Usability testing as a tool for e-dictionary design: collocations as a case in point

Ulrich Heid & Jan Timo Zimmermann

Keywords: *electronic dictionaries, usability testing, collocations, access to lexicographic data*.

Abstract

We report about the application of usability tests to electronic dictionaries; our examples concern the design of dictionary interfaces that allow the user to access lexicographic data about collocations. We thus first summarize options for collocation retrieval, in terms of search criteria and types of data displayed as search results. We then present usability testing methods in general, as well as their application to electronic dictionaries, and we report about two tests, one with existing e-dictionaries, the other with custom-built mock-ups. We interpret this work as a first step towards usability design of electronic dictionaries: we suggest that new concepts for e-dictionary interfaces could be developed by rapid prototyping and tested with users before being integrated into dictionary products.

1. Introduction

1.1. *Motivation – Objectives*

The design of most dictionaries, paper or electronic, is based on what the designer assumes to be the best (affordable) way to present lexicographic data. While data description is a task for the sciences dealing with the contents of the dictionary, data presentation is a genuine task of lexicographers. In electronic dictionaries, lexicographic data presentation includes (i) modelling of search pages, (ii) modelling of data display in reply to searches carried out by the users and (iii) general user interface design.

As electronic dictionaries may be seen as software tools (cf. Bergenholtz/Bergenholtz 2011, Heid 2011), it seems natural to verify the quality of their user interfaces, i.e. the above aspects (i) to (iii), by using the same means as used for other software tools, namely usability tests as they are known from information science (Dumas/Redish 1999). Information scientists have developed a range of usability criteria (cf. ISO 9241) and usability testing procedures (cf., section 2).

We applied usability tests and the pertaining pre- and post-test questionnaires, both to existing electronic dictionaries (Bank 2010) and to mock-ups of possible electronic dictionaries. We report mainly on the second type of tests, as these serve also as a starting point for improved dictionary design. We focus on collocations, and we mainly test usability issues of a very simple dictionary that follows the interface metaphor of search engines, as compared with profile-based dictionaries of the type of the *ordbøger over faste vendinger* (lemma.com). Contrary to many studies on "general purpose" dictionaries (e.g. Tiberius/Niestad 2011), we found that our users prefer in the particular situation at hand the profile-based dictionaries over the search-engine-like ones.

We briefly introduce usability, usability testing and usability design in section 2, describe usability testing of electronic dictionaries (section 3), and report about results of tests of existing (section 4) and proposed (mock-up) dictionaries (in section 5). The remainder of this introduction is about collocation presentation in electronic dictionaries.

1.2. *Presenting lexicographic data about collocations in electronic dictionaries*

Notion of 'collocation': We use a notion of 'collocation' that is anchored in Hausmann's approach to collocations (cf. Hausmann 2004, etc.) and which assumes the existence of a collocation base (autosemantic element) and a collocate (synsemantic element), which are in a syntactic relationship, e.g. verb+object, noun+adjective, etc. (cf. Bartsch 2004:76). This view is shared by a number of collocation dictionaries, most prominently by the *Oxford Collocations Dictionary for Students of English*, OCDSE (McIntosh Ed., 2009).

Access to collocations: In the following, we adopt a user-oriented view on electronic dictionaries that is inspired by the Function Theory of lexicography (cf. e.g. Tarp 2008). We make reference to cognitive as well as to communicative situations in which (electronic) dictionaries may be used. Among the latter, we follow the theory in distinguishing text production and text reception functions.

As collocations show the above mentioned semantic polarity, a production-oriented dictionary will give access from the base to all collocations which are typical for that base. To ease search in the dictionary, the collocates may be sorted by category (e.g. the base *line*$_{N}$ plus all its adjectives, such as *straight line*, *dotted line*, ...) or by meaning (onomasiological collocation dictionary). The explanatory combinatorial dictionaries (ECDs) of the tradition of Mel'čuks Meaning ↔ Text Theory are typical production-oriented collocation dictionaries, where the Lexical Functions used to explain collocations can be seen as a device supporting access by meaning: they are generalizations of collocate meaning (e.g. causative, inchoative, etc.) that allow the user to identify appropriate collocations of a given base by searching for these abstract meanings.

For text reception, the situation is different. If the user has an understanding problem, (s)he likely does not understand the reading of the collocate, and probably, if the base is polysemous, also not the reading of the base. (S)he may in addition not even be aware of being in presence of a collocation. To optimally support a user in such a situation, the electronic dictionary should give access to collocations (i) from the collocate, (ii) from the base and (iii) from the word combination as a whole.

Lexicographic data types: The lexicographic data types that need to be given in a dictionary also differ according to the usage situation: for production, a user needs to know how to insert the collocation into a sentence (s)he is constructing, i.e. diasystematic marks and in particular syntactic properties must be given. A detailed account of linguistic properties that might need to be described in such a situation is given e.g. in Heid/Gouws (2006). Syntax is less important for text reception. Here the meaning explanation (or: gloss) is the main piece of information to be provided.

Phases of the search processes: Electronic dictionaries are quite similar to other information tools, in so far as their use typically involves two phases: the formulation of a query, i.e. searching, and, secondly, the output of a search result. For users, the tasks involved in these two phases are (i) formulating queries, which may include the selection of properties of the searched item; (ii) interpreting the search result, which may include the selection of the appropriate result, if several results are provided. Dictionary users need to go from their information need via both of the above phases to lexicographic data that fulfil their information need (cf. Tarp 2008). In entries of printed dictionaries, the first phase consists in macrostructural search, typically in the alphabetical lemma list of the dictionary, and the second phase in the reading and checking of the microstructural indications, e.g. for appropriate readings of the lemma, where the targeted collocations are included.

The Van Dale dictionaries of the 1990s used specific devices to guide users to collocation on the basis of a semantic profile of the readings of the headword, combined with category-based access to the collocations: the cijfer-punt-cijfer-code used in these dictionaries allowed users

to first identify the reading of the base they were interested in, and to then only check the collocations made up of the selected readings plus a collocate of a selected word category. As a consequence, users had to perform three steps in he first phase of the look-up process: (i) alphabetical search for the headword, (ii) identification of the appropriate headword reading, and (iii) selection of the appropriate word class of the collocate. This precise search led to a precise result: typically between one and five collocations from which it was easy to choose.

In long entries of other printed dictionaries, which do not use the Van Dale system, typically the first phase only consists of alphabetical search for the headword, and is consequently quick, while the second phase involves detailed checking of numerous indications and may thus require quite some time.

For electronic dictionaries, and for the two functions of text production vs. text reception, we summarize options for the two phases of the look-up process in table 1: different ways of stepwise entering collocation-related searches in an electronic dictionary, and the system feedback we expect, in terms of lexicographic data types. As a good collocation dictionary should contain a substantial amount of collocations, it may be necessary to present these in several steps; displaying them all on one screen may be confusing for the user. We number different options of access (cf. "type" column in table 1), for later reference.

Table 1. Possible search and output steps in collocation retrieval from electronic dictionaries

Function	**Type**	**Step**	**Search Criterion**	**Output**
Production	1	1 2	base lemma, reading semantic type	available semantic types collocation+properties
	2	1 2 3	base lemma category of collocate semantic type	available category types available semantic types collocation+properties
Reception	3	1	collocation	meaning of collocation
	4	1 2 3	collocate base category collocation	available base categories list of collocations meaning of collocation
	5	1 2 3	base category of collocate collocation	categories of collocates list of collocations meaning of collocation

2. Usability, usability testing, usability design

According to the ISO standard ISO 9241-11, t

2.1. *Usability of software*

Developers of software tools for the consumer market have a tradition in usability testing and usability design of their products. Both are methods from information science aimed at ensuring that users are indeed able to successfully work with information-related products. The key aspects of usability are the following:

- Effectiveness: does a product provide the service it is supposed to?
- Efficacy: is the time and effort a user must invest to get the product to provide its

service commensurate to the task?
- User satisfaction: does the product perform to the users' delight, and perhaps better than expected?

If the first two properties are measurable more or less objectively, the last one is subjective. Obviously, it is correlated with the first two ones: for example, a product which is not effective and/or not efficient will likely not cause users to be satisfied. For more details on usability, see e.g. Dumas/Redish (1999).

2.2. *Usability testing*

On the basis of the above rough outline definition of usability, information scientists have developed usability testing methods, as well as usability design techniques. The former include tests with experts and tests with subjects from the prospective user group, while the latter takes care to feed the results of the tests back into the process of the design of functions and user interaction aspects of the product under study.

The usability testing discussed in this paper is based on sessions with student users in a usability laboratory; such a laboratory consists of a standard computer on which the tested software is run, and of specialized software (i) to minute users' mousing and keystroke patterns, (ii) to record their oral statements during the use of the software (e.g. via think-aloud protocols) and (iii) to possibly record their gaze by means of an eye-tracker.

Usability laboratory tests typically are based on hypotheses about the software tested, and on three steps: a pre-test questionnaire for gathering data about the subjects and their expectations about the product, the actual laboratory test, and a post-test questionnaire. The laboratory test is based on tasks which the subjects have to carry out with the help of the software product, and which should be as close to the "real" use of the product as possible.
Post-test questionnaires are mostly used to capture aspects of user satisfaction: users are asked to rate features of the tools they tested, individually or comparatively.

2.3. *Interpreting usability tests*

The objective of laboratory tests is to find major difficulties in the use of the given software: if, e.g., more than half of the users independently run into trouble with a feature of a software product, this cannot be their fault, but it is then rather a design deficit. Thus laboratory tests can serve to identify problems of the effectiveness and efficiency of the tested product; for the latter, time to task completion is often a good indicator.

Usability tests are typically performed with small numbers of subjects, between 15 and 30. They are not meant to be interpreted quantitatively, beyond the simple identification of tendencies concerning design errors. Thus, it does not matter whether 46% or 54% of the users have a problem, but the fact that about half of the users do experience the same problem, indicates that the problem is real, and design-related.

3. Usability tests of electronic dictionaries

Above, we noted that electronic dictionaries are a type of software tools. They can thus be tested for usability issues like any other software product (cf. Heid 2011). Our basic motivation for doing so is that some electronic dictionaries we used were comparatively

complex, and it was not a priori clear how well users would be able to manipulate this complexity. In a second step, we intended to test the working methodology of usability design: we wanted to understand whether the development of interfaces of electronic dictionaries could be carried out as a cycle of prototyping and usability testing. Both are inspired by the observation that electronic dictionaries should evolve more towards their users, e.g. by investing into devices that make them appropriate for certain types of users.

3.1. *Dictionary usability testing and dictionary use studies*

Usability testing has to our knowledge so far not been applied to electronic dictionaries; its objective is related with that of studies on dictionary use, but it differs from these in a number of points. It is task-based, as dictionary use studies are, but the objective is more focused: not to understand in general how users work with a dictionary, but to test one or more dictionaries for problems at the level of functionality, of the graphical user interface (i.e. of data presentation) or of the metalanguage used.

Typical usability criteria are conformity to user expectations, consistency, error tolerance, learnability and memorability etc., which are much less in the focus of "traditional" studies on dictionary use. As mentioned, usability studies are mostly interpreted qualitatively, more than quantitatively. And usability testing starts from the assumption of a (close to) perfect tool and identifies possible imperfections. It can be used on single tools or in a setup where several dictionaries are compared.

3.2. *Tasks in usability tests of dictionaries*

In our tests, we started from a simplistic version of the typology of dictionary use situations discussed by Tarp 2008. We worked with text production and text understanding tasks. Text production tasks concern, for example, the choice of collocates for a given base (reading), or a situation where users must decide which of two possible collocation candidates is better (or the only adequate one). Reception-oriented tasks concern the identification of readings, or the translation from a foreign language to the mother tongue, etc.

3.3. *Subjects*

In our tests, we had German students of second and third year university courses in language-related fields working with the dictionaries. The first series of tests (cf. section 4) involved 31 students from different study programmes, while the second one (sect. 5) was administered to 13 students of translation science and specialized communication.

4. Usability tests of three different electronic dictionaries

Christina Bank has carried out a comparative usability evaluation of three electronic dictionaries (cf. Bank 2010):

- the German and Italian learners' dictionary ELDIT;
- the French learners' dictionary BLF, *Base lexicale du français* (now ILT; Bank used the version of January 2010);

– the German scientific online dictionary OWID.

Access to collocations differs in these dictionaries: ELDIT lists all collocations of a base lemma; BLF allows for access according to type 3 (cf. Table 1), and OWID according to type 1. In BLF, multiword items needed, at the time of the tests, to be entered by the user into two different search fields[1]; despite related indications in BLF's GUI, most users did not manage to handle this device – likely because it is contrary to their expectations (which are likely based on their experience with web search interfaces). We tested it on *c'est une question de vie ou de mort* (Bank 2010).

In OWID, to get access to collocations, several search steps need to be performed: entering the base lemma, selecting the appropriate reading, opening co-occurrence data. We tested users on the choice between *aus gutem Grund* and *mit gutem Grund*. The depth of the search path again affected the users' success with the tools: here phase 1 of the search process (see section 1.2) was complex, as several steps needed to be performed; their sequencing seemed not to be self-evident to all users.

Overall, the impression from the tests was that multi-step search massively reduced users' performance with the dictionaries and that there was a preference for simple "one-shot" solutions, similar to the use of a web search engine: enter the search term, hit "return" and get results which then are to be interpreted interactively.

5. From usability testing to usability design: Comparing mock-ups of a collocation dictionary

Methodological conclusions from the first set of tests were the following:

– in order to use the tests as a comparative diagnostic tool, as many parameters need to be kept identical as possible, such that differences in user behaviour and performance can be clearly attributed to the properties under review; in Bank's (2010) tests, a comparison between the three dictionaries tested was almost impossible: they concern different languages, are intended for different types of public, and they cover collocations with rather different devices;
– the tasks should be relatively homogeneous, to ensure the test persons have no unexpected difficulties with the tasks;
– testing effectiveness and efficiency separately from more subjective features is useful, to avoid interference.

5.1. *Prototypes of collocation dictionaries*

An assumption that emerged from the first set of tests, as well as from other user research on electronic dictionaries (e.g. Tiberius/Niestadt 2011) is that many users are so much used to web search engines that they expect electronic dictionaries to work the same way as these: the user inserts a word or word sequence, and the dictionary gives a set of answers from which the user selects. This is opposed to the lexicographic approach of profile-based dictionaries, as discussed by e.g. Tarp (2011) and realized, among others, by the Aarhus electronic dictionaries (lemma.com), where text production and text reception are different profiles. As the distinction between text production and text reception makes a major difference for the presentation of collocations, we decided to compare the following types of dictionary mock-ups:

– a “one-shot” dictionary that follows the search engine metaphor; it implements a simplified version of type 1 search (cf. table 1): the user inserts a base and gets all available collocations to choose from;
– a production-oriented profile-based dictionary, which exactly implements type 1 access;
– a reception-oriented profile-based dictionary, which implements type 4 access.

The illustrations in figure 1 and 2 show the interfaces of the one-shot dictionary and of the production-oriented profile-based dictionary, respectively.

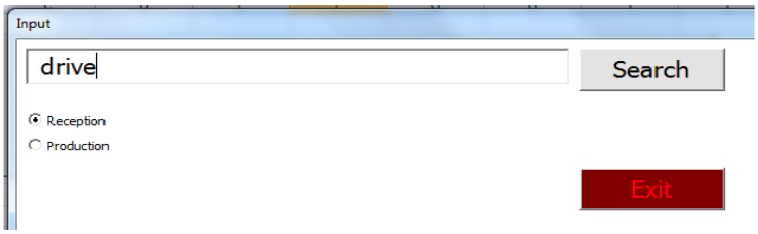

Figure 1. Interface of the mock-up one-shot dictionary.

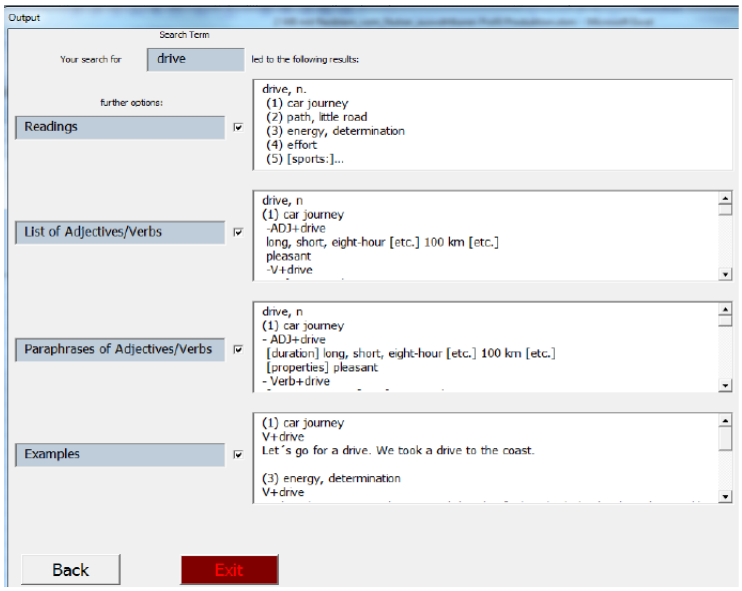

Figure 2. Interface of the mock-up production-oriented collocation dictionary.

5.2. *Test results*

A first result of this test suggests that, for the specific task of getting data about collocations, and for the specific public of translation students (who are rather oriented towards the production of high-quality texts), the profile-based dictionaries work better than the search-engine-like one. This is shown by success rates and time to task completion for both production and reception tasks. An example of measurement for a production task is given in figure 3; the task is to find out the correct verb (and its valency) for a given reading of a noun, e.g. *go for a drive* vs. *go on a drive*. In figure 3, the leftmost graph shows the one-shot dictionary, the middle and right graph the profile-based dictionaries.

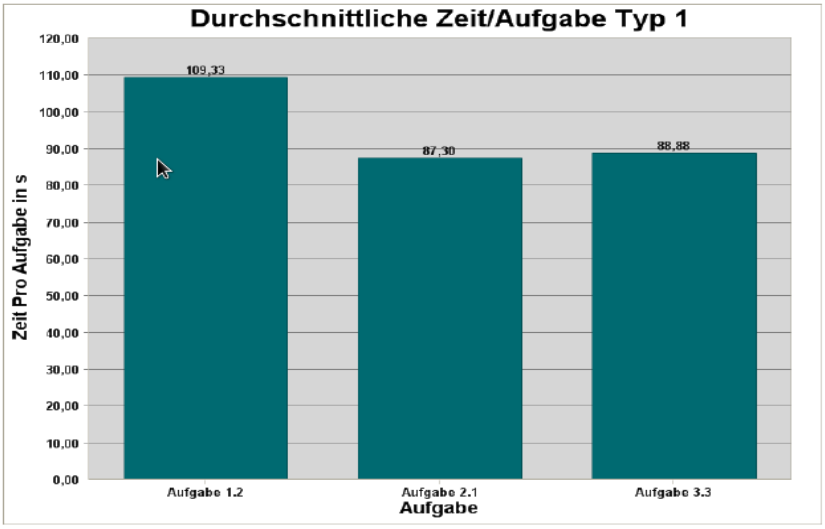

Figure 3. Time to task completion for a production-oriented collocation selection task: left to right: one-shot dictionary – production-dictionary – reception-dictionary.

These figures are consistent with user views on the possibilities of focused search (fig. 4) and on the clarity of the result presentation (fig. 5) of the three dictionaries, measured on a five-point Likert scale (from left (=very good) to right (very bad)). Users prefer the profile-based dictionaries (marked as No. 2 and 3, in green and red in figures 4 and 5) over the one-shot dictionary in these respects.

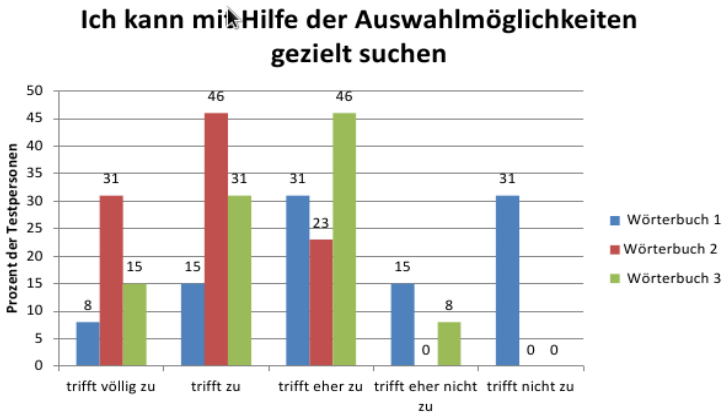

Figure 4. User opinions on possibilities of focused search in the three mock-up collocation dictionaries.

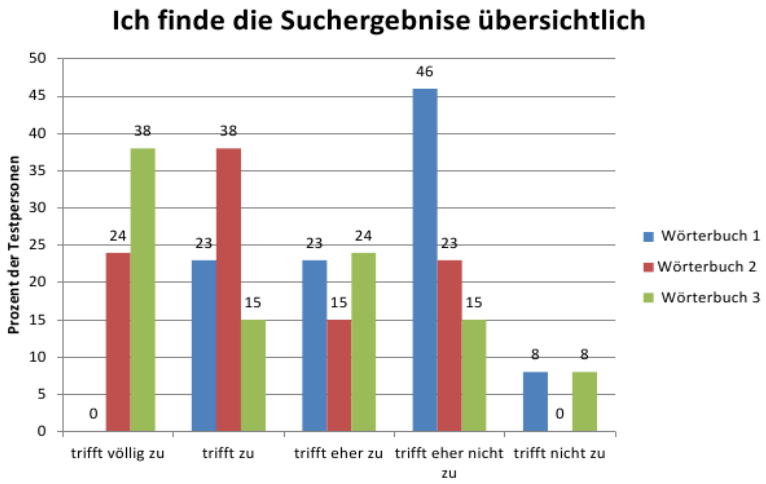

Figure 5. User opinions on the presentation of search results (clarity) in the three mock-up collocation dictionaries.

However, the fact that our production dictionary mock-up offered different user-definable degrees of detail (e.g. collocation candidates only by category or along with paraphrases, cf. figure 2) was perceived as a complication in the use of the tool rather than as an advantage, by about one third of the subjects. A comment made by most participants to the study was that they had needed some time to get acquainted with the profile-based dictionaries; this shows that these are not exactly conformant to users' a priori expectations. Our users however also noted that once the principles had been understood, the profile-based dictionaries were indeed more effective and efficient than the one-shot dictionary (learnability, memorability).

6. Conclusions and future research

We tested (among other features) the way in which collocation data are made accessible in three different "styles" of electronic dictionaries. Access to collocations typically presupposes a number of decisions on the side of the user: production vs. reception, readings of bases, usage properties of the collocations etc. Usability testing along the lines of information scientific theory and practice provides insight into the usability of lexicographic presentation devices: the complexity of the decisions that lead to appropriate data on collocations is to be distributed over both, the search interface and the design of the display of search results. In the case of a web search engine-like dictionary, all decision and selection tasks are left to the user. Our test subjects found this unhandy and thus inappropriate. Our profile-oriented production dictionary offers several levels of detail of the output, thereby easing the output interpretation, but at the price of complicating data input in the search step (cf. Types 3, 4 and 5 in table 1).

It seems to us that still better and more easy to use devices for searching and retrieving collocation data from electronic dictionaries are needed. These also should be as intuitive to use (expectation- conformant) as possible.

We also have the impression that the design of electronic dictionaries, especially for non-trivial tasks, like collocation search, profits from focused comparative usability tests applied to different variants of user interfaces. Our provisional conclusion is thus that a methodology whereby mock-ups of interfaces for electronic dictionaries are produced by means of rapid prototyping, followed by focused usability tests, may indeed give relatively precise answers to questions about the adequateness of certain concepts for lexicographic data presentation, at least for certain user(type)s.

As the collocation issue is not yet solved, we intend to analyse further presentational constellations. We also plan to enhance our approach to usability testing and to apply it also to other design issues in the domain of electronic dictionaries.

Note

[1] The way to enter collocation search into BLF and its successor ILT has since then been completely modified, cf. Verlinde 2011.

References

A. Dictionaries

***Base lexicale du français*.** 15 November 2011. http://ilt.kuleuven.be/blf/. (BLF)

***Elektronisches Lernerwörterbuch Deutsch-Italienisch*.** 15 November 2011. http://dev.eurac.edu:8081/MakeEldit1/Eldit.html. (ELDIT)

McIntosh, C. (ed.) 2009. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press. (OCDSE)

***Lemma A/S*.** 15 November 2011. http://www.lemma.com.

***Online-Wortschatz-Informationsystem Deutsch*.** 15 November 2011**.** http://www.owid.de/. (OWID)

B. Other literature

Bank, C. 2010. *Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen.* MA Dissertation, Universität Hildesheim.

Bartsch, S. 2004. *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Tübingen: Narr.

Bergenholtz, H. and I. Bergenholtz 2011. 'A Dictionary is a Tool, a Good Dictionary is a Monofunctional Tool.' In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), *e-Lexicography. The Internet, Digital Initiatives and Lexicography.* New york: Continuum, 187–207.

Dumas, J. and J. C. Redish 1999. *A practical guide to usability testing.* (Revised edition.) Norwood: NJ Ablex Publishing.

Hausmann, F. J. 2004. 'Was sind eigentlich Kollokationen?' In K. Steyer (ed.). *Wortverbindungen –mehr oder weniger fest.* Berlin: DeGruyter, 309–334.

Heid, U. 2011. 'Electronic Dictionaries as Tools: Towards an Assessment of Usability.' In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), *e-Lexicography. The Internet, Digital Initiatives and Lexicography.* New York: Continuum, 287–304.

Tarp, S. 2011. 'Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction.' In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), *e-Lexicography. The Internet, Digital Initiatives and Lexicography.* New York: Continuum, 54–70.

Tiberius, C. and J. Niestadt 2011. 'Dictionary use: A case study of the ANW.' In I. Kosem and K. Kosem (eds.), *Electronic lexicography in the 21st century: New applications for new users: proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia.* Ljubljana: Trojina.

Verlinde, S. 2011. 'Lexicographers' 'have-tos' in the electronic dictionary age.' In I. Kosem and K. Kosem (eds.), *Electronic lexicography in the 21st century: New applications for new users: proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia.*

Ljubljana: Trojina.