EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki

Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

LBC-Dictionary: a Multilingual Cultural Heritage Dictionary. Data Collection and Data Preparation

Farina A.[1], **Flinz C.**[2]

[1] *University of Florence, Italy*
[2] *University of Milan, Italy*

Abstract

An increasing number of a wide variety of texts on Italian cultural heritage are available today, both online and on paper. However, there are no specific tools (dictionaries, reference materials on technical translations) that can train and support specialists involved in cultural tourism. Mainly focusing on Florence and its cultural heritage, the LBC project (Farina 2016) will try to fill this gap by providing tools for those who have to write/translate for dissemination in various languages: in a first step by building monolingual corpora (English, French, German, Italian, Russian, Spanish) that the user can freely search; in a second step by developing a plurilingual LSP internet dictionary on cultural heritage which uses the above-mentioned corpora as a primary source. The aim of this paper is to give an insight in the lexicographical process of the LBC-Dictionary, concentrating in particular on data collection and data preparation, which, as is usual for dynamic dictionaries, are open-ended and ever ongoing (Klosa 2013). In particular, we will illustrate the main characteristics of the French and German LBC Corpora and reflect on the provisional French and German entry list, also illustrating the procedure adopted, an alternation of corpus-driven and corpus-based steps (Tognini-Bonelli 2001), for their extraction.

Keywords: Corpora; cultural heritage; internet dictionary

1 Introduction[1]

An increasing number and a wide variety of texts on Italian cultural heritage are available today, both online and in print, from tourist guidebooks to museum web sites, from art catalogues to critical essays. Provided in different languages, these works attempt to satisfy an international public increasingly in need of information on Italian cultural heritage. However, at present across Europe, there are no specific tools (dictionaries, reference materials on technical translations) that are able to convey such knowledge in an appropriate way (Billero/Nicolas Martinez 2017: 203), or specialised institutions that can train and support specialist translators and other specialists involved in cultural tourism (tourist guides, tourist information centres, museum staff, etc.).

The LBC project (Farina 2016), which involves experts from different disciplines (among others lexicography, corpus linguistics etc.) and universities (Florence, Bologna, Lisbon, Milan, Paris, Pisa etc.), tries to fill this gap mainly by focusing on Florence and its cultural heritage and providing tools for those who have to write/translate for its dissemination in the various languages. Our principal aim is to create monolingual dictionaries of Italian Heritage in all the languages involved in our project, which could be used as plurilingual tools thanks to translation links created among them.

In a first step, we have built monolingual comparable corpora (English, French, German, Italian, Russian, Spanish, see Figure 1)[2] that could also serve the principal target user of our dictionaries (persons who must write or translate texts about Tuscan Cultural Heritage). We decided not to set limits of time and place, but to use each text by referring to the cultural heritage of the city of Florence[3] in each language featured in the project:

> …the city of Florence as it has appeared in the actual use of language over the centuries and in the discourses of the people who have described it in the seven languages featured in the project, thinking that in this way the cultural basis might emerge and lead us to design lexicographical articles which shed light on the cultural and historical connotations of the words actually used to describe it. (Farina 2015: 125)

[1] The present contribution was conceived jointly by the two authors and discussed in detail in its individual parts, in particular Annick Farina the French section, and Carolina Flinz the German one.
[2] In a next step, the corpus platform will also contain parallel corpora (see Zotti 2017).
[3] Its extension to other cities and to the entire region of Tuscany is in the planning stage.

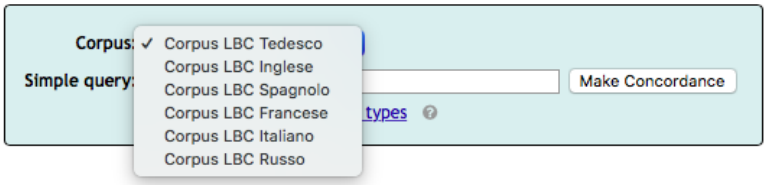

Figure 1: Screenshot of the available corpora (04.2.2020).

The corpora make use of open-source *no-sketch engine*[4] software, which provides users with several functionalities (including the search for a word in its context, the sorting of references according to various criteria, the filtering of texts according to text types and authors) and thus enables intra- and interlingual applications, from both a synchronic and diachronic perspective (see Ballestracci/Buffagni/Flinz in prep.).

As a second step, these corpora will be used as the primary source (Engelberg/Lemnitzer 2009: 235-237) for the plurilingual LSP internet dictionary on cultural heritage, which is currently in the planning stage (Farina/Billero 2018): with an interplay of automated procedures and manual selection/interpretation (Geyken/Lemnitzer 2016: 208), the data will be extracted from the above-mentioned corpora to construct the provisional entry list.

The aim of this paper is to provide an insight into the lexicographical process of the LBC-Dictionary, by concentrating in particular on data collection (1) and data preparation (2), which, as usual for dynamic dictionaries, are open-ended and ever ongoing (Klosa 2013). We will:

(1) concentrate in particular on the LBC French and German Corpora, illustrating their main characteristics: size, text types (among others popular, technical, literary texts), time period involved (from the Renaissance to the present), authors etc.;

(2) focus on the provisional French and German entry list, also illustrating the procedure adopted, an alternation of corpus-driven and corpus-based steps (Tognini-Bonelli 2001) for their extraction. The main characteristics of the list will be presented in order to reflect on their items.

We will conclude by proposing perspectives, for example the extraction of the concordances related to headword lists.

2 Data collection

In the data collection phase, the sources for the dictionary base have been compiled. Corpora are the primary sources[5] of many contemporary dictionaries (Klosa 2020: 11), and their use in the lexicographical process, mostly with a quantitative-qualitative approach, has opened up a variety of new possibilities (Lemnitzer/Zinsmeister 2015: 170) that were previously unthinkable with traditional collections of documents, and opportunities impossible with any other type of source, since they are accessible regardless of location and provide an authentic picture of the language depicted (Geyken/Lemnitzer 2016: 203).

Even if we were advised to use existing corpora for a variety of reasons (i.e. they realise the criteria of size and representativeness, see Lemnitzer/Zinsmeister 2015: 137), in our project we could only partially follow this procedure because there are no existing LSP-Corpora on art and it was impossible to create virtual LSP-Corpora from existing ones. So, we decided to adopt a combined procedure of using both *ad hoc* created monolingual LSP-Corpora and existing reference corpora[6]. As secondary sources we used existing monolingual and bilingual dictionaries (among others TLFi, Duden online, Zanichelli 2009) and, as tertiary sources, manuals and grammars.

2.1 LBC French and German Corpora

The creation of corpora is associated with methodological problems that, however, can be solved with careful planning (see Flinz 2019; Hunston 2008; Lemnitzer/Zinsmeister 2015). For example, among others: 1. The choice of the type of corpus must be carefully considered, since not all corpus types are suitable for all lexicographical purposes; 2. The requirements of the corpus, as for example its size (see Kupietz/Schmidt 2015: 302f), must be taken into consideration, because the larger a corpus is, the higher the probability of finding rare constructions or obtaining good results from statistical analyses will be (see Geyken 2007: 37); 3. The origin and quality of the texts, which should not be chosen

[4] While our researches depend on *public found (public Universities)* we publish all tools in Open Access: corpora can be freely searched by its above-intended users. See the LBC-Platform, http://corpora.lessicobeniculturali.net (04.02.2020).
[5] For the division in primary, secondary, tertiary sources see Wiegand 1998: 140.
[6] See §2.1

arbitrarily[7]; 4. The documentation of primary sources to ensure the value of a corpus.
In our project the basic jointly determined criterion for selecting works and authors was their importance for Florentine Renaissance art and culture, considering both translations and original texts: The German and French LBC corpora in fact - like the other LBC corpora - consist of original language texts as well as texts translated from the other languages of the project (Italian, French and English). In their nature as monitor corpora (Lemnitzer/Zinsmeister 2015: 140) they can be constantly expanded, so what we present in this paper is only a snapshot of the actual situation, but when we decided to extract the provisional entry list, we fixed a minimum of 1,000,000 words (Table 1):

	French LBC-Corpus	German LBC-Corpus
Tokens	3,818,747	1,183,484

Table 1: Size of the French and German LBC-Corpora

Each LBC-Corpus contains texts that belong mainly to two macro categories: technical and literary texts[8] (see Figure 2).

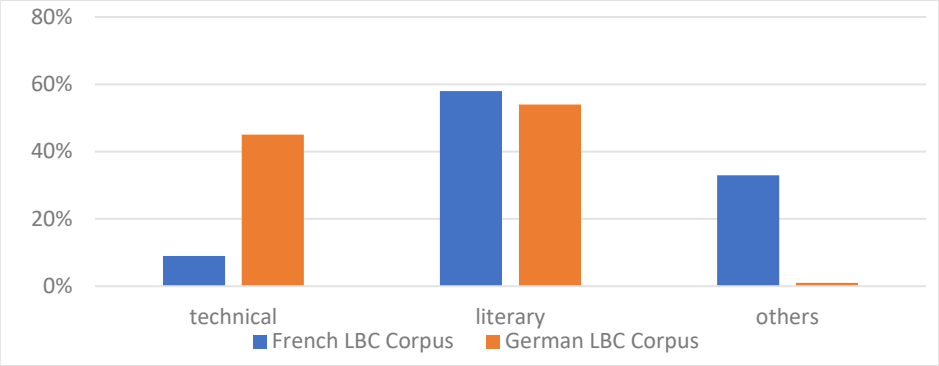

Figure 2: Diagram of the French and German LBC-Corpora

As we can see from the diagram, the most important difference is the higher number of technical texts in the German Corpus and of the category 'others' in the French Corpus. In this latter group we can find mainly informative texts. The aim of the research group is now to integrate texts into the corpus of any language involved, especially in the above-mentioned macro-categories, in order to increase their comparability (see Billero/Farina/Nicolas in prep.).
Among the literary texts, we can find different texts genres, such as biographies (for example, travel diaries and travelogues), fictional narrative works (i.e. short stories, novels and other literary narrative texts with Florence or Tuscany as their main theme or locale) and essays. The technical texts comprise LSP-texts focusing especially on art and on architecture. The texts are mostly original ones. Our purpose was in fact to guarantee a certain variety for all text types, while giving preference to those that were particularly representative of Italian culture and art and its international dissemination and reception[9]. Concerning the involved authors[10] and the diachronic variation, we decided to offer a certain variety: 74 authors[11] for French and 16 authors[12] for German. The time laps covered by the texts (date of writing

[7] See also Farina/Billero 2018 for the semi-automatic evaluation of text translations.
[8] For a detailed description of the corpora see Farina in prep. for the French LBC Corpus and Ballestracci/Buffagni/Flinz in prep. for the German LBC Corpus.
[9] As for example: Giorgio Vasari's *Vite* (1550, 1568) - a work that is fundamental for the art and culture of the Renaissance, and that contributed to spread the myth of the Italian Renaissance in most European countries; non-Italian authors, who played a major role in spreading the Italian Renaissance culture in foreign countries (John Ruskin, Jacob Burkhardt); and famous authors or writers who travelled to Italy and written about it (among others Dumas and Stendhal for the French Corpus and Johann Wolfgang Goethe, Karl Philipp Moritz for the German one).
[10] We have mentioned only the authors belonging to the literary and technical field.
[11] Allais, Auzias, Bard, Bazin, Beaugrand, Bertheroy, Brosses, Camus, Cellini, Chateaubriand, Colet, Colin, Colombari, Creuzé de Lesser, da Vinci, De la Borie, De Navenne, Delacroix, Dufay, Dumas, Erdan, Faudre, Favre le Bret, Félibien, Fernandez, Feuillet, France, Fréville, Gaboury, Gautier, Giono, Goncourt, Goupil, Grandgeorge, Grimaldi, Jaucourt, Klaczko, La Sizeranne, Labourdette, Lafenestre, Lang, Le Routard, Lescure, Libri, Machiavelli, Mallarmé, Mallet, Maurel, Méry, Meyer, Michel, Michel-Ange, Montaigne, Moran, Musset, Nobecourt, Palustre, Pasquin, Perrot, Pommier, Powell, Prieur, Renan, Revel, Rosov, Schmitz, Staël-Holstein, Stendhal, Taillasson, Taine, Vasari, Viollet Le Duc, Wyzewa.

in both languages or date of translation) is from the 16th to the 21st century.

3 Data preparation

Central steps in the data preparation phase are extracting the provisional dictionary entry list and modelling the lexicographical data into a database structure. In our paper we concentrate on the first aspect by illustrating the alternation of the corpus-driven and corpus-based procedures (Tognini-Bonelli 2001) used for extracting the list.

3.1 LBC French and German entry list

We could not create the provisional lemma list on existing lemma lists of other dictionaries, since no lexicographic resources of this type exist, so instead we used a combination of different corpora as our primary source:
- *ad hoc* created monolingual LSP-Corpora (see table 1);
- reference corpora of the involved languages: for French we used *L'Est Républicain* (15,000,000 tokens) and in the Sketch Engine integrated *French Web 2017* (frTenTen17) with 6,845,630,573 tokens; for German we chose das *Deutsche Referenzkorpus DeReKo* (2017-I, Release of 08.03.2017) and the in the *Sketch Engine* integrated *German Web 2013* (deTenTen13) with 19,808,173,163 tokens.

We first extracted automatically and manually different types of word lists for each language (see Table 2):

	List-Name	Corpora	Measure	Automatic/Manual	N.
French	K-LBC (fr)	LBC-Korpus (fr) / frTenTen17	keyness Score	automatic	2000 single units 2000 multiple units
	L-LBC (fr)	LBC-Korpus (fr)	absolute frequency	automatic	25,337
	K-L-RIF (fr)	L'Est Républicain	absolute and relative frequency	automatic	145,644
	G-LEX (fr)	Dictionaries		manual	1806
German	K-LBC (de)	LBC-Korpus (de) / deTenTen13	keyness Score	automatic	2000 single units 2000 multiple units
	L-LBC (de)	LBC-Korpus (de)	absolute frequency	automatic	45,029
	K-L-RIF (de)	LBC-Korpus (de) / DeReKo	*chi2* e *LLR*	automatic	10,402
	G-LEX (de)	Dictionaries		manual	2547

Table 2: Size of the French and German LBC-Corpora

1) a keyword list based on the reference corpora integrated in the *Sketch Engine* (K-LBC). The K-LBC Lists were automatically driven by using the function *Keywords* of *Sketch Engine*. We extracted 2000 keywords and 2000 multi-words expressions (see Figure 3 and 4, which show the first ten German single and multiword keywords), representing the most typical items of both corpora.

[12] Alberti, Brandi, Burckhardt, Cellini, da Vinci, Gass, Goethe, Heine, Kurz, Machiavelli, Moritz, Ruskin, Stendhal, Vasari.

	Word	Frequency? Focus	Frequency? Reference	Relative freq.? Focus	Relative freq.? Reference	Score ?
1	verfertigen	512	6,194	426.647	0.313	325.78
2	disegno	348	239	289.986	0.012	287.52
3	Florenz	1,475	79,843	1,229.109	4.031	244.52
4	Medici	495	14,431	412.481	0.729	239.21
5	Florentiner	373	7,634	310.819	0.385	225.08
6	Cosimo	342	6,170	284.987	0.311	218.06
7	Filippo	367	8,189	305.819	0.413	217.08
8	Lorenzo	750	37,771	624.971	1.907	215.34
9	diligenza	249	3	207.49	0	208.46
10	florentinisch	264	3,629	219.99	0.183	186.77

	Word	Frequency? Focus	Frequency? Reference	Relative freq.? Focus	Relative freq.? Reference	Score ?
1	Santa Maria	261	0	217.49	0	218.49
2	Mutter Gottes	115	0	95.829	0	96.83
3	großer Teil	111	0	92.496	0	93.5
4	anderes Ding	106	0	88.329	0	89.33
5	Maria Del	91	13	75.83	0.056	72.75
6	kleine Figur	86	0	71.663	0	72.66
7	ganzes Werk	82	0	68.33	0	69.33
8	ausgezeichneter Maler	75	0	62.497	0	63.5
9	Vasari Giorgio	69	0	57.497	0	58.5
10	Leben der Ausgezeichnetsten Maler	69	0	57.497	0	58.5

Figure 3 and 4: Screenshots of first 10 keywords of the K-LBC (de) (single and multi-words expressions)

As typical items of both corpora we can see LSP-Words (*assise*/base; *verfertigen*/produrre), Italian words (*disegno*/drawing; *diligenza*/diligence for German LBC), the name of the city of Florence (*Florence-Florenz*) and the connected adjective (*florentin*/Florentine; *florentinisch*/Florentine), proper names (e.g. Medici, Cosimo, Filippo, Lorenzo, Giotto). But in going down the list we also find obsolete or sophisticated words, such as *brasses*/fathom (obsolete) in K-LBC (fr) and *woselbst*/where (sophisticated) or *heutetags*/today (obsolete) in K-LBC (de). The extracted multiword expressions again show proper names of artists and monuments (Santa Croce; Vasari Giorgio) or possible candidates for collocations (*marbre blanc*/white marble; *grande chapelle*/big chapel; *Mutter Gottes*/Blessed Virgin Mary; *ausgezeichneter Maler*/excellent painter). Differences can be seen in the greater presence of Italian words considered as Keywords in the German Corpus.

2) a lemmatized frequency list (L-LBC). The L-LBC Lists were automatically extracted through the functionality *Word List*[13] of *Sketch Engine*. We decided to set the minimum frequency of x >1, since not only the most frequent terms, but also the terms recurring only once (i.e. the *hapax legomena*) could be of interest for our final entry list. Both lists, L-LBC (fr) and L-LBC (de) have in common the fact that articles, conjunctions, prepositions, auxiliary verbs etc. occupy the first positions, while LSP-Terms (*art*/art; *peinture*/picture; *artiste*/artist; *Skulpturensammlung*/sculpture collection; *Marzocco-Löwen*/Marzocco-Lions; *unpoliert*/unpolished) occupy the lowest ones. The same holds true for Italian proper names (Cosimo, Medici, Bargello etc.) and Italian words (among others *palazzo*/building, *loggia*/lodge in the French one and *non-finito*/not finisched, *chiesa*/church in the German one). Even if there are many similarities, we also note some differences, such as the greater incidence of French Equivalents of Italian names (Michel-Ange, Médicis, Raphaël).

3) a keyword list based on the reference corpora of the languages involved (K-L-RIF). First, for French a reference list was extracted from *L'Est Républicain* (L-RIF) (fr) by using the *AntConc* software. The list obtained of lemmatized forms arranged by frequency (145.644) was then compared with our L-LBC (fr), calculating their relative frequency. For German we used a slightly different procedure: the K-L-RIF (de) was automatically extracted[14] by comparing our *ad hoc* compiled LBC-Corpus (de) and *DeReKo* with the reference corpus for the German language. *DeReKo* is the world's largest linguistically motivated collection of electronic corpora for German and contains different types of corpora from the present and the recent past, corresponding to different types of texts (including articles from daily newspapers and magazines, literary texts, specialised texts)[15]. The result of the procedure was an excel-list[16] (figure 5), whose ranking can be changed on the basis of two statistical measurements ($\chi 2$ und LLR, see Dunning 1993).

[13] We chose the option ‚Lemma'.

[14] At his point we want to thank the *Leibniz-Institut für Deutsche Sprache* and in particular Rainer Perkuhn for his support. The use of IDS internal tools was fundamental for the comparison between the two German corpora.

[15] See https://www1.ids-mannheim.de/kl/projekte/korpora/ (02.02.2020)

[16] For this procedure see also Flinz/Perkuhn 2018: 962.

		List N.2	Only G-LEX			
rank	dereko-iso	LBC-Lemma-iso	winner	llr	chi2	
1.	6176	10080	LBC-Lemm	185.688,89	181.550.018,58	seine
2.	3172	3951	LBC-Lemm	71.432,76	63.655.268,36	ihre
3.	903	1801	LBC-Lemm	33.575,37	34.842.795,42	ander
4.	7906	3106	LBC-Lemm	50.747,12	25.442.971,33	meine
5.	0	497	LBC-Lemm	10.215,30	14.436.557,14	e
6.	0	348	LBC-Lemm	7.152,71	10.108.494,74	disegno
7.	0	324	LBC-Lemm	6.659,42	9.411.357,17	andern
8.	0	262	LBC-Lemm	5.385,07	7.610.418,45	della
9.	0	262	LBC-Lemm	5.385,07	7.610.418,45	con
10.	0	249	LBC-Lemm	5.117,87	7.232.802,27	diligenza

Figure 5: Screenshot of the K-L-RIF (de) List

A very high degree of association with simultaneously low frequency in *DeReKo* suggests that it can be an artefact in the procedure or even in the primary source (such as for typing errors). High associative measurements (especially LLR) show good candidates for keywords of our corpus. If we consider the first ten positions, there was a particularly striking presence of possessive pronouns (*seine*/his; *ihre*/hers; *meine*/my) and adjectives in the comparative form (*lieber*/nicer; *besser*/better); however, these could be explained by a lemmatization error of the *Sketch engine*. As keywords of our German LBC-Corpus Italian words (*e*/and; *disegno*/drawing; *della*/of; *con*/with; *diligenza*/diligence), proper names (Jacopo; Arezzo; Giovan; Vasari), obsolete spelling variants (*seyn* vs. *sein*/to be; *giebt* vs. *gibt*/gives) and LSP-items (*Bauten*/buildings; *mediceisch*/as Medici etc.) were also signalled.

4) a technical word list from a monolingual lexicographical resource (G-LEX). These lists were extrapolated from central monolingual and bilingual dictionaries (TLFi for French, Duden online and Zanichelli 2009 for German[17]). All entries (1,806 for French and 2,439 for German) were entered into an Excel table.

In a second step, all mentioned lists (K-LBC; L-LBC; L-RIF; G-LEX) were automatically compared and merged[18] with formulas and functions from Excel (including CERCA.VERT):

- From the comparison of the keywords lists K-LBC and the dictionary lists G-LEX two lists resulted (see figure 6):
  - list 1, which comprises all items present both in our corpus and in a lexicographical resource
  - 'ONLY in K-LBC' with items that only appear in LBC. From this list additional technical terms could be identified.

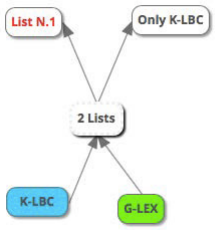

Figure 6: Lists resulting from the merging of K-LBC and G-LEX

- From the comparison of the frequency L-LBC list and the G-LEX dictionary list we obtained two lists (see figure 7):
  - list 2, with items present in both lists
  - 'Only in G-LEX': items which are uniquely present in the lexicographical resource but which are missing from our list

[17] The entry lists of these lexicographical resources were used also as secondary sources in our project.
[18] This procedure was adopted for each language.

Figure 7: Lists resulting from the merging of L-LBC and G-LEX

- From the comparison of the L-LBC frequency list and the reference corpus of each language we obtained two lists (see figure 8):
  - list 3, with items which appear in both lists
  - 'LBC not in L-RIF', with items that occur only in the corpus LBC but not in the reference corpus

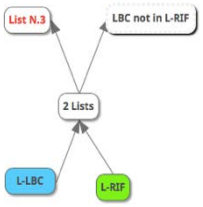

Figure 8: Resulting lists from the merging of L-LBC and L-RIF

From list 3 it was possible to filter out specialized lexemes which also occur in common language; from 'LBC not in L-RIF' all words could be extracted which are not included in the reference corpus of the German language, because they are e.g. archaisms, foreign words, proper names or errors.

All lists were then manually refined thanks to a fine-grained qualitative analysis: for example, all common language terms and other 'irrelevant terms' (such as place names, literary works, religious names etc.) were removed. To create the provisional entry list, Lists 1-3 were merged. Duplicates, and triplicates were deleted; variants of a term (among others *Perugino – Pérugin*, *Pisa – Pise*, *empattement – empatement*, *bastimens – bâtiment; terre-cuite - terracotta* for the French List; *chiaroscuro – chiaro-scuro*; *Lorenzo il Magnifico - Lorenzo de' Medici - Lorenzo Vecchio de' Medici*; *Mahlerei - Malerei; Piazza della Signoria - Piazza der Signoria - Piazza der Signoren*; *Scultur - Skultpur* for the German List etc.) were noted, and the most usual variant was chosen as the lemma.
With this procedure we were able to create the French and German provisional entry lists, which were rechecked and recleaned (table 3):

Language	Provisional entry list
French	1,200 entries
German	1,355 entries

Table 3: Merging of the lists to create the final French and German provisional entry list

Even if the two lists are different for numbers of items (see table 3) we can also see some similarities. Beyond the fact that

we can find mostly LSP-terms, they also have following aspects in common[19]:
1. the presence of many Italian names:

- in both lists: among others Agostino, Ambrogio, Andrea, Angelo, Angelico, Annunziata, Antonio, Arnolfo
- in only one list: among others Agnolo (fr), Alessandro (fr), Alessio (de), Alesso (de), Ambrogio (fr), Annibale (fr), Antonello (fr), Ascanio (fr)

2. the presence of many Italian surnames:

- in both lists: among others Alberti
- in only one list: among others Acciaioli/Acciaiuoli (de), Albizzi (fr), Aldobrandini (fr), Allori (fr), Antinori (fr)

3. the presence of many Italian city names, rivers:

- in both lists: among others Arno (in German we find the compound *Arnobrücke*/bridge of the Arno)
- in only one list: among others Arezzo (fr)

4. the presence of denomination of Italian institutions and monuments:

- in both lists: among others *Accademia*
- in only one list: among others *Academia* (de)

5. the presence of multiword items, which are both first names and surnames of Italian artists, such as Agostino Chigi (with the variant Agostino Chisi), Alesso Baldovinetti, Andrea di Cione, Angelo Poliziano, Arnolfo di Cambio (with the variants Arnolfo di Lapo, Arnolfo Lapi)

6. the presence of collocations such as: among others *adoration des bergers*/adoration of the Shepherds (fr), *adoration des mages*/adoration of th Magi (fr), *Auferstehung Christi*/ Christ's resurrection (de), *Ausgießung des Heiligen Geistes*/ outpouring of the Holy Spirit (de).

The provisional entry lists are the result of the current situation, but will be extended and supplemented in the future. For each entry list we also extracted keywords in context (KWICs), which after a qualitative analysis will be offered to the user (see Billero/Cetro/Farina et al. in prep. for French and Flinz/Ballestracci/Buffagni et al. in prep. for German). In the LBC-Dictionary the extracted KWICs will be useful in different ways: for determining the collocations and usual word combinations of the entries, as examples and as translations of collocations in case of equivalence.

4 Conclusions

The lexicographical process of the planned internet LBC dictionary has gotten under way, and the aim of this paper is to reflect on the data collection and the data preparation phase, which, as is usual for internet dictionaries, are open-ended, so what we present in this paper is only a snapshot of the actual situation.
Primary sources of our LBC dictionary are *ad hoc*-created comparable LSP-corpora, which are also freely accessible for other aims, in addition to the lexicographical one presented in this paper (see LBC-Platform). For their construction we used works from major Renaissance authors (see 2.1), both text in original language and translations, because the future aim of this research group is also to set up parallel corpora.
With an alternation of corpus-driven and corpus-based procedures we were able to extract the above discussed provisional entry list for French and German (3.1); the combination and merging of different types of lists and the consequent fine-grained qualitative analysis enable us to focus not only on the most frequent lexemes of our corpora but also on the lesser ones (i.d. *hapax legomena*), not only on the typical terms and multiword expressions according to web corpora (TenTen corpora) but also according to reference corpora. In addition, we also involved existing monolingual and bilingual dictionaries as secondary sources. KWICs of the entries have already been automatically extracted, and after a meticulous qualitive work which aimed at removing all the non LSP-ones will be freely accessible (see Billero/Cetro/Farina et al. in prep. for French and Flinz/Ballestracci/Buffagni et al. in prep. for German). A selection of them will also play a central role in modelling the lexicographical data in the database structure: they will be used for filtering out typical collocations and their examples, but also to reflect on equivalent structures.

5 References

Analyse et traitement informatique de la langue française - UMR 7118 (ATILF), Cognition, Langue, Langages, Ergonomie - UMR 5263 (CLLE-ERSS) (2020). *Corpus journalistique issu de l'Est Républicain* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). www.ortolang.fr, https://hdl.handle.net/11403/est_republicain/v3 [04.05.2020].

Ballestracci, S., Buffagni, C., Flinz, C. (in prep.). Das deutsche LBC-Korpus: Zusammenstellung und Anwendung. In A. Farina, C. Nicolás Martínez, R. Billero (eds.) *Corpora LBC*. Firenze: Firenze University Press.

Billero, R., Cetro R., Farina, A. et al. (in prep.). *Lexique français de l'art basé sur le corpus LBC (Lessico dei Beni*

[19] For this article we decided to focus, as examples, only on the lemmas beginning with the letter A.

Culturali). Firenze: Firenze University Press.
Billero, R., Farina, A., Nicolás Martinez, C. (in prep.). *Conclusioni: Dati numerici attuali e bilanciamento dei corpora*. In A. Farina, C. Nicolás Martínez, R. Billero *Corpora LBC*. Firenze: Firenze University Press.
Billero R., Nicolás Martinez, M.C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In *CHIMERA Romance Corpora and Linguistic Studies*, Vol.4, No. 2, pp. 203-216.
Duden online. Accessed at: https://www.duden.de/woerterbuch [04.05.2020]
Dunning, T. (1993). Accurate methods for statistics of surprise and coincidence. In *Computational Linguistics*, 19(1), pp. 61-74.
Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. 4. überarb. u. erw. Aufl. Tübingen: Stauffenburg.
Farina, A. (2015). Guideline Proposal for the Description and Translation of Proper Nouns in a Multilingual Cultural Heritage Dictionary of Florence. In O. M. Karpova, Faina I. Kartashkova (eds.), *Life Beyond Dictionaries*, Newcastle: Cambridge Scholars Publishing, pp. 122-132.
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*, vol. <24 http://publifarum.farum.it/ezine_articles.php?id=335> [04.05.2020]
Farina, A. (in prep.). *Le corpus LBC français.* In A. Farina, C. Nicolás Martínez, R. Billero *Corpora LBC*. Firenze: Firenze University Press.
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. In *JADT 2018 - International Conference on Statistical Analysis of Textual Data*, Roma, 12-15 giugno 2018. Rome: UniversItalia, pp. 108-116.
Flinz, C. (2019). Der lexikographische Prozess bei Tourlex (ein deutsch-italienisches Fachwörterbuch zur Tourismussprache) für italienische DaF-Lerner. In A. Klosa, A. Storrer, J. Taborek (Hrsg.) *Internetlexikographie und Sprachvermittlung. Jahrbuch Lexicographica*. Berlin: de Gruyter, pp. 9-35.
Flinz, C., Ballestracci, S., Buffagni, C. et al. (in prep.). *Deutsche Lexik der Kunst auf der Basis des Korpus LBC (Lessico dei Beni Culturali)*. Firenze: Firenze University Press.
Flinz, C., Perkuhn, R. (2018). Wortschatz und Kollokationen in ‚Allgemeine Reisebedingungen'. Eine intralinguale und interlinguale Studie. In S. Krek, et al. (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Context*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 959-967.
Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: C. Fellbaum (Hg.*) Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum, pp. 23-41.
Geyken, A., Lemnitzer, L. (2016). Automatische Gewinnung von lexikografischen Angaben. In A. Klosa, C. Müller-Spitzer (Hrsg.), *Internetlexikografie: Ein Kompendium*. Berlin/Boston: de Gruyter, pp. 195-241.
Hunston, S. (2008). Collection strategies and design decisions. In A. Lüdeling, M. Kytö, (Hrsg.) *Corpus Linguistics. An International Handbook*. Volume 1. Berlin/New York: de Gruyter, pp. 154-168.
Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release vom 08.03.2017). Mannheim: Institut für Deutsche Sprache. PID: 10932/00-0373-23CD-C58F-FF01-3 [04.05.2020].
Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. Gouws et al. (Hrsg.) *Dictionaries. An International Encyclopaedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin u.a.: de Gruyter Mouton, pp. 517-524.
Klosa, A. (2020). The lexicography of German. In P. Hanks, G.M. de Schryver (Hrsg.) *International handbook of modern lexis and lexicography*. Berlin: Springer, pp. 1-21.
Kupietz, M., Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L.M. Eichinger (Hrsg.) *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven* (Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/Boston: de Gruyter, pp. 297-322.
Il nuovo dizionario di tedesco. Dizionario tedesco-italiano, italiano-tedesco (2009). Ediz. bilingue. Con CD-ROM (Italiano). Torino: Zanichelli.
Lemnitzer, L., Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung*. 3. Aufl. Tübingen: Narr.
Lessico dei Beni Culturali. Accessed at: http://www.lessicobeniculturali.net [04.05.2020].
Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing.
Trésor de la Langue française informatise (TLFi). Accessed at: http://atilf.atilf.fr [04.05.2020].
Wiegand, H.E. (1998). Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York: de Gruyter.
Zotti, V. (2017). L'integrazione di corpora paralleli di traduzione alla descrizione lessicografica della lingua dell'arte: l'esempio delle traduzioni francesi delle Vite di Vasari. In V. Zotti, A. P. Alamán (cur.) *Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali*. Firenze: Firenze University Press, pp. 105-134.