DE GRUYTER

INTERNET LEXICOGRAPHY

AN INTRODUCTION

Edited by Annette Klosa-Kückelhaus

LEXICOGRAPHICA. SERIES MAIOR

Internet Lexicography

LEXICOGRAPHICA
Series Maior

Supplementary Volumes to the International Annual
for Lexicography
Suppléments à la Revue Internationale
de Lexicographie
Supplementbände zum Internationalen Jahrbuch
für Lexikographie

Edited by
Rufus Hjalmar Gouws, Ulrich Heid, Thomas Herbst,
Anja Lobenstein-Reichmann, Oskar Reichmann,
Stefan J. Schierholz and Wolfgang Schweickard

Volume 164

Internet Lexicography

An Introduction

Edited by
Annette Klosa-Kückelhaus

DE GRUYTER

This publication has received funding from Leibniz-Gemeinschaft (Publikationsfonds für Monografien) and Leibniz-Institut für Deutsche Sprache (Mannheim). / Diese Publikation wurde gefördert durch den Publikationsfonds für Monografien der Leibniz-Gemeinschaft und durch das Leibniz-Institut für Deutsche Sprache (Mannheim).

ISBN 978-3-11-120521-2
e-ISBN (PDF) 978-3-11-123375-8
e-ISBN (EPUB) 978-3-11-123392-5
ISSN 0175-9264
DOI https://doi.org/10.1515/9783111233758

This work is licensed under the Creative Commons Attribution 4.0 International License.
For details go to https://creativecommons.org/licenses/by/4.0/.

Creative Commons license terms for re-use do not apply to any content that is not part of the Open Access publication (such as graphs, figures, photos, excerpts, etc.). These may require obtaining further permission from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

Library of Congress Control Number: 2024940019

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available on the internet at http://dnb.dnb.de.

© 2024 the author(s), editing © 2024 Annette Klosa-Kückelhaus, published by Walter de Gruyter GmbH, Berlin/Boston
The book is published open access at www.degruyter.com.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface

This introduction to Internet lexicography is based on the German publication "Kompendium Internetlexikografie" (edited by Annette Klosa and Carolin Müller-Spitzer and published by de Gruyter in 2016). As almost 10 years have passed between writing the original text and the English version in 2024, not only has the content changed considerably, but the group of authors and publication mode (now: open access) have changed as well. I am glad that almost all of the original authors were once again able to give their time and expertise to update their chapters; my thanks also go to those colleagues who joined the team for this edition.

We (the authors and I) are grateful to de Gruyter publishing house for their willingness to publish this introduction as part of the series "Lexicographica. Series Maior". More importantly, we would like to thank the series editors Rufus Hjalmar Gouws, Ulrich Heid, Thomas Herbst, Anja Lobenstein-Reichmann, Oskar Reichmann, Stefan J. Schierholz, and Wolfgang Schweickard for their valuable feedback on the text and for accepting this title into the series. We are especially grateful to the Leibniz Association and the Leibniz Institute for the German Language (IDS, Mannheim) who granted funding for this open access publication. IDS also provided generous funds for the professional translation and proofreading of this book. We very much appreciate this support and the excellent work which our translator Philipp Matthews and our proofreader Helen Heaney provided.

We are indebted to Bloomsbury Academic (an imprint of Bloomsbury Publishing Plc), who granted permission for the reuse of material from the chapter "The design of Internet Dictionaries" by Annette Klosa-Kückelhaus and Frank Michaelis in "The Bloomsbury Handbook of Lexicography" edited by Howard Jackson (2022) for this volume. We also thank INFORMA UK Ltd. for granting permission to use material from the chapter "User participation in the Internet era" by Andrea Abel and Christian M. Meyer in "The Routledge Handbook of Lexicography" edited by Pedro A. Fuertes-Olivera (2018) for this volume.

This Introduction could not have been produced in its present form if the German Research Foundation DFG had not approved the application for funding a scientific network on the subject of "Internet lexicography" in 2010, which carried out its work between 2011 and 2015. The funding made it possible to set up the network and to advance research into and discussion of key issues relating to the creation and publication of dictionaries on the Internet. Many topics were raised, presented, and discussed at the network meetings as well as at conferences and workshops on (electronic) lexicography such as the EURALEX (https://euralex.org/conferences/) and eLex – Electronic lexicography in the 21st century series (https://elex.link/), which are now reflected in condensed form in the chapters of this volume. Thank you to all of those colleagues who shared their knowledge and expertise with us, many of whom have also been associated with us through the projects "eNeL – European Network of e-Lexicography"

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/10.1515/9783111233758-202

(COST Action IS1305, 2013–2017, https://www.elexicography.eu/) and “ELEXIS – European Lexicographic Infrastructure” (Horizon 2020 project no. 731015, 2018–2022, https://elex.is).

On behalf of all of the authors, I hope that this introduction will not only be used in university teaching but will also provide an impetus for further professional exchange in the Internet lexicography community.

Annette Klosa-Kückelhaus
(Mannheim 2024)

Contents

Annette Klosa-Kückelhaus

Introduction

Lexicography is a long-established academic and cultural practice going back many hundreds of years. However, the dramatic growth of the Internet since the late 1990s has led to fundamental changes in this practice. In the meantime, it has become possible to locate and browse through a wide range of lexicographic content relating to almost all of the major languages and many smaller languages and endangered languages in the world in a matter of minutes and free of charge. Just one generation ago, you would have had to undertake an elaborate library search and possibly order a book through an inter-library loan. Many historical dictionaries that were previously only accessible in specialist libraries have also been made freely available as part of comprehensive digitisation projects. This availability of more and more lexicographic content and in new formats is undoubtedly the principal change that users of reference works have seen.

However, behind the scenes of lexicographic practice and research, a great deal more has changed. These changes began as early as the mid-1990s, when the use of computers already began to radically modify the processes of lexicography (cf. Storrer 2001), with the publication of dictionaries in other media not far behind. But the dictionary landscape was altered much more decisively by the advent of free dictionaries on the Internet that were not produced by prestigious publishers. The range of freely accessible lexicographic content online may not have been able to entirely match academic and published dictionaries in terms of quality, but they still drew very high numbers of users and led to a collapse in the sales of publishers' print dictionaries. At the same time, many publishers found it difficult to identify a business model suitable for marketing digital lexicographic data on the Internet for money. Equally, academic dictionaries took a very long time to adjust to the altered media context. However, many of the lexicographers and researchers involved in these dictionaries were able to see the numerous opportunities offered by digital media and the Internet as a publication platform for lexicography (cf., e.g., de Schryver 2003). And yet, naturally, long-established practice does not change overnight, and even now much remains in flux, especially since Artificial Intelligence (AI) and Large Language Models took off in lexicography in the third decade of the 21st century.

This introduction is devoted to the opportunities and perspectives provided for lexicography by digital media and the Internet. Its aim is to communicate to students and academics at universities the central aspects of the research and practice of Internet lexicography. The emphasis lies less on unresolved research questions and specialist technical aspects of Internet lexicography and more on an easily accessible,

Annette Klosa-Kückelhaus, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: klosa@ids-mannheim.de

Open Access. © 2024 the author(s), published by De Gruyter. cc BY This work is licensed under the Creative Commons Attribution 4.0 International License.

https://doi.org/10.1515/9783111233758-001

introductory thematic overview of the individual areas of work, enriched through references to further, more in-depth reading. In the process, we have concentrated on key areas of Internet lexicography, with the goal of sharing fundamental concepts and methods in a way that is readily understandable, thereby embedding this important and innovative field of research and practice in university teaching and, above all, in the training of language teachers as well as future lexicographers.

More specifically, our compendium covers the following areas of Internet lexicography: first of all, Peter Meyer, Axel Herold, and Frank Wiegand provide an introduction to **The Technological Context for Internet Lexicography** (→ Chapter 1), explaining the most important technical requirements and processes that enable a dictionary to be provided and used online. Issues to do with logging, versioning, and persistence/identity are also discussed in this chapter. In this way, their contribution makes it easier to understand the technical questions addressed in other chapters (e.g. in relation to the processes involved in editing and publishing an Internet dictionary, different ways of accessing the lexicographic data, and possible approaches to researching dictionary use).

In the chapter on **A Typology of Internet Dictionaries and Dictionary Portals** (→ Chapter 2), Stefan Engelberg and Angelika Storrer develop the criteria for classifying online reference works that are applied in subsequent chapters of this volume. They discuss typological features of Internet dictionaries that are both specific to the medium and independent of it and also propose a typology of dictionary portals (which include several Internet dictionaries).

Chapters 3 to 8 provide insights into the development of an Internet dictionary: in **The Lexicographic Process** (→ Chapter 3), Annette Klosa-Kückelhaus and Carole Tiberius explain how the preparation and publication of an Internet dictionary (or dictionary portal or central lexicographic database) proceed. After introducing and providing an overview of research into the lexicographic process in general, they describe the particular details of the digital lexicographic process for Internet dictionaries, giving specific examples. In addition to discussing software that supports the lexicographic process, the question arises about the process that has to be described in order to develop lexicographic portals and central lexicographic databases.

Axel Herold, Peter Meyer, and Frank Wiegand then investigate the central question of modelling in the chapter on **Data Modelling** (→ Chapter 4), exploring a number of different possible options. They provide an introduction to data structures and formats of representation (e.g. XML documents), different data models (e.g. conceptual-semantic models), and attempts to standardise data modelling for Internet dictionaries (e.g. the Text Encoding Initiative, TEI).

There are also various strategies for linking lexicographic data and providing access to those linked data. These are presented by Stefan Engelberg, Carolin Müller-Spitzer, and Thomas Schmidt in the chapter on **Linking and Access Structures** (→ Chapter 5). They show how lexicographic information in Internet dictionaries can be interconnected and describe onomasiological and semasiological structures for accessing data alongside other methods (e.g. grapheme-based searches). In the process, the differences between Internet dictionaries and print dictionaries become particularly clear.

In → Chapter 6 on **The Design of Internet Dictionaries**, Annette Klosa-Kückelhaus and Frank Michaelis present some general thoughts on the design of dictionaries and discuss differences between print and online publications. They also explain design dependencies (e.g. on data modelling, on the user) and elaborate on specific aspects of Internet dictionary design such as content-centric presentation vs. user-/human-centric design. Ideas on the design of search functions and the design process as a whole are discussed as well.

Alexander Geyken and Lothar Lemnitzer provide an introduction to one particular aspect of compiling lexicographic content in their chapter on **The Automatic Extraction of Lexicographic Data** (→ Chapter 7), where they explore the different possibilities of extracting word-based information from electronic corpora. Corpora are central in the typology of possible data sources, and the chapter shows in detail what information can be extracted from them to generate particular lexicographic data. The limits of automatic processes are also discussed, in addition to desirable future developments, such as access for users to the primary sources themselves.

In the chapter on **User Participation** (→ Chapter 8), Andrea Abel and Christian M. Meyer report on how users can be involved in the lexicographic process. They distinguish between direct user participation (e.g. forms for entering new word entries), indirect user participation (e.g. feedback forms), and complementary participation (e.g. dictionary blogs), using a range of specific examples to discuss their specific advantages and disadvantages as well as their effects on the lexicographic process involved in creating dictionaries.

A published Internet dictionary can be the subject of **Research into Dictionary Use**, a topic which is introduced by Carolin Müller-Spitzer and Sascha Wolfer in → Chapter 9. Empirical research into dictionary use concerns itself with actual instances of use or, more generally, with observations and experience of dictionary use. As such, it must draw on methods of empirical research in the social sciences, the basic elements of which are elaborated in the chapter. The main part is dedicated to user research in relation to Internet dictionaries, which are the focal point of this introduction.

This book (including the extensive → Index) gives insights into the state of research and its development since the Internet's first phase of popularisation and up to 2024. We have sought to position these developments in the wider tradition of lexicography as a cultural practice and also to illuminate its connections to dictionary research in the typographical age. However, the focus lies on innovations that are connected to digital media and the Internet. Today, lexicography is once again standing "at a turning point in its history" (Granger 2012: 10). We can certainly assume that human beings will always have linguistic questions and needs in the distant future and that some form of tool will be required to deal with them. It is less clear, however, whether dictionaries as we know them today will continue to exist or whether they will be increasingly integrated into the context of smart reading and writing tools and other digital resources (cf. Lew 2015: 7) and disappear as such. The role that digital dictionaries can and will play for Large Language Models and the other way round is now (in 2024) in the process of being researched and defined as well. This introductory volume should provide the foundations to be able to trace future developments in practice and research.

Our experience of the last three decades of digital lexicography has demonstrated that a cultural practice like lexicography only changes slowly at its core and mostly only as a result of external pressure. As such, we have strong grounds to assume that the present volume will provide a good overview of the field, at least for the coming years. And yet, at some point, this volume, too, will represent but a historical snapshot of Internet lexicography in the mid-2020s.

Bibliography

de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic Dictionary Age. In: *International Journal of Lexicography* 16, 143–199.

Granger, Sylviane (2012): Introduction: Electronic lexicography – from challenge to opportunity. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 1–11.

Lew, Robert (2015): Dictionaries and Their Users. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer, 1–9 (manuscript version).

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

Axel Herold, Peter Meyer, and Frank Wiegand

1 The Technological Context for Internet Lexicography

Fig. 1.1: This modern submarine cable trencher, a special machine for laying undersea cables offshore, weighs more than 100 tonnes.

There have only been Internet dictionaries for a few decades – compared to the thousands of years of dictionary writing history, this is a vanishingly small period of time. The photograph illustrates one of the many technological and infrastructural require-

Note: The German version of 2016 on which this chapter is based was written by Axel Herold, Lothar Lemnitzer, and Peter Meyer.

Axel Herold, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: herold@bbaw.de
Peter Meyer, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: meyer@ids-mannheim.de
Frank Wiegand, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: wiegand@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/10.1515/9783111233758-002

ments of modern Internet lexicography: by far the largest proportion of international data transfers is handled by a network of glass fibre cables measuring many hundreds of thousands of kilometres, which often cross the oceans at great depths.

Computer technology is becoming ever smaller and cheaper, both to acquire and operate, and its processing and storage performance is increasing exponentially. This is one of the technological requirements for making dictionaries available online, but so too is the infrastructure of the Internet, which makes it possible to exchange information and data simply and reliably between billions of interconnected computers. This chapter is devoted to the fundamental technological preconditions for present-day Internet lexicography. First, we outline what actually happens "behind" the user interfaces that are visible on the screen when a user accesses a dictionary online and how these processes can be recorded in log data for the purposes of documenting them. Second, we discuss how the identity and long-term availability of content can be maintained in view of the possibility of online material being constantly updated.

1.1 Introduction

The digital revolution in the 20th century has completely transformed the ways in which dictionaries are compiled and used. Just like the resources connected to them, such as textual corpora and multimedia, dictionary texts can be represented in digital form, that is, ultimately as sequences of 0s and 1s. Digital data of this kind can be processed at ever greater speeds by computers, stored in ever greater quantities so as to be downloaded rapidly anywhere, quickly transferred to a worldwide network of computers, and presented flexibly in audiovisual form to be viewed and manipulated by humans. For both lexicographers and dictionary users, this opens up a broad spectrum of possibilities; these are the subject of the present volume, including in particular:

- the managing, searching, and exploring of dictionary data, including the large textual corpora connected to them (→ Chapter 3),
- the (semi-)automatic creation of particular dictionary content (→ Chapter 7),
- the collaborative, ubiquitous compilation of dictionaries (→ Chapter 8),
- the removal of the constraints of the print medium (→ Chapter 5).

It is an essential prerequisite when engaging with the topic of Internet lexicography to have a basic understanding of the technologies required for the technical development of Internet dictionaries, their functioning, and use. This applies in particular to the associated requirements for structuring and representing the dictionary content, as is shown in detail in → Chapter 4 on data modelling. However, even in the realm of web development, an enormous variety of technologies is employed so that this introductory chapter can only provide an overview of knowledge in selected areas of particular relevance to lexicographic work. Furthermore, in order to make the discussion more accessible to newcomers, the presentation in → Section 1.2 very deliberately oversimplifies the reality, focusing only on the aspects essential for lexicography. The result is that technical details are sometimes knowingly described in a manner that is incomplete or formally not entirely correct.

1.2 Internet technology in the context of Internet dictionaries and lexical information systems

1.2.1 Network communication on the Internet

The notional starting point for our short tour of the most important web technologies is the typical case in which a user of an Internet dictionary would like to view a word entry in the browser on their computer. Let us take a toy example. The user would like to be able to see the entry for the English noun *disproof* in the monolingual English dictionary "MyEnglishDict". To do that, they must tell their browser where "on the Internet" the website with the required information can be found. For that, the browser needs an *Internet address*, more formally a URL (uniform resource locator) that indicates where exactly this site can be found. In our example, this URL might look as follows:

```
https://www.my-english-dict.com/entry/disproof
```

A URL like this can be entered directly into the address bar of the browser. The browser then retrieves the resource (website) identified by the URL from the Internet and displays it on the screen. Normally, though, users do not enter such complex URLs manually themselves but rather click on a *hyperlink* (usually abbreviated to *link*) that is located on another web page, say, a list of results generated by a search engine like Google or Bing. Such a link leads the browser to the appropriate web page: when the user clicks on the link, this prompts the browser to load from the Internet the website with the URL that is connected to the visible text of the link. In the most basic case, the technical process that follows after a URL link has been clicked on is identical to that prompted by entering the same URL manually in the address bar. In a similar way, the main web page of the dictionary "MyEnglishDict" may offer a list of headwords that are hyperlinks to the web pages belonging to the dictionary entries concerned. The user may also use the search functionality of the Internet dictionary, for example, to search for lemmas beginning with "dispr"; the results are then presented as a further list of links on a search results page.

What does the process look like by which the browser retrieves the information from the desired website and displays it?

First, a few general points. A *web browser* is a program that runs on a device connected to the Internet (PC, smartphone, etc.) and that is in a position to download information from the Internet and display it on a screen. The *Internet* is a complex worldwide network of electronic hubs (so-called routers) mostly connected to one another by cables; essentially, every device connected to a hub in this network can send information to every other connected device via these hubs in a way that is extremely failure resistant. This functions by virtue of every device on the Internet being allocated a unique identifying combination of numbers, its *IP address*. The dictionary data (web pages, etc.) to be retrieved are stored on a particular computer managed, for example, by the provider of the dictionary or an external third party. Thus, the web browser has to have the data from the desired web page sent from that computer over the Internet. To do this, the browser must send a request over the Internet to the relevant computer and, hence, has to know the latter's IP address.

However, the URL given above does not contain an IP address, which may even change from time to time for any given device, but rather an alternative name for the computer that is easy for people to read and recognise, its so-called *host name*, i.e. www.my-english-dict.com. Through communication with specific computers (so-called *name servers*) on the Internet, the browser can find the current IP address of the computer (say, 93.184.216.34) for this host name. In fact, it is sometimes even possible to use the IP address directly in a URL instead of the host name. For example:

```
https://93.184.216.34/entry/disproof
```

The browser then sends its request for a web page as a message to the computer with the IP address 93.184.216.34. This message consists simply of a sequence of characters (numbers and letters as well as some specific control characters), which are ultimately coded as sequences of 0s and 1s. A strict system of rules, a so-called *network protocol*, determines how the message has to be constructed; that is, it provides formal rules for the language through which the computers communicate with one another. Which protocol is used is also given in the URL: the prefix “https://” indicates that the usual protocol for transferring web pages, *HTTPS* (Hypertext Transfer Protocol Secure), is being used. In the past, and in a few cases today, web pages used the variant HTTP, which provides no data encryption; it corresponds to the URL prefix “http://”. The protocol prefix can usually be omitted when the URL is entered manually into the browser's address bar. The message sent by the browser over the Internet after the URL has been entered is itself a short text that specifically contains a line with the actual request, in addition to some further lines with meta information, the HTTP(S) headers (→ Section 1.3). In our case, the line containing the actual request looks as follows:

```
GET /entry/disproof HTTP/1.1
```

The keyword GET in the HTTP(S) protocol designates the method; in this case it simply requests the transfer of data from the remote computer, as opposed to, say, the modification of data on the remote system. GET is followed by the *URL path*, which is, in a sense, the actual designation of the required digital resource, here the requested web page. Next, the version of the HTTP(S) network protocol to be used is given, here 1.1. Note that this is always indicated as HTTP/1.1, even with HTTPS, since the underlying message exchange is the same in HTTP as in HTTPS, the only difference being data encryption in the HTTPS variant.

The URL path can also be derived from the URL: in the present elementary case, it is obviously just the part of the URL that follows the host name. It consists of individual segments (series of characters) that are separated from one another with slashes. There are no generally binding rules as to what the URL for a specific resource must look like. In this example, it could have read "/dictionary/entry/3325" or "/dict/disproof/showentry" instead of "/entry/disproof"; ultimately, the programmer of the Internet dictionary makes the relevant decision. In many cases, paths are chosen so that they give a rough impression of the structure of the online content being made available.

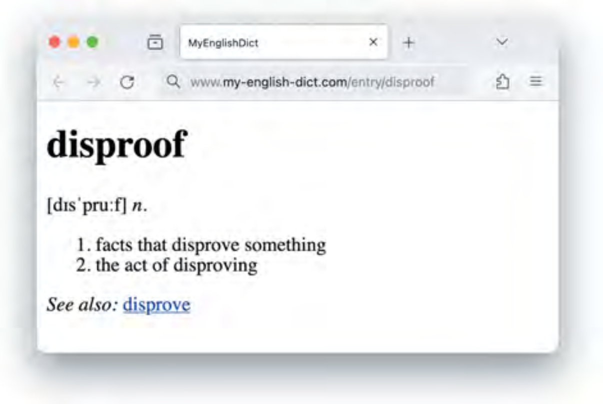
The only kinds of URLs that users normally enter manually into a browser are those with an *empty path*, that is, those whose URL consists only of a prefix like https:// and the host name: "www.google.de". The empty path is indicated in the HTTP(S) request with a simple forward slash: "/":

```
GET / HTTP/1.1
```

In typical cases, the empty path corresponds to the *home page* of an Internet presence from which the desired pages are reached through links or search functions.

A technical note for those who are interested and have prior knowledge: it may well be the case that the URL path corresponds to an actual data path on the computer responding to the request so that a path such as "/dictionary/entry/3325" refers to a piece of data with the name "3325" in the subdirectory "entry" in the directory "dictionary" on a hard disk drive, the content of which is sent back in response to the browser making the request. This is the reason for the hierarchical form of URL paths. Generally, though, there is no correspondence between the URL and the location of the data on the remote computer because the answer to a request is usually only "constructed" after the request and is not already waiting, prepared in advance, on a hard drive.

In order for the computer with the address 93.184.216.34 to be able to process the request at all, there must be a program running on it that is in a position to receive and respond to requests from other computers over the Internet. In very general terms, this kind of program is known as a *web server*. The web server then passes the request to another program, the *web application*, that is responsible for delivering the web pages of the MyEnglishDict dictionary. So it is ultimately the web application that services the *request* “GET /entry/disproof”, providing a specific description of the requested resource, in this case the code for the web page with the dictionary entry on *disproof*. This web page could, in the simplest case, look like → Fig. 1.2.[1] The code for the page, on which more below, is passed to the web server, which sends it as a *response* to the *client*, i.e. to the browser on the computer where the request originated. The response sent by the web server also follows the rules of the HTTP(S) protocol and again contains meta information (the response headers) beside the returned content proper. Note that the terms *client* and *server* are also used to refer to the computers on which client or server programs run. In the present example, we may say that the device with the web browser is a client that is making a request to the web application on the server computer with the IP address 93.184.216.34.



disproof

[dɪsˈpruːf] *n.*

1. facts that disprove something
2. the act of disproving

See also: disprove

Fig. 1.2: Minimal example of the view of an entry in an Internet dictionary.

1.2.2 HTML, CSS, and JavaScript

Yet how exactly does a web application, in its response, describe a website to a client, that is, to a web browser? The description is written in a particular language, namely Hypertext Markup Language (*HTML*). The central task of the browser is to transfer this description into the required presentation (to *render* the HTML source code, usually on a screen). The mini web page with the entry on *disproof* looks as follows:

[1] This example draws on one given in the “Guidelines for Electronic Text Encoding and Interchange” (TEI 2023) on the dictionary module of the Text Encoding Initiative. The example is also used in Chapter 4.4.1 on TEI.

```
<!DOCTYPE html>
<html>
    <head>
        <meta charset="utf-8">
        <title>MyEnglishDict</title>
    </head>
    <body>
        <h1>disproof</h1>
        <p>[dɪs'pruːf] <i>n.</i></p>
        <ol>
            <li>facts that disprove something</li>
            <li>the act of disproving</li>
        </ol>
        <p><i>See also:</i> <a href="/entry/disprove">
        disprove</a></p>
     </body>
</html>
```

The basic idea behind HTML is a strictly descriptive and hierarchically structured markup of sections of text by means of structuring information in angular brackets, so-called *tags*. Thus, the word *disproof* is marked here as a heading at the first – i.e. the highest – organisational level, by virtue of a *start tag* <**h1**> (meaning: “level 1 heading”) placed in front of the word and a corresponding *end tag* </**h1**> after the word. End tags are marked by a forward slash immediately after the opening angular bracket. How exactly this structural information is rendered is a matter for the browser. Headings at level 1 are usually represented in a larger, boldface font on a separate line. The example code for the miniature web page contains further illustrations of typical HTML tags:

- a *paragraph* of text: <**p**> . . . </**p**>;
- a span of text “in an alternate voice or mood”,[2] usually rendered in *italics*: <i> . . . </i>;
- an *ordered list*: <**ol**> . . . </**ol**>;
- a *list item* in that list: <**li**> . . . </**li**>;
- the web page *title*, shown in a browser’s title bar or a page’s tab: <**title**> . . . </**title**>;
- a hyperlink (*anchor*): <**a href=**" . . . "> . . . </**a**>. Here, the text that is actually shown in the browser is placed between the start and end tags (recognisable as a link in → Fig. 1.2 by underlining and a different colour), and the URL (or URL path) for the web page that is brought up by clicking on the link is given as a so-

[2] From the HTML specification, 4.5.20 “The i element”, https://html.spec.whatwg.org/#the-i-element.

called *attribute*, an additional piece of information, inside the start tag (**href** stands for *hypertext reference* and is the *name* of the attribute; the text in quotes, here the URL path, is its *value*).

Further tags structure the HTML document as a whole; thus, the whole document has to be enclosed in the <**html**> . . . </**html**> pair of tags. The initial line <**!DOCTYPE html**> is the *document type declaration* and has a special syntax; it marks the document as being written in the current version of HTML, which is HTML5. The actual content of the page to be shown in the browser is the "body" of the document and is marked by <**body**> . . . </**body**>. Core information about the web page is found in the "head" of the document and is indicated by <**head**> . . . </**head**>: in the example above, the head only contains the **title** of the page, which is shown in the tab of the browser window, plus information about the so-called *text encoding* used in the document, that is, the set of characters used and how each character is represented by a certain number. UTF-8 encoding is the most widely used encoding today, covering the characters of most of today's writing systems and being part of an ongoing standardisation effort known as the Unicode Standard. A start tag and an end tag, together with all of the content between them, represent what is known as an *element* in HTML. The tags indicate the *name* of the element while the content of the element consists of text and/or subordinate elements. Some elements cannot have content. The **meta** tag that is used here to specify the character encoding is one such *void* element; therefore, as a special syntax rule in HTML5, it must not have an end tag. It still conveys information, though, through its attribute **charset** (i.e. 'character set').

HTML code describes only the textual structure of a web page in a hierarchically structured way. Normally, HTML is combined with two further languages: The graphic and colour structure of the content is described using *Cascading Style Sheets* (CSS), including more complex aspects like animations and the definition of different presentations of site content, for example, on printers or small screens.

JavaScript is a programming language available on all modern browsers through which all of the interactive processes of a web page can be implemented directly in the browser, including comprehensive manipulation of graphics, data processing, and communication with other computers on the Internet, etc.

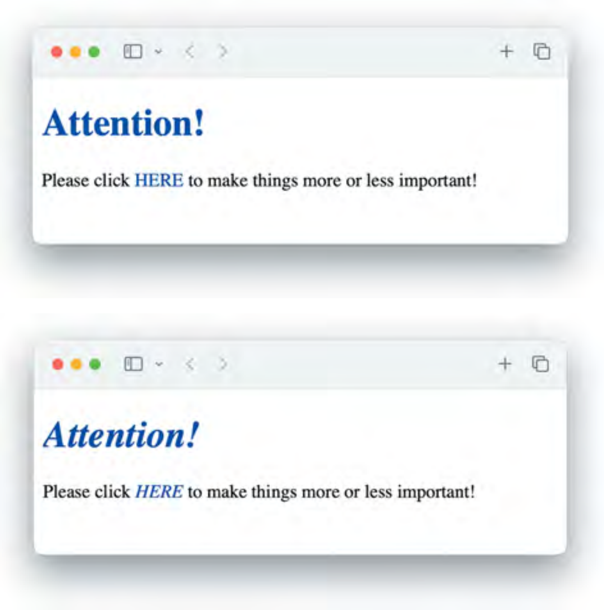
At this point, we have to be content with a miniature example to illustrate some basic ideas. In the following HTML code, which can be tested directly with a browser, CSS and JavaScript code is integrated directly:

```
<!DOCTYPE html>
<html>
   <head>
      <meta charset="utf-8">
      <title>CSS and JavaScript Demo</title>
      <style>
```

```
            .teaser {
                color: blue;
            }
            .alert {
                font-style: italic;
            }
        </style>
    </head>
    <body>
        <h1 class="teaser">Attention!</h1>
        <p>
            Please click
            <span onclick="toggleEmphasis()"
class="teaser">HERE</span>
                to make things more or less important.
        </p>
        <script>
            function toggleEmphasis(){
                for (teaserElement of
document.getElementsByClassName("teaser")) {
                    teaserElement.classList.toggle("alert");
                }
            }
        </script>
    </body>
</html>
```

Two HTML elements have a **class** attribute with the value “teaser”: the **h1** heading and a **span** element, which simply delimits a stretch of running text containing the text *HERE*. The **class** attribute assigns the *CSS class* ‘teaser’ to these elements. Such a CSS class is simply a kind of custom marker that can be used to define presentation-related aspects pertaining to the elements it is assigned to. In our example, this definition is done in the **style** element. We will not discuss the finer points of the CSS language here but the first CSS ‘instruction’ basically says that any element marked with the ‘teaser’ class gets a ‘blue’ text colour, where the predefined keyword ‘blue’ actually represents a certain, pure shade of blue. As a result, both elements with the ‘teaser’ class are indeed rendered blue by the browser, as shown in the screenshot in → Fig. 1.3. If the user clicks on the word *HERE*, all of the blue text is additionally italicised; on clicking again, the italics are removed again. This interactive behaviour is governed by the **onclick** attribute of the *HERE* **span**: If the user clicks somewhere on the text inside the **span**, the JavaScript *function* ‘toggleEmphasis’, which is defined in the **script** element, is executed. A function is basically a block of programming code.

The code in the ‘toggleEmphasis’ function looks at each HTML element with the class ‘teaser’ and assigns to it – or removes if already present – the CSS class ‘alert’ that, according to the CSS code in **style**, triggers italic text.



Attention!

Please click HERE to make things more or less important!

Attention!

Please click *HERE* to make things more or less important!

Fig. 1.3: Toy example of an interactive web page using CSS and JavaScript. If the user clicks on *HERE*, all of the blue text becomes italic; on clicking again, the italics are removed.

The CSS and JavaScript code is normally put in separate files so that the design and interactivity of the web page can, as far as possible, be maintained independently of the textual content. In this way, you could alter the colour and the interactive behaviour of *HERE* just by modifying these external files, which must be retrieved by the browser from the web server using dedicated URLs. In our example, the code inside the **style** and the **script** elements could alternatively be put in text files **mystylesheet.css** and **myscript.js**, respectively and referenced as follows in the head of the HTML code:

```
<link rel="stylesheet" href="/stylesheets/mystylesheet.css">
<script src="/scripts/myscript.js"></script>
```

While processing this HTML code, further HTTP(S) requests are initiated to retrieve the referenced files. This is also a common way of integrating multimedia content: The HTML element

```
<img src="/images/mypicture.jpg">
```

initiates a new request to load the image with the URL path "/images/mypicture.jpg" from the server. In this way, a request for a single complex HTML web page can prompt dozens of additional requests to download further data that are needed to render the page and enable its functionality.

Linking HTML documents to separate, external CSS and JavaScript files allows web applications to be developed in a modular way. Developers can take advantage of a huge number of software *libraries*, often freely available as open source software. In the realm of web page development, libraries are essentially just CSS and/or JavaScript files that assist in the complex task of creating standards-compliant web design, providing ready-made interactive visual components for web pages, or simplifying the implementation of complex functionality. *CSS frameworks* provide a large number of predefined CSS classes that developers can use in their HTML in order to achieve a professional and consistent look, including what is known as *responsive web design* that adapts the page layout automatically to different screen sizes and device types. So-called *front-end frameworks* have simplified web development considerably by improving on the approach that was used in our toy example: roughly speaking, instead of writing 'imperative' code that, depending on external circumstances, explicitly changes the structure of the web page and the properties of its elements, the programmer describes 'declaratively' what the page should look like depending on a set of data that defines the overall 'state' of the page.

1.2.3 Outlook

As we have seen, the web application sends HTML code over the Internet, in response to the request from the client, to the web browser where it is rendered. Interested users can trace the process described here in detail at any time on their own computer. On the one hand, browsers usually offer the option of displaying the HTML code of a page (often referred to as source code). On the other hand, most modern browsers assist programmers with inbuilt developer tools that, for example, let you view the exact content of the HTTP(S) request and response and even show details such as how long it took to determine the IP address of the web server by consulting a name server.

But where does a web application take the HTML code for an entry? Generally, this code does not remain fixed and complete ("static") on the hard drive of the web server but rather is built "dynamically" from abstract lexicographic data structures only when the client request is answered. This is explained in more detail in → Chapter 4.

Finally, it is important to emphasise that we have only examined in some detail the simplest example of web content being accessed, namely the “classic” request-response cycle, in which the user initiates a browser request to the server with an operating action, such as a mouse click on a link, thereby subsequently fetching a new HTML web page through the server’s response. The limitations of this approach can be overcome in different ways. Here are three important examples:

- With a set of technologies collectively known as *Ajax*, the program code (JavaScript) on a web page can request data from a web server via HTTP(S) asynchronously, that is, in the background and without blocking any additional user interaction on the web page. The code can then process these data and modify the content of the page in any way necessary. In complex applications, this removes the slow and non-intuitive loading of a whole new web page, for example, after a button or icon has been clicked. Thus, navigating Ajax-based websites approaches the user experience of conventional desktop applications.
- A web server can deliver data to a client in the HTTP(S) protocol only when a corresponding request has previously been made by the client. The *WebSocket* protocol makes genuine bidirectional communication between the client and server possible such that a server can send data at any time to clients “of its own accord” following a particular event. For example, whenever some participant in an online chat posts a message, everybody else should receive an update of the chat history immediately; with pure HTTP(S), the only way to implement this would be through letting the browser issue requests every few seconds in order to check for updates.
- A lexicographical web application on a server can do much more than just deliver HTML pages (and associated web resources) to web browsers. There are many types of client applications that may need to use lexicographical data. Typical examples are dictionary apps on a mobile device such as a smartphone or tablet or, more generally, word processors, language learning apps, language-related games, or programs that researchers implement to process large amounts of data. Such programs might fetch lexical information from a server on a per-need basis, if storing a complete lexicographical database on the computer itself is not a viable option. In many cases, the data delivered by the server are not formatted in HTML. Instead, formats are used that are better suited for machine processing than for direct rendering in a web browser. Server applications that serve these types of clients are typically called *web services*. A common scenario for the exploitation of such web services is an aggregating server for lexicographical content that itself draws its data from a range of lexicographical resources located on other computers around the world. A request to the server about a specific word would lead to the server fetching relevant pieces of information from all of the other computers, bundling them, and then forwarding them to the client. This approach is called a *federated search.*

In all of these cases, a set of rules is needed which a client program can use to retrieve or even modify data on a specific server. This rule set or protocol is called the server's *API* (application programming interface). It has become widespread practice to simply use HTTP(S) for that purpose: individual resources – e.g. dictionary entries or collections thereof – can then be accessed through specific URLs using a variety of access modes, including the GET mode mentioned above. The machine-readable content returned in the response is typically delivered by the server in a highly structured data format, such as XML (→ Chapter 4) or JSON (a notation for structured data that JavaScript can directly understand and process), instead of HTML.

Even in the classic case of an online dictionary running in a browser, it would not be unusual to use the Ajax approach sketched above, using JavaScript code to fetch the lexicographical data currently requested by the user through an API provided by the server. The data obtained this way are then processed by JavaScript code on the web page to construct HTML elements that are inserted into the currently shown web page in order to render a human-readable view of the entry without even loading a completely new web page.

In all of this, the increasingly ubiquitous availability of the Internet is erasing the boundaries between online and offline content. Specifically, this could mean that a core set of data is available on a local device while an application can automatically search online for updates and other associated content depending on the availability of an Internet connection – without the user knowing the origin of the data.

1.3 Logging

In what follows, the term "logging" summarises, very generally, the recording of information about the internal state of a technical system as well as the interaction of users (or of other technical systems) with the system.

The information recorded is stored as *log data* in the form of individual datasets (often called "records"), usually with an exact timestamp so that the chronological sequence of actions and the state of the system can be reconstructed for relevant aspects. Additional metadata may supplement these datasets, for example, a classification of the meaning of the datasets (debug information, warning, serious error, etc.) or the name of the system component that generated the dataset. The log data can be grouped and filtered using this metadata to allow for better informed analyses of the system's behaviour, e.g. for debugging purposes.

For an Internet dictionary, two technical systems that generate log messages are of principal interest: the actual dictionary web application and the web server through which the web application communicates with computers making requests. In the concrete technical realisation of the overall system, both can also be subsystems of a single integrated system. In the following illustration, we shall proceed like in → Section 1.2

from the latter situation in order to be able to provide a concise overview. As such, we shall treat the Internet dictionary as a monolithic (server) system that communicates with a human user, mediated through their web browser. In a process of communication of this kind, data are transferred by a variety of protocols.

During interactions with web applications, metadata are inevitably created as an integral part of the protocols that the interactions rely upon. In addition to the requested URL, each request to the Internet dictionary involves a whole range of further information being transferred by the web browser in different HTTP(S) headers, for example:

- its own IP address (*Host*),
- an identification of the browser type (*User-Agent*),
- preferred data formats for direct display (*Accept*),
- preferred language (*Accept-Language*),
- the URL of the last retrieved page (*Referer*),
- a wish (not) to leave a user profile on the server (*DNT*, “do not track”).

When retrieving a URL – for example, by entering a search term in a search field or by clicking on a link – different *parameters* may be sent to the server. In this way, the values that users have entered in a form on a web page (date of access, search criteria, personal settings) can be transferred to the web application. Depending on the method of the HTTP(S) request being used, either these parameters appear in a so-called query string as part of the URL in the address line of the browser (GET method) or they are sent opaquely for the dictionary user as part of the actual HTTP(S) message (POST method).

The dictionary web application can also send further data to the browser in addition to the information explicitly requested by the user. These data – so-called *cookies* – are stored locally by the browser and transferred back, on request, and usually unnoticed. Often cookies serve to identify the user through a unique token, typically after they have registered on a website to have a list saved of the entries they have already searched for or other information that has to be made available once another page has been retrieved. As such, cookies can be understood in many cases as a form of logging in the browser, but with the particular characteristic that these logged datasets can be evaluated by the dictionary web application itself while it is running. There are a variety of processes by which to send cookies to a browser, for example, using the HTTP(S) protocol.

The two most important uses for log information are, first, to analyse problems when program errors or general technical errors occur in the functioning of the dictionary application and, second, to analyse user behaviour and their interactions with the dictionary web application. The analysis of technical problems will not be discussed further here since it depends very strongly on the specific implementation of particular web applications. However, a whole chapter in this volume is dedicated to the analysis of user behaviour (→ Chapter 9). For that reason, the emphasis in what follows will be more on details about the type of metadata that can be gained for this purpose.

The dictionary application communicates first of all with a technical system that is identifiable through its IP address (HTTP(S) 'Host' request header). However, these IP addresses are often allocated dynamically and are, therefore, not associated permanently with a particular device. Many devices also operate behind a so-called shared gateway, which, for example, processes the whole outgoing communication of an organisation through a single IP address. In this way, a simple reference to an IP address does not make it possible to reliably identify a particular device (and therefore a single user). This uncertainty can be countered in part through further log information. In addition to the client's IP address, information about the type of the user's browser can be taken into account. Conclusions can also be drawn from the URL of the last page visited and the time of retrieval. While this approach generally works well for small groups of users sharing a common IP address, it often fails to reliably identify users from larger groups. For some research questions on user interaction, it may not be necessary, though, to actually identify specific users. Instead, it may suffice to focus on the behaviour of groups of users (identified as a group by common metadata features) or on single-step interactions such as the consecutive retrieval of two pages regardless of which user interacted in this scenario.

Reliable observations of a specific user (*tracking*) become possible when the user has registered on the dictionary application (i.e. they are assigned a unique identifier) or when the application silently assigns a unique identifier (e.g. a cookie) to the browser used. Using either of these unique identifiers, all of the interactions of the user can be read from the log data as long as the identifiers are stored in the logs.

Of course, not everything that is technically possible in terms of tracking users is legally permitted. For example, the specific tracking of user behaviour outlined above is generally not allowed in the European Union without the explicit and conscious consent of the user. Various legal regulations describe and limit the types of communication data gathered and their use, above all:

- the EU's General Data Protection Regulation (Regulation 2016/679, GDPR),
- the EU's Privacy and Electronic Communications Directive (Directive 2002/58/EC, ePrivacy Directive),
- data protection acts in EU member states.

In an institutional context, there are often additional and, in part, more specific provisions and guidelines (based on the aforementioned laws) determining which interaction data can be legally and ethically stored and analysed as log data and in which form. There are also appointed individuals with mandates for data protection who can provide help and support.

1.4 Versioning

The reader of a print dictionary is not dependent as a matter of principle on the support of a technical system to be able to use the storage medium of the book. However, the perception of a dictionary that appears in electronic form is not possible without recourse to a suitable device to display it and to navigate through it. The specific type of device – whether it is an electronic translation device, a mobile phone, or, more generally, a computer system – plays no role in our considerations here. What is important is the basic principle common to them all, namely that the presentation of the stored information that can be read by people always has to be generated first from the stored representation of the data. Unlike a book, the content of which is fixed and immutable after the printing process, the underlying data that are stored electronically can be changed dynamically or be replaced relatively easily. The display device will then show the user the updated information (e.g. a revised dictionary entry). There is a series of processes and technologies designed to deal with the new challenges arising from this variability, which will be presented briefly in this section and in → Section 1.5. We begin with the problem that systematic access to different versions of dictionary entries needs to be possible for dictionary creators and dictionary users alike.

Even if this so-called versioning is not a specific Internet technology or a widespread concept in lexicography, it does play a certain role in Internet lexicography, which justifies our treatment of the ideas that lie behind it.

Versioning of digital data is an idea that originated in software development. There, we talk of the life cycle of a piece of software. A program is developed, tested, launched on the market, used, and revised. The revision results in various versions of the same program. The "life" of the program comes to an end when its further development and support are discontinued – which does not mean that the program is no longer in use. In software development, tools that support the managing of versions – in particular of a program's source code – have the following purposes and functions: all changes are recorded and possibly commented on, as appropriate, and earlier stages of development (versions) of the software are archived automatically. It is then possible to go back to them as needed (cf. Baerisch 2005).

The idea of a life cycle has been transferred to documents in the digital world (cf. Lobin 2004). The typical document conceived in this way is a product description or instruction manual that keeps pace with the further development of the product; that is, it must be adapted without being written completely afresh. In this case, we can speak of multiple versions of this document that have to be managed so that the authors of the document retain an overview of them.

In the world of printed texts, there is a comparable concept: the *edition*. A text can appear in several editions. It can be reproduced unchanged from edition to edition but it can also be changed to a greater or lesser extent. The authors usually give very brief information in a foreword about the changes to the text that characterise the new edition.

The most important differences between an edition (of a book) and a version (of software or a document) are as follows:

- The time period between two editions normally amounts to one or more years. The gaps between two versions of a piece of software or a document accessible online are typically considerably shorter.
- The scope of an edition usually extends to the whole printed work (the book is in its fifth edition); in software development, a whole software package can be versioned but also a single module. As far as documents are concerned, the versioning may apply to single chapters, sections, or – in the case of lexicographic works – entries, or even just parts of entries.
- The documentation of changes in a new version (the so-called "change log") is usually more detailed than the "foreword to the new edition" in a book.

Internet lexicography involves two different types of documents and two different types of "users". On the one hand, it is possible to view a whole dictionary as a single document; on the other hand, a single module, typically a dictionary entry, but also supporting texts can each be conceived as individual documents. These different perspectives on granularity correspond to different user perspectives: while dictionary users will generally consider a dictionary as a fixed set of entries, for lexicographers the focus is often on a single entry (and possibly on closely related entries) and on the entry's individual stages of development over time.

As a result, the following applies to managing versions when compiling an Internet dictionary:

- It makes sense for lexicographers to version at the level of individual entries. Indeed, this is necessary when several lexicographers are working on the same entry. It has to be possible to recreate older stages of development of an entry and to compare the different versions with one another. A particular version has to be accessible with an unambiguous name or identifier, for example, a version number. Further metadata can be helpful in addition to this name, including the name of the individual who created this version, the time at which the version was created, and a description of the change to this version compared to the previous one.
- For the user of the "finished" product, i.e. a dictionary, which is typically accessed through a browser, versioning at the level of the whole work is often sufficient. In this case, a description should be provided of the important changes from the previous version – and cumulatively from the version before last, and so on. An indication of when this version was made available is also helpful. Corresponding supporting texts should also belong to the "product contents" of the dictionary. As a rule, there is no expectation that earlier versions of the dictionary or individual entries should be accessible since the resources required for that are very exten-

sive. So-called “wikis” (e.g. Wikipedia and Wiktionary) remain the exception. The version management of individual entries is an integral part of these systems and, thus, is available to all users (on the grounds that users can, in principle, also create or edit entries; → Chapter 8).

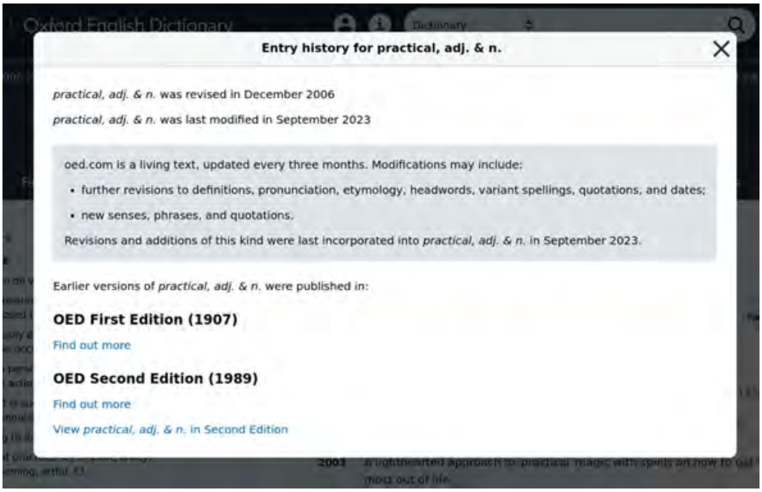
From this, we can derive the following recommendations for the planning of a dictionary to be published on the Internet.

Whenever a lexicographic process is to be planned or undertaken (→ Chapter 3), fundamental decisions need to be taken about data management. One of these decisions is whether different instances of a document should be created at the individual stages of the lexicographic process and how they should be dealt with. With every new instance of a document, previous iterations can either be discarded or conserved. If a database management system is to be employed for conserving data (→ Chapter 4.2.2), then it is important to be aware that this type of software does not automatically support the management of different versions. Each change to the stored data overwrites the previous version. For version management in these cases, the storing of data has to be conceived in such a way that any changes made to the data result in new datasets with appropriate metadata instead of simply modifying existing datasets. However, this requires greater technical effort, which has to be taken into account when planning the project. Alternatively, it is possible to employ a wiki system, where, as we have seen, version management is already built in. A third alternative is to combine an editing system with a version control system (VCS), as is typically the case in software development. Subversion (https://subversion.apache.org/) and Git (https://git-scm.com/) are common examples of such VCS among many others. Using dedicated VCSs can require extensive technical knowledge. For example, so-called “version conflicts” may have to be resolved if the VCS does not provide exclusive locking of resources to prevent two lexicographers from simultaneously and independently working on the same entry. Should version conflicts arise, they would have to be resolved in such a way that a single common version exists after merging the conflicting entry versions.

Editing cycles also have to be taken into account when planning the publication of Internet dictionaries. Here, wiki systems are again the simplest solution. Each change is immediately visible online, and changes can be undone relatively easily. However, the process of checking can be very complicated and time-consuming when a large number of changes are involved. In online reference works that are compiled with a limited set of editors, a dictionary entry will only be published after rigorous checking and approval. Updates to individual entries may not be made public as soon as they are approved. Rather, updates to the dictionary could be made in bulk at certain intervals. The model of the edition in print lexicography is an extreme case of this: the whole work is published afresh after a period of several years or even decades. The other extreme is the publication of individual updated and approved entries. The practice for updating most online dictionaries will probably lie somewhere in between. Possible options are updates at particular intervals (e.g. monthly or weekly) or when a specific number of approved entries, either new or revised, are ready for publication.

Finally, we provide two examples of versioning in large online lexicographic projects:

- OED Online: in the online version of the "Oxford English Dictionary" (https://www.oed.com/), a link is provided in the top left corner of each entry stating the most recent year of its revision. When following the link, a more detailed summary of the entry's revision history appears, summarising all major revisions and the last minor revision of the entry. The information in the revision history refers either to editions of a volume (i.e. "OED First Edition 1907"), to one of the supplementary (published) volumes, or to an "online version" (→ Fig. 1.4). We are unaware of the exact internal version management practised by the OED editors.
- DWDS: the "Digital Dictionary of the German Language" ("Digitales Wörterbuch der deutschen Sprache") is conceived as a lexical information system that encompasses several dictionaries, linguistic corpora, and statistical tools (cf. Klein/Geyken 2010). Some of these dictionaries are regularly updated, including the "Etymological Dictionary of the German Language" ("Etymologisches Wörterbuch des Deutschen", also known as the Pfeifer-DWDS). Work on the print version has long since been completed, with three editions published between 1989 and 1995. However, the principal author, Wolfgang Pfeifer, worked continuously on revising existing entries and compiling new ones for the online version until his death in 2020. Until that point, his dictionary in the DWDS was updated around twice a month, with each version of the Pfeifer-DWDS being assigned its own version



Entry history for practical, adj. & n.

practical, adj. & n. was revised in December 2006

practical, adj. & n. was last modified in September 2023

oed.com is a living text, updated every three months. Modifications may include:

- further revisions to definitions, pronunciation, etymology, headwords, variant spellings, quotations, and dates;
- new senses, phrases, and quotations.

Revisions and additions of this kind were last incorporated into *practical, adj. & n.* in September 2023.

Earlier versions of *practical, adj. & n.* were published in:

OED First Edition (1907)

Find out more

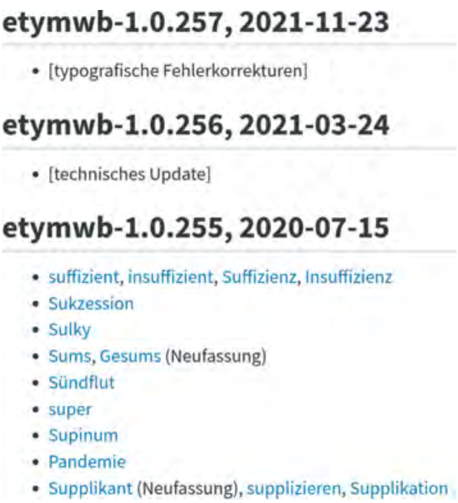
OED Second Edition (1989)

Find out more

View *practical, adj. & n.* in Second Edition

Fig. 1.4: Publication history of the entry for *practical* in the OED.

number. The version number has three parts so that more significant changes can be distinguished from more minor ones. A change in the format of the data (e.g. more detailed tagging of information within the entries) led to a new version, even if these changes might not always be visible to the user. All versions are archived and can be retrieved as necessary, although this option is not offered on the website. The change log for this resource is documented on a separate page (https://www.dwds.de/wb/etymwb/changes; → Fig. 1.5). An editing system is used in the DWDS to compile the entries of the main dictionary, which relies on the Git VCS. In this way, all versions of an entry are automatically archived and can be consulted on demand.



etymwb-1.0.257, 2021-11-23

- [typografische Fehlerkorrekturen]

etymwb-1.0.256, 2021-03-24

- [technisches Update]

etymwb-1.0.255, 2020-07-15

- suffizient, insuffizient, Suffizienz, Insuffizienz
- Sukzession
- Sulky
- Sums, Gesums (Neufassung)
- Sündflut
- super
- Supinum
- Pandemie
- Supplikant (Neufassung), supplizieren, Supplikation

Fig. 1.5: Version information for the Pfeifer-DWDS.

1.5 Persistence and identity

Moving beyond project-internal version management, the possibility of updating the data available on the server at any time raises questions about the longevity (*persistence*) and identity of electronic data. The term persistence designates the property of an object to remain unchanged over a long period of time. This property applies to print dictionaries by their very nature. The carrier medium of paper can last for several centuries when stored appropriately without the fixed written form changing. However, the usual storage media for electronic documents typically demonstrate a much shorter lifespan. In part, this is due to the materials used. Chemical changes can occur in the synthetic materials that are used as the carrier medium or as a protective layer (e.g. in magnetic tapes and optical storage media, such as CDs or DVDs). The storage cells in some non-volatile semiconductors (e.g. flash storage in memory sticks and cards) degenerate when they are written onto; their likelihood of failure increases every time they are written onto. For this reason, this type of semiconductor storage has integrated components for recognising and correcting errors as well as reserve storage for rescuing data from defective areas.

Understood more broadly, the concept of persistence can also be applied to the processes used for storing data, that is, to the technological methods, tools, and agreed conventions for storing and receiving data on the storage medium. In the case of a printed or handwritten book, the process consists of mechanically fixing the (agreed) written symbols on paper (writing, printing, stamping with an appropriate tool) and the direct recognition of these written symbols (optical and haptic reading). In electronic storage processes, storage and reception take place with the aid of technical devices. Consequently, stored data cannot be accessed by humans without technological tools. This results in the persistence of the storage process being strongly dependent on the availability of the necessary storage and reading devices as well as on the data encoding used (that is, the representation conventions agreed upon) being supported by the display device. In order to be able to read an old magnetic tape with typesetting instructions for a particular dictionary, not only does the magnetic tape need to be available (and as intact as possible) but also a suitable tape reader and, in some circumstances, a further device or program to extract the typesetting data from the data stream of the tape player.

As a rule, no detailed distinction is made between the aspects of persistence outlined above. Instead, the term persistence generally refers to the theoretical and temporally non-specific availability and usability of a dataset. Here, there is a tendency to abstract from the specific storage technology being used (storage medium and process). In particular, the storage technology actually used on the server side is ultimately irrelevant for the user accessing data over a network.

Because of the ease with which electronic data can be altered, it becomes possible to publish corrections, revisions, or new entries at any time; improve the access structures dynamically; or even extend the types of lexicographic information (→ Chapter 3). Nonetheless, if these possibilities are used as part of versioning (→ Section 1.4), this has far-reaching consequences for the way the dictionary and its parts can be cited. Depending on the time when they access it, a dictionary user will see a very particular version of a dictionary entry. In order to cite that version of the entry, they could provide the URL and the exact time at which the page was retrieved. Taken together, this information would represent a version-specific indicator. Nonetheless, both details are arbitrary: a URL is not a fixed indicator (it can, in theory, be changed by the provider of the dictionary at any time) and the time will, as a rule, be one of any number of times that all refer to the same entry version since edited entries are not updated that frequently. In addition, Internet dictionaries do not usually provide for a time-specific query that would make it possible to download the entry again in the form in which it appeared at a particular moment in time, unlike collaborative platforms like WIKTIONARY, as already explained, which offer a version history for every entry. In order to avoid the arbitrariness of the retrieval date, it is preferable to indicate the date and time on which the specific version of the dictionary entry was published.

Yet this does not resolve the problem of the URL lacking persistence. For this reason, there are now several services that provide *persistent identifiers (PIDs)*, including:

- DOI (Digital Object Identifier, https://www.doi.org/),
- ePIC handles (persistent identifiers for eResearch, https://www.pidconsortium.eu/),
- URN (Uniform Resource Name, e.g. URN:NBN at the German National Library, https://nbn-resolving.org/),
- PURL (Persistent Uniform Resource Locators, https://purl.archive.org/).

PIDs have the function of providing a stable abstract address for an electronic resource. The PID can be resolved in order to derive the actual address from the abstract one. This is done by looking up the correct allocation of the PID in a directory that lists the allocation of all PIDs to "traditional" URLs. As such, the persistence of a PID is based on the guarantee that the consortia or institutions concerned will ensure a reliable correspondence to "classical" URLs. The dictionary providers themselves must, in turn, take responsibility for the accuracy and accessibility of this URL if they wish to offer PIDs to their users. If they alter the URL for their dictionary, they must ensure that the corresponding mapping of PIDs to URLs in the PID directory also changes.

PIDs are agnostic when it comes to changes in the content of the resources to which they refer. They can be used to identify single versions of the whole dictionary in a persistent way, or an individual dictionary entry, or a dynamic resource, that is, one that changes with time or depending on context. In the case of individual entries, the same number of PIDs are needed as there are versions of an entry, with each PID referring to a single version of the entry. The version number of the entry (e.g. the date of publication) must be retained in the URL when it is resolved. In the case of a dynamic resource, only one PID is needed per entry. When it is resolved, the URL should lead to the most up-to-date version of the entry. A versioning of the entry can be provided on the web page for the Internet dictionary independently of the PID, as discussed in → Section 1.4; it would then not be possible to provide direct addresses for individual versions using a PID.

It is also possible to mix the two ways of working with PIDs that we discussed in relation to versioning and unique identification. Thus, a PID might refer to the dictionary as a whole in its current form while provision is made for individual entries to have version-specific PIDs to allow for more accurate citation.

Embracing an Internet archive such as www.archive.org brings forth a plethora of advantages compared to being solely reliant on persistent URLs. The major benefit lies in the preservation of web content over time. While persistent URLs may succumb to the inevitability of “link rot”, an Internet archive acts as a digital time capsule, capturing and storing historical versions of websites. This capability not only safeguards against broken links but also allows users to access and reference content that may have undergone alterations or been removed entirely from its original source.

In addition to mitigating the risks of link deterioration, Internet archives provide a robust solution for the continuity of access. Persistent URLs are effective only as long as the original website remains available. In cases of temporary downtime or permanent cessation, Internet archives often serve as a reliable alternative, ensuring that users can still retrieve valuable information from archived versions of web pages. This accessibility during downtimes contributes to the resilience of information dissemination and proves especially beneficial for researchers, educators, and the general public seeking reliable sources beyond the limitations of persistent URLs.

1.6 Concluding remarks

At the beginning of this chapter, we addressed the radical new possibilities available to producers and users of lexicographic products as a result of computer and network technology. These were set out with great clarity at a very early stage in their development (cf. de Schryver 2003). This also includes the observation that it makes little sense in many contexts to distinguish between the relatively traditional use of such products as a technological continuation of print dictionaries and other ways of using digital lexicographic resources. Thus, the following definition by Nesi (2000: 839) has remained valid for over two decades:

> The term *electronic dictionary* (or ED) can be used to refer to any reference material stored in electronic form that gives information about the spelling, meaning, or use of words. Thus a spell-checker in a word-processing program, a device that scans and translates printed words, a glossary for on-line teaching materials, or an electronic version of a respected hardcopy dictionary are all EDs of a sort, characterised by the same system of storage and retrieval.

Increasingly, lexicographic and encyclopaedic information about words is presented by search engines in a particular part of the screen – often next to links to corresponding entries in various Internet dictionaries – or it is even available at any time in applications, for example, by double-clicking on any word. Dictionaries are being used in ever greater numbers of applications from the field of natural language processing, unseen by users, including in language-learning and correction programs.

In all of these contexts, the authors of lexicographic works, together with their specialist authority, and the works themselves are becoming increasingly invisible to users. Googling words is now a frequent substitute for consulting a recognised dictionary. Even the physical handling of dictionaries is disappearing from our consciousness thanks to computer programs presenting results from the process of “consulting” digital dictionary data in a form that has already been further edited. In this way, the digital revolution of lexicography can be viewed not only from a technological perspective but also from a sociological one. As early as 1997, Niklas Luhmann was already offering a diagnosis in the context of computer-aided communication technology in his abstract systems-theoretical distinction between loosely coupled elements of a *medium* (Luhmann 1997: 309–310):

> Mit all dem ist die soziale Entkopplung des medialen Substrats der Kommunikation ins Extrem getrieben. In unserer Begrifflichkeit muß das heißen, daß ein neues Medium im Entstehen ist, dessen Formen nun von den Computerprogrammen abhängig sind. [With all this, the social decoupling of the medial substrate of communication is pushed to the extreme. In our conceptualisation, it must mean that a new medium is coming into being, the form of which is dependent on computer programs.]

Bibliography

Further reading

Fischer, Peter/Witt, Andreas (2014): Best Practices on Long-Term Archiving of Spoken Language Data. In: Ruhi, Şükriye/Haugh, Michael/Schmidt, Thomas/Wörner, Kai (eds.): *Best Practices for Spoken Corpora in Linguistic Research*. Newcastle: Cambridge Scholars Publishing, 162–182. *Despite the divergent thematic focus, a good overview of the problems of and possible solutions to the long-term storage and availability of language-related data with extensive further reading*.

MDN Web Docs. Mountain View, California, USA: Mozilla Foundation. Online: https://developer.mozilla.org/. *Very comprehensive and up-to-date open-source, collaborative project that documents web technologies. Alongside detailed technical reference documentation for experts, MDN web docs provides extensive learning resources for students and beginners. The documentation is primarily available in English, but translations of most documents are available in a number of large languages (at the time of writing, Chinese, French, Japanese, Korean, Portuguese, Russian, and Spanish).*

Literature

Academic literature

Baerisch, Stefan (2005): *Versionskontrollsysteme in der Softwareentwicklung*. Bonn. https://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/iz_arbeitsberichte/ab_36.pdf [last access: April 25, 2024].

de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic Dictionary Age. In: *International Journal of Lexicography* 16, 143–199.

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

Lobin, Henning (2004): Textauszeichnungssprachen und Dokumentgrammatiken. In: Lobin, Henning/Lemnitzer, Lothar (eds.): *Texttechnologie. Perspektiven und Anwendungen*. Tübingen: Stauffenburg, 51–82.

Luhmann, Niklas (1997): *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp.

Nesi, Hilary (2000): Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art. In: Heid, Ulrich, et al. (eds.): *Proceedings of the Ninth Euralex International Congress, EURALEX 2000, Stuttgart, Germany, August 8th–12th, 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 839–847.

TEI Consortium (eds.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.7.0. Last updated November 16, 2023, revision e5dd73ed0. TEI Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [last access: April 15, 2024].

Dictionaries and reference works

DWDS = *Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. Ed. by Berlin-Brandenburgische Akademie der Wissenschaften*. Online publishing. https://www.dwds.de/ [last access: April 25, 2024].

OED Online = *Oxford English Dictionary online*. Ed. by Michael Proffitt. Online publishing. https://www.oed.com/ [last access: April 25, 2024].

Pfeifer-DWDS = *Etymologisches Wörterbuch des Deutschen. Digitalisierte und von Wolfgang Pfeifer überarbeitete Version im Digitalen Wörterbuch der deutschen Sprache*. Online publishing. https://www.dwds.de/d/wb-etymwb [last access: April 25, 2024].

Wikipedia = *Wikipedia, The Free Encyclopedia. San Francisco*. Ed. by Wikimedia Foundation. Online publishing. https://www.wikipedia.org/ [last access: April 25, 2024].

Wiktionary = *Wiktionary, the free dictionary*. Ed. by Wikimedia Foundation. Online publishing. https://en.wiktionary.org/ [last access: April 25, 2024].

Images

Fig. 1.1 “Nessie II”: https://upload.wikimedia.org/wikipedia/commons/thumb/c/cd/Nessie_II.JPG/800px-Nessie_II.JPG. Friflash, Wikimedia Commons, licensed under Creative Commons Attribution-ShareAlike 4.0, URL: https://creativecommons.org/licenses/by-sa/4.0/legalcode.

Stefan Engelberg and Angelika Storrer

2 A Typology of Internet Dictionaries and Portals

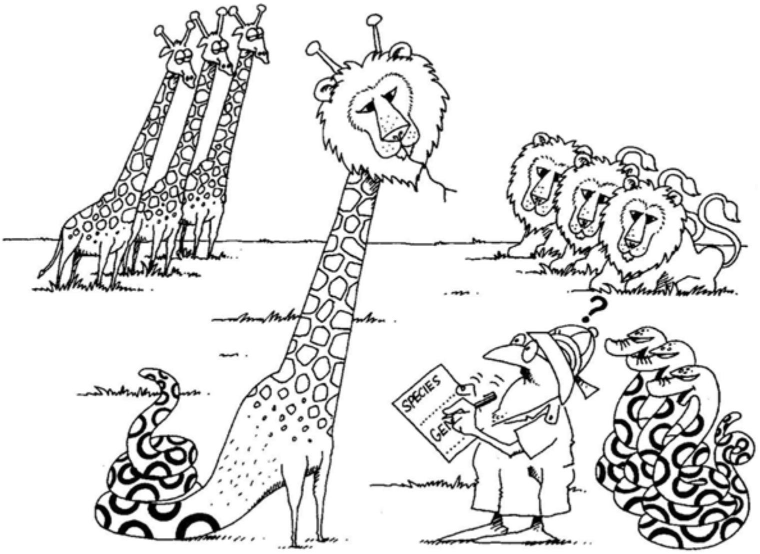


Fig. 2.1: Classification as the academic's task.

There used to be a time when an expert could still easily distinguish a spelling dictionary from a frequency dictionary, a collocation dictionary from a valency dictionary, and a thesaurus from an illustrated dictionary. Rightly enough differentiating between a dictionary, a corpus, an atlas, and a frequency list would have posed not the slightest difficulty. The combination of lexicography and the Internet have made these tasks more difficult: Internet dictionaries are able to bring together many different types of information in new ways and present them in a way that adapts to the user. Networks of dictionaries, blended with corpora, multimedia extensions, and automatic language analysis tools create new types of lexical information systems and dictionary portals.

In this section, we shall attempt to shine a little light in the gloom of different types of dictionaries, dictionary portals, and lexical information systems. In the process, we

Stefan Engelberg, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany,
e-mail: engelberg@ids-mannheim.de
Angelika Storrer, Universität Mannheim, Schloss Ehrenhof Ost, 68161 Mannheim, Germany,
e-mail: angelika.storrer@uni-mannheim.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/10.1515/9783111233758-003

aim to demonstrate that the disorder brought to the classification of the animal kingdom by a “sniraffion” can be worthwhile.

2.1 Introduction

Dictionaries can be categorised according to very different perspectives. A classification according to which each individual dictionary could be assigned precisely to a single class would not do justice to their variety. Therefore, most category systems are typologies, defining the characteristic features of dictionary types. An individual dictionary is then regarded as being more or less representative of a particular type depending on which of those features it exhibits. Already well-elaborated suggestions exist for print dictionaries, for example Hausmann (1989), Engelberg/Lemnitzer (2009), and Wiegand et al. (2010). The features selected for a typology depend on the purpose of the categorisation: Wiegand et al. (2010: 202ff.) distinguish between typologies according to user relationships, according to the dictionary subject matter, according to the dictionary structure, and according to the medium of storage and publication. Typologies according to user relationships are based on features that capture the dictionary function and typical situations of dictionary usage. Kühn (1989) suggests a typology of this kind for print dictionaries. Storrer/Freese (1996), Nesi (2000), Lew (2011), de Schryver (2003), Tono (2004), and Wiegand et al. (2010: 208ff.) present reflections on how to typologize digital dictionaries published on the Internet.

The foundation for our proposal in this chapter is the typology proposed in Engelberg/Lemnitzer (2009), which takes into account both print and digital dictionaries. The most important features are elaborated in → Section 2.2 and → Section 2.3. We focus on the media-specific characteristics of Internet dictionaries and deal with typological features related to the lexicographic processes which give rise to Internet dictionaries. The features introduced here are elaborated in more detail and further differentiated in → Chapter 3 and in → Chapter 8. Several different dictionaries can often be retrieved on the Internet in a single user interface; Engelberg/Lemnitzer (2009: 73) designate these kinds of resources as dictionary portals. → Section 2.4 introduces the typological features of dictionary portals and illustrates the basic types with relevant examples. This discussion is extended in → Chapter 3 and in → Chapter 5.

2.2 Media-specific typological features of Internet dictionaries

2.2.1 Digitised dictionaries vs. newly designed dictionaries

When it comes to the source or origin of lexicographic data, a distinction can be drawn between digitised and newly designed Internet dictionaries. *Newly designed* dictionaries are planned from the very beginning for digital publication and online use. Examples include the 'Algemeen Nederlands Woordenboek' (ANW), and WIKTIONARY-EN. *Digitised dictionaries* are based on print dictionaries, which are then transferred into a digital format. Retrospective digitisation is primarily about making the dictionary accessible on the Internet with its original text; no alterations in content are made. Nonetheless, flexible search tools as well as hyperlinks to sources or to other dictionaries create "added value" compared to the book publication. Examples are the WEBSTER-1828 and the German 'Deutsches Wörterbuch' (DWB-ONLINE). In other Internet lexicography projects the digitisation of print dictionaries is simply the first step in a lexicographic process in which that resulting lexical resource is successively further developed and updated.

Crucial for the distinction between *digitised* and *newly created* is the starting-point of the lexicographic process. If the lexicographic process involves the digitisation of dictionaries that have only previously existed in print form, then it is a *digitised* dictionary. If the data is presented in a digital format from the very beginning of the lexicographic process, then it is a newly designed dictionary. The design of print dictionaries was often influenced by the challenge to accommodate as much information as possible in the smallest possible print space. Abbreviations and other solutions for text condensation are often taken over in the digitisation process. Newly designed dictionaries do not have to cope with this challenge. From the beginning, they can exploit the potential of digital media – e.g., flexible search and access provision, hyperlinks, or multimodal enhancement (→ Chapter 6).

2.2.2 Extension vs. closed version dictionaries

The distinction between extension dictionaries and closed version dictionaries, introduced by Schröder (1997: 16), relates to the completeness or incompleteness of the lexicographic process (→ Chapter 3).

Closed version dictionaries are created in a lexicographic process over a defined time period. Dictionary entries are not altered after the completion of this process. This includes Internet dictionaries that were constructed as part of a time-limited project, e.g., the German IDIOMDATENBANK.

In contrast, *extension dictionaries* are oriented towards continuous addition and revision. Many newly designed Internet dictionaries belong to this type, e.g., the different language versions of the WIKTIONARY or the bilingual dictionaries included in the portal LEO.

2.2.3 Dictionaries without user participation vs. dictionaries with user participation

The Internet is not only a publication medium for dictionaries. Through Internet-based communication services, it also offers multiple ways to get in contact with users and involve them in the lexicographic process. As a result, we can categorise Internet dictionaries according to whether, and in what form, they make use of these possibilities (cf. also Lew 2011).

Internet dictionaries without user participation make lexicographic information available on the Internet but they offer users only very limited opportunities to be involved, or none at all.

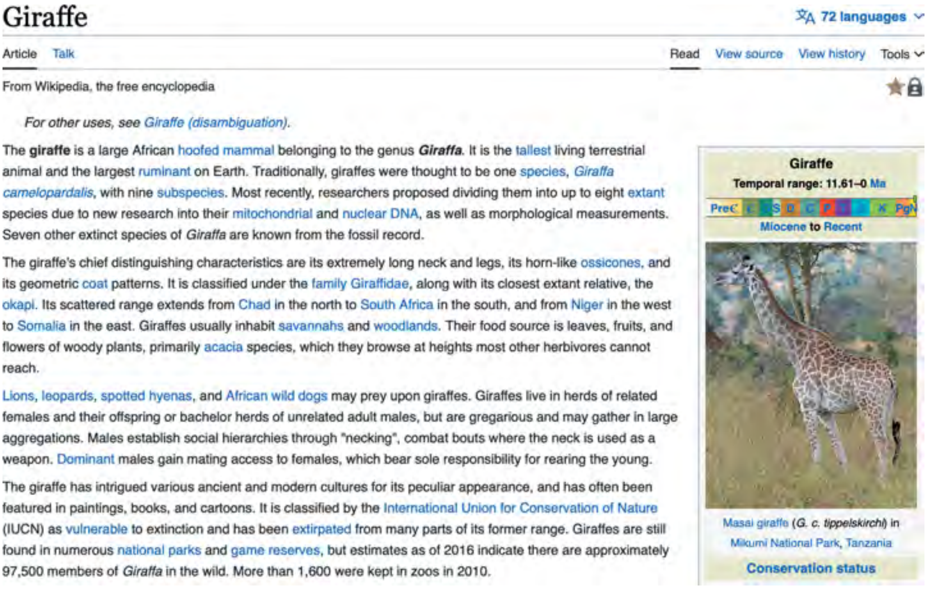
Dictionaries with user participation provide functions which help the users to participate in the content of the lexicographic process. Already in early Internet dictionaries, users could fill in online forms to correct or supplement dictionary articles, pose questions to the dictionary creators, or exchange information in discussion forums (cf. Storrer/Freese 1996). The development of the World Wide Web into the "social web" has led to the forms of participation multiplying further. In projects like WIKTIONARY or URBAN-DICT dictionaries are constructed collectively by users on a voluntary basis and made available free of charge. → Chapter 3 includes a section on the lexicographic process involved in WIKTIONARY projects. → Chapter 8 provides a detailed typology for forms of user involvement (direct, indirect, and associated) and explains these with relevant examples.

2.3 Non media-specific typological features of dictionaries

Research into dictionaries distinguishes, on a very general level, between *language lexicography*, which concentrates on describing linguistic features and *encyclopedic lexicography*, which conveys knowledge about the world. We follow Engelberg/Lemnitzer (2009, ch. 1.2.2) and describe the outputs of *encyclopedic* lexicography as *encyclopedias* and those of language lexicography as *dictionaries*.

On closer inspection, the distinction between *linguistic* and *encyclopedic* knowledge is not straightforward, a point that is fiercely debated in dictionary research.

By comparing the encyclopedic article in the English WIKIPEDIA-EN for the lemma *giraffe* (→ Fig. 2.2) with the dictionary article for *giraffe* in the English WIKTIONARY-EN (→ Fig. 2.3), we can see both, the differences and the overlap. The encyclopedic article conveys knowledge about the giraffe as a species. Where do giraffes live? What do they eat? What are the subspecies of giraffe? etc. In contrast, the dictionary article provides information on the English word *giraffe*. How is the word spelt correctly? How is it pronounced? To which part of speech does the word belong? etc. Overlap exists in the information given about the meaning of the word and its etymology. However, the language-oriented WIKTIONARY-EN aims to list all meanings of the word "giraffe", including metaphorical uses in slang varieties. In contrast, the encyclopedic WIKIPEDIA-EN is focussed on the species *giraffe*.



Giraffe

From Wikipedia, the free encyclopedia

For other uses, see Giraffe (disambiguation).

The **giraffe** is a large African hoofed mammal belonging to the genus ***Giraffa***. It is the tallest living terrestrial animal and the largest ruminant on Earth. Traditionally, giraffes were thought to be one species, *Giraffa camelopardalis*, with nine subspecies. Most recently, researchers proposed dividing them into up to eight extant species due to new research into their mitochondrial and nuclear DNA, as well as morphological measurements. Seven other extinct species of *Giraffa* are known from the fossil record.

The giraffe's chief distinguishing characteristics are its extremely long neck and legs, its horn-like ossicones, and its geometric coat patterns. It is classified under the family Giraffidae, along with its closest extant relative, the okapi. Its scattered range extends from Chad in the north to South Africa in the south, and from Niger in the west to Somalia in the east. Giraffes usually inhabit savannahs and woodlands. Their food source is leaves, fruits, and flowers of woody plants, primarily acacia species, which they browse at heights most other herbivores cannot reach.

Lions, leopards, spotted hyenas, and African wild dogs may prey upon giraffes. Giraffes live in herds of related females and their offspring or bachelor herds of unrelated adult males, but are gregarious and may gather in large aggregations. Males establish social hierarchies through "necking", combat bouts where the neck is used as a weapon. Dominant males gain mating access to females, which bear sole responsibility for rearing the young.

The giraffe has intrigued various ancient and modern cultures for its peculiar appearance, and has often been featured in paintings, books, and cartoons. It is classified by the International Union for Conservation of Nature (IUCN) as vulnerable to extinction and has been extirpated from many parts of its former range. Giraffes are still found in numerous national parks and game reserves, but estimates as of 2016 indicate there are approximately 97,500 members of *Giraffa* in the wild. More than 1,600 were kept in zoos in 2010.

Giraffe

Temporal range: 11.61–0 Ma

Miocene to Recent

Masai giraffe (*G. c. tippelskirchi*) in Mikumi National Park, Tanzania

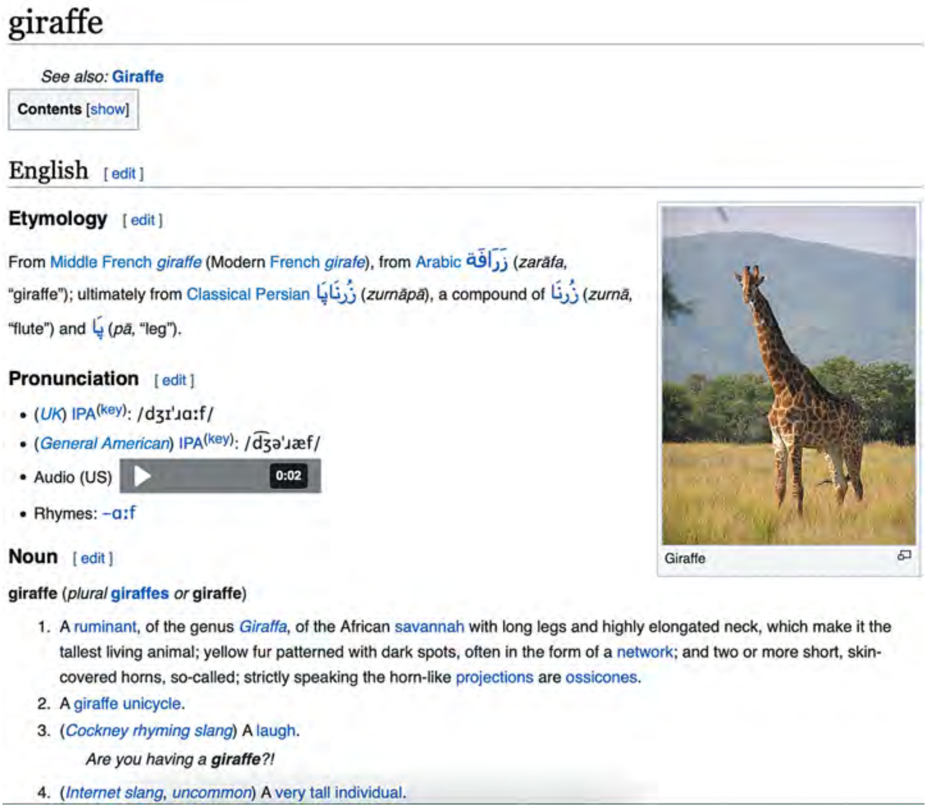
Conservation status

Fig. 2.2: Extract from the article on the lemma *giraffe* in WIKIPEDIA-EN.

Our categorisation for Internet dictionaries concentrates on language lexicography, that is, on dictionaries. Nevertheless, there are dictionaries that incorporate both linguistic and encyclopedic knowledge, in particular dictionaries describing technical terms or special-field vocabulary.

2.3.1 Monolingual – bilingual – multilingual

Language dictionaries can be divided into *monolingual dictionaries*, which take as their lexicographic subject matter a single language, and bilingual dictionaries, which



giraffe

See also: **Giraffe**

Contents [show]

English [edit]

Etymology [edit]

From Middle French *giraffe* (Modern French *girafe*), from Arabic زَرَافَة (*zarāfa*, “giraffe”); ultimately from Classical Persian زُرنَاپَا (*zurnāpā*), a compound of زُرنَا (*zurnā*, “flute”) and پَا (*pā*, “leg”).

Pronunciation [edit]

- (*UK*) IPA(key): /dʒɪˈɹɑːf/
- (*General American*) IPA(key): /d͡ʒəˈɹæf/
- Audio (US) 0:02
- Rhymes: -ɑːf

Noun [edit]

giraffe (*plural* **giraffes** *or* **giraffe**)

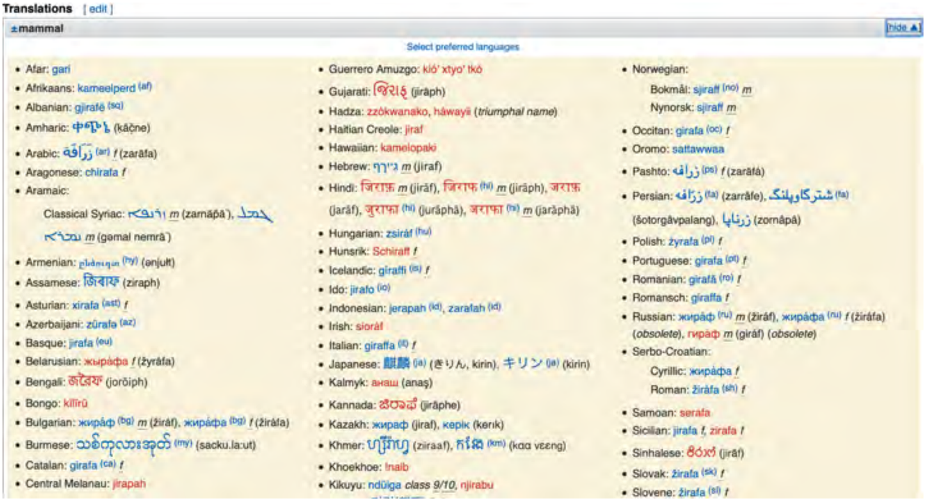
1. A ruminant, of the genus *Giraffa*, of the African savannah with long legs and highly elongated neck, which make it the tallest living animal; yellow fur patterned with dark spots, often in the form of a network; and two or more short, skin-covered horns, so-called; strictly speaking the horn-like projections are ossicones.
2. A giraffe unicycle.
3. (*Cockney rhyming slang*) A laugh.
   Are you having a ***giraffe****?!*
4. (*Internet slang, uncommon*) A very tall individual.

Fig. 2.3: Extract from the article on the lemma *giraffe* in WIKTIONARY-EN.

assign equivalents to the lemmas (headwords) in a target language. Bilingual dictionaries are important above all for language learning and teaching, for the reception and production of texts in a foreign language, and for translation.

There are a variety of portals on the Internet that provide bilingual dictionaries for multiple pairs of languages, e.g., the portals LEO, LINGUEE, COLLINS, and LAROUSSE. Many of such portals are also available as apps for mobile Internet access.

In *multilingual* dictionaries, the lemmas are assigned equivalents from multiple target languages. One example of this is WIKTIONARY-EN, which includes comprehensive lists of equivalents in many languages (→ Fig. 2.4). The ‘Unisa Multilingual Proverb Dictionary’ (UNISA-PROVERB) is an example of a multilingual dictionary, in which equivalents for proverbs are available in four different languages (→ Fig. 2.5).



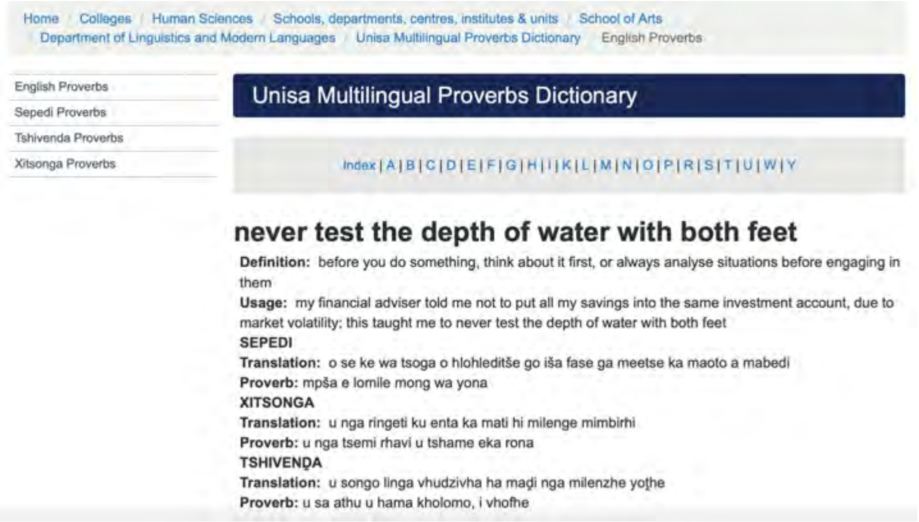
Translations [edit]

±mammal [hide ▲]

Select preferred languages

- Afar: gari
- Afrikaans: kameelperd (af)
- Albanian: gjirafë (sq)
- Amharic: ቀጭኔ (käčne)
- Arabic: زَرَافَة (ar) *f* (zarāfa)
- Aragonese: chirafa *f*
- Aramaic:
  Classical Syriac: ܙܪܢܦܐ *m* (zarnāpā), ܓܡܠ ܢܡܪܐ *m* (gəmal nemrā)
- Armenian: ընձուղտ (hy) (ənjult)
- Assamese: জিৰাফ (ziraph)
- Asturian: xirafa (ast) *f*
- Azerbaijani: zürafə (az)
- Basque: jirafa (eu)
- Belarusian: жыра́фа *f* (žyrāfa)
- Bengali: জিরাফ (jorōiph)
- Bongo: kīlīrū
- Bulgarian: жира́ф (bg) *m* (žiráf), жира́фа (bg) *f* (žiráfa)
- Burmese: သစ်ကုလားအုတ် (my) (sacku.la:ut)
- Catalan: girafa (ca) *f*
- Central Melanau: jirapah
- Guerrero Amuzgo: kió' xtyo' tkó
- Gujarati: જિરાફ (jirāph)
- Hadza: zzókwanako, háwayii (*triumphal name*)
- Haitian Creole: jiraf
- Hawaiian: kamelopaki
- Hebrew: ג׳ירף *m* (jiraf)
- Hindi: जिराफ़ *m* (jirāf), जिराफ (hi) *m* (jirāph), जराफ़ (jarāf), जुराफा (hi) (jurāphā), जराफा (hi) *m* (jarāphā)
- Hungarian: zsiráf (hu)
- Hunsrik: Schiraff *f*
- Icelandic: gíraffi (is) *f*
- Ido: jirafo (io)
- Indonesian: jerapah (id), zarafah (id)
- Irish: sioráf
- Italian: giraffa (it) *f*
- Japanese: 麒麟 (ja) (きりん, kirin), キリン (ja) (kirin)
- Kalmyk: анаш (anaş)
- Kannada: ಜಿರಾಫೆ (jirāphe)
- Kazakh: жираф (jiraf), керік (kerık)
- Khmer: ហ្សីរ៉ាហ្វ (ziiraaf), កវែង (km) (kɑɑ vɛɛng)
- Khoekhoe: !naib
- Kikuyu: ndũiga *class 9/10*, njirabu
- Norwegian:
  Bokmål: sjiraff (no) *m*
  Nynorsk: sjiraff *m*
- Occitan: girafa (oc) *f*
- Oromo: sattawwaa
- Pashto: زرافه (ps) *f* (zarāfá)
- Persian: زرّافه (fa) (zarrâfe), شترگاوپلنگ (fa) (šotorgâvpalang), زرناپا (zornâpâ)
- Polish: żyrafa (pl) *f*
- Portuguese: girafa (pt) *f*
- Romanian: girafă (ro) *f*
- Romansch: giraffa *f*
- Russian: жира́ф (ru) *m* (žiráf), жира́фа (ru) *f* (žiráfa) (*obsolete*), гира́ф *m* (giráf) (*obsolete*)
- Serbo-Croatian:
  Cyrillic: жира̏фа *f*
  Roman: žìrāfa (sh) *f*
- Samoan: serafa
- Sicilian: jirafa *f*, zirafa *f*
- Sinhalese: ජිරාෆ් (jirāf)
- Slovak: žirafa (sk) *f*
- Slovene: žirafa (sl) *f*

Fig. 2.4: Extract from the equivalent list for *giraffe* in the English WIKTIONARY-EN.



Home / Colleges / Human Sciences / Schools, departments, centres, institutes & units / School of Arts / Department of Linguistics and Modern Languages / Unisa Multilingual Proverbs Dictionary / English Proverbs

English Proverbs
Sepedi Proverbs
Tshivenda Proverbs
Xitsonga Proverbs

Unisa Multilingual Proverbs Dictionary

Index | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | R | S | T | U | W | Y

never test the depth of water with both feet

Definition: before you do something, think about it first, or always analyse situations before engaging in them

Usage: my financial adviser told me not to put all my savings into the same investment account, due to market volatility; this taught me to never test the depth of water with both feet

SEPEDI
Translation: o se ke wa tsoga o hlohleditše go iša fase ga meetse ka maoto a mabedi
Proverb: mpša e lomile mong wa yona

XITSONGA
Translation: u nga ringeti ku enta ka mati hi milenge mimbirhi
Proverb: u nga tsemi rhavi u tshame eka rona

TSHIVENḒA
Translation: u songo linga vhudzivha ha maḓi nga milenzhe yoṱhe
Proverb: u sa athu u hama kholomo, i vhofhe

Fig. 2.5: Extract from the article for the proverb *never test the depth of water with both feet* in UNISA-PROVERB.

2.3.2 General dictionaries – special dictionaries

The section of language described in a dictionary is also referred to as its *subject matter*. In the course of the lexicographic process, lemmas (headwords) belonging to the subject matter of the dictionary are selected and described in dictionary articles.

Those compiling the dictionary establish which language features are included in these articles as items, e.g., items giving the phonetics, items giving the meaning, items giving etymology etc. *General dictionaries* place no restrictions on the lexical signs that may be selected as lemmas from the subject matter of the dictionary, and they include as many relevant item classes as possible in their articles. Examples of general dictionaries on the Internet are Wiktionary-En or the online version of the 'Oxford English Dictionary' (OED-Online).

Special dictionaries diverge from these "generalists" in different ways. *Lemma-type-specific* dictionaries focus on particular types of lemmas, e.g., abbreviations, loanwords, place names, or swearwords. *Information-type-specific* dictionaries focus on particular language features, e.g., spelling, pronunciation, word formations, or etymology. *User-group-specific* dictionaries are directed towards special types of user groups and usage situations, e.g., learners' dictionaries, or childrens' dictionaries. *Variety-specific* dictionaries focus on language varieties, such as dialects, language stages, sociolects, and special-field vocabulary. Text-related dictionaries, that is author dictionaries, concordances, and dictionaries of quotations, also belong to this subtype. The online version of the Goethe-Dict or the dictionary ShakespearesWords are examples of author dictionaries available on the Internet.

The overview in → Fig. 2.6 is based on Engelberg/Lemnitzer (2009: 22), where these types are described using a wide range of examples. In the following we present a small selection of special dictionaries on the Internet, to demonstrate some of their media-specific properties.

Lemma-Type-Specific Dictionaries: In the field of onomastic lexicography there are innovative forms of presentation. The place name dictionary Schweizer-Ortsnamen or the database English-Place-Names are examples, which offer comprehensive and flexible search options as well as links to automatically generated map snippets (→ Fig. 2.7).

Information-Type-Specific Dictionaries: Examples for Internet dictionaries specialising in syntagmatic information like idioms, collocations and valency are e-VALBU, a database describing the valency of German verbs, or the 'Oxford Collocations Dictionary' (OCD). In addition, corpus-based dictionary projects have made available the results of automatic co-occurrence and collocation analyses for search terms (→ Chapter 7.4.2). Examples of lexical resources specialising in paradigmatic information are the lexical database WordNet, or the dictionary of synonyms and associated words OpenThesaurus-de, integrated into the German DWDS lexical information system.

User-Group-Specific Dictionaries: The portal 'Oxford Learners' Dictionaries' (OED-Learner) supports learners of the English language with monolingual and bilingual articles, providing simple definitions and examples as well as audio clips to learn the pronunciation.

Variety-specific Dictionaries: The portal Wörterbuchnetz provides access to several digitised dictionaries of German language stages and dialects. They preserve the content of their print versions, but are enhanced with links between articles, or with geo-referencing functions, e.g., the palatine dialect dictionary (Pf-Dict) (→ Fig. 2.8).

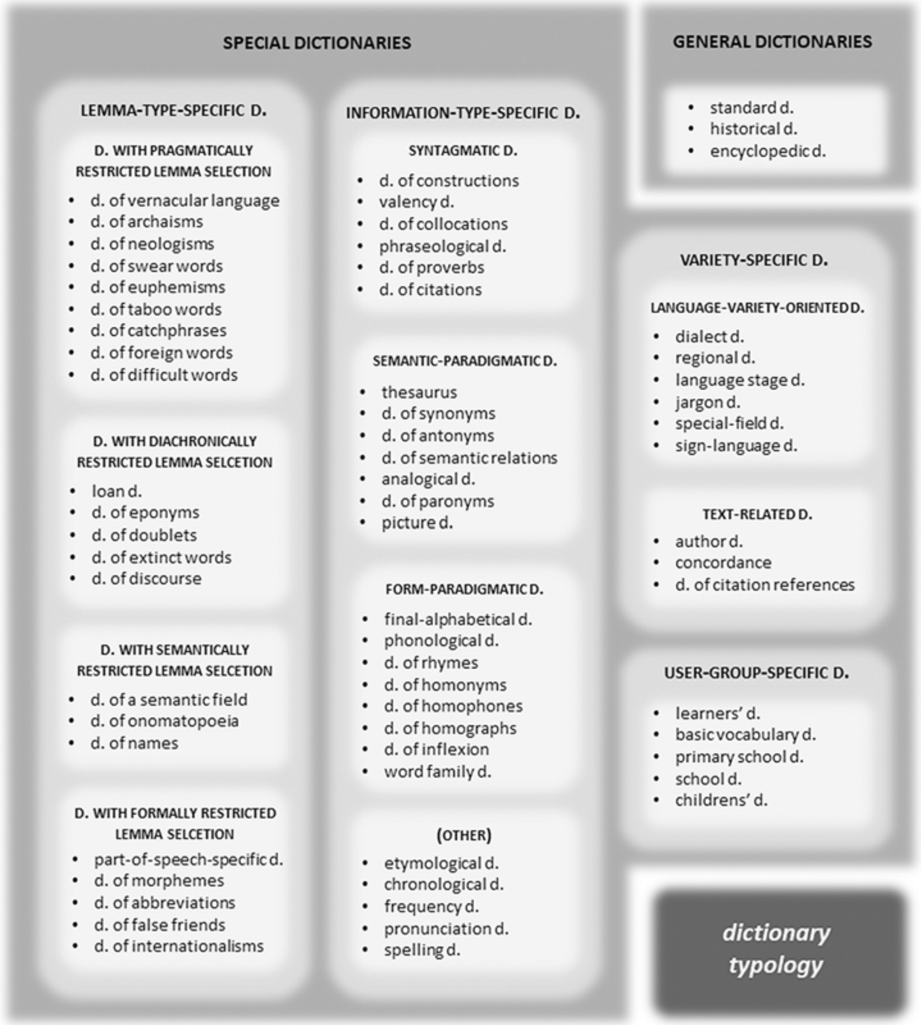
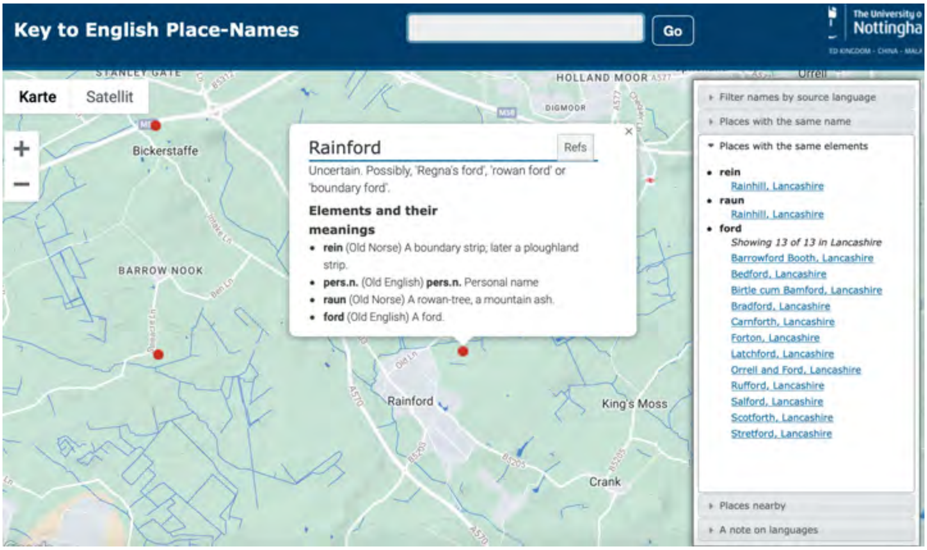


Fig. 2.6: Typology of general and special dictionaries (adapted and translated from Engelberg/Lemnitzer 2009: 22).

Sign language dictionaries, like the 'ASL Sign Language Dictionary' (ASL-DICT), take profit from the possibility of combining videos, text and images in their dictionary articles (→ Fig. 2.9). There are a large number of monolingual and bilingual special-field dictionaries on the Internet. Examples are the MERRIAM-WEBSTER-MEDICAL specialising in medical terms and abbreviations, or the user-generated GLOTTOPEDIA specialising in linguistic terminology. The multilingual dictionary of football language, KICKTIONARY, links words with types of situation on the pitch (e.g., shot, goal, pass) and is structured by semantic relations like hyperonymy, holonymy etc. (→ Chapter 5).



Rainford

Refs

Uncertain. Possibly, 'Regna's ford', 'rowan ford' or 'boundary ford'.

Elements and their meanings

- **rein** (Old Norse) A boundary strip; later a ploughland strip.
- **pers.n.** (Old English) **pers.n.** Personal name
- **raun** (Old Norse) A rowan-tree, a mountain ash.
- **ford** (Old English) A ford.

Filter names by source language

Places with the same name

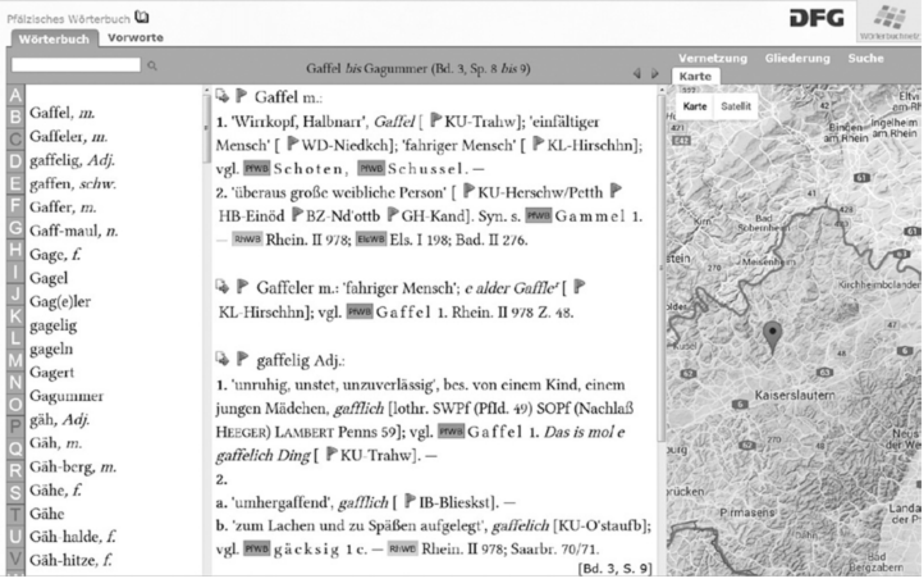
Places with the same elements

- **rein**
  Rainhill, Lancashire
- **raun**
  Rainhill, Lancashire
- **ford**
  Showing 13 of 13 in Lancashire
  Barrowford Booth, Lancashire
  Bedford, Lancashire
  Birtle cum Bamford, Lancashire
  Bradford, Lancashire
  Carnforth, Lancashire
  Forton, Lancashire
  Latchford, Lancashire
  Orrell and Ford, Lancashire
  Rufford, Lancashire
  Salford, Lancashire
  Scotforth, Lancashire
  Stretford, Lancashire

Places nearby

A note on languages

Fig. 2.7: Extract from the article on the lemma *Rainford* in the database ENGLISH-PLACE-NAMES.



Gaffel bis Gagummer (Bd. 3, Sp. 8 bis 9)

Gaffel, *m.*
Gaffeler, *m.*
gaffelig, *Adj.*
gaffen, *schw.*
Gaffer, *m.*
Gaff-maul, *n.*
Gage, *f.*
Gagel
Gag(e)ler
gagelig
gageln
Gagert
Gagummer
gäh, *Adj.*
Gäh, *m.*
Gäh-berg, *m.*
Gähe, *f.*
Gähe
Gäh-halde, *f.*
Gäh-hitze, *f.*

Gaffel m.:

1. 'Wirrkopf, Halbnarr', *Gaffel* [KU-Trahw]; 'einfältiger Mensch' [WD-Niedkch]; 'fahriger Mensch' [KL-Hirschhn]; vgl. PfWB Schoten, PfWB Schussel. –
2. 'überaus große weibliche Person' [KU-Herschw/Petth HB-Einöd BZ-Nd'ottb GH-Kand]. Syn. s. PfWB Gammel 1. – RhWB Rhein. II 978; ElsWB Els. I 198; Bad. II 276.

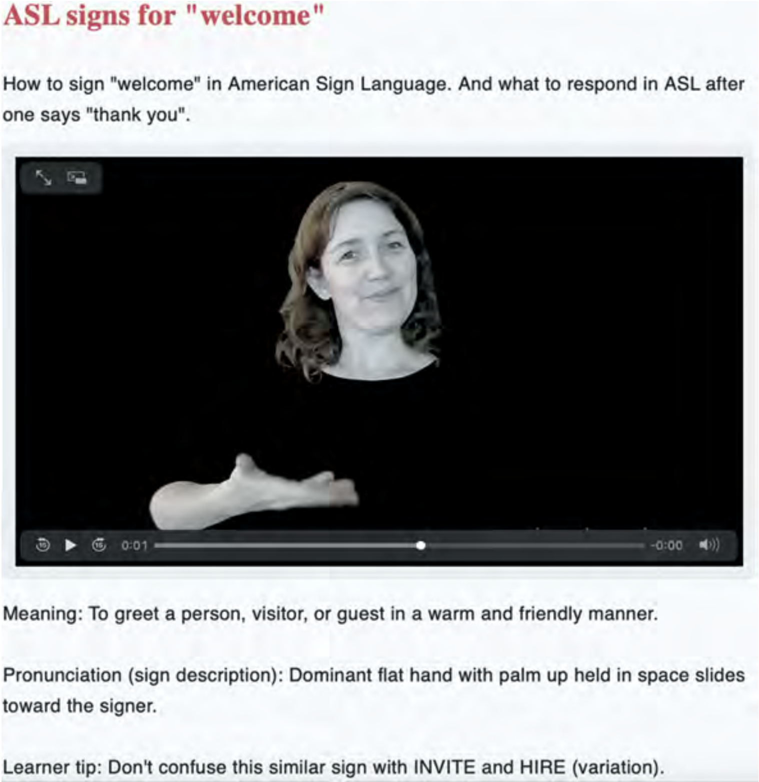
Gaffeler m.: 'fahriger Mensch'; *e alder Gaffle'* [KL-Hirschhn]; vgl. PfWB Gaffel 1. Rhein. II 978 Z. 48.

gaffelig Adj.:

1. 'unruhig, unstet, unzuverlässig', bes. von einem Kind, einem jungen Mädchen, *gafflich* [lothr. SWPf (Pfld. 49) SOPf (Nachlaß HEEGER) LAMBERT Penns 59]; vgl. PfWB Gaffel 1. *Das is mol e gaffelich Ding* [KU-Trahw]. –
2.
a. 'umhergaffend', *gafflich* [IB-Blieskst]. –
b. 'zum Lachen und zu Späßen aufgelegt', *gaffelich* [KU-O'staufb]; vgl. PfWB gäcksig 1 c. – RhWB Rhein. II 978; Saarbr. 70/71.

[Bd. 3, S. 9]

Fig. 2.8: Extract from the PF-DICT in the WÖRTERBUCHNETZ.



ASL signs for "welcome"

How to sign "welcome" in American Sign Language. And what to respond in ASL after one says "thank you".

Meaning: To greet a person, visitor, or guest in a warm and friendly manner.

Pronunciation (sign description): Dominant flat hand with palm up held in space slides toward the signer.

Learner tip: Don't confuse this similar sign with INVITE and HIRE (variation).

Fig. 2.9: Extract from the article on the lemma *welcome* in the ASL-DICT.

2.4 Dictionary portals

2.4.1 Criteria for a description of dictionary portals

The fundamental idea of the WWW immediately suggests devising lexicographic concepts that integrate a variety of different Internet dictionaries. *Dictionary portals* provide access to lexicographic information across dictionaries. A portal – as we learn in Webster's dictionary from 1828 (WEBSTER-1828) – is a "gate; an opening for entrance; as the portals of heaven". As we shall see on the following pages, the entrance to the heaven of lexical information provided by dictionary portals gains its particular quality by its access structures and the way of integration of dictionaries and their lexicographic information. Access structure and data integration are therefore also the basis for a typology of dictionary portals and related platforms. The typology presented in section 2.4.2 is a revision of Engelberg/Müller-Spitzer (2013), containing alterations and additions, due above all to the developments in the field of Internet lexicography in recent years.

A dictionary portal is an Internet site or a set of linked Internet sites that provides access to multiple Internet dictionaries or the information contained in them, where the dictionaries and the original information obtained from them remain reconstructible from the output of search queries to the portal. A range of criteria can be used to categorise dictionary portals. To single out the most important ones here: (i) the integrity of the dictionary, (ii) the access structures, (iii) the way of integration of dictionaries, and (iv) the digital layout of the portal.

The above definition of a dictionary portal involves the dictionaries integrated into the portal also having a separate existence, independent of the portal. This criterion needs to be understood in a more graded way than it initially appears. The dictionaries accessible through the OneLook portal fulfil this criterion fully. They are conceived and compiled separately from the portal and their digital form is not determined by the operators of the dictionary portal.[1]

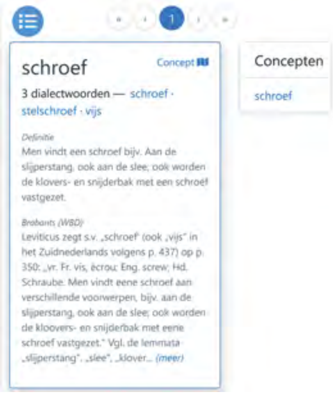
By contrast, the individual dictionaries in portals such as the Slovenian portal FRAN (Perdih/Ježovnik 2016) or the German portals Wörterbuchnetz (→ Fig. 2.18; Hildenbrandt/Moulin 2012) or OWID (→ Fig. 2.19; Engelberg et al. 2020) are the result of lexicographic projects independent of the portals, but their digital form is in essence the result of the portal creators' work. Other lexical Internet platforms are based on independent dictionaries, but the dictionaries and their original information are not prominent on the portals' interface. The 'Database of the Southern Dutch Dialects' (DSDD) extracts information from dialect dictionaries, enriches it with additional information, and presents the information in a way in which the original lexicographic information is only given in the form of quotes (→ Fig. 2.10). Thus, as its name suggests, the platform is more like a database than a dictionary portal. Lexical platforms of this sort will be treated in section 2.4.3.

The blending of portal and dictionaries was even more pronounced in the 'Base lexicale du français' (BLF, → Fig. 2.11). From the perspective of user functionality, the portal proceeds from the assumption that the reception of foreign language texts requires different information than that required for translation. Thus, the user can enter the relevant function, and the portal will then generate from its underlying resources different dictionaries or dictionary entries tailored to the specific function (cf. Verlinde 2011).[2]

Bilingual portals such as LEO and Linguee present icons for language pairs on the user interface without making it transparent whether the entities behind these icons are just configurations generated from a single database or stand-alone Internet dictionaries (→ Fig. 2.12).

[1] On different types of operators of dictionary portals cf. Boelhouwer/Dijkstra/Sijens (2018: 756f.).

[2] As a result of detailed user studies, the BLF has since been replaced by the 'Interactive Language Toolbox' (ILT). This foregoes the option where the user selects the function of the dictionary (cf. Verlinde/Peeters 2012).



schroef

Concept

3 dialectwoorden — schroef · stelschroef · vijs

Definitie

Men vindt een schroef bijv. Aan de slijperstang, ook aan de slee; ook worden de klovers- en snijderbak met een schroef vastgezet.

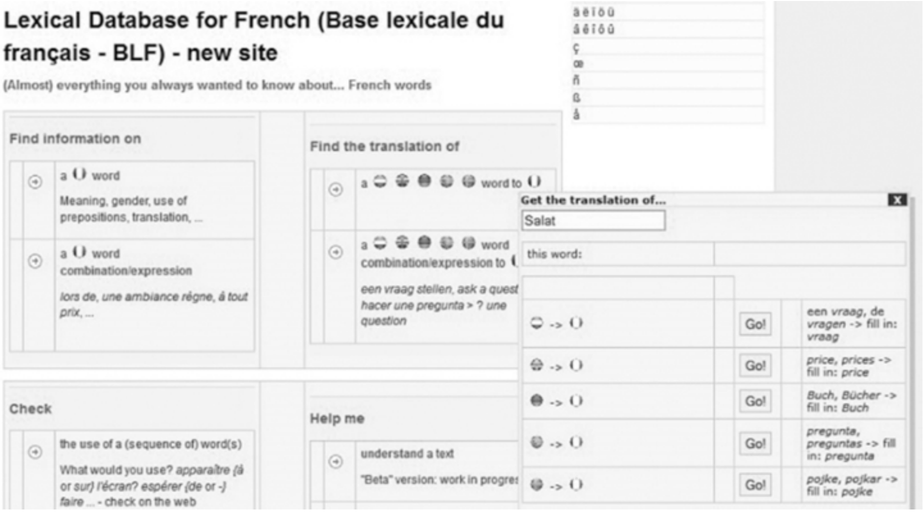
Brabants (WBD)

Leviticus zegt s.v. „schroef" (ook „vijs" in het Zuidnederlands volgens p. 437) op p. 350: „vr. Fr. vis, écrou; Eng. screw; Hd. Schraube. Men vindt eene schroef aan verschillende voorwerpen, bijv. aan de slijperstang, ook aan de slee; ook worden de kloovers- en snijderbak met eene schroef vastgezet." Vgl. de lemmata „slijperstang", „slee", „klover... *(meer)*

Concepten

schroef

Fig. 2.10: Result from the search for Dutch *schrouf* ('screw') in the 'Database of the Southern Dutch Dialects' (DSDD).



Lexical Database for French (Base lexicale du français - BLF) - new site

(Almost) everything you always wanted to know about... French words

Find information on

a word

Meaning, gender, use of prepositions, translation, ...

a word combination/expression

lors de, une ambiance règne, à tout prix, ...

Check

the use of a (sequence of) word(s)

What would you use? *apparaître (à or sur) l'écran? espérer (de or -) faire ...* - check on the web

Find the translation of

a word to

a word combination/expression to

een vraag stellen, ask a quest hacer une pregunta > ? une question

Help me

understand a text

"Beta" version: work in progres

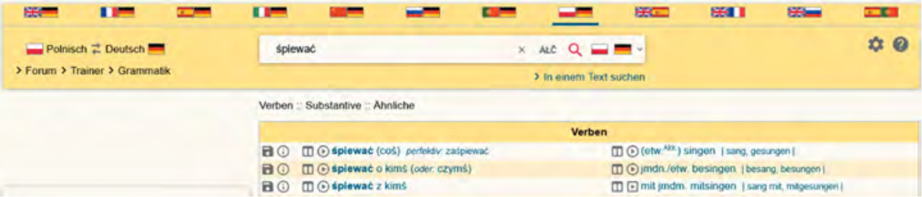
Get the translation of...

Salat

this word:

->	Go!	*een vraag, de vragen -> fill in: vraag*
->	Go!	*price, prices -> fill in: price*
->	Go!	*Buch, Bücher -> fill in: Buch*
->	Go!	*pregunta, preguntas -> fill in: pregunta*
->	Go!	*pojke, pojkar -> fill in: pojke*

Fig. 2.11: Extract from the BLF; translation is selected as the user function.



Polnisch ⇄ Deutsch

› Forum › Trainer › Grammatik

śpiewać

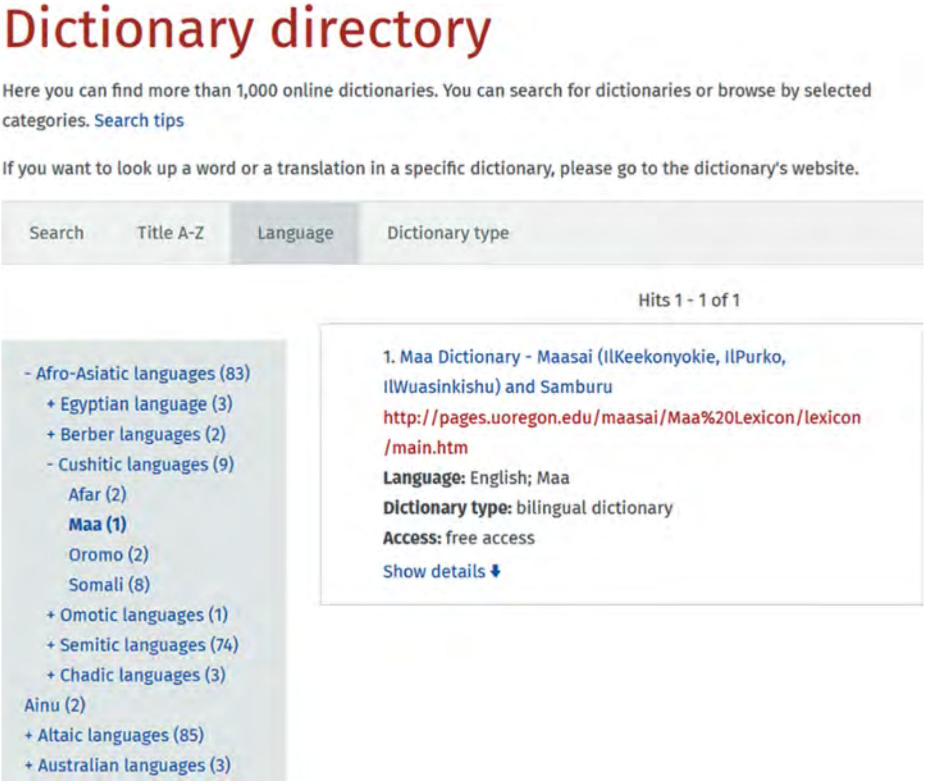
› In einem Text suchen

Verben :: Substantive :: Ähnliche

Verben	
śpiewać (coś) *perfektiv: zaśpiewać*	(etw.$^{Akk.}$) singen | sang, gesungen |
śpiewać o kimś (*oder:* czymś)	jmdn./etw. besingen | besang, besungen |
śpiewać z kimś	mit jmdm. mitsingen | sang mit, mitgesungen |

Fig. 2.12: Search for the German equivalent of Polish *śpiewać* 'to sing' in LEO.

Dictionary portals differ in terms of the types of access that they offer to their lexicographic data (cf. Boelhouwer/Dijkstra/Sijens 2018: 761f). Here, we distinguish between external access, outer access, and inner access. A portal with external access provides hyperlinks that allow access to the start page of the integrated dictionaries. If this is the only means of dictionary access, then the portal is often little more than a list of links. These kinds of portals emerged in the early 1990s right after the first dictionaries appeared on the net (cf. Storrer/Freese 1996: 106ff), and they can still be found, in particular if a systematic overview of a large number of dictionaries shall be given, sometimes provided with a multilayered access structure to the set of dictionaries. An example is the dictionary directory of the 'Lin|gu|is|tik portal' (DD-LINGUISTIKPORTAL) (→ Fig. 2.13) with descriptions of more than 1.000 linked online dictionaries, that can be searched alphabetically via dictionary title, dictionary type, a filter-based search, or a hierarchically structured representation of language families.



Dictionary directory

Here you can find more than 1,000 online dictionaries. You can search for dictionaries or browse by selected categories. Search tips

If you want to look up a word or a translation in a specific dictionary, please go to the dictionary's website.

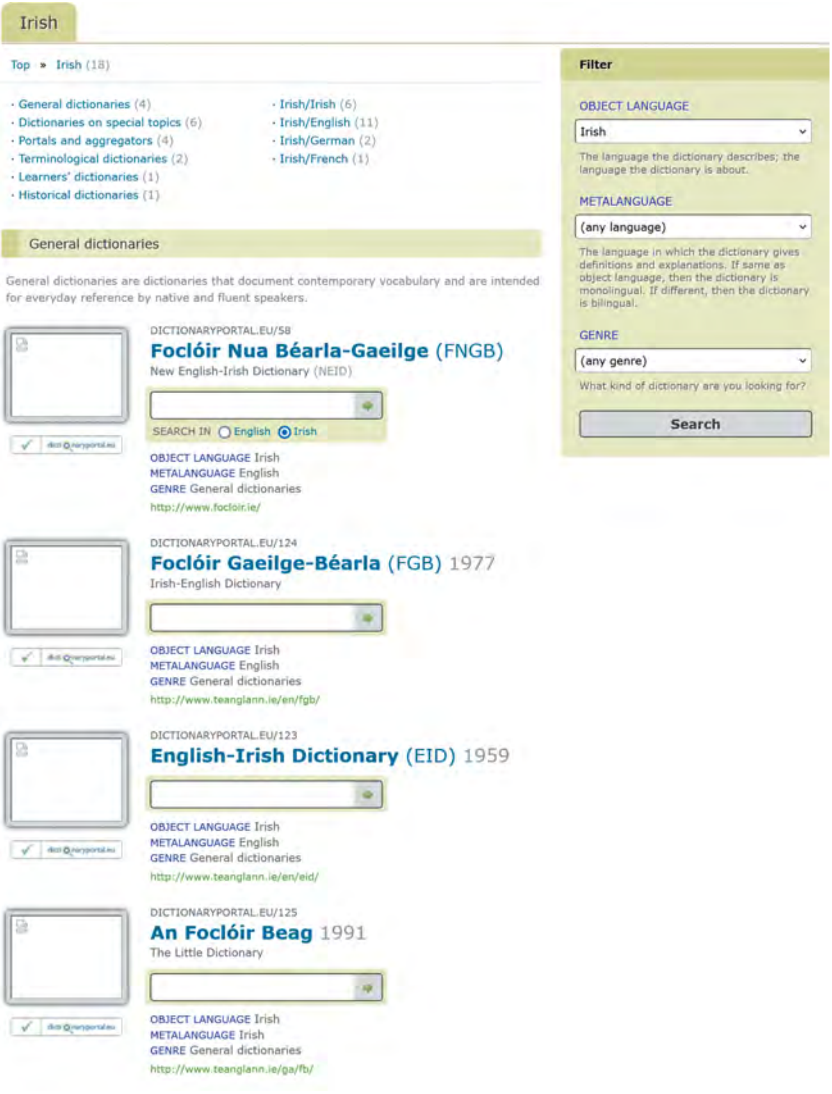
Hits 1 - 1 of 1

- Afro-Asiatic languages (83)
  - + Egyptian language (3)
  - + Berber languages (2)
  - - Cushitic languages (9)
    - Afar (2)
    - **Maa (1)**
    - Oromo (2)
    - Somali (8)
  - + Omotic languages (1)
  - + Semitic languages (74)
  - + Chadic languages (3)
- Ainu (2)
- + Altaic languages (85)
- + Australian languages (3)

1. Maa Dictionary - Maasai (IlKeekonyokie, IlPurko, IlWuasinkishu) and Samburu
http://pages.uoregon.edu/maasai/Maa%20Lexicon/lexicon/main.htm
Language: English; Maa
Dictionary type: bilingual dictionary
Access: free access
Show details ⬇

Fig. 2.13: Different forms of external access in the DD-LINGUISTIKPORTAL; here access by language families.

If the portal offers outer access, then it is possible to directly access the lemmas of the embedded dictionaries from the portal page. In case of a single outer search each listed dictionary is complemented by a lemma search field, as can be seen in the European Dictionary Portal EDP in → Fig. 2.14. If a multiple outer search is provided, entering a search term in the lemma search field brings up either a list of all the dictionary articles that correspond to the search term in the embedded dictionaries, as in FREEDICT or ETYMOLOGIEBANK (→ Fig. 2.15), or a list of all the lemmas whose associated articles can then be reached via a link, as in ONELOOK (→ Fig. 2.16).



Irish

Top » Irish (18)

- General dictionaries (4)
- Dictionaries on special topics (6)
- Portals and aggregators (4)
- Terminological dictionaries (2)
- Learners' dictionaries (1)
- Historical dictionaries (1)

- Irish/Irish (6)
- Irish/English (11)
- Irish/German (2)
- Irish/French (1)

General dictionaries

General dictionaries are dictionaries that document contemporary vocabulary and are intended for everyday reference by native and fluent speakers.

DICTIONARYPORTAL.EU/58
Foclóir Nua Béarla-Gaeilge (FNGB)
New English-Irish Dictionary (NEID)
SEARCH IN ○ English ● Irish
OBJECT LANGUAGE Irish
METALANGUAGE English
GENRE General dictionaries
http://www.focloir.ie/

DICTIONARYPORTAL.EU/124
Foclóir Gaeilge-Béarla (FGB) 1977
Irish-English Dictionary
OBJECT LANGUAGE Irish
METALANGUAGE English
GENRE General dictionaries
http://www.teanglann.ie/en/fgb/

DICTIONARYPORTAL.EU/123
English-Irish Dictionary (EID) 1959
OBJECT LANGUAGE Irish
METALANGUAGE English
GENRE General dictionaries
http://www.teanglann.ie/en/eid/

DICTIONARYPORTAL.EU/125
An Foclóir Beag 1991
The Little Dictionary
OBJECT LANGUAGE Irish
METALANGUAGE Irish
GENRE General dictionaries
http://www.teanglann.ie/ga/fb/

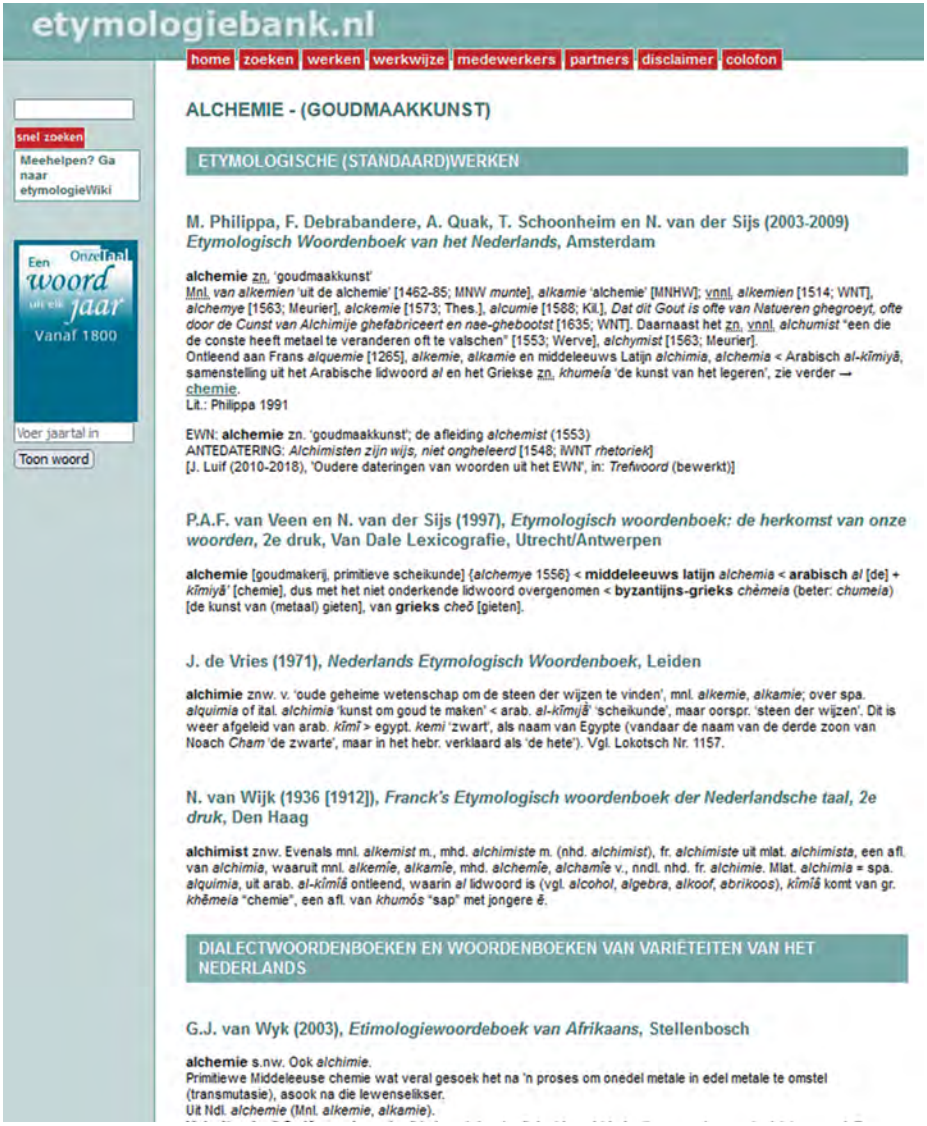
Filter

OBJECT LANGUAGE
Irish
The language the dictionary describes; the language the dictionary is about.

METALANGUAGE
(any language)
The language in which the dictionary gives definitions and explanations. If same as object language, then the dictionary is monolingual. If different, then the dictionary is bilingual.

GENRE
(any genre)
What kind of dictionary are you looking for?

Fig. 2.14: Single outer access for each Irish dictionary included in the 'European Dictionary Portal' (EDP).



ALCHEMIE - (GOUDMAAKKUNST)

ETYMOLOGISCHE (STANDAARD)WERKEN

M. Philippa, F. Debrabandere, A. Quak, T. Schoonheim en N. van der Sijs (2003-2009) *Etymologisch Woordenboek van het Nederlands*, Amsterdam

alchemie zn. 'goudmaakkunst'
Mnl. *van alkemien* 'uit de alchemie' [1462-85; MNW *munte*], *alkamie* 'alchemie' [MNHW]; vnnl. *alkemien* [1514; WNT], *alchemye* [1563; Meurier], *alckemie* [1573; Thes.], *alcumie* [1588; Kil.], *Dat dit Gout is ofte van Natueren ghegroeyt, ofte door de Cunst van Alchimije ghefabriceert en nae-ghebootst* [1635; WNT]. Daarnaast het zn. vnnl. *alchumist* "een die de conste heeft metael te veranderen oft te valschen" [1553; Werve], *alchymist* [1563; Meurier].
Ontleend aan Frans *alquemie* [1265], *alkemie*, *alkamie* en middeleeuws Latijn *alchimia*, *alchemia* < Arabisch *al-kīmiyā*, samenstelling uit het Arabische lidwoord *al* en het Griekse zn. *khumeía* 'de kunst van het legeren', zie verder → chemie.
Lit.: Philippa 1991

EWN: **alchemie** zn. 'goudmaakkunst'; de afleiding *alchemist* (1553)
ANTEDATERING: *Alchimisten zijn wijs, niet ongheleerd* [1548; iWNT *rhetoriek*]
[J. Luif (2010-2018), 'Oudere dateringen van woorden uit het EWN', in: *Trefwoord* (bewerkt)]

P.A.F. van Veen en N. van der Sijs (1997), *Etymologisch woordenboek: de herkomst van onze woorden*, 2e druk, Van Dale Lexicografie, Utrecht/Antwerpen

alchemie [goudmakerij, primitieve scheikunde] {*alchemye* 1556} < **middeleeuws latijn** *alchemia* < **arabisch** *al* [de] + *kīmiyā'* [chemie], dus met het niet onderkende lidwoord overgenomen < **byzantijns-grieks** *chèmeia* (beter: *chumeia*) [de kunst van (metaal) gieten], van **grieks** *cheō* [gieten].

J. de Vries (1971), *Nederlands Etymologisch Woordenboek*, Leiden

alchimie znw. v. 'oude geheime wetenschap om de steen der wijzen te vinden', mnl. *alkemie*, *alkamie*; over spa. *alquimia* of ital. *alchimia* 'kunst om goud te maken' < arab. *al-kīmijā̀* 'scheikunde', maar oorspr. 'steen der wijzen'. Dit is weer afgeleid van arab. *kīmī* > egypt. *kemi* 'zwart', als naam van Egypte (vandaar de naam van de derde zoon van Noach *Cham* 'de zwarte', maar in het hebr. verklaard als 'de hete'). Vgl. Lokotsch Nr. 1157.

N. van Wijk (1936 [1912]), *Franck's Etymologisch woordenboek der Nederlandsche taal, 2e druk*, Den Haag

alchimist znw. Evenals mnl. *alkemist* m., mhd. *alchimiste* m. (nhd. *alchimist*), fr. *alchimiste* uit mlat. *alchimista*, een afl. van *alchimia*, waaruit mnl. *alkemîe*, *alkamîe*, mhd. *alchemîe*, *alchamîe* v., nndl. nhd. fr. *alchimie*. Mlat. *alchimia* = spa. *alquimia*, uit arab. *al-kîmîâ* ontleend, waarin *al* lidwoord is (vgl. *alcohol*, *algebra*, *alkoof*, *abrikoos*), *kîmîâ* komt van gr. *khēmeia* "chemie", een afl. van *khumós* "sap" met jongere *ē*.

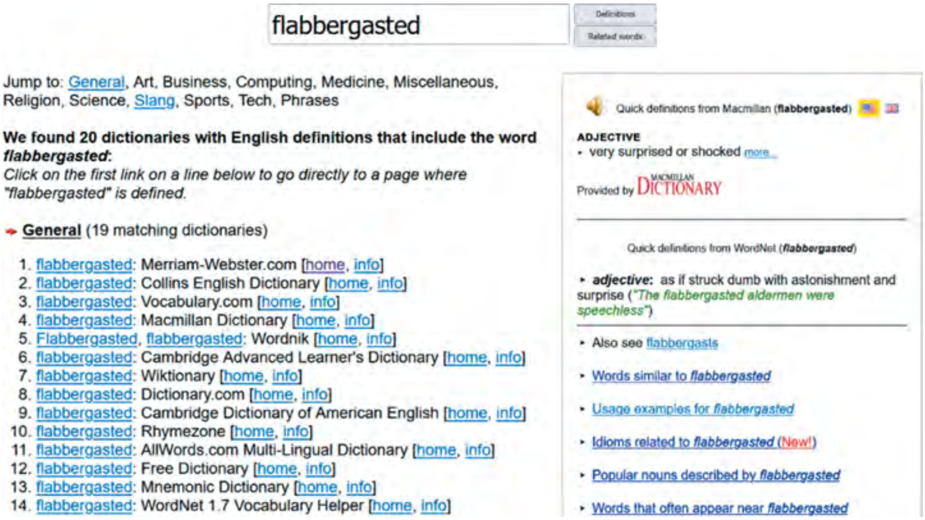
DIALECTWOORDENBOEKEN EN WOORDENBOEKEN VAN VARIËTEITEN VAN HET NEDERLANDS

G.J. van Wyk (2003), *Etimologiewoordeboek van Afrikaans*, Stellenbosch

alchemie s.nw. Ook *alchimie*.
Primitiewe Middeleeuse chemie wat veral gesoek het na 'n proses om onedel metale in edel metale te omstel (transmutasie), asook na die lewenselikser.
Uit Ndl. *alchemie* (Mnl. *alkemie*, *alkamie*).

Fig. 2.15: Multiple outer access to the dictionaries in the Dutch Etymologiebank, a dictionary portal for dictionaries of etymological interest (etymological dictionaries, dialect dictionaries, loanword dictionaries, etc.); the search for *alchemie* produces a list of corresponding articles sorted according to the types of dictionaries they were found in.

Less frequently, portals offer the option of inner access (cf. Müller-Spitzer 2010 on inner access in OWID). Inner access makes it possible to directly access particular information in the embedded dictionaries from a search function on the portal page. Inner access is available, for example, in the Dutch/Frisian portal for historical dictio-



flabbergasted

Jump to: General, Art, Business, Computing, Medicine, Miscellaneous, Religion, Science, Slang, Sports, Tech, Phrases

We found 20 dictionaries with English definitions that include the word *flabbergasted*:
Click on the first link on a line below to go directly to a page where "flabbergasted" is defined.

General (19 matching dictionaries)

1. flabbergasted: Merriam-Webster.com [home, info]
2. flabbergasted: Collins English Dictionary [home, info]
3. flabbergasted: Vocabulary.com [home, info]
4. flabbergasted: Macmillan Dictionary [home, info]
5. Flabbergasted, flabbergasted: Wordnik [home, info]
6. flabbergasted: Cambridge Advanced Learner's Dictionary [home, info]
7. flabbergasted: Wiktionary [home, info]
8. flabbergasted: Dictionary.com [home, info]
9. flabbergasted: Cambridge Dictionary of American English [home, info]
10. flabbergasted: Rhymezone [home, info]
11. flabbergasted: AllWords.com Multi-Lingual Dictionary [home, info]
12. flabbergasted: Free Dictionary [home, info]
13. flabbergasted: Mnemonic Dictionary [home, info]
14. flabbergasted: WordNet 1.7 Vocabulary Helper [home, info]

Quick definitions from Macmillan (**flabbergasted**)

ADJECTIVE
- very surprised or shocked more...

Provided by MACMILLAN DICTIONARY

Quick definitions from WordNet (***flabbergasted***)

- ***adjective:*** as if struck dumb with astonishment and surprise (*"The flabbergasted aldermen were speechless"*)

- Also see flabbergasts
- Words similar to *flabbergasted*
- Usage examples for *flabbergasted*
- Idioms related to *flabbergasted* (New!)
- Popular nouns described by *flabbergasted*
- Words that often appear near *flabbergasted*

Fig. 2.16: Multiple outer access to the dictionaries in OneLook: on the left, a list of keywords from the integrated dictionaries related to the search term *flabbergasted*; on the right, an extract from the article on the keyword in one of the dictionaries.

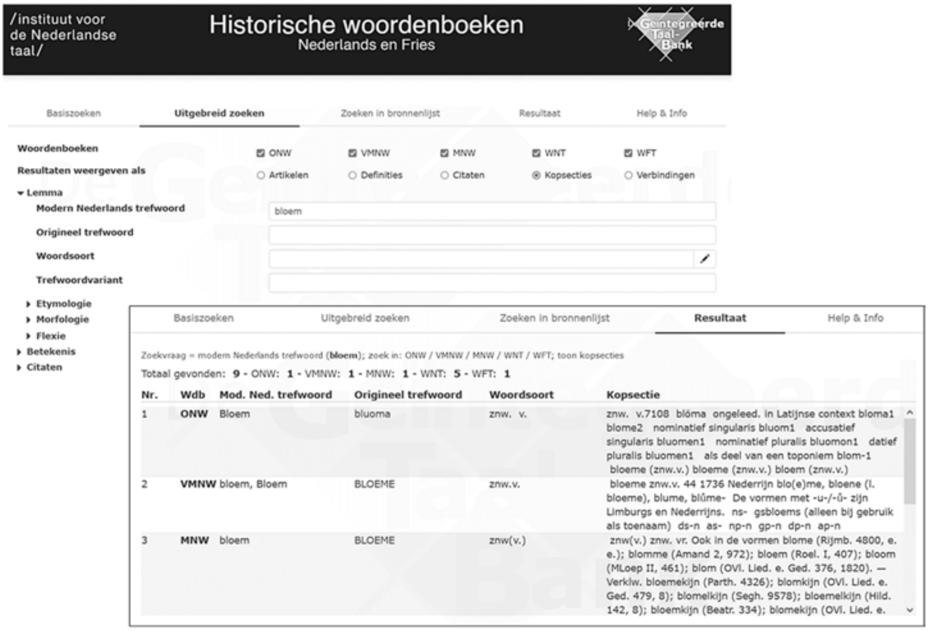
naries HWNF (→ Fig. 2.17). The inner access structure in the HWNF operates over all integrated dictionaries. The recent version of the German portal OWID and the portal AISRI for Northern Caddoan languages allow inner access to meaning paraphrases in single dictionaries.

Whether navigation across dictionaries is also possible beyond the search functions provided on the portal page depends on how strongly the portal dictionaries are integrated with one another. Integration presupposes that the operators of the portal also have access to the digital structure of the embedded dictionaries. In the Wörterbuchnetz, dictionary entries are linked inside the portal to corresponding entries from dictionaries on other German varieties included in the portal (→ Fig. 2.18).

A single layout for the portal and its dictionaries promotes the visual integration of the portal's content and facilitates the user's orientation. If the operator of the portal has access to the digital structure of the dictionaries, a consistent design concept (→ Chapter 6) for the layout of the portal and its dictionaries can be created, as in OWID (→ Fig. 2.19), Fran, Linguee, Etymologiebank (→ Fig. 2.15), or Wörterbuchnetz (→ Fig. 2.18).

2.4.2 Typology of dictionary portals

The above criteria make it possible to construct a typology of dictionary portals in relation to two gradable dimensions: the degree of autonomy of the dictionaries con-



Woordenboeken ☑ ONW ☑ VMNW ☑ MNW ☑ WNT ☑ WFT

Resultaten weergeven als ○ Artikelen ○ Definities ○ Citaten ● Kopsecties ○ Verbindingen

▾ Lemma

Modern Nederlands trefwoord: bloem

Origineel trefwoord:

Woordsoort:

Trefwoordvariant:

▸ Etymologie
▸ Morfologie
▸ Flexie
▸ Betekenis
▸ Citaten

Zoekvraag = modem Nederlands trefwoord (**bloem**); zoek in: ONW / VMNW / MNW / WNT / WFT; toon kopsecties

Totaal gevonden: **9** - ONW: **1** - VMNW: **1** - MNW: **1** - WNT: **5** - WFT: **1**

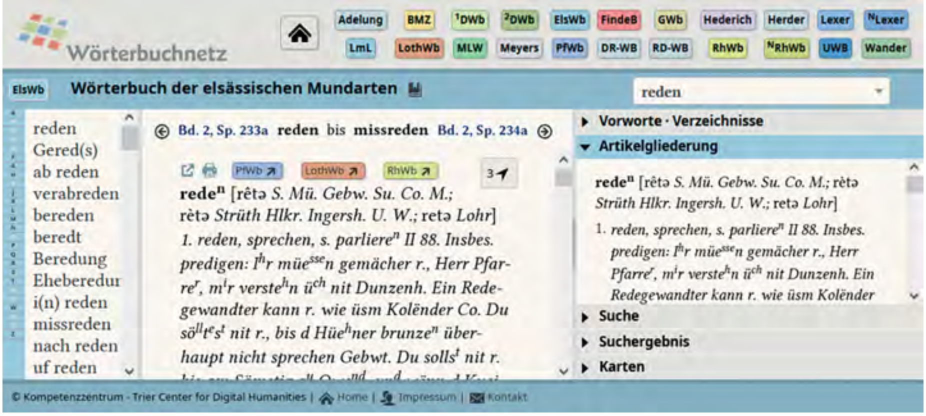
Nr.	Wdb	Mod. Ned. trefwoord	Origineel trefwoord	Woordsoort	Kopsectie
1	**ONW**	Bloem	bluoma	znw. v.	znw. v.7108 blōma ongeleed. in Latijnse context bloma1 blome2 nominatief singularis bluom1 accusatief singularis bluomen1 nominatief pluralis bluomon1 datief pluralis bluomen1 als deel van een toponiem blom-1 bloeme (znw.v.) bloeme (znw.v.) bloem (znw.v.)
2	**VMNW**	bloem, Bloem	BLOEME	znw.v.	bloeme znw.v. 44 1736 Nederrijn blo(e)me, bloene (l. bloeme), blume, blûme- De vormen met -u-/-û- zijn Limburgs en Nederrijns. ns- gsbloems (alleen bij gebruik als toenaam) ds-n as- np-n gp-n dp-n ap-n
3	**MNW**	bloem	BLOEME	znw(v.)	znw(v.) znw. vr. Ook in de vormen blome (Rijmb. 4800, e. e.); blomme (Amand 2, 972); bloem (Roel. I, 407); bloom (MLoep II, 461); blom (OVl. Lied. e. Ged. 376, 1820). — Verklw. bloemekijn (Parth. 4326); blomkijn (OVl. Lied. e. Ged. 479, 8); blomelkijn (Segh. 9578); bloemelkijn (Hild. 142, 8); bloemkijn (Beatr. 334); blomekijn (OVl. Lied. e.

Fig. 2.17: Inner access in the portal for historical dictionaries of Dutch and Frisian (HWNF). Selecting etymological information (under "kopsectie") for the search term *bloem* 'flower' in the extended search ("uitgebreid zoek") will address the etymological information in the relevant articles and present – apart from core information (modern Dutch keyword, "Mod. Ned. trefwoord", original keyword, "Origineel trefwoord"; part of speech, "Woordsoort") – only etymological information from all integrated dictionaries, skipping other information items like meaning or corpus examples.

tained in the portal; and the degree of integration of the dictionaries (access structure, cross-dictionary referencing, layout) (→ Fig. 2.20).

Dictionary collections maintain complete autonomy of their dictionaries and make no attempts to integrate dictionaries. They provide only external access to separate dictionaries via hyperlinks. There is no integration by cross-dictionary references and no uniform layout. The DD-LINGUISTIKPORTAL is an example for a well-designed dictionary collection (→ Fig. 2.13); the SLANG-PORTAL provides access to slang dictionaries in many languages. The database OBELEXDICT, which offers annotated links to more than 10,000 Internet dictionaries via multi-dimensional search queries can also be considered a dictionary collection (Möhrs/Töpel 2011).

Dictionary search engines are also characterized by a high autonomy of the included dictionaries; however, they provide outer access to the lemmas of their dictionaries, without further integrating these dictionaries; attempts to unify the layout are at best moderate. Examples are ONELOOK with multiple outer access (→ Fig. 2.16), ORDNET with single outer access, and the European Dictionary Portal (EDP) (→ Fig. 2.14) with external and outer access.



rede^{n} [rêtə *S. Mü. Gebw. Su. Co. M.*; rètə *Strüth Hlkr. Ingersh. U. W.*; retə *Lohr*] *1. reden, sprechen, s. parliere^{n} II 88. Insbes. predigen: I^{h}r müe^{sse}n gemächer r., Herr Pfarre^{r}, m^{i}r verste^{h}n ü^{ch} nit Dunzenh. Ein Redegewandter kann r. wie üsm Kolënder Co. Du sö^{ll}t^{e}s^{t} nit r., bis d Hüe^{h}ner brunze^{n} überhaupt nicht sprechen Gebwt. Du solls^{t} nit r.*

rede^{n} [rêtə *S. Mü. Gebw. Su. Co. M.*; rètə *Strüth Hlkr. Ingersh. U. W.*; retə *Lohr*]
1. *reden, sprechen, s. parliere^{n} II 88. Insbes. predigen: I^{h}r müe^{sse}n gemächer r., Herr Pfarre^{r}, m^{i}r verste^{h}n ü^{ch} nit Dunzenh. Ein Redegewandter kann r. wie üsm Kolënder*

Fig. 2.18: Dictionary integration in the WÖRTERBUCHNETZ; hyperlinks above the entry for *reden* in the Alsatian dictionary (ELS-DICT) to the corresponding entries in the Palatian (PF-DICT), Rhenisch (RH-DICT), and Lorraine dictonary (LOTH-DICT).

Dictionary networks offer a high degree of networked integration, sophisticated access structures, and layout uniformity, and do not interfere with the integrity of the dictionaries, or do so only moderately. Examples of dictionary networks are AISRI (Northern Caddoan, multilingual), FRAN (Slovenian), HWNF (Dutch, Frisian; → Fig. 2.17), LEHNWORTPORTAL-DEUTSCH (German, multilingual), OWID (German; → Fig. 2.19), WÖRTERBUCHNETZ (German; → Fig. 2.18), and WORDREFERENCE (English, multlingual). The LEHNWORTPORTAL-DEUTSCH (→ Fig. 2.21) embeds dictionaries of German loanwords in other languages (Meyer/Engelberg 2011). The integration arises here in that a "reciprocal loan dictionary" of German words borrowed in other languages is produced out of the German etyma in the dictionary articles (Meyer 2022).

In some ways, dictionary networks represent the prototype for dictionary portals. They seek to combine a maximum of integration with a maximum of autonomy for the dictionaries, to the extent that is possible given criteria that conflict with one another on certain points.[3]

Dictionary simulations are integrated to a high degree and standardised in terms of layout and access structures, while the autonomy of the integrated dictionaries is limited. Examples are bilingual portals such as LEO (→ Fig. 2.12) or LINGUEE.

We designate the grouping of portals presented here as a typology rather than a classification, because naturally a variety of transitional forms can also be found. For example, EDP presents itself in some areas as a search engine with outer access, but other dictionaries are only connected via external access. CAMBRIDGE-ONLINE and PONS-

[3] Cf. also Lew (2011) on types of dictionary portals and Boelhouwer/Dijkstra/Sijens (2018) for a survey and investigation of dictionary collections, search engines, and nets for different languages.

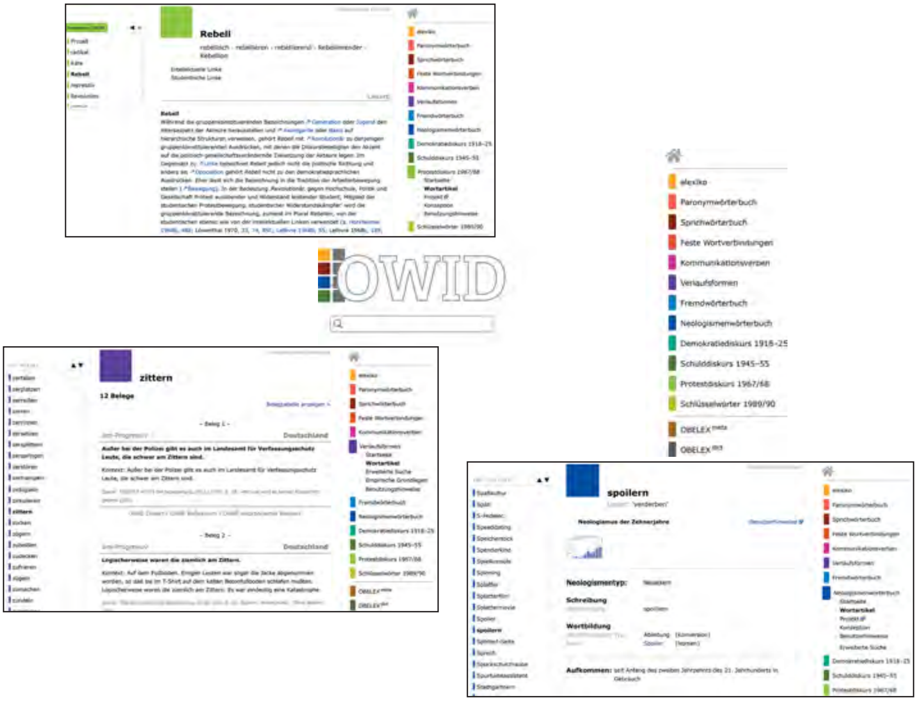
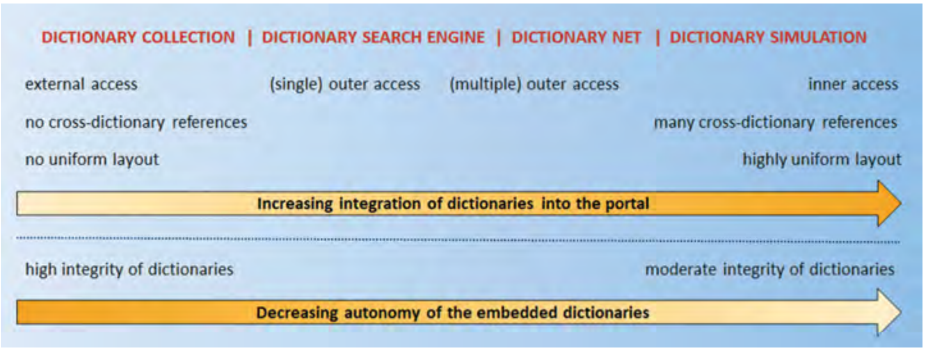


Fig. 2.19: Uniform layout of dictionaries in the German portal OWID; lemmas *Rebell* 'rebel' in a discourse dictionary (PROTESTDISKURS-DICT), *zittern* 'tremble' in a dictionary of progressive forms (VERLAUFSFORMEN-DICT) (ii), *spoilern* 'to spoil' in a dictionary of neologisms (NEO-DICT).


DICTIONARY COLLECTION | DICTIONARY SEARCH ENGINE | DICTIONARY NET | DICTIONARY SIMULATION

external access
(single) outer access
(multiple) outer access
inner access

no cross-dictionary references
many cross-dictionary references

no uniform layout
highly uniform layout

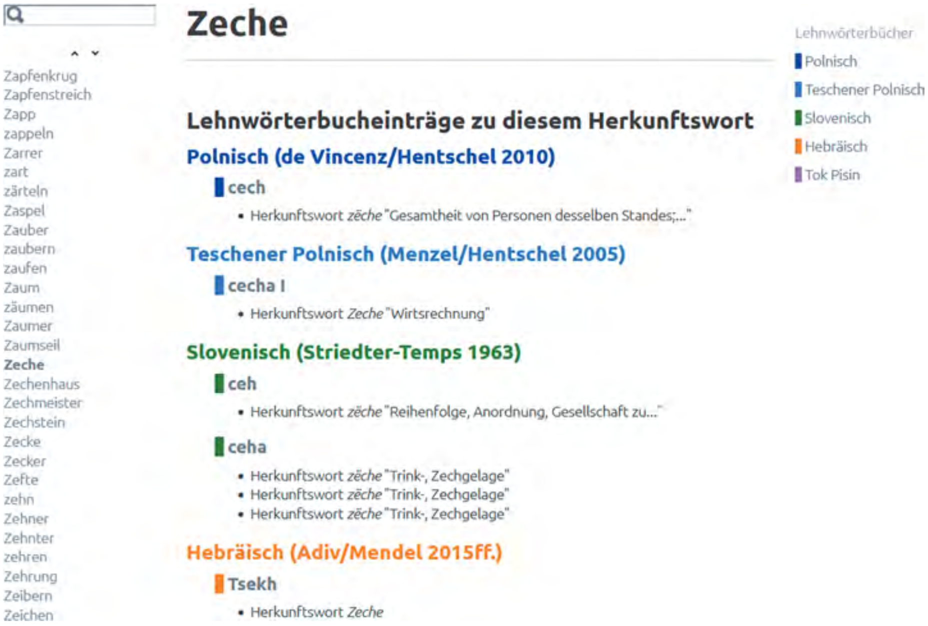
Increasing integration of dictionaries into the portal

high integrity of dictionaries
moderate integrity of dictionaries

Decreasing autonomy of the embedded dictionaries

Fig. 2.20: Typology of dictionary portals.

ONLINE appear to involve no integration of embedded dictionaries – as in dictionary search engines – but present themselves in their layout and access in a manner similar to what we would expect of dictionary nets.



Zapfenkrug
Zapfenstreich
Zapp
zappeln
Zarrer
zart
zärteln
Zaspel
Zauber
zaubern
zaufen
Zaum
zäumen
Zaumer
Zaumseil
Zeche
Zechenhaus
Zechmeister
Zechstein
Zecke
Zecker
Zefte
zehn
Zehner
Zehnter
zehren
Zehrung
Zeibern
Zeichen

Zeche

Lehnwörterbucheinträge zu diesem Herkunftswort

Polnisch (de Vincenz/Hentschel 2010)

cech

- Herkunftswort *zěche* "Gesamtheit von Personen desselben Standes;..."

Teschener Polnisch (Menzel/Hentschel 2005)

cecha I

- Herkunftswort *Zeche* "Wirtsrechnung"

Slovenisch (Striedter-Temps 1963)

ceh

- Herkunftswort *zëche* "Reihenfolge, Anordnung, Gesellschaft zu..."

ceha

- Herkunftswort *zëche* "Trink-, Zechgelage"
- Herkunftswort *zëche* "Trink-, Zechgelage"
- Herkunftswort *zëche* "Trink-, Zechgelage"

Hebräisch (Adiv/Mendel 2015ff.)

Tsekh

- Herkunftswort *Zeche*

Lehnwörterbücher
Polnisch
Teschener Polnisch
Slovenisch
Hebräisch
Tok Pisin

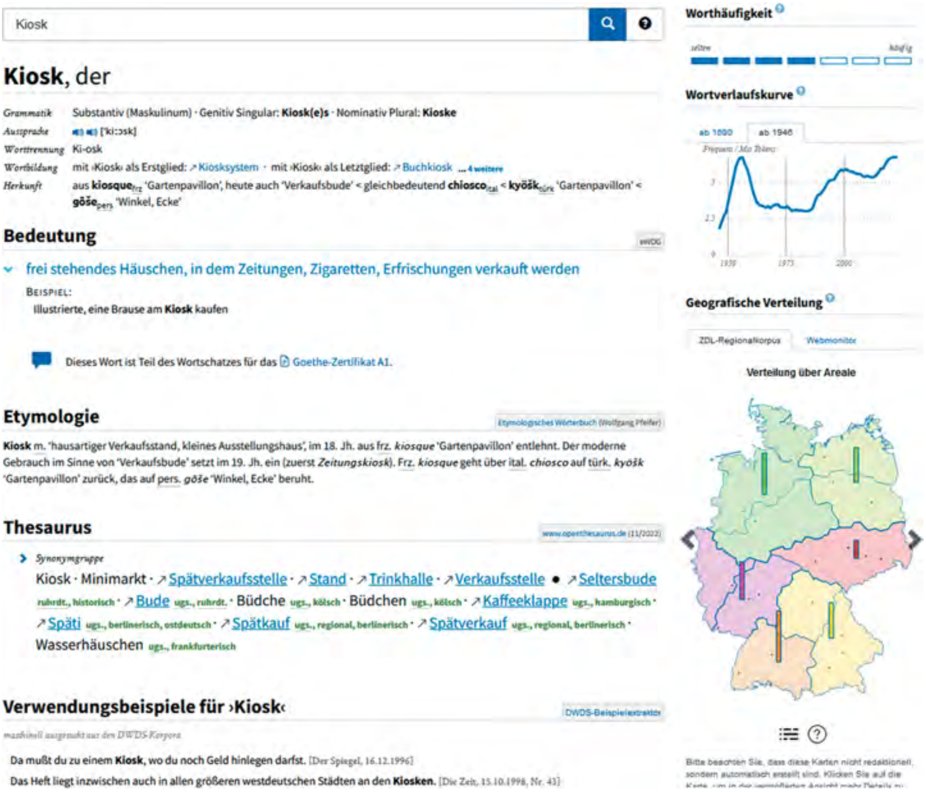
Fig. 2.21: Result of a search on the search term *Zeche*, ‘mine’, in the portal-generated reciprocal German loanword dictionary in the LEHNWORTPORTAL-DEUTSCH; the corresponding lemmas and the beginnings of articles are displayed in the integrated loanword dictionary.

Viewed historically, dictionary collections and search engines stand at the beginning of portal lexicography, while dictionary networks and dictionary simulations are developments of the last two decades. Interlinking of dictionaries and lexicographic data are still the way to go. Recent surveys conducted by the project ‘European Lexicographic Infrastructure’ (ELEXIS) among affiliated institutions have revealed an increasing need for “increased interoperability, linking and sharing of resources” and for the aggregation of “stand-alone lexicographic (and also terminological) resources into dictionary portals” (Tiberius et al. 2022: 518).

2.4.3 Dictionaries on other language-related platforms

Dictionaries can not only form the lexicographic basis for dictionary portals, but also be part of Internet platforms that are more general in nature and in which dictionaries are combined with other digital resources.

Lexical portals provide information about words, their meaning, grammar, and use, as well as about cultural aspects and vocabulary learning. In particular, corpus-based platforms tend to complement lexicographic information from dictionaries with all kind of statistical information: automatically obtained co-occurrences pre-



Kiosk

Kiosk, der

Grammatik Substantiv (Maskulinum) · Genitiv Singular: **Kiosk(e)s** · Nominativ Plural: **Kioske**
Aussprache ['kiːɔsk]
Worttrennung Ki-osk
Wortbildung mit ›Kiosk‹ als Erstglied: ↗ Kiosksystem · mit ›Kiosk‹ als Letztglied: ↗ Buchkiosk ... 4 weitere
Herkunft aus **kiosque**frz 'Gartenpavillon', heute auch 'Verkaufsbude' < gleichbedeutend **chiosco**ital < **kyöšk**türk 'Gartenpavillon' < **gōše**pers 'Winkel, Ecke'

Bedeutung eWDG

frei stehendes Häuschen, in dem Zeitungen, Zigaretten, Erfrischungen verkauft werden

BEISPIEL:
Illustrierte, eine Brause am **Kiosk** kaufen

Dieses Wort ist Teil des Wortschatzes für das Goethe-Zertifikat A1.

Etymologie Etymologisches Wörterbuch (Wolfgang Pfeifer)

Kiosk m. 'hausartiger Verkaufsstand, kleines Ausstellungshaus', im 18. Jh. aus frz. ***kiosque*** 'Gartenpavillon' entlehnt. Der moderne Gebrauch im Sinne von 'Verkaufsbude' setzt im 19. Jh. ein (zuerst ***Zeitungskiosk***). Frz. ***kiosque*** geht über ital. ***chiosco*** auf türk. ***kyöšk*** 'Gartenpavillon' zurück, das auf pers. ***gōše*** 'Winkel, Ecke' beruht.

Thesaurus www.openthesaurus.de (11/2022)

Synonymgruppe
Kiosk · Minimarkt · ↗ Spätverkaufsstelle · ↗ Stand · ↗ Trinkhalle · ↗ Verkaufsstelle ● ↗ Seltersbude ruhrdt., historisch · ↗ Bude ugs., ruhrdt. · Büdche ugs., kölsch · Büdchen ugs., kölsch · ↗ Kaffeeklappe ugs., hamburgisch · ↗ Späti ugs., berlinerisch, ostdeutsch · ↗ Spätkauf ugs., regional, berlinerisch · ↗ Spätverkauf ugs., regional, berlinerisch · Wasserhäuschen ugs., frankfurterisch

Verwendungsbeispiele für ›Kiosk‹ DWDS-Beispielextraktor

maschinell ausgesucht aus den DWDS-Korpora

Da mußt du zu einem **Kiosk**, wo du noch Geld hinlegen darfst. [Der Spiegel, 16.12.1996]
Das Heft liegt inzwischen auch in allen größeren westdeutschen Städten an den **Kiosken**. [Die Zeit, 15.10.1998, Nr. 43]

Worthäufigkeit

Wortverlaufskurve

Geografische Verteilung

Verteilung über Areale

Bitte beachten Sie, dass diese Karten nicht redaktionell, sondern automatisch erstellt sind. Klicken Sie auf die Karte, um in der vergrößerten Ansicht mehr Details zu

Fig. 2.22: Information on the German word *Kiosk* on the lexical platform DWDS, containing information from three dictionaries on meaning (“Bedeutung”, E-WDG), etymology (“Etymologie”, PFEIFER-DWDS), and synonyms (“Thesaurus”, OPENTHESAURUS-DE), statistical information on word frequency (“Worthäufigkeit”), word frequency over time (“Wortverlaufskurve”), as well as information about regional distribution and automatically extracted corpus examples (“Verwendungsbeispiele für ›Kiosk‹”).

sented as lists, graphs, or word clouds, data on frequency of use, charts documenting the time course of usage, etc. (→ Chapter 7). A prominent example is the DWDS portal (→ Fig. 2.22).

In addition to statistical and other corpus-based information about words, lexical portals may also include thematic word lists, word-related quizzes and games, stories about words presented as blogs, videos or “word-of-the-day” sections, and resources to support vocabulary learning and teaching (cf. also Boelhouwer/Dijkstra/Sijens 2018: 758f). In particular, commercial platforms like MERRIAM-WEBSTER (→ Fig. 2.23) are characterized by these features.

Language platforms emerge when providers move to expand their information offerings beyond the presentation of lexical knowledge to include linguistic knowledge in general. This includes sketch grammars, information on grammatical pat-

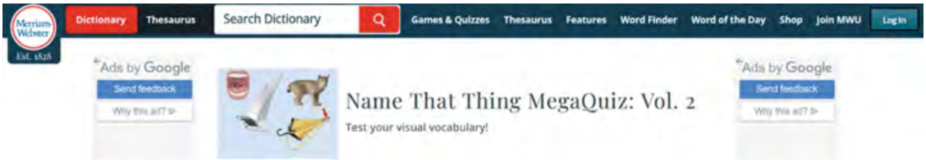


Fig. 2.23: Naming quiz on the Merriam-Webster lexical platform.

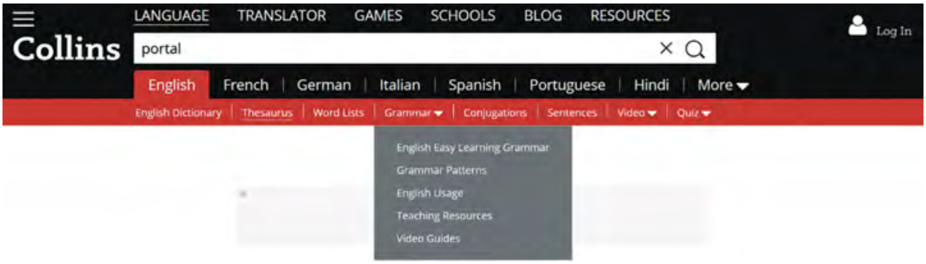
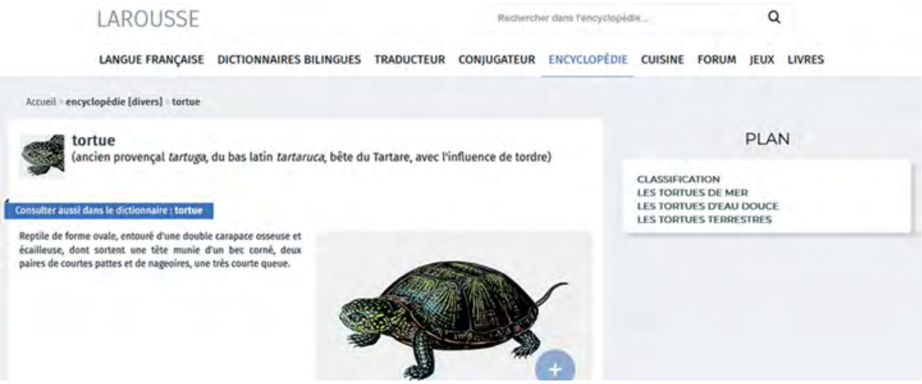


Fig. 2.24: Information about grammar on the Collins language platform.

terns, or inflection tables. This again can be observed in commercial platforms of publishing houses such as Collins (→ Fig. 2.24).

Reference platforms extend the content of language-related platforms by further including encyclopedias or other resources containing non-linguistic information. The French reference portal Larousse not only provides dictionaries, inflection tables, and an automatic translation device but also access to an encyclopedia (→ Fig. 2.25) and a list of names of dishes interlinked to the corresponding recipes ("cuisine").



tortue
(ancien provençal *tartuga*, du bas latin *tartaruca*, bête du Tartare, avec l'influence de tordre)

Reptile de forme ovale, entouré d'une double carapace osseuse et écailleuse, dont sortent une tête munie d'un bec corné, deux paires de courtes pattes et de nageoires, une très courte queue.

Fig. 2.25: Enyclopedic article for French *tortue* 'turtle' on the Larousse reference platform.

The chart in → Fig. 2.26 summarises the typology of platforms presented in the preceding sections.

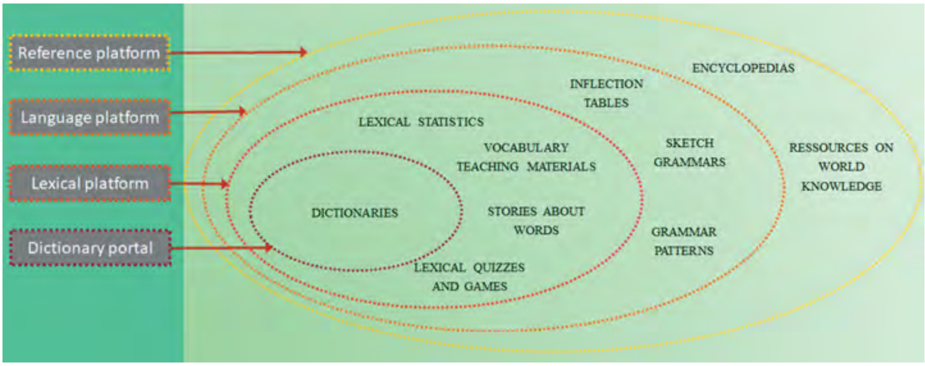


Fig. 2.26: Content of language-related platforms.

Pure encyclopedic portals can of course also be found on the Internet. They relate to encyclopedic reference works as dictionary portals relate to language dictionaries. Examples are ENCYCLOPEDIA with outer access and the list of encyclopedias in 'Wikipedia' (ENZYKLOPÄDIENLISTE-WIKIPEDIA) with external access to the integrated encyclopedias. Encyclopedic portals can be typologised in a way similar do dictionary portals but are not the subject matter of this chapter. In the domain of reference portals, mixed forms of dictionary and encyclopedic portals can also be found, such as the WÖRTERBUCHNETZ.

Bibliography

Further reading

Boelhouwer, Bob/Dykstra, Anna/Sijens, Hindrik (2018): Dictionary portals. In: Fuertes-Olivera, Pedro A. (ed.): *The Routledge Handbook of Lexicography*. London/New York: Routledge, Taylor & Francis, 754–765. *Based on a survey, the article presents an inventory of dictionary portals and what they offer to the user.*

de Schryver, Gilles-Maurice (2003): Lexicographer's Dreams in the Electronic Dictionary Age. In: *International Journal of Lexicography* 16, 143–199. *Early reflections on the impact of computers on dictionary making with a section on typologies for "electronic dictionaries".*

Hausmann, Franz Josef (1989): Wörterbuchtypologie. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst/Zgusta, Ladislav(eds.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie. 1. Teilband*. Berlin/New York: De Gruyter, 968–980. *General foundations for the typologisation of dictionaries (still limited to print dictionaries).*

Wiegand, Herbert Ernst, et al. (2010): Systematic Introduction. In: Wiegand, Herbert Ernst, et al. (eds.): *Wörterbuch zur Lexikographie und Wörterbuchforschung*, vol. 1*: A–C*. Berlin/New York: De Gruyter, 123–225. *Introduction to central concepts of lexicography (including Internet lexicography) with a section on typologies and typological features.*

Literature

Academic literature

Boelhouwer, Bob/Dykstra, Anna/Sijens, Hindrik (2018): Dictionary portals. In: Fuertes-Olivera, Pedro A. (ed.): *The Routledge Handbook of Lexicography*. London/New York: Routledge, Taylor & Francis, 754–765.

de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic-Dictionary Age. In: *International Journal of Lexicography* 16, 143–199.

Engelberg, Stefan/Klosa-Kückelhaus, Annette/Müller-Spitzer, Carolin (2020): Internet lexicography at the Leibniz-Institute for the German Language. In: *K Lexical News* 28, 54–77.

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

Engelberg, Stefan/Müller-Spitzer, Carolin (2013): Dictionary portals. In: Gouws, Rufus, et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin/Boston: De Gruyter, 1023–1035.

Hausmann, Franz Josef (1989): Wörterbuchtypologie. In: Hausmann, Franz Josef et al. (eds.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie. 1. Teilband*. Berlin/New York: De Gruyter, 968–980.

Hildenbrandt, Vera/Moulin, Claudine (2012): Das Trierer Wörterbuchnetz. Vom Einzelwörterbuch zum lexikographischen Informationssystem. In: *Korrespondenzblatt des Vereins für niederdeutsche Sprachforschung* 119, 73–81.

Kühn, Peter (1989): Typologie der Wörterbücher nach Benutzungssituationen. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie. 1. Teilband*. Berlin/New York: De Gruyter, 111–127.

Lew, Robert (2011): Online Dictionaries of English. In: Fuertes-Olivera, Pedro Antonio/Bergenholtz, Henning (eds.): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 230–250.

Meyer, Peter (2022): Lehnwortportal Deutsch: a new architecture for resources on lexical borrowings. In: Klosa-Kückelhaus, Annette, et al. (eds.): *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, *12–16 July 2022*, *Mannheim, Germany*. Mannheim: IDS-Verlag, 577–583.

Meyer, Peter/Engelberg, Stefan (2011): Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In: Hedeland, Hanna/Schmidt, Thomas/Wörner, Kai (eds.): *Multilingual Resources and Multilingual Applications*. Hamburg: Universität Hamburg, 169–174. http://www.corpora.uni-hamburg.de/gscl2011/downloads/AZM96.pdf [last access: April 25, 2024].

Möhrs, Christine/Töpel, Antje (2011): The "Online Bibliography of Electronic Lexicography" (OBELEX). In: Kosem, Iztok/Kosem, Karmen (eds.): *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex2011, Bled, Slowenien*, *10–12 November 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 199–202. http://elex2011.trojina.si/Vsebine/proceedings.html [last access: April 25, 2024].

Müller-Spitzer, Carolin (2010): OWID – A dictionary net for corpus-based lexicography of contemporary German. In: Dykstra, Anne/Schoonheim, Tanneke (eds.): *Proceedings of the XIV Euralex International Congress. Leeuwarden*, *6–10 July 2010*. Leeuwarden: Fryske Akademy, 445–452.

Nesi, Hilary (2000): Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art. In: Heid, Ulrich, et al. (eds.): *Proceedings of the Ninth EURALEX International Congress*, *Stuttgart, Germany*, *August 8th–12th*, *2000*. Stuttgart: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, 839–847.

Perdih, Andrej/Ježovnik, Janoš (2016): Designing and developing the Slovenian dictionary portal Fran. In: COST ENeL WG3 meeting (organized with WG1) Barcelona, Spain, 31 March–1 April 2016.

https://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-barcelona-2016/ [last access: April 25, 2024].

Schröder, Martin (1979): Brauchen wir ein neues Wörterbuchkartell? Zu den Perspektiven einer computerunterstützten Dialektlexikografie und eines Projekts "Deutsches Dialektwörterbuch". In: *Zeitschrift für Dialektologie und Linguistik* LXIV/1, 57–66.

Storrer, Angelika/Freese, Katrin (1996): Wörterbücher im Internet. In: *Deutsche Sprache* 24:2, 97–153.

Tiberius, Carole et al. (2022): An insight into lexicographic practices in Europe. Results of the extended ELEXIS Survey on User Needs. In: Klosa-Kückelhaus, et al. (eds.): *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, *12–16 July 2022*, *Mannheim, Germany*. Mannheim: IDS-Verlag, 509–521.

Tono, Yukio (2004): Research on the Use of Electronic Dictionaries for Language Learning: Methodological Considerations. In: Campoy Cubillo, Maria Carmen/Safont Jordá, Maria Pilar (eds.): *Computer-Mediated Lexicography in the Foreign Language Learning Context*. Castelló de la Plana: Universitat Jaume I, 13–27.

Verlinde, Serge (2011): Modelling Interactive Reading, Translation and Writing Assistants. In: Fuertes-Olivera, Pedro Antonio/Bergenholtz, Henning (eds.): *e-Lexikography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 275–286.

Verlinde, Serge/Peeters, Geert (2012): Data access revisited: The Interactive Language Toolbox. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 147–162.

Wiegand, Herbert Ernst, et al. (2010): Systematic Introduction. In: Wiegand, Herbert Ernst et al. (eds.): *Wörterbuch zur Lexikographie und Wörterbuchforschung*. vol. 1: *A–C*. Berlin/New York: De Gruyter, 123–225.

Dictionaries, portals and other reference works

AISRI = *AISRI Dictionary Portal*. Indiana University: American Indian Studies Research Institute. https://zia.aisri.indiana.edu/~dictsearch/ [last access: April 25, 2024].

ANW = *Algemeen Nederlands Woordenboek*. Leiden: Instituut voor de Nederlandse Taal. https://anw.ivdnt.org/search [last access: April 25, 2024].

ASL-Dict = *ASL Sign Language Dictionary*. Jolanta Lapiak. https://www.handspeak.com/word/ [last access: April 25, 2024].

BLF = *Lexical Database for French (Base lexicale du français – BLF)*. Leuven: Katholieke Universiteit Leuven. http://ilt.kuleuven.be/blf/ [*no longer accessible*].

Cambridge-Online = *Cambridge Dictionaries Online*. Cambridge: Cambridge University Press. http://dictionary.cambridge.org/ [last access: April 25, 2024].

Collins = *Collins*. Glasgow: HarperCollins Publishers. https://www.collinsdictionary.com/ [last access: April 25, 2024].

DD-Linguistikportal = *Dictionary directory of Lin|gu|is|tik portal*. Frankfurt am Main: Goethe-Universität. Fachinformationsdienst Linguistik. https://www.linguistik.de/en/search/dictionary-directory [last access: April 25, 2024].

DSDD = *Database of Southern Dutch Dialects*. Ghent: Ghent Centre for Digital Humanities, Ghent University. https://www.ghentcdh.ugent.be/projects/database-southern-dutch-dialects-dsdd [last access: April 25, 2024].

DWB-Online = Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/DWB/ [last access: April 25, 2024].

DWDS = *Das Digitale Wörterbuch der deutschen Sprache*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de/ [last access: April 25, 2024].

EDP = *European Dictionary Portal*. European Network of e-Lexicography. http://www.dictionaryportal.eu/de/ [last access: April 25, 2024].

Els-Dict = Wörterbuch der elsässischen Mundarten. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/ElsWB/ [last access: April 25, 2024].

Encyclopedia = *Encyclopedia.com*. Chicago: HighBeam Research. http://www.encyclopedia.com/ [last access: April 25, 2024].

English-Place-Names = *Key to English Place Names*. Nottingham: The Institute for Name-Studies, University of Nottingham. http://kepn.nottingham.ac.uk [last access: April 25, 2024].

Enzyklopädienliste-Wikipedia = *Liste von Online-Enzyklopädien.* Wikipedia – Die freie Enzyklopädie. https://de.wikipedia.org/wiki/Liste_von_Online-Enzyklopädien [last access: April 25, 2024].

Etymologiebank = *Etymologiebank.nl*. Nicoline van der Sijs, Instititutt vor de Nederlandse taal, 2010. https://etymologiebank.nl/ [last access: April 25, 2024]

e-VALBU = Elektronisches Valenzwörterbuch deutscher Verben. In: *GRAMMIS*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/evalbu/index.html [last access: April 25, 2024].

e-WDG = Wörterbuch der deutschen Gegenwartssprache online. In: *DWDS*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de [last access: April 25, 2024].

FRAN = *Fran*. Ljubljana: Slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. https://www.fran.si/ [last access: April 25, 2024].

Free-Dict = *The Free Dictionary*. Huntingdon Valley, PA: Farlex Inc. https://www.thefreedictionary.com/ [last access: April 25, 2024].

Glottopedia = *Glottopedia, the free encyclopedia of linguistics*. http://www.glottopedia.org/index.php/Main_Page [last access: April 25, 2024].

Goethe-Dict = Goethe-Wörterbuch. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/GWB/ [last access: April 25, 2024].

HWNF = *Historische woordenboeken Nederlands en Fries*. Leiden: Instituut voor de Nederlandse Taal. https://gtb.ivdnt.org/search/ [last access: April 25, 2024].

Idiomdatenbank = *Idiomdatenbank*. Project “Kollokationen im Wörterbuch”, Christiane Fellbaum. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://kollokationen.bbaw.de/htm/idb_de.html [last access: April 25, 2024].

ILT = *Interactive Language Toolbox*. Leuven: Katholieke Universiteit Leuven. http://ilt.kuleuven.be/blf/ [last access: April 25, 2024].

Kicktionary = Schmidt, Thomas: Kicktionary. Mehrsprachiges digitales Wörterbuch zur Fachsprache des Fußballs. In: *OWIDplus*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.kicktionary.de/index_de.html [last access: April 25, 2024].

Larousse = *Larousse*. Paris. https://www.larousse.fr/ [last access: April 25, 2024].

Lehnwortportal-Deutsch = *Lehnwortportal Deutsch*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://lwp.ids-mannheim.de/ [last access: April 25, 2024].

LEO = *LEO*. Sauerlach: LEO GmbH. http://www.leo.org/ [last access: April 25, 2024].

Linguee = *Linguee*. Köln: Linguee GmbH. http://www.linguee.de/ [last access: April 25, 2024].

Merriam-Webster = *Merriam.Webster*. Springfield: Merriam-Webster Inc. https://www.merriam-webster.com/ [last access: April 25, 2024].

Merriam-Webster-Medical: *Medical Dictionary*. Springfield: Merriam-Webster Inc. https://www.merriam-webster.com/medical [last access: April 25, 2024].

Loth-Dict = Wörterbuch der deutsch-lothringischen Mundarten. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/LothWB/ [last access: April 25, 2024].

Neo-Dict = Neologismenwörterbuch. In: *OWID*. Mannheim: Leibniz-Institut für Deutsche Sprache. https://www.owid.de/docs/neo/start.jsp [last access: April 25, 2024].

OBELEXdict = Online-Bibliografie zur elektronischen Lexikografie – Wörterbücher. In: *OWID*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.owid.de/obelex/dict [last access: April 25, 2024; maintenance until 2020].

OCD = *Oxford Collocations Dictionary*. Oxford: Oxford University Press. https://www.oxfordlearnersdictionaries.com/about/collocations/introduction.html [last access: April 25, 2024].

OED-Online = *Oxford English Dictionary online*. Oxford: Oxford University Press. http://www.dictionary.oed.com [last access: April 25, 2024].

OED-Learner = *Oxford Learners' Dictionaries*. Oxford: Oxford University Press. http://www.oxfordlearnersdictionaries.com/ [last access: April 25, 2024].

OneLook = *OneLook Dictionary Search*. Datamuse. http://www.onelook.com/ [last access: April 25, 2024].

OpenThesaurus-de = *OpenThesaurus.de – Synonyme und Assoziationen*. Daniel Naber. Potsdam. http://www.openthesaurus.de/ [also integrated into DWDS].

Ordnet = *Ordnet.dk*. København: Det Danske Sprog- og Litteraturselskab. https://ordnet.dk/ [last access: April 25, 2024].

OWID = *OWID – Online-Wortschatz-Informationssystem Deutsch*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.owid.de [last access: April 25, 2024].

Pf-Dict = Pfälzisches Wörterbuch. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/PfWB/ [last access: April 25, 2024].

Pfeifer-DWDS = Pfeifer, Wolfgang: Etymologisches Wörterbuch des Deutschen. In: *DWDS*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://www.dwds.de/wb/etymwb/search?q= [last access: April 25, 2024].

Pons-Online = *PONS Online-Wörterbuch*. Stuttgart: PONS GmbH. http://de.pons.com/ [last access: April 25, 2024].

Protestdiskurs-Dict = Protestdiskurs 1967/68. In: *OWID*. Mannheim: Leibniz-Institut für Deutsche Sprache. https://www.owid.de/wb/disk68/start.html [last access: April 25, 2024].

Rh-Dict = Rheinisches Wörterbuch. In: *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/RhWB/ [last access: April 25, 2024].

Schweizer-Ortsnamen = *ortsnamen.ch – Das Portal der schweizerischen Ortsnamenforschung*. Zürich: Schweizerisches Idiotikon. https://www.ortsnamen.ch/de/ [last access: April 25, 2024].

ShakespearesWords = *ShakespearesWords.com. Explore Shakespeare's works like never before*. David Crystal, Ben Crystal. https://www.shakespeareswords.com/ [last access: April 25, 2024].

Slang-Portal = *Slang Portal*. Oslo: Norsk Språkservice. http://www.spraakservice.net/slangportal/ [last access: April 25, 2024].

Unisa-Proverb = *Unisa Multilingual Proverbs Dictionary*. Pretoria: National Institute for the Humanities and Social Sciences, University of South Africa. https://www.unisa.ac.za/sites/corporate/default/Colleges/Human-Sciences/Schools,-departments,-centres,-institutes-&-units/School-of-Arts/Department-of-Linguistics-and-Modern-Languages/Unisa-Multilingual-Proverbs-Dictionary [last access: April 25, 2024].

Urban-Dict = *Urban Dictionary*. San Francisco: Urban Dictionary LLC. https://www.urbandictionary.com/ [last access: April 25, 2024].

Verlaufsformen-Dict = Kleines Wörterbuch der Verlaufsformen im Deutschen. In: *OWID*. Mannheim: Leibniz-Institut für Deutsche Sprache. https://www.owid.de/wb/progdb/start.html [last access: April 25, 2024].

WEBSTER-1828 = *Webster’s Dictionary 1828*. MasonSoft Technology Ltd. https://webstersdictionary1828.com/ [last access: April 25, 2024].

WIKIPEDIA-EN = *Wikipedia, the Free Encyclopedia*. San Francisco: Wikimedia Foundation. https://en.wikipedia.org/wiki/Main_Page [last access: April 25, 2024].

WIKTIONARY = *Wiktionary*. San Francisco: Wikimedia Foundation. https://www.wiktionary.org [last access: April 25, 2024].

WIKTIONARY-EN = *Wiktionary, the Free Dictionary*. San Francisco: Wikimedia Foundation. https://en.wiktionary.org/wiki/Wiktionary:Main_Page [last access: April 25, 2024].

WORDNET = *WordNet. A Lexical Database for English*. Princeton: Princeton University. https://wordnet.princeton.edu/ [last access: April 25, 2024].

WORDREFERENCE = *WordReference.com*. Vienna, VA (USA). http://www.wordreference.com/ [last access: April 25, 2024].

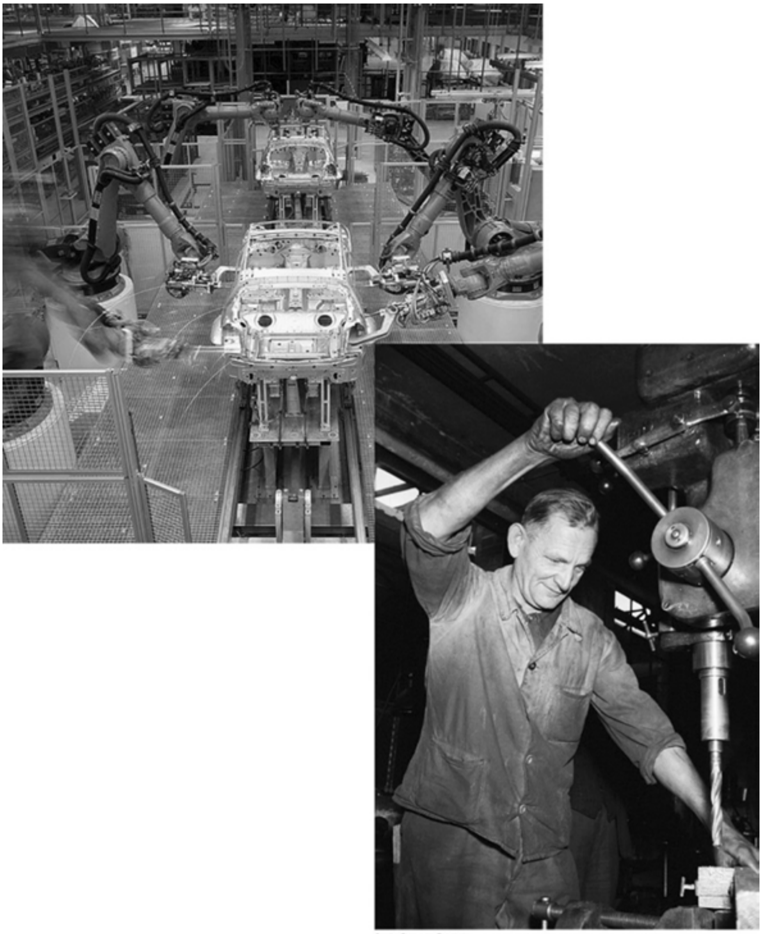
WÖRTERBUCHNETZ = *Wörterbuchnetz*. Trier: Kompetenzzentrum – Trier Center for Digital Humanities, Universität Trier. http://woerterbuchnetz.de/ [last access: April 25, 2024].

Images

Fig. 2.1 “Klassifizierung zur Wissensanordnung”. http://www.wiki-hilfe.de/attach/Pic/klassifizieren_zur_wissensordnung.jpg [last access: April 25, 2024].

Annette Klosa-Kückelhaus and Carole Tiberius

3 The Lexicographic Process



Quelle: Deutsche Fotothek

Fig. 3.1: Robots employed on an assembly line – machinist at work.

For many people nowadays, the assembly line symbolises a production process that runs according to a particular sequence of individual activities and was first introduced for the production of cars. Robots have now replaced people for frequently repeated steps but in

Annette Klosa-Kückelhaus, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: klosa@ids-mannheim.de

Carole Tiberius, Instituut voor de Nederlandse Taal, Rapenburg 61, 2311 GJ Leiden, The Netherlands, e-mail: carole.tiberius@ivdnt.org

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/10.1515/9783111233758-004

this type of manufacturing process, it is – still – a human being, here a machinist, who assembles and repairs the machines and whose specialist skills and expertise are indispensable. It is precisely within this tension – between automated work on the one hand and lexicographic activity on the other – that the lexicographic processes for born-digital dictionaries take place.

Until the mid-2010s, the lexicographic process was examined almost exclusively in relation to print dictionaries. But for Internet dictionaries, this process takes on an altogether different form: here, it is not a question of describing a linear series of individual production phases but rather individual tasks which are permanently intertwined and run in parallel with one another. A whole series of questions present themselves in this context, such as how subsections of the lexicon are chosen for editing, how new ways of gathering data from electronic text corpora influence the lexicographic process, what software can be used to support lexicographic processes, and what impact all of these changes have on users consulting dictionaries. There are also lexicographic portals in which different dictionaries are combined (→ Chapter 2.4) as well as centralised lexicographic databases from which single dictionaries can be generated, both of which have their own lexicographic processes.

3.1 Introduction

Since time immemorial, the form and content of dictionaries have been the focus of academic dictionary criticism and academic studies of dictionaries. Form and content are easily accessible from the outside because both the structure and substance are visible on paper in a printed dictionary or on the screen in the case of an electronic dictionary. However, the process by which a dictionary comes into being tends to take place much more in the background. By this process, we mean all of the tasks necessary to create a single dictionary or a (centralised) lexicographic database as well as all of the steps that need to be taken to compile a product (a dictionary or other language application) out of the data. These processes involve the participation of people with different skill sets and the use of different technical resources for different periods of time and in a particular order. Traditionally, dictionary editors scarcely provided any detailed information about these internal processes, something which applies equally to commercial and academic lexicography.

Nonetheless, these processes are worth investigating and describing for a variety of reasons:

- The form and content of a dictionary can be judged more appropriately against the background of knowledge about the circumstances in which it originated; this is important, for example, in the context of critical reviews of existing dictionaries. For instance, knowing that there were no financial resources in a dictionary project to be able to buy illustrations or produce their own can explain why recourse had to be made to freely available images that may not have fulfilled their purpose to optimum effect. A more valid comparison can also be drawn between different dictionaries when information exists about their development timelines and lexicographic teams.

- When the individual processes and participants in the overall production process of a dictionary are known, meaningful suggestions can be developed on this basis for future improvements to the dictionary. For example, if possible ways of presenting data graphically are hardly exploited in an Internet dictionary, or not at all, this might be because the necessary technical know-how was missing from the project or the corresponding work stages were not planned into it. Accordingly, the dictionary editors can change their planning to include a possible visualisation of data if demand arises from users or dictionary researchers, for example.
- The impact of lexicographic processes on published dictionaries, (centralised) lexicographic databases, or lexicographic portals can make the planning of completely new lexicographic projects easier; new dictionaries, databases, or portals can be compiled more efficiently and consistently by learning from mistakes in the planning of existing projects. This may be the case, for example, when an existing dictionary team does not have any corpus-linguistic competence of their own at their disposal so that they are dependent on buying in this expertise from the outside in order to automatically identify corpus evidence. When similar evidence is planned for a new dictionary, this can be avoided if that project has its own corpus linguists in its team from the beginning.
- For users, a good understanding is necessary of specialised aspects of the lexicographic process for lexicographic portals and Internet dictionaries so that they can understand the new elements involved when using them. If they are interested, users should be able to find out about the lexicographic process so that they can assess, for example, to what extent a lexicographic resource is current and can be cited.
- Knowledge about the lexicographic process makes it possible to understand lexicographic resources as the product of many different cognitive, linguistic, computational-linguistic, corpus-linguistic, and editorial tasks and, therefore, as complex cultural goods. In this way, users can learn to assess the quality of edited Internet dictionaries and portals and support their continued existence by using them frequently.

Therefore, we are devoting a whole chapter of this book to the lexicographic processes behind Internet dictionaries, centralised lexicographic databases, and lexicographic portals. Some of what was previously known and well established concerning the lexicographic process in practical lexicography and dictionary research may no longer apply to dictionaries compiled for publication on the Internet. Thus, it is necessary “to unlearn a great deal of what we know” (Gouws 2011: 29; → Section 3.3). The initial foundation for our topic is provided by a brief summary of academic research into the production process (→ Section 3.2) and a description of various lexicographic processes for Internet dictionaries of different types (→ Section 3.4). In addition, we consider the process of networks of Internet dictionaries (so-called dictionary portals, → Section 3.6) and centralised lexicographic databases (→ Section 3.7). Since electronic dictionaries cannot be produced without using computers, we show what software can be employed for this in → Section 3.5.

3.2 Research into the lexicographic process

Descriptions of the lexicographic process did not begin in essence until the 1980s, at which time they were restricted entirely to print dictionaries (e.g. Dubois 1990, Riedel/Wille 1979, and Schaeder 1987). To radically simplify and abbreviate matters, Landau (1984: 227) identified three phases that occur in every lexicographic process: "planning (30%), writing (50%), and producing (20%)". Here, the writing phase is said to last considerably longer for academic dictionaries in general than for commercial dictionaries.

Zgusta (1971: 223) defined the following stages in lexicographic work: "(1) the collection of material; (2) the selection of entries; (3) the construction of entries; and (4) the arrangement of the entries". According to Landau (1984), the following tasks are part of the core process, the writing phase, for print dictionaries: drafting a definition, editing it, preparing the text for typesetting (including typographical markups), or proofreading different stages of typesetting (proofs and wraps). As such, these activities include, on the one hand, core lexicographic tasks (explaining the meaning or meanings of a lemma) and, on the other, tasks that generally arise in the production of print media (proofreading). Generally, it is clear that the individual stages of work for print dictionaries are assumed to be linear, at least to some extent. However, it is less clear which individual tasks arise in the planning and production of a dictionary and who undertakes them.

It is to Wiegand that we owe the first complete description and theoretical conceptualisation of the lexicographic process. He defined the lexicographic process as follows (1998: 134):

> Ein abgeschlossener lexikographischer Prozeß [. . .] ist die Menge derjenigen prozeßzugehörigen Tätigkeiten, welche ausgeführt wurden, damit ein bestimmtes Wörterbuch entsteht. [A completed lexicographic process [. . .] is the set of process activities carried out to create a particular dictionary.]

Wiegand (1998: 135) divided the lexicographic process for producing print dictionaries into five phases: preparation, acquisition of material, editing of material, evaluation of material, and typesetting and print preparation. He distinguished generally between lexicographic processes without the use of computers and computer-aided lexicographic processes, i.e. those in which all phases of the process involve the use of computers (Wiegand 1998: 233; for more on the deployment of computers in the lexicographic process, cf. also Knowles 1990: 1648). However, the goal of both processes is to compile a print dictionary. By contrast, the goal of a digital lexicographic process (involving the use of computers throughout the process) is to compile a dictionary that is not published in print but on an electronic data carrier (Wiegand 1998: 239).

Through critical engagement with Wiegand's proposed model, Müller-Spitzer (2003: 161) arrived at a further elaboration of different lexicographic processes based, first, on whether the dictionary is intended to serve human users or as a resource for language technology. A second distinction is made between a dictionary intended to be published purely on an electronic data medium and a body of data that is medium-neutral and from which both print and electronic dictionaries can be published. As such, all of the lexicographic processes for Internet dictionaries are either digital lexicographic processes or lexicographic processes that are conceived as medium-neutral.[1]

Reflection on involving the human user systematically in the lexicographic process (→ Chapter 8), and more specifically the digital lexicographic process, started at the beginning of the 21st century. Feedback from dictionary users during the development of a dictionary can bring about a clear improvement in quality (e.g. in relation to lemma coverage; cf. de Schryver/Prinsloo 2000a, 2000b). However, the involvement of users (→ Chapter 8) is also useful, for example, in the material acquisition phase if the dictionary is looking for first attestations or examples from sources which are difficult to find. Lexicographers then take on a stronger organisational role in the digital lexicographic process, especially in dictionaries that are compiled semi-collaboratively (Abel/Klosa 2014: 7). Careful planning and supervision of the process is essential, but it is also necessary to make this procedure transparent to dictionary users.

With more extensive dictionary teams, there is a change in the specific expertise needed, as indicated by Wiegand (1998). Thus, a series of other experts are involved in digital lexicographic processes in addition to lexicographers, for example, corpus linguists, computational linguists, text technologists, IT specialists, and designers (Klosa 2013: 504). Against this background, norms and encoding formats for digital lexicography have been established (cf. TEI Lex-0, ONTOLEX-LEMON, and the ISO LMF (Lexical Markup Framework) standard as well as ISO 1951:2007, and, most recently, DMLEX).

3.3 The digital lexicographic process for Internet dictionaries and its particularities

If we use the media-specific characteristics of different types of dictionary to distinguish between them (→ Chapter 2), we can differentiate between the types of Internet dictionary listed in → Tab. 3.1. The original form of an Internet dictionary particularly (but

[1] In what follows, we use the shorthand "digital lexicographic process".

not only) influences the dictionary's online structure, the density of hypertexts, the number of multimedia elements, and the means of access available. The criterion of completeness has a particular influence on the lexicographic process and it does so continuously until the dictionary is (possibly, but not necessarily) finished. Schröder (1997) thus introduced the distinction between "Abschlusswörterbuch" (completed dictionary) and "Ausbauwörterbuch" (dictionary under construction) while Lemberg (2001) referred to "statisch" (static) as opposed to "dynamisch" (dynamic) dictionaries. As far as dictionaries under construction are concerned (→ Section 3.3.2), we can further differentiate between those that initially appeared in print and were then digitised retrospectively before being (continuously) extended and revised (e.g. the Oxford English Dictionary [OED] or the Deutsches Rechtswörterbuch [DRWB]) and those that were planned directly for online publication and are continuously extended and revised (e.g. the Algemeen Nederlands Woordenboek [ANW] or elexiko, which, however, stopped as a project in 2017).

Tab. 3.1: Possible classification of Internet dictionaries (following Storrer/Freese 1996; Storrer 1998; Storrer 2001).

Characteristic	Type of Internet dictionary
Original form of publication	– first appeared as a print dictionary – first appeared as an electronic offline dictionary – appeared directly as an online dictionary
Completeness	– completed dictionary – dictionary under construction
Hypertexts	– dictionary with hypertexts – dictionary without hypertexts
User interaction	– dictionary with user interaction – dictionary without user interaction
Multimedia	– dictionary with text, illustrations, tables, diagrams – dictionary with text and audio data – dictionary with text, illustrations, and audio data – dictionary without multimedia
Access to the dictionary	– access by scrolling through the list of lemmas – access via a list of lemmas with specific characteristics in hypertext form – access via search options – combined forms of access

Finally, user interaction influences the lexicographic process insofar as dictionaries with user participation have to account for users' feedback in specific phases of their process (→ Section 3.3.3, → Chapter 8). Before examining the particularities of the lexicographic process for these different types of dictionaries, we can describe the digital lexicographic process more generally in seven phases, each with many individual tasks (→ Tab. 3.2), using Wiegand (1998: 233ff.) as a starting point.

Tab. 3.2: Phases and tasks in the digital lexicographic process for Internet dictionaries.

Phase	Tasks (selected)
Preparatory phase	Dictionary outline, organisational plan (finances, workflow, timetable, staffing), pilot studies on lexicographic information types and the list of lemmas, plan for the dictionary (rough modelling of the data structure, editorial guidelines, model entry, planning the user interface and access structures, planning technical support, planning the versioning and archiving of dictionary data, planning user involvement and user studies)
Data acquisition phase	Acquiring primary lexicographic sources (corpus construction) and additional sources (e.g. reference dictionaries), acquiring further data (e.g. illustrations, videos, audio data)
Computerisation phase	Preparing corpus texts (tagging, lemmatisation), programming/acquiring a corpus research and analysis tool, programming/acquiring a dictionary writing system and implementing the data model, programming for the versioning and archiving of dictionary contents, programming for the user interface, acquiring and installing necessary hardware and any further software
Data processing phase	Compiling potential lemmas and frequency lists, defining frequency levels and classes, analysing co-occurrences, labelling image and audio data, integration of automatically identified data into dictionary entries (e.g. frequency, collocations)
Data evaluation phase	Linking dictionary entries and corpus, writing entries, inserting hyperlinks, integrating illustrations and other multimedia elements
Preparation phase for online release	Proofing content and form, testing the user interface (hyperlinks, multimedia elements, search options, etc.), writing user manuals for the dictionary, developing a “guided tour” of the dictionary
Maintenance and preservation phrase	Archiving different versions of the data and version control; maintenance of the online application

Different technical qualifications are needed to undertake the tasks listed. Ideally, everybody involved in the dictionary project participates in the preparatory phase, i.e. lexicographers and corpus linguists when planning the design of the corpus, or lexicographers and graphic designers or specialists in web design when conceiving the layout of the dictionary entries. Corpus linguists (acquiring corpus texts) and lexicographers (inspecting and gathering other sources) also work together in the data acquisition phase. The tasks involved in the computerisation phase lie primarily with corpus and computational linguists and, possibly, programmers, but most of these tasks cannot proceed without the agreement of the team of lexicographers.

The phase that takes up the most amount of time in the lexicographic process is the data evaluation phase in most dictionary projects. This is the lexicographers' main area of responsibility, but this only partially applies to Internet dictionaries if information compiled automatically from corpora, or with natural language processing software, or generated with the help of Large Language Models (→ Chapter 6) appears in the dictionary alongside lexicographic information prepared by editors. For example, it is the responsibility of corpus and computational linguists to ensure that examples are extracted automatically from the corpus and information about syllabification is automatically generated for the lemmas. Nonetheless, in many dictionary projects, information that has been compiled automatically is checked by the editors before it appears in the dictionary as part of what is called "post-editing lexicography" (cf. Jakubíček 2017). This checking and correction takes place in the phase when the dictionary is being prepared for online release, during which the lexicographers work together with the programmers to test the user interface. Once the dictionary (or dictionary portal) has been published, its website needs to be maintained and preserved as one of the last phases of the process (cf. Svensén 2009: 413; Tiberius/Krek 2014: 1).

Below we consider finished dictionaries that have been retrospectively digitised in → Section 3.3.1 and then two types of dictionaries under construction in → Section 3.3.2 before reflecting on the implications of dictionary users' feedback for the digital lexicographic process in → Section 3.3.3.

3.3.1 The digital lexicographic process in retro-digitised Internet dictionaries

Retro-digitised dictionaries are finished dictionaries originally published in print whose lexicographic process has already run its complete course. However, they are to be made newly available in an electronic medium (here, the Internet). As such, of the phases of the lexicographic process described above, only the following apply in the case of retrospective digitisation of a dictionary that is to be published online (for an example → Section 3.4.3):

- Preparatory phase: organisational plan (finances, workflow, timetable, staffing) and plan for the dictionary (rough modelling of the data structure, possibly in alignment with a centralised lexicographic database if needed; planning the user interface and access structures, planning technical support, and potentially planning for user involvement);
- Computerisation phase: acquiring and installing the necessary hardware and software, implementing the data model and the user interface;

- Preparation phase for online release: testing the user interface, developing a "guided tour" of the dictionary;
- Maintenance and preservation phase: → Tab. 3.2.

However, the phase with the greatest overall significance for the lexicographic process – the data evaluation phase including the core lexicographic, corpus linguistic, and computational linguistic tasks – is missing altogether, as is the data acquisition phase. As such, it is worth asking whether the process leading to the publication of a retrospectively digitised dictionary can be defined as a digital lexicographic process at all (→ Section 3.6). If a retro-digitised Internet dictionary is not simply a 1:1 replica of the underlying print dictionary but rather has an added lexicographic value online, we consider it legitimate to refer to at least a digital lexicographic sub-process. This added value exists, for example, if:

- the lexicographic information is presented in a different way (e.g., distributed on seperate windows instead of a simple reproduction of the print image),
- the dictionary can be searched in an innovative way,
- the mediostructure is extended systematically with hyperlinks.

Retro-digitisation projects primarily require the expertise of computational linguists and programmers as the tasks are mostly computational. However, lexicographic expertise is indispensable in a successful retro-digitisation dictionary project for planning the user interface and access structures as well as for modelling the data structure and extending the content.

3.3.2 The digital lexicographic process in dictionaries under construction

As mentioned above, when it comes to dictionaries under construction we must first of all distinguish between whether the dictionary is a completely new creation (for examples → Section 3.4.1) or whether a print dictionary is first being retro-digitised (in parts or as a whole), then published online, and continuously extended in that form. The latter raise the interesting case of a lexicographic process for a print dictionary, which may or may not be computer-aided, that has been completed and after which a digital lexicographic process is initiated for the Internet dictionary. In this scenario, certain phases in these processes may be omitted, others may replace them, and new tasks may be added:

- Preparatory phase: only questions specific to the medium of the Internet have to be clarified e.g. planning the retrospective digitisation, planning the further development of the dictionary in particular subsets of the corpus, potentially planning the development of the underlying data, and planning the user interface.

- Data acquisition phase: this can be omitted if only the material from the print dictionary (not edited electronically) actually serves as the primary source for the further development of the dictionary.
- Computerisation phase: the development of the Internet dictionary cannot be realised without using computers in a consistent way, even if the print dictionary has already been compiled with the help of computers. The user interface also has to be implemented.
- Data processing phase: this phase is also omitted if no new sources are to be used for the dictionary.
- Data evaluation phase: the tasks already undertaken by lexicographers when evaluating material for the print dictionary are also necessary for the further development of the Internet dictionary. Further tasks might be added, e.g. editing and preparing hyperlinks.
- Preparation phase for online release: this phase replaces the typesetting and print preparation phase for a printed dictionary.
- Maintenance and preservation phase: this phase did not exist for the printed dictionary but is of major importance for the online dictionary under construction, which needs updates as well as version control.

For print dictionaries and finished (retro-digitised) electronic dictionaries (→ Section 3.3.1), the typesetting and print preparation phase, or the preparation phase for online release, only begins once the preceding phases in the process have been completed. By contrast, dictionaries under construction that are intended for Internet publication from the very start are published gradually, bit by bit. Furthermore, dictionaries under construction are not necessarily compiled in alphabetical order but may appear instead in modules since the alphabetical sorting and listing of lemmas has become redundant as an access structure (→ Chapter 4). The body of entries included in a module can be defined in different ways (e.g. on the basis of frequency or according to word class; → Section 3.3). Working on a module (or partial lexicon) chosen according to particular criteria is an advantage for the lexicographers because in this way the lemmas can often be edited more consistently (Storrer 2001: 61f.).

Within a dictionary under construction project, it is also possible to work on multiple modules in parallel. Hence, it might be the case that one module is still in the preparatory phase, another is under construction, and a third has already appeared online. Depending on the decisions taken about the publication cycle in the dictionary project, the different phases can overlap even more. For example, if a project decides to release each completed entry immediately instead of posting larger finished collections of entries online (e.g. every quarter), the following situation can arise: the data for some of the lemmas in a module are still being evaluated, other lemmas are nearly ready to be published online, and others have already been published. If a dictionary combines the automatic compilation of lexicographic information with the manual editing of lemmas, the boundaries between the phases in the process will shift still further: a lemma may already have appeared online with its automatically compiled data while subsequent editorial treatment of the same lemma may only be entering the data evaluation phase. As such, the same lemma may find itself in two different phases of the lexicographic process at one and the same time.

These reflections make clear that dictionaries under construction are "offene Systeme" (open systems; Schröder 1997: 16). As such, the digital lexicographic process for these dictionaries is to be seen as more circular than linear. It is essential that this process is carefully planned and continuously checked; in addition, all of those involved must be in a position to undertake different tasks from different phases of the process in parallel. This is carried out more easily when a dictionary under construction is generated from an existing (centralised) lexicographic database. In such a scenario, data from retro-digitised resources, corpus data, or other linguistic information can be combined with newly compiled lexicographic information.

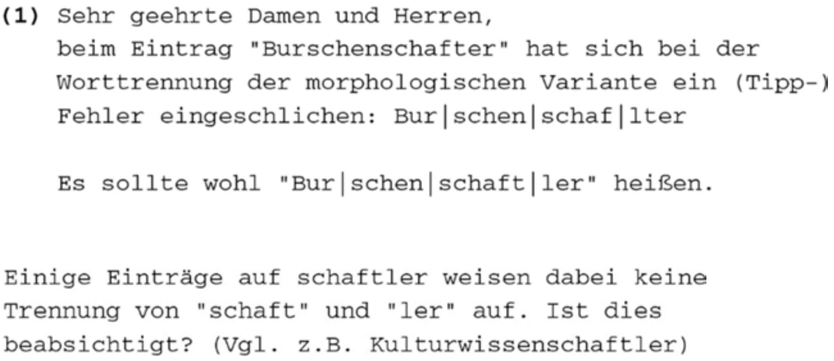
3.3.3 The digital lexicographic process and dictionary users

There is a range of options for involving users directly in the work on dictionaries (→ Chapter 8) and this can have an impact on the lexicographic process for Internet dictionaries:[2]

- Users are asked to report errors and/or to make suggestions for existing entries and new entries;
- Users can provide comments on a word entry that are answered by the editors (both the comments and answers are available online for other users);
- Users privately ask questions on the content of the dictionary that are answered by the editorial team;
- Users give each other advice (e.g. in forums) or are asked to assess the content provided by other contributors.

What happens after errors have been logged by users depends on the type of error reported. In → example (1) a user of the German online dictionary ELEXIKO informed the editors that there was an orthographic mistake in the morphological variant *Burschenschaftler* given for the lemma *Burschenschafter* 'member of a student fraternity'. The user also pointed out that not all entries for German words ending with -*schaftler* showed the correct syllabification ". . . schaft|ler" and asked whether this had been done on purpose:

[2] For more on the digital lexicographic process in collaborative dictionaries such as WIKTIONARY, → Section 3.4.2.



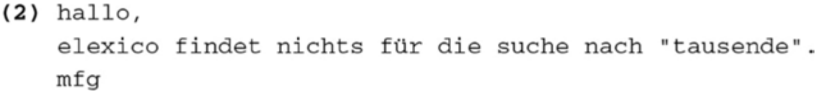
(1) Sehr geehrte Damen und Herren,
beim Eintrag "Burschenschafter" hat sich bei der
Worttrennung der morphologischen Variante ein (Tipp-)
Fehler eingeschlichen: Bur|schen|schaf|lter

Es sollte wohl "Bur|schen|schaft|ler" heißen.

Einige Einträge auf schaftler weisen dabei keine
Trennung von "schaft" und "ler" auf. Ist dies
beabsichtigt? (Vgl. z.B. Kulturwissenschaftler)

The mistake in this example requires making a correction in a single entry and then systematically checking a whole series of words for corresponding corrections. In a case like this, a lexicographer has to put the relevant entry or the relevant groups of headwords back into the data evaluation phase, undertake the corrections, check them again in the online preparation phase, and finally re-release the relevant entry or entries online.

In → example (2) a user of ELEXIKO noticed that a search for *tausende* 'thousands' did not yield any result.



(2) hallo,
elexico findet nichts für die suche nach "tausende".
mfg

This error relates to the list of lemmas or the dictionary's search functionality rather than to individual lemmas or a group of lemmas. If the missing lemma really is absent, even though it should be on the list according to the original concept behind that list, the lexicographers will also return to the data preparation and evaluation phases here. For example, they will check in the corpus whether the missing word is present in the correct spelling in sufficient frequency. It is possible that all of the other stages of editing and revision will follow on here before the lemma can appear online. The process will be different if the word could not be found because of an error in the search functionality or the way the search functionality was realised. In this case, the relevant technical colleagues will have to return to the computerisation phase. If the programming of the search options needs to be corrected, this, of course, will need to be tested again before the improved version is released online.

In general, dealing both with errors reported by users and with their suggestions for improvements, additions, new lemmas, or the revision of selected lemmas and similar means that for dynamic Internet dictionaries, certain phases of the digital lexicographic process are undertaken again. Requests from dictionary users or discussions in dictionary blogs may have a similar effect, namely when these serve as an impetus for the dictionary team to further develop the dictionary in a particular way or to undertake systematic corrections or additions.

Research into the use of the dictionary (→ Chapter 8) can have similar consequences. For example, unsuccessful searches in the log files may lead to the conclusion that potential headwords have been omitted and the dictionary project may decide for these words to be worked on retrospectively. In this case, the project has to go back to the phases of data processing and analysis as well as preparation for publication in the lexicographic process. If the results of a user survey indicate that the dictionary should be urgently updated with information about pronunciation in the form of audio data, but these were not originally envisaged and therefore have to be acquired only now, the project needs to go back to the data acquisition phase. Or the dictionary project has to return to the preparatory phase because an evaluation of the user interface in user studies requires a thorough revision of the interface.

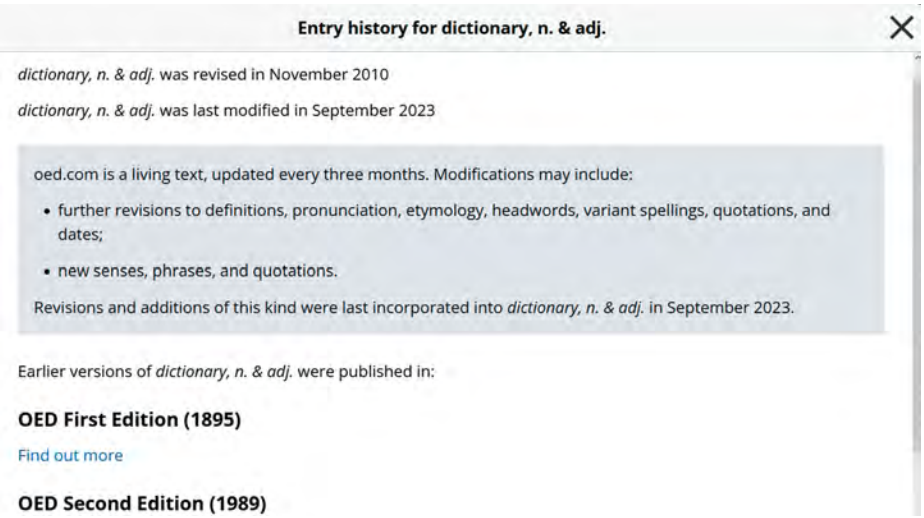
To some extent this circularity is characteristic for the lexicographic process in Internet dictionaries under construction because of the involvement and consideration of dictionary users. However, it need not only be burdensome in terms of the dictionary team being forced to permanently correct and add to the dictionary; it can also be an opportunity because the dictionary can be developed and improved in terms of its user friendliness and overall usefulness. Furthermore, these steps can also make it possible for those consulting the dictionary to come to terms better with the changing state of the dictionary. However, the latter can only work if there is a careful versioning of the dictionary.

In the ideal case, all revisions, additions, or deletions in a dictionary under construction have to be marked in a transparent way for users (or at least for users in an academic context, where exact bibliographic details are needed for citations from the dictionary). However, as a bare minimum, the different stages of revision should be recorded (as in → Fig. 3.2 in the example from the OED) or the relevant most recent versions or date of revision (as in → Fig. 3.3 in the example from the DIGITALES WÖRTERBUCH DER DEUTSCHEN SPRACHE [DWDS]).

It must be borne in mind that it is very costly in technical terms to keep older editions of the dictionary genuinely available. As such, the project team must weigh up whether the effort is worthwhile for information that may be accessed very rarely.

3.4 Examples of lexicographic processes for Internet dictionaries

In this section we consider and discuss the lexicographic practice and process for four Internet dictionaries: the ALGEMEEN NEDERLANDS WOORDENBOEK (ANW), ELEXIKO, and the German-language WIKTIONARY, which were planned for online publication only, and a retro-digitised dictionary, the ORDBOG OVER DET DANSKE SPROG (ODS).



Entry history for dictionary, n. & adj.

dictionary, n. & adj. was revised in November 2010

dictionary, n. & adj. was last modified in September 2023

oed.com is a living text, updated every three months. Modifications may include:

- further revisions to definitions, pronunciation, etymology, headwords, variant spellings, quotations, and dates;
- new senses, phrases, and quotations.

Revisions and additions of this kind were last incorporated into *dictionary, n. & adj.* in September 2023.

Earlier versions of *dictionary, n. & adj.* were published in:

OED First Edition (1895)

Find out more

OED Second Edition (1989)

Fig. 3.2: Record of the publication history of the entry *dictionary* in the OED.[3]

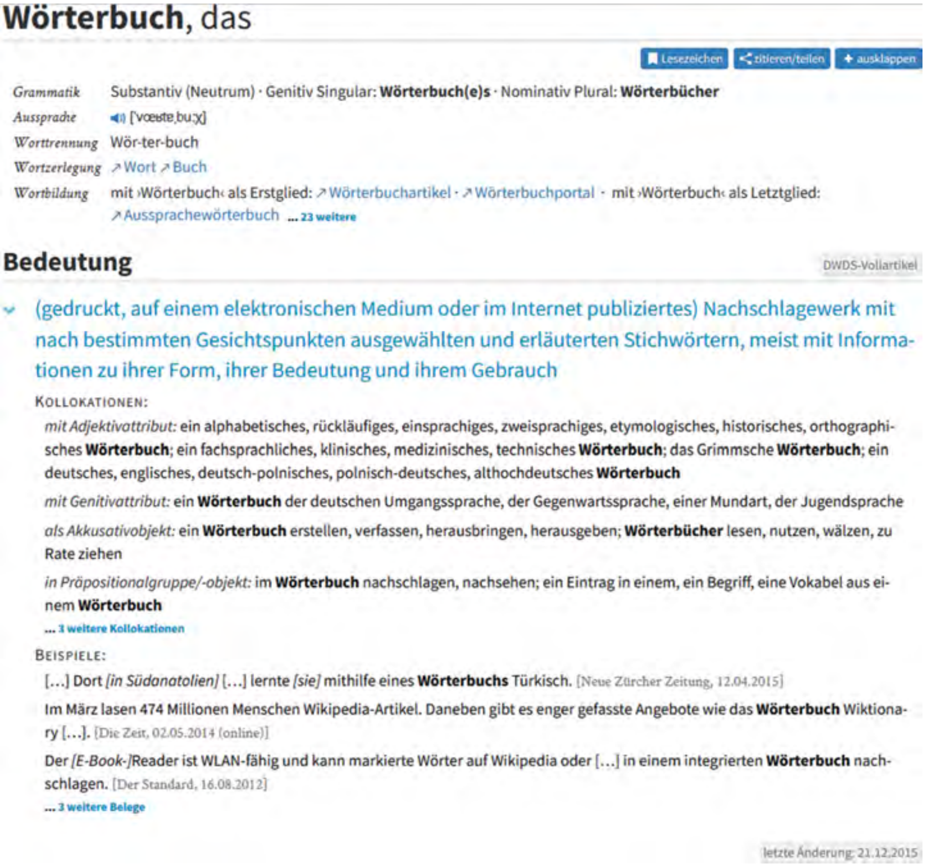
3.4.1 The lexicographic process for the ANW and ELEXIKO

Here, we combine our discussion of the lexicographic processes for the ANW and ELEXIKO since the two projects were relatively similar from the outset. Both are academic, corpus-based dictionaries that describe contemporary language use, the ANW for Dutch and ELEXIKO for German. The ANW is one of the projects run by the Instituut voor de Nederlandse Taal (INT) in Leiden, and ELEXIKO is a project realised at the Leibniz Institute for the German Language (IDS) in Mannheim. The ANW and ELEXIKO are Internet dictionaries that are (or were) open in the truest sense of the expression. From the very beginning, both were planned for online publication including continuous further development (in the case of ELEXIKO, until 2017, when the project ended).

→ Table 3.3 shows the different phases of the digital lexicographic processes involved in the ANW and ELEXIKO between 2001 and 2017. The table makes clear that the different phases no longer take place neatly one after the other, as they would for a print dictionary, but rather that they run in parallel and that it can be difficult to separate the individual phases from each another. In general, we notice many overlaps between the individual phases, except in the data preparation phase.

Determining the end of a phase is not always unequivocal. For instance, the phase of data processing may be reactivated whenever new technology becomes available, like for the automatic extraction of examples from a corpus or the com-

[3] Webpage last accessed on 24 March 2024.



Wörterbuch, das

Grammatik Substantiv (Neutrum) · Genitiv Singular: **Wörterbuch(e)s** · Nominativ Plural: **Wörterbücher**
Aussprache [ˈvœʁtɐˌbuːx]
Worttrennung Wör-ter-buch
Wortzerlegung ↗Wort ↗Buch
Wortbildung mit ›Wörterbuch‹ als Erstglied: ↗Wörterbuchartikel · ↗Wörterbuchportal · mit ›Wörterbuch‹ als Letztglied: ↗Aussprachewörterbuch … 23 weitere

Bedeutung

(gedruckt, auf einem elektronischen Medium oder im Internet publiziertes) Nachschlagewerk mit nach bestimmten Gesichtspunkten ausgewählten und erläuterten Stichwörtern, meist mit Informationen zu ihrer Form, ihrer Bedeutung und ihrem Gebrauch

KOLLOKATIONEN:

mit Adjektivattribut: ein alphabetisches, rückläufiges, einsprachiges, zweisprachiges, etymologisches, historisches, orthographisches **Wörterbuch**; ein fachsprachliches, klinisches, medizinisches, technisches **Wörterbuch**; das Grimmsche **Wörterbuch**; ein deutsches, englisches, deutsch-polnisches, polnisch-deutsches, althochdeutsches **Wörterbuch**

mit Genitivattribut: ein **Wörterbuch** der deutschen Umgangssprache, der Gegenwartssprache, einer Mundart, der Jugendsprache

als Akkusativobjekt: ein **Wörterbuch** erstellen, verfassen, herausbringen, herausgeben; **Wörterbücher** lesen, nutzen, wälzen, zu Rate ziehen

in Präpositionalgruppe/-objekt: im **Wörterbuch** nachschlagen, nachsehen; ein Eintrag in einem, ein Begriff, eine Vokabel aus einem **Wörterbuch**

… 3 weitere Kollokationen

BEISPIELE:

[…] Dort *[in Südanatolien]* […] lernte *[sie]* mithilfe eines **Wörterbuchs** Türkisch. [Neue Zürcher Zeitung, 12.04.2015]

Im März lasen 474 Millionen Menschen Wikipedia-Artikel. Daneben gibt es enger gefasste Angebote wie das **Wörterbuch** Wiktionary […]. [Die Zeit, 02.05.2014 (online)]

Der *[E-Book-]*Reader ist WLAN-fähig und kann markierte Wörter auf Wikipedia oder […] in einem integrierten **Wörterbuch** nachschlagen. [Der Standard, 16.08.2012]

… 3 weitere Belege

letzte Änderung: 21.12.2015

Fig. 3.3: Information on the last update (bottom right) for the entry *Wörterbuch* ‘dictionary’ in the DWDS.[4]

puter-aided compilation of word families. The data in → Tab. 3.3 also indicate that the ANW was more dependent on computer support than ELEXIKO.

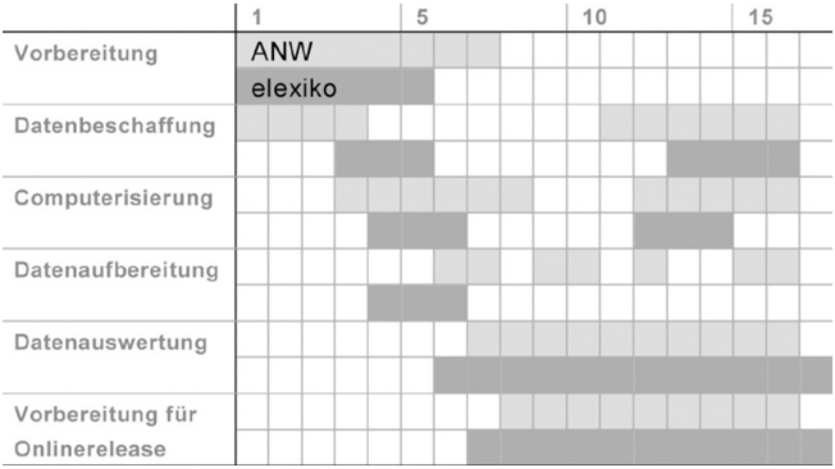
In what follows we provide a comparative analysis of the different phases in the lexicographic processes of the two projects.

Preparatory phase

In the field of academic lexicography, both the ANW and ELEXIKO broke new ground, and essentially the only useful experience that the projects could build on came from print lexicography. For that reason, some essential steps were missed out on or incorrectly im-

[4] Webpage last accessed on 24 March 2024.

Tab. 3.3: Process phases (over the years) for the ANW and ELEXIKO (left column: preparatory phase – data acquisition phase – computerisation phase – data processing phase – data evaluation phase – preparation phase for online release; the last phase for maintenance and preservation is not shown but started as soon as the dictionaries were released online).



	1					5					10					15		
Vorbereitung	ANW	░	░	░	░	░	░	░										
	elexiko	▓	▓	▓	▓	▓												
Datenbeschaffung	░	░	░	░								░	░	░	░	░	░	
				▓	▓	▓								▓	▓	▓	▓	
Computerisierung				░	░	░	░	░	░				░	░	░	░	░	
					▓	▓	▓						▓	▓	▓			
Datenaufbereitung							░	░		░	░		░			░	░	
					▓	▓	▓											
Datenauswertung								░	░	░	░	░	░	░	░	░	░	
							▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓
Vorbereitung für Onlinerelease									░	░	░	░	░	░	░	░	░	
								▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓

plemented in both projects. In the preparatory phase, both projects concentrated on developing the concept of the dictionary's content and on pilot studies, underestimating the importance of constructing a detailed organisational plan including information about finances, staffing, timetable, and workflow. For example, user studies were not initially envisaged or carried out, with log data being stored since the beginning of the online publication of the ANW in 2009 but not analysed systematically; only incidental detailed log analyses have been undertaken (cf. Tiberius/Niestadt 2015). Furthermore, as Abel/Klosa (2012) concluded, no proper market analysis of existing dictionary writing systems was carried out for ELEXIKO. As a result, both projects incurred unnecessary expenditure of time and money in order to improve or retrospectively undertake tasks that were initially poorly planned or not planned at all.

Data acquisition phase

Since the ANW and ELEXIKO are corpus-based dictionaries, the data acquisition phase was devoted above all to assembling the corpus. The ANW corpus was originally a closed corpus of 100 million words that was specially compiled for the project. This idea has since been reconsidered, and new material is being added to the corpus. The ELEXIKO corpus was a dynamic corpus constituted virtually from the German reference corpus (DEREKO) at the IDS Mannheim and subsequently updated continuously until the end of the project.

Both projects also invested much time in gathering image material (in particular, image material that can be used free of charge). The ANW uses online databases of public domain photographs, such as WIKIMEDIA, as well as other illustrations from the Internet (always with reference to the original source. The illustrations in ELEXIKO also came from online databases of public domain photographs such as WIKIMEDIA or PIXELIO and were only systematically added to the edited lemmas between 2012 and 2017 so that the project had to return to the data acquisition phase for illustrations after it had already been running for a long time.

The ELEXIKO project also planned, successfully, to illustrate the natural pronunciation of lemmas (in context) with the help of examples from sound recordings. Here, up to three examples per lemma had been selected for this purpose from the "Archiv für gesprochenes Deutsch" (AGD) held by the Leibniz Institute for the German Language (IDS), although the examples were only available to download and listen to from 2012 onwards.

These examples from the data acquisition phase demonstrate that the phases do not really run one after another but can repeatedly overlap. For most dictionaries under construction, the data acquisition phase runs until the end of the project, although the main period lies at the beginning of the project with the original conceptualisation and constitution of the corpus.

Computerisation phase

In the computerisation phase, both projects prepared and processed the corpus texts so that they could be used lexicographically and established an editorial system (also known as a dictionary writing system). The preparation and processing of the corpus involved, first, tokenising and enriching the corpus with lemmas and part-of-speech tags and, second, loading the corpus into a system in which it could be searched. For the ANW, an in-house corpus query system was used originally but this was replaced in 2007 by the commercial system SKETCH ENGINE (Kilgarriff et al. 2004) and now a combination of SKETCH ENGINE and an in-house system is used. ELEXIKO used corpus tools already available at the Leibniz Institute for the German Language (IDS).

A lexicographic editing system was developed in-house in 2007/2008 for the ANW and the system has been used since 2008 (Niestadt 2009; Tiberius et al. 2014). The INT's editing system is regularly improved on the basis of new insights provided not only by lexicographers but also by computational linguists and programmers. It is also used for other projects at the institute as well as for a dictionary project at the Frisian Academy. The ELEXIKO project did not use a dictionary writing system in the strict sense, but instead an existing XML editor (OXYGEN[5]) was adapted to the needs of the project. The edi-

[5] OXYGEN is a commercial XML editor: http://www.oxygenxml.com/.

tor was supplemented with additional software, namely an in-house linking tool called "Vernetziko" (cf. Meyer 2011) and an administrative tool.

Backing up data and versioning are important in the computerisation phase. For example, the Norwegian dictionary project NORSK ORDBOK established that if the daily backup of data did not work, this basically meant losing data that would take six person-weeks to compile (Grønvik/Smith Ore 2013: 255). In the ANW, Git[6] software is used for versioning the source code of the editing system and the web application. The dictionary entries are stored in a MySQL database. Backups are made every day that are stored for three months. Each quarter a backup is made that is stored for a year, and once a year a backup is made that is stored for ten years. This all takes place in order to avoid delays in the project.

Data preparation phase

As a rule, the first task that has to be completed in the data preparation phase is compiling a list of potential lemmas. This is a semi-automatic task as often lexicographers have to manually check the candidates. Although the possibilities for automatically extracting lexical data from corpora have improved significantly in the last decades through improvements in language technology (→ Chapter 6), the ANW still is and ELEXIKO was always relatively conservative when it came to including automatically extracted data in the dictionary, and their corpus data still are/were predominantly analysed manually (although, of course, the analysis is computer aided). However, both projects contain automatically extracted data. The ANW contains data that is dynamically derived from other lexical resources, such as information on orthographic form from the official guidelines on Dutch spelling. In the ELEXIKO project, for example, three examples identified in the ELEXIKO corpus are displayed for all the lemmas that were not edited before the project came to an end.

Data evaluation phase

Originally, the ANW had a larger lexicographic team at its disposal than the ELEXIKO project. The ANW had five full-time lexicographers until mid-2015 in addition to three full-time lexicographic assistants, an editor-in-chief, and a project manager. This has now been reduced to one full-time lexicographer assisted by two lexicographers who dedicate a small amount of their time to the ANW. Four full-time lexicographers (at one time, five) were employed on the ELEXIKO project (including the editor-in-chief). Sometimes, students work(ed) as interns with the ANW or ELEXIKO.

[6] Git is a free, open source tool for version control: http://git-scm.com.

The ANW had a more complex workflow in the data evaluation phase than ELEXIKO since the lexicographers and lexicographic assistants worked closely together as can be seen in → Fig. 3.4. On the ELEXIKO project there was no support from lexicographic assistants, and a lexicographer edited the complete entry for any word. The situation is now the same for the ANW.

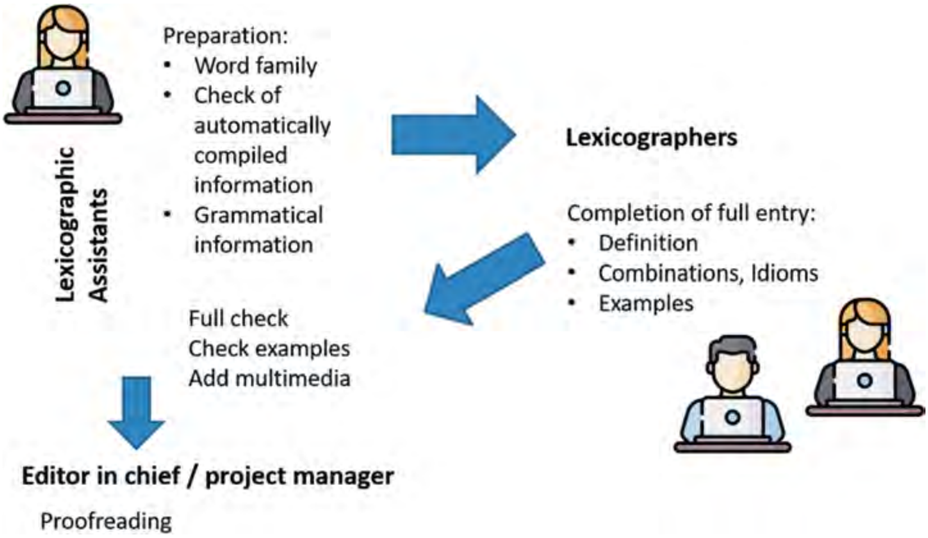


Fig. 3.4: The original ANW workflow.[7]

In this phase, the work of the lexicographers is not fundamentally different from the data evaluation phase in a print dictionary. However, in all cases the analysis is computer aided; that is, functions such as concordancing, filtering, and sorting make it possible for lexicographers to analyse large amounts of data efficiently.

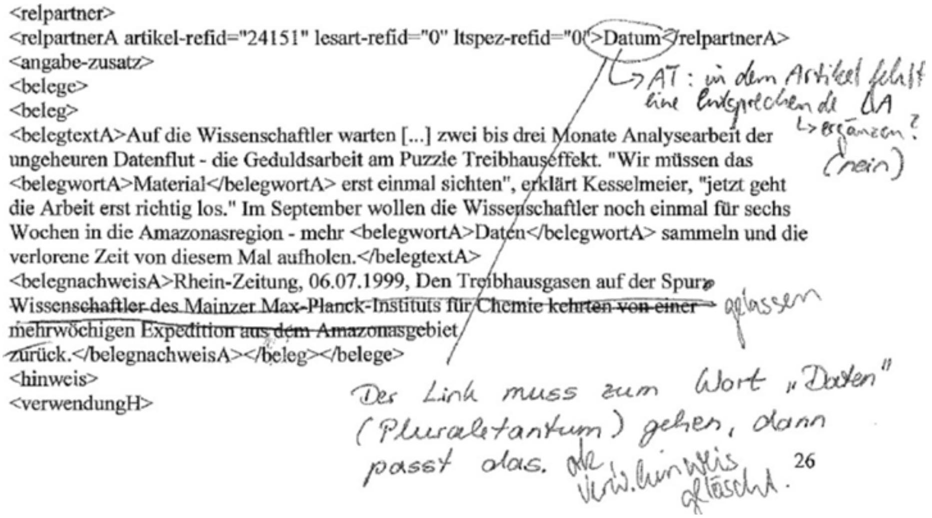
Preparation phase for online release

In this pre-final phase the dictionary entries are proofread. In the early years of the ANW, a new online version of the dictionary was compiled every three months. The editor-in-chief and project director would do the proofreading after which the articles would go back to the lexicographers for final revisions. As soon as that was done, the editor-in-chief and project director would check the entries again and then change the status of the dictionary entry to “going online”. Spelling and hyperlinks were

[7] Icons designed by Freepik: https://www.freepik.com/.

checked automatically and then corrected manually. After that, a new version of the dictionary was uploaded into a test environment for one week. Errors and inconsistencies, etc. could still be corrected. If, after a week, the version in the test environment was approved, an updated version of the dictionary was published on the public website. Nowadays, the ANW is much more dynamic, and updates are done overnight so that changes to the data are visible online the very next day.

In ELEXIKO, all entries first underwent a double reading for content (by the editor-in-chief and a different lexicographer than the one who compiled the entry, as seen from the different handwriting in → Fig. 3.5) followed by formal proofreading (e.g. for orthographic errors). All the revisions were checked again, and generally all the hyperlinks in the entry were checked, illustrations were opened to test them, and audio data played, etc. before the finished article was published online overnight.



<relpartner>
<relpartnerA artikel-refid="24151" lesart-refid="0" ltspez-refid="0">Datum</relpartnerA>
<angabe-zusatz>
<belege>
<beleg>
<belegtextA>Auf die Wissenschaftler warten [...] zwei bis drei Monate Analysearbeit der ungeheuren Datenflut - die Geduldsarbeit am Puzzle Treibhauseffekt. "Wir müssen das <belegwortA>Material</belegwortA> erst einmal sichten", erklärt Kesselmeier, "jetzt geht die Arbeit erst richtig los." Im September wollen die Wissenschaftler noch einmal für sechs Wochen in die Amazonasregion - mehr <belegwortA>Daten</belegwortA> sammeln und die verlorene Zeit von diesem Mal aufholen.</belegtextA>
<belegnachweisA>Rhein-Zeitung, 06.07.1999, Den Treibhausgasen auf der Spur.
~~Wissenschaftler des Mainzer Max-Planck-Instituts für Chemie kehrten von einer~~
~~mehrwöchigen Expedition aus dem Amazonasgebiet~~
~~zurück.~~</belegnachweisA></beleg></belege>
<hinweis>
<verwendungH>

↳ AT: in dem Artikel fehlt eine entsprechende LA
↳ ergänzen? (nein)

gelassen

Der Link muss zum Wort „Daten" (Pluraletantum) gehen, dann passt das. die Verw.hinweis gelöscht.

26

Fig. 3.5: Extract from the double content-editing of a manuscript version (with XML markup) of the entry for *Material* ‘material’ in ELEXIKO (excerpt from paradigmatic relationships with corrections hinting at wrong linking).

Before a dictionary is first published on the web, the outer texts for the dictionary also need to be written like user instructions, information about the content, and information about the corpus, etc. For the ANW, these dictionary outer texts have remained largely unchanged since first published in 2009, and are only changed when necessary to reflect the current status of the project. For ELEXIKO, these texts had been online since 2003 in their initial, very brief form before being completely revised and extended in 2007.

Maintenance and preservation phase

Because ELEXIKO is part of the dictionary portal OWID at the Leibniz Institute for the German Language (IDS), all technical tasks of maintenance and preservation are carried out in this context. However, at different times, editors had to update entries, for example in 2011, when a new version of the official German orthographic rules was released. These comprised new rules for the syllabification of loanwords so that this information in many entries in ELEXIKO had to be changed accordingly.

Maintenance of the ANW is an integral part of the workflow of the project to ensure that the online application is continuously up and running. To this end software updates are carried out regularly and the data is updated on a daily basis.

Summary

As we noted at the outset, work on the ANW and the ELEXIKO project began without having a complete overview of the digital lexicographic process necessary to realise the projects successfully. As a result, important planning stages were overlooked. Because both dictionaries were compiled completely from scratch and were conceived exclusively for the Internet medium, the necessary experience could only be acquired in practice in order to ultimately set out the lexicographic process in full. In doing so, it was possible to translate existing experience from print lexicography.

3.4.2 The lexicographic process for the German WIKTIONARY

In this section we consider the digital lexicographic process for another type of dictionary under construction, namely for WIKTIONARY, the collaboratively compiled online dictionary. WIKTIONARY is a freely available, multilingual dictionary for the vocabulary of different languages, the content of which is compiled collaboratively. Meyer/Gurevych (2016) describe the lexicographic process for the German WIKTIONARY and compare it to the lexicographic process for dictionaries compiled by an editorial team. In their study, they concluded that the phases recognisable from print dictionaries are only replicated to a small extent in WIKTIONARY. The data acquisition and data preparation phases merge strongly with the data evaluation phase while the data evaluation phase is almost impossible to distinguish from the preparation for online publication phase since the markup language is automatically translated into fully formatted dictionary entries.

Another important difference is that the lexicographic process for WIKTIONARY is shaped strongly by revisions and by discussion. The writing process for entries in the dictionary is based on multiple revisions by different authors. The revisions include additions, more specific details, re-formulations, corrections of errors, the integration of examples and sources, and deletions of what some editors believe is irrelevant material. There are special discussion pages relating to the overall concept of the dictionary as well as to individual entries that can be used for preparing, assessing, and implementing changes and that represent an important tool for tracking changes. Meyer/Gurevych (2016) have proposed a new model to better describe the lexicographic process for WIKTIONARY (→ Fig. 3.6): in this model, different collaborating users take already part in the preparatory phase with the conceptualisation of the dictionary. In the compilation phase, articles are written, and the lexicographic instructions are continuously updated. Entries and instructions alike are discussed by users, which may lead to revisions of entries. Revised entries can be the topic of user discussions as well. Thus, collaborating users are part of every step in WIKTIONARY'S lexicographic process.

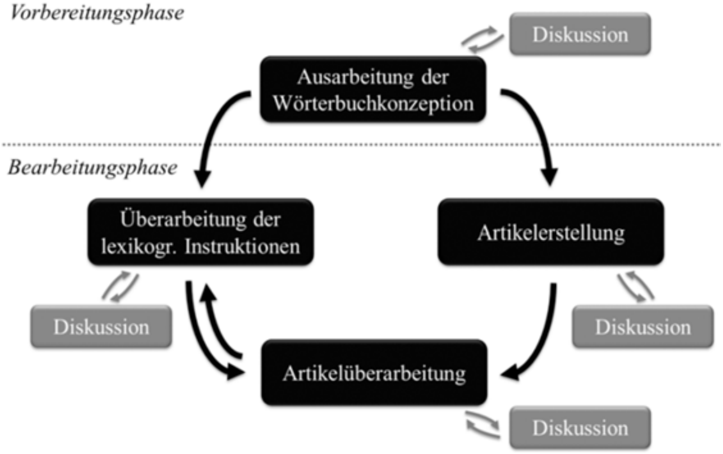


Fig. 3.6: Process model for WIKTIONARY (following Meyer/Gurevych 2016: 67); top: preparatory phase: development of the dictionary concept with discussion; bottom: editing phase: revision of lexicographic instructions (left) with discussion, compiling entries (right) with discussion, revision of entries (bottom) with discussion.

3.4.3 The lexicographic process for the ORDBOG OVER DET DANSKE SPROG

Finally, we consider the lexicographic process for the ORDBOG OVER DET DANSKE SPROG (ODS), the retrospectively digitised Danish language dictionary. The ODS is a historical dictionary comparable to the great national dictionaries like Jacob and Wilhelm Grimm's DEUTSCHES WÖRTERBUCH (DWB), the WOORDENBOEK DER NEDERLANDSCHE TAAL (WNT), the OXFORD ENGLISH DICTIONARY (OED), and the SVENSKA AKADEMIENS ORDBOK (SAOB). It was originally published between 1918 and 1956 in 28 volumes and has since been extended with five additional volumes (1992 to 2005). There has been a digital version since November 2005 available at ORDNET.DK, a portal of dictionaries for the Danish language.

The digitisation of the ODS began relatively late compared with other projects. The SAOB began being digitised as early as 1983, the OED and WNT in 1984, and the DWB between 1998 and 2003/4. As a result, the ODS was able to benefit from the digitisation experiences of other projects. The process began with a preparatory phase in 2004 and was supposed to last for five years. In broad terms, two main tasks could be distinguished in the subsequent data preparation phase: raw digitisation and structural markup.

The first step in the digitisation process was carried out in collaboration with the Center for Digital Humanities at the University of Trier, which was also responsible for digitising the DWB. Dual text capture without proofreading was the method used for the raw digitisation, that is, the print version of the dictionary was typed up by two people (in Asia). In order to achieve good results, a handbook was drawn up to ensure consistent coding for special characters and symbols. After the text capture, the two versions were compared automatically in Trier, and a list of differences was generated, checked, and corrected. This process of text capture took nine months.

In the second stage, the digitised version was marked up structurally using scripts. First, a rough markup was created in which only lemmas, homograph numbers (if present), and word classes were explicitly indicated as such. This version took around two years to complete and appeared online in November 2005. The markup was subsequently refined in order to identify other units of information in the microstructure of the dictionary text that could be derived from typographical features (examples, for instance, are always coded in italics). In addition, the supplementary volumes were to be integrated into the digitised dictionary. For this, the project returned to the data preparation phase following the online release of a first stage of data processing.

In contrast to the original plan to complete the retro-digitisation within five and a half years with the online release of the ODS, it has to be recognised that, although an end point for the necessary processes is theoretically possible, it has not been achieved in practice. This is undoubtedly the result of the Internet as a medium and its general capacity to enable quick corrections and additions to websites. We can also see from this example that further optimisation of a dictionary like the ODS – and also, of course, other types of Internet dictionary – can, in principle be undertaken, ad infinitum so that the digital lexicographic process in this kind of scenario is never complete.

3.5 Software to support lexicographic processes

Nowadays, lexicographic work is characterised by increasing automated support (cf., among others, Abel 2022, Abel/Klosa 2012, Rundell 2023, Rundell/Kilgarriff 2011). Modern lexicographic work is inconceivable without two tools, namely a dictionary writing system and a corpus query system.

3.5.1 Dictionary writing systems

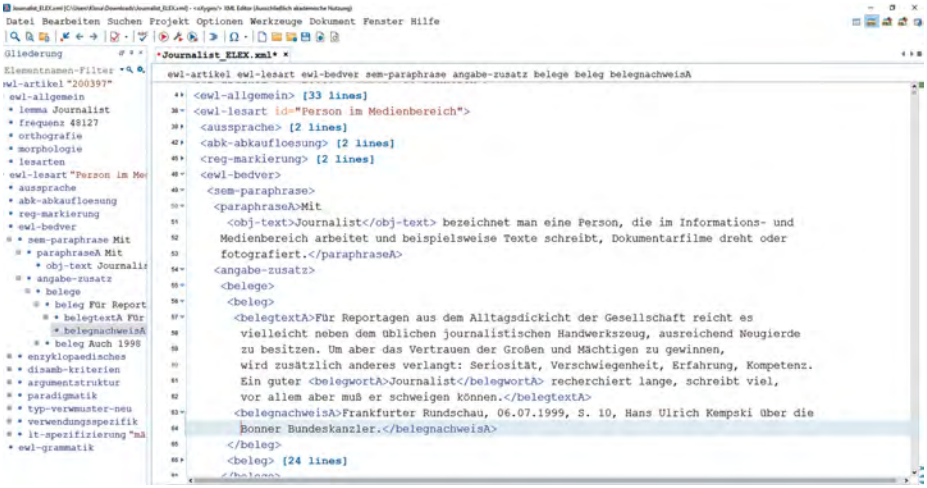
A dictionary writing system is a software application that facilitates the lexicographic process and, preferably, optimises and streamlines it. It allows lexicographers to write a dictionary entry (data evaluation phase) and it makes both project management and publication easier (preparation phase for online release). A dictionary writing system usually has three components:

- a text-editing interface that allows the lexicographers to edit dictionary entries – this can be either a WYSISWYG ('what-you-see-is-what-you-get') view or simply an XML view (→ Fig. 3.7);
- a database to keep the data secure; relational databases like ORACLE, MySQL, and PostgreSQL as well as native XML databases are often used in a lexicographic context;
- a series of administration tools for project management and publication.

A dictionary writing system can be developed in house during the computerisation phase of a particular project (as is often the case for academic dictionaries such as the ANW), or a commercial system can be acquired. There is also the possibility to work with open-source systems, e.g. LEXONYMY (cf. Měchura 2017; Váradi et al. 2022).

The advantage of an in-house system is that it is easier to adapt it to the lexicographic process of a particular project and improvements can be made in-house, if necessary. However, this assumes that the project has the necessary human and financial resources at its disposal. The advantage of a commercial system is that many users will have contributed collectively to improving the system, which allows for the rapid development of new functions (such as a component added in 2023 to the system TLex that facilitates collaboration with ChatGPT, cf. de Schryver 2023: 2). Open-source systems also profit from the feedback that its many users give to the developers. Nonetheless, in commercial and open-source systems, the project is tied to the available data model (cf. Tiberius/Niestadt/Schoonheim 2014 and Abel 2022 for further discussion of the advantages and disadvantages of all options).

The more complex the administration and publication tools are that are offered by the system, the greater the control that the editors have over the lexicographic process. In this respect, the dictionary writing system developed in house for the *Norsk Ordbok* (NOB) offers an interesting feature. It makes it possible to monitor how much



```
<ewl-allgemein> [33 lines]
<ewl-lesart id="Person im Medienbereich">
 <aussprache> [2 lines]
 <abk-abkaufloesung> [2 lines]
 <reg-markierung> [2 lines]
 <ewl-bedver>
  <sem-paraphrase>
   <paraphraseA>Mit
     <obj-text>Journalist</obj-text> bezeichnet man eine Person, die im Informations- und
    Medienbereich arbeitet und beispielsweise Texte schreibt, Dokumentarfilme dreht oder
    fotografiert.</paraphraseA>
   <angabe-zusatz>
    <belege>
     <beleg>
      <belegtextA>Für Reportagen aus dem Alltagsdickicht der Gesellschaft reicht es
       vielleicht neben dem üblichen journalistischen Handwerkszeug, ausreichend Neugierde
       zu besitzen. Um aber das Vertrauen der Großen und Mächtigen zu gewinnen,
       wird zusätzlich anderes verlangt: Seriosität, Verschwiegenheit, Erfahrung, Kompetenz.
       Ein guter <belegwortA>Journalist</belegwortA> recherchiert lange, schreibt viel,
       vor allem aber muß er schweigen können.</belegtextA>
      <belegnachweisA>Frankfurter Rundschau, 06.07.1999, S. 10, Hans Ulrich Kempski über die
       Bonner Bundeskanzler.</belegnachweisA>
     </beleg>
     <beleg> [24 lines]
    </belege>
```

Fig. 3.7: The XML editor OXYGEN, which was used in the ELEXIKO project to prepare dictionary entries (here an extract from the lemma *Journalist* ‘journalist’).

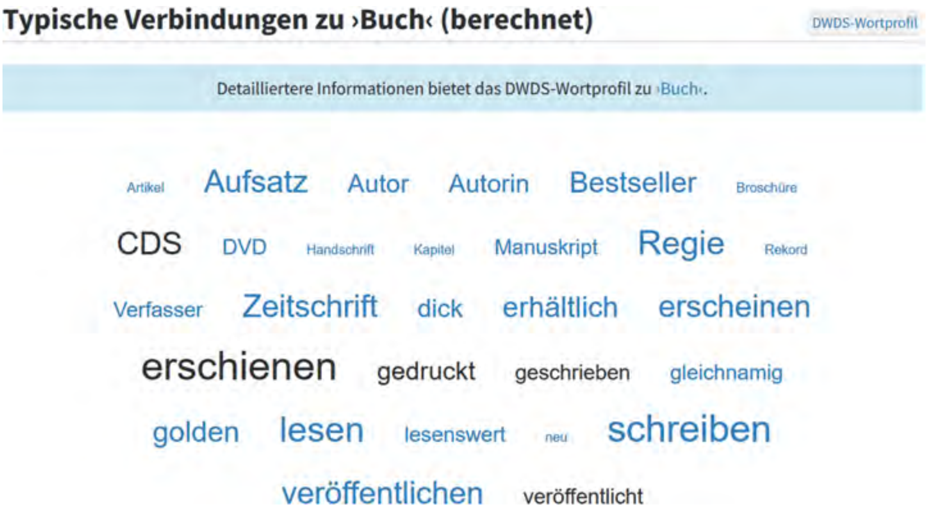
text should be written for an article in order to maintain an appropriate distribution of article lengths in a particular alphabetical category. When a new word is selected for inclusion, a maximum length for the entry is suggested, based on the amount of data that is available at that point in time for compiling the whole dictionary. While the entry is being written and edited, the actual length of the entry is constantly compared with the maximum length that has been calculated so that the lexicographers can see precisely whether they have remained within the proposed length of the entry (Grønvik/Smith Ore 2013: 254).

3.5.2 Corpus query systems

Corpus query systems are used in the data preparation and analysis phase of the digital lexicographic process. Corpus query systems are tools with which corpus texts can be queried in ways that are linguistically relevant. Lexicographers are probably the most demanding corpus users – they need and regularly use the highest number of features in corpus tools. In fact, several features of corpus tools were originally designed especially for lexicographic purposes, only to be found useful by linguists, teachers, and others as well (Tiberius et al. 2022). SKETCH ENGINE is an example of a corpus query system which is often used in lexicographic projects (Kilgarriff et al. 2004), but there are many other corpus tools available (a comprehensive up-to-date list of these tools can be found at https://corpus-analysis.com/tag/concordancer).

The basic functionality that a corpus query system provides to support lexicography is (KWIC) concordancing, which displays all the occurrences of a keyword found in the corpus with around 20 words of surrounding context. KWIC stands for "Keyword in Context" and refers to the concordance being displayed with the keyword in the middle of the screen. Most corpus query systems make it possible to sort and filter concordance lines and also, when necessary, to display more context.

Most systems also support a variety of powerful search queries for the lemma itself, a particular word form, or a phrase (in combination with a particular word class), right down to searching for all of the occurrences of a word in a specific lexical context. For example, it is possible to search for all occurrences of the verb *to speak* that can be found in a range of five words before or after the keyword *language* (as in *Which languages do you speak?*). A further feature that is particularly useful for lexicography (and that is available in advanced corpus tools) is the so-called lexical profile. A lexical profile is a statistical overview of the most important facts about a word and its customary combinations with other words (Atkins/Rundell 2008: 109). The word sketches in SKETCH ENGINE are a type of lexical profile, providing collocate lists per grammatical relation. In the German DWDS, it is not only lexicographers who use the "Wortprofil" ('word profile', its name for lexical profiles), but also users who are able to display the word profile of a word as part of the dictionary entry (→ Fig. 3.8).



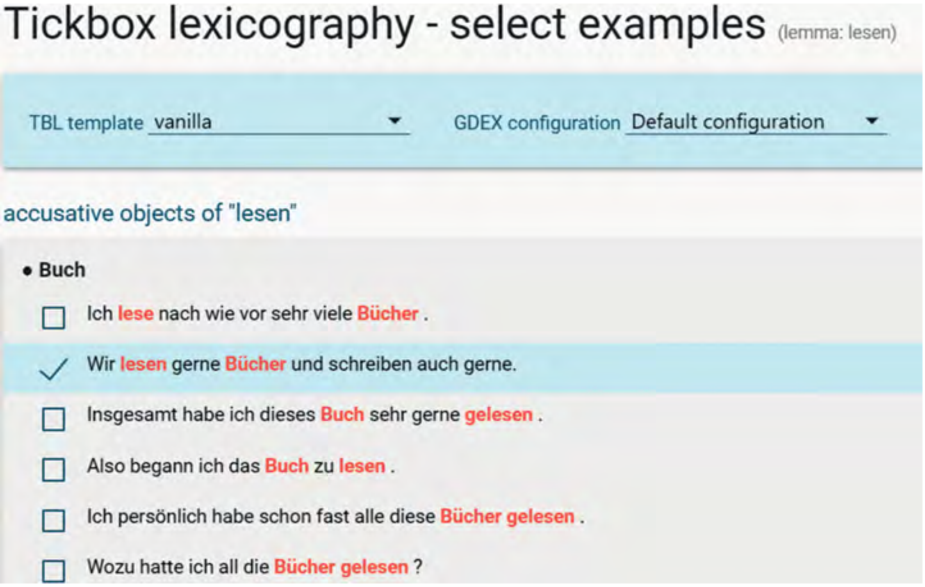
Typische Verbindungen zu ›Buch‹ (berechnet)

Detailliertere Informationen bietet das DWDS-Wortprofil zu ›Buch‹.

Fig. 3.8: Word profile for the word *Buch* 'book' in the DWDS with collocates such as *lesen* 'read', *schreiben* 'write', *veröffentlichen* 'publish' or *dick* 'fat', *neu* 'new', etc.

Most functions in a corpus query system only work if the corpus data have been correctly processed in the computerisation phase. Processing corpus texts involves two steps: preparing the metadata and preparing the texts. The metadata contains information about the source such as the author, date, genre, and domain, which enables lexicographers to assign labels like "primarily spoken language" to particular lemmas with greater confidence. Preparing the texts means adding linguistic annotations to the raw corpus texts, i.e. lemmatisation, tagging (annotation according to word class), or parsing (annotation according to syntactic structures) (→ Chapter 5).

Corpus software is improved and updated regularly on the basis of new insights and requirements that come from users. *Good Dictionary Examples* (GDEX), *Tickbox Lexicography* (TBL), *one-click copying*, *trend* analysis, and *word sense induction* in the word sketch are examples of this kind of functionality in SKETCH ENGINE which were added over time. GDEX is a function that seeks to automatically sort sentences in a concordance according to their usability as an example sentence in the dictionary based on a set of quantifiable heuristics, such as sentence length, frequency of words, and lists of vulgar words (Kilgarriff et al. 2008). In this way, the best example sentences will appear at the top of the concordance list, and these are the ones that lexicographers see first. TBL is a function through which lexicographers are able to select examples from a list of (good) candidates and export them directly into the editing system (→ Fig. 3.9).


Tickbox lexicography - select examples (lemma: lesen)

TBL template vanilla ▾ GDEX configuration Default configuration ▾

accusative objects of "lesen"

- Buch

□ Ich lese nach wie vor sehr viele Bücher .
✓ Wir lesen gerne Bücher und schreiben auch gerne.
□ Insgesamt habe ich dieses Buch sehr gerne gelesen .
□ Also begann ich das Buch zu lesen .
□ Ich persönlich habe schon fast alle diese Bücher gelesen .
□ Wozu hatte ich all die Bücher gelesen ?

Fig. 3.9: Example of TickBox lexicography (for the verb *lesen* 'to read') in SKETCH ENGINE.

One-click copying is another useful feature for lexicographers. It supports easy transfer of data from the corpus query system into the editing system. Not only is the concordance line copied, but also the whole sentence (and potentially its metadata) is transferred onto the clipboard to be pasted into the editing system. *Trends* is a feature for detecting words that undergo changes in the frequency of use in time. It can be used to identify words whose use increases or decreases over time. *Word sense induction* is a functionality that has recently been added to the word sketch tool and that identifies word senses automatically. This function categorises the collocations identified by a word sketch into groups corresponding to the different senses of a word.

Although it would be beneficial to have the dictionary writing system and the corpus query system integrated in one tool, they are often separate tools (although they are able to communicate with one another, as Sketch Engine and the editing system used in the ANW project). There are only a few examples of systems in which the editing system and corpus query system are integrated in a single tool e.g. the TLex system (Joffe/de Schryver 2004).

Choosing the most appropriate software tool to use for a new dictionary project may not be straightforward as there are many different dictionary writing systems and even more corpus query systems. See Kallas et al. (2019) for a list of systems that were mentioned in the context of the surveys that were carried out within the Elexis project to obtain an overview of existing lexicographic practices across Europe. These surveys also showed that in-house solutions are still very common for dictionary writing systems whereas for corpus query systems, commercial systems tend to be used most. The fundamental consideration in the choice of software is whether or not it meets all of the necessary requirements of the project. Basic aspects like price and the availability of academic licences can also play a role in choosing a system. Particularly in the planning phase, a thorough analysis of the available software is important in close collaboration between the lexicographic team and the IT experts in order to evaluate which work tasks can be carried out with existing software and which technical developments can be implemented in house under certain circumstances. Unfortunately, this kind of analysis is often absent, even now, in the planning stages of many dictionary projects.

3.6 The lexicographic process for dictionary portals

In this section, we consider the lexicographic process for dictionary portals (→ Chapter 2.4). Engelberg/Müller-Spitzer (2013) define a dictionary portal as a data structure that is represented as a website or a series of interlinked pages, that provides access to a series of electronic dictionaries, and where these dictionaries can be consulted as independent products. They distinguish between three main types that they refer to as dictionary collections, dictionary search engines, and dictionary networks.

Dictionary collections only give external access to the dictionaries in the portal. This means that they consist of links to the home pages of the dictionaries in the portal, and as a rule these dictionaries are not owned by the institution that runs the portal. The Erlanger Liste is one such dictionary collection.

Dictionary search engines are a little more sophisticated. They provide the option of searching all of the lemmas in the integrated dictionaries. However, the data in the various dictionaries are not interconnected with links. No dictionary content is presented within the portal, and the owners of the portal and the dictionaries are not usually the same. The website ONELOOK and the EUROPEAN DICTIONARY PORTAL are examples of this kind of dictionary search engine.

Dictionary networks go one step further and make it possible for users to search for particular information in the entries in the dictionaries that are contained in the portal. Examples of a portal like this are OWID and the TRIERER WÖRTERBUCHNETZ.

It is clear that the lexicographic processes for these three types of dictionary portal differ from one another. The lexicographic process for dictionary collections is the simplest. There is a preparatory phase, a data acquisition phase, possibly a short computerisation phase to set up the website, and a preparation phase for online release that involves writing the outer texts (what the portal is and what it is not). For a dictionary collection, the phases follow one another in a more or less linear way. First, it must be decided which dictionaries are to be included, then the website must be planned, and finally the portal has to be made available and maintained online. Once the dictionary collection is online, there are very few changes to make.

The lexicographic process for dictionary search engines is similar to the process for dictionary collections. However, the computerisation phase is more elaborate since a combined list of lemmas has to be compiled for all of the dictionaries in the portal, a search function implemented, and the data indexed. By clicking on a link, users move into an individual dictionary and leave the portal.

Dictionary networks have the most complex lexicographic process among dictionary portals. During the preparatory phase, decisions have to be taken about which dictionaries should be included in the portal and about the network connections between the dictionaries in the network. The dictionaries that are integrated in the network may be existing published dictionaries or dictionaries under construction. This influences the tasks that have to be completed in the data preparation and data evaluation phases. Here the computerisation phase involves interlinking the search engines and website. The preparation phase for online release involves writing the user guide.

Generally, consideration must be given as to whether we can refer to a digital lexicographic process at all in relation to dictionary portals and networks. Wiegand (1998: 134) defines such a process as being carried out so that one particular dictionary is created. However, compiling dictionary portals and networks does not involve a dictionary coming into existence but rather a website through which a variety of dictionaries can be accessed. Central tasks in the digital lexicographic process, e.g. corpus construction (data acquisition phase), processing corpus texts (computerisation phase), compiling a list of lemmas (data processing phase), and in particular writing dictionary entries (data evaluation phase) are absent here. Nevertheless, creating dictionary search engines and networks does involve tasks that require lexicographic competences, e.g. combining different lists of lemmas in an overarching access structure or conceiving and realising the linking between the individual contents of a dictionary network.

In this sense we tend towards extending Wiegand's (1998) definition: according to our understanding, a digital lexicographic process is the collection of activities belonging to the process that have to be undertaken so that a particular dictionary or a group of dictionaries comes into being on or in an electronic carrier medium. This process can be divided into different phases in which computers are used consistently.

3.7 The lexicographic process of centralised lexicographic databases

The production of dictionaries is a meticulous and detail-oriented task that requires a great deal of time, effort, and expertise. In the light of shorter project running times as well as the changing role of lexicographic institutions and publishing houses (cf. Tiberius et al. 2023), which are becoming more of a data provider and less of a dictionary publisher, dictionary projects and publishers have started to move away from the production of stand-alone Internet dictionaries towards centralised lexicographic knowledge bases from which different end products (e.g. Internet dictionaries, dictionary apps, data for NLP tools) can be derived (for example, Duden dictionaries, cf. Alexa et al. 2002, or lexicographic institutions such as the Estonian Language Institute, cf. Tavast et al. 2018).

Having a single pool of data has several practical advantages. It supports the reusability of the lexicographic data for different end products, both dictionaries and language technology applications. It avoids having multiple dictionaries with duplicated and/or conflicting information and helps to minimise redundancies. Furthermore, it ensures consistency, which again leads to more efficient maintenance of the lexicographic data in a homogenous manner. On the other side of the coin, having a single pool of data also presents new challenges for lexicography. Integrating all lexicographic data into one centralised database naturally results in a more complex data model (cf. e.g. Tavast et al. 2018, Depuydt et al. 2019, and Gantar 2020 for the challenges faced at different institutions when creating a centralised lexicographic database).

In this new constellation, there seems to be not just one lexicographic process but at least two: one for the centralised lexicographic database and one for each end product that is compiled out of the data. The compilation of a central database is by default media-neutral, and requires a media-neutral lexicographic process in which all of the phases outlined in → Tab. 3.2 are still relevant, except for the phase of preparation for online release as publication is expected to occur in the process for individual end products. For the creation of the individual end products, a subset of the phases is needed. In particular, for each individual end product there will be a preparatory phase (to develop a concept for a specific end product), a preparation phase for online release, and a maintenance and preservation phase.

Furthermore, it should be noted that compiling data for a centralised lexicographic database changes the traditional organisation of the work: instead of editing per headword, generic content creation requires more task-based editing (e.g. editing morphological information for a set of headwords in the database). This requires a turnaround in the way of working for lexicographers. New editing tools are also needed to support this task-based editing, allowing different views on the data and supporting a more relational encoding of lexicographic data (cf. Měchura et al. 2023, Meyer/Eppinger 2018). Finally, it remains to be seen how much work is needed to customise the content from such centralised lexicographic knowledge bases for specific end products.

In the context of recent developments, we tend towards extending Wiegand's (1998) definition of the lexicographic process to cover lexicographic databases (instead of dictionaries alone) to better reflect the digital aspects: a (completed) lexicographic process is the set of process activities carried out to create a particular lexicographic database. Lexicographic databases can be used in language technology, for creating dictionaries for human users, or can be integrated in a dictionary portal.

Bibliography

Further reading

Abel, Andrea (2022): Dictionary writing systems. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45369-4_111-1. *An overview of lexicographic editing systems.*

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

Müller-Spitzer, Carolin (2003): Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: Lexicographica 19, 140–168. *A continuation and development of H.E. Wiegand's thinking for electronic dictionaries*.

Tiberius, Carole et al. (2023): *A Lexicographic Practice Map of Europe.* In: International Journal of Lexicography. https://doi.org/10.1093/ijl/ecad023. *The results of various surveys that give an insight into lexicographic processes*.

Wiegand, Herbert Ernst (1998): *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1. Teilband. Berlin/New York: De Gruyter. *Chapter "1.5. Computer, wissenschaftliche Lexikographie und Wörterbuchforschung" (pp. 133–246) provides the first analysis and description of digital lexicographic processes*.

Literature

Academic literature

Abel, Andrea (2022): Dictionary writing systems. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45369-4_111-1 [last access: April 26, 2024]

Abel, Andrea/Klosa, Annette (2012): Der lexikographische Arbeitsplatz – Theorie und Praxis. In: Fjeld, Ruth Vatvedt/Torjusen, Julie Matilde (eds.): *Proceedings of the 15th EURALEX International Congress in Oslo 2012*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 413–421.

Abel, Andrea/Klosa, Annette (2014): *Einleitung: "Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess"*. Mannheim: Institut für Deutsche Sprache, 3–8.

Alexa, Melina, et al. (2002): The Duden Ontology: An Integrated Representation of Lexical and Ontological Information. In: *LREC Workshop on WordNet Structures and Standardisation, and How These Affect Wordnet Applications and Evaluation 2002*. Las Palmas, Gran Canaria. https://publica.fraunhofer.de/entities/publication/efe658dc-eda0-4915-ab40-1aeafb79c2c6/details [last access: April 26, 2024].

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

Culy, Christopher/Lyding, Verena (2010): Double Tree: An Advanced KWIC Visualization for Expert Users. In: *Information Visualization, Proceedings of IV 2010, 2010 14th International Conference Information Visualization*, 98–103.

Depuydt, Katrien/Schoonheim, Tanneke/de Does, Jesse: *Towards a More Efficient Workflow for the Lexical Description of the Dutch Language*. http://videolectures.net/elexisconference2019_depuydt_dutch_language/ [last access: April 26, 2024].

de Schryver, Gilles-Maurice (2023): Generative AI and Lexicography: The Current State of the Art Using ChatGPT. In: *International Journal of Lexicography* 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021 [last access: April 26, 2024].

de Schryver, Gilles-Maurice/Prinsloo, Daniel Jacobus (2000a): The Concept of "Simultaneous Feedback": Towards a New Methodology for Compiling Dictionaries. In: *Lexikos* 10, 1–31.

de Schryver, Gilles-Maurice/Prinsloo, Daniel Jacobus (2000b): Dictionary Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In: Heid, Ulrich, et al. (eds.): *Proceedings of the Ninth EURALEX International Congress*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, 197–209.

Dubois, Claude (1990): Considérations générales sur l'organisation du travail lexicographique. In: Hausmann, Franz-Josef, et al. (eds.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: De Gruyter, 1574–1588.

Engelberg, Stefan/Müller-Spitzer, Carolin (2013): Dictionary Portals. In: Gouws, Rufus H., et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Computational Lexicography*. Berlin/Boston: De Gruyter, 1023–1035.

Gantar, Polona (2020): *Dictionary of Modern Slovene: From Slovene Lexical Database to Digital Dictionary Database*. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 46 (2), 589–602. https://doi.org/10.31724/rihjj.46.2.7 [last access: April 26, 2024].

Gouws, Rufus H. (2011): Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In: Fuertes-Olivera, Pedro Antonio/Bergenholtz, Henning (eds.): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 17–29.

Grønvik, Oddrun/Smith Ore, Christian-Emil (2013): What should the electronic dictionary do for you – and how? In: Kosem, Iztok, et al. (eds.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, *17–19 October 2013*, *Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 243–260.

Jakubíček, Miloš (2017): The advent of post-editing lexicography. In: *Kernerman Dictionary News*, 14–15.

Joffe, David/de Schryver, Gilles-Maurice (2004): TshwaneLex: A State-of-the-Art Dictionary Compilation Program. In: Williams, Geoffrey/Vessier, Sandra (eds.): *Proceedings of Eleventh EURALEX International Conference*. Le Paquebot: Lorient, 99–104.

Kallas, Jelena et al. (2019): *ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs*. https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf [last access: April 26, 2024]

Kilgarriff, Adam (2008): GDEX: Automatically finding good dictionary examples in a corpus. In: Bernal, Elisande/De Cesaris, Janet (eds.): *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, 425–432.

Kilgarriff, Adam, et al. (2004): The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra (eds.): *Proceedings of Eleventh EURALEX International Conference*. Le Paquebot: Lorient, 105–116.

Klosa, Annette (2013): The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H., et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Deve lopments with Focus on Computational Lexicography*. Berlin/Boston: De Gruyter, 501–508.

Knowles, Francis E. (1990): The Computer in Lexicography. In: Hausmann, Franz Josef, et al. (eds.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin/New York: De Gruyter, 1645–1672.

Landau, Sydney (1984): *Dictionaries. The Art and Craft of Lexicography*. New York: Scribner's.

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

Měchura, Michal (2017): Introducing Lexonomy: an open-source dictionary writing and publishing system. In: Kozem, Iztok, et al. (eds.): *Electronic lexicography in the 21st century: Proceedings of the eLex 2017 conference*. Brno: Lexical Computing, 662–679.

Měchura, Michal, et al. (2023): *Relations, relations everywhere: an introduction to the DMLex data model. Video presentation at eLex conference 2023, Brno, Czech Republic, 27–29 June 2023*. https://www.youtube.com/watch?v=b6Lkv-3D5C0 [last access: April 26, 2024].

Meyer, Peter (2011): 'vernetziko': A Cross-Reference Management Tool for the Lexicographer's Workbench. In: Kosem, Iztok/Kosem, Karmen (eds.): *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, Slovenia, 10.–12. November 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 191–198.

Meyer, Peter/Eppinger, Mirjam (2018): fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In: Čibej, Jaka et al. (eds.): *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018*. Ljubljana University Press, Faculty of Arts, 1017–1022.

Meyer, Christian M./Gurevych, Iryna (2016): Der lexikografische Prozess im deutschen Wiktionary. In: Hildenbrandt, Vera/Klosa, Annette (eds.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache, 55–68.

Müller-Spitzer, Carolin (2013): Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: *Lexicographica* 19, 140–168.

Niestadt, Jan (2019): De ANW-artikeleditor: software als strategie. In: Beijk, Egbert, et al. (eds.): *Fons verborum. Feestbundel voor prof. dr. A.F.M.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie*. Leiden/Amsterdam: Instituut voor Nederlandse Lexicologie/Gopher BV, 215–222.

Riedel, Hans/Wille, Margit (1979): *Über die Erarbeitung von Lexika*. Leipzig: Bibliographisches Institut.

Rundell, Michael/Kilgarriff, Adam (2011): Automating the creation of dictionaries: where will it all end? In: Meunier, Fanny, et al. (eds.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: Benjamins, 257–281.

Rundell, Michael (2023): Automating the creation of dictionaries: are we nearly there? In: *Proceedings of Asialex 2023, June 22–24*. Seoul: Yonsei University Seoul, 9–17.

Schaeder, Burkhard (1987): *Germanistische Lexikographie*. Tübingen: Narr.

Schröder, Martin (1997): Brauchen wir ein neues Wörterbuchkartell? Zu den Perspektiven einer computerunterstützten Dialektlexikographie und eines Projektes "Deutsches Dialektwörterbuch". In: *Zeitschrift für Dialektologie und Linguistik* 64/1, 57–65.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, Herbert Ernst (ed.): *Wörterbücher in der Diskussion III*. Tübingen: Niemeyer, 106–131.

Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/Schröder, Bernd/Storrer, Angelika (eds.): *Chancen und Perspektiven computergestützter Lexikographie*. Tübingen: Niemeyer, 54–69.

Storrer, Angelika/Freese, Karin (1996): Wörterbücher im Internet. In: *Deutsche Sprache* 24, 97–136.

Svensén, Bo (2009): *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

Tavast, Arvi, et al. (2018): Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In: Čibej, Jaka, et al. (eds.): *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018*. Ljubljana University Press, Faculty of Arts, 749‒761.

Tiberius, Carole/Krek, Simon (2014): *Workflow of Corpus-Based Lexicography*. *Deliverable COST-ENeL-WG3 meeting*. https://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf [last access: April 26, 2024].

Tiberius, Carole/Niestadt, Jan (2015): Dictionary Use: A Case Study of the ANW dictionary. In: Tiberius, Carole/Müller-Spitzer, Carolin (eds.): *Research into dictionary use/Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache, 28–35.

Tiberius, Carole/Niestadt, Jan/Schoonheim, Tanneke (2014): The INL Dictionary Writing System. In: Kosem, Iztok/Rundell, Michael (eds.): *Slovenščina 2.0: Lexicography* 2:2, 72–93.

Váradi et al. (2022): Váradi, Tamás et al. (2022): *Lexonomy: Mastering the ELEXIS Dictionary Writing System. Version 1.1.0*. DARIAH-Campus [training module]. https://campus.dariah.eu/id/Qm_2SzS_rGB-Py9YTCHYm [last access: April 27, 2024].

Wiegand, Herbert Ernst (2018): *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband*. Berlin/New York: De Gruyter.

Zgusta, Ladislav (1971): *Manual of Lexicography*. The Hague/Paris: Mouton.

Dictionaries

ANW = *Algemeen Nederlands Woordenboek*. Leiden: Instituut voor Nederlandse Lexicologie. https://anw.ivdnt.org/ [last access: April 27, 2024].

DRWB = *Deutsches Rechtswörterbuch*. Heidelberg: Heidelberger Akademie der Wissenschaften. http://drw-www.adw.uni-heidelberg.de/drw/ [last access: April 27, 2024].

DWB = Deutsches Wörterbuch von Jacob und Wilhelm Grimm online. In: *Wörterbuchnetz des Trier Center for Digital Humanities/Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier*. http://woerterbuchnetz.de/DWB/ [last access: April 27, 2024].

DWDS = *Das Digitale Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de [last access: April 27, 2024].

elexiko = Online-Wörterbuch zur deutschen Gegenwartssprache. In: *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. http://www.elexiko.de [last access: April 27, 2024].

Erlanger Liste = *Lexika & Wörterbücher*. http://www.erlangerliste.de/ressourc/lex.html [last access: April 27, 2024].

NOB = *Norsk Ordbok*. Oslo: Universitetet i Oslo, Institutt for lingvistikke og nordiske studier. http://no2014.uio.no/l/ordbok/no2014.cgi [last access: April 27, 2024].

ODS = Ordbog over det Danske Sprog. In: *ordnet.dk*. *Dansk Sprog i Ordbøger og Korpus*. *Den Danske Sprog- og Litteraturselskab*. http://ordnet.dk/ods [last access: April 27, 2024].

OED = *Oxford English Dictionary online*. Oxford: Oxford University Press. http://dictionary.oed.com [last access: April 27, 2024].

OneLook = *OneLook Dictionary Search*. http://www.onelook.com [last access: April 27, 2024].

ordnet.dk = *Dansk Sprog i Ordbøger og Korpus. Den Danske Sprog- og Litteraturselskab*. http://ordnet.dk [last access: April 27, 2024].

OWID = *OWID Online-Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. Online: http://www.owid.de/ [last access: April 27, 2024].

SAOB = *Svenska Akademiens Ordbok*. Stockholm: Svenska Akademien. http://g3.spraakdata.gu.se/saob/index.html [last access: April 27, 2024].

Trierer Wörterbuchnetz = *Wörterbuchnetz des Trier Center for Digital Humanities/Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier*. http://woerterbuchnetz.de/ [last access: April 27, 2024].

Wiktionary = *Wiktionary, das freie Wörterbuch*. http://de.wiktionary.org/wiki/Wiktionary:Hauptseite [last access: April 27, 2024].

WNT = *Woordenboek der Nederlandsche Taal*. Leiden: Instituut voor Nederlandse Lexicologie. https://gtb.ivdnt.org/ [last access: April 27, 2024].

Internet sources

AGD = *Archiv für Gesprochenes Deutsch*. Mannheim: Institut für Deutsche Sprache. http://agd.ids-mannheim.de/index.shtml.

DeReKo = Deutsches Referenzkorpus. Mannheim: Institut für Deutsche Sprache. Online: http://www.ids-mannheim.de/kl/projekte/korpora.html 8last access: April 27, 2024].

DMLex = *OASIS Lexicographic Infrastructure Data Model and API (LEXIDMA)*. https://groups.oasis-open.org/communities/tc-community-home2?CommunityKey=0fd41fbb-72be-4771-8faf-018dc7d3f419 [last access: April 27, 2024].

Double Tree JS = *Double TreeJS: A compact view or words in context*. http://linguistics.chrisculy.net/lx/software/DoubleTreeJS/index.html [last access: April 27, 2024].

Elexis = *European lexicographic infrastructure*. https://elex.is/ [last access: April 27, 2024].

European Dictionary Portal = *European Dictionary Portal*. http://www.dictionaryportal.eu [last access: April 27, 2024].

Git = *Git Tool zur Versionenkontrolle*. https://git-scm.com [last access: April 27, 2024].

ISO LMF = *ISO Lexical Markup Framework*. https://www.iso.org/standard/68516.html [last access: April 27, 2024].

ISO 1951:2007 = *ISO 1951:2007. Presentation/representation of entries in dictionaries*. https://www.iso.org/standard/36609.html [last access: April 27, 2024].

Lexonymy = *Lexonymy – online, open-source platform for writing and publishing dictionaries*. https://www.lexonomy.eu/ [last access: April 27, 2024].

MySQL = *MySQL Open-Source-Datenbank*. https://www.mysql.de/ [last access: April 27, 2024].

Ontolex-Lemon = *Lexicon Model for Ontologies*. https://www.w3.org/2016/05/ontolex/ [last access: April 27, 2024].

ORACLE = *ORACLE Datenbank*. https://www.oracle.com/de/index.html [last access: April 27, 2024].

oXygen = *oXygen XML Editor*. https://www.oxygenxml.com/ [last access: April 27, 2024].

Pixelio = *pixelio.de – Deine kostenlose Bilddatenbank für lizenzfreie Fotos*. http://www.pixelio.de [last access: April 27, 2024].

PostgreSQL = *PostgreSQL Datenbankmanagementsystem*. http://www.postgresql.org/ [last access: April 27, 2024].

Sketch Engine = *Sketch Engine*. https://www.sketchengine.co.uk/ [last access: April 27, 2024].
TEI Lex-0 = *TEI Lex-0. A baseline encoding for lexicographic data*. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html [last access: April 27, 2024].
TLex = *TLex Suite: Dictionary Compilation Software*. https://tshwanedje.com/tshwanelex/ [last access: April 27, 2024].
Wikimedia = *Wikimedia Commons, das freie Medienarchiv*. http://commons.wikimedia.org [last access: April 27, 2024].

Images

Fig. 3.1 "Robotereinsatz am Fließband": KUKA Systems GmbH [CC BY-SA 3.0-de (http://creativecommons.org/licenses/by-sa/3.0/de/deed.elegen.)], via Wikimedia Commons (http://commons.wikimedia.org/wiki/Category:KUKA_robots?uselang=de).

Fig. 3.1 "Werkzeugschlosser bei der Arbeit": Deutsche Fotothek [CC-BY-SA-3.0-de (http://creativecommons.org/licenses/by-sa/3.0/de/deed.en)], via Wikimedia Commons (http://upload.wikimedia.org/wikipedia/commons/7/75/Fotothek_df_roe-neg_0006486_016_Portr%C3%A4t_eines_Arbeiters_an_einer_Bohrmaschine%2C_W.jpg).

Axel Herold, Peter Meyer, and Frank Wiegand

4 Data Modelling

Fig. 4.1: Lego bricks.

A huge pile of Lego bricks: a great deal of material that makes it possible to construct numerous buildings. But what’s the best way to start? Do you pick out individual bricks piece by piece in order to build a house? At the very latest, when you’re struggling to find a red three for the third time, it might be worth thinking about whether you should have sorted the bricks first. But what’s the best system to use? The reds in one box and the blues in another; the same for the yellows and whites? Or is it better to sort all the 2 × 1s, 2 × 2s, and 2 × 3s together, irrespective of colour? Whichever organising system you choose, after the bricks are sorted, you can “access” them in a more targeted way, i.e. building becomes no trouble at all.

Note: The German version of 2016 on which this chapter is based was written by Axel Herold, Lothar Lemnitzer, and Peter Meyer.

Axel Herold, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: herold@bbaw.de
Peter Meyer, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: meyer@ids-mannheim.de
Frank Wiegand, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22–23, 10117 Berlin, Germany, e-mail: wiegand@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/10.1515/9783111233758-005

4.1 Introduction

Data modelling is also concerned with sorting and structuring, but of data rather than Lego bricks. In relation to lexicography, the task of data modelling is to structure the lexicographic content so that the computer can grasp it in a targeted way. Translated into the Lego example, it is possible to attach labels to the individually sorted Lego boxes that enable a machine to grab all of the red bricks or all of the blue ones in a targeted way, or – when sorted differently – to target only the 2 × 2 and 2 × 3 bricks. This programming is considerably easier than developing a machine that automatically distinguishes between the singles and doubles and can deliberately pick them out. It is exactly the same with lexicographic data: here we can also proceed more flexibly with the data when the content can be distinguished easily by a machine.

Just as different houses can be built out of the same Lego bricks, it is a normal requirement nowadays to present the same lexicographic content in different ways, e.g. in a print and an electronic dictionary. The prerequisite for this – and for so-called advanced searches in digital dictionaries, where users are able to enter complex combinations of search options (→ Chapter 5) – is appropriate data modelling.

In order to understand this process, we must first take a look at the different "levels" that have to be taken into account in the production of a print or electronic dictionary (→ Fig. 4.2 and → Chapter 3). The basis of any dictionary is, first and foremost, a lexicographic database. Various product-related excerpts can be derived from this database. A good example for this approach is the shared database called "Duden – Wissensnetz deutsche Sprache" (Alexa 2011) that serves as the basis for the content of various reference dictionaries on Standard High German where all of the dictionaries are collectively known under the brand name Duden. The "Wissensnetz" database includes the data contained in the dictionaries on German orthography, loanwords and synonyms, for example. To compile a dictionary, a specific excerpt is generated from this database targeted for a specific output, such as the 148,000 headwords in Duden 1 – Die deutsche Rechtschreibung (DDRS) with the associated information on spelling and grammar and brief explanations of meaning. Then, either a dictionary can be printed from this output-specific database or an app can be developed for smartphones and tablets or an Internet dictionary. The data modelling is done at the level of the database since all of the prerequisites for the following steps are created there. Returning to the Lego analogy, the individual bricks can be found at the database level. The finished outputs are located at the external presentational level, i.e. the houses in the Lego example or the individual dictionaries when applied to lexicography. The lexicographers work directly on the lexicographic database and the users interact with and read the outputs.

It may be worth stressing at this point that the underlying lexicographic database might not only include textual descriptions of linguistic phenomena. It might also store information such as images or even short video sequences that are used in the description of meaning within entries. This is most obviously the case for dedicated picture dictionaries. Another common type of data stored in lexicographic databases are sound files that capture the pronunciation of headwords and possibly other parts of the entries such as related multi-word expressions or even citations. Thus, when we talk about lexicographical data throughout this chapter, we take it to mean data in a very broad sense.

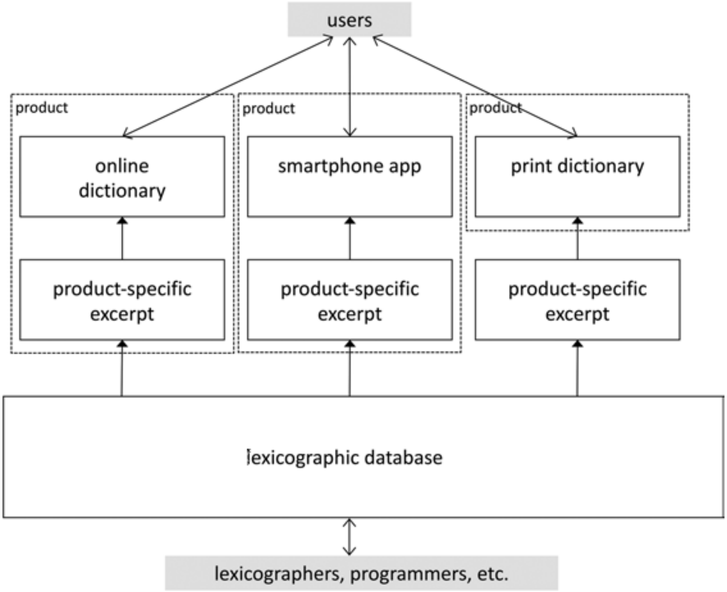


Fig. 4.2: Levels in the lexicographic process.

In order to be able to present the same lexicographic content in different ways, that content must be machine accessible in specific ways. The foundation for this is a suitable data model. Just like with the Lego bricks, an ordering principle has to be chosen: is the leading element the content of the information (e.g. whether it describes the word class or meaning of a headword), or only the part of the entry to which it belongs (the general information at the beginning of an entry or specifically a section on a particular meaning), or a different aspect altogether? It is important to have clear guidelines: as with the Lego bricks, choosing a new sorting system in the middle of the process is “expensive”. For example, if we had already sorted half of the Lego bricks according to colour and then decided we wanted to sort them according to size, then all of the work done so far would have been in vain. It is no different when modelling data.

In this chapter, we will first discuss different data formats in which structured content can be formally represented, explaining their respective advantages and disadvantages, and how suitable query languages can be used to retrieve information from these data structures. The third section covers the core issues of data modelling – how to describe the structure of specific lexicographical content, e.g. which “boxes” the lexicographic content should be put in – both in abstract terms and with reference to the data structures introduced before. put in, including the associated advantages and disadvantages. There are many lexicographic projects that face largely similar challenges. For this reason, initiatives have been launched oriented towards developing standardised solutions for modelling lexicographic data, similar to a set of guidelines for sorting Lego bricks. We report on these in the fourth section.

4.2 Data structures and representation formats

In this chapter, we pick up on concepts, explanations, and examples from → Chapter 1 on the technical foundations of Internet dictionaries. The starting point now is how the lexicographic information for a dictionary should be stored in a sensible way on a server. Let us revisit the HTML code already discussed in → Chapter 1 used to create a basic web page for a dictionary entry on the English lemma *disproof*, replicated here for convenience.

```
<html>
   <head>
     <meta charset="utf-8">
     <title>MyEnglishDict</title>
   </head>
   <body>
     <h1>disproof</h1>
     <p>[dɪs'pru:f] <i>n.</i></p>
     <ol>
       <li>facts that disprove something</li>
       <li>the act of disproving</li>
     </ol>
     <p><i>See also:</i> <a href="/entry/disprove">
        disprove</a></p>
   </body>
</html>
```

In our explanation of how web servers and web applications generally work in → Chapter 1, we deliberately left out the central issue of where exactly the web application obtains the HTML code for the relevant dictionary entry required by the user. One obvious and straightforward answer would be to simply integrate the code for the dictionary entry in the web application. A web application is a program that runs on the web server and responds to requests from the client. As such, the following instruction could, for example, be integrated into this program: “If the request reads ‘`GET /entry/disproof`’, then send the following HTML code back as the response:

(see code as shown above)". However, this is not a viable approach since any change in the appearance of the web page for an entry would require a change in the code of the computer program; this means that programming the web application and editing lexicographic content would become inextricably intertwined.

In order to separate the storage of data for lexicographic information from the programming and administration of the web application, an individual text file could, quite simply, be deposited for each dictionary entry and named according to the relevant lemma in a particular directory on the hard drive of the web server. Each text file would then contain the HTML code of the relevant web page. In this scenario, when the web application receives the request "`GET /entry/[LEMMA]`" (where `[LEMMA]` is a placeholder for the required lemma), it searches in the aforementioned data folder for a text file with the name `[LEMMA].txt`. If such a file is found, the program reads the content of the text file (which is the HTML code for the entry's web page) and sends it as an HTTP(S) response to the web browser making the request. The obvious advantage of this solution is that lexicographers can change or even delete the text files completely independently of the programmers and deposit new text files as the dictionary is extended; the web application itself remains unaffected by these changes and can simply continue to run since the program code contains no information at all about the content of the web pages.

In this approach – or, typically, a more sophisticated variant using a database, for example – the more "technical" aspects of compiling an Internet dictionary are somewhat decoupled from the more "content-related" aspects. However, the decoupling is neither as complete nor as far reaching as necessary: the lexicographic data that are made available to the program still consist of HTML code. Thus, these data are stored from the beginning in a specific format for Internet dictionaries and determine the appearance of individual dictionary entries. As such, lexicographers writing entries for an Internet dictionary like this find themselves at the level of data presentation from the beginning. Let us assume that the dictionary editors decide to make changes to the presentation of an entry at a later date:

- "Information about the word class must sit right next to the lemma and be written out in full, e.g. *noun* instead of *n.*"
- "Information on pronunciation should appear between |vertical slashes| rather than between [square brackets]."
- . . .

In such a case, a simple change to some CSS code does not do the trick. All of the HTML pages have to be altered manually or with the help of suitable programming, even though the lexicographic content will not have changed at all. Therefore, what is also required is a separation between the *lexicographic data* proper and the properties of *data presentation*: the text files for individual dictionary entries on the web server should not contain any HTML code but rather just indicate the lexicographic information in a data format that makes sense for the lexicographic work and that is abstracted as far as possible from the details of its final presentation. It is then the task of the web application to translate this data format into suitable HTML code. If the editors decide to make changes to the presentation of the entries at any point, only the program code for that process of translation has to be changed; the original data with the lexicographic information remain unchanged.

Yet what does a suitable format for representing lexicographic information look like? Answering this question brings us to the problems of data modelling and the data structures associated with it.

One obvious textual representation that is suited to machine processing is to separate the individual *types of information* from one another in a hierarchically structured form and to apply corresponding headings or "labels" to them. This can be done in many different ways, as in this sketch of a possible approach:

```
entry (id: MED.disproof):
  form:
    spelling: disproof
    pronunciation: dɪs'pru:f
  grammar:
    part-of-speech: noun
  senses:
    sense (numbering: 1):
      definition: facts that disprove something
    sense (numbering: 2):
      definition: the act of disproving
  cross-reference (refid: MED.disprove): disprove
```

As we can see, the content structure of the entry is represented here in blocks that are hierarchically nested inside one another and marked with "headings" and indents. Each block contains either a series of further subordinated blocks (the **senses** block contains two individual **sense** blocks and these each contain, in turn, a **definition** block) or simply text (the **pronunciation** block contains the text "dɪs'pru:f"). Some blocks have additional meta information that is recorded as what we will call *attributes* in parentheses after the block name, instead of in its own subordinate blocks. Thus, the whole block with the name **entry** is assigned the attribute **id**, here with a *value* that is supposed to specify a uniquely identifying ID character string for this particular entry; the block **cross-reference** contains the lemma of the referenced entry and has an attribute **refid**, whose value is the ID of the entry that is being referred to. The **sense** blocks feature an additional attribute **numbering** that indicates numbering labels such as "2.", "2.c", "(ii)", or similar. Such an attribute could be useful in the context of digitising a print dictionary. Note that the names of the blocks and attributes can, in principle, be chosen arbitrarily; the hierarchical structure could also have been designed in a different way. In our oversimplified example, it would have been possible, for instance, to just have the blocks named **spelling**, **pronunciation**, **part of speech**, **definition**, and **cross-reference** as pieces of information directly subordinate to **entry**. Even the textual order of the data could have been different, one possible exception being the ordering of the **sense** blocks, which might reflect a lexicographical assessment of the frequency, importance, or relatedness of the senses. Finally, there is no logical necessity to introduce attributes as a separate syntactic device for marking meta information; it would have been sufficient to use ordinary blocks for that purpose. The particular way in which the data are structured in our toy example actually foreshadows standard ways of modelling and digitally encoding lexicographical content, which are to be discussed in what follows.

4.2.1 XML documents

Formally, hierarchical structures of this type can be described as trees and, accordingly, can be represented as tree diagrams, which do, in fact, look like a tree standing on its head. Such trees consist of individual positions (*nodes*) connected by "arrows" (*edges*) (→ Fig. 4.3).

The *root node* at the top of the "inverted" tree represents the entire structure, and the *child nodes* linked with it by edges represent the blocks of the highest structural level. From a formal and informatics perspective, trees are simple structures that are easy to describe and process. They have been used for a long time in metalexicography to systematically describe entry structures in print and digital dictionaries (Kunze/Lemnitzer 2007: 77–93; cf. also Wiegand 1989). Every node in the tree – each content block – has precisely one parent node – a block that contains it – with the exception of the root node. Accordingly, trees can be stored in a simple way in a computer in that each node is basically represented by referencing the storage addresses for its child nodes, with the exception of "childless" nodes or *leaves* at the bottom of the hierarchy, which are simply stored as text.

A very common way of encoding such tree structures in a textual form that is both human and machine readable is *XML* (Extensible Markup Language), which is strongly reminiscent of the HTML discussed in → Chapter 1 and which works "according" to the same basic principles. The individual content blocks are each enclosed in a start tag and a corresponding end tag and contain further "blocks", formally known in HTML and XML as *elements*, and/or plain text, that is, a sequence of characters, especially letters and numbers. XML is syntactically more rigid than HTML: for example, unlike HTML, elements always must have an end tag, even if they do not have any content at all. While the "vocabulary", i.e. the range of available element and attribute names, and the "grammar", i.e. the rules that govern where elements are allowed to occur in the document, are mostly predefined in HTML, in XML all these aspects can be defined individually for each concrete *application* – e.g. for encoding entries in a particular dictionary – in a so-called *schema*.

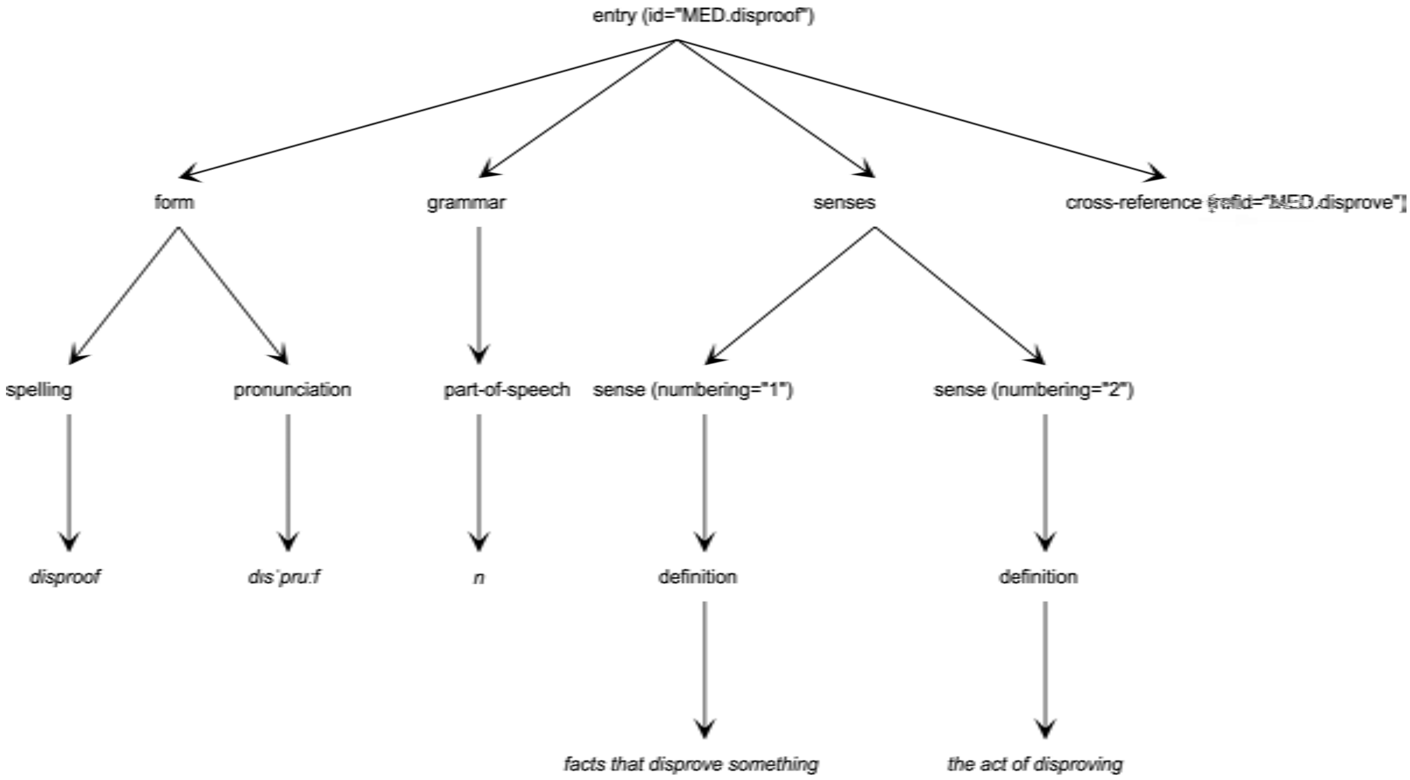


Fig. 4.3: Representation of the microstructure of an entry as a tree diagram.

Represented in XML, the toy entry could appear as:

```
<?xml version="1.0" encoding="UTF-8"?>
<entry id="MED.disproof">
  <form>
    <spelling>disproof</spelling>
    <pronunciation>dis'pru:f</pronunciation>
  </form>
  <grammar>
    <part-of-speech>n</part-of-speech>
   </grammar>
  <senses>
    <sense numbering="1">
      <definition>facts that disprove something
      </definition>
    </sense>
    <sense numbering="2">
      <definition>the act of disproving</definition>
    </sense>
  </senses>
  <cross-reference refid="MED.disprove">disprove</cross-
  reference>
</entry>
```

The first line is the *XML declaration*, which has a special syntax and specifies the character encoding used. (For more about this concept and the somewhat similar document type declaration in HTML, → Chapter 1) Just like with HTML, XML has the concept of attributes. Attributes are typically used to describe information that is in some sense descriptive of a specific element occurrence but is not regarded as part of its content. However, there are no clear-cut rules when to use attributes and when to use elements with text content. For example, the part of speech could also be coded as an attribute.

If the XML-based representation shown above is to be used for a web application, the application must contain program code capable of *parsing*, i.e. analysing the structure of this XML document and creating, in the simplest case, the HTML document shown at the beginning of → Section 4.2. This code can take advantage of the fact that the different types of lexicographic information in the XML document are each marked up semantically with their own tags. A widely used technology for generating other XML documents or HTML documents out of given XML documents is called an *XSL transformation*. As a side note, the program code of XSL transformations, often referred to as an XSLT *stylesheet*, is itself written using XML syntax. Using different stylesheets, an XML document can be “translated” into HTML pages in completely different ways, depending, for example, on the user's preference, or, alternatively, into other types of documents, such as a PDF file for printing. In the transformation process, any information contained in the XML document can be omitted or reorganised; in this way, the same XML document can be used to generate, for example, an overview presentation of the most important information as well as a complete, detailed view of a dictionary entry.

4.2.2 Relational databases

Another form of representation for lexicographic data, and one that has been around for much longer, are *relational databases*. By this, we mean a structured system of data tables, somewhat comparable to the tables in spreadsheet software. These data tables can be saved in an extremely efficient form on the hard drive of a server computer and be read, altered, and managed with great speed by a program known as a *database management system* (DBMS). Programs – e.g. web applications – that receive information from a relational database or wish to modify it "call" the database management system using *SQL*, a specialised query and data manipulation language. The database management system can run on the same computer as the web application or on another server, usually connected via a fast internal computer network. It is not possible here to go into the complex details of relational database technology; instead, by virtue of the miniature entry used as a toy example above, we shall demonstrate how dictionary entries can be described in a relational database. In order to keep the example simple, we initially assume that all of the entries in the dictionary include only one indication of pronunciation and, at most, one cross-reference to another entry. Only the number of word senses will vary. Then, the main table of dictionary entries might look as it does in → Table 4.1.

Table 4.1: Relational table ENTRYTABLE of dictionary entries.

ID	Spelling	Pronunciation	PartOfSpeech	CrossReference
. . .	. . .	. . .	. . .	. . .
MED.disproof	disproof	dɪs'pruːf	noun	MED.disprove
MED.disprove	disprove	dɪs'pruːv	verb	NULL
. . .	. . .	. . .	. . .	. . .

Since the number of word senses varies and, in theory, any number of meanings can belong to any one dictionary entry (the so-called *1:n-relation*), senses require their own table, which contains the sense definitions and the numbering label but also, crucially, the reference to the relevant entry. The references use the entry IDs, and the ID column acts as the so-called *key* to unambiguously identifying a table row (a specific dataset or “record” as in → Table 4.2). The database management system can automatically guarantee the *referential integrity* of these references to datasets in other tables; that is, it prevents a dataset for an entry from being deleted in the entry table, if there is still a reference to the ID of this entry in the senses table.[1]

Table 4.2: Relational table SENSETABLE of word senses.

Entry	Numbering	Definition
. . .	. . .	. . .
MED.disproof	1	facts that disprove something
MED.disproof	2	the act of disproving
. . .	. . .	. . .

In a more realistic scenario, an entry can contain any number of cross-references to other entries. In that case, the “CrossReference” column is omitted from the main table, and a further table is needed for cross-references, as shown in → Table 4.3. Each row (record) in this table contains the ID of the source entry (which contains the reference) and the ID of the target entry (to which the reference refers). This is an m:n-relation: in theory, any number of target entries can belong to each source entry – each entry can refer to any number of others. At the same time, any entry can be referred to by any number of other entries. If the order in which multiple cross-references are to be presented in an entry matters lexicographically, it must be encoded as a separate column in the table because the rows of a relational table have no intrinsic ordering (→ Fn. 1).

Table 4.3: Relational table REFERENCETABLE of cross-references.

Source	Target	Position
. . .	. . .	. . .
MED.disproof	MED.disprove	1
MED.disproof	MED.proof	2
. . .	. . .	. . .

The web application can now make requests to the database management system in the aforementioned query language, SQL, in order to receive the lexicographic data

[1] Note that the rows in a relational database table are not ordered in a technical sense (since, mathematically speaking, they are elements of a set), even if a specific order has to be chosen in a diagrammatic representation like the one in → Table 4.2. Thus, the order of word senses present in the XML document has no equivalent in the relational database although it might be recoverable from the **Numbering** column.

for the entry on *disproof* with the ID “MED.disproof”. The following SQL query returns all of the column values belonging to the row with the ID “MED.disproof” in the main entry table:

```
SELECT * FROM ENTRYTABLE WHERE ID="MED.disproof";
```

However, all of the corresponding rows in the other two tables must be retrieved as well:

```
SELECT * FROM SENSETABLE WHERE Entry="MED.disproof";
SELECT * FROM REFERENCETABLE WHERE Source="MED.disproof";
```

With the help of the data acquired in this way, the web application or web service can, in turn, construct its response to the client (e.g. an HTML page).

4.2.3 Other types of databases

While XML documents and relational databases continue to be the dominant representation forms for large lexical resources, other types of databases are being explored and used in different contexts. Conventionally, such non-relational databases are collectively referred to as NoSQL databases. Conceptually, one strain of development is focusing strongly on the notion of documents (resulting in a *document store*) and another strain is focusing on generalising the tree model underlying many traditional lexicographic databases to a graph model (resulting in a *graph database*).[2]

In a document store, entries are typically managed as individual documents by the database management system. For example, in an XML-based store, entries such as those described in → Section 4.2.1 are stored by the DBMS as separate and “individual” entities without mapping them explicitly onto a (complex) table structure. The indexing system of an XML database then allows for the direct retrieval of sets of documents or parts of documents using established query devices such as XPath (XML Path Language) or XQuery (XML Query Language).

XPath expressions allow the specification of the location of individual nodes within an XML document by specifying their properties in terms of their name (i.e. a *node test*), contextual constraints on the node (*predicates*), and an indication as to the *axes* that need to be followed when traversing the tree (e.g. following the edges vs.

[2] There are other NoSQL approaches too, such as key-value stores, but we will not discuss them here because there has been little uptake so far in the domain of lexicography.

moving laterally across sibling nodes). Consider the following XPath expression and its application on the XML representation presented in → Section 4.2.1:

```
/child::entry/child::form/child::spelling
```

The forward slash separates individual steps of the path expression, the axis specification (child::) tells us to move along the line of descendants of the nodes (i.e. along the edges of the tree), and the node tests specify the names of the nodes to be expected along the path. There are no predicates in this example. When the expression is "processed" (evaluated), the result returned will be a set of all the nodes (*node set*) that are reached when traversing the tree as follows: start at the root node <entry>, from there proceed to a child node <form>, and from there to another child node <spelling>. Given the single document in the example, a node set containing the node <spelling>disproof</spelling> would be returned. As more and more entries are added to the database that are structured like the *disproof* entry, the resulting node set would grow accordingly, effectively providing a headword list derived from all of the spellings in all of the entries. The expression in our example can be abbreviated in two regards: syntactically and semantically. As the child:: axis is considered the default, it can be omitted, resulting in the syntactically equivalent expression:

```
/entry/form/spelling
```

If we are only interested in creating a headword list from <spelling> nodes, it is not strictly necessary to specify that each such node needs to have a parent node called <form>. In this case we can change the axis that needs to be traversed from child:: to the broader descendant::, resulting in the semantically equivalent expression:

```
/entry/descendant::spelling
```

or even:

```
/descendant::spelling
```

denoting all <spelling> nodes that can be reached when moving along the edges of the tree downwards from the root node.

As an example for the application of predicates, consider the following expression:

```
/entry[descendant::sense[@numbering]]
```

Here, the square brackets enclose predicates, i.e. constraints that the nodes in the path expression have to satisfy, with the @ symbol denoting the name of an XML attribute. Thus, the node set returned by this expression is the set of all `<entry>` nodes that have one or more descendant `<sense>` nodes, which, in turn, must meet the condition to carry an attribute called `numbering`. In this way, we can determine the set of all entries that describe polysemous words.

While XPath expressions allow for the selection of nodes that meet certain criteria so they can be retrieved and returned by the DBMS, they do not provide a means of modifying or storing data in the DBMS. For complex query, storage, and retrieval tasks, XML databases typically provide XQuery-based interfaces. XQuery uses XPath expressions to create node sets that can then be used in complex expressions. Consider the following example for such a query. It generates a fragment of HTML code consisting of an `<ol>` element that contains `<li>` child elements with all the entries' headwords, alphabetically sorted, as their text content. Note that the `collection('/db/dict')` part serves to illustrate a locator for the dictionary in the DBMS, which will often be stored as a collection of documents:

```
<ol>
   {
   for $headword in collection('/db/dict')/
   descendant::spelling
    order by $headword
    return <li>{ $headword} </li>
   }
</ol>
```

The evaluation of this XQuery expression goes beyond the application of XPath in two ways. First, it provides a template for HTML markups (`<ol>` for ordered lists and `<li>` for list items therein) that enables a direct rendering of the query result. "Second", much like with the SQL queries on relational databases, XQuery engines allow for further modifications of result sets, such as sorting (`order by $headword` in our example). In XQuery, users may define and invoke custom functions, and also the DBMS will provide its own interface via special functions so that the database can not only be queried but also modified, updated, and added to.

Let us conclude this section with a brief description of graph databases, which, at their core, rely on the mathematical concept of a graph. We will not go into the finer details of graph theory here but rather focus on the essence needed to understand its possible applications in the context of lexicography.

Generally, a graph consists of a set of nodes (called *vertices* in graph theory; we will use the term *node* here to underline the relation with the description of the trees above) and a set of edges, which are essentially pairs of nodes where the nodes are related to one another in a specific way. Graphs can be classified according to certain properties, such as whether the edges are directed or undirected, whether all nodes need to be connected or not, whether they may contain loops as opposed to only containing a single path between any two given nodes, or whether the edges may carry additional information (such as weights), to name but a few.

Graph databases differ from each other in their restrictions on and assumptions about the features of nodes and edges. In the case of labelled-property graphs, nodes and/or edges may have explicitly specified, named (i.e. labelled) properties that can be used to store additional data directly without the need to model these properties as nodes and edges as well. As a consequence, nodes and edges in a labelled-property graph may have different “data types”, i.e. sets of mandatory or optional properties. The opposite approach is adopted by the data model in the *Research Description Framework* (RDF). Here, the graph consists of a set of *triples* that each comprise two unlabelled nodes and an unlabelled edge connecting them. The edge (the *predicate*) always points from one node (the *subject*) to the other node (the *object*); thus the graph is directed. The triples constructed in this way can be considered statements about two *resources* for which the relation expressed by the predicate holds. The term *resource* is used very generically in RDF. In the domain of lexicography, a resource may be a single dictionary, an entry within a dictionary, or any constituent that entries are constructed from. To refer to resources, RDF relies on *uniform resource identifiers* (URIs) that unambiguously identify resources. While the subject and the predicate always need to be URIs, the object may be either a URI or a literal (i.e. a character string). Several notations are used for RDF triples, among them XML- and JSON-based serialisations as well as RDF specific formats such as N-Triples or Turtle. To provide a practical example, we use the easily readable N-Triple notation. The triple is given on a single line and terminated by a full stop:

```
<http://example.com/entry/disproof>
<http://example.com/has_headword> "disproof" .
```

This triple states that a dictionary entry referred to by its URI `http://example.com/entry/disproof` has a headword (`http://example.com/has_headword` – the relation is also referred to by its URI) that is given by the literal string “disproof”. Statements regarding senses could be formalised accordingly:

```
<http://example.com/entry/disproof> <http://example.com/
means> <http://example.com/sense/disproof_facts_sense> .

<http://example.com/entry/disproof> <http://example.com/
means> <http://example.com/sense/disproof_act_sense> .
```

Note how the subject is identically referred to by its URI twice and how the object in each statement is also referred to by a URI this time. With a triple representation of our dictionary stored in a graph database (which in this case would be called a *triple store*), triples with subjects referred to by the URI `http://example.com/sense/disproof_facts_sense` would allow us to retrieve further information on the first sense of the entry. To query an RDF triple store, SPARQL (the SPARQL Protocol and RDF Query Language) is used. The following SPARQL query retrieves the senses that are associated with the entry "disproof" (a line starting with `PREFIX` describes a prefix that is used to shorten the URIs):

```
PREFIX ex: <http://example.com/>
PREFIX entry: <http://example.com/entry/>
SELECT ?sense
WHERE
{
   entry:disproof ex:means ?sense .
}
```

For graph databases, many efficient algorithms have been described and implemented (cf. Robinson/Eifrem/Webber 2013), which makes it possible to quickly search for paths in graphs, that is, to locate routes from one node to another running along multiple edges. Especially in the context of Linked Open Data (LOD), graph databases have become hugely popular recently. The types of data considered in the LOD paradigm go far beyond lexicographic data. There is a strong focus on general knowledge bases such as Wikidata and DBpedia, two projects that automatically extract facts from WIKIPEDIA and convert them into *knowledge graphs*. Another common type of LOD resources are ontologies that model – often domain-specific – conceptual hierarchies. LOD resources form the basis for the *Semantic Web*, thus named to highlight its overarching goal, which is to provide the data and infrastructure needed to create semantic annotations for resources on the Internet. Early on, ideas were proposed to also include lexicographical resources (cf. Spohr 2012). Dictionaries that rely heavily on relations (such as the lexical-semantic wordnets discussed in → Section 4.4.2) are ideal for graph-based representations because of the close resemblance of their internal organisation and the modelling assumptions imposed by graph databases. Nevertheless, in principle, all lexical resources can be represented in graph databases.

4.3 Data modelling

4.3.1 Conceptual (semantic) data models

The discussion so far has shown how lexicographic information can be represented in very different data formats – textual or tabular – independently of presentational aspects, facilitating further machine processing and flexible presentation of the data. In the process, we also raised the problem that, when developing an Internet dictionary, it must first be decided in very general terms how the data will be structured that need to be stored and processed. Particular questions that arise here are which types of lexicographic detail we need in our dictionary entries, which hierarchical relationships exist between them, which are obligatory, and which can occur more than once. As the example of cross-references between entries demonstrated above, these fundamental decisions about structure are necessary in the case of relational databases in order to determine the number and structure of data tables and their relationships to one another. But these decisions are also a prerequisite for determining which XML elements are needed for an XML-based dictionary and how they are to be nested inside each other; thus, it would make little sense to distinguish, as shown in the example in → Section 4.2.1, between a superordinate 'container' or 'wrapper' element <senses> and subordinate <sense> elements if there was a maximum of one meaning per entry.

If developing a dictionary involves specifying the required lexicographic indications and their relationships in an abstract way without already deciding, for example, on whether to use a relational database or XML, then we have entered the territory of *conceptual data modelling*. There are established and formalised diagrammatic formats for formulating conceptual data models, in particular the *entity-relationship model* and the *Unified Modelling Language* (UML). As an illustration, we shall present only a very simple example based on UML modelling applied to the toy dictionary entry discussed in → Section 4.2.

A large number of different types of diagrams are associated with UML. → Fig. 4.4 shows a *class diagram*. The rectangles represent *classes*, that is, types of *entities* that need to be modelled. This example sets out two types of entities, namely dictionary entries as a whole and word sense information within these entries. The names of *attributes* are located underneath the names of the classes, separated by a horizontal line. In UML, attributes are the properties that jointly characterise each entity (entry, word sense) of the relevant class. For actual entries, these properties in our example are an ID, the orthographic form of the lemma sign, its pronunciation, and its part of speech. Word senses have a definition and (assumed here for demonstration purposes) a numbering within the entry. In more detailed modelling, the *data type* of the individual attributes could also be given, for example, the pronunciation is a string of symbols of any length or the part-of-speech indication is one of several predetermined sets of symbols such as "n", "v", "prep", etc.

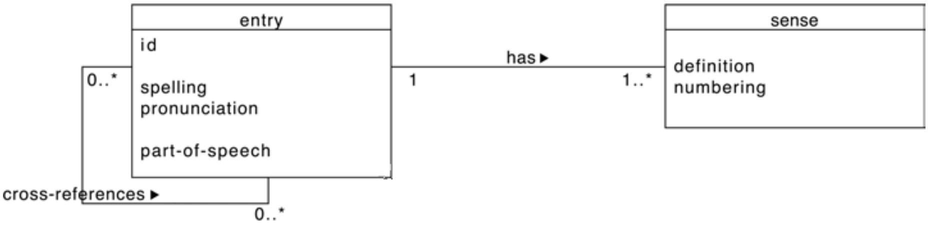


Fig. 4.4: Simple UML modelling of an example entry.

The relationships between the entities in classes are represented by associations, which are lines drawn between the relevant class rectangles to link them. The *multiplicities* of the association are given at each end of such a line. This is explained in the modelling requirements that the diagram above expresses (the asterisk symbol * designates in general terms an arbitrary non-negative integer number):

- A given entry has at least one but otherwise any number of senses (multiplicity 1..*).
- Conversely, each meaning only "belongs" to exactly one entry (multiplicity 1 or, in more detailed notation, 1..1). This is not a trivial point; one might wish to model relationships of synonymy explicitly in this way so that one and the same "meaning" can be assigned to multiple entries.[3]
- A given entry *cross-references* any number of other entries (multiplicity 0..*); since the entities that relate to one another are instances of the same class (entries), we speak of a *reflexive association*.
- Conversely, a given entry can be referred to by any number of other entries (multiplicity 0..*).

It is clear that the principles and terminology of conceptual data modelling sketched out here can generally be applied without alterations to describe the microstructure and mediostructure of print dictionaries. However, especially older print dictionaries tend not to have rigid, formalisable structures since they were conceived for human users rather than for machine searching and processing. As such, the fundamental difference between digital dictionaries and print dictionaries is not the manner in which information can be structured as such. Instead it is the necessity to actually store and prepare the data in some kind of rigorously structured way as well as the possibility of presenting this structured content in a flexible way and of making it possible for it to be searched accordingly. Here, the *granularity* of the data modelling can vary considerably. Especially in older dictionaries, there are often sections within entries that cannot be structured in a consistent way when they are retrospectively prepared in a digital form because of their narrative character; a typical example would be discursive explanations of etymology. In the most extreme case, this kind of section has to be modelled as an entity, the sole attribute of which is simply the whole text of the section as a non-structured series of characters that can only be accessed in a full text search. Freshly conceived digital dictionaries are the opposite extreme, since in this case it is possible to model the lexicographic data in a very granular way, that is, to store the individual types of information (indications) in a very fine-grained way, each as a different attribute of an entity. In the case of actively edited and maintained dictionaries, the modelling and overall lexicographic process must be flexible and ex-

[3] In that case, the position numbering would have to be dealt with in a different way since the numbering assigned to a sense in one entry might differ from that in another entry that features the same word sense. The solution would basically be to encode the numbering as an attribute of the association itself, using what is called an association class.

tensible to guarantee that each entry may be revised at any time. This often makes it necessary for the conceptual modelling to adjust to new requirements that arise while the dictionary is already operational.

4.3.2 Logical data models

It is striking that the manner in which one entry is supposed to cross-reference another is not specified in the UML diagram in → Fig. 4.4. In the XML document shown in → Section 4.2.1, the cross-reference is achieved by providing the ID of the entry which is being referred to. However, implementing the actual “cross-referencing mechanism” assumes that the data(base) format is already known. On the conceptual modelling level, though, these issues are generally dealt with in abstract terms. The focus is essentially on content-related decisions such as the types and properties of entities that will be described and the types and properties of the relationships “between” these entities. Questions arising from the actual implementation of cross-referencing structures are addressed instead in the domain of *logical data modelling*, which involves “spelling out” the conceptual data model for a specific data format and the database system associated with it. The process of spelling out the data model is not a process that can be carried out mechanically since the conceptual and logical data models do not exist in a simple correspondence to one another. For example, the grouping of the **spelling** and **pronunciation** elements under the superordinate **form** element in the XML document in → Section 4.2.1 does not have any formal correspondence in the class diagram in → Fig. 4.4.

Logical data modelling with XML documents is again captured through suitable formal descriptions, so-called *schema languages*. There are several established formal schema languages for XML documents, including *DTD* (Document Type Definition), *XSD* (XML Schema Definition), and *RELAX NG* (REgular LAnguage for XML Next Generation). For illustrative purposes we show here a simple, almost self-explanatory RELAX-NG modelling applied to the toy example XML document in → Section 4.2.1:

```
element entry {
 attribute id {  text } ,
 element form {
  element spelling {  text } ,
  element pronunciation {  text }
 } ,
 element grammar {
   element part-of-speech {  string "n" | string "v" |
 string "adj" }
 } ,
 element senses {
   element sense {
    attribute numbering { text },
    element definition { text }
   }+
  },
  element cross-reference {
   attribute refid { text },
   text
  }*
  }
```

The specified modelling determines exactly which elements are permitted to appear in a generic XML document for our fictional dictionary, with which attributes, in which position, and how many times. In the miniature modelling provided, the word class element **part-of-speech** can only contain one of the three labels “n”, “v”, or “adj”. In the example, the equivalents for the multiplicities from the conceptual modelling are the symbols “*” (corresponds to 0..* in UML; so “any number, including none”) and “+” (corresponds to 1..*, so “at least one”).

The various schema languages differ from one another in terms of their expressive power, i.e. they permit constraints to varying degrees and of varying complexity to be formulated. But their purpose is the same: to describe with precision the desired structure of a class of XML documents. Then a computer can check in a purely formal way whether or not a given XML document really matches this required structure. This process is called *validation*. The validity of XML documents is a fundamental prerequisite for any form of further machine processing of the documents. Thus, a program to translate any dictionary entry represented in XML into an HTML representation (e.g. an XSL transformation) can only be developed if it knows the structure of the XML documents and, therefore, where in these documents to find which indications.

Of course, there are also formal techniques for specifying the desired data structures for relational databases. A relational database schema determines which tables there are, which columns they have, which types of data can be entered into the different columns, which relationships exist between the tables, and which keys have to refer to a specific row in another table (→ Section 4.2.2). It is also possible to determine further restrictions in a database, so-called *constraints*, which prescribe, for example, the range of values allowed in a particular column or certain complex conditions for the permissibility of whole datasets (rows), the maintenance of which is automatically protected by the database management system.

4.3.3 Technical implications of logical data modelling

In principle, any given conceptual data modelling can be realised using any of the technological methods for representing and manipulating data introduced in this chapter. However, the choice of a representation format has far-reaching practical consequences, especially when it comes to the tools required for processing the data and the necessary technical equipment as well as the lexicographic work process and the compatibility of data with the output and requirements of other projects as well. A further criterion is the flexibility and expandability of the chosen form of representation in the event of new requirements for the lexicographic information represented in the dictionary concerned. Here, relational databases are often at a disadvantage since changes to the data modelling can bring about a complex reorganisation of the table structures. Finally, in certain circumstances a justifiable balance needs to be found between the desired complexity of data modelling and the speed of data retrieval.

In a relational database, the data are distributed across many tables in the optimum form for machine processing. In order for a human processor to be able to do anything with these data, a program designed for lexicographers to edit the data has to read the desired information from the various tables using queries to the database management system and then present it as readable text. Conversely, any changes or additions to the data input via the editing application must be “translated” again by this program into SQL commands to change the datasets in the various tables. Because the input program cannot randomly change the database schema (that is, the number or structure of tables, for example) and because the database management system itself systematically prevents formal inconsistencies in the data, adherence to the chosen conceptual data modelling and the integrity of the data are guaranteed, even if several people revise the lexicographic information in an entry at the same time.

At present, XML is still the de facto standard for representing lexicographic data. Unlike a relational database, where the lexicographic information for an entry is stored in a clever way so as to be dispersed across multiple tables, XML documents are initially nothing more than plain text documents that can be read by a human being, that contain all the lexicographic information for an entry in one place, and that can, in principle, be viewed and edited in any simple text editor or word processor.

However, in practice, specialised *XML editors* are used to edit XML documents. These automatically ensure, for example, that the documents are syntactically well formed. In other words, when changes are made, the editors prevent the general rules for constructing and structuring XML documents from being inadvertently contravened, such as the end tag being forgotten after its associated start tag. Professional XML editors can present XML documents in a way familiar from word processors so that a lexicographer working on the document shown in → Section 4.2.1 sees it in a similar way to that in → Section 4.2. Such a convenience view must first of all be configured for a given XML schema. During editing, one very important function of an XML editor is constant automatic validation of an XML document with respect to a given XML schema. In this way, if an XML editor is set to use the schema from → Section 4.3.2, it can automatically prevent an additional **part-of-speech** element from being added in the XML source text of → Section 4.2.1. Nonetheless, in contrast to relational database systems, no standard solution exists for managing as well as simultaneously and collaboratively editing what might be a huge collection of large XML documents.

Numerous established technologies exist for the machine processing of XML documents. There are specialised query languages that make it possible to read information in a targeted way, leveraging the hierarchical structure of XML documents: the query language *XPath*, which makes it possible to systematically address elements and attributes, and the powerful programming language *XQuery*, which is built on the former (→ Section 4.2.3).

In view of the considerable technological differences between relational and XML representations, it is vitally important that the two formats can essentially be translated into one another. Some XML databases can even transform XML documents automatically into relational database tables with the help of a specified XML schema in order to efficiently store, search, and retrieve the data. Conversely, XML can be used as an easy textual conversion format if the content of a whole relational database (or just the lexicographic information of a single entry) has to be transferred from one system to another or has to be further processed in a different way.

Because of the extensive translatability of representation formats into one another, special data formats that are tailored to the workflows and existing, often historically developed, technical infrastructure are often used internally in lexicographic projects. Thus, it makes sense for collaborative or partially collaborative dictionaries (→ Chapter 2.2.3) to use a markup language for revising entries that is much simpler than XML or HTML. A well-known example is the markup languages used in Wiki systems like WIKIPEDIA; these systems are also used for extensive collaborative lexicographic projects (cf. Hämäläinen/Rueter 2018; also → Chapter 8). The disadvantage of using these formats is that they are often ill-suited for modelling complex and hierarchically structured information.

Data formats used internally are often not published systematically. If it is planned to transfer the data to other projects or institutions, they are typically "translated" into standardised data formats, as discussed in the following section.

4.4 Attempts at standardisation

Over time, typical forms of presentation have emerged for the contents of dictionaries so that users of a print dictionary can find the information they are looking for quickly and easily. Thus, pieces of information that belong together normally appear grouped next to one another and the headword to which the information refers is usually highlighted by a particular font or by its position at the beginning of the grouping. Of course, lexical information does not have to be presented in this particular way, but a targeted search for specific content would be made substantially more difficult if a dictionary diverged from these conventions.

While conventionalised forms of presentation are sufficient for human dictionary users to search for and find the desired information, machine production and interpretation of lexical data require that a specific form of representation is identified and agreed as binding – that is, standardised. In particular, the exact specification of the formats used is a necessary precondition for the practical implementation of software tools. Work processes (for example, the entire lexicographic process, as described in → Chapter 3) can also be standardised. However, in this section we restrict ourselves to discussing the standardisation of lexicographic models and data formats.

Generally, there are many reasons for modelling and storing lexical data in a standardised form:

- Using standard formats ensures that different datasets are compatible with one another. Information of the same type appears in the same form of presentation. For example, a standard format can specify the precise form in which pieces of data have to be stored. In this way, it becomes possible to process, change, and present data from different sources with the same software tools. Above all, specific access to entries and individual bits of information within these entries is made easier in situations where lexicographic data from different sources are aggregated and merged according to users' preferences (→ Chapter 7).
- The lexicographic process is often supported by different software tools (→ Chapter 3 and Abel 2012). Using standard models and formats ensures that these tools are interoperable, both conceptually and technically. Here, the standard format represents a defined interface between the tools. The agreed output format of one tool serves as the input format for another tool. In this way, exchanging data becomes technically possible beyond the boundaries of individual work groups.
- Publicly documented standards are an important prerequisite for long-lasting and sustainable storage of lexical data, i.e. for their long-term archiving. They can be understood as explicit and detailed format documentation. On this basis, software tools can also be (re)implemented at a later date, even when the programs used originally cannot be used any more for technical reasons.
- Alongside the advantages already listed, which are primarily of a technical nature, the consistent use of standard formats also supports the internal consistency and coherence of a lexical resource. For dictionaries modelled according to a specific format, the rules specified in that format take on the role of the dictionary's grammar. With the help of corresponding schema descriptions, it is very easy to check whether an electronic version of a dictionary – an instance of this schema – corresponds to this grammar. Here, the specification of the format can be formulated in a very detailed way and can, for example, lay down exactly which values may be used for detailing specialist domains. Some grammars such as

Schematron also allow rules to be formulated that determine properties of elements that refer to properties of elements positioned elsewhere in a dictionary entry. One such rule might, for example, state that an entry containing a synonym reference must not contain an antonym reference to the same target reference as well while still allowing antonym references to other targets.

– The explicit and detailed modelling and storage of data structures and data elements that is required by most standard lexical formats – in particular, when using XML technologies – means that the storage volume of electronic "dictionaries" is often relatively large. Nevertheless, this effect is negligible considering the availability of increasingly inexpensive computer storage.

The idea of modelling lexical resources on a *common* model in order to ensure the compatibility and interoperability of electronic dictionaries is certainly not a new one. Indeed, Kanngießer (1996) already considered the question in relation to the growing range of electronic lexical resources at that time from the perspective of the (integrated) re-use of those resources. Starting from the observation that the challenge for standardisation consists in depicting very heterogeneous lexical models in a single representation, he sets out the central problem: "lexical re-use [. . .] is therefore possible to the precise extent that it is possible to unify grammars and their underlying theories" (p. 92). Because lexicographic description cannot proceed in a theory-neutral way and, at the same time, grammatical theories can take incompatible or contradictory basic assumptions as their starting points, any standardisation would necessarily lead to inconsistent forms of modelling within a particular theory. However, this does not apply equally to all lexicographic descriptions. Rather it is possible to identify invariant elements, that is, elements modelled in the same way independently of the grammatical theory underlying them, which can quite probably be modelled on one another (cf. also Romary/Wegstein (2012), who refer to these elements as *crystals* in relation to electronic dictionaries). If a model is restricted to these invariant descriptive parts and specified dynamically on the basis of the concrete resource to be modelled, then at least a valid partial model can be achieved. This approach has been supported more recently by the introduction of a standardised lexical metamodel, the Lexical Markup Framework (→ Section 4.4.3).

Standards for electronic dictionaries are often distinguished by a high degree of variability and modularity, meaning that the formats and guidelines for actual lexicographic processes can be adapted to project-specific needs. Therefore, they typically provide modelling frameworks rather than strictly fixed rules for lexicographic descriptions. Nonetheless, it can happen that – independently of the dictionary – there is no suitable model in a standard for a particular type of information. In particular, innovative dictionaries of contemporary language like ELEXIKO or the DWDS find themselves confronted with this problem. As a rule, project-specific data models are developed for these purposes that focus on the necessary types of information. Still, standard formats can be used to exchange lexicographic data with third parties, although the transformation then necessarily involves some loss of lexicographic information. Another possibility, albeit one that is only practicable in the long term, would involve influencing the standardisation process, leading to more specialised data models that can be adopted in later versions of a standard.

The standardisation of lexical models and data formats does not have to be limited to the formal data structures themselves. In the ideal case, it also encompasses an explicit semantic description of these data structures and the elements from which they are constructed. One possible way of explicitly describing the semantics of data elements is by referring to an index of data categories and concepts that includes "definitions" for all of the elements (often in various languages), permissible values, and relationships between classes of elements. This is often achieved by referring to common ontologies.

Generally, broader technical standards underlie the modelling of the various lexicographic "standards" described in the following sections. For example, in many cases the individual letters and symbols that appear are coded using the UNICODE standard. In order to ultimately store abstract data models as data on the computer, they are often transformed according to a family of XML standards (https://www.w3.org/XML/; → Section 4.2.1). This process is known as *serialisation*. However, in what follows, we will not explore these kinds of base standards any further. Instead, we focus on the higher-level lexicographic standards.

Organisationally, attempts at standardisation can be located at different levels. The boundaries are never sharply defined, but it is possible to distinguish three prototypical organisational levels on which standards are located with differing degrees of obligatoriness.

In the simplest case, a standard only applies to a single dictionary project or work team. Initially, such *ad-hoc standards* have little relevance outside a relatively narrow project context. They are used exclusively for working and organisational processes within a specific project and often undergo changes and adaptations in relation to the specific requirements of the project.

Standardising models and formats in larger project contexts necessitates agreement between different actors, who may have different requirements. Because of their larger community of users and, in particular, when they continue to be actively developed, they emerge as a *de-facto standard* in the field in which they are employed. De-facto standards often establish themselves by being implemented in a wide range of computer programs. The Multi-Dictionary Formatter format (MDF; → Section 4.4.4) used by linguists in the Shoebox/Toolbox working environment when they are undertaking fieldwork to document endangered languages is one example of this kind of development.

Finally, attempts at standardisation can be pursued on an international level and can culminate, for example, in the adoption of an ISO standard. Multinational consortia like the Unicode Consortium or the Text Encoding Initiative (→ Section 4.4.1) play a role similar to that of the International Organization for Standardization (ISO). One advantage of international standardisation lies in the associated convergence towards a stable standard. Models and formats are no longer subject to short-term changes because the standardisation process on this level takes a very long time. Another advantage is that the organisational structure of international standardisation bodies guarantees a reliable, long-term point of reference, which individual time-limited lexicographic projects are unable to provide in this form.

In the following sections, we present a selected number of lexicographic formats and models in more detail. In the process, we attempt to provide a cross-section of different types of dictionary, different groups of users, and different fields in which these dictionaries are used. Furthermore, the different standards are situated on different organisational levels. Nevertheless, we shall restrict ourselves to presenting formats and models for resources intended to be consulted by human users. We will not examine specialised dictionaries and lexical databases that are developed for automatic language processing applications.

4.4.1 Text Encoding Initiative

First formulated at the end of the 1980s and continuously developed since then, the guidelines for the Text Encoding Initiative (TEI) were conceived very generally from the outset, focusing on the standardised description of texts of any kind. These guidelines have detailed, ready-made ways of describing many different types of text. With their help, it is possible to model printed literature, handwritten texts, inscriptions on gravestones, transcribed dialogues, for example, with a high degree of accuracy. Thus, they can also be used to model dictionaries. Nowadays, the TEI guidelines are the most widely used text markup standard in the (digital) humanities, and there is a vast array of resources available in this form.

The main area where the TEI guidelines are applied in lexicography is in the retro-digitisation of existing print dictionaries from one of the three main perspectives identified in the guidelines: typographical, editorial, and lexical (cf. TEI 2023). Ideally, these different perspectives are modelled in such a way that they are cleanly separate from one another. However, the TEI model also allows for hybrid forms. Let us briefly elaborate on each of these three perspectives:

The *typographical perspective* reflects the surface form of a printed page that is determined by technical (typesetting) and typographical factors. It captures information on the fonts and emphasis used, on line breaks, and on the layout of areas of text on the page as well as further medium-specific properties of the actual two-dimensional representation.

The *editorial perspective* involves an abstraction from the two-dimensional positioning of textual symbols in that it constitutes a stream of letters, punctuation symbols, and possible processing instructions for a hypothetical typesetting process. Medium-specific artefacts from the typographical perspective (such as hyphenation at the end of a line) no longer occur in this textual model. The lexicographic information is thus modelled conceptionally as a one-dimensional sequence of symbols.

Just like the editorial perspective, the *lexical perspective* is an abstraction from the two-dimensional typographical perspective. With the help of a semantically determined inventory of categories, the lexical information is assigned to specific lexical categories. This results in a semantic annotation for each piece of lexicographic information. Furthermore, the relationship of one piece of information to another models the scope of this information and what it addresses. For example, the lexical perspective makes it possible to indicate exactly which entry each sense description belongs to or which citation is an attestation for a specific sense.

Below, the (a) typographical and (b) lexical perspectives will be compared in detail using as a starting point a short entry on the lemma *nachtlied* (night song) from the first edition of Jacob and Wilhelm Grimm's Deutsches Wörterbuch (German Dictionary, DWB-Online). While the typographical perspective reproduces many technical typesetting details (the comma after the lemma, the colon after the meaning paraphrase, the indent at the beginning of the entry, the line breaks, hyphens, and so on), no information is provided about the lexicographic status of individual sections of text – even the boundaries between pieces of information are not clearly recognisable by virtue of the markup (for example, between the indication of gender – "n." – and the beginning of the definition – "abends oder nachts gesungenes . . . lied"). By contrast, the purely lexical perspective does not indicate how to present the lexicographical information. Punctuation marks that delimit pieces of information have to be derived in a hypothetical typesetting process following rules from the sequence of information ("in the event that further information follows, a colon follows the definition"; "authors' names are set in small caps"). A normalisation of values can also take place. For example, the indication of gender in the lexical perspective appears in the form "neuter", while – again following rules – the form "n." is used in the hypothetical typesetting process in order to shorten the text. Finally, the lexical perspective can encode information that does not appear in print at all, as is the case with the indication of the word class "noun".

```
(a)
<hi rend="capitalized indented">nachtlied</hi>,
<hi rend="italics">n. abends oder nachts gesungenes oder zu
  sin-
<lb/>gendes lied:</hi> nachtlieder dichten. <hi
rend="smallcaps">Petr.</hi>
40ª ; wanderers nacht-<lb/>lied. <hi rend="smallcaps">Göthe
</hi> 1,109;
```

(b)

```
<entry>
  <form>
    <orth>nachtlied</orth>
    <gramGrp>
      <gen>neuter</gen>
      <pos>noun</pos>
    </gramGrp>
  </form>
  <sense>
    <def>abends oder nachts gesungenes
      oder zu singendes lied</def>
    <cit>
      <quote>nachtlieder dichten</quote>
      <bibl>
        <author>Petrarca</author>
        <biblScope>40ª</biblScope>
      </bibl>
    </cit>
    <cit>
      <quote>wanderers nachtlied</quote>
      <bibl>
        <author>Göthe</author>
        <biblScope>1,109</biblScope>
      </bibl>
    </cit>
```

The two perspectives each have their own specific fields of application. However, for lexicographic (and metalexicographic) work, only modelling from a lexical perspective is of use since it is capable of directly reflecting the inherent tree structure of dictionary entries (→ Section 4.2.1), which results from the relation between the entry's individual pieces of lexical information.

The inventory of concepts in the TEI is organised in a modular fashion. Because the TEI model can be adapted in very specific and far-reaching ways by those who use it and because it retains the option of subcategorisation, the inventory of categories can be extended practically at will. In the TEI world, such adaptations are called *customisations*.

One notable customisation that specifically aims at the representation of dictionaries is provided by the Lex-0 initiative (TEI Lex-0 2023). The main focus of TEI Lex-0 is on interoperability across different dictionaries and, thereby, on fostering tool reuse across lexical resources. This goal is pursued by streamlining the number of elements allowed in dictionary-specific contexts. For example, the different entry-like objects allowed in the general TEI framework (`entry` – general entry, `re` – related entry, `superEntry` – groups of entries, `entryFree` – unstructured entry, `hom` – homograph) are collapsed into a single `entry` object that may be used recursively and may be associated with a type attribute if needed. Other constraints introduced by the TEI Lex-0 guidelines concern attributes that are made mandatory as opposed to their optional status in the general framework of the TEI (e.g. the `id` attribute on `entry` and `sense` elements), or tighter restrictions for contexts in which certain elements may occur. We provide a serialisation in TEI Lex-0 for our toy example *disproof* below. Note the `xml:id` attribute on the `entry` and `sense` elements as well as the `type` attribute on the `gram` element – all of which are optional in the general framework but are obligatory in TEI Lex-0.

```
<entry xml:id="MED.disproof" xml:lang="en">
  <form type="lemma">
    <orth>disproof</orth>
    <pron>dɪs'pruːf</pron>
  </form>
  <gramGrp>
    <gram type="pos">n</gram>
  </gramGrp>
  <sense xml:id="MED.disproof.1" n="1">
    <def>facts that disprove something</def>
  </sense>
  <sense xml:id="MED.disproof.2" n="2">
    <def>the act of disproving</def>
  </sense>
  <xr type="related">
    <ref target="#MED.disprove" type="entry">disprove
    </ref>
  </xr>
</entry>
```

4.4.2 Lexical-semantic wordnets

The first large-scale lexical-semantic wordnet has been developed from the mid-1980s onwards at Princeton University under the name WordNet. It was originally conceived as a model of a section of linguistic knowledge inspired by psycholinguistics and cognitive science, namely the mental lexicon. Here, mental concepts that extend across individual languages are modelled (STONE, GO, RED), which are represented by *synsets* (collections of synonyms realised in individual languages: {rock, stone}, {go, go away, depart}, {red, reddish, ruddy, blood-red, . . . }). As such, wordnets belong to the category of onomasiological dictionaries, that is, dictionaries that assign linguistic forms to lexical meanings.

A variety of lexical-semantic relationships exist between synsets. Formally, a wordnet represents a graph for which the synsets form the set of nodes. The lexical-semantic relationships of the synsets between one another are produced through a series of relations across the collection of nodes. They can thus be conceived as the set of vertices for the graph. Such a graph is not necessarily connected; nor does it have to be free of loops. Fellbaum (1998) provides a good overview of the construction and many early applications of the English-language WordNet.

Wordnets have enjoyed great popularity up to the present time, especially in the context of computational linguistic applications. A wordnet provides a good foundation for the automatic semantic annotation and analysis of texts. If wordnets in different languages are interoperably modelled or translated into a common form of representation, this approach can be extended to cover different languages. Human users employ wordnets first and foremost as thesauri or as synonym dictionaries. Princeton's WordNet exists in two storage formats: a proprietary text-based database version (*lexicographer files*, see below) and as a Prolog knowledge base, a way of representing knowledge that has traditionally been used in the research field of artificial intelligence. Many subsequent monolingual wordnet projects have also used text-based database representations as an exchange format or have made proprietary XML-based formats available.

```
{  [ rock1, adj.all:rough^rocky,+  ] [ stone, adj.all:
  rough^stony,+ verb.contact:stone,+  ] noun.Tops:
  natural_object, (a lump or mass of hard consolidated
  mineral matter; “he threw a rock at me”) }
```

At the moment there is no single, standard format used by all wordnet projects. Nonetheless, WordNet-LMF does provide a suggested LMF model for wordnets and equivalence relations between synsets (→ Section 4.4.3; cf. also Soria et al. 2009). This suggested model was implemented as an example for a series of wordnet projects but has scarcely been adopted outside the original project context so far. It has therefore remained an example of an ad-hoc standard to date.

4.4.3 The Lexical Markup Framework – a model for all types of dictionaries

The Lexical Markup Framework (LMF) was adopted in 2008 as international standard ISO 24613:2008. This standard includes a modular metamodel to describe the actual models of a variety of types of lexical resources. The most important modelling principles are the consolidation of the elements on individual levels of linguistic description in modules (syntax, phonology, etc.) and the hierarchical arrangement of those units. In order to do justice to the issues discussed above concerning the theoretically informed genesis of a dictionary, LMF contains a reference mechanism which can be used to describe explicitly the semantics of lexical concepts. This is validated by reference to a further international standard (ISO 12620:2009), which describes a data category registry.

Using a range of examples, Romary/Wegstein (2012) demonstrate that, under certain prior assumptions, lexical modelling in the TEI framework can be understood as a realisation of the LMF model. Their core argument is the way the model is limited to "crystals", which form semantically autonomous units in an entry.

Since LMF has been available as an integrative, internationally standardised (meta)model, a series of specific formats derived from it have been proposed, for example: WordNet-LMF (→ Section 4.4.2), UBY-LMF (Eckle-Kohler et al. 2012), or the lemon lexicon model (McCrae et al. 2017). It remains to be seen whether one of these proposals does indeed develop into a de-facto standard format for the LMF model or whether reference to the common metamodel already suffices in order to represent lexical resources so that they are interoperable, i.e. they are able to communicate with one another. However, what we can conclude is that existing resources can demonstrably be modelled within the LMF model in many areas of electronic lexicography. Clearly LMF provides a sufficiently wide framework in order to model lexical resources of the most varied kinds (cf. Francopoulo 2013).

4.4.4 Toolbox and Multi-Dictionary Formatter

Toolbox is a computer program provided by SIL International for documenting and managing linguistic and, specifically, lexical data. It has been used especially by linguists working on the documentation of endangered languages for many years. Because of its widespread use in this group, the Multi-Dictionary Formatter (MDF) format used by Toolbox to store data represents a de-facto standard in this field of research. Users can employ a collection of around 100 lexicographic information types and also supplement this collection with custom types.

The MDF standard format is represented as an example below. The entry for the lemma *alabanja* is part of a dictionary of Iwaidja, an Australian language (presented in Ringersma/Drude/Kemp-Snijders 2010). Lexicographic information is introduced by field labels (in the example, among others, by: \lx – "lexeme", the form of the symbol for the lemma; \sn – "sense number", semantic classification mark; \ps – "part-of-speech”, word class; \de – “definition”). The lexical model underlying the MDF standard model is that of a semasiological dictionary, that is, a dictionary starting from lexical signs and assigning meanings to them.

```
\lx alabanja
\sn 1
\ps n
\de beach hibiscus. Rope for harpoons and tying up
  canoes is made from this tree species, and the
  timber is used to make \fv{larrwa} smoking pipes
\ge hibiscus
\re hibiscus, beach
\rfs 205,410; IE 84
\sd plant
\sd material
\rf Iwa05.Feb2
\xv alabanja alhurdu
\xe hibiscus string/rope
\sn 2
\ps n
\de short-finned batfish
...
```

The addressing of information remains, for the most part, implicit in the MDF format. Although the lexical categories are clearly identified, their relationships with one another are not. Individual conventionalised classification functions constitute an exception, such as those assigned to the \sn and \ps fields in the documentation. There is no explicit hierarchical categorisation of the entry. Using the different perspectives introduced above in our discussion of the TEI model (→ Section 4.4.1), the MDF format models a mixed form between the editorial and lexical perspectives. If we consider the main field in which the format is used, this becomes immediately clear. First, a typesetting process can be derived directly from the data since the information is already stored sequentially. The field labels then acquire the role of simple typographical processing instructions. Second, targeted access to lexicographic categories is made possible for linguists so that they can retrieve and analyse the data on the basis of specific linguistic phenomena that are addressable by the field labels.

4.5 Outlook

It is customary practice nowadays that standards are used in data modelling. For example, it is almost impossible to find a relatively large dictionary project that does not rely on the use of XML-based technology. However, the picture is slightly different when it comes to the application of the lexicographic standards or guidelines discussed in → Section 4.4. On the one hand, there are numerous initiatives and infrastructure projects working to promote and refine linguistic, lexicographic, and metadata standards, such as the European CLARIN and DARIAH consortia. On the other hand, the most important requirement for individual lexicographic projects is typically to develop a data model that best suits the needs of these projects. This often results in a data model tailored to a specific dictionary. Understandably, the applicability of the data model for everyday work within the project plays a crucial role – and is sometimes more important than data exchange and interoperability with other projects. It is always possible to transform a finely granulated, tailor-made data model into a representation using more general lexicographic standards, for example one that conforms to the TEI. Nonetheless, as discussed in → Section 4.3, this kind of conversion can, at times, be fraught with the loss of highly specific annotations due to generalisations imposed by the standard and also due to deviating interpretations of certain data categories. Thus, it remains to be seen to what extent international attempts at standardisation are embraced across the board.

The most compelling question for the future will be whether the highly granular markup of lexicographic content remains a prerequisite for data to be machine accessible in the first place. It is a long-standing belief in the lexicographic community that the granular, standard-based modelling of lexicographic data fosters their usefulness and applicability and ultimately leads to the data being much easier to process. Alas, for many tasks in the field of automatic natural language processing (NLP), the best results are often achieved by machine learning (ML) approaches based on manually annotated (lexical) data. After being trained on a high-quality standard-based dataset, the computer can then annotate, analyse, and retrieve unstructured, unannotated data (the supervised ML approach). Today, with the advent of Large Language Models (LLM), purely automatic, unsupervised ML approaches are gaining ground fast, i.e. algorithms that are based on current neural network techniques and trained on huge amounts of unstructured data. As such applications increasingly reduce the need for manually prepared data, lexicography might, in the long term, lose its significance in the field of the NLP. In fact, initial attempts to let LLMs create dictionary entries are already promising (Lew 2023). However, the central role of data modelling in "producing" innovative digital lexicographic tools and resources that can be analysed and understood by humans (as opposed to the black boxes that LLMs constitute) as well as its role in the sustainable archiving of lexicographic content remains unaffected by these developments for now.

To return to the Lego analogy from the beginning of this chapter: at the moment, it is (still) more effective to ensure that the red and blue bricks, and the 2 × 1s and 2 × 2s are labelled so that the computer can grab them in a targeted way. Perhaps at some point in the future, it will be more effective to either train the computer to identify the different types of bricks among the unsorted mass of Lego – or let the computer figure out the solution entirely on its own.

Bibliography

Further reading

DARIAH-Campus. Paris: DARIAH ERIC (Digital Research Infrastructure for the Arts and Humanities European Research Infrastructure Consortium). Online: https://campus.dariah.eu/. *Online open-source platform for learning resources on topics in the digital humanities*.

Lemnitzer, Lothar/Romary, Laurent/Witt, Andreas (2013): Representing human and machine dictionaries in Markup languages. In: Gouws, Rufus H., et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Computational Lexicography*. Berlin/Boston: De Gruyter, 1195–1208. *In-depth summary of XML-based lexicographic data modelling.*

Romary, Laurent (2011): Stabilizing knowledge through standards – A perspective for the humanities. In: Grandin, Karl (ed.): *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. New York: Science History Publications. *Good, accessible introduction to standardisation issues in relation to lexicography*. https://doi.org/10.48550/arXiv.1011.0519 [last access: April 27, 2024].

Literature

Academic literature

Abel, Andrea (2012): Dictionary writing systems and beyond. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 81–106.

Alexa, Melina (2011): Modellierung eines semantischen Wissensnetzes für lexikographische Anwendungen am Beispiel der Duden-Ontologie. In: Klosa, Annette/Müller-Spitzer, Carolin (eds.): *Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim, 61–70. https://pub.ids-mannheim.de/laufend/opal/pdf/opal2011-2.pdf [last access: April 27, 2024].

Eckle-Kohler, Judith, et al. (2012): UBY-LMF – A uniform model for standardizing heterogeneous lexical-semantic resources in ISO-LMF. In: *Proceedings of LREC 2012*. Istanbul, 275–282.

Fellbaum, Christiane (1998): *WordNet. An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Francopoulo, Gil (2013) (ed.): *LMF Lexical Markup Framework*. Oxford: Wiley.

Hämäläinen, Mika/Rueter, Jack (2018): Advances in Synchronized XML-media Wiki Dictionary Development in the Context of Endangered Uralic Languages. In: Čibej, Jaka, et al. (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts: 17–21 July 2018*, *Ljubljana*. Ljubljana: Ljubljana University Press, 967–978.

Kanngießer, Siegfried (1996): Zwei Prinzipien des Lexikonimports und Lexikonexports. In: Hötker, Wilfried/ Ludewig, Petra (eds.): *Lexikonimport, Lexikonexport. Studien zur Wiederverwertung lexikalischer Informationen*. Tübingen: Niemeyer.

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

Lew, Robert (2023): ChatGPT as a COBUILD lexicographer. In: *Humanities and Social Sciences Communications* 10, 705. https://doi.org/10.1057/s41599-023-02119-6 [last access: April 27, 2024].

McCrae, John Philip, et al. (2017). TheOntoLex-Lemon Model: development and applications. In: *Proceedings of eLex 2017*, 587–597.

Ringersma, Jacqueline/Drude, Sebastian/Kemp-Snijders, Marc (2010): *Lexicon standards: From de facto standard Toolbox MDF to ISO standard LMF. Talk presented at LRT standard workshop, LREC'2010*, Max Planck Institute for Psycholinguistics, Nijmegen/Goethe-Universität, Frankfurt. https://pure.mpg.de/rest/items/item_446072_8/component/file_446073/content [last access: April 27, 2024].

Robinson, Ian/Eifrem, Emil/Webber, Jim (2013): *Graph Databases*. Sebastopol, CA: O'Reilly & Associates.

Romary, Laurent/Wegstein, Werner (2012): Consistent Modeling of Heterogeneous Lexical Structures. In: *Journal of the Text Encoding Initiative [online]* 3. https://doi.org/10.4000/jtei.540 [last access: April 27, 2024].

Soria, Claudia/Monacchini, Monica/Vossen, Piek (2009): Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In: *Proceedings of IWIC*, Stanford.

Spohr, Dennis (2012): *Towards a Multifunctional Lexical Resource. Design and Implementation of a Graph-based Lexicon Model*. Berlin/Boston: De Gruyter.

Wiegand, Herbert Ernst (1989): Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Hausmann, Franz Josef/Reichmann, Oskar/Wiegand, Herbert Ernst (eds.): *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. 1. Teilbd*. Berlin/New York: De Gruyter, 409–462.

Dictionaries

DDRS = *Duden – Die deutsche Rechtschreibung*: *Das umfassende Standardwerk auf der Grundlage der aktuellen amtlichen Regeln. 28*., *völlig neu bearbeitete und erweiterte Auflage*. Berlin 2020: Bibliographisches Institut.

DWB-Online = Deutsches Wörterbuch von Jacob und Wilhelm Grimm online. In: *Wörterbuchnetz des Trier Center for Digital Humanities/Kompentenzzentrum für elektronische Erschließungs- und Publikationsverfahrens in den Geisteswissenschaften an der Universität Trier*. https://woerterbuchnetz.de/DWB/ [last access: April 27, 2024].

DWDS = *Das Digitale Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. https://www.dwds.de/ [last access: April 27, 2024].

elexiko = Online-Wörterbuch zur deutschen Gegenwartssprache. In: *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. http://www.elexiko.de/ [last access: April 27, 2024].

WordNet = *WordNet*. Princeton, NJ: Princeton University. https://wordnet.princeton.edu/ [last access: April 27, 2024].

Internet Sources

CLARIN = *Common Language Resources and Technology Infrastructure*. https://www.clarin.eu/ [last access: April 27, 2024].

DARIAH = *Digital Research Infrastructure for the Arts and Humanities*. https://www.dariah.eu/ [last access: April 27, 2024].

DBPEDIA = *DBpedia Open Knowledge Graph*. https://www.dbpedia.org/ [last access: April 27, 2024].

SIL = *SIL International*. https://www.sil.org/ [last access: April 27, 2024].

TEI (2023) = TEI Consortium (2023) (eds.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.7.0. Last updated November 16*, *2023*, *revision e5dd73ed0*. TEI Consortium. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html [last access: April 27, 2024].

TEI LEX-0 (2023) = Tasovac, Toma/Romary, Laurent, et al. (2023): *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.2. DARIAH Working Group on Lexical Resources*. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html [last access: April 27, 2024].

UNICODE = *The Unicode Consortium*. Online: https://www.unicode.org/.

WIKIPEDIA = *Wikipedia*, *the free Encyclopaedia*. https://www.wikipedia.org/ [last access: April 27, 2024].

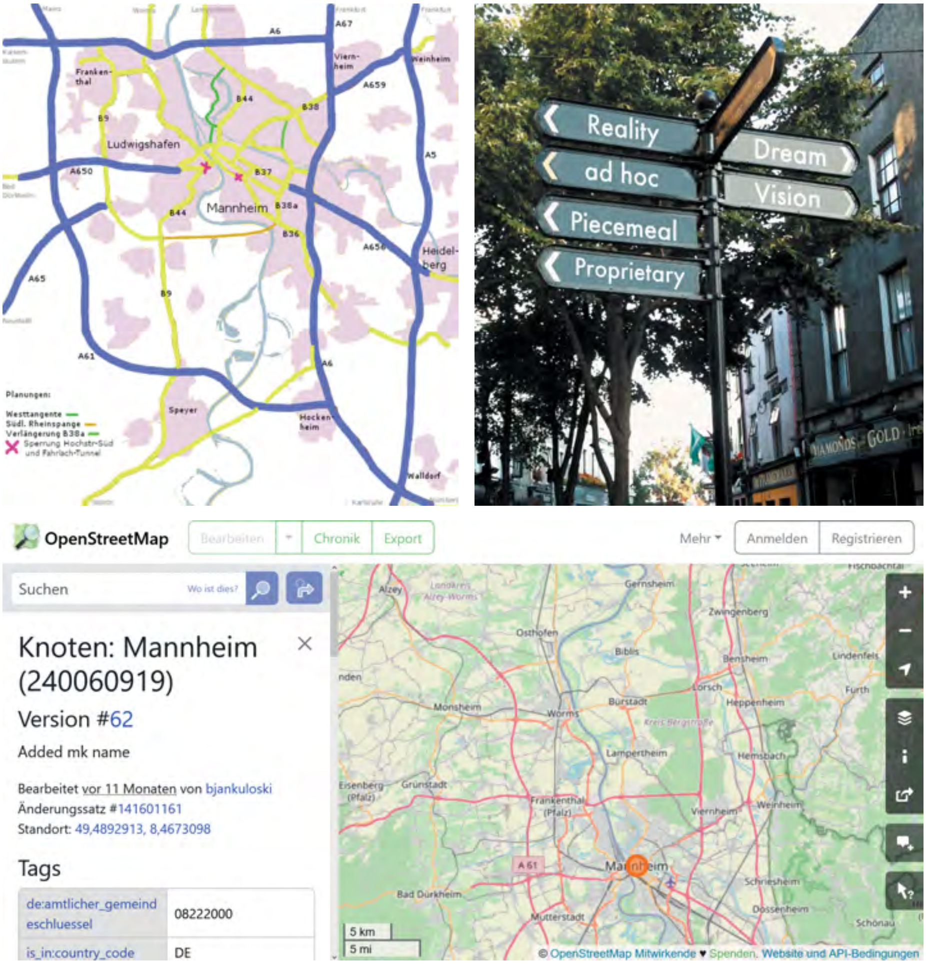
WIKIDATA = *Wikidata Knowledge Base*. https://www.wikidata.org/ [last access: April 27, 2024].

Images

Fig. 4.1 private.

Stefan Engelberg, Carolin Müller-Spitzer, and Thomas Schmidt

5 Linking and Access Structures



Knoten: Mannheim (240060919)

Version #62

Added mk name

Bearbeitet vor 11 Monaten von bjankuloski
Änderungssatz #141601161
Standort: 49,4892913, 8,4673098

Tags

de:amtlicher_gemeindeschluessel	08222000
is_in:country_code	DE

Fig. 5.1: Maps, town plans, and street signs.

Stefan Engelberg, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: engelberg@ids-mannheim.de
Carolin Müller-Spitzer, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: mueller-spitzer@ids-mannheim.de
Thomas Schmidt, LinguisticBits GmbH, Nahestraße 28, 55411 Bingen, Germany, e-mail: thomas@linguisticbits.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/10.1515/9783111233758-006

Lexicographic content interlinked within and between Internet dictionaries can be thought of as a network of streets. The streets connect different pieces of content in Internet dictionaries, thus forming the digital street network between different dictionary entries. The links in the user interface of a dictionary, from one headword to its associated synonyms for example, play an important role here as the main signposts by which users arrive, ideally, directly at their destination. Admittedly, this is not quite the same as reading a signpost in a street (cf. also Blumenthal et al. 1988). If you want to look up a place in an atlas, this can be done very conveniently online nowadays with the search function in digital maps. Direct access to dictionary content works in the same way. Here, there is a wide range of options for dictionary users to access individual pieces of lexicographic content.

5.1 Introduction

This chapter describes linking and access structures in Internet dictionaries. Linking refers to the navigation routes through a dictionary created by lexicographers. Hence, in many language dictionaries, headwords that can be used as synonyms in certain contexts are linked to one another, such as *smart* with *intelligent* or *bright*. This linking of content is mostly realised as hyperlinks, through which users are able to arrive directly at the destination of the networked connection. There have always been cross-references in print dictionaries; what is new in Internet dictionaries is that, in the best case, we only have to click once in order to reach our goal, instead of spending time leafing through pages. Whether in print or online, it is important for users of dictionaries to be able to find particular information in a direct way and as quickly as possible. Indeed, what all reference works have in common is that they are not read in a linear way, but that information is sought selectively. In this regard too, the digital medium offers a whole spectrum of possibilities.

This chapter is intended to provide insights into the whole field of linking and access structures. In the process, our aim is not to provide an exhaustive overview but rather to demonstrate, by way of example, which basic possibilities exist. In the first section, we explain what can be understood by linking in Internet dictionaries and how the level of data management differs from the presentational level. In the second part, we present the options for both semasiological (→ Section 5.3.1) and onomasiological access structures (→ Section 5.3.2). Finally, in → Section 5.4, we show what new impulses electronic cross-linking and access structure can offer for modern dictionary research.

5.2 Linking structures

The vocabulary of a language does not consist of individual words that exist as independent units, detached from one another. Rather, all of the elements of the lexicon are interconnected in multiple ways. Some words are used frequently with one another (like *dog* and *leash* or *smart* and *choice*); they can have (almost) the same meaning (like *smart* and *intelligent*) or are typically used in particular constructions (like *to make a smart move*). Yet, this web of words and the connections between them are difficult to represent in a general language dictionary, especially in two-dimensional print space. For that reason, the practice has developed over the centuries that in so-called semasiological dictionaries, the graphical form of individual words forms the access structure: that is, if you want to know something about the meaning of *smart move*, you know that, as a rule, you should look under either *smart* or *move* – at least in a general, monolingual print dictionary. In this way, the content relationships between the words are depicted by cross-references between individual dictionary entries (cf. Nielsen 1999; Engelberg/Lemnitzer 2009: 177f.). This type of organisation is not necessarily the "natural order of things" but rather a form of cultural practice.

While semasiological dictionaries start from individual words or groups of words, onomasiological dictionaries sit at the opposite end of the spectrum as they proceed from concepts or objects. For this kind of reference work, an alphabetically arranged index of words has to provide access to the concepts, at least if the work exists in print form. In Internet dictionaries, these different ways of accessing content are generally implemented digitally as search options. Hence, we will return to the distinction between semasiological and onomasiological access structures again in → Section 5.3.

The content-based relationships in the lexicon are represented in a language dictionary through cross-references based on the dictionary object. The term 'dictionary object' refers to the language and subsection of the language described by the dictionary (cf. Engelberg/Lemnitzer 2009: 272). These cross-references arise very frequently in print lexicography since, for reasons of space, some information is only marked in one place in the dictionary, even though it would be relevant in many places (cf. Wiegand 2002: 173). These formal cross-referencing requirements should occur relatively rarely in Internet dictionaries since the space for presenting data is substantially less restricted. Another type of cross-references is based on the intended dictionary functions (cf. Tarp 1999; Wiegand 2001).

All aspects of cross-referencing phenomena in print dictionaries are treated under the heading of *mediostructure* in dictionary research (Wiegand/Smit 2013). For digital dictionaries, we talk more generally of the linking structure (Müller-Spitzer 2007: 169f.; Meyer 2014). As a rule, the mediostructure of print dictionaries is analysed by inspecting example entries from one or more dictionaries.[1] The basis of the data for this kind of analysis is a printed book from which information is gathered and classified by reading and cognitive analysis. The analysis of dictionary structure often proceeds in a similar way for digital dictionaries.[2] For example, this kind of research analyses which types of cross-references appear in a particular dictionary and how these are presented in the user interface, etc. However, it can also proceed in a completely different way when the basis of the data is formed of the entire digital database of a dictionary and when this data is evaluated using statistical methods. We show a brief example of this in our "Outlook" (→ Section 5.4).

The prerequisites for how many cross-references can be presented in an Internet dictionary are set in its data modelling (→ Chapter 4). Already at the end of the 1980s, the two different levels – data modelling and presentation – were illustrated in an essay using the analogy of maps vs road signs (Blumenthal et al. 1988: 356f.). In this analogy, data modelling is equivalent to drawing a map, that is, to defining, on an abstract level, which elements can be linked at all. Individual cross-references in the actual dictionary are then the individually placed signposts.

Cross-references are mostly rendered by links on the user interface. For the most part, we do not distinguish between the terms *link* (the element of an Internet site that can be activated) and *hyperlink* (the connection between this element and other content, managed by the computer). However, when analysing linking structures, it is often useful to be able to distinguish terminologically between these two uses of a link. For this purpose, we use the term *link marker* for the element that can be activated on the presentational level, the term *link target* for the element the link marker points to, and the term *link relation* for the computer connection between the content (text) units on the data modelling level. The link relation is not visible on the user interface. On the presentational level, we can only see the link markers that can be activated with a mouse, a keyboard, or a touchpad/touchscreen in order to call up other units of information (for further information, cf. Storrer 2013).

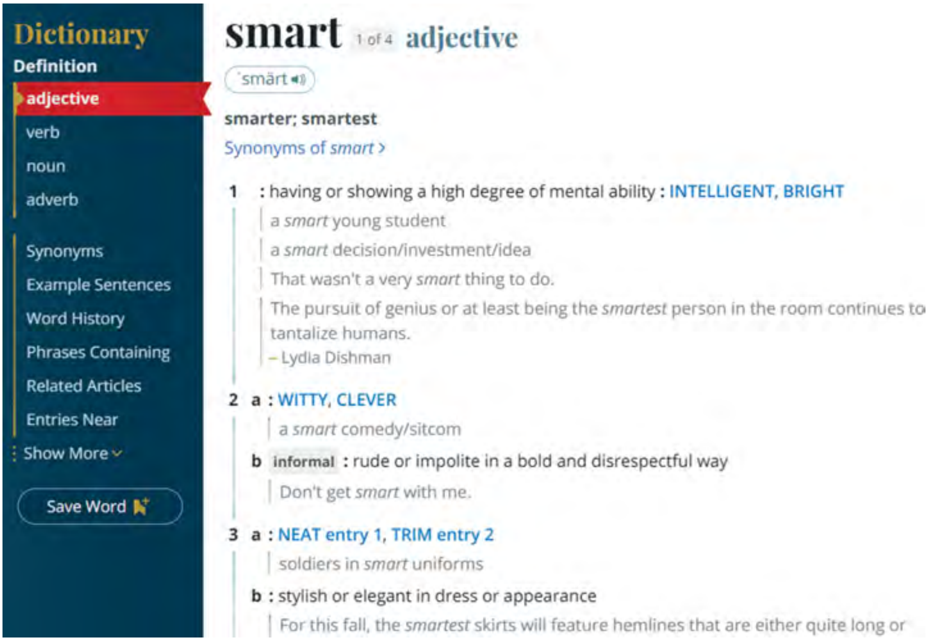
Consider an example. In the article *smart* in MERRIAM-WEBSTER, various link markers to synonyms can be found under the different subsenses of the word (and under the heading "Synonyms of smart"). In this dictionary, these are rendered in small capitals and in blue font colour. Underneath the keyword, however, there is another form of link marker: a loudspeaker icon, which takes the user to audio examples of the keyword. As such, link markers need not necessarily be units of written language; other graphical elements can also function as link markers. Various types of data also come into play as link targets in Internet dictionaries – text and images as well as audio and video data. In the online entry for *smart*, there are entries in the left sidebar with yet another form of

[1] Cf. Kammerer (1998: 325); for other examples of this kind of study, cf. Lindemann (1999) or Müller (2002).

[2] Cf., for example, Mann (2010: 28f., 36f.); on questions of the possible transferability of concepts, cf., among others, Tarp (2008: 102) and Müller-Spitzer (2013).

link: “synonyms” and “example sentences”, etc. will each lead the user to different groups of information pertaining to the headword. These kinds of links will be referred to as *structural links* and belong to the so-called internal access structure: in other words, they serve to provide access to individual parts of the word entry. By contrast, we refer to cross-references to synonyms, audio examples, or translations as *content links* since the connection is motivated by properties of the dictionary object.

At least in the form in which they are shown in → Fig. 5.2, structural links are part of the internal access structure of an entry. The external structure in Internet dictionaries is realised through search functions. These are the subject of the following section.



smart 1 of 4 adjective

ˈsmärt

smarter; smartest

Synonyms of *smart* ›

1 : having or showing a high degree of mental ability : INTELLIGENT, BRIGHT

a *smart* young student

a *smart* decision/investment/idea

That wasn't a very *smart* thing to do.

The pursuit of genius or at least being the *smartest* person in the room continues to tantalize humans.
– Lydia Dishman

2 a : WITTY, CLEVER

a *smart* comedy/sitcom

b informal : rude or impolite in a bold and disrespectful way

Don't get *smart* with me.

3 a : NEAT entry 1, TRIM entry 2

soldiers in *smart* uniforms

b : stylish or elegant in dress or appearance

For this fall, the *smartest* skirts will feature hemlines that are either quite long or

Fig. 5.2: Link markers in the field for related words in the entry for the lemma *smart* in Merriam-Webster.

5.3 Access structures

The linking structures described in the previous section make it possible for dictionary users to use a cross-reference to move from any given entry in the dictionary to another with which the former has a connection. However, a variety of access structures stand at the user’s disposal to facilitate their access to the dictionary “from the outside” so that they can find relevant entries in the first place. Typically, this involves different kinds of searches.

In a print dictionary, there are two principal types of search: first, a semasiological search in an alphabetical dictionary by means of successively flicking through pages, forwards and backwards, until the location is found; second, an onomasiological search shaped from a content perspective, for example, in a hierarchically structured ontology. This division is reflected in the two following sections, which present the access structures in Internet dictionaries.

The digital medium and digital methods for processing lexical information multiply the possible ways of accessing dictionary content. Some of these new access structures cannot be classified unambiguously as either semasiological or onomasiological. These are the subject of → Section 5.3.3.[3]

5.3.1 Semasiological access structures

In the following, we characterise the different types of searches in Internet dictionaries. To this end, lexicographical Internet searches are considered according to four criteria, each of which relate to aspects of the search action. These four aspects are (1) the starting point of the search action, (2) the type of search action, (3) the complexity of the search action, and (4) the target of the search action (→ Fig. 5.3).

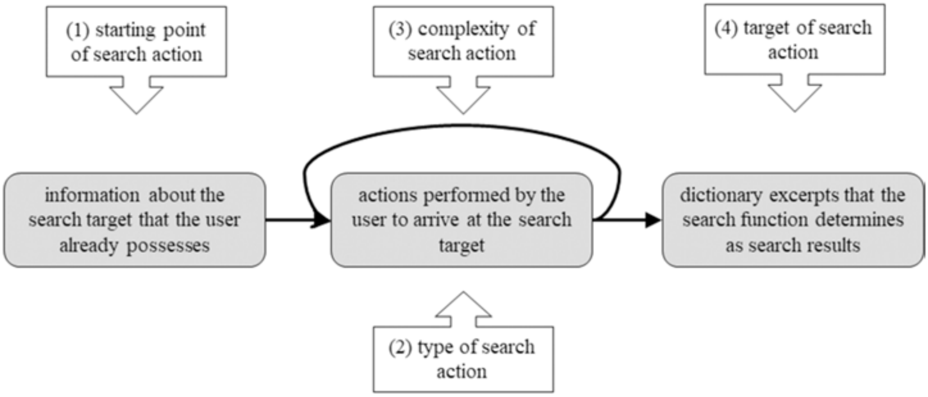

(1) starting point of search action
(3) complexity of search action
(4) target of search action
information about the search target that the user already possesses
actions performed by the user to arrive at the search target
dictionary excerpts that the search function determines as search results
(2) type of search action

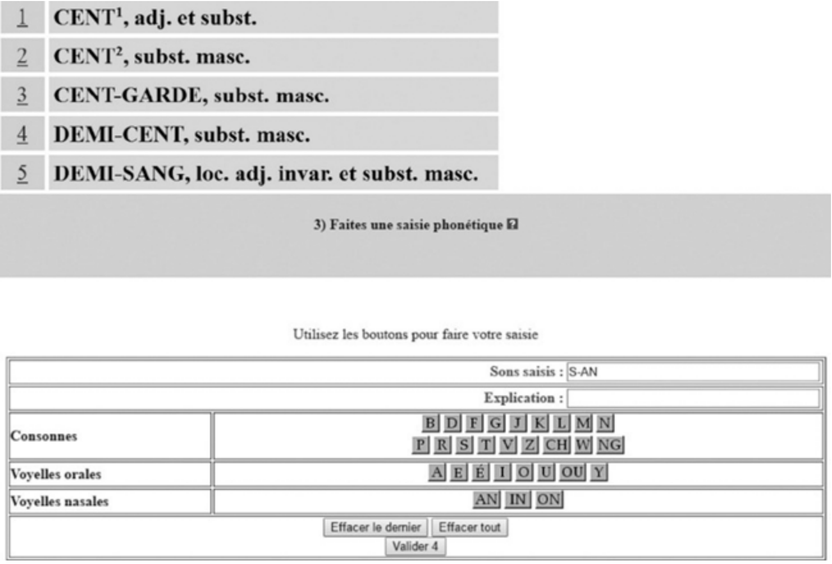
Fig. 5.3: The search action.

(1) *Starting point of the search action.* In order to find a needle in a haystack, you can take the haystack apart, stalk by stalk, until the needle turns up. However, searching in a dictionary does not normally involve such a time-consuming path through the whole

[3] For overviews of search functions in electronic dictionaries cf. Engelberg/Lemnitzer (2009: 99f.), Lew (2012), Dziemianko (2018: 667ff.), Pastor/Alcina (2022), Klosa-Kückelhaus/Michaelis (2022: 416f.). See Giacomini (2015) on access structures in LSP lexicography.

search space; rather, it takes as its starting point certain information about the goal of the search that is already at the user's disposal. This information could relate to the written form of the lemma symbol that is being sought, its sound form, its (intensional) meaning, or its typical objects of reference (extensional meaning). As is the case for a print dictionary, a search that starts with the written or spoken form of a linguistic sign is referred to as a semasiological search, and one that starts with the intensional or extensional meaning of a linguistic sign is referred to as an onomasiological search. → Section 5.3.2 is devoted to the latter; here, we concentrate on semasiological searches.

Searches by written form are implemented in as good as every Internet dictionary and represent by far the most common form of dictionary search. In what follows they are presented in detail. Conversely, the option to search by spoken form is realised much more rarely (cf. Lew 2012: 346; Dziemianko 2018: 669). In principle, the latter can take two forms: in a search based on phonetic transcription, the user chooses the transcription symbols (e.g., IPA) that correspond to the sound form of the lemma symbol as the search term; in a speech-input search, the search proceeds from the inputting of spoken language which is then processed by a speech recognition module. A transcription-based search is possible, for example, in the Trésor de la langue française informatisé (TLFi) (→ Fig. 5.4).



1 CENT[1], adj. et subst.

2 CENT[2], subst. masc.

3 CENT-GARDE, subst. masc.

4 DEMI-CENT, subst. masc.

5 DEMI-SANG, loc. adj. invar. et subst. masc.

3) Faites une saisie phonétique

Utilisez les boutons pour faire votre saisie

Sons saisis : S-AN

Explication :

Fig. 5.4: Search based on sound form in the TLFi.

Voice input options are now widely used in all kinds of systems, such as speech-to-text conversion or automatic translation, and they are also used in lexicographic products, especially in dictionary apps for mobile devices.

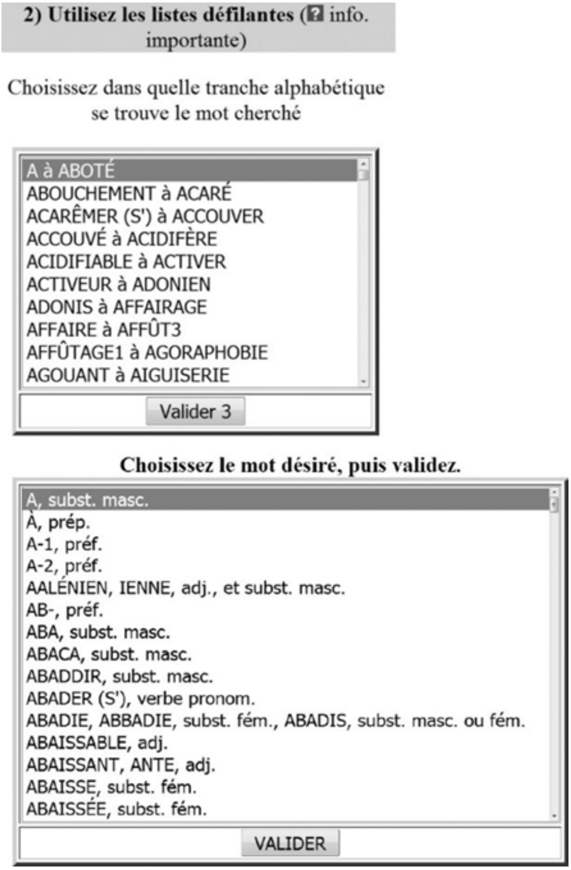
(2) *Type of search action*. Basic search actions are oriented towards the medium and are familiar from other Internet-based forms of communication. Above all, they are based on inputting text, clicking on links, or moving the cursor. The basic lexicographic Internet search actions include:

- typing in a search term (input-based search);
- clicking on a linguistic expression, for which a corresponding dictionary entry can be found (index-based search);
- clicking on a selection field or making a selection from a drop-down menu in order to limit the volume of hits (filter-based search);
- reading a linguistic expression for which a corresponding dictionary entry can be found, for example, using the scanning function of a mobile phone (scanner-based search);
- the spoken inputting of a search term already mentioned above (speech input search).

There are some particular aspects of input-based, index-based, and filter-based searches that are worth mentioning now. An input-based search by means of typing into a search field is often supported by a series of specific options:

- When a search term begins to be entered, suggestions are made to complete it that can be selected; these correspond to the characters already entered and to lemmas in the dictionary (incremental search, type-ahead search) (cf. Engelberg et al. 2020: 64f.; Lew 2012: 351f.).
- An option is available to decide whether the search should be case sensitive or not (case-sensitive search).
- In order to offer a suitable target search term for users who are uncertain about spelling, lemmas are shown that are similar phonetically or graphemically to the search term (fuzzy search, spelling-tolerant, or phonetic search) (cf. Engelberg/Lemnitzer 2009: 106f.; Lew 2012: 347f.).
- Parts of the search term are kept variable by certain operators; these placeholders can stand for individual letters or for a chain of letters (placeholder search); in this way, for example, all the entries can be found that describe lemma symbols with particular morphological elements, such as all words ending in the German suffix -*ung* or all words with the component part -*moon*- (cf. Pastor/Alcina 2022: 96).
- The inflected form of a word is entered into the search field, which leads back to the root form by automatic lemmatisation and for which the corresponding lemma is then sought (search by inflected form) (cf. Pastor/Alcina 2022: 97).

An index-based search involves lemmas being searched by means of lemma lists and lemma range indicators.[4] Searching in lemma lists usually involves navigating through moving lemma lists, in which the required lemma can be chosen by clicking. Navigating in lemma lists is often supported by lemma range indicators, i.e., letter bars or lemmas listed by their start sequences, which limits the range of lemmas within which the required lemmas can be located (→ Fig. 5.5). Here, the search often involves successive navigation from wider to narrower ranges of lemmas. At the end of navigation via lemma range indicators, there is normally a section of a lemma list within which the required lexeme can be found.



2) Utilisez les listes défilantes (? info. importante)

Choisissez dans quelle tranche alphabétique se trouve le mot cherché

A à ABOTÉ
ABOUCHEMENT à ACARÉ
ACARÊMER (S') à ACCOUVER
ACCOUVÉ à ACIDIFÈRE
ACIDIFIABLE à ACTIVER
ACTIVEUR à ADONIEN
ADONIS à AFFAIRAGE
AFFAIRE à AFFÛT3
AFFÛTAGE1 à AGORAPHOBIE
AGOUANT à AIGUISERIE

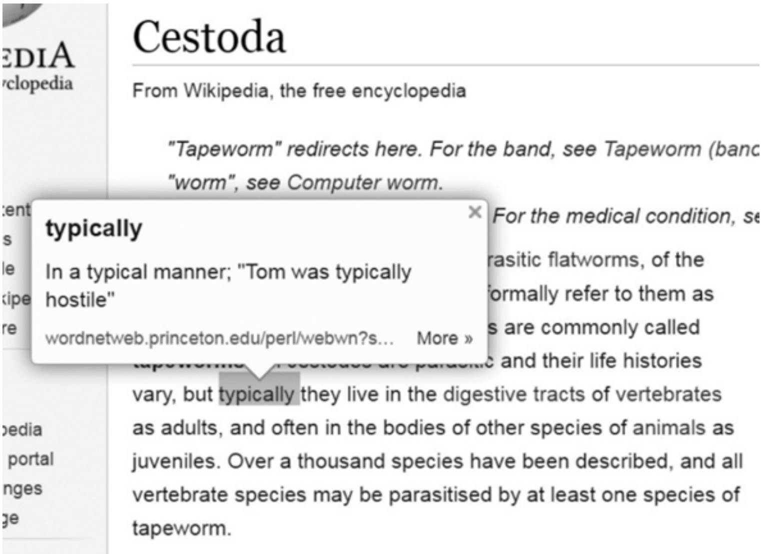
Choisissez le mot désiré, puis validez.

A, subst. masc.
À, prép.
A-1, préf.
A-2, préf.
AALÉNIEN, IENNE, adj., et subst. masc.
AB-, préf.
ABA, subst. masc.
ABACA, subst. masc.
ABADDIR, subst. masc.
ABADER (S'), verbe pronom.
ABADIE, ABBADIE, subst. fém., ABADIS, subst. masc. ou fém.
ABAISSABLE, adj.
ABAISSANT, ANTE, adj.
ABAISSE, subst. fém.
ABAISSÉE, subst. fém.

Fig. 5.5: Lemma range indicator in the TLFi.

[4] A lemma range is an uninterrupted sequence of entries in a dictionary. They can be referred to in the form of lemma range indicators, e.g., by giving the first and last lemma of the range.

Navigating by clicking on particular expressions is also the basis for two other forms of search. In a text-based search, it is not lemmas that are clicked from the dictionary's lemma list but rather words from electronic texts external to the dictionary (→ Fig. 5.6). Then, potentially following automatic lemmatisation, the clicked word is matched against the dictionary's lemma list. In this way, the user can call up a dictionary entry directly from the text editor or text display. The scanner-based search mentioned above is also a form of text-based search. In the ideal case, connecting a text-based search with context-sensitive analysis even makes it possible to identify the specific interpretation of the word (Seretan/Wehrli 2013).



Cestoda

From Wikipedia, the free encyclopedia

"Tapeworm" redirects here. For the band, see Tapeworm (banc
"worm", see Computer worm.
For the medical condition, se

typically

In a typical manner; "Tom was typically hostile"

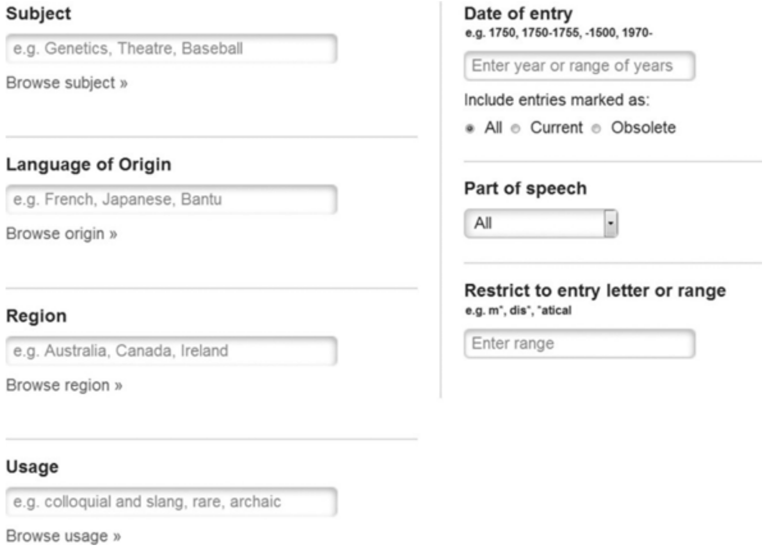
wordnetweb.princeton.edu/perl/webwn?s… More »

rasitic flatworms, of the
ormally refer to them as
s are commonly called
c and their life histories
vary, but typically they live in the digestive tracts of vertebrates as adults, and often in the bodies of other species of animals as juveniles. Over a thousand species have been described, and all vertebrate species may be parasitised by at least one species of tapeworm.

Fig. 5.6: Text-based search in the GOOGLE dictionary (starting from a WIKIPEDIA article).

A filter-based search is particularly suitable when it is not an individual word that is being sought but rather a number of lemmas, sublemmas, compound words listed for a lemma within a dictionary entry, or semantically related words. This makes it possible to filter out those lemmas with particular properties (formal, semantic, etymological). Here, the search process can include clicking on checkboxes or selecting from a drop-down menu (→ Fig. 5.7). A particular type of filter-based searches is a faceted search. It allows a progressive refinement of a search using one filter after the other while the search output is continuously reduced (cf., e.g., Porta-Zamorano 2018: 926f., Engelberg et al. 2020: 61f.).

(3) *Complexity of the search action*. One-dimensional search actions only require a single one of the search processes outlined above, or a short sequence of them: that is, entering a single search term, clicking one lemma in a lemma list, or applying a single filter. Multidimensional search actions, in contrast, combine several individual actions simul-



Subject

e.g. Genetics, Theatre, Baseball

Browse subject »

Language of Origin

e.g. French, Japanese, Bantu

Browse origin »

Region

e.g. Australia, Canada, Ireland

Browse region »

Usage

e.g. colloquial and slang, rare, archaic

Browse usage »

Date of entry
e.g. 1750, 1750-1755, -1500, 1970-

Enter year or range of years

Include entries marked as:

● All ○ Current ○ Obsolete

Part of speech

All

Restrict to entry letter or range
e.g. m*, dis*, *atical

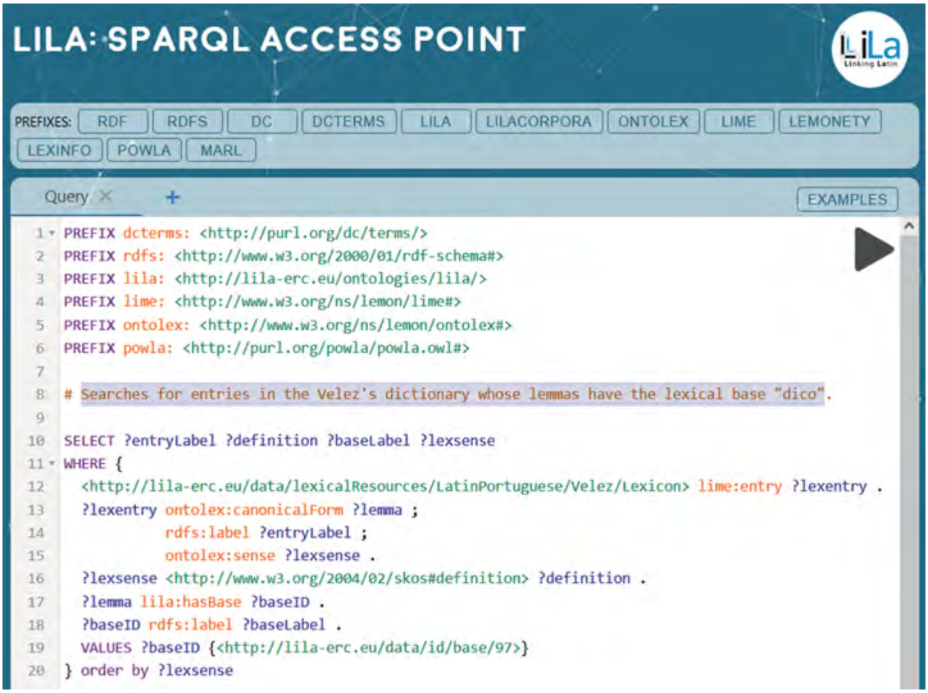
Enter range

Fig. 5.7: Filter-based search in the Oxford English Dictionary (OED) via the selected entry of a term (e.g., “subject”, “region”), selecting a radio button (e.g., “all/current/obsolete”), or making a selection in a drop-down menu (“part of speech”).

taneously into a complex search query. For the most part, they do not serve to locate a single lemma but rather a number of expressions that satisfy particular criteria. This applies, for example, to the “advanced search” in the OED (→ Fig. 5.7).

Individual academic language platforms sometimes allow searches in dictionaries using query languages like SPARQL or CQP, e.g., BabelNet or LiLa, a knowledge base of linguistic resources for Latin (→ Fig. 5.8).

(4) *Target of the search action.* The target of a search action can be a specific lemma, a number of lemmas, or a particular information item in one or more dictionary articles: for example, all of the sense items whose paraphrase contains a particular content word. The most common case is, indisputably, a search for an individual lemma and its associated dictionary entry. Most one-dimensional search actions lead to this kind of result. Conversely, complex search actions, and also some simple placeholder or filter searches, serve for the most part to identify a number of lemmas that satisfy particular syntactic (→ Fig. 5.9), morphological, semantic-pragmatic (→ Fig. 5.21), etymological, or other criteria. Searches of this kind lead either directly to a dictionary entry or to a lemma, which is then clicked to reach the entry.


LILA: SPARQL ACCESS POINT
LiLa Linking Latin
PREFIXES: RDF RDFS DC DCTERMS LILA LILACORPORA ONTOLEX LIME LEMONETY LEXINFO POWLA MARL
Query
EXAMPLES

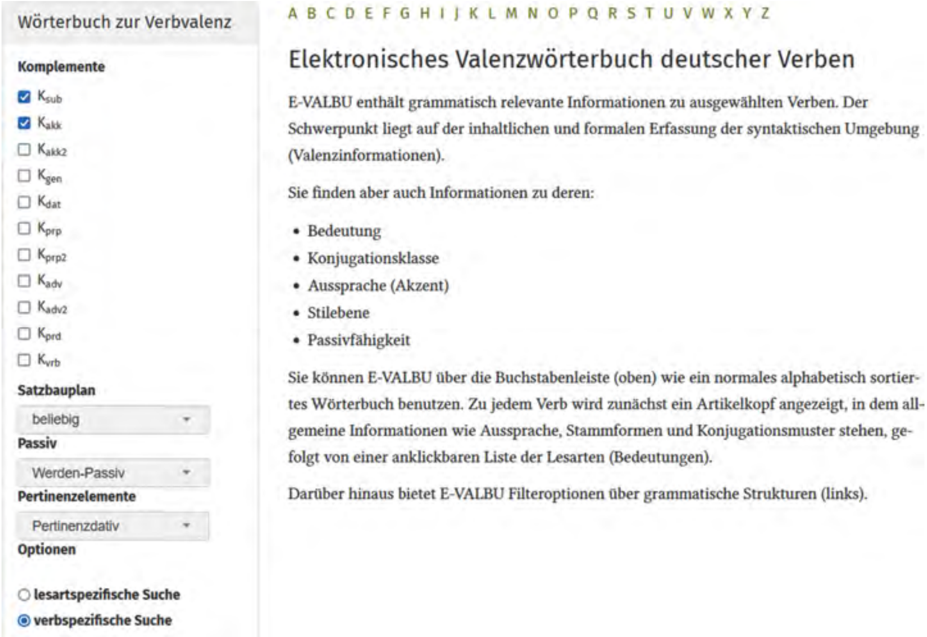
```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

# Searches for entries in the Velez's dictionary whose lemmas have the lexical base "dico".

SELECT ?entryLabel ?definition ?baseLabel ?lexsense
WHERE {
  <http://lila-erc.eu/data/lexicalResources/LatinPortuguese/Velez/Lexicon> lime:entry ?lexentry .
  ?lexentry ontolex:canonicalForm ?lemma ;
            rdfs:label ?entryLabel ;
            ontolex:sense ?lexsense .
  ?lexsense <http://www.w3.org/2004/02/skos#definition> ?definition .
  ?lemma lila:hasBase ?baseID .
  ?baseID rdfs:label ?baseLabel .
  VALUES ?baseID {<http://lila-erc.eu/data/id/base/97>}
} order by ?lexsense
```

Fig. 5.8: Searches in the Latin knowledge base LiLa for entries in one of the included dictionaries whose lemmas have the lexical base “dico”, using SPARQL as a query language.

5.3.2 Onomasiological access structures

Semasiological dictionary access proceeds from linguistic forms and leads to information about the meaning and use of these forms. By contrast, *onomasiological dictionary access* has its starting point in a meaning (an idea, a concept, a piece of content) and refers to associated linguistic forms. Onomasiological access structures can be helpful for productive dictionary use, for example, when the dictionary is being used to help write a text, but also when a language learner wants to open up and explore a section of foreign language vocabulary or specialised terminology.

As a rule, onomasiological access structures exist in addition to semasiological structures, that is, as a complement or supplement to them: printed illustrated dictionaries normally contain an alphabetical keyword index that cross-refers back from the linguistic form to the content depicted in pictorial form. Digital dictionaries open up new, extended ways to provide onomasiological access structures. For one thing, being liberated from the print form facilitates notably more flexible forms of presentation. If dictionary data are first modelled separately from their form of presentation, according to purely content-based aspects (→ Chapter 4), the individual components of the dictionary



Wörterbuch zur Verbvalenz

Komplemente

☑ K_{sub}
☑ K_{akk}
□ K_{akk2}
□ K_{gen}
□ K_{dat}
□ K_{prp}
□ K_{prp2}
□ K_{adv}
□ K_{adv2}
□ K_{prd}
□ K_{vrb}

Satzbauplan

beliebig

Passiv

Werden-Passiv

Pertinenzelemente

Pertinenzdativ

Optionen

○ **lesartspezifische Suche**
● **verbspezifische Suche**

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Elektronisches Valenzwörterbuch deutscher Verben

E-VALBU enthält grammatisch relevante Informationen zu ausgewählten Verben. Der Schwerpunkt liegt auf der inhaltlichen und formalen Erfassung der syntaktischen Umgebung (Valenzinformationen).

Sie finden aber auch Informationen zu deren:

- Bedeutung
- Konjugationsklasse
- Aussprache (Akzent)
- Stilebene
- Passivfähigkeit

Sie können E-VALBU über die Buchstabenleiste (oben) wie ein normales alphabetisch sortiertes Wörterbuch benutzen. Zu jedem Verb wird zunächst ein Artikelkopf angezeigt, in dem allgemeine Informationen wie Aussprache, Stammformen und Konjugationsmuster stehen, gefolgt von einer anklickbaren Liste der Lesarten (Bedeutungen).

Darüber hinaus bietet E-VALBU Filteroptionen über grammatische Strukturen (links).

Fig. 5.9: Multidimensional filter-based search in the E-VALBU ("Electronic Valency Dictionary of German Verbs"); the search is for all verbs that require an obligatory accusative object in addition to a subject and that also allow a dative of possession and a *werden*-passive.

can be assembled and (re)organised according to any criteria at all for presentational purposes.[5] In this way, one or more onomasiological access options (e.g., in the form of an image or a hierarchically organised ontology) can be set alongside an alphabetical lemma list (as the classic semasiological access structure), both of which point to the same dictionary entries. For another, the multimedia capabilities of computers open up new possibilities for presenting and illustrating the content aspects of an expression for the user. In addition to static images, which could already be used as the starting point for an onomasiological approach to accessing a dictionary in the print medium (albeit at a relatively high cost), moving images (video clips) can also be integrated into the dictionary in the digital medium to illustrate an action or audio data (audio clips) to illustrate sounds.

When it comes to semasiological access structures (→ Section 5.3.1), orthography acts as a system familiar to almost all dictionary users for representing linguistic forms. This system is not only standardised as widely as possible and applicable across the entire lexicon (every word has an orthographic form) but it also includes a distinct system for order-

[5] Meyer/Tu (2021) show how an onomasiological search can be implemented post hoc based on existing word senses and multilingual pre-trained word embeddings.

ing different forms by placing them in relationships with one another (the alphabet and the classification of root forms and inflected forms). This is fundamentally different for onomasiological access structures. First of all, it is not at all obvious how a given meaning (an idea, a concept, a piece of content) can be presented as the starting point for onomasiological access on the part of dictionary users, and there is no distinct system by which the different meanings can be organised exhaustively and put into relationships with one another. While pictures, for example, might often be a suitable way of representing concrete objects, involving the part-whole relationship (partonymy) as an inherent organisational system (→ Fig. 5.10), the meanings of more complex actions (e.g., "exmatriculate") or more abstract content (e.g., "shy") cannot be illustrated well through images.

The basis of onomasiological access structures is thus more diverse and less clearly defined than for semasiological structures; furthermore, any given onomasiological access structure often does not cover the whole lexicon but only the part of it for which that particular form of representing meaning is well suited. Fillmore (1978) argues that it can be entirely adequate to select the access structures in this way, dependent on "semantic domains":

> I think that semantic theory must reject the suggestion that all meanings need to be described in the same terms. I think, in fact, that semantic domains are going to differ from each other according to the kind of 'definitional base' which is most appropriate to them. (p. 148)

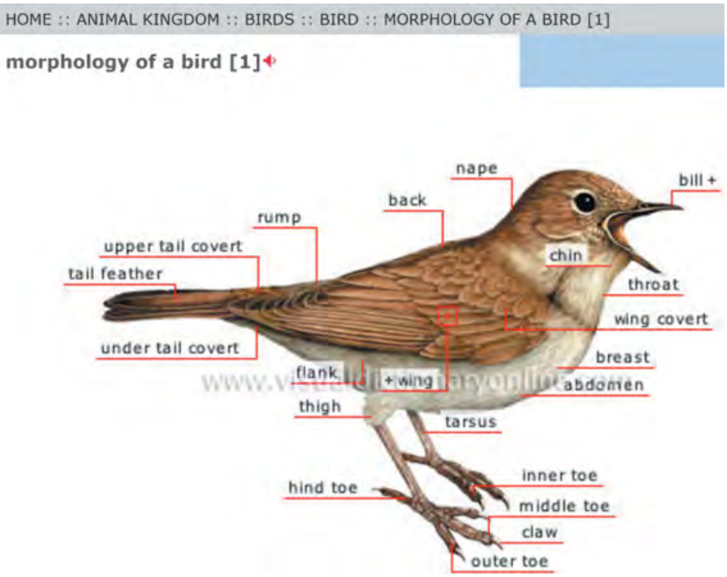
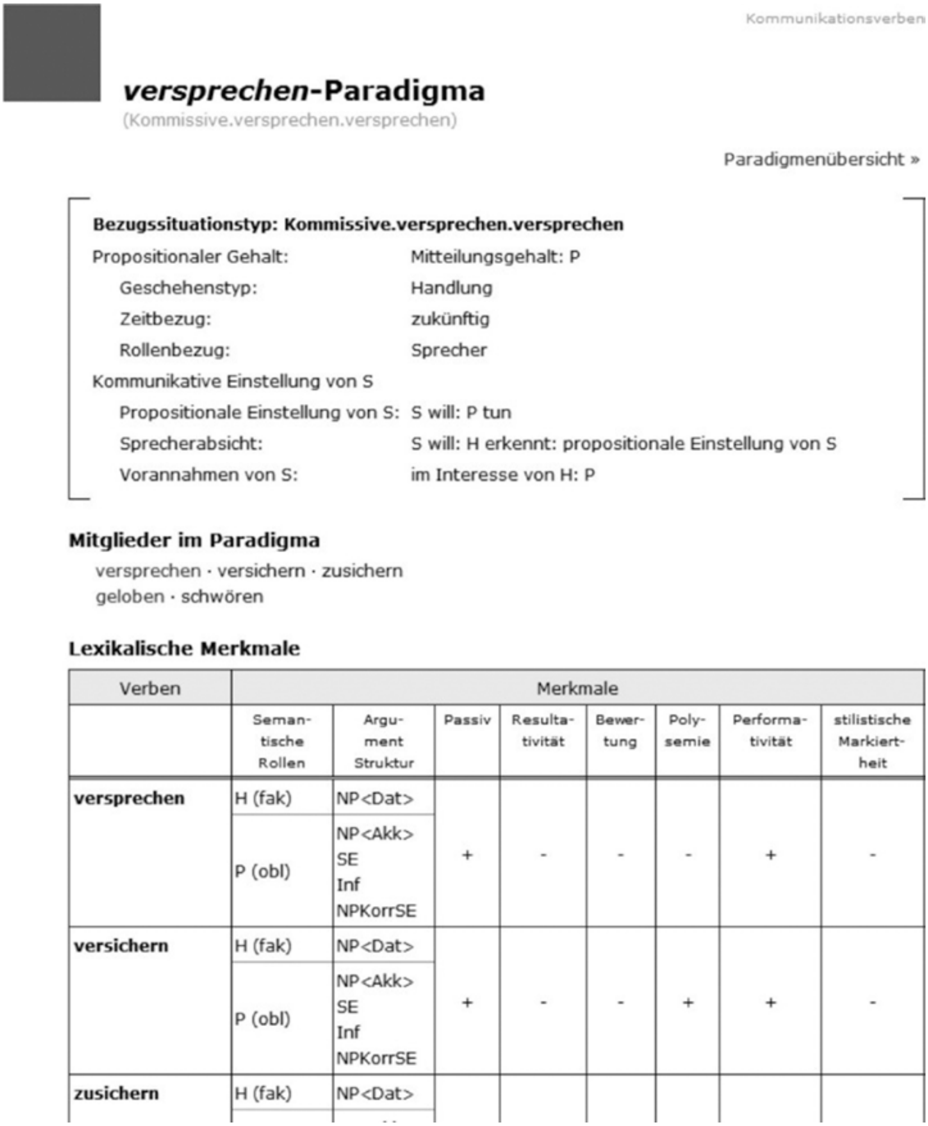


Fig. 5.10: *Bird* in the Merriam-Webster Visual Dictionary (MWVDO).

As far as the presentation of meanings for onomasiological access is concerned, we can initially draw an essential distinction between linguistic and non-linguistic forms.

When it comes to linguistic representation, (intensional) meanings are described by linguistic forms: for example, if verbs – as in the dictionary of verbs of communication (KOMMUNIKATIONSVERBEN) in OWID (→ Fig. 5.11) – are collected into paradigms listed according to the semantically dominating verb, if terms relating to linguistically named concepts are assigned to an ontology (see below) (cf. Pastor/Alcina 2022: 113f.),



Kommunikationsverben

versprechen-Paradigma

(Kommissive.versprechen.versprechen)

Paradigmenübersicht »

Bezugssituationstyp: Kommissive.versprechen.versprechen

Propositionaler Gehalt:	Mitteilungsgehalt: P
Geschehenstyp:	Handlung
Zeitbezug:	zukünftig
Rollenbezug:	Sprecher
Kommunikative Einstellung von S	
Propositionale Einstellung von S:	S will: P tun
Sprecherabsicht:	S will: H erkennt: propositionale Einstellung von S
Vorannahmen von S:	im Interesse von H: P

Mitglieder im Paradigma

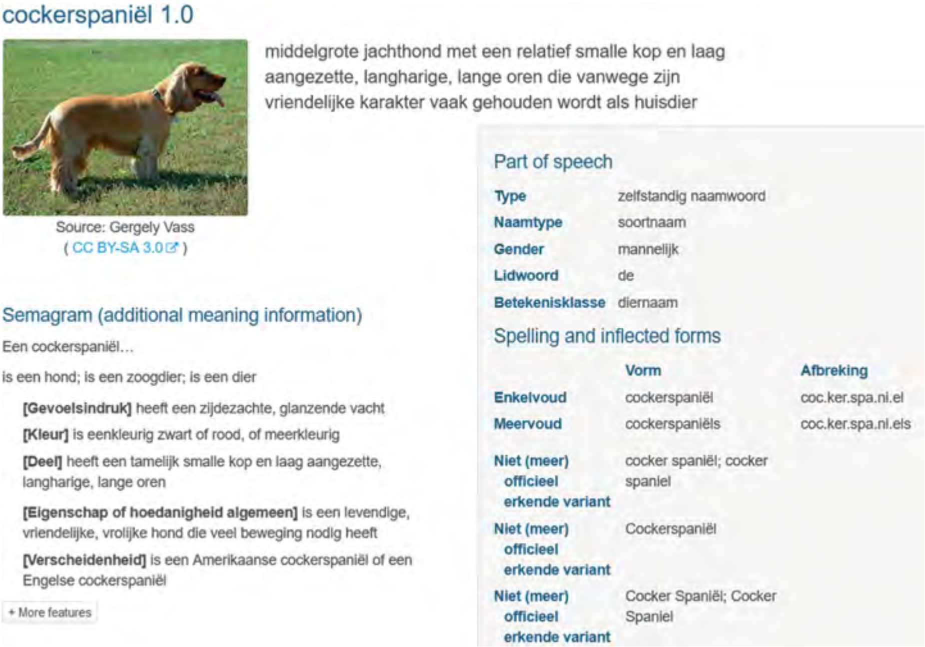
versprechen · versichern · zusichern
geloben · schwören

Lexikalische Merkmale

Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
versprechen	H (fak)	NP<Dat>	+	-	-	-	+	-
	P (obl)	NP<Akk> SE Inf NPKorrSE						
versichern	H (fak)	NP<Dat>	+	-	-	+	+	-
	P (obl)	NP<Akk> SE Inf NPKorrSE						
zusichern	H (fak)	NP<Dat>						

Fig. 5.11: Verbs belonging to the paradigm of verbs of promise (German: *versprechen*) in the dictionary of verbs of communication (KOMMUNIKATIONSVERBEN) in OWID containing information about their valency and semantic-pragmatic features.

or if – as in the Algemeen Nederlands Woordenboek (ANW; → Fig. 5.12) – the meaning of a lexeme is described by semagrams with linguistically named properties. Meanings are also represented in linguistic form in a full-text search in a dictionary, which, as outlined above, can equally be seen as an onomasiological form of access.



cockerspaniël 1.0

Source: Gergely Vass
(CC BY-SA 3.0)

middelgrote jachthond met een relatief smalle kop en laag aangezette, langharige, lange oren die vanwege zijn vriendelijke karakter vaak gehouden wordt als huisdier

Semagram (additional meaning information)

Een cockerspaniël…

is een hond; is een zoogdier; is een dier

[Gevoelsindruk] heeft een zijdezachte, glanzende vacht

[Kleur] is eenkleurig zwart of rood, of meerkleurig

[Deel] heeft een tamelijk smalle kop en laag aangezette, langharige, lange oren

[Eigenschap of hoedanigheid algemeen] is een levendige, vriendelijke, vrolijke hond die veel beweging nodig heeft

[Verscheidenheid] is een Amerikaanse cockerspaniël of een Engelse cockerspaniël

+ More features

Part of speech

Type	zelfstandig naamwoord
Naamtype	soortnaam
Gender	mannelijk
Lidwoord	de
Betekenisklasse	diernaam

Spelling and inflected forms

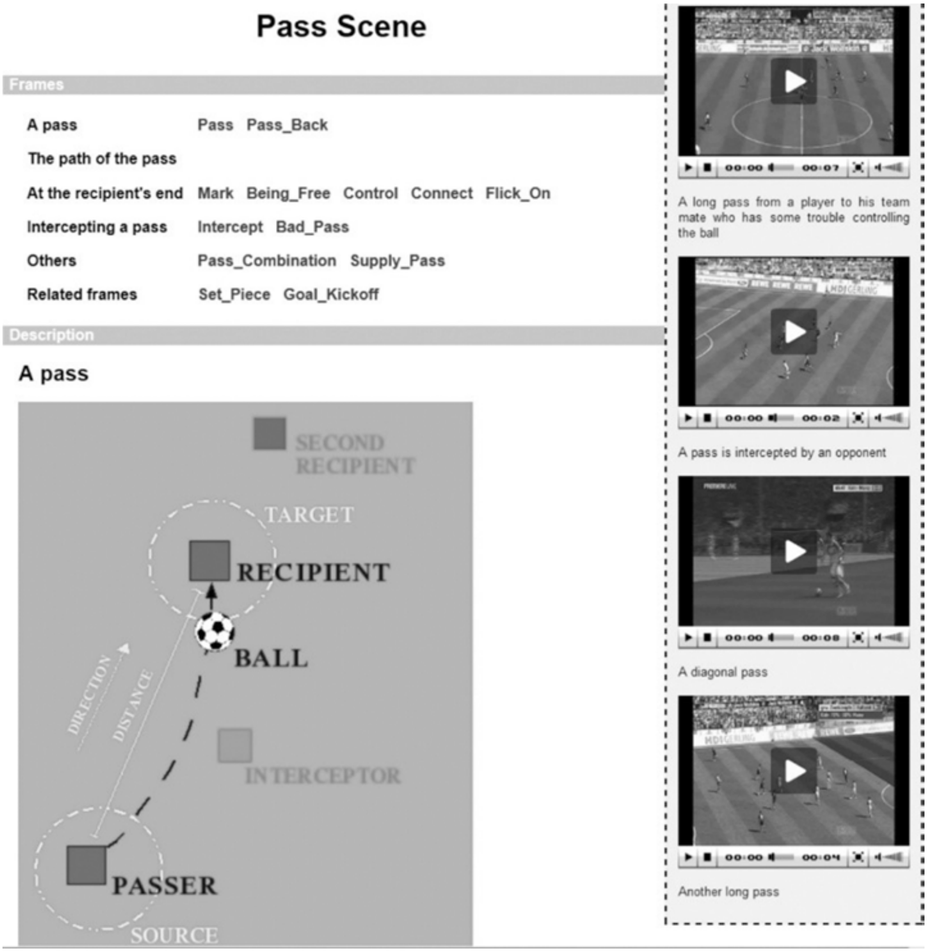
	Vorm	Afbreking
Enkelvoud	cockerspaniël	coc.ker.spa.ni.el
Meervoud	cockerspaniëls	coc.ker.spa.ni.els
Niet (meer) officieel erkende variant	cocker spaniël; cocker spaniel	
Niet (meer) officieel erkende variant	Cockerspaniël	
Niet (meer) officieel erkende variant	Cocker Spaniël; Cocker Spaniel	

Fig. 5.12: Semagram for Dutch *cockerspaniël* ('cocker spaniel') in the ANW.

By contrast, images serve to describe (extensional) meanings in a non-linguistic representation. Examples for this kind of illustration-based representation can be found in → Fig. 5.10 and → Fig. 5.12, in which a typical reference object is represented for each in either a drawing (*bird*) or a photograph (*cockerspaniël*). Schematic drawings or moving images (or potentially sounds) are other conceivable methods for representing or illustrating meanings in non-linguistic form. For example, Kicktionary (→ Fig. 5.13) makes use of diagrams and video clips, among other things, in order to show users the meaning of actions ("scenes") in football matches.

However, in order to facilitate onomasiological access to a dictionary, it is not sufficient to make individual meanings available as the starting point for locating linguistic forms. Rather, these individual meanings have to be organised and related to one another in a comprehensible way so that the user is in a position to find them in the first place as the starting point for an onomasiological search in the dictionary.

In terms of the form of this organisation, we can distinguish between hierarchical and non-hierarchical structures and between top-down and bottom-up processes for



Pass Scene

Frames

A pass	Pass Pass_Back
The path of the pass	
At the recipient's end	Mark Being_Free Control Connect Flick_On
Intercepting a pass	Intercept Bad_Pass
Others	Pass_Combination Supply_Pass
Related frames	Set_Piece Goal_Kickoff

Description

A pass

A long pass from a player to his team mate who has some trouble controlling the ball

A pass is intercepted by an opponent

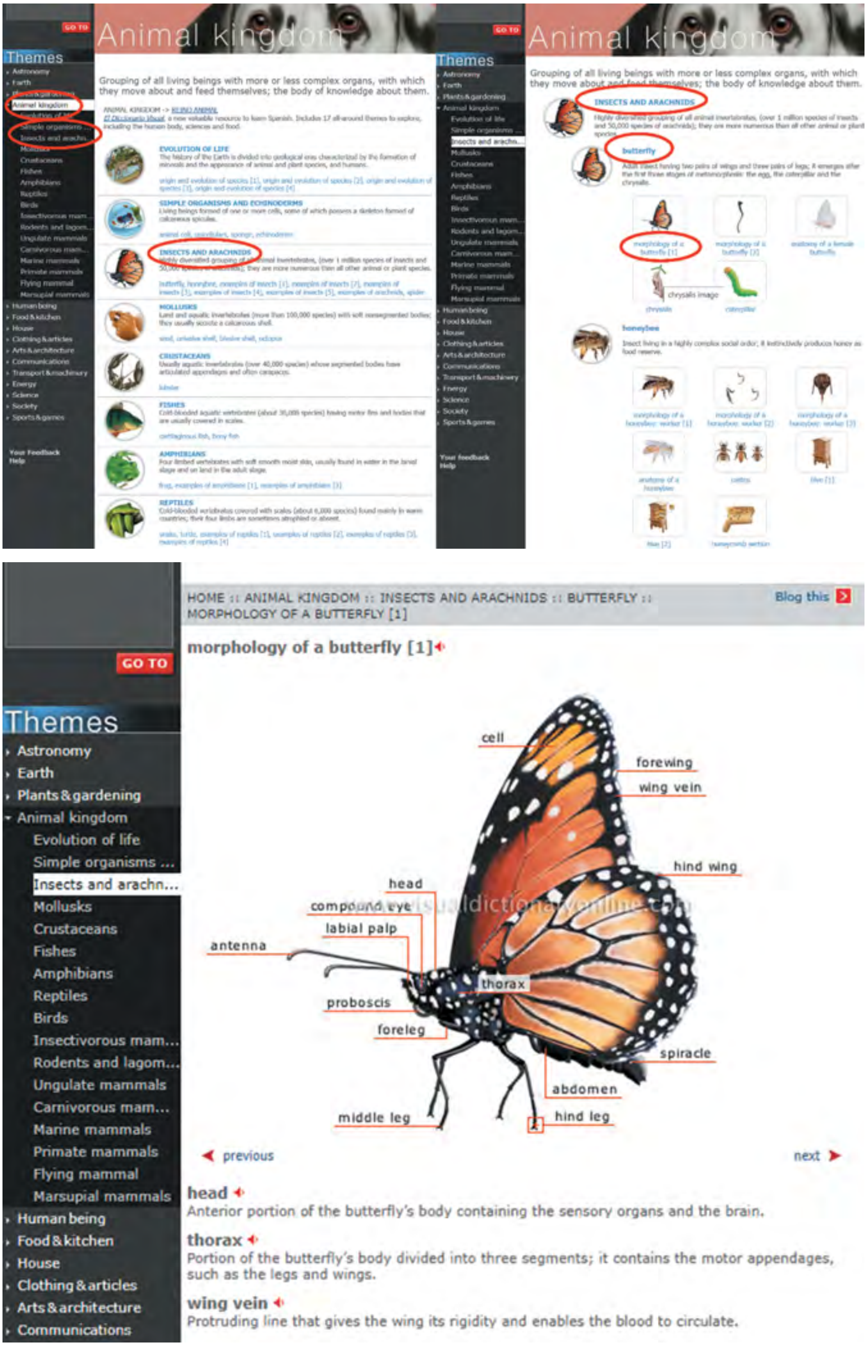
A diagonal pass

Another long pass

Fig. 5.13: “Pass scene” in KICKTIONARY.

constructing them. In the following, this will be explained using four examples of onomasiological access structures.

Example 1 (Pictorial dictionary MWVDO): The basic components of onomasiological access (that is, images or linguistic signs, etc. that stand for a given meaning) are often organised in a hierarchical structure. For example, the Merriam-Webster Visual Dictionary (MWVDO) (→ Fig. 5.14) initially starts with 17 different thematic areas that are then each subdivided into further subareas on multiple levels (here: animal kingdom > insects and arachnids > butterfly > morphology of a butterfly) until the actual linguistic forms appear as the caption for an image at the lowest level.



Grouping of all living beings with more or less complex organs, with which they move about and feed themselves; the body of knowledge about them.

ANIMAL KINGDOM -> REINO ANIMAL
El Diccionario Visual, a new valuable resource to learn Spanish. Includes 17 all-around themes to explore, including the human body, sciences and food.

EVOLUTION OF LIFE
The history of the Earth is divided into geological eras characterized by the formation of minerals and the appearance of animal and plant species, and humans.
origin and evolution of species [1], origin and evolution of species [2], origin and evolution of species [3], origin and evolution of species [4]

SIMPLE ORGANISMS AND ECHINODERMS
Living beings formed of one or more cells, some of which possess a skeleton formed of calcareous spicules.
animal cell, unicellulars, sponge, echinoderm

INSECTS AND ARACHNIDS
Highly diversified grouping of all animal invertebrates, (over 1 million species of insects and 50,000 species of arachnids); they are more numerous than all other animal or plant species.
butterfly, honeybee, examples of insects [1], examples of insects [2], examples of insects [3], examples of insects [4], examples of insects [5], examples of arachnids, spider

MOLLUSKS
Land and aquatic invertebrates (more than 100,000 species) with soft nonsegmented bodies; they usually secrete a calcareous shell.
snail, univalve shell, bivalve shell, octopus

CRUSTACEANS
Usually aquatic invertebrates (over 40,000 species) whose segmented bodies have articulated appendages and often carapaces.
lobster

FISHES
Cold-blooded aquatic vertebrates (about 30,000 species) having motor fins and bodies that are usually covered in scales.
cartilaginous fish, bony fish

AMPHIBIANS
Four limbed vertebrates with soft smooth moist skin, usually found in water in the larval stage and on land in the adult stage.
frog, examples of amphibians [1], examples of amphibians [2]

REPTILES
Cold-blooded vertebrates covered with scales (about 6,000 species) found mainly in warm countries; their four limbs are sometimes atrophied or absent.
snake, turtle, examples of reptiles [1], examples of reptiles [2], examples of reptiles [3], examples of reptiles [4]

Grouping of all living beings with more or less complex organs, with which they move about and feed themselves; the body of knowledge about them.

INSECTS AND ARACHNIDS
Highly diversified grouping of all animal invertebrates, (over 1 million species of insects and 50,000 species of arachnids); they are more numerous than all other animal or plant species.

butterfly
Adult insect having two pairs of wings and three pairs of legs; it emerges after the first three stages of metamorphosis: the egg, the caterpillar and the chrysalis.

morphology of a butterfly [1]
morphology of a butterfly [2]
anatomy of a female butterfly
chrysalis image
chrysalis
caterpillar

honeybee
Insect living in a highly complex social order; it instinctively produces honey as food reserve.

morphology of a honeybee: worker [1]
morphology of a honeybee: worker [2]
morphology of a honeybee: worker [3]
anatomy of a honeybee
castes
hive [1]
hive [2]
honeycomb section

HOME :: ANIMAL KINGDOM :: INSECTS AND ARACHNIDS :: BUTTERFLY :: MORPHOLOGY OF A BUTTERFLY [1]

head
Anterior portion of the butterfly's body containing the sensory organs and the brain.

thorax
Portion of the butterfly's body divided into three segments; it contains the motor appendages, such as the legs and wings.

wing vein
Protruding line that gives the wing its rigidity and enables the blood to circulate.

Fig. 5.14: Hierarchical construction of a pictorial dictionary (exemplified by Merriam-Webster Visual Dictionary; MWVDO).

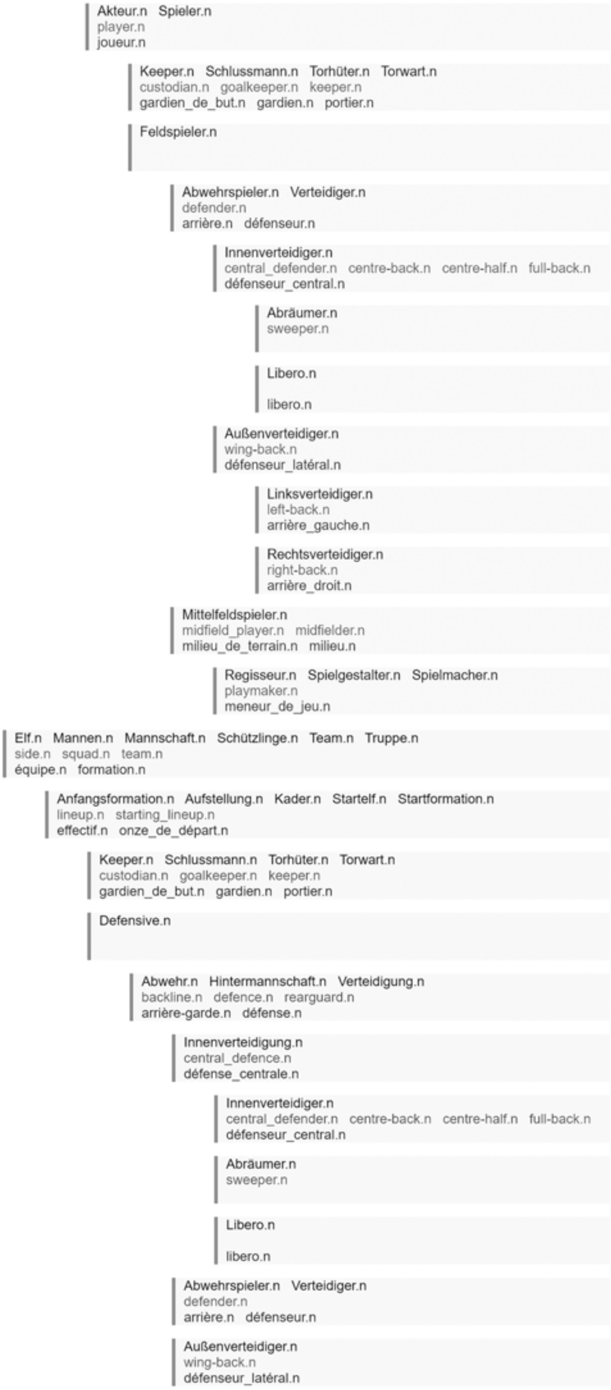
Example 2 (Semantic Relations in Kicktionary): The so-called concept hierarchies in Kicktionary are also organised hierarchically; however, meanings are not represented pictorially but directly through synonyms or the linguistic forms of translation equivalents. The relationship between the individual entries in the hierarchy is a semantic relation like the ones used in the organisation of wordnets (e.g., WordNet or GermaNet) (Schmidt 2009).

In this way, there is synonymy between entries from the same language at a particular level, such as *goalkeeper, keeper, custodian* (all = 'goalkeeper'). The entirety of synonymous forms is referred to as a SynSet (→ Chapter 4.4.2) and represents the meaning they have in common. In Kicktionary this principle also extends across languages: along with {*Torwart, Torhüter, Schlussmann*} for German and {*gardien de but, gardien, portier*} for French, the result is a multilingual SynSet that stands for the meaning (the "concept") 'goalkeeper'.

Further semantic relations can exist between SynSets, which then lead to the hierarchies that are depicted in the dictionary. The hierarchy shown in → Fig. 5.15 is based on the semantic relation of hyponymy (or its converse, hyperonymy), which denotes the relationship between a subordinate and superordinate term – if X is a type of Y (a goalkeeper is a player, a sweeper is a defender), then X is a hyponym of Y, and the SynSet containing X is subordinate to the SynSet containing Y. The hierarchy shown below in → Fig. 5.15 is based on the semantic relation of partonymy (converse: holonymy), which denotes a part-whole relationship. If X is a part of Y (a goalkeeper is part of the lineup, the lineup is part of the team), then X is a partonym of Y. In this way, a dictionary user can start with a meaning and arrive at various linguistic forms that denote this meaning, and they can also navigate in the relevant hierarchy to find linguistic forms which have a related (i.e., more general or more specific) meaning.

Example 3 (Frames in the Berkeley FrameNet): A notably more complex onomasiological organisation is used in dictionaries based on frames. Here, the frame is the starting point for the dictionary's structure – a structure in which knowledge about prototypical courses of action and their actors and objects is represented.

For example, the frame *Commerce buy* from FrameNet in → Fig. 5.16 provides a structure in which different linguistic expressions to do with buying (*buy*, *purchase*, *buyer*) can be organised. The definition explains the relevant action in an abstract way and specifies the so-called frame elements involved (in this case, among others, a buyer, a seller, goods, and money). The description of individual linguistic elements ("lexical units", e.g., the verb *buy*) can then have recourse to this superordinate structure, for example by annotating the frame elements with corresponding labels in an example sentence. In this way, different linguistic forms can be assigned to a common meaning, thereby facilitating onomasiological access. Additional possible forms of dictionary navigation arise because individual frames are assigned to one another in frame-to-frame relations. For example, the frame *Rent* constitutes a special case of the frame *Commerce_buy* and thereby "inherits" its properties. Likewise, *Commerce_buy*



Akteur.n Spieler.n
player.n
joueur.n

Keeper.n Schlussmann.n Torhüter.n Torwart.n
custodian.n goalkeeper.n keeper.n
gardien_de_but.n gardien.n portier.n

Feldspieler.n

Abwehrspieler.n Verteidiger.n
defender.n
arrière.n défenseur.n

Innenverteidiger.n
central_defender.n centre-back.n centre-half.n full-back.n
défenseur_central.n

Abräumer.n
sweeper.n

Libero.n

libero.n

Außenverteidiger.n
wing-back.n
défenseur_latéral.n

Linksverteidiger.n
left-back.n
arrière_gauche.n

Rechtsverteidiger.n
right-back.n
arrière_droit.n

Mittelfeldspieler.n
midfield_player.n midfielder.n
milieu_de_terrain.n milieu.n

Regisseur.n Spielgestalter.n Spielmacher.n
playmaker.n
meneur_de_jeu.n

Elf.n Mannen.n Mannschaft.n Schützlinge.n Team.n Truppe.n
side.n squad.n team.n
équipe.n formation.n

Anfangsformation.n Aufstellung.n Kader.n Startelf.n Startformation.n
lineup.n starting_lineup.n
effectif.n onze_de_départ.n

Keeper.n Schlussmann.n Torhüter.n Torwart.n
custodian.n goalkeeper.n keeper.n
gardien_de_but.n gardien.n portier.n

Defensive.n

Abwehr.n Hintermannschaft.n Verteidigung.n
backline.n defence.n rearguard.n
arrière-garde.n défense.n

Innenverteidigung.n
central_defence.n
défense_centrale.n

Innenverteidiger.n
central_defender.n centre-back.n centre-half.n full-back.n
défenseur_central.n

Abräumer.n
sweeper.n

Libero.n

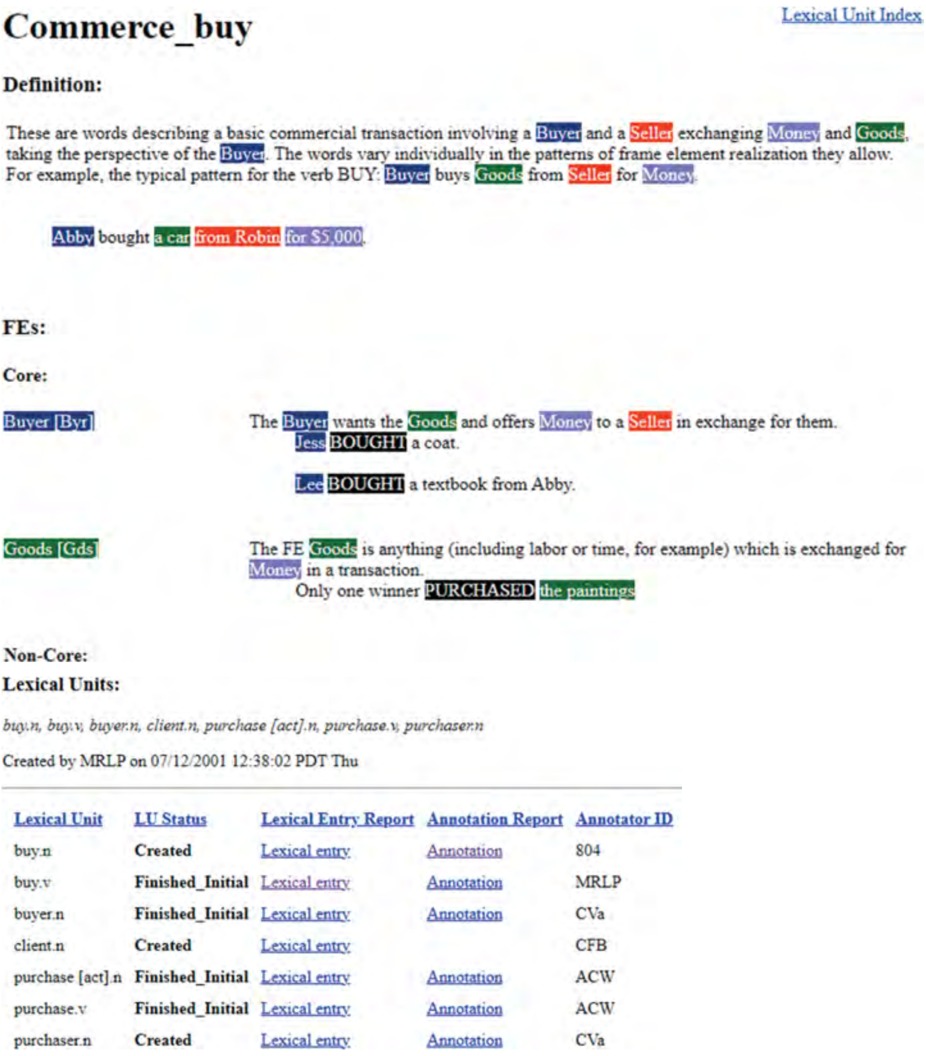
libero.n

Abwehrspieler.n Verteidiger.n
defender.n
arrière.n défenseur.n

Außenverteidiger.n
wing-back.n
défenseur_latéral.n

Fig. 5.15: Concept hierarchies in KICKTIONARY.

and *Commerce_sell* constitute opposing perspectives on the same superordinate frame *Commerce_goods-Transfer* and therefore share its core frame elements. In this way, a complex network of frames related to one another develops (→ Fig. 5.17), which makes it possible for the dictionary user to explore relationships between meanings and the linguistic forms that belong to them.



Commerce_buy

Lexical Unit Index

Definition:

These are words describing a basic commercial transaction involving a Buyer and a Seller exchanging Money and Goods, taking the perspective of the Buyer. The words vary individually in the patterns of frame element realization they allow. For example, the typical pattern for the verb BUY: Buyer buys Goods from Seller for Money.

Abby bought a car from Robin for $5,000.

FEs:

Core:

Buyer [Byr]

The Buyer wants the Goods and offers Money to a Seller in exchange for them.
Jess BOUGHT a coat.

Lee BOUGHT a textbook from Abby.

Goods [Gds]

The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
Only one winner PURCHASED the paintings

Non-Core:

Lexical Units:

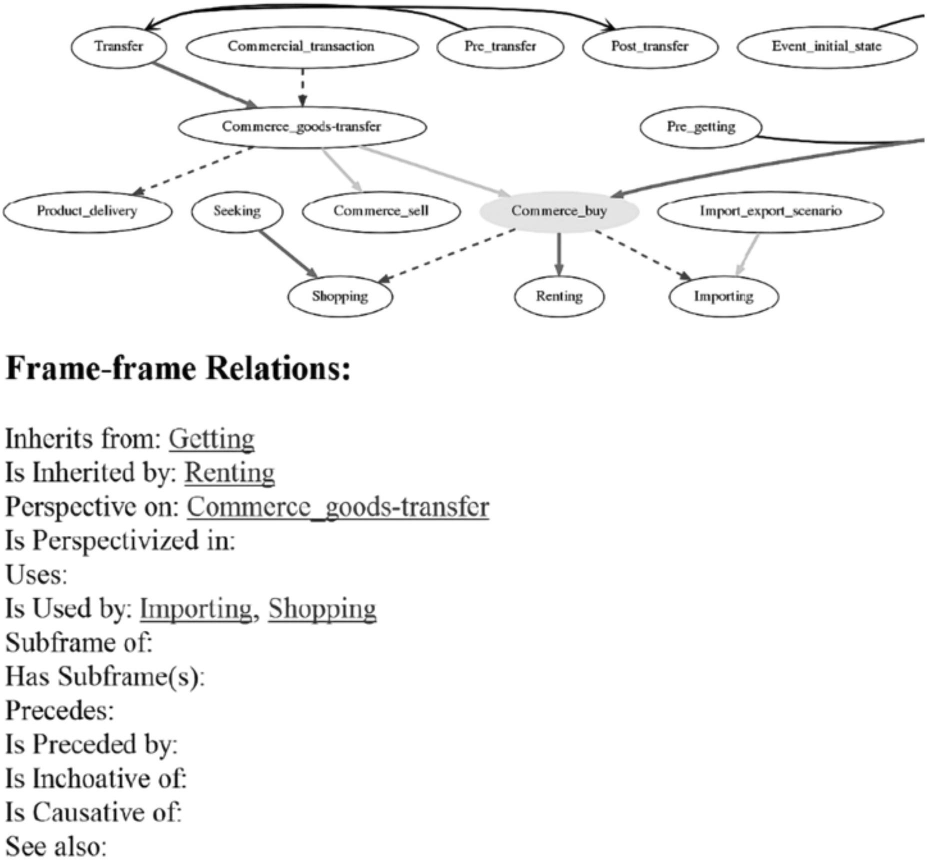
buy.n, buy.v, buyer.n, client.n, purchase [act].n, purchase.v, purchaser.n

Created by MRLP on 07/12/2001 12:38:02 PDT Thu

Lexical Unit	LU Status	Lexical Entry Report	Annotation Report	Annotator ID
buy.n	Created	Lexical entry	Annotation	804
buy.v	Finished_Initial	Lexical entry	Annotation	MRLP
buyer.n	Finished_Initial	Lexical entry	Annotation	CVa
client.n	Created	Lexical entry		CFB
purchase [act].n	Finished_Initial	Lexical entry	Annotation	ACW
purchase.v	Finished_Initial	Lexical entry	Annotation	ACW
purchaser.n	Created	Lexical entry	Annotation	CVa

Fig. 5.16: Description of the frames *Commerce_buy* (above) with the associated lexical units (below) in FRAMENET.

Example 4 (Semagrams in the ANW): While pictorial dictionaries and frames explicitly create onomasiological access structures as a macrostructure – a lexicographer selects



Frame-frame Relations:

Inherits from: Getting
Is Inherited by: Renting
Perspective on: Commerce_goods-transfer
Is Perspectivized in:
Uses:
Is Used by: Importing, Shopping
Subframe of:
Has Subframe(s):
Precedes:
Is Preceded by:
Is Inchoative of:
Is Causative of:
See also:

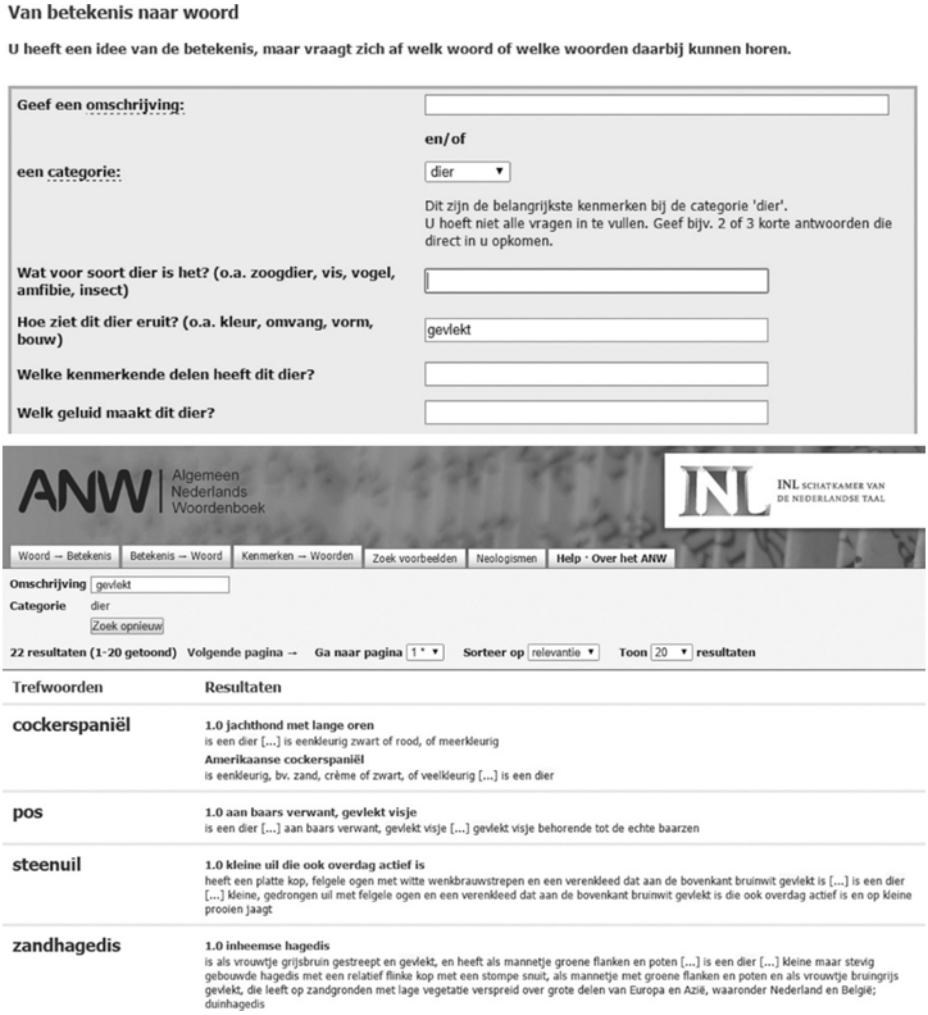
Fig. 5.17: Frame-Frame relations in FrameNet.

images or defines frames to which linguistic forms are then assigned – in the case of concept hierarchies, they result implicitly from mediostructural elements, namely the relations of linguistic forms to one another. The former method can be classed as “top-down” since it specifies the superordinate structures that are then “filled” with lexical units; the latter are classed as “bottom-up” because here the superordinate categories result from the information which is assigned to the lexical units – in this case, the superordinate categories are “emergent”.

The semagrams in the ANW constitute a further bottom-up method for constructing onomasiological access structures (cf. Tiberius/Declerck 2017). A semagram represents knowledge that belongs to a word:

> A semagram is the representation of knowledge associated with a word in a frame of ‘slots’ and ‘fillers’. ‘Slots’ are conceptual structure elements which characterise the properties and relations of the semantic class of a word meaning. (Moerdijk et al. 2008: 19)

As shown in → Fig. 5.12, for example, semagrams belonging to the word *cockerspaniël* (‘cocker spaniel’) record superordinate and subordinate terms (“dog” or “English cocker spaniel”) for this word but also those that denote particular characteristics of this species (e.g., “spotted”).



Van betekenis naar woord

U heeft een idee van de betekenis, maar vraagt zich af welk woord of welke woorden daarbij kunnen horen.

Geef een omschrijving:

en/of

een categorie: dier ▾

Dit zijn de belangrijkste kenmerken bij de categorie 'dier'.
U hoeft niet alle vragen in te vullen. Geef bijv. 2 of 3 korte antwoorden die direct in u opkomen.

Wat voor soort dier is het? (o.a. zoogdier, vis, vogel, amfibie, insect)

Hoe ziet dit dier eruit? (o.a. kleur, omvang, vorm, bouw) gevlekt

Welke kenmerkende delen heeft dit dier?

Welk geluid maakt dit dier?

Omschrijving gevlekt
Categorie dier

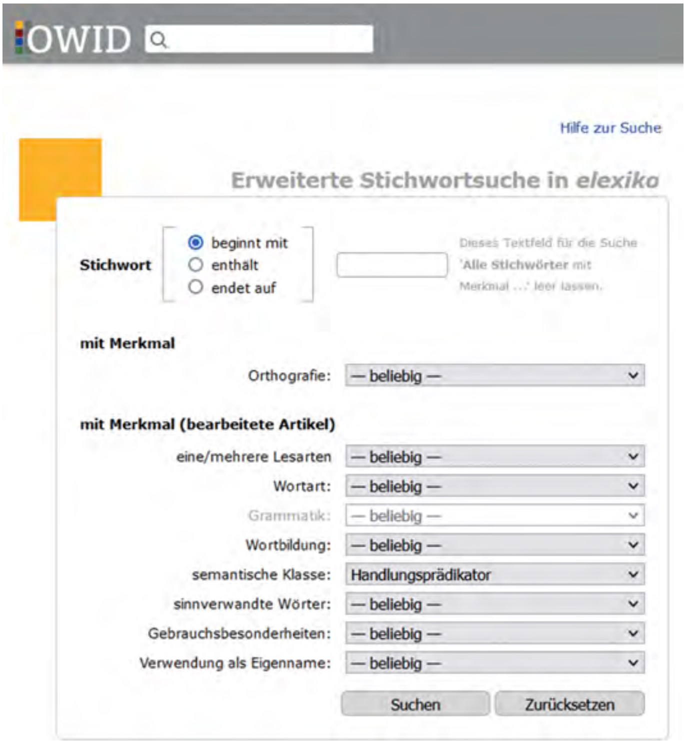
22 resultaten (1-20 getoond) **Volgende pagina →** **Ga naar pagina** 1 ▾ **Sorteer op** relevantie ▾ **Toon** 20 ▾ **resultaten**

Trefwoorden	Resultaten
cockerspaniël	**1.0 jachthond met lange oren** is een dier [...] is eenkleurig zwart of rood, of meerkleurig **Amerikaanse cockerspaniël** is eenkleurig, bv. zand, crème of zwart, of veelkleurig [...] is een dier
pos	**1.0 aan baars verwant, gevlekt visje** is een dier [...] aan baars verwant, gevlekt visje [...] gevlekt visje behorende tot de echte baarzen
steenuil	**1.0 kleine uil die ook overdag actief is** heeft een platte kop, felgele ogen met witte wenkbrauwstrepen en een verenkleed dat aan de bovenkant bruinwit gevlekt is [...] is een dier [...] kleine, gedrongen uil met felgele ogen en een verenkleed dat aan de bovenkant bruinwit gevlekt is die ook overdag actief is en op kleine prooien jaagt
zandhagedis	**1.0 inheemse hagedis** is als vrouwtje grijsbruin gestreept en gevlekt, en heeft als mannetje groene flanken en poten [...] is een dier [...] kleine maar stevig gebouwde hagedis met een relatief flinke kop met een stompe snuit, als mannetje met groene flanken en poten en als vrouwtje bruingrijs gevlekt, die leeft op zandgronden met lage vegetatie verspreid over grote delen van Europa en Azië, waaronder Nederland en België; duinhagedis

Fig. 5.18: Semagram-based search in the ANW.

Semagrams provide dictionary users with a way to navigate through the dictionary based on meanings: for example, to display all of the words to which the semagram “spotted” is assigned. As illustrated in → Fig. 5.18, it is possible, for example, to search in the superordinate category “animal” for the keyword “gevlekt” (‘spotted’), which returns the hits *pos* (‘chub’), *steenuil* (‘little owl’), and *zandhagedis* (‘sand lizard’) as responses as well as *cockerspaniël*.

Onomasiological searches can take very different forms. Some Internet dictionaries provide the option of filtering hits semantically. In this way, the “advanced search” in ELEXIKO allows the user to restrict the desired lemmas to those in particular semantic classes, in → Fig. 5.19, for instance, to words that denote actions.



Erweiterte Stichwortsuche in *elexiko*

Stichwort ● beginnt mit ○ enthält ○ endet auf

Dieses Textfeld für die Suche 'Alle Stichwörter mit Merkmal ...' leer lassen.

mit Merkmal

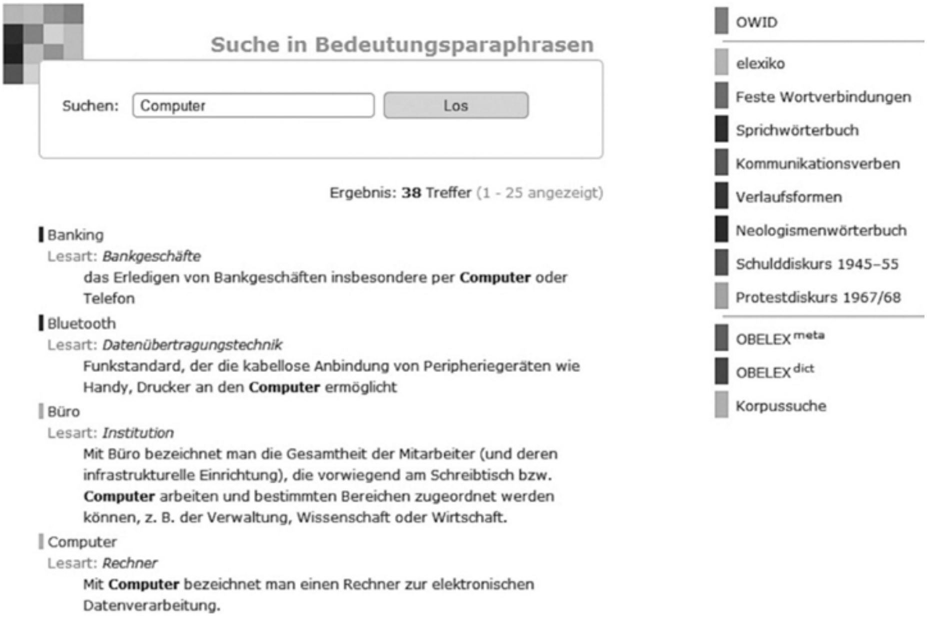
Orthografie:	— beliebig —

mit Merkmal (bearbeitete Artikel)

eine/mehrere Lesarten	— beliebig —
Wortart:	— beliebig —
Grammatik:	— beliebig —
Wortbildung:	— beliebig —
semantische Klasse:	Handlungsprädikator
sinnverwandte Wörter:	— beliebig —
Gebrauchsbesonderheiten:	— beliebig —
Verwendung als Eigenname:	— beliebig —

Fig. 5.19: “Advanced search” in ELEXIKO, searching for words denoting actions (“Handlungsprädikator”).

The full text search is actually conceived in its core function as a semasiological search but when used skilfully and verbalised consistently in the entry texts it can also be employed as an onomasiological search (Engelberg et al. 2020: 61; Pastor/Alcina 2022: 98f., 108f.). Here, the entries in which the search term corresponds to the lemma are not sought but rather the entries in which the search term appears in the entry text or its meaning. For example, the OWID dictionary portal allows a “search in meaning paraphrases” in all its integrated dictionaries; for the search term “Computer”, this would list all of the entries that stand in a semantic relationship with the German word *Computer* (→ Fig. 5.20).



Ergebnis: **38** Treffer (1 - 25 angezeigt)

Banking
Lesart: *Bankgeschäfte*
das Erledigen von Bankgeschäften insbesondere per **Computer** oder Telefon

Bluetooth
Lesart: *Datenübertragungstechnik*
Funkstandard, der die kabellose Anbindung von Peripheriegeräten wie Handy, Drucker an den **Computer** ermöglicht

Büro
Lesart: *Institution*
Mit Büro bezeichnet man die Gesamtheit der Mitarbeiter (und deren infrastrukturelle Einrichtung), die vorwiegend am Schreibtisch bzw. **Computer** arbeiten und bestimmten Bereichen zugeordnet werden können, z. B. der Verwaltung, Wissenschaft oder Wirtschaft.

Computer
Lesart: *Rechner*
Mit **Computer** bezeichnet man einen Rechner zur elektronischen Datenverarbeitung.

Fig. 5.20: “Search in meaning paraphrases” in OWID.

As in a semasiological search, navigating through successive clicks also plays a role in onomasiological searches, when, for example, the user navigates through thematic trees and ontologies. Using a pictorial dictionary, for instance, as in → Fig. 5.14, requires first of all navigating through the thematic tree for “animal kingdom” to “butterfly”, before a lemma is chosen by clicking in the illustration. This is referred to as illustration-based searches.

The representation of meaning relationships in graphs (also → Section 5.3.3) can facilitate access to onomasiological structures. For example, various lexemes that have a semantic relation to the adjective “happy” are represented in a graph in → Fig. 5.21.

5.3.3 Other access structures

Graph-based searches represent a new form of visually supported access to dictionary data that cannot always be classified clearly as semasiological or onomasiological. Here, a graph which represents relations to other lemmas is produced and visualised for a particular lemma. It is possible to access the lemmas visualised in the graph by clicking (→ Fig. 5.22 and → Fig. 5.23), or the user can display a compact form of the article by hovering the mouse over it (→ Fig. 5.21 and → Fig. 5.23) (cf. Meyer 2013, Pastor/Alcina 2022: 116f.; Torner/Arias-Badia 2019 on collocation networks in dictionaries).

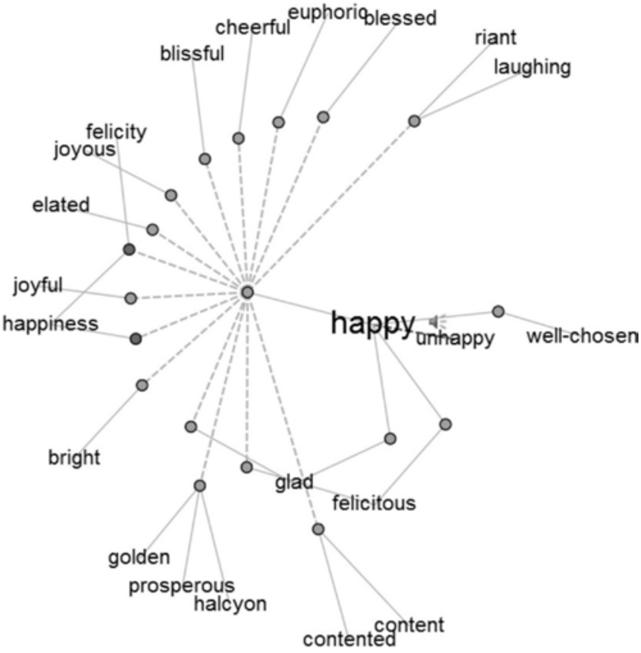


Fig. 5.21: Graph representing semantic relations in the Visual Thesaurus using the option of a graph-based search.

In addition to graph-based search structures, further access structures that are based on various visual associations of lexemes have been popular. So, for example, it is possible to call up lemmas by clicking in word clouds that are generated from co-occurrence analyses (→ Fig. 5.24).

Finally, the boundaries between accessing dictionaries and accessing other types of Internet-based language resources, particularly corpora, become blurred in the digital medium (→ Chapter 2). After all, input-based searches are used not only in dictionary searches but also in corpus queries. In advanced digital lexical systems, individual input-based search queries are used to reach not only dictionary entries but also an array of corpus examples. These searches are realised in both the monolingual DWDS (→ Fig. 5.25) and the bilingual Linguee dictionaries (→ Fig. 5.26).[6]

Finally, it also has to be mentioned that the apparently paradoxical form of arbitrary searching has also been realised in Internet lexicography. In this way, it is possible to have an entry chosen for you by a random generator in the Wiktionary dictionaries. This is more comparable to randomly exploring dictionary content than the targeted accessing of information.

[6] Cf. also Granger/Paquot (2015, pp. 134f.). A dictionary that provides direct access to a corpus of spoken language is described in Meliss et al. (2019).

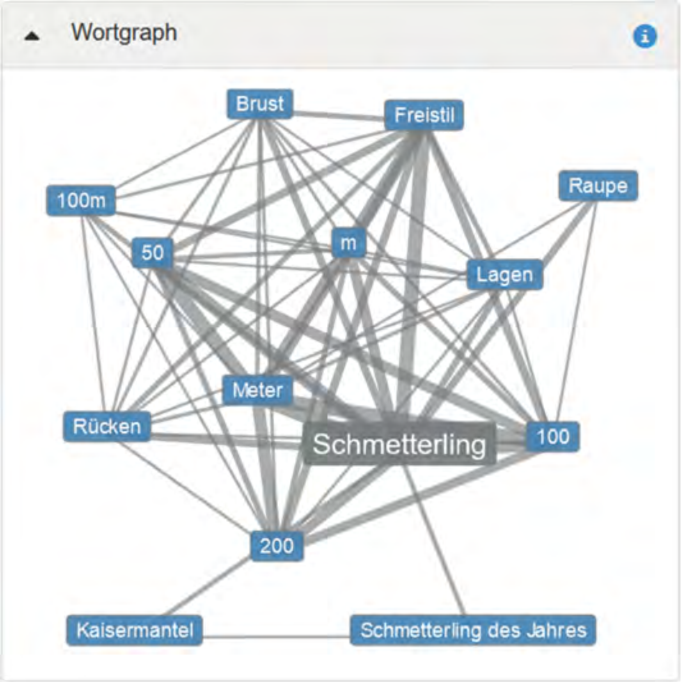
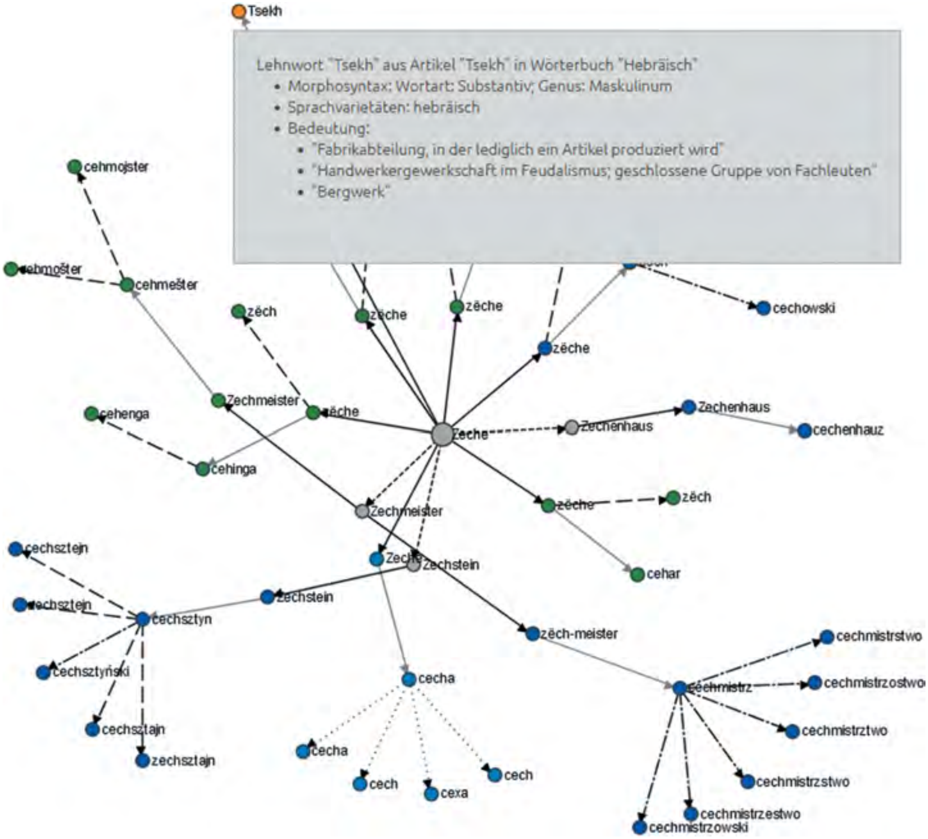


Fig. 5.22: Graph representing co-occurrence relationships of the German word *Schmetterling* 'butterfly' in WORTSCHATZ using the option for graph-based searches.

5.4 New perspectives for dictionary research

The strengths of the digital medium are the possibilities for linking data and the options available to access it in a targeted way. This is reflected in the multiple forms of linking and access structures in Internet dictionaries. However, this not only offers increased room for manoeuvre on the part of dictionary users; it also opens up new perspectives for dictionary research. At the outset, we wrote that it was possible to analyse the linking of data in Internet dictionaries in a similar way to the mediostructures of print dictionaries, in other words by inspecting individual entries as examples. However, we can also proceed in a completely different way when the whole digital basis of data of a dictionary provides the underlying data and when these data are analysed using statistical methods. At the end of this contribution, therefore, we present an example of this kind of novel analysis of the "linking roadmap" for an Internet dictionary using the example of paradigmatic information in the German WIKTIONARY (cf. in more detail Müller-Spitzer/Wolfer 2015) about synonyms, antonyms, hyponyms, hyperonyms, and words related in terms of reference or meaning.



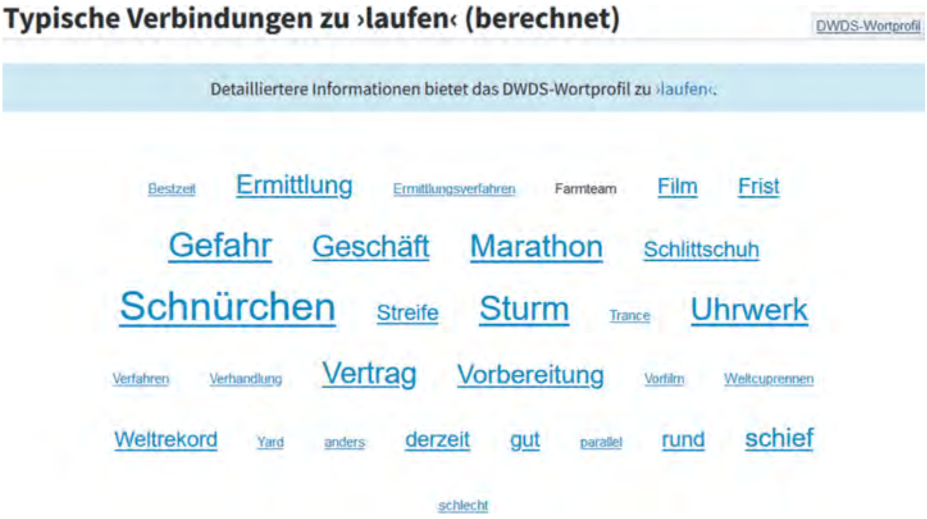
Lehnwort "Tsekh" aus Artikel "Tsekh" in Wörterbuch "Hebräisch"

- Morphosyntax: Wortart: Substantiv; Genus: Maskulinum
- Sprachvarietäten: hebräisch
- Bedeutung:
  - "Fabrikabteilung, in der lediglich ein Artikel produziert wird"
  - "Handwerkergewerkschaft im Feudalismus; geschlossene Gruppe von Fachleuten"
  - "Bergwerk"

Fig. 5.23: Graph representing loanword relationships and morphological relationships in the LWPD using the option of graph-based searches; article for Hebrew *Tsekh* as a borrowing from German *Zeche* ('mine').

It is possible to download the entire basis of data of WIKTIONARY and, thus, to analyse it as a whole body of data.[7] For example, it is possible to visualise all of the relevant information about paradigmatic linking in WIKTIONARY in a single overall representation, drawing an atlas, as it were, of the paradigmatic information in the dictionary (→ Fig. 5.27). The basis for → Fig. 5.27 is provided by all of the incoming and outgoing edges for all five of the relevant classes of information (synonyms, antonyms, factually related words and words related by meaning, superordinate terms, and subordinate terms), represented as a single graph. To aid clarity, only the nodes (keywords) and not the connections between them (edges) are represented. In the process, three clear groups emerge: verbs, nouns, and adjectives. Here, nouns are the largest group. The visualisation routine that is used to create the graph organises the headwords with many con-

[7] https://dumps.wikimedia.org/ [last access: October 14, 2023].



Typische Verbindungen zu ›laufen‹ (berechnet)

DWDS-Wortprofil

Detailliertere Informationen bietet das DWDS-Wortprofil zu ›laufen‹.

Bestzeit Ermittlung Ermittlungsverfahren Farmteam Film Frist

Gefahr Geschäft Marathon Schlittschuh

Schnürchen Streife Sturm Trance Uhrwerk

Verfahren Verhandlung Vertrag Vorbereitung Vorfilm Weltcuprennen

Weltrekord Yard anders derzeit gut parallel rund schief

schlecht

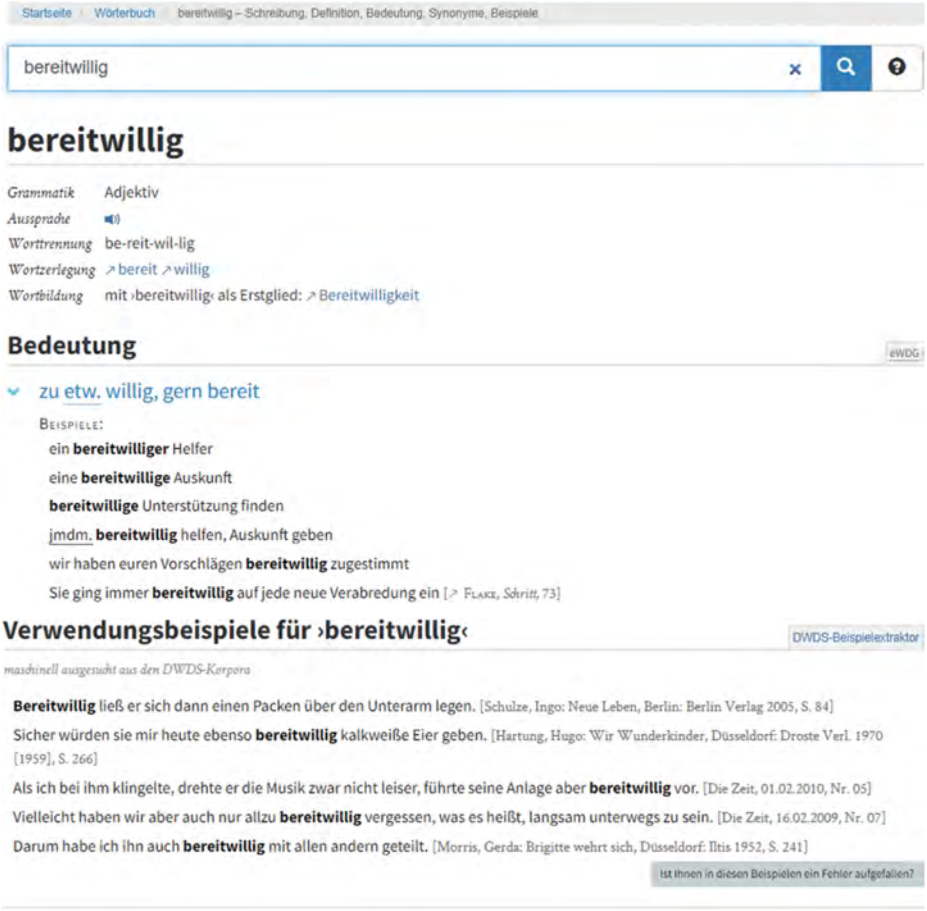
Fig. 5.24: Word cloud with automatically derived collocations for the German lexeme *laufen* (‘run’) in the DWDS; corresponding dictionary articles are called up by clicking on words.

nections between them close to one another spatially. As we would expect, the whole graph shows that paradigmatic linking exists above all between headwords from the same word class. Furthermore, the image as a whole makes it possible to see that a large group of headwords are positioned at the periphery of the graph. These are headwords that are only linked in a very weak way with other headwords. This is the case, for example, when two headwords are connected with one another, but no connection exists in the rest of the graph. A digital version of this graph has been made available online, which allows enlarged sections to be viewed by “zooming in”.[8] This kind of global map does not make it possible to see any details of linking, but it offers a completely different view of the linking structure of the dictionary.

Furthermore, the analysis of the whole basis of data makes it possible to determine their quantitative distributions. Are there more cross-references to synonyms or antonyms? On average, how many nouns, verbs, or adjectives are reported? In this study, for example, we learn that around 25% of the whole inventory of headwords in the German Wiktionary are linked paradigmatically, that these linkages exist above all among headwords of the same class, and that the overwhelming majority of instances of paradigmatic information are in entries for nouns, while for verbs the average number of relational partners is higher than for nouns.

In addition, this kind of global analysis of all of the paradigmatic linking makes it possible to detect particularly strongly linked groups of headwords, for example, by analysing whether there is a group of headwords in the graph where all of the mem-

[8] http://www.ids-mannheim.de/fileadmin/lexik/bilder/all.links.pdf [last access: October 14, 2023).


Startseite
Wörterbuch
bereitwillig – Schreibung, Definition, Bedeutung, Synonyme, Beispiele
bereitwillig
eWDG
DWDS-Beispielextraktor
Ist Ihnen in diesen Beispielen ein Fehler aufgefallen?

bereitwillig

Grammatik Adjektiv
Aussprache
Worttrennung be-reit-wil-lig
Wortzerlegung ↗bereit ↗willig
Wortbildung mit ›bereitwillig‹ als Erstglied: ↗ Bereitwilligkeit

Bedeutung

zu etw. willig, gern bereit

Beispiele:

ein **bereitwilliger** Helfer
eine **bereitwillige** Auskunft
bereitwillige Unterstützung finden
jmdm. **bereitwillig** helfen, Auskunft geben
wir haben euren Vorschlägen **bereitwillig** zugestimmt
Sie ging immer **bereitwillig** auf jede neue Verabredung ein [↗ Flake, *Schritt*, 73]

Verwendungsbeispiele für ›bereitwillig‹

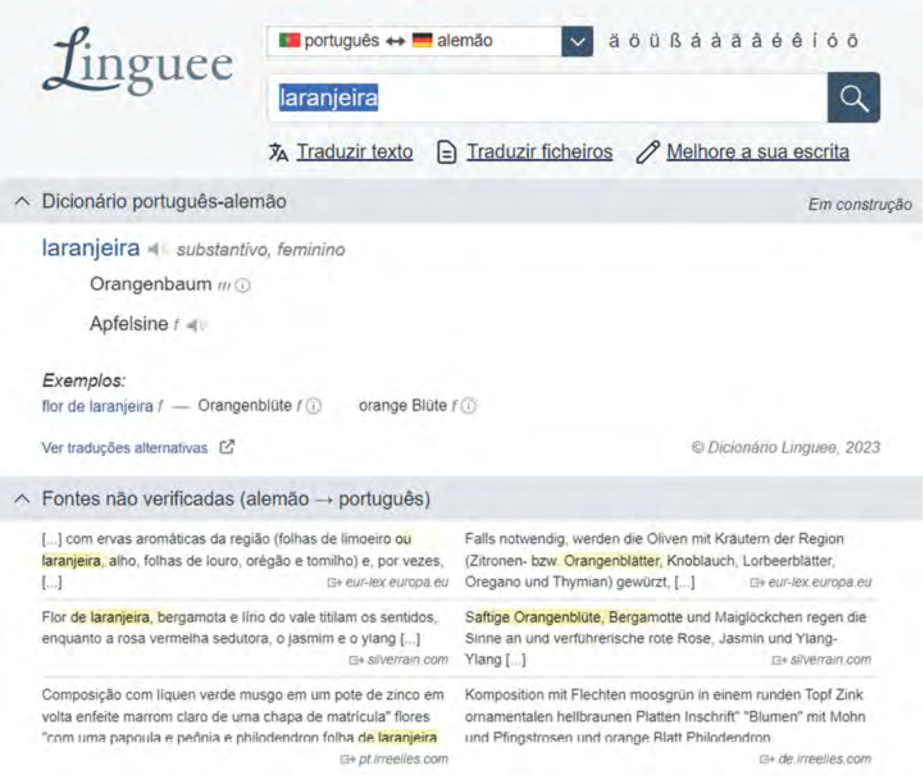
maschinell ausgesucht aus den DWDS-Korpora

Bereitwillig ließ er sich dann einen Packen über den Unterarm legen. [Schulze, Ingo: Neue Leben, Berlin: Berlin Verlag 2005, S. 84]
Sicher würden sie mir heute ebenso **bereitwillig** kalkweiße Eier geben. [Hartung, Hugo: Wir Wunderkinder, Düsseldorf: Droste Verl. 1970 [1959], S. 266]
Als ich bei ihm klingelte, drehte er die Musik zwar nicht leiser, führte seine Anlage aber **bereitwillig** vor. [Die Zeit, 01.02.2010, Nr. 05]
Vielleicht haben wir aber auch nur allzu **bereitwillig** vergessen, was es heißt, langsam unterwegs zu sein. [Die Zeit, 16.02.2009, Nr. 07]
Darum habe ich ihn auch **bereitwillig** mit allen andern geteilt. [Morris, Gerda: Brigitte wehrt sich, Düsseldorf: Iltis 1952, S. 241]

Fig. 5.25: Search for German *bereitwillig* ('willing') in the DWDS and in its integrated corpora.

bers of that group are linked with all of the other members. For instance, this was the case in the German Wiktionary for the causal connectors around *deswegen* ("therefore"), where all of the members of the headword group were connected with all of the others (→ Fig. 5.28). In a second step, this kind of data can be pulled together with further (meta)data about these words. For example, we investigated whether paradigmatically linked words are also frequently looked up. The results for the headword group around *deswegen* can be seen in → Fig. 5.29: here, it is above all the headword *ergo* that is looked up particularly frequently (cf. Müller-Spitzer 2015).[9]

[9] We thank Sascha Wolfer for providing us with these figures in printable quality.



laranjeira

Dicionário português-alemão — *Em construção*

laranjeira *substantivo, feminino*

Orangenbaum *m*

Apfelsine *f*

Exemplos:

flor de laranjeira *f* — Orangenblüte *f* orange Blüte *f*

Ver traduções alternativas

© Dicionário Linguee, 2023

Fontes não verificadas (alemão → português)

[...] com ervas aromáticas da região (folhas de limoeiro ou laranjeira, alho, folhas de louro, orégão e tomilho) e, por vezes, [...] *eur-lex.europa.eu*	Falls notwendig, werden die Oliven mit Kräutern der Region (Zitronen- bzw. Orangenblätter, Knoblauch, Lorbeerblätter, Oregano und Thymian) gewürzt, [...] *eur-lex.europa.eu*
Flor de laranjeira, bergamota e lírio do vale titilam os sentidos, enquanto a rosa vermelha sedutora, o jasmim e o ylang [...] *silverrain.com*	Saftige Orangenblüte, Bergamotte und Maiglöckchen regen die Sinne an und verführerische rote Rose, Jasmin und Ylang-Ylang [...] *silverrain.com*
Composição com líquen verde musgo em um pote de zinco em volta enfeite marrom claro de uma chapa de matrícula" flores "com uma papoula e peônia e philodendron folha de laranjeira *pt.irreelles.com*	Komposition mit Flechten moosgrün in einem runden Topf Zink ornamentalen hellbraunen Platten Inschrift" "Blumen" mit Mohn und Pfingstrosen und orange Blatt Philodendron *de.irreelles.com*

Fig. 5.26: Search for Portuguese *laranjeira* ‘orange tree’ in the Portuguese-German LINGUEE and in its parallel corpus.

This kind of new approach to analysing linking structures may not only provide new impetus for describing linking structures but may also be used to create new access structures. For example, users could have the option to be shown groups of keywords that are closely linked paradigmatically and to be able to access them directly (which may be more useful than showing words that are close to each other in the alphabet, as in printed dictionaries). This is just one example for the way in which so much could still change in the field of linking and access structures.

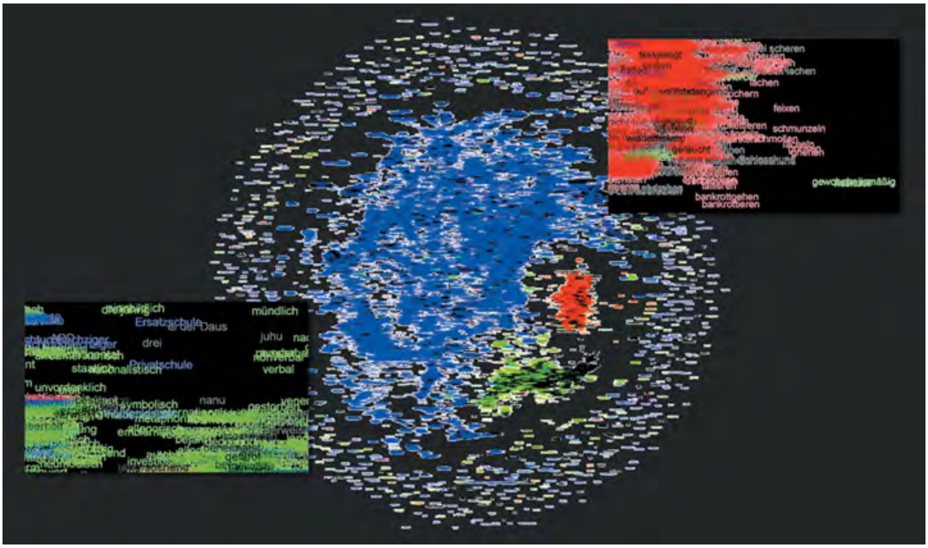


Fig. 5.27: Paradigmatic linking in the German WIKTIONARY as a complete graph; colours indicate different parts of speech.

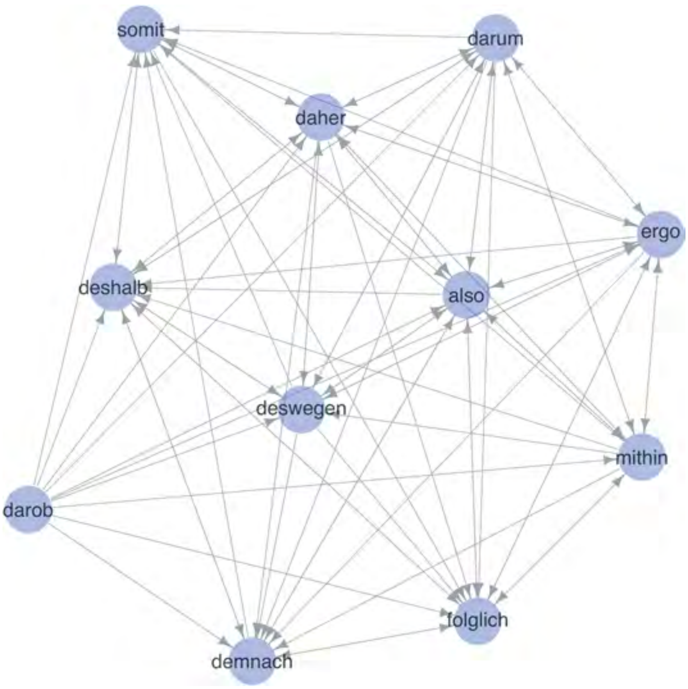


Fig. 5.28: Clique *deswegen*.

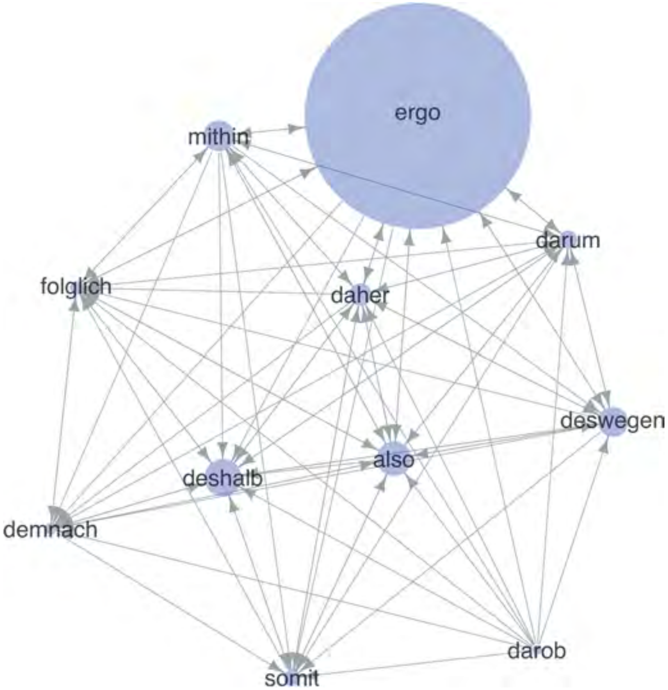


Fig. 5.29: Clique *deswegen*, labelled for frequency of consultation (the size of the circle indicates the frequency of consultation in 2014).

Bibliography

Further reading

Gouws, Rufus H. (2018): Internet lexicography in the 21st century. Berlin/Boston: de Gruyter. In: Engelberg, Stefan/Kämper, Heidrun/Storjohann, Petra (eds.): *Wortschatz: Theorie, Empirie, Dokumentation*, 215–236. https://doi.org/10.1515/9783110538588-010 [last access: April 27, 2024]. *This article relates questions of linking and access structure in digital dictionaries to structural features of printed dictionaries, thus providing a helpful (terminological) link to earlier metalexicographic research.*

Pastor, Verónica/Alcina, Amparo (2022): Researching the use of electronic dictionaries. In: Jackson, Howard (ed.): *The Bloomsbury Handbook of Lexicography*. London et al.: Bloomsbury Academic, 89–124. *This article also provides an overview of linking and access structures, but with a particular focus on developing a standardised nomenclature for describing the different types of searches found in different dictionaries.*

Literature

Academic literature

Blumenthal, Andreas/Lemnitzer, Lothar/Storrer, Angelika (1998): Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computerunterstützter Verweisbehandlung. In: Harras, Gisela (ed.): *Das Wörterbuch. Artikel und Verweisstrukturen. Jahrbuch 1987 des Instituts für deutsche Sprache*. Düsseldorf: Schwann, 351–373.

Dziemianko, Anna (2018): Electronic dictionaries. In: Pedro A. Fuertes-Olivera (ed.): *The Routledge Handbook of Lexicography*. London/New York: Routledge, Taylor & Francis, 663–681.

Engelberg, Stefan/Klosa-Kückelhaus, Annette/Müller-Spitzer, Carolin (2020): Internet lexicography at the Leibniz-Institute for the German Language. In: *K Lexical News* 28, 54–77.

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

Fillmore, Charles J. (1978): On the Organization of Semantic Information in the Lexicon. In: Farkas, Donka/ Jacobsen, Wesley M./Todrys, Karol W. (eds.): *Papers from the Parasession on the Lexicon*. Chicago: Chicago Linguistic Society, 148–173.

Giacomini, Laura (2015): Macrostructural properties and access structures of LSP e-dictionaries for translation: the technical domain. In: *Lexicographica* 31, 90–117.

Granger, Sylviane/Paquot, Magali (2015): Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. In: *Lexicographica* 31, 118–141.

Kammerer, Matthias (1998): Die Mediostruktur in Langenscheidts Großwörterbuch Deutsch als Fremdsprache. In: Wiegand, Herbert Ernst (ed.): *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von "Langenscheidts Großwörterbuch Deutsch als Fremdsprache"*. Tübingen: Niemeyer, 315–330.

Klosa-Kückelhaus, Annette/Michaelis, Frank (2022): The design of internet dictionaries. In: Howard Jackson (ed.): *The Bloomsbury Handbook of Lexicography*. London et al.: Bloomsbury Academic, 405–421.

Lew, Robert (2012): How can we make electronic dictionaries more effective? In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 343–361.

Lindemann, Margarete (1999): Mediostrukturen in modernen italienischen Wörterbüchern. In: *Lexicographica* 15, 38–65.

Mann, Michael (2010): Internet-Wörterbücher am Ende der "Nullerjahre": Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien unter besonderer Berücksichtigung der Fachlexikographie. In: *Lexicographica* 26, 19–45.

Meliss, Meike/Möhrs, Christine/Ribeiro Silveira, Maria/Schmidt, Thomas (2019): A corpus-based lexical resource of spoken German in interaction. In: Kosem, Iztok, et al. (eds.): *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 conference*, *Sintra, Portugal*, *1–3 October 2019*. Brno: Lexical Computing CZ s.r.o., 783–804.

Meyer, Peter (2013): Advanced graph-based searches in an Internet dictionary portal. In: Kosem, Iztok, et al. (eds.): *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, *17–19 October 2013*, *Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 488–502.

Meyer, Peter (2014): Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von elexiko. In: Abel, Andrea/Lemnitzer, Lothar (eds.): *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*. Mannheim: Institut für Deutsche Sprache, 9–21. https://pub.ids-mannheim.de/laufend/opal/opal14-2.html [last access: April 27, 2024].

Meyer, Peter/Tu, Ngoc Duyen Tanja (2021): A word embedding approach to onomasiological search in multilingual loanword lexicography. In: Kosem, Iztok, et al. (eds.): *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 conference. 5–7 July 2021*, *virtual*. Brno: Lexical Computing CZ, s.r.o., 78–91.

Moerdijk, Fons/Tiberius, Carole/Niestadt, Jan (2008): Accessing the ANW dictionary. In: *Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 18–24.

Müller, Peter O. (2002): Die Mediostruktur im De Gruyter Wörterbuch Deutsch als Fremdsprache. In: Wiegand, Herbert Ernst (ed.): *Perspektiven der pädagogischen Lexikographie des Deutschen II*. Tübingen: Niemeyer, 485–496.

Müller-Spitzer, Carolin (2007): *Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis*. Tübingen: Narr.

Müller-Spitzer, Carolin (2013): Textual structures in electronic dictionaries compared with printed dictionaries. A short general survey. 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: De Gruyter, 367–381.

Müller-Spitzer, Carolin/Wolfer, Sascha (2015): Vernetzungsstrukturen digitaler Wörterbücher. Neue Ansätze zur Analyse. In: *Lexicographica* 31, 173–199.

Nielsen, Sandro (1999): Mediostructures in Bilingual LSP Dictionaries. In: *Lexicographica* 15, 90–113.

Pastor, Verónica/Alcina, Amparo (2022): Researching the use of electronic dictionaries. In: Jackson, Howard (ed.): *The Bloomsbury Handbook of Lexicography*. London et al.: Bloomsbury Academic, 89–124.

Porta-Zamorano, Jordi (2018): Exploratory and text searching support in the Dictionary of the Spanish Language. In: Čibej, Jaka et al. (eds.): *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, 925–929.

Schmidt, Thomas (2009): The Kicktionary – A multilingual lexical resource of football language. In: Boas, Hans C. (ed.): *Multilingual Frame-Nets in Computational Lexicography*. Berlin/Boston: De Gruyter, 101–132.

Seretan, Violeta/Wehrli, Eric (2009): Context-sensitive look-up in electronic dictionaries. 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: De Gruyter, 1056–1062.

Storrer, Angelika (2013): Representing (computational) dictionaries in hypertextual form. 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: De Gruyter, 1244–1253.

Tarp, Sven (1999): Theoretical foundations of the so-called crossreference structures. In: *Lexicographica* 15, 114–137.

Tarp, Sven (2008): *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Niemeyer.

Tiberius, Carole/Declerck, Thierry: A lemon model for the ANW Dictionary. In: Kosem, Iztok, et al. (eds.): *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 conference. 19–21 September 2017*, *Leiden, Netherlands*. Brno: Lexical Computing CZ, s.r.o., 237–251.

Torner, Sergi/Arias-Badia, Blanca (2019): Visual networks as a means of representing collocational information in electronic dictionaries. In: *International Journal of Lexicography* 32, 3, 270–295.

Wiegand, Herbert Ernst (2001): Was eigentlich sind Wörterbuchfunktionen? Kritische Anmerkungen zur neueren und neuesten Wörterbuchforschung. In: *Lexicographica* 17, 217–248.

Wiegand, Herbert Ernst (2002): Altes und Neues zur Mediostruktur in Printwörterbüchern. In: *Lexicographica* 18, 168–252.

Wiegand, Herbert Ernst/Smit, Maria (2013): Mediostructures in printed dictionaries. 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: De Gruyter, 214–253.

Dictionaries and other reference works

ANW = *Algemeen Nederlands Woordenboek*. Leiden: Instituut voor de Nederlandse Taal. https://anw.ivdnt.org/search [last access: April 27, 2024].

BabelNet = *BabelNet. Version 5.2*. https://babelnet.org [last access: April 27, 2024].

DWDS = *Digitales Wörterbuch der deutschen Sprache*. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de [last access: April 27, 2024].

elexiko = elexiko. In: *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.owid.de/wb/elexiko/start.html [last access: April 27, 2024].

e-VALBU = Das elektronische Valenzwörterbuch deutscher Verben. In: *Grammis*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://hypermedia.ids-mannheim.de/evalbu/index.html [last access: April 27, 2024].

FrameNet = *FrameNet*. Berkeley: International Computer Science Institute. http://framenet.icsi.berkeley.edu/ [last access: April 27, 2024].

GermaNet = *Germanet*. *A Lexical-semantic Net for German*. Universität Tübingen, Allgemeine Sprachwissenschaft und Computerlinguistik. www.sfs.uni-tuebingen.de/GermaNet/index.shtml [last access: April 27, 2024].

Google = *Google Dictionary*. Google Inc. [available as an app for Chrome; otherwise integrated into the Google search engine].

Kicktionary = Schmidt, Thomas: *Kicktionary*. *Mehrsprachiges digitales Wörterbuch zur Fachsprache des Fußballs*. http://www.kicktionary.de/index_de.html [last access: April 27, 2024].

Kommunikationsverben = Kommunikationsverben. In: *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.owid.de/docs/komvb/start.jsp [last access: April 27, 2024].

LiLa = *LiLa: Linking Latin*. *Building a Knowledge Base of Linguistic Resources for Latin*. Milan: Università Cattolica del Sacro Cuore. https://lila-erc.eu [last access: April 27, 2024].

Linguee = *Linguee Wörterbuch*. Köln: Linguee GmbH. http://www.linguee.de [last access: April 27, 2024].

LWPD = Meyer, Peter/Engelberg, Stefan: *Lehnwortportal Deutsch*. Unter Mitarbeit von Friederike Appel, Frank Michaelis und Simona Štavbar. Mannheim: Leibniz-Institut für Deutsche Sprache. Online: http://lwp.idsmannheim.de/ [last access: April 27, 2024].

Merriam-Webster = *Merriam.Webster*. Springfield: Merriam-Webster Inc. https://www.merriam-webster.com/ [last access: April 27, 2024].

MWVDO = *Merriam-Webster Visual Dictionary Online*. QA International. https://www.visualdictionaryonline.com [last access: April 27, 2024].

OED = *Oxford English Dictionary*. Oxford: Oxford University Press. http://www.oed.com/ [last access: April 27, 2024].

OWID = *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Leibniz-Institut für Deutsche Sprache. http://www.owid.de/.

TLFi = *Trésor de la langue française informatisé*. Nancy: ATILF. http://atilf.atilf.fr/ [last access: April 27, 2024].

VISUAL THESAURUS = *Thinkmap Visual Thesaurus*. New York: Thinkmap, Inc. http://www.visualthesaurus.com [last access: April 27, 2024].

WIKIPEDIA = *Wikipedia, die freie Enzyklopädie*. San Francisco, CA: Wikimedia Foundation. https://www.wikipedia.org [last access: April 27, 2024].

WIKTIONARY = *Wikitionary, das freie Wörterbuch*. https://de.wiktionary.org/wiki/Wiktionary:Hauptseite [last access: April 27, 2024].

WORDNET = *WordNet*. Princeton, NJ: Princeton University. https://wordnet.princeton.edu/ [last access: April 27, 2024].

WORTSCHATZ = *Wortschatz-Portal*. Universität Leipzig. http://wortschatz.uni-leipzig.de/ [last access: April 27, 2024].

Images

Fig. 5.1 Mannheim-Straßenverkehr: https://upload.wikimedia.org/wikipedia/commons/c/c7/Mannheim-Strassenverkehr.png?uselang=de> Knoten Mannheim: OpenStreetMap <http://www.openstreetmap.org/node/240060919#map=11/49.4898/8.4670 Frank, Wikimedia Commons, licenced under CreativeCommons-Lizenz BY-SA 3.0, URL: https://creativecommons.org/licenses/by-sa/3.0/legalcode. Straßenschild: Street sign with ideas https://upload.wikimedia.org/wikipedia/commons/3/33/Street_Sign_with_ideas.jpg. Tom Murphy, Wikimedia Commons, licenced under CreativeCommons-Lizenz BY-SA 3.0, URL: https://creativecommons.org/licenses/by-sa/3.0/legalcode.

Annette Klosa-Kückelhaus and Frank Michaelis

6 The Design of Internet Dictionaries

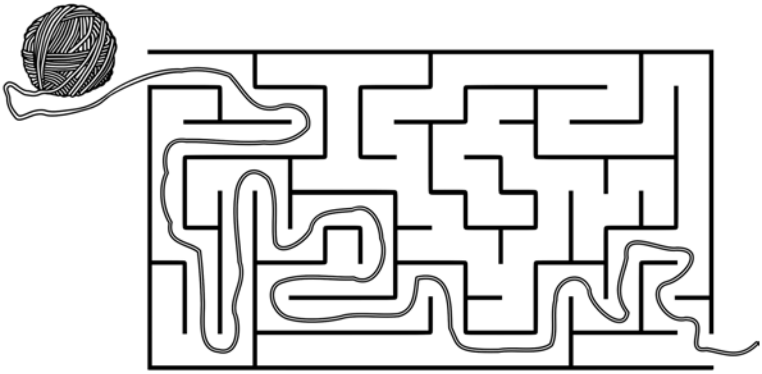


Fig. 6.1: Finding the best possible design solution.

Design can be so much more than creating something pleasing to the eye. The right choice of design tools can support the essential functions of a product. In the case of dictionaries, with their overwhelming number of word entries and sometimes confusing internal article structure, good design can create a "guiding thread" through the maze of information, allowing users to orient themselves and not lose sight of their path.

6.1 Introduction

This chapter will provide an overview of the essential role played by design in both the form of dictionaries and their usability and will also examine the different traditions that exist in the design of (print and electronic) dictionaries (→ Section 6.2.1). The development of dictionary design depends on the intended context in which the dictionary will be used, its potential users, and its data modelling (→ Section 6.2.2). Usage studies (→ Section 6.3) can help delve deeper into user needs concerning dictionary design. Design practice is dependent on a number of elements that are not unique to dictionaries but also on many dictionary-specific factors, for example whether the dictionary is a retrospective digitalisation project or whether the design

Annette Klosa-Kückelhaus, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: klosa@ids-mannheim.de
Frank Michaelis, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161 Mannheim, Germany, e-mail: michaelis@ids-mannheim.de

Open Access. © 2024 the author(s), published by De Gruyter. (cc) BY This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111233758-007

takes the content of the dictionary or its intended users as its starting point (→ Section 6.4). Search functionality is what provides access to an Internet dictionary, so the design of this functionality also has to be planned carefully (→ Section 6.5). Finally, the role of established design guidelines and frameworks will be considered, including how templates are employed and how the lexicographic process should be informed by the interconnected development of content and design (→ Section 6.6).

Design is much more than mere aesthetic eye candy, added on top of the core conception of the dictionary. The design of practical everyday objects, including tools like dictionaries, involves a wide range of aims and requirements. As such, functional, economic, and aesthetic factors all need to be taken into account, and even psychological aspects, such as the emotions that users associate with the object. In this context, design means developing the best possible solution for a product so that these potentially competing requirements are combined in an effective whole.

6.2 General thoughts on the design of (Internet) dictionaries

6.2.1 Similarities and differences between print and online design

Print dictionaries are created according to the principles of graphic design, where typography plays the most important role. Some of the familiar design elements for print dictionaries have been passed down over many hundreds of years, including the alphabetical order of the headwords, which often appear in bold at the beginning of the entry; the layout of the headwords in columns (usually two per page); or the range of entries on a page indicated by column headings at the top of that page. A long-standing problem is that print space has to be used as economically as possible, leading to high text density and increased reading difficulty. For this reason, design decisions for print dictionaries mostly seek to achieve a balance between the need to optimise use of limited print space and the need to present the text in a readable manner. In the online medium, different conditions tend to apply so that different design decisions can be reached.

Website content is described in a hierarchically structured fashion using HTML (Hypertext Markup Language; → Chapter 1.2.2), which a browser converts into the desired form of presentation. In the early years of the worldwide web, HTML left it up to individual browsers to determine how particular elements, such as text, were represented, so that the design of websites left much to be desired. Nowadays, the combination of HTML and Cascading Style Sheets (CSS; → Chapter 1.2.2) gives web developers greater control over how browsers render websites as CSS provides the vocabulary for describing the presentation of a document such as fonts, colours, margins, and even animations as well as the layout for different screen sizes and for printing. Recent enhancements, such as web fonts and rules for complex grid layouts, mean that HTML and CSS come pretty close to print in the possible forms of visual representation that they offer. However, this wide range of possibilities also requires correct usage. As such, the demands placed on web designers' skills and the resources that need to be invested in the design of dictionaries have also increased.

In the digital medium, the new aspect of user interface or application design adds further complications to the design of the text itself, including the wide range of interactions that users have with a dictionary website. For example, Internet dictionaries contain links (the defining characteristic of hypertext; → Chapter 1.2.1 and 1.2.2) and they have a number of standardised interactive elements, such as buttons, text fields, or menus (all of which are already included in HTML to create simple input forms). Finally, JavaScript (→ Chapter 1.2.2) can be used to change the content of a website dynamically, allowing components (also known as widgets) such as tabs and menus to be added that are not (yet) included in the HTML standard. This facilitates complex interactions between users and Internet dictionaries. If implemented correctly, users do not have to learn specially how to look things up or how to navigate in an Internet dictionary. Rather, the dictionary "functions" in the same way as other websites and familiar native desktop applications.

At its best, the design of digital dictionaries draws on both traditional graphic design and user-interface design. Depending on how interactive the design for an Internet dictionary needs to be, a greater or lesser number of application design elements have to be incorporated. While dictionary text and its word entries are still at the heart of the overall design, dictionary-specific components such as headword lists, indexes, extended search functions, or data visualisations could provide the user with quicker access to relevant dictionary entries or with links to collated information otherwise scattered over many pages.

6.2.2 Design dependencies

In most design decisions, it is possible to distinguish between three sets of dependencies: first, regarding the context in which an Internet dictionary is used; second, regarding the data modelling chosen for the dictionary data (→ Chapter 4); and third, regarding the dictionary's users.

Context of use

In the case of a stand-alone dictionary, design decisions may have fewer constraints than when part of a dictionary portal or embedded within another application, such as a text editor or a language-learning platform. In such cases, the design standards of the environment in which the dictionary is embedded must be implemented first. In the most extreme cases, the dictionary in its own right may disappear almost entirely from the user interface and is visible only, for example, in a text editor through the wavy underlining of an incorrectly spelled word and the suggestion provided for how the word should be spelled.

A dictionary intended to be used on a mobile device is subject to different constraints than a dictionary for a desktop browser. This includes not only the space available on the screen where the content is to be displayed but also a variety of control elements. While a mouse can be used in a desktop browser, interaction on mobile devices works by touching the surface of the screen with a stylus or fingers. For example, controls on mobile pages have to be designed to be large enough to allow them to be operated reliably, and some functions, such as "mouseover" effects, are absent altogether from mobile sites. While mouse clicks are the primary form of interaction for desktop browsers, mobile platforms offer a wider range of interactions, such as swiping, pinching, or zooming. Location and light conditions also have a role to play. For example, an Internet dictionary that is to be used primarily outside, on a smartphone, in bright sunlight has to use contrast differently than one that is used mostly indoors. For this reason, many websites now have two design variants – a light mode and a dark mode – and allow the user to adjust them accordingly.

Data modelling

The structure of the dictionary content itself has a decisive effect on design. A fundamental distinction exists between textual data and structured data (→ Chapter 4). Textual data consists of continuous discursive, narrative, or argumentative text in natural language (in contrast to artificial language). In addition, this form of data may contain an internal informational structure distinguished in semantic terms (e.g. headings, quotations, references). In contrast, data structures or records can be thought of as pairs of information called keys and values: in the context of dictionaries, for example, a key called "lemma" could have the values *hand*, *run*, *diligent*, or *you*, and the "word class" key could have the values *noun*, *verb*, *adjective*, or *pronoun*. These pairs of keys and values can be assembled into groups or objects, combined into more complex structures such as lists and hierarchical trees, and stored in databases.

In XML (Extensible Markup Language; → Chapter 4.2.1), which is the most common metalanguage in lexicography, structured textual data is also referred to as "mixed content". Keys correspond to the names of elements or attributes and values to the specific values of the elements or attributes. In most cases, dictionaries can be characterised as hybrid forms of textual and structured data; in other words, data structures containing additional information (e.g. metadata) may be embedded in the text. These embedded data structures may also break down information represented as discursive text into a formal representation or model that can be interpreted by a computer. Conversely, data structures may be supplemented by textual data, for example, in the form of detailed commentary fields.

As far as textual data are concerned, the emphasis in design rests primarily on typography and legibility. For data structures, the design often reflects the tree structure of the data in list form or in a hierarchically organised form, reminiscent of a table of contents. However, data structures can also be presented in a form similar to continuous text, for example, when entries from a list are arranged one after another in the same line separated by commas. In any case, planning the graphic display of dictionary data in the design of Internet dictionaries involves combining both principles (continuous text and structured data) in a manner appropriate to the data model as well as the context of use and the user.

The user

Focusing on the user in the design process – in other words, user-centric or human-centric design – has its origins in industrial product design. Applied to the use of a dictionary, this means that the elements in the dictionary and its contents are organised and designed in such a way that the user is able to successfully look up what they need to while expending the minimum possible time and cognitive effort. If the user is to be the starting point for design decisions, a number of questions have to be answered. For example: who is the (typical) user? Which problems do they typically want to solve? What is the (typical) search behaviour adopted to answer the problem? We can begin to answer these kinds of questions through so-called “user stories”. These are case study scenarios involving fictional users (who are conceived in as concrete and realistic a way as possible), which give designers a framework for the development process. User testing and dictionary usage studies (→ Chapter 9) can then be employed to establish how effective these scenarios and planning strategies prove to be in reality.

By contrast, many Internet dictionaries continue to adopt a content-centric approach to design: that is to say, they list their information in a more or less condensed fashion, organising it according to their internal structure (which is primarily motivated by lexicological or lexicographical principles). As such, it is left up to the user to extract the information relevant to them in a particular situation from the Internet dictionary. This is particularly the case for general monolingual or multilingual dictionaries that are not integrated into other applications. However, if a dictionary is embedded in an application and a specific context of use, as might be expected, the user and their aims should exert a strong influence on the design. Unfortunately, embedded dictionaries and those intended for specific purposes have tended to play a lesser role in academic lexicography to date.

6.3 Usage studies on design

Although there is now a relatively long tradition of research into the use of print and Internet dictionaries (→ Chapter 9), there are not many usage studies that deal specifically with questions of design. Research in metalexicography has not tended to concentrate on design issues for Internet dictionaries either. Exceptions include publications by Almind (2005), Debus-Gregor/Heid (2013), Oppentocht/Schutz (2003), Spohr (2008), and Swanepoel (2001), which focused on the connection between the modelling of data and its online presentation while studies by Corréard (2002), Hollós (2018), Lew (in press), and Schmitz (2016) looked, above all, at the arrangement of the lexicographical information on the screen. Other researchers, notably Dziemianko (2014, 2015, and 2016), examined the positioning of particular kinds of information or the use of colour. Finally, Michaelis/Müller-Spitzer/Wolfer (2019), Storjohann (2018), and Torner/Arias-Badia (2019) among others concerned themselves with possible new forms of data presentation.

In relation to usage, Heid/Zimmermann (2012) proposed usability testing as a method to develop the design of Internet dictionaries and Koplenig/Müller-Spitzer (2014) outlined the results from a usage study on various possibilities for presenting data. Usage studies on Internet dictionaries involving eye-tracking experiments were undertaken notably by Lew (2010), Lew et al. (2013), Lew/Tokarek (2010), Nesi/Tan (2011), and Tono (2000 and 2011) while Müller-Spitzer/Michaelis/Koplenig (2014) used this method to test a new design for a dictionary portal. Eye-tracking studies, in particular, enable a detailed assessment of whether the arrangement of information on the screen, the typographical design, and the use of colour, etc. are understood by the study participants in the way that was planned and whether they are used to orient the way they look at the screen (→ Chapter 9).

6.4 Design practice for Internet dictionaries

6.4.1 Design fundamentals

If we view Internet dictionaries more generally as a subset of websites, the design options and rules that have been developed in this field will also apply to them. For designers of Internet dictionaries, this has the crucial advantage that they can draw on a wealth of existing design practice and experience. As explained in → Section 6.2.1, web design is influenced by print and graphic design, and their traditions reach back centuries. This should not surprise us: for all that our technology and media may have changed humans' cognitive capacities when interacting with text and image cannot have changed in any fundamental way in what is, in evolutionary terms, a relatively short period of time. Something that was easy or difficult to read 200 years ago will continue to be so today.

It is beyond the scope of this chapter to provide a comprehensive overview of the wide variety of design traditions and schools. However, we would like to present a selection of basic principles as they apply to Internet dictionaries, before addressing more dictionary-specific issues.

Questions about the design goals of a project cannot be answered in general terms. A specific text design or page design is intended to put the user in a particular mood and make them associate the content with a particular experience, usually an emotional one. This is the domain of UX design (user experience design), and although this aim seems to be of greater importance for marketing and product pages, it also plays a role in Internet dictionaries. For reference works, for example, an appearance that communicates "reliability" and "credibility" might be appropriate, comparable with news broadcasting. A dictionary that addresses a very specialist group of users – for example, sportspeople or computer enthusiasts – might prefer to adopt a "modern" or "fresh" look. However, conveying information quickly and simply should be a common goal of most dictionaries so that design principles such as readability, consistency, and visual hierarchy play a significant role in most dictionary design decisions.

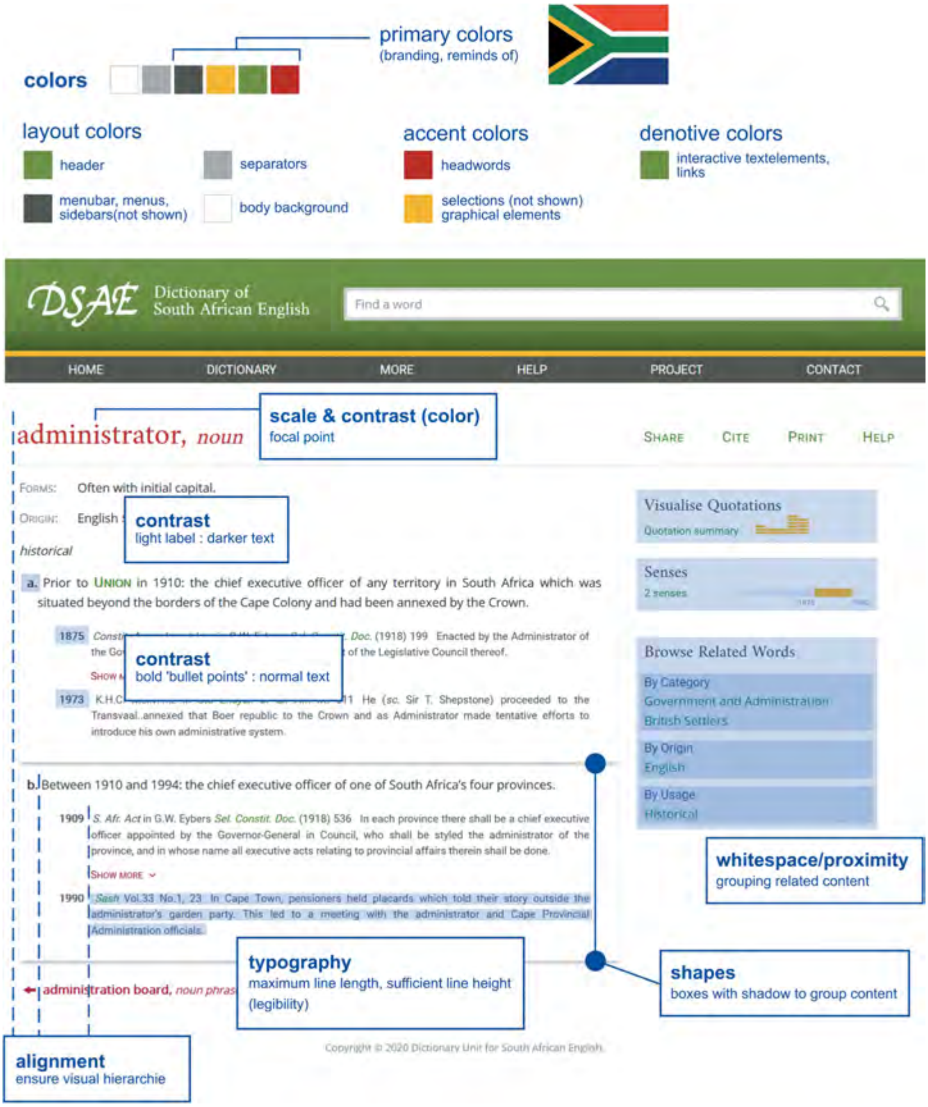
Here, readability means the extent to which a text can be read easily and without tiring the eyes. Decisive design techniques in this context are line length, line spacing, font size, choice of font, and the contrast between the colour of the font and the background.

Consistency (and repetition) refers to the uniform design of recurring elements, reducing the cognitive effort on the part of the user, who does not have to learn the position and use of control elements of the interface time and again. The rule "less is more" also has a place here since any newly created and different element must be (re-)learnt, and understood afresh, by the user.

The principle of visual hierarchy means that every element on the page possesses a specific level of importance. If all of them were of the same importance, the user would not know where to look first. The visual hierarchy of the page should establish a structure to deliberately direct the user's attention towards particular focal points. The use of colour and scale are relevant design techniques in this context, as are animations, which are particularly effective at attracting and retaining the user's attention.

→ Figure 6.2 demonstrates how design techniques such as white space and proximity, colour, contrast, scale, alignment, shapes, and typography can be used in a dictionary text in different ways, and in combination with one another, in order to support the principles outlined above.

On websites, the traditional design elements are supplemented by elements that originated in the field of application interface design, like input masks, which facilitate the user's interaction with the computer. In user interface design, components (also known as widgets) are the basic building blocks that are used to assemble more complex structures, such as the individual views of an application or the application



administrator, *noun*

FORMS: Often with initial capital.

ORIGIN: English

historical

a. Prior to UNION in 1910: the chief executive officer of any territory in South Africa which was situated beyond the borders of the Cape Colony and had been annexed by the Crown.

1875 *Consti*… *Doc.* (1918) 199 Enacted by the Administrator of the Go… of the Legislative Council thereof.

SHOW M…

1973 K.H.C… 11 He (*sc.* Sir T. Shepstone) proceeded to the Transvaal..annexed that Boer republic to the Crown and as Administrator made tentative efforts to introduce his own administrative system.

b. Between 1910 and 1994: the chief executive officer of one of South Africa's four provinces.

1909 *S. Afr. Act* in G.W. Eybers *Sel. Constit. Doc.* (1918) 536 In each province there shall be a chief executive officer appointed by the Governor-General in Council, who shall be styled the administrator of the province, and in whose name all executive acts relating to provincial affairs therein shall be done.

SHOW MORE

1990 *Sash* Vol.33 No.1, 23 In Cape Town, pensioners held placards which told their story outside the administrator's garden party. This led to a meeting with the administrator and Cape Provincial Administration officials.

← administration board, *noun phras*…

Copyright © 2020 Dictionary Unit for South African English.

Fig. 6.2: Entry "administrator" in the *Dictionary of South African English*.

as a whole. Components themselves are, in turn, made up of smaller components, or design primitives (lines, shapes, text; → Fig. 6.3).

In addition, it is possible to distinguish these components according to their function. Hence, there are components:

1. for grouping and organising content, e.g. cards, lists, text sections, accordions;
2. for navigating within content, e.g. tabs, navigation drawers, navigation bars (top, side, bottom);

3. for performing tasks or giving commands, e.g. buttons, menus;
4. for user input or selections, e.g. text input fields, select boxes, check boxes;
5. for messages or responses from the application, e.g. popups, progress bars, dialogue boxes, status bars.

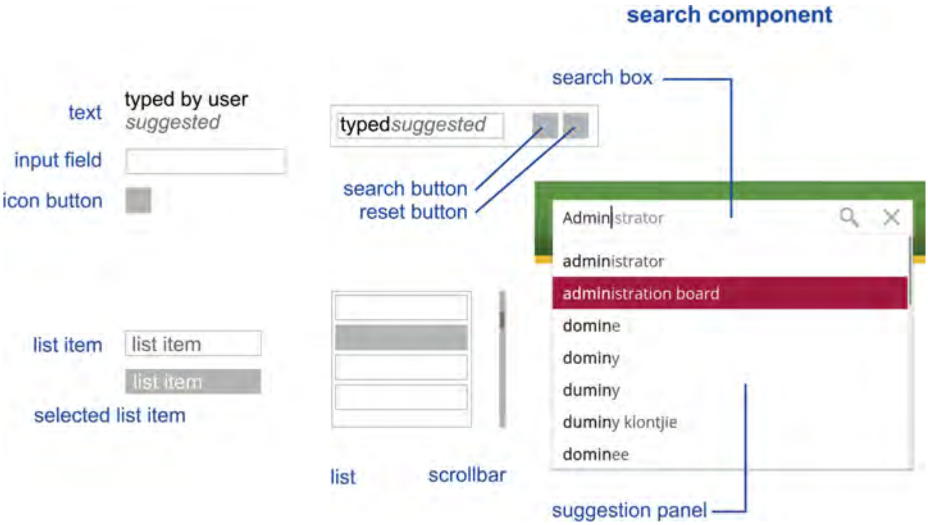


Fig. 6.3: Examples for design primitives in the *Dictionary of South African English*.

A particular challenge for user interface design is that these components also have to be (repeatedly) recognised as such by the user. Hence, these components tend to exist in a similar form in all operating systems (Windows, Linux, Android, iOS). However, they intentionally diverge from one another in their specific design in order to create an individual look and feel unique to the particular product. Websites, including Internet dictionaries, make use of the same techniques and are able to design their own look and feel. If the design of the user interface diverges too far from the conventions of the operating system that is most familiar to the user, however, there is a real danger that they will no longer recognise the components as interface components and will not know how to operate them.

Moreover, the implementation of the user interface design and interactive components is more demanding than that of static content. Components often possess several states, which have to be distinguished visually from one another. A button, for example, can be "normal", "pressed", "focused", "active", or "disabled". The principles that govern the design of these states must be well thought out to ensure that they can be easily distinguished from one another and conform to product or branding guidelines as well.

Users also require direct visual feedback to show whether their action has been successful or not. For example, a button that does not change its state when the user clicks on it means they do not know whether the computer has recognised the click or not and whether it will perform the required action. In this respect, modern user interface design (as of 2024) seeks to be as unobtrusive as possible. Instead of using text to provide lengthy status messages, an action button will change colour: for example, if the action has been successful, the button will change to green and its label to a tick; if not, it will turn red and the label will become a cross. Implementing these kinds of animated microinteractions assumes at least basic knowledge about animation techniques on the part of the dictionary designer.

Another complex area is accessibility, that is, design that ensures access without any barriers. The technical possibilities for accessible design have improved over the years as far as browsers are concerned but (as of 2024) designers often still lack knowledge and experience in implementing these recommendations and guidelines. Standardisation organisations such as *W3C* provide assistance in this area and are driving developments forward, for example with their *Web Content Accessibility Guidelines (WCAG) 2.0*. Nowadays, development tools in browsers indicate to designers whether, for instance, the contrast they have chosen between the foreground and background meets these guidelines. HTML itself allows for additional markups, which make it easier for text-to-speech programs to read an HTML page. However, planning for all of these technologies implicates a discernible increase in design effort, and it is essential that these be taken into account in the conception of an Internet dictionary (→ Chapter 3).

6.4.2 Specific aspects of Internet dictionary design

Retrospective digital dictionaries

There are considerable overlaps with the field of textual studies in the presentation of retrospective digital dictionaries, that is, print dictionaries, usually older ones that are subsequently digitalised. One common characteristic of these projects is to achieve as exact a reproduction as possible of the original text. Hence, the pagination of the print version is frequently retained to ensure that the online version can still be cited in the same way. Editorial interventions have to be marked and created in such a way that they are recognisable, and so on.

One recurring design issue pertains to the relationship between the “modern” dictionary application and the “old” dictionary pages. There is a particularly striking discontinuity in the case of image digitalisation, where the user is presented with scanned images of the original dictionary. But that discontinuity can also be intentional, as a reminder to the user that they are reading a historical source rather than a contemporary reference work.

Conversely, digital transcriptions of older print dictionaries can take the opportunity to re-evaluate the original print design, improving its clarity, for example, by introducing a clearer visual hierarchy or by replacing an old-fashioned typeface such as *Fraktur* (a blackletter typescript) with a modern font in order to ensure legibility for 21st-century readers. If users come across a historical dictionary with a contemporary design, there is, of course, an increased risk that the user will confuse it with a contemporary dictionary. Unfortunately, there are limited design options available to counteract such a misunderstanding.

Content-centric presentation

On a very abstract level (and from a design perspective), many dictionary entries can be described as a structure in which the lexicographical information about a headword is organised in thematically related groups (→ Chapter 4); then, alongside that information, these groups may contain further subordinate groups (e.g. primary meaning and secondary meaning). In a content-centric design, the dictionary interface reflects, in a more or less one-to-one manner, this tree-like structure, nested in as many levels as necessary.

This hierarchical structure is intended to enable the user to quickly grasp the structural organisation of the entry so that they can direct their attention to the relevant block. Of course, one prerequisite for this is that the user has prior expectations as to what type of information they can find in which group and how this information can help them solve their problem. Whether these expectations of user behaviour on the part of lexicographers are realistic is the object of enquiry in user research (→ Chapter 9). → Fig. 6.2 shows the design techniques employed to translate this hierarchical lexicographical structure into a visual hierarchy.

User-/Human-centric design

In user-centric design, the lexicographical structure no longer stands at the centre; rather the design is oriented towards the actual task the user is undertaking or the problem to be solved. The dictionary *Paronyme – Dynamisch im Kontrast*, for example, is a dictionary that is meant to help the user deal with uncertainty about the meaning and usage of German paronyms. In many of the views in this dictionary, the design attempts to assist in the task of “comparing and contrasting”. Partial meanings are presented to the user in a sortable overview; they are able to choose up to three of them, receiving the corresponding detailed views presented alongside one another in an overlay. This allows similarities and differences between the words to be compared, down to the level of individual examples of usage.

If the user’s tasks and questions are placed at the centre of the design, the question arises as to why those tasks and questions should not be resolved at the point at which they arise. A logical step would be, for example, to integrate dictionaries in text editing programs to assist in the production of texts, or in digital editions of texts to aid user comprehension. Here, dictionaries no longer appear as independent entities; rather, as far as possible, they fit seamlessly into the user's working environment in order to support them in their actual work, such as writing or reading texts. This is already standard today for very simple lexicographical questions, such as spelling or hyphenation. In these kinds of applications, the challenge for design lies more in the area of functional integration than in visual design.

Other features of online dictionaries

In addition to entries for individual words, Internet dictionaries can provide a range of further texts, illustrations, or applications that, above all, make it easier to access information relating to the words in the dictionary (→ Chapter 5). One example is overviews of word entries that satisfy particular criteria: for example, in a dictionary of neologisms a list of words that emerged in a particular time period; in a dictionary of loanwords lists of words borrowed from a particular language; or in a general dictionary a list of all of the words derived from proper nouns, and so on. The word entries included in the lists are created as hyperlinks so that these kinds of lists not only have an informational value referring to the content of the dictionary but also provide possible points of access to that content.

Visualisations such as word clouds can also be used as navigation tools, inviting users to explore the content of the dictionary, all the more so if these are interactive visualisations. For example, if allowed by the corresponding data model, chains of loanwords from one language into a series of other languages can be represented as an interactive graph in which users can navigate. Nevertheless, such complex representations are more appropriate for illustrative purposes and to encourage exploration of dictionary content; they are not suitable for quickly looking something up.

Finally, it is possible to integrate static illustrations, videos, or audio data alongside text and visualisations. Dictionary design has to plan for these kinds of elements: for example, decisions need to be taken as to whether photographs, film, or audio clips should only be opened or started by clicking on them, whether they should be integrated into the dictionary interface or open in a new window, or whether hyperlinks should link to content hosted elsewhere. In conceptual terms, it is important, in each case, to ensure a close interconnection between the word entry and these kinds of features.

6.5 The design of search functions

Users of Internet dictionaries are familiar with three different search options (for more details, → Chapter 5), which they recognise from other websites: a simple search for a search term, a search by characteristics or attributes, and a full-text search. Each of these search options comes with advantages and disadvantages for the user and poses challenges for the design of an Internet dictionary.

The simplest way to search in an Internet dictionary is to enter a search term into a search field (the positioning of the search field on the page should follow the usual expectations for websites). If only one entry for the search term is found, this entry is usually shown directly on the screen. If a search generates multiple search results, the situation is different, and a list of entries is displayed on a separate page of search results.

The main purpose of a search by characteristics is to limit the number of hits returned to the user, something that is particularly common on the websites of online retailers. Shoppers in an online shop can, for example, restrict their search to blue sweaters made of cotton with long sleeves and a V-neck costing between $30 and $50. This is not easy to translate to dictionaries since, when searching for a particular word, it does not usually help to limit that search according to word class, number of syllables, inflectability, and so on. However, these kinds of “faceted searches” do exist in Internet dictionaries, allowing the dictionary to be used like a database. For instance, in the context of lexicological research, it is possible to search for examples of verbs borrowed in the 18th century from French into Italian, word entries in which a quotation from Jane Austen provides the first attested usage in English, or German neologisms from the 1990s that do not originate in English. In design, faceted searches frequently draw on menus and dropdown lists, among other techniques. The results of these searches are often displayed on a separate page on which the results can be further sorted or filtered before the user is able to either follow the hyperlink to an individual word entry or export or print the search results as a whole.

In a full-text search, a search term is generally searched for in the visible dictionary text, that is, in all word entries and, where applicable, also in the surrounding text, irrespective of whether the dictionary consists of textual data or structured data (→ Section 6.2.2). In terms of design, search results are displayed according to well-known models from other applications (e.g. Google) whereby a small snippet of the text is shown with the highlighted result. In cases with very high numbers of hits, the search results are distributed across several pages, so-called “pagination”. A hyperlink leads from each snippet to the original dictionary entry.

6.6 The design process

At the end of this presentation of the design of Internet dictionaries, it is worth including some reflections on the design process. Where possible and appropriate, these should draw on well-known design frameworks and should, at least, give consideration to the use of templates. Finally, when planning a dictionary project, the design process should be integrated into the lexicographical process at an early stage in order to facilitate the development of a form of presentation that is attractive, intuitive to use, and appropriate to the subject area of the dictionary and its intended function (→ Chapter 3).

6.6.1 Established design frameworks

Engaging with the design guidelines and frameworks developed by the major producers of operating systems (Google/Android, Microsoft, and Apple) can bring particular benefits: as has already been mentioned in → Section 6.4.1, they convey the "native look and feel" of the surrounding operating system to which users are most accustomed. Users already have certain expectations about how the elements on their screen should behave, and applications that do not hold to those conventions can discourage them, or even cause annoyance. On top of that comes the not inconsiderable effort and complexity involved in the development of a new design system. Adopting existing designs allows designers to focus on the development of the components specific to the application.

In addition to technical documentation and tutorials on web development, corporations such as Google and Microsoft provide detailed documentation and, above all, explanations of their design guidelines, for example, Google's *Material Design*. The design systems or guidelines describe what has evolved over the years into "good practice". They contain collections of standard components and colour schemes as well as standard navigation and interaction models (for their platform). Pairs of "do's" and "don'ts" illustrations help designers avoid errors that can irritate users.

However, this consolidation of design conventions through market success does not always lead to the best possible design solution. A prominent example of this is our standard keyboard layout, which still follows that of typewriters and which is far from optimal in ergonomic terms. For this reason, user research (→ Chapter 9) and creative experiments are important in order to question and challenge existing conventions.

Alongside the more gradual general developments in design, there are also design fashions and trends, with the best known being Web 2.0 with its glossy image buttons (early 2000s). Nowadays (as of 2024), so-called "flat design" tends to dominate. However, these are more stylistic elements than design elements in the strictest sense. Nonetheless, as is the case in fashion, what was once the latest style quickly appears old fashioned, if not downright ridiculous. Since Internet dictionaries are mostly long-term undertakings, elements that are characteristic of a particular fashion should be used with caution. Use of such elements can draw unnecessary attention to them and quickly make what is actually a well-designed and well-functioning site appear old fashioned.

6.6.2 Templates

There are numerous resources on the Internet that offer website templates, frequently as open source material, free for anyone to use. These can be implementations of existing design frameworks by the manufacturer or by third parties, such as Google's *Material Design*, or implementations of original designs. Many prominent websites make their own framework available, such as *Bootstrap,* a framework originating from X, formerly known as Twitter. If a framework is used in a great number of other projects, as *Bootstrap* has been, the design acquires a certain prominence and familiarity. This degree of familiarity is an advantage in terms of usability. However, it becomes more difficult to distinguish one project from another visually.

A further definite advantage of using existing frameworks is the possibility of drawing on the work of professional designers and developers. However, because designers are often oriented towards what is popular on the market, these templates tend to be conceived more for blogs, portfolios, and commercial or marketing sites rather than for the particular requirements of Internet dictionaries. Depending on the framework, extending and modifying an existing templates to the lexicographer's special needs can be expensive and can, in certain circumstances, require just as much prior knowledge as implementing one's own design from scratch.

6.6.3 Processes

The following lexicographical processes (→ Chapter 3) would be involved in producing an Internet dictionary according to the waterfall model. Starting with the planning and conception of the dictionary, the process would move on to the preparation and provision of the dictionary sources for the compilation of the word entries. Next the web application would be implemented, followed by the proofreading and testing of the interface. Finally, the Internet dictionary would be released or would go on sale. However, this linear process can be problematic in some circumstances: for example, problems that were not identified during the planning phase, or were dealt with inadequately, can only be resolved later in a very time- and cost-intensive way. Moreover, feedback from users that is gathered only after its release or delivery cannot be taken into consideration during the development of the dictionary.

When applied to the design of Internet dictionaries in particular, it is important to consider that the linear planning and realisation of a dictionary project results in particular dependencies between content and presentation being identified only when it is too late and being reworked only at great cost, if at all. For example, later in the “application implementation” step, an Internet dictionary project wants to offer brief lexicographical commentaries in small pop-up windows, but their content cannot be automatically derived from information already found in the entry, so the whole project has to move back to the “compilation” step to create and edit this information. It would have been better if, instead, the data model had provided this information type from the very beginning.

For these reasons, an iterative design process should be chosen for Internet dictionaries in which developmental phases focusing on specific areas can be run on numerous occasions. In this kind of process, prototypes can be developed at an early stage, or specific elements of the application can be tested so that feedback from users can also be taken into account in early planning stages. In this way, the conception of content and design should be interconnected from the outset so that, at best, the team working on an Internet dictionary project involves not only lexicographical expertise but also expertise in IT and web design.

6.7 Conclusion

Whether in print or online, dictionaries comprise not only content but also the form in which this content is presented to users. For Internet dictionaries in particular, it is worth planning this presentation carefully, adopting in the process the best of both worlds, print lexicography and web design, in order to facilitate a successful user experience. To this end, specific technical and design expertise is required in order to take a wide variety of decisions in the design process in consultation with the lexicographers responsible for the content. The fact that this is being accomplished increasingly frequently nowadays demonstrates how far the design of Internet dictionaries has developed over the last few decades.

Bibliography

Further reading

Krug, Steve (2006): *Don’t Make Me Think! A Common Sense Approach to Web Usability, Second Edition*. Berkeley: New Riders. *A classic introduction to usability.*

Wathan, Adam/Schoger, Steve (no date): *Refactoring UI*. https://www.refactoringui.com/#get-refactoring-ui” [last access: April 27, 2024]. *Wathan and Schoger are the creators of tailwind.css, an extraordinarily popular css-framework as of 2024. “Refactoring UI” was published before “tailwind.css” and by many examples it is demonstrating the application of design principles and how these principles informed design decisions behind a framework like “tailwind.css”.*

Literature

Academic literature

Almind, Richard (2005): Designing Internet Dictionaries. In: *Hermes. Journal of Linguistics* 18 (34), 37–54.

Corréard, Marie-Hélène (2002): Are space-saving strategies relevant in electronic dictionaries? In: Braasch, Anna/Povlsen, Claus (eds.): *Proceedings of the 10th EURALEX International Congress*, *Copenhagen, Denmark*, *13–17 August 2002*. København: Center for Sprogteknologi, 463–470.

Debus-Gregor, Esther/Heid, Ulrich (2013): Design criteria and 'added value' of electronic dictionaries for human users. In: Gouws, Rufus H., et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires: Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie international de lexicographie*. Berlin/New York: De Gruyter, 1001–1012.

Dziemianko, Anna (2014): On the Presentation and Placement of Collocations in Monolingual English Learners' Dictionaries: Insights into Encoding and Retention. In: *International Journal of Lexicography* 27 (3), 259–279.

Dziemianko, Anna (2015): Colours in Online Dictionaries: A Case of Functional Labels. In: *International Journal of Lexicography* 28 (1), 27–61.

Dziemianko, Anna (2016): An insight into the visual presentation of signposts in English learners' dictionaries online. In: *International Journal of Lexicography* 29 (4), 490–524.

Heid, Ulrich/Zimmermann, Jan Timo (2012): Usability testing as a tool for e-dictionary design: collocations as a case in point. In: Fjeld, Ruth V./Torjusen, Julie M. (eds.): *Proceedings of the 15th EURALEX International Congress 2012*, *Oslo, Norway*, *7–11 August 2012.* Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier, 661–671.

Hollós, Zita (2018): Datendistribution relativ zum Webdesign. Der erste Prototyp des E-KOLLEX. In: Jesenšek, Vida/Enčeva, Milka: *Wörterbuchstrukturen zwischen Theorie und Praxis*. Berlin/Boston: De Gruyter, 151–171.

Koplenig, Alexander/Müller-Spitzer, Carolin (2014): Questions of design. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/New York: De Gruyter, 189–204.

Lew, Robert (2010): Users Take Shortcuts: Navigating Dictionary Entries. In: Dykstra, Anne/Schoonheim, Tanneke (eds.): *Proceedings of the XIV Euralex International Congress*. Ljoufwert: Afuk, 1121–1132.

Lew, Robert (2011): Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries. https://repozytorium.amu.edu.pl/items/519b2f5d-d065-47f5-913c-70336f43ce34. [Last access: April 27, 2024].

Lew, Robert/Tokarek, Patryk (2010): Entry Menus in Bilingual Electronic Dictionaires. In: Granger, Sylviane/Paquot, Magali (eds.): *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve: Cahiers du Central, 145–146.

Lew, Robert/Grzelak, Marcin/Leszkowicsz, Mateusz (2013): How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. In: *Lexikos. Journal of the African Association for Lexicography* 23, 228–254.

Michaelis, Frank/Müller-Spitzer, Carolin/Wolfer, Sascha (2019): The Sintra Variations – Thinking Outside the Box in Designing Online Dictionaries. In: Kosem, Iztok/Zingano Kuhn, Tanara (eds.): *Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts. Sintra, Portugal*, *1–3 October 2019*. Brno: Lexical Computing CZ s.r.o., 43–44.

Müller-Spitzer, Carolin/Michaelis, Frank/Koplenig, Alexander (2014): Evaluation of a new web design for the dictionary portal OWID. An attempt at using eye-tracking technology. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 207–228.

Nesi, Hilary/Hua Tan, Kim (2011): The Effect of Menus and Signposting on the Speed and Accuracy of Sense Selection. In: *International Journal of Lexicography* 24(1), 79–96.

Oppentocht, Lineke/Schutz, Rik (2003): Developments in electronic dictionary design. In: van Sterkenburg, Piet (ed.): *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, 215–227.

Schmitz, Ulrich (2016): Wörterbücher als Sehflächen. In: Schierholz, Stefan, et al. (eds.): *Wörterbuchforschung und Lexikographie*. Berlin/Boston: De Gruyter, 207–225.

Spohr, Dennis (2008): Requirements for the Design of Electronic Dictionaries and a Proposal for their Formalisation. In: Bernal, Elisenda/DeCesaris, Janet (eds.): *Proceedings of the 13th EURALEX International Congress*, *Barcelona, Spain*, *15–19 July 2008*. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, 617–629.

Storjohann, Petra (2018): Commonly Confused Words in Contrastive and Dynamic Dictionary Entries. In: Čibej, Jaka, et al. (eds.): *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, Slovenia 17–21 July 2018.* Ljubljana: Ljubljana University Press, Faculty of Arts.

Swanepoel, Piet (2001): Dictionary Quality and Dictionary Design: A Methodology for Improving the Functional Quality of Dictionaries. In: *Lexikos. Journal of the African Association for Lexicography* 11, 160–190.

Tono, Yukio (2000): On the Effects of Different Types of Electronic Dictionary Interfaces on L 2 Learners' Reference Behaviour in Productive/Receptive Tasks. In: Heid, Ulrich, et al. (eds.): *Proceedings of the Ninth EURALEX International Congress. Stuttgart, Germany, August 8th–12th*. Stuttgart: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, 855–861.

Tono, Yukio (2011): Application of Eye-Tracking in EFL Learners' Dictionary Look-Up Process Research. In: *International Journal of Lexicography* 24 (1), 124–153.

Torner, Sergi/Arias-Badia, Blanca (2019): Visual Networks as a Means of Representing Collocational Information in Electronic Dictionaries. In: *International Journal of Lexicography* 32 (3), 270–295.

Dictionaries, portals and guidelines

Bootstrap Version 4.5. 2020. Introduction. Introduction. https://getbootstrap.com/docs/4.5/getting-started/introduction/ [last access: March 12, 2024].

Dictionary of South African English, s.v. "administrator, n.". https://dsae.co.za/entry/administrator/e00073 [last access: March 12, 2024].

Material Design. "Introduction." Ed. by Google. https://material.io/design/introduction [last access: March 12, 2024].

OWID – Online-Wortschatz-informationssystem Deutsch. Ed. by Leibniz-Institut für Deutsche Sprache Mannheim, 2008–. https://www.owid.de [last access: March 12, 2024].

Paronyme – Dynamisch im Kontrast. Ed. by Leibniz-Institut für Deutsche Sprache Mannheim. 2018–. https://www.owid.de/parowb [last access: March 12, 2024].

Web Content Accessibility Guidelines (WCAG) 2.0. Ed. by Bend Caldwell, Micheal Cooper, Loretta Guarino Reid, and Gregg Vanderheiden. Last modified December 11, 2008. https://www.w3.org/TR/WCAG20/ [last access: March 12, 2024].

Images

Fig. 6.1 copyright by authors.

Fig. 6.2 Entry *administrator* in the *Dictionary of South African English*. https://dsae.co.za/entryentry/administrator/e00073 [last access: April 4, 2024].

Fig. 6.3 Examples of design primitives in the website of the *Dictonary of South African English*. https://dsae.co.za/ [last access: April 4, 2024].

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

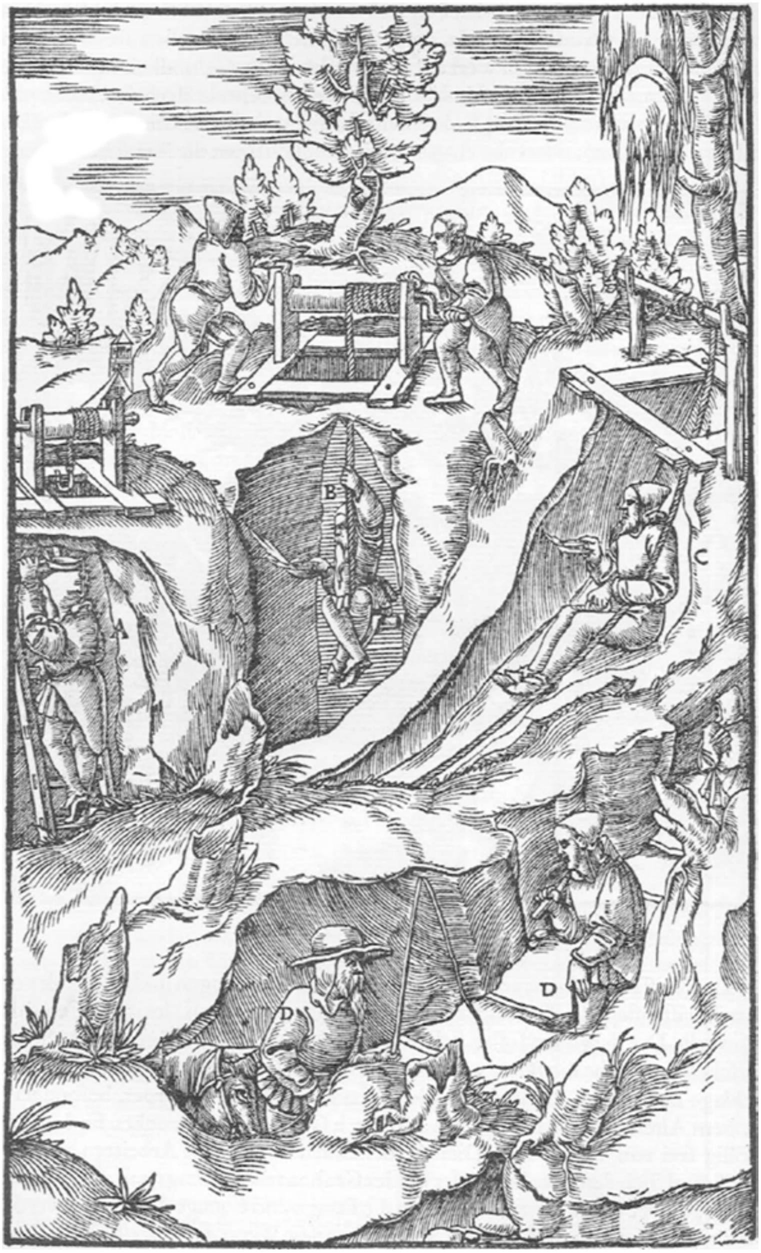
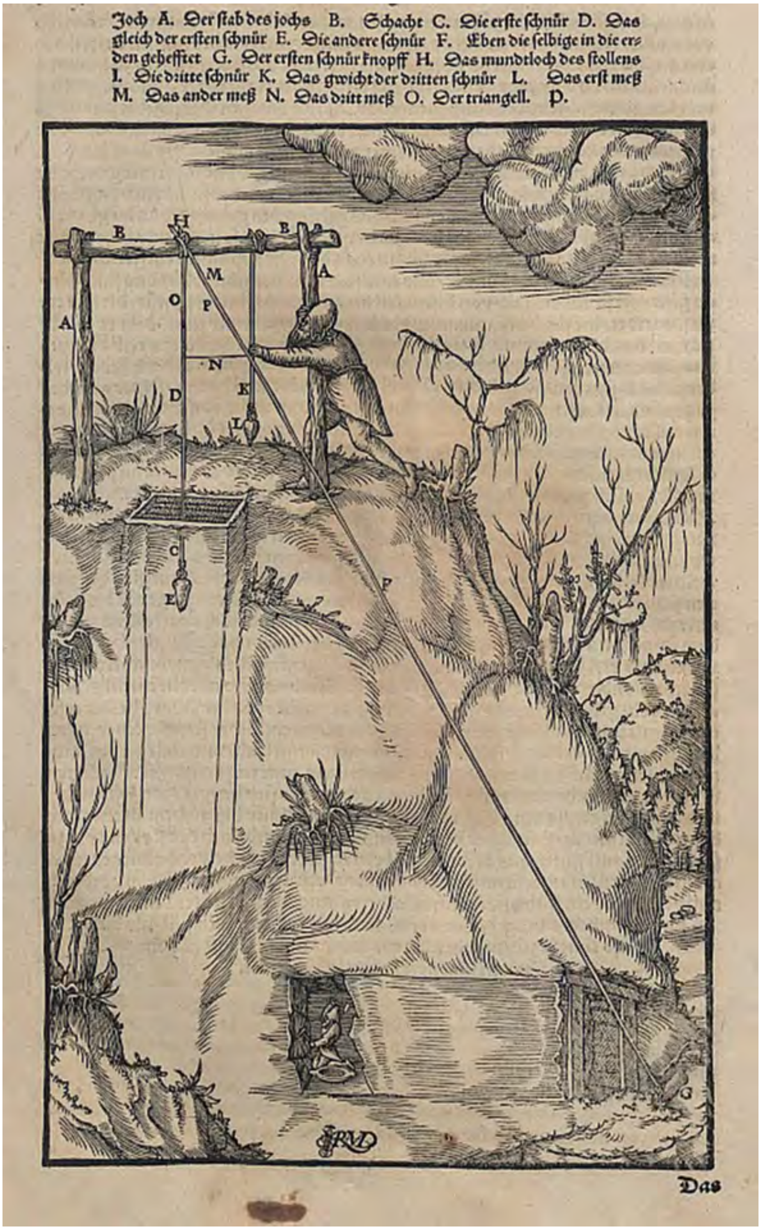


Fig. 7.1: Different forms of manriding.

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



Joch A. Der ſtab des jochs B. Schacht C. Die erſte ſchnůr D. Das gleich der erſten ſchnůr E. Die andere ſchnůr F. Eben die ſelbige in die erden gehefftet G. Der erſten ſchnůr knopff H. Das mundtloch des ſtollens I. Die dꝛitte ſchnůr K. Das gwicht der dꝛitten ſchnůr L. Das erſt meß M. Das ander meß N. Das dꝛitt meß O. Der triangell. P.

Das

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 critical 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 20th century, which is balanced chronologically and by text types across the whole of the 20th 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). However, 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 occurrences 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.

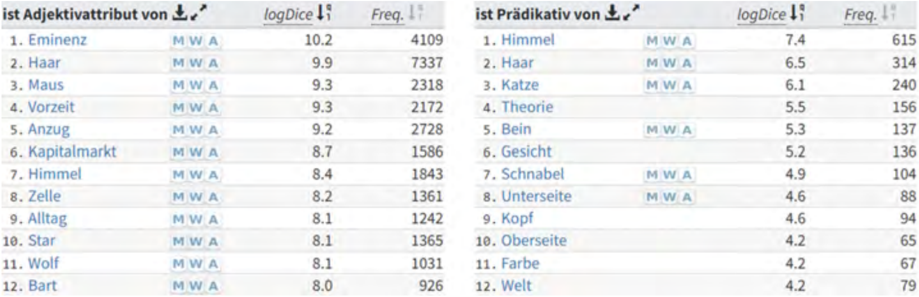
Probably the best known method for extracting syntactic co-occurrences is SKETCH ENGINE (Kilgarriff et al. 2004), a method which can extract and classify co-occurrences in a targeted way following grammatical patterns. In other words, only those co-occurrences are considered that exist in a pre-defined 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.

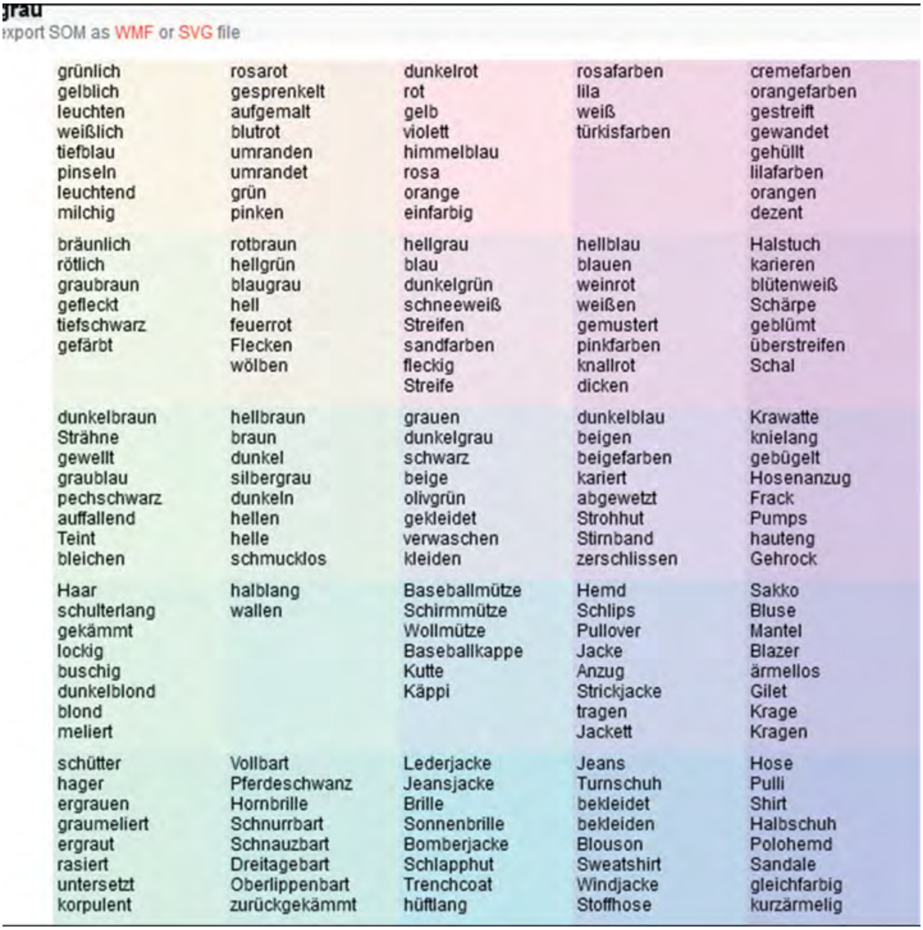
A feature of the CCDB not available in other tools is the attempt to group the collocators of a basis word (automatically) according to meaning nuance. The results for the keyword *grau* are shown in → Fig. 7.4.

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'.



rau

xport SOM as WMF or SVG file

grünlich gelblich leuchten weißlich tiefblau pinseln leuchtend milchig	rosarot gesprenkelt aufgemalt blutrot umranden umrandet grün pinken	dunkelrot rot gelb violett himmelblau rosa orange einfarbig	rosafarben lila weiß türkisfarben	cremefarben orangefarben gestreift gewandet gehüllt lilafarben orangen dezent
bräunlich rötlich graubraun gefleckt tiefschwarz gefärbt	rotbraun hellgrün blaugrau hell feuerrot Flecken wölben	hellgrau blau dunkelgrün schneeweiß Streifen sandfarben fleckig Streife	hellblau blauen weinrot weißen gemustert pinkfarben knallrot dicken	Halstuch karieren blütenweiß Schärpe geblümt überstreifen Schal
dunkelbraun Strähne gewellt graublau pechschwarz auffallend Teint bleichen	hellbraun braun dunkel silbergrau dunkeln hellen helle schmucklos	grauen dunkelgrau schwarz beige olivgrün gekleidet verwaschen kleiden	dunkelblau beigen beigefarben kariert abgewetzt Strohhut Stirnband zerschlissen	Krawatte knielang gebügelt Hosenanzug Frack Pumps hauteng Gehrock
Haar schulterlang gekämmt lockig buschig dunkelblond blond meliert	halblang wallen	Baseballmütze Schirmmütze Wollmütze Baseballkappe Kutte Käppi	Hemd Schlips Pullover Jacke Anzug Strickjacke tragen Jackett	Sakko Bluse Mantel Blazer ärmellos Gilet Krage Kragen
schütter hager ergrauen graumeliert ergraut rasiert untersetzt korpulent	Vollbart Pferdeschwanz Hornbrille Schnurrbart Schnauzbart Dreitagebart Oberlippenbart zurückgekämmt	Lederjacke Jeansjacke Brille Sonnenbrille Bomberjacke Schlapphut Trenchcoat hüftlang	Jeans Turnschuh bekleidet bekleiden Blouson Sweatshirt Windjacke Stoffhose	Hose Pulli Shirt Halbschuh Polohemd Sandale gleichfarbig kurzärmelig

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 provided 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 20th 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 aspects, 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 dictionary 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 automatically 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 Lexikografie. 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.

Andrea Abel and Christian M. Meyer

8 User Participation

Election of councillors to Coventry City Council
Vote for only one candidate by putting a cross ☒ in the box next to your choice
DONNELLY, Jim
(address in Coventry)
Trade Unionist and Socialist Coalition
TUSC AGAINST CUTS
JOBBAR, Abdul
(address in Coventry)
The Labour Party Candidate
Labour
MARSHALL, Jess
39 Westcotes, Tile Hill, Coventry CV4 9BB
Green Party
Green Party
MASIH, Asha
(address in Coventry)
Conservative and Unionist Party
Conservatives
SIMPSON, Jamie
(address in Coventry)
Liberal Democrat
Liberal Democrats
STRINGFELLOW, Danielle
Studiengebühren sind unsozial
FSI INFORMATIK
Josh Olson ("A History of Violence"), who sold a pitch just
[See Scriptland, Page E16]
in black and white, a
the 2nd Amendment.

Fig. 8.1: Direct, indirect, and complementary forms of political participation.

Andrea Abel, Free University of Bozen-Bolzano, Faculty of Education, Regensburger Allee 16/Viale Ratisbona 16, 39042 Brixen/Bressanone, Italy, e-mail: andrea.abel@unibz.it; EURAC Research, Institute for Applied Linguistics, Drususallee 1/Viale Druso 1, 39100 Bozen/Bolzano, Italy, e-mail: andrea.abel@eurac.edu

Christian M. Meyer, Technische Universität Darmstadt, Ubiquitous Knowledge Processing Lab, 64289 Darmstadt, Germany, e-mail: research@chmeyer.de

Open Access. © 2024 the author(s), published by De Gruyter. CC BY This work is licensed under the Creative Commons Attribution 4.0 International License.

https://doi.org/10.1515/9783111233758-009

Democracy lives in the participation of citizens, who can contribute to the politics of a state, both directly by electing representatives and also indirectly by expressing their ideas, suggestions, or wishes. In addition to these ways of directly and indirectly shaping politics, people inform themselves about political happenings and exchange their views about them with one another. For dictionaries, a similar breadth of options exists for participation: users swap ideas with those providing the dictionaries and among themselves. They formulate requests, give feedback, or play an active role in creating dictionary entries.

The Internet is increasingly shaping our society and connecting us ever more strongly with one another. By the mid-1990s, the use of so-called social media technologies, such as blogs, wikis, or social networks, had changed the Internet from a repository of curated expert information into an interactive platform for exchanging user-generated content. In this changed environment, referred to as "Web 2.0", Internet dictionaries increasingly involve dictionary users in lexicographic activities. The degree of user participation ranges from user-driven compilations of entire dictionaries and feedback about the quality of individual entries to dialogues between lexicographers and users or among users themselves.

These new forms of user participation on the Web have hardly been researched. In this chapter, we provide an overview of the different possibilities for involving dictionary users directly, indirectly, or in accessory ways when compiling a dictionary. Above all, the assessment of quality and clarification of legal issues are of paramount interest in order to be able to evaluate the potential of user-generated content. In addition to a systematic categorisation of user contributions, we discuss several practical examples of individual organisational forms and examine possible motivations for active involvement. Furthermore, we seek to prompt a critical discussion about the strengths and weaknesses of different forms of user participation.

8.1 Introduction

The participation of users in compiling a dictionary is hardly a new topic in lexicography, but dates back to the pre-electronic era. Already in the 19th century, the Oxford English Dictionary (OED) established reader programmes in which the public was asked to read books, collect examples of the customary ways in which a word was used, and then send them in. Initially, these examples were unsystematic but they were then compiled more and more deliberately by prescribing lists of words, literature, and particular thematic areas for the volunteers (cf. Thier 2014).

However, the development of the Internet and the World Wide Web permitted completely new ways for involving dictionary users so that the question of user participation has become an increasingly important factor in the planning, development, and use of a dictionary. For one thing, the Web allows dictionary users to communicate with each other, which was a very laborious and time-consuming activity before – if at all possible. For another, it offers totally new possibilities for interaction, putting users in a position to compose dictionary entries independently and to improve them collaboratively. This user-driven creation of dictionary content represents a fundamental change in the lexicographic process (→ Chapter 3.4.2). Carr (1997: 214) describes this method as "bottom-up lexicography" since dictionaries are assembled "bottom up" from individual entries and user contributions into a whole. By contrast, in the traditional model of commercial or academic lexicography, dictionaries are systematically planned as a whole and then compiled "top down" by expert lexicographers.

The new forms of user participation mean that the boundary between dictionary users and dictionary editors is becoming increasingly blurred. In this context, Lew (2014: 9) proposed the portmanteau "*prosumer*" since a user is both actively involved in compiling the dictionary as a "*pro*ducer" and interested in the compiled content as a "con*sumer*".

Research into user participation in Internet dictionaries is still a fairly new topic area. For example, Wiegand et al. (2010: 17) observed:

> Allerdings sind die lexikographischen Prozesse, wie man sie bei der Entstehung von gemeinschaftlich erstellten Online-Wörterbüchern, wie dem Wiktionary, beobachten kann, mit der traditionellen Phaseneinteilung nicht mehr adäquat beschreibbar; ihre Abläufe sind bislang auch erst ansatzweise erforscht. [In any case, the lexicographic processes that we can observe in the creation of collaboratively compiled online dictionaries like Wiktionary can no longer be adequately described with the traditional division of phases; as yet, research into the way these processes work is only rudimentary.]

Storrer/Freese (1996) and Storrer (1998) were among the first to attempt a classification targeted at correcting errors, suggesting new headwords, obtaining expert contributions on certain topics, and collecting contributions by laypeople. Køhler Simonsen (2005) distinguished between *active user involvement*, by which he means feedback on the design and development of a dictionary with the help of surveys or test groups, and *lexicographic democracy*, which he describes as feedback on final articles (e.g. error corrections). While this definition is limited to giving feedback to the editorial staff, Fuertes-Olivera (2009) used the term *democratisation* in a different way, focussing on "*collective free multiple-language Internet dictionaries*" (ibid.: p. 101) such as Wikipedia and Wiktionary, which are compiled entirely by users without editorial control. Storrer (2010) introduced a similar distinction between user contributions controlled by professional editors and those created by the users themselves in a collaborative effort. Further works by Lew (2011, 2014) suggested a more detailed classification of dictionaries that allow for direct user contributions, additionally introducing *collaborative-institutional dictionaries* and dictionaries making use of external *user-generated content*. Melchior (2012, 2014) used the term *semi-collaborative dictionary* for his analysis of the Leo dictionary portal as being supported by users rather than generated by users.

In a different strand of research, de Schryver/Prinsloo (2001) coined the term *fuzzy simultaneous feedback* to point out user feedback which is available during the development of a lexicographic product. For electronic dictionaries, de Schryver/Joffe (2004) focussed primarily on log file analyses, which are a way of exploring a user's behaviour and demands without additional effort by the users (→ Chapter 1.3).

Insight into the variety of forms of user participation is of great relevance. Dictionary entries and background material that are contributed voluntarily as well as feedback on new and existing content have the potential to speed up the production of a dictionary, enhance its quality, and enrich its content. Publishers can save money, and users can acquire knowledge about the structure of a dictionary and its use, thereby identifying more strongly with the product. Conversely, assessing user contributions often means more work when a dictionary is being compiled. A large number of poor quality or inappropriate contributions can also lead to disruptions in the lexicographic process or to wrong and inconsistent lexicographic products.

The aim of this chapter is to describe the different types of user participation and organise them systematically. We discuss several practical examples for each type of participation, examining in particular the motivation of users, legal questions, quality issues, and the processes for submitting user contributions.

At the highest level, we distinguish between three basic types of user participation, which we shall consider in more detail below:

1. Direct user participation (→ Section 8.2);
2. Indirect user participation (→ Section 8.3);
3. Accessory user participation (→ Section 8.4).

This categorisation and the corresponding descriptions are based on Abel/Meyer (2013, 2016) and Meyer/Abel (2017), taking as their starting point the reflections by Mann (2010). → Tab. 8.1 gives an overview of the three types of user participation and their characteristics. As we shall see, this does not prevent multiple forms of user participation from being used in parallel within a single dictionary project. Our categorisation is suitable for describing user participation in individual Internet dictionaries (e.g. OED Online, Duden online, Wiktionary) and dictionary portals (cf. Storrer 2010;

Tab. 8.1: Overview of different types of user participation.

Direct user participation:	– Contributions to open-collaborative dictionaries – Contributions to collaborative-institutional dictionaries – Contributions to semi-collaborative dictionaries
Indirect user participation:	– Explicit feedback – Implicit feedback
Accessory user participation:	– Exchanges between dictionary makers and dictionary users – Exchanges between dictionary users

Engelberg/Müller-Spitzer 2013; → Chapter 2), i.e. interfaces that permit access to a whole series of dictionaries (e.g. Leo, dict.cc).

8.2 Direct user participation

Direct user participation denotes the compiling, changing, or deleting of dictionary entries by dictionary users. We distinguish between contributions to open-collaborative dictionaries, collaborative-institutional dictionaries, and semi-collaborative dictionaries. These categories will be described in more detail below.

User contributions to open-collaborative dictionaries are not subject to any editorial supervision by a fixed group of experts. Rather, the dictionary is based on entries composed and revised by a potentially unlimited number of volunteer users. All changes are directly visible in the dictionary and can, therefore, be immediately examined by other users and, if necessary, revised again. The collective knowledge of the participants – frequently referred to as "swarm intelligence" (cf. Krause/Krause 2011, Rosenberg 2015), as "collective intelligence" (cf. Malone et al. 2010), or as the "wisdom of crowds" (cf. Surowiecki 2005) – takes the place of expert knowledge. The basic assumption of this approach is that the different subjective perspectives and knowledge of the individuals involved is consolidated into a communal group dynamic and is thereby bound together into a larger whole. The open-collaborative process has especially gained popularity through the free online encyclopaedia Wikipedia and its sister project Wiktionary. With versions in 168 languages and a total of 38.7 million dictionary entries, Wiktionary is currently the largest open-collaborative dictionary.[1]

Malone et al. (2010) distinguish between economic factors (direct financial advantages, future earning potential, and practising personal competences), pleasure (enjoyment, altruism, sociability), and reputation (recognition by peers) as fundamental motivational factors behind contributing to open-collaborative resources. Other studies that deal with the possible driving forces behind contributions to online communities point to similar, and also wider, factors, such as acquiring and exchanging information, identifying with particular groups, or a sense of belonging (cf. Lampe et al. 2010, Rafaeli/Ariel 2008).

The contents of the dictionary are not tied to a particular provider or publisher, and many such resources use free licences, also known as open content licences, such as the Creative Commons series of licences through which – unlike in the classic copyright model – the distribution of content in unchanged form is generally possible for anyone as well as, to some extent, commercial use and modification – depending on the specific licence (cf. Kreutzer 2011).

[1] https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries [last access: March 22, 2024].

In addition to the licence under which the contents are published, the source or origin of the contributions is also a relevant issue. Uncovering and preventing the copying of dictionary information from other works protected by copyright pose great challenges for the providers of collaborative dictionaries. Plagiarism is hardly a new phenomenon in lexicography (cf. Hausmann 1989), where, on the whole, transcribing or copying from existing dictionaries seems to have been a common and long known practice, albeit one that has been judged differently in different contexts (cf. Landau 2001). However, this aspect has to be considered anew given the high number of participants in collaborative resources. The user communities around WIKIPEDIA and WIKTIONARY have defined comprehensive guidelines and workflows in relation to copyright issues.[2] Here, the attempt is made to provide as much information as possible through references to sources. Data of questionable origin are first put up for discussion or, possibly, removed from the dictionary.

At the same time, there were phases when large bodies of lexicographic data or entire dictionaries from sources that are freely licensed or whose copyright had already expired were integrated into WIKTIONARY. Hanks (2012) and Hanks/Franklin (2019) noted that numerous outdated meaning paraphrases were found in the English WIKTIONARY that could be traced back, in particular, to the adoption of sources that were copyright free. For example, some parts of the digitised version of WEBSTER'S REVISED UNABRIDGED DICTIONARY (WEBSTER) from 1913 were used in the English WIKTIONARY, with sometimes negative consequences for the quality of entries. Lew (2014) demonstrated this in relation to the English verb *handle* in WIKTIONARY, for which an uncommon interpretation in contemporary English was listed as the first meaning (i.e. "To use the hands"). This was one result of copying content from the old edition of WEBSTER, along with numerous uncommented archaic quotations from the bible or classical literature that were provided as lexicographic examples. In the meantime, the entry *handle* has been changed: The meaning is no longer in first place, but it is still there and still with a quotation from the bible.[3]

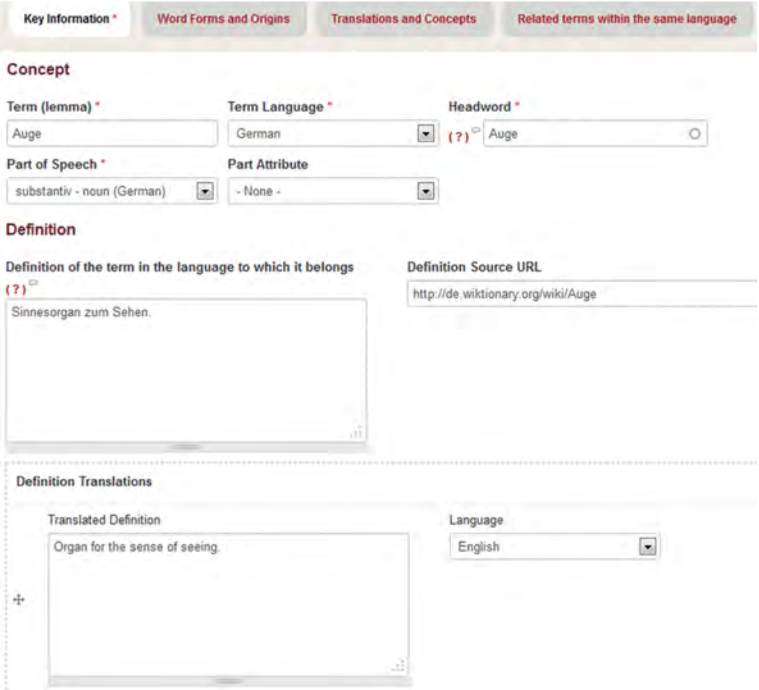
A large proportion of open-collaborative dictionaries is based on clearly prescribed lexicographic instructions that describe, at the macrostructural level, the choice of headwords to be included and, at the microstructural level, the construction of dictionary entries and the information classes to be included in them, such as pronunciation, meanings, or example sentences. The URBAN DICTIONARY, for example, focuses on English colloquial language and nonce words. It collects the associated information through an online form that permits the headword to be entered, along with an explanation of its meaning, an example of usage, and further freely selected tags (such as synonyms, misspellings, etc.). SPRACHNUDEL is a German equivalent that also covers slang and neologisms.

[2] See https://en.wiktionary.org/wiki/Wiktionary:Copyrights [last access: July 28, 2023].

[3] https://en.wiktionary.org/wiki/handle#Verb [last access: March 28, 2024].

Many of the collaborative dictionaries with fixed lexicographic instructions are translation dictionaries such as bab.la or Glosbe, whose entries are often very simply structured. For these dictionaries, the input form only covers the headword in the source and target language, which makes participation possible without great effort. Because of the large variety of language pairs and their wide potential user community, these kinds of bilingual or multilingual dictionaries benefit particularly from direct user contributions (cf. Meyer/Gurevych 2012).

More comprehensive input forms are required for more complex entry structures in order to capture all of the classes of information, for example separate text fields for meaning paraphrases, usage examples, sources, a tabular input for synonyms and translations, or a selection field for the word class. For instance, the multilingual dictionary project Kamusi allows explanations of meaning to be given in multiple languages. The input forms are adapted for exactly this case, and compiling or editing an entry is modelled as a multi-stage, interactive process. → Fig. 8.2 shows an extract from these input options. Benjamin (2014) discusses contributions to Kamusi and also discusses the challenges of such a complex article structure, such as users frequently typing information in the wrong field.


Key Information *
Word Forms and Origins
Translations and Concepts
Related terms within the same language

Concept

Term (lemma) *: Auge

Term Language *: German

Headword * (?): Auge

Part of Speech *: substantiv - noun (German)

Part Attribute: - None -

Definition

Definition of the term in the language to which it belongs (?): Sinnesorgan zum Sehen.

Definition Source URL: http://de.wiktionary.org/wiki/Auge

Definition Translations

Translated Definition: Organ for the sense of seeing.

Language: English

Fig. 8.2: Extract from the input forms in Kamusi.

Beside these form-based input options, we can also find dictionaries whose entries are composed in a markup language. Here, the dictionary content contributed is not distributed across several predefined fields belonging to particular information classes but is rather written using a specific syntax that defines the formatting (e.g. bold face, italics, coloured text) and logical structure (e.g. headlines or the beginning and end of a meaning explanation) for individual information classes.

The XML markup language that is often employed for expert-built lexicographic products was already introduced in → Chapter 4. XML and the dictionary standards based on it, such as the Text Encoding Initiative (TEI)[4] and the Lexical Markup Framework (LMF; Francopoulo 2013), make it possible to represent very complex lexicographic data structures. However, this expert markup requires a high degree of knowledge to master it. As a result, dictionary standards of this type are predominantly aimed at professional lexicographers and are hardly ever employed for collaboratively compiled dictionaries.

Instead, these dictionaries are often based on wiki technology, which provides fairly simple ways of writing and revising content. A wiki is an online platform through which texts can be collaboratively composed and edited by users themselves. The texts are written with the help of a special markup language, the so-called wiki markup, which provides both the established formatting options (e.g. bold face) and the definition of reusable text blocks (e.g. a table of inflected word forms) while also being easy to learn. Above all, the wiki markup language should reduce the inhibitions of users who are less comfortable with technology. The English-language Rap Dictionary is an example of a wiki-based dictionary of this kind; its entry for *Cheeser* is structured as follows in the Wiki markup language:[5]

```
===noun===
'''Cheeser'''

A person that trys to become closer to you using all ways for the
purpose of having your money.

''Becareful of the cheesers, the teasers''-- [[ Grand Pupa ]](Song:I
like It, Album: 2000 - 1995)

 [[ Category:Terms ]]
 [[ Category:Nouns ]]
```

[4] https://tei-c.org/ [last access: September 9, 2023].

[5] https://web.archive.org/web/20140429154439/http://www.rapdict.org/Cheeser [last access: March 29, 2024].

The text set within three equals signs produces a heading. Italics are activated by two inverted commas and bold text by three inverted commas. Cross-references to other headwords can be marked with square brackets, as are classifications of the entry into the categories “terms” and “nouns”.

In contrast to form-based input, markup languages make it possible to define, organise, and position lexicographic information freely. Thus, wiki-based dictionaries are not limited to particular, pre-defined lexicographic instructions but instead allow participants to determine these and change them themselves, thereby becoming involved in the planning and conceptual development of the dictionary. Matuschek et al. (2018) compare Omega Wiki and Wiktionary, two open-collaborative dictionaries with more rigid vs more flexible microstructures and with prescribed vs variable lexicographic instructions. Here, it becomes clear that a more flexible approach, like that in Wiktionary, offers noticeably more organisational options for an entry, for example, through hierarchical division of the explanations of meanings for entries with many possible meanings. At the same time, inconsistencies between the various entries arise very easily in this kind of dictionary, which can, in turn, hinder efficient use of the dictionary.

Since direct contributions to collaborative dictionaries are not checked by experts, we find two types of quality issues: first spam and vandalism; and second descriptions that are vague, incorrect, old-fashioned, too general, and/or too complicated. By spam and vandalism, we mean nonsensical and crude content, such as clearly false information, swearwords, or derogatory comments in existing texts as well as deleting actually useful dictionary content without reason or without making at least a correction. As a result, there is a need for quality control mechanisms, particularly in large dictionary projects. In the German Wiktionary, for example, individual stages of a dictionary entry are marked as so-called flagged revisions once a certain quality standard has been reached.[6] While the label is only aimed at the absence of spam and vandalism, there were discussions as to whether a second label should be assigned for reaching minimum quality requirements in relation to the second type of quality issues. However, defining these requirements is notably more difficult than for spam and vandalism, and the questions of quality and defining quality criteria are constant points of discussion even for expert-built dictionaries (cf. Kemmer 2010).[7]

Only active users who had worked on at least 200 entries were assigned the right to flag revisions in the German Wiktionary. This prevented the label from being misused. So-called “construction site” labels were a further measure to ensure quality. Anyone who noticed a quality issue in an entry but who could not or did not want to

[6] https://en.wikipedia.org/wiki/Wikipedia:Flagged_revisions/fact_sheet [last access: March 22, 2024].

[7] The extent of research still needed, in particular in relation to Internet dictionaries, was shown by the workshop “Was ist ein gutes (Internet-)Wörterbuch? – Alte und neue Fragen zur Qualität lexikographischer Produkte im ‘digitalen Zeitalter’” [What is a good (Internet) dictionary? – Old and new questions on the quality of lexicographic products in the ‘digital age’”] at the XVI International EURALEX Congress in 2014 (cf. http://internetlexikografie.de).

correct the issue themselves was able to assign a pre-defined text block to draw attention to missing sources, necessary or useful additions, or inconsistent structure, among other things. Other contributors could then revise the entry or further discuss the quality issue. If an entry was judged to be unsuitable or irrelevant for the planned dictionary content, a separate label could be used to suggest the deletion of the entire entry (cf. the entry on the plural form *Erdoberflächen*[8] which was flagged for some time due to there being no plural for this lemma). Meyer/Gurevych (2014) discuss quality control measures in collaborative online dictionaries in more detail.

Contributions to collaborative-institutional dictionaries constitute a second kind of direct user participation (cf. Lew 2011). These collaborative-institutional dictionaries used to be provided by established publishers; examples include the former MERRIAM-WEBSTER OPEN DICTIONARY and MACMILLAN OPEN DICTIONARY projects. The user contributions in these dictionaries mostly took the form of complete, submitted dictionary entries that were checked by the editors of the dictionary for vandalism, insults, or defamation. In contrast to semi-collaborative dictionaries and explicit feedback (see below), the contributions to collaborative-institutional dictionaries remained by and large unedited. At the same time, however, there is a close connection to these two types of user contributions, which we shall consider in more detail in what follows. One difference to open-collaborative resources is that users had no way of changing or deleting someone else's contributions.

While contributors to these dictionaries presumably had similar motivations to those for open-collaborative dictionaries, providers of collaborative-institutional dictionaries had two aims: first, to gather suggestions for preparing professionally and editorially compiled dictionaries; and, second, to advertise the publishers' own activities and products. The contributions could be collected without precise guidelines on the type and scope of the entries, as with the MACMILLAN OPEN DICTIONARY, or with a particular section of language in mind, as was intended for youth language in the former Duden SZENESPRACHENWIKI. Since collaborative-institutional dictionaries were, for the most part, accompanied by dictionaries provided by professional editorial teams, they tended to contain entries that were not included in those expert-built resources. Hence, the resulting dictionaries were usually smaller than open-collaborative dictionaries. The MACMILLAN OPEN DICTIONARY, for example, contained about 11,700 entries in 2023, the year when it closed down[9]

Unlike open-collaborative resources, collaborative-institutional dictionaries mostly did not use free licences to publish their contributions. The rights of use remained either entirely with the contributing user or they were transferred entirely, or in part, to the dictionary.

[8] https://de.wiktionary.org/w/index.php?title=Erdoberfl%C3%A4chen&oldid=3753791 [last access: March 22, 2024].

[9] https://web.archive.org/web/20230216132652/https://www.macmillandictionary.com/open-dictionary/index-chronological-order_page-1.htm [last access: March 28, 2023].

Given that well-known collaborative-institutional dictionaries such as the ones listed above have been closed down and we are not aware of any active projects of this kind, this indicates that this type of direct user participation was not successful. This can be due to the economic situation of the institution, the costs of running such a service, the quality of the submissions, or the amount of work that would be needed to make professional use of these contributions (i.e. considering them as explicit feedback for an expert-built dictionary, as discussed in the next section).

Contributions to semi-collaborative dictionaries constitute the third kind of direct user participation. They are carefully checked by professional lexicographers or other language experts before being integrated into the dictionary. The TechDictionary, for example, is based on contributions on topics related to computers and technology that are written by users and only included in the dictionary after being checked. Leo, a portal of 12 bilingual dictionaries, is a prominent example of semi-collaborative dictionaries. Its central components are translation entries contributed by users as well as lists of words, terminology, and glossaries donated to the portal. After being carefully checked, the contributions are generally added directly to the dictionary but are not substantially revised, as is the case with explicit feedback (→ Section 8.3). Nevertheless, the decision whether, and how, to include a contribution in the dictionary always remains the responsibility of the dictionary publishers, so that quality control and a consistent dictionary structure are made possible.

In semi-collaborative dictionaries, the rights of use are either transferred to the providers of the resource, which is usually the case with commercial providers (e.g. Leo), or are channelled into a dictionary with a free licence, as is the case, for example, in the semi-collaborative synonym dictionary OpenThesaurus.

While these kinds of resources often enjoy high numbers of visitors, Naber (2005) found that only a small proportion of the registered users were actually actively involved in the writing of entries in the case of the synonym dictionary OpenThesaurus and that most contributions represented newly suggested synonyms, even though alterations and deletions would also be possible. Similar findings were reported for Wikipedia (cf. Javanmardi et al. 2009) and Wiktionary (cf. Meyer 2013). This kind of distribution, with extremely few very active users, on the one hand, and a very high number of users who only make a small contribution, on the other, can be found with virtually all types of user participation. This distribution can be described as a power law, which became very familiar in linguistics as Zipf's Law, for example, also in relation to the distribution of word frequency in corpora. Furthermore, it is well known that online communities have high numbers of lurkers, that is, members who only observe, without being active, for example, by making contributions or revisions (cf. Rafaeli/Ariel 2008 for a summary).

What is common to all three types of direct user participation is that user contributions are integrated directly into the dictionary. This mode of compiling dictionaries is referred to as *collaborative lexicography*. The dictionaries discussed benefit particularly from the diversity of the participants, which is, in principle, high. This applies both to the areas of knowledge covered and the forms of use of the linguistic units represented by the participants. For language varieties (e.g. youth language, technical languages, dialects) and translation dictionaries, this provides clear additional value (cf. Meyer/Gurevych 2012).

Direct user participation as has been described in this section has only become possible with the advent of Internet dictionaries and the corresponding technology since user contributions are based on the new possibilities for interaction available on the Internet. By contrast, previous options for user participation almost exclusively involved forms of indirect participation, which we consider in more detail in the next section.

8.3 Indirect user participation

Indirect user participation denotes feedback from dictionary users on existing or missing lexicographic content, on the use of the dictionary, and on the dictionary project as a whole. Among its characteristic forms are suggestions, additions, corrections, requests and opinions, externally generated content, and dictionary usage data. What is common to all of these user participation forms is that the dictionary users have no possibility of directly changing the dictionary but only the possibility of effecting an indirect change through their feedback. In the rest of this section, we distinguish between explicit and implicit feedback as the two main forms of indirect user participation.

Explicit feedback refers to contributions that users express explicitly and that they intentionally make available to the dictionary providers, e.g. suggestions for new words, corrections of errors, or comments on the organisation or presentation of the entries. Such contributions may address both new and existing dictionary content.

Submitting explicit feedback is popular, and the motivations for engaging as an active user are similar to those for direct contributions. Above all, the motivational factors referred to by Malone et al. (2010) as pleasure and reputation play a large role. In an online survey on DUDEN ONLINE, Rautmann (2014) noted that just under half of the respondents were interested in feedback options for dictionary entries. For OED ONLINE, Thier (2014: 70) found:

> Die Beiträge stammen bei weitem nicht nur von Akademikern, sondern von Menschen aus allen Teilen der Bevölkerung, die sich für ihre Sprache interessieren. [The contributions come very much not only from academics but also from people from all sections of the population who are interested in their language.]

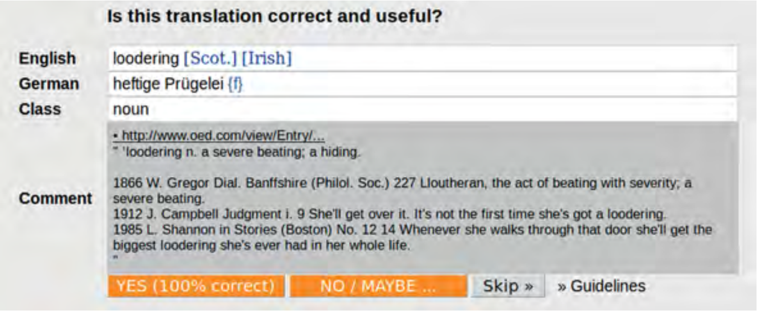
When it was launched, for example, DUDEN ONLINE offered a button for “Suggested Words” through which users could propose new headwords to be included.[10] However, additions and corrections can still be submitted by email. Rautmann’s (2014) analysis showed that more than half of the words suggested in this way met the inclu-

[10] This function is no longer available.

sion criteria of the dictionary and were envisaged to be included in a new edition, e.g. *Burgerbude* ("burger stall"). Overall, the quality and usefulness of the explicit feedback on DUDEN ONLINE was perceived by Duden's editors as predominantly positive (cf. Rautmann 2014).

We already mentioned in the introduction that submitting additional material has a long tradition at the OED, especially concerning examples of attestation. Since the mid-19th century, reader programmes have involved volunteers being encouraged to send in citations. The "Wordhunt" campaign between 2007 and 2008 and the "Science Fiction Citations" initiative (cf. Thier 2014) constitute more recent examples. The "Wordhunt" involved a BBC television programme in which viewers were encouraged to submit examples of words from a list that could be dated to an earlier point in time than given in the dictionary. By contrast, the "Science Fiction Citations" call is framed more openly; although it is no longer an official OED project, it still aims to receive submissions of examples of any concepts from science fiction literature.[11] The OED has continued with participatory campaigns in the recent past. As part of the "OED M-R antedatings" initiative launched in 2020, members of the public should find the earliest possible evidence for dictionary entries in the alphabetical range from M to R and submit their findings via an online form.[12]

In addition to new and supplementary information, user-driven assessments of quality are also included in the form of explicit feedback. For example, the DICT.CC Internet dictionary asks users to judge the accuracy of translation equivalents. Questionable equivalents and their word class are displayed on the screen and users are able to choose between "YES (100% correct)" und "NO/MAYBE" or skip to the next translation without making a decision. For example, → Fig. 8.3 shows the translation *loodering – heftige Prügelei*. In order to integrate only high-quality translations in the dictionary, the labels on the buttons have been chosen so that translations are only marked as correct if the user is certain about their decision ("YES (100% correct)").



Is this translation correct and useful?

English	loodering [Scot.] [Irish]
German	heftige Prügelei {f}
Class	noun
Comment	• http://www.oed.com/view/Entry/... " 'loodering n. a severe beating; a hiding. 1866 W. Gregor Dial. Banffshire (Philol. Soc.) 227 Lloutheran, the act of beating with severity; a severe beating. 1912 J. Campbell Judgment i. 9 She'll get over it. It's not the first time she's got a loodering. 1985 L. Shannon in Stories (Boston) No. 12 14 Whenever she walks through that door she'll get the biggest loodering she's ever had in her whole life. "

Fig. 8.3: Evaluation of quality in dict.cc.

[11] https://sfdictionary.com/about [last access: March 28, 2024].

[12] https://pages.oup.com/ol/cus/1646166399178702002/oed-m-r-antedatings [last access: March 22, 2024].

It is not unusual to find this kind of evaluation task in the field of (paid) crowdsourcing, a common strategy of companies to outsource certain tasks to volunteer participants on the Internet and thereby benefit from the "wisdom of many" or "crowd intelligence" (→ Section 8.2). Reviewing a newly developed product or online service is one example of this kind of evaluation task, often described as a Human Intelligence Task (HIT) since those asked are bringing their intuitions and intelligence to bear on solving the task, which would be impossible or difficult to complete with a machine. For example, businesses set out HITs in which participants have to indicate the best shop category for a particular product that is perhaps difficult to categorise. Designers of user interfaces can use HITs to test whether, for example, the colour is felt to be pleasant and whether users are able to find their way around quickly. Equally, product developers can survey a wide user group to assess the importance of particular product features. Crowdsourcing is also used in the field of computational linguistics research to generate training and evaluation data on, for example, whether an automatically created summary has been successful or not. In order to find participants for these kinds of tasks, the HITs are posted on crowdsourcing platforms like Crowdflower or Amazon Mechanical Turk and renumerated with small sums of money (e.g. USD 0.05; cf. Fort et al. 2011). From the perspective of dictionary research, we can consider not only quality evaluations but also user research questioning (→ Chapter 9) as crowdsourcing activities. As far as we are aware, though, crowdsourcing platforms have not yet been used for this kind of questioning.

However, the basic idea behind crowdsourcing is not limited to paid evaluation tasks. In the broadest sense, all forms of "crowd intelligence" can be understood as crowdsourcing, including the volunteer contributions in collaborative dictionaries (→ Section 8.2). A particular form of crowdsourcing is crowdfunding, a way of fundraising on the Internet, in which a project is intended to be financed by small payments from as many users as possible (cf. Howe 2008). In the field of dictionaries, crowdfunding could be used to finance new dictionaries or existing active dictionaries that are under development. Meyer/Gurevych (2014) discussed this form of user participation with the example of Nitty Grits: a crowdfunding campaign run by the Southern Food and Beverage Museum was intended to raise the resources necessary to revise a dictionary of food and culinary terms in order to make it the definitive International Culinary Dictionary, but this was not achieved.[13]

We distinguish form-based feedback, where user submissions take place through an online form with fixed, pre-determined fields, and free-text feedback, where no further restrictions on the form of the feedback are provided, such as an email with an arbitrary text. To a certain extent, form-based feedback makes it possible to guide the type and volume of submissions received. For example, the Leo dictionaries provide

[13] https://www.indiegogo.com/projects/nitty-grits-the-international-culinary-dictionary#/ [last access: October 24, 2023].

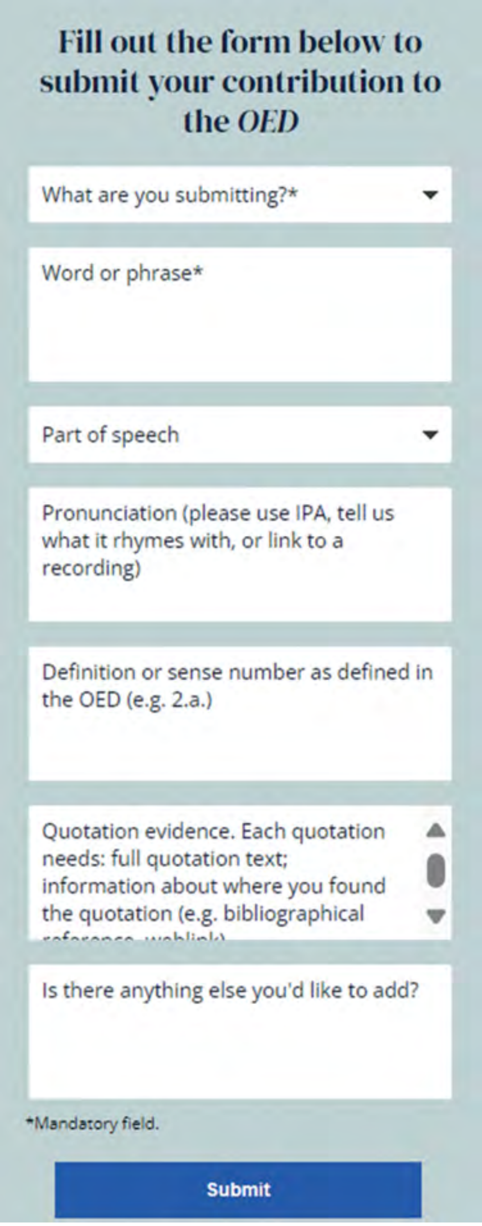
different forms for corrections and for suggesting new entries or translations, meaning that submissions from users are pre-sorted and relevant details can be requested in a targeted way. The forms used by LEO contain easily understood fields and can therefore be completed by contributors with little effort. It should also be noted that the suggestions are submitted in the forum area so that other users can also comment on or add to the suggestions. The OED ONLINE uses one single form, which is shown in part in → Fig. 8.4. Detailed information can be requested in a form of this kind (cf. Thier 2014), e.g. the bibliographic data of the sources indicated. Complex forms can, however, inhibit users. As a result, fewer, but possibly more accurate, user contributions can be expected when using complex rather than simple forms. For dictionary providers, this can be a way to steer the quality and volume of user feedback.

In addition, the OED ONLINE offers the option to submit free-text feedback by post or email. These contributions first have to be checked and categorised by the editors and so sometimes must represent considerable additional work. When it comes to error corrections, for example, they must check whether the problem listed can actually be found in the dictionary with the information provided. However, the free-text feedback evaluated by DUDEN ONLINE shows that the majority of submissions were useful for the editorial work on the dictionary (cf. Rautmann 2014).

As well as providing explicit feedback on particular dictionary entries, users can comment on the dictionary as a whole. This includes both content-related aspects (e.g. the choice of headwords) and layout or organisational aspects. Melchior (2012: 359–367) analysed these kinds of user submissions for the LEO German-Italian dictionary and characterised eight different types of users on this basis. Tensions arise when different types of users come into contact with one another, for example, users who wish for neologisms and nonce words to be included promptly and users who view the dictionary as a “moral compass”, demanding that vulgar expressions are removed.

Feedback on the structure and organisation of a dictionary can also be sought by lexicographers by publishing beta or advance versions (cf. Melchior 2014). This enables different layout versions to be tested at the same time or one after the other, without compromising access to the actual dictionary. This kind of beta version was, for example, made available for the DIGITALES WÖRTERBUCH DER DEUTSCHEN SPRACHE (DWDS) (cf. Klein/Geyken 2010).

The boundary between direct contributions to semi-collaborative dictionaries and indirect contributions in the form of explicit feedback is fluid. For example, the submission of a new translation to one of the LEO dictionaries can be integrated into the dictionary without extensive editorial work (as long as the translation is accurate). In this case, we are talking about a direct user contribution to a semi-collaborative dictionary. However, we count a citation submitted to supplement an entry in the OED ONLINE as explicit feedback since the submission is neither a complete dictionary entry nor will the submission be immediately integrated into the dictionary. Rather, the editorial team has to decide whether the citation is relevant and informative for the existing



Fill out the form below to submit your contribution to the *OED*

What are you submitting?*

Word or phrase*

Part of speech

Pronunciation (please use IPA, tell us what it rhymes with, or link to a recording)

Definition or sense number as defined in the OED (e.g. 2.a.)

Quotation evidence. Each quotation needs: full quotation text; information about where you found the quotation (e.g. bibliographical reference, weblink)

Is there anything else you'd like to add?

*Mandatory field.

Fig. 8.4: Input form for submitting examples to the OED Online.

entry, whether it can be verified, and in which form it can be integrated into the entry (e.g. which context is required).

In contrast to explicit feedback, *implicit feedback* arises without any input from individuals; it is often, in fact, unintentional and without dictionary users being aware that they are providing feedback. This kind of user contribution includes records about dictionary usage and external contributions that are integrated into a dictionary without being compiled specially for this purpose.

Records of dictionary usage are employed in dictionary research as an instrument to understand the behaviour of users and thereby adapt the dictionary more effectively to their information needs. Log data (→ Chapter 1.3) often form the basis of these kinds of analysis. These logs automatically capture every access to a dictionary entry along with the access date and time and potentially the retention time, search terms, and navigation history. Ready-made software solutions are available to analyse log data, e.g. GOOGLE ANALYTICS or MATOMO. Such tools process the raw log files and report the most frequently visited pages, the average time spent by users, and frequent navigational patterns. At the same time, the data protection requirements in the countries concerned always have to be taken into consideration when recoding and analysing log data.

This kind of evaluation is known in the context of DUDEN ONLINE for example (cf. Rautmann 2014). In the process, the Duden editors receive access to the list of the most frequently read entries. In addition to optimising the dictionary towards the entries that are regularly consulted, log data can be used to improve access to dictionary contents. To achieve this, they are filtered to show unsuccessful searches so that the users' search strategies can be analysed more closely or potential gaps may be revealed. It has been shown, for example, that expressions of more than one word, such as *im Folgenden* ('in the following') or *des Weiteren* ('furthermore') are often entered into the DUDEN ONLINE search field. For reasons of space in print dictionaries, information on these constructions are primarily found in the examples section for the relevant lemma. However, high demand indicated by the log data analysis has prompted the editors to broaden their headword guidelines so that these frequently consulted multi-word expressions appear as separate dictionary entries in addition to the existing descriptions.

Some dictionaries provide returning users with a log-in screen in order to personalise their use of the dictionary, for example by being able to view a list of their own previous search requests. For these dictionaries, more extensive log data can be captured and user behaviour evaluated over a longer time period. Already in the early 2000s, profiles were generated for the ELEKTRONISCHES LERNERWÖRTERBUCH DEUTSCH–ITALIENISCH (ELDIT) based on user log-ins; these are characterised by the headwords and information classes that users consult (cf. Abel et al. 2003). A similar analysis was conducted for the BASE LEXICALE DU FRANÇAIS (BLF) in which the search and consultation behaviour of the participating users was analysed in addition to the headwords and multi-word expressions in their search requests. Among other things, this showed that users were mostly seeking information about meanings and grammatical gender, the latter being a typical problem for learners of French (Verlinde/Binon 2010).

However, analysing log data usually does not provide precise results: for one thing, access by automated computer programs and search engines cannot be filtered out well enough; for another, there is no exact record of reading time or whether an attempt to look something up was successful. For example, Verlinde/Binon (2010) observed that over 90% of the page visits were caused by automated search engines checking the website for new or updated content. However, these automated visits cannot always be distinguished clearly from a human visit to the page, which leads to so-called noise, that is, inaccuracies in measurement in the generated data. This noise can be reduced to a certain extent through automatic procedures to clean the data, but log data analysis is often criticised for being superficial and limited in its meaningfulness (cf. Müller-Spitzer/Möhrs 2008; Verlinde/Binon 2010). Newer works rely on data cleaning and statistical measures in order to analyse the relation of look-up frequency and corpus frequency (Müller-Spitzer et al. 2015; de Schryver et al. 2019).

At the same time, users' registered accounts supply dictionary providers with additional implicit feedback about the use of the dictionary. On MERRIAM-WEBSTER ONLINE users can, for example, mark individual dictionary entries as favourites; in this way the editors receive additional information on particularly popular entries so that they can cultivate and develop these. Similar options are available in DICTIONARY.COM and WORDNIK, where favourites can also be organised in user-defined lists. The title and composition of this kind of word list provide further information on users' needs and their behaviour when looking up words. In WORDNIK, for example, we can find a user-generated word list of about 3,000 academic terms, a list of 100 colour names and a list of about 600 words that a user has marked as "learned".[14]

Indirect user contributions are not limited exclusively to the dictionary itself but can also be drawn from external sources and displayed as part of dictionary entries. This form of implicit feedback involves external user-generated content. This external content includes messages or blog posts about a particular headword as well as illustrations, videos, and audio data that have been contributed by users to other online sites. For example, WORDNIK allows photographs from Flickr to be included in their dictionary entries (cf. McKean 2017: 473). As with direct user contributions, adhering to copyright for external user-generated content is also an important aspect of dictionary planning. For example, when incorporating Flickr images, WORDNIK indicates that the photographs are subject to a CREATIVE COMMONS licence. Avoiding inappropriate content is another important issue. Lew (2014), for example, discussed how inappropriate images were displayed in the retired GOOGLE DICTIONARY, which showed automatally retrieved illustrations from the Google image search in its dictionary entries until 2011. As external content is continuously changing and as the images were integrated into the display for a dictionary article in a fully automated way, it was

[14] https://www.wordnik.com/lists/academic-words–4, https://www.wordnik.com/lists/great-color-names, https://www.wordnik.com/lists/learned-words–1 [last access: March 29, 2024].

almost impossible to manually check whether they improve the lexicographical descriptions and respect copyright and social norms. This is why publishers declare limitations on liability when (external) user-generated content is used, e.g. Wikipedia.[15] Another approach is using artificial intelligence methods to filter out unsuitable contributions. For example, Wang/McKeown (2010) employed language technology to detect vandalising changes in WIKIPEDIA. To do this, they modelled and automatically analysed different forms of vandalism with particular attention paid to syntactic features (e.g. syntactically incorrect sentences), lexical features (e.g. certain lexical elements, including Web slang like "LOL", "haha", etc., often accompanied by noticeable repetition of punctuation, such as "!!!!!!", and comments on revisions), and semantic features, which is a particularly challenging task (e.g. words or word meanings that do not fit in the given context or are thematically unsuitable), as well as the editing history of individual authors.

8.4 Accessory user participation

Accessory user participation denotes exchanges between dictionary compilers and users or among dictionary users themselves. In this way, it describes a kind of integration that is located beyond the contents of the dictionary but focussing on the macrostructure (i.e. the selection and organisation of the lemmas) or microstructure (i.e., the organisation and format of individual dictionary entries).

If these are exchanges in which the dictionary compilers address the users and provide them with information, without a reaction being demanded or being possible, we can refer to *unidirectional communication*. Blogs represent a typical example of this kind of communication. For example, some dictionary publishers post blogs in which they report on interesting, surprising, or amusing topics about language use or language history. The MACMILLAN DICTIONARY blog, for example, used to contain a collection of dictionary-related resources that is now partially accessible on Macmillan Education's website.[16] In 2013, the publisher launched the rubric "Stories behind Words",[17] in which teachers, authors, linguists, and general language enthusiasts were asked to write about anecdotes or experiences relating to words. In this case, the publishers employed user contributions to address their audience in a unidirectional manner.

Blog contributions often contain hyperlinks to dictionary entries and are thereby intended to help advertise the publishers' own products as well as to bind users and

[15] https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use#16._Limitation_on_Liability [last access: March 29, 2024].

[16] https://www.onestopenglish.com/adults/vocabulary/macmillan-dictionary-blog [last access: October 24, 2023].

[17] https://www.onestopenglish.com/stories-behind-words/552993.article [last access: October 24, 2023].

customers to their brand. Using newsletters or social networks like Facebook or microblogging services like X, formerly known as Twitter, to disseminate product information represents a similar approach. For example, the OED ONLINE – like other publishers – uses a whole spectrum of unidirectional communication options in order to reach users. Services including blogs, social media, and video platforms are driven by the marketing department but they make available content created by dictionary staff (Thier 2014).

Language games are another type of popular service offered by different publishers or institutions. In 2010, for example, the Dutch ALGEMEEN NEDERLANDS WOORDENBOEK (ANW) invited users to search for "the lost word" in their game "Het Verloren Woord". Those interested received a series of cryptic descriptions at set intervals: for example, the phrase *niet vroeg* ('not early') led to the word *laat* ('late') and from this palindrome read backwards, the word *taal* ('language') had to be deduced (Schoonheim et al. 2012: 975). Here, participants were able to exchange ideas with other users and receive feedback from the organisers. However, in order to solve the task, it was necessary more than anything to use the dictionaries of the Instituut voor de Nederlandse Taal; this not only raised awareness of those dictionaries but also encouraged the use of the dictionaries in a playful way, thereby achieving an educational goal (Schoonheim et al. 2012).

In 2023, the Danish Dictionary DDO launched a quiz in collaboration with the magazine "DM Akademikerbladet". Under the motto "Test dig selv: Fostår du ordbogens nye ord?" ["Test yourself: do you understand the dictionary's new word?"] participants could playfully find out how well they understood Danish neologisms already included in the dictionary. Such initiatives not only contribute to the visibility of dictionaries, but also raise awareness of the fact that dictionaries adapt to changes in the language.[18]

In many cases, users also have the option to engage more actively in these forms of communication, for example, by commenting on an announcement, evaluating contributions or suggesting new topics, thereby helping publishers to orient their offer more effectively to the demand. If this kind of mutual exchange between dictionary makers and dictionary users takes place, we can talk of *bidirectional communication*. The boundaries between unidirectional and bidirectional communication are fluid in many ways since users may also respond to forms of unidirectional communication (e.g. in an email or phone call) and, likewise, there might be no response at all, even if bidirectional communication were technically possible.

Language advice services constitute a particular kind of bidirectional communication. Since the 1960s, the Duden editors have offered telephone help, providing further assistance on language-related questions to users who, for example, have been unable to find what they need in one of the publisher's dictionaries. In keeping with the motto "There are no stupid questions! – Every question is answered", users can

[18] https://www.akademikerbladet.dk/aktuelt/2023/november/test-dig-selv-forstaar-du-ordbogens-nye-ord [last access: March 22, 2024].

direct language-related questions by email to an expert in the CANOONET language blog "Ask Dr. Bopp", which moved to the LEO language blog in 2020.[19] In addition to a personalised answer to the specific language question, recurring or interesting examples are often made available on the blog for a wider number of users. These resources offer useful insights into the information needs of users and, thus, can contribute to improving and adapting dictionary content. Furthermore, the expert answers in the particular case of Dr. Bopp often refer users to dictionary content or other content from the website so that these are indirectly promoted.

Accessory user contributions are not limited to communication between dictionary providers and users. Thanks to the technology of Web 2.0, the opportunities for users to communicate among themselves are also increasing. One popular option in this context is the forum in the LEO portal. If we take the German compound *Nutzerbindung* (literally: 'user binding') as an example, there was still no English translation given in the German-English LEO dictionary when this chapter was first written.[20] One user posted their query about a suitable equivalent in the forum, describing the meaning of the term in German as follows: "It means binding users to a website (e.g. with an interesting offer) and motivating them to return to the website". The user wanted to know if the literal translation "user binding" could be used in English. In response, another user suggested ". . . to build a loyal customer base . . . to get repeat business (or customers)". This example brings home to us that reciprocal user participation sometimes represents an important addition to the dictionary content itself, above all by allowing users to explore specific language questions in a very specific way.

Discussion pages and comments are a further form of mutual exchange among users. On WORDNIK, users can comment on dictionary entries. This function is meant to be used to react to entries, to ask questions, or simply to express one's own opinion on words but it is also used to express views on content that is hardly related to the content of the entry at all. For example, comments on the headword *dictionary* range from preferences for particular Internet dictionaries to descriptions of terms like *lexicography*.[21]

Discussion pages in WIKTIONARY make it possible to discuss each individual dictionary entry on its own page. Unlike comments or forums, user contributions on these pages are not tied to a chronological structure but can be placed anywhere. As a result, different aspects can be discussed in parallel (cf. Ferschke et al. 2013). → Fig. 8.5 shows an extract from a discussion on the meaning description in WIKTIONARY for the headword *Kreuzung* ('crossroads, junction, intersection'). A core question being discussed is whether a road, by definition, ends at an intersection or continues across it, which has implications when defining the term as a crossing of four or only two

[19] https://blog.leo.org/ [last access: October 24, 2023].

[20] https://dict.leo.org/forum/viewUnsolvedquery.php?idThread=88976&lang=en [last access: October 24, 2023].

[21] https://www.wordnik.com/words/dictionary [last access: October 24, 2023].

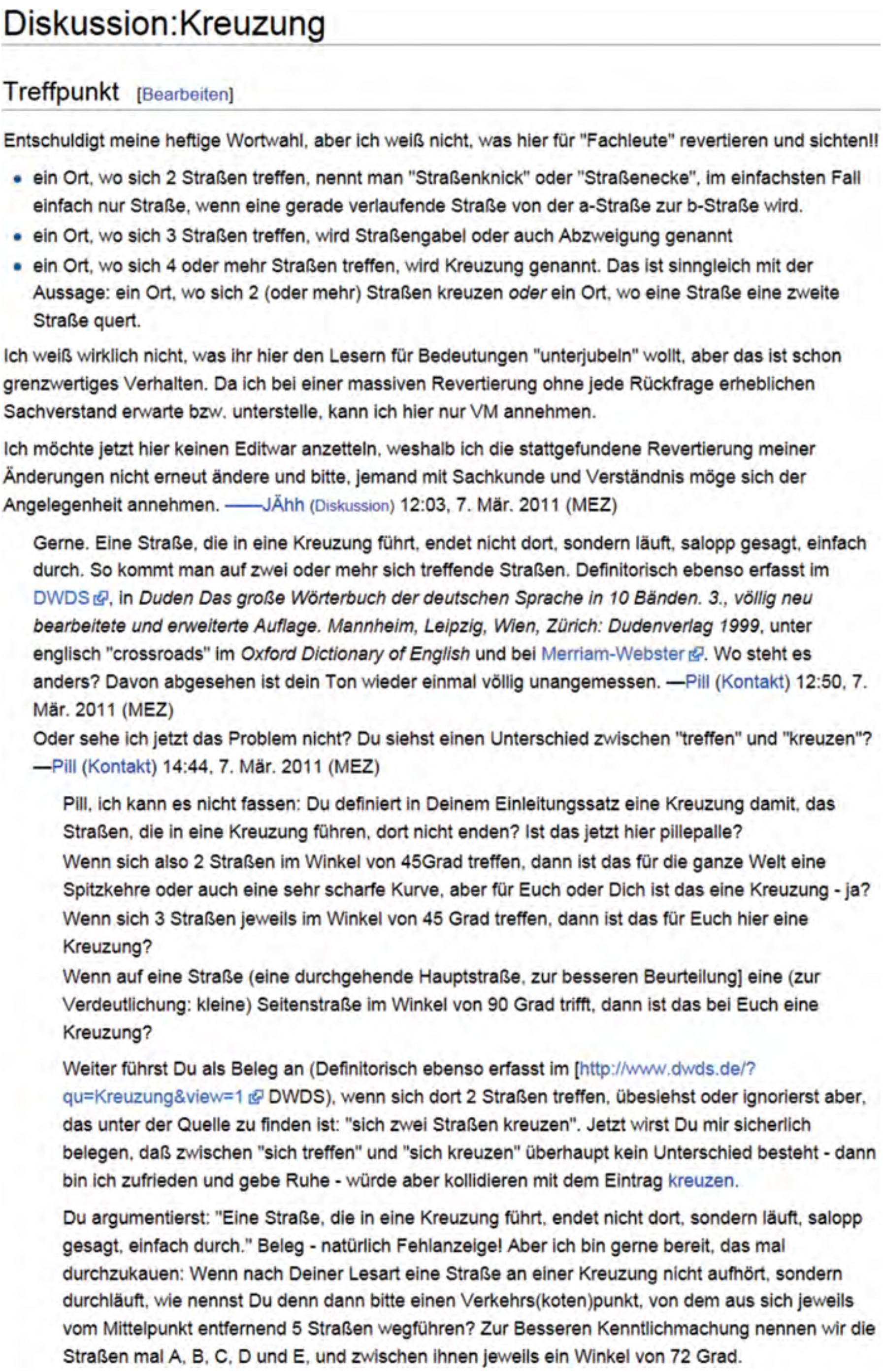
roads. Among other things, the extract makes clear what an important role sources perceived as authorities, like DUDEN or the DWDS, play in users' argumentation but also how vehemently these discussions can be conducted, particular if, as in this specific example, an "edit war" is to be averted. In the specific example (→ Fig. 8.5), a registered user is annoyed about the reversion of a change he/she made to the entry *Kreuzung*: "Sorry for my strong choice of words, but I don't know what kind of 'experts' are reverting and reviewing here!" She/he emphasises that only "a place where 4 or more roads meet is called a crossroads. This is equivalent to saying: a place where 2 (or more) roads intersect OR a place where one road crosses a second road. I really don't know what meanings you are trying to 'palm off' on the readers here, but this is borderline behaviour. [. . .] I don't want to start an edit war here, which is why I won't change the reversion of my changes again and ask someone with expertise and understanding to take care of the matter." Another user replies: "Gladly. A road that leads to a crossroad doesn't end there but simply runs through it. So you end up with two or more roads meeting. Defined in the same way in the DWDS, in Duden Das große Wörterbuch der deutschen Sprache [. . .]. Where does it say otherwise? Apart from that, your tone is once again completely inappropriate. [. . .] Or am I not seeing the problem now? You see a difference between 'meet' and 'cross'?" And the dispute continues in this tone of voice.

In essence, accessory user contributions are affected by the same quality criteria that were discussed above for direct and indirect contributions. Removing inappropriate content manually is possible, especially in smaller projects, while larger initiatives make use of collaborative engagement or automatic systems like spam filters. In WORDNIK, for example, the option exists for every comment to send a feedback email to the editors. This is activated by clicking on the symbol of a downturned thumb as the end of each comment; the editors can then remove anything unsuitable if necessary. In open-collaborative resources like WIKTIONARY, however, the removal or correction of misplaced or false content rests in the hands of the contributors alone.

As long as a discussion about relevant lexicographic issues actually takes place (in contrast, for example, to vague comments or comments largely not relevant to the topic in environments like WORDNIK), comments and discussion pages like those in WIKTIONARY can also constitute a quality measure for the development of the relevant dictionaries. However, this applies not only to purely collaboratively compiled dictionaries like WIKTIONARY but also to the field of professional and, in part, commercial lexicography, which can gather qualitative and quantitative feedback and information in this way to develop their own dictionary outputs.

Ensuring quality is a major incentive for publishers to provide opportunities for communication or exchange associated with dictionaries, together with the opportunity to advertise their products and bind users to their brand, for which a wider variety of online channels are used. Educational initiatives around dictionaries can also play a role in this context, as we saw through the example of the ANW.

On the part of the users, the motivations for being involved can be as varied as the ways of making the contributions themselves. One reason undoubtedly lies in the desire



Diskussion:Kreuzung

Treffpunkt [Bearbeiten]

Entschuldigt meine heftige Wortwahl, aber ich weiß nicht, was hier für "Fachleute" revertieren und sichten!!

- ein Ort, wo sich 2 Straßen treffen, nennt man "Straßenknick" oder "Straßenecke", im einfachsten Fall einfach nur Straße, wenn eine gerade verlaufende Straße von der a-Straße zur b-Straße wird.
- ein Ort, wo sich 3 Straßen treffen, wird Straßengabel oder auch Abzweigung genannt
- ein Ort, wo sich 4 oder mehr Straßen treffen, wird Kreuzung genannt. Das ist sinngleich mit der Aussage: ein Ort, wo sich 2 (oder mehr) Straßen kreuzen *oder* ein Ort, wo eine Straße eine zweite Straße quert.

Ich weiß wirklich nicht, was ihr hier den Lesern für Bedeutungen "unterjubeln" wollt, aber das ist schon grenzwertiges Verhalten. Da ich bei einer massiven Revertierung ohne jede Rückfrage erheblichen Sachverstand erwarte bzw. unterstelle, kann ich hier nur VM annehmen.

Ich möchte jetzt hier keinen Editwar anzetteln, weshalb ich die stattgefundene Revertierung meiner Änderungen nicht erneut ändere und bitte, jemand mit Sachkunde und Verständnis möge sich der Angelegenheit annehmen. ——JÄhh (Diskussion) 12:03, 7. Mär. 2011 (MEZ)

> Gerne. Eine Straße, die in eine Kreuzung führt, endet nicht dort, sondern läuft, salopp gesagt, einfach durch. So kommt man auf zwei oder mehr sich treffende Straßen. Definitorisch ebenso erfasst im DWDS, in *Duden Das große Wörterbuch der deutschen Sprache in 10 Bänden. 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim, Leipzig, Wien, Zürich: Dudenverlag 1999*, unter englisch "crossroads" im *Oxford Dictionary of English* und bei Merriam-Webster. Wo steht es anders? Davon abgesehen ist dein Ton wieder einmal völlig unangemessen. —Pill (Kontakt) 12:50, 7. Mär. 2011 (MEZ)
> Oder sehe ich jetzt das Problem nicht? Du siehst einen Unterschied zwischen "treffen" und "kreuzen"? —Pill (Kontakt) 14:44, 7. Mär. 2011 (MEZ)

>> Pill, ich kann es nicht fassen: Du definiert in Deinem Einleitungssatz eine Kreuzung damit, das Straßen, die in eine Kreuzung führen, dort nicht enden? Ist das jetzt hier pillepalle?
>> Wenn sich also 2 Straßen im Winkel von 45Grad treffen, dann ist das für die ganze Welt eine Spitzkehre oder auch eine sehr scharfe Kurve, aber für Euch oder Dich ist das eine Kreuzung - ja?
>> Wenn sich 3 Straßen jeweils im Winkel von 45 Grad treffen, dann ist das für Euch hier eine Kreuzung?
>> Wenn auf eine Straße (eine durchgehende Hauptstraße, zur besseren Beurteilung] eine (zur Verdeutlichung: kleine) Seitenstraße im Winkel von 90 Grad trifft, dann ist das bei Euch eine Kreuzung?
>>
>> Weiter führst Du als Beleg an (Definitorisch ebenso erfasst im [http://www.dwds.de/?qu=Kreuzung&view=1 DWDS), wenn sich dort 2 Straßen treffen, übesiehst oder ignorierst aber, das unter der Quelle zu finden ist: "sich zwei Straßen kreuzen". Jetzt wirst Du mir sicherlich belegen, daß zwischen "sich treffen" und "sich kreuzen" überhaupt kein Unterschied besteht - dann bin ich zufrieden und gebe Ruhe - würde aber kollidieren mit dem Eintrag kreuzen.
>>
>> Du argumentierst: "Eine Straße, die in eine Kreuzung führt, endet nicht dort, sondern läuft, salopp gesagt, einfach durch." Beleg - natürlich Fehlanzeige! Aber ich bin gerne bereit, das mal durchzukauen: Wenn nach Deiner Lesart eine Straße an einer Kreuzung nicht aufhört, sondern durchläuft, wie nennst Du denn dann bitte einen Verkehrs(koten)punkt, von dem aus sich jeweils vom Mittelpunkt entfernend 5 Straßen wegführen? Zur Besseren Kenntlichmachung nennen wir die Straßen mal A, B, C, D und E, und zwischen ihnen jeweils ein Winkel von 72 Grad.

Fig. 8.5: Extract from the discussion contribution on the headword *Kreuzung* in WIKTIONARY.

to fill gaps in information quickly, for which forms of bidirectional communication appear to be particularly well suited (for more on possible motivations → Section 8.2).

8.5 Discussion

A classification of different forms of user participation like the one presented above serves, first, as an instrument to describe existing dictionaries and as a basis for further research into user participation in lexicographic contexts. Second, it is helpful when planning new resources and platforms or when revising existing ones.

A thorough discussion of user participation has also shown that having recourse to the potential power of collective intelligence is in no way a particularly new phenomenon in the field of lexicography nor one that has scarcely been used before. Above all, explicit feedback has been encouraged by dictionary compilers from their early days, for example by Duden's editors or by Oxford University Press, in the form of postal submissions. However, what is new is social interaction via social media and associated technologies, which have paved the way for user participation to become a mass phenomenon in its current scale and format. In particular, the forms of direct user participation were not – or only barely – possible before the emergence of the Internet.

All forms of user participation exhibit specific strengths and weaknesses, which have to be recognised and balanced for a dictionary to be planned effectively. The potential of collaboratively compiled dictionaries lies in the fact that, in theory, there are an unlimited number of participants – instead of single individuals or teams of a clearly defined size – with varying expertise who can devote themselves to these dictionaries for an unlimited time and in very particular ways. Not only the compilation of these dictionaries can be essentially unrestricted and free of charge but also access to them.

First and foremost, added value arises for dictionary content through direct user participation and explicit feedback. In open-collaborative dictionaries, users and providers are one and the same to some extent, and all content is compiled and revised in a participatory manner. Particularly in the case of contributions to collaborative-institutional dictionaries and semi-collaborative dictionaries or in the form of explicit feedback, this added value can extend from closing individual gaps in lemmas via supplementing important examples of usage to whole dictionary entries and the supply of larger bodies of material. This not only means that the coverage of a dictionary can be extended and content gaps closed but also that the lexicographic work can be undertaken more quickly and at lower cost. Lexicographers and language experts can save time and money when research tasks or the draft formulation of whole entries can be given to users. Dictionary providers and users benefit equally if content is available more quickly and in a more up-to-date form.

Furthermore, the strengths of collaborative lexicography lie in the diversity of the user group, which facilitates a wide-ranging description of different speech varieties and language pairs. This includes numerous dialect and regional expressions and phrases (e.g. *bostitchen*: Swiss *tackern* ‘to staple’[22]), slang terms and Internet jargon (e.g. *Karen*: a pejorative term used to refer to a middle-aged and middle-class white woman who puts herself first, is rude, insensitive, pushy, and whiny;[23] *ROFL*: rolling on the floor laughing[24]) and technical language/jargon (e.g. *ageotype*: a category of ageing biomarkers;[25] *shewee*: a portable female urinary device[26]). Among the languages and translation equivalents, we can find languages with only a few speakers and endangered languages (*siissisoq*: nose horn in Greenlandic) as well as language combinations that are scarcely of any commercial interest (e.g. Greenlandic–Italian; cf.; Matuschek et al. 2018; Meyer 2013; Meyer/Gurevych 2012; Rundell 2012).

However, it is not only newly contributed descriptions that bring added value but also the reporting or correcting of errors, which can raise the quality of a dictionary considerably. On the one hand, this kind of collaborative checking of quality can be used to perfect information that has been professionally compiled; on the other, it can fulfil its own purpose in selecting the best user entries. Here, it is the large number of users that is, first and foremost, an advantage since inappropriate user contributions (e.g. inappropriate comments and discussion contributions but also external user-generated content) can hardly be checked by single individuals or a few moderators. The example of Wordnik shows some of the possible ways of employing users to monitor comments. While many forms of user participation express the opinion or understanding of an individual user, there are multiple efforts to consolidate the different perspectives of a larger group of speakers, for example by collaboratively formulating a dictionary article in Wiktionary or by jointly evaluating the usefulness of a particular translation equivalent in dict.cc (→ Fig. 8.3).

However, dictionary users also benefit directly from the different forms of user participation. The use of open licences in collaborative dictionaries makes lexicographic content accessible to a large body of users. Furthermore, accessory forms of user participation increase the popularity of dictionaries while direct and indirect forms of participation provide the opportunity to actively shape the dictionary as a resource and to have a stake in the final product. In addition, binding users closely into the lexicographic process can serve an educational purpose and help to develop

[22] https://de.wiktionary.org/wiki/bostitchen [last access: February 22, 2024]. The verb derives from *Bostitch*, the name of a company producing staplers.

[23] https://www.urbandictionary.com/define.php?term=Karen&page=2 [last access: February 22, 2024].

[24] https://en.wiktionary.org/wiki/ROFL [last access: February 22, 2024].

[25] https://en.wiktionary.org/wiki/ageotype [last access: February 22, 2024].

[26] https://www.urbandictionary.com/define.php?term=shewee [last access: February 22, 2024]. The noun derives from *Shewee*, the company producing the devices.

important competences in using dictionaries. Complementary services and products achieve this in particular, for example, in playful ways or through engaging blog posts that prompt users to consult the dictionary. Direct user participation also has a contribution to make in this respect since checks have to be undertaken to see whether information is already contained in the dictionary and to see how language descriptions can be most effectively formulated. In turn, the exchange of views among dictionary users and language advice services represent added value for users if language questions are discussed that are not answered in the dictionary or at least not for a specific, given context.

Overall, user contributions lead to a negotiation of content according to the principles of supply and demand, from which both users and providers can, in theory, benefit. Implicit feedback reveals what is actually looked up by users. Explicit feedback and user comments provide information about the wishes and expectations of users in relation to the dictionary. On the one hand, direct user contributions reflect the usage of language on the part of the users (i.e., the user's "supply" of content). On the other hand, newly created content may be oriented towards demand in cases where users come across a gap while consulting the dictionary and then research and add the relevant material.

In contrast to this, the forms of participation also entail numerous challenges and weaknesses, which is an argument against planning a particular participatory resource or which demand further lexicographic, technological, or educational solutions. For example, the potential to save time and money described above is in no way clear cut. For collaborative-institutional and semi-collaborative forms of participation as well as with explicit feedback, the editorial checking of user contributions leads, in the first instance, to an increase in work for dictionary providers. The extent to which the usefulness and quality of contributions exceeds the time invested in checking them undoubtedly varies between individual dictionary projects and the different modes of participation. While explicit feedback from providers has been predominantly judged as positive (cf. Rautmann 2014; Thier 2014), it has also been shown that implicit feedback from log data only has a limited significance, no matter how much effort and expense is devoted to organising the analysis of the results. Dealing with plagiarism is also a particular problem. It is a challenging task to identify what appear to be high-quality user contributions as direct, unacknowledged use of data from other secondary sources, a task that can bring with it greater effort than would be involved in compiling a whole new lexicographic description based on primary sources.

Questions concerning the quality of user contributions in comparison with resources maintained purely by editors require particular reflection. User-generated dictionaries contain information about extremely varied language varieties and specialist vocabularies or about rarely occurring lemmas, while, to draw on examples from WIKTIONARY, common German words like *Fehlalarm* ('false alarm') or *Einzugsgebiet* ('catchment area') are missing. Frequent interpretations of lemmas are sometimes also not captured, such as the interpretation of the German lemma *Favorit* as “preferred object”[27].

In addition, if poor quality, inappropriate, or false descriptions are posted in a dictionary, this carries with it the danger of users’ language questions no longer being able to be answered reliably and the reference work thereby becoming unusable. The survey described in → Chapter 9.3.1 on the importance of criteria for Internet dictionaries demonstrates that users assign the highest priority to the reliability of the information available in dictionaries. Thus, it remains to be determined whether, and in what ways, user contributions really provide added value, something which has only been the subject of rudimentary research thus far. The studies by Fuertes-Olivera (2009), Hanks (2012), Meyer/Gurevych (2012), Rundell (2012), and Lew (2014) exhibit qualitative shortcomings in collaboratively compiled sources that can be traced back to mistaken, non-specific, old-fashioned, and partly obsolete descriptions. Whether user contributions bring anything new at all in qualitative or quantitative terms is of central importance for evaluating their potential. Meyer/Gurevych (2014) demonstrated that edited sources (e.g. DWDS or Duden online) were often listed in descriptions in the German Wiktionary that had been contributed collaboratively. This, and also a look at the traditional microstructure of dictionary entries, point to a comparatively conservative lexicographic approach in this collaborative space while collaborative innovation tends instead to be found in wide-ranging collections of material and new ways of integrating existing material. Overall, user-generated dictionaries seem to have considerable gaps and shortcomings in quality, which limits their usefulness. As a result, the expertise of professional lexicographers is indispensable if Internet dictionaries of high quality are to be created.

8.6 Summary and outlook

In this chapter, we addressed a relevant topic area in Internet lexicography, namely user participation, which should not be underestimated in its relevance. Indeed, it constitutes an important basis for enriching the quality and quantity of dictionary resources and in some cases is even the sole source for their entire construction. Using specific examples, we discussed three basic types of user participation in a systematic overview.

Forms of direct user participation encompass communal efforts in the construction and development of open-collaborative, collaborative-institutional, and semi-collaborative dictionaries, albeit with different degrees of editorial control and input

[27] For example, Wiktionary captures only the traditional meanings of the lemma in German, i.e. a) living being that is favoured by someone, b) participants in a competition with the best chance of winning. https://de.wiktionary.org/wiki/Favorit [last access: February 22, 2024].

options but with the common characteristic of directly integrating user contributions into the relevant dictionary. To a large degree, this type of user participation has only become possible with the advent of social media technologies.

We have to distinguish forms of indirect participation from direct participation. These are based on the principle of feedback or a mediated influence on the content and form of dictionaries. These include, on the one hand, form-based and free-text feedback, which dictionary users make available knowingly and of their own volition, and, on the other hand, implicit feedback gathered through lexicographically motivated analyses of log data or the integration of external user-generated content not intended a priori for lexicographic purposes.

Finally, the concept of accessory user participation covers different forms of exchange between dictionary compilers and users beyond the actual dictionary contents on the macro- and microstructural levels, which can proceed in a unidirectional or bidirectional manner.

It has also become clear, among other things, that different forms of user participation do not rule out one another but rather can be applied in parallel or in combination within the same dictionary or dictionary portal. This can be seen through the example of the Leo portal, which facilitates all three types of user participation presented in this chapter: it makes strong use of translation contributions and donated word lists, which are generally included directly in the relevant dictionaries, after they have been checked by editors, as is customary for semi-collaborative contexts. In addition, there are opportunities for users to contribute indirectly to the dictionary via corrections or suggestions for headwords. Feedback on provisional versions, for example, with different layouts can also be gathered. Finally, Leo provides forums through which users can exchange views with one another and the “Ask Dr. Bopp” language blog, which constitute forms of accessory user participation.

In a final discussion we illustrated the relevance of the present classification of types of user participation and explored their strengths and weaknesses. As the online publication of dictionary content increases, the issue of how to structure resources for user participation is increasingly gaining in significance for dictionary providers. This chapter should serve as an orientation for asking these questions and, at the same time, serve as a basis for further discussion on user contributions. In particular, the quality of contributions from the various characteristic forms of user participation has not been extensively researched to date.

Bibliography

Further reading

Abel, Andrea/Klosa, Annette (2014) (ed.): *Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache. *The volume addresses user participation in Internet dictionaries in a targeted way and discusses different forms of participation for individual dictionary services or products*.

Carr, Michael (1997): Internet Dictionaries and Lexicography. In: *International Journal of Lexicography* 10:3, 209–230. *The article is relevant to the field of electronic lexicography from a historical perspective, especially with regard to the term "bottom-up lexicography"*.

Melchior, Luca (2012): Halbkollaborativität und Online-Lexikographie. Ansätze und Überlegungen zu Wörterbuchredaktion und Wörterbuchforschung am Beispiel LEO Deutsch–Italienisch. In: *Lexicographica* 28, 337–372. *This article describes different types of users and the ways they are involved in the LEO dictionary portal*.

Meyer, Christian M. (2013): Wiktionary: The Metalexicographic and Natural Language Processing Perspective. Dissertation, Darmstadt: Technische Universität Darmstadt. *Chapters 2–4 explore in detail user contributions, the organisation, and the content of the collaboratively compiled WIKTIONARY*.

Rundell, Michael (2016): Dictionaries and crowdsourcing, wikis and user-generated content. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer, 1–16. *This article critically discusses various forms of crowdsourcing in the context of dictionary projects.*

Literature

Academic literature

Abel, Andrea, et al. (2003): Evaluation of the Web-based Learners Dictionary ELDIT. In: Lassner, David/McNaught, Carmel (eds.): *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, *Chesapeake, VA, USA*, 1210–1217.

Abel, Andrea/Meyer, Christian M. (2013): The dynamics outside the paper: user contributions to online dictionaries. In: Kosem, Iztok, et al. (eds.): *Proceedings of the 3rd eLex conference 'Electronic lexicography in the 21st century: thinking outside the paper'*. Ljubljana/Tallinn, 179–194.

Abel, Andrea/Meyer, Christian M. (2016): Nutzerbeteiligung. In: Klosa, Annette/Müller-Spitzer, Carolin (eds.): *Internetlexikografie. Ein Kompendium*. Berlin/Boston: De Gruyter, 249–290.

Benjamin, Martin (2014): Collaboration in the Production of a Massively Multilingual Lexicon, In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland, 211–215.

Carr, Michael (1997): Internet Dictionaries and Lexicography. In: *International Journal of Lexicography* 10:3, 209–230.

de Schryver, Gilles-Maurice/Joffe, David (2004): On how electronic dictionaries are really used. In: *Proceedings of the 11th EURALEX International Congress*. Lorient, France, 187–196.

de Schryver, Gilles-Maurice/Prinsloo, Danie J. (2001): Fuzzy SF: Towards the ultimate customised dictionary. In: *Studies in Lexicography* 11:1, 97–111.

de Schryver, Gilles-Maurice/Wolfer, Sascha/Lew, Robert (2019): The relationship between dictionary look-up frequency and corpus frequency revisited: a log-file analysis of a decade of user interaction with a Swahili-English dictionary. In: *GEMA Online Journal of Language Studies* 19:4, 1–27.

Engelberg, Stefan/Müller-Spitzer, Carolin (2013): Dictionary Portals. 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: De Gruyter, 1023–1035.

Ferschke, Oliver/Daxenberger, Johannes/Gurevych, Iryna (2013): A Survey of NLP Methods and Resources for Analyzing the CollaborativeWriting Process in Wikipedia. In: Gurevych, Iryna/Kim, Jungi (eds.): *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Berlin/Heidelberg: Springer, 121–160.

Fort, Karën/Adda, Gilles/Cohen, K. Bretonnel (2011): Amazon Mechanical Turk: Gold Mine or Coal Mine? In: *Journal of Computational Linguistics* 37:2, 413–420.

Francopoulo, Gil (2009) (ed.): *LMF: Lexical Markup Framework*. London: Wiley-ISTE.

Fuertes-Olivera, Pedro A. (2009): The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In: Bergenholtz, Henning/Nielsen, Sandro/Tarp, Sven (eds.): *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, 99–134.

Hanks, Patrick (2012): Word Meaning and Word Use: Corpus evidence and electronic lexicography. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 57–82.

Hanks, Patrick/Franklin, Emma (2019): Do Online Resources Give Satisfactory Answers to Questions About Meaning and Phraseology? In: Corpas Pastor, Gloria/Mitkov, Ruslan (eds.): *Computational and Corpus-Based Phraseology*. Cham: Springer International Publishing, 159–172.

Hausmann, Franz Josef (1989): Dictionary Criminality. In: Hausmann, Franz J., et al. (eds.): *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, Berlin/New York: De Gruyter, 97–101.

Howe, Jeff (2008): *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York: Three Rivers Press.

Javanmardi, Sara et al. (2009): User contribution and trust in Wikipedia. In: *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, 1–6.

Kemmer, Katharina (2010): *Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien*. Mannheim 2010.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: *Lexicographica* 26, 79–96.

Køhler Simonsen, Henrik (2005): User Involvement in Corporate LSP Intranet Lexicography. In: Gottlieb, Henrik/Mogensen, Jens Erik/Zettersten, Arne (eds.): *Symposium on Lexicography XI: Proceedings of the Eleventh International Symposium on Lexicography.* Tübingen: Niemeyer, 489–510.

Krause, Jens/Krause, Stefan (2011): Kollektives Verhalten und Schwarmintelligenz. In: Otto, Klaus-Stephan/Speck, Thomas (eds.): *Darwin meets Business: Evolutionäre und bionische Lösungen für die Wirtschaft*. Wiesbaden: Springer, 127–134.

Kreutzer, Till (2011): *Open Content Lizenzen. Ein Leitfaden für die Praxis*, Bonn: UNESCO.

Lampe, Cliff et al. (2010): Motivations to Participate in Online Communities. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, 1927–1936.

Landau, Sidney I. (2001): *Dictionaries: The art and craft of lexicograph*y. Cambridge: Cambridge University Press.

Lew, Robert (2011): Online dictionaries of English. In: Fuertes-Olivera, Pedro Antonio/Bergenholtz, Henning (eds.): *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 230–250.

Lew, Robert (2014): User-generated content (UGC) in English online dictionaries. In: Abel, Andrea/Klosa, Annette (eds.): *Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim, 7–25.

Malone, Thomas W./Laubacher, Robert/Dellarocas, Chrysanthos (2010): Harnessing Crowds: Mapping the Genome of Collective Intelligence. In: *Social Science Research Network Electronic Paper Collection*. Rochester, NY. http://ssrn.com/abstract=1381502 [last access: May 2, 2024].

Mann, Michael (2010): Internet-Wörterbücher am Ende der "Nullerjahre": Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien. In: *Lexicographica* 26, 19–46.

Matuschek, Michael/Meyer, Christian M./Gurevych, Iryna (2018): Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications. In: Rehm, Georg, et al. (eds.): *Language technologies for a multilingual Europe. TC 3 III*. Berlin: Language Science Press, 139–180.

McKean, Erin (2017): Wordnik. In: Fuertes-Olivera, Pedro A. (ed.): *The Routledge Handbook of Lexicography*. London: Routledge, 473–484.

Melchior, Luca (2012): Halbkollaborativität und Online-Lexikographie. Ansätze und Überlegungen zu Wörterbuchredaktion und Wörterbuchforschung am Beispiel LEO Deutsch–Italienisch. In: *Lexicographica* 28, 337–372.

Melchior, Luca (2014): Ansätze einer halbkollaborativen Lexikographie. In: Abel, Andrea/Klosa, Annette (eds.): *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim, 26–47.

Meyer, Christian M. (2013): Wiktionary: The Metalexicographic and the Natural Language Processing Perspective. Dissertation, Darmstadt: Technische Universität Darmstadt. http://tuprints.ulb.tu-darmstadt.de/3654/ [last access: May 2, 2024].

Meyer, Christian M./Andrea, Abel (2017): User Participation in the Internet Era. In: Fuertes-Olivera, Pedro A (ed.): *The Routledge Handbook of Lexicography*. London/New York: Routledge, 735–753.

Meyer, Christian M./Gurevych, Iryna (2012): Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 259–291.

Meyer, Christian M./Gurevych, Iryna (2014): Methoden bei kollaborativen Wörterbüchern. In: *Lexicographica* 30, 187–212.

Müller-Spitzer, Carolin/Möhrs, Christine (2008): First ideas of user-adapted views of lexicographic data exemplified on OWID and ELEXIKO. In: Zock, Michael/Huang, Chu-Ren (eds.): *Proceedings of the COLING Workshop on 'Cognitive Aspects on the Lexicon'*. Manchester, 39–46.

Müller-Spitzer, Carolin/Wolfer, Sascha/Koplenig, Alexander (2015): Observing Online Dictionary Users: Studies Using Wiktionary Log Files. In: *International Journal of Lexicography* 28:1, 1–26.

Naber, Daniel (2005): OpenThesaurus: ein offenes deutsches Wortnetz. In: Fisseni, Bernhard, et al. (eds.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung*. Frankfurt, 422–433.

Rafaeli, Sheizaf/Ariel, Yaron (2008): Online Motivational Factors: Incentives for Participation and Contribution in Wikipedia. In: Barak, Azy (ed.): *Psychological Aspects of Cyberspace: Theory, Research, Applications*. Cambridge: Cambridge University Press, 234–267.

Rautmann, Karin (2014): Duden online und seine Nutzer. In: Abel, Andrea/Klosa, Annette (eds.): *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*. Mannheim, 48–61.

Rosenberg, Louis B. (2015): Human Swarms: a real-time paradigm for collective intelligence. In: Andrews, Paul, et al. (eds.): *Proceedings of the European Conference on Artificial Life 2015 (ECAL)*. Cambridge: MIT Press, 658–659.

Rundell, Michael (2012): ‘It works in practice but will it work in theory?’ The uneasy relationship between lexicography and matters theoretical. In: Vatdedt Fjeld, Ruth/Torjusen, Julie Matilde (eds.): *Proceedings of the 15th EURALEX International Congress*. Oslo, 47–92.

Schoonheim, Tanneke, et al. (2012): Dictionary Use and Language Games: Getting to Know the Dictionary as Part of the Game. In: Vatdedt Fjeld, Ruth/Torjusen, Julie Matilde (eds.): *Proceedings of the 15th EURALEX International Congress*. Oslo, 974–979.

Storrer, Angelika (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, Herbert Ernst (ed.): *Wörterbücher in der Diskussion III*. Tübingen: Niemeyer, 107–135.

Storrer, Angelika (2010): Deutsche Internet-Wörterbücher: Ein Überblick. In: *Lexicographica* 26, 155–164.

Storrer, Angelika/Freese, Katrin (1996): Wörterbücher im Internet. In: *Deutsche Sprache* 24:2, 97–153.

Surowiecki, James (2005): *The Wisdom of Crowds*. New York: Anchor Books.

Thier, Katrin (2014): Das Oxford English Dictionary und seine Nutzer. In: Abel, Andrea/Klosa, Annette (eds.): *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess. 3. Arbeitsbericht des wissenschaftlichen Netzwerks “Internetlexikografie”*. Mannheim, 62–69.

Verlinde, Serge/Binon, Jean (2010): Monitoring Dictionary Use in the Electronic Age. In: Dykstra, Anne/Schoonheim, Tanneke (eds.): *Proceedings of the 14th EURALEX International Congress*. Ljouwert, 1144–1151.

Wang, William Yang/McKeown, Kathleen (2010): “Got You!”: Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling. In: Huang, Chu-Ren/Jurafsky, Dan (eds.): *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, 1146–1154.

Wiegand, Herbert Ernst, et al. (2010) (eds.): *Wörterbuch zur Lexikographie und Wörterbuchforschung/Dictionary of Lexicography and Dictionary Research*, vol. 1: *A–C*. Berlin/New York: De Gruyter.

Dictionaries

ANW = *Algemeen Nederlands Woordenboek*. Leiden: Instituut voor Nederlandse Lexicologie. http://anw.inl.nl [last acccess: May 2, 2024].

BAB.LA = *Bab.La. Online Wörterbuch für 24 Sprachen*. Hamburg: bab.la GmbH. http://bab.la [last access: May 2, 2024].

BLF = *Base lexicale du français*. Leuven: Katholieke Universiteit. Archived at https://web.archive.org/web/20131025210135/http://ilt.kuleuven.be/blf/ [last access: May 2, 2024].

CANOONET = Deutsche Wörterbücher und Grammatik. Basel: Canoo Engineering. Archived at https://web.archive.org/web/20181001082324/http://www.canoo.net:80/ [last access: May 2, 2024].

DDO = *Den Danske Ordbog (The Danish Dictionary)*. København: Det Danske Sprog- og Litteraturselskab. https://ordnet.dk/ddo [last access: May 2, 2024].

DICT.CC = *dict.cc*. Deutsch-Englisch Wörterbuch. Wien: dict.cc GmbH. http://www.dict.cc [last access: May 2, 2024].

DICTIONARY.COM = *Dictionary.com*. Oakland, CA: Dictionary.com. http://www.dictionary.com [last access: May 2, 2024].

DUDEN ONLINE = *Duden*. Berlin: Bibliographisches Institut/Dudenverlag. http://www.duden.de [last access: May 2, 2024].

DWDS = *Digitales Wörterbuch der deutschen Sprache*. Berlin: Berlin-Brandenburgischen Akademie der Wissenschaften. http://dwds.de [last access: May 2, 2024].

ELDIT = *Elektronisches Lernerwörterbuch Deutsch–Italienisch*. Bozen: Europäische Akademie. http://eldit.eurac.edu/ [last access: May 2, 2024].

GLOSBE = *Glosbe – das mehrsprachige Online-Wörterbuch*. Warschau: Cloud Inside. http://glosbe.com [last access: May 2, 2024].

GOOGLE DICTIONARY = *Google Dictionary*. Mountain View: Google. Archived at https://web.archive.org/web/20101127085248/http://www.google.com:80/dictionary [last access: May 2, 2024].

KAMUSI = *The Kamusi Project, Global Online Living Dictionary*. Genf: Kamusi Project International/Delaware: Kamusi Project USA. http://kamusi.org [last access: May 2, 2024].

LEO = *LEO*. München: LEO GmbH. http://dict.leo.org [last access: May 2, 2024].

MACMILLAN DICTIONARY ONLINE = *Macmillan Dictionary Online*. London: Macmillan Publishes Ltd. Archived at https://web.archive.org/web/20220111230227/https://www.macmillandictionary.com/ [last access: May 2, 2024].

MACMILLAN OPEN DICTIONARY = *Macmillan Open Dictionary*. London: Macmillan Publishers Ltd. Archived at https://web.archive.org/web/20211129035202/https://www.macmillandictionary.com/open-dictionary/latestEntries.html [last access: May 2, 2024].

MERRIAM-WEBSTER ONLINE = *Merriam-Webster Online*. Springfield, MA: Merriam-Webster. http://www.merriam-webster.com [last access: May 2, 2024].

MERRIAM-WEBSTER OPEN DICTIONARY = *The Open Dictionary*. Springfield, MA: Merriam-Webster. Archived at https://web.archive.org/web/20180624155024/http://nws.merriam-webster.com/opendictionary/ [last access: May 2, 2024].

NITTY GRITS = *Nitty Grits: International Culinary Dictionary* (Suzy Oakes, ed.). New Orleans: Southern Food and Beverage Museum, 2011. http://www.nittygrits.org [last access: May 2, 2024].

OED = *Oxford English Dictionary* (John A. Simpson/Edmund S.C. Weiner, ed.). 2nd edition. Oxford: Oxford University Press 1989.

OED ONLINE = *Oxford English Dictionary Online*. Oxford: Oxford University Press. http://www.oed.com [last access: May 2, 2024].

OMEGAWIKI = *OmegaWiki, a dictionary in all languages*. San Francisco: Wikimedia Foundation. Archived at https://web.archive.org/web/20230709193706/http://www.omegawiki.org/ [last access: May 2, 2024].

OPENTHESAURUS = OpenThesaurus, Synonyme und Assoziationen. Potsdam: Daniel Naber. Online: http://www.openthesaurus.de.

RAP DICTIONARY = The Rap Dictionary. Nijmegen: Patrick Atoon. http://www.rapdict.org [last access: May 2, 2024].

SPRACHNUDEL = *Sprachnudel.de*, Wörterbuch der Jetztsprache. Berlin: WEB'arbyte. http://www.sprachnudel.de [last access: May 2, 2024].

SZENESPRACHENWIKI = *Duden Szenesprachenwiki*. Mannheim: Bibliographisches Institut/Dudenverlag. http://szenesprachenwiki.de [now offline].

TECHDICTIONARY = *TechDictionary, the Online Computer Dictionary*. Chesterbrook: techdictionary.com. Offline [archived at https://web.archive.org/web/20190605111602/http://www.techdictionary.com/].

URBAN DICTIONARY = *Urban Dictionary*. San Francisco: Urban Dictionary. http://www.urbandictionary.com [last access: May 2, 2024].

WEBSTER = *Webster's Revised Unabridged Dictionary* (Noah Porter, ed.). Springfield: G & C. Merriam Co., 1913.

WIKTIONARY = *Wiktionary, das freie Wörterbuch*. San Francisco, CA: Wikimedia Foundation. http://www.wiktionary.org [last access: May 2, 2024].

WORDNIK = *Wordnik. All the words*. San Mateo, CA: Wordnik. http://www.wordnik.com [last access: May 2, 2024].

Internet sources

Amazon Mechanical Turk = https://www.mturk.com/mturk/welcome [last access: May 2, 2024].

Canoonet-Sprachblog = archived at https://web.archive.org/web/20160724003445/http://canoo.net/blog/ [last access: May 2, 2024].

Creative Commons = http://de.creativecommons.org/ [last access: May 2, 2024].

Crowdflower = http://www.crowdflower.com/.

Google Analytics = https://www.google.com/intl/de_de/analytics/ [last access: May 2, 2024].

Leo-Sprachblog = https://blog.leo.org/ [last access: May 2, 2024].

LMF = *Lexical Markup Framework*. http://www.lexicalmarkupframework.org/ [last access: May 2, 2024].

Macmillan-Dictionary-Blog = http://www.macmillandictionaryblog.com/ [last access: May 2, 2024].

Matomo = https://matomo.org/ [last access: May 2, 2024].

TEI = *Text Encoding Initiative*. www.tei-c.org [last access: May 2, 2024].

Wikipedia = *Wikipedia, The Free Encyclopedia*. San Francisco, CA: Wikimedia Foundation. www.wikipedia.org [last access: May 2, 2024].

Images

Fig. 8.1 (top left) “Ballot paper for the 2021 United Kingdom local elections (Coventry, Westwood ward)”, licensed under the “Creative Commons Attribution 4.0 International” licence (CC BY 4.0) by Wikimedia Commons user domdomegg. https://commons.wikimedia.org/w/index.php?oldid=815943269 [last access: September 9, 2023].

Fig. 8.1 (top right) “Eine Debatte im Plenarsaal des Bayerischen Landtages. Fotografiert im Rahmen des Landtagsprojektes Bayern 2012.”, licensed under the “Creative Commons Attribution-Share Alike 3.0 Unported” licence (CC BY-SA 3.0) by Wikipedia user Tobias Klenze. http://commons.wikimedia.org/w/index.php?oldid=74775027 [last access: July 18, 2012].

Fig. 8.1 (bottom left) “A demonstration in Erlangen, Germany, against tuition fees”, licensed under the “Creative Commons Attribution-Share Alike 2.0 Germany” licence (CC BY-SA 2.0) by Stefan Wagner (http://trumpkin.de). http://commons.wikimedia.org/w/index.php?oldid=112937438 [last access: December 31, 2013].

Fig. 8.1 (bottom right) “LA Times”, licensed under the “Creative Commons Attribution-Share Alike 2.0 Generic” licence (CC BY-SA 2.0) by Flickr user Daniel R. Blume. http://commons.wikimedia.org/w/index.php?oldid=117616015 [last access: February 28, 2014].

Carolin Müller-Spitzer and Sascha Wolfer

9 Research into Dictionary Use

Fig. 9.1: Scenes involving spoken production and reception.

People produce and receive language in multiple ways, whether through gestures, oral utterances in direct speech or on the phone, or in writing. Both when composing linguistic utterances and when trying to understand them, as well as when simply thinking about language, questions can arise, for example when the meaning of a word is unknown, when we do not know how to spell a word, when we wish to achieve language variation, or when language is being taught. These questions are particularly relevant when we are communicating in different languages or different terminological registers.

As a rule, dictionaries are compiled to facilitate communication between people speaking different languages or language varieties as well as to provide information on individual linguistic phenomena when there is a need to look things up. In this way, dictionaries count as functional objects; in other words, their actual purpose is to be used to deal with language tasks. *Research into dictionary use*, which is the topic of this chapter, is concerned with the practice of using lexicographic reference works and also, more generally, with the solving of linguistic problems with the help of reference works. The goal of research into dictionary use is to discover more accurately in which

Carolin Müller-Spitzer, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161, Mannheim, Germany, e-mail: mueller-spitzer@ids-mannheim.de
Sascha Wolfer, Leibniz-Institut für Deutsche Sprache, R5, 6–13, 68161, Mannheim, Germany, e-mail: wolfer@ids-mannheim.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/10.1515/9783111233758-010

situations, in what way, to what success, etc. lexicographic tools are used. This knowledge can then serve to adapt future dictionaries better to the needs of users.

This chapter is structured as follows. In the first part, we provide an introduction to the topic. User research concerns itself with actual user activity or, to put it more generally, with the experience and observations of dictionary use and is, as such, empirically oriented. As a result, user research has to look to methods from empirical social research, and the foundations of this are the subject of the second section. The third part is devoted to user research in relation to Internet dictionaries, the subject which stands at the heart of this volume.

9.1 Introduction

User research in relation to dictionaries is a very recent branch in the whole field of dictionary research. It is to the credit of many lexicographers and dictionary researchers that the importance of this branch of research has increased in recent years. It has certainly been emphasised for a long time in individual publications that users should be a central factor when planning lexicographic processes (→ Chapter 3); however, now it is no longer questioned – unlike 30 years ago – that dictionaries are functional objects. As such, users should be the central factor in the planning and production of dictionaries (Bogaards 2003: 26–33; Sharifi 2012: 626; Tarp 2008: 33–43; Wiegand 1998: 259–260; Wiegand et al. 2010: 680). As Lew (2011: 1) puts it, "[M]ost experts now agree that dictionaries should be compiled with the users' needs foremost in mind". Nonetheless, we can ask ourselves why this reference to users is emphasised in this particular way for lexicography when in reality every text is oriented towards its addressee. However, what is special about lexicographic texts compared to most other texts is that the *genuine purpose* of dictionaries is, for the most part, to be employed as a tool. In this respect, the focus on the practical user is stronger than with other sorts of texts. As we have already indicated, user research serves not only to discover more about the practice of dictionary use but also to improve dictionaries on the basis of the knowledge acquired from it and to shape them in a more user-friendly way.

In addition to dictionaries that are primarily conceived as functional tools, there has always been a form of lexicography oriented towards documentation as well. Users did not have the same importance for this branch of lexicography because these dictionaries were concerned, above all, with documenting the state of the language and its lexicon, maybe for posterity, or to "purify the language", or to "construct the language". For example, the GOETHE-WÖRTERBUCH was founded in the period after the Second World War to contribute to "rehumanising" society. We can read this in the dedication to the dictionary, which included the following:

> Der individuelle Sprachschatz eines Menschen ist stets zugleich Abbild und Ausdruck der Welt, wie diese sich gerade in diesem Kopf und Herzen spiegelt. Bei der besonderen Weltgemäßheit

> von Goethes Sehen, Denken, Sprechen muß dies Verhältnis jedoch eine ganz besondere Bedeutung gewinnen. Die Aufbereitung der Sprache Goethes in einem Wörterbuch wird nicht nur Goethes Sprache, sondern damit zugleich auch Goethes Welt erschließen. [The individual language and vocabulary of a person is always at once an illustration and expression of the world as it is reflected in his mind and body. However, in the particular measure of the world embodied through Goethe's sight, thought, and language this relationship had to acquire a particular meaning. Editing Goethe's language in a dictionary will not only make Goethe's language accessible but also his world.] (Schadewaldt 1949: 297)

Nevertheless, an overwhelming number of dictionaries are considered to be good if they serve as adequate tools for particular users in particular situations. This orientation towards particular groups or situations can also be partly extracted from the titles of these works. There are "learners' dictionaries", "primary school dictionaries", or more unusual titles as well such as "Döskopp, Saudepp, Zickzackpisser: Schimpfwörter aus deutschen Regionen" ("The Best Swearwords from the German Regions"), "Ohne-Wörter-Buch: 550 Zeigebilder für Weltenbummler" ("Word-less Dictionary: 550 Illustrative Pictures for Globetrotters"), and many more. In order to find out whether these dictionaries really correspond to the needs of their target users, we must examine empirically whether a language question can actually be resolved by using the dictionary and if so, how the dictionary is used, what users value or criticise about the dictionary, and which areas for improvement can be identified. However, there are also empirical studies in dictionary research that are detached from individual dictionaries, for example on individual dictionary types such as Internet vs print dictionaries or spelling dictionaries vs synonym dictionaries. The results of general questions such as these, then, do not serve, for the most part, to improve individual dictionaries but they do provide different dictionary projects with indications as to the direction in which their work might best proceed.

User research can, in theory, take place at completely different stages in the lexicographic process (→ Chapter 3): in the preparatory phase, to test different draft ideas for the dictionary in a pilot study of their user suitability; after the online release is ready, in order to check how the dictionary is used; or also to prepare new functionality, for example, to test the usability of different search functions. However, as we have already emphasised, dictionary user research can be undertaken without being connected to a specific lexicographic product. First of all, though, let us briefly consider the "tools of the trade" necessary to do empirical studies.

9.2 Methodological foundations

The following guide to methodological foundations (based on Koplenig 2014 and Diekmann 2011) provides an initial overview of the steps that have to be considered when undertaking an empirical study. The following sections provide insights into the following questions:

- How can a research problem be formulated and specified? (→ Section 9.2.1)
- How are the relevant variables measured? (→ Section 9.2.2)
- Which study design is appropriate to elicit the data? (→ Section 9.2.3)
- Which research design is best suited to answer the research question with regards to controlling variation? (→ Section 9.2.4)
- How should the data be gathered (→ Section 9.2.5)
- What needs to be taken into consideration for the data analysis? (→ Section 9.2.6)
- What has to be considered when reporting the study? (→ Section 9.2.7)

To illustrate these questions, we not only give examples from dictionary user research but also present some from empirical social research from completely different areas of life in order to illustrate the broad application area of this kind of research.[1]

9.2.1 Formulating and specifying the research problem

Every empirical project begins with a question. The more precisely this question is formulated, the easier the steps become to develop an empirical study. Karl Popper illustrated this as follows: we can only meaningfully follow the instruction “Observe” if we know *what* we are supposed to observe. For example, if we sat in a classroom and observed a year four class in a German lesson, we would not be able to identify any patterns through this observation alone without having previously formulated a problem; in other words, observations are not a reliable foundation for acquiring insight. Thus, Popper advocates the thesis: “no observation without a problem”. So if we first pose a precise question such as “Do girls raise their hands more frequently than boys?” or “Does the number of spoken answers relate to how far forward in the classroom a student sits?”, we can gather data on these questions and, as a consequence, also acquire new insights into these problems (Popper 1994: 19f.). All subsequent steps in an empirical enquiry depend on the nature of the research question, the research aim associated with it, and the corresponding hypotheses. For this reason, it is particularly important to formulate this research question clearly:

> Manche Studie krankt daran, daß *irgendetwas* in einem sozialen Bereich untersucht werden soll, ohne daß das Forschungsziel auch nur annähernd klar umrissen wird. Auch mangelt es häufig an der sorgfältigen, auf das Forschungsziel hin abgestimmten Planung und Auswahl des Forschungsdesign, der Variablenmessung, der Stichprobe und des Erhebungsverfahrens. Das Resultat unüberlegter und mangelhaft geplanter empirischer ‚Forschung‘ sind nicht selten ein kaum noch genießbarer Datensalat und aufs äußerste frustrierte Forscher oder Forscherinnen. [‘Many studies suffer because *something* in a social field is supposed to be being investigated without the research goal being outlined even remotely clearly. Often studies lack careful planning and selec-

[1] There are now good Wikipedia entries for most of the terms used below (such as usability test, log files, etc.).

> tion in line with the research aim, a research design, measurement of variables, sampling, and the survey process. Frequently, the result of empirical 'research' that has not been thought through and has been inadequately planned is a scarcely palatable mess of data and some extremely frustrated researchers.'] (Diekmann 2011: 187; cf. on lexicography, also Lew 2011: 8)

Formulating the research question also involves being clear about what data need to be collected in order to answer the question so that it can be measured, or *operationalised*, accordingly.

9.2.2 Operationalisation

Once the research question and, with it, the theoretical conception of the study have been specified, the researchers must decide how they wish to measure the variables involved. To take an example to illustrate this: a project team that has developed a new Internet dictionary would like to investigate how this dictionary is used. To this end, a so-called usability test is to be carried out in a laboratory. A usability test serves to assess the suitability for use of a piece of software or hardware with potential users; in the process, the test subjects are prompted to complete typical tasks with the test object, in this example, the Internet dictionary. We do this to investigate at which points problems arise in the use of the dictionary, for example that a user cannot find the appropriate search option, cannot orient themselves accurately or quickly enough in the dictionary, or cannot find their way back to a previously viewed entry. In the subsequent data analysis for the new dictionary, the test participants who have already used many types of language dictionaries (→ Chapter 2) should be distinguished from those who can be classified as inexperienced users. Thus, the planning of the study must consider how this experience or inexperience can be measured. For example, if the researchers were to ask a question before the usability test such as "Have you ever used a general dictionary?" and then proceed on the basis that the test participants enter the types of dictionary in a free-text field, they could be in for an unpleasant surprise. If the participants only enter "Langenscheidt" or "Duden", that is, the name of the publisher and not the dictionary type (as we experienced once in a pilot study), it is not possible to operationalise their experience with regard to different types of language dictionaries. Thus, it would be better to provide a fixed list of dictionary types here and, perhaps in addition, to create a free-text field for participants who wish to give more information.

9.2.3 Study design

The study design specifies the temporal mode by which the data are generated. Here we can distinguish between three types of study design:

- cross-sectional design;
- trend design;
- panel design.

A cross-sectional design denotes data being collected once, at a particular point in time or over a short period of time, with any number of participants. Thus, a cross-sectional study makes it possible to compare different entities at a particular point in time. It is not possible to measure changes over time in this way.

A typical example for a cross-sectional design is the so-called *Sonntagsfrage* or "Sunday question", in other words, the question that asks which party the respondent would vote for if there were a Federal election in Germany the following Sunday (→ Fig. 9.2). A single Sunday question makes it possible to compare the voting intentions of the individual study participants in that calendar week.

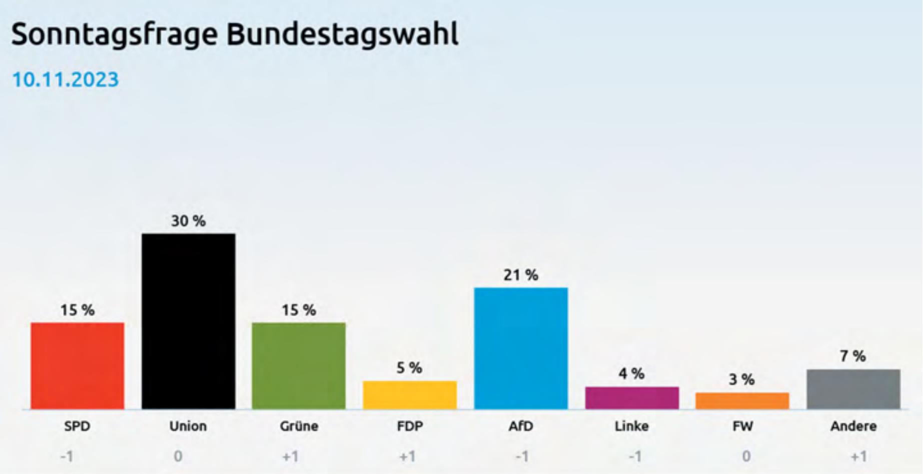


Fig. 9.2: Sample cross-sectional study: Sonntagsfrage – Deutschland 10.11.2023: infratest dimap for ARD-DeutschlandTREND.[2]

Trend or panel designs, in contrast, are longitudinal designs. We speak of a trend design when multiple horizontal studies on the same topic are carried out at multiple points in time and these are then summarised into a trend. More specifically, a trend design involves eliciting (a) values of the same variable (b) at multiple points in time with (c) different sampling, i.e. different participants. An example of a trend study can be seen in → Fig. 9.3: here the results of the horizontal studies of voting intentions elicited by the *Sonntagsfragen* are summarised into a trend from January 1991 to January 2013.

[2] https://www.infratest-dimap.de/umfragen-analysen/bundesweit/sonntagsfrage/. For more information about political parties in German, cf. https://en.wikipedia.org/wiki/List_of_political_parties_in_Germany.

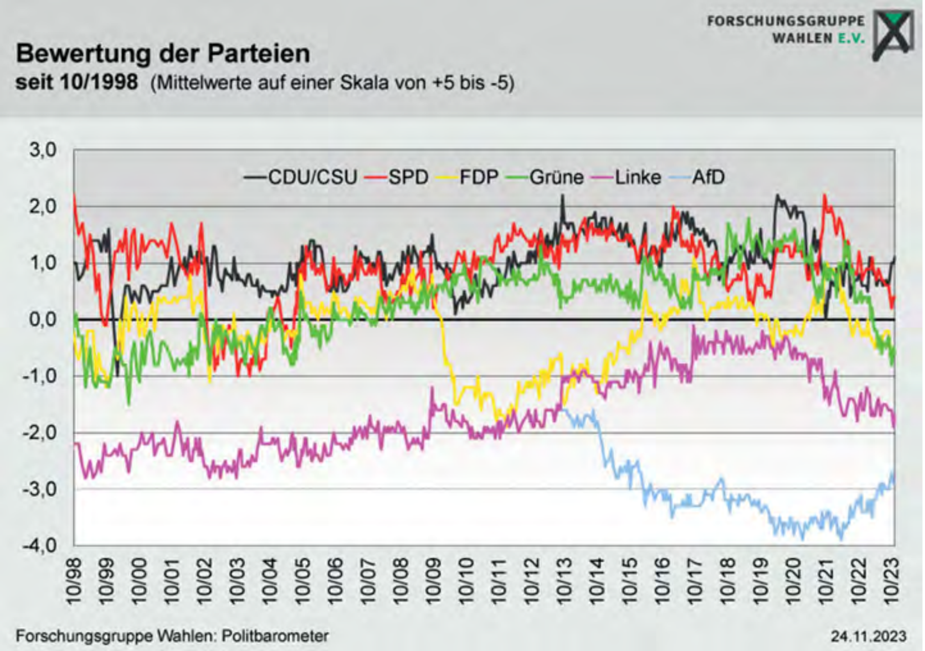


Fig. 9.3: Sample trend design; Forschungsgruppe Wahlen Politbarometer (24.11.23).[3]

In contrast to a trend design, a *panel design* involves eliciting (a) values of the same variables at (b) different points in time but with (c) the same sampling, i.e. the same participants. This small formal difference is very significant in practice because, unlike trend designs, panel studies make it possible to understand developments on an individual level. However, a panel study involves considerably more effort. A great deal of time needs to be invested in maintaining contact with the participants and ensuring that they are available for future *panel waves*.

One example of a large panel study in Germany is the National Educational Panel Study (NEPS[4]) on educational paths in Germany. The so-called marshmallow study, initiated by Walter Mischel at Stanford in the 1960s is another well-known example of a panel study.[5] Let us present this one in more detail. Mischel conducted the first part of the study between 1968 and 1974 with children aged about four years old attending

[3] https://www.forschungsgruppe.de/Umfragen/Politbarometer/Langzeitentwicklung_-_Themen_im_Ueberblick/Politik_II/.

[4] https://www.neps-data.de/ [last access: February 9, 2024].

[5] More information on the marshmallow study can be found on an archived version of Walter Mischel's home page (https://web.archive.org/web/20140424191957/https://www.columbia.edu/cu/psychology/indiv_pages/mischel/Walter_Mischel.html) and also in the associated publications (Mischel et al. 1972, Shoda et al. 1990); a follow-up study by Kidd is documented in Kidd et al. (2013). A popular sci-

the nursery school on the Stanford campus. The research question was whether the ability to delay gratification could predict a variety of subsequent developments and consequences in an individual's life, particularly in relation to social competence, the capacity to learn, and chronic weaknesses, such as a particular sensitivity to rejection.

The ability to delay gratification in early childhood was measured in a laboratory setting as follows: the children were shown a desired object in individual laboratory sessions, for example, a marshmallow (biscuits, pretzels, and plastic poker chips were used in other versions of the experiment). The experimenter told the particular child that they were going to leave the room and made it clear to the child that they could call them back by ringing a bell and then receive a marshmallow or the other object on offer. However, if the child would wait until the experimenter returned by themselves, they would immediately receive two objects; in other words, they would be rewarded for waiting. If the child did not ring the bell, the experimenter returned after 15 minutes.

In further panel waves from 1980 to 1981, Mischel and his team found that the longer the children had waited in the original experiment, the more competent were they described as being at school and in social settings (according to their parents' statements) and the better they were able to deal with frustration and stress while also tending to exhibit higher performance in school. Proceeding from these research results, the marshmallow task was perceived to be a significant tool, capable of measuring an important personal ability or characteristic that can predict long-term success in many areas of life. This kind of study can only be performed on an individual level through a panel design. According to information on Walter Mischel's home page, contact is still being maintained with this cohort (i.e. the participants who took part in the first round) and even the children of those participants are now involved in the study.

An interesting follow-up investigation to the now legendary marshmallow test was undertaken around 40 years later. The psychologist Celeste Kidd, who had worked in a home for homeless families for some time, developed the hypothesis that for children who came from socially less secure backgrounds, it was not a rational decision to wait for a second marshmallow in the marshmallow task but that it was more reasonable to immediately eat the one that was directly available. She was able to demonstrate in her experiment that the reliability or trustworthiness of the researcher in the first task could halve or double the waiting time in the marshmallow task (Kidd et al. 2013). The study suggested that the ability for delayed gratification is more strongly influenced by the social milieu than had been accepted up to then. Kidd's follow-up study outlined above was carried out in the form of an experiment, one of three different types of research design that will be presented in the following section.

One short concluding observation needs to be made on panel studies in user research: the marshmallow study should have made it clear why panel studies have not been under-

ence article on the study can be found at https://www.psychologytoday.com/us/blog/beyond-school-walls/202304/10-ways-life-is-a-marshmallow-test [last access: February 9, 2024].

taken in lexicography. This kind of study is very labour intensive and therefore expensive. In comparison, the use of dictionaries is a research area that is not nearly as fundamental to human life as, for example, educational trajectories. However, in theory, a panel study could be used in the field of user research to investigate, for example, how dictionary training, in a university context, say, affects the use of dictionaries in the long term.

9.2.4 Types of research design and controlling variance

The choice of a horizontal or longitudinal design specifies the temporal dimension of the data. Planning an empirical study involves another aspect that relates to the constitution of comparison groups and the way participants are divided between these groups. This aspect is known as variance control (Diekmann 2011: 329). Here, we can distinguish between three types of design:

- experimental design;
- quasi-experimental design;
- ex-post-facto design.

In an *experimental design*, at least two groups are formed according to a random process ("randomisation") whereby the researcher manipulates the independent variables. A typical example are drugs trials in which the independent variables (medicine or placebo) are decided by the researcher and the participants are allocated randomly to a group (the treatment group or the control group). In this case, the treatment group is the one with the drug and the control group consists of the participants who receive the placebo. Another example is Kidd's study described above in which the children were assigned to a reliable or unreliable condition. The terms independent or dependent variable relates to their position in the hypothesis. In general terms, the independent variables are the variables that are generated (experimental) or given (ex-post-facto); the dependent variables are then the variables calculated as depending on them, that is, the measured value that is of interest for the study. Using the Kidd study as an illustration, the variable of reliability or unreliability was generated by the researcher and was therefore the independent variable. Dependent on this, the researcher then investigated how long the children waited in the marshmallow task, i.e. the waiting time was the dependent variable.

The same preconditions apply to a *quasi-experimental design* as to an experimental design but with the difference that the conditions are not distributed at random. That is, the comparison groups are determined explicitly and for the most part in advance, while planning the study, but the participants are not allocated to the comparison groups at random. One example of this kind of design could be staff interviews about job satisfaction that are undertaken before and after a business is restructured. The given independent variables would then be the time before vs. after the restructuring and the dependent variable the degree of satisfaction. In the field of dictionary user research, the usefulness of new features could be evaluated in this way: for example, the number of searches in a dictionary could be recorded that were unsuccessful before and after the implementation of a search feature that tolerates errors. The difference in these values can then be interpreted as the usefulness of the feature.

An *ex-post-facto design* is a research design without random allocation to experimental conditions and without manipulation of the independent variables, i.e. groups of participants are differentiated on the basis of characteristics that existed before the study and that will continue to exist independently of the study. This design is very common among studies that seek to investigate the influence of socio-economic and socio-demographic factors on upbringing, education, or professional success. The studies on potential differences between groups of users (translators/linguists) in dictionary user research discussed in → Section 9.3.1 also fall into the category of ex-post-facto design since the test participants were translators or linguists before our research study and will continue to be so afterwards. It is different in drugs trials: belonging either to a test group or a control group is a variable that exists only in the context of the study and not before or after.

9.2.5 Data collection methods

Empirical social sciences distinguish between four methods of data collection:

- surveys (in person, by phone, written);
- observation;
- content analysis.

In addition to this categorisation, two groups are distinguished from one another: reactive and non-reactive methods. Non-reactive methods are those where an empirical study is conducted without the knowledge of the participant. As such, a survey is a reactive method since the interview situation can influence the answers because the participant naturally knows that they are being asked questions. Diekmann provides an example to illustrate the general distinction between reactive and non-reactive methods. If the nutritional habits of households are being investigated using a questionnaire, this is a reactive method. However, if the same outcome is studied by looking at household waste, this is a non-reactive data collection method (Diekmann 2011: 195–196). The strength of non-reactive methods is that they provide unbiased results and data about real behaviour. At the same time, the possibilities for using these methods are severely restricted since researchers only have control over the process in few cases. One example of a non-reactive method from dictionary user research is the analysis of log files (→ Chapter 3). Log files are records that contain information about some or all of the actions and processes in a computer system. For example, for Internet dictionaries, these log files can store which headwords have been looked up by users. This makes it possible to conduct interesting studies (→ Section 9.3.4) but it typically does not allow us to compare the behaviour of different user groups with one another since most log files have no additional information about them. It is not possible, for example, to determine the reasons why users cancel a search or whether their query was successfully answered. This means that we have no non-reactive procedures for generating data at our disposal for many research questions where the answer depends on background information about the participants (cf. Trochim 2006 and, in relation to dictionary user research, Wiegand 1998: 574).

Surveys are the method used most frequently in social science research. Knowledge about social structures, social classes, or educational opportunities are primarily the result of quantitative population surveys. Critics take issue, above all, with the reactivity of this method in relation to the problem of social desirability. This refers to the fact that participants (might) tend to answer questions in a way that is socially desirable. For example, we would find few people who would answer "yes" to the question "Do you discriminate against marginalised groups in everyday life?" Diekmann demonstrated one example of this phenomenon with his colleague Preisendörfer in the "Sansal Drugstore Study" (Diekmann 1994). The first part of the study consisted of telephone surveys with more than 1,000 participants on various aspects of environmental behaviour. The results revealed a very high sensitivity towards upcoming environmental problems. In a second part of the study, three months later, a sub-section of the participants were sent a professionally produced brochure for the fictional drugstore "Sansal" in which heavily discounted brand products were on offer for the following reason: "Wegen der zu erwartenden strengeren Umweltschutzgesetzgebung müssen die Lager mit FCKW-haltigen Artikeln geräumt werden" [Because we expect stricter environmental laws, our warehouses have to be cleared of products containing CFCs] (Diekmann 1994: 20). A subsequent catalogue order was interpreted in the study as an intention to buy. What was interesting was the comparison between the actual reactions and the answers in the preceding telephone interviews since those who placed catalogue orders were not predominantly people who were ambivalent about environmental issues. For example, according to the survey, the vast majority of those interested in making a purchase (75%) knew about the damaging consequences of using CFCs. As a result, this study demonstrates how certain social issues are difficult to investigate using survey methods.

However, the problem of social desirability is not equally relevant for all areas of life. For example, it is difficult to imagine that social desirability would play a role in answering a question about dictionary use in situations of text production and reception. Insofar as the use of questionnaires in dictionary user research is criticised (e.g. by Tarp 2008), it relates to observations about the potential shortcomings of questionnaires, rather than focussing on the weaknesses of this form of data collection in general. Developing a good questionnaire involves a great deal of background knowledge, or – as Trochim puts it – is "an art in itself" (Trochim 2006[6]).

In a general sense, all empirical methods are observational procedures in that observation identifies which point is circled on a rating scale. However, as a data

[6] https://conjointly.com/kb/constructing-survey/ [last access: March 23, 2024].

method in the social sciences, *observation* means more specifically the direct observation of human actions, spoken utterances, non-verbal reactions (e.g. body language), or also the observation of social characteristics (clothing, furnishing, status symbols). Ethnological field research is one example of a research area in which the observational method is widespread. Here, the boundary between social reportage and academic observational studies is fluid. The prerequisite for the latter is a clear reference to research hypotheses and a systematic approach to observation under strict supervision. The observational method is superior to survey techniques for gathering up-to-date data, since information from surveys is of limited validity in this respect. Regarding this, Diekmann gives the example of a survey and a subsequent observational study of traffic behaviour (Diekmann 2011: 572): while 72% of the respondents in a survey claimed to always give drivers a hand signal before crossing the road, in reality only 10% of the participants in an observational study actually did so.

Content analysis is concerned with the systematic collection and evaluation of texts, images, and films (Mayring 2011). The designation "content analysis" is, in a certain sense, too narrow since the formal aspects of texts (e.g. the length of sentences) may play a role in the method of content analysis as well. Data for this method are abundant, for example, letters, marriage announcements, school books from various time periods, party manifestos, and much more. As Diekmann puts it, because the potential volume of material is so extensive, "[ist,] wie generell in der empirischen Sozialforschung die disziplinierende Wirkung expliziter Fragestellungen und Hypothesen zu betonen" [as is generally the case in social research, the emphasis rests on the disciplining effect of explicit questions and hypotheses] (Diekmann 2011: 580).

The method of content analysis was already employed to analyse propaganda in World War II, for instance. A more recent example for an empirical project that uses content analysis, among other methods, is one led by Thomas Chadefaux that seeks to predict armed conflicts by developing a kind of risk barometer that could give early warning to diplomats about regions in the world where armed engagements are particularly likely. For this purpose, masses of newspaper articles (based on the "Google News Archive") are searched for keywords (like *Spannung* 'tension', *Krise* 'crisis', *Konflikt* 'conflict', and *Militärausgaben* 'military spending') that point towards conflicts. If they appear noticeably often in reports about a particular country, this is interpreted as a sign that that the risk of war is growing for that country. The method has also been evaluated historically, ascertaining the likelihood with which past wars could have been predicted with this form of content analysis. This example shows that whole new studies can be conceived using large-scale data that are now freely available and which make use of content analysis as a data method.[7]

[7] The risk barometer for predicting armed conflict is documented in Chadefaux (2014); it was also reported on Deutschlandradio (https://www.deutschlandfunk.de/krieg-mit-vorwarnung-100.html) [last access: July 12, 2024].

In (almost) every kind of data collection method, it is important to conduct a kind of “rehearsal” as well, also known as a pre-test, before the actual data are generated in order to uncover formulations that might possibly be misunderstood or unclear instructions, etc. so that the problems can be corrected before the start of the study. Pre-tests are typically conducted with a few test participants whose data are not analysed along with those of the main study. Pre-tests are extremely important to avoid the risk of collecting a lot of data with a problematically designed study. In the worst case, it is only after collection that one realises that the data are useless. Pre-tests help to prevent this.

9.2.6 Data analysis

Once data have been generated for an empirical research study, they have to be analysed. The more carefully the preceding steps of an empirical study have been conducted, the better the data analysis will work. In the best case, a rough idea of how the data will be analysed is already sketched out during the planning phase of the study. In the worst case, the researchers will realise during the data analysis that variables required to answer the research question have not been included in the data collection. As such, knowledge of data analysis is indispensable for conducting an empirical study. This knowledge is also important in order to understand other studies and be able to identify questionable findings or potentially mistaken sources. However, a few pages here are not enough to provide a solid introduction to statistical data analysis. Introductions to statistical data analysis in the linguistic context are provided by Baayen (2008) and Gries (2021); Diekmann (2011: 659) also mentions general introductions on statistical data analysis.

9.2.7 Reporting

As a rule, the final part of an empirical study is the reporting. In basic terms, the type of reporting in empirical studies does not differ from other research results. Nonetheless, a particular model has been established for presenting empirical studies that is used in most publications, the so-called IMRAD structure (an abbreviation for “introduction, method, results, and discussion”; Sollaci/Pereira 2004). According to this structure, the introductory section usually presents the research question alongside relevant literature; in the methods section, the structure of the study is explained, including the participants, the data collection procedure, how it was conducted, etc.; and the results section presents the descriptive results, which are then discussed in the discussion section and situated in the research context. This relatively fixed structure enables experienced readers to replicate and critique the research, since they know where to find particular types of information in the report.

9.3 User research in relation to Internet dictionaries

As mentioned at the beginning of this chapter, dictionary user research is a relatively young field of research. Bogaards was still able to claim in 2003 that "nevertheless, uses and users of dictionaries remain for the moment relatively unknown" (Bogaards 2003: 33). Here, the group of non-native speakers, so-called L2 users, is still the one that has been researched most thoroughly. By contrast, little is known about the use of monolingual dictionaries by L1 users and other more or less unspecified user groups, such as 'interested lay users'. There are more studies comparing print and electronic dictionaries (cf. Dziemanko 2012). Yet, even if some studies have been published in the last ten years in the field of dictionary use, the need for research remains as great as ever (cf., among others, Bowker 2012; Lew 2015; Kosem et al 2018; Welker 2010, 2013). In particular, there were few comprehensive studies dealing with the use of Internet dictionaries before Müller-Spitzer's work (2014) (cf. Töpel 2014 for an overview of studies on Internet dictionaries).

When we wrote the original article in 2014, according to many experts, Internet dictionaries were the dictionaries of the future. Already then, the Internet was the central platform for many publishers and academic dictionary projects. This situation immediately suggested that we should concentrate user research on Internet dictionaries. At the same time, this posed risks because the dictionary landscape was and is changing rapidly in this area, and empirical studies require a great deal of time for analysis. In this way, it is possible for studies to have already been overtaken by their object of enquiry by the time they were published (cf. Lew 2012: 343). For example, if we had investigated which devices were being used to access Internet dictionaries in 2011 and the study had taken 18 months to publish, the market could have changed considerably because of the spread of smartphones and tablets. All the same, these kinds of results can be interesting and relevant in the longer term as a sort of historical snapshot.

In what follows, we shall present five examples of research questions and the studies constructed from them (cf. also Müller-Spitzer et al. 2018). The examples have been chosen so that, in terms of both content and, above all, methodology, they illustrate a wide range, thereby allowing connections back to the methods section above. All examples come from studies conducted at the Leibniz Institute for the German Language (IDS) in Mannheim, partly with external partners. The first three studies are described in detail in the edited volume "Using Online Dictionaries" (Müller-Spitzer 2014), the last two in other publications referred to in the respective sections. In order to permit a more concise presentation, the IMRAD structure is not used here.[8] We have to admit that we actually have quite different questions for lexicography, which are not yet re-

[8] In addition, not all of the possibly unfamiliar terms in the following discussion, such as *box plot* or *median*, can be fully explained. For a basic understanding, it is sufficient to consult Wikipedia.

flected in this article, such as: Will traditional dictionaries still exist in the future? What linguistic questions can be answered by AI systems? But presumably there will also be user research for more or less classical dictionaries in the future and for them, the following chapters can serve as an introduction and illustration of possible studies.

9.3.1 What makes a good Internet dictionary?

Digital dictionaries can and now clearly do differ from printed ones. It is not only that collaborative lexicographic resources are now being compiled (→ Chapter 8) but also that direct connections between lexicographic data and their underlying corpora have been implemented (→ Chapter 7) along with new forms of design. The online medium also makes it possible to represent lexicographic data more flexibly than in a printed book (Atkins 1992; de Schryver 2003; Rundell 2012: 29). Print dictionaries always have a fixed form determined by the medium, in other words, the lexicographic data and their typographical appearance are connected with one another in an inseparable way. By contrast, the electronic medium makes it possible to separate the lexicographic data from its presentation. The same lexicographic data can be presented in different ways – assuming the corresponding data modelling and data structure (→ Chapter 4) – so that the user is only shown the data relevant to them in their usage situation. These are only some examples of many potential changes (for further discussion, cf. Engelberg 2014; Granger 2012; Rundell 2012).

Simultaneously, the talk is of an existential crisis in lexicography. It is safe to assume that more language-related information is being looked up than ever before since people have vastly more freely accessible language resources at their disposal than, say, 20 years ago and, as a result, even those who would have hardly ever used dictionaries are now “googling” language questions. At the same time, these information searches do not lead them primarily to lexicographic resources, at least not in the sense of the paid use of such resources. Many Internet dictionaries can certainly not complain about access figures being too low but this operating model is certainly not economically viable.

Here, it is questionable whether fewer dictionaries are really being used only because there are fewer buyers. Previously, schoolchildren, students, and language learners were often obliged to buy dictionaries as learning materials because there was no alternative. However, it is unclear how often and how intensively they were actually used. Still, the crisis is existential in nature because it is increasingly difficult to earn money with lexicographic content. This raises the question as to whether lexicography can maintain an important position in the future even if Internet dictionaries develop “light years away” (Atkins 1992: 521) from print dictionaries, as other researchers demand.

However, if digital dictionaries develop in a direction which clearly diverges from print dictionaries, established models are brought into question and priorities have to be determined afresh. To put it in more general terms, to develop a good service, it is first of all necessary to find out which features of a product or service are particularly important for customer satisfaction and which are of secondary importance. These features can be formulated initially in abstract terms; for example, it could be about a group of products where the packaging is more important than the contents. This still does not tell individual producers how, specifically, their packaging should look, but it can give an indication that particular value should be placed on the design of the packaging.

The criteria for a good Internet dictionary, which we had participants assess and evaluate in an online study in 2010 and which we then investigated in more detail in a second online study later that year, also need to be taken into account on this level.[9] It is equally relevant for Internet dictionary projects to assess which criteria are thought to be particularly important since not everything that we would wish to include in the possible design of an Internet dictionary can be realised in practice. As Atkins pointed out (1996: 9):

> the greatest obstacle to the production of the ideal bilingual dictionary is undoubtedly cost. While we are now, I believe, in a position to produce a truly multidimensional, multilingual dictionary, the problem of financing such an enterprise is as yet unsolved. (cf. also de Schryver 2003: 188)

Evaluating the basic characteristics of dictionaries in the way that we did in our study still does not give lexicographers any specific indications about how exactly to design their dictionary. However, the results can give an indication as to which areas they should concentrate on because they are judged to be important by users.

Methodologically, our study was a cross-sectional ex-post-facto design where survey data was collected using an online questionnaire. The first study ran from February to March 2010 and the second from August to September 2010. A total of 684 people took part in the first study and 390 in the second. Our research question was “What makes a good online dictionary?” We wanted our participants to answer this question using ten basic criteria, which we put up for discussion. Because the study was not to last longer than 25 minutes and each criterion was to be evaluated individually, ten criteria were the maximum possible number. Furthermore, the complex of questions relating to the features of good Internet dictionaries was only one of many in this study. The chosen criteria extended from “traditional” properties of dictionaries, such as the reliability of content or clarity, to specific features of Internet dictionaries like animations for browsing or linking with a corpus.

First, the study sought to test how the participants evaluated each individual criterion by itself. The hypothesis there was that each criterion would be judged as im-

[9] For a detailed presentation of this study, cf. Müller-Spitzer/Koplenig (2014).

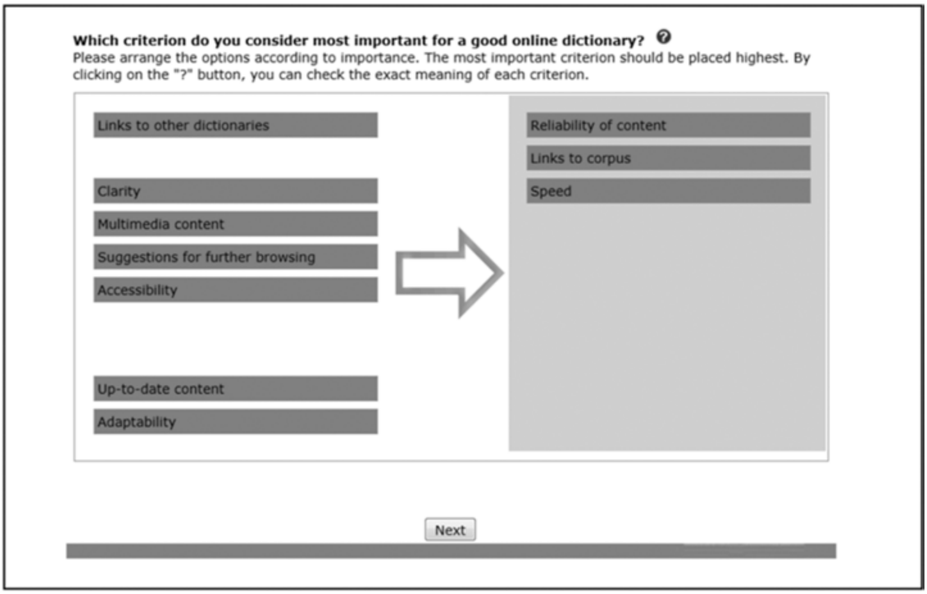
portant by itself, since all of the criteria together perhaps represented the ideal Internet dictionary. However, in order to find out how the participants judged the features in comparison to one another, an additional ranking exercise was undertaken in which the criteria had to distributed across positions 1–10.

An important issue in evaluating the features was to see whether they would reveal differences between groups. That is to say, we were interested in the influence of the personal (professional/technical) background of the participants on their individual evaluation of the criteria. So we also had to collect information about this personal background as a set of independent variables. The dependent variables were the preferences expressed by the participants for different characteristics of an Internet dictionary. That is, starting from the information about their personal backgrounds, we were able to analyse whether the preferences for the criteria changed depending on that background. These independent variables (like professional background, L1, etc.) were collected in one section of the demographic data in the questionnaire.

The first step was to evaluate each individual criterion on a five-point Likert scale. A Likert scale (named after Rensis Likert, a US social scientist) is a procedure to measure personal opinions by means of so-called items. Accordingly, a three-point Likert scale has three items that, one of which can be chosen to represent one of the following standpoints on a given statement: “agree”, “don’t know”, “disagree”. In this way, our participants were able to say how important they though each criterion was on a five-point scale that extended from “very important” to “not important at all”. They then had to rank the ten criteria (→ Fig. 9.4). The results can be seen in → Fig. 9.5. The position of the criteria in the ranking exercise is plotted on the left y-axis and the evaluations on the Likert scale on the right y-axis. As the lines shows, the two judgements correlate very clearly with one another; in other words, the criterion of content reliability was ranked in first place most frequently in the ranking exercise and received the highest average score on the Likert scale.

Contrary to our expectations, the participants evaluated the individual criteria very differently in the separate judgements on the Likert scale. The criterion that was judged to be the most important by some distance was the reliability of the content of an Internet dictionary. By contrast, media-specific criteria, like the integration of multimedia elements or possible user-adaptive customisation, were judged to be less important (a value of “2” corresponds to an evaluation as “not important”). Contrary to expectations, there were no statistically significant differences between the participant groups either. For example, we had expected that translators and linguists would find a connection to corpora particularly important. However, this was not supported by our data (→ Fig. 9.6; for more detail, cf. Müller-Spitzer/Koplenig 2014 and for a replication with a broader group of participants cf. Kosem et al. 2019).

Because the evaluation of the criteria in the first study turned out to be considerably more uniform than expected, we attempted to investigate the four most important characteristics (reliability of content, regular updates, clarity, and long-term accessibility) more precisely in a second online questionnaire. We also followed up on



Which criterion do you consider most important for a good online dictionary?

Please arrange the options according to importance. The most important criterion should be placed highest. By clicking on the "?" button, you can check the exact meaning of each criterion.

Fig. 9.4: Ranking of the criteria in the online questionnaire.

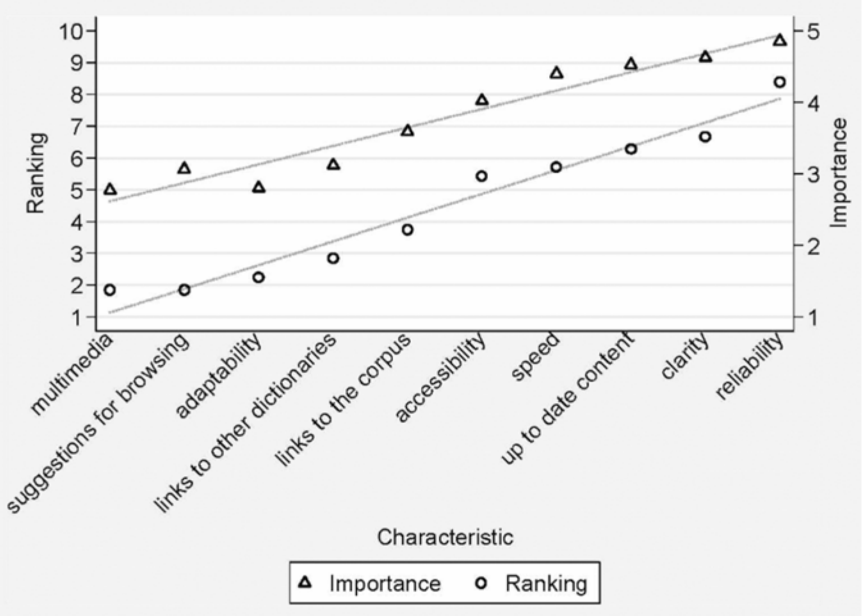


Fig. 9.5: Correlation between the mean rank and mean importance of criteria in the use of an Internet dictionary.

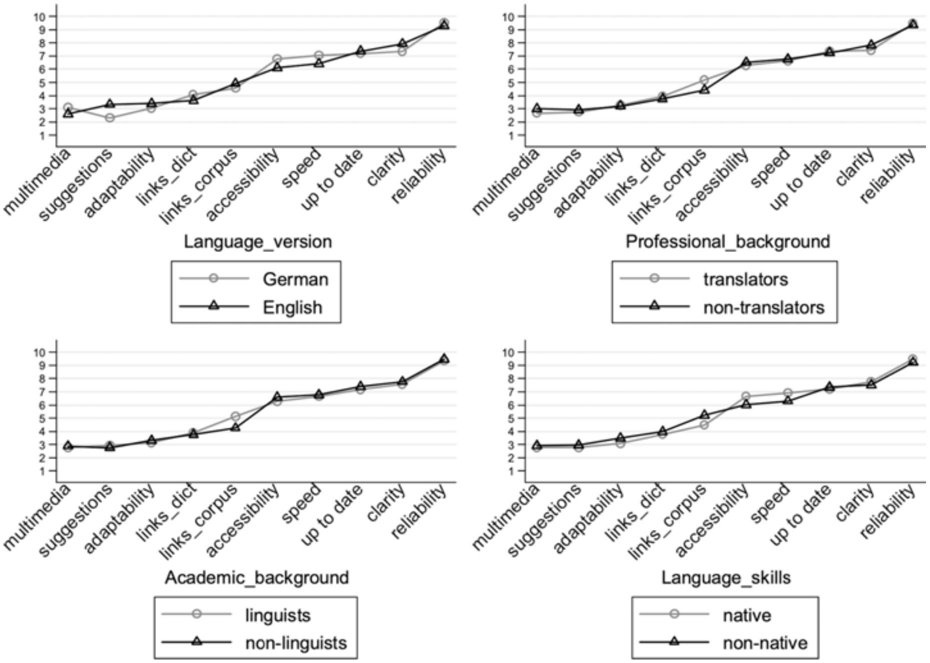


Fig. 9.6: Group-specific analyses of the rank orders.

the two features generally judged to be least important – multimedia and user adaptivity.

The results of two out of the four criteria judged to be most important will be elaborated here. We were interested, above all, in discovering what exactly participants understood by the very general terms like "reliability of content" or "updates" in more detail. After all, the first study may have shown, for example, that the reliability of lexicographic data was judged by some distance to be the most important feature of a good Internet dictionary, but we also know that collaboratively compiled dictionaries like Wiktionary and semi-collaboratively compiled dictionaries like LEO have a lot of users (→ Chapter 8). And it is precisely those dictionaries that were judged by specialists to be not particularly reliable in terms of their content (cf., e.g., Hanks 2012: 77–82). In the process we tried to list four characteristics for each criterion, so for the reliability of content:

- A well-known publisher or institution is behind the dictionary project.
- All of the information reflects different text types and usage across regions.
- All of the information reflects actual language use, i.e. the details have been checked in a corpus.
- All of the information has been checked by (lexicographic) experts.

Precisely in relation to collaboratively or automatically compiled (parts of) dictionaries, it would be interesting to find out how highly the participants would judge the criteria of a well-known creator and checking by experts (cf. Sharifi 2012: 637, who demonstrates that in the field of Persian dictionaries, the users surveyed by him saw “the author’s reputation as the most important factor when buying a dictionary”).

In part, we also tried to list individual criteria where we thought that they would perhaps demonstrate differences in groups between linguists and translators, on the one hand, and non-language specialists, on the other, such as the criteria for “updates”:

- Current developments in the language (e.g. changes to German spelling or new typical contexts) find their way quickly into the Internet dictionary.
- Words processed by editors appear online immediately.
- Current research finds its way into the lexicographic work.
- New words are described promptly in the Internet dictionary.

The hypothesis here was that the criterion of integrating current research into a dictionary would only be chosen by specialists. The results can be seen in → Fig. 9.7 and → Fig. 9.8.

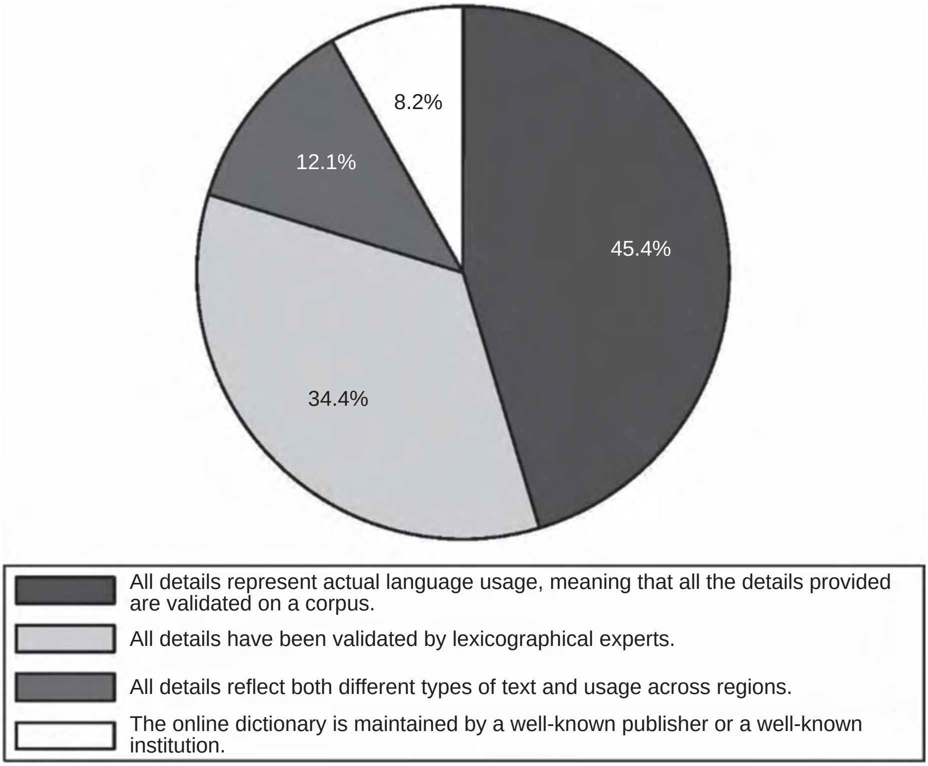


Fig. 9.7: Pie chart: aspects of content reliability.

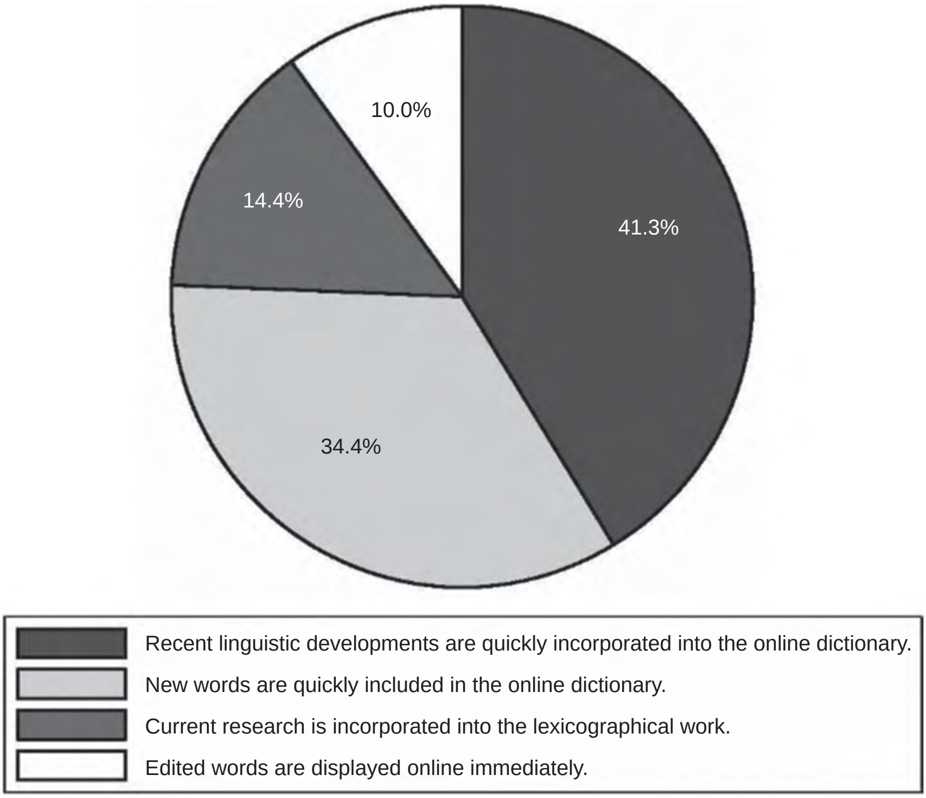


Fig. 9.8: Pie chart: aspects of dictionary updates.

In addition, for each aspect, we asked participants to list any further aspects that were perhaps also important in an open question. These will not be shown in detail here (cf. Müller-Spitzer/Koplenig 2014: 156–168). However, these free-text fields can sometimes provide indications that something was not understood. For example, some participants indicated to us in this field that they had not understood the formulation “words processed by editors appear immediately”:

- What are “words processed by editors”? Why should they not appear online? Did not understand the question.
- The user can contribute new words themselves and also potentially discuss them. In addition: I do not understand the option “words processed by editors appear immediately online”. As a result, I’ve rated it as less important.
- Comment on the above “Words processed by editors appear immediately online” – what does that mean? Everything is “immediately online”, isn’t it? And hopefully also processed by editors . . .

The aspect “words processed by editors . . .” relates to online dictionaries that publish their data online when they become available, such as ELEXIKO or the ALGEMEEN NEDERLANDS WOORDENBOEK (ANW; → Chapter 3.4.1). In these projects, the question arose whether the Internet dictionary should be updated from day to day, i.e. edited words are displayed online immediately, or whether a whole group of headwords should be released together every three months. Apparently, though, this problem was unfamiliar to many participants so they did not understand this option as an answer. As such, these open response fields provide the opportunity to identify problems in the clarity of the questionnaire, in addition to the standardised selection options.

The research question posed at the outset was which criteria characterise a good Internet dictionary in the opinion of our participants. What can our data tell us about that? Our studies showed that the classic features of dictionaries were very highly valued, especially the reliability of content. And this was not only the case in competition with the other criteria but also generally. That means that our participants expected an Internet dictionary to be a reliable reference work, above all, and that enriching it in a medium-specific way with innovative features was clearly subordinate to that. Here, there were no significant differences between groups: neither for age, nor professional background, nor language version. The hypothesis that linguists or translators would tend towards other judgements was also not confirmed. How can we interpret that? One possible explanation is that our participant group was too homogenous. However, we can refute that: the number of participants was high enough in both studies so that if there had been differences between participants with a linguistic background and those without, it is very likely that this would have shown up, especially because we were able to reach students as participants who were not studying linguistics. The same holds for age groups: the group sizes were sufficient to reveal differences if there had been any. As such, the much more plausible interpretation is that the participants – no matter what professional background they had, whether they lived in English-speaking or German-speaking countries, whether they were young or old – were surprisingly in agreement about which features make a good Internet dictionary. And those are features that have characterised good reference works for centuries: tools that are reliable in their content, clear to understand, and as up-to-date as possible with up-to-date knowledge. Thus, it is not the case that a user-friendly dictionary has to be one that is, above all, flexible (de Schryver 2003: 182) or fast (Almind 2005: 39; Bergenholtz 2011), as claimed in the publications just cited. Our empirical data demonstrate a different emphasis.

Does that mean that only those classic features count for digital dictionaries and that innovative features are unimportant, even though it is precisely those features that exploit the potential of the new medium and have great appeal? We would not necessarily draw this conclusion: in our studies innovative features may have been judged to be unimportant, but we were able to demonstrate in an experiment that this could lie in the fact that the participants were not familiar with enough examples to be able to evaluate these features. This experiment is the subject of the next section.

9.3.2 Does the evaluation of the innovative features of Internet dictionaries depend on previous knowledge?

In the last section we showed that, in contrast to the classic characteristics of good reference works (reliability of content, clarity), medium-specific possibilities for digital dictionaries (multimedia, user-adaptive customisation) were rated as unimportant. On the one hand, this is not surprising since a reference work with great multimedia components but unreliable content makes no sense. We also showed that these judgements were made not only in competition with one another but also independently of one another, in other words that this explanation was insufficient. Another interpretation was that our participants were perhaps not familiar with enough useful examples of these kinds of innovative features.

Thus, the research question here was whether the participants judged the usefulness of multimedia features or possible user-adaptive customisation more favourably when they were informed about these features first.[10] Our hypothesis was that the participants would judge their usefulness to be higher when they were informed about the options open up by these features beforehand, the underlying idea being that they were probably not familiar with enough examples from their everyday dictionary practice to be able to really judge how helpful these innovative features could be without this demonstration. In order to test this hypothesis, we integrated an experiment into the second online study (N=390). First, we showed the participants the possibilities of multimedia and user-adaptive features and then asked them how useful they thought these features were. The participants in the control group did not have any examples shown to them and were asked immediately how useful they thought these features were. The participants were allocated at random to one of the groups.

The result was that the participants in the test group judged the usefulness of these features to be significantly higher than the control group (→ Fig. 9.9). The graph is to be read as follows: The participants were asked to judge the usefulness of the features on a seven-point Likert scale. These values can be found on the y-axis. The distribution of data can be seen in the box plots. The shaded box corresponds to the region in which the middle 50% of the data points lie. The white horizontal line in the box shows the median (M = 5.02 in the condition with the learning effect (left) and 4.50 in the condition without the learning effect). The values lying outside the box are represented by the whiskers (i.e. the lines extending out of the box), which lie at a maximum distance of one and a half times the size of the box. Outliers would be represented in this kind of box plot as circles beyond the whiskers; however, in this case there were no outliers. The learning effect shown here is moderate but highly significant, which is the most important characteristic for the reliability of a statistical claim. Expressed in numbers:

[10] For a detailed presentation of this study, cf. Müller-Spitzer/Koplenig (2014).

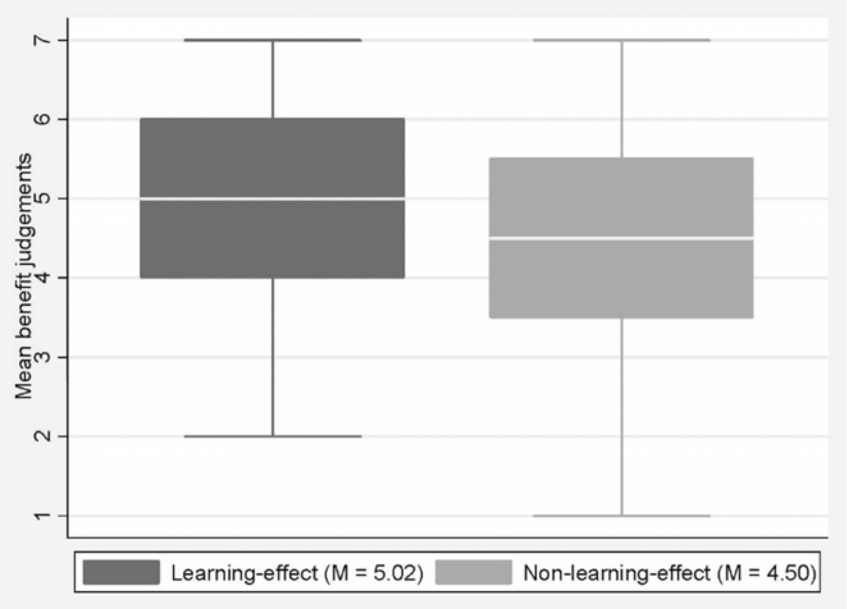


Fig. 9.9: Box plots: Evaluating multimedia and adaptive features depending on learning effect vs no learning effect.

the p-value is less than $p < 0.005$, i.e. the probability that the different judgements are a matter of chance is less than 1:1,000.

Thus, our hypothesis was confirmed in this experiment: participants who had innovative features shown to them first judged these as being more useful than participants who were not shown these examples. Our data show that it is worthwhile integrating innovative features into Internet dictionaries but also that the providers of these dictionaries have to understand that users can only be persuaded gradually to adopt these features. Or – as Trap-Jensen puts it – we "must make an effort" to bring innovative features closer to users:

> The lesson to learn is probably that both lexicographers and dictionary users must make an effort. Dictionary-makers cannot use the introduction of user profiles as a pretext for leaning back and do nothing but should be concerned with finding ways to improve presentation. (Trap-Jensen 2010: 1142; cf. also Heid/Zimmermann 2012: 669; Tarp 2011: 59; Verlinde/Peeters 2012: 151)

In any case, the issue is how this might look in practice since lexicographers do not generally have any direct contact with their users. One possibility could be to use situations in educational institutions, such as school or university classes, to establish contact with users, with the chance to educate them. This would certainly not reach the users who want to quickly check the spelling of a word but perhaps it would reach those who are interested in more extensive forms of dictionary use, such as more in-depth information about the range of meanings of headwords.

9.3.3 How do potential users cope with individual aspects of the new version of the OWID dictionary portal?

In this section, we present a further form of observation in the context of dictionary use, namely collecting data in the form of eye tracking. Eye tracking means recording a person's eye movements, primarily fixations (points which they look at closely), saccades (rapid eye movements between fixations), and regressions (backward jumps of the eye to a previous fixation point, for example); the devices used to record this are known as eye trackers. → Fig. 9.10 shows a PR image for an eye tracker like the one we used in our study.

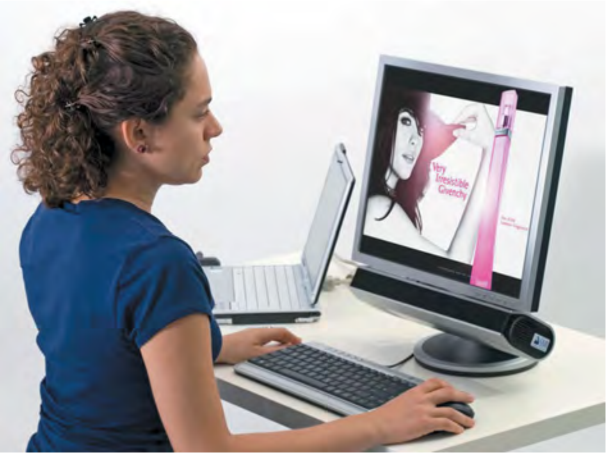


Fig. 9.10: PR image for the SMI Eye Tracker (http://www.gizmag.com/smired500-500hz-remote-eye-tracker/16957/picture/124519/ [last access: June 10, 2016].

In particular, Lew et al. (2013) present eye-tracking studies on users finding individual meanings in print dictionaries (for a summary of the results of other studies, cf. Lew et al. 2013, especially pp. 4–6; Lew 2010; Lew/Tokarek 2010; Nesi/Tan 2011; Tono 2001, 2011). The aims of our eye-tracking study were, first, to test this method of generating data in the context of our research project on dictionary user research and to gather experience in this area and, second, to evaluate the new version of the OWID dictionary portal which we had completed but not yet released online.[11]

[11] For a detailed presentation of this study, cf. Müller-Spitzer et al. (2014).

A suitably equipped laboratory is needed to conduct an eye-tracking study. For that reason, we carried out our study in collaboration with the University of Mannheim (Professor Rosemarie Tracy). The laboratory there is equipped with different computer work stations with an eye tracker suitable for reading-time experiments (a very high resolution is needed for these since they have to be able to see exactly which parts of a text are being read at the level of individual lines and words) and an SMI RED Remote Eye Tracker where a small box under the screen records the eye movements (→ Fig. 9.10). Each test subject sat in front of the eye tracker; the person conducting the test sat in the same room, separated by a partition screen. During the test, they had to check that the participants did not move out of the "field of view" of the eye tracker. Thus, the setup, or the design of the experiment, was relatively natural for the test subjects since no complicated equipment had to be used, unlike in the earliest eye-tracking studies (cf. for example the illustrations in the WIKIPEDIA article on *Eye tracking*[12]).

Thirty-eight people aged between 20 and 30 took part in our study, which was conducted in August/September 2011. All of the participants received a compensation of EUR10. Nearly 40 participants are a relatively high number for an eye-tracking study; other eye-tracking studies in dictionary user research only had 6 to 8 test subjects.

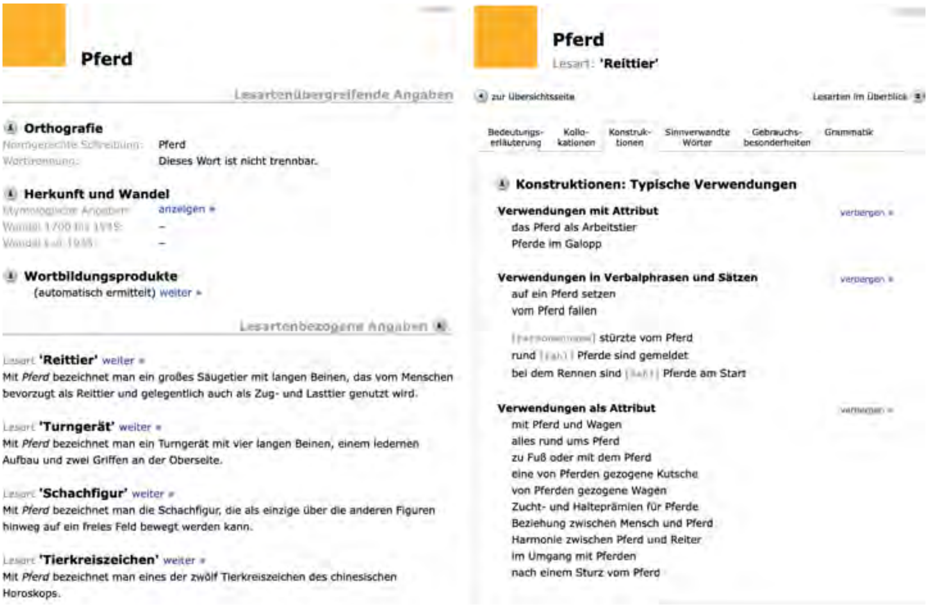
In our eye-tracking study, we wanted to study particular elements of the internal structure that we had changed in the new web design. One of these was navigation to the individual meanings in ELEXIKO, one of the dictionaries in OWID. In what follows, we will present the research question and the results of the study.

The information on a headword in ELEXIKO is distributed across two areas on the screen. The first page contains information that extends beyond individual meanings, such as the spelling of the word, syllabification, word formation, etc. while the information on individual meanings (referred to as *Lesarten* in ELEXIKO), typical usage, and related words follows on a second screen when an individual meaning is selected through the corresponding label. In turn, the information on individual meanings is distributed in individual tabs (→ Fig. 9.11, right-hand side).

In the old OWID layout, the individual meanings were listed on the first page of a word entry, each with the help of a word or short phrase, so-called labelling. This was changed in the new layout. Here, we added the paraphrases to the labels on the first screen, that is, the descriptions of the individual meanings. This was intended to help users gain a faster impression of the range of meanings of the word and the individual meaning relevant to each situation in which it is used (→ Fig. 9.11).

In the eye-tracking study, we wanted to investigate how the participants perceived this information. Or, more specifically: What did the patterns of eye movement

[12] https://en.wikipedia.org/wiki/Eye_tracking#/media/File:Yarbus_eye_tracker.jpg [last access: March 23, 2024]. .



Pferd

Orthografie
Normgerechte Schreibung: Pferd
Worttrennung: Dieses Wort ist nicht trennbar.

Herkunft und Wandel
Etymologische Angaben: anzeigen »
Wandel 1700 bis 1945: –
Wandel seit 1945: –

Wortbildungsprodukte
(automatisch ermittelt) weiter »

Lesart **'Reittier'** weiter »
Mit *Pferd* bezeichnet man ein großes Säugetier mit langen Beinen, das vom Menschen bevorzugt als Reittier und gelegentlich auch als Zug- und Lasttier genutzt wird.

Lesart **'Turngerät'** weiter »
Mit *Pferd* bezeichnet man ein Turngerät mit vier langen Beinen, einem ledernen Aufbau und zwei Griffen an der Oberseite.

Lesart **'Schachfigur'** weiter »
Mit *Pferd* bezeichnet man die Schachfigur, die als einzige über die anderen Figuren hinweg auf ein freies Feld bewegt werden kann.

Lesart **'Tierkreiszeichen'** weiter »
Mit *Pferd* bezeichnet man eines der zwölf Tierkreiszeichen des chinesischen Horoskops.

Pferd
Lesart: **'Reittier'**

Konstruktionen: Typische Verwendungen

Verwendungen mit Attribut
das Pferd als Arbeitstier
Pferde im Galopp

Verwendungen in Verbalphrasen und Sätzen
auf ein Pferd setzen
vom Pferd fallen

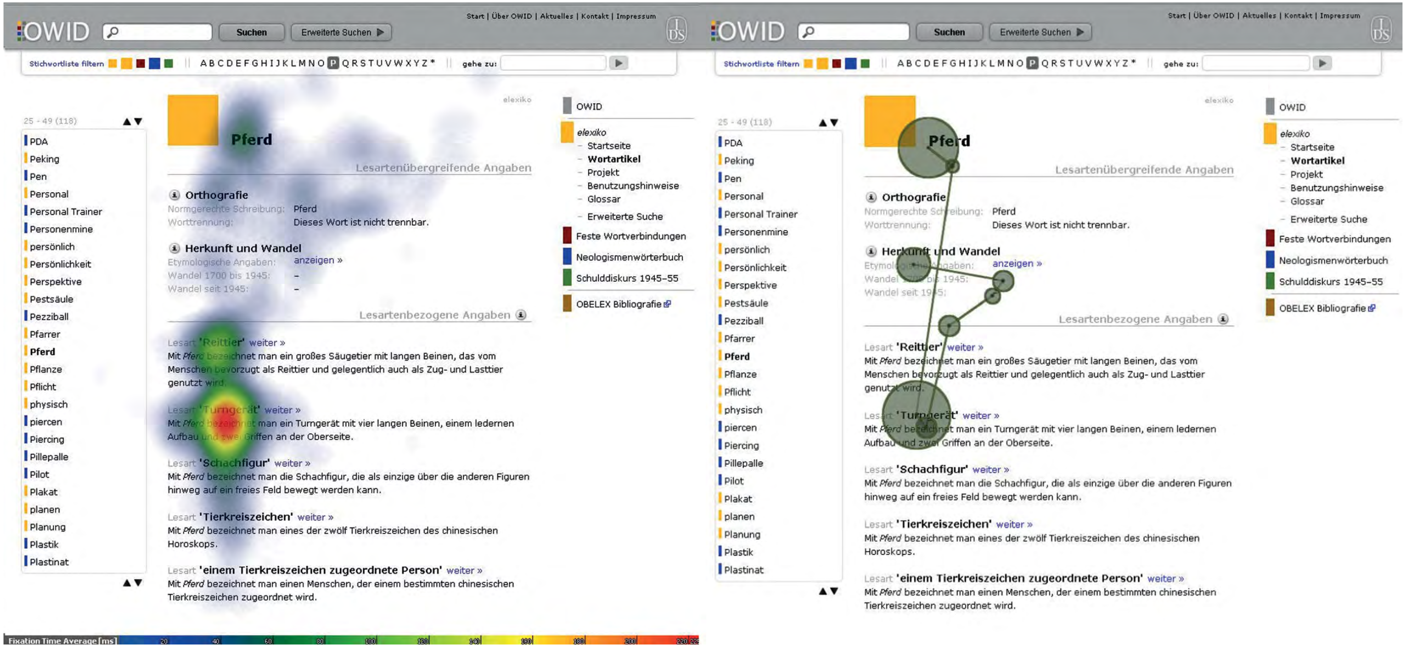
|Personenname| stürzte vom Pferd
rund |Zahl| Pferde sind gemeldet
bei dem Rennen sind |Zahl| Pferde am Start

Verwendungen als Attribut
mit Pferd und Wagen
alles rund ums Pferd
zu Fuß oder mit dem Pferd
eine von Pferden gezogene Kutsche
von Pferden gezogene Wagen
Zucht- und Halteprämien für Pferde
Beziehung zwischen Mensch und Pferd
Harmonie zwischen Pferd und Reiter
im Umgang mit Pferden
nach einem Sturz vom Pferd

Fig. 9.11: General information (left) and meaning-specific information (right) in ELEXIKO.

look like when we asked the participants about individual meanings? Did they find the relevant meanings? Did they read or scan all the labels first and only then read the paraphrases? Or was it a linear reading process (even though that is very unlikely)? When developing the new design, our intention was that the labelling would “catch the eye” first and the full paraphrase would only be read if necessary. If this was reproduced in the scan paths of our participants, we would be able to see this as confirmation of our design.

The procedure for the study was as follows. In the first task, participants were asked to check whether the headword *Pferd* ‘horse’ had the meaning *Turngerät* ‘gym equipment’: “On the next page you will see an entry from ELEXIKO. Please try to find out whether the word has a meaning in the sense of ‘gym equipment’”. This was to enable us to test whether the participants could find the relevant meaning quickly. The results can be seen in → Fig. 9.12. On the left-hand page we can see a so-called *heat map*, which displays the cumulative viewing of an area by all participants; the *fixation duration* is illustrated by a corresponding colour. We can see that attention was concentrated on the relevant individual meaning. The *scan path* of one individual participant can be seen on the right-hand side of → Fig. 9.12. Here, it is possible to see the fixation steps taken by the test subject in their search. Overall, the eye-tracking data show that the relevant individual meaning was found quickly in this relatively simple task.



OWID
Suchen
Erweiterte Suchen ▶
Start | Über OWID | Aktuelles | Kontakt | Impressum
Stichwortliste filtern
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z *
gehe zu:
elexiko
Fixation Time Average [ms]
20
40
60
80
100
120
140
160
180
200
220

25 - 49 (118)

PDA
Peking
Pen
Personal
Personal Trainer
Personenmine
persönlich
Persönlichkeit
Perspektive
Pestsäule
Pezziball
Pfarrer
Pferd
Pflanze
Pflicht
physisch
piercen
Piercing
Pillepalle
Pilot
Plakat
planen
Planung
Plastik
Plastinat

Pferd

Lesartenübergreifende Angaben

Orthografie
Normgerechte Schreibung: Pferd
Worttrennung: Dieses Wort ist nicht trennbar.

Herkunft und Wandel
Etymologische Angaben: anzeigen »
Wandel 1700 bis 1945: –
Wandel seit 1945: –

Lesartenbezogene Angaben

Lesart **'Reittier'** weiter »
Mit *Pferd* bezeichnet man ein großes Säugetier mit langen Beinen, das vom Menschen bevorzugt als Reittier und gelegentlich auch als Zug- und Lasttier genutzt wird.

Lesart **'Turngerät'** weiter »
Mit *Pferd* bezeichnet man ein Turngerät mit vier langen Beinen, einem ledernen Aufbau und zwei Griffen an der Oberseite.

Lesart **'Schachfigur'** weiter »
Mit *Pferd* bezeichnet man die Schachfigur, die als einzige über die anderen Figuren hinweg auf ein freies Feld bewegt werden kann.

Lesart **'Tierkreiszeichen'** weiter »
Mit *Pferd* bezeichnet man eines der zwölf Tierkreiszeichen des chinesischen Horoskops.

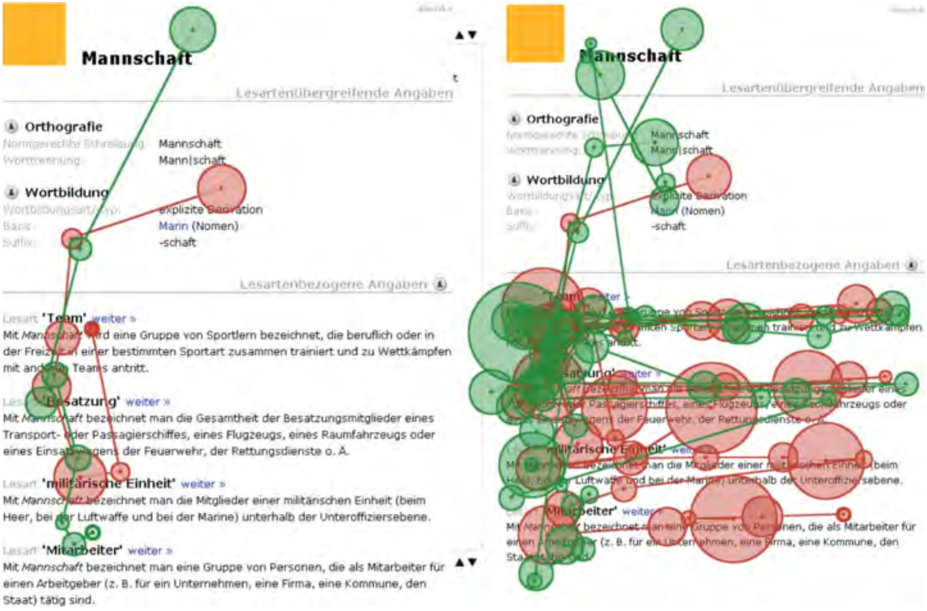
Lesart **'einem Tierkreiszeichen zugeordnete Person'** weiter »
Mit *Pferd* bezeichnet man einen Menschen, der einem bestimmten chinesischen Tierkreiszeichen zugeordnet wird.

OWID
elexiko
- Startseite
- **Wortartikel**
- Projekt
- Benutzungshinweise
- Glossar
- Erweiterte Suche

Feste Wortverbindungen
Neologismenwörterbuch
Schulddiskurs 1945–55
OBELEX Bibliografie

Fig. 9.12: Heat map of all of the participants (left); scan path of one participant (finding the individual meaning ‘Turngerät’; right).

In a second stage, we asked participants to find a particular meaning of the headword *Mannschaft* ‘team’: “Please try to find out whether the following entry contains a meaning which is explained as ‘members of a group of people active in an organisation’. If so, which one?” The results are shown in → Fig. 9.13.



Mannschaft

Lesartenübergreifende Angaben

Orthografie
Normgerechte Schreibung: Mannschaft
Worttrennung: Mann|schaft

Wortbildung
Wortbildungsart/-typ: explizite Derivation
Basis: Mann (Nomen)
Suffix: -schaft

Lesartenbezogene Angaben

Lesart 'Team' weiter »
Mit *Mannschaft* wird eine Gruppe von Sportlern bezeichnet, die beruflich oder in der Freizeit in einer bestimmten Sportart zusammen trainiert und zu Wettkämpfen mit anderen Teams antritt.

Lesart 'Besatzung' weiter »
Mit *Mannschaft* bezeichnet man die Gesamtheit der Besatzungsmitglieder eines Transport- oder Passagierschiffes, eines Flugzeugs, eines Raumfahrzeugs oder eines Einsatzwagens der Feuerwehr, der Rettungsdienste o. Ä.

Lesart 'militärische Einheit' weiter »
Mit *Mannschaft* bezeichnet man die Mitglieder einer militärischen Einheit (beim Heer, bei der Luftwaffe und bei der Marine) unterhalb der Unteroffiziersebene.

Lesart 'Mitarbeiter' weiter »
Mit *Mannschaft* bezeichnet man eine Gruppe von Personen, die als Mitarbeiter für einen Arbeitgeber (z. B. für ein Unternehmen, eine Firma, eine Kommune, den Staat) tätig sind.

Fig. 9.13: Scan paths of two participants (stored on film); one snapshot at 00:01 seconds (left) and the other at 00:14 seconds (right).

What is interesting here is that the participants obviously first scanned the labels very quickly (both participants here had already scanned all of the labels after one second) and only then turned their attention to the paraphrases. This corresponds to the process that we had intended with the new design. Overall, we can conclude from this section of the eye-tracking study that the participants found the relevant meanings and that the different functions of the labels and paraphrases were clear in practice, in the way they had been conceived in the new design.

One supplementary note: nobody in our team had had experience with this method of collecting data before we ran this study. Only in the analysis, for example, did we realise that it would have been better to use more comparative views of the old layout compared to the new layout in order to really be able to conclude that the new layout worked better than the old one. In the way that we conducted the study, it was often only possible to conclude that the new layout worked well – as in the case above – but the old layout might also have done exactly the same. Of course, these learning processes are part of research.

9.3.4 Do lexicographic resources really help with linguistic problems?

At the beginning of this chapter, we claimed that the primary purpose of dictionaries is to be used as tools to work on language tasks or to solve linguistic problems. But can dictionaries or, more generally, lexicographic resources even satisfy this expectation? To find out, we conducted a user study in which they asked native speakers of German to solve a realistic language task, namely revising a text in their L1.[13] The linguistic problems contained in the two presented texts were not real errors (e.g. spelling mistakes) but rather something we referred to internally as "stumbling blocks". These were problems like an inappropriate choice of words (e.g. a regional variant instead of Standard German), too condensed a formulation (e.g. the German equivalent of "the most important phase of a human" instead of "the most important phase in the life of a human"), poor collocational choices, or inappropriate use of prepositions.

To isolate the effect that the presence of lexicographic resources had on the solution process, we worked with an experimental paradigm, that is, we assigned our participants randomly to one of three experimental groups. The first group, which we called "only text", received no help at all and were simply presented with the plain texts. This group served as a baseline condition to see what would happen if participants received no help at all. The second group ("highlighted") received versions of the texts where all of the problems were highlighted in yellow. Only the third group ("full") saw the text with the highlighted problems and lexicographic material suitable for solving the linguistic "stumbling block" (see the original publication for an overview of the resources used). Note that we have already solved an important task for the participants in this group: finding the appropriate lexicographic resource for a particular linguistic problem. This was intentional because our primary research question was whether linguistic problems would be solved better with the appropriate lexicographic resource at hand – assuming this resource had already been found.[14]

Our participants were 105 undergraduate students of German linguistics at the University of Mannheim and participation in the study was a course requirement. After excluding participants from the analyses who stated that German was not their native language as well as participants who took less than five minutes on the texts, data from 78 participants entered the final analyses. These were distributed roughly equally over the experimental conditions (26 for "only text", 25 for "highlighted", and 27 for "full"). We also asked the participants how often they used monolingual dictionaries, and there was no tendency for participants in one experimental condition to

[13] For a detailed presentation of this study, cf. Wolfer et al. (2016).

[14] In another, more explorative study (Müller-Spitzer et al. 2018), we presented another group of participants (learners of German with Spanish, Portuguese, Galician, or Italian as their L1) with a different linguistic task without giving them any lexicographic resources at all. This study, however, was based on a different research question.

use dictionaries more often than in another. Hence, none of the effects of the experimental condition that are reported below is attributable to the participants' different levels of experience using general dictionaries. Each participant received two texts (in randomised order) and a total of 35 language problems that we had identified beforehand. Taken together, all participants saw 78 * 35 = 2,730 language problems.

After all of the participants had revised their texts, we noted whether the problems we had identified beforehand had been changed. We ignored all of the other changes that the participants made to the texts. If a problem had been changed, we further noted if this change solved the problem ("improvement") or actually made it worse, for example by altering the meaning of the text ("semantic distortion").

We found that the participants in the "only text" condition, who received no help at all and only saw the texts without any highlighting or resources, changed 36% of the problems. This stands in sharp contrast to the "full" condition where 89% of the problems were changed. The "highlighted" condition was in an intermediate position at 75%. All of these differences were statistically significant. However, the more relevant question is actually whether the participants with lexicographic resources *improved* more of the problems. So, we only looked at the 1,838 problems that had been changed and saw that for the "full" condition, 76% of the problems had been improved. This is a statistically significant difference to the 59% in the "only text" condition. Again, the "highlighted" condition lay in between at 64%. Not only did the participants in the "full" condition improve more problems, they also introduced fewer semantic distortions (13% vs 20% for "highlighted" and 28% for "only text"). To sum up, the participants who got help with appropriate lexicographic resources changed and improved linguistic problems more often and introduced fewer semantic distortions than the participants in the other two experimental groups. The results for improved and semantically distorted problems are visualised in → Fig. 9.14.

We can also look at these results from another perspective: if we give each participant one point when improving a problem and subtract one point for each inappropriate revision, each participant can receive a maximum score of 35 (all problems changed and all improved) and a minimum score of -35 (all problems changed but all made worse). The average scores over the experimental conditions give a pretty clear impression of how successful the three groups were at revising the texts. The mean score was 10.4 for the "highlighted" condition and 3.6 for the "only text" condition. Participants in the "full" condition reached an average score of 18.6, which was significantly better than both of the other groups (→ Fig. 9.15). Not only did the participants with the lexicographic resources score higher but they also achieved more points per minute (0.62) than both the "highlighted" (0.46) and the "only text" (0.19) groups. That means that although the "full" group took longer to work on the task (an average of 31.6 minutes compared to 26.9 minutes for the "highlighted group" and 24.8 minutes for the "only text" group) because they had to integrate the lexicographic resources into their task, it was worth it because they achieved more successful results per minute.

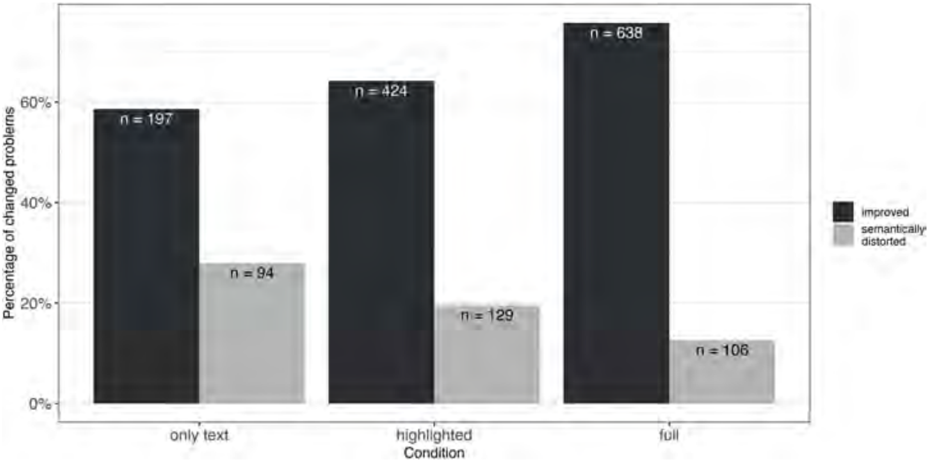


Fig. 9.14: Improved and semantically distorted problems under the three conditions. On the y-axis, the percentage of improved vs. semantically distorted problems is indicated (100% represents all problems). The figures in the bars give the raw number of linguistic problems for each category.

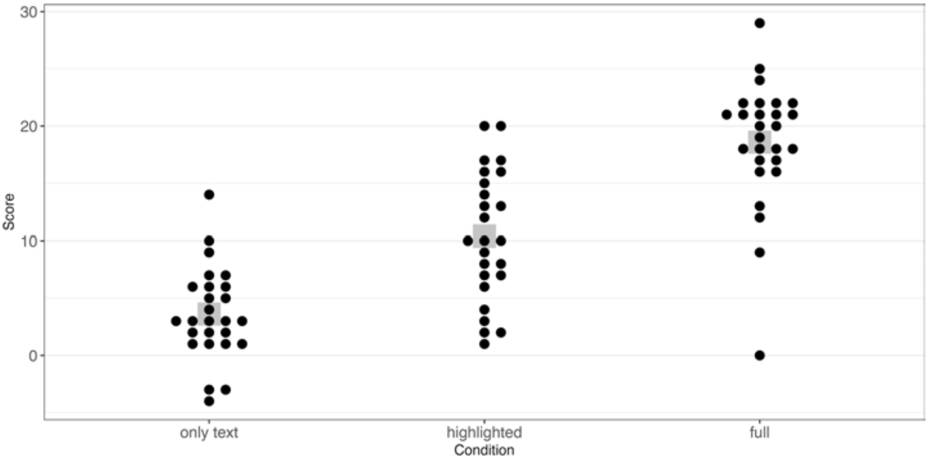


Fig. 9.15: Scores for all of the participants in the three experimental conditions. One black dot stands for one participant. Grey squares indicate the mean values of the experimental conditions.

Taken together, these results paint a fairly clear picture of the benefits of working with lexicographic resources: the experimental group which received the most assistance indeed made more changes, more improvements, and fewer wrong revisions. Moreover, they achieved more points and worked more efficiently.

However, it must also be noted that although the results improved considerably, the participants did not perform perfectly when provided with lexicographic resources. Even though we maximised the helpfulness of the resources by handpicking the relevant information for certain language problems, the participants still had to understand them and put them to good use when revising a text. For practical applications, this poses two major challenges: selecting a suitable resource for a given problem and selecting the relevant parts of this resource (e.g. a dictionary entry). This implies two things: First, if we could manage to create electronic writing environments that automatically provide users with the appropriate lexicographic resources, this would most likely significantly improve writing and/or revision products. Second, language users should be trained to find and use the appropriate lexicographic resources for their specific problems. Only then can they exploit the full potential of these resources.

Overall, the changes in writing conditions that have taken place since the time of the study must also be taken into account. AI-based systems such as DeepLWrite can now do a very good job of correcting a text which has already been written, at least for certain languages. Even ChatGPT could be put to good use when formulating text if it is prompted accordingly.[15] Of course, these systems might also be easier and faster for the user to work with than most "traditional" dictionaries. It remains to be seen what effects such systems will have on assisted writing in the future and whether dictionaries or lexicographic resources will be relevant at all.

9.3.5 Are frequent words in the corpus also consulted frequently in Internet dictionaries?

We conclude this section with an example of a study in which we made use of a non-reactive method to collect data, namely the analysis of log files from the German WIKTIONARY and the DIGITALES WÖRTERBUCH DER DEUTSCHEN SPRACHE (DWDS).[16]

The research question we pursued in this study was as follows: "Are words that occur frequently in the corpus also frequently consulted in a dictionary?" This question is particularly interesting if a new dictionary is to be compiled and we do not have a precise target group for which the appropriate selection of headwords is already clear (e.g. for a terminological dictionary or a dictionary designed for learners at a particular level). A relevant question in that process is which words should be prepared first. As a rule, it is desirable to first focus lexicographic work on the words that are looked up frequently in order to spare users unsuccessful searches. However, previous studies (de Schryver et al. 2006; Verlinde/Binon 2010) have shown that the frequency of a word in the corpus has little influence on whether it will be looked up frequently. For de Schryver and his colleagues this led to the conclusion that basing

[15] For a recent study investigating the performance of learners of English using a dictionary vs. Chat GPT see Ptasznik et al. (2024).

[16] For a detailed discussion of this study, cf. Koplenig et al. (2014).

the selection of headwords on the underlying corpus was overvalued in lexicography (the title of their article is “On the Overestimation of the Value of Corpus-based Lexicography”). However, the members of our team who are versed in statistics noticed that their research used an approach for data analysis which could prove problematic to prove their point. Thus, this is an example of how important it is to have the relevant knowledge of data analysis in order to be able to identify weaknesses in previous research and find better ways of approaching it.

The approach to data analysis in previous studies seems problematic for the following reasons. Linguistic data are, for the most part, distributed according to *Zipf's law*; in other words, there are a very small number of very frequent words and a very large number of very rare words. One example for a Zipfian distribution can be found in → Fig. 9.16. Data in text corpora are also distributed according to this pattern: we find a small number of very frequent words, like *der* ‘the’, *die* ‘the’, or *in* ‘in’, and a very large number of words that only occur very rarely, like *Amaryllis* ‘amaryllis’ or *Studienbuch* ‘text book’. In order to examine whether the corpus frequency of a word has any bearing on the frequency with which a word is looked up, de Schryver et al. examined whether the frequency rank of words correlated with the rank order with which they were looked up. The problem in the kind of analysis that was applied in their study is that the differences between individual ranks were treated as the same; in other words, the difference between the first and second positions was seen as being the same as that between numbers 100,001 and 100,002. However, a Zipfian distribution of data points means that these places are not equidistant. For example, in the frequency lists of the Deutschen Referenzkorpus (DeReKo), which we used in our study, the frequency difference between the first two positions is 251,480 (i.e. the word in the top position occurs more than 250,000 times more than the second one), while the difference in frequency between positions 3,000 and 3,001 is only five. Yet this difference is not taken into account by de Schryver et al. in their correlation analysis. It may be, then, that this analytical approach led to the conclusion that there was no strong correlation between corpus frequency and the frequency of a word being looked up.

Hence, we took a different approach in our study. As data, we used the absolute and relative frequencies of the 100,000 most common words in the DeReKo and the log files of the DWDS and the German Wiktionary for the whole of 2012. We chose the following method for our analysis. First of all, we had to make the log files from the two dictionaries comparable with one another. We achieved this by introducing the value *poms*. Here a value of 8 poms, for example, means that the term in question was searched for 8 times “*p*er *o*ne *m*illion” search queries. Then, we created the following categories: if a word has the value of 1 poms, or occurs at least once in every 1,000,000 search queries, we state that the word is searched for *regularly*. If the poms is at least 2, then the word is searched for *frequently*. If the poms value is greater than 10, we talk of the term being searched for *very frequently*. In this way, we get around the problem of individual ranks being compared to one another when the gaps between them are not actually comparable. → Tab. 9.1 summarises the results of this analysis of our log files.

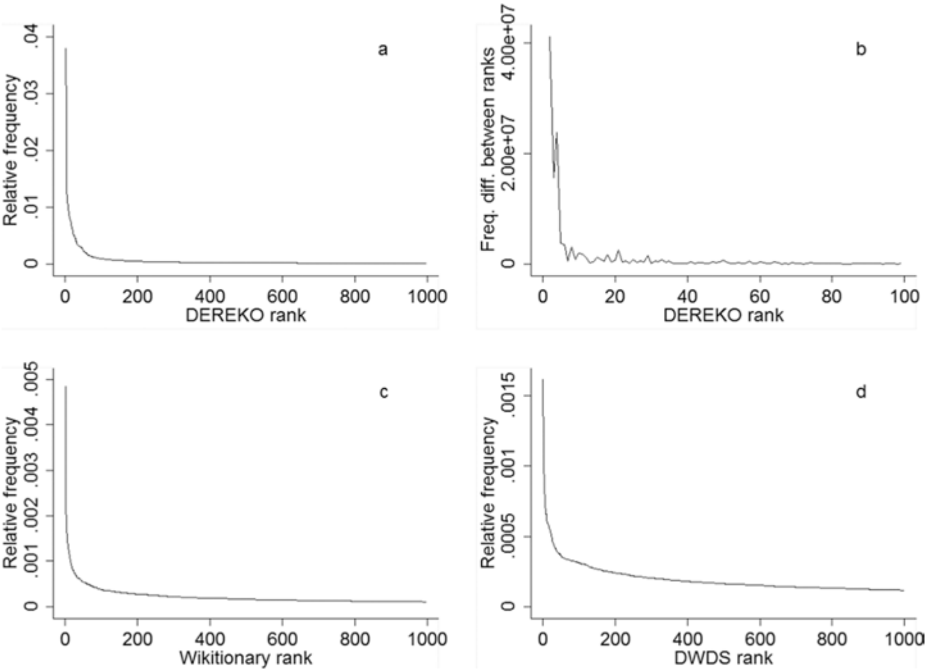


Fig. 9.16: Distribution of corpus and log file data (from the DeReKo and Wiktionary/DWDS) as examples of a Zipf distribution (Koplenig et al. 2014: 238).

Tab. 9.1: Relationship between corpus rank and log file data.

DeReKo rankings	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.0	7.5	99.5	99.5	86.5
2,000	96.9	91.0	67.6	98.4	96.0	64.9
10,000	85.5	72.9	47.5	86.3	75.3	40.2
15,000	80.3	66.5	41.8	77.4	66.1	33.7
30,000	69.4	54.6	31.3	62.7	50.9	23.4

The relationship between corpus ranking and frequency of consultation becomes apparent in this table: the more DeReKo ranks are included in the analysis, the smaller the percentage of words that are consulted normally/frequently/very frequently, both in the DWDS and in Wiktionary. For example, if we imagine compiling a dictionary with the 2,000 most frequently occurring words in a corpus, this table tells us the following: 96.9% of these words are regularly searched for in the DWDS, 91% are frequently searched for, and nearly 67% very frequently. Thus, there does seem to be a relationship between corpus frequency and frequency of consultation. This also becomes clear in a second analysis. de Schryver et al. claim that, "beyond the top few thousand words" (de Schryver et al. 2006: 79), it would make no difference which words to select next (whether the next ten thousand or very rare ones). To check this, we removed the 10,000 most frequent words from the analysis and then created a random sample from log files of 10,000 other words. The analysis revealed that 34% of these were consulted in Wiktionary and 45% in the DWDS. As a comparison we took the words with frequency ranks 10,001–20,000 in the DeReKo. If the claim made by de Schryver et al. were confirmed by our analysis, we would expect there to be similar percentages for these 10,000 words. However, this was not the case: in this case 56% (instead of 34%) were looked up in Wiktionary and 67% (instead of 45%) in the DWDS. That is, our results suggest that users very probably look up frequent words but also words outside the top 10,000. As such, this study is also an example of a case where replicating studies, but with other statistical methods, can lead to different results.

In the meantime, the effect of frequency on dictionary look-ups has been replicated for other dictionaries and other languages. De Schryver et al. (2019) found the same relationship for a Swahili-English dictionary. They used the method we introduced above and applied it to log files of a whole decade of user interaction with both the Swahili and English entries in the dictionary. Frequency effects on dictionary look-ups can be shown for both Swahili and English queries and also for less frequent words (beyond frequency rank 5,000 and 10,000). Lew and Wolfer (2022) show similar effects for the English Wiktionary. They demonstrated that corpus frequency is a better predictor of dictionary look-ups than polysemy (words with multiple meanings are looked up more often), age-of-acquisition (words that are acquired later in life are looked up more often), and prevalence (words that are known to more people are looked up less often). All of these other factors are indeed relevant in predicting dictionary look-ups, but corpus frequency is by far the most important one.

In another log file study (Wolfer et al. 2014), we investigated whether there was anything else which stood out in the behaviour of users, beyond the effects of frequency. To do this, we again analysed the log files of the German-language Wiktionary (this time from January to August 2013). What is striking here is that, first, words that are the subject of general lexical-semantic discussion are consulted noticeably more often. One word that was notable in this respect was the headword *Furor* 'furore'. At the beginning of March, Joachim Gauck (then the president of Germany) had used the word *Tugendfuror* 'virtue furore' in relation to the debate on everyday sexism, thereby sparking a debate about whether this was an appropriate way to phrase it. It came as no surprise that a word like this was subsequently looked up frequently – it was, at least temporarily, a word of great social relevance.

Surprisingly we found that the word *larmoyant* 'lachrymose' was looked up particularly frequently on one day. Our search revealed that the TV commentator on a football match involving the men's German national football team had noted (on 6.2.2013): "Der [Joachim Löw] ist jetzt aber richtig sauer. Das ist ihm ein bisschen zu larmoyant . . ." (Literally: "[Joachim Löw] is really angry now. That was just a little too lachrymose for him . . ."). Within the hour, this led to a statistically noticeable increase in queries for this word. This seemed noteworthy to us because there was such a direct connection between watching a football match and searching in WIKTIONARY – a relationship that probably never existed for print dictionaries. In exactly the same way, the word *Borussia* was looked up more and more frequently the further the German football team Borussia Dortmund got in the Champions League (→ Fig. 9.17). This is also not necessarily to be expected because the word *Borussia* is not the subject of a discussion about its meaning in the narrow sense and it is perhaps also not to be expected that during a football match, or immediately after it, the correct spelling of *Borussia* would be checked. Further research questions that can be investigated with this type of analysis are, for example, whether the ambiguity of a word correlates with its frequency of consultation (i.e. whether polysemous words are looked up more frequently in the dictionary, cf. Müller-Spitzer et al. 2015 and Lew and Wolfer 2022) or whether there are groups of words that are often looked up together. To take these kinds of observations and analyses further is certainly an exciting task for future research.

9.4 Outlook

An argument is sometimes raised against making current dictionaries the object of user research because this method of research could hinder innovation since it takes as its starting point existing dictionaries, thereby making it impossible to imagine possible innovations. No matter how sensible or useful they are in the long run, innovations are unfamiliar at the beginning and, therefore, a hurdle to overcome. However, the criticism is only partially valid because dictionary user research does not always mean taking already existing dictionaries as the starting point. For example, it is possible to make the evaluation of innovative features that do not yet exist in practice the subject of a study as demonstrated in → Section 9.3.2.

At the same time, it is important in user research not to lose sight of dictionary use as the starting point, that is, situations in which language difficulties occur and from which the need to consult a dictionary arises. In essence, if we wish user re-

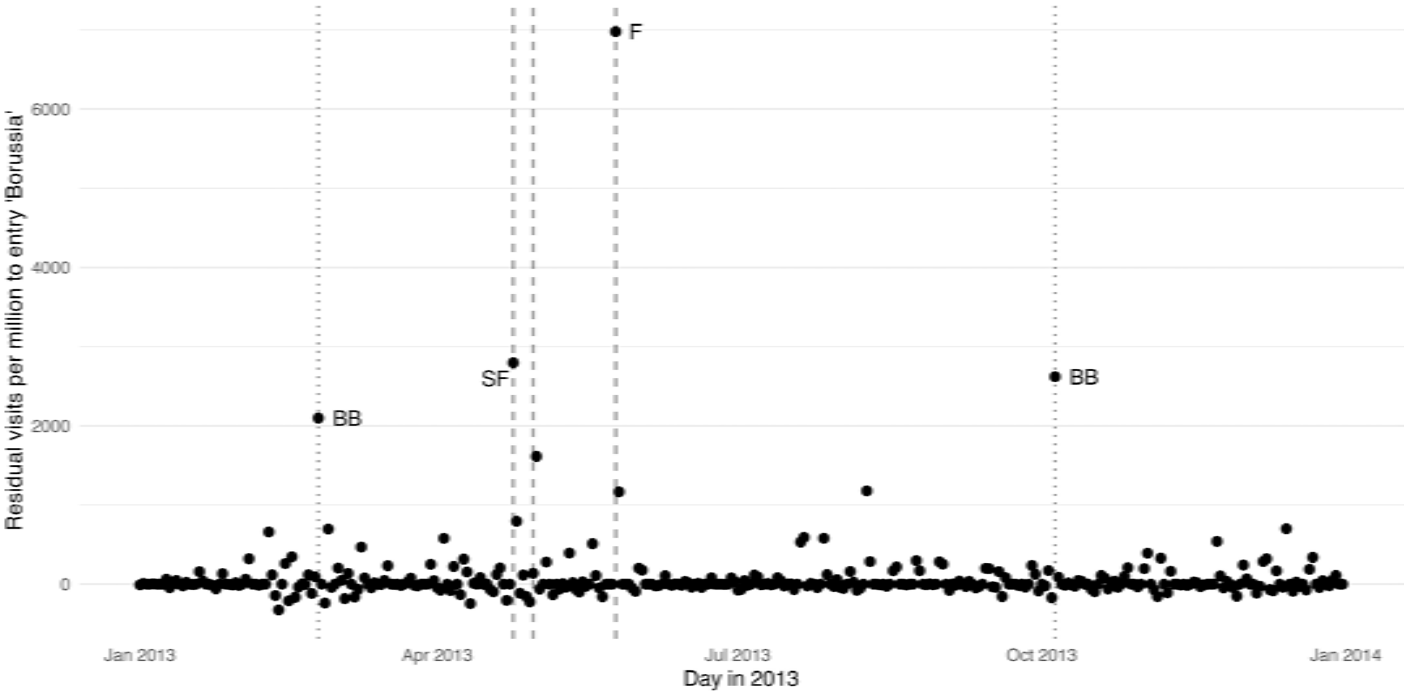


Fig. 9.17: Access to the word *Borussia* cleaned of trends (Jan.–Dec. 2013). Dotted vertical lines (BB) mark football matches between the German teams Borussia Mönchengladbach and Borussia Dortmund. Dashed lines indicate the Champions League 2023 semi-final matches (SF) and the final match (F).

search to ensure that dictionary use corresponds more closely to actual user needs, we should begin precisely with those user needs. Theodore Levitt, a US economist, wrote an influential article in the 1960s entitled “Marketing Myopia”, in which he pointed to exactly this aspect, namely that industry is not about limiting itself to one product or one type of product either but about concentrating on the purpose for which the product was developed:

> The railroads did not stop growing because the need for passenger and freight transportation declined. That grew. The railroads are in trouble today not because the need was filled by others (cars, trucks, airplanes, even telephones), but because it was not filled by the railroads themselves. They let others take customers away from them because they assumed themselves to be in the railroad business rather than in the transportation business. The reason they defined their industry wrong was because they were railroad-oriented instead of transportation-oriented; they were product-oriented instead of customer-oriented (Levitt 1960: 24)

Applied to dictionary user research, this means that it should extend its perspective beyond its examination of the use of dictionaries that exist today and on to the language problems in which the need to consult them arose (cf. for such an approach Müller-Spitzer et al. 2018). Lexicography finds itself in a difficult situation today: in the era of free Internet dictionaries, fewer and fewer dictionaries are being bought so that publishers are having great difficulty maintaining their staff and resources. And the public purse is hardly funding lexicographic projects any more that extend across decades. At the same time, very many language questions are being researched on the Internet – perhaps, or very probably – more than were ever looked up in print dictionaries. As such, the question is how we can integrate this activity more effectively with the available lexicographic resources. A question to which user research can contribute a great deal if it explores this wider field.

Bibliography

Further reading

Dziemanko, Anna (2012): On the use(fulness) of paper and electronic dictionaries. In: Granger, Sylviane/ Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 320–341. https://doi.org/10.1093/acprof:oso/9780199654864.003.0015 [last access: May 2, 2024]. *Offers a good summary of relevant studies comparing print and digital dictionaries, a topic which is hardly dealt with in this chapter*.

Lew, Robert (2015): Dictionaries and Their Users. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45369-4_11-1 [last access: May 2, 2024]. *Offers another good general introduction on the topic of “dictionary user research”.*

Müller-Spitzer, Carolin, et al. (2018): Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. In: *Lexikos* 28, 287–315. https://doi.org/10.5788/28-1-1466 [last access: May 2, 2024]. *Provides an insight into a study that combines quantitative and qualitative methods, something that is also less represented in this chapter*.

Töpel, Antje (2014): Review of research into the use of electronic dictionaries. In: Müller-Spitzer, Carolin (ed.): Using Online Dictionaries. Berlin/Boston: De Gruyter, 145. *Contains an extensive overview of user studies already conducted on digital dictionaries (until 2014).*

Literature

Academic literature

Almind, Richard (2005): Designing Internet Dictionaries. In: *Hermes* 34, 37–54.

Atkins, B. T. Sue (1992): Putting lexicography on the professional map. Training needs and qualifications of lexicographers. In: Alvar Ezquerra, Manuel (ed.): *Proceedings of the 4th Euralex Conference 1990*, Barcelona, 519–526.

Atkins, B. T. Sue (1996): Bilingual dictionaries: Past, present and future. In: Corréard, Marie-Hélène (ed.): *Lexicography and natural language processing* 96. Huddersfield: Euralex, 1–29.

Baayen, R. Harald (2008): *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bergenholtz, Henning (2011): Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries. In: Bergenholtz, Henning/Fuertes-Olivera, Pedro Antonio (eds.): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 30–53.

Bogaards, Paul (2003): Uses and users of dictionaries. In: van Sterkenburg, Piet (ed.): *A Practical Guide to Lexikography*. Amsterdam/Philadelphia: Benjamins, 26–33.

Bowker, Lynne (2012): Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities. In: Granger, Sylviane/Paquot, Magali (eds.): Electronic lexicography. Oxford: Oxford University Press, 379–397.

Chadefaux, Thomas (2014): Early warning signals for war in the news. In: *Journal of Peace Research* 51:1, 5–18.

de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic-Dictionary Age. In: *International Journal of Lexicography* 16/2, 143–199. https://doi.org/10.1093/ijl/16.2.143 [last access: May 2, 2024].

de Schryver, Gilles-Maurice et al. (2006): Do dictionary users really look up frequent words? – on the overestimation of the value of corpus-based lexicography. In: *Lexikos* 16, 67–83.

de Schryver, Gilles-Maurice/Lew, Robert/Wolfer, Sascha (2019): The relationship between dictionary look-up frequency and corpus frequency revisited: A log-file analysis of a decade of user interaction with a Swahili-English dictionary. In: *GEMA Online Journal of Language Studies* 19, 1–27. https://doi.org/10.17576/gema-2019-1904-01 [last access: May 2, 2024].

Diekmann, Andreas (1994): Umweltverhalten zwischen Egoismus und Kooperation. In: *Spektrum der Wissenschaft* 6/1994, 20–24.

Diekmann, Andreas (2011): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. Hamburg: Rowohlt.

Dziemanko, Anna (2012): On the use(fulness) of paper and electronic dictionaries. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 320–341.

Engelberg, Stefan (2014): Gegenwart und Zukunft der Abteilung Lexik am IDS: Plädoyer für eine Lexikographie der Sprachdynamik. In: *50 Jahre IDS*. Mannheim: Institut für Deutsche Sprache, 243–253.

Granger, Sylviane (2012): Introduction: Electronic lexicography – from challenge to opportunity. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 1–11.

Gries, Stefan Thomas (2021): *Statistics for linguistics with R: a practical introduction*. Berlin/Boston: De Gruyter Mouton.

Hanks, Patrick (2012): Corpus evidence and electronic lexicography. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 57–82.

Heid, Ulrich/Zimmermann, Jan Timo (2012): Usability testing as a tool for e-dictionary design: collocations as a case in point. In: Torjusen, Julie Matilde/Fjeld, Ruth V. (eds.): *Proceedings of the 15th EURALEX International Congress 2012*, *Oslo, Norway*, *7–11 August 2012*. Oslo: Universitetet Oslo, 661–671.

Kidd, Celeste/Palmeri, Holly/Aslin, Richard N. (2013): Rational snacking: young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. In: *Cognition* 126/1, 109–114.

Koplenig, Alexander (2014): Empirical research into dictionary use. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 55–76.

Koplenig, Alexander/Meyer, Peter/Müller-Spitzer, Carolin (2014): Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 229–249.

Kosem, Iztok, et al. (2019): The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture. In: *International Journal of Lexicography* 32:1, 92–114. https://doi.org/10.1093/ijl/ecy022 [last access: May 2, 2024].

Levitt, Theodore (1960): Marketing Myopia. In: *Harvard Business Review* 38, 24–47.

Lew, Robert (2010): Users Take Shortcuts: Navigating Dictionary Entries. In: Dykstra, Anna/Schoonheim, Tanneke (eds.): *Proceedings of the 14th Euralex International Congress*. Ljouwert: Afûk, 1121–1132.

Lew, Robert (2011): Studies in Dictionary Use: Recent Developments. In: *International Journal of Lexicography* 24:1, 1–4.

Lew, Robert (2012): How can we make electronic dictionaries more effective? In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 343–361.

Lew, Robert (2015): Dictionaries and Their Users. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45369-4_11-2 [last access: May 2, 2024].

Lew, Robert/Grzelak, Marcin/Leszkowicz, Mateusz (2013): How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye- Tracking Study. In: *Lexikos* 23, 228–254.

Lew, Robert/Tokarek, Patryk (2010): Entrymenus in bilingual electronic dictionaries. In: Granger, Sylviane/Paquot, Magali (eds.): *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-La-Neuve: Cahiers Du Cental, 193–202.

Lew, Robert/Wolfer, Sascha (2022): Predicting English Wiktionary Consulations. In: Klosa-Kückelhaus, Annette, et al. (eds.): *Dictionaries and Society. Book of Abstracts of the 20th EURALEX International Congress*. Mannheim: IDS-Verlag, 146–148.

Mayring, Philipp (2011): *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Beltz.

Mischel, Walter et al. (1972): Cognitive and attentional mechanisms in delay of gratification. In: *Journal of Personality and Social Psychology* 21:2, 204–218.

Müller-Spitzer, Carolin (2014) (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter.

Müller-Spitzer, Carolin (2018): Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. In: *Lexikos* 28, 287–315. https://doi.org/10.5788/28-1-1466 [last access: May 2, 2024].

Müller-Spitzer, Carolin/Koplenig, Alexander (2014): Online dictionaries: expectations and demands. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 143–188.

Müller-Spitzer, Carolin/Koplenig, Alexander/Wolfer, Sascha (2018): Dictionary usage research in the era of the Internet. In: Fuertes-Olivera, Pedro Antonio (Hrsg.): The Routledge Handbook of Lexicography. London et al.: Routledge, 715–734.

Müller-Spitzer, Carolin/Michaelis, Frank/Koplenig, Alexander (2014): Evaluation of a New Web Design for the Dictionary Portal OWID. An Attempt at Using Eye-Tracking Technology. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 207–228.

Müller-Spitzer, Carolin/Wolfer, Sascha/Koplenig, Alexander (2015): Observing Online Dictionary Users. Studies Using Wiktionary Logfiles. In: *International Journal of Lexicography* 28:1, 1–26.

Nesi, Hilary/Tan, Kim Hua (2011): The Effect Of Menus And Signposting On The Speed And Accuracy Of Sense Selection. In: *International Journal of Lexicography* 24:1, 79–96.

Popper, Karl (1994): *Alles Leben ist Problemlösen*. München: Piper.

Ptasznik, Bartosz/Wolfer, Sascha/Lew, Robert (2024): A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. In: *International Journal of Lexicography*, ecae011.

Rundell, Michael (2012): The road to automated lexicography: An editor's viewpoint. In: Granger, Sylviane/ Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 15–30.

Schadewaldt, Wolfgang (1949): *Schadewaldt-Denkschrift zum Goethe-Wörterbuch*. http://www.uni-tuebingen.de/gwb/denkschr.html [last access: May 2, 2024].

Sharifi, Saghar (2012): General Monolingual Persian Dictionaries and Their Users: A Case Study. In: Torjusen, Julie Marie/Fjeld, Ruth V. (eds.): *Proceedings of the 15th EURALEX International Congress 2012, Oslo, Norway, 7–11 August 2012*. Oslo: Universitetet i Oslo, 626–639.

Shoda, Yuichi/Mischel, Walter/Peake, Philip K. (1990): Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: identifying diagnostic conditions. In: *Developmental Psychology* 26:6, 978–986.

Sollaci, Luciana B./Pereira, Mauricio G. (2004): The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. In: *Journal of the Medical Library Association* 92:3, 364–371.

Tarp, Sven (2008): *Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography*. Tübingen: Niemeyer.

Tarp, Sven (2011): Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. In: Bergenholtz, Henning/Fuertes-Olivera, Pedro Antonio (eds.): e-*Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 54–70.

Tono, Yukio (2001): *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Tübingen: Niemeyer.

Tono, Yukio (2011): Application of Eye-Tracking in EFL Learners'. Dictionary Look-up Process Research. In: *International Journal of Lexicography* 24:1, 124–153.

Töpel, Antje (2014): Review of research into the use of electronic dictionaries. In: Müller-Spitzer, Carolin (ed.): *Using Online Dictionaries*. Berlin/Boston: De Gruyter, 145.

Trap-Jensen, Lars (2010): One, Two, Many: Customization and User Profiles in Internet Dictionaries. In: Dykstra, Anna/Schoonheim, Tanneke (eds.): *Proceedings of the 14th EURALEX International Congress*. Ljouwert: Afûk, 1133–1143.

Trochim, William (2006): *“Design”. Research Methods Knowledge Base*. http://www.socialresearchmethods.net/kb/design.php [last access: May 2, 2024].

Verlinde, Serge/Binon, Jean (2010): Monitoring Dictionary Use in the Electronic Age. In: Dykstra, Anna/Schoonheim, Tanneke (eds.): *Proceedings of the 14th Euralex International Congress*. Ljouwert: Afûk, 1144–1151.

Verlinde, Serge/Peeters, Geert (2012): Data access revisited: The Interactive Language Toolbox. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic lexicography*. Oxford: Oxford University Press, 147–162.

Welker, Herbert Andreas (2010): *Dictionary use: a general survey of empirical studies*. Brasília: self-publishing.

Welker, Herbert Andreas (2013): Empirical research into dictionary use since 1990. 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: De Gruyter, 531–540.

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

Wiegand, Herbert Ernst, et al. (2010) (eds.): *Wörterbuch zur Lexikographie und Wörterbuchforschung: mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neuen Sprachen*. Berlin/New York: De Gruyter.

Wolfer, Sascha, et al. (2014): Dictionary users do look up frequent and socially relevant words. Two log file analyses. In: Abel, Andrea/Vettori, Chiara/Ralli, Natascia (eds.): *Proceedings of the 16th EURALEX International Congress: The User in Focus*. Bolzano/Bozen, 281–290.

Wolfer, Sascha et al. (2016): The Effectiveness of Lexicographic Tools for Optimising Written L1-Texts. In: *International Journal of Lexicography* 31:1, 1–128.

Dictionaries

ANW = *Algemeen Nederlands Woordenboek*. Leiden: Instituut voor Nederlandse Lexicologie. www.anw.inl.nl [last access: May 2, 2024].

DWDS = *Das Digitale Wörterbuch der deutschen Sprache*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. 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. www.owid.de/elexiko_/index.html [last access: May 2, 2024].

GOETHE-WÖRTERBUCH = *Goethe-Wörterbuch*. Online abrufbar im Trierer Wörterbuchnetz: www.woerterbuchnetz.de/GWB/ [last access: May 2, 2024].

LEO = *LEO*. Sauerlach: LEO GmbH. www.leo.org/ [last access: May 2, 2024].

OWID = *Online-Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. www.owid.de [last access: May 2, 2024].

WIKTIONARY = *Das deutsche Wiktionary*. https://de.wiktionary.org/wiki/Wiktionary:Hauptseite [last access: May 2, 2024].

Internet sources

DeReKo = *Deutsches Referenzkorpus*. Mannheim: Institut für Deutsche Sprache. www1.ids-mannheim.de/kl/projekte/korpora/ [last access: May 2, 2024].

Forschungsgruppe Wahlen = *Politbarometer.* https://www.forschungsgruppe.de/Umfragen/Politbarometer/ [last access: May 2, 2024].

Wikipedia = *Wikipedia, die freie Enzyklopädie*. San Francisco, CA: Wikimedia Foundation. https://www.wikipedia.org [last access: May 2, 2024].

Images

Image 9.1 private.

Index

Open Access. © 2024 the author(s), published by De Gruyter. (cc) BY This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111233758-011