Presentation of the LBC database[1]

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

The LBC Database is one of the support tools in Open Access developed by the *Lessico multilingue dei Beni Culturali* project (Multilingual Lexicon of Cultural Heritage). The Research Unit aims to provide a corpora with such specific lexical information as to enable lexicographic and translation research. We created, for this purpose, a digital space with various tools, which will spread the knowledge of Tuscany's artistic and cultural heritage at an international level (Farina 2016). The database permits searches within the texts of published corpora in six different languages (French, English, Italian, Russian, Spanish and German) through the project's platform, which contains various tools including the corpora and information about them[2].

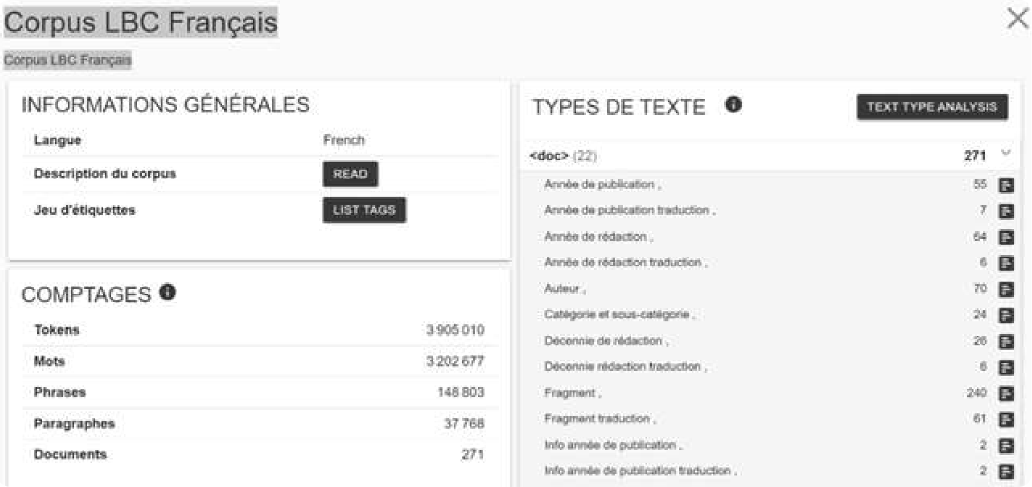
The corpora encompass texts of various genres, such as classical literary works, travel novels or correspondence, scientific and technical texts, tourist guides, textbooks, etc., all written over an extended period of time; these sources have been organized and managed through a software with functions suited for responding to the needs of multiple users. The main target groups of the corpora are: linguists, scholars, and humanities and social sciences researchers, whose work requires data on the lexical information of an author, chronological period, genre, etc.; translators who need to consult specific lexical resources; specialists in the tourism sector; or tourists interested in deepening their knowledge of the territory and its culture.

For each language of the project, the texts were chosen based on two priorities: first, the recognised prestige of the text/author in the source culture (Billero, Nicolás 2017: 208); secondly, the ease of conversion into an editable format, avoiding texts that were difficult to digitise. The translated texts were drawn from a list created by the project's members containing the texts in Italian and other languages considered essential for the international knowledge of Tuscany's artistic and cultural heritage: the fundamental art history texts referring to Tuscany, such as Vasari's *The Lives of the Most Excellent Painters, Sculptors, and Architects*; the architecture books by Alberti, Palladio and Sellio; some writings by Machiavelli and Leonardo Da Vinci; well-known travel books, such as the travels of Stendhal and Ruskin; and art books like Burckhardt's.

However, the different types of texts in the corpora were given different priorities and proportions, for various reasons: the criterion of accessibility to sources is obviously different according from country to country and so is the interest in Tuscan heritage, which varies according to historical periods and textual genres in the various languages/cultures represented in the project. In the future development of the project, we hope to limit the resulting disparity among the corpora. In fact, at the end of this initial corpora-building phase, an analysis of the distribution of the types of texts in each corpus, as well as their time periods, will allow for greater uniformity in the future, thus enabling more comparative work on texts. In this first phase, instead, priority was given to including reference texts which provide a sufficiently comprehensive base of texts for searches in each single language.

After a careful analysis of the various software available for consulting the corpora, our choice fell on NoSketchEngine, as it offers several interesting features which met the project's purposes, namely allowing concordance queries and filters with various features (Billero 2020).

Specifically, you can access information on the nature of the contents of each corpus by accessing "Corpus info" from the NoSketchEngine menu (Figure 1).



Corpus LBC Français

Corpus LBC Français

INFORMATIONS GÉNÉRALES

Langue	French
Description du corpus	READ
Jeu d'étiquettes	LIST TAGS

COMPTAGES

Tokens	3 905 010
Mots	3 202 677
Phrases	148 803
Paragraphes	37 768
Documents	271

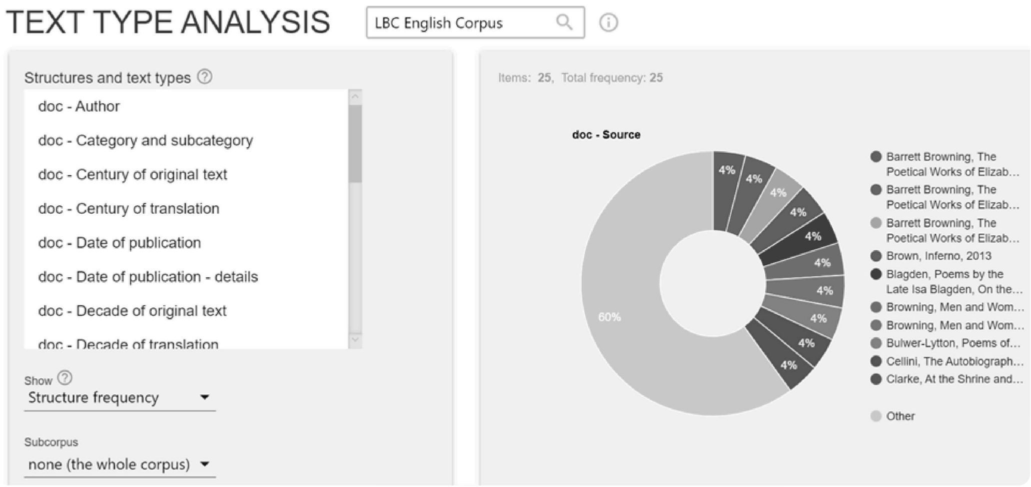
TYPES DE TEXTE

TEXT TYPE ANALYSIS

<doc> (22)	271
Année de publication ,	55
Année de publication traduction ,	7
Année de rédaction ,	64
Année de rédaction traduction ,	6
Auteur ,	70
Catégorie et sous-catégorie ,	24
Déconnie de rédaction ,	26
Déconnie rédaction traduction ,	6
Fragment ,	240
Fragment traduction ,	61
Info année de publication ,	2
Info année de publication traduction ,	2

Figure 1 – Detailed information on the French corpus available under "Corpus info" [Nov. 2022].

This same page also provides information about the quantities of the various documents in each of the categories provided, as shown in Figure 2 for the English corpus



TEXT TYPE ANALYSIS

LBC English Corpus

Structures and text types

- doc - Author
- doc - Category and subcategory
- doc - Century of original text
- doc - Century of translation
- doc - Date of publication
- doc - Date of publication - details
- doc - Decade of original text
- doc - Decade of translation

Show: Structure frequency

Subcorpus: none (the whole corpus)

Items: 25, Total frequency: 25

doc - Source
4%
4%
4%
4%
4%
4%
4%
4%
4%
4%
60%
Barrett Browning, The Poetical Works of Elizab...
Barrett Browning, The Poetical Works of Elizab...
Barrett Browning, The Poetical Works of Elizab...
Brown, Inferno, 2013
Blagden, Poems by the Late Isa Blagden, On the...
Browning, Men and Wom...
Browning, Men and Wom...
Bulwer-Lytton, Poems of...
Cellini, The Autobiograph...
Clarke, At the Shrine and...
Other

Figure 2 – Structures and attributes of the documents inserted in the English corpus [Nov. 2022].

The structure of the corpora follows the traditional rules concerning shared metadata management, as can be seen in a "Search" for text types ("Text types"[3], Figure 3).

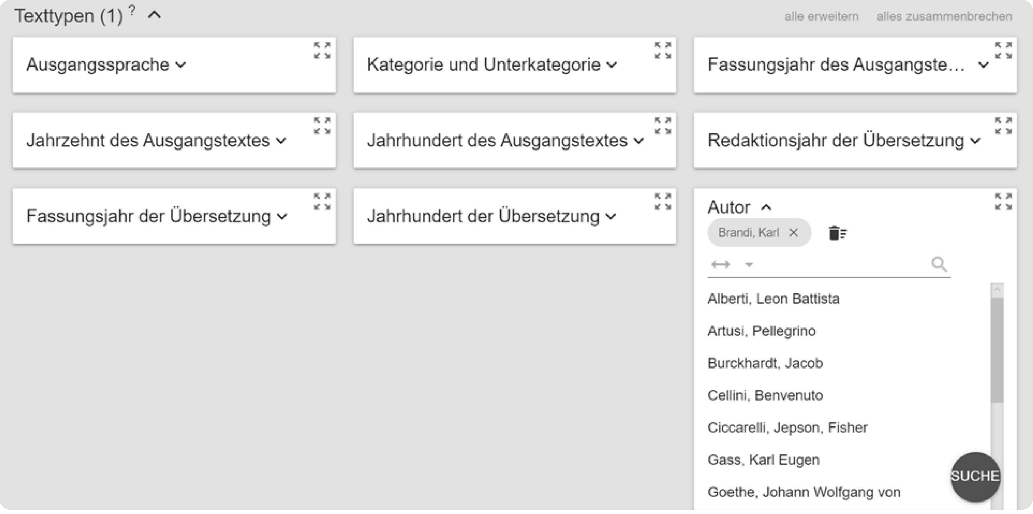

Texttypen (1) ?
alle erweitern
alles zusammenbrechen
Ausgangssprache
Kategorie und Unterkategorie
Fassungsjahr des Ausgangste…
Jahrzehnt des Ausgangstextes
Jahrhundert des Ausgangstextes
Redaktionsjahr der Übersetzung
Fassungsjahr der Übersetzung
Jahrhundert der Übersetzung
Autor
Brandi, Karl ×
Alberti, Leon Battista
Artusi, Pellegrino
Burckhardt, Jacob
Cellini, Benvenuto
Ciccarelli, Jepson, Fisher
Gass, Karl Eugen
Goethe, Johann Wolfgang von
SUCHE

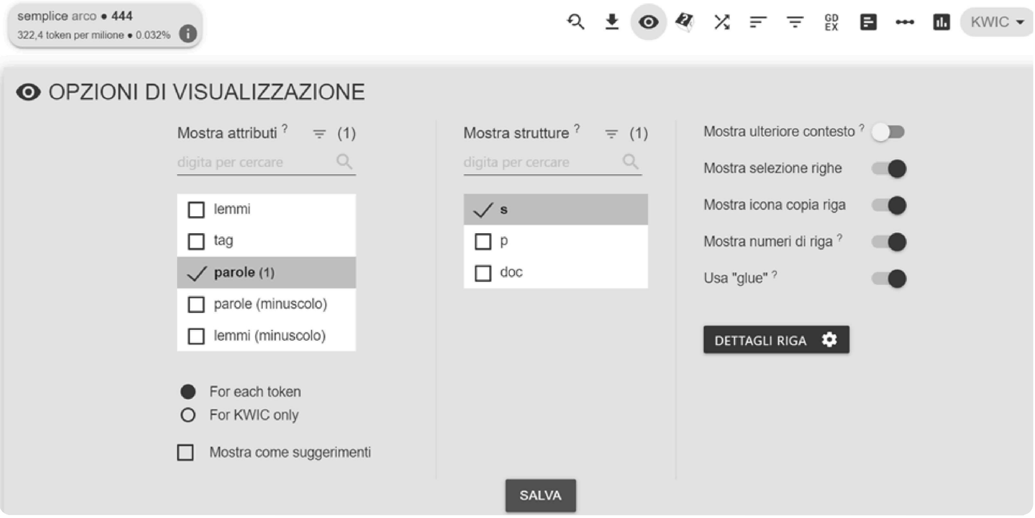
Figure 3 – Search in the German corpus through the "Text types" window.

The metadata you can filter the concordance search with are:

- Original language: both the language of the text and the source language for translated texts appear;
- Translation language: allows a search for all translations within the corpus language;
- Category and subcategory: indicates the various types of texts. All the texts have as their subject the artistic heritage and its lexicon, in particular a broad survey of Florence and Tuscany from different points of view. There are four macro-categories (Informative, Technical, Dictionary and Literary), each with their relative sub-categories (Informative: Blog, Guide, Magazine; Technical: Architecture, Art, Food and Wine; Literary: Biographical, Fiction, Essay; Dictionary: Monolingual, Bilingual / Multilingual). In deciding on these categories, the main uses and users of the work were taken into account; this conditioned the type of language involved and its level of specialisation[4], specifically:
- Author: surname and first name are indicated , with the designation "sa" when non-existent;
- Title and fragment: we chose the introduction of both whole texts and of fragments that correspond to a textual unit, such as book chapter, complete letter, journal article, etc. This was due to the fact that in many cases the entire book did not coincide with the interests of the project. Furthermore, this facilitates the future creation of parallel versions of translated texts. For the

translated texts both the original and translated titles are included;
- Year of writing / year of publication / year of translation: the chronological information differentiates between the period of writing of the texts (where possible) and the date of publication; for translated texts the same information was inserted both for the source text and for the translated one[5]. For online publications, the date of consultation is indicated;
- Source: allows to search within a single document of the corpus (book or fragment);
- Geographical delimitation[6]: for texts regarding a specific city or region, the name of the location has been inserted. This indication is present mainly for travel books and correspondence.

Further, more complete bibliographic details are visible by accessing the concordances by clicking on the reference (file name, document number, author name, etc. according to the options chosen in "View options", Figure 4).


semplice arco • 444
322,4 token per milione • 0.032%

OPZIONI DI VISUALIZZAZIONE

Mostra attributi ? (1)
digita per cercare
☐ lemmi
☐ tag
✓ parole (1)
☐ parole (minuscolo)
☐ lemmi (minuscolo)

● For each token
○ For KWIC only

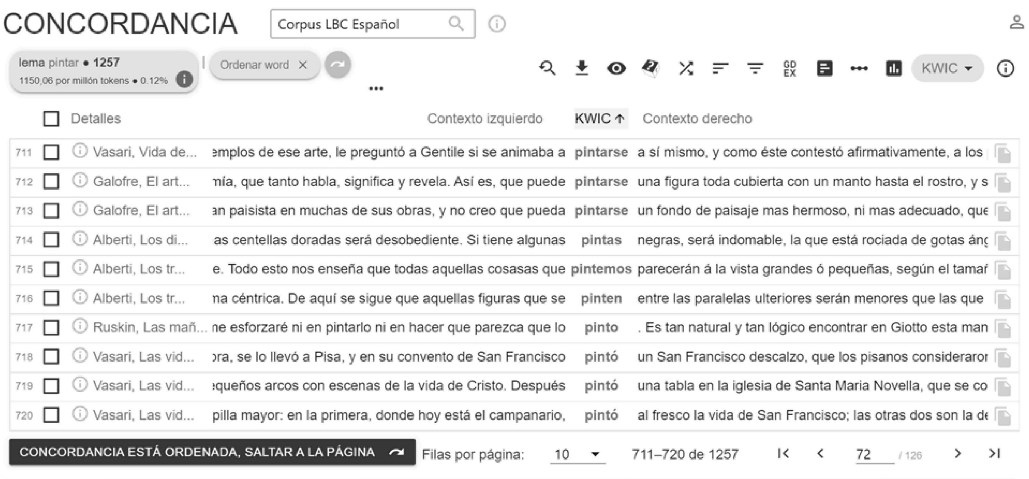
☐ Mostra come suggerimenti

Mostra strutture ? (1)
digita per cercare
✓ s
☐ p
☐ doc

Mostra ulteriore contesto ? ○
Mostra selezione righe ●
Mostra icona copia riga ●
Mostra numeri di riga ? ●
Usa "glue" ? ●

Figure 4 – Choices available for viewing the textual references in "View options".

Using the "Search" function, you can access the concordances displayed in random order (on the number of documents) as in Figure 5, or in alphabetical order in relation to the queried word or to its right or left context, using the "Sort left/right" function (Figure 6).



CONCORDANCIA　Corpus LBC Español

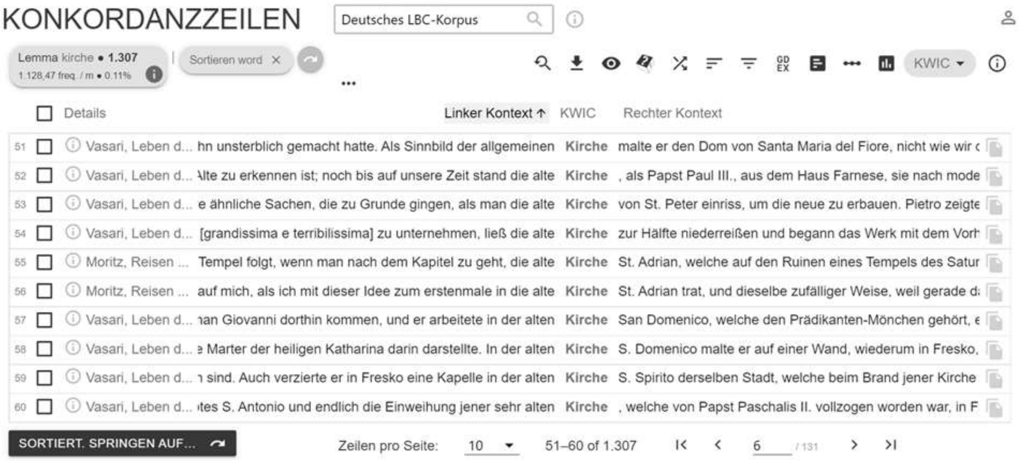
lema pintar • 1257
1150,06 por millón tokens • 0.12%

Ordenar word ×

		☐ Detalles	Contexto izquierdo	KWIC ↑	Contexto derecho
711	☐	Vasari, Vida de...	emplos de ese arte, le preguntó a Gentile si se animaba a	pintarse	a sí mismo, y como éste contestó afirmativamente, a los
712	☐	Galofre, El art...	nía, que tanto habla, significa y revela. Así es, que puede	pintarse	una figura toda cubierta con un manto hasta el rostro, y s
713	☐	Galofre, El art...	an paisista en muchas de sus obras, y no creo que pueda	pintarse	un fondo de paisaje mas hermoso, ni mas adecuado, que
714	☐	Alberti, Los di...	as centellas doradas será desobediente. Si tiene algunas	pintas	negras, será indomable, la que está rociada de gotas ánç
715	☐	Alberti, Los tr...	e. Todo esto nos enseña que todas aquellas cosasas que	pintemos	parecerán á la vista grandes ó pequeñas, según el tamañ
716	☐	Alberti, Los tr...	na céntrica. De aquí se sigue que aquellas figuras que se	pinten	entre las paralelas ulteriores serán menores que las que
717	☐	Ruskin, Las mañ...	ne esforzaré ni en pintarlo ni en hacer que parezca que lo	pinto	. Es tan natural y tan lógico encontrar en Giotto esta man
718	☐	Vasari, Las vid...	ora, se lo llevó a Pisa, y en su convento de San Francisco	pintó	un San Francisco descalzo, que los pisanos consideraron
719	☐	Vasari, Las vid...	equeños arcos con escenas de la vida de Cristo. Después	pintó	una tabla en la iglesia de Santa Maria Novella, que se co
720	☐	Vasari, Las vid...	pilla mayor: en la primera, donde hoy está el campanario,	pintó	al fresco la vida de San Francisco; las otras dos son la de

CONCORDANCIA ESTÁ ORDENADA, SALTAR A LA PÁGINA　Filas por página: 10　711–720 de 1257　72 / 126

Figure 5 – Search for concordances for the lemma *pintar* in the Spanish corpus without choice of order.



KONKORDANZZEILEN　Deutsches LBC-Korpus

Lemma kirche • 1.307
1.128,47 freq. / m • 0.11%

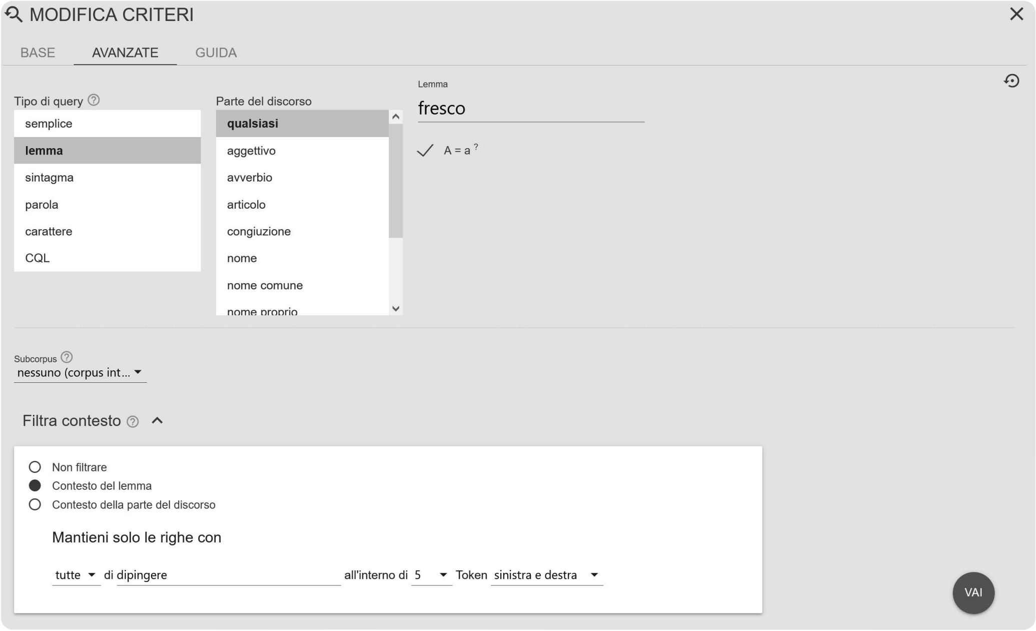
Sortieren word ×

		☐ Details	Linker Kontext ↑	KWIC	Rechter Kontext
51	☐	Vasari, Leben d...	hn unsterblich gemacht hatte. Als Sinnbild der allgemeinen	Kirche	malte er den Dom von Santa Maria del Fiore, nicht wie wir c
52	☐	Vasari, Leben d...	Alte zu erkennen ist; noch bis auf unsere Zeit stand die alte	Kirche	, als Papst Paul III., aus dem Haus Farnese, sie nach mode
53	☐	Vasari, Leben d...	e ähnliche Sachen, die zu Grunde gingen, als man die alte	Kirche	von St. Peter einriss, um die neue zu erbauen. Pietro zeigte
54	☐	Vasari, Leben d...	[grandissima e terribilissima] zu unternehmen, ließ die alte	Kirche	zur Hälfte niederreißen und begann das Werk mit dem Vorh
55	☐	Moritz, Reisen ...	Tempel folgt, wenn man nach dem Kapitel zu geht, die alte	Kirche	St. Adrian, welche auf den Ruinen eines Tempels des Satur
56	☐	Moritz, Reisen ...	auf mich, als ich mit dieser Idee zum erstenmale in die alte	Kirche	St. Adrian trat, und dieselbe zufälliger Weise, weil gerade da
57	☐	Vasari, Leben d...	nan Giovanni dorthin kommen, und er arbeitete in der alten	Kirche	San Domenico, welche den Prädikanten-Mönchen gehört, e
58	☐	Vasari, Leben d...	e Marter der heiligen Katharina darin darstellte. In der alten	Kirche	S. Domenico malte er auf einer Wand, wiederum in Fresko,
59	☐	Vasari, Leben d...	n sind. Auch verzierte er in Fresko eine Kapelle in der alten	Kirche	S. Spirito derselben Stadt, welche beim Brand jener Kirche
60	☐	Vasari, Leben d...	tes S. Antonio und endlich die Einweihung jener sehr alten	Kirche	, welche von Papst Paschalis II. vollzogen worden war, in F

SORTIERT. SPRINGEN AUF...　Zeilen pro Seite: 10　51–60 of 1.307　6 / 131

Figure 6 – Search for concordances to the left of the lemma *Kirche* in the German corpus.

It is also possible to search for the presence of two words or lemmas in the same context at a chosen distance of *tokens* by using the “Context" function in the “Search” menu, as shown in Figure 7A. This would allow, for example, verification of the certified uses of various collocations (*dipingere a fresco / in fresco* in Italian in Figure 7B).



MODIFICA CRITERI

BASE | AVANZATE | GUIDA

Tipo di query: semplice / **lemma** / sintagma / parola / carattere / CQL

Parte del discorso: **qualsiasi** / aggettivo / avverbio / articolo / congiuzione / nome / nome comune / nome proprio

Lemma: fresco

✓ A = a ?

Subcorpus: nessuno (corpus int...

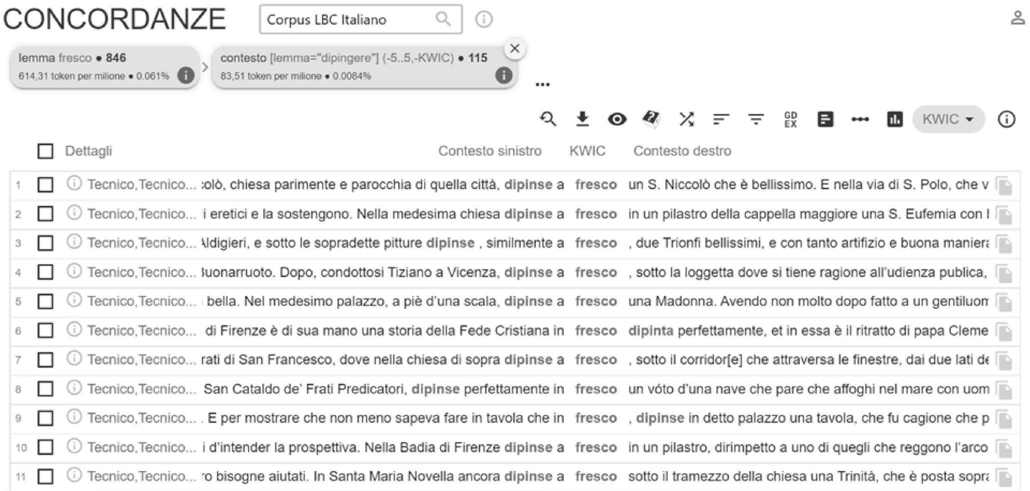
Filtra contesto

○ Non filtrare
● Contesto del lemma
○ Contesto della parte del discorso

Mantieni solo le righe con

tutte di dipingere all'interno di 5 Token sinistra e destra

Figure 7A - Search for the *dipingere* and *fresco* lemmas at 5 *tokens* distance in the Italian corpus.



CONCORDANZE

Corpus LBC Italiano

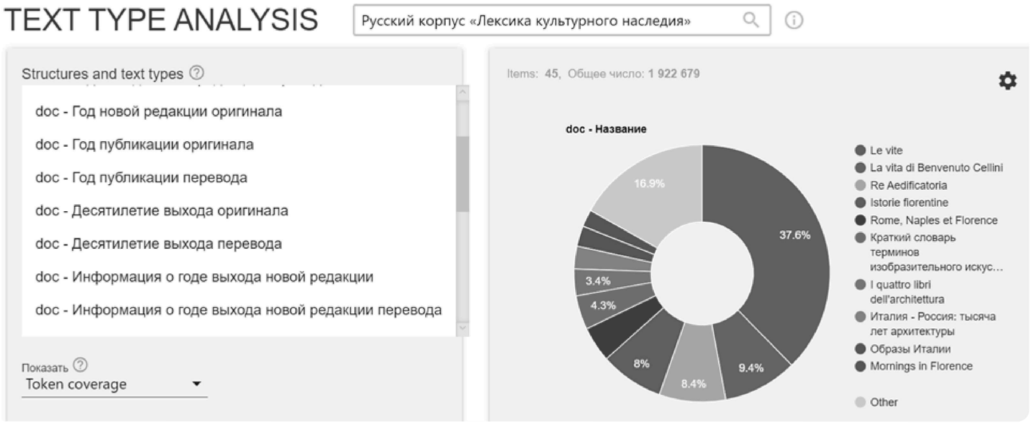
lemma fresco • 846
614,31 token per milione • 0.061%

contesto [lemma="dipingere"] (-5..5,-KWIC) • 115
83,51 token per milione • 0.0084%

	☐	Dettagli	Contesto sinistro	KWIC	Contesto destro
1	☐	Tecnico,Tecnico...	:olò, chiesa parimente e parocchia di quella città, **dipinse** a	**fresco**	un S. Niccolò che è bellissimo. E nella via di S. Polo, che v
2	☐	Tecnico,Tecnico...	i eretici e la sostengono. Nella medesima chiesa **dipinse** a	**fresco**	in un pilastro della cappella maggiore una S. Eufemia con l
3	☐	Tecnico,Tecnico...	Aldigieri, e sotto le sopradette pitture **dipinse** , similmente a	**fresco**	, due Trionfi bellissimi, e con tanto artifizio e buona maniera
4	☐	Tecnico,Tecnico...	Buonarruoto. Dopo, condottosi Tiziano a Vicenza, **dipinse** a	**fresco**	, sotto la loggetta dove si tiene ragione all'udienza publica,
5	☐	Tecnico,Tecnico...	bella. Nel medesimo palazzo, a piè d'una scala, **dipinse** a	**fresco**	una Madonna. Avendo non molto dopo fatto a un gentiluom
6	☐	Tecnico,Tecnico...	di Firenze è di sua mano una storia della Fede Cristiana in	**fresco**	**dipinta** perfettamente, et in essa è il ritratto di papa Cleme
7	☐	Tecnico,Tecnico...	rati di San Francesco, dove nella chiesa di sopra **dipinse** a	**fresco**	, sotto il corridor[e] che attraversa le finestre, dai due lati de
8	☐	Tecnico,Tecnico...	San Cataldo de' Frati Predicatori, **dipinse** perfettamente in	**fresco**	un vóto d'una nave che pare che affoghi nel mare con uom
9	☐	Tecnico,Tecnico...	. E per mostrare che non meno sapeva fare in tavola che in	**fresco**	, **dipinse** in detto palazzo una tavola, che fu cagione che p
10	☐	Tecnico,Tecnico...	i d'intender la prospettiva. Nella Badia di Firenze **dipinse** a	**fresco**	in un pilastro, dirimpetto a uno di quegli che reggono l'arco
11	☐	Tecnico,Tecnico...	ro bisogne aiutati. In Santa Maria Novella ancora **dipinse** a	**fresco**	sotto il tramezzo della chiesa una Trinità, che è posta sopra

Figure 7B - Concordances relative to the research of *dipingere* and *fresco* in the same context in the Italian corpus.

The "Word list" function allows users to obtain numerical results on the frequencies present in a corpus both according to the sources, by querying for example the frequencies of headwords attributable to each author (Figure 8), and according to the headwords of a corpus (Figures 9 and 10).



TEXT TYPE ANALYSIS

Русский корпус «Лексика культурного наследия»

Structures and text types

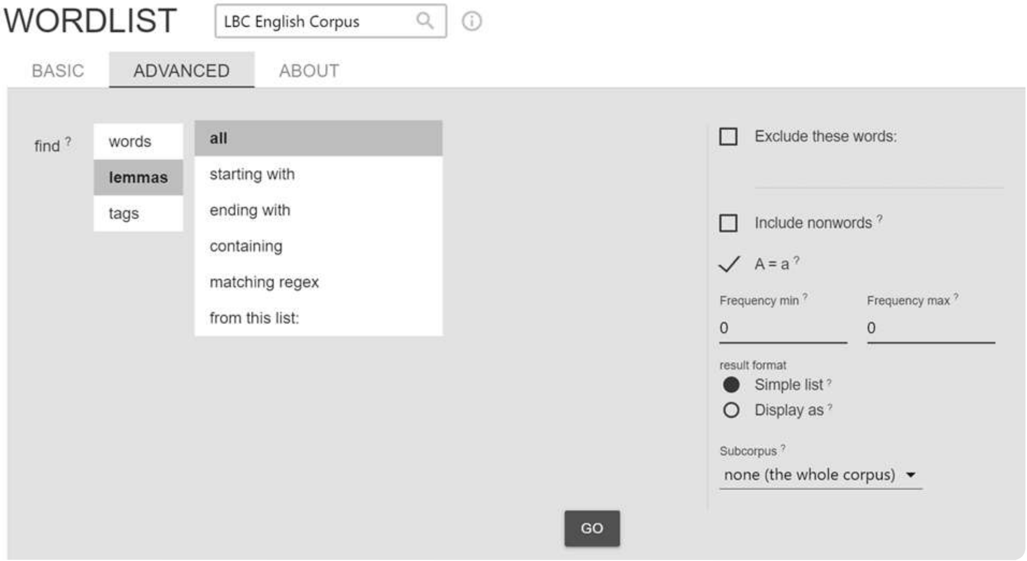
- doc - Год новой редакции оригинала
- doc - Год публикации оригинала
- doc - Год публикации перевода
- doc - Десятилетие выхода оригинала
- doc - Десятилетие выхода перевода
- doc - Информация о годе выхода новой редакции
- doc - Информация о годе выхода новой редакции перевода

Показать: Token coverage

Items: 45, Общее число: 1 922 679

doc - Название
37.6%
16.9%
3.4%
4.3%
8%
8.4%
9.4%
● Le vite
● La vita di Benvenuto Cellini
● Re Aedificatoria
● Istorie fiorentine
● Rome, Naples et Florence
● Краткий словарь терминов изобразительного искус…
● I quattro libri dell'architettura
● Италия - Россия: тысяча лет архитектуры
● Образы Италии
● Mornings in Florence
● Other

Figure 8 - Frequencies of *tokens* present for authors in the Russian corpus.



WORDLIST

LBC English Corpus

find: lemmas, all

□ Exclude these words:

□ Include nonwords

✓ A = a

Frequency min: 0

Frequency max: 0

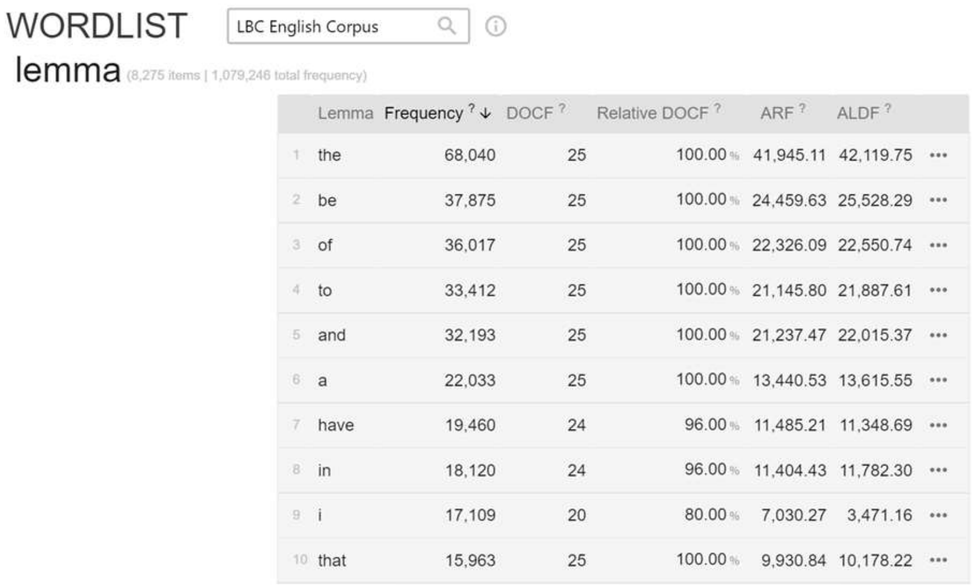
result format

● Simple list

○ Display as

Subcorpus: none (the whole corpus)

Figure 9 - Word list search of lemmas in the 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 10 - Result of the Word list search of lemmas in the English corpus [Nov. 2022].

The completion of this first phase of our corpora is to be considered satisfactory as it created the necessary foundations for initial works and research for our group (Carpi 2017; Farina, Billero, Carpi 2018; Garzaniti 2020; Farina, Flinz 2020). The first lists of entries of each language have already been performed; this will be complemented by concordances extracted from the corpora to be published on the platform by 2021. At that point it can be used for producing future dictionaries.

The main objective of this preliminary work, carried out by each *linguistic* team, was to validate the corpora, recognising that only by actually using it would they be able to identify problems that would otherwise remain latent.

In the future, expanding both the number of languages (currently there are still no Chinese, Portuguese or Turkish corpora, languages that are part of the LBC project) and the number of texts (to increase uniformity as mentioned above) are planned, to try to make the corpora as comparable as possible.

Bibliography

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] This text is a translation by Carlo Garavaglia of the Italian introduction to the LBC corpora published in http://corpora.lessicobeniculturali.net/it/

[2] For comprehensive data on the LBC corpora, please refer to the group’s publication (Farina, Nicolás Martínez, Billero 2020).

[3] The "Text Type" search interface is currently in Italian , but we will shortly edit it so that it can be consulted in the language of each corpus.

[4] In the next phase of the project the classification will be reviewed in order to overcome the problems encountered by some groups with texts that could be belong to more than one category, e.g. texts by classical authors whose style is clearly literary but who wrote texts that can be considered specialised for their content and vocabulary (e.g. Stendhal's *History of Painting* is classified for now in the literary/ essay category).

[5] The texts in the corpora range from the Renaissance to the present day. Although both dates are present, the year of publication is secondary to the year of writing. The latter, in fact, is the most interesting date when extracting information, since it is representative of the linguistic characteristics of the period being examined; in fact, the texts, as entered into the database, remain faithful to the edition used, without any kind of modernisation or spelling correction.

[6] This option will be available from 2023.

UNIVERSITÀ
DEGLI STUDI
FIRENZE
FORLILPSI
DIPARTIMENTO DI FORMAZIONE, LINGUE, INTERCULTURA, LETTERATURE E PSICOLOGIA

fup
FIRENZE
UNIVERSITY
PRESS

Natali, Ilaria. The LBC English Corpus

e-ISBN: 979-12-215-0307-4 | DOI: 10.36253/979-12-215-0307-4