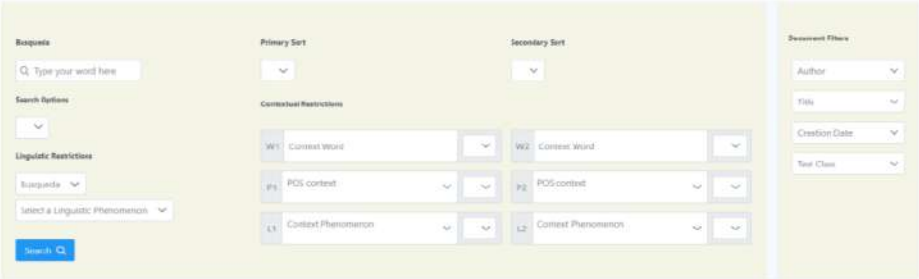

Figura 16: motor de búsqueda. Elaboración propia.

En el segundo nivel se encuentran las herramientas de búsqueda dentro del corpus y estas se organizan en dos bloques diferenciados. En el apartado izquierdo se ubica el motor de búsqueda principal y en el apartado derecho encontramos las restricciones creadas a partir de los metadatos del documento. A continuación, detallaremos cada uno de los parámetros.

5.3.2.1. Busqueda

Busqueda [*sic*, sin tilde, algo que conviene corregir] permite hacer consultas simples a partir de la búsqueda de una única palabra. Las consultas simples no admiten la búsqueda de cadenas de palabras ni expresiones lógicas como (y, o, y no, etc.) y, además, *Busqueda* hace distinción entre mayúsculas/minúsculas y acentos gráficos y no hace consultas vacías. Por este motivo, habrá que realizar tantas búsquedas como variantes de una palabra haya si procede. Todos los resultados se mostrarán en la parte de visualización

de los resultados en formato *KWIC* (Key Word in Context). Las consultas complejas se realizan mediante la combinación de *Busqueda* y los parámetros de búsqueda.

5.3.2.2. Primary Sort y Secondary Sort

Primary Sort Key y *Secondary Sort Key* permiten elegir en qué posición se inserta nuestra muestra respecto del párrafo en que se encuentra.

5.3.2.3. Search options

Search options permite elegir cómo queremos que aparezcan los caracteres que busquemos en las consultas en las muestras. Se puede elegir una opción y no son combinables.

- *Case insensitive*: distinción entre mayúsculas o minúsculas.
- *Regex*: la palabra forma parte de una expresión regular.
- *Contains*: la palabra contiene esos caracteres.
- *Starts With*: la palabra empieza con esos caracteres.
- *Ends With*: la palabra termina con esos caracteres.
- *Lemma*: se recupera el lema, es decir, la forma lexicográfica de la palabra que se busca[25].

5.3.2.4. Linguistic Restrictions

Linguistic Restrictions permite elegir tanto la categoría gramatical de la palabra buscada como el fenómeno lingüístico que sea. Una de las innovaciones de la herramienta en línea de PROLEGRAMES respecto a otras es que permite buscar los fenómenos lingüísticos que se han etiquetado en los textos del corpus. A continuación, mostraremos una tabla que recoge todas las categorías gramaticales:

Abreviatura	Categoría gramatical	Abreviatura	Categoría gramatical
ADJ	Adjetivo	NUM	Numeral
ADP	Aposición	PART	Partícula
ADV	Adverbio	PRON	Pronombre
AUX	Auxiliar	PROPN	Nombre propio

[25] En CORPES XXI se hace una diferenciación entre las búsquedas con lema y con forma que ejemplifica de la siguiente forma: "Si se escribe llegar en la casilla de Forma se obtendrán únicamente los casos del infinitivo de ese verbo. En cambio, si se escribe *llegar* en la casilla de Lema, la aplicación devolverá todas las formas que pertenecen al paradigma de ese verbo; *llego, llegaron, llegaría, llegue, llegar, llegando*, etc." (RAE "CORPES" s.p)

CCONJ	Conjunción coordinante	PUNCT	Puntuación
DET	Determinante	SCONJ	Conjunción subordinante
INTJ	Interjección	SYM	Símbolo
NOUN	Sustantivo	VERB	Verbo

Tabla 6: categorías gramaticales. Elaboración propia.

En cuanto a los fenómenos lingüísticos, cabe mencionar que se pueden consultar en la Tabla 3, y, además, se han ampliado los fenómenos lingüísticos con aquellas etiquetas que se eliminaron en las fases de 2003-2005 del proyecto PROLEGRAMES:

Etiqueta	Descripción
<cd></cd>	Complemento directo
<ci></ ci>	Complemento indirecto
<ppas></ppas>	Participio de pasado
<ra></ra>	Forma verbal en *-ra*
<reg></reg>	Régimen preposicional de los verbos
<rel></rel>	Relativo
<relart></relart>	Artículo + relativo
<relcomp></relcomp	Relativo compuesto

Tabla 7: etiquetas ampliadas de fenómenos lingüísticos. Elaboración propia.

5.3.2.5. Contextual Restrictions

Contextual Restrictions permite al usuario elegir tres restricciones contextuales, es decir, criterios que afectan al contexto. Cabe mencionar que tanto las categorías gramaticales como los fenómenos lingüísticos de P1/P2 y L1/L2 respectivamente se corresponden con los de *Linguistic Restrictions*.

- *W1/W2:* permiten elegir una Context Word (palabra contextual) para que aparezca en las posiciones de L1-L5 y/o de R1-R5.
- *P1/P2:* permiten elegir el POS Context[26] (categoría gramatical) que queramos que aparezca en las posiciones de L1-L5 y/o R1-R5.
- *L1/L2:* permiten elegir el Context Phenomenon[27] (fenómeno lingüístico) que queremos que aparezca en las posiciones de L1-L5 y/o R1-R5.

[26] Consultar las categorías gramaticales disponibles en la Tabla 6.
[27] Consultar los fenómenos lingüísticos disponibles en las Tabla 3 y Tabla 7.

5.3.2.6. Document Filters

En el apartado derecho se encuentra *Document Filters*, que contiene los parámetros *Autor, Title, Creation Date* y *Text Class*, los cuale permiten filtrar las consultas a partir de los metadatos. Estos filtros pueden aplicarse a partir del marcaje de las casillas o ingresando directamente el nombre del autor o el título de la obra. Es por esto por lo que incidimos en la importancia de la normalización de los nombres de las obras y de los autores, pues la falta de normalización puede ocasionar la duplicación de elementos. Un ejemplo de ello es el caso de los autores, donde un mismo autor podría aparecer varias veces por la variación de las grafías de su nombre.

5.3.2.7. Herramientas de análisis de corpus: concordancia

En comparación con las herramientas de análisis de corpus que presentamos en el apartado 3.2.1, la herramienta de PROLEGRAMES presenta limitaciones significativas respecto a otros corpus debido a la carencia de ciertas funciones relativas a la falta de lematización efectiva de los textos, es decir, no se pueden realizar análisis estadísticos dentro del corpus ni búsquedas efectivas por lemas. Los documentos se están lematizando con el método de *POS Tagging* por el desarrollador de Avantopy, aunque esta funcionalidad como está en proceso todavía falla a menudo, menos con los verbos, por lo que las búsquedas por categorías gramaticales no son del todo fiables.

Por el momento, la herramienta en línea solo cuenta con las funciones *Concordancia* (3.2.1.4) y *Búsqueda por anotación lingüística* (3.2.1.6). Antes de continuar con el siguiente apartado, cabe mencionar que las búsquedas de *Concordancia* se visualizan en formato tabular KWIC. Por su parte, la *Búsqueda por anotación lingüística* se puede realizar sin introducir ningún carácter en la búsqueda y, además, como recientemente se han implementado los fenómenos lingüísticos que se dejaron de etiquetar durante la fase del proyecto 2003-2005, las búsquedas de los mismos serán erróneas y marcarán caracteres que no corresponden, pues los textos no están provistos todavía con esas etiquetas.

A continuación, mostraremos dos ejemplos con estas búsquedas:

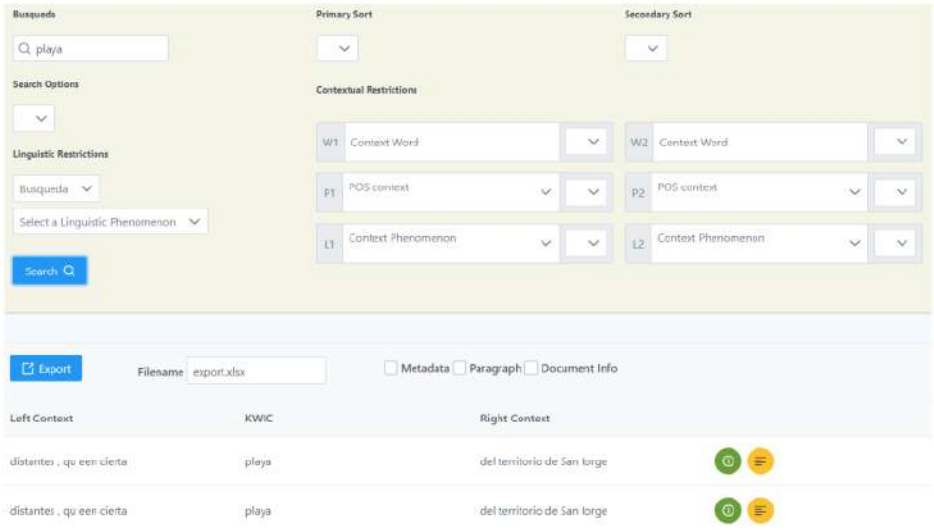

Figura 17: ejemplo de búsqueda con concordancia (playa). Elaboración propia.

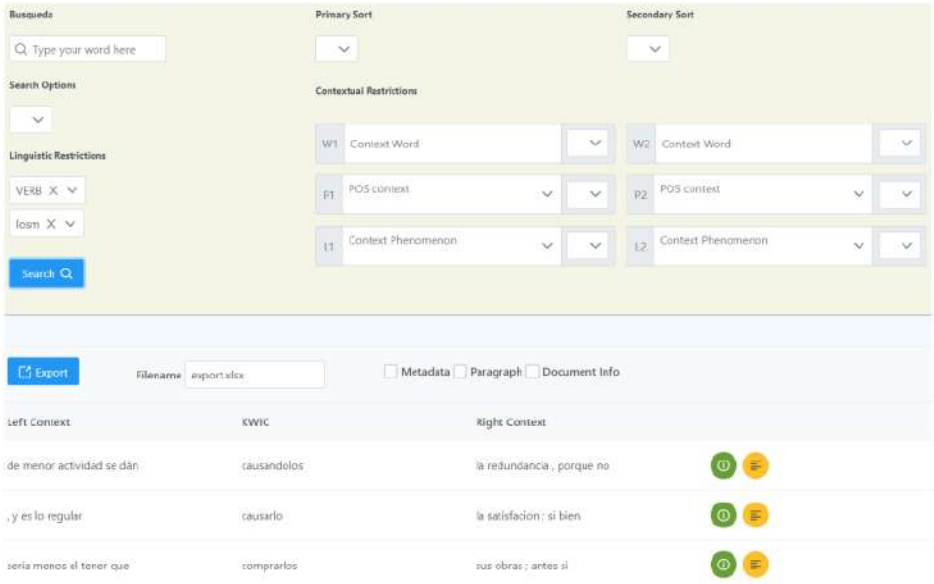

Figura 18: ejemplo de búsqueda por anotación lingüística (loísmo). Elaboración propia.

5.3.3. *Visualización de los resultados*

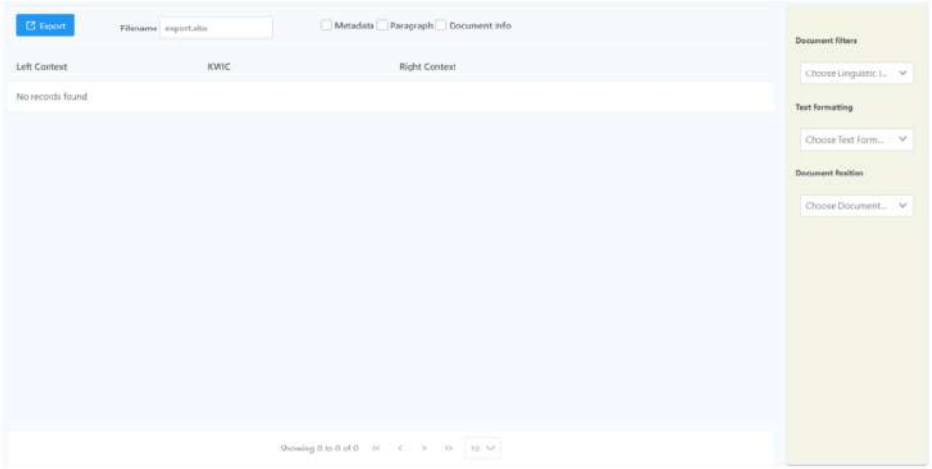

Figura 19: visualización de los datos. Elaboración propia.

En el tercer nivel de la interfaz se encuentra la visualización de los resultados que se divide en apartados izquierdo y derecho.

La sección izquierda se corresponde con la visualización de los resultados y se divide en partes. En la parte de arriba se muestran las opciones de exportación que son:

- *Export*: un botón para exportar los resultados de nuestras consultas en formato .xlxs, es decir, se exporta en una hoja de Excel.
- *Filename*: es un cuadro de texto que permite nombrar el archivo de la exportación.
- *Metadata*: es una casilla que permite incluir con la exportación los metadatos de la muestra.
- *Paragraph*: es una casilla que permite incluir con la exportación el párrafo donde se encuentra la muestra.
- *Document Info*: es una casilla que permite incluir con la exportación información acerca del texto.

En el medio se encuentra la visualización de las búsquedas en formato tabular KWIC que muestra el *Left context* y el *Right context* de cada ejemplo, o sea, el contexto a la izquierda y a la derecha. En la misma línea se ofrecen dos botones, uno verde que es *Document Details* y uno amarillo, *Paragraph*. Al pulsarlos aparecerá una ventana emergente con el título, el autor, la fecha de creación y la clase de texto y el párrafo donde se inserta la muestra, respectivamente. En la parte de debajo se encuentra *Showing* que

indica el número de resultados por página y el total de páginas disponibles, y, además, permite ajustar la cantidad de resultados visibles mediante un menú desplegable cuyas opciones son 10, 20 y 50.

Por último, en el segundo apartado se encuentran los *Documents Filters* similares al que aparece en el motor de búsqueda. Sin embargo, estos parámetros no afectan a los metadatos, sino que permiten resaltar etiquetas de las muestras según tres criterios:

- *Linguistic information*: se refiere a las etiquetas de los fenómenos lingüísticos[28].
- *Text Formating*: se refiere a las etiquetas de formato del texto. Las opciones son: "sic" para los errores, "foreign" para los extranjerismos y "curs" para las cursivas.
- *Document Position*: se refiere a las etiquetas de formato del texto relativa a la posición donde se encuentre este. Las opciones son: "col1/col2/col3" para las columnas, "pie" para el pie de página, "marg" para el margen, "tach" para texto tachado, "colf" para el colofón, "ms" para la anotación manuscrita y "superp" para la superposición.

A continuación, mostraremos un ejemplo de los resultados con un filtro:

Figura 20: ejemplo de resultado con un filtro (loísmo). Elaboración propia.

Desafortunadamente no tenemos una figura que muestre un ejemplo de la visualización de los resultados exportados mediante *Export* porque esta función aún no se encuentra disponible.

[28] Para consultar los fenómenos lingüísticos consultar las Tabla 3 y Tabla 7.

5.4. Análisis heurístico de la interfaz web

En este apartado procederemos a realizar un análisis heurístico de la interfaz web siguiendo los principios de la heurística de Molich y Nielsen de la nueva herramienta en línea del corpus PROLEGRAMES.

5.4.1. Visibilidad del estado del sistema

Observamos que la herramienta mantiene informado al usuario mientras procesa las solicitudes y esto se muestra a partir del cambio del puntero del ratón al modo ocupado si la tarea, en este caso búsqueda, tarda más de lo esperado. Además, el puntero resalta y cambia a un color más oscuro las casillas que seleccionamos, lo que permite diferenciarla del resto y reducir el riesgo de equivocación.

5.4.2. Relación entre el sistema y el mundo real

Se observa un uso de términos y conceptos propios del ámbito del estudio de la lingüística de corpus como *Linguistic Restrictions, POS Context* o *Context Words* que refleja su carácter técnico y específico. Así mismo, observamos un uso adecuado de los iconos que facilitan la comprensión de la navegación de forma visual y que existe una opción en la cabecera de la página para cambiar la lengua de la interfaz entre el inglés y el español que contribuye a mejorar la accesibilidad para usuarios no hispanohablantes.

5.4.3. Control y libertad del usuario

Cualquier usuario puede acceder al corpus y utilizar todas sus funciones sin ninguna restricción. Además, como menciona Nielsen cuando recalca la importancia de existencia de opciones de salida, la herramienta de PROLEGRAMES ofrece una opción para volver la página principal a partir de *Home*. Algo que, sin embargo, no se puede realizar es el borrado de una búsqueda a menos que se reinicie la página principal o se hagan otras búsquedas. Por otra parte, en los menús desplegables observamos que hay una opción que es "X" que permite borrar los filtros seleccionados. Sin embargo, esto no es algo que está implementado en las funciones de *Busqueda, Primary Sort, Secondary Sort* y la elección de las posiciones de *Contextual Restrictions*.

5.4.4. Consistencia y estándares

En la herramienta PROLEGRAMES presenta consistencia en su interfaz y sigue una estructura similar a otras herramientas de análisis de corpus como CORDIAM, ODE o *Sketch Engine*. Los usuarios familiarizados con el uso de las herramientas de análisis

de corpus no tendrán dificultad para utilizar PROLEGRAMES. Una ventaja que podemos destacar de la herramienta de PROLEGRAMES es que sus criterios de búsqueda se muestran muy visibles en lugar de estar en menús desplegables como es el caso de otras plataformas. Quizá, uno de los aspectos que podrían causar algo más de problemas serían los fenómenos lingüísticos, pues se emplea una nomenclatura propia de PROLEGRAMES.

5.4.5. Prevención de errores

La plataforma no restringe ningún tipo de consulta, incluso si son incorrectas. Por este motivo consideramos necesario que haya un mensaje de aviso para que notifique al usuario que la consulta no es válida si no hay una palabra clave. Los filtros *Primary Sort, Secondary Sort, Search options, Contextual Restrictions* no funcionan por sí mismos. Otra cuestión que hay que tener en cuenta es que la búsqueda de funciones gramaticales de *Linguistic Restrictions* puede mostrar resultados incorrectos sin una palabra clave. También sería recomendable que la herramienta notifique cuando se intente descargar una consulta vacía, lo que sí se advierte con la frase "*No records found*" en el apartado de visualización de resultados.

5.4.6. Reconocer antes que recordar

La herramienta en línea PROLEGRAMES no permite recuperar consultas realizadas anteriormente y tampoco tiene un historial. No obstante, como solución se podría utilizar la función ***Export*** para guardar cada consulta.

5.4.7. Flexibilidad y eficiencia de uso

Los únicos atajos de teclado o *shortcuts* que permite son los que afectan a todo el navegador como "Ctrl+C", "Ctrl+V", "Ctrl+S" o las flechas de navegación entre otros ejemplos. Además, cabe mencionar que la tecla "Enter" no funciona como sustituto del botón de *Search*.

5.4.8. Estética y diseño minimalista

Complutense PROLEGRAMES presenta un estilo y diseño minimalista, con secciones claramente delimitadas por espacios, títulos y funciones, lo que facilita su uso y comprensión. Además, cabe destacar el uso de los colores como en los botones de *Search* y *Export* y el uso de iconos que mejoran la navegación por medio de elementos

visuales. También es fundamental destacar que el corpus no tiene un diseño web adaptable o *responsive*[29], lo que provoca que se deforme en cuanto se reduce la vista.

5.4.9. Ayudar a los usuarios a reconocer, diagnosticar y solucionar errores

Cuando se muestran los errores en la interfaz se acompañan por mensajes de error, que, por el momento, siguen los protocolos por defecto. Por ello, es recomendable modificar los mensajes para que presenten advertencias más amigables y accesibles, sin códigos de error para que puedan entenderlos cualquier usuario. Por ejemplo. En caso de que una consulta falle porque ha tardado demasiado o ha dado error el envío, se podría utilizar el mensaje: "Su consulta no ha podido efectuarse, por favor, refresque la página web."

5.4.10. Ayuda y documentación

Actualmente, la herramienta de PROLEGRAMES no cuenta con ningún apartado de *Ayuda* dado que por el momento sigue en desarrollo. No obstante, podríamos considerar la sección *Information* como esa supuesta sección de ayuda, pero de todas sería recomendable mejorar su señalización. Además, consideramos que se necesita una guía documentada para la utilización de la herramienta, una sección de preguntas frecuentes e indicar las abreviaturas que se utilizan para las funciones gramaticales o los fenómenos lingüísticos, pues se pueden intuir, pero lo óptimo sería proporcionar esa información al usuario.

6. Desarrollo y líneas de trabajo futuro

Tras haber realizado un análisis exhaustivo de la situación actual del corpus PROLEGRAMES, en este apartado se propondrán nuevas vías de desarrollo para su mejora.

Empezando por las debilidades señaladas en el análisis DAFO (5.1), se destacan aspectos clave fundamentales para la mejora del proyecto. En primer lugar, encontramos la normalización gráfica, que ya se está implementando en los nombres de los autores para mejorar la usabilidad de la herramienta en línea, además de la creación de una

[29] "El diseño web adaptable o responsive es una filosofía de diseño y desarrollo cuyo objetivo es adaptar la apariencia de las páginas web al dispositivo que se esté utilizando para visitarlas." Diseño web adaptable. *Wikipedia, la enciclopedia libre*, es.wikipedia.org/wiki/Diseño_web_adaptable

tipología textual formalizada para guiar a los etiquetadores. Con respecto a la normalización gráfica de los nombres de las obras se limitará a los nombres de los archivos de descarga y visualización, manteniendo las grafías originales en los títulos y los textos conforme a la filosofía de PROLEGRAMES de mantener "fidelidad absoluta al texto original" (Sáez Rivera 149a). Para abordar la problemática de tener un equipo pequeño y en rotación, así como la dificultad para la fidelización de los etiquetadores, sería necesario tener una financiación estable para poder compensarlos económicamente.

En cuanto a las versiones facsimilares, su implementación es difícil porque para poder ofrecerlas primero se necesita el permiso de los archivos o bibliotecas y resulta complicado que los concedan sin pagar. Además, como hemos visto, PROLEGRAMES cuenta con una financiación escasa y, actualmente no dispone de los recursos financieros necesarios, pero si en un futuro consiguieran los permisos, se podría proponer la implementación de estas versiones en formato de imagen o en formato .pdf en la nueva herramienta en línea. No obstante, otra alternativa sería proporcionar a los usuarios enlaces de los facsímiles que se encuentran en línea y son de acceso libre, aunque no sería tan cómodo como consultarlos simultáneamente. Un ejemplo de ello sería: "*Vocabvlario de los vocablos que mas comunmente se suelen usar* [...] de Llaño en https://books.google.es/books?id=GMce6MDbndcC&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false". Siguiendo con esta línea, se podría subir en la herramienta en línea los textos del corpus en formato de texto plano .txt. Sin embargo, con la nueva interfaz ya no sería de vital importancia, pues se podrán hacer consultas de forma directa en el corpus sin necesidad de recurrir a ninguna herramienta externa.

Las amenazas identificadas se relacionan con la dependencia de PROLEGRAMES de las organizaciones externas para la financiación y son difíciles de eliminar por completo. No obstante, siempre se pueden mitigar optimizando las fortalezas del proyecto y así asegurar las subvenciones. En relación con esto, la amenaza de la fragilidad de los servidores podría evitarse, ya que Complutense PROLEGRAMES, al estar alojado en el IUMP, este instituto podría hacerse cargo del pago del *hosting* del corpus y de la herramienta en línea.

Siguiendo con las oportunidades, identificadas como automatización del etiquetado y creación de una herramienta en línea, observamos que tienen un gran potencial de optimización del proyecto. Esto es porque la automatización por medio de

IA, de poder desarrollarse e implementarse, no solo mejoraría la eficiencia de los procesos de etiquetado, pues el proyecto no dependería tanto de los etiquetadores, sino que también permitiría implementar nuevas funciones como la lematización automática, la transcripción semiautomática de los textos y la inclusión de etiquetas complejas que fueron eliminadas por su carga de trabajo en fases anteriores de PROLEGRAMES. No obstante, la prioridad inicial sería automatizar el etiquetado de fenómenos lingüísticos y de formato. Con relación a la creación de una herramienta en línea, en los apartados 5.3 y 5.4 se realizaron unos análisis de dicha plataforma y se observó que la herramienta posee ciertas limitaciones que no permiten la óptima explotación del corpus. La lematización de los textos es un aspecto fundamental que podría abrir nuevas oportunidades para el corpus y la interfaz de usuario, pues gran parte de la investigación de los corpus digitales se centra en el análisis estadístico. La lematización facilitaría la implementación de nuevas funcionalidades relativas a la estadística como la creación de listas que permitirían un análisis aún mayor de los textos. Así mismo, se consideraría importante añadir funcionalidades avanzadas como las consultas complejas con operadores lógicos y expresiones regulares, búsquedas con CQL o la creación de vistas, para mejorar tanto la herramienta como la experiencia de usuario.

Finalmente, se recomienda la creación de un apartado de *Ayuda* que facilitara una guía o manual que ayudara con el manejo de la herramienta en línea.

7. Conclusiones

El corpus PROLEGRAMES es un corpus diacrónico especializado en el estudio de la lengua española desde sus orígenes hasta su gramaticalización y lexicalización que se centra sobre todo en las etapas menos estudiadas del español como el paso del español clásico al moderno o la segunda mitad del siglo XVII. Además, también contribuye a la visibilización de autores y a tradiciones discursivas que no han sido tan estudiadas.

El análisis DAFO del proyecto PROLEGRAMES, fundamentado en los estudios previos de Sáez Rivera (2018) y Sánchez (2020), ha permitido identificar el estado actual del proyecto y han servido de base para fundamentar las nuevas vías de desarrollo y explotación propuestas. Las debilidades identificadas se relacionan con la falta de normalización gráfica, la falta de personal y la carencia de las versiones facsimilares que inciden en el progreso del corpus. Por su parte, las fortalezas radican en su etiquetado

riguroso de fenómenos lingüísticos y de formato a falta de una versión facsimilar, la recuperación de autores y tipologías poco estudiadas, lo que contribuye significativamente al campo de la investigación de la lengua. Entre las amenazas más críticas se destaca la dependencia financiera que podría poner en peligro la continuidad del proyecto además de afectar directamente sobre los recursos del corpus como podría ser no poder pagar al personal o no poder pagar los servidores. A pesar de estas limitaciones, encontramos oportunidades significativas como la creación de la herramienta de consulta en línea del corpus y la automatización del etiquetado mediante inteligencia artificial que se muestran como soluciones que podrían resolver varios problemas como la normalización de los nombres o la reducción de la dependencia de los etiquetadores en continua rotación.

Siguiendo con esto, la creación y el desarrollo de la nueva herramienta en línea de PROLEGRAMES ha supuesto una mejora significativa en las posibilidades de explotación del corpus y de sus textos. Esto es porque ya no se necesita recurrir a herramientas de análisis externas y concentra tanto los textos como las herramientas de análisis en una misma plataforma. La interfaz de usuario de PROLEGRAMES se organiza en tres niveles: cabecera, motor de búsqueda y visualización de los resultados. La herramienta en línea ofrece las funcionalidades de consulta simple, consulta con filtros, búsqueda por anotación lingüística y concordancia, todo ello se muestra en formato KWIC. Es una herramienta sencilla de entender y sigue los principios de la heurística de interfaces web de Nielsen. No obstante, señalamos la necesidad de la creación de una guía o manual detallado que facilite la usabilidad de la herramienta.

Por último, hemos visto que la herramienta, aunque todavía está en desarrollo, carece de algunas funcionalidades claves para su mejora como la necesidad de una normalización enfocada en los nombres de los autores y las tipologías textuales, lo que facilitará el etiquetado y transcripciones a los miembros del proyecto. También es fundamental la lematización efectiva del corpus, ya que su ausencia limita el potencial de análisis al impedir la inclusión de ciertas funcionalidades imprescindibles para el análisis estadístico como las listas estadísticas. Otra cuestión que mejoraría la explotación, además de optimizar el tiempo de los miembros del proyecto y que solucionaría el problema de la continua rotación del equipo, sería la implementación de una inteligencia artificial que pudiera realizar los etiquetados de formato y de fenómenos lingüísticos, y, en el mejor de los casos, mejorar las funcionalidades que ya existen como la lematización

e introducir nuevas como la transcripción semiautomática con revisión final humana. Esta última tarea podría ser realizada en un futuro con *Transkribus*, una plataforma de pago que permite reconocer textos y editarlos, además de que cuenta con una función de entrenamiento personalizado de una IA para que sirva para digitalizar e interpretar documentos (Transkribus s.p).

En definitiva, el corpus PROLEGRAMES está desarrollando una herramienta prometedora que con las mejoras y adaptaciones adecuadas puede convertirse en un recurso importante para el estudio de estas etapas no tan estudiadas de la lengua española.

8. Bibliografía

Academia Mexicana de la Lengua (AML). *Corpus Diacrónico y Diatónico del Español de América (CORDIAM),* www.cordiam.org. Consultado por última vez: 01/09/2024.

Azevedo, Andrés de. "Reseña de: IGNACIO BOSQUE, SYLVIA COSTA y MARISA MALCUORI (eds.). *Palabras en lluvia minuciosa. Veinte visitas a la gramática del español inspiradas por Ángela Di Tullio*", Lingüística, vol. 34-2, 2018, págs. 177-189 doi: 10.5935/2079-312X.20180022.

Calderón Campos, Miguel. "Los corpus del español clásico y moderno: entre la filología y la lingüística computacional". *RLA: Revista de Lingüística Teórica y Aplicada,* vol. 57(2), 2019, pásg. 41-64.

"Corpus" *REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española*, 23.ª ed., [versión 23.7 en línea], dle.rae.es/ Consultado por última vez: 01/09/2024.

De Beni, Matteo; Hourani-Martín, Dunia; Sartor Elisa: *DIACOM-es,* dh.dlls.univr.it/corpora/diacomes Consultado por última vez: 14/08/2024.

De Beni, Matteo; Hourani-Martín, Dunia; Sartor Elisa: *OCCOR-es,* dh.dlls.univr.it/corpora/occores/ Consultado por última vez: 14/08/2024.

Infoautónomos. "Guía fundamental del Análisis DAFO". *Infoautónomos*, www.infoautonomos.com/plan-de-negocio/analisis-dafo/. Consultado por última vez: 14/08/2024.

García Miguel, José M. "Lingüística de corpus: de los datos textuales a la teoría lingüística". *Estudios de Lingüística del Español,* vol.45, 2022, págs. 11-42

Jakubíček, Miloš; Kilgarriff Adam, McCarthy Diana, Rychlý Pavel. "CQL - Corpus Query Language" *Sketch Engine, www.sketchengine.eu/documentation/corpus-querying/*, Consultado por última vez: 01/09/2024.

"Lingüística de corpus". *Diccionario de términos clave de ELE,* cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/linguisticacorpus.htm. Consultado por última vez: 01/09/2024.

Menéndez-Barzanallana Asensio, Rafael. "Capítulo 2. Herramientas de análisis del corpus", *Rafael Barzanallana,* www.um.es/docencia/barzana/TEI/Informatica-Aplicada-

a-la-Traduccion-Herramientas-de-analisis-del-corpus-2013-14.html. Consultado por última vez: 01/09/2024.

Ministerio de Industria y Turismo. "Herramienta DAFO". *Herramienta DAFO,* dafo.ipyme.org. Consultado por última vez: 14/08/2024.

Nielsen, Jakob. "10 Usability Heuristics for User Interface Design" *Nielsen Norman Group,* www.nngroup.com/articles/ten-usability-heuristics/. Consultado por última vez: 14/08/2024.

Nielsen, Jakob y Molich, Rolf. "Heuristic evaluation of user interfaces". CARRASCO, J.; WHITESIDE, J. (ed.). *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 1990, págs. 249–256, doi: doi.org/10.1145/97243.97281. Consultado por última vez: 14/08/2024.

Octavio de Toledo y Huerta, Álvaro S. "Varia lectio y variación morfosintáctica el caso del Crotalón". *Historia de la lengua y la crítica textual*, coord. Lola Pons Rodríguez, Iberoamericana Vervuert, 2006, págs. 195-264.

Pérez, Chantal. "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". *Estudios de Lingüística del Español (ELiES),* elies.rediris.es/elies18/26.html. Consultado por última vez: 14/08/2024.

Pérez Capdevila, Javier. "Óbito y resurrección del análisis DAFO". *Revista Avanzada Científica,* vol.14, nº2, 2011.

PROLEGRAMES. *Complutense PROLEGRAMES,* prolegrames-iump.ucm.es. Consultado por última vez: 01/09/2024.

Real Academia Española (RAE). *Corpus Diacrónico del Español (CORDE),* corpus.rae.es/cordenet.html. Consultado por última vez: 01/09/2024.

Real Academia Española (RAE). *Corpus del Español del Siglo XXI (CORPES),* www.rae.es/corpes/. Consultado por última vez: 01/09/2024.

Sáez Rivera, Daniel M. "Algunas posibilidades de investigación del proyecto Procesos de gramaticalización en la historia del español (Documentos PROGRAMES)". María Dolores Romero / Manuel Salamanca (eds.), *Entornos digitales: Humanidades y Ciencias sociales en la Universidad Complutense de Madrid*, *Red de Humanidades Digitales de México*, 2018, págs. 141-157.

Sáez Rivera, Daniel. M. "La lengua de las gramáticas y métodos de español como lengua extranjera en Europa (1640-1726)". Tesis doctoral dirigida por Dr. José Luís Girón Alconchel. *Universidad Complutense de Madrid,* 2007.

Sáez Rivera, Daniel M. y Octavio de Toledo y Huerta, Álvaro S. *Textos españoles de la primera mitad del siglo XVIII para la historia gramatical y discursiva. Vientos de arrastre y de cambio en la historia del español* (con un prólogo de José Luis Girón Alconchel). Editorial síntesis, 2020.

Sánchez Vera, Nayra. "Guía de etiquetado Documentos Programes" dirigido por Daniel M. Sáez Rivera, 2020. *Universidad Complutense de Madrid,* www.ucm.es/data/cont/docs/821-2020-09-03-guia%20programes%20final%2020200903.pdf. Consultado por última vez: 14/08/2024.

Sánchez Vera, Nayra. "Repercusión de TEI (Text Encoding Initiative) en España: Evolución, situación actual y caso de Corpus en español (Documentos PROGRAMES)". TFM dirigido por Daniel M. Sáez Rivera, 2020. *Universidad Complutense de Madrid.*

Santamaría Urbieta, Alexandra y Peñalver Alcalde, Elena. "Autocrítica de publicaciones previas basadas en corpus: Análisis DAFO". Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. MonTI vol.13, 2021, págs. 280-300, doi: doi.org/10.6035/MonTI.2021.13.09 Consultado por última vez: 14/08/2024.

Scott, Mike. "WordSmith Tools". *Stroud: Lexical Analysis Sorftware,* lexically.net/wordsmith/. Consultado por última vez: 01/09/2024.

TEI. "Text Encoding Initiative". *TEI. Text Encoding Initiative,* tei-c.org. Consultado por última vez: 14/08/2024.

TEITOK. "About" *TEITOK – The tokenized TEI Enviroment,* www.teitok.org/index.php?action=about. Consultado por última vez: 01/09/2024.

"Transkribus" *Transkribus,* www.transkribus.org/es. Consultado por última vez: 01/09/2024.

Ueda, Hiroto. "LYNEAL, Letras y Números en Análisis Lingüísticos". *LYNEAL,* h-ueda.sakura.ne.jp/lyneal/. Consultado por última vez: 01/09/2024.

Vaamonde Dos Santos, Gael. "Diseño y explotación de un corpus histórico de textos oralizantes para el estudio del español clásico y moderno". *Revista de Humanidades Digitales,* vol.9, 2024, págs. 41-70, doi: doi.org/10.5944/rhd.vol.9.2024.39834. Consultado por última vez: 14/08/2024.

WIKIPEDIA "Diseño web adaptable." *Wikipedia, la enciclopedia libre*, es.wikipedia.org/wiki/Diseño_web_adaptable. Consultado por última vez: 01/09/2024.