Erschienen in: Erschienen in: Müller-Spitzer, Carolin (Hrsg.): Using Online Dictionaries. Berlin/Boston: de Gruyter, 2014. (Lexicographica: Series Maior 145), S. 55-76.

Alexander Koplenig

Empirical research into dictionary use

A brief guide

Abstract: This chapter summarizes the typical steps of an empirical investigation. Every step is illustrated using examples from our research project into online dictionary use or other relevant studies. This chapter does not claim to contain anything new, but presents a brief guideline for lexicographical researchers who are interested in conducting their own empirical research.

Keywords: research question, operationalization, research design, methods of data collection, data analysis

Alexander Koplenig: Institut für Deutsche Sprache, R 5, 6-13, D-68161 Mannheim, +49-(0)621-1581435, koplenig@ids-mannheim.de

1 Introduction

On the subject of the methodology of user studies in the context of dictionary research, Lew (2011, p. 8) argues that

> "[u]ser studies can answer a number of questions that are relevant to (mostly) practical lexicography. However, to be maximally useful, researchers need to be really careful about the exact form of the question they actually want to ask. Having settled on this part, they need to think long and hard about what are the best possible means to tackle the specific questions that they want answered."

To "tackle the specific questions" (for example about how online dictionaries are actually being used or how they could be made more user-friendly), many researchers have called for a more intense focus on empirical research (Atkins & Varantola, 1997; Hartmann, 2000; Hulstijn & Atkins, 1998). When referring to empirical social research, Hartmann (1987, p. 155, 1989, pp. 106–107), Ripfel & Wiegand (1988, pp. 493–520), Tono (1998, pp. 102–105) and Zöfgen (1994, p. 39 seqq.) list experiments and surveys as distinct methods of dictionary usage research. However, as I will try to show below, these authors seem to mix up two distinct elements of empirical research that it is important to distinguish: on the one hand the research design, and on the other hand the instrument of data collection. In this chapter, I will therefore try to describe the typical steps in an empirical investigation as defined by Diekmann (2002), Babbie (2008) and Trochim & Land (1982) that seem to be im-

portant for empirical research into dictionary usage. Every step will be illustrated using examples from our research project or other relevant studies.

I hope that this brief guide will be of some help – in Lew's words – in maximizing the usefulness of user studies in dictionary research by helping the investigator to answer the following questions:

- What is the relationship of interest? (cf. Section 2.1)
- How will the variables involved be measured? (cf. Section 2.2)
- What type of design is the most appropriate for collecting the data? (cf. Section 2.3.1)
- What kind of structure is best suited for answering the research question? (cf. Section 2.3.2)
- How should the data be collected? (cf. Section 2.3.3)
- How should the collected data be analyzed? (cf. Section 2.4)

2 Research Methodology

The following five steps on how to conduct an empirical study are closely based on Diekmann's guidelines (2002, pp. 161–191).

Formulating the Research Question

It may seem trivial, but nevertheless it is worth mentioning: each (empirical) research project starts with the formulation of a question. If there is no question, then there is nothing to research. Popper (1972) most notably demonstrated this by asking his audience to "observe". Of course, the only logical reply to his demand is "what?" (or "why", or "when", or "how"). The better the research question is articulated, the easier all subsequent steps will be. In relation to dictionary usage research, it is also first of all necessary to clarify what exactly the focus of interest is.

On the one hand, framing a research question can also mean specifying the hypotheses to be tested in the study. Ideally, the researcher already knows at this point which variables are dependent and which ones are independent (Diekmann, 2002, p. 162; cf. Example 1).

Example 1
In our project, we tried to find out whether different user groups have different preferences regarding the use of an online dictionary (Koplenig, 2011). This means we were interested in the influence of a person's background on his/her individual preference. In this case, the dependent variable is the preference relation of the user, whereas professional background and academic background serve as independent variables.

Example 1 highlights an important point: if the researcher is interested in the influence of one or more background variables on a response variable, it is essential that all the information needed to answer the question is collected. In this case, using log files as "the electronic trace of the dictionary user's consultation behaviour" (Lew, 2011, p. 6) does not seem to be a promising research strategy (cf. the following sections for further arguments supporting this view).

On the other hand, if the general aim of the research is to gain an initial insight into a topic, no hypotheses can be formed *ex ante*. In this situation, the purpose of the investigation is to develop hypotheses by exploring a field (Diekmann, 2002, p. 163, cf. Example 2).

Example 2
As Tarp (2009a) has pointed out, the "close relation between specific types of social needs and the solutions given by means of dictionaries" (Tarp, 2009a, p. 19) have not yet been thoroughly investigated. To explore the contexts of dictionary use, we decided to include an open-ended question in our first survey: *In which contexts or situations would you use a dictionary? Please use the field below to answer this question by providing as much information as possible.* (cf. Müller-Spitzer: Contexts of dictionary use, this volume).

After the theoretical concept of the investigation has been decided, the researcher needs to decide how to measure this concept. This is called operationalization.

Operationalization

Testing a hypothesis usually means that the researcher first has to clarify how to measure the variables involved (Babbie, 2008, p. 46; Diekmann, 2002, p. 168). Hartmann (1989, p. 103) hypothesizes that "[d]ifferent user groups have different needs", and therefore ,"[t]he design of any dictionary cannot be considered realistic unless it takes into account the likely needs of various users in various situations" (Hartmann, 1989, p. 104). But what is a user group (Wiegand et al., 2010, p. 678)? For example, if it is assumed that the groups are determined by classical sociodemographical variables such as gender or age, the operationalization is easy. But it is also a reasonable assumption that relevant group variables in this context could be the professional or academic background (cf. Example 1) or the usage experience.

However, it is not clear a priori what is meant by an experienced dictionary user. So if one of the research hypotheses states that the amount of experience in dictionary use is a determinant of a successful dictionary consultation, it is first of all necessary to define how experience is measured in this context (cf. Example 3).

Example 3
In our project, we asked the respondents to one of our surveys to estimate on how many days per week they use online dictionaries (0 – 7). This estimation served as a proxy for dictionary usage experience: we assumed that people who use a dictionary every day (7) are on average more experienced than people who use the dictionary less often.

Our operationalization of dictionary usage experience is rather simple (for a different approach, see Wiegand, 1998, p. 506 seqq.). Whether it fully captures the essential nature of experience in this context is open to debate. Indeed, in our analysis of the data, we found no correlation between this simple proxy and user preferences. Unfortunately, we cannot deduce from this result that in reality there is no correlation between these two variables, because it is (maybe even more) likely that our operationalization of usage experience was not successful. We might therefore have obtained a different result if we had used another operationalization. In most cases, it is not possible to avoid this problem completely, but using multiple indicators for one construct whenever possible is a reasonable strategy. If several indicators point in the same direction, the convergent validity of the construct increases and as a result the problem levels off (cf. Example 4).

Example 4
To identify different user demands (cf. Example 1), we decided to ask the respondents both to rate ten aspects of usability regarding the use of an online dictionary, and to create a personal ranking of those aspects according to importance. Analysis of (Spearman's rank) correlation revealed a significant association between importance and ranking. At this point, we were fairly confident that it would be possible to use the individual ranking as a reliable indicator of users' demands.

After the meanings of all the variables involved have been defined, that is, operationalized, the researcher has to decide on the mode of study.

Research Design

Diekmann (2002) argues that 'the function of research designs is to provide meaningful data'[1] (Diekmann, 2002, p. 274). Any research design has two dimensions (Diekmann, 2002, pp. 267–304):
– A temporal dimension (cf. Section 2.3.1)
– A methodological dimension (cf. Section 2.3.2).

1 „Erhebungsdesigns sind Mittel zum Zweck der Sammlung aussagekräftiger Daten".

The design type is concerned with the temporal dimension of the research, while the methodological dimension of a research design affects the control of variance. Both dimensions will now be briefly outlined.

Research Design Type

In general, there a three distinct classes of design types:

(1) Cross-sectional design
(2) Trend design
(3) Panel design

A cross-sectional design refers to a one-dimensional process of data collection. This means collecting the data of a sample of a number of subjects at the same point in time. On this basis, it is not feasible to measure (intra-individual) change over time with a cross-sectional design without adjusting the research design accordingly (cf. Section 2.3.2). With this type of design, it is only possible to compare different entities, such as subjects, at one moment in time (cf. Example 5).

Example 5
All of the four online surveys carried out throughout our project were designed as cross-section surveys. In each survey, our subjects were asked to answer a questionnaire. We then used the collected data to compare subjects with different characteristics, for instance, those who work as translators and those who do not.

If a researcher is interested in change over time, it is more appropriate to use a longitudinal design. Both trends and panels are longitudinal designs. Correspondingly, a trend design is like a cross-sectional design with more than one temporal dimension. This means collecting the data of different samples of subjects at several points in time. By aggregating the data, it is possible to observe temporal changes (Diekmann, 2002, p. 268). An example of this type of design is the study by De Schryver & Joffe (2004). The authors analyze log files and argue that

> "[w]ith specific reference to a Sesotho sa Leboa dictionary, it was indicated that the general trend during the first six months has been one of a growing number of lookups by growing number of users."(De Schryver & Joffe, 2004, p. 194)

However, it is debatable whether it is adequate to classify this study as a pure trend design, because De Schryver & Joffe are also able to collect data on the individual level. Thus they are also able to draw conclusions about individual users:

> "While the distribution of the number of lookups per visitor is Zipfian, most visitors tend to look up frequent items on the one hand, and sexual/offensive items on the other" (De Schryver & Joffe, 2004, p. 194)

Statements on this level are typical of a panel design. This means collecting the data of one sample of subjects at several points in time. By measuring the same variables for the same individuals or units at multiple points in time, it becomes possible to model change on the individual or the unit level, in contrast to a cross-sectional design (Diekmann, 2002, p. 267).

One objection against categorizing the investigation by De Schryver & Joffe (2004) as a panel design is raised by the authors:

> "One can also not distinguish between multiple users who share a computer, or determine when a single user has made use of multiple computers (e.g., a student who uses a computer lab). Nonetheless, the technique is reliable in the majority of cases, providing an error margin of probably not more than 15%." (De Schryver & Joffe, 2004, p. 194)

In contrast to De Schryver and Joffe (2004), it can be argued that this is a strong methodological objection against drawing any conclusions on the individual level. An error margin of 15% is problematic in itself, especially because it is reasonable to assume that this error is non-random: imagine a public institution such a library or a school where the visitors can use the dictionary (Bergenholtz & Johnson, 2005, p. 125). This institution would be classified as one particularly heavy user. This could, in turn, lead to a systematic over-estimation of heavy users.

Panel designs could also be interesting for research into dictionary use, because dictionary skills are very likely to develop over time. Lew (2011) addresses this question:

> "As users work with a dictionary over time, they learn some of the structure, conventions; the learn how to cut corners. Humans exhibit a natural and generally healthy cognitive tendency to economize on the amount of attention assigned to a task at hand. So in the course of interaction with dictionaries, users' habits adjust, and their reference skills evolve." (Lew, 2011, p. 3)

Furthermore, it would be interesting to investigate what kind of lexical information L2 learners actually look up in a dictionary at successive moments in time, because it is reasonable to assume that with growing language skills, dictionary users have different needs. In order to investigate the influence of the language acquisition process on users' needs, a sample of fresh L2 learners could be drawn and their look up processes measured at several points in time.

Research Design Structure

Creating the structure of the investigation means deciding how the units of interest will be assigned to the categories of the independent variables (Diekmann, 2002, p. 289). In this section, three common design structures will be presented and discussed:

(1) Experimental designs
(2) Quasi-experimental designs
(3) Ex-post-facto designs

Experimental Design

Three conditions need to be satisfied for an experimental design (Diekmann, 2002, p. 291):

(1) At least two experimental conditions are formed: one treatment group and one control group.
(2) The respondents (or the units of interest) are randomly assigned to either the treatment group or the control group.
(3) The independent variables are manipulated by the researcher.

Conducting an experiment is the best way of making causal interferences, because this design structure guarantees internal validity (Campbell & Stanley, 1966; Pearl, 2009). This can be best understood in terms of placebo-controlled medical trials, where the respondents in the control group receive a treatment without any active substance. The measured effect in this group is then compared with the respondents in the experimental group who received an actual treatment (Shang et al., 2005). Randomly assigning the units of interest to the experimental conditions eliminates the selection bias, which is the potential influence of confounding variables on an outcome of interest. Balancing the "subject's characteristics across the different treatment groups" (Angrist & Pischke, 2008, p. 14) ensures that the experimental condition and all possible (even unidentified) variables that could affect the outcome are uncorrelated. Through the manipulation of the independent variables, it can be inferred that the treatment is the cause of the outcome, random fluctuations aside (Diekmann, 2002, p. 297): if the effect in the treatment group differs significantly (either positively or negatively) from the effect in the control group, then the only logical explanation for this is a causal effect of treatment on outcome. Therefore, I believe that Angrist and Pischke (2008) are right to say that "[t]he most credible and influential research designs use random assignment. "(Angrist & Pischke, 2008, p. 9).

In dictionary usage research, this paradigm has been fruitfully applied in several studies, such as Lew (2010) and Dziemianko (2010). Example 6 illustrates one of our experimental approaches.

Example 6
In Müller-Spitzer & Koplenig (Expectations and demands, this volume), we argue that when the users of online dictionaries are thoroughly informed about possible multimedia and adaptable features, they will come to judge these characteristics to

be more useful than users that do not have that information. To test this assumption, we included an experimental element in our second survey: participants in the treatment condition were first presented with examples of multimedia and adaptable features. After that, they were asked to indicate their opinion about the application of multimedia and adaptable features in online dictionaries. Participants in the control group first had to answer the questions regarding the usefulness of multimedia and adaptable features followed by the presentation of the actual examples. As predicted, the results revealed a learning effect. This effect turned out to be modest in size (about a half a point on a 7-point scale), but highly significant.

By direct analogy with social research, real controlled experimental trials are often not feasible in dictionary usage research. The reason for this is quite simple: imagine a researcher, who is trying to ascertain whether the dictionary look up process (dependent variable) is determined by the language skills of a respondent (independent variable). For reasons of simplification, let us assume that the researcher believes that native speakers differ from L2 speakers of a language. Conducting an experiment in this situation would involve the random assignment of the participants to one of the two experimental conditions. Of course this is not possible, as potential respondents either are native speakers or are not (Diekmann, 2002, p. 303). Subsequently, the researcher would not be able to eliminate the fundamental problem of selection bias: for instance, it is quite likely that the native speakers would be better at understanding the experimental instructions. And this advantage, in turn, could affect the look up process. Similar instances could be multiplied. Thus, we cannot infer, on logical grounds, a causal effect of language skills on the look up process.[2]

Two alternative design structures will be presented in the following two sections.

2 The problem of selection bias is also important in the case of log file investigations. If a research project is directed at the "exact needs of the users" (Bergenholtz & Johnson, 2005, p. 117) it must be borne in mind that the there is an error that is – again – non-random (cf. Section 2.3.1): the sample only includes data for people who use (or have used) the dictionary. Take for instance the following hypothetical but plausible situation (cf. Koplenig, 2011): Alex does not know the spelling of a particular word. To solve this problem, he visits an online dictionary. However, when trying to find the search window, he stumbles across various types of innovative buttons, hyperlinks and other distracting features. Instead of further using this online dictionary, he decides to switch to a well-known search engine, because he prefers websites that enable him to find the information he needs easily. In this example, there would not be any data to log (except for an unspecified and discontinued visit on the website). Therefore, the external validity of the investigation can be questioned (Diekmann, 2002, pp. 301–302).

Non-experimental Design

Quasi-experimental design
In principle, quasi-experimental designs are experiments without the random assignments to the experimental conditions. Typical examples are the evaluation of the effects of specific actions, such as political or legal reforms or social interventions (Diekmann, 2002, p. 309). In those contexts, the variables of interest are measured before and after the implementation of the action. The difference between the two measurements represents the effect induced by the action.

In dictionary usage research, a quasi-experimental design could be applied to measure the effectiveness of new dictionary features. For example, the researcher might wish to know whether the implementation of an error-tolerant search function makes the dictionary more user-friendly. One could measure how many look ups are successful before and after the implementation of this feature. The difference could be considered to be the usefulness of the feature.

Ex-post-facto design
An ex-post-facto group design can be classified as a research design both without the random assignments to the experimental conditions, and without the manipulation of the independent variables by the researcher. In fact, groups are compared because of shared differences that exist prior to the investigation. As a result, the formation of groups is independent of the research design. In this case, the comparison group is not equivalent to the control group in an experimental design. In social research, this type of design is very common. Typical examples are the influence of socio-economic or socio-demographic factors on various types of outcomes, such as educational achievement or occupational career.

Background factors of this kind that could affect the use of dictionaries and the look up process, could be the language skills of the user (cf. the example given at the end of Section 3.2.2.1), as well as his/her academic or professional background.

Example 7
An extension of Example 6 demonstrates that even an experimental design does not replace a careful examination of the collected data: a closer ex-post-facto inspection of the data showed that the effect mentioned in Example 6 is mediated by linguistic background and the language version of the questionnaire: while there is a significant learning effect in the German version of the questionnaire but only for non-linguists, there is a highly significant learning effect in the English version of the questionnaire but only for linguists.

Indeed, this type of design was very important in our project, as we tried to find out whether different user groups have different preferences regarding the use of an online dictionary:

> "[m]ore specifically we need to ask: Should a user (i.e. while using a dictionary) create a profile at the beginning of a session (e.g. user type: nonnative speaker, situation of use: reception of a text) and should s/he navigate in all articles with this profile?" (Müller-Spitzer & Möhrs, 2008, pp. 44–45)

This is an example of an important lexicographical question that seems hard to answer using log file analyses alone. Hartmann (1989) hypothesizes that "[d]ifferent user groups have different needs" (Hartmann, 1989, p. 103), therefore , "[t]he design of any dictionary cannot be considered realistic unless it takes into account the likely needs of various users in various situations" (Hartmann, 1989, p. 104). Of course, log files do not contain information about the individual dictionary user, such as his or her academic background, age, usage experience and language skills, but it is reasonable to assume that these factors influence the dictionary usage process.

Since the "user type" is a precondition that is – of course – not determined by the investigator, we applied an ex-post-facto design (cf. Example 1, Example 7). Verlinde & Binon (2010) argue in this context:

> "[I]t will almost be impossible to conceive smart adaptive interfaces for dictionaries, unless more detailed data combining tracking data and other information as age or language level for instance, would eventually infirm this conclusion." (Verlinde & Binon, 2010, p. 1150)

It is important to bear in mind that as a result of the missing randomization in both quasi-experimental and ex-post-facto design, the problems of selection bias and confounding variables cannot be solved. This is why, in principle, both types of design permit no causal interpretations.[3]

As I noted at the beginning of this section, it is important to distinguish between the research design and the instrument of data collection. In the next paragraph, I will explain why.

3 In recent years, several refined strategies have been proposed to approach this problem, for example matching, instrumental variables, difference-in-difference designs, regression-discontinuity designs or quantile regressions (cf. Angrist & Pischke, 2008). None of these models will ultimately overcome the shortcomings of non-experimental data. Nevertheless they prove to be a valuable basis for cautious (counterfactual) causal reasoning without experimental data (Pearl, 2009). By all means, it is important to control for variables that are assumed to be correlated with the relationship of interest.

Data collection

In principle, data collection means any systematic method of gathering the information needed to answer the research question on the basis research design. Following Kellehear (1993), Diekmann (2002) and Trochim (2006), I distinguish between obtrusive (sometimes referred to as reactive) and unobtrusive methods of data collection.

In general, an unobtrusive method can be understood as a method of data collection without the knowledge of the participant or the unit-of-interest. In contrast, obtrusive measurement means that the researcher has "to intrude in the research context" (Trochim, 2006).

As interviews or laboratory tests are also social interactions between the respondents and the researcher, respondents tend to present themselves in a favorable light. This is called social desirability (Diekmann, 2002, pp. 382–386). Furthermore, filling out a questionnaire or taking part in a laboratory test can be exhausting or boring, which can also lead to biased results. Zwane et. al. (2011) even present (field-)experimental evidence that under certain circumstances, participation in a survey can change subsequent behavior:

> "Methodologically, our results suggest that researchers should rely on the use of unobtrusive data collection when possible and consider the tradeoffs between potential biases introduced from surveying and the benefits from having baseline data to identify heterogeneous treatment effects not possible to estimate without implementation of a baseline survey." (Zwane et al., 2011, p. 1824)

Thus, the great advantage of unobtrusive methods is that the measurement does not influence what is being measured. Without the knowledge of a participant, it is possible to measure his or her "actual behaviour as opposed to self-reported behaviour" (Kellehear, 1993, p. 5). At the same time, this strength is also the biggest limitation of the method, because the researcher loses much of the control of the research process. Whenever the researcher needs to collect information about background factors assumed to influence the outcome of interest, e.g. the user type (cf. Section 2.3.2.1-2), s/he must accept that:

> "[f]or some constructs there may simply not be any available unobtrusive measures." (Trochim, 2006; regarding dictionary usage research, cf. Wiegand, 1998, p. 574)

Consequently, the question "what is better: unobtrusive or obtrusive methods?" cannot be answered in a meaningful way, since the answer always depends on the concrete research question. Whenever possible, it is best to combine both types of method, in order to increase both the reliability and the validity of the results.

Surveys

Surveys – whether administered by means of a questionnaire or an interview – involve collecting the data by asking questions. In Müller-Spitzer, Koplenig & Töpel (2012: 449f) we argue that the critique of Bergenholtz & Johnson (2005) regarding the usefulness of conducting a survey for empirical research into dictionary usage is inadequate, because it is based on a somewhat biased picture of this method. Thus the examples in Bergenholtz & Johnsen (2005, pp. 119–120) only provide good examples of how a questionnaire should not be prepared. For example, the first question ("Under which headword would you look for the following collocations?") implies that every respondent knows the definition of 'collocation', which is certainly not the case. Furthermore, a cleverly designed survey neither rests on the assumption "that the informants remember exactly how they have used dictionaries in the past", nor expects the respondent to "be able to predict how they will do it in the future" (Bergenholtz & Johnson, 2005, pp. 119–120), but uses proxies to measure the construct of interest (cf. Krosnick, 1999). Survey methods only deliver reliable information if the survey is constructed in a comprehensible and precise way. Accordingly, there is a special branch of the social sciences the aim of which is to evaluate the quality of survey questions and identify potential flaws experimentally (cf. Madans, Miller, Maitland, & Willis, 2011).

The critique of Tarp (2009b) falls prey to the same sort of counter-argument: the problem is not the method but its application. Tarp argues that

> "many lexicographers still carry out user research by means of questionnaires, arriving at conclusions which even a modest sociological knowledge would show to have no scientific warranty." (2009b, p. 285)

I am quite certain that many scientists with a "modest sociological knowledge" would question the validity of this argument, because its premise is false, since it is based on a biased description of the method. Let me give you two examples:

Instead of just asking "how do judge your own medical ability" Das and Hammer (2005; cf. Banerjee & Duflo, 2011, pp. 52–53) constructed five test scenarios ("vignettes") of hypothetical patients with different symptoms, each containing several questions, to measure the quality of doctors in Delhi, India. The vignettes were presented to a random sample of 205 local doctors. In principle, the competence of each doctor was measured by comparing the responses of the participants with the "ideal" responses. Even if this was not a real situation in Tarp's terms (2009b, p. 285), the findings plausibly show that the quality of doctors in poorer neighborhoods is significantly lower than that of those in richer neighborhoods.

Instead of just asking "Which of the following alternatives is best suited to capture lexicographical information about sense-relations?", we constructed a multilevel test scenario in our third study to evaluate how well users understood the ter-

minology of the user interface of *elexiko* (cf. Klosa et al., this volume). In *elexiko*, the sense-related information is structured into tabs. We wanted to find out whether the labels on the tabs were easy to understand. Or in other words, given that a user needs a specific type of information, for example a synonym, does s/he know which tab to click on and, if not, are there better labels (which are more user-friendly)? Therefore, for every type of information, several different types of labels were prepared. For example, the following four labels were prepared for the sense-related information:

- Synonyme und mehr ("synonyms and more")
- Sinnverwandte Wörter ("sense-related words")
- Wortbeziehungen ("word-relations")
- Paradigmatik ("paradigmatics")

Amongst other things, the participants were presented with different questions, such as: "Imagine the following situation: you are writing a text. Because you do not want to use the same word every time, you are trying to find an alternative for the word address. Please click on the item, where you think you would find the information you are looking for." Each participant answered two of these vignettes (for two different kinds of information). For each vignette, the participant (randomly) received one of 25 different combinations of differently labeled tabs. In principle, the quality of label was measured by relative frequency of correct clicks. For example, our results show that "paradigmatics" is not really an appropriate label: only 8.33% of our participants (N = 685) were able to answer the question correctly, if this label was chosen, whereas both "synonyms and more" (100.00%) and "synonymous words" (92.59%) proved to be more successful. The information gathered was used to rename the tabs in *elexiko* accordingly (if necessary). Again, these results are not based on a "real usage situation", but they show that questionnaires can be applied in a fruitful way to empirical research into dictionary usage.

Apart from that, I believe that even if Tarp's premises were right, the conclusion that questionnaires are not useful for dictionary research would not follow. Tarp's critique is based on the argument that answers to (retrospective) questions (e.g. "Which information do you think was most helpful when you used the dictionary X") are unreliable, because they "only reveal the users' perception of their consultation, not the real usage" (Tarp, 2009b, p. 285). This seems to imply that for Tarp "the perception of the users" is not important at all. For example, if many users mention having trouble with a certain kind of information in a dictionary, this may not be identical to a "real usage" situation; nevertheless I think – contrary to Tarp - that this would at least be a result to think about and not just a negligible detail.

Thus, the bottom line is that it is important to bear in mind that "[c]onstructing a survey instrument is an art in itself" (Trochim, 2006), but this art does not have to be reinvented from scratch for the purpose of research into dictionary use, because there is already a vast body of literature on the proper construction of question-

naires (e.g. Krosnick, 1999; Krosnick & Fabrigar, forthcoming; Rea & Parker, 2005; or Diekmann, 2002, pp. 371–443).

(Direct) Observation

Whenever an instrument of measurement, such as a watch, a photon counter, or a survey is used, the reading of the instrument is an observation. As Diekmann (2002) points out:

> 'Generally speaking, all empirical methods are observational.'[4] (Diekmann, 2002, p. 456)

However, in this context, in terms of social research, observation can be defined as: directly and systematically gathering data about the unit(s)-of interest. In contrast to a survey design, the relevant information is not based upon the self-assessment or the answers of the participant. Thus, direct observation can take place both in an artificial setting (e.g., in a laboratory) or in a natural setting (e.g. a class room). Of course, observation can also be hidden, meaning that the subject is not aware of the observation (e.g., a log file analysis). In this case, the observation is indirect and has to be classified as an unobtrusive method (cf. Section 2.3.3). In social research, it has become a common strategy to measure the response latency, i.e. the duration between the presentation of a stimulus, for example a question, and the response. This measurement is then used as a proxy for various constructs, such as the accessibility of an attitude or the level of difficulty of a question (Mayerl, 2008). As the survey respondents do not necessarily have to be aware of the fact that their response time is being measured, this mode of observation has to be classified as a hybrid of an unobtrusive and an obtrusive method.

In dictionary research, several studies have used direct observational methods. For example, Aust, Kelley & Roby (1993) used the "raw number of words the subjects looked up in the dictionary" (Aust et al., 1993, p. 67) as a measurement for dictionary consultation and the "[n]umber of consultations per minute" (Aust et al., 1993, p. 68) as a measurement for efficiency. In a similar manner, Tono (2000) recorded "[t]he subjects' look up process [...] to obtain the list of words looked up. For each look-up word or phrase, the time taken for look-up and accuracy rate were calculated" (Tono, 2000, p. 858). Dziemianko (2010) carried out an "unexpected vocabulary retention test" (Dziemanko, 2010, p. 261) as one way of assessing "the usefulness of a monolingual English learners' dictionary in paper and electronic form" (Dziemanko, 2010, p. 259). Example 8 illustrates one of our observational approaches.

4 „In einem allgemeinen Sinne sind sämtliche empirische Methoden Beobachtungsverfahren."

Example 8
In our project, we tried to evaluate how users navigate their way around electronic dictionaries, especially in a dictionary portal. The concrete navigation process is hard to measure with a survey. In collaboration with the University of Mannheim, we therefore used an eye-tracker to record the respondent navigation behavior in "the lexicographic internet portal OWID, an Online Vocabulary Information System of German", hosted at the Institute for German Language (IDS) (Müller-Spitzer et al., this volume).

Indirect methods[5]

In dictionary usage research, the analysis of log files seems to be the best example of an indirect method. Other applications of this type of method are at least imaginable:

- A researcher could monitor the library lending figures of different dictionaries. This measure could serve as an indicator of the importance of the particular dictionary.
- A researcher could ask participants to translate texts using a dictionary. The resulting translated texts are then analyzed for lexical choices (especially erroneous ones). This analysis can then be used to "recreate the scenario that led to choosing the wrong equivalent" (Lew, personal communication).

However the application of this method in dictionary research seems to be mainly restricted to the analysis of log files.

Content analysis

By analyzing any kind of existing written material, the aim of content analyses is to find patterns in texts (Trochim, 2006). Both Ripfel and Wiegand (1988), Zöfgen (1994), and Wiegand (1998) list content analysis as one distinct method of dictionary usage research. But to my knowledge, no study applying this method has been pub-

5 The discovery of a special empirical distribution of digits is an intuitive example of an indirect method of data collection: to detect fraud in statistical data, Newcomb-Benford's law (Diekmann & Jann, 2010; Diekmann, 2007) can be used. This law states that the digits in empirical data are often distributed in a specific manner. So, if the published statistical results do not follow this distribution, this is an indicator for faked data (e.g. Roukema (2009) analyzed the results of the 2009 Iranian Election). The distribution was first discovered by the astronomer Simon Newcomb (1881). Without the assistance of calculating devices or a computer, the only option in those days was to rely on precalculated logarithm tables. Newcomb made an interesting observation: he noticed that the early pages of the books containing those tables were far more worn out than the pages in the rest of the book. This observation led, in turn, to the formulation of this law.

lished so far. This is somewhat surprising, as in neighboring disciplines, such as corpus linguistics, the same techniques are applied, e.g. analyzing keywords (in context) or word frequencies (e.g. Baayen, 2001; Lemnitzer & Zinsmeister, 2006). Example 9 demonstrates how we used a content analysis to investigate the answers given in an open-ended question.[6]

Example 9
In Example 2, the open-ended question has already been presented: *In which contexts or situations would you use a dictionary? Please use the field below to answer this question by providing as much information as possible*. To analyze the answers given to that question, we used the TEXTPACK program (cf. Mohler & Züll, 2001; Diekmann, 2002, pp. 504–510). On average, our respondents wrote down 37 words. There are no noteworthy differences between the German language version of the survey and the English version. As shown in Müller-Spitzer's chapter about usage opportunities and contexts of dictionary use, our results indicate that active usage situations (e.g. translating or writing) are mentioned more often than passive situations (e.g. reading) (Müller-Spitzer: Contexts of dictionary use, this volume).

Secondary analysis of data

To be precise, this type is not an actual method of collecting data, since it uses or re-analyzes existing data (Diekmann, 2002, pp. 164–165). In the natural sciences, this is common practice. Unfortunately, as Trochim (2006) points out:

> "In social research we generally do a terrible job of documenting and archiving the data from individual studies and making these available in electronic form for others to re-analyze." (Trochim, 2006)

This seems to hold for dictionary research, too. In our project, we have decided to publish the raw data on which our findings are based on our website www.using-dictionaries.info/ including supplementary material, such as the questionnaires.

Data analysis

Since the answer to this question is beyond the scope of this chapter, the purpose of the next section is to briefly outline the relationship between the planning process of an empirical study, the data collection and the subsequent data analysis. Angrist

6 Of course, as this question was part of a survey, it is not appropriate to classify this as an unobtrusive method. This example is just for illustration purposes.

& Pischke (2008), Baayen (2008), Fox & Long (1990), Gries (2009), Kohler & Kreuter (2005), and Scott & Xie (2005) provide useful introductions to the statistical analysis of data. At this point, it is important to emphasize that if the previous steps have been carefully and thoroughly followed, the statistical analysis of the data can be quite easy to manage.

In the best case scenario, an initial idea of how to analyze the data is developed during the early planning stages of the study, while the worst case scenario is a situation where the investigator starts to analyze the data and finds out that s/he cannot answer his/her research questions, because necessary data on confounding variables (cf . Section 2.3.2.2) have not been collected, or, maybe even worse, plenty of data have been collected but no proper research questions have been articulated, so the data analysis ends with a Popperian "what" (cf. Section 2.1).

In addition to a graphical numerical description of the collected data, the purpose of quantitative methods is to use the distributional information from a sample to estimate the characteristics of the population that the sample was taken from (statistical inference). For research into the use of (online) dictionaries, the relevant populations can be overlapping but need not be the same, depending on the research question.

A researcher, for example, who wants to understand the specific needs of the users of the online version of the OED, could choose a population such as *everyone who has ever used the OED Online*. The sample then only includes data from people who use (or have used) this specific dictionary. However, this sample would be inappropriate, of course, if the researcher wanted to compare the needs of experienced OED Online users with the needs of potential new users. In this case, the researcher first has to decide which subjects the population actually consists of. While the population in common political opinion polls is usually all eligible voters, the actual population in dictionary usage research has to be determined on a methodological basis. As previously mentioned in Section 2.3.2.2, it is not possible to learn anything about the needs of potential online dictionary users by conducting a log file study, because the sample only includes data from people who have actually used the dictionary:

> "For example, if log files show that someone has typed in *Powerpuff Girls* into our online dictionary, what do we do with this information? For all we know, this could be an 8-year old trying to print a colouring page of her favourite cartoon characters. So where do we go from here?" (Lew, 2011, p. 7)

In our research project, this also turned out to be one of Lew's hard questions, as representativeness is an important issue in research into the use of dictionaries and empirical quantitative research in general (Lew, 2011, p. 5). Roughly speaking, our target population consisted of all (!) internet users. For financial and technical reasons, it was of course not feasible to draw a random sample of all internet users. Since it is also rarely possible to conduct real controlled experimental trials (see

Section 2.3.2.1 for an explanation), we decided on the one hand to collect data on the respondents' academic, professional, and socio-demographic background, and, on the other hand, to obtain a huge number of respondents by distributing the surveys through multiple channels such as "Forschung erleben" ("experience research"), which is an online platform for the distribution of empirical surveys run and maintained by the chairs of social psychology at the University of Mannheim and visited by students of various disciplines, mailing lists (including the Linguist List (a list for students of linguistics and linguists all over the world hosted by the Eastern Michigan University), the Euralex List (a list from the European Association of Lexicography), and U-Forum (a German mailing list for professional translators)), and various disseminators (e.g. lecturers at educational institutions). While it is not possible to rule out any selection bias with a non-experimental design for principal reasons (cf. Section 2.3.2), we used an ex-post-facto design to control for potential group differences (cf. Example 7). For example, it could be argued that our survey results are somewhat biased towards lexicographical experts. In order to respond to this justified criticism, we could (and in fact we did) compare the data of respondents who were invited to take the survey via the online platform "Forschung erleben" with that of the rest of our respondents, because it is unlikely that the former group mainly consists of typical "dictionary experts".

However, results from a non-representative sample are problematic if and only if the traits of people taking part are correlated with the outcome of interest (cf. Section 2.3.2.1), because in this case – statistically speaking – the estimators are no longer efficient. Essentially, this just means that it is not possible to infer from the sample to the population (cf. Yeager et al., 2011).

3 Conclusion

To summarize, let me refer to Lew's (2010) keynote mentioned in the introduction. Lew defends the hypothesis that

> "[m]uch of the available body of user research appears to have invested the better part of time and effort into data collection and analysis, to the detriment of careful planning and reflection. But, arguably, more benefit might have come from redirecting this time and effort to the more careful planning of the study design." (2010, p. 1 f)

I think Lew made an important point, not only for empirical research into the use of (online) dictionaries, but in general for any empirical investigation. In a similar vein, Diekmann (2002, p. 162) states:

> 'Some studies have to face the problem that "any old thing" in the social field is supposed to be investigated, without the research objective being even roughly defined. At the same time, there is a lack of careful planning and selection of a research design, operationalization, sam-

pling and data collection. The result of ill-considered and insufficiently planned empirical "research" is quite often a barely edible data salad mixed with extremely frustrated researchers.'[7]

I hope that this chapter shows that, while the planning of empirical research into dictionary use and empirical research in general can be quite demanding, this additional effort pays off, because it helps enormously to answer many questions relevant for research into the use of online dictionaries.

Bibliography

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton NJ: Princeton University Press.

Atkins, S. B. T., & Varantola, K. (1997). Monitoring dictionary use. *International Journal of Lexicography*, *10*(1), 1–45.

Aust, R., Kelley, M. J., & Roby, W. (1993). The Use of Hyper-Reference and Conventional Dictionaries. *Educational Technology Research and Development*, *41*(4), 63–73.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.

Babbie, E. (2008). *The Basics of Social Research* (4th ed.). Belmont, CA: Wadsworth.

Banerjee, A. V., & Duflo, E. (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Bergenholtz, H., & Johnson, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. *Hermes*, (34), 117–141.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Skokie, Ill: Rand McNally.

Das, J., & Hammer, J. (2005). Which Doctor? Combining Vignettes and Item-Response to Measure Doctor Quality. *Journal of Development Economics*, *78*(2), 348–383.

De Schryver, G.-M., & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, Lorient, France, July 6th–10th* (pp. 187–196). Lorient: Université de Bretagne Sud.

Diekmann, A. (2002). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (8th ed.). Reinbek: Rowohlt Taschenbuch Verlag.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics*, *34*(3), 221–229.

Diekmann, A., & Jann, B. (2010). Law and Fraud Detection: Facts and Legends. *German Economic Review*, *11*(3), 397–401.

7 „Manche Studie krankt daran, daß ‚irgend etwas' in einem sozialen Bereich untersucht werden soll, ohne daß das Forschungsziel auch nur annähernd klar umrissen wird. Auch mangelt es häufig an der sorgfältigen, auf das Forschungsziel hin abgestimmten Planung und Auswahl des Forschungsdesign, der Variablenmessung, der Stichprobe und des Erhebungsverfahrens. Das Resultat unüberlegter und mangelhaft geplanter empirischer ‚Forschung' sind nicht selten ein kaum noch genießbarer Datensalat und aufs äußerste frustrierte Forscher oder Forscherinnen." (Diekmann (2002, p. 162).

Dziemanko, A. (2010). Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. *International Journal of Lexicography*, *23*(3), 257–273.

Fox, J., & Long, S. A. (1990). *Modern Methods of Data Analysis*. Thousand Oaks, CA: Sage.

Gries, S. T. (2009). *Statistics for Linguistics with R: A Practical Introduction* (1st ed.). Berlin, New York: De Gruyter Mouton.

Hartmann, R. R. K. (1987). Four Perspectives on Dictionary Use: A Critical Review of Research Methods. In A. P. Cowie (Ed.), *The Dictionary and the Language Learner* (pp. 11–28). Tübingen: Niemeyer. Retrieved June 12, 2011, from http://search.ebscohost.com/login.aspx?direct=true&db=mzh&AN=1987017474&site=ehost-live

Hartmann, R. R. K. (1989). Sociology of the Dictionary User: Hypotheses and Empirical Studies. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher – Dictionaries – Dictionnaires. Ein internationales Handbuch zur Lexikographie* (Vol. 1, pp. 102–111). Berlin, New York: de Gruyter.

Hartmann, R. R. K. (2000). European Dictionary Culture. The Exeter Case Study of Dictionary Use among University Students, against the Wider Context of the Reports and Recommendations of the Thematic Network Project in the Area of Languagew 1996-1999. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *IX EURALEX International Congress* (pp. 385–391). Stuttgart.

Hulstijn, J. H., & Atkins, B. T. S. (1998). Empirical research on dictionary use in foreign-language learning: survey and discussion. In B. T. S. Atkins (Ed.), *Using Dictionaries* (pp. 7–19). Tübingen: Max Niemeyer Verlag.

Kellehear, A. (1993). *The Unobtrusive Researcher: A guide to methods*. St. Leonards, NSW: Allen & Unwin Pty LTD.

Kohler, U., & Kreuter, F. (2005). *Data Analysis Using Stata*. College Station: Stata Press.

Koplenig, A. (2011). Understanding How Users Evaluate Innovative Features of Online Dictionaries – An Experimental Approach (Poster). Presented at the eLexicography in the 21st century: new applications for new users, organized by the Trojina, Institute for Applied Slovene Studies, Bled, November 10-12.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

Krosnick, J. A., & Fabrigar, L. R. (forthcoming). *The handbook of questionnaire design*. New York: Oxford University Press.

Lemnitzer, L., & Zinsmeister, H. (2006). *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Lew, R. (2010). Users Take Shortcuts: Navigating Dictionary Entries. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1121–1132). Ljouwert: Afûk.

Lew, R. (2011). User studies: Opportunities and limitations. In K. Akasu & U. Satoru (Eds.), *ASIALEX2011 Proceedings Lexicography: Theoretical and practical perspectives* (pp. 7–16). Kyoto: Asian Association for Lexicography.

Ludwig-Mayerhofer, W. (2011). Ilmes – Internet Lexikon der Methoden der empirischen Sozialforschung. *ILMES – Internet-Lexikon der Methoden der empirischen Sozialforschung*. Retrieved September 14, 2013, from http://wlm.userweb.mwn.de/ilmes.htm

Madans, J., Miller, K., Maitland, A., & Willis, G. (Eds.). (2011). *Experiments for evaluating survey questions*. New York: John Wiley & Sons.

Mayerl, J. (2008). Response effects and mode of information processing. Analysing acquiescence bias and question order effects using survey-based response latencies. Presented at the 7th International Conference on Social Science Methodology, Napoli.

Mohler, P. P., & Züll, C. (2001). Applied Text Theory: Quantitative Analysis of Answers to Open-Ended Questions. In M. D. West (Ed.), *Applications of Computer Content Analysis*. Conneticut: Aplex Publishing.

Müller-Spitzer, C., & Möhrs, C. (2008). First ideas of user-adapted views of lexicographic data exemplified on OWID and elexiko. In M. Zock & C.-R. Huang (Eds.), *Coling 2008: Proceedings of the workshop on Cognitive Aspects on the Lexicon (COGALEX 2008)* (pp. 39–46). Manchester. Retrieved 14 September, 2013, from http://aclweb.org/anthology-new/W/W08/W08-1906.pdf

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

Popper, K. (1972). *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford Univ Press.

Rea, L. M., & Parker, R. A. (2005). *Designing and Conducting Survey Research. A Comprehensive Guide* (3rd ed.). San Francisco: Jossey-Bass.

Ripfel, M., & Wiegand, H. E. (1988). Wörterbuchbenutzungsforschung. Ein kritischer Bericht. In H. E. Wiegand (Ed.), *Studien zur neuhochdeutschen Lexikographie VI* (Vol. 2, pp. 491–520). Hildesheim: Georg Olms Verlag.

Roukema, B. F. (2009). Benford's Law anomalies in the 2009 Iranian presidential election. Retrieved September 14, 2011, from http://arxiv.org/abs/0906.2789

Scott, J., & Xie, Y. (2005). *Quantitative Social Science*. Thousand Oaks, CA: Sage.

Shang, A., Huwiler-Müntener, K., Nartey, L., Jüni, P., Stephan Dörig, Sterne, J. A. C., … Egger, M. (2005). Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet*, *366*, 726–732.

Tarp, S. (2009a). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen, & S. Tarp (Eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (pp. 17–32). Frankfurt a.M./Berlin/Bern/Bruxelles/NewYork/Oxford/Wien: Peter Lang.

Tarp, S. (2009b). Reflections on Lexicographical User Research. *Lexikos*, *19*, 275–296.

Tono, Y. (1998). Interacting with the users: research findings in EFL dictionary user studies. In T. McArthur & I. Kernerman (Eds.), *Lexicography in Asia: selected papers form the Dictionaries in Asia Conference, Hong Kong University of Science and Technology* (pp. 97–118). Jerusalem: Password Publishers Ltd.

Tono, Y. (2000). On the Effects of Different Types of Electronic Dictionary Interfaces on L 2 Learners' Reference Behaviour in Productive/Receptive Tasks. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress, Stuttgart, Germany, August 8th–12th* (pp. 855–861). Stuttgart: Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Trochim, W. (2006). Design. *Research Methods Knowledge Base*. Retrieved September 14, 2013, from http://www.socialresearchmethods.net/kb/design.php.

Trochim, W., & Land, D. (1982). Designing designs for research. *The Researcher*, *1*(1), 1–6.

Verlinde, S., & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Ljouwert: Afûk.

Wiegand, H. E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlin, New York: de Gruyter.

Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Storrer, A., & Wolski, W. (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung: mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neuen Sprachen*. de Gruyter. Retrieved September 14, 2013, from http://books.google.de/books?id=Bg9tcgAACAAJ.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quaterly*, *75*(4), 709–747.

Zöfgen, E. (1994). *Lernerwörterbücher in Theorie und Praxis. Ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*. Tübingen: Max Niemeyer.

Zwane, A. P., Zinman, J., Dusen, E. V., Pariente, W., Null, C., Miguel, E., … Banerjee, A. (2011). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 1821,1826.