Presentazione della banca dati LBC

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

La Banca Dati LBC fa parte degli strumenti di supporto in Open Access sviluppati dall'Unità di Ricerca *Lessico multilingue dei Beni Culturali* al fine di rendere fruibile la consultazione di corpora che forniscono informazioni lessicali specifiche necessarie nello svolgimento di ricerche lessicografiche e di traduzione. L'Unità di Ricerca intende infatti disporre di uno spazio digitale con vari strumenti utili a diffondere a livello internazionale la conoscenza del patrimonio artistico e culturale toscano (Farina 2016).

La Banca Dati permette di effettuare ricerche all'interno dei corpora testuali delle diverse lingue pubblicate (francese, inglese, italiano, russo, spagnolo, tedesco) tramite la piattaforma del progetto che contiene vari strumenti fra cui i corpora e informazioni su di essi[1].

I corpora sono stati originati da testi di vari generi quali opere letterarie classiche, romanzi di viaggio o corrispondenze, testi scientifici e tecnici, guide turistiche, manuali, ecc. scritti in un arco temporale ampio, e le fonti sono state strutturate e gestite tramite un software con funzioni adeguate, rispondendo alle necessità di un'utenza multipla. In particolare, i principali destinatari cui i corpora vengono rivolti sono: linguisti, letterati, ricercatori in scienze umane e sociali, il cui lavoro necessita di ricerche per ottenere informazioni sul lessico per autore, periodo cronologico, genere ecc.; traduttori che hanno necessità di consultare risorse lessicali specifiche; e infine

specialisti del settore turistico, o turisti interessati ad approfondire la propria conoscenza del territorio e della cultura legata ad esso.

Per ogni lingua del progetto, sono presenti i testi corrispondenti per tematica e genere a quelli dell'intero progetto, scelti con due criteri di priorità per i testi in lingua originale: autorità riconosciuta del testo/autore nella cultura di appartenenza e diffusione (Billero, Nicolás 2017: 208); facilità di conversione in formato editabile, evitando testi di difficile digitalizzazione nella prima fase. Per i testi in traduzione, la scelta si basa su un elenco redatto dal gruppo contenente i testi in italiano e in altre lingue ritenuti primari per la conoscenza a livello internazionale del patrimonio artistico-culturale toscano: i testi di base di Storia dell'Arte riferiti alla Toscana quali *Le Vite* del Vasari, i libri di architettura di Alberti, Palladio, Sellio, alcuni scritti di Machiavelli e di Leonardo; i libri di viaggio di nota fama, come i viaggi di Stendhal e Ruskin, e libri d'arte come il Burckhardt.

Tuttavia, in questa fase, nei vari corpora non è stata data la stessa priorità e proporzione alle varie tipologie di testi, a causa di vari motivi: il criterio dell'accessibilità alle fonti è ovviamente diverso secondo i paesi e anche l'interesse per il patrimonio toscano, che varia secondo i periodi storici e i generi testuali nelle varie lingue/culture rappresentate nel progetto.

Da queste osservazioni deriva una eterogeneità fra corpora che vorremmo limitare negli sviluppi futuri del progetto. Infatti, l'analisi della distribuzione delle tipologie di testi scelti in ogni corpus e dei secoli rappresentati alla fine di questa prima fase di costituzione dei corpora potrà permettere una più ampia omogeneizzazione in futuro, consentendo lavori di comparazione dei testi. Nella prima fase la priorità data all'inserimento di testi di riferimento della propria lingua ha permesso di ottenere una base di testi consistente e sufficiente per ricerche in un'unica lingua.

Dopo una attenta analisi dei vari software utilizzabili per la consultazione dei corpora, la scelta è ricaduta su NoSketchEngine (Billero, 2020), per la presenza di diverse funzionalità interessanti per gli scopi del progetto, permettendo ricerche di concordanze e filtri in base a varie caratteristiche.

Si può attingere alle informazioni sulla natura dei contenuti di ogni corpus accedendo alla parte "Corpus info" disponibile nel menu di NoSketchEngine (Figura 1).

Figura 1 – Informazione dettagliata sul corpus francese disponibile su "Corpus info" [nov. 2022].

In questa pagina sono disponibili anche informazioni relative ai diversi valori presenti per i vari documenti in ognuna delle categorie individuate, come illustrato nella Figura 2 per il corpus inglese:

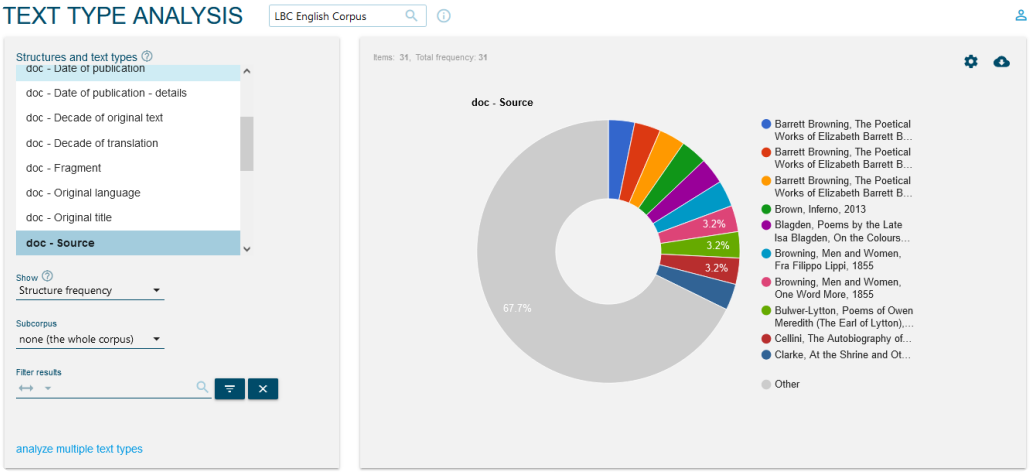

Figura 2 - Struttura e caratteristiche dei documenti inseriti nel corpus inglese [nov. 2022].

La struttura dei corpora segue le regole tradizionali rispettando criteri condivisi di gestione dei metadati che si rispecchia nella ricerca

via “Search” sui tipi testuali (“Text types”[2], Figura 3).

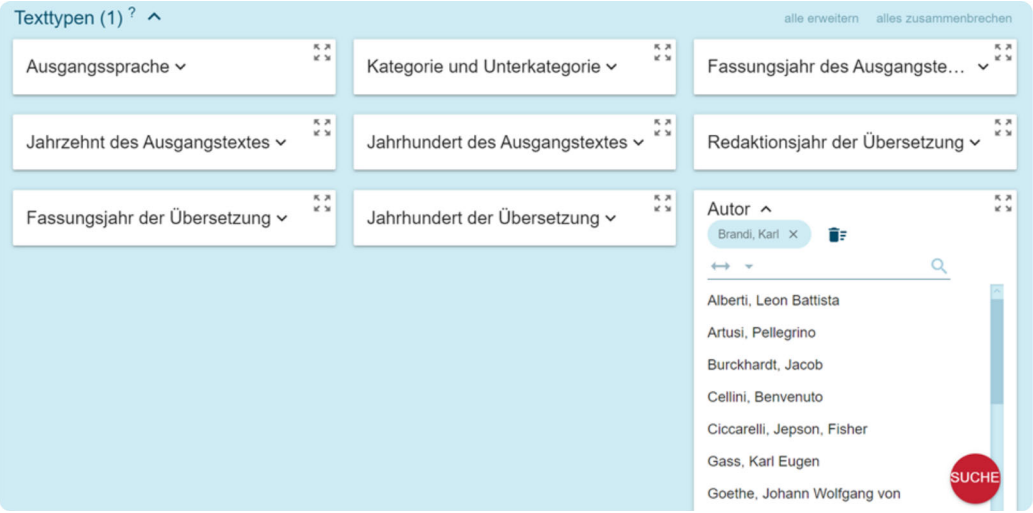

Figura 3- Ricerca nel corpus tedesco tramite la finestra “Text types”.

I metadati con cui si può filtrare la ricerca di concordanze sono:

- Lingua originale: appare sia la lingua del testo sia la lingua di origine per i testi in traduzione;
- Lingua di traduzione: permette una ricerca su tutte le traduzioni nella lingua del corpus;
- Categoria e sottocategoria: indica le varie tipologie di testi. Tutti i testi hanno come argomento il patrimonio artistico e il suo lessico, in particolare un’ampia visione di Firenze e della Toscana descritta da diversi punti di vista.

Sono state distinte quattro macrocategorie (Divulgativo, Tecnico, Dizionario e Letterario) e le loro relative sottocategorie (Divulgativo: Blog, Guida, Rivista; Tecnico: Architettura, Arte, Enogastronomia; Letterario: Biografico, Fiction, Saggistica; Dizionario: Monolingue, Bilingue/plurilingue). Si è tenuto conto per individuare queste categorie della destinazione principale dell’opera e del tipo di lettore a cui è rivolta, dati che condizionano il tipo di lingua usata e il suo livello di specializzazione[3]:

- Autore: sono indicati cognome e nome e l’indicazione “sa” (senza autore) quando inesistente;
- Titolo e frammento: si è scelto l’introduzione sia di testi interi sia di frammenti che corrispondono ad un’unità testuale perché provvisti di titoli, quali capitolo di libro, lettera completa, articolo di rivista ecc. Tale scelta è stata effettuata poiché in molti casi

l'intero libro non coincideva con gli interessi del progetto ma anche per facilitare la futura realizzazione di versioni in parallelo di testi tradotti. Per i testi tradotti sono stati inseriti sia titoli originali sia titoli tradotti;

- Anno di redazione / anno di pubblicazione / anno di traduzione: l'informazione cronologica fa una differenziazione tra data di redazione dei testi (laddove possibile) e data di edizione; per i testi tradotti sono state inserite le stesse informazioni sia sul testo di origine che sul testo tradotto[4]. Per le pubblicazioni online viene indicata la data di consultazione;
- Fonte: permette di fare una ricerca su un unico documento del corpus (libro o frammento);
- Area geografica[5]: per testi che hanno come oggetto una città o regione definita si è inserito il nome della città o regione. Questa indicazione è presente principalmente per i libri di viaggio e per le corrispondenze.

A queste informazioni si aggiungono dettagli bibliografici più completi quando si accede alle concordanze, cliccando sulla referenza (nome file, numero documento, nome autore, ecc. secondo le opzioni scelte in "View options", Figura 4)

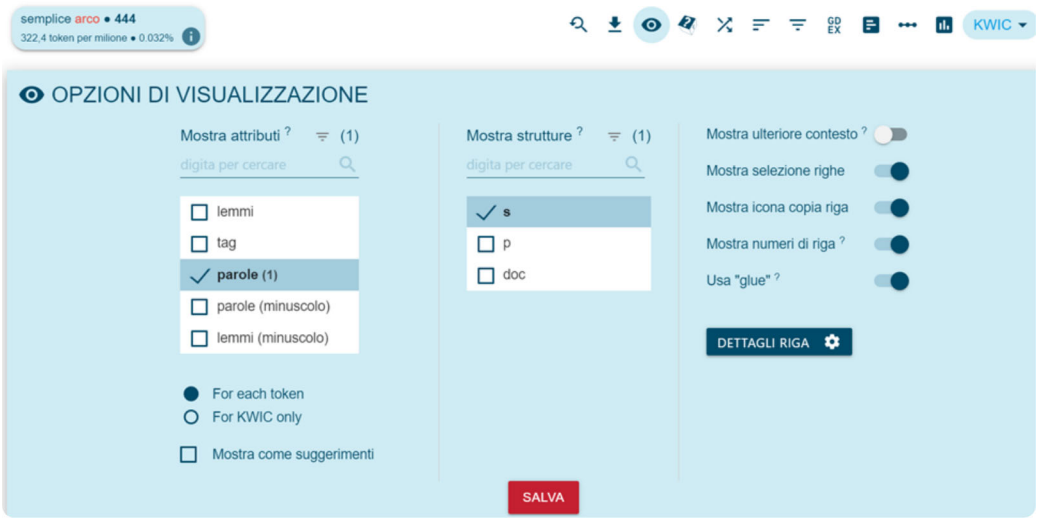

Figura 4 - Scelte disponibili per la visualizzazione del riferimento testuale in "View options".

Tramite l'opzione "Search", si accede alle concordanze visualizzate in ordine aleatorio (sul numero dei documenti) come nella Figura 5 oppure in ordine alfabetico rispetto alla parola considerata o al suo contesto destro o sinistro, mediante l'opzione "Sort left/right" (Figura 6).

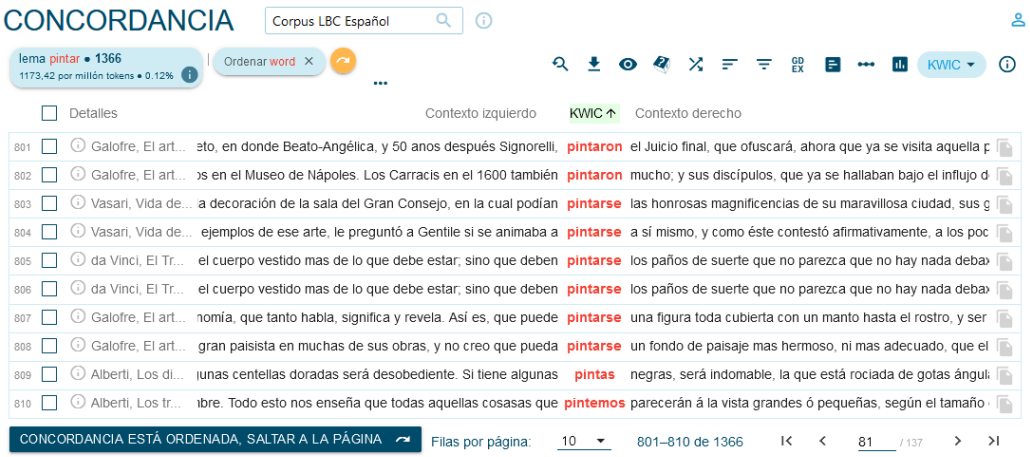

Figura 5 - Ricerca concordanze sul lemma *pintar* nel corpus spagnolo senza scelta di ordine.

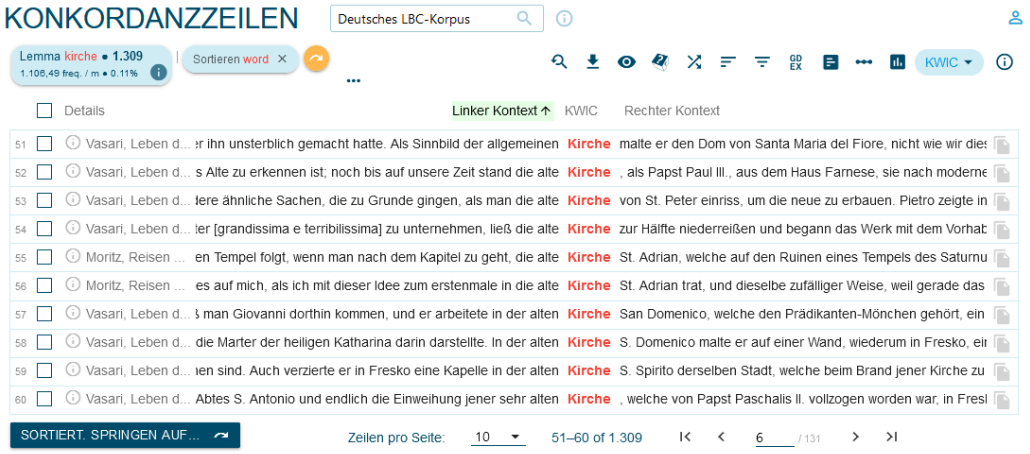

Figura 6 - Ricerca concordanze sul lemma *Kirche* nel corpus tedesco con ordine a sinistra del lemma.

È inoltre possibile effettuare una ricerca sulla presenza di due parole o lemmi nello stesso contesto a una distanza prescelta di *tokens* usando l'opzione "Context" nel menu di ricerca "Search", come da Figura 7, permettendo ad esempio di verificare gli usi attestati di varie collocazioni (*dipingere a fresco / in fresco* in italiano nella Figura 8).

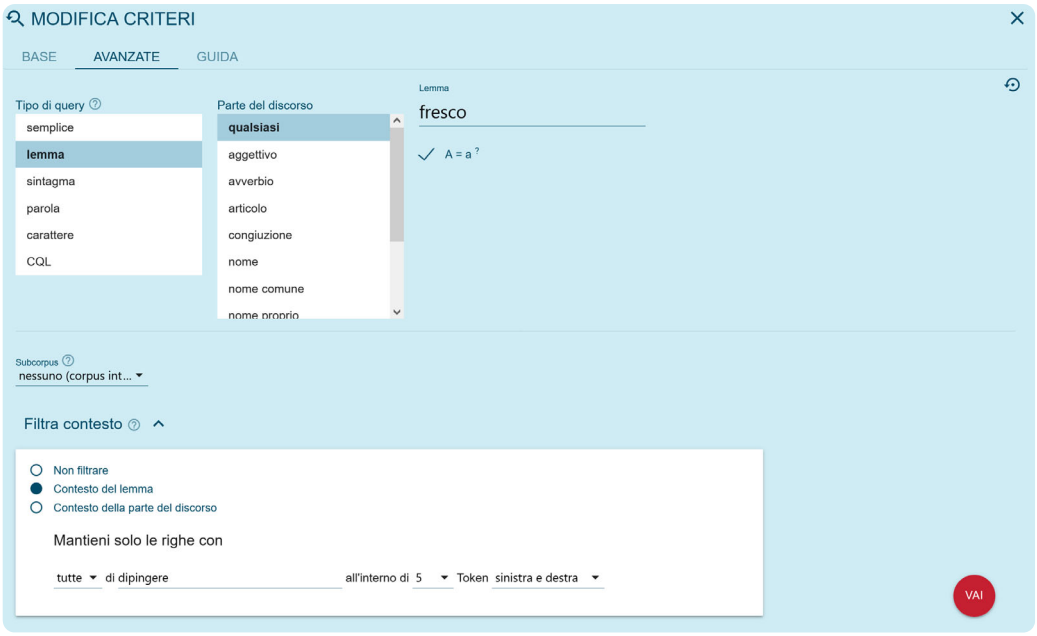

Figura 7 - Ricerca dei lemmi *dipingere* e *fresco* a 5 *token* di distanza nel corpus italiano.

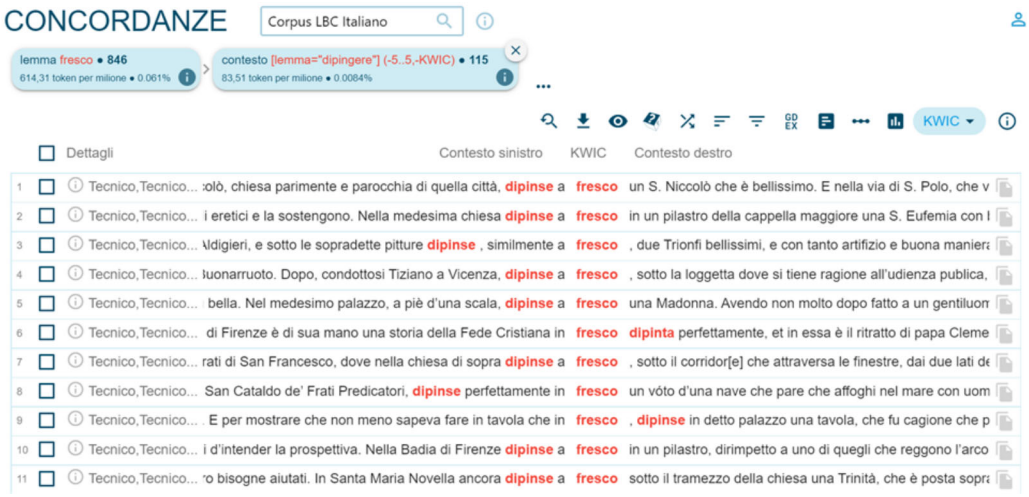

Figura 8 – Concordanze relative alla ricerca di *dipingere* e *fresco* nello stesso contesto nel corpus italiano.

L'opzione "Word list" permette di ottenere risultati numerici sulle frequenze presenti in un corpus sia sulle fonti, cercando ad esempio le frequenze di lemmi attribuibili ad ogni autore (Figura 9), sia sui lemmi di un corpus (Figura 10 e Figura 11).

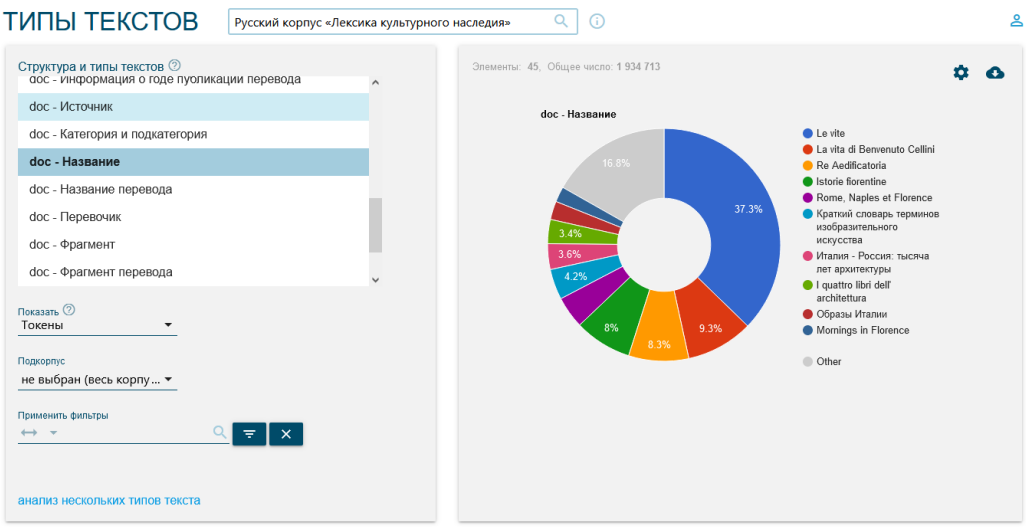

Figura 9 - Frequenze dei *token* presenti per autori nel corpus russo.

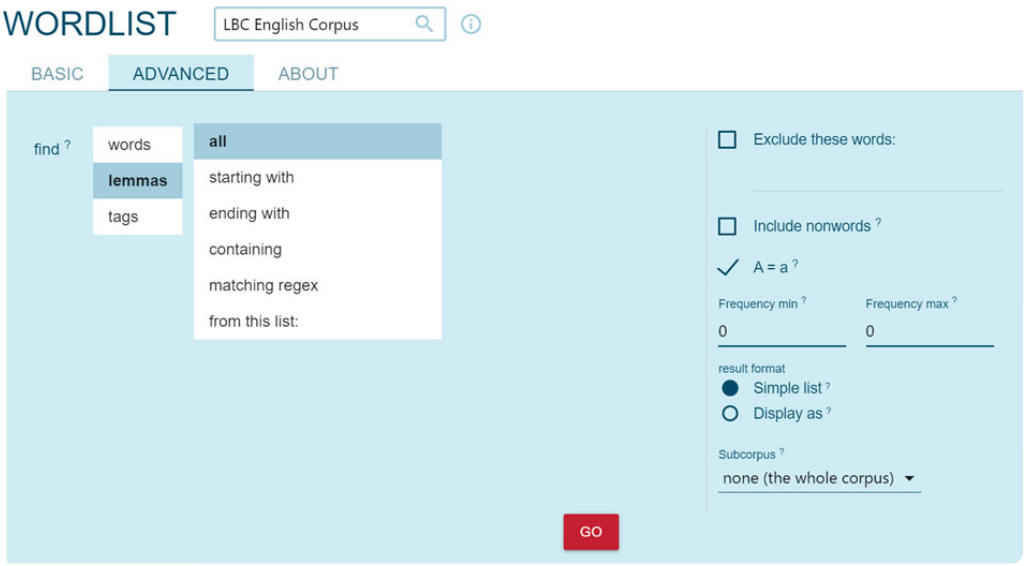

Figura 10 - Ricerca su Word list dei lemmi presenti nel corpus inglese

WORDLIST LBC English Corpus

lemma (8,275 items | 1,079,246 total frequency)

	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 - Risultato della ricerca su Word list sui lemmi presenti nel corpus inglese [nov. 2022].

La realizzazione di questa prima fase dei nostri corpora ha raggiunto gli obiettivi che ci eravamo proposti creando le basi necessarie per i primi lavori e per le ricerche del nostro gruppo (Carpi 2017; Farina, Billero 2018; Billero, Carpi 2018; Garzaniti 2020; Farina, Flinz 2020). Sono stati già realizzati i primi lemmari di ogni lingua corredati di concordanze estratte dai corpora che verranno pubblicate sulla piattaforma entro il 2022 e potranno essere usati per l'elaborazione di futuri dizionari.

Il principale obiettivo di questo primo lavoro, realizzato da ogni gruppo linguistico, era quello di effettuare una validazione dei corpora nella consapevolezza che soltanto il loro effettivo utilizzo avrebbe premesso di individuare problematiche che altrimenti sarebbero rimaste latenti.

Nel futuro si pensa di ampliare sia il numero di lingue (attualmente sono ancora assenti i corpora di cinese, portoghese e turco, lingue facenti parte del progetto LBC) sia quello dei testi con l'idea di omogeneizzazione già descritta, per cercare di rendere i corpora quanto più possibile comparabili fra loro.

Bibliografia

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.

Note

[1] Per dati esaustivi sui corpora LBC, si rimanda alla pubblicazione del gruppo (Farina, Nicolás Martínez, Billero 2020).

[2] La denominazione di ogni tipologia testuale presenti sull'opzione "Text Type" è per ora in lingua italiana ma verrà modificata a breve.

[3] Nella successiva fase di progetto la classificazione sarà rivista sulla base dei problemi riscontrati da alcuni gruppi con testi che potevano essere considerati come appartenenti a più categorie, come ad es. i testi di autori classici il cui stile è chiaramente letterario ma che scrivono testi che possono essere considerati specialistici per le tematiche e alcuni vocaboli (ad esempio *l'Histoire de la Peinture en Italie* di Stendhal classificato per ora nella categoria letterario/ saggistica).

[4] I testi contenuti vanno dal Rinascimento ai giorni nostri. Sebbene siano presenti entrambe le datazioni, l'anno di pubblicazione è secondario rispetto a quello di redazione. Quest'ultimo, infatti, è il dato di maggiore interesse per l'estrazione di informazioni, poiché rappresentativo delle caratteristiche linguistiche del periodo considerato; infatti, i testi sono stati inseriti nella banca dati rimanendo fedeli all'edizione usata, senza produrre alcun tipo di modernizzazione o di correzione ortografica.

[5] Questa opzione sarà disponibile prossimamente.

Lanini, Ludovica. Corpus LBC Italiano

e-ISBN: 979-12-215-0305-0 | DOI: 10.36253/979-12-215-0305-0