Präsentation der LBC-Datenbank

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

1. Einleitung

Die LBC-Datenbank ist eines der Open Access-Instrumente, das von der Forschungseinheit *Lessico multilingue dei Beni Culturali (LBC)* entwickelt worden ist, um die Konsultation von Korpora zu ermöglichen. Spezifische lexikalische Informationen, die für die Durchführung lexikographischer und übersetzungswissenschaftlicher Untersuchungen notwendig sind, werden somit zur Verfügung gestellt.

Die Forschungseinheit möchte der Öffentlichkeit einen digitalen Raum mit verschiedenen Werkzeugen anbieten, um das Wissen über das künstlerische und kulturelle Erbe der Toskana auf internationaler Ebene zu verbreiten (Farina 2016).

Die Datenbank, die über die Projektplattform zugänglich ist, ermöglicht die Recherche in den Textkorpora verschiedener Sprachen (Englisch, Französisch, Italienisch, Russisch, Spanisch, Deutsch). Dort sind verschiedene Tools, die oben genannten Korpora sowie ihre Metadaten frei zugänglich[2].

In den Korpora sind Texte aus verschiedenen Textgenres enthalten, wie u. a. klassische literarische Werke, Reiseromane, Briefwechsel,

wissenschaftliche und fachliche Texte, Reiseführer und Handbücher, die eine große Zeitspanne umfassen. Die Primärquellen wurden mit Hilfe einer Software mit entsprechenden Funktionen strukturiert und verwaltet, um den Bedürfnissen verschiedener Benutzer gerecht zu werden. Hauptzielgruppen der Korpora sind dabei Linguisten, Literaten, Forscher in den Geistes- und Sozialwissenschaften, deren Arbeit Recherchen erfordert, um Informationen über den Wortschatz nach Autor, Datum, Textgenres- bzw. -typ usw. zu erhalten, Übersetzer, die spezifische lexikalische Ressourcen konsultieren müssen, und schließlich Spezialisten im Tourismussektor oder Touristen, die daran interessiert sind, ihre Kenntnisse über das Gebiet und die damit verbundene Kultur zu vertiefen.

Für jede Sprache des Projekts sind Texte enthalten, die dem Thema und dem Texttyp des Gesamtprojekts entsprechen. Für die Auswahl der Texte in der Originalsprache waren zwei Kriterien vorrangig: die anerkannte Autorität des Textes/Autors in der Ausgangskultur und seine Verbreitung (Billero, Nicolás 2017: 208) sowie die einfache Konvertierung in ein editierbares Format, wobei in der ersten Phase schwer zu digitalisierende Texte vermieden wurden. Für die Übersetzungen stützte sich die Auswahl auf eine von der Gruppe erstellte Liste, die italienische und anderssprachige Texte enthielt, die für die internationale Kenntnis des toskanischen künstlerisch-kulturellen Erbes als vorrangig betrachtet werden können: die grundlegenden Texte der Kunstgeschichte, die sich auf die Toskana beziehen, wie *Le Vite* von Vasari, die Architekturbücher von Alberti, Palladio, Serlio, einige Schriften von Machiavelli und Leonardo; bekannte Werke der Reiseliteratur, wie die Reisebeschreibungen von Stendhal und Ruskin, und kunstgeschichtliche Abhandlungen, wie die von Burckhardt.

Zum gegenwärtigen Zeitpunkt kann jedoch festgestellt werden, dass die verschiedenen Gruppen im Projekt die verschiedenen Texte nicht gleich priorisiert und auch in unterschiedlichen Anteilen behandelt haben, und zwar aus verschiedenen Gründen: Die Zugänglichkeit der Quellen ist von Land zu Land unterschiedlich und auch das Interesse am toskanischen Erbe, das je nach historischen Epochen und Textgenres in den verschiedenen im Projekt vertretenen Sprachen/ Kulturen nicht immer identisch ist. Aus diesen Beobachtungen wird die Heterogenität unter den Korpora verständlich, die wir allerdings in der zukünftigen Entwicklung des Projekts einschränken werden. Die Analyse der Verteilung der gewählten Texttypen in jedem Korpus und die Analyse der vertretenen Jahrhunderte am Ende dieser ersten

Phase wird in der Zukunft eine umfassendere Homogenisierung ermöglichen, so dass auch Textvergleiche leichter durchgeführt werden können. Zum jetzigen Zeitpunkt ermöglicht die prioritäre Einführung von Referenzwerken in der eigenen Sprache eine konsistente und ausreichende Basis für intralinguale Untersuchungen in den weiteren Sprachen.

Nach einer sorgfältigen Analyse der verschiedenen Software-Programme, mit denen die Korpora konsultiert werden können, fiel die Wahl auf *NoSketchEngine* (Billero, 2020), da es für die Zwecke des Projekts verschiedene interessante Funktionen anbieten kann, wie die Suche nach Konkordanzen und ihre Sortierung auf der Basis verschiedener Merkmale.

Informationen über jedes Korpus können aus der Sektion „Corpus-Info“ entnommen werden, die im Menü *NoSketchEngine* enthalten ist (Abbildung 1).

Abbildung 1 – Detaillierte Informationen über das französische Korpus unter „Corpus info“

Auf dieser Seite finden sich auch Angaben in Prozent zu den Anteilen der unterschiedlichen im Korpus enthaltenen Texttypen (vgl. Abbildung 2 zum englischen Korpus):

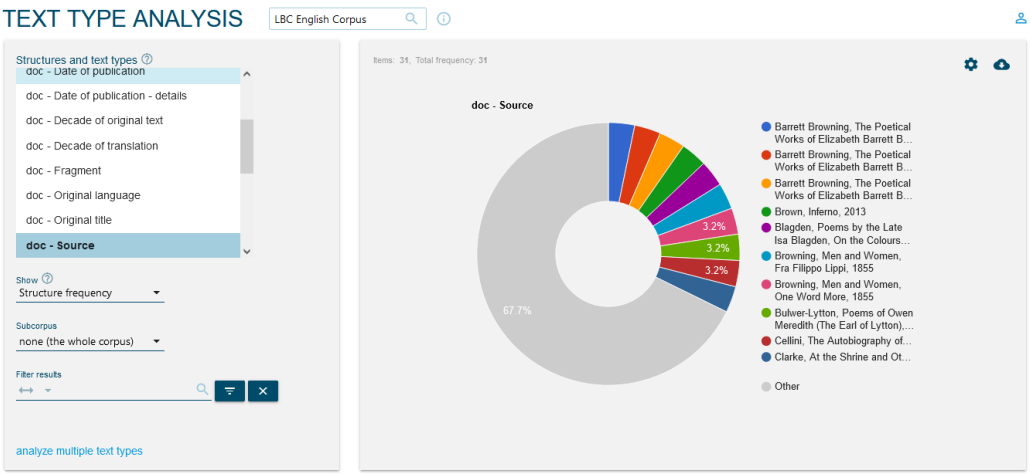


Abbildung 2 – Struktur und Merkmale der Dateien im englischen Korpus

Die Struktur der Korpora folgt den traditionellen Regeln mit Berücksichtigung von gemeinsamen Kriterien für die Metadatenverwaltung. Vgl. die Suche nach Texttypen („Text types“, Abbildung 3):

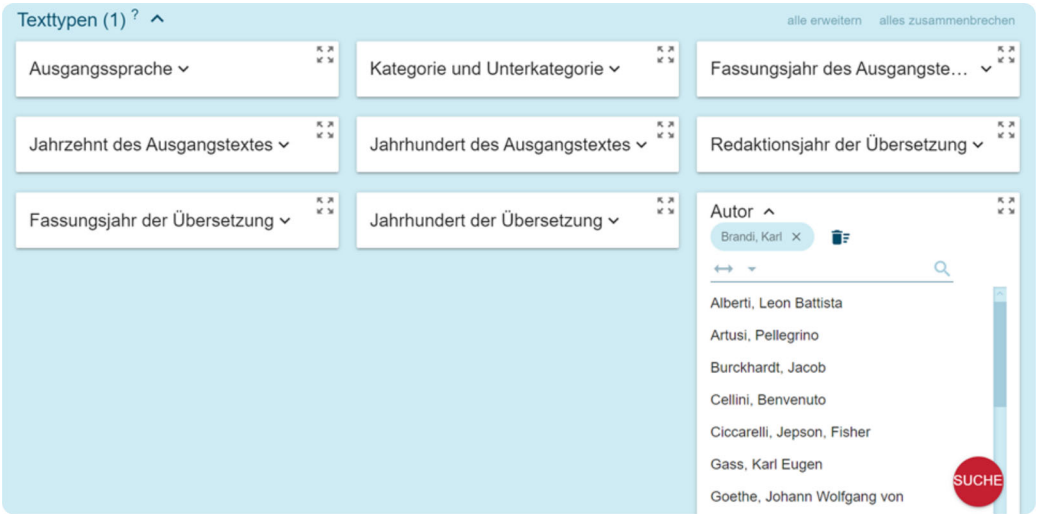


Abbildung 3 – Suche im deutschen Korpus über das Fenster „Texttypen“.

Die Metadaten, mit denen die Konkordanzzeilen gefiltert werden können, sind:

- die Originalsprache: Es wird sowohl die Sprache des Textes als auch die Ausgangssprache des Textes im Falle einer Übersetzung angegeben;
- die Übersetzungssprache: Die Suche in allen vorhandenen Übersetzungen wird ermöglicht;
- die Kategorie und die Unterkategorie: Die verschiedenen

Texttypen werden angegeben. Alle Texte haben die Kulturgüter und ihr Lexikon zum Gegenstand, insbesondere eine umfassende Vision von Florenz und der Toskana, die aus verschiedenen Blickwinkeln beschrieben wird. Dabei werden vier Makrokategorien unterschieden (populärwissenschaftlich, technisch/fachlich, lexikographisch und literarisch), die jeweils in Unterkategorien aufgeteilt sind (populärwissenschaftlich: Blogs, Reiseführer, Zeitschriften; technisch/fachlich: Architektur, Kunst, Weingastronomie; literarisch: Biografie, Fiktion, Essays; lexikographisch: einsprachige, zweisprachige/mehrsprachige Ressourcen). Diese Kategorien entsprechen dem Hauptziel des Projektes und den Benutzertypen, an die es sich richtet. Es handelt sich um Informationen, die die Art der verwendeten Sprache und den Grad der Spezialisierung beeinflussen[3];

- Autor: Nachname und Vorname werden angegeben. Wenn nicht vorhanden, findet man die Angabe „*sa*“ (ohne Autor);
- Titel und Fragment: Es wurden sowohl die Einleitung von ganzen Texten als auch von Fragmenten, die einer Texteinheit entsprechen (z.B. ein Buchkapitel, vollständige Briefe, Zeitschriftenartikel usw.) ausgewählt. Diese Wahl wurde getroffen, weil in vielen Fällen nicht das gesamte Buch für das Projekt von Interesse war, aber auch, um die zukünftige Erstellung von Parallelversionen zu erleichtern. Bei Übersetzungen wurden sowohl der Originaltitel als auch der Titel der Übersetzung angegeben;
- Redaktionsjahr / Erscheinungsjahr / Jahr der Übersetzung: Die zeitlichen Angaben unterscheiden zwischen dem Datum der Abfassung der Texte (wenn vorhanden) und dem Veröffentlichungsdatum; bei übersetzten Texten wurden die gleichen Informationen sowohl für den Originaltext als auch für den übersetzten Text angegeben. Bei Online-Publikationen ist das Abrufdatum angegeben;
- Quelle: Ermöglicht die Suche in einem einzelnen Dokument des Korpus (Buch oder Fragment);
- Geografische Abgrenzung: Bei Texten mit einer definierten Stadt oder Region als Thema wurde der Name der Stadt oder der Region eingegeben. Dieser Hinweis ist vor allem bei Reisebüchern und Korrespondenzen vorhanden.

Vollständige bibliographische Angaben werden beim Zugriff auf die Konkordanzen durch Anklicken der Metadaten visualisiert (Dateiname, Dokumentnummer, Autorenname usw. auf der Basis der ausgewählten „View options“, vgl. Abbildung 4).

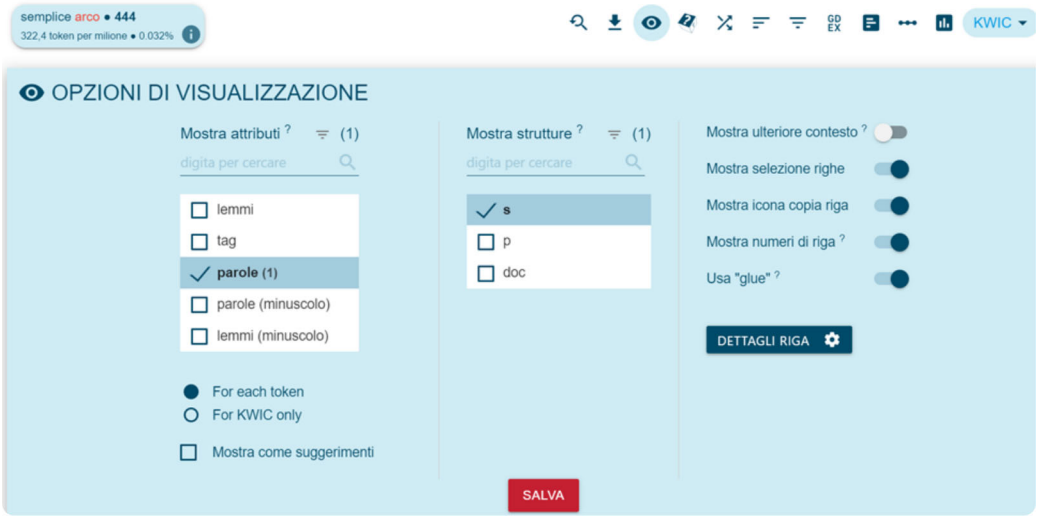


Abbildung 4 – Verfügbare Optionen für die Auswahl von Textinformation unter „View options“.

Über die Funktion „Search“ wird der Zugriff auf die Konkordanzen ermöglicht: Sie können in zufälliger Reihenfolge (nach der Anzahl der Dokumente) wie in Abbildung 5 angezeigt werden, oder in alphabetischer Reihenfolge in Bezug auf das betreffende Wort oder nach ihrem rechten oder linken Kontext über die Funktion „Sortieren nach links/rechts“ (Abbildung 6).

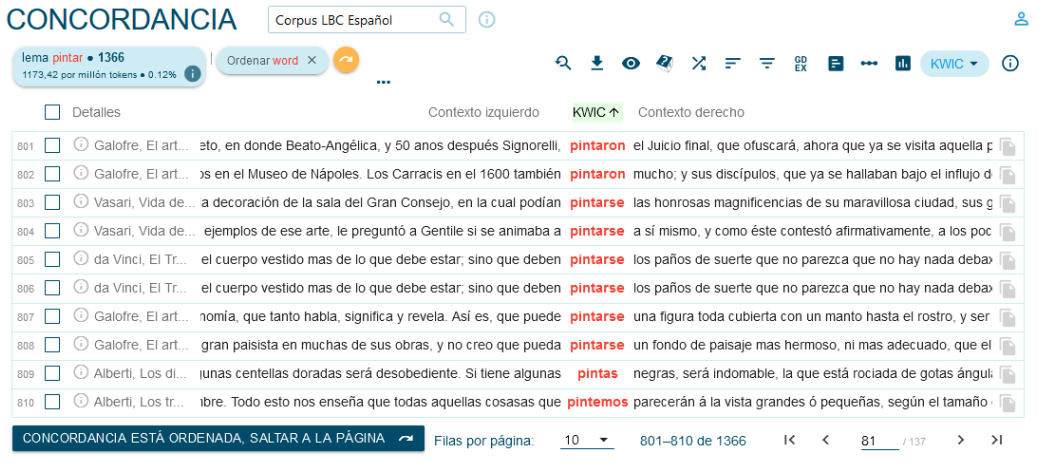


Abbildung 5 – Suche nach den KWICs des Lemmas *pintar* im spanischen Korpus ohne Auswahl der Reihenfolge.

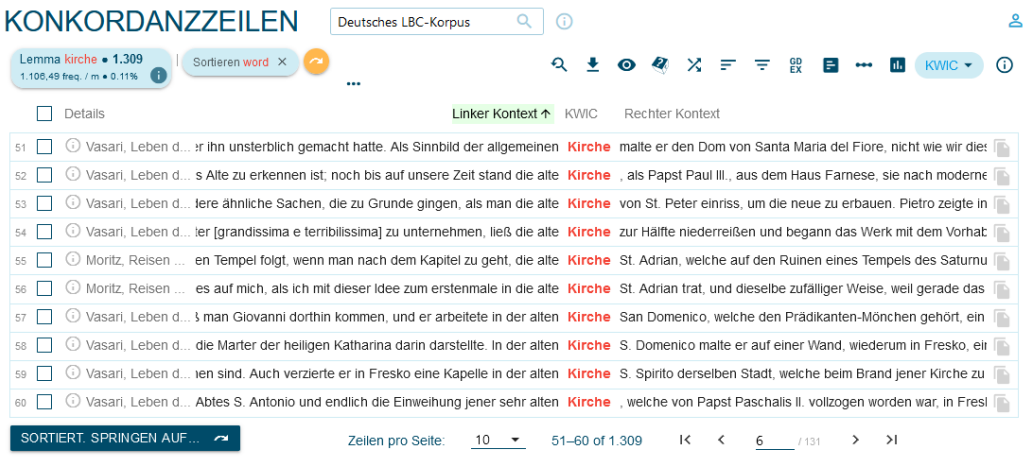


Abbildung 6 – Suche nach den KWICs des Lemmas *Kirche* im deutschen Korpus. Filter ‚nach links'.

Es ist auch möglich, nach dem Vorhandensein von zwei Wörtern oder Lemmata im gleichen Kontext in einem gewählten Abstand von Tokens zu suchen, indem man die Funktion „Context" im Menü „Search" verwendet, wie in Abbildung 7, so dass beispielsweise die Verwendung verschiedener Kollokationen überprüft werden kann (*dipingere a fresco / in fresco* im Italienischen, vgl. Abbildung 8).

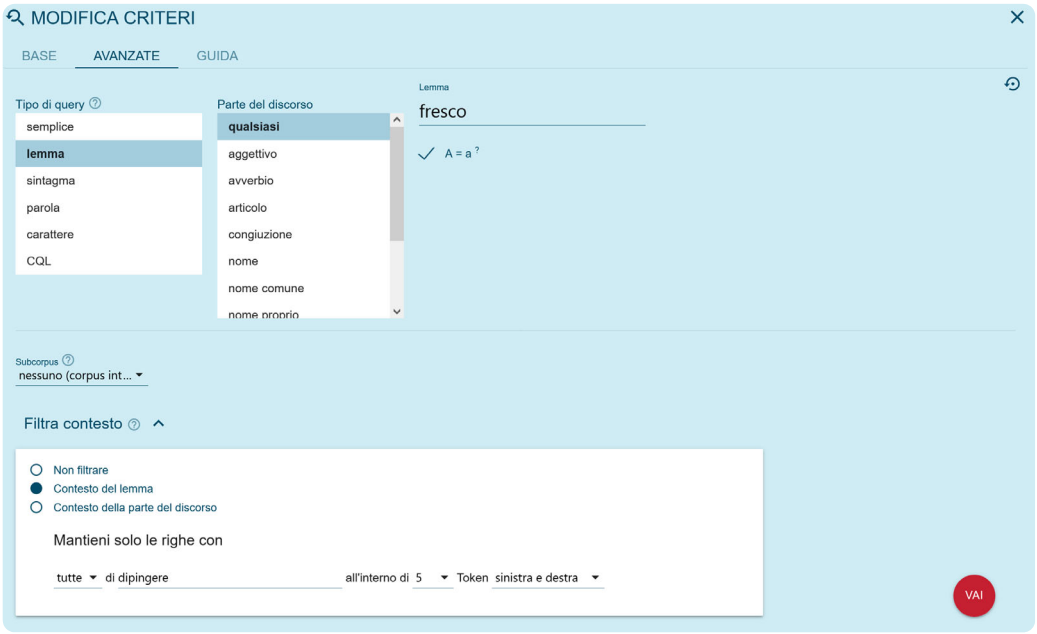


Abbildung 7 – Suche der Lemmata *dipingere* e *fresco* im italienischen Korpus. Range: 5 tokens.

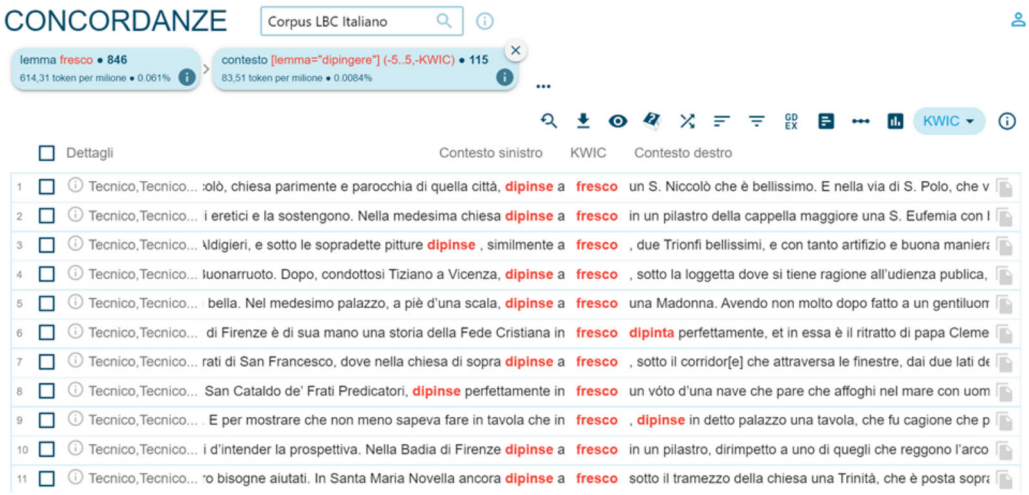


Abbildung 8 – KWICs von *dipingere* e *fresco* im gleichen Kontext im italienischen Korpus.

Die Funktion „Word List“ ermöglicht es, Informationen über die Häufigkeit von bestimmten Elementen in einem Korpus zu entnehmen (Abb. 10-11): Es können sowohl die häufigsten Lemmata in einem Korpus herausgefiltert werden als auch die häufigsten Lemmata pro Autor (Abbildung 9).

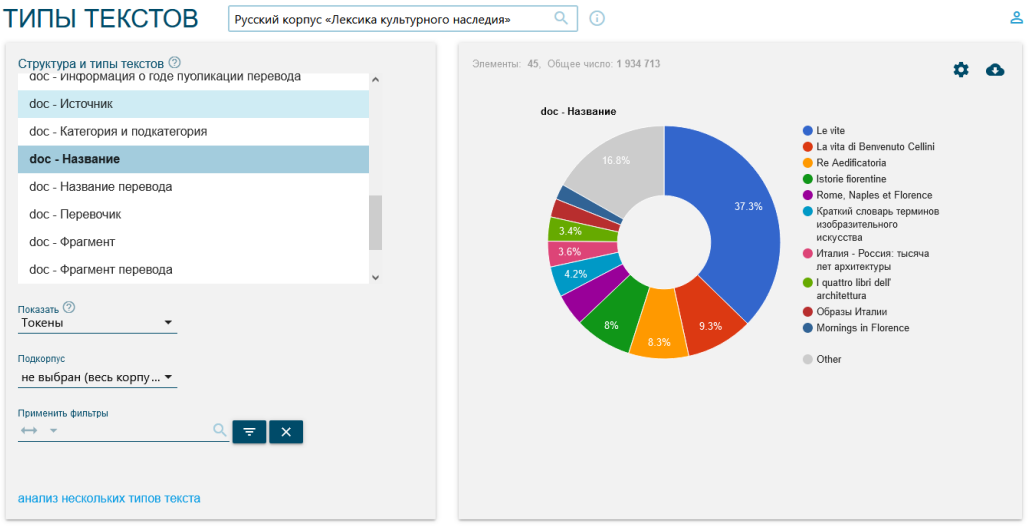


Abbildung 9 – Frequenz der Token pro Autor im russischen Korpus.

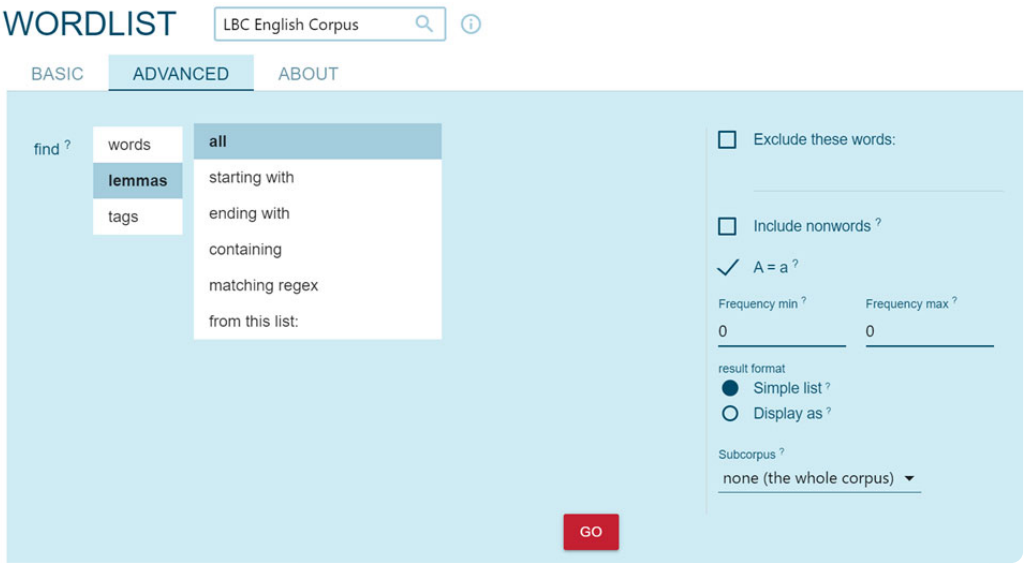
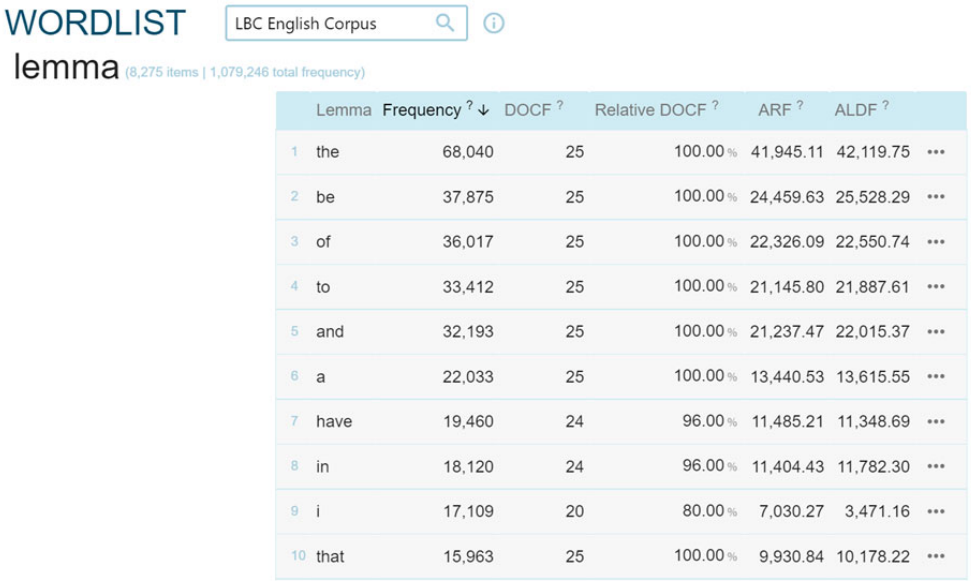


Abbildung 10 – Extraktion einer Word List nach Lemmata im englischen Korpus.



	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	•••

Abbildung 11 – Word List nach Lemmata im englischen Korpus.

Nach dem Abschluss dieser ersten Phase der Arbeit an den Korpora lässt sich eine positive Bilanz ziehen, denn durch sie konnte die notwendige Grundlage für die ersten Arbeiten und Untersuchungen unserer Gruppe geschaffen werden (Carpi 2017; Farina, Billero, Carpi 2018; Garzaniti 2020; Farina, Flinz 2020). Die ersten provisorischen Stichwortlisten jeder Sprache wurden bereits erstellt, zusammen mit den aus den Korpora extrahierten Konkordanzen, die noch vor 2022 auf der Plattform veröffentlicht werden sollen und für die Fertigstellung künftiger Wörterbücher verwendet werden können.

Das Hauptziel dieser ersten Arbeit, die von jeder Sprachgruppe durchgeführt wurde, bestand darin, eine Validierung der Korpora durchzuführen, mit dem Wissen, dass nur ihre tatsächliche Verwendung Probleme aufdecken würde, die ansonsten latent bleiben würden.

Für die Zukunft ist geplant, sowohl die Anzahl der Sprachen (derzeit gibt es noch keine Korpora der ebenfalls am LBC-Projekt beteiligten Sprachen Chinesisch, Portugiesisch und Türkisch) als auch die Anzahl der Texte mit der bereits beschriebenen Idee der Homogenisierung zu erweitern, um zu versuchen, die Korpora so vergleichbar wie möglich zu machen.

Literatur

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.

Fußnoten

[1] Der vorliegende Text ist eine von Carolina Flinz, Anna Nissen und Sabrina Ballestracci angefertigte Übersetzung der italienischen Einleitung zu den LBC-Korpora (vgl. http://corpora.lessicobeniculturali.net/it/).

[2] Für umfassende Daten zu den LBC-Korpora vgl. Farina, Nicolás Martínez, Billero 2020.

[3] In der nächsten Phase des Projektes soll die jetzige Textklassifikation überprüft werden, da sich das Problem gezeigt hat, dass einige Textexemplare zu mehr als einer Kategorie zugeordnet werden können. Texte von klassischen Autoren, deren Stil eindeutig literarisch ist, können, wenn sie sich mit bestimmten Themen befassen, auch als Fachtexte angesehen werden. Ein Beispiel dafür ist Stendhals *Histoire de la Peinture en Italie*, die zum jetzigen Zeitpunkt als Literatur/Fiktion klassifiziert worden ist.

e-ISBN: 979-12-215-0311-1 | DOI: 10.36253/979-12-215-0311-1