Alexander Geyken and Lothar Lemnitzer

# 7 The Automatic Extraction of Lexicographic Information

*In the mining industry, the art of surveying mines (die Markscheidekunst; Mark = boundary; scheiden = to divide; kunst = art) has always been vital for prospecting and finding one's way around underground (→ Fig. 7.1).*[1] *By analogy with the art of mining mineral deposits, processes are depicted here that corpus linguists use to describe precious deposits of words (i.e. corpora). Our finds, lexical information in this case, must still be brought up to the surface if they are worth it (→ Fig. 7.2) and must potentially be refined. This process is also presented in this chapter.*

## 7.1 Introduction

The focus of this chapter is the processes used to extract relevant lexicographic information from large collections of authentic language data, typically corpora, which are well suited to representing the usage of a language or language variety in a particular time period because of their size and the way they are documented with metadata. In the rest of this chapter, we will proceed from the assumption that the goal of our lexicographic work is to compile entries for a general monolingual dictionary of, say, contemporary German. The most important characteristics of a dictionary of this type are to capture the vocabulary of the language that is currently in use and to describe as many features as possible of the lexical units of this language, including formal properties, grammatical properties, and meanings (for more on the typology of dictionaries and, specifically, on this type of dictionary, cf. Engelberg/Lemnitzer 2009: 25–27). Deviations from this model assumption will be mentioned where appropriate. Lexicographic processes that fall outside the remit of this chapter are those required for compiling dictionaries that are bilingual or multilingual, specialist and technical, or document older stages of the language. This model is an abstract one in the sense that it says nothing about the presentation of the entries; in other words, a dictionary of this type could be published as a print dictionary, an electronic dictionary, or an Internet dictionary. Nevertheless, publication online

1 Source for the quotation: Geo- und Umweltportal Freiberg, http://tufreiberg.de/geo/gupf.

**Alexander Geyken,** Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: geyken@bbaw.de
**Lothar Lemnitzer,** Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: lemnitzer@bbaw.de

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111233758-008

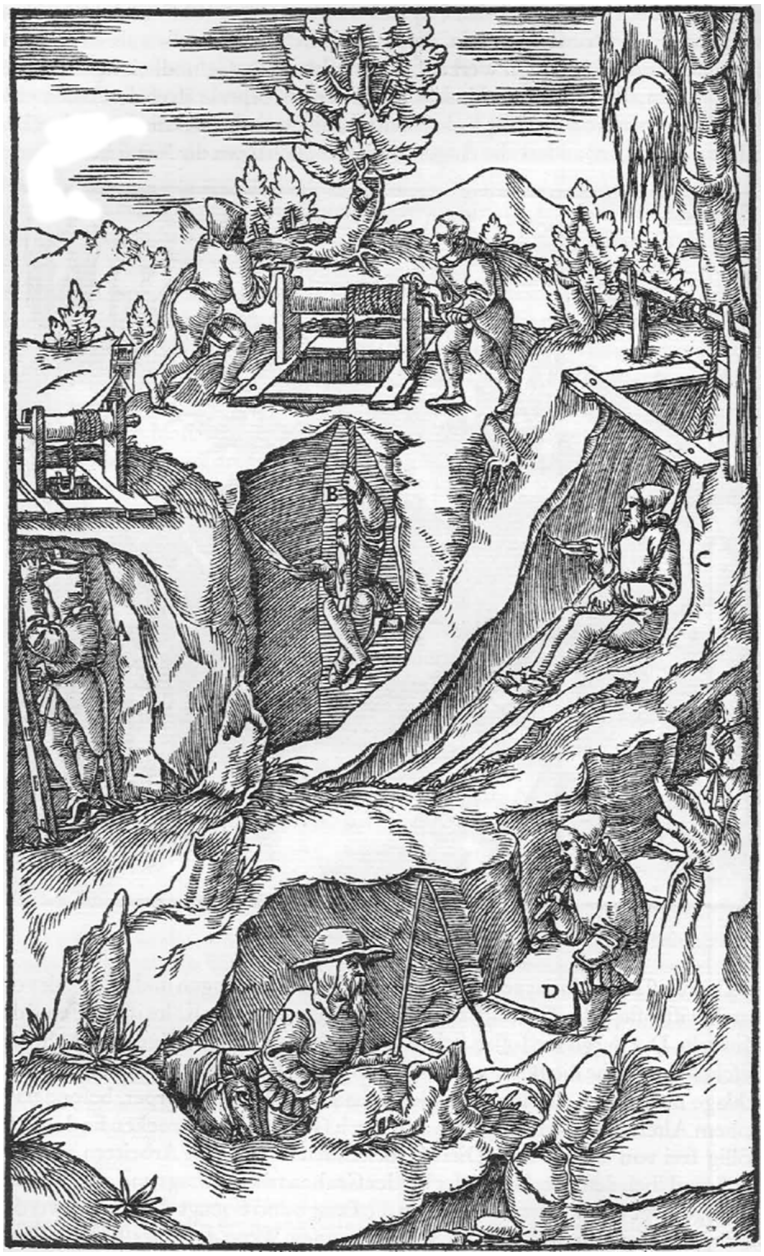


**Fig. 7.1:** Different forms of manriding.

does offer a whole range of possibilities for linking it in with other resources. Later, we shall examine the opportunities and risks of directly linking a dictionary with primary sources in more detail. We shall also introduce some classes of information for


**Fig. 7.2:** A mine surveyor taking measurements.

which the (semi-)automatic extraction of information is particularly fruitful. There is, however, no attempt on our part to be exhaustive concerning the microstructure of a typical, standardised word entry in our model dictionary.

In → Section 7.2, we provide information about the different sources used in the process of compiling the dictionary, followed by a more detailed exploration of corpora as a particularly interesting source for our purposes in → Section 7.3. In → Section 7.4, we consider some types of information to establish whether and how (i.e. with which tools) lexicographers – and users in the case of linked dictionaries and sources in the Internet – can extract data that lead to reliable and empirically secure judgements about the character of the word under consideration. In → Section 7.5, we then demonstrate the limits imposed by the current state of technology on the automatic extraction of lexicographic information. Finally, we explore a problem that arises specifically in Internet dictionaries: digital lexical systems make it possible to consult lexicographic information in dictionary entries and the sources on which this information is based simultaneously. In the process, inconsistencies between the base data and the lexicographic description become visible. We briefly present strategies for dealing with this problem from a lexicographical perspective.

## 7.2 The base data of a dictionary project: a typology of data sources

Whether we are talking about the pre-digital or digital period, a variety of types of sources have always been consulted to compile dictionary entries. In their totality, these sources are referred to in the lexicographic literature as the *dictionary basis*.

In the research on lexicographic processes (→ Chapter 3), namely the editorial processing of linguistic findings in dictionary entries and information, a systematic distinction is made between three types of sources. Primary sources include those texts that originate from natural communication situations. In what follows, we shall refer to collections of such texts as "(text) corpora". Secondary sources encompass those dictionaries that are consulted and analysed during the lexicographic process while tertiary sources cover all other linguistic sources, including grammars. The language competence or linguistic intuition of the editors and compilers also falls in this last category (Wiegand 1998; for further details, see Engelberg/Lemnitzer 2009: 235–237).

As a collection of authentic statements, or excerpts from them, *lexicographic cards indexes* count as the earliest type of primary source. As a rule, the notes or collections of attested examples referred to in this way are the result of work by many excerpters, who have taken excerpts out of texts and annotated them with details about their source. Hence, they reflect the choices and biases of the excerpters, although they do give lexicographers access to the primary text by virtue of a precise citation to the source example.

On the one hand, these collections are the result of well-considered and planned selections from a wealth of material that would otherwise be unmanageable, at least in the pre-digital era. On the other hand, Atkins and Rundell, among others, are criti-

cal of this type of source.[2] In addition, access to collections of attestations is cumbersome once they exceed a certain size. If we approach a large number of examples with a new enquiry, as a rule, that will involve re-sorting a pile of cards. Simple questions like "Is word X attested in the masculine gender later than 1800?" require a time-consuming search in large piles of cards, and some questions that would require examples to be aggregated simply cannot be answered at all in this way. A further difficulty is that lexicographic card indexes are tied to a particular physical location. Examples of "paper" collections of attestations can be found above all in long established historical dictionaries of a language, such as the Oxford English Dictionary (OED) and the Deutsches Wörterbuch (DWB) founded by Jacob and Wilhelm Grimm. An example of a collection of examples oriented towards contemporary language is the Duden language card index.

In the era of digitisation, the use of *(text) corpora* is opening up possibilities for analysing current language use that, as shown above, are not possible with any other kind of source. In the context of a project, digital corpora are accessible regardless of their location and they provide an unbiased picture of the language they illustrate in the sense that they also offer evidence of apparently trivial (i.e. ordinary, common) phenomena. Nowadays, the task of extracting data for specific queries has been taken on by flexible search engines, often purpose built for lexicographic needs. Examples include the search engines on the websites https://www.collinsdictionary.com/ and https://dictionary.cambridge.org/. As a rule, the search engines themselves are not visible to the user and only accessible by inputting a search term or terms into a text field.

According to the classification above, *other dictionaries* count as secondary sources. Older dictionaries of the same type as the reference work being compiled as well as specialist dictionaries of all kinds are important sources for a project's own work. However, as lexicographers, we must be constantly aware that a dictionary text is always an interpretation made by our predecessors or their colleagues of the source material available to them, which will have been limited in one way or another. As a rule, experienced lexicographers can judge the general quality and reliability of the lexicographic descriptions that have been consulted. In any case, healthy scepticism and, ideally, checks in other sources are advisable before adopting information from other dictionaries. In the Digitales Wörterbuch der deutschen Sprache (DWDS), on which the authors of this chapter work, an attempt is made to connect historical examples – of which there are sufficient in the Wörterbuch der deutschen Gegenwartssprache (WDG), which underpins the digital project – with their sources, insofar as these are available in digital form and accessible via the Internet. The textual basis

2 Atkins and Rundell 2008: 52: "As Noah Webster and James Murray both observed, human readers tend to notice what is remarkable and ignore what is typical, and this creates a bias towards the novel or idiosyncratic usages which inevitably catch the reader's eye . . .".

used here is the DEUTSCHES TEXTARCHIV. However, dictionaries can be used as more than simply a source of inspiration in the process of compiling entries. Insofar as another related dictionary is well structured and available electronically, it can also be used for comparing data on a larger scale, like for comparing lemma lists or the meaning of a particular headword.

Individual *language competence* or the linguistic intuition of the staff working on the dictionary or of the excerpters belongs to the group of tertiary sources, together with a well-stocked linguistic reference library, which ought to be at the disposal of any large project. Linguistic intuition is available throughout the lexicographic process, but is not necessarily reliable. Individual judgements are difficult to generalise to the degree that is required for reliable lexicographic work. In some areas, linguistic intuition is even systematically unreliable, for example when estimating frequency of occurrence (cf. Rapp 2003), or inadequate, for example when capturing relevant connections to other words for a given lexical sign (e.g. collocations, cf. Geyken 2011, who compared the collocations for several headwords in the "Dictionary of Contemporary German Language" with the results from an analysis of large corpora, and, more generally, Hanks 2012). Our own linguistic intuition can be an important corrective when interpreting other sources but it must always be questioned critically.

In the next section we examine corpora in more detail. As with the other data sources, using corpora to compile dictionaries has to involve awareness of the following limitations. Firstly, no corpus, however large, can illustrate or represent a living language as a whole. However, the bigger the corpus that is used and the more balanced it is in terms of a number of dimensions, such as text types or the geographical and temporal distribution of texts (cf. Geyken 2007), the higher its illustrative value. Many large corpora consist to a large extent, or even exclusively, of newspaper texts. Other corpora systematically capture other text types as well, such as functional texts. Transcripts of the spoken language are limited practically to specialist corpora, such as the ARCHIV FÜR GESPROCHENES DEUTSCH (AGD) at the Leibniz Institute for the German Language (IDS). Secondly, caution is needed when abstracting from observational data in corpora to systematic descriptions of the language, especially when the number of attested examples of a phenomenon is very small. Finally, all secondary linguistic analyses of large volumes of textual data are prone to error; when data have been manually annotated or checked, the result will contain a multitude of subjective decisions that are difficult to monitor (for more detail on these three aspects, cf. Lemnitzer/Zinsmeister 2010: 50–57, and Lemnitzer 2022).

Despite these limitations, we shall relate this chapter, which is devoted to the automatic extraction of lexicographic information, to textual corpora as a source of data. As shown above, lexicographic data cannot be extracted automatically from any of the other data sources. In the following section, we first consider in more detail the relevant features of digital text corpora.

## 7.3 Corpora

Drawing on Lemnitzer/Zinsmeister (2010: 8), we *define* "corpus" as a collection of written or spoken statements. The data in the corpus are typically digitised and machine readable. A corpus consists of primary data (that is, the texts), as well as possibly also metadata that describe these data and linguistic annotations that are assigned to these data.[3]

From a lexicographic standpoint, an important requirement for a corpus relates to scale. For one thing, as its scale increases, so does the probability of finding a rare construction that can nonetheless be formed according to the grammatical rules of the language. As we shall see below, a certain scale – measured as the number of words – is actually obligatory for aggregating statistical analyses; in other words, under a certain size of corpus, the results of statistical analyses are poor for lexicographic purposes (Geyken 2007: 37). By way of comparison, the English corpus underpinning the first edition of the COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (CCELD, 1987) encompassed 20 million words; the original reference corpus for British English, the BRITISH NATIONAL CORPUS (BNC), extended to 100 million tokens in 1993, as does the core corpus of the DWDS. Currently, the number of words in corpora of contemporary language hover in the region of double-digit billions: a widely used example is the TenTen Corpus Family,[4] where corpora of an average size of 10 billion words are still being crawled from web data for more than 35 languages (Jakubíček et al. 2013: 125–127, cf. also the website of sketchengine[5]).

The origins of the corpus texts and the quality of the digitised copies are further requirements. Other requirements for lexicography, especially when the corpus is supposed to serve to document language through attested examples, are the selection of texts and their documentation, that is, the metadata of the corpus data themselves and the accompanying texts, which, for example, provide information about the compilation of the texts. While "100 million word" reference corpora are ideal in this respect, the "billions of words" corpora exhibit considerable shortcomings, the selection of texts often being "opportunistic" and the documentation about their origin inadequate.

The quality of primary data often leaves much to be desired as well insofar as they involve scans of text documents that have not been subject to any further checks. This does not mean that these corpora cannot be used – quite the opposite, as they are often the only available source for rare linguistic phenomena. From a lexico-

**3** "Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind" (p. 8).

**4** https://www.sketchengine.eu/.

**5** https://www.sketchengine.eu/corpora-and-languages/corpus-list/.

graphic perspective, however, this kind of corpus should only be used as a supplementary source.

Following Lemnitzer/Zinsmeister (2010: 44–50), we distinguish between three levels in (textual) corpora: primary data, metadata, and structural and linguistic annotations of primary data. As we shall demonstrate below, information from all three levels is essential in different ways for different types of lexicographic analysis.

Apart from a few exceptions, the *primary data* of a corpus are directly available: one exception worth mentioning is collections of so-called tweets, posts on X, the platform formerly known as Twitter as only links to the data can be provided, not the data themselves (for further details, cf. Moreno-Ortiz 2024).

Metadata are essential for almost any reuse of corpus data in linguistics or lexicography. Those data comprise minimal information like author, title and date of publication. The correct date is of crucial importance for reuse in lexicography, including time series as well as the correct date of dictionary quotations.

Additionally, corpora are annotated with structural document data, i.e. the division of the document into chapters, sections, titles of chapters or sections and, of course, page numbers in the original document if it is not born digitally. These details are necessary for lexicographic attestation (and the associated details of the source of the evidence). Linguistic annotations, which typically describe morphosyntactic features of the words and, more rarely, semantic features, are useful primarily for searching for examples in a more targeted way. Thirdly, high-quality metadata should provide reliable information, above all, about the date and source. They are indispensable for lexicographic use of corpora for attestation. Further below, we shall demonstrate that some information from corpora cannot be identified at all without the availability of suitable metadata. Schmidt (2004) also deals in detail with the topic of metadata in relation to corpora.

Finally, it is important to distinguish between different *types of corpora* since they vary in their suitability for different lexicographic requirements. Some of the relevant differences for lexicographic work are:

– Differences between *reference corpora* and *specialist corpora*. The former aspire to give an overall picture of the documented language while the latter only cover a specific field. Since 2004, there has been a general core corpus of contemporary German language of the 20$^{th}$ century, which is balanced chronologically and by text types across the whole of the 20$^{th}$ century: the DWDS-Kernkorpus (Geyken 2007). Corpora of technical language are a good example for specialist corpora that are relevant for lexicography since they serve as the source for specialist technical lexicography. We can also view a corpus that encompasses the texts of a single author (e.g. Kant-Korpus) or a magazine (the corpus of the magazine “Die Fackel”) as a specialist corpus.
– Differences in the *modality* of the corpus. In addition to the well-established distinction between written and spoken language (also mentioned above), a third type has emerged that is called computer-mediated communication (CMC; for fur-

ther details, see https://cmc-corpora.org/). A reason for the distinction is that, linguistically, computer-mediated communication bears the characteristics of both written and spoken language.

- Differences between *monolingual corpora* and *multilingual corpora*. Monolingual corpora are essential for the lexicographic work described here. Multilingual corpora are often parallel corpora in the sense that a sentence from the section of the corpus in language B is a translation of a sentence from the part of the corpus in language A. However, sometimes multilingual corpora are not aligned precisely but consist of texts originating from a similar language field. In this case, we talk about comparative multilingual corpora. Multilingual corpora are not very relevant for monolingual lexicography.
- Differences between contemporary (*synchronic*) *corpora* and (*diachronic*) *corpora* relating to earlier stages of the language. This distinction relates to the object being described. Corpora of the first type illustrate a window in time for the language that we can describe as "contemporary language", mostly going back several decades before the corpus was compiled. Corpora of the second type document language use in a particular well-defined period, such as the language of Old High German or Middle High German. We can view corpora that cover several stages in the language, including contemporary language use, as a hybrid form between these two types. If the metadata make it possible, this kind of corpus can be divided as required into a synchronic contemporary part and a diachronic historical part.
- Differences between *static corpora* and *dynamic corpora*. Corpora of the former type are permanently available and it is possible to reliably refer to them when searching for and documenting lexicographic or linguistic findings; in other words, the primary data can be found again. Dynamic corpora, in contrast, change continuously, mostly by regularly adding further texts, often on a daily basis. The strength of dynamic corpora lies in how up to date the data are and the fact that particular phenomena can be observed over a longer period of time thanks to ever newer data. In an extreme case of a dynamic corpus, a so-called monitor corpus, language data are available from a very small window of time and only for a very short period of time, after which they are deleted again. The data from X (formerly known as Twitter) represent such a case (for more information on monitor corpora and their lexicographic use, cf. Sinclair 1991).[6]

Once the project team on our nominal general monolingual dictionary have settled on one or more corpora as primary sources, the work of data extraction and data analy-

**6** An extensive and up-to-date collection of links to all sorts of corpora is given in the "Virtual Language Observatory" (VLO; https://www.clarin.eu/content/virtual-language-observatory-vlo).

sis can begin. At present the following modes of data extraction predominate in lexicographic practice.

- For a particular keyword, possibly further specified through linguistic information on that keyword, examples are extracted in which that keyword appears and are displayed. The resulting list of examples is called a *concordance*. This is the selection method used for exploring the different ways in which a keyword is used. We can further distinguish so-called *Keyword in Context* (KWIC) concordances, where, in addition to the keyword, a certain number of words to the left or right are displayed, from concordances where a whole sentence or an even larger context is shown.
- Statistical data are identified for a keyword covering, for example, the frequency of occurrence of the keyword in the corpus (important, for example, for selecting lemmas), the distribution of the keyword in different texts or parts of the corpus (these can be interesting for identifying pragmatic usage characteristics), or typical word combinations (this is important for identifying collocations and phrasemes, etc. that have the keyword as a component).

In lexicographic work with corpora, there is almost always an interplay between automatic or automatised extraction processes and the process of selecting and interpreting data that follows. In this respect, it is more accurate to refer to the partially automatic extraction of information. Irrespective of whether the data are extracted automatically or partially automatically, lexicographers have to interpret the extracted data in the case of a dictionary compiled by editors, classifying them and incorporating them into their evaluation of the issues. In a case where corpora and their partially automatic analyses are directly accessible in the context of a lexical information system, the interpretation and evaluation of the data are the users' responsibility.

Taking as our starting point the set of information that is typically provided by large general language dictionary (examples for this type are for German: the WDG or DUDEN – DEUTSCHES UNIVERSALWÖRTERBUCH [DDUW]; for English: Cambridge English Dictionaries [CaED] or Merriam-Webster [MW]), we shall demonstrate in the following what information can be extracted systematically from corpora (→ Section 7.4) and what problems need to be reckoned with (→ Section 7.5).

## 7.4 Information classes in dictionaries

As mentioned above, our starting point in what follows is the model structure of a standardised entry,[7] or article, in a *comprehensive monolingual dictionary of general*

7 The terms *entry* and *article* are used as synonyms.

*language*. This is independent of the medium in which the data for this kind of dictionary will be presented: in print, as an electronic dictionary app, or on the Internet.

In identifying and listing *information classes*, we follow the formal description of standardised article structures for dictionaries that was developed in detail above all by Wiegand and Hausmann (cf., among others, Hausmann/Wiegand 1989). According to this, the abstract microstructure of the entries in a particular dictionary consists of a series or hierarchy of *information classes* clustered into larger groups. Some of the information types are obligatory; others are optional. Some of these information types – at the very least all of those that are obligatory – will be realised in the concrete microstructure of a particular article.

Since we are not dealing with a specific dictionary in this chapter but rather with the *information programme* of a general, model dictionary, we shall always refer in the following to the information classes and to the contribution that corpora and extraction tools can bring to identifying concrete data for a particular information class in an individual entry.

Terminologically, our reference point is the "tree of information types" in Hausmann/Wiegand (1989, Fig. 36.9) and the list of information types in Wiegand (1989, Fig. 39.3). The "tree" makes it possible to organise and group information types hierarchically and the table in Wiegand (1989) allows us to label the types with the correct terminology.

## 7.4.1 Form-based information classes

#### Form of the lemma sign and variants

In terms of the external form of the written word, that is, the representation of its form and spelling in the dictionary, Wiegand (1989: 468) lists the following *information classes*: form of the lemma sign, syllables, spelling and spelling variants. Relevant lexicographic insights cannot be extracted for all of this information by analysing corpora, however. Syllabification, for example, is normative in many languages, overwhelmingly facilitated technically nowadays using corresponding software modules in word-processing programs, and mostly removed at the end of a line during digitisation since it is a typographic strategy related to line length that makes finding words more difficult or altogether impossible, for example, with a search engine.

In contrast, information where *orthographic norms* are not prescriptive or leave room for interpretation is particularly interesting for lexicographic work; here, different *language usages* can be established. For the German-speaking countries, the official institution where orthographic norms are dealt with is the Council for German Orthography (Rat für deutsche Rechtschreibung: https://www.rechtschreibrat.com/). More specifically, its goals are to monitor the development of German spelling on the

basis of large reference corpora, to maintain the uniformity of spelling in the German-speaking world, and, finally, to clarify cases of doubt in German spelling.

Representing these different conventions can be one aim of a dictionary project with a primarily descriptive orientation. We illustrate this below with some examples where *spelling variant information* can be gleaned from corpora.

- Competing spellings of compounds with and without a hyphen. The rules of the Rat für deutsche Rechtschreibung (§§40f.; for the current version of the rule(s), see https://grammis.ids-mannheim.de/rechtschreibung/6159#) allow for some flexibility here, especially §45: "Man kann einen Bindestrich setzen zur Hervorhebung einzelner Bestandteile, zur Gliederung unübersichtlicher Zusammensetzungen, zur Vermeidung von Missverständnissen oder beim Zusammentreffen von drei gleichen Buchstaben". [It is possible to insert a hyphen in order to emphasise individual parts, to divide confusing compounds, and to avoid misunderstandings or runs of three identical letters]. Considerable variation can be found, above all, in compounds with a non-native component (*Musik-Download* vs. *Musikdownload* 'music download') and also, for example, in copulative compounds (*rot-grün* vs. *Rotgrün* 'red-green').
- Competing spellings in the use of a joining morpheme in a compound (or not). Here, there can be one variant with a joining morpheme and one without (*Vertragrecht* vs. *Vertragsrecht* 'contractual law') or two variants with different joining morphemes (*Schweinebraten* vs. *Schweinsbraten* 'roast pork').
- Competing spellings due to the liberalisation of norms in new spelling rules. This relates in particular to the degree of integration of loanwords into the system of native spelling (see Deutsche Rechtschreibung, §32[2], https://grammis.ids-mannheim.de/rechtschreibung/6151); for example *Portemonnaie* vs. *Portmonee* 'wallet').
- Competing spellings for other reasons. These include the variation between *-oxid* and *-oxyd* (in *Eisenoxid* vs. *Eisenoxyd* 'iron oxide') or between *Ski-* and *Schi-* (both simplex and in compounds like *Schigebiet* vs. *Skigebiet* 'ski resort').

If the editors decide to mark spelling variants like this when compiling the entry, this raises the question as to the order of the different variants. This problem can be solved in three different ways. First, a rule is stipulated in the lexicographic manual, for example, that (for case 1 above) the variant without a hyphen is given before the variant with a hyphen, but this might contradict current writing practice and is therefore misleading. Corpora come into play for the second option: the variant that is more frequent in the underlying corpus is always presented first. However, this kind of ordering, where one variant or another is "preferred" for each particular headword depending on the evidence, has, at least, to be explained in the supporting texts for the dictionary; better still is a relative or absolute indication of frequency for the variants. The third possibility is to mark a variant with its possible usage restrictions (for example, as "technical", "southern German", or "old-fashioned"), if this can be established from the metadata for the texts in which each variant occurs (for further details on this → Section 7.4.3).

In certain circumstances, a change in *usage preference* over time has to be taken into account when observing spelling variants. One example of this concerns the variant *Ski*, where 141 examples can be found in the DWDS-KERNKORPUS in documents from the first half of the 20th century (= Z1, Z standing for Zeitraum 'time period') and examples in documents from the second half of the 20th century (= Z2). For the variant *Schi*, 77 examples can be found in Z1 but only 6 examples in Z2. This empirical finding indicates a change in usage preference in favour of the first variant. If the second variant (*Schi* or *Schi-*) is included in a dictionary of contemporary language, it can justifiably be marked as "rare(r)" following evidence from the corpus.

### Grammatical information

This is not the place to present the specifics and functions of grammatical information in dictionaries or in *dictionary grammars* in detail. By way of introduction, we recommend Mugdan (1989). However, there are still some points to cover which are of relevance for our monolingual dictionary of German.

Unlike details on the form of the headword being described, grammatical information cannot be extracted directly from corpora as it is structural in nature. In order to be able to find the necessary information in a targeted way, additional details may be required in certain circumstances going beyond the surface form of individual words, for example concerning word class, or abstract linguistic categories such as "prepositional phrase" or "subordinate clause". Hence, a successful search requires either prior linguistic analysis or subsequent selection and interpretation of data.

More specifically, three aspects play a role in successful searches for linguistic structures in large bodies of text: the corpus and its linguistic pre-processing; the tool which can be used to put search queries to the corpus; and the researcher or lexicographer and their interpretation of the data.

We already mentioned that there is a second layer in a linguistic corpus in addition to the primary data, namely linguistic annotations (→ Section 7.3). To understand the following, it is only important to know that language technology tools can add information on the morphology and part of speech of a word (token) in the text (usually a word class tagger) and also mark up and analyse structures beyond individual words (i.e. clauses and sentences; this is the task of syntactic parsers). While word-related annotation is the standard in most corpora of contemporary language, annotation beyond individual words is not very widespread since it requires more resources and is prone to errors. Corpora that are completely and reliably annotated at the sentence level are known as tree banks (cf. Lemnitzer/Zinsmeister 2010: 75–84).

The second aspect is the search engine that linguists or lexicographers use to submit their queries. The following search options are possible with linguistic search engines, although not all of these options are realised everywhere.

- Searching for a surface form (*gibt* 'gives') or for a lemma (*geben* 'to give'). In the second query, all of the surface forms in the paradigm of the lemma are evaluated as hits and the corresponding text extracts are displayed (*geben* → *gebe*, *gibst*, *gibt*, *gab*, *gegeben*, etc.).
- Searching for a *lexical form* or word class (*Entscheidung treffen* 'to take a decision' or *Entscheidung* $p[8]=verb). In the second case, phrases with *Entscheidung* followed by a verb such as *Entscheidung fällen* 'to draw a decision', *Entscheidung drängen* 'to push for a decison' are outputted. The potential of this kind of query becomes clearer if we note that this kind of search machine can formulate concepts such as "keyword and preceding/following verb" or "verb at a maximum distance of 3 words from the keyword".
- Searching for or in a syntactic structure (e.g. "*schnell* 'quick' in an adverbial phrase" or "adverbial phrase in the pre-field"). This kind of query requires specialised search tools for tree banks (e.g. TIGERSEARCH for German and DACT for the Dutch Alpino corpus; for further details, cf. Lemnitzer/Zinsmeister 2010, section 4).

The last example, in particular, shows that a query in a corpus that has been annotated linguistically and that registers the desired hits makes certain demands on those undertaking the search. Common linguistic concepts such as "subordinate clause" or "imperative sentence" are not usually available in the corpus query and, if required, can only be formulated approximately. In this way, successful queries for grammatical structures assume good knowledge of the query language and of (the linguistic annotations in) the underlying corpora. We shall demonstrate this using some simple and some somewhat more complicated examples relating to grammatical information in dictionaries.

In standardised entries in comprehensive dictionaries of general language, the form section gives *information about inflection*. In print dictionaries, this is mostly achieved by giving those variant forms of the headword that enable educated users to determine all other paradigmatic forms. In German, these are the genitive singular and nominative plural (*Schuh*, *-s*, *-e* 'shoe'). In an Internet dictionary, where there is more space, the full forms in the inflectional paradigm can be given (*Schuh*, *Schuhs*, *Schuhe*, etc., like in the German WIKTIONARY, for example), which is presumably more user friendly. In certain circumstances, generic information in the form section has to be restricted to individual meanings (e.g. in the case of *Sand* 'sand', the plural *Sände* 'sands' cannot be formed for all of its meanings). Alternatively, the exact form can be given for individual meanings. See, for example, *Wasser* 'water' in ELEXIKO and the "Grammatische Angaben" for individual meanings: the entry immediately leads us to a

**8** "$p" refers to the underlying corpus representation of the part-of-speech of an individual word (token).

further piece of information that is relevant for the inflection of nouns: *number restrictions*. Some nouns are used exclusively in the plural (e.g. *Kosten* 'costs') and others primarily in the plural (e.g. *Süßwaren* 'sweets') while some words are used only in the singular (e.g. *Plastizität* 'plasticity') or primarily in the singular. Let us illustrate the last of these cases with an example from the *-politik* 'politics' group of compounds. The plural of *Agrarpolitik* 'agricultural policy' is certainly rare but it is still attested on multiple occasions in the DWDS-KERNKORPUS, where the plural is used to designate 'comparable fields of political action in multiple states'.

In the case of some nouns, loanwords in particular, there are *inflectional variants* in the singular or plural. For example, the typical English suffix *-ing* demonstrates variation in the genitive singular as in *Outing*/*Outings* (in its meaning of revealing the sexual or gender identify of a person); and the older loanword *Bonus* 'bonus' has variation in its plural forms: in addition to the more usual plural *Boni,* the native plural formation *Bonusse* exists, albeit rarely.

Corpora can reveal the existence of these singular/plural forms and singular/plural variants. A precise search based on form leads to initial results. These results then have to be interpreted and assessed with necessary caution, first of all because if a particular form is attested only once in a very large corpus, it might be, for example, an idiosyncrasy specific to an individual speaker or a straightforward error. The situation has to be assumed to be similar if there are multiple examples but these all appear in only one text. In this way, it is necessary to look out for a minimum number of occurrences and a sufficient distribution of examples across texts or text types. Rare findings which we distrust should be verified by research in other (reference) corpora. Second, the search for a particular form in a paradigm is made more difficult by the fact that this form can "occupy" multiple positions in that paradigm. So, for example, *Outings* is the form for both the genitive singular and all cases in the plural. A more precise enquiry than *Outings* alone that includes the relevant article mostly leads to the desired hits. However, it should be kept in mind that a narrower query will not find all of the occurrences in the corpus, such as when it occurs in the genitive in a noun phrase with a pre-head modifier between the article and the head (e.g. *des längst fälligen Outings* 'the long overdue outing'). Third, with the current state of technology, a linguistic search engine can only find (word) forms and not the individual meanings of a keyword. As such, when searching for a particular form with a particular meaning, we are faced in certain circumstances with many irrelevant examples (cf. also Lemnitzer 2022: 355–356).

Thus, a large corpus enables us to establish whether a particular form is used in the paradigm of a lexical unit or not; it is also possible to establish whether a particular form is used frequently or rarely relative to another form. This is interesting because rare phenomena (rare plural forms like the *-politik* compounds) should, in any case, be indicated in a dictionary. However, it is also possible to use a scale of relative frequencies to mark all notable *frequencies*. In ELEXIKO, this is attempted at the level of individual meanings (see www.owid.de/wb/elexiko/glossar/Grammatik.html). How-

ever, when we make a judgement such as "occurs (relatively) rarely" or read that as a user, we have to be aware that the dictionary basis can only ever illustrate the dictionary object imperfectly. For example, it might be the case that a form that appears relatively rarely in the section of language represented by the corpus occurs noticeably more frequently in other sections or varieties (technical language, youth language, Internet-based communication, etc.). As a result, lexicographic judgements like this are always limited to the dictionary basis and thereby susceptible to possible revision if the base data are extended.

To round off the topic of grammatical information, let us consider noun + preposition combinations and subordinate clauses with *ob* 'whether' or *dass* 'that' as two examples that require structural searches. First, an *analysis of the prepositions* used after the noun *Anfangsverdacht* 'initial suspicion' produces combinations with the preposition *auf*. This is to be expected since the root word *Verdacht* also allows this type of prepositional connection. What is not expected is a combination with the preposition *für* since this is not inherited from the root word. This is due to a specific legal use of the word, the typical combination for which is as follows: *Anfangsverdacht für eine Straftat* 'reasonable suspicion for a crime'. It is absolutely essential to include this collocational information for this keyword in a dictionary. Data of this kind can be identified in the DWDS-Wortprofil (Didakowski/Geyken 2014), for example, which draws on corpora that have been analysed and annotated syntactically.

Second, *subordinate clauses* introduced by *ob* have a propositional content whose facticity is put into question (*sie fragten mich, ob ich den Unfall gesehen habe → ?Ich habe den Unfall gesehen* 'they asked me whether I had seen the accident → ?I saw the accident'). In contrast, subordinate clauses introduced by *dass* have a propositional content that is assumed as given (*ich sagte ihnen, dass ich den Unfall gesehen habe → Ich habe den Unfall gesehen* 'I told them that I had seen the accident → I saw the accident'). With verbs of a propositional attitude that assume the facticity of that proposition (e.g. *wissen* 'to know'), the combination with a subordinate clause introduced by *ob* ought to be excluded. Corpus research produces counterexamples, however. The verb *wissen* 'know' can govern an *ob* subordinate clause if the verb itself is used in the matrix sentence in the preterite, in combination with a modal verb (*möchte wissen* 'would like to know'), or with a negator (*weiß nicht, niemand weiß* 'do not know, nobody knows', etc.). These dependencies between the verb in the matrix sentence, the resp. conjunction and the negator in the subordindate clause can only be understood by scrutinizing all examples from the DWDS corpora; here, the DWDS query reads as follows: "wissen #5 ob", that is, *ob* at a maximum distance of five words from (a form of) *wissen.* Unfortunately, the search query "*wissen* in the main clause with a modal verb or negator" cannot be posed to corpora in that way. Here we reach the limits of corpus annotation and search facilities.

## 7.4.2 Content-based information classes

### Meaning paraphrase/definition

One of the most difficult types of information to compile for a given entry in a monolingual dictionary is the *meaning* paraphrase or *definition*. Its text must be informative but should not be too long nor expressed in vocabulary that is too complicated. The last aspect applies in particular to meaning paraphrases in dictionaries for learners.

To what extent can corpora assist in dealing with this difficult task? One way relates to the fact that words are often defined in many texts, e.g. in text books and journalistic texts. That is to say, their meaning is described when the author assumes that a word (in a specific meaning) is unknown to the reader. For a number of years, computational linguistics has focused on automatically identifying definitions in texts (cf. the dissertations written by Cramer (2011) and Walter (2011), which both relate to German texts). The usual approach here is to search for grammatical and lexical patterns that are typically used to define word meanings ("*Unter* X *versteht man*" 'X is considered as', "*ein* X *ist* NP" 'X is an NP', "*Sei* X" 'Let X be', etc.). As such, we can talk about typical *definitional contexts*. Of course, not all definitions are found in this way, and not all the extracted text locations are really definitions. Nevertheless, an *automatically extracted definitional context* can be helpful for the lexicographer, for one thing as an aid to understanding the word being described; for another as an aid to formulating the definition being written.

### Collocations

Collocations have been an object of research for decades in theoretical linguistics, lexicology, lexicography, and corpus linguistics. An early definition of the concept comes from the school of British Contextualism (Firth 1957), where the concept of a collocation applied to typical co-occurrences. The concept was taken up by continental European lexicography and its content made more precise in order to work on the practical lexicon and dictionaries (cf. Hausmann 1984; 2007). Collocations were characterised here as:

> normtypische phraseologische Wortverbindungen, die aus einer Basis und einem Kollokator bestehen. Die Basis ist ein Wort, das ohne Kotext definiert, gelernt und übersetzt werden kann. Der Kollokator ist ein Wort, das beim Formulieren in Abhängigkeit von der Basis gewählt wird und das folglich nicht ohne Basis definiert, verwendet und übersetzt werden kann [norm-typical phraseological word combinations that consist of a basis and a collocator. The basis is a word that can be defined, learned, and translated without context. The collocator is a word that, in its formulation, is chosen in dependency on the basis and that, it follows, cannot be defined, learned, or translated without the basis.] (Hausmann 2007: 218).

Word pairs such as *Tisch*;*decken* (table;lay) or *Haar*;*dichtes* (hair;thick) are examples of collocations. The first word in each of these example pairs denotes the basis (*Tisch*, *Haar*), the second the collocator (*decken*, *dichtes*). An important characteristic of the collocation is the directedness of the basis to the collocator. In a situation where language is being formulated, we start from the basis in order to find the appropriate collocator – not the other way round. For example, we would not search for all the nouns that one can *commit* but rather we would proceed from the nouns, that is, from *crime*, *sin, murder,* etc., in order to find the correct verb.

Collocations can be *semantically* fully *transparent* but, as part of language norms, they are not arbitrary. This becomes apparent when collocations are translated into another language. For example, the adjective *dicht* 'dense' in the sense of *dichtes Haar* is rendered as *thick* (*hair*). We say *Tisch decken* 'to cover the table' but not *legen* 'lay' (as in the French *mettre la table* and English *to lay the table*). However, they can also be partially transparent. Examples here are *schwarzer Kaffee* (black coffee) or *blinder Passagier* (blind passenger), whereby the base words *Kaffee* or *Passagier* retain their literal meaning, but the meanings of *schwarz* in the sense of 'without milk' and *blind* in the sense of 'non-paying' do not follow on logically from the meaning of the collocators. The distinction between collocations and idioms, or idiomatic phrases, stems from the fact that the former are transparent or partially transparent whereas idioms are semantically opaque. Another distinguishing feature lies in the fact that collocations always possess a transparent basis whereas the semantic reach of an idiomatic phrase can only apply to the expression as a whole. This applies to phrases like *den Löffel abgeben* (literally: to give up the spoon; to die) or polyleximatic phrases, like *schwarzes Gold* for crude oil. For these reasons, it is important that collocations are described in a dictionary of general language. This information is needed and looked up primarily for text production and language learning.

The corpus-based description of collocations can be traced back to the late 1980s. Based on the larger volumes of textual language data that became available at this time, it was possible to evidence and describe collocations in language usage (Sinclair 1991).

In the process, simple *statistical processes* were used for the first time in order to identify collocations based on their frequency (Dunning 1993). The so-called Mutual Information Measure rates the co-occurrences of two words A and B more highly if these occur together more frequently than would be expected statistically. The various statistical measures underwent refinement and systematic comparison in the late 1990s and early 2000s (cf. Evert 2005). Here, two fundamental problems emerged. First, the accuracy of hits in the processes for recognising collocations was unsatisfactory. In the statistically highly significant cases, the collocations extracted in this way corresponded overwhelmingly with collocations, but in the broader range of word combinations with very low statistical significance, the number of word pairs that cannot judged to be collocations in the narrow sense and that would not be included in a dictionary was very high. What is striking here is the high number of banal oc-

currences such as *große Stadt* (big town), *Bier kaufen* (buy beer), or *neues Hemd* (new shirt). Second, with a size of 10–50 million tokens, the corpora used at the time were too small to achieve coverage appropriate for a dictionary of general language. Many current collocations included in dictionaries simply did not appear in corpora of this size and so could not be captured by the statistical models. The problem of insufficient coverage has been remedied in recent years by the construction of very large linguistic corpora (→ Section 7.3). At the moment there is not enough empirical evidence to answer how large text corpora have to be in order to achieve an appropriate coverage of collocations, but there are some relevant empirical values to draw on. Various studies report that a statistically valid and secure co-occurrence profile can only be extracted for words with an occurrence frequency of over 1,000 (Kilgarriff et al. 2004; Ivanova et al. 2008; Geyken 2011). For corpora consisting of billions of words, that means that a sufficiently large coverage would exist for around 20,000 keywords (Kilgarriff/Kosem 2012). Using a random selection of 231 low-frequency headwords from the Oxford Advanced Learner's Dictionary (OALD), the same study demonstrated that corpora extending to at least 10 billion words would be needed to describe the collocations of these words.

This problem of a lack of accuracy in hits in the automatic extraction of collocations from corpora has not yet been resolved satisfactorily. This is connected to that fact that the concept of a collocation is too broadly defined for an automated process. Already in the framework of British Contextualism, the very broad notion of collocations was made more precise through the term colligation (combinations of lexical items and grammatical features, cf. Greenbaum 1970). As a result, many of the common tools for automatically extracting possible collocations actually work by extracting colligations.

Probably the best known method for extracting syntactic co-occurrences is Sketch Engine (Kilgarriff et al. 2004), a method which can extract and classify co-occurrences in a targeted way following grammatical patterns. In other words, only those co-occurrences are considered that exist in a pre-defined syntactic relation. These relations could, for example, be adjective-noun, verb-object, noun and genitive attribute, or verb and prepositional object combinations. Although Sketch-Engine platforms exist for a large variety of languages, including English, Czech, Japanese, and Chinese, a straightforward transfer of the Sketch-Engine approach from English to German (and potentially other languages as well) is difficult. There are two main reasons for this: free word order in German and case syncretism. Both mean that, unlike with English, extracting syntactic relations on the basis of word classes and the sentence patterns dependent on them does not lead to satisfactory results. Experiments with Sketch Engine for German have shown that, depending on the parameters set, either the accuracy of analysis is insufficient or the coverage, or the proportion of texts that can be analysed, is too small (Kilgarriff et al. 2004; Ivanova et al. 2008). For this reason, the two existing approaches for extracting syntactic relations from large German text corpora are based on a general formalism that can recognise syntactic sentence functions and resolve

local ambiguity of meaning (Ivanova et al. 2008; Geyken et al. 2009). The first is the approach developed at the University of Stuttgart to extract "significant word pairs as a web service" (Fritzinger et al. 2009), which is based on the dependency parser FSPAR (Schiehlen 2003); the second process is the DWDS-WORTPROFIL (Geyken et al. 2009; Didakowski et al. 2012), developed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).

In the following, we shall describe the tool based on the second approach, the DWDS-WORTPROFIL, in more detail. This process is integrated into the standard view of the DWDS website and serves first and foremost as the basis for lexicographic work on the DWDS project. However, it is also available to external users for other purposes when consulting the corpus. Let us first outline the process before exploring the current coverage of the word profile.

The calculations of the DWDS-WORTPROFIL proceed in three stages, which are described in full elsewhere (Didakowski/Geyken, 2014):

1. Deciding which syntactic relations are to be extracted. Twelve types of relation are used including ATTR (adjective-noun), GMOD (noun-noun in genitive), OBJA (verb-noun as direct object), PRED (predicative), or VPP (verb-preposition-noun).
2. Using a syntax parser to annotate the syntactic relations. Until 2021 DWDS-WORTPROFIL was based on SynCoP (Syntactic Constraint Parser, Didakowski 2007), a parsing formalism founded on syntactic tagging. In 2022, SynCoP was replaced by transformer-based methods (Nguyen et al. 2021). Simple filter techniques are used to extract the relevant syntactic relations from the full dependency parse tree.
3. Using two statistical measures for the DWDS-WORTPROFIL to rank the results by their relevance: logDice (Rychlý 2008) and MI-log (Kilgarriff/Rundell 2002). These statistical measures are used as a quantitative measure of the cohesiveness of word tuples (pairs or triples): the higher the value (salience value, or sal for short), the stronger the association. A negative value (sal<0) stands for a negative strength of association (Evert 2005). A frequency threshold (default f=5) is introduced for the minimum frequency of occurrence in order to improve the quality of the results. This is based on the experience that word tuples with too low an absolute frequency can reduce the quality of the results (for more information, see also Kilgarriff/Kosem 2012).

The basis of the DWDS-cf. (as of 2023, https://www.dwds.de/b/dwds-wortprofil-in-neuer-version/) is a corpus of 6 billion words (essentially a newspaper corpus). From this, stages 1–3 above were carried out to compile a database of 56 million different syntactic co-occurrences. With this, queries can be run for co-occurrences of around 900,000 different lemmas. → Fig. 7.3 shows a screenshot for the word *grau* (grey) with two example relations, ATTR and PRED. In turn, the collocators are connected to the corpus examples so that it is possible to go back directly to the basis of the result.

| ist Adjektivattribut von | | logDice | Freq. |
|---|---|---|---|
| 1. Eminenz | M W A | 10.2 | 4109 |
| 2. Haar | M W A | 9.9 | 7337 |
| 3. Maus | M W A | 9.3 | 2318 |
| 4. Vorzeit | M W A | 9.3 | 2172 |
| 5. Anzug | M W A | 9.2 | 2728 |
| 6. Kapitalmarkt | M W A | 8.7 | 1586 |
| 7. Himmel | M W A | 8.4 | 1843 |
| 8. Zelle | M W A | 8.2 | 1361 |
| 9. Alltag | M W A | 8.1 | 1242 |
| 10. Star | M W A | 8.1 | 1365 |
| 11. Wolf | M W A | 8.1 | 1031 |
| 12. Bart | M W A | 8.0 | 926 |

| ist Prädikativ von | | logDice | Freq. |
|---|---|---|---|
| 1. Himmel | M W A | 7.4 | 615 |
| 2. Haar | M W A | 6.5 | 314 |
| 3. Katze | M W A | 6.1 | 240 |
| 4. Theorie | | 5.5 | 156 |
| 5. Bein | M W A | 5.3 | 137 |
| 6. Gesicht | | 5.2 | 136 |
| 7. Schnabel | M W A | 4.9 | 104 |
| 8. Unterseite | M W A | 4.6 | 88 |
| 9. Kopf | | 4.6 | 94 |
| 10. Oberseite | | 4.2 | 65 |
| 11. Farbe | | 4.2 | 67 |
| 12. Welt | | 4.2 | 79 |

**Fig. 7.3:** Table view in DWDS-Wortprofil: the word *grau* 'grey' used attributively with nouns such as *Haar* 'hair', *Maus* 'mouse', *Vorzeit* 'prehistory', *Anzug* 'suit', etc. (e.g., *graues Haar* 'grey hair') and predicatively with nouns such as *Himmel* 'sky', *Katze* 'cat', *Theorie* 'theory', etc. (e.g., *der Himmel ist grau* 'the sky is grey').

Geyken (2011) includes a first attempt to undertake a comparison of the results of the word profile with the "Wörterbuch der deutschen Gegenwartssprache" (WDG). This is interesting insofar as the WDG has always been valued for providing a good coverage of collocations (cf. Kramer 2011). Using the example above of the adjective *grau*, we can show in an exemplary fashion how the automatically extracted relations can be ranked in terms of quality. First of all, the quantitative comparison shows the following: in the DWDS-Wortprofil[9] 7,727 relations were extracted, with 398 different relations (f > 4, sal > 0). The corresponding dictionary entry in the WDG contains 39 different typical word combinations. There are 30 collocations that overlap between the two sets of results. The remaining 9 that do not appear in the word profile results are combinations like *grauer Stoff* 'grey fabric' or expressions like *in Ehren grau geworden* 'turned grey in honour'. Interestingly, there are current, semantically near-equivalent alternatives for these in the corpora that form the base of the Wortprofil. such as *grauer Flanell* 'grey flannel' or *graue Wolle* 'grey wool' and *in Ehren ergraut* 'greyed in honour', none of which is included in the WDG. On the other hand, the word profile results include a whole range of salient and current combinations that have the status of collocations but are not included in the dictionary. There are 44 (or 132) co-occurrences with a salience <10 (or >5), of which quite a few have the status of a collocation. Examples include: *graue Eminenz* 'eminence grise', *graue Zellen* 'grey cells', *graue Schläfen* 'grey temples', *graue Asche* 'grey ash', or *grauer Markt* 'grey market'. This example is representative of many others categorised as being very frequent words, that is, those that are attested with a frequency of more than 1,000 in the corpus. Of course, in defence of print dictionaries like the WDG, it must be considered that print space is limited and that the selection of collocations therefore had to be very restrictive. When presenting collocations in Internet dictionaries, a lexicographi-

**9** This prototype was based on a 500-million word corpus. The database contained 2 million co-occurrences for 90,000 lemmas.

cally informed selection must be made from the array of co-occurring word pairs (cf. Klosa/Storjohann 2011).

The CCDB (co-occurrence database) developed at the Leibniz Institute for the German Language is comparable to the approaches described above but with some differences. This service is comparable insofar as very large corpora of contemporary language form the foundation for the data, which are also analysed statistically in order to find salient word combinations. A fundamental difference is that the corpus base is not tagged and, thus, the co-occurrence pairs extracted in the results cannot be sorted by syntactic relations.

A feature of the CCDB not available in other tools is the attempt to group the collocators of a basis word (automatically) according to meaning nuance. The results for the keyword *grau* are shown in → Fig. 7.4.

The connection of *grau* with items of clothing (top right), body hair (bottom right), and with other shades (bottom left) can be seen clearly. A detailed presentation of this service with further examples can be found in Perkuhn et al. (2012: 132–136).

## Examples

Whether and in what way the corpora and extraction tools available today are of use in gathering examples for the information category of *attested examples* depends on the function that this information has for the dictionary entry. We distinguish examples that illustrate the meaning from examples are given to prove a statement made elsewhere about the word being described.

Given the current state of technology, the examples that a search engine identifies in a corpus as “hits” and then displays in a larger or smaller context (KWIC lines, sentences, multiple sentences, whole text) are not separated according to the *different meanings* of the keyword. For the most part, the differentiation between multiple meanings of a headword is a genuine accomplishment on the part of the lexicographer when describing the word and can only be applied retrospectively to the corpus or the extracted examples. However, processes of automatic recognition for different meanings of a lexeme (cf. Henrich/Hinrichs 2012) can provide the lexicographer with valuable indications for differentiating meanings in that these methods group (or *cluster*) examples in “similar” contexts of use. Again, given the current state of technology, the results of these clustering methods do not match the intuition of lexicographers sufficiently. As such, there remains the option of a manual search for examples with a particular meaning in what is often a large number of examples. This can prove to be difficult and time consuming when one meaning is clearly attested more frequently than all of the others. However, it is possible to make it easier by making the search more specific. If we search for *Avatar* in the DWDS corpora, we overwhelmingly find examples in which the word denotes a ‘representative of a real person in the virtual world’.


**Fig. 7.4:** Grouping of collocators for the keyword *grau* 'grey' according to meaning nuance in a "self-organising map" in CCDB.

However, if we know that *Avatar* originally denoted something like a god, we can search for shared occurrences of *Avatar* and *Gott* 'God' in a sentence and obtain (a few) examples that allow a second meaning to be formulated.

The flipside of the scarcity of examples for a particular keyword or meaning is high frequency examples for others. Many keywords occur so frequently in large corpora that reviewing them from a lexicographic standpoint goes beyond the limited time that is usually available in a project to process a headword. In these cases, an informed *pre-selection* of examples makes the work considerably easier. Kilgarriff et al. (2008) first introduced an automated method to extract good examples from corpora, called "Gdex" (good dictionary examples). The underlying algorithm sorts all the concordance sentences for a given headword according to a "goodness score". Each sentence is pro-

vided with a goodness score that depends on several parameters, including length, use of complicated vocabulary, and absence of pronouns and absence of named entities (proper nouns). The algorithm was subsequently refined (e.g. Kosem et al. 2019) and can be parametrised by its users. We employ this method on the basis of an adaptation of Gdex to German that was developed by Didakowski et al. (2012) and that is used in the DWDS project for selecting examples for a given keyword. The results can be accessed on the DWDS project website in the section headed "Gute Beispiele" 'good examples'. If we want to attest a particular statement or claim and if the statement relates to a *rare*, but precisely notable, property of the word, the search can very quickly become extremely complicated and resemble looking for a needle in a haystack. As examples, we can take those already listed in → Section 7.4.1: the rare singular forms (*die Süßware* 'a piece, item of sweet', in contrast to the non-countable plural *Süßwaren 'confectionary'*), rare plural forms (*die Wässer* 'waters'), or rare variants (genitive of the word *Outing* in German).

The limitations of print space no longer apply to dictionaries compiled for publication on the Internet. This becomes significant in other ways as well but particularly in the number and length of examples. In a print dictionary, the examples have to be strictly chosen and edited with the limited print space in mind. The latter can occur at the expense of comprehension if the relevant word is not able to be presented with sufficient context. The fact that these restrictions are removed in the online medium, therefore, has implications for the lexicographic process (→ Chapter 3), in this case, in the selection and *processing of examples*. However, when processing examples, aspects of user friendliness also have to be taken into account. Examples that are too numerous and too long could possibly discourage users from reading them or distract them from the aspect of language use that is actually being documented. This is an area that should be investigated more closely by research into dictionary use, although Klosa et al. (2014) have already been able to relieve some of these concerns in their user studies.

### Lexical-semantic relations

In many dictionaries of general language we find information about words with which the headword has a *lexical-semantic relation*. In what follows, we restrict ourselves to paradigmatic relationships and, in particular, to information about antonyms and synonyms. Lexical-semantic relationships between lexical signs are a structural feature of the language system and, more specifically, of the lexicon. These relationships can be presumed to structure the (individual) mental lexicon of each speaker of a language as well. For each language, including German, there are specialist lexical resources known as *word nets* that use these lexical-semantic relationships as their primary structuring feature. Further details are presented in Kunze/Lemnitzer (2007: 135–141).

In this case, it is not obvious that relationships of a language-system nature between lexical units can be "found" in texts and extracted from them. However, there have been a pleasantly high number of attempts in computational linguistics to operationalise the concept of *semantic relations* to the extent that examples for pairs of semantically linked lexical units can be extracted from textual corpora. The method involves defining structural patterns within which pairs of lexical units with a lexical-semantic relationship to each other typically occur. Jones (2010) chose this approach in order to locate *antonym pairs* in English data in relation to English extraction patterns (which he calls *frames*). Some aspects can be transferred to German: antonym adjective pairs often appear in the "*weder* ADJ *noch* ADJ" 'neither ADJ nor ADJ' pattern that signals a contrast.

Of course, as well as true antonym pairs, we also end up with a variety of occasional contrastive expressions as the result of an appropriate corpus query so that careful selection and checking of the data are necessary. It is also possible, of course, to orient the search in a targeted way on a particular lexical sign. A search for the adjective *groß* 'big' (DWDS search: "weder groß noch $p=ADJD") results in numerous hits for *weder groß noch klein* 'neither big nor small' and *weder größer noch kleiner* 'neither bigger nor smaller' in addition to some more occasional formulations.

Textual patterns for the synonyms of lexical signs are notably more difficult to find. Storjohann (2010) proposed some examples, attesting them with data from the corpus she used, but the "patterns" are either impossible to operationalise as corpus queries or they are too imprecise to identify synonyms in the narrower sense. The team at the WORTSCHATZ-LEIPZIG project (cf. Biemann et al. 2004) followed a more general approach to identifying paradigmatic relations. They also made reference to the contexts in which a keyword occurs but considered the words that occur with the keyword with a frequency greater than chance ("co-occurrences") and, in a further step, also the co-occurrences of these co-occurring words. The expectation was that these words will exist in a semantic relationship with the original keyword. For example, 25 synonyms are given for the word *fleißig* 'hard-working'. The results of the automatic synonym extraction can be examined on the project website.[10] Synonym data of this quality is certainly helpful starting material for compiling the corresponding information in a dictionary, but it most certainly requires selection and evaluation by lexicographers.

Overall, it is worthwhile further pursuing research and development in extracting lexical-semantic relations from large textual corpora. At the moment, this is a very active research field. Even if not all of the approaches and methods are suitable for lexicographic purposes, it can be assumed that one or the other impulses can be taken from there to shape our own corpus searches.

**10** E.g. https://corpora.uni-leipzig.de/de/res?corpusId=deu_news_2023&word=flei%C3%9Fig for the word *fleißig* 'industrious'.

### 7.4.3 (Pragmatic) use-based information classes

The class of pragmatic information involves details that indicate particularities or restrictions in the use of a word with a particular meaning. As a whole this information is referred to as *diasystematic information*. In many dictionaries, the following types of details are created: information about temporal restrictions of use (= diachronic), information about spatial restrictions on use (= diatopic), information about restrictions on use to a particular (technical) discourse (diatechnical), information about the use of the word on a particular stylistic level (= diastratic). In addition, there is information about frequency (= diafrequent).

Diasystematic information can fulfil three functions in the process of compiling a dictionary. First, words marked with diasystematic information play a role in deciding on the selection of lemmas. Some caution has to be exercised when including words marked diasystematically in a learners' dictionary. If it is decided to include words marked as technical language, for example, it is necessary to take care to achieve a certain balance in these selections. Second, diasystematic information can be used when compiling an entry in order to delimit a meaning or spelling variant used in more specialised contexts from a more general meaning or spelling variant. Third, it can be used to define subsections of vocabulary with this markup, for example when dividing up work in the lexicographic process (→ Chapter 3) or as a search option (→ Chapter 5.3) for users of the dictionary if it is available in digitised form (for more information on this, cf. Atkins/Rundell 2008: 182f. and 227).

We already established above, in relation to the group of grammatical and meaning-related details, that the corpus, or more accurately the primary data of the corpus, does not give direct answers to these questions, even more so in relation to information on the context of the utterances in which a particular word is used.

We also demonstrated in → Section 7.3 that a linguistic corpus is a structure with multiple levels, involving not only primary data but also metadata. With appropriate quality and detail, metadata describes, among other things, the situatedness of the text in time and space; it can also provide information about the type of discourse and the stylistic level of the texts in which a word is attested. Let us demonstrate with some examples the possibilities for diasystematic information opened up by metadata:

1. *Diachronic information*. In the DWDS there is a "word history graph" for which the section of metadata related to time (= the date a text appeared) is analysed. From that, we can learn that the word *Droschke* (cab, carriage) only occurs rarely in the second half of the 20$^{th}$ century, but the word *Streß* or *Stress* does not find widespread use until the 1960s and beyond. This kind of information is also provided in other places, for example, in relation to neologisms: cf. Steffens/al-Wadi (2013), the German Neologismenwörterbuch (NEO-OWID), and the dictionary of neologisms compiled by Quasthoff (2007, NEO-WB).
2. *Diatopic information*. Information about regional restrictions or preferences in the use of a word or variant, etc. can only be established indirectly from the cor-

pus metadata. Indications about these tendencies could be the provenance of a newspaper in which a word is predominantly used or the origin of authors who prefer to use a certain word. However, these indications have to be treated with caution and are best verified by speakers of the corresponding regiolect.

3. *Diatechnical information*. The use of a word in particular technical discourses can, in certain circumstances, be deduced from the author and title of the texts in which this word appears. Particular terms such as *discourse ethics* or *unconcealing* can be assigned not only to a particular discourse, but even to the characteristic wording of a particular author. However, deriving a technical area from these findings has to come with reservations since every corpus, even a large one, is incomplete compared to the language that it documents and cannot ever be representative. Similar considerations to those applied to diatechnical information also apply to information about the use of a word predominantly within a particular social or professional group (youth language, military language, etc.).
4. *Diastratic information* can also not be deduced directly from the metadata but requires careful analysis of many examples or recourse to the language competence of mother-tongue speakers. This applies to both stylistic level and tone.
5. *Diafrequent information* seems to be the information about usage where it should be possible to extract it most easily from a corpus: counting words is one of the easier exercises if the corpus is digitised. However, representing frequency values in corpora as frequency information in dictionaries is problematic in two respects: first, the frequency data in many dictionaries are not scalar but rather comparative ("more frequent in the plural") or nominal ("frequent/rare in the plural"). Second, many occurrence figures have to be considered relative to other figures such that if a word occurs only twice in the plural, can we refer to this as "rare(r)" if the occurrence figure for the singular lies in the region of three or four? Consequently, the frequency information in ELEXIKO is given on the basis of a quantifiable relative occurrence frequency in the underlying corpus, as described in → Section 7.4.1 in relation to grammatical information.

English lexicography, and especially learner lexicography, has also gone over to working with scalar values or frequency classes, visualising these in ways that are easily understood (e.g. in order to compare them to the frequency of quasi-synonymous words), cf. Bogaards 2008.

## 7.5 The limits of automatic methods and desirable future developments

In the previous sections, we have shown that the opportunities afforded by corpora to generate high quality information in a dictionary entry depend on the following as-

pects, i.e. the quality and detail of the metadata and the linguistic annotation of the primary corpus data and the options provided by (linguistic) search engines.

Not only the scale of a corpus, measured as the number of words, plays a role as a criterion for its suitability for lexicographic purposes but also the *diversity* of texts in it, for example, its distribution across different (technical) domains and different time periods as well as the coverage of different genres and stylistic levels. Lexicographically, niche areas in language development can be captured and recorded on the basis specialised corpora. This applies not only to genre-specific vocabulary but much more to genre-specific idiosyncrasies in the use of existing words, from peculiarities on the orthographic level to new meanings (cf., for example, the genre-specific use of the word *troll* to designate a person who tries to systematically disturb the discussion in online forums).

Past developments to extend corpora indicate that the future construction of corpora cannot proceed as a single project but in a coordinated way and thereby across institutions. The texts must be held in such a way that they can be corrected and annotated in an ongoing way and the metadata must contain statements about *quality*. Suggestions for how to compile a wider corpus infrastructure in this way can be found in Geyken et al. (2012), Krek et al. (2018), as well as in the language data infrastructure projects, including CLARIN-EU or elexis, with its federated content search infrastructure.

The *language technology tools* for applying linguistic annotations to corpus data will also develop further. This means better quality, that is, higher accuracy in linguistic annotations; improving the quality of analysis for texts not in the standard language, such as Internet-based communication; and capturing further levels of linguistic analysis. We can also expect a qualitative leap for the use of corpora in lexicography, in which different uses of words (in different classes) in different contexts of use are distinguished and annotated. It is worth keeping an eye on these developments and, above all, the resources that will be created through these efforts. How much effort will be needed to be able to extract rare or complex grammatical properties will also depend on the quality of annotations. We described examples of this in → Section 7.4.1.

## 7.6 Integrating primary sources into lexicographic resources

Up until now we have described the extraction of lexicographic information from corpora from the perspective of a traditional lexicographic process. Lexicographers are mediators between the primary data, which represent language use, and the users, who can – and must – rely on the selection and judgement of lexicographers.

In Internet-based *lexical systems*, the practice has become to publish primary data alongside the actual lexicographic data, that is the edited and compiled dictio-

nary entries (cf. Asmussen 2013), or to integrate primary sources directly into the lexicographic resource. In this case, we can talk of (heterogenous) word information systems or of *digital lexical systems* (Klein/Geyken 2010).

The integrated publication of dictionaries and primary sources has the advantage that the users of the dictionary can understand the decisions made by the lexicographers in relation to the primary data and can undertake their own research into the primary data in cases where there are gaps in the dictionary, form their own picture. This integration can be more or less complete. For example, in the DWDS, data that have been checked by lexicographers and automatically generated data are displayed in different windows or "panels". In ELEXIKO, automatically generated information, for example about the division of a word into written syllables is found integrated into the lexicographic product itself.

The advantages of integrating primary sources are accompanied by several disadvantages (for more information, cf. Asmussen 2013: 1082f.). As documents of language use, corpora are full of idiosyncrasies and errors; when reviewing this data, lexicographers abstract from those inconsistencies. Further *errors* arise during the process of automatically annotating and analysing the data since no language technology tool can operate without making mistakes. Statistical tools like the DWDS-WORTPROFIL produce correct – that is, statistically significant – data from a statistical point of view, but this can be irrelevant for describing a headword.

As such, users find themselves confronted by a mixture of reliable information (interpretations of the raw data by the lexicographers that are recorded in dictionary entries) and less reliable raw data (from the primary sources and analysis tools). It is no small achievement to be able to draw the line between what is reliable and what is unreliable. In particular, the following discrepancies can arise between the data from the dictionary basis (raw data) and the dictionary data (processed data):

1. Forms of use for a word can be found in the dictionary basis that, for whatever reason, the lexicographers did not take into account.
2. The data contain words that are not described in the dictionary. This is the result of word or lemma selection during the lexicographic processing of the data. No dictionary of general language can process all the words that occur in a corpus, especially as the range of words covered by a corpus grows with every text that is added to it (for more on the relationship between corpus size and the size of the lexicon, cf. Kunze/Lemnitzer 2007: 189–191; Geyken 2008).
3. The data of the dictionary basis contain usage that deviates from prescribed norms, for example spellings that do not correspond to the norm described in the dictionary.
4. Processing the dictionary basis with language technology tools introduces further errors that are not always apparent to users. For example, during lemmatisation, full forms may be mapped onto a false root form. As a result, when searching for examples relating to a root form ("lemmatised search"), the user also obtains forms that do not belong to the root form. The ambiguity of word forms also pro-

duces allocations that seem bizarre but are systematically correct. Under certain circumstances, for example, all occurrences of *heute* are assigned to the root form of the verb *heuen* 'to make hay', rather than to the adverb *heute* 'today'. This makes searching for examples for the keyword *heuen* difficult, if not impossible.

While expert users of language corpora who are using a lexical system for their research know how to deal with these discrepancies, they can cause confusion for users who expect to be presented with entirely reliable information about linguistic norms and about "correct" usage when they "look things up". An extreme reaction to this confusion, but one which is presumably not unusual, is to dismiss the resource altogether, since it (apparently) "provides false information" (more on the user's perspective in relation to these hybrid information systems can be found in → Chapter 9).

Users' reactions represent a particular challenge for designing a lexical information system. There are several possible ways to clarify the differing quality and reliability of different parts of these resources to users.

1. When entering the digital lexical system, the user is presented initially only with verified lexical information, that is, with the dictionary, but at the same time access to further sources is also made possible.
2. The editorial texts indicate the different provenance and, therefore, quality of the data; however, it is well known that attention is hardly ever paid to editorial texts.
3. Automatically generated, unverified information or its source is presented in a different way graphically than verified information. The website LINGUEE displays pairs of equivalent (translated) sentences for German or English keywords, marking the sentence pairs that have not been verified with a small warning triangle. An alternative is to distinguish windows containing unverified data from those displaying verified data by using another colour or, as is the practice in ELEXIKO, by providing an explicit warning about their status. A similar strategy is to choose to differentiate between *verified* and *unverified information zones*.

Overall, the way dictionaries users manage the mixture of more and less reliable information has not been investigated sufficiently. An attempt to do this is presented in Klosa et al. (2014), where the difficulties of such studies are also reported more fully. However, to conclude our contribution, let us issue an appeal to researchers into dictionary use and dictionary education not to abandon their efforts in this area.

## 7.7 Epilogue

With the advent of Large Language Models (LLMs) such as OpenAI ChatGPT (where GPT stands for Generative Pre-Trained Transformer), many of the approaches to auto-

matically extract lexicographic information described in this chapter have to be revisited. Indeed, a recent survey by de Schryver (2023) provided evidence that many dictionary production tasks can be carried out by GPTs in an astonishing quality. One example that he cites is a study carried out by Rees and Lew (2023) in which they could show that GPT generated definitions were found the least satisfying compared to their hand-crafted counterparts (CCLED), both in terms of quality ratings and free-text comments. On other parts like the generation of dictionary examples the GPTs performed less well but this may be temporary as their GPT system was not yet trained on example generation. Is this "the end of lexicography", as the title of a publication of Jakubíček and Rundell indicates (Jakubíček/Rundell 2023)? At the moment, the authors state that this is not the conclusion to be drawn, especially for reliable, comprehensive, large reference dictionaries where polysemy has to be dealt with appropriately and rare and unusual meanings have to be included. Another very likely consequence, however, is one proposed by Nichols in an invited talk at the elex 2023 conference (Nichols 2023), where she recommends that dictionary producers train "AI on their good content and retrieve information in imaginative new ways to improve the customer's experience".

# Bibliography

## Further reading

Kilgarriff, Adam/Kosem, Iztok (2012): Corpus Tools for lexicographers. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 31–55. *Overview of corpus tools for lexicographers.*

Wiegand, Herbert Ernst (1989): Formen von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*. 1. Teilband, Berlin/New York: De Gruyter, 462–501. *This handbook entry presents a model of abstract microstructures and their realisation in the form of specific microstructures with series of information. We orient ourselves on this model in our presentation of the information classes for which information from corpora might be helpful*.

## Bibliography

### Academic literature

Asmussen, Jörg (2013): Combined products: dictionary and corpus. In: Gouws, Rufus H., et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/Boston: Mouton de Gruyter, 1081–1090.

Atkins, B. T. Sue/Rundell, Michael (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Biemann, Chris/Bordag, Stefan/Quasthoff, Uwe (2004): Automatic Acquisition of Paradigmatic Relations using Iterated Cooccurrences. In: *Proceedings of LREC2004, Lisboa, Portugal*. Lisbon: European Language Resources Association (ELRA), 967–970. http://www.lrec-conf.org/proceedings/lrec2004/pdf/549.pdf [last access: May 2, 2024]

Bogaards (2008) = Bogaards, Paul (2008): Frequency in Learners' Dictionaries. In: *Proceedings of the 13th EURALEX 2008 conference.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 1231–1236.

Cramer, Irene (2011): *Definitionen in Wörterbuch und Text*. Dissertation. TU Dortmund. http://hdl.handle.net/2003/27628 [last access: May 2, 2024].

de Schryver, Gilles-Maurice (2013): Generative AI and Lexicography: The Current State of the Art Using ChatGPT. In: *International Journal of Lexicography* 36:4, 355–387.

Didakowski, Jörg (2007): SynCoP – Combining syntactic tagging with chunking using WFSTs. In: *Proceedings of FSMNLP 2007*. Potsdam: Universitätsverlag, 107–118. https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/2526/file/fsmnlp07proc10.pdf [last access: May 2, 2024].

Didakowski, Jörg/Geyken, Alexander (2014): From DWDS corpora to a GermanWord Profile – methodological problems and solutions. In: Abel, Andrea/Lemnitzer, Lothar (eds.): *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*. Mannheim: Institut für Deutsche Sprache, 43–52.

Didakowski, Jörg/Geyken, Alexander/Lemnitzer, Lothar (2012): Automatic example sentence extraction for a contemporary German dictionary. In: *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 343–349.

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

Engelberg, Stefan/Lemnitzer, Lothar (2009): *Lexikografie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Evert, Stefan (2005): The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. http://dx.doi.org/10.18419/opus-2556 [last access: May 2, 2024].

Firth, John Rupert (1957): Modes of Meaning. In: *Papers in Linguistics 1934–1952*. London: Longmans, 190–215.

Fritzinger, Fabienne (2009): Werkzeuge zur Extraktion von signifikanten Wortpaaren als Web Service. In: *GSCL Symposium Sprachtechnologie und eHumanities, Duisburg, 26.–27. Februar 2009*, 32–43.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (ed.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum, 23–41.

Geyken, Alexander (2008): Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus. In: Cori, Marcel/Léon, Jacqueline/David, Sophie (eds.): *Langages, Construction des faits en linguistique: la place des corpus* 171, 77–94.

Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In: Abel, Andrea/Zanin, Renata (eds.): *Korpora in Lehre und Forschung*. Bolzano: University Press, 115–137.

Geyken Alexander/Didakowski, Jörg/Siebert, Alexander (2009): Generation of word profiles for large German corpora. In: Kawaguchi, Yuji/Minegishi, Makoto/Durand, Jacques (eds.): *Corpus Analysis and Variation in Linguistics*. Tokyo: Benjamins, 141–157.

Geyken, Alexander/Gloning, Thomas/Stäcker, Thomas (2012): *Panel: Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints, Digital Humanities Conference*. Hamburg, Video Lecture.

Greenbaum, Sidney (1970): *Verb-Intensifier Collocations in English. An experimental approach*. Den Haag/Paris: Mouton.

Hanks, Patrick (2012): The Corpus Revolution in Lexicography. In: *International Journal of Lexicography* 25:4, 398–436.

Hausmann, Franz Josef (1984): Wortschatzlernen ist Kollokationslernen. In: *Praxis des neusprachlichen Unterrichts* 31, 395–406.

Hausmann, Franz Josef (2007): Die Kollokationen im Rahmen der Phraseologie – Systematische und historische Darstellung. In: *Zeitschrift für Anglistik und Amerikanistik* 55:3, 217–234.

Hausmann, Franz Josef/Wiegand, Herbert Ernst (1989): Component Parts and Structures of General Monolingual Dictionaries. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*. 1. Teilband. Berlin/New York: De Gruyter, 328–360.

Henrich, Verena/Hinrichs, Erhard (2012): Word Sense Disambiguation Algorithms for German. In: *Proceedings of the 8th conference on International Language Resources and Evaluation LREC 2012*. Istanbul: European Language Resources Association (ELRA), 576–583.

Ivanova, Kremena (2008): Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In: *Proceedings of the 6th Conference on Language Resources and Evaluation*. Marrakech: European Language Resources Association (ELRA).

Jakubíček, Miloš, et al. (2013): The Tenten Corpus Family. In: *7th International Corpus Linguistics Conference CL*. Lancaster: Lancaster University, 125–127.

Jakubíček, Miloš/Rundell, Michael (2023): The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In: Medvěď, Marek, et al. (eds.): *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023*. Brno: Lexical Computing CZ s.r.o., 518–533.

Jones, Steven (2010): Using web data to explore lexico-semantic relations. In: Storjohann, Petra (ed.): *Lexical-Semantic Relations. Theoretical and practical perspectives*. Amsterdam: Benjamins, 49–67.

Kilgarriff, Adam (2004): The Sketch Engine. In: *Proceedings Euralex 2004*. Lorient: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, 105–116.

Kilgarriff, Adam (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In: *Proceedings of the 8th EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, 425–433.

Kilgarriff, Adam/Kosem, Iztok (2012): Corpus Tools for Lexicographers. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 31–55.

Kilgarriff, Adam/Rundell, Michael (2002): Lexical Profiling Software and its Lexicographic Applications – a Case Study. In: *Proceedings of the 10th EURALEX International Congress*. København: Center for Sprogteknologi, 807–818.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: *Lexicographica* 26, 79–93.

Klosa, Annette/Koplenig, Alexander/Töpel, Antje (2014): Benutzerwünsche und -meinungen zu dem monolingualen deutschen Onlinewörterbuch ELEXIKO. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. BerlinBoston: De Gruyter, 281–384.

Klosa, Annette/Storjohann, Petra (2011): Neue Überlegungen und Erfahrungen zu den lexikalischen Mitspielern. In: Klosa, Annette (ed.): *ELEXIKO. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs*. Tübingen: Narr, 49–80.

Kosem, Iztok et al. (2019): Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. In: *International Journal of Lexicography* 32:2, 119–137.

Kramer, Undine (2011): Klappenbach/Steinitz, Wörterbuch der deutschen Gegenwartssprache. In: Haß, Ulrike (ed.): *Große Lexika und Wörterbücher Europas*. Berlin/Boston: De Gruyter, 449–476.

Krek, Simon, et al. (2018): European Lexicographic Infrastructure (ELEXIS). In: Čibej, Jaka et al. (eds.): *Proceedings of the 16th EURALEX International Congress*: *Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, 881–891.

Kunze, Claudia/Lemnitzer, Lothar (2007): *Computerlexikographie*. Tübingen: Narr.

Lemnitzer, Lothar (2022): *Erhebung, Aufbereitung und Auswertung von Korpusdaten*. In: Beißwenger, Michael/Lemnitzer, Lothar/Müller-Spitzer, Carolin (eds.): *Forschen in der Linguistik*. Paderborn: Brill/Fink, 350–360.

Lemnitzer, Lothar/Diewald, Nils (2022): Abfrage und Analyse von Korpusbelegen. In: Beißwenger, Michael/Lemnitzer, Lothar/Müller-Spitzer, Carolin (eds.): *Forschen in der Linguistik*. Paderborn: Brill/Fink, 374–390.

Lemnitzer, Lothar/Zinsmeister, Heike (2010): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Moreno-Ortiz, Antonio (2024): *Making Sense of Large Social Media Corpora. Keywords, Topics, Sentiment, and Hashtags in the Coronavirus Twitter Corpus*. London et al.: Palgrave Macmillan.

Mugdan, Joachim (1989): Grundzüge der Konzeption einer Wörterbuchgrammatik. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*. 1. Teilband. Berlin/New York: De Gruyter, 462–501.

Nguyen, Minh Van, et al. (2021): Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 80–90.

Nichols, Wendalyn (2023): 'Invisible Lexicographers, AI, and the Future of the Dictionary. In: *eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno, Czech Republic.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): *Korpuslinguistik*. Paderborn: Fink.

Quasthoff, Uwe (2007): *Neologismenwörterbuch*. Berlin/New York: De Gruyter.

Rapp, Reinhard (2003): Computersimulation sprachlicher Intuition. In: Cyrus, Lea, et al. (eds.): *Sprache zwischen Theorie und Technologie/Language between Theory and Technology*. Wiesbaden: Deutscher Universitätsverlag, 237–255.

Rees, Geraint Paul/Lew, Roibert (2024): The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. To appear in: *International Journal of Lexicography* 37.

Rychlý, Pavel (2008): A lexicographer-friendly association score. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 6–9.

Schäfer, Roland/Bildhauer, Felix (2012): Building large corpora from the web using a new efficient tool chain. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul: European Language Resources Association (ELRA), 486–493.

Schiehlen, Michael (2003): A cascaded finite-state parser for German. In: *Proceedings of the 10th EACL*. Budapest: Association for Computational Linguistics, 163–166.

Schmidt, Ingrid (2024): Modellierung von Metadaten. In: Lobin, Henning/Lemnitzer, Lothar (eds.): *Texttechnologie. Anwendungen und Perspektiven*. Tübingen: Stauffenburg, 143–164.

Sierra, Gerardo et al. (2008): Definitional verbal patterns for semantic relation extraction. In: *Terminology* 14:1, 74–98.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Steffens, Doris/al-Wadi, Doris (2013): *Neuer Wortschatz. Neologismen im Deutschen 2001–2010*. Mannheim: Institut für Deutsche Sprache.

Storjohann, Petra (2010): Synonyms in corpus texts. Conceptualisation and construction. In: Storjohann, Petra (ed.): *Lexical-Semantic Relations. Theoretical and practical perspectives*. Amsterdam: Benjamins, 69–94.

Walter, Stephan (2011): *Definitionsextraktion aus Urteilstexten*, PhD Thesis, Universität des Saarlandes. http://www.coli.unisaarland.de/~stwa/publications/DissertationStephanWalter.pdf [last access: May 2, 2024].
Wiegand, Herbert Ernst (1989): Formen von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*. 1. Teilband, Berlin/New York: De Gruyter, 462–501.
Wiegand, Herbert Ernst (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York: De Gruyter.

## Dictionaries

CaED = *Cambridge English Dictionaries*. https://dictionary.cambridge.org/ [last access: May 2, 2024].
CCELD = Sinclair, John, et al. (eds.): *Collins Cobuild English Language Dictionary*. London/Glasgow: Collins, 1987.
CED = *Collins English Dictonary*. https://www.collinsdictionary.com/dictionary/english [last access: May 2, 2024].
DDUW = *Duden – Deutsches Universalwörterbuch*. 8. Auflage. Berlin 2015: Dudenverlag.
DUDEN ONLINE = *Duden*. Berlin: Bibliographisches Institut/Dudenverlag. www.duden.de [last access: May 2, 2024].
DWB = *Deutsches Wörterbuch von Jacob und Wilhelm Grimm*. Leipzig: Hirzel.
DWDS = *Das Digitale Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de [last access: May 2, 2024].
ELEXIKO = Online-Wörterbuch zur deutschen Gegenwartssprache. In: *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. http://www.elexiko.de [last access: May 2, 2024].
LINGUEE = *Linguee Wörterbuch Englisch–Deutsch*. www.linguee.de [last access: May 2, 2024].
MW= *Merriam-Webster*. https://www.merriam-webster.com/ [last access: May 2, 2024].
NEO-OWID = Neologismenwörterbuch. In: *OWID-Online-WortschatzInformationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. http://www.owid.de/wb/neo/start.html [last access: May 2, 2024].
NEO-WB = Quasthoff, Uwe (ed.): *Deutsches Neologismenwörterbuch*. Berlin/New York: De Gruyter, 2007.
OALD = *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press. http://www.oxfordlearners dictionaries.com [last access: May 2, 2024].
OED = *Oxford English Dictionary online*. Oxford: Oxford University Press. http://dictionary.oed.com [last access: May 2, 2024].
WDG = Klappenbach, Ruth (ed.): *Wörterbuch der deutschen Gegenwartssprache*. Berlin: Akademie-Verlag.
WIKTIONARY = *Das deutsche Wiktionary*. de.wiktionary.org [last access: May 2, 2024].
WORTSCHATZ LEIPZIG = *Wortschatz*. Universität Leipzig. http://wortschatz.uni-leipzig.de/ [last access: May 2, 2024].

## Internet sources

AGD = *Archiv für Gesprochenes Deutsch*. Mannheim: Institut für Deutsche Sprache. www.agd.ids-mannheim.de [last access: May 2, 2024].
BNC = *British National Corpus*. www.natcorp.ox.ac.uk [last access: May 2, 2024].

CCDB = *Kookkurrenzdatenbank*. Mannheim: Institut für Deutsche Sprache. http://corpora.ids-mannheim.de/ccdb/ [last access: May 2, 2024].
CLARIN-EU = *CLARIN – The research infrastructure for language as social and cultural data*. https://www.clarin.eu/ [last access: May 2, 2024].
COW = *Corpora from the Web*. Freie Universität Berlin. http://corporafromtheweb.org/ [last access: May 2, 2024].
DACT = *Dact Werkzeug für die Analyse von Alpino Korpora*. Danïel de Koh. Online: www.rug-compling.github.io/dact/.
DeReKo = *Deutsches Referenzkorpus*. Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/kl/projekte/korpora/ [last access: May 2, 2024].
Deutsche Rechtschreibung = *Deutsche Rechtschreibung*. Regeln und Wörterverzeichnis. IDS-Mannheim. https://grammis.ids-mannheim.de/rechtschreibung [last access: May 2, 2024].
Deutsches Textarchiv = *Deutsches Textarchiv*. Berlin-Brandenburgische Akademie der Wissenschaften. www.deutsches-textarchiv.de [last access: May 2, 2024].
DWDS-Kernkorpus = *Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de/ressourcen/kernkorpus/ [last access: May 2, 2024].
DWDS-Wortprofil = *DWDS-Wortprofil*. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de/ressourcen/wortprofil/ [last access: May 2, 2024].
FSPar = *FSPar – a cascaded finite-state parser for German*. Universität Stuttgart: Institut für Maschinelle Sprachverarbeitung. http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/fspar.html [last access: May 2, 2024].
Kant-Korpus = *Bonner Kant-Korpus*. Universität Duisburg-Essen. https://korpora.zim.uni-duisburg-essen.de/kant/ [last access: May 2, 2024].
Korpus der Zeitschrift "Die Fackel" = *Korpus der Zeitschrift "Die Fackel"*. http://corpus1.aac.ac.at/fackel// [last access: May 2, 2024].
Sketch Engine = *Sketch Engine. Lexical Computing*. https://www.sketchengine.eu// [last access: May 2, 2024].
TigerSearch = *TIGERSearch*. Universität Stuttgart: Institut für Maschinelle Sprachverarbeitung. http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html [last access: May 2, 2024].
VLO = *Virtual Language Observatory*. www.clarin.eu/vlo [last access: May 2, 2024].

## Images

**Fig. 7.1** "Die verschiedenen Arten der Fahrung": Georg Agricola, 1556 (Source: https://de.m.wikipedia.org/wiki/Datei:Die_verschiedenen_Arten_der_Fahrung.png).
**Fig. 7.2** "Vermessung im Bergbau durch einen Markscheider": Source: Deutsche Fotothek via Wikipedia: https://commons.wikimedia.org/wiki/File:Fotothek_df_tg_0000341_Bergwerk_%5E_Bergbau_%5E_Markscheider_%5E_Vermessung.jpg.