Présentation de la base de données LBC

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

La Base de données LBC fait partie des instruments de support, en *Open Access*, développés par l'Unité de recherche *Lessico multilingue dei Beni Culturali* de manière à permettre la consultation de corpus qui fournissent des informations lexicales spécifiques, nécessaires pour effectuer des recherches lexicographiques et de traduction. L'Unité de Recherche souhaite en effet disposer d'un espace digital avec différents instruments utiles pour diffuser la connaissance du patrimoine artistique et culturel toscan à un niveau international (Farina 2016).

La Base de données permet d'effectuer des recherches à l'intérieur des corpus textuels des différentes langues publiées (français, anglais, italien, russe, espagnol, allemand) à partir de la plateforme du projet qui contient différents outils comprenant les corpus et des informations sur ceux-ci[2].

Les corpus ont été créés à partir de textes de différents genres dont des œuvres littéraires classiques, des récits de voyage ou des correspondances, des textes scientifiques et techniques, des guides touristiques, des manuels, etc., couvrant un laps de temps étendu, et les sources ont été structurées et gérées par le biais d'un logiciel avec des fonctions ciblées, pour répondre à des utilisations

diversifiées. En particulier, les principaux utilisateurs auxquels les corpus s'adressent sont : des linguistes, des chercheurs en littérature et en sciences humaines et sociales, qui ont besoin d'obtenir des informations sur le lexique par auteur, période, genre, etc. ; des traducteurs qui doivent consulter des ressources lexicales de spécialité ; et enfin des spécialistes du secteur touristique, ou des touristes qui souhaitent approfondir leur connaissance du territoire et de la culture qui s'y rattache.

Pour chaque langue du projet, des textes dont l'objet et le genre correspondent à ceux du projet sont présents, qui ont été choisis en priorité selon deux critères : pour les textes en langue originale, d'une part l'autorité reconnue du texte/auteur dans la culture d'appartenance et sa diffusion (Billero, Nicolás 2017 : 208), d'autre part la facilité de conversion dans un format éditable, en évitant des textes difficiles à numériser pour la première phase. Pour les textes en traduction, le choix se base sur une bibliographie rédigée par le groupe qui contient les textes en italien et dans les autres langues, considérés comme primordiaux pour la connaissance du patrimoine artistique et culturel toscan dans le monde. Il s'agit des textes de référence d'histoire de l'art qui se réfèrent à la Toscane comme les *Vies* de Vasari, les livres d'architecture d'Alberti, Palladio, Sellio, certains ouvrages de Machiavel et de Léonard de Vinci ; les récits de voyage les plus célèbres comme ceux de Stendhal et Ruskin, des livres sur l'art comme ceux de Burckhardt.

Toutefois, dans la phase actuelle, les corpus ne donnent pas les mêmes priorités et proportions aux différents types de textes, et ce, pour plusieurs raisons : le critère de l'accessibilité des sources ne s'applique pas au mêmes textes selon les pays, et l'intérêt envers le patrimoine toscan varie tant selon les périodes historiques que selon les genres textuels dans les différentes langues/cultures représentées dans le projet.

De cela dérive une hétérogénéité entre les corpus que nous voudrions limiter dans les futurs développements du projet. En effet, l'analyse de la distribution des types de textes choisis pour chaque corpus et des siècles représentés à la fin de cette première phase de création des corpus pourra permettre d'obtenir une plus grande homogénéisation dans le futur, permettant des travaux de comparaison plus étendus entre les corpus. Dans la première phase,

priorité a été donnée à l'introduction des textes de référence de chaque langue, ce qui a permis d'obtenir une base de textes consistante et suffisante pour des recherches à l'intérieur d'une langue particulière.

Après une analyse attentive des différents logiciels utilisables pour la consultation de corpus, notre choix s'est porté sur NoSketchEngine (Billero 2020), en raison de la présence de différentes fonctions intéressantes en relation avec les buts du projet, qui permettaient des recherches de concordances et des filtres sur la base de différentes caractéristiques pour lesquelles nous donnerons quelques exemples ici.

Il est possible d'accéder aux informations sur la nature des contenus de chaque corpus en accédant à la partie « Corpus info » disponible dans le menu de NoSketchEngine (figure 1).

Figure 1 – Informations détaillées sur le corpus français disponibles sur « Corpus info »

Sur cette page, on peut consulter aussi des informations relatives aux différentes données chiffrées sur les documents de chacune des catégories définies, comme illustré dans la figure 2 pour le corpus anglais :

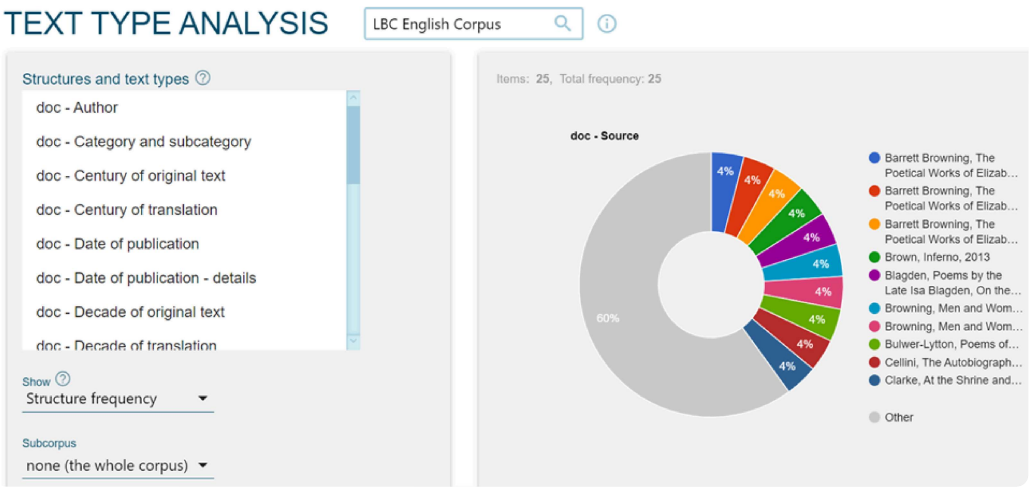


Figure 2 – Structure et caractéristiques des documents du corpus anglais

La structure des corpus LBC suit les règles traditionnelles, respectant des critères partagés de gestion des métadonnées, qui se reflètent dans la recherche par le biais de 'Search' sur les différents types de textes ('Text types'[3], figure 3).

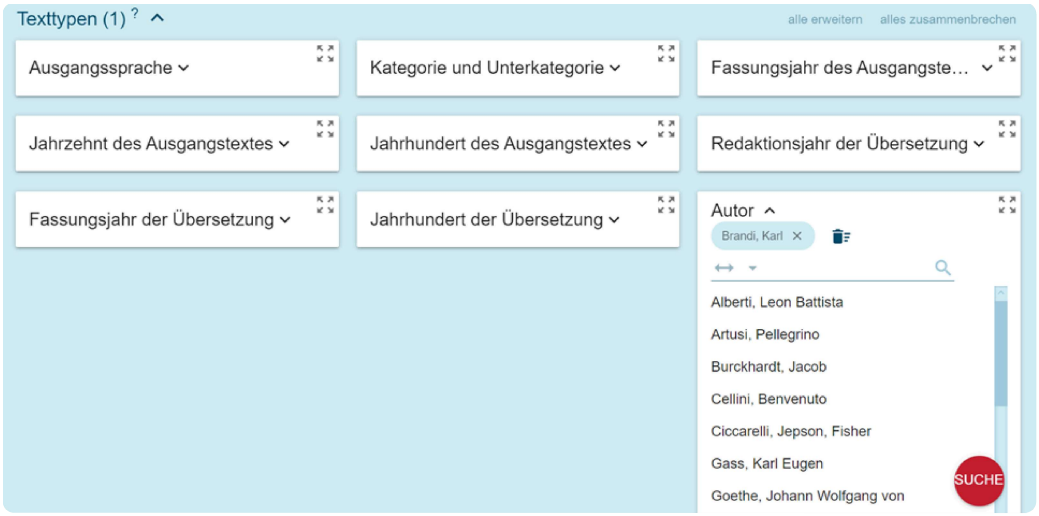


Figure 3 – Recherche dans le corpus allemand par le biais de la fenêtre 'Text types'

Les métadonnées par le biais desquelles on peut filtrer la recherche de concordances sont :

- Langue originale : apparaissent tant la langue du texte que celle du texte en traduction ;

- Langue de traduction : permet une recherche sur tous les textes traduits dans le corpus ;
- Catégorie et sous-catégorie : indique les différents types de textes. Tous les textes s'intéressent au patrimoine artistique, et offrent une ample vision de Florence et de la Toscane selon différents points de vue. Nous avons distingué quatre grandes catégories textuelles : vulgarisation (DIV pour *divulgativo*), technique (TEC pour *tecnico*), dictionnaires (DIZ pour *dizionario*) et littéraire (LET pour *letterario*). Chacune de ses catégorie admet des sous-catégories : blog (BLG pour *blog*), guides (GUI pour *guide*), revues (RIV pour *riviste*) pour les textes de vulgarisation ; architecture (ARC pour *architettura*), art (ART pour *arte*), œno-gatronomie (ENO pour *enograstronomia*) pour les textes techniques ; biographique (BIO pour *biografico*), fiction (FICT pour *fiction*), essais (SAG pour *saggistica*) pour les textes littéraires ; monolingue (MON pour *monolingue*) et bilingue (BIL pour *bilingue*) pour les dictionnaires. Pour distinguer les textes appartenant à ces différentes catégories, nous avons tenu compte de la destination principale de l'ouvrage et du type de lecteur auquel il s'adresse, données qui conditionnent le type de langue utilisée et son niveau de spécialisation[4] ;
- Auteur : sont indiqués le nom et le prénom de l'auteur ou 'sa' pour 'sans auteur' ;
- Titre et fragment : nous avons choisi d'introduire tant des textes entiers que des fragments qui correspondent à une unité textuelle complète : chapitres de livre, lettres, articles de revues, etc. Nous avons fait ce choix parce que dans de nombreux cas seule une partie des ouvrages coïncidait avec les intérêts du projet mais aussi pour faciliter la future réalisation de base de données parallèles avec les textes traduits intégrés à nos corpus. Pour les textes traduits nous avons indiqués tant les titres/titres de fragments originaux que leurs traductions ;
- Année de rédaction / année de publication / année de traduction : l'information chronologique propose une différence entre date de rédaction des textes (lorsque c'est possible) et la date d'édition ; pour les textes traduits, les mêmes données sont fournies tant pour le texte original que pour le texte traduit[5]. Pour les publications en ligne, c'est la date de consultation qui est indiquée ;

- Source : permet de faire une recherche sur un document unique du corpus (livre ou fragment) ;
- Zone géographique[6] : pour les documents qui ont une ville ou une région définie comme objet, nous avons indiqué le nom de la ville ou région. Cette indication est présente principalement pour les récits de voyages et pour les correspondances.

À ces informations s'ajoutent des détails bibliographiques plus complets que l'on peut consulter depuis la page des concordances, en cliquant sur la partie de référence (nom du fichier, numéro du document, nom de l'auteur, etc. selon les options choisies dans 'View options', fig. 3).

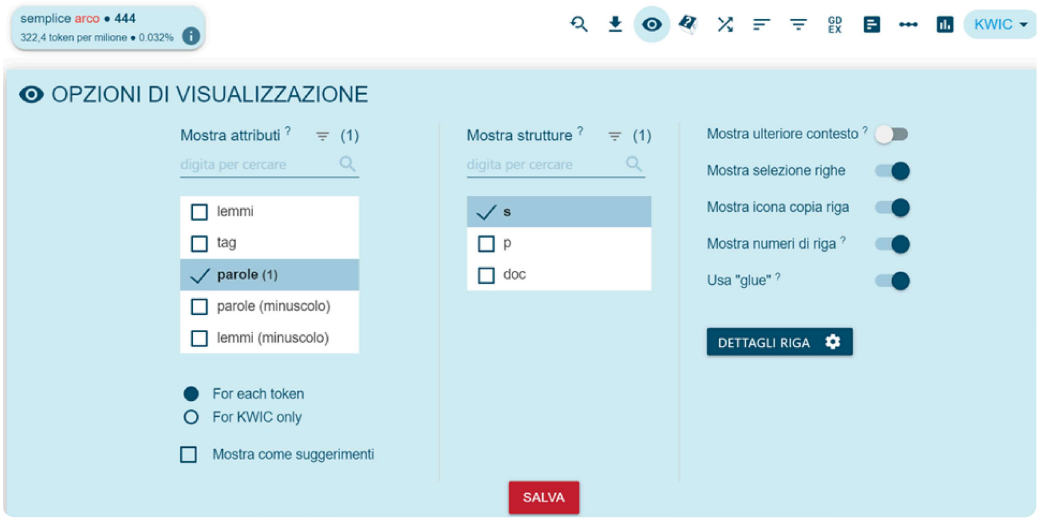


Figure 4 – Choix disponibles pour l'affichage des références textuelles dans le menu 'View options'.

Grâce à la fonction 'Search', on peut accéder aux concordances affichées dans un ordre casuel (numéro du document) comme dans la figure 5 ou bien par ordre alphabétique par rapport au mot considéré ou à son contexte droit ou gauche, par le biais de l'option 'Sort left/right' (figure 6).

Figure 5 – Recherche de concordances sur le lemme *pintar* dans le corpus espagnol sans ordre particulier.

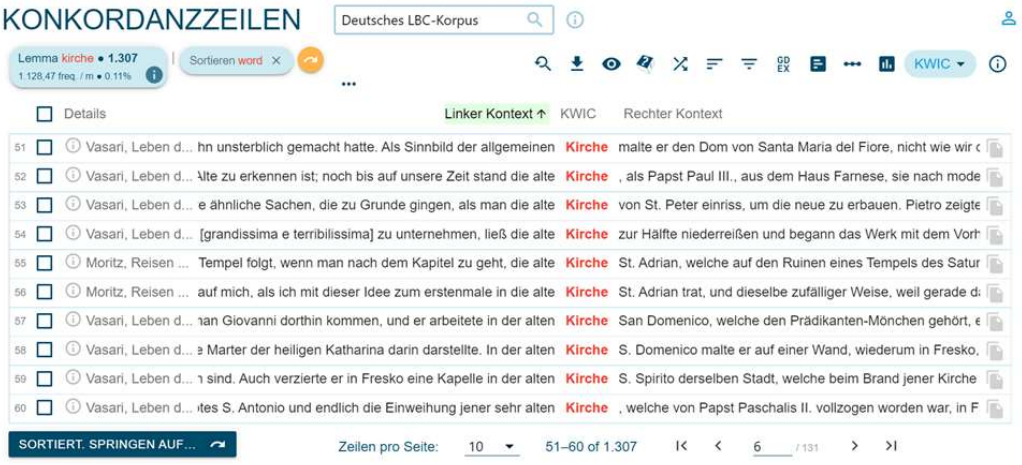


Figure 6 – Recherche de concordances sur le lemme *Kirche* dans le corpus allemand avec ordre alphabétique à gauche du lemme.

On peut aussi effectuer une recherche pour vérifier la co-présence de deux mots ou lemmes dans un même contexte, à une distance définie de *tokens* en utilisant la fonction 'Context' dans le menu de recherche 'Search', comme sur la figure 7. Cela nous permet par exemple dans ce cas de vérifier les utilisations attestées de différentes collocations contenant les lemmes *dipingere* et *fresco* (*dipingere a fresco / in fresco* que l'on voit dans les résultats, sur la figure 8).

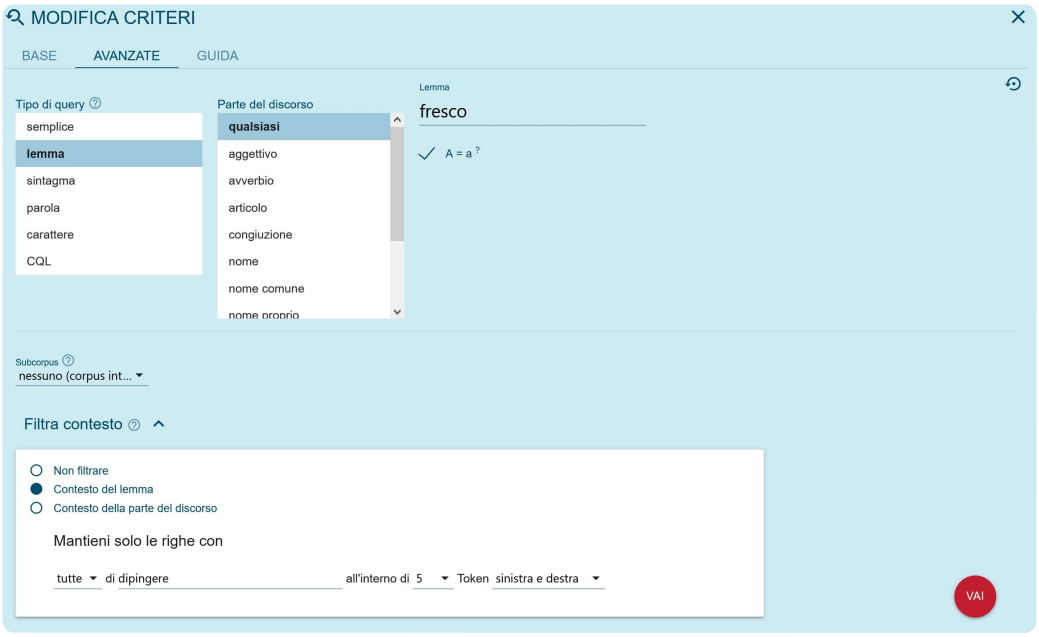


Figure 7 – Recherche des lemmes *dipingere* et *fresco* à 5 *tokens* de distance dans le corpus italien.

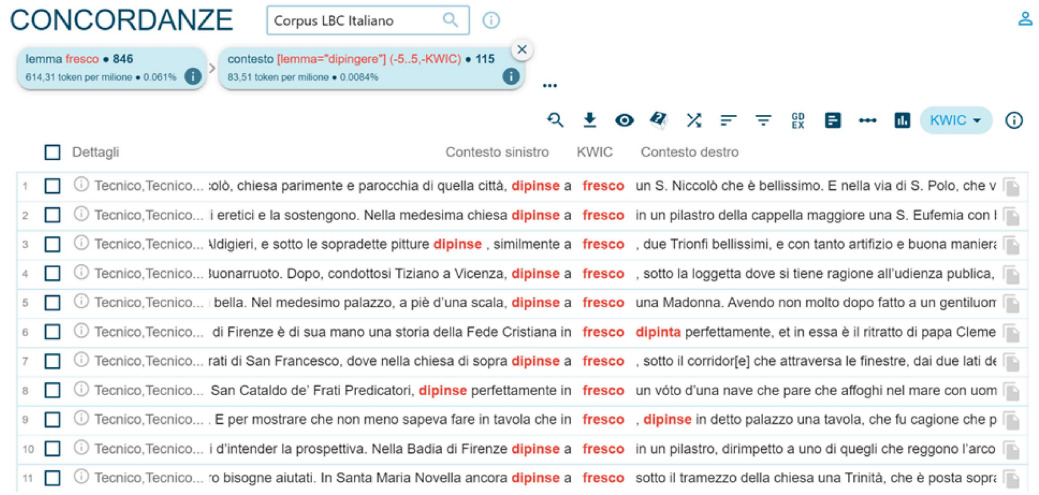


Figure 8 – Concordances relatives à la recherche de *dipingere* et *fresco* dans un même contexte dans le corpus italien.

La fonction ‘Word list’ permet d’obtenir des résultats numériques sur les fréquences présentes dans un corpus tant sur les sources, en cherchant par exemple la fréquence de mots attribuables à chaque auteur (figure 9), que sur les lemmes d’un corpus (figures 10-11).

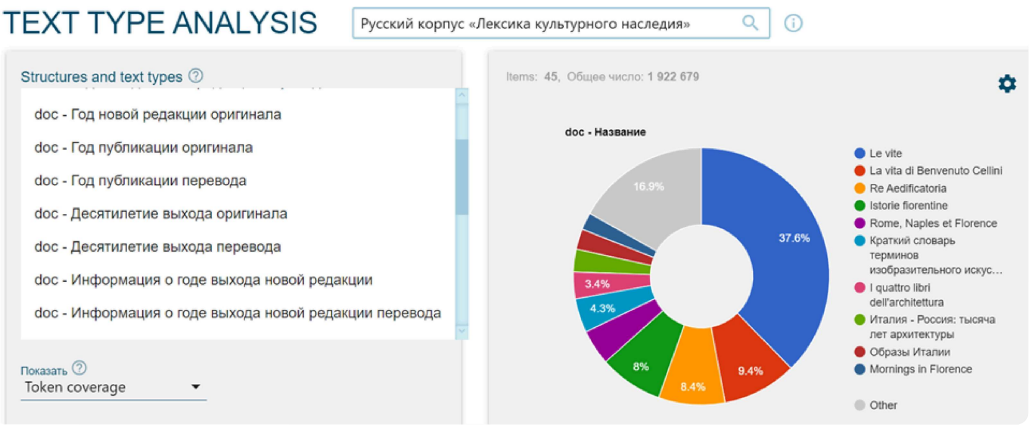


Figure 9 – Fréquence des *tokens* présents par auteur dans le corpus russe.

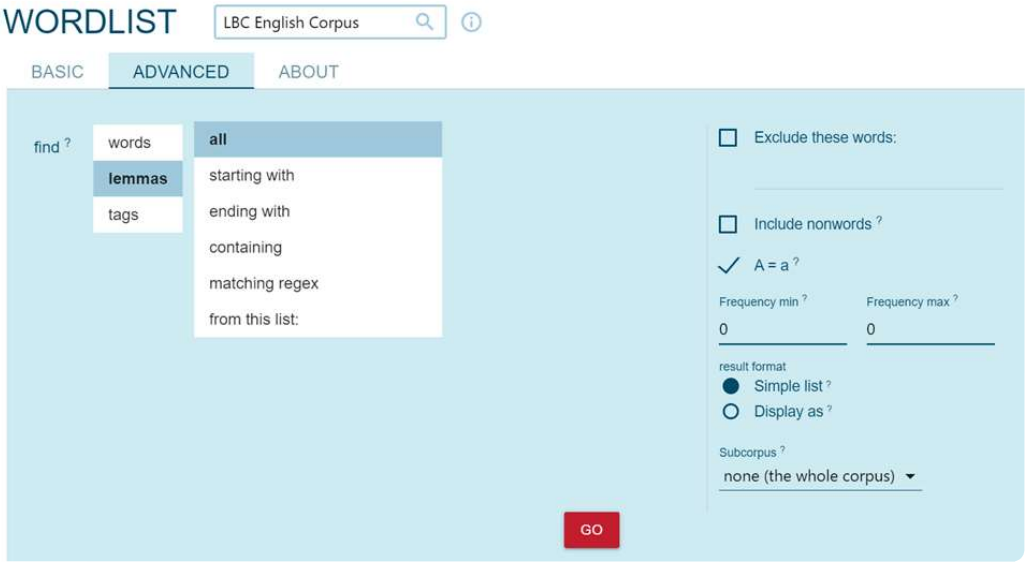
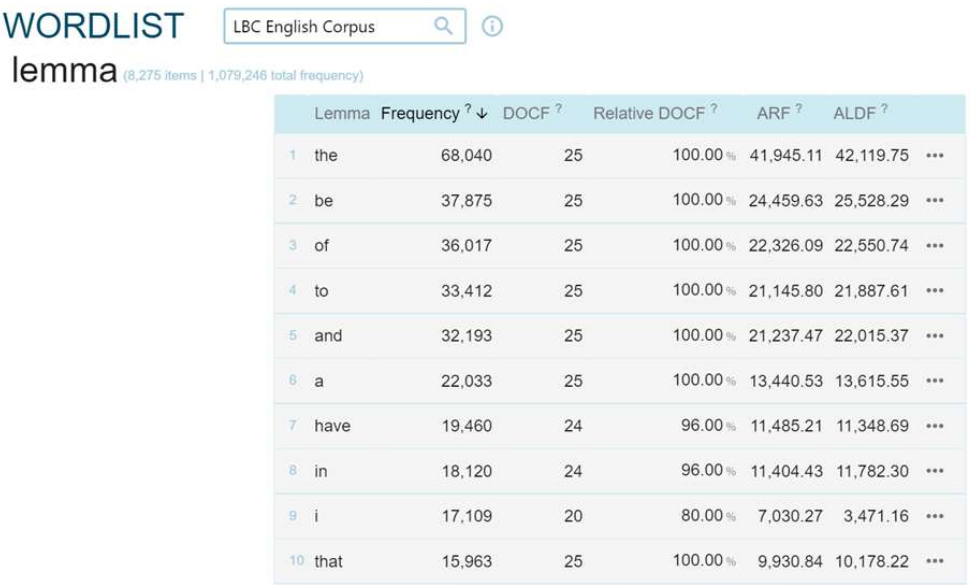


Figure 10 – Recherche sur 'Word list' des lemmes présents dans le corpus anglais.


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

Figure 11 – Résultat de la recherche sur 'Word list' sur les lemmes présents dans le corpus anglais

L'achèvement de cette première phase de travail sur nos corpus est très satisfaisant dans la mesure où il a permis de créer les bases nécessaires pour pouvoir effectuer les premiers travaux de notre équipe (Carpi 2017 ; Farina, Billero 2018 ; Billero, Carpi 2018 ; Garzaniti 2020 ; Farina, Flinz 2020). Nous avons déjà réalisé des 'Lexiques essentiels' de toutes les langues qui ont publié leur corpus accompagnés de concordances extraites des corpus qui seront publiés sur notre plateforme d'ici 2021 et qui pourront servir de base pour l'élaboration des futurs dictionnaires.

L'objectif principal de ce premier travail, réalisé par les différentes équipes linguistiques, était de permettre de valider les corpus. Nous sommes conscients en effet que c'est en les utilisant que nous pouvons repérer des problèmes qui ne se seraient pas manifestés autrement.

Dans le futur, nous pensons augmenter tant le nombre de langues (les corpus de chinois, portugais et turc, langues qui font partie du projet, sont encore en cours de création) que celui des textes présents dans les différents corpus, dans un but d'homogénéisation déjà décrit, pour essayer de faire en sorte qu'ils soient le plus comparable possible entre eux.

Bibliographie

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.

Notes

[1] Ce texte est une traduction faite par Annick Farina de l'introduction en italien du corpus LBC publiée sur http://corpora.lessicobeniculturali.net/it/

[2] Nous renvoyons à la publication du groupe de recherche sur l'ensemble des corpus (Farina, Nicolás Martínez, Billero 2020) pour avoir de plus amples informations sur chacun d'eux.

[3] L'interface de recherche que l'on trouve avec l'option 'Text types' est en italien pour le moment, mais nous la modifierons d'ici peu pour qu'on puisse la consulter dans la langue de chaque corpus.

[4] Dans la prochaine phase du projet, la classification sera revue sur la base des problèmes rencontrés par certains groupes avec des textes qui pouvaient être considérés comme appartenant à plusieurs catégories, comme par exemple des textes d'auteurs classiques dont le style est clairement littéraire mais qui peuvent être considérés comme spécialisés tant en raison des thèmes abordés que de leur vocabulaire (par exemple *l'Histoire de la Peinture en Italie* de Stendhal classifié actuellement dans la catégorie Littéraire/essais (LET_SAG).

[5] Les textes contenus vont de la Renaissance à nos jours. Même si les deux dates sont disponibles pour la recherche, l'année de publication est secondaire par rapport à celle de rédaction. Cette dernière date, en effet, est la donnée la plus intéressante lors de l'extraction d'informations sur les textes, parce que représentative des caractéristiques linguistiques de la période considérée ; en effet, les textes ont été introduits dans la base de données en respectant fidèlement l'édition utilisée, sans correction ou modernisation orthographique.

[6] Cette option sera bientôt disponible.

Farina, Annick. Corpus LBC Français

e-ISBN: 979-12-215-0309-8 | DOI: 10.36253/979-12-215-0309-8