

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

Carolyn 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

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 Forschungsdesigns, 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.

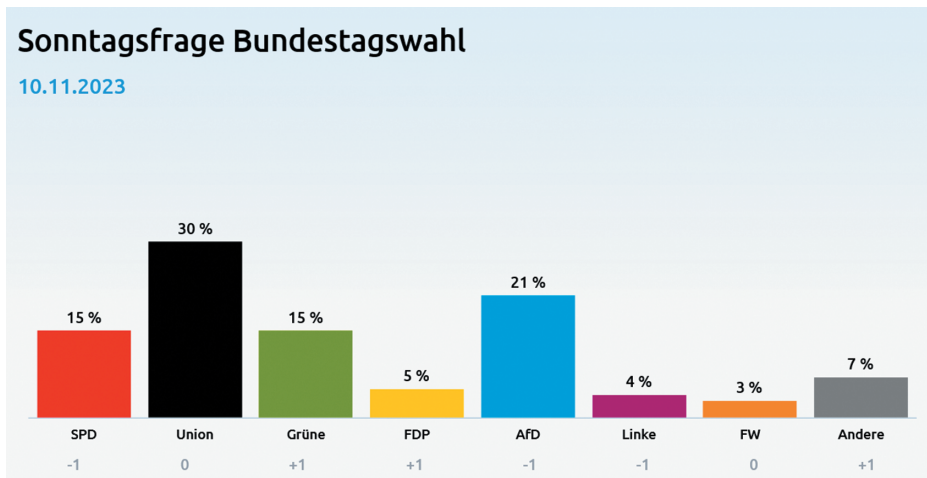


Fig. 9.2: Sample cross-sectional study: Sonntagsfrage – Deutschland 10.11.2023: infratest dimap for ARD-DeutschlandTREND.²

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.

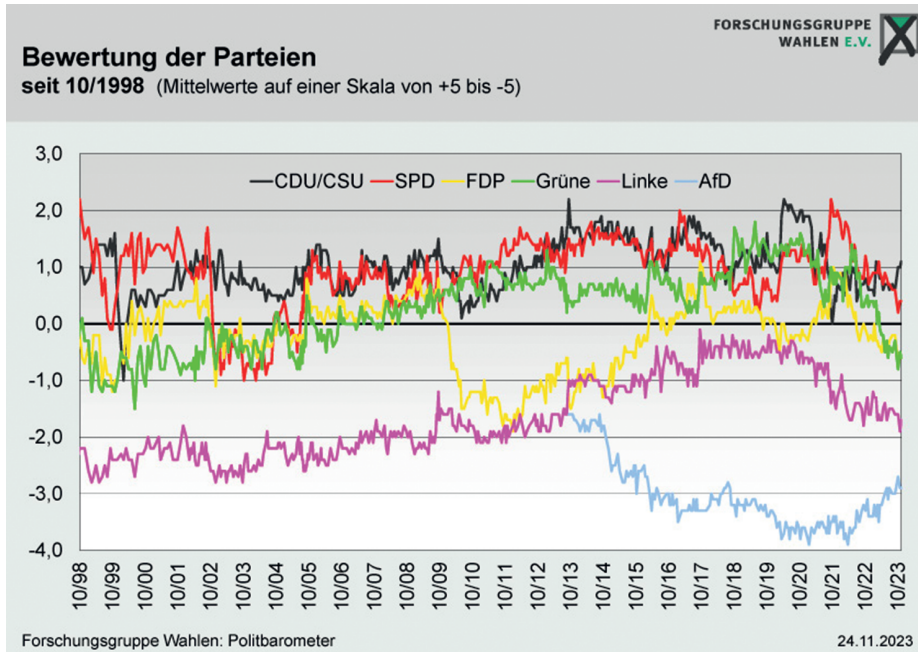


Fig. 9.3: Sample trend design; Forschungsgruppe Wahlen Politbarometer (24.11.23).³

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 (NEPS⁴) 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 ex-

ample, 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 2006⁶).

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 Chadeaux 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 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 Chadeaux (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 be 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 thought 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 show, 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.

Links to other dictionaries

Clarity


Multimedia content

Suggestions for further browsing

Accessibility

Up-to-date content

Adaptability



Reliability of content

Links to corpus

Speed

Next

Fig. 9.4: Ranking of the criteria in the online questionnaire.

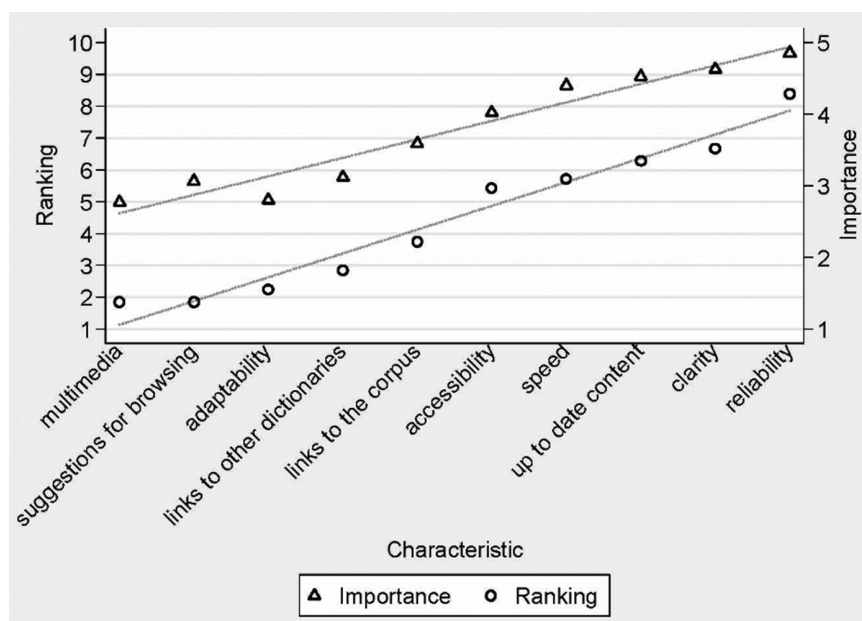


Fig. 9.5: Correlation between the mean rank and mean importance of criteria in the use of an Internet dictionary.

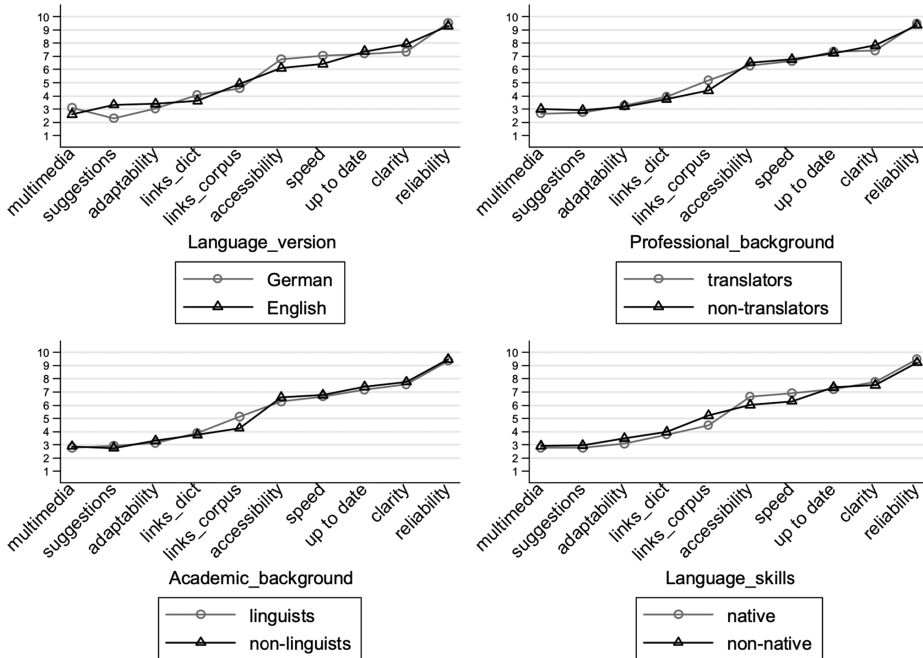


Fig. 9.6: Group-specific analyses of the rank orders.

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.

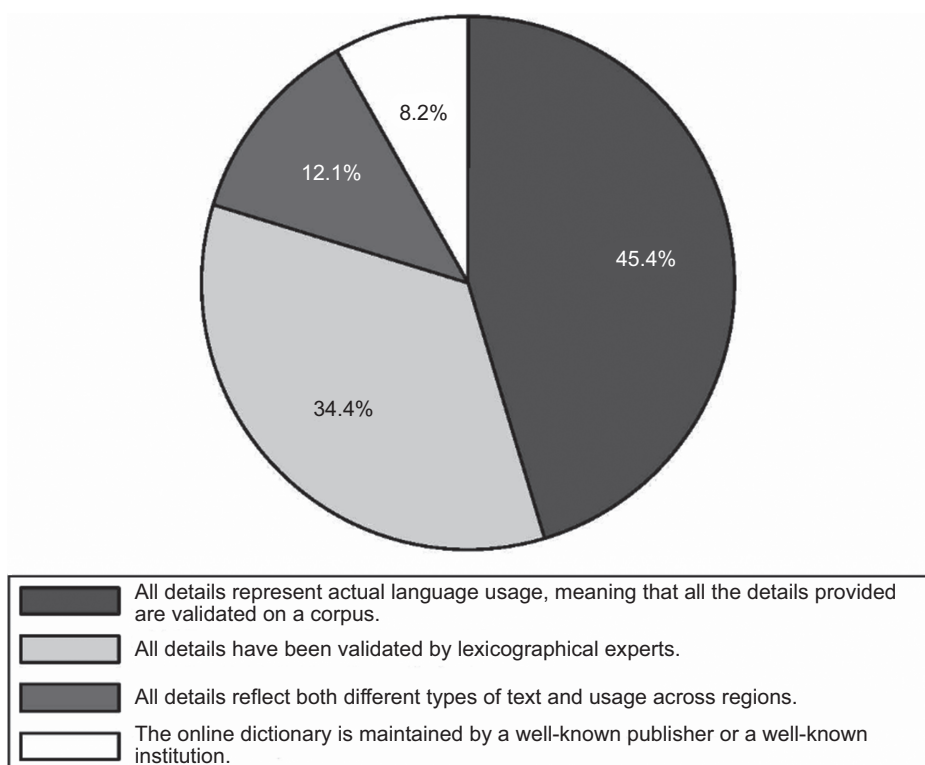


Fig. 9.7: Pie chart: aspects of content reliability.

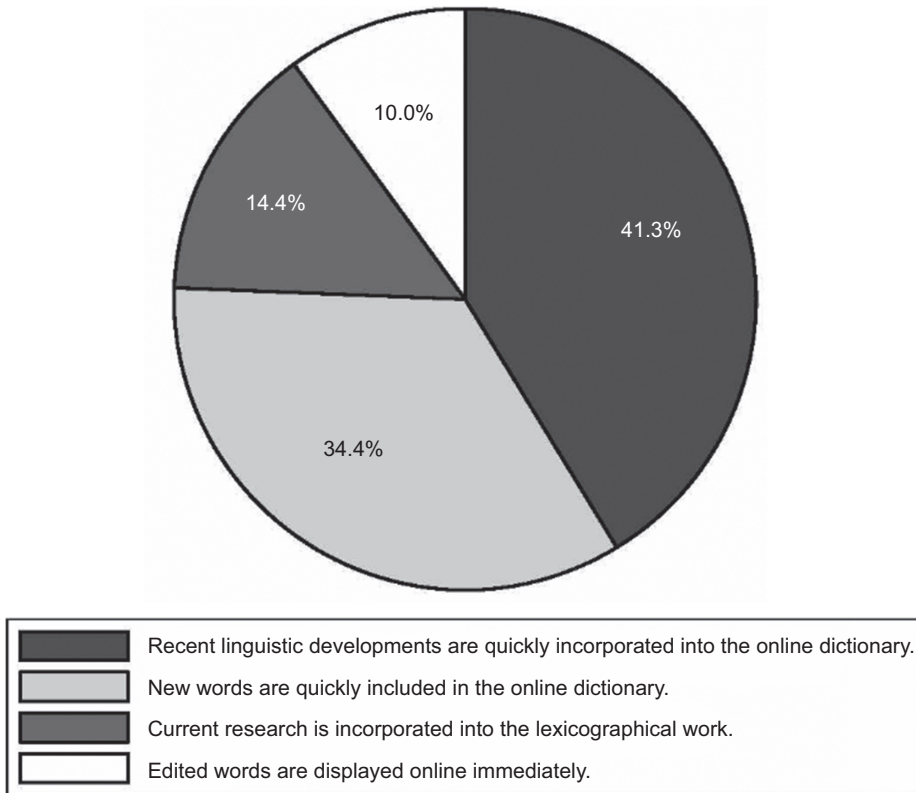


Fig. 9.8: Pie chart: aspects of dictionary updates.

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).

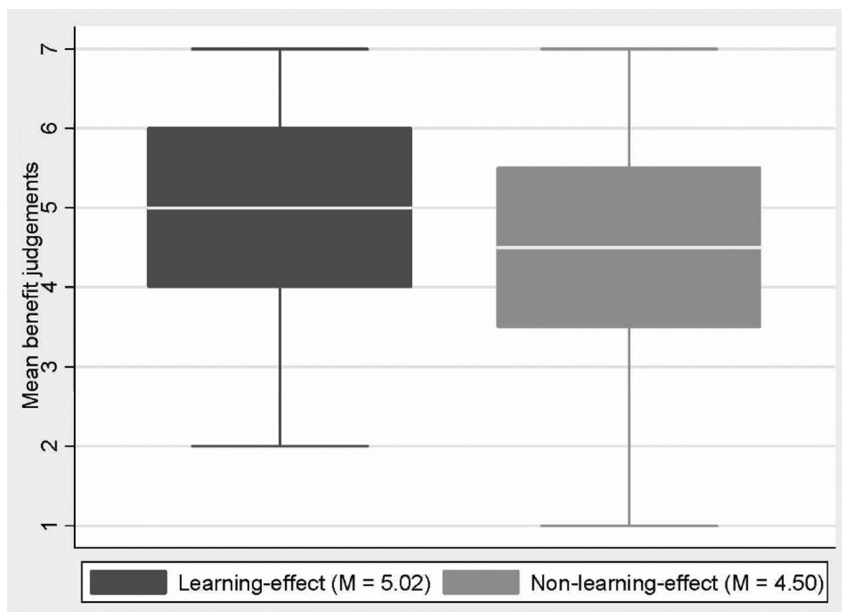


Fig. 9.9: Box plots: Evaluating multimedia and adaptive features depending on learning effect vs no learning effect.

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.

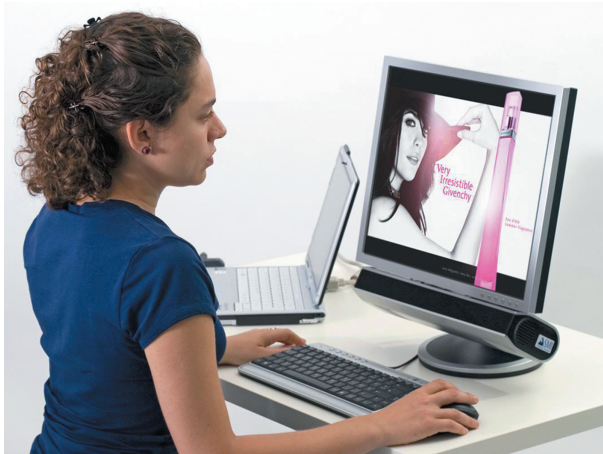


Fig. 9.10: PR image for the SMI Eye Tracker (<http://www.gizmag.com/smired500-500hz-remote-eye-tracker/16957/picture/124519/> [last access: June 10, 2016]).

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

Lesarten: 'Reittier'

Lesartenübergreifende Angaben

zur Übersichtsseite

Lesarten im Überblick

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 >

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.

Konstruktionen: Typische Verwendungen

Verwendungen mit Attribut verbergen x
 das Pferd als Arbeitstier
 Pferde im Galopp

Verwendungen in Verbalphrasen und Sätzen verbergen x
 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 verbergen x
 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.

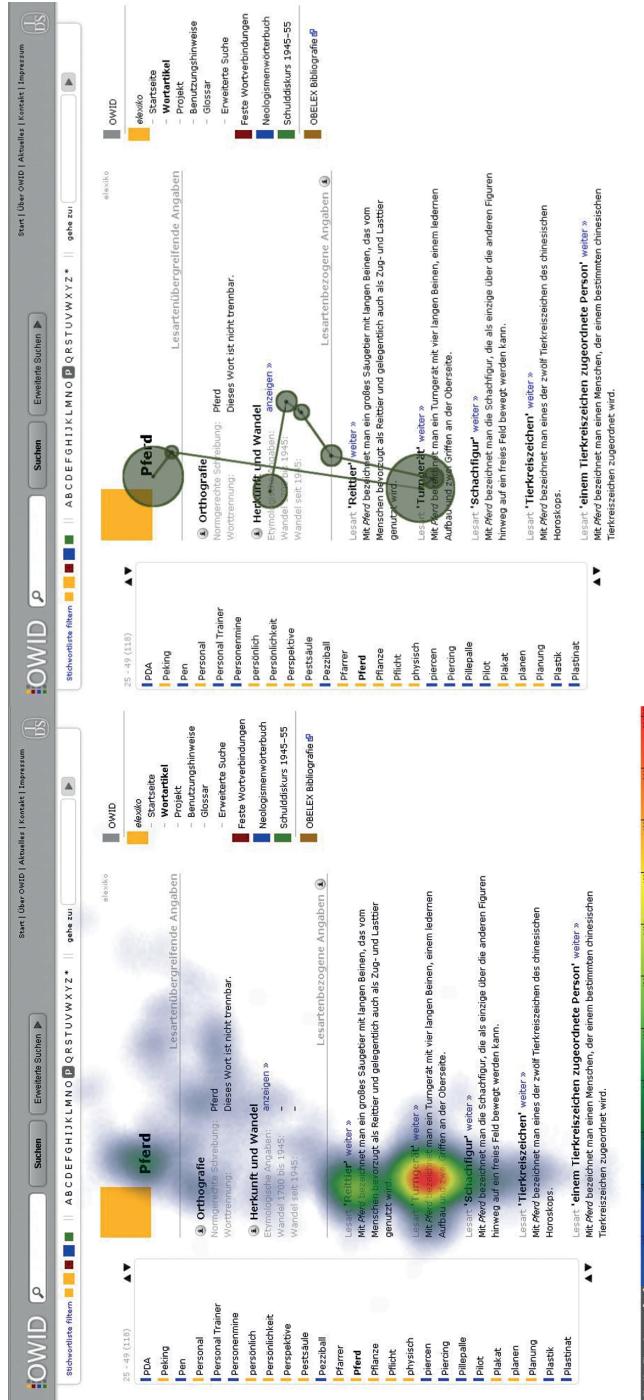


Fig. 9.12: Heat map of all of the participants (left); scan path of one participant (finding the individual meaning 'Turngerät'; right).

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.

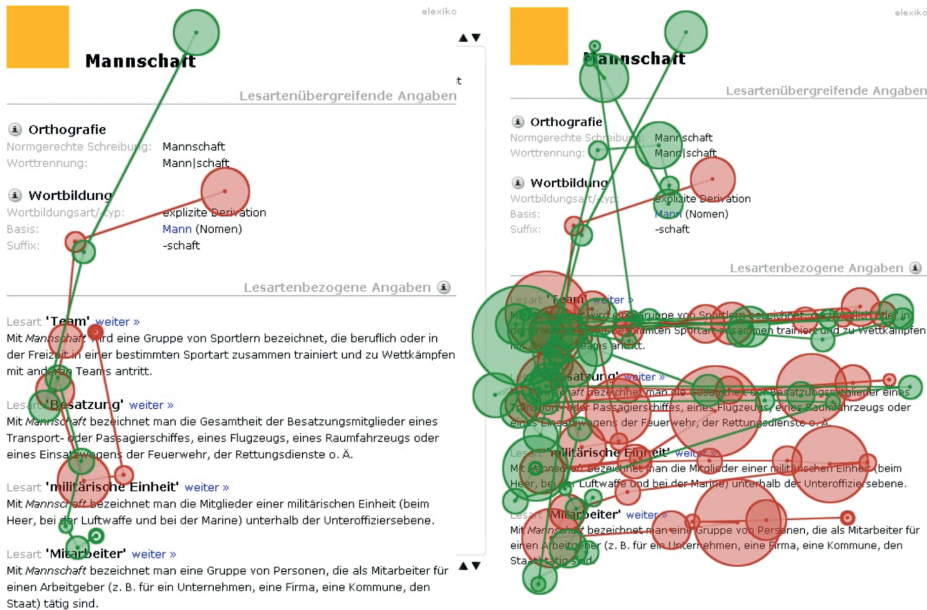


Fig. 9.13: Scan paths of two participants (stored on film); one snapshot at 00:01 seconds (left) and the other at 00:14 seconds (right).

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.

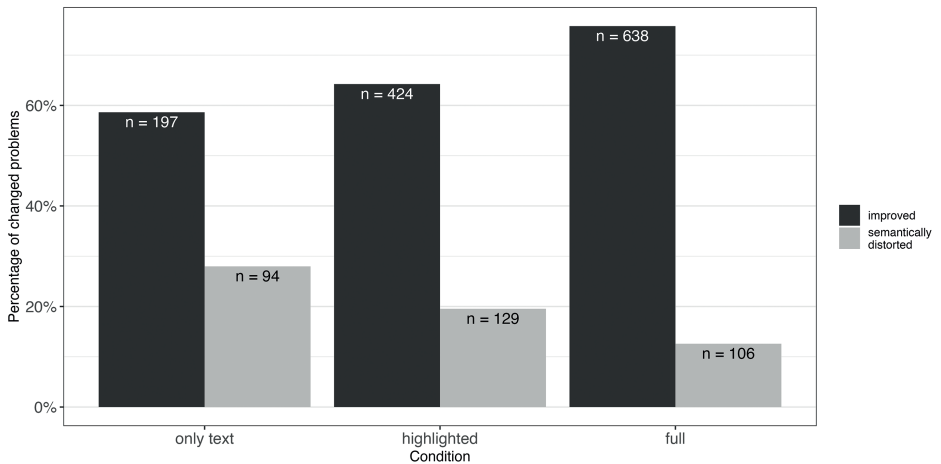


Fig. 9.14: Improved and semantically distorted problems under the three conditions. On the y-axis, the percentage of improved vs. semantically distorted problems is indicated (100% represents all problems). The figures in the bars give the raw number of linguistic problems for each category.

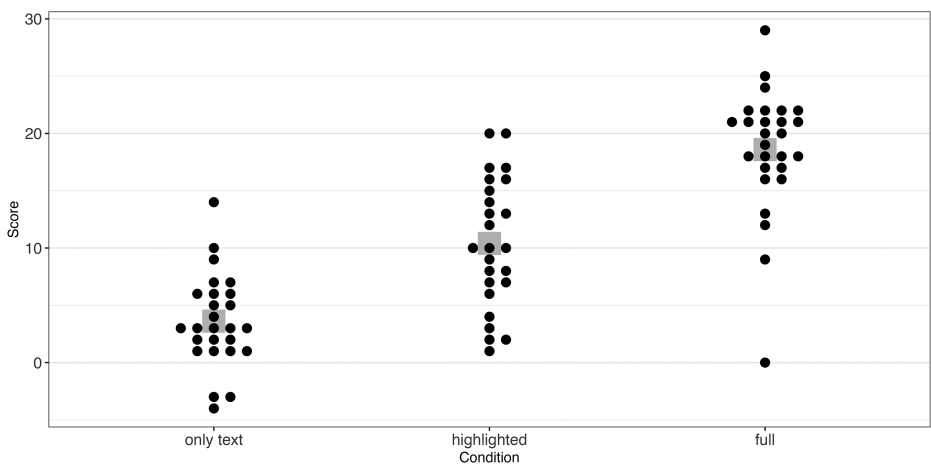


Fig. 9.15: Scores for all of the participants in the three experimental conditions. One black dot stands for one participant. Grey squares indicate the mean values of the experimental conditions.

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. Kopenig 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 “per one million” 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 be-

tween them are not actually comparable. → Tab. 9.1 summarises the results of this analysis of our log files.

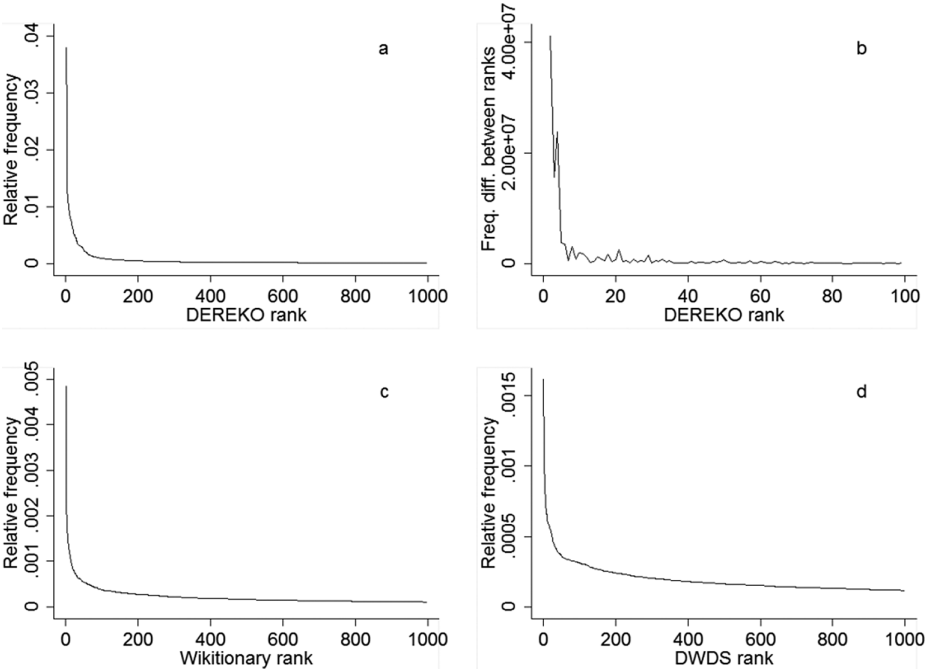


Fig. 9.16: Distribution of corpus and log file data (from the DEREKO and WIKTIONARY/DWDS) as examples of a Zipf distribution (Koplenig et al. 2014: 238).

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 *Tugendfurore* ‘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-

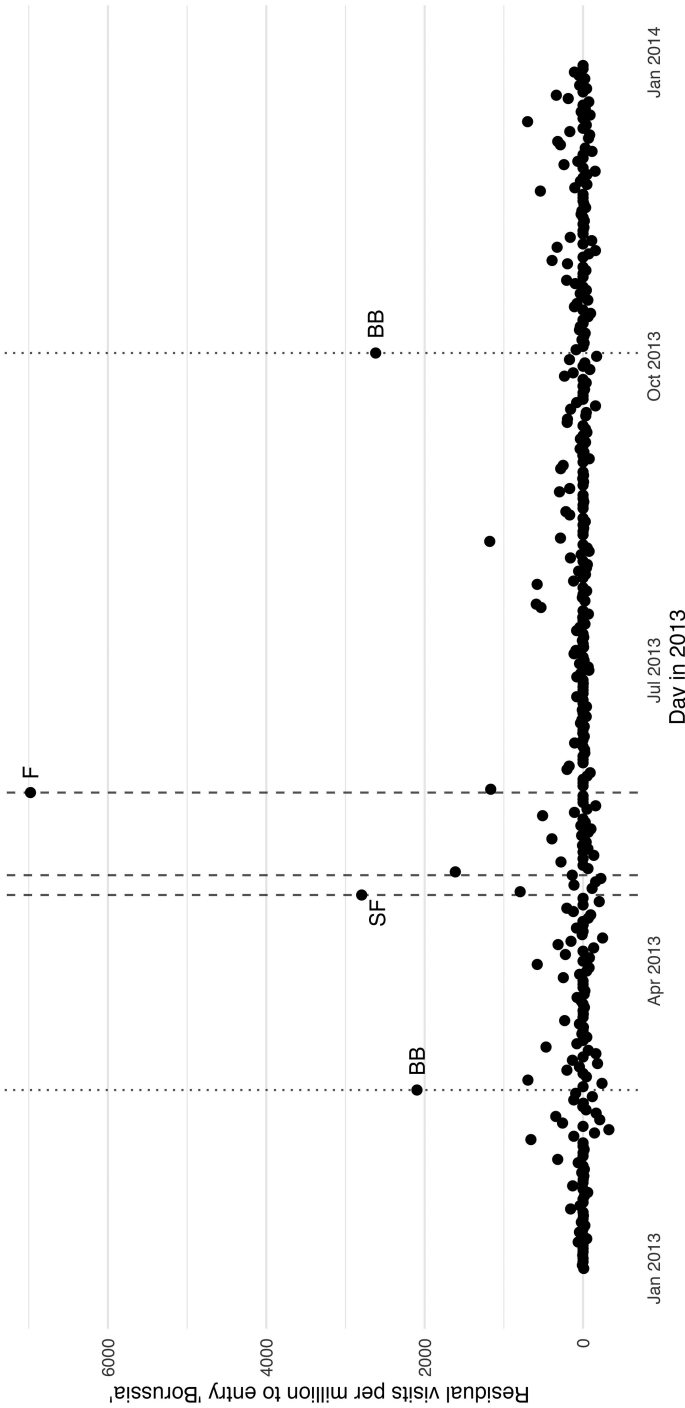


Fig. 9.17: Access to the word *Borussia* cleaned of trends (Jan.–Dec. 2013). Dotted vertical lines (BB) mark football matches between the German teams Borussia Mönchengladbach and Borussia Dortmund. Dashed lines indicate the Champions League 2013 semi-final matches (SF) and the final match (F).

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 Ezquerro, 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 Lexicography*. 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: Torjusén, 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/lec022> [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/Leszczkiewicz, 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): Entry menus 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 Consultations. 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: Afök, 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.