



МЕНЮ

Презентация базы данных LBC (Лексики культурного наследия)^[1]

Annick Farina (Университет Флоренции),
Riccardo Billero (Университет Флоренции),
Carlota Nicolás Martínez (Университет Флоренции)

База данных LBC (Лексики культурного наследия) является одним из инструментов поддержки Open Access (открытого доступа), разработанных исследовательским объединением «Многоязычный лексикон культурного наследия», чтобы иметь возможность обращаться к корпусам, которые предоставляют конкретную лексическую информацию, необходимую для проведения лексикографических и переводческих исследований. Это объединение намеревается создать цифровое пространство с различными инструментами, используемыми для распространения и популяризации знаний о художественном и культурном наследии Тосканы на международном уровне (Farina 2016).

База данных позволяет выполнять поиск в корпусах текстов на всех заявленных языках: французском, английском, итальянском, русском, испанском и немецком, осуществляемый через платформу проекта, которая содержит различные инструменты, включая корпуса и информацию о них^[2].

Корпуса содержат тексты различных жанров: классические литературные произведения, романы о путешествиях или переписку, научно-технические тексты, туристические справочники, учебные пособия, создаваемые в течение длительного периода времени. Все источники были систематизированы и управляются с помощью многофункционального программного обеспечения, отвечающего потребностям множества пользователей.

Основной целевой группой, которой адресованы корпуса, являются: лингвисты, писатели, исследователи в области гуманитарных и социальных наук, чья работа требует разысканий для получения информации о лексике, систематизированной по авторам, хронологическому периоду, жанру и т. д., переводчики, которым необходимо обращаться к определенным лексическим ресурсам, и, наконец, специалисты в сфере туризма или туристы, заинтересованные в углублении своих знаний о территории и связанной с ней культуре.

Каждый язык проекта представлен текстами на языке оригинала, соответствующими по теме и жанру текстам всего проекта, и отобранными по двум основным критериям: общепризнанный авторитет текста или его автора в соответствующей культуре и его распространение (Billero, Nicolas 2017: 208); а также простота преобразования текста в редактируемый формат, что было особенно важно на первом этапе формирования корпуса.

Выбор переводных текстов основан на списке, составленном группой, и он содержит тексты как на итальянском, так и на тех языках, которые внесли весомый вклад в интернациональное изучение и оценку художественного и культурного наследия Тосканы: основополагающие тексты по истории искусства, как, например, «Жизнеописания» Вазари, труды по архитектуре Альберти, Палладио, Селлио, некоторые сочинения Макиавелли и Леонардо; знаменитые книги о путешествиях Стендаля и Раскина, и книги по искусству Буркхардта.

Надо отметить, что на данном этапе в каждом языковом корпусе уделяется различный приоритет разным типам текстов, что влияет на их соотношение в национальных корпусах. Эта асимметрия зависит от многообразных факторов, среди которых немаловажную роль играет критерий доступности источников,

варьирующийся от страны к стране, а также от интереса к наследию Тосканы, который находится в тесной связи с особенностями исторических периодов и текстовых жанров на языках и культурах, представленных в проекте.

Из этих обстоятельств вытекает неоднородность корпусов, которую мы хотели бы ограничить в будущем. Анализ распределения типов текстов, отобранных для каждого корпуса, и временных периодов, представленных в конце первой фазы формирования корпусов, даст возможность в будущем достичь большего однородности, позволяющей сравнивать тексты. Приоритет, отдаваемый на начальном этапе справочным текстам на каждом из участвующих в проекте языках, позволил создать представительную базу текстов для поиска на одном языке.

После тщательного анализа различного программного обеспечения, которое можно использовать для работы с корпусом, выбор пал на NoSketchEngine (Billero, 2020) благодаря наличию в нем нескольких полезных функций, необходимых для целей проекта: оно позволяет искать конкордансы и задавать фильтры на основе различных характеристик.

Информацию о характере содержания каждого корпуса можно получить, обратившись к части *“Corpus info”*, доступной в меню NoSketchEngine (Рис. 1).

The screenshot displays the 'Corpus LBC Français' interface. It is divided into several sections:

- 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:**
 - <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
 - Décennie de rédaction: 26
 - Décennie rédaction traduction: 6
 - Fragment: 240
 - Fragment traduction: 61
 - Info année de publication: 2
 - Info année de publication traduction: 2

Рис. 1 - Подробная информация о французском корпусе, доступная в «Corpus info» [ноябрь, 2022].

Эта страница содержит также информацию о количестве различных документов, представленных в каждой из описанных категорий, как показано на рисунке 2 для корпуса английского языка:

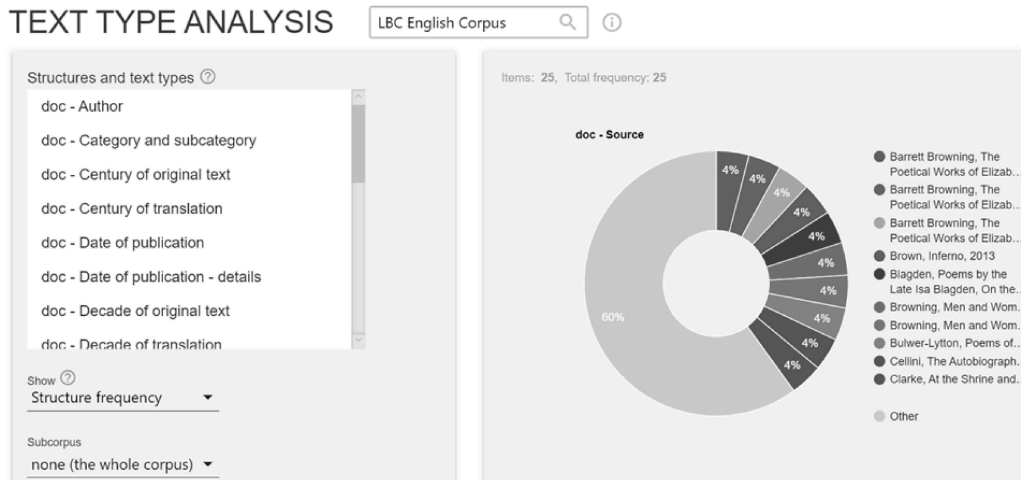


Рис. 2 - Структура и характеристика документов, входящих в английский корпус [ноябрь, 2022].

Структура корпуса соответствует традиционным правилам, касающимся общих критериев управления метаданными, что отражается в поиске "Search" по типам текста ("Text types"^[3], Рис. 3).

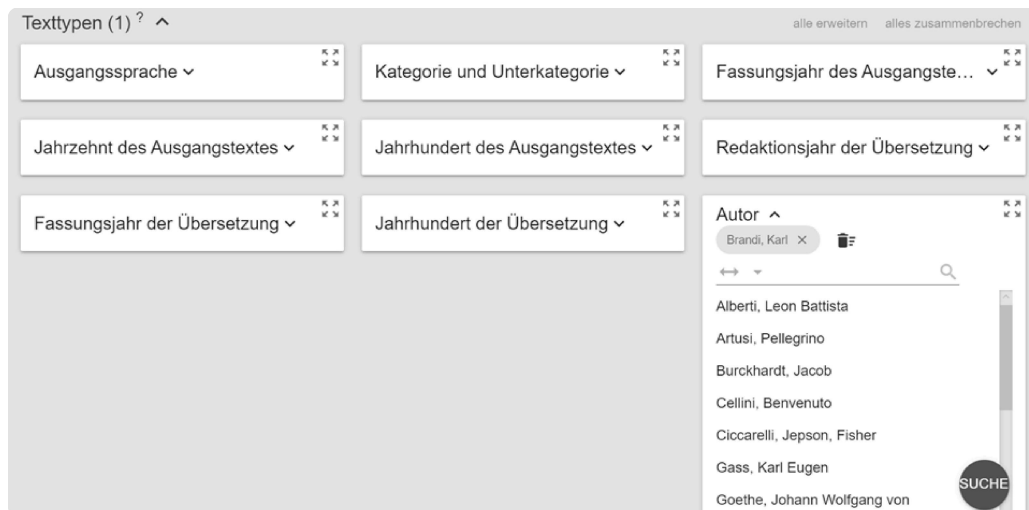


Рис. 3 - Поиск в немецком корпусе через окно "Text types" / «Типы текста».

Метаданные, с помощью которых можно фильтровать поиск конкордансов:

- Исходный язык: отображаются как язык текста, так и язык оригинала для переводных текстов;
- Язык перевода: позволяет искать все переводы на языке корпуса;
- Категория и подкатегория: указывает на различные типы текстов. Все тексты посвящены основной теме: художественному наследию и его лексике, в основном представленной через видение Флоренции и Тосканы, описанных с разных точек зрения. Выделены четыре макрокатегории текстов (Популярный, Специальный/Технический, Словарь и Литературный) и связанные с ними подкатегории (Популярный: Блог, Путеводитель, Журнал; Специальный/Технический: Архитектура, Искусство, Гастрономия и вино; Литературный: Биография, Художественная литература, Нехудожественная литература; Словарь: Одноязычный, Двужызычный / многоязычный). При определении этих категорий учитывались основное предназначение произведения и тип читательской аудитории, которому оно адресовано, данные, влияющие на тип используемого языка и уровень его специализации^[4];
- Автор: указываются фамилия, имя и указание «sa» / бз (без автора) при его отсутствии;
- Название и фрагмент: введение как полных текстов, так и фрагментов, соответствующих текстовой единице, и имеющих заголовки, такие как глава книги, полное письмо, статья в журнале и т. д. Этот выбор был сделан потому, что во многих случаях вся книга не совпадала с направлением проекта, но также для облегчения в будущем создания параллельных версий переводных текстов. В переводные тексты включены как оригинальные, так и переведенные названия;
- Год подготовки / год публикации / год перевода: хронологическая информация позволяет различать дату составления текстов (где возможно) и дату издания; для переведенных текстов была введена та же информация, что и для исходных^[5]. Для интернет-публикаций указывается дата обращения;

- Источник: позволяет выполнять поиск по отдельному документу корпуса (книге или фрагменту);
- Географическая зона^[6]: для текстов, в качестве объекта которых указан город или регион, вводится название города или региона. Это указание особенно важно для книг-путешествий и писем.

Более полные библиографические данные добавляются к этой информации при доступе к конкордансам путем нажатия на ссылку (имя файла, номер документа, имя автора и т. д. в соответствии с параметрами, заданными в *“View options”* / «Параметрах просмотра», рис. 4)

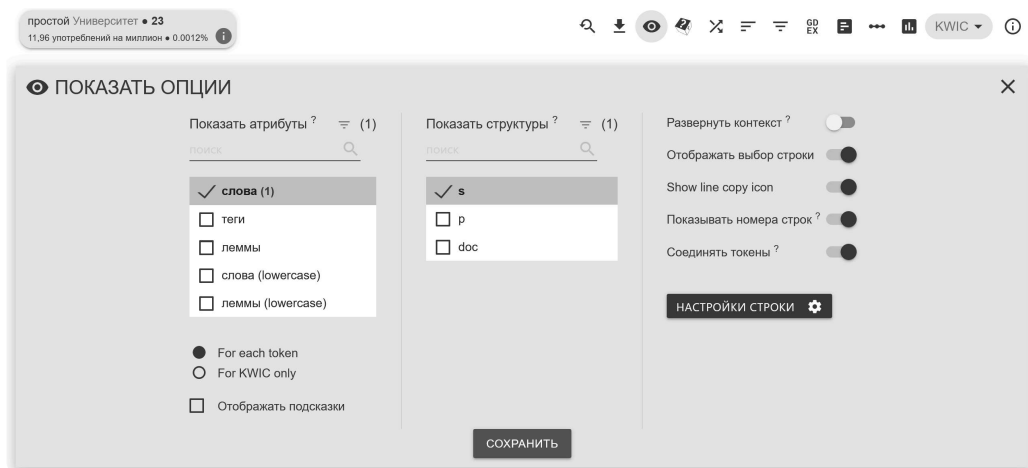


Рис. 4. - Доступные варианты просмотра текстовой ссылки в *“View options”* / «Параметрах просмотра».

Используя функцию *“Search”* / «Поиск», можно получить доступ к конкордансам, отображаемым в случайном порядке (по номеру документов), как на рисунке 5, в алфавитном порядке относительно рассматриваемого слова или его правого или левого контекста, используя функцию *“Sort left/right”* / «Сортировка влево / вправо» (рис. 6).

CONCORDANCIA Corpus LBC Español

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

Ordenar word

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...	nia, 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 áng
715	Alberti, Los tr...	e. Todo esto nos enseña que todas aquellas cosas 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 superiores 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...	ra, 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...	queños 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

Рис. 5. - Поиск конкордансов по лемме “pintar” в испанском корпусе без выбора порядка.

KONKORDANZZEILEN Deutsches LBC-Korpus

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

Sortieren word

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...	lte 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 d
57	Vasari, Leben d...	tan Giovanni dorthin kommen, und er arbeitete in der alten Kirche San Domenico, welche den Prädikanten-Mönchen gehört, €
58	Vasari, Leben d...	er Marter der heiligen Katharina darin darstellte. In der alten Kirche S. Domenico malte er auf einer Wand, wiederum in Fresko,
59	Vasari, Leben d...	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

Рис. 6. - Поиск конкордансов по лемме “Kirche” в немецком корпусе с заданным порядком слева от леммы.

Также можно осуществить поиск на наличие двух слов или лемм в одном и том же контексте на заданном расстоянии токенов, используя функцию “Context” / «Контекст» в меню “Search” / Поиск, как показано на рисунке 7, что позволяет, например, проверить засвидетельствованное использование различных словосочетаний (*dipingere a fresco* / *in fresco* на итальянском языке на рис. 8).

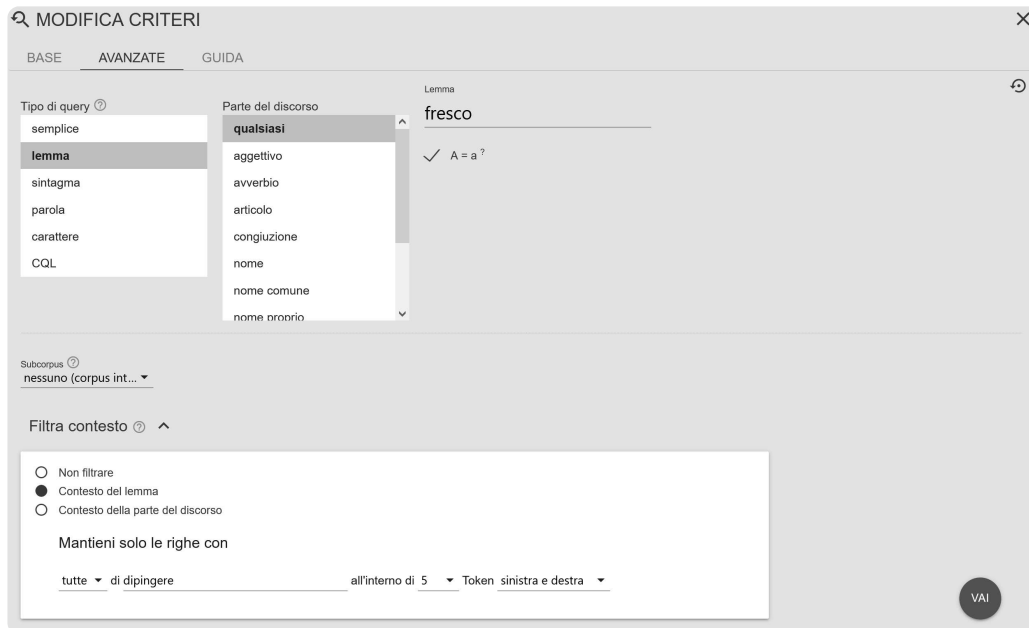


Рис. 7. - Поиск лемм *dipingere* и *fresco* на расстоянии 5 токенов в итальянском корпусе.

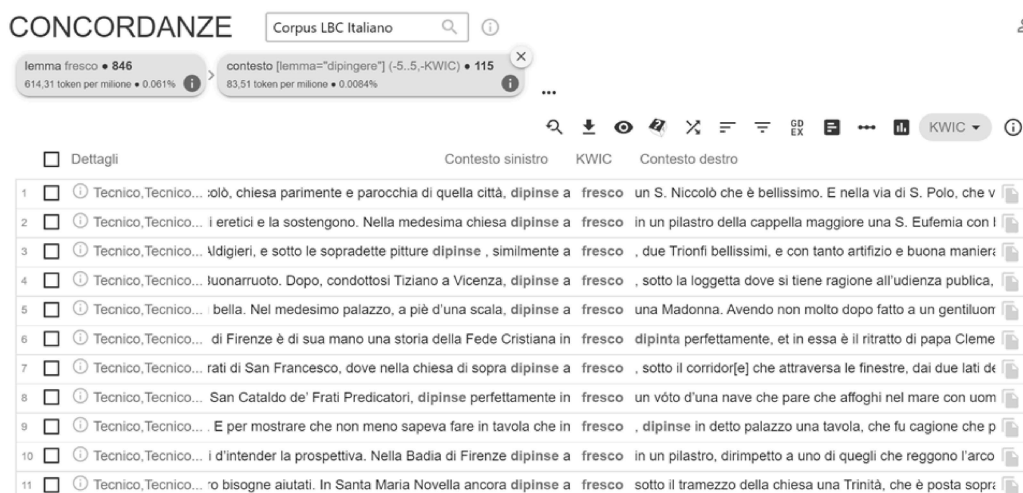


Рис. 8. - Конкордансы, относящиеся к поиску "*dipingere*" и "*fresco*" в том же контексте в итальянском корпусе.

Функция "*Word list*" / «Список слов» позволяет получать результаты о частотности лемм, присутствующих в корпусе, как по источникам, задавая в поиске, например, частотность лемм, относящихся к отдельному автору (рис. 9), так и по леммам корпуса (рис. 10—11).



Рис. 9. - Частотность токенов, присутствующих по авторам в русском корпусе.

WORDLIST

LBC English Corpus

BASIC **ADVANCED** ABOUT

find ?

- words
- lemmas**
- tags

- all
- starting with
- ending with
- containing
- matching regex
- from this list:

Exclude these words:

Include nonwords ?

A = a ?

Frequency min ? Frequency max ?

result format



- Simple list ?
- Display as ?

Subcorpus ?

none (the whole corpus)

GO

Рис. 10. - Поиск по Word list лемм, присутствующих в корпусе английского языка.

WORDLIST  

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	...

Рис. 11. - Результат поиска в списке слов по леммам, присутствующим в корпусе английского языка [ноябрь, 2022].

Реализация первого этапа по созданию наших корпусов достигла целей, которые мы поставили перед собой, создав необходимую основу для первых работ и исследований нашей группы (Carpì 2017; Farina, Billero 2018; Billero, Carpì 2018; Garzaniti 2020; Farina, Флинц 2020). Уже созданы первые леммари для каждого языка, снабженные конкордансами, извлеченными из корпусов, которые будут опубликованы на платформе к 2021 году и могут быть использованы для разработки будущих словарей.

Основная цель первого этапа работы, выполненного исследовательскими группами, заключалась в том, чтобы провести оценку корпусов с четким пониманием того, что только их фактическое использование позволит выявить проблемы, которые в противном случае остались бы скрытыми.

В будущем планируется расширить как количество языков (в настоящее время корпуса китайского, португальского и турецкого языков, участвующих в проекте LBC, все еще отсутствуют), так и количество текстов с уже описанной идеей большей однородности, чтобы попытаться сделать Корпуса максимально сопоставимыми.

Библиография

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.

Примечание

[1] Этот текст является переводом итальянского *Введения* к корпусу LBC, опубликованного на портале <http://corpora.lessicobeniculturali.net/it/> Перевод Н. Жуковой.

[2] Исчерпывающие сведения о корпусах LBC можно найти в разделе Публикации (*Farina, Nicolás Martínez, Billero 2020*).

[3] Название каждого типа текста, присутствующего в опции "*Text Type*" / «Тип текста», представлено пока только на итальянском языке, но вскоре будет переведено и на другие языки.

[4] На следующем этапе проекта классификация текстов будет пересмотрена в виду тех проблем, с которыми сталкиваются некоторые исследовательские группы при определении этой классификации. Некоторые тексты можно рассматривать как принадлежащие к большему количеству категорий, например, тексты классических авторов, чей стиль явно литературный, но созданные ими тексты могут считаться специальными благодаря темам и употребляемой лексике (например, «*Histoire de la Peinture en Italie*» Стендаля, в настоящее время относящаяся к категории нехудожественной литературы).

[5] Содержащиеся тексты относятся к периоду от эпохи Возрождения до наших дней. Хотя присутствуют обе даты, год публикации вторичен по сравнению с годом редактирования. Последние, по сути, представляют наибольший интерес для извлечения информации, поскольку они представляют лингвистические характеристики рассматриваемого периода. Фактически, тексты были внесены в базу данных, оставаясь верными используемому изданию, без какой-либо модернизации или исправления орфографии.

[6] Эта опция будет доступна с 2023 г.



Rossi, Valentina; Zhukova, Natalia. Русский корпус «Лексика культурного наследия»

© 2024 - Author(s) | Published by Firenze University Press

e-ISBN: 979-12-215-0313-5 | DOI: 10.36253/979-12-215-0313-5

Content license: CC BY-SA 4.0 International | Metadata license: CC0 1.0 Universal