User studies: Opportunities and limitations

Robert Lew, Adam Mickiewicz University

rlew@amu.edu.pl

Abstract

In this keynote address I reflect on the role of user studies in dictionary research. Starting with a distinction between positivistic and naturalistic general methodological paradigms, I discuss a number of specific methods and techniques, pointing to their respective strengths and weaknesses. I introduce the audience to a related research paradigm which has recently emerged within computer science, namely usability studies. Finally, I single out several key issues having to do with construct validity and operationalization, demonstrating on concrete examples why I see them as topical problems.

1. Introduction

Methodological reflection on research into dictionary use has not had a very long history, but is marked by a handful of meaningful contributors, notably Reinhard Hartmann and Yukio Tono: both very familiar to this audience.

A list of techniques with which to investigate dictionary look-up processes includes, but need not to limited to:

- observation (participant/non-participant)
- self-accounts
- think-aloud protocols (TAP)
- videotaping
- screen recorders
- server logging
- eye tracking

With modern advances in technology, some of the traditional techniques have found their counterparts in various technology-enhanced techniques. For example, direct visual observation by the investigator can to some extent be replaced with videotaping.

2. Methodological approaches to the study of dictionary use

Cohen et al. (2007) make a broad distinction between two general research paradigms: positivistic and naturalistic, which may serve as a convenient framework for discussing the different strains of research into dictionary use. Both paradigms have their advocates and skeptics, but to get the most complete picture possible, I believe there is room for engaging both approaches, as in fact they do not exclude, but rather complement one another. Dictionary use is a complex activity, with some aspects more quantifiable and thus amenable to the positivistic approach, others more qualitative, attitudinal, less tangible and thus not readily reducible to simple numbers.

I would therefore argue here that a felicitous division of labour between the different methods and techniques should be sought, and to achieve this goal we would need greater methodological reflection at the planning stage than has routinely been the case so far. Much

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. A somewhat similar point has recently been made by Tarp (2009b). One major limitation of the "hard science" approach, particularly the true experiment (and this point is valid more generally, not just within dictionary use research) is its "small steps" characteristic: the fact that planning and executing exact experimental designs takes a lot of time and resources, and can only cover a tiny slice of the whole range of research problems at a time. Therefore, it makes good sense resource-wise to split the work to be done so that the "softer" approaches may be used for the initial job of judiciously restricting the alternatives which could later be subjected to more rigorous experimental scrutiny. In the real world, where time and resources are limited, we should think twice before using too many resources on expensive procedures only to confirm the all but obvious.

Exploratory qualitative methods can be used to "populate" the field with questions, problems and hypotheses. This is the approach advocated by Kwary (2010), where he suggests that the "softer" methods be used to a greater extent in the initial hypothesis formulation. In his specific proposal, Kwary singles out two interesting methodological approaches so far hardly utilized in dictionary use research. One is the focus group approach, which can alternatively be described as a semi-structured group interview, often involving an element of hands-on practice which is later followed up by retrospective sessions. A fairly similar methodological approach is adopted in the recent research by Chan (in press, 2011) involving Hong Kong ESL learners. The other qualitative methodology proposed by Kwary (2010) is the Delphi method. The key element of the Delphi method is a series of evaluations by a panel of experts, proceeding in a stepwise fashion in at least two rounds, with expert opinions being summarized and disseminated to the remaining panel members. The method has clear philosophical grounding in the rationalist tradition.

2.1. The positivistic approach and its problems

The positivistic approach attempts to isolate and control variables, and standardize conditions. The *randomized controlled trial* is seen as the gold standard in many domains of research. The attraction of the randomized controlled trial lies in the belief that it comes with a reasonable guarantee of *causality* and *generalizability*. The paradigm grew out of laboratory-based natural sciences, and there has been continued controversy as to the degree of its applicability in the humanities. Skeptics argue that the fixation on causality is unwarranted, and propose instead to focus more on dynamic, developing relationships, networks and interconnections. It may be claimed that isolating out a handful of variables for manipulation and measurement may violate those complex relationships and networks, leading to conclusions of dubious value.

In this vein, it is worthwhile to reflect on the nature of the interaction between human users and dictionaries, and the role of user research in this schema (see Figure 1). Through their normal contact with lexicographic products, the reference habits of dictionary users are being shaped (I am ignoring here, in this simplified model, the impact of any instruction in dictionary use, be it formal training or casual guidance offered by the teacher, family member, friend, etc.). As a result of such repeated interaction, dictionary reference skills arise. It is these habits and skills, and more directly what actually takes place in acts of dictionary consultation, that come into the purview of lexicographic user research. Provided that publishers are willing and interested, they can feed the results of user research into the *content* and *presentation* of the dictionaries they subsequently produce.

All this patterns into a picture, if not exactly of a vicious circle, then at least one that includes a distinct self-perpetuating element. The cyclic pattern dovetails quite well with the story of past lexicographic practice as we know it, which has been marked by repetition and circularity. User research has only been in the picture for what is but a blip in the history of lexicography, and it is not easy to break with the old ways. For one thing, publishers have for centuries copiously drawn on previous lexicographic works. For another thing, and more recently, user expectations have been shaped by existing dictionaries, but those expectations have themselves formed a basis for creating the next generation of dictionaries.

In this scheme of things, user research tends to set its sights on the description of the users and their interaction with dictionaries, and as such would be classified as *contemplative* in the sense of Tarp (2009a): focusing on the present and past lexicographic practice rather than on what can be done to improve it in the future. However, user research can also address Tarp's *transformative* agenda, in that it can be employed to field-test alternative lexicographic solutions: those that have already been adopted in some actual dictionaries, as well as novel, experimental ones. This is in fact what some studies have attempted, but are such endeavours fully immune to the circularity fallacy? Unfortunately, they are not, and here is the reason why.

As users work with a dictionary over time, they learn some of the structure, conventions; they learn how to cut corners. Humans exhibit a natural and generally healthy cognitive tendency to economize on the amount of attention assigned to the task at hand. So in the course of interaction with dictionaries, users' habits adjust, and their reference skills evolve. The process is driven through users getting accustomed to the particular features of the dictionary. A realization of this fact is central to grasping an important limitation on experimental user research. The efficiency and effectiveness of a given feature as measured by experimental means will not just be the function of its inherent *fitness for purpose*; it will inevitably reflect the degree to which the users tested have become *habituated* to the specific solutions, and their resulting dexterity in making use of them.

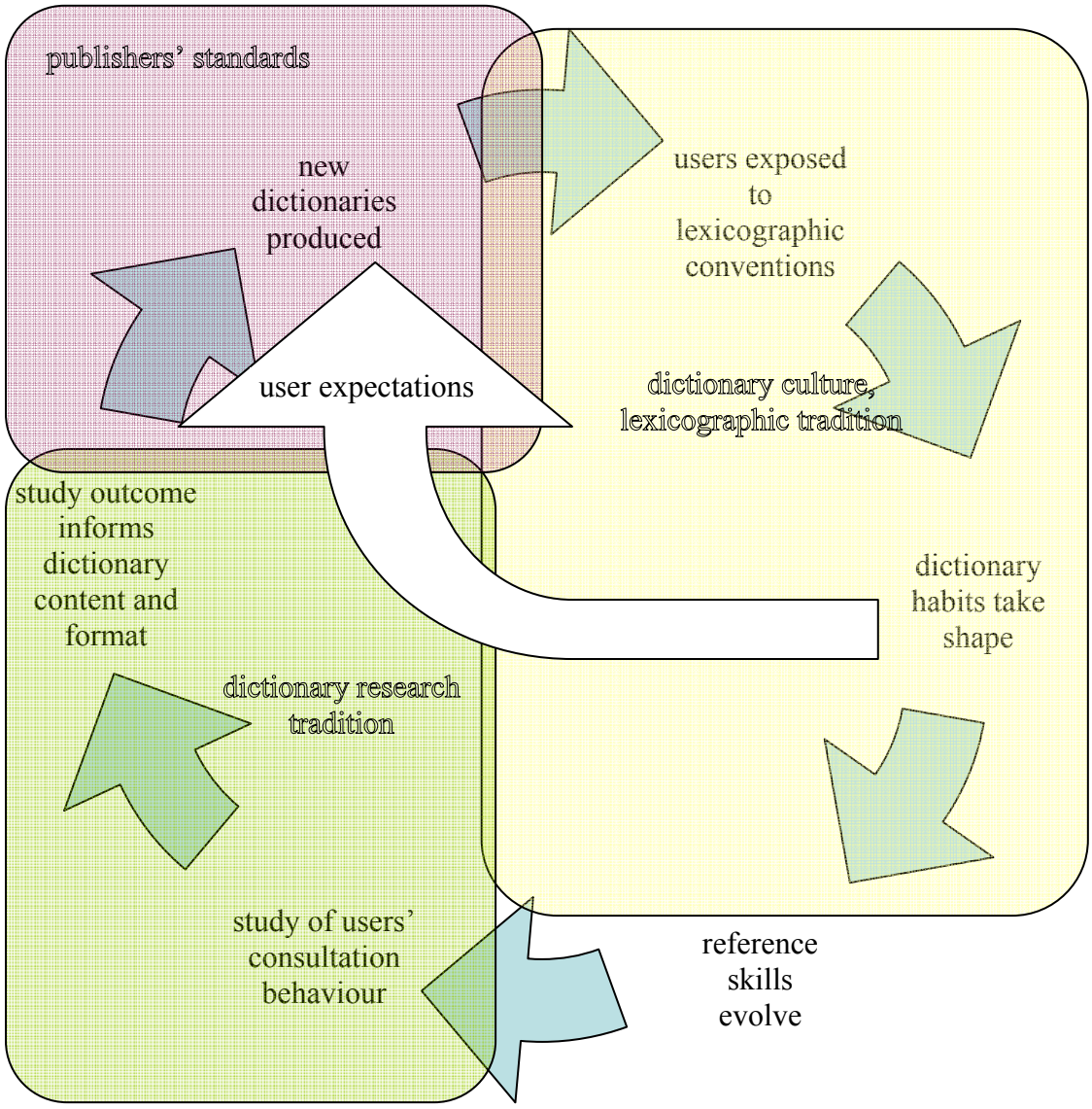

Figure 1: The cycle of dictionary use and user research

This may not be much of a problem if the user's experience with the specific lexicographic solutions being tested is relatively balanced or uniform. But if a solution is unknown to the users, as is necessarily the case with any experimental feature we would like to test, their performance is likely to be negatively affected by the novelty of the feature. Depending on how steep a learning curve the new feature has, it may take more or less time and practice before users get more familiar with the innovation tested, and before the benefits, if any, get a chance to come to the surface.

In cases such as these, one solution might be to conduct longer-term studies. Yet longitudinal or time-series studies are almost unheard of in dictionary research on users, one notable exception being Jim Ronald's (2002) study of vocabulary acquisition from dictionary-assisted book reading. A related and also unexplored format that could be adapted for dictionary use studies is the *single-case research design* (Kazdin 1982), in particular *phase design*, where a single subject would be observed over a prolonged period under alternating conditions, such as, say, the use of two different dictionaries, or dictionaries of different formats.

2.2. Sampling the dictionary-using population

Representativeness is an issue that deserves serious attention, too, but the bulk of studies adopt a somewhat cavalier approach to sampling. Convenience sampling prevails. Metalexicography shares this rather carefree approach to sampling issues with other research domains, notably linguistics. Judging by some of the comments made by lexicographic researchers in their sections on methodological limitations, this is perhaps not so much primarily a question of insufficient awareness of the issue, as of limited resources. Nevertheless, clearly a more determined attempt should be made to reach beyond the most common setting of university-level education. While it is not entirely true that lower-level educational stages have not been studied at all, there certainly is an imbalance in how well the dictionary-using population has been covered by past studies. In particular, dictionary use beyond the setting of educational institutions remains a severely underresearched black hole. Isolated exceptions—such as Diab's (1990) study of dictionary use by nurses, or Łakomski's (2001) investigation into the dictionary habits and skills of Swedish border services staff—only underscore the deficit of data on dictionary use by various types of users in a range of possible situations and settings (e.g. families doing crossword puzzles at home, tourists, and many others).

2.3. The naturalistic paradigm: study types

Moving on now to the *naturalistic paradigm*, it is fair to say that it tends to be qualitative and interpretive, but may include the occasional quantitative element as well. The approach places less emphasis on representativeness and generalizability, as these concepts are of questionable value if each event is viewed as one-time and unique. The naturalistic approach tries to respect the natural wider context of the phenomena under study.

Within metalexicography, naturalism means focusing on the context of dictionary use as well as details of the process of dictionary use. These details are worthy of being recorded and reflected on. One way to achieve this is through careful case studies.

2.3.1. Case studies

Case studies are underestimated and underutilized. However, when properly executed, case studies have many advantages. Case studies—in contrast to surveys or experiments—are *strong on reality* (Nisbet and Watt 1984): they are rich in authentic data, preserving their natural sequencing. This gives the case study a potential to reveal the rich texture and complexity of the interaction between the user and the dictionary, identifying user strategies, their genuine problems and the way those problems are tackled by the users: all this with a level of granularity normally unattainable in survey-type studies. Given sufficient duration, a case study could also be useful in tracking the acquisition and development of dictionary reference skills. Case study reports are normally more immediately intelligible than reports on experimental research and, as such, are potentially more accessible to a greater variety of non-academic audiences, including practical lexicographers, but also teachers, publishers and other stakeholders in the dictionary-using context; or indeed various dictionary users themselves. Case studies may employ data-collection techniques that focus more on the details of the process of dictionary use. A variety of *observation protocols* may be involved.

2.3.2. Observation

Observing what dictionary users do while immersed in actual dictionary use can shed light on the process of dictionary use, the types of difficulties experienced by users, and the strategies they adopt when interacting with dictionaries. Various technological enhancements may assist in the process of observing dictionary use, mostly to make a permanent record of

what happens, so that it can be subject to systematic and repeated review at a later time. These include audio and video recording devices, as well as—for computer-based dictionary use—screen recorders. Eye tracking is a new technique capable of tracing the user's eye gaze across the screen. Some types of equipment can perform eye-tracking when interacting with physical objects, such as a printed dictionary. One of the first applications of eye-tracking in dictionary use is Tono (2011). For a pioneering study of this type, it covers an impressive range of issues and detail. In another very recent eye-tracking study, Simonsen (2011) has studied the consultation of an online Danish-English accounting dictionary by professional translators, which he followed up with interviews. His data bring into light differences in the patterns of use of online entries, where translators with experience in the field tend to spend less time on examining the lexicographic data than on interacting with the search field, while less experienced translators do the opposite.

However, it would be naive to believe that the eye-tracker will instantly answer all questions as to how users interact with dictionaries. The technique uses the user's gaze as a proxy for attention. Even if this simplification is accepted, there are still questions open to various interpretations. One basic question is how long a user needs to fixate in an area of the screen—and what size an area—for this to count as a genuine instance of examining the data found there. Some guidelines in this matter are available from past studies in the areas of reading research and web page usability, but Simonsen (2011) cautiously states that they may not be directly applicable to dictionary consultation due to a different nature of the interaction.

2.3.3. Interviews

Another technique relatively little used in dictionary use research is that of the interview. A recent call to expand the interviewing techniques to include *focus group* interviews (Kwary 2010) has resulted in at least one interesting follower (Chan 2011). Interviews may be particularly useful for probing the field. Oppenheim (1992), in a classic manual focusing on questionnaires and interviews, writes of *exploratory interviews* as a type of heuristic for hypothesis generation rather than data collection.

2.4. Log files

As dictionaries became more frequently offered, and consequently consulted, in electronic rather than printed form, tapping the electronic trace of the dictionary user's consultation behaviour as a window into patterns of dictionary use became a possibility. Although this potential of computers was noted already in the 1980's (Hatherall 1984; Hartmann 1987), for many years virtually no studies followed up on the idea (Knight 1994 being one worthwhile exception).

Greater interest in the use of computer logging as a way of recording dictionary users' consultation behaviour came with the proliferation of internet dictionaries. Internet servers typically log their interaction with their remote "clients" in machine-readable text files. The data so acquired may be of use in various tasks related to maintenance, quality control, or security. But log files can also reveal how visitors are interacting with the website.

Log files is, in fact, a rather ambiguous term, referring, as it does, to a particular form of data record, yet, on a literal level at least, neutral with respect to the way and circumstances in which the data itself were generated. And so, the term can well refer to the electronic records of just about any computer-based investigation of an actual or experimental dictionary (e.g. Lew and Doroszewska 2009; Lew and Tokarek 2010), this irrespective of the exact data collection methodology. It might, for instance include a study done within the *usability paradigm*, of which more will be said shortly. In the context of online dictionaries, however, *log files* would simply refer to run-of-the-mill web server logs.

If used in the latter sense, log files have a number of limitations. The context of dictionary use is completely unknown (with the possible exception of the user having been redirected from a well-known site, as indicated by *http referrer logging*). Furthermore, we typically know precious little about the user. We cannot be sure that the user has selected an even remotely appropriate tool for the job. 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? Should we let ourselves be tempted to modify the dictionary to dutifully and indiscriminately serve all types of oddball queries? Whatever happened to the mantra of *genuine purpose* here? And how is this better than letting user questionnaires shape the dictionary: an approach criticized (among others) on the grounds that users tend to be limited by their particular experience, imagination, and preconceptions of what a dictionary is supposed to be like? Leaving these questions in the air for now, I would like to urge the researchers employing online log files to reflect on the limitations of this admittedly convenient data source.

2.4.1. Usability studies

An alternative *usability study* paradigm, which has evolved quite independently of the metalexicographic and applied-linguistics traditions, can also be applied in the study of dictionary use, most directly electronic dictionaries. This research paradigm has arisen within the domain of information science and (more specifically) human-computer interaction. There is now lively interest in the quality of interaction between humans and machines, through interface and software. A general concept of *usability* is applicable here, even though there is no single agreed definition of usability.

One fairly popular view is due to Nielsen (1993), who defines usability by five quality components:

- **Learnability**: How easy is it for users to accomplish basic tasks the first time they encounter the design?
- **Efficiency**: Once users have learned the design, how quickly can they perform tasks?
- **Memorability**: When users return to the design after a period of not using it, how easily can they reestablish proficiency?
- **Errors**: How many errors do users make, how severe are these errors, and how easily can they recover from the errors?
- **Satisfaction**: How pleasant is it to use the design?

Some authors separate out *utility* from usability, where utility would gauge how *useful* for the task at hand the software tool is. A view more compatible with the dictionary use tradition is that of Heid (2011), who singles out three major aspects of usability: effectiveness, efficiency, and satisfaction. This understanding may be inclusive of utility, as the latter concept appears to be related to effectiveness. Heid's article is based on data obtained in the course of Christina Bank's M.A. project conducted under his direction at Universität Hildesheim (Bank 2010). It is reassuring to see what might be the beginnings of a convergence between the two historically separate research paradigms. Bringing together the two approaches could result in healthy cross-pollination between the two disciplines, which do in fact have a lot in common; after all, they study similar types of objects, and the repertoires of data-collection techniques overlap. This brings us to our next point: problem areas in the methodology of empirical user research.

3. Some key issues in the methodology of user research

3.1. Due attention to user variables

Proper reflection and attention should always be given to the relevant qualities of the user (such as their age, education, command of the language(s) involved, orientation within the pertinent subject domain, dictionary reference skills), as well as the context and nature of the task which has prompted dictionary use in the first place.

3.2. Construct validity and operationalization

Researchers need to ask the right questions and have to select the right instruments to obtain the best answers. This may sound very much like a commonplace, but the reality is that a surprisingly high proportion of user studies display problems with construct validity. I hope to be able to talk you through a few concrete examples during the plenary lecture itself. This is not included here due to space restrictions, but I will at least highlight the headings below:

Example 1: The construct of *dictionary type*: how to compare dictionary types by using specific titles

Example 2: Operationalizing the recognition of syntactic category (When-definitions and POS information)

Example 3: Test design in operationalizing the comprehension of low-frequency vocabulary items by native speakers

Example 4: Measuring success in dictionary use

4. The role of user studies in dictionary reviews

I would be amiss not to mention the role that user studies can play in the context of dictionary evaluation and criticism. Even though research on dictionary criticism is set apart from user research by the leading theoreticians of the field, such as Hartmann (e.g. 1999) or Wiegand (throughout, *Kritische Wörterbuchforschung*), it is in fact a natural consequence of viewing dictionaries as tools driven by user needs to try to assess how well such tools work in practice when evaluating their quality. In this paradigm, what stands out prominently is the user studies published in *Lexicon*, the journal of the Iwasaki Linguistic Circle. In fact, it has been ten years since the first one in a series of empirical user studies was conducted by Takashi Kanazashi and published in 2002. These reviews formed an important integral component of each of a series of comprehensive analyses of a number of EFL dictionaries.

5. Conclusion

User 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.

References

Bank, C. (2010). *Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen*. M.A. Thesis, Universität Hildesheim,

Chan, A. Y. W. (2011). “Bilingualised or monolingual dictionaries? Preferences and practices of advanced ESL learners in Hong Kong”. *Language, Culture and Curriculum 24*(1). 1-21.

Chan, A. Y. W. (in press). “The use of grammatical information in a monolingual dictionary in helping ESL learners determine the correct use of a target word”. *International Journal of Lexicography 25*.

Cohen, L., Manion, L. and Morrison, K. (2007). *Research methods in education* (6th ed.). London - New York: Routledge.

Diab, T. A. A. (1990). *Pedagogical lexicography. A Case study of Arab nurses as dictionary users* (Lexicographica Series Maior 31). Tübingen: Niemeyer.

Hartmann, R. R. K. (1987). “Four perspectives on dictionary use: A critical review of research methods”. In Cowie, A. P. (ed.), *The dictionary and the language learner. Papers from the EURALEX Seminar at the University of Leeds, 1-3 Apr. 1985* (Lexicographica Series Maior 17). Tübingen: Niemeyer. 11-28.

Hartmann, R. R. K. (1999). “What is 'dictionary research'?”. [Review article]. *International Journal of Lexicography 12*(2). 155-161.

Hatherall, G. (1984). “Studying dictionary use: Some findings and proposals”. In Hartmann, R. R. K. (ed.), *LEXeter '83 Proceedings: Papers from International Conference on Lexicography at Exeter, 9- 12 Sept. 1983* (Lexicographica Series Maior 1). Tübingen: Niemeyer. 183-189.

Heid, U. (2011). “Electronic dictionaries as tools: Towards an assessment of usability”. In Fuertes-Olivera, P. A. and Bergenholtz, H. (eds.). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Knight, S. (1994). “Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities”. *Modern Language Journal 78*(3). 285-299.

Kwary, D. (2010). “The Mixed Methods Design in User-Centred Lexicography”. In Zhang, Y. (ed.), *Learner's lexicography and second language teaching*. Shanghai: Shanghai Foreign Language Education Press. 160-172.

Lew, R. and Doroszewska, J. (2009). “Electronic dictionary entries with animated pictures: Lookup preferences and word retention”. *International Journal of Lexicography 22*(3). 239-257.

Lew, R. and Tokarek, P. (2010). “Entry menus in bilingual electronic dictionaries”. In Granger, S. and Paquot, M. (eds.). *eLexicography in the 21st century: New challenges, new applications*. Louvain-la-Neuve: Cahiers du CENTAL. 193-202.

Łakomski, T. (2001). *The image of the dictionary for Polish high school students and Scandinavian border workers: A comparison*. M.A., Adam Mickiewicz University, Poznań.

Nielsen, J. (1993). *Usability engineering*. San Diego: Academic Press.

Nisbet, J. and Watt, J. (1984). “Case study”. In Bell, J., et al. (eds.). *Conducting small-scale investigations in educational management*. London: Harper & Row. 79-92.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement*. London - New York: Pinter Publishers.

Ronald, J. (2002). “L2 lexical growth through extensive reading and dictionary use: A case study”. In Braasch, A. and Povlsen, C. (eds.). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 12-17, 2002, Vol.2*. Copenhagen: Center for Sprogteknologi, Copenhagen University. 765-771.

Simonsen, H. K. (2011). “User consultation behaviour in internet dictionaries: An eye-tracking study”. *Hermes 46*. 75-101.
Tarp, S. (2009a). “The foundations of a theory of learners' dictionaries”. *Lexicographica 25*. 155-168.
Tarp, S. (2009b). “Reflections on lexicographical user research”. *Lexikos 19*. 275-296.
Tono, Y. (2011). “Application of eye-tracking in EFL learners’ dictionary look-up process research”. *International Journal of Lexicography 24*(1). 124-153.