Carolina Flinz

# The multifunctional LBC-Corpora: different aims depending on the user



**Abstract:** Corpora are nowadays the primary source of many dictionaries and the core element of various platforms and information systems. Lexicographers have therefore a variety of new possibilities which were unthinkable in the past with different types of corpora available for use in the data collection phase of the lexicographic process (Flinz 2021). Not only lexicographers, but also other types of users (academics, translators, teachers, students etc.) can profit from them, especially when corpora are public and can be accessed using corpus linguistic tools (Ballestracci/Buffagni/Flinz 2020; Flinz/Farina 2020). The LBC-corpora are monolingual specialized comparable corpora, already online (http://corpora.lessicobeniculturali.net/) and, as monitor corpora (Lemnitzer/Zinsmeister 2015: 140), they will be augmented over time. They can be analysed using the open source tool *NoSketchEngine* (Billero 2020). The LBC-Corpora are also the lexicographic primary source of the LBC multilingual dictionary, which is in preparation: the provisional entry lists of different languages (Spanish, German, and French) are now ready (Billero/Farina/Nicolás Martínez 2020) and together with a selection of KWICS, which have been carefully selected following a quantitative-qualitative procedure (for German see Buffagni/Flinz/Ballestracci in prep.), will soon be online (Flinz et al. in prep.). The purpose of this paper is to reflect on the LBC-corpora from a double perspective: from the user of the LBC-platform and from the lexicographic team. In the first case following an overview of the principal characteristics of the LBC-Platform, the focus will be on the accessible corpora showing the tools which can be used (§ 2). In the second case the LBC-corpora will be examined in their function as a data basis for the LBC-Dictionary (§ 3). The attention will be on the data preparation phase: after discussing the procedure for the realization of the LBC-provisional lemma candidate lists, the focus will be on the adopted procedure for finding equivalence relations and for the individuation of other types of relations between the entries (synonymy, belonging to the same semantic field etc.). In § 4 the focus will be on the LBC-provisional lemma

**Carolina Flinz**, Università degli Studi di Milano, Dipartimento di Lingue, Letterature, Culture e Mediazioni, Piazza S. Alessandro, 1, 20122 Milano (MI), Italy, E-Mail: carolina.flinz@unimi.it

https://doi.org/10.1515/lex-2023-0010

candidate lists and their related KWICs. Conclusions and an outlook to the future can be found in the last section (§ 5).

**Keywords:** lexical information system, cultural heritage, LSP-dictionary, corpora, lemma candidate list, KWICs

# 1 Introduction

Corpora[1] have long found their way into lexicography and are also primary sources[2] of many contemporary dictionaries. Their use for lexicographic purposes is diverse (see among others the descriptions in Wiegand 1998 and Engelberg/Lemnitzer 2009), but for the lexicographic process they are nowadays central, offering possibilities which were previously inconceivable using traditional collections of documents. If we concentrate on the different phases of the internet lexicographic process[3] (cf. Klosa 2013; Lemnitzer/Zinsmeister 2015; Flinz 2018), which is a dynamic process (cf. Klosa 2016) as the lexicographic resources are never complete and always open to new entries,[4] corpora can be used in the following ways (cf. Lemnitzer/Zinsmeister 2015: 170; Flinz 2018):

1. In the planning of the dictionary (*Vorbereitungsphase*), in which a great deal of time is invested in the conception of the content of the dictionary, the platform used for its maintenance and the exploration of the possibilities of computer technology. First reflections are on the primary sources, if they should be accessible to the user, and how this should be managed.
2. In the data collection phase (*Phase der Datenbeschaffung*), in which the decision of either using existing corpora or constructing a new one should be taken.[5] The determination of the correct existing corpus for the dictionary or the construction of the adequate corpus/corpora is strongly related to the type of dictionary,[6] its functionalities, and its intended users. The choice of the methodology/approach for the investigation and data extraction (corpus driven vs. corpus based; quantitative

**1** A corpus is a collection of written or spoken utterances in digital form (Lemnitzer/Zinsmeister 2015: 13).

**2** For the distinction between primary, secondary and tertiary sources see Wiegand (1998: 140).

**3** The main feature of the lexicographic process of internet dictionaries is the convergence and circular overlapping of the phases as the publication of the data begins even before the dictionary is completed (Klosa 2013: 518).

**4** These types of dictionaries are therefore called dynamic dictionaries (Lemberg 2001: 85) or dictionaries under construction (Storrer 2001: 65).

**5** The use of existing free accessible corpora is preferable (Lemnitzer/Zinsmeister 2015: 170), because of their size and representativity, but there are LSP-lexicographic projects in which this isn't possible (cf. Scherer 2006; Flinz 2018).

**6** For an overview of dictionary types in the online medium see Engelberg/Storrer (2016).

approach vs. quantitative-qualitative approach), as well as the tools that should be used, are fundamental in this phase.

3. In the data preparation phase (*Phase der Datenaufbereitung*): corpora are the primary source for the creation of the entry list, together with criteria[7] to determine if a lexical unit should be included or not in the lemma list of a dictionary. Corpora are also used for the individuation of the relation of equivalence of single and multiple items in the case of bilingual dictionaries: with the help of comparable corpora, which are fundamental in their focus on syntactic, semantic, morphological, and lexical relations, equivalents can be identified.
4. In the data analysis phase (*Phase der Datenauswertung*): one of the main tasks is the writing of new dictionary articles, in which the corpus data[8] in the form of Keywords in Context (KWICs) and word profiles etc. are interpreted and analyzed to determine possible collocations or other types of multiword units. Corpus data are also used to identify examples of use which, as authentic material, are valuable data for the dictionary user. Corpora can also help to determine the relations between the entries (synonymy, entries with the same collocational profile, entries belonging to same semantic field etc.). These relations are also relevant for the planning of hyperlinks inside the dictionary.

The LBC-Dictionary is a dynamic LSP-dictionary,[9] which is still in preparation. A lexicographic resource of this kind is strongly needed today, since a variety of print and online texts in different languages on Italian cultural heritage belonging to different text genres are readily available (among others tourist guidebooks, museum websites, art catalogues, critical essays), but there is still a lack of specific lexicographic resources able to support the different users (translators, teachers, tourist guides, tourist information centres, museum staff, students etc.) (cf. Billero/Nicolas Martinez 2017: 203).

The purpose of this paper is to reflect on the corpora[10], core of the LBC-platform (http://corpora.lessicobeniculturali.net/) and their possible applications for the user of the platform and for the lexicographer. Following an overview of the principal characteristics of LBC-Platform, the accessible corpora will be focused on, illustrating the tools which can be used (§ 2). The data preparation phase is described in the third section (§ 3): after discussing the LBC-provisional lemma candidate lists, the focus will be on

**7** The criteria chosen for the creation of the provisional or final entry list are various: single criteria based on frequency, typicality, keyness etc., using different statistical parameters or a mixture of criteria (cf. Geyken/Lemnitzer 2012; Flinz 2018; Farina/Flinz 2020; Flinz 2021).

**8** For possibilities and limits of the extraction of lexicographic information see Lemnitzer/Geyken (2016).

**9** For the beginning of the LBC project see Farina (2016). The project involves experts from different disciplines (among other lexicographers, corpus linguistics etc.) and universities (Florence, Bologna, Lisbon, Milan, Paris, Pisa etc.).

**10** For an overview of corpora types and on the parameters used to classify them, see Lemnitzer/Zinsmeister (2015: 137).

the adopted procedure for finding equivalence relations and for the individuation of other types of relations (synonym relations, belonging to the same semantic field). In § 4 the focus will be on the LBC-provisional lemma candidate lists and their related KWICs. Conclusions and an outlook to the future are presented in the last section (§ 5).

## 2 The LBC-Platform and its free resources/tools for different types of analyses

The LBC-Platform is a lexical information system[11] (https://www.lessicobeniculturali.net) created for different types of users (translators, teachers, lexicographers, tourist guides, tourist information centres, museum staff, students etc.) interested in Florence and the Tuscan cultural heritage. It provides different types of tools and resources (Figure 1), which are freely accessible for the user:


**Figure 1:** Screenshot of resources of the LBC-Platform (Dictionary, Corpus, Wordlist)

1. Monolingual comparable corpora: monolingual comparable corpora of English, French, German, Italian, Russian and Spanish[12] are already present in the platform and can be consulted using the functionalities of *NoSketch Engine.*[13] The texts were selected from all time periods according to their role in the dissemination of the cultural heritage of the city of Florence. They should provide an authentic overview of the type of language related to this theme (Geyken/Lemnitzer 2016: 203).
2. Monolingual LSP-dictionaries in the languages of the project (English, French, German, Italian, Russian, Spanish) (Farina/Billero 2018): the LBC-dictionaries are in preparation and will be dynamic. The lexicographic process is in progress and the data preparation phase has begun for different languages: the provisional entry lists of the LBC-German, French and Spanish dictionary are now ready (see point 3) and the first reflections on the possible relation of equivalence between the entries have started. The data analysis phase for these languages is running since Autumn 2021.

**11** See Klosa (2016).

**12** The corpora were published when a minimum of one million words was reached.

**13** *NoSketch Engine* is the free version of the professional tool *Sketch Engine*. *Sketch Engine* is normally a commercial tool, but thanks to a European project it has been free until March 2022. The *NoSketch Engine* interface used for LBC will have from 2024 a different graphic, more similar to the tool *Sketch Engine*.

3. Provisional entry lists with a selection of KWICs: the entry lists are ready for German, French and Spanish and the Spanish list was published at the end of 2021. For each entry selected, KWICs showing the use of the term in context will be linked.
4. Parallel corpora: Parallel corpora are in preparation and will also be available to the user. The aim is to widen the original project, focusing not only on the cultural heritage of Florence and Tuscany, but on Italian culture in general. There will be four types of corpora:
   – *The Vasari corpus,* in which the original Italian text of Giorgio Vasari (*Vite*, second edition) and its translations in the available languages of the project will be aligned.
   – *The Literature corpus*, which will include such texts such as *Corinne ou l'Italie* by Madame de Staël and *Morning in Florence* by John Ruskin. The translations will be available in the project languages.
   – *The Museum Web-Sites-Corpus,* in which texts from important museum websites and their translations in the available languages of the project will be aligned, starting with the *Musei Vaticani* website[14].
   – The *Travel Guide Corpus*, in which texts from travel guides and their translations in the available languages of the projects will be aligned.

## 3 The LBC-Comparable corpora: different types of analyses with NoSketch Engine

The LBC-Corpora are collections of written texts (8 Mil. Words) in English, French, German, Italian, Russian, and Spanish and are freely accessible from the LBC-Platform (http://corpora.lessicobeniculturali.net). During construction the most important criteria[15], such as corpus type, size, origin and quality of the texts, the documentation of the primary date and the annotation, were taken into consideration.

The main characteristics of the LBC-Corpora are summed up in table 1, but as monitor corpora the numbers of words can be increased:[16]

**14** Cf. https://www.museivaticani.va/content/museivaticani/it.html.

**15** For problems related to the construction of a new corpus see Hunston (2008), Lemnitzer/Zinsmeister (2015: 48–55) and Flinz (2021).

**16** The different language groups are in the process of collecting and preparing texts. For the adopted procedure see Billero (2020).

**Table 1:** Information about the corpora in the LBC-Platform

| | English | French | German | Italian | Russian | Spanish |
|---|---|---|---|---|---|---|
| Words | 1.036.000 | 3.165.000 | 1.018.000 | 1.009.000 | 1.900.000 | 1.020.000 |

For the selection of works and authors the criterion of their importance for the Florentine Renaissance in art and culture was followed. One of the central texts for the project was, for example, Giorgio Vasari's *Vite* (1550, 1568), which contributed to the spread of Italian Renaissance culture to most European countries, but also other works from non-Italian authors such as John Ruskin, Jacob Burkhardt etc., who promulgated Renaissance culture throughout the world, or works from authors such as Dumas, Stendhal, Goethe etc., who loved Italy and wrote about their journeys.

Both original texts and translations are part of the corpus, which consists therefore of three text categories:

– Technical texts, which are mainly LSP-texts (art and architecture).
– Literary texts (biographies, travel diaries, fictional narrative works such as short stories, novels etc., and essays).
– Other types of texts (mainly informative texts).

The proportion of each category varies from language to language (for the comparison of French and German see Farina/Flinz 2020), but the aim of the research groups is to increase the different categories in order to create more balanced corpora and improve their comparability (cf. Billero/Farina/Nicolá Martinez 2020). The featured authors are many and the covered timespan is from the 16th to the 21st century. Each corpus has detailed metadata[17] designed to designed to support different ways of searching.

The corpora can be searched and analyzed by the user through *NoSketch Engine*, which allows different types of analyses: intralingual (cf. Ballestracci/Buffagni/Flinz 2020, Ballestracci 2022) and interlingual (cf. Ballestracci/Buffagni/Flinz 2020; Flinz/Farina 2020; Ballestracci 2022). The tool offers two different functionalities: 1) the query in the corpus and 2) the extraction of word lists.

1. The query in the German LBC-Corpus (*Corpus LBC Tedesco*) can concentrate on the searched item ('Query types'), its context ('Context') and also on the text types ('Text Types') (Figure 2).

**17** Metadata for original texts are, among others, original language, text category, author, title, year of writing, year of publication (cf. Billero 2020).

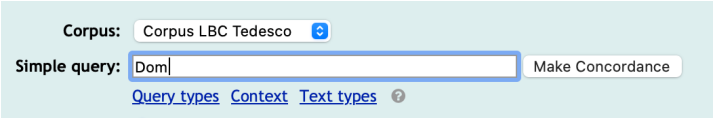


**Figure 2:** Search mask for corpus search (German LBC-Corpus, Simple Query of *Dom*)

In the first case an item can be searched in different ways (Figure 3):

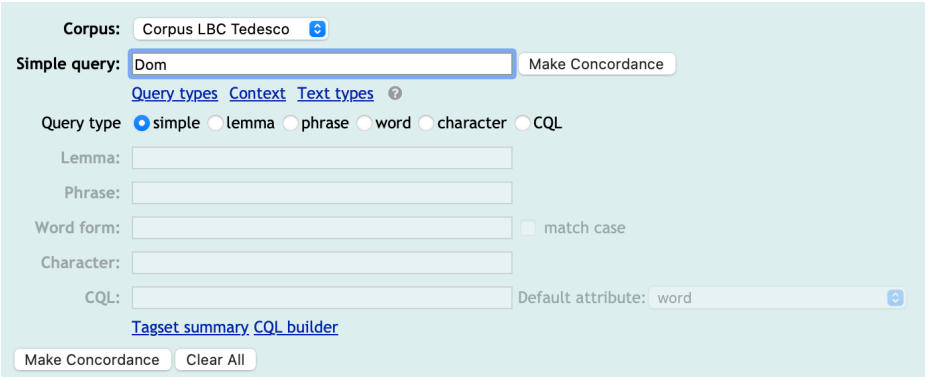


**Figure 3:** Search mask when the focus is on "Query Types" (German LBC-Corpus, Simple Query of *Dom*)

a. As a simple item (option "simple"), which is the default choice. If the query *Dom*[18] is entered the extracted concordances (219 occurrences, 185.05 per Million[19]) (Figure 4) show the lemma *Dom*. Occurrences of the singular form *Dom* and its genitive (*Doms*) are displayed:

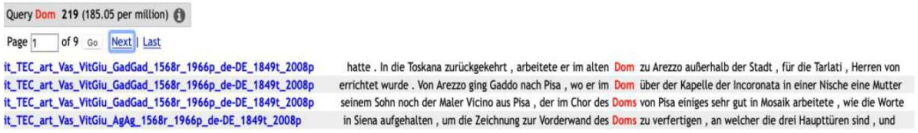


**Figure 4:** Screenshot of the results of the search query *Dom* (option "simple")

**18** The search is not case sensitive, so if the user types "dom" or "dOm" the same results are given.
**19** The absolute frequency is useful when working with only a corpus, for example intralingual analyses. If the interest is in comparing two corpora the frequency per million is important information, because corpora never have the same number of tokens.

The user has different possibilities of visualization and can choose to view the results as KWICs or sentences. He/she can also choose how many words should be shown, how many lines should be displayed on the page and if the metadata string (on the left in blue) should be visible or not. The KWICs can then be sorted by "the node", by "the left context" or by "the right context", so that different aspects can be focused on: for German, for example, if the KWICs are sorted by "the left context" the attention will be, among others, on adjectives used as attributes to *Dom*, as *alter Dom* ('old cathedral'), *majestätischer Dom* ('majestic cathedral'), *neuer Dom* ('new cathedral'), *schöner Dom* ('beautiful cathedral') etc. (cf. Flinz/Farina 2020) or on the prepositions used with Dom as *im Dom* ('in the cathedral'), *am Dom* ('at the cathedral'), *nach dem Dom* ('after the cathedral'), *für den Dom* ('for the cathedral'), *in den Dom* ('into the cathedral') etc.[20]

Collocation[21] candidates of the search term can also be extracted, using different statistical measures[22] (Figure 5). The focus on possible collocations of a term is an important element for the learning of a foreign language, because "Wortschatzlernen ist Kollokationslernen" ('the learning of a lexicon is the learning of collocations', Hausmann 1984).

**Collocation candidates**

Page 1 Go Next >

| | Cooccurrence count | Candidate count | T-score | MI | logDice |
|---|---|---|---|---|---|
| P \| N Pisa | 11 | 158 | 3.307 | 8.555 | 9.901 |
| P \| N Arezzo | 11 | 203 | 3.305 | 8.193 | 9.738 |
| P \| N Im | 12 | 368 | 3.444 | 7.461 | 9.387 |
| P \| N alten | 14 | 596 | 3.712 | 6.988 | 9.136 |
| P \| N im | 60 | 3,631 | 7.659 | 6.480 | 8.996 |
| P \| N Hauptaltar | 5 | 102 | 2.227 | 8.049 | 8.995 |
| P \| N außerhalb | 5 | 108 | 2.227 | 7.966 | 8.968 |
| P \| N Kuppel | 5 | 148 | 2.223 | 7.512 | 8.802 |

**Figure 5:** Collocation candidates of *Dom* using LogDice

**20** These kinds of analyses are also very useful also for the Teaching/Learning of German as a Foreign Language (for other examples cf. also Ballestracci/Buffagni/Flinz 2020; Ballestracci 2022; Flinz 2021a; 2021c). For other works about *Data Driven Learning* (DDL) see also Flinz (2021b).

**21** In the following example the empirical notion of collocation is evidenced (Firth 1957: 194; Evert 2009: 1213; Steyer 2013: 76), but with a quantitative-qualitative analysis. The theoretical view (Hausmann 1984: 401) can also be focused on.

**22** For example, the T-Score, MI, log likelihood, logDice. For the advantages and disadvantages of the different measures see Stefanowitsch (2020).

b. As lemma (option “lemma”), which is the same search as “simple” (219 occurrences).
c. As phrase (option “phrases”) as *dem alten Dom*.
d. As word form (option “word form”): for example, entering *Dom*, the results will be the occurrences of the word form *Dom* (136 occurrences). Only the occurrences of the entered word form *Dom* are displayed and no other forms (for example *Doms*).
e. As a sequence of characters (option “character”): entering *lein*, which is a German suffix, all words containing the sequence of characters “lein” are extracted.
f. Using the corpus query language (option “CQL”): for example, entering the query [tag=“ADJA”][lemma=“Dom”], all combinations “Adjective + Dom” are extracted (Figure 6). In the system there is also a CQL builder, which can support the user.


Query ADJA, Dom 25 (21.12 per million)

| | |
|---|---|
| it_TEC_art_Vas_VitGiu_GadGad_1568r_1966p_de-DE_1849t_2008p | arbeitete er im alten Dom zu Arezzo außerhalb |
| it_TEC_art_Vas_VitGiu_AgAg_1568r_1966p_de-DE_1849t_2008p | Maria neben dem alten Dom gegen den Platz |
| it_TEC_art_Vas_VitGiu_Marg_1568r_1966p_de-DE_1849t_2008p | Er ward im alten Dom außerhalb Arezzo in |
| it_TEC_art_Vas_VitGiu_Marg_1568r_1966p_de-DE_1849t_2008p | war in jenem alten Dom von Spinello gemalt |
| it_TEC_art_Vas_VitGiu_BuoBuf_1568r_1966p_de-DE_1849t_2008p | ließ ihn im alten Dom viele Bilder malen |
| it_TEC_art_Vas_VitGiu_JacCas_1568r_1966p_de-DE_1849t_2008p | und führte im alten Dom , der nunmehr |
| it_TEC_art_Vas_VitGiu_JacCas_1568r_1966p_de-DE_1849t_2008p | fand sich im alten Dom in einer Anbetung |
| it_TEC_art_Vas_VitGiu_PaSpin_1568r_1966p_de-DE_1849t_2008p | die er im alten Dom in Fresko malte7 |
| it_TEC_art_Vas_VitGiu_PaSpin_1568r_1966p_de-DE_1849t_2008p | einem Pfeiler des alten Domes den genannten heiligen |

**Figure 6:** Screenshot of the results of a query with CQL ([tag=“ADJA”][lemma=“Dom”])

In the case where the focus should also be on the context, the “size” of the displayed window of the searched entry can be adapted according to different needs (how many tokens left, how many tokens right, or both). In the case where the attention is on the text type, the user can decide if he/she wishes to work with a saved subcorpus[23] (for example with the subcorpus “literary texts”) or with the whole corpus. Other options include: selecting the language of the original text, the language of the translation (if it’s a translated text), the type of text, the author, the title, the fragment, the year of writing, the year of publishing etc.

2. With *NoSketch Engine* word lists (absolute frequency of words, tags and lemmas) of the whole corpus or of subcorpora can be extracted, as well as keyword lists, selecting a reference corpus and a focus corpus (for example the keywords of a subcor-

**23** Users can create a subcorpus of their own, making, for example, a search query which can be saved.

pus in comparison with the entire corpus). These lists are important for identifying the most frequent and typical words/lemmas of a corpus and to compare them to other corpora.

## 4 The LBC-Dictionaries: the data preparation phase

Corpora are the primary source for many dictionaries and play a central role among others for the following lexicographic purposes: for the creation of the lemma candidate list; for the exploration of the usage of a headword with the help of KWICs; for the determination of word meaning and its pragmatic properties; for the extraction of collocations and multiword units; for the identification of neologisms and new word meanings; as a source of examples; for identifying lexical-semantic and syntagmatic relations; for the determination of equivalents; for establishing reference structures (see among others Geyken/Lemnitzer 2016; Flinz 2021). However, only with accurate analyses can lexicographers filter and select the information they need and only by drawing on all available sources can a high-quality lexicographic product be created (cf. Ďurčo 2010: 131).

The LBC-Corpora are also the primary source for the LBC-Dictionaries. There are no other lexicographic resources of this kind nor are there freely accessible LSP-Corpora in existence.

The LBC-dictionaries are already in the data preparation phase for many languages. The provisional entry lists have been extracted and refined with an interplay of automated procedures and manual selection/interpretation of data (Geyken/Lemnitzer 2016: 208). For their extraction an alternation of corpus-driven and corpus-based procedures[24] can be used. In the first case the corpus is evaluated as a whole, and the data obtained are described in full (cf. Tognini-Bonelli 2001: 65), while in the second the corpus is used to look for examples that support certain arguments. In the lexicographic practice, both approaches can be combined (cf. Storjohann 2005: 254; Flinz 2021). Lemnitzer calls it a corpus-based quantitative-qualitative approach (Lemnitzer/Zinsmeister 2015: 37): the data are extracted from corpora, but additionally interpreted and lexicographically processed.

The methodology and the principal tool (*Sketch Engine*) used for the investigation of the LCB-Corpora were the same for all languages. Other tools have also been integrated according to the language and the reference corpus used for comparison.

The adopted procedure saw the automatic or manual extraction of the following lists:[25]

**24** Klosa (2007: 106; 112; 2010: 104) distinguishes between *korpusvalidierend* and *korpusgesteuert*.

**25** For the German Corpus see also Farina/Flinz (2020); Buffagni/Flinz/Ballestracci in prep.

1. The L-LBC-List based on the absolute frequency of the lemmas in the monolingual corpora (*functionality* "Word List" of *Sketch Engine*). It was decided to set the parameter "frequency ≥1" to also include *hapax legomena*. For German, for example, 45,029 items were extracted.
2. The K-LBC-Lists, which are automatically extracted single and multiword keyword lists based on the comparison with integrated web-corpora[26] (functionality "Keywords" of *Sketch Engine*). The 2000 single and 2000 multiword keywords have been determined through the Keyness Score and represent the most typical items of the focus corpora.
3. The K-L-RIF-List, based on the automatic comparison[27] between the LBC-Corpus and the Reference Corpus of the language. For German *DeReKo* was used, the German Reference Corpus hosted by the *Leibniz-Institut für Deutsche Sprache* (https://www.ids-mannheim.de). Two statistical measurements (χ2 und LogLikelihood Ratio Test, see Dunning 1993) were used for the analyses. The extracted list for German contained 10,402 items.
4. The G-LEX-List, which is a technical word list, compiled manually from the consultation of monolingual and bilingual dictionaries (2439 items for German).
   The mentioned lists were then automatically compared and merged by means of the functionalities of Excel[28] using a multiple step procedure (see Figure 7, for details concerning the merging procedure see Billero 2020; Farina/Flinz 2020 and Buffagni/Flinz/Ballestracci in prep.).

Thanks to a fine-grained manual qualitative analysis, common language and other irrelevant terms such as Italian words were removed as well as duplicated words. If there were variants of a term, only the most frequent variant was entered in the final list. The German provisional entry list has for example 1466 entries:[29] 84 % are nouns, 10 % verbs and 6 % adjectives. Multiword units and proper names are also present among the items.

The provisional entry lists of the LBC-Dictionaries contain the most frequent, typical, and relevant terms of our monolingual LBC-Corpora. Despite their provisional character, these lists are of great importance for the lexicographic process, as they will be the base for the data processing phase – the modelling of the lexicographic data. The entry lists as well as a selection of KWICs will soon be available online (see section 5).

**26** The TenTen Corpus Family (TenTen Corpora) is integrated in the *Sketch Engine* platform. The TenTen Corpora are text corpora automatically collected from the Web and have a corpus size of more than 10 billion words per language. They are available in more than 40 languages (cf. Kilgarriff 2004; 2008).

**27** This comparison was possible thanks to the support of the *Leibniz-Institut für Deutsche Sprache*. IDS internal tools were used for the comparison between the two German corpora.

**28** As for example CERCA.VERT

**29** The number of entries will change after the KWIC-analyses (see section 5).

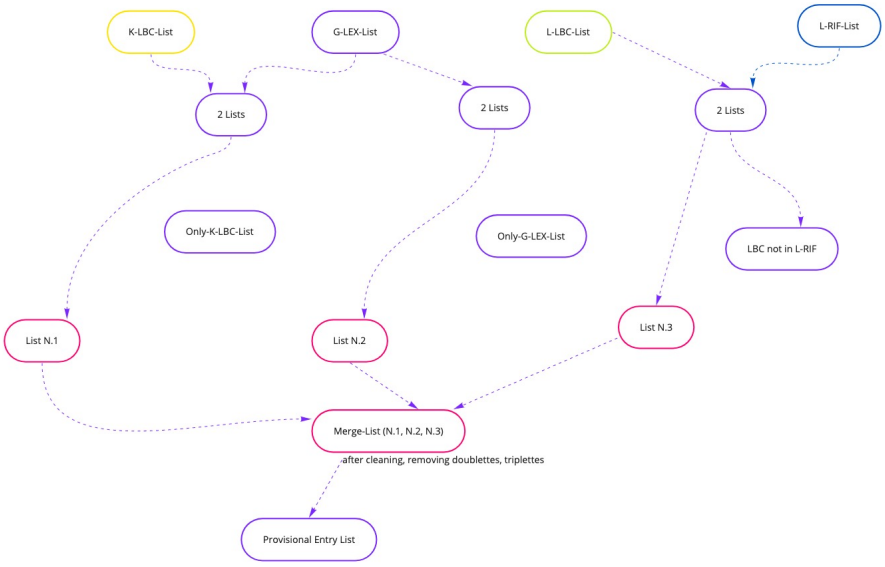


**Figure 7:** The multiple step procedure used for comparing and merging of the extracted lists

The monolingual LBC-Corpora will also be used for the individuation of relations of equivalence of single and multiword items. The starting point will be the comparison of the lemma candidate lists, focusing on their similarities. With attention directed to the first entries of the German and French lemma candidate lists possible equivalents can be immediately identified: *Abacus – abaque*; *Amphitheater – amphithéâtre*; *Antiquität – antiquité* etc. Starting from this observation, after further controls, hyperlinks from one LBC-Dictionary to another can be set. For other terms the lists will be inadequate and only detailed analyses of the syntactic, semantic, morphological, and lexical relations of the terms will help to identify the possible equivalents.

The LBC-Corpora will also be used to identify words which have the same collocation profile, words which are potential synonyms of an item and words which simply belong to the same semantic field. The tool used for this purpose is "Thesaurus" (*Sketch Engine*), which, starting from selected terms, extracts possible related items from the corpus. For example, if the searched entry in the German LBC-Corpus is *Kirche* ('church'), possible related items in the corpus could be *Kapelle* ('chapel'), *Palast* ('palace'), *Kloster* ('abbey'), *Dom* ('cathedral'), whereas if the item is the verb *malen* ('to paint'), related words could be *arbeiten* ('to work'), *verfertigen* ('to fabricate'), *ausführen* ('to execute').

The identified terms are important for the microstructure of the dictionary entry and can be added and linked together.

# 5 The LBC-provisional entry lists and the linked KWICs

The LBC-provisional entry lists are now complete for French, German, and Spanish. On the basis of these lists, KWICs for each term were automatically extracted and after filtering the context "on the left/right node" according to language, they were selected for the fine-grained qualitative analysis that followed (for German cf. Buffagni/Flinz/Ballestracci in prep.).

The selected KWICs are focused on the art and culture of the Renaissance, the city of Florence and the province of Tuscany; only the concordances which illustrated the use of a given lexeme as an LSP-word were chosen independently from the text genre in which they were used. For example, in the case where an LSP-term was not used in our corpus (there were no LSP-KWICs), it was decided to temporarily delete the entry from the provisional candidate list. With the expansion of the corpus, we hope to find subject-specific concordances and thus be able to include the lemma in the definitive word list. In some cases, the introspection and linguistic sensitivity of the lexicographic team played an important role.

The platform *Nextcloud* (https://nextcloud.com/) and the editor *Notepad++* (https://notepad-plus-plus.org/) were used for the selection of the KWICs. The fine-grained qualitative analyses temporarily reduced the provisional entry lists: in the German entry list the number of lemmas was reduced from 1466 to 1321 lemmas.[30] The related KWICs show the LSP-use of a term with the number varying from entry to entry.

After the editing procedure the LBC-Lists and their KWICs will be published online. As an illustration see the Spanish LBC-entry list, which contains 792 lemmas (http://lexicon.lessicobeniculturali.net/es/lemmario).

The user can search the entries either through the search field or via the alphabetically ordered entries (Figure 8):

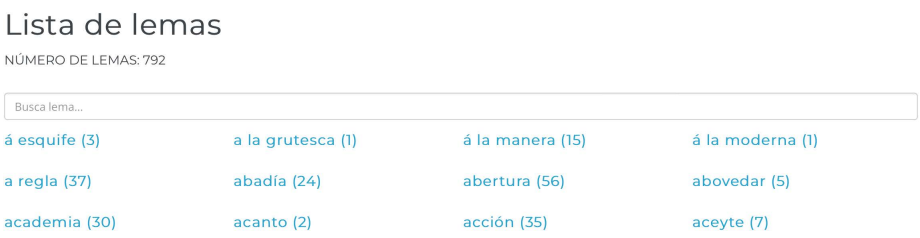


**Figure 8:** Screenshot of the Spanish LBC-Entry list

**30** A corrective formal comparison with other dictionary lists could not be carried out, as no other dictionaries of this type exist.

In the first case the search field has the function of helping the user to find the searched entry guiding him or her with the writing of the word. By typing "al" for example, different lemmas which begin with "al" or contain the character sequence are displayed: *al desnudo* ('naked'), *almohadillado* ('padded'), *alabastro* ('alabastro'), but also *balaustrada* ('balaustrada'). In the second case the entry can be searched by scrolling the alphabetically ordered entries, which are followed by the number of contained KWICs lines.

After finding the search entry, for *alabastro* (16 KWICs), the KWICs are listed: the keyword *alabastro*, in bold and evidenced in yellow is central. The KWICs are divided also per primary source. The primary source is given in the complete form (Author's Surname, Title, Book or Chapter, Date, Figure 9).

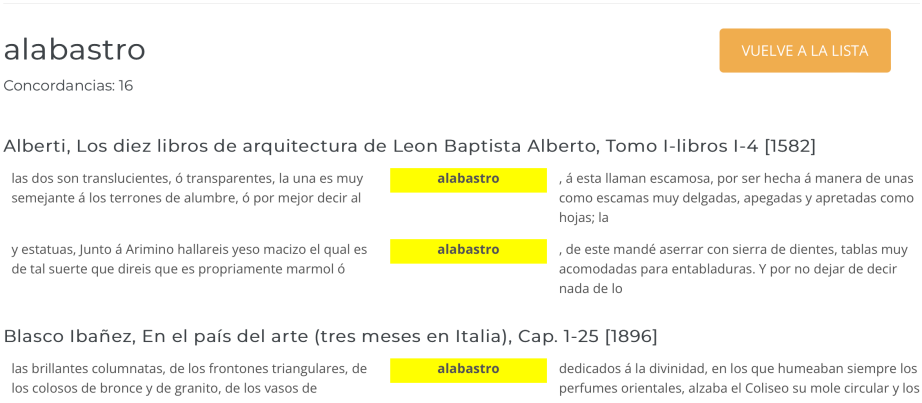


**Figure 9:** Screenshot of the Spanish LBC-KWICs

Entry lists and concordances can be very useful in identifying the relation of equivalence not only between the entries but also between collocations. The six-phase model[31] used for other LSP lexicographic projects (cf. Flinz 2021) could have several advantages for keywords that are not included in other resources or could not be assigned to the technical language of the cultural heritage. The analysis of syntactic, semantic, morphological and lexical relations is for such cases extremely important in describing collo-

**31** The model is based on preliminary work in the field of corpus-based contrastive analysis. The phases (in this case only 3 to 6 could be followed) are: 1. Metalexicographic analysis of the headword in bilingual dictionaries (L1<->L2) with regard to the equivalent lexeme, the equivalent collocation and the usage examples; 2. Analysis of the contexts in bilingual parallel corpora (L1<->L2) with regard to the equivalent lexeme, the equivalent collocation and the usage examples; 3. Extraction of the co-occurrences in the L1 corpus to determine the collocations; 4. Extraction of the co-occurrences in the L2 corpus to determine the collocations; 5. Comparison of the results of 1., 2., 3., 4.; and conclusions on equivalence relations; 6. Further research on the internet to find or check equivalents and contexts.

cational behaviour both intralingually and interlingually.[32] Collocations can be inductively reconstructed and tested for strength, variance, and pattern (Steyer 2013: 76).

## 6 Conclusions and future outlooks

The LBC-Platform is a lexical information system which offers different kinds of resources and tools to different user types (translators, teachers, lexicographers, tourist guides, tourist information centres, museum staff, students etc.):

- Monolingual comparable corpora which are online and can be analysed using *NoSketch Engine*. The texts were selected according to their role in the dissemination of the cultural heritage of the city of Florence over the centuries and include detailed metadata. The LBC monolingual corpora were also the primary source of the LBC-dictionaries which are in the data preparation phase. The provisional entry lists of the LBC-German, French and Spanish dictionary are now ready and the first reflections have begun on the possible relation of equivalence with a six-phase model between the entries (Flinz 2021). The data analysis phase for these languages is running since September 2021.
- The LBC-lemma lists and their KWICs. The entry lists are ready for German, French and Spanish and were published at the end of 2021. For each entry selected KWICs showing the use of the term in context will be linked. Both lists and KWICs will be accessible to the user, who will be able to search the entries and their KWICs by scrolling the list or using the search tool.
- The LBC-parallel corpora (*The Vasari corpus, The Literature corpus, The Museum Web-Sites-Corpus, The Travel Guide Corpus*) are in preparation and will also be available to the user. The aim is to widen the original project by focusing not only on the cultural heritage of Florence and Tuscany, but on Italian culture in general.

The LBC-Platform is a dynamic platform, which will be expanded and revised. It is an extremely important resource in the field of cultural heritage and a long awaited and much welcomed project.

**32** On the necessity of comparison corpora for the identification of equivalents (cf. inter alia Sinclair/Payne/Pérez 1996: 177; Calzolari 1996: 6; Prinsloo 2013: 1346).

# 7 References

## 7.1 Research literature

Ballestracci, Sabrina (2022): Für die universitäre DaF-Didaktik sind wissenschaftlich konzipierte korpusbasierte online-Ressourcen eine Ressource! Grammatikvermittlungsstrategien am Beispiel der open-access-Datenbank LBC. In: Cantarini, Sibilla/Missaglia, Federica/Bertollo, Sabrina (eds.): *Digitale Lehr-, Lern- und Forschungsressourcen für die deutsche Sprache*. Volume monografico di L'Analisi Linguistica e Letteraria, XXX, I–2022, 173–192.

Ballestracci, Sabrina/Buffagni, Claudia/Flinz, Carolina (2020): Il corpus LBC tedesco: costruzione e possibili applicazioni. In: Billero, Riccardo/Farina, Annick/Nicolás Martínez, María Carlota (eds.): *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Firenze: FUP, 55–75.

Billero, Riccardo (2020): Implementazione di software per la gestione dei corpora LBC. In: Billero, Riccardo/Farina, Annick/Nicolás Martínez, María Carlota (eds.): *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Firenze: FUP, 19–32.

Billero, Riccardo/Farina, Annick/Nicolás Martínez, María Carlota (2020) (eds.): *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali.* Firenze: FUP.

Billero, Riccardo/Nicolás Martínez, María Carlota (2017): Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In: *CHIMERA Romance Corpora and Linguistic Studies* 4:2, 203–216.

Buffagni, Claudia/Flinz, Carolina/Ballestracci, Sabrina (in prep.): Das deutsche LBC-Korpus: Provisorische Stichwortliste und Konkordanzen. In: Flinz Carolina, et al.: *Deutsche Lexik der Kunst auf der Basis des Korpus LBC (Lessico die Beni Culturali)*. FUP: Firenze.

Calzolari, Nicoletta (1996): Lexicon and corpus: a Multi-faceted Interaction. In: Gellerstam, Martin, et al. (eds.): Euralex *'96 Proceedings I*. Göteborg: Göteborg University, 3–16.

Dunning, Ted (1993): Accurate methods for the statistics of surprise and coincidence. In: *Journal of Computational Linguistics* 19:1, 61–74.

Ďurčo, Peter (2010): Einsatz von Sketch Engine im Korpus – Vorteile und Mängel. In: Ptashnyk, Stefaniya/Hallsteinsdóttir, Erla/Bubenhofer, Noah (eds.): *Korpora, Web und Datenbanken: computergestützte Methoden in der modernen Phraseologie und Lexikographie. Corpora, web and databases.* Baltmannsweiler: Hohengehren Schneider-Verlag, 119–131.

Engelberg, Stefan/Lemnitzer, Lothar (2009): *Lexikographie und Wörterbuchbenutzung*. 4. überarb. u. erw. Aufl. Tübingen: Stauffenburg.

Engelberg, Stefan/Storrer, Angelika (2016): Typologie von Internetwörterbüchern und -portalen. In: Klosa, Annette/Müller-Spitzer, Carolin (eds.): *Internetlexikografie. Ein Kompendium*. Berlin/Boston: De Gruyter, 31–63.

Evert, Stefan (2009): Corpora and Collocations. In: Lüdeling, Anke/Merja, Kytö (eds.): *Corpus Linguistics. An International Handbook*, vol. 2. Berlin/New York: De Gruyter, 1212–1248.

Farina, Annick (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* 24, <http://www.farum.it/publifarum/ezine_articles.php?art_id=335>.

Farina, Annick/Billero, Riccardo (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*. UniversItalia, 108–116.

Farina, Annick/Flinz, Carolina (2020): LBC-Dictionary: a Multilingual Cultural Heritage Dictionary. Data collection and data preparation. In: Gavriilidou Z./Mitsiaki M./Asimakis F. (eds.): *Lexicography for inclusion. Euralex-Proceedings*, vol. 1, 371–379, <https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p371-379.pdf>.

Firth, John Rupert (1957): Modes of meaning. In: *Papers in Linguistics 1934–1951*. London: Oxford University Press, 190–215.

Flinz, Carolina (2018): Der lexikographische Prozess bei Tourlex (ein deutsch-italienisches Fachwörterbuch zur Tourismussprache) für italienische DaF-Lerner. In: Klosa, Annette/Storrer, Angelika/Taborek, Janusz (eds.): *Internetlexikographie und Sprachvermittlung*. Jahrbuch Lexicographica. Berlin/Boston: De Gruyter, 9–35.

Flinz, Carolina/Farina Annick (2020): Analisi comparativa dei corpora LBC. La visione del patrimonio fiorentino tedesco e francese: l'esempio del Duomo di Firenze. In: Funari, Fernando/Farina, Annick (eds.): *Le présent dans le passé/Past in Present/Il passato nel presente*. Firenze: Firenze University Press, 77–100.

Flinz, Carolina (2021): Korpora als primäre Quellen von Tourlex. In: Taborek, Janusz/Piosok Michal/Woznicka Marta (eds.): *Korpora in der Lexikographie*. Berlin/Boston: De Gruyter, 57–83.

Flinz, Carolina (2021a): *Vergleichbare Spezialkorpora für den Tourismus: eine Chance für den Fachsprachenunterricht*. In: Hepp, Marianne/Salzmann Katharina (eds.): *Sprachvergleich in der mehrsprachig orientierten DaF-Didaktik. Theorie und Praxis*, Roma: Istituto Italiano di Studi Germanici, 133–151.

Flinz, Carolina (2021b): Korpora in DaF und DaZ: Theorie und Praxis. In: Flinz, Carolina/Hufeisen, Britta (eds.): *Korpora in DaF und DaZ: Theorie und Praxis. (Themenheft). Zeitschrift für Interkulturellen Fremdsprachenunterricht. Didaktik und Methodik im Bereich Deutsch als Fremdsprache* 26:1 (April), 1–43. <https://tujournals.ulb.tu-darmstadt.de/index.php/zif/>.

Flinz, Carolina (2021c): Attributive Funktion und weitere Funktionen von ganz. Vorschläge für den DaF-Unterricht polyfunktionaler Wörter anhand von Korpora. In: Fandrych, Christian/Foschi Albert, Marina/Hepp, Marianne/Thurmair, Maria (eds.): *Attribution in Text, Grammatik, Sprachdidaktik*. Berlin: Verlag Erich Schmidt, 275–301.

Flinz Carolina, et al. (in prep.): *Deutsche Lexik der Kunst auf der Basis des Korpus LBC (Lessico dei Beni Culturali)*. Firenze: Firenze University Press.

Flinz, Carolina/Farina, Annick (2020): Analisi comparativa dei corpora LBC. La visione del patrimonio fiorentino tedesco e francese: l'esempio del Duomo di Firenze. In: Funari, Fernando/Farina, Annick (eds.): *Le présent dans le passé/Past in Present/Il passato nel presente*. Firenze: Firenze University Press, 77–100.

Geyken, Alexander/Lemnitzer, Lothar (2012): Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In: Vatvedt Fjeld, Ruth/Torjusen, Julie Matilde (eds.): *Proceedings of the 15th EURALEX International Congress. 7–11 August 2012*. Oslo: Department of Linguistics and Scandinavian Studies, 362–366.

Geyken, Alexander/Lemnitzer, Lothar (2016): Automatische Gewinnung von lexikografischen Angaben. In Klosa, Annette/Müller-Spitzer, Carolin (eds.): *Internetlexikografie: Ein Kompendium*. Berlin/Boston: De Gruyter, 195–241.

Hausmann, Franz J. (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: *Praxis des neusprachlichen Unterrichts* 31, 385–406.

Hunston, Susan (2008): Collection strategies and design decisions. In: Lüdeling, Anke/Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook*, vol. 1. Berlin/New York: De Gruyter, 154–168.

Kilgarriff, Adam, et al. (2004): The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra (eds.): *Proceedings of the 11th* Euralex *International Congress, Lorient, France, July 6-10*, vol. 1. Lorient: Université de Bretagne Sud, 105–115.

Kilgarriff, Adam, et al. (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In: *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universidad Pompeu Fabra, 425–433.

Klosa, Annette (2007): Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In: Kallmeyer, Werner/Zifonun, Gisela (eds.): *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Berlin/New York: De Gruyter, 2007, 105–122.

Klosa, Annette (2010): On the combination of automated information and lexicographically interpreted information in two German online dictionaries. In: Granger, Sylviane/Paquot, Magali (eds.): *eLexicography in the 21st century. New challenges, new applications. Proceedings of eLex 2009. Louvain-la-Neuve, 22–24 October 2009*. Louvain: UCL Presses Universitaires, 157–163.

Klosa, Annette (2013): The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus, et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/Boston: De Gruyter Mouton, 517–524.

Klosa, Annette (2016): Der lexikographische Prozess im Projekt elexiko. In: Hildenbrandt, Vera/Klosa, Annette (eds.): *Lexikographische Prozesse bei Internetwörterbüchern. OPAL 1/201*6. Mannheim: Institut für Deutsche Sprache, 29–38.

Lemberg, Ingrid (2001): Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In: Lemberg, Ingrid/Schröder, Bernd/Storrer, Angelika (eds.): *Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. Tübingen: Niemeyer, 71–91.

Lemnitzer, Lothar/Zinsmeister, Heike (2015): *Korpuslinguistik. Eine Einführung*. 3. Aufl. Tübingen: Narr.

Prinsloo, Danie J. (2013): The utilization of bilingual corpora for the creation of bilingual dictionaries. In: Gouws, Rufus H., et al.: *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin/Boston: De Gruyter Mouton, 1344–1356.

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg: Winter.

Sinclair, John/Payne, Jonathan/Pérez Hernández, Chantal (1996): Corpus to Corpus: a study of Translation Equivalence. In: *International Journal of Lexicography* 9:3, 171–178.

Stefanowitsch, Anatol (2020): *Corpus Linguistics. A guide to the methodology*. Berlin: Language Science Press.

Steyer, Kathrin (2013): *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Storjohann, Petra (2005): Paradigmatische Relationen. In: Haß, Ulrike (ed.): *Grundfragen der elektronischen Lexikographie. Elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin/New York: De Gruyter, 249–264.

Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/Schröder, Bernd/Storrer, Angelika (eds.): *Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. – Tübingen: Niemeyer, 53–69.

Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Wiegand, Herbert Ernst (1998): *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*, 1. Teilband. Berlin/New York: De Gruyter.

## 7.2 Internet Resources

*Leibniz-Institut für Deutsche Sprache* = https://www.ids-mannheim.de
*Lemmario LBC (Es)* = http://lexicon.lessicobeniculturali.net/es/lemmario
*Lessico dei Beni Culturali (LBC)* = http://corpora.lessicobeniculturali.net/
*Musei Vaticani* = https://www.museivaticani.va/content/museivaticani/it.html
*Nextcloud* = https://nextcloud.com
*Editor Notepad*++ = https://notepad-plus-plus.org/