Scriptum digital
Vol. 8 (2019), pp. 1-21

# LOS CORPUS DEL ESPAÑOL DESDE LA PERSPECTIVA DEL USUARIO LINGÜISTA

CARLOTA DE BENITO MORENO
*Universidad de Zúrich*
*carlota.debenitomoreno@uzh.ch*
ORCID-iD: https://orcid.org/0000-0001-9112-5471

**RESUMEN**

En este trabajo se realiza una comparación de los corpus más importantes del español (Biblia Medieval, Post Scriptum, CORDE, CREA, CDH, CORDIAM, Corpus del Español, CODEA+ 2015, COSER, PRESEEA, ESLORA, CORPES XXI y Val.Es.Co) en lo que se refiere a su interfaz web, abordando dos aspectos fundamentales: la herramienta de búsqueda y el acceso a las concordancias y textos. La perspectiva adoptada es la del investigador lingüista y usuario de dichos corpus, con el objetivo de proponer caminos para el futuro.

**PALABRAS CLAVE:** diseño de corpus electrónicos, interfaz de usuario, datos ordenados, anotación manual

## SPANISH CORPORA FROM A LINGUIST USER PERSPECTIVE

**ABSTRACT**

This work compares the most important corpora of Spanish (Biblia Medieval, Post Scriptum, CORDE, CREA, CDH, CORDIAM, Corpus del Español, CODEA+ 2015, COSER, PRESEEA, ESLORA, CORPES XXI and Val.Es.Co) with regard to their web interface, addressing two main aspects: the search tool and the access provided to both concordances and texts. The comparison takes the perspective of researchers who are both linguists and users of these corpora, with the goal of proposing future ways.

**KEY WORDS:** Electronic Corpus Design, user interface, tidy data, manual annotation

## 1. INTRODUCCIÓN

En el XI Congreso Internacional de Historia de la Lengua Española, celebrado en Lima en agosto de 2018, tuvo lugar una mesa redonda que llevaba por título "Fuentes y métodos para el estudio de la variación sintáctica", en la que Virginia Bertolotti le preguntó a Andreas Dufter cuál sería su corpus ideal para estudiar el latinismo sintáctico.[1] Esta pregunta es indicativa de las preocupaciones que asaltan constantemente al creador de corpus: ¿cómo puedo mejorar su utilidad para otros usuarios? Con esta pregunta, Bertolotti y Dufter pensaban tanto en tipos de textos abarcados por el corpus como en niveles de anotación de estos, cuestiones que dependen frecuentemente de los intereses específicos de los creadores de corpus así como de las posibilidades de financiación. Sin embargo, existe otro

[1] Este artículo se ha visto muy enriquecido con los comentarios de Ana Estrada Arráez y de dos revisores anónimos, que agradezco profundamente. Por supuesto, cualquier inexactitud que persista solo puede achacárseme a mí.

Fecha de recepción: 30/04/2019
Fecha de aceptación: 27/05/2019
ISSN: 2014-640X

ámbito, más técnico, que determina absolutamente la relación del usuario con el corpus: la interfaz de búsqueda y el acceso a los resultados. En este breve trabajo pretendo hacer un repaso sobre las distintas posibilidades con las que cuenta actualmente el usuario interesado en la historia (y sincronía) de la lengua española, así como proponer una serie de estándares que pueden mejorar esa relación corpus-usuario.

Es fundamental partir de la base de que la lingüística española actual se encuentra en una situación envidiable respecto a otras lenguas en cuanto al número de corpus, la amplitud de tipos de textos representados y la calidad filológica de estos, gracias al esfuerzo y trabajo de muchos investigadores. Proyectos como el CORDIAM o el CORPES XXI, que aumentan sustancialmente los textos americanos de que disponíamos, o la red CHARTA, que promueve unos estándares filológicos de selección y presentación de textos de gran calidad, son indicativos de la buena salud de nuestra lingüística de corpus y del constante progreso que presenciamos. Desde la perspectiva del lingüista-usuario de corpus, que acude a estos para realizar sus investigaciones, creo que la mayor necesidad de mejora no se encuentra en cuestiones del diseño de corpus, de su fiabilidad lingüística o de la cantidad de datos, sino en las interfaces de búsqueda y en la forma de acceder a los resultados.

En la era de la ciencia de datos (*data science*) y el acceso abierto (*open access*), el acceso digital a los excelentes datos de que disponemos para poder tratarlos computacionalmente es esencial.[2] Mi objetivo en este trabajo es acercarme, por lo tanto, a uno de los aspectos del diseño de corpus que menos atención suele recibir en la bibliografía: el acceso a los datos por parte del usuario investigador.[3] Mi punto de partida es la idea de que el análisis lingüístico se realiza con cada vez más frecuencia por medio de herramientas computacionales, como hojas de cálculo para la anotación manual de las concordancias de forma externa (Smith, Hoffman & Rayson 2008) y lenguajes de programación como R para el análisis cuantitativo.[4] En este sentido, este trabajo no se dedica a los corpus como herramientas, sino a las herramientas de los corpus (Anthony 2015). A mi parecer, es fundamental que los corpus lingüísticos mejoren tanto estas como sus interfaces para facilitar el trabajo de anotación manual de los datos que realiza el lingüista empírico. Dicha mejora implica dos aspectos: por un lado, la flexibilización de las herramientas del corpus para mejorar su usabilidad y, por el otro, el poner a disposición de los usuarios, siempre que sea posible, los textos completos (con cuantas capas de anotación estén disponibles), para que aquellos puedan explotarlos creativamente con ayuda de herramientas externas.

El trabajo está basado en la comparación de trece corpus del español, tanto históricos como sincrónicos, a saber, Biblia Medieval (y su versión Biblias Hispánicas),[5] Post Scriptum, CORDE, CREA, CDH, CORDIAM, Corpus del Español (en su interfaz actual), CODEA+ 2015,[6]

[2] Igual que lo es la adecuada formación académica en este aspecto de los jóvenes investigadores. Es urgente la inclusión de curso de ciencia de datos en los másteres de investigación, en las que los jóvenes se formen en el manejo de programas de hojas de cálculo, en lenguajes de programación como R o Python y en estadística, todas ellas herramientas fundamentales de la investigación actual en ciencias sociales.

[3] La bibliografía sobre el diseño y compilación de corpus es muy abundante. Buenas introducciones se encuentran en Torruella & Llisterri (1999), Lüdeling & Kytö (2008), O'Keeffe & McCarthy (2010), Biber & Reppen (2012), Crawford & Csomay (2016), Torruella (2016).

[4] Me centro, además, en la perspectiva del lingüista y no la del lexicógrafo. Para una panorámica de las herramientas útiles para el lexicógrafo se encuentra en Kilgarriff & Kosem (2013). Sobre las distintas necesidades de otros especialistas de las Humanidades, véase Mambrini, Passarotti & Sporleder (2011).

[5] Puesto que muchos aspectos de las interfaces de ambas versiones son similares, todo aquello que se diga sobre Biblia Medieval es aplicable a Biblias Hispánicas (la versión más nueva), pero no viceversa.

[6] Aunque me centro únicamente en el CODEA+ 2015, por ser el corpus de mayor tamaño de toda la red, mis observaciones son en general aplicables a todos los corpus de la red CHARTA, que emplean la misma interfaz.

COSER, PRESEEA, ESLORA, CORPES XXI y Val.Es.Co[7]. En lo que sigue discutiré sobre la interfaz de búsqueda (apartado 2), el tamaño del contexto ofrecido (apartado 3), la herramienta de exportación (apartado 4), el formato de la exportación (apartado 5), la posibilidad de acceso a los textos (apartado 6), la información sobre los corpus (apartado 7) y la posibilidad de que los usuarios colaboren en la creación de los corpus (apartado 8). Se ofrecen unas breves conclusiones en el apartado 9.

## 2. LA INTERFAZ DE BÚSQUEDA

La posibilidad de realizar búsquedas en el interior de un corpus es una de las utilidades básicas de un corpus digital (y lo que marca la diferencia más importante con los corpus en papel). Las opciones para llevar a cabo dichas búsquedas son muy numerosas y tienen que ver con una de las cuestiones más tratadas de la lingüística de corpus: la anotación de los textos.

El formato más básico consiste en la realización de búsquedas por cadenas de texto literal. Si bien todos los corpus examinados ofrecen esta posibilidad, no todos lo hacen de la misma manera. Aunque estas búsquedas pueden corresponder a una simple búsqueda de una cadena de caracteres, generalmente están basadas en un proceso previo de tokenización que ha permitido la identificación de unidades básicas (es decir, de palabras) (véase Kilgarriff & Kosem 2013). La tokenización es fundamental para la correcta identificación de los resultados buscados, puesto que existe una diferencia entre cadena de caracteres y "palabras" (en el sentido gráfico): la separación entre espacios no es suficiente para identificar todas las palabras gráficas de un texto, que pueden estar circundadas por signos de puntuación.[8] Igualmente, los textos de un corpus, especialmente si se trata de transcripciones paleográficas o de audio, pueden contener signos o interrupciones en el interior de las palabras, que impedirían la obtención de los resultados deseados a partir de búsquedas basadas en los textos sin procesar. En el mismo sentido, para el usuario es crucial saber si la herramienta de búsqueda ofrece un tratamiento diferenciado de las mayúsculas o las minúsculas. Un ejemplo de buena práctica es el corpus Biblia Medieval, que informa explícitamente a los usuarios de cómo se ha producido el proceso de tokenización —algo no tan frecuente, pero crucial— y explica que las búsquedas ignoran marcas paleográficas como las barras horizontales que marcan los saltos de línea. Además, permite marcar en la caja de búsqueda si la distinción entre mayúsculas y minúsculas es o no relevante. De esta manera, el usuario conoce todos los detalles de los resultados que producen sus búsquedas.

La tokenización también permite distinguir entre las búsquedas de palabras enteras o las búsquedas por fragmentos de palabras. Prácticamente todos los corpus analizados permiten estos dos tipos de búsqueda, ya sea por medio del uso de comodines (Post Scriptum, CORDE y CREA, CDH, CODEA+ 2015, CORDIAM, Corpus del Español, ESLORA, CORPES XXI), indicaciones ortotipográficas (Biblia Medieval) o por opciones que deben indicarse en la caja de búsqueda (COSER, Val.Es.Co).

La única excepción es el corpus PRESEEA, que no parece estar tokenizado (o no siguiendo criterios lingüísticos). Las búsquedas en este corpus se realizan en las

[7] Aunque me refiera al corpus únicamente como Val.Es.Co en lo que sigue, me centro en la última versión (Val.Es.Co 2.0), que es la que permite acceso digital a todo el corpus.

[8] Véase Anthony (2015) para algunos ejemplos de la manera en que los distintos métodos de tokenización inciden en los resultados del análisis.

transcripciones sin procesar, lo que implica que no existe la posibilidad de buscar por palabra gráfica (la búsqueda entre espacios elimina los resultados que estén precedidos o seguidos de signos de puntuación, por ejemplo). Más aún, las marcas que pueden aparecer en el interior de una palabra gráfica no se eliminan: es decir, una búsqueda como "*que*" no devuelve los resultados de la preposición con apócope de la *-e* final, transcritos como "*qu<[e]>*" (véanse las figuras 1 y 2, que muestran como el primer resultado de la búsqueda de la imagen 1 no puede recuperarse por medio de la búsqueda de la figura 2). Así, una de las pocas excepciones a la literalidad en este corpus parece ser la indistinción entre mayúsculas y minúsculas.

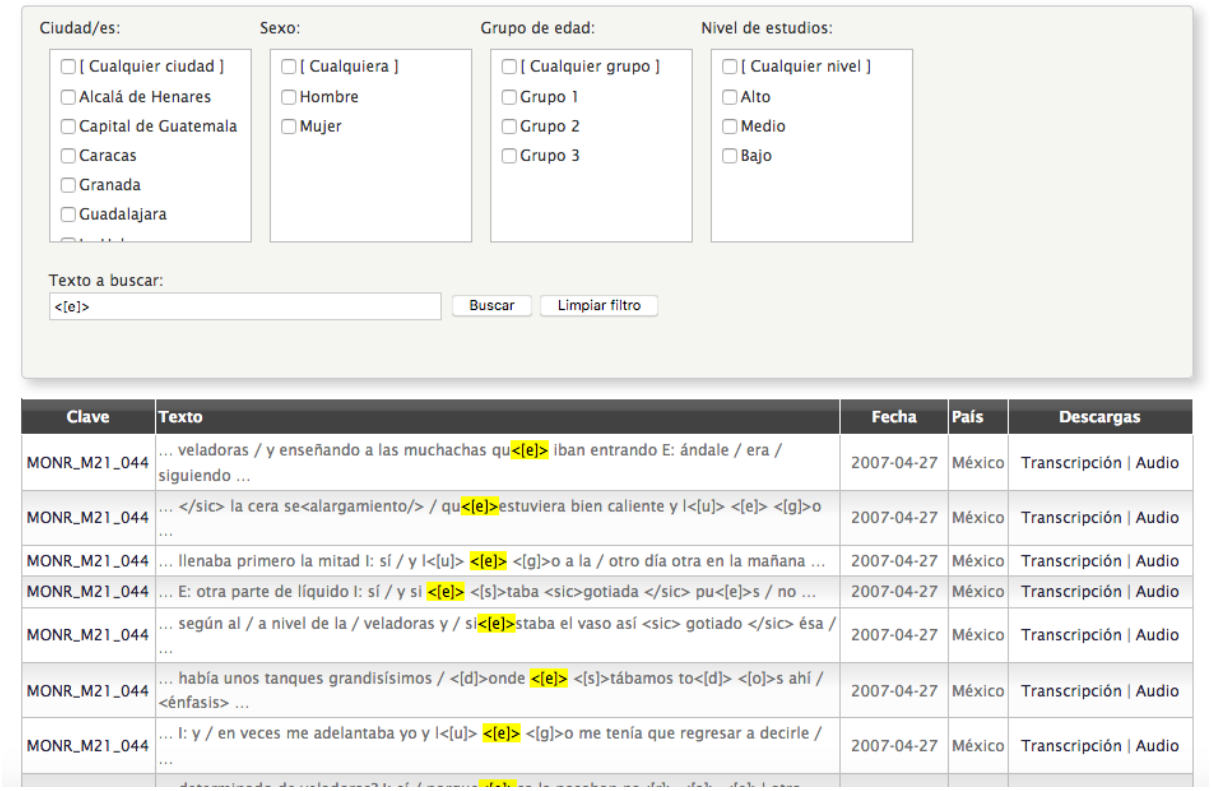



Figura 1. Búsquedas de caída de *e* en PRESEEA

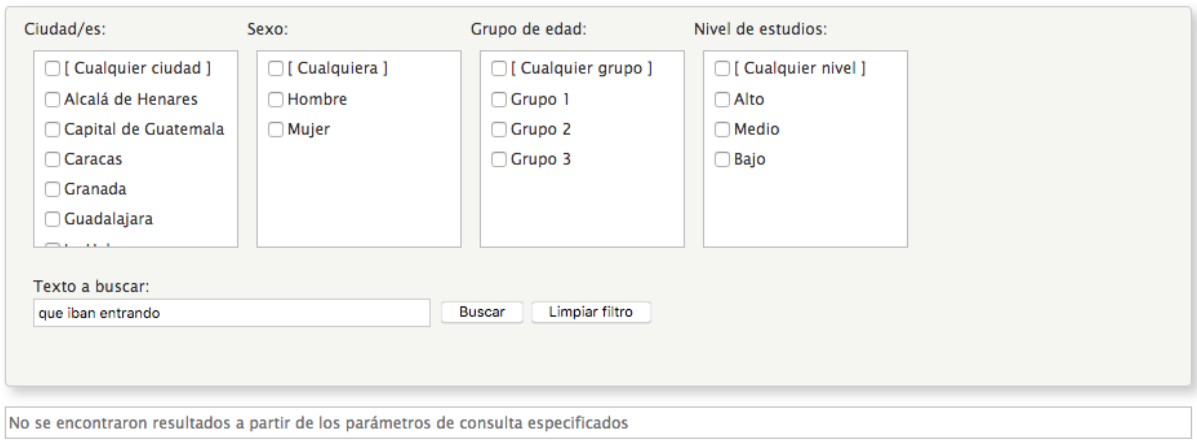



Figura 2. Búsqueda de *que* en un contexto en el que hay caída de *-e*

Más sofisticadas son las búsquedas lematizadas o, incluso, por etiquetas morfosintácticas (véanse Reppen 2010 y Kübler & Zinsmeister 2015 para una enumeración de los distintos niveles de anotación lingüística).[9] La lematización de un corpus consiste en la asignación a cada palabra de su correspondiente forma de diccionario, para poder recuperar con una simple búsqueda todas las formas de un paradigma, mientras que con etiquetado nos referimos a la asignación de rasgos gramaticales a cada palabra. El etiquetado puede ser de varios niveles como, por ejemplo, morfológico, en el que puede anotarse la categoría gramatical de la palabra, también en diversos grados de detalle, o sintáctico, en el que se anota la función sintáctica de una palabra (o un sintagma) en la oración.

Estos procesos, que, lógicamente, se realizan de forma automática,[10] presentan distintos niveles de complejidad, pero la dificultad siempre reside en el mismo aspecto: en la ambigüedad lingüística. En la lematización deben resolverse ambigüedades léxicas (¿es *como* en una oración dada una forma verbal del lema *comer* o una forma invariable que debe englobarse bajo el lema *como*?), mientras que en el etiquetado morfológico deben resolverse ambigüedades categoriales (¿es la forma invariable *como* en una oración dada una conjunción, un relativo o un interrogativo?), y en el etiquetado sintáctico deben resolverse ambigüedades sintácticas (¿es *el perro del vecino* el sujeto o el objeto de una oración dada?).[11]

Resulta fundamental establecer una diferenciación entre lematizado y etiquetado en lo que se refiere a sus objetivos. Mientras que el lematizado permite ofrecer más resultados de los que ofrecería una simple búsqueda por forma, el etiquetado permite devolver menos resultados de los que ofrecería una simple búsqueda por forma. En ambos hay, sin embargo, margen de error (véase también Kübler & Zinsmeister 2015). Por un lado, dado el elevado nivel de acierto de los softwares más utilizados (al menos en lo que se refiere al lematizado y la anotación morfológica), este margen de error nos puede parecer despreciable. Por otro lado, los errores se producirán generalmente en los casos de usos más sorprendentes, que pueden también ser muy interesantes —incluso los más interesantes— para el investigador, por lo que en este caso la tecnología puede volverse en nuestra contra, impidiéndonos encontrar resultados dignos de ser considerados (generando falsos negativos, es decir, resultados deseados no encontrados).[12]

Este problema puede paliarse ofreciendo un lematizado totalmente inclusivo, que devuelva todas las formas potencialmente correspondientes al lema, sin desambiguar. En un

[9] Consúltese también Davies (2009) para una demostración de las posibilidades que abren los diferentes niveles de etiquetado. Una cuestión debatible puede ser, sin embargo, el grado de confianza que debe prestar el lingüista a estos procesos cuando no los ha llevado a cabo él mismo y no cuenta con una descripción detallada de ellos. En mi opinión, para la fiabiiidad de nuestros resultados es crucial entender cómo se han obtenido los datos, para poder ser conscientes de posibles sesgos o carencias.

[10] Los avances que produce la lingüística computacional en este aspecto son constantes y se enfocan a reducir el margen de error de los distintos métodos disponibles, ya estén basados en reglas, en procedimientos estadísticos o en la combinación de ambos. Mientras que los lematizadores y los etiquetadores morfosintácticos pueden llegar a alcanzar porcentajes de acierto muy cercanos al 100% (Padró et al. 2010, Le Roux, Sagot & Seddah 2012, Straka, Hajič & Straková 2016), los anotadores sintácticos están todavía lejos de estas cifras, rondando porcentajes de entre el 79 % y el 88 % (Straka, Hajič & Straková 2016, Le Roux, Sagot & Seddah 2012).

[11] Puesto que la separación de estas ambigüedades es más teórica que práctica, no es raro que los distintos procesos de anotación se lleven a cabo de forma conjunta (Straka, Hajič & Straková 2016).

[12] Véase Anthony (2015) para un resumen de las posturas teóricas a favor y en contra de la lematización de los corpus.

sistema así, la búsqueda del lema *comer* devolvería todas las formas de *como*, que corresponde a dos lemas: *comer* y *como*. De esta manera aumenta el número de falsos positivos (resultados encontrados no deseados), que deben ser identificados manualmente por el investigador.

Más complejo técnicamente sería disponer de herramientas de lematización y etiquetado que permitan una identificación más refinada de formas ambiguas, que asigne un etiquetado múltiple solo a estas. Es decir, a la hora de analizar la forma *cosa* podemos encontrar casos que inequívocamente corresponden a la forma verbal (por ejemplo, si el sujeto es explícito: *ella cosa*) y casos que inequívocamente corresponden al sustantivo (*gran cosa*), pero también casos ambiguos, como *la cosa*, en el cual *la* puede ser determinante o pronombre. Si solo a este último caso le asignamos una etiqueta doble (referida al sustantivo *cosa* y al verbo *coser)* minimizamos tanto los falsos positivos como los falsos negativos.

De los corpus tomados en consideración, están lematizados y etiquetados el Post Scriptum, el CORPES XXI, el CDH, el Corpus del Español, el ESLORA, Val.Es.Co 2.0, el COSER (en su acceso a través de la consulta avanzada) y Biblias Hispánicas. El CORDIAM está solo lematizado parcialmente por el momento . Los dos recursos habituales para permitir estos diferentes tipos de búsquedas son habilitar diferentes cajas de búsqueda (por forma, por lema o por etiqueta morfosintáctica) o una única caja de búsqueda en la que puede marcarse la categoría correspondiente (lo que hace el CORDIAM): de esta manera el usuario puede diseñar fácilmente su búsqueda. El uso de diferente cajas de búsqueda permite la combinación de los distintos tipos de búsquedas, que puede ayudar a acotar estas (por ejemplo, se podría buscar la forma *fueron* en combinación con el lexema *ir*, con el objetivo de evitar los casos en que *fueron* es una forma del verbo *ser*). Esto se puede conseguir en el COSER, el CORPES XXI, el CDH, el ESLORA, el Post Scriptum y en Val.Es.Co 2.0. Esta posibilidad no existe en Biblias Hispánicas, donde el lema se distingue de la forma al escribirse entre corchetes, ni en el Corpus del Español, donde parece que es la tipografía la que marca la diferencia entre forma (búsqueda en minúsculas) y lema (búsqueda en mayúsculas) —aunque esto no parece estar explicitado en la ayuda de la búsqueda, sino que puede deducirse de los ejemplos que se dan en el apartado *Dialects*, por ejemplo (véase el ejemplo con *GUSTAR* en la figura 3)—. En cuanto a la búsqueda por etiquetas, puesto que cada corpus suele contar con sus propias convenciones, lo habitual es contar con menús desplegables que permiten seleccionar las categorías deseadas y así se hace en Biblias Hispánicas, Post Scriptum, el CDH, el CORPES XXI, el CODEA+ 2015, el Corpus del Español, el COSER y el ESLORA. Val.Es.Co 2.0 no cuenta con este sistema, sino que exige que el usuario conozca las etiquetas EAGLE, que pueden escribirse en la caja de búsqueda correspondiente.

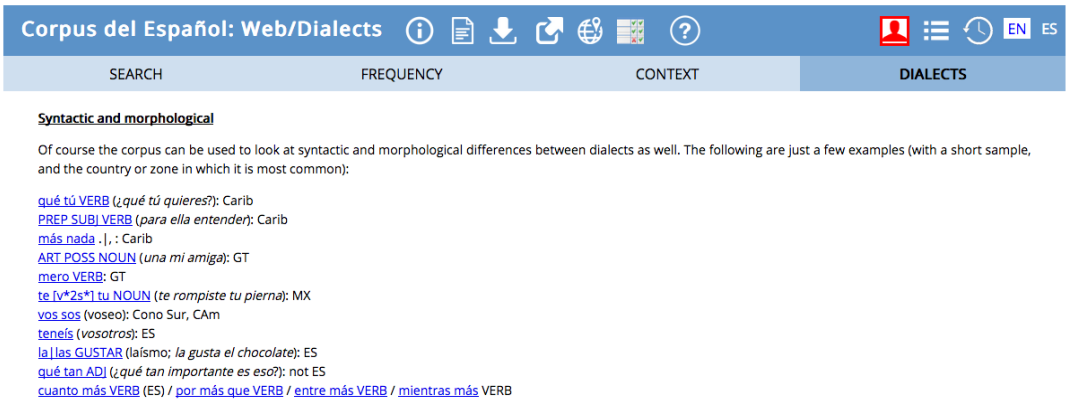



Figura 3. Ejemplos ofrecidos por el Corpus del Español

Antes de pasar a las búsquedas complejas, me gustaría mencionar la posibilidad de un nivel de etiquetado de complejidad inferior al lematizado (y que, de hecho, es un requisito para este) y que, por lo que sé, solo se ofrece en el Post Scriptum: la devolución de todas las formas ortográficas de una forma determinada. Cuando nos interesa una única forma y no un lema entero (por ejemplo, *hay*), poder obtener todas las formas ortográficas con una simple búsqueda puede resultar extremadamente útil —especialmente en textos históricos, pero también en otros casos de escritura no estándar, como cartas privadas, comunicación mediada por ordenador o transcripciones semiortográficas—. Post Scriptum ofrece transcripciones estandarizadas además de las literales y permite hacer búsquedas en cualquiera de ellas, con la ventaja de que los resultados siempre se devuelven en la transcripción literal (por lo que siempre accedemos a "la realidad" de los textos). Así, la búsqueda de "*se*" en las transcripciones estandarizadas devuelve casos de *ce*, de *sse* y de formas enclíticas (como *irse*, *darssele*, etc.), siendo un excelente ejemplo de exhaustividad (véase la figura 4). Por supuesto, cuando este etiquetado se realiza de forma automática, aparecen problemas de ambigüedades, por lo que las observaciones realizadas más arriba son también aplicables en este caso.

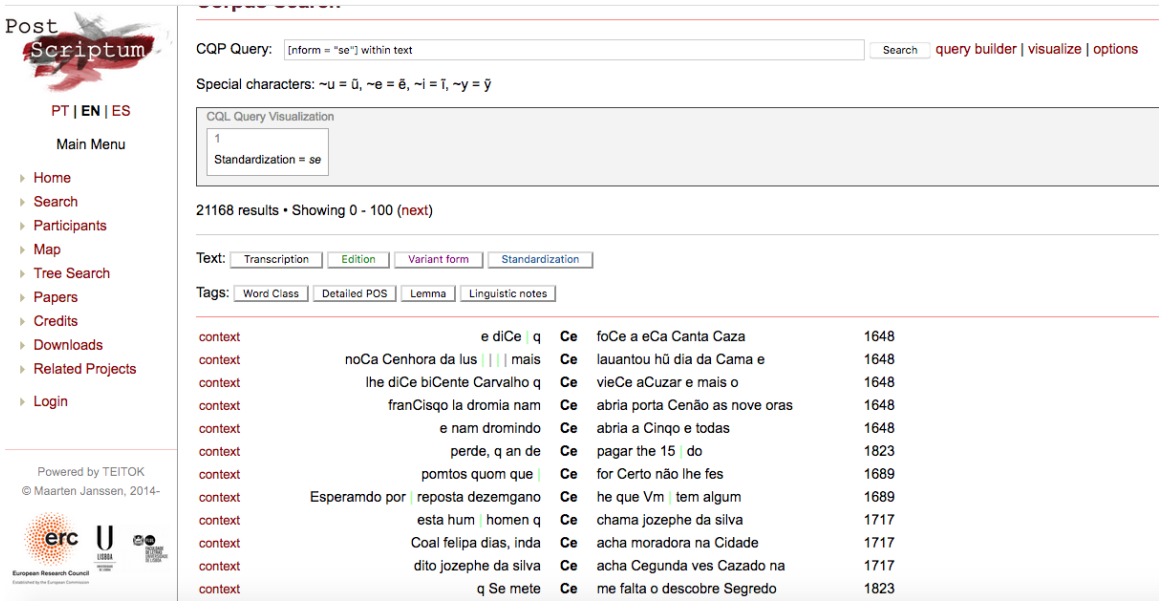



Figura 4. Búsquedas de la forma estandarizada *se* en Post Scriptum

Una utilidad esencial de los corpus es la posibilidad de realizar búsquedas complejas, es decir, búsquedas que involucran más de una palabra. Estas pueden consistir tanto en la búsqueda de palabras que aparecen a cierta distancia entre sí como en búsquedas de una palabra que tengan en cuenta la aparición o ausencia de otra palabra en el contexto cercano o, incluso, en la búsqueda de palabras alternativas. Estas últimas suelen realizarse por medio de la inserción de los operadores lógicos Y, NO y O respectivamente, cuya notación puede variar de corpus a corpus. Esta posibilidad la ofrecen todos los corpus académicos (salvo el CORDIAM), así como Biblia Medieval. El COSER permite la búsqueda de palabras alternativas en su búsqueda avanzada (en la búsqueda básica pueden emplearse expresiones regulares, con lo que el usuario ducho en estas puede aplicar estas posibilidades). La búsqueda de secuencias de palabras (contiguas o a cierta distancia) se habilita generalmente a través de

diversas cajas de búsqueda en las que puede explicitarse la distancia y la posición respecto de la primera palabra de la búsqueda (derecha o izquierda) (CDH, CORPES XXI, CODEA+ 2015, CORDIAM, COSER, ESLORA, Val.Es.Co 2.0). Post Scriptum usa la misma caja de búsqueda, que se vacía cada vez que se añade una palabra a la término de búsqueda, pero no permite escoger la distancia a la que estas deben encontrarse. El CORDE, el CREA, Biblias Hispánicas y el PRESEEA permiten la adición de distintos términos en la misma caja de búsqueda (véase la figura 5). En los tres primeros, que están tokenizados y permiten el uso de comodines, estos se pueden emplear para indicar las distancias a las que deben estar los distintos términos insertados, pero no ocurre lo mismo en PRESEEA. El Corpus del español tiene la posibilidad de buscar *collocates* (permitiendo solo la búsqueda de dos términos), para los cuáles se puede indicar una distancia mínima (pero no exacta).

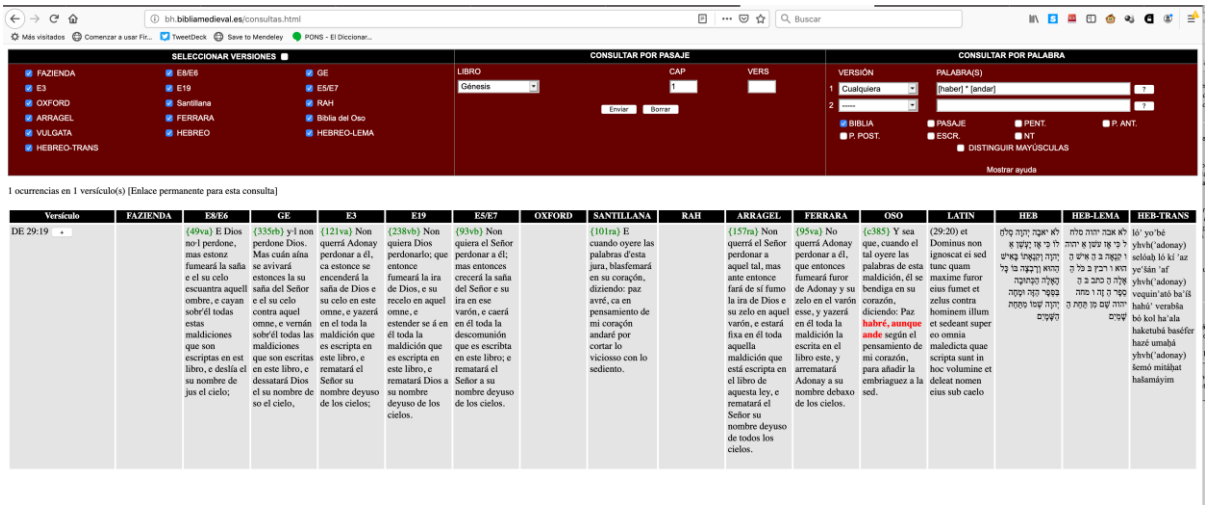



Figura 5. Búsquedas de lemas combinados en Biblias Hispánicas

Por último, la interfaz de búsqueda suele contener también una serie de opciones de filtrado que permite obtener únicamente los resultados de una parte del corpus completo. Esta opciones dependen, claro está, del contenido del corpus: mientras que en Biblias Hispánicas se puede filtrar por obra, libro e incluso pasaje bíblico, en el COSER se puede filtrar por provincia, década de nacimiento, hablante (informante o cualquiera), tema o marcas de la conversación. En corpus escritos, son parámetros habituales el año, el autor, la obra, el país, el tipo de texto, etc. Debemos congratularnos de la posibilidad de distinguir año de creación y testimonio en el CDH, a diferencia de lo que permite el CORDE, lo que permite una consulta mejor informada filológicamente (Rodríguez Molina y Octavio de Toledo 2017). Mientras que la mayoría de los corpus ofrecen esta opción también por medio de cajas de búsqueda (con menús desplegables o no) dentro de la propia interfaz, en el Corpus del Español debe crearse un subcorpus (lo que es menos intuitivo, pero tiene la ventaja de permitir guardar ese subcorpus para otras búsquedas) y Val.Es.Co lo ofrece en pantallas distintas (lo cual también resulta menos intuitivo). La solución de CODEA, que presenta las cajas de filtrado encima de cada columna de los resultados, se nos antoja excelente: es de una gran flexibilidad, ya que permite tanto hacer como deshacer filtros sucesivos, y elegantemente intuitiva.

## 3. EL CONTEXTO DE LAS CONCORDANCIAS

Una vez realizada la(s) búsqueda(s) deseada(s), el usuario obtiene acceso a los resultados o concordancias, que generalmente se ofrecen siguiendo el sistema *Key Word In Context* (KWIC): con la secuencia buscada en el centro de la concordancia y algo de contexto previo y posterior. En este apartado me detengo brevemente en el tamaño de dicho contexto, otro aspecto en el que existe mucha disparidad según el corpus.

Por supuesto, que los resultados tengan un contexto suficiente para que podamos descifrar bien el significado de las formas que nos interesan es absolutamente esencial. Podemos establecer una primera diferencia entre aquellos corpus que ofrecen unidades de sentido completas y aquellos que ofrecen contextos de longitud más o menos definida, sin tener en unidades de sentido superiores a la palabra. Al primer grupo pertenecen Biblia Medieval por un lado (que ofrece el versículo que contiene la forma buscada) y, por otro, COSER y Val.Es.Co (que ofrecen intervenciones de los hablantes). En el resto de casos el criterio para limitar los contextos parece ser una combinación entre número de palabras y longitud total de la secuencia. Muchas veces estos contextos son extremadamente breves (un caso extremo es Post Scriptum, que por defecto devuelve secuencias de cinco palabras antes y después del término buscado, aunque puede ampliarse hasta siete).

La opción de poder acceder a más contexto si así se desea está presente en muchos de los corpus (en CORDE, CREA, CDH, CORDIAM, Corpus del Español, CORPES XXI, ESLORA y COSER). Sin embargo, puesto que esta opción casi siempre supone que el contexto ampliado se ofrece en una ventana nueva (es decir, no insertado en la tabla en que se devuelven los resultados), la ampliación del contexto generalmente no se incluye en la exportación de los resultados (cuando esta opción existe) o no se puede copiar junto con el resto de resultados, como sería deseable. Al contrario, debe ampliarse cada contexto individualmente y añadirlo de forma manual a los resultados exportados, sobre los que habitualmente realizamos la anotación lingüística. Una excepción es el COSER, que en la búsqueda avanzada devuelve únicamente la intervención en que se encuentra la forma buscada (en la búsqueda sencilla se devuelven varias intervenciones antes y después, por lo que el contexto suele ser suficiente), pero permite añadir más contexto allí donde se desee pulsando un icono con un + (véase la figura 6) e incluye dicho contexto en la exportación.

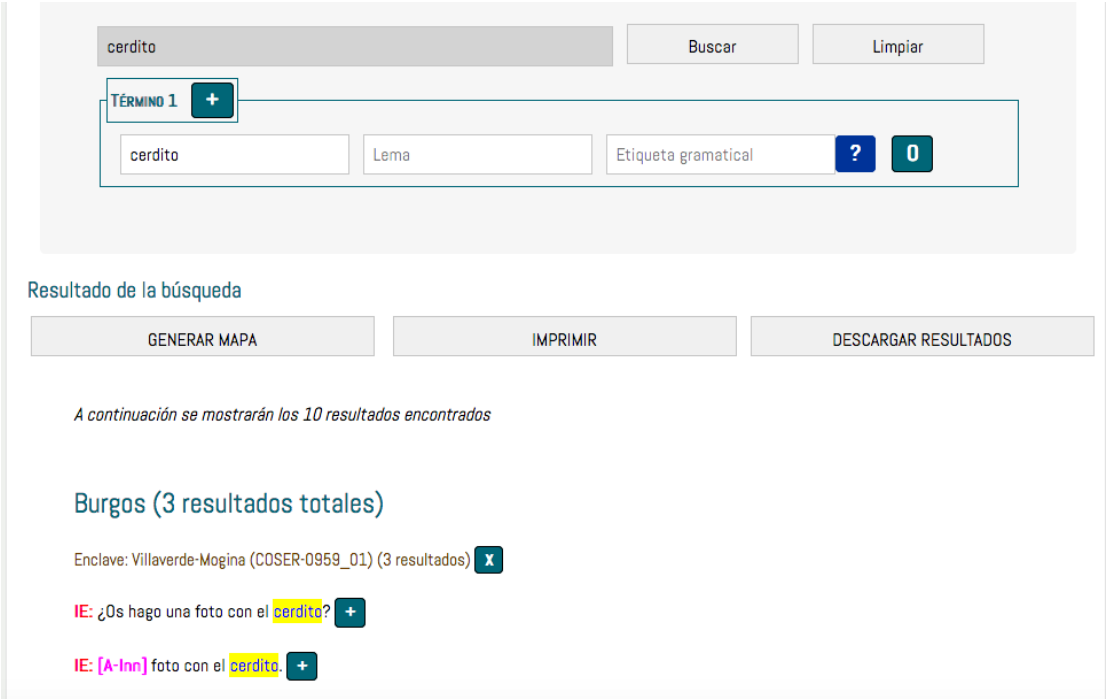



Figura 6. Resultados devueltos en el COSER

Lo ideal, en mi opinión, consiste en devolver un contexto basado en unidades significativas superiores a la palabra (oraciones, párrafos, versículos, intervenciones, etc.), para facilitar la comprensión del fragmento. Asimismo, un contexto abundante es preferible a un contexto escaso, puesto que las opciones para solucionar dicha escasez suponen generalmente mucho tiempo y no se producen de forma automática, como acabamos de ver. Si no quiere (o puede) incrementarse el coste computacional que pueda suponer la devolución de un contexto amplio por defecto, puede incluirse el tamaño del contexto como una opción más de la búsqueda: el usuario puede entonces decidir si quiere un contexto más o menos abundante, a cambio de un tiempo mayor de espera en la recuperación de las concordancias.

Por último, debe mencionarse una opción muy útil que presentan CORPES XXI, el CDH y Post Scriptum: la de ordenar los resultados obtenidos por el contexto precedente o por el contexto posterior (véase la figura 7). Puesto que las palabras inmediatamente anteriores o posteriores son estadísticamente muy informativas a la hora de analizar lingüísticamente un elemento, esta posibilidad supone una gran ayuda para el lingüista.[13]

[13] Especialmente porque al menos la ordenación por palabra inmediatamente anterior exige unos conocimientos de procesamiento del lenguaje natural elevados.


Figura 7. Los resultados del CORPES XXI ordenados por el lema inmediatamente a la izquierda del término buscado

## 4. LA EXPORTACIÓN DE LAS CONCORDANCIAS

Una vez obtenidas las concordancias, la cuestión clave para el acceso a los resultados es la posibilidad de exportar los resultados, es decir, de extraerlos automáticamente para poder trabajar con ellos desde herramientas externas. Por lo tanto, un aspecto fundamental de la usabilidad de un corpus tiene que ver con el formato de dicha exportación, que debe permitir trabajar con las ocurrencias obtenidas de forma sencilla. En este apartado me detengo en la posibilidad de exportar dichos resultados y en cómo se materializa en los distintos corpus investigados, mientras que en el siguiente me centro en la información (los metadatos) que contiene la exportación.

Tanto la anotación manual como el posterior análisis (contar ocurrencias, realizar estadísticas, etc.) de las ocurrencias recuperadas debe realizarse óptimamente en formato tabular. Para ello, la solución más sencilla es ofrecer la exportación de los resultados en una hoja de cálculo. Sin embargo, muchos de los corpus del español no permiten esta opción. Por un lado, encontramos corpus que no permiten la exportación en ningún formato. Es lo que ocurre en el CORDE, el CREA, el Corpus del Español y el PRESEEA: si queremos extraer los resultados de estos corpus, debemos copiar y pegar directamente desde la interfaz web, una tarea que exige mucho tiempo del investigador —hay que copiar cada página—. Mientras que los dos últimos corpus mantienen el formato tabulado cuando se copian, no ocurre lo mismo con los dos primeros, lo que genera dificultades extra. El CDH, por su parte, solo permite "imprimir" los resultados de cada página: para ello, genera un archivo html con las ocurrencias en el que se eliminan todos los metadatos (véase el apartado 5) que aparecen en la interfaz web: es justo decir que esta función no tiene ninguna utilidad para el investigador.

Biblia Medieval y Post Scriptum permiten exportar los resultados a un archivo de texto (.txt) sin tabular, aunque la tabulación solo sería de utilidad en el primer caso, ya que Post Scriptum tampoco ofrece información sobre los metadatos de las ocurrencias. El resto de corpus analizados sí permiten exportar las ocurrencias en formato tabulado: el CORDIAM genera un archivo XML que puede abrirse directamente desde un editor de hojas de cálculo;

el COSER, un archivo .xls; Val.Es.Co, un archivo .xlsx, y tanto el CORPES XXI como el ESLORA, un archivo de texto plano separado por tabulaciones (.tsv).[14] El CODEA+ 2015 dispone de una función de exportar a Excel (que por el momento exige tener el complemento de Adobe Flash Player activado y autorizado), así como de una función de copiar todos los resultados en formato tabulado que puede pegarse cómodamente en cualquier programa, incluidas hojas de cálculo, de manera que es el usuario el que puede decidir el formato en que desea guardarlos.

Además del formato en que se pueden exportar los resultados, que determina crucialmente la cantidad de trabajo que debe invertirse en preprocesar los datos para luego trabajar en una hoja de cálculo, existe otro aspecto importante en la exportación: ¿cuántas ocurrencias pueden descargarse de una vez? Otra vez, los corpus difieren sustancialmente a la hora de solucionar esta cuestión. Por supuesto, lo más cómodo para el usuario es poder descargar todas las ocurrencias de una vez, ya que supone un importante ahorro de tiempo. Esto lo permiten Biblia Medieval, Post Scriptum y Val.Es.Co. El CODEA+ 2015 vuelve a presentar una situación peculiar, pues, si hacemos una búsqueda con comodines que devuelve varias formas, el resultado de la búsqueda aparece dividido en dichas formas: si bien es posible descargar todos los resultados de cada forma de una única vez, no podemos descargar todos los resultados de todas las formas de una única vez. Por lo tanto, con una búsqueda como "*pod**", que devuelve 113 formas distintas, debemos abrir 113 ventanas diferentes para poder descargar todos los resultados. Esto supone un gasto de tiempo innecesario, especialmente teniendo en cuenta que el CODEA+ 2015 ofrece la forma buscada en una columna separada en sus resultados, facilitando enormemente el filtrado de las formas no deseadas de un hipotético documento unitario. Las herramientas de exportación de otros corpus solo descargan los resultados de la página que está abierta, lo que significa que debemos descargar cada página individualmente. Es el caso del CORDIAM, que presenta solo 20 resultados por página, o del ESLORA, que devuelve 50 ocurrencias por página. Por su parte, el COSER y el CORPES XXI descargan un máximo de 1000 ocurrencias, sin que esté claro cómo puede accederse a las restantes: para obtener todos los resultados debemos combinar búsquedas más restringidas que den menos resultados: cuanto mayor es el corpus, más gravosa es la tarea para el investigador.

El caso de los corpus que no ofrecen posibilidad de exportación y que, por tanto, exigen que copiemos los resultados directamente de la web, supone necesariamente que debemos copiar los resultados página a página. En este caso, el PRESEEA es el que resulta más cómodo, pues muestra todos los resultados de la búsqueda en la misma página, que luego pueden copiarse ya tabulados. Este corpus, por lo tanto, funciona a efectos prácticos como si tuviera la posibilidad de exportación, con la salvedad de que la existencia de esta le ahorra al usuario la necesidad de seleccionar todos los resultados de la tabla, lo que también puede ser fatigante, sobre todo si las ocurrencias son abundantes. El CORDE y el CREA solo ofrecen 25 resultados por página y el CDH, 20, por lo que extraer sus ocurrencias es extraordinariamente pesado. El Corpus del Español presenta un máximo de 1000 resultados en su interfaz de búsqueda y, si esta se hizo por anotación morfológica (*PoS tagging*) o por lema, exige que se pinche en cada forma por separado, lo que vuelve a exigir un tedioso esfuerzo por parte del usuario.

Solucionar estas cuestiones supone una gran mejora en la usabilidad de los corpus, aunque implican llegar a un compromiso entre la cantidad de resultados descargables y el

[14] Es especialmente de lamentar la falta de una funcionalidad parecida en el CDH, teniendo en cuenta que el CORPES XXI sí la tiene.

coste que esto exige. En mi opinión, la gran velocidad con la que dichos resultados se ofrecen podría incluso sacrificarse a favor de la posibilidad de descargar los resultados en uno o pocos clics (que no necesariamente ha de ser la opción por defecto de cada corpus, sino que puede ofrecerse al usuario). En el siguiente apartado profundizo en el formato que deben tener los resultados exportados para que el investigador pueda sacarles provecho.

## 5. EL FORMATO DE LA EXPORTACIÓN

Para poder analizar correctamente los datos, la exportación de los resultados, además de contener las ocurrencias encontradas, necesita los metadatos asociados con dichas ocurrencias: los datos del texto en que aparecen, su localización geográfica, la fecha, las características del hablante, etc. (véase también Kilgarriff & Kosem 2013). Por supuesto, lo ideal es que los corpus ofrezcan todos los metadatos disponibles, para que sea el usuario el que pueda decidir la información que le interesa. Otra vez nuestros corpus presentan extremos muy diferentes en este aspecto. Por un lado, el Post Scriptum no ofrece *ninguna* información sobre el origen de sus ejemplos ni en las búsquedas ni en la exportación, lo que hace que esta sea inservible a muchos efectos (véase la figura 8).[15] El CDH tampoco ofrece ninguna información en la funcionalidad de "Imprimir", pero ofrece país y año en la búsqueda que devuelve la interfaz web. Para obtener más información hay que pinchar en los ejemplos, lo cual otra vez resulta poco conveniente por la cantidad de tiempo que exige al investigador interesado. El Corpus del Español solo ofrece información sobre siglo, género textual (o eso deducimos, ya que no se explicita) y un identificador del texto, aunque en ocasiones esto se indica con etiquetas de poca utilidad como "entrevista".

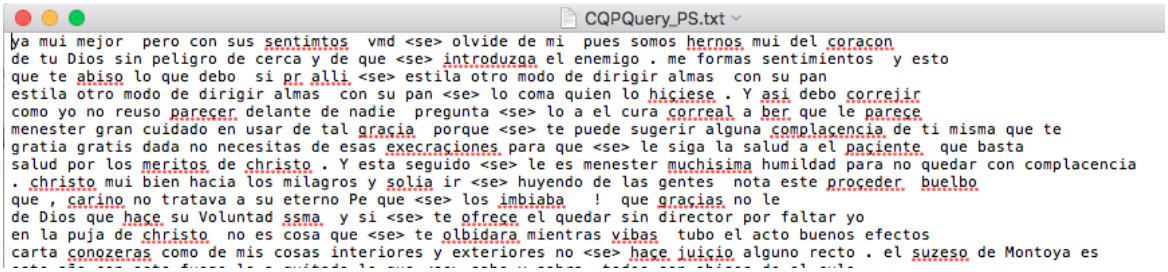



Figura 8. Resultados exportados de la búsqueda "se" en Post Scriptum

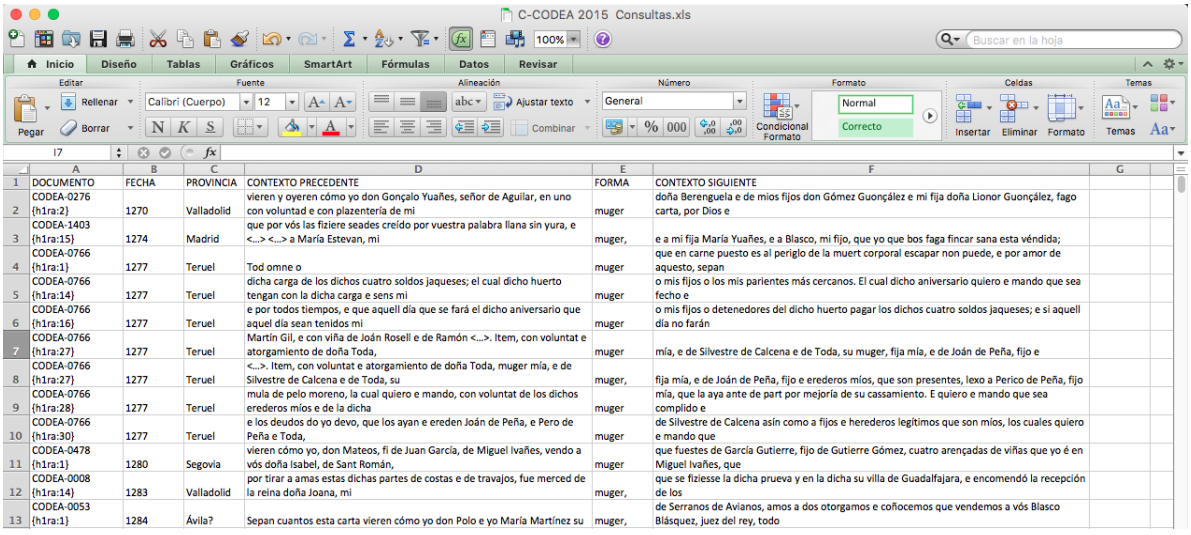
La mayoría de los corpus académicos (incluido el CORDIAM, pero con la excepción ya mencionada del CDH) son ejemplares, igual que el ESLORA, puesto que ofrecen una amplia lista de información sobre los textos y/o hablantes. Muchos de los corpus ofrecen solo algunos metadatos, pero contienen información suficiente para buscar el resto de información en otro lugar del corpus. Así, el corpus Val.Es.Co solamente empareja los resultados con el identificador de la conversación, por lo que el resto de información (año y número de hablantes) debe buscarse en la información sobre las conversaciones; Biblia Medieval da texto y versículo, pero no fecha o procedencia geográfica, que son aspectos que se encuentran en la información sobre el corpus; el COSER da información sobre el enclave y

[15] La información puede recuperarse, porque el corpus es descargable en su totalidad, pero exige mucho tiempo del usuario sin conocimientos de programación.

la provincia, así como el identificador de la entrevista: la información sobre los hablantes puede encontrarse en la ficha técnica de cada entrevista.

Al usuario le sería de gran utilidad en estos casos contar con tablas con los metadatos de cada texto o entrevista. De esta manera, bastaría con indicar el identificador del texto o entrevista del que proceden cada ocurrencia en la exportación de los resultados: el resto de información puede obtenerse fácilmente uniendo la tabla de ocurrencias con la de metadatos (algo que permiten la mayoría de herramientas externas, por medio de la operación *join*). En corpus que ponen el foco en la distribución geográfica de los datos, como el CODEA+ 2015 o el COSER, poner a disposición del usuario una tabla con las coordenadas de cada localidad sería extremadamente útil —y fácil, puesto que los corpus que permiten cartografiar los resultados de forma dinámica ya disponen de esta información, que puede llegar a ser muy costosa de obtener —especialmente porque algunos puntos son difíciles de localizar—. Post Scriptum sí ofrece esta información, aunque no en forma tabulada: se encuentra en la cabecera de cada documento cuando los descargamos en formato XML-TEI P5.

Por otro lado, resulta conveniente que los resultados (ya junto con sus metadatos) se ofrezcan como lo que se conoce como "datos ordenados" (*tidy data*) —un buen ejemplo es el formato del CODEA+ 2015, en la figura 9—. Esto implica dos características fundamentales: en primer lugar, cada ocurrencia debe presentarse en una fila distinta y, en segundo lugar, a cada tipo de metadato le debe corresponder una columna distinta. Los datos ordenados presentan siempre forma tabular (por lo que un formato .txt sin delimitar, como el ofrecido por Biblia Medieval no cumple con esta característica). Asimismo, etiquetas como "19-OR" (del Corpus del Español), "MADR_H23_033" (del PRESEEA) o "x3^00GE1^1^1^0^1ra^" (de Biblia Medieval), que contienen distintos tipos de información (siglo y género textual; ciudad, sexo, edad y entrevista; testimonio, libro, capítulo y versículo respectivamente) deberían separase en distintas columnas para cumplir con los requisitos de los datos ordenados. Idealmente, las columnas deberían tener un nombre que explique la naturaleza de la información que indican y, si se emplean abreviaturas, explicar su significado.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | DOCUMENTO | FECHA | PROVINCIA | CONTEXTO PRECEDENTE | FORMA | CONTEXTO SIGUIENTE | |
| 2 | CODEA-0276 {h1ra:2} | 1270 | Valladolid | vieren y oyeren cómo yo don Gonçalo Yuañes, señor de Aguilar, en uno con voluntad e con plazentería de mi | muger | doña Berenguela e de mios fijos don Gómez Guonçález e mi fija doña Lionor Guonçález, fago carta, por Dios e | |
| 3 | CODEA-1403 {h1ra:15} | 1274 | Madrid | que por vós las fiziere seades creído por vuestra palabra llana sin yura, e <...> <...> a María Estevan, mi | muger, | e a mi fija María Yuañes, e a Blasco, mi fijo, que yo que bos faga fincar sana esta véndida; | |
| 4 | CODEA-0766 {h1ra:1} | 1277 | Teruel | Tod omne o | muger | que en carne puesto es al periglo de la muert corporal escapar non puede, e por amor de aquesto, sepan | |
| 5 | CODEA-0766 {h1ra:14} | 1277 | Teruel | dicha carga de los dichos cuatro soldos jaqueses; el cual dicho huerto tengan con la dicha carga e sens mi | muger | o mis fijos o los mis parientes más cercanos. El cual dicho aniversario quiero e mando que sea fecho e | |
| 6 | CODEA-0766 {h1ra:16} | 1277 | Teruel | e por todos tiempos, e que aquell día que se fará el dicho aniversario que aquel día sean tenidos mi | muger | o mis fijos o detenedores del dicho huerto pagar los dichos cuatro soldos jaqueses; e si aquell día no farán | |
| 7 | CODEA-0766 {h1ra:27} | 1277 | Teruel | Martín Gil, e con viña de Joán Rosell e de Ramón <...>. Item, con voluntat e atorgamiento de doña Toda, | muger | mía, e de Silvestre de Calcena e de Toda, su muger, fija mía, e de Joán de Peña, fijo e | |
| 8 | CODEA-0766 {h1ra:27} | 1277 | Teruel | <...>. Item, con voluntat e atorgamiento de doña Toda, muger mía, e de Silvestre de Calcena e de Toda, su | muger, | fija mía, e de Joán de Peña, fijo e erederos míos, que son presentes, lexo a Perico de Peña, fijo | |
| 9 | CODEA-0766 {h1ra:28} | 1277 | Teruel | mula de pelo moreno, la cual quiero e mando, con voluntat de los dichos erederos míos e de la dicha | muger | mía, que la aya ante de part por mejoría de su cassamiento. E quiero e mando que sea complido e | |
| 10 | CODEA-0766 {h1ra:30} | 1277 | Teruel | e los deudos do yo devo, que los ayan e ereden Joán de Peña, e Pero de Peña e Toda, | muger | de Silvestre de Calcena asín como a fijos e herederos legítimos que son míos, los cuales quiero e mando que | |
| 11 | CODEA-0478 {h1ra:1} | 1280 | Segovia | vieren cómo yo, don Mateos, fi de Juan García, de Miguel Ivañes, vendo a vós doña Isabel, de Sant Román, | muger | que fuestes de García Gutierre, fijo de Gutierre Gómez, cuatro arençadas de viñas que yo é en Miguel Ivañes, que | |
| 12 | CODEA-0008 {h1ra:14} | 1283 | Valladolid | por tirar a amas estas dichas partes de costas e de travajos, fue merced de la reina doña Joana, mi | muger, | que se fiziesse la dicha prueva y en la dicha su villa de Guadalfajara, e encomendó la recepción de los | |
| 13 | CODEA-0053 {h1ra:1} | 1284 | Ávila? | Sepan cuantos esta carta vieren cómo yo don Polo e yo María Martínez su | muger, | de Serranos de Avianos, amos a dos otorgamos e conocemos que vendemos a vós Blasco Blásquez, juez del rey, todo | |

Figura 9. Los resultados de la búsqueda en CODEA+ 2015 exportados en formato .xls

Finalmente, quiero hablar de una última característica que puede mejorar la experiencia del usuario con respecto a la exportación de los resultados. Se trata de resaltar de alguna manera los términos de la búsqueda, puesto que así se facilita la localización de aquello que se está buscando (especialmente si el contexto es abundante). Esto puede hacerse de varias maneras: por ejemplo, Post Scriptum enmarca el término de la búsqueda con comillas angulares simples; el COSER con un doble asterisco; el CORDIAM en negrita y de color rojo. El resalte (subrayado y en azul) que emplean el CORDE y el CREA en la visualización en web se mantiene cuando lo copiamos y pegamos, junto con el enlace que nos permite consultar más contexto, y también se mantiene el resalte en negrita y verde que emplea el Corpus del Español en su visualización web. No ocurre lo mismo con el PRESEEA: en la web se emplea un resalte en amarillo que se pierde al copiar los resultados. Biblia Medieval resalta en negrita las búsquedas en la web, pero no lo mantiene en el archivo de descarga. Resulta excelente, en mi opinión, la opción elegida por el CODEA+ 2015 y el ESLORA de separar el término de la búsqueda en una columna aparte —por lo que la columna de la izquierda será el contexto previo y la de la derecha, el posterior—. Este método no solo ayuda a la rápida identificación de la búsqueda, sino que permite también aplicar filtros a los datos y ordenarlos, opciones muy útiles para la anotación manual de nuestros ejemplos —posibilidad que también existe cuando el corpus ofrece en una columna separada el contenido de la búsqueda, como hacen el COSER y Val.Es.Co—. Es, por ello, una lástima la decisión del CORPES XXI de perder esta funcionalidad en el resultado de la exportación, que sí se emplea en la presentación de las ocurrencias en la interfaz web (como en el CDH).

## 6. EL ACCESO A LOS TEXTOS

Además de las herramientas empleadas por cada corpus para diseñar las búsquedas en su interior y ofrecer los resultados de estas, existen herramientas externas —ya sean programas de concordancias, como AntConc o WordSmith Tools (Anthony 2015), ya sean librerías dedicadas a la minería de textos, como *tidytext* en R o *spacy, sklearn* o *nltk* en Python— que permiten que el investigador con los conocimientos básicos del procesamiento de lenguaje natural (como las expresiones regulares) realice tareas semejantes si tiene acceso a los textos completos en el formato adecuado.

Los corpus del español gozan de una situación envidiable en lo que se refiere a dicho acceso a los textos. Así, todos ellos permiten la descarga de las transcripciones, con la salvedad de los corpus de referencia académicos (CORDE, CREA, CDH y CORPES XXI) y el Corpus del Español (en su versión histórica), que por las características de sus textos —generalmente protegidos por derechos de autor— no pueden ofrecer un acceso completo.

Las diferencias más fundamentales entre los distintos corpus se refiere a cómo se realiza dicha descarga y al formato en que se ofrecen los textos. En cuanto a la primera cuestión, el CORDIAM y el PRESEEA son los que presentan un acceso más restringido, pues se realiza a través de las búsquedas. Es decir, se ofrece acceso a los textos completos que contienen resultados de una búsqueda dada, pero no se da la posibilidad de descargar todos los textos del corpus (que podría realizarse, teóricamente, a través de una búsqueda que devuelva todos los documentos). Por su parte, Biblia Medieval, el CODEA+ 2015, el COSER y Val.Es.CO sí ofrecen un listado —en distintos formatos— de todos sus textos, que pueden descargarse, pero también de uno en uno, lo cual, según el tamaño del corpus, puede

suponer una importante inversión de tiempo para el usuario que quiere obtener todo el corpus. Post Scriptum permite descargar todos los textos de cada siglo de una única vez, pero no da la opción de descargarlos todos simultáneamente, independientemente de su fecha de producción. Solamente el ESLORA permite descargar todas las transcripciones juntas de una única vez.

En cuanto al formato en que se ofrecen los textos, esta es una cuestión crucial para su manejo con herramientas computacionales. Los archivos en formato PDF (que ofrecen el CORDIAM o el COSER) son los menos adecuados, pues no son legibles por estas. Los formatos en texto plano (.txt), que ofrecen Biblia Medieval, Post Scriptum, el PRESEEA y el ESLORA, son mucho más adecuados —y mejores que el formato DOC, por ser un sistema cerrado nativo de Word, ofrecido por Val.Es.Co, aunque la conversión de .doc a .txt es más sencilla y menos problemática que de .pdf a .txt—. Algunos corpus, como Post Scritpum y Val.Es.Co, ofrecen varios formatos: XML —ofrecido por ambos corpus, aunque Val.Es.Co lo ofrece incrustado en la web, no como exportación— también es un formato sencillo de leer por lenguajes de programación, mientras que los archivos de Excel (.xls) —también ofrecido por Val.Es.Co— pueden presentar más dificultades: el formato CSV es más adecuado, por ser un formato abierto.

Un paso más avanzado en la descarga de los textos supone poner a disposición de los usuarios también las transcripciones anotadas.[16] Esto solo lo permiten el ESLORA, que ofrece —previa solicitud— el acceso a las transcripciones lematizadas, y Post Scriptum, que, junto con la descarga de las transcripciones literales, presenta la posibilidad de descargar las transcripciones estandarizadas ortográficamente, las transcripciones anotadas morfológicamente y las transcripciones siguiendo los estándares TEI.[17]

Por otro lado, la cuestión de los metadatos mencionada en el apartado anterior también es de crucial importancia en la descarga del corpus entero. Lo habitual es que dicha información se ofrezca en las cabeceras de cada transcripción. Esta información puede recuperarse de forma automática, aunque el proceso no es necesariamente sencillo (especialmente si las cabeceras no siguen un formato estandarizado). Como ya sugeríamos en el apartado anterior, poner a disposición del usuario una tabla (o varias) que reúna los metadatos de cada documento y siga el formato de los datos ordenados facilita mucho el acceso a dicha información[18]: el único corpus que lo hace es Post Scriptum, con una tabla muy completa que, lamentablemente, no se puede descargar. Sí se puede copiar y, al pegarla, conserva el formato tabla. Sin embargo, se ofrecen 100 filas por página: puesto que la tabla contiene 5071 filas, debemos copiar manualmente los resultados de 51 páginas distintas.

Por último, es destacable el compromiso de muchos de nuestros corpus con ofrecer acceso a los textos originales, ya sea en formato imagen o en formato audio. Salvo los corpus académicos (incluido esta vez el CORDIAM) y el Corpus del español, todos los demás corpus analizados permiten la descarga de los originales (a veces previa solicitud, como en el caso de Val.Es.Co). De esta forma el investigador puede comprobar personalmente las lecturas de las transcripciones cuando así lo desee. En este sentido, debemos congratularnos de la

[16] Empleo un concepto más restringido que Reppen (2010), que considera las cabeceras como formas de anotación.

[17] El uso de dichos estándares es siempre recomendable, debido a que están bien documentados, lo que facilita el procesamiento de los textos. Se puede encontrar más información sobre la iniciativa en su página web: https://tei-c.org/

[18] Otra buena práctica es nombrar a cada documento con el identificador con que se referencia en dicha tabla.

iniciativa de la red CHARTA (seguida por Biblia Medieval y el CODEA+ 2015 de los corpus aquí analizados) de acompañar sus ediciones críticas de transcripciones paleográficas e imágenes del manuscrito, aumentando así la fiabilidad de los datos de que dispone la comunidad.

## 7. LA INFORMACIÓN SOBRE EL CORPUS

A lo largo del trabajo hemos insistido varias veces en la importancia de que el investigador tenga acceso a la mayor cantidad de información posible, puesto que dicha información es crucial para la correcta interpretación de los resultados, a lo largo de todo el proceso de análisis. Así, la información sobre el contenido y el diseño del corpus, al igual que las convenciones de transcripción, les resulta de interés a todos los investigadores, mientras que la información sobre los distintos mecanismos de procesamiento y anotación de las transcripciones empleados en la herramienta de búsqueda (tokenización incluida) les resulta de interés a todos los investigadores que utilicen la interfaz de búsqueda del corpus (que son, indudablemente, la mayoría).

En general, todos los corpus aquí analizados contienen información sobre el contenido del corpus, ofreciendo un listado de todos los textos y explicitando el diseño del corpus, es decir, las causas y/o métodos de la selección de los textos incluidos. Una llamativa excepción es el caso del PRESEEA, en cuya web no se ofrece información sobre la composición del corpus disponible online —por lo tanto, el investigador ni siquiera conoce el tamaño del corpus con el que trabaja—. Este corpus sí ofrece acceso a los documentos del proyecto que explican el diseño del corpus.[19]

En lo que se refiere a las normas de transcripción, la mayoría de los corpus analizados son muy transparentes en este sentido, informando al usuario de las convenciones. No lo hacen los corpus de referencia académicos (CORDE, CREA, CDH, CORPES XXI), que emplean los textos de las ediciones sin unificar (a diferencia del CORDIAM), lo que, a efectos prácticos, implica que el investigador no tiene acceso a estas. Val.Es.Co no ofrece tampoco esta información en su versión 2.0, aunque sí lo hace en la web original, todavía disponible (¡aunque no enlazada en la versión 2.0!).

Por lo que atañe a la información sobre los procesos de anotación morfológica, también resulta esencial que el investigador tenga acceso a información sobre ellos, para que el usuario pueda evaluar los problemas que se mencionaron en el apartado 2 y discernir así el método de búsqueda que le conviene. En los corpus analizados que están lematizados y etiquetados (Post Scriptum, CDH, Corpus del Español, COSER, ESLORA, CORPES XXI, Biblias Hispánicas, Val.Es.Co y CORDIAM, que está únicamente lematizado parcialmente por ahora) existe bastante disparidad a este respecto. Post Scriptum contiene toda la información sobre las etiquetas empleadas y el proceso de anotación, mientras que otros corpus (como ESLORA, COSER, Biblias Hispánicas o el Corpus del Español) solo dan información de las etiquetas empleadas, pero no de cómo se han asignado. En cuanto a los corpus de la Academia, mientras que el CORPES XXI tiene un breve documento en que se indica el porcentaje de acierto del modelo estadístico empleado para la anotación (sin más detalles), el CDH no explica nada al respecto (y las etiquetas se deducen de la interfaz de búsqueda).

[19] Por otra parte, es frecuente que los responsables de los corpus publiquen detalles sobre la composición y diseño de estos en los medios habituales de difusión lingüística, por lo que el usuario puede familiarizarse con estos detalles (Davies 2002, Moreno Fernández 2005, Bertolotti y Company 2014, de Benito, Pueyo y Fernández-Ordóñez 2016, Enrique-Arias 2016, Miguel y Sánchez-Prieto 2016, Rojo 2016, Vaamonde 2018).

Una solución intermedia es la de Val.Es.Co, en la que se indica que se ha empleado la herramienta Freeling, por lo que uno puede familiarizarse con esta si le interesan más detalles.

## 8. LA COLABORACIÓN DEL USUARIO

Tras seis apartados dedicados a las necesidades del usuario de corpus, creo justo dedicar este último apartado a las responsabilidades de este. La creación de un corpus supone un esfuerzo ímprobo, generalmente de años (incluso décadas) y los usuarios no podemos estar más que agradecidos a todos aquellos que han dedicado parte de su vida laboral a dicha tarea. En la época del consumo colaborativo, es razonable esperar que el usuario también se comprometa a colaborar en la mejora de los corpus.

En el largo y complejo camino que implica partir del texto original para ponerlo a disposición de la comunidad en formato digital se realizan varios procesos (transcripción, distintos niveles de anotación) cuya fiabilidad no es absoluta. Especialmente en los corpus más voluminosos, no es esperable que se alcance la infalibilidad, pues la revisión manual de las anotaciones automáticas, por ejemplo, resulta una tarea descomunal. Sin embargo, el usuario está en la posición adecuada para detectar algunos de estos errores, cuando, por ejemplo, los resultados de las búsquedas devuelven concordancias mal anotadas. En mi opinión, establecer un sistema que permita que los creadores de corpus puedan aprovechar ese conocimiento de los usuarios puede ayudar a la mejora de las herramientas disponibles.

Semejante sistema puede adoptar varias formas, con diversos grados de integración en la interfaz del corpus. Es imaginable una herramienta similar a la que en el COSER permite descartar resultados, pero más completa (indicando, por ejemplo, si están mal lematizados o anotados e, incluso, proponiendo un lema o etiqueta alternativos; lo mismo es aplicable para errores de transcripción). Una herramienta así debería, por supuesto, estar coordinada con la herramienta de exportación: de esta manera, el trabajo realizado por el usuario es aprovechable tanto para los creadores del corpus como para el propio investigador, que podría exportar únicamente los resultados que le interesan. La ventaja de este método es que permite fácilmente la identificación del lugar en el que se ha encontrado el error.

El mismo efecto se conseguiría si, por ejemplo, en la exportación de los resultados se añadiera una columna con un identificador de cada concordancia encontrada (que facilite que los usuarios reporten el lugar del error adecuadamente y que luego los creadores encuentren el error para subsanarlo) y una columna llamada *¿Resultado inesperado?* —o similar—, que recuerde al usuario su deber de colaboración con el corpus. Así, mientras el usuario anota manualmente sus datos, puede también ir reportando los errores que encuentre, que luego puede enviar a los creadores del corpus.

Algunos de los corpus analizados ya prevén la colaboración de los usuarios. Así, Biblia Medieval indica en las condiciones de uso el deber de los usuarios de informar de los errores que detecten para contribuir a la mejora del recurso. El COSER, por su parte, tiene un formulario rubricado “Colabore con el proyecto” en el que se anima al usuario a informar de los errores que se detecten, incluyendo campos específicos para el enclave afectado. Por último, el CORPES XXI tiene un buzón de sugerencias, que entendemos se puede utilizar a tal

efecto, aunque no se explicita, mientras que otros corpus (el PRESEEA, el ESLORA, el CORPES XXI o Val.Es.Co) ofrecen formularios o información de contacto.[20]

## 9. CONCLUSIÓN

En este breve trabajo he analizado las interfaces de los corpus del español desde la perspectiva del lingüista que emplea los corpus para sus investigaciones, poniendo el foco en la tarea de anotar los resultados manualmente, necesidad que afronta la mayoría de investigadores. Centrándome en la interfaz de búsqueda y en el acceso a los resultados, he comparado las interfaces de trece corpus del español, identificando las soluciones que, a mi parecer, resultan más convenientes para el trabajo con herramientas externas, por suponer un ahorro de tiempo al investigador, abogando en general por la mayor transparencia posible y la posibilidad de extraer la información de forma sencilla y en formatos abiertos. Puesto que la tecnología está en constante avance, los creadores de corpus se enfrentan al reto de mantenerse al día en estos progresos: la conversación con los usuarios puede ser de gran ayuda para identificar mejoras y soluciones.

## REFERENCIAS BIBLIOGRÁFICAS

ANTHONY, Laurence (2015): «A Critical Look at Software Tools in Corpus Linguistics», *Linguistic Research*, 30, 2, pp. 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

BARCALA, Mario, Eva DOMÍNGUEZ, Alba FERNÁNDEZ, Raquel RIVAS, María Paula SANTALLA, Victoria VÁZQUEZ y Rebeca VILLAPOL (2018): «El corpus ESLORA de español oral: diseño, desarrollo y explotación», *CHIMERA: Romance Corpora And Linguistic Studies,* 5, 2, pp. 217-237. doi:http://dx.doi.org/10.15366/chimera2018.5.2.003

DE BENITO MORENO, Carlota, F. Javier PUEYO MENA e Inés FERNÁNDEZ-ORDÓÑEZ (2016): «Creating and designing a corpus of rural Spanish», en *Procceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), pp. 78-83.

BERTOLOTTI, Virginia y Concepción COMPANY COMPANY (2014): «El corpus diacrónico y diatópico del español de América (CORDIAM). Propuesta de tipología textual», *Cuadernos del ALFAL*, 6, pp. 130-148.

BIBER, Douglas y Randi REPPEN (eds.) (2012): *Corpus Linguistics.* London: SAGE Publications.

CRAWFORD, William J. y Eniko CSOMAY (2016): *Doing Corpus Linguistics*. New York/London: Routledge.

DAVIES, Mark (2002): «Un corpus anotado de 100.000.000 palabras del español histórico y moderno», *Procesamiento del Lenguaje Natural*, 29, pp. 21-27.

DAVIES, Mark (2009): «Creating Useful Historical Corpora: a Comparison of CORDE, the Corpus del Español and the Corpus do Português», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus.* Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 137–166.

ENRIQUE-ARIAS, Andrés (2016). «Ventajas e inconvenientes del uso de Biblia medieval (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español» en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 269–284.

[20] Merece la pena también mencionar la existencia de proyectos que prevén corpus colaborativos en su creación y diseño, como el del Corpus judeoespañol digital (Stulic-Etchevers y Rouissi 2009) o el del corpus de comunicación digital CoDiCE (Vela Delfa y Cantamutto 2015).

KILGARRIFF, Adam y Iztok KOSEM (2013): «Corpus Tools for Lexicographers», *Electronic Lexicography*, pp. 1–37. https://doi.org/10.1093/acprof:oso/9780199654864.003.0003

KÜBLER, Sandra y Heike ZINSMEISTER (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Academic.

LE ROUX, Joseph, Benoît SAGOT y Djamé SEDDAH (2012): «Statistical Parsing of Spanish and Data Driven Lemmatization», en *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*. Jeju, pp. 56-61.

LÜDELING, Anke y Merja KYTÖ (2008): *Corpus linguistics: An international handbook*. Berlin: De Gruyter.

MAMBRINI, Francesco, Marco PASSAROTTI y Caroline SPORLEDER (eds.) (2011): «Annotation of Corpora for Research in the Humanities», *Journal for Language Technology and Computational Linguistics*, 26.

MIGUEL FRANCO, Ruth y Pedro SÁNCHEZ-PRIETO BORJA (2016): «CODEA: A "Primary" Corpus of Spanish Historical Documents», *Variants*, 12-13, pp. 211-230.

MORENO FERNÁNDEZ, Francisco (2005): «Corpus para el estudio del español en su variación geográfica y social. El corpus PRESEEA», *Oralia*, *8*, pp. 123–140.

O'KEEFFE, Anne y Michael MCCARTHY (2010): *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

PADRÓ, Lluís, Miquel COLLADO, Samuel REESE, Marina LLOBERES e Irene CASTELLÓN (2010): «FreeLing 2.1: Five years of open-source language processing tools», en *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp. 931–936.

REPPEN, Randi (2010): «Building a Corpus. What Are the Key Considerations?», en Anne O'Keeffe y Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, pp. 31–37.

RODRÍGUEZ MOLINA, Javier y Álvaro OCTAVIO DE TOLEDO Y HUERTA (2017): «La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística», *Scriptum Digital*, 6, 5-68.

ROJO, Guillermo (2016): «Citius, maius, melius: del CREA al CORPES XXI», en Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlin/Boston: De Gruyter, pp. 197–212.

SMITH, Nicholas, Sebastian HOFFMANN y Paul RAYSON (2008): «Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations», *Literary and Linguistic Computing*, 23, 2, pp. 163–180. https://doi.org/10.1093/llc/fqn004

STRAKA, Milan, Jan HAJIČ y Jana STRAKOVÁ (2016): «UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing», en *Proceedings of LREC 2016*, pp. 4290–4297.

STULIC-ETCHEVERS, Ana y Soufiane ROUISSI (2009): «Pensando un corpus en modo colaborativo: hacia el prototipo del corpus judeoespañol digital», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus.* Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 117–134.

TORRUELLA, Joan (2016): «Tres propuestas en el ámbito de la lingüística de corpus», en Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica.* Berlín/Boston: De Gruyter, pp. 90-112.

TORRUELLA, Joan y Joaquim LLISTERRI (1999): «Diseño de corpus textuales y orales», en José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.), *Filología e Informática. Nuevas Tecnologías En Los Estudios Lingüísticos*. Barcelona: Milenio, pp. 45–77.

VAAMONDE, Gael (2018): «Escritura epistolar, edición digital y anotación de corpus», *Cuadernos del Instituto Historia de la Lengua*, 11, pp. 139-164.

VELA DELFA, Cristina y Lucía CANTAMUTTO (2015): «Problemas de recogida y fijación de muestras del discurso digital», *CHIMERA. Romance Corpora and Linguistic Studies,* 2, pp. 131–155.

**CORPUS**

Biblia Medieval = ENRIQUE-ARIAS, Andrés y F. Javier PUEYO MENA (2008-): *Biblia Medieval*. <http://www.bibliamedieval.es>

Biblias Hispánicas = ENRIQUE-ARIAS, Andrés y F. Javier PUEYO MENA (2008-): *Biblia Medieval*. <http://bh.bibliamedieval.es/>

CDH = INSTITUTO DE INVESTIGACIÓN RAFAEL LAPESA DE LA REAL ACADEMIA ESPAÑOLA (2013): *Corpus del Nuevo diccionario histórico (CDH)* [en linea]. <http://web.frl.es/CNDHE>

CODEA+ 2015 = GITHE (Grupo de Investigación Textos para la Historia del Español): *CODEA+ 2015 (Corpus de documentos españoles anteriores a 1800)*. <www.corpuscodea.es>

CORDE = REAL ACADEMIA ESPAÑOLA: Banco de datos (CORDE) [en línea]: *Corpus diacrónico del español.* <http://www.rae.es>

CORDIAM = ACADEMIA MEXICANA DE LA LENGUA: *Corpus Diacrónico y Diatópico del Español de América*. <www.cordiam.org>

CORPES XXI = REAL ACADEMIA ESPAÑOLA: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI (CORPES).* <http://www.rae.es>

Corpus del Español = DAVIES, Mark: *El corpus del español*. <https://www.corpusdelespanol.org/>

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

CREA = REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es>

ESLORA = ESLORA: *Corpus para el estudio del español oral*. <http://eslora.usc.es>, versión 1.2.2 de noviembre de 2018, ISSN: 2444-1430.

Post Scriptum = CLUL (ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt.>

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

Val.Es.Co = CABEDO, Adrián y Salvador PONS (eds.): *Corpus Val.Es.Co 2.0.* <http://www.valesco.es>