Presentación de la base de datos LBC

Annick Farina (Università di Firenze), Riccardo Billero (Università di Firenze), Carlota Nicolás Martínez (Università di Firenze)[1]

La base de datos LBC es uno de los recursos de libre acceso que ofrece el grupo de investigación Unità di Ricerca *Lessico multilingue dei Beni Culturali*. El objetivo es poder realizar consultas en los corpus diseñados para estudios del léxico especializado en el ámbito de la traducción y de la lexicografía. De hecho, la Unità di Ricerca pretende ofrecer de un espacio digital con diversas herramientas útiles para difundir el conocimiento del patrimonio artístico y cultural toscano a nivel internacional (Farina 2016).

En la base de datos están estructurados los corpus de texto de las distintas lenguas publicadas (francés, inglés, italiano, ruso, español, alemán). Todo ello está alojado en la plataforma del proyecto que contiene varias herramientas entre los que se encuentran los corpus e información sobre ellos[2].

Los corpus se crean a partir de textos de diversos géneros como obras literarias clásicas, novelas de viaje o epistolarios, textos científicos y técnicos, guías turísticas, manuales, etc. Textos comprendidos en un largo período de tiempo. Las fuentes se estructuraron y gestionaron a través de programas informáticos con funciones adecuadas, que respondieran a las necesidades de múltiples usuarios. Concretamente, los principales destinatarios a los que se dirigen los corpus son: lingüistas, escritores, investigadores en ciencias humanas y sociales, cuya labor requiere de investigación

para obtener información sobre el léxico por autor, período cronológico, género, etc.; traductores que necesiten consultar recursos léxicos específicos; y finalmente especialistas en el sector turístico, o turistas interesados en profundizar en el conocimiento del territorio y la cultura a él vinculada.

Para cada lengua del proyecto, hay textos que coinciden en tema y género las demás lenguas del proyecto, estos textos en la lengua original han sido elegidos con dos criterios de prioridad: autoridad reconocida del texto y del autor en la cultura de pertenencia y amplia difusión (Billero, Nicolás 2017: 208); facilidad de conversión en un formato de edición, evitando textos difíciles de digitalizar en la primera fase. Para los textos traducidos, la elección se basa en una lista elaborada por el grupo que contiene los textos en italiano y otras lenguas considerados esenciales para el conocimiento internacional del patrimonio artístico-cultural toscano: los textos de referencia de la Historia del Arte para Toscana como son *Las vidas* de Vasari, los libros de arquitectura de Alberti, Palladio, Sellio, algunos escritos de Maquiavelo y Leonardo; libros de viajes famosos, como los viajes de Stendhal y Ruskin, y libros de arte como Burckhardt.

Sin embargo, en esta etapa, en los distintos corpus no se ha dado la misma prioridad y proporción a los distintos tipos de textos, por diversas razones: el criterio de accesibilidad a las fuentes es obviamente diferente según los países, como lo es el interés por la herencia cultural toscana, que varía según períodos históricos y géneros textuales en las distintas lenguas y culturas presentes en el proyecto.

Estas razones explican la heterogeneidad entre corpus, si bien en el futuro desarrollo del proyecto nos gustaría limitarla, con este objetivo, el análisis de la distribución de los tipos de textos elegidos en cada corpus, y de los siglos representados al final de esta primera fase de constitución de corpus, permitirá una homogeneización más amplia en el futuro. Esta futura homogeneidad permitirá la comparación de textos. En esta primera fase, la prioridad ha sido dada a la inserción de textos de referencia de la propia lengua lo que ha permitido obtener una base textual consistente y suficiente para las búsquedas en una sola lengua.

Respecto a las cuestiones técnicas, tras un cuidadoso análisis de los distintos programas que se pueden utilizar para la consulta de corpus, la elección recayó en NoSketchEngine (Billero, 2020), debido que cuenta con varias características interesantes para los

propósitos del proyecto, pues permite búsquedas de concordancias y uso de filtros basados en diversas características.

A la información sobre la naturaleza de los contenidos de cada corpus se puede acceder en *Corpus info* en el menú NoSketchEngine (figura 1).

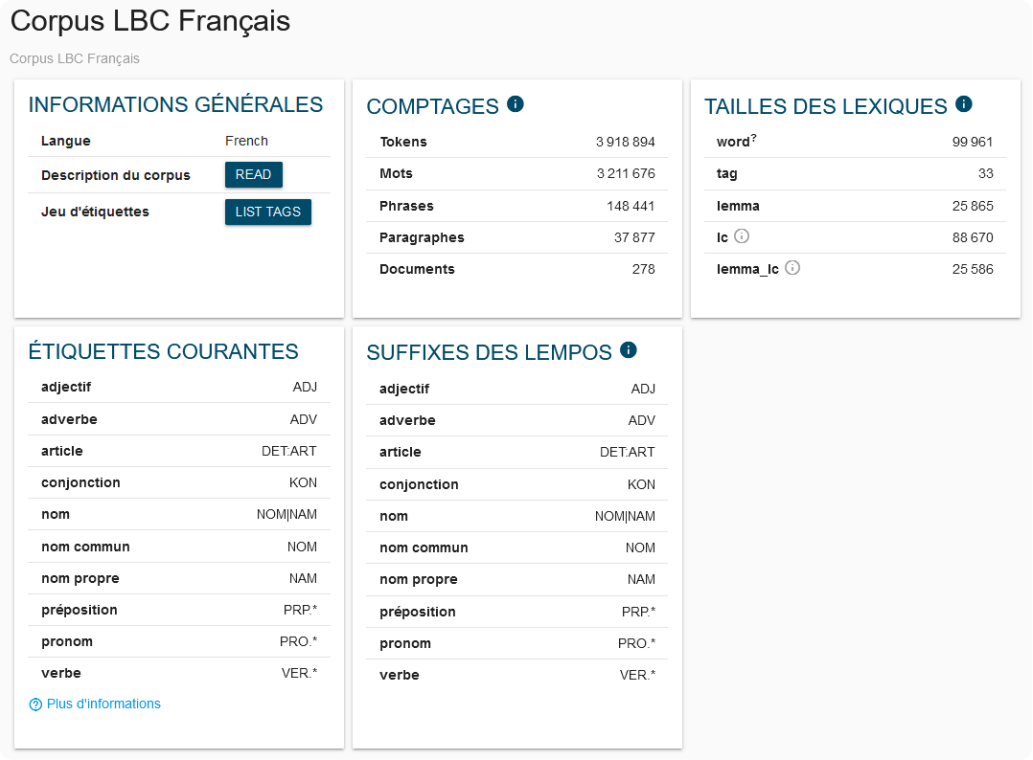


Fig. 1 – Información detallada sobre el corpus francés disponible en *Corpus info* [nov. 2020].

Esta página también contiene información sobre los valores cuantitativos atribuidos a lo0 documentos en cada una de las categorías utilizadas, como se ve en el corpus inglés en la Figura 2:

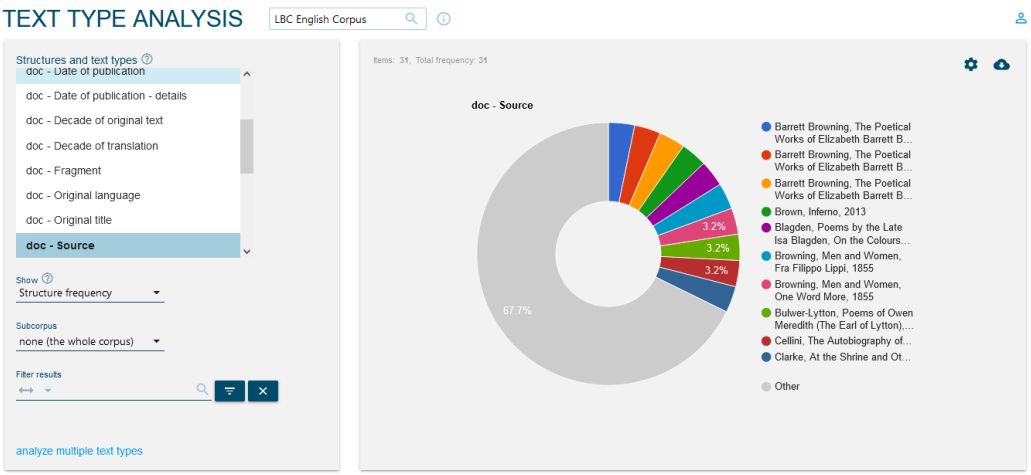


Fig. 2 - Estructura y características de los documentos incluidos en el corpus en inglés [nov. 2020].

La estructura de los corpus sigue las reglas tradicionales respetando los criterios de gestión de metadatos que se puede observar en la búsqueda con *Search* en la opción que ofrece los tipos de texto (*Text Types*, Figura 3).

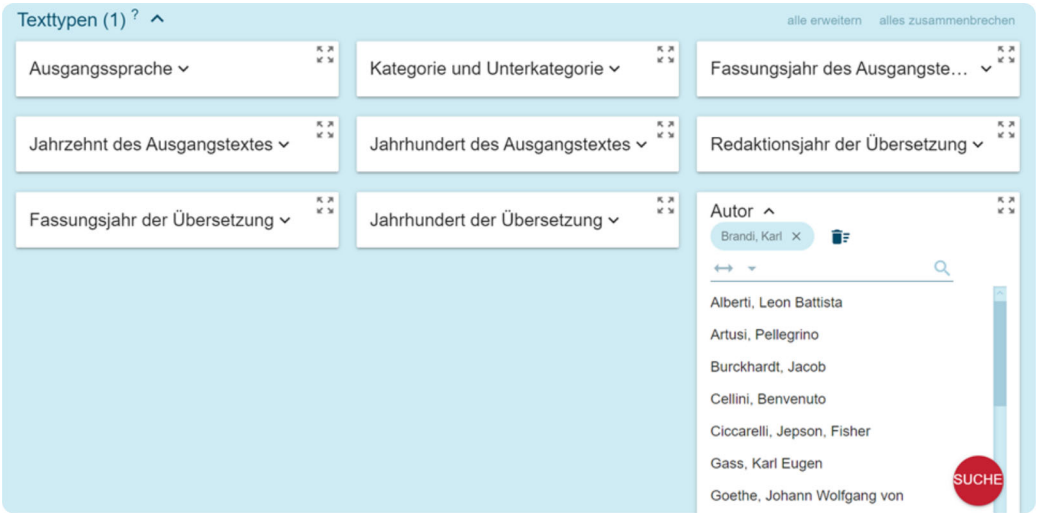


Figura 3 - Búsqueda en el corpus alemán utilizando la opción *Tipos de texto*.

Los metadatos con los que se puede filtrar la búsqueda de concordancias son:

- Lengua original: donde se presentan las opciones tanto de la lengua del texto como la lengua de origen como de los textos que se traducen;
- Lengua de traducción: permite buscar todas las traducciones en la lengua del corpus;

- Categoría y subcategoría: indica los distintos tipos de textos. Todos los textos tienen como tema el patrimonio artístico y su léxico, con una visión amplia de Florencia y Toscana descrita desde diferentes puntos de vista.
  Se han distinguido cuatro macrocategorías (Divulgación, Técnica, Diccionario y Literaria) y sus subcategorías (Divulgación: Blog, Guía, Revista; Técnica: Arquitectura, Arte, Gastronomía y Vino; Literario: Biografía, Ficción, No ficción; Diccionario: Monolingüe, Bilingüe/plurilingüe). Para la identificación de estas categorías se tuvo en cuenta el destino principal de la obra y el tipo de lector al que se dirige, datos que inciden en el tipo de lenguaje utilizado y en su nivel de especialización;
- Autor: se indican apellido y nombre, y la indicación *sa* (sin autor) cuando no existe;
- Título y fragmento: se eligió la introducción tanto de textos completos como de fragmentos que corresponden a una unidad textual, entendiendo como tal todo texto que tiene título, como son el capítulo de un libro, una carta completa, un artículo de revista, etc. Esta elección se ha hecho porque en muchos casos el libro completo no coincidía con los intereses del proyecto, y también para facilitar la futura creación de versiones paralelas de los textos traducidos. Se han incluido tanto títulos originales como títulos traducidos para los textos traducidos;
- Año de redacción/año de publicación/año de traducción: la información cronológica diferencia entre la fecha de redacción de los textos (cuando es posible) y la fecha de edición; para los textos traducidos se ha introducido la misma información tanto en el texto fuente como en el texto traducido. Para publicaciones en la red, se indica la fecha de consulta;
- Fuente: permite buscar en un solo documento del corpus (libro o fragmento);
- Delimitación geográfica: para los textos que tienen como objeto una ciudad o región definida, se ha introducido el nombre de la ciudad o región. Esta indicación está presente principalmente en libros de viajes y los epistolarios.

Esta información es accesible, con detalles bibliográficos más completos, desde el acceso a las concordancias haciendo clic en la referencia (nombre de archivo, número de documento, nombre del autor, etc. según las opciones elegidas en *View options*, fig. 4

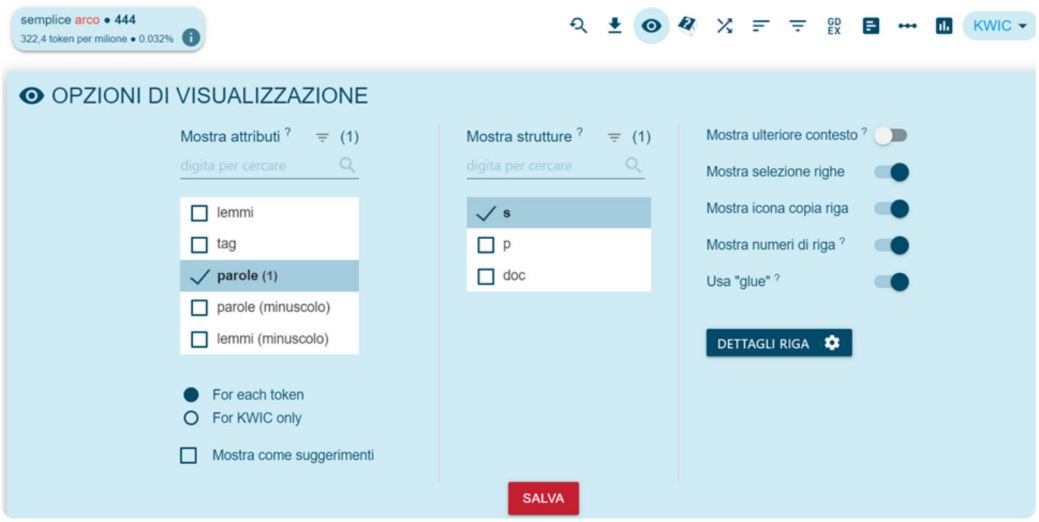


Figura 4 - Opciones disponibles para ver la referencia textual en *View options*.

Mediante la opción *Search*, se puede acceder a las concordancias cuyo pivote se muestra en orden aleatorio (pues depende del número de documento del que procede) como se ve en la figura 5 o en orden alfabético con respecto al pivote o nudo con su contexto derecho o izquierdo, utilizando la opción *Sort: Right / Left* (Figura 6).

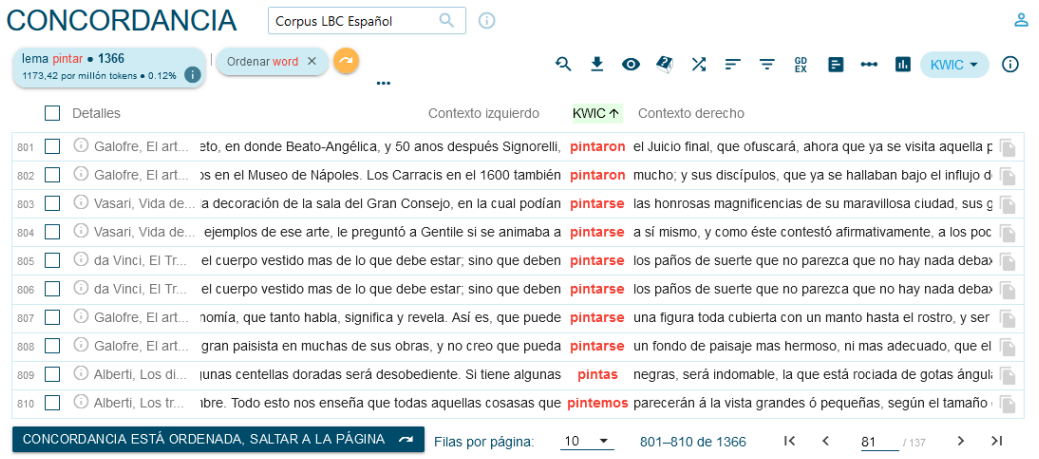


Figura 5 - Búsqueda de concordancias sobre el lema *pintar* en el corpus español sin elección de orden del pivote o nudo.

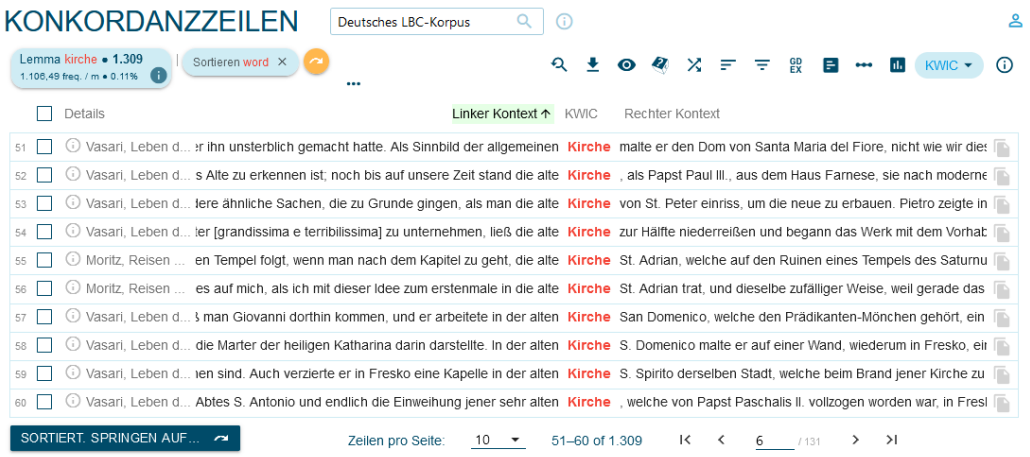


Figura 6 - Búsqueda de concordancias sobre el lema *Kirche* en el corpus alemán con el orden a la izquierda del lema seleccionado.

También es posible buscar la presencia de dos palabras o lemas en el mismo contexto a una distancia seleccionada de *tokens*utilizando la opción *Context* en el menú de búsqueda *Search*, como se muestra en la figura 7, que permiten por ejemplo verificar los usos presentes en varios contextos (*pintura al fresco* / *al fresco* en italiano en la figura 8).

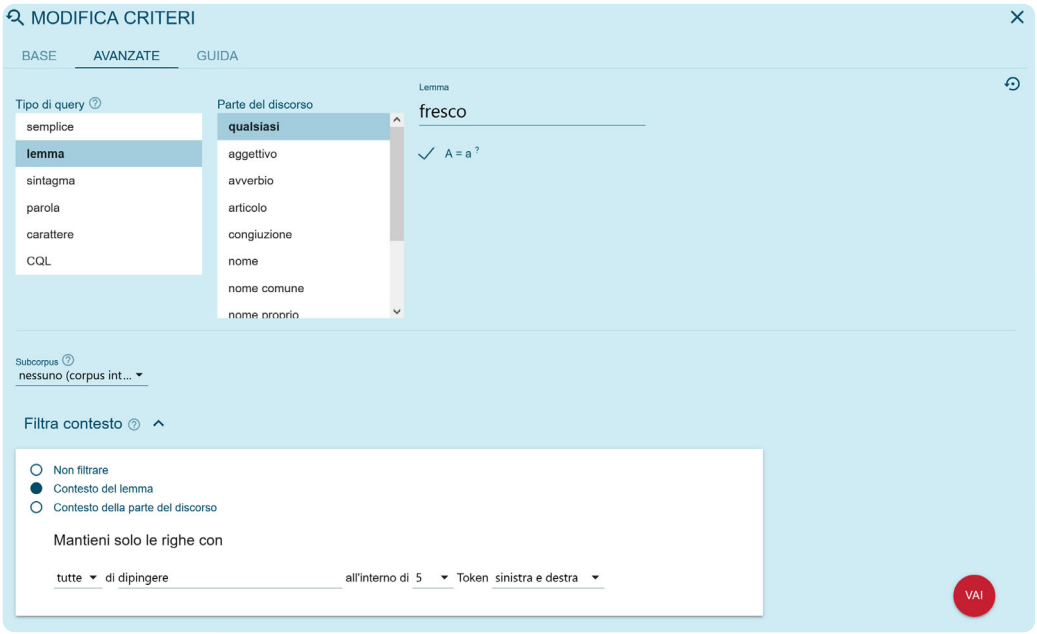


Figura 7 - Búsqueda en el corpus italiano de las palabras para *pintar* y *fresco* a 5 *tokens* de distancia.

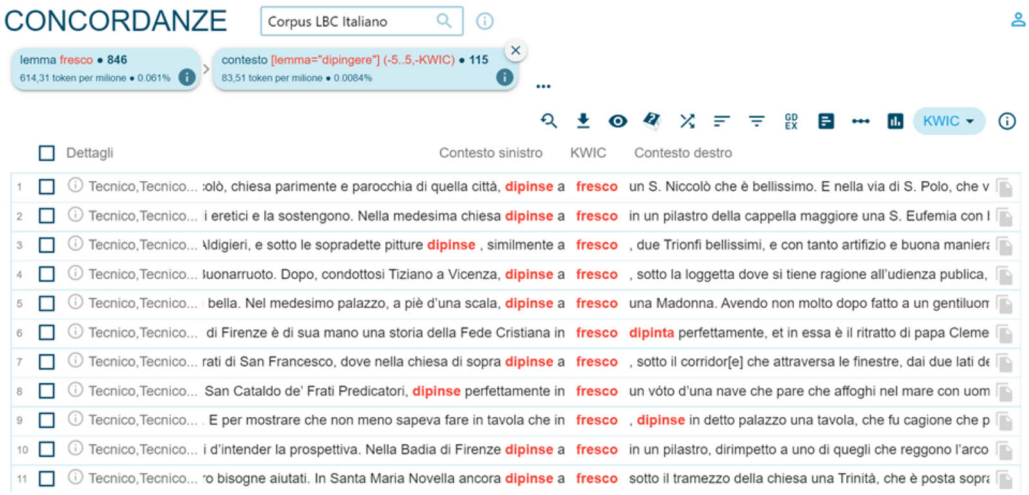


Figura 8 - Concordancias en el corpus italiano fruto de la búsqueda de *pintura* y *fresco* en el mismo contexto.

La selección de *Word list* permite obtener resultados numéricos, entre otros vemos por ejemplo el número de formas presentes en un corpus en cada una de sus fuentes (fig. 9); O también el número de lemas de un corpus (fig. 10-11).

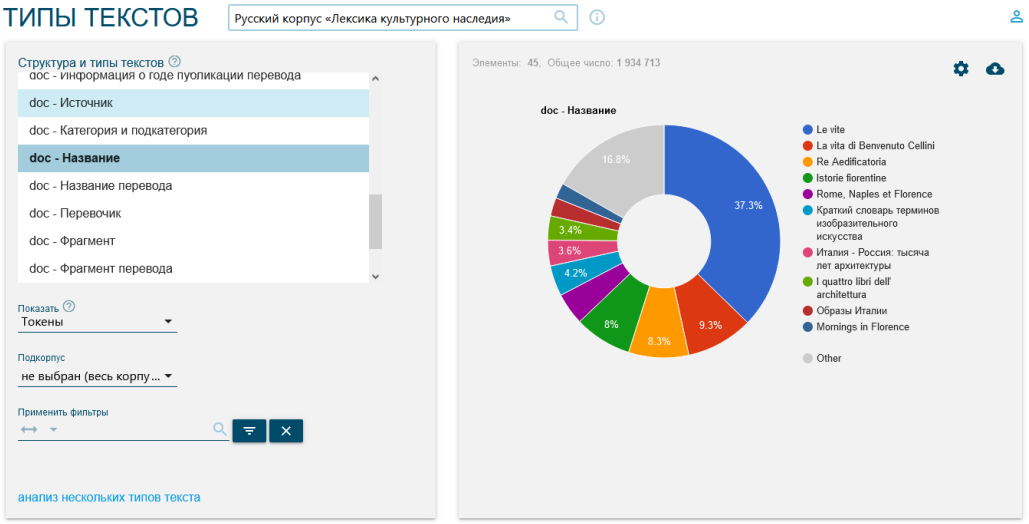


Figura 9 - Frecuencias en el corpus ruso de *tokens* presentes en cada uno de los autores.

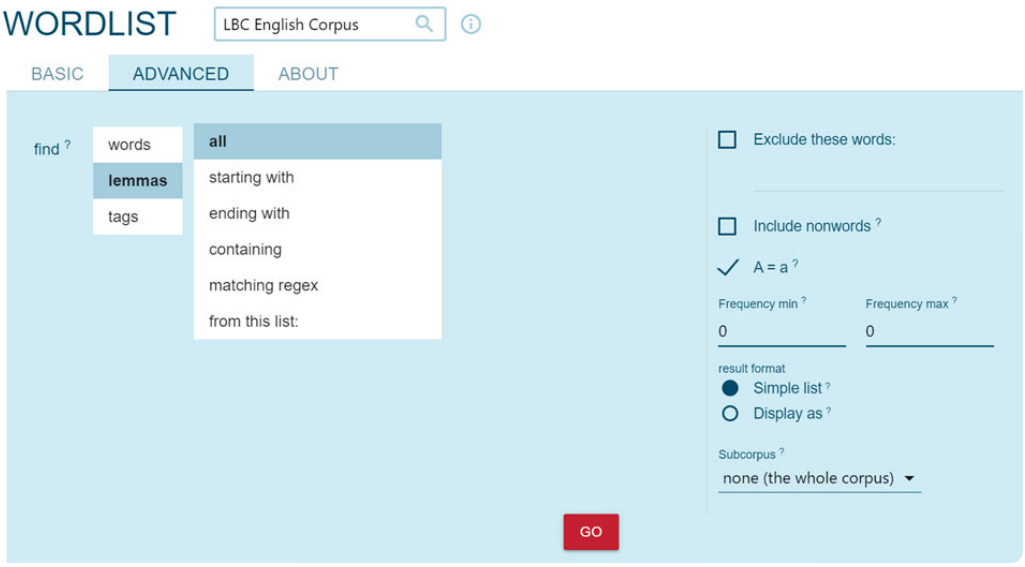
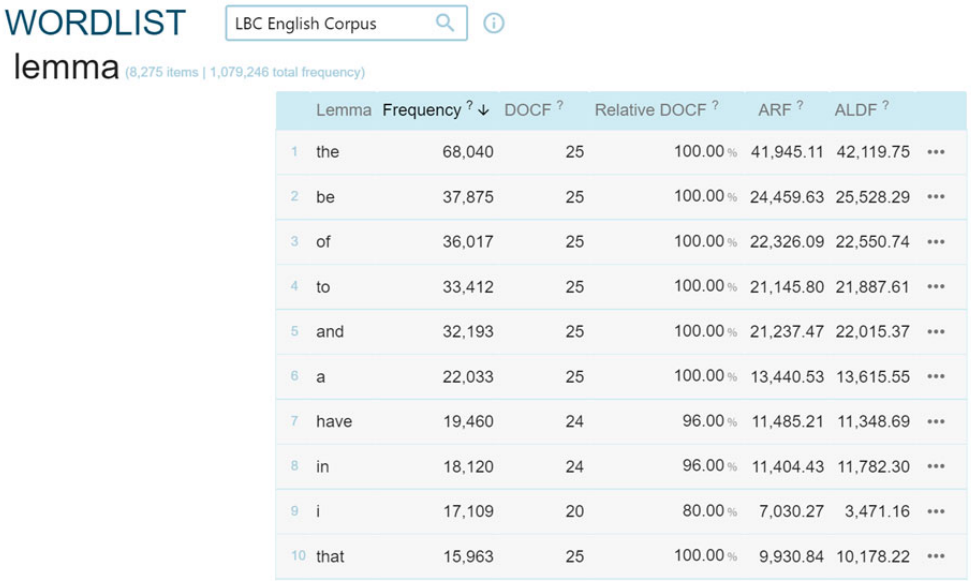


Figura 10 - Búsqueda en la lista de palabras de los lemas presentes en el corpus en inglés.



	Lemma	Frequency	DOCF	Relative DOCF	ARF	ALDF
1	the	68,040	25	100.00 %	41,945.11	42,119.75
2	be	37,875	25	100.00 %	24,459.63	25,528.29
3	of	36,017	25	100.00 %	22,326.09	22,550.74
4	to	33,412	25	100.00 %	21,145.80	21,887.61
5	and	32,193	25	100.00 %	21,237.47	22,015.37
6	a	22,033	25	100.00 %	13,440.53	13,615.55
7	have	19,460	24	96.00 %	11,485.21	11,348.69
8	in	18,120	24	96.00 %	11,404.43	11,782.30
9	i	17,109	20	80.00 %	7,030.27	3,471.16
10	that	15,963	25	100.00 %	9,930.84	10,178.22

Figura 11 - Resultado de la búsqueda de la lista de lemas presentes en el corpus en inglés [nov. 2020].

La realización de esta primera fase de nuestros corpus debe considerarse satisfactoria ya que ha creado las bases necesarias para los primeros trabajos y para los primeros trabajos de investigación de nuestro grupo (Carpi 2017; Farina, Billero 2018; Billero, Carpi 2018; Garzaniti 2020; Farina, Flinz 2020). Ya se han creado los primeros léxicos para cada lengua, acompañados de concordancias extraídas de los corpus que se publicarán en la plataforma en 2021 y se podrán utilizar para el desarrollo de futuros

diccionarios.

El principal objetivo de este primer trabajo, realizado por cada grupo lingüístico, ha sido realizar una validación de los corpus conscientes de que solo su uso real habría permitido identificar los problemas que de otro modo habrían permanecido ocultos.

En el futuro se prevé ampliar tanto el número de lenguas (actualmente los corpus de chino, portugués y turco, lenguas pertenecientes al proyecto LBC) como el de los textos con la idea de homogeneización ya descritos, para intentar hacer que los corpus sean comparables.

Bibliografía

Billero R. (2020), Cultural Heritage Lexicon: A Case Study. In Ana Pano Alamán, Valeria Zotti, *The language of art and culture heritage: a plurilingual and digital perspective*, Cambridge Scholars Publishing, pp. 86-103.

Billero R., Carpi E. (2018), Corpora e terminologia artistica: il caso del corpus spagnolo LBC. In *CHIMERA Romance Corpora and Linguistic Studies*, Madrid, UAM, 5, no. 1, pp. 85-91.

Billero R., Nicolás Martínez M.C. (2017), Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In *CHIMERA Romance Corpora and Linguistic Studies*, Madrid, UAM, 4.2, pp. 203-216.

Carpi E. (2017), El lenguaje para fines artísticos: traducciones de tondo al español. In Alejandro Curado (ed.), *LSP in Multi-disciplinary contexts of Teaching and Research. Papers from the 16th International AELFE Conference*, vol. 3, pp. 79–84. https://doi.org/10.29007/wx3m

Farina A., Nicolás Martínez C., Billero R. (eds.) (2020), *I Corpora LBC*, Firenze University Press, Firenze.

Farina A., Flinz C. (2020), Analisi comparativa dei corpora LBC. La visione del patrimonio fiorentino francese e tedesco: l'esempio del Duomo. In Fernando Funari, Annick Farina (eds.), *Le présent dans le passé / Past in Present/ Il passato nel presente*, Firenze University Press, Firenze.

Farina A., Billero R. (2018), Comparaison de corpus de langue «

naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues, *JADT'18 Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, UniversItalia, pp. 108-116.

Farina A. (2016), Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique. In *Publif@rum*, n. 24, 2016. http://www.farum.it/publifarum/ezine_articles.php?art_id=335

Garzaniti M. (2020), Il termine russo *friag* e le sue radici nelle relazioni culturali e artistiche fra la Russia e l'Italia. In Ana Pano Alamán, Valeria Zotti, *The language of art and culture heritage: a plurilingual and digital perspective*, Cambridge Scholars Publishing. pp 104-119.

Notas

[1] Este texto es una traducción hecha por [Carlota Nicolás Martínez] de la introducción italiana a los corpus LBC publicada en http://corpora.lessicobeniculturali.net/it/

[2] Para obtener datos exhaustivos sobre los corpus LBC, ver la publicación del grupo (Farina, Nicolás Martínez, Billero 2020).

Carpi, Elena; Pano Alamán, Ana. Corpus LBC Español

e-ISBN: 978-88-5518-035-1 | DOI: 10.36253/978-88-5518-035-1