THE LEXICOGRAPHICAL PROCESS: ONLINE DICTIONARIES

ANNETTE KLOSA, INSTITUT FÜR DEUTSCHE SPRACHE MANNHEIM

1.   PREFATORY REMARKS

Lexicographic theory has only recently reflected on the lexicographical process as one of the major questions in research on the production and use of dictionaries. The lexicographical process has been almost exclusively described for printed dictionaries as in Dubois (1990), Landau (1984), Riedel/Wille (1979), Zgusta (1971), and (summarizing them) Schaeder (1987), while the process of producing encyclopedias has been widely ignored. Wiegand (1999: 134-246) analyzes, defines, and segments the lexicographical process for both (academic) printed dictionaries and electronic dictionaries (for a classification cf. 2.). He introduces the differentiation between a lexicographical process (with or without computer support) aiming at producing a paper dictionary and a computer-lexicographical process aiming at producing an electronic dictionary (Wiegand 1999: 242), and he regards the role of the computer in the lexicographical process in detail, pursuing initial ideas introduced in Knowles (1990: 1648 pp.). Müller-Spitzer (2003: 155 pp.) gives an advanced overview on different lexicographical processes and distinguishes them clearly by referring to the intended users of the product. She also suggests consideration of the lexicographical process when criticizing a dictionary. Engelberg/Lemnitzer (forthcoming) follow older descriptions for a lexicographical process, but hint also at electronic dictionaries. de Schryver/Prinsloo (2000a and b) suggest obtaining feedback from the target users of a dictionary during the process of compiling it to enhance the quality, which is possible especially for online dictionaries.

Further stimulation is to be expected from the ISO project on "Lexicographical production, marketing concepts and vocabulary" (launched in 2007) as discussed in Derouin/Le Meur (2008). The ISO project addresses every type of dictionary and will contain, among other things, a description of the "process of dictionary making (setting up a product project, compilation of data, data and corpora management, editorial work, product naming, publishing etc.)".

2.   CLASSIFICATION OF ONLINE DICTIONARIES

Different classifications of electronic dictionaries have been summarized by de Schryver (2003: 146 pp.). He suggests a primary distinction between dictionaries on a stand-alone computer and dictionaries on a networked computer (intranet dictionaries and internet dictionaries) (cf. Table 1).

ELECTRONIC DICTIONARY			
Dictionary on a stand-alone computer (one user at a time)		Dictionary on a network computer (many users at a time)	
Handheld dictionary	Robust-machine dictionary	Intranet dictionary	Online dictionary
One user uses a palmtop to access a dictionary stored on a small disk	One user uses a laptop/desktop to access a dictionary stored on a CD-ROM	A group of users use laptops /desktops to access a dictionary stored on a local mainframe	Users worldwide use laptops/desktops/ netbooks/mobile phones to access a dictionary stored on an online server

Table 1: Classification of electronic dictionaries (based on de Schryver 2003: 151).

As opposed to other electronic dictionaries, online (or Internet) dictionaries can be classified mainly according to their medial characteristics (cf. Table 2)

CHARACTERISTIC FEATURES	DICTIONARY TYPES
Original form of publication	First published as printed dictionary
	First published as electronic offline dictionary
	First published as online dictionary
Completeness	Completed dictionary
	Dictionary under construction
Hypertextualization	Hypertextualized dictionary
	Dictionary without hypertextualization
Interaction with the user	Dictionary with interaction
	Dictionary without interaction
Multimedia	Dictionary with text, illustrations, charts, or diagrams
	Dictionary with text and audio files
	Dictionary with text, illustrations, charts, diagrams, and audio files
	Dictionary without multimedia
Dictionary access	Dictionary with access through scrolling in an entry list
	Dictionary with access via hyperlinked list of headwords
	Dictionary with access via search options
	Dictionary with combined access
	etc.

Table 2: Classification of online dictionaries (based on Storrer/Freese 1996, Storrer 1998, and Storrer 2001).

The original form of publication affects to a great extent the online design, the degree and methods of hypertextualization, the degree of interaction with the user, the integration of multimedia, and the creation of ways of access to the dictionary. The criterion of completeness mainly has an effect

on the computer-lexicographical process: With online dictionaries under construction, this process is not completed, but continuously running until the intended end of the project.

An online dictionary under construction is not "a fixed object", but "an organic, changing database" (Prinsloo 2001: 141). Schröder (1997: 60) introduces the terms "Abschlusswörterbuch" [completed dictionary] vs. "Ausbauwörterbuch" [dictionary under construction], Lemberg (2001: 81) prefers "statisches Wörterbuch" [static dictionary] vs. "dynamisches Wörterbuch" [dynamic dictionary]. Online dictionaries under construction can be first published in print, then digitalized, and continually enlarged (e.g. OED, DRW), or first published online (e.g. DDN, ELEXIKO). In the following, we will focus on the computer-lexicographical process for online dictionaries under construction because other online dictionaries are not published until they are completed. Thus, their lexicographical process resembles the one for other electronic dictionaries as described in Wiegand (1999: 233 pp.).

3. COMPUTER-LEXICOGRAPHICAL PROCESS FOR ONLINE DICTIONARIES UNDER CONSTRUCTION

3.1 GENERAL REMARKS

Publishing a dictionary under construction on the Internet affects the computer-lexicographical process as a whole and also the phase of writing the dictionary.

3.1.1 MERGING OF PHASES

Theoretically, three phases of any lexicographical process can be distinguished: planning, writing, producing (Landau 1984: 227), each following one another. In practice, working on a dictionary usually (but especially for long-term academic projects) means that while still writing the dictionary, new material may be added to the corpus, corrections for entries already published are gathered, headwords and cross-references will be supplemented, and modifications in the original concept will become necessary. Permanent corrections and actualizations are usually part of a dictionary's writing phase. This is why Landau (1984: 267) suggests making plans for a revised edition of a paper dictionary even before the dictionary has been published. The phases of planning and writing, thus, do not really follow each other: For some time they run parallel. Only the production phase of paper dictionaries or completed electronic dictionaries will not start before the other phases are completed. Wiegand (1999: 141) observes that variations in the lexicographical process cannot change the borderlines between its main phases.

This is different for online dictionaries under construction (see Klosa 2007: 80 pp.). Producing an online dictionary may begin before the phase of writing is finished: online dictionaries can be published step-by-step. Thus, all phases of the computer-lexicographical process (planning – writing – producing) merge (in detail cf. 3.2.7), giving yet unknown flexibility to the lexicographer (Klein 2004: 286).

3.1.2   LEXICOGRAPHIC WORK MODULES

The phase of writing an online dictionary under construction is also affected by the incompleteness of the dictionary before publication. While most dictionaries are written following the alphabetic order of the entries, this is not necessarily so for online dictionaries. In an electronic dictionary, the presentation of the headwords must not follow the alphabet: Entries can be searched, found, and read without an alphabetical access structure. Instead, other search criteria (e.g. all derived nouns, all verbs in a specific semantic field; for more examples see Engelberg/Lemnitzer [forthcoming]) may be applied. This implies that writing the dictionary may follow specific types of headwords (see Storrer 2001: 61 p.) or may happen in "modules" (Haß 2005: 13 pp.; Klein 2004: 300 p.). Modules are components of lexicographic work, comprising words with specific criteria. Modules can also be defined by the frequency of a headword in the corpus. To work on words of the same word class may lead to consistent descriptions (see Storrer 2001: 63); to work on words with a certain frequency is especially interesting because of their relevance for learners or other groups of dictionary users.

3.2   DETAILS OF A COMPUTER-LEXICOGRAPHICAL PROCESS

Adapting Wiegand (1999: 233 pp.), a computer-lexicographical process for a corpus-based online dictionary under construction will be described in detail below, naming the phases of production, listing actions in those phases, and describing the expertise needed for the actions (cf. also Table 3). Since hypertextualization, interaction with the user, and multimedia are characteristic features of online dictionaries (cf. 2.), there will be special focus on actions related to those features.
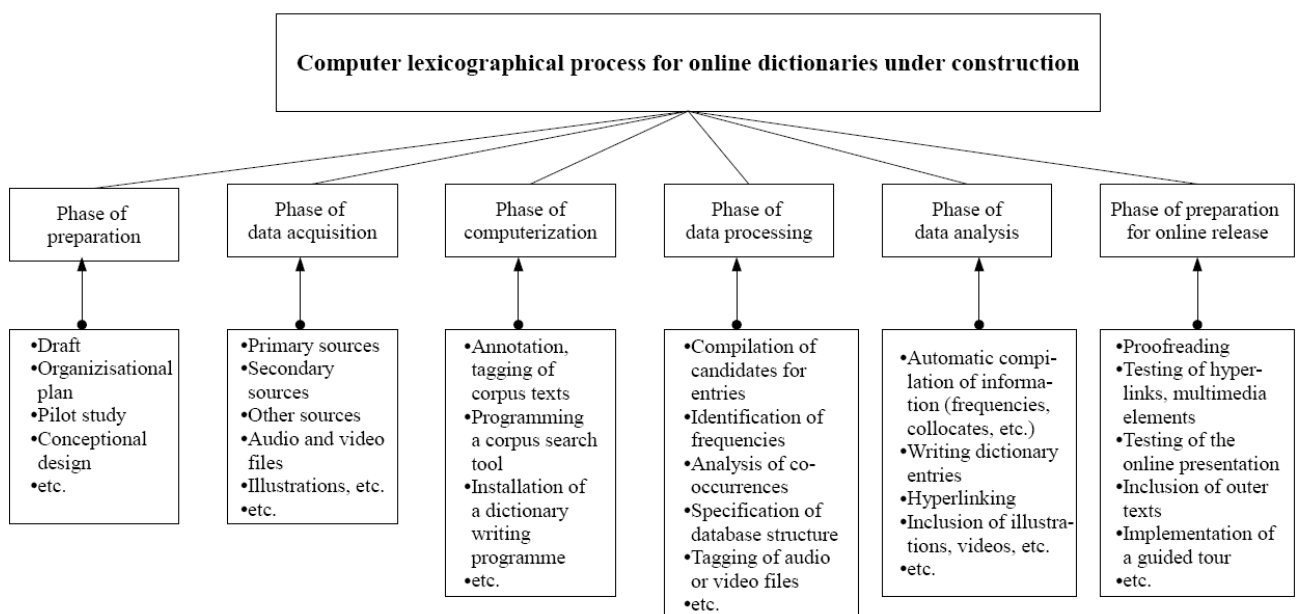

Table 3: Computer-lexicographical process for online dictionaries under construction

3.2.1   PHASE OF PREPARATION

As for any other dictionary, the computer-lexicographical process for an online dictionary starts with writing a draft, followed by the phase of planning. A dictionary plan for an online dictionary consists of an organizational plan and the conceptional design. The organizational plan should contain details on financing, the work flow, a schedule, and information on personnel. An online dictionary (as any

other electronic dictionary) cannot be produced only by lexicographers, but corpus linguists, computational linguists, computer scientists, text technologists, graphic artists, and others probably participate in the process. A pilot study can be prepared containing a preliminary programme of lexicographic information or details on the intended list of headwords. As a characteristic for online dictionaries under construction, the organizational plan should also contain a schedule for publication, describing which modules (cf. 3.1.2) will be published when.

The dictionary's conceptional design contains the lexicographer's manual (most suitably in electronic form), describing the corpus, the dictionary matter, how entries are written, and how information has to be tagged in order to allow content-oriented access to the dictionary. Corpus linguists have to work on the corpus design together with lexicographers; computational linguists should be responsible for mapping the data structure upon consultation with lexicographers. The dictionary's conceptional design also contains samples for entries. It is important to plan the technical support for the project (hardware and software, personnel). This should be discussed between the participating lexicographers, computational and corpus linguists, and computer scientists. Besides all that, for online dictionaries, a graphic artist should (upon consultation with the lexicographers) draft the online design and plan different views adaptable to different user needs, discussing it with lexicographers and Internet experts. Planning access structures, ways of hypertextualization, interaction with the users, and the assignment of multimedia elements is also something worth discussing with all staff members. Lexicographic modules have to be defined. Finally, the conceptional design should contain a plan on how to monitor users and how to reattach them to the dictionary, discussing the concept with sociologists and the implementation with computer scientists.

3.2.2   PHASE OF DATA ACQUISITION

The basis for an online dictionary (as for any other dictionary) includes primary sources, that is, an electronic text corpus, secondary sources, for example, other paper or electronic dictionaries, and other sources, that is, grammar books, lexical databases. The selection of the sources depends, of course, on the decision to publish a corpus-driven or corpus-based dictionary. While corpus linguists should be responsible for acquisition of corpus texts, it is the lexicographers' task to obtain secondary and other sources. Besides providing those, audio and video files, illustrations, diagrams, etc. have to be supplied in the phase of data acquisition of an online dictionary (as of any other electronic dictionary).

3.2.3   PHASE OF COMPUTERIZATION

As for any lexicographical process using computers, in a computer-lexicographical process for online dictionaries there is a phase of computerization (see Wiegand 1999: 238) when corpus texts will be annotated, tagged, and lemmatized, when a corpus search tool has to be programmed, and when a dictionary writing system has to be installed. The technical equipment for data storage and retrieval has to be supplied. In all this, lexicographers can only have an advisory position, while corpus linguists and computer scientists perform these tasks.

3.2.4   PHASE OF DATA PROCESSING

In this phase of writing a corpus-based online dictionary, the corpus texts can be processed by corpus linguists in different ways, for example, compilation of a list of candidates for entries, identification of frequencies of headwords, definition of frequency layers, and analysis of co-occurrences for headwords. The database structure is specified by computer scientists. Other sources such as audio-

files or videos need to be recorded and tagged for online and other electronic dictionaries (lexicographers should be responsible for this).

3.2.5 PHASE OF DATA ANALYSIS

In this phase, lexicographers, corpus linguists, and computational linguists participate since information can be compiled lexicographically or automatically. Many online dictionaries contain automatically compiled information (usually combined with lexicographically compiled information, e.g. DWDS), for example, information on frequency, information on collocations, information on paradigmatic partner words, and automatically extracted corpus citations. This information is often displayed in tables, lists, or graphs. Computer and corpus linguists are responsible for this information, retrieving it from the corpus while employing different ways of programming. They are also responsible for linking entries to the corpus.

Lexicographers analyze the data in basically the same way as for a paper dictionary. They check the headword candidates and decide on headword types and lemmatization, thus, creating the list of headwords. For online dictionaries under construction, lexicographers also define lexicographic modules. Most of their work, of course, goes into writing dictionary entries, analyzing corpus data (concordances, KWICs, citations, etc.), specifying different meanings, defining, choosing citations, checking secondary sources, etc. While this basically does not differ from writing entries for paper dictionaries, lexicographers working on online dictionaries have to do more: Cross-references may not only be placed within one entry, between different entries in the dictionary, or between entries and the dictionary's outer texts, but can also be used to hyperlink entries with other online sources (encyclopedias, dictionaries). Lexicographers also have to sift illustrations, charts, etc., audio or video files and choose those to be either incorporated into an entry or linked with it.

3.2.6 PHASE OF PREPARATION FOR ONLINE RELEASE

In this phase, it is the lexicographers' task to ensure the dictionary's quality by extensive proofreading and testing. This comprises in particular: content related and formal proof reading (of printed versions of the articles) and testing (in a test environment) of hyperlinks, multimedia elements, the online presentation of each entry and the possibilities of information retrieval and interactive elements offered. Lexicographers in this phase still may add entries, correct entries (errors, poor definitions, ambiguities, inconsistencies, etc.), complete entries (hyperlinks, new senses, etc.), or delete entries (for details see Landau 1984: 256). The dictionary's outer texts have to be written or complemented and linked with the entries. A guided tour through the dictionary may be designed, implemented, and tested. In this, lexicographers are supported by computer scientists, who provide the test environment and also test their programming.

3.2.7 RECAPITULATION

A computer-lexicographical process is a process in which lexicographers as well as computer specialists (corpus linguists, computational linguists, and computer scientists) participate. Other expertise on questions of Internet design and usability of Internet sites has to be sought and included to be able to publish an online dictionary of good quality and reliability. To talk of a computer-lexicographical process rather than a lexicographical process is, therefore, justified when the process of making an online dictionary (either completed or under construction) is described. In all the phases of production, the computer plays an essential role, not only, as advocated in Knowles (1990:

1648), on the level of logistics for writing and presenting the dictionary (either on paper or electronically).

It is important to understand that the phases described above do not necessarily and always pass by one after the other when working on an online dictionaries under construction (see Lemberg 2001: 86). One module (e.g. all speech verbs) may still be in the phase of planning; another module (e.g. all words in a word family) may already be in the phase of data analysis, and a third module (e.g. all words with a specific frequency) may already be published online. It is also possible that more data has to be obtained and processed for a new module, while for other modules the phases of data acquisition and processing are over. Modules not only as a whole but also single entries within a module can be situated in different phases if the dictionary allows online release of single entries. Thus, a number of entries in the same module may already be online; others may be prepared for online release while still others are in the phase of data analysis. Finally, entries of a module may be released even if all lexicographic information is not gathered yet. When, for example, automatically and lexicographically compiled information is combined in a module, the entries may initially be released online with only the automatically compiled information (e.g. information on orthography, collocations). The lexicographic analysis will be added later so that one entry is simultaneously in the phase of data analysis and already online.

If an online dictionary asks for user feedback when online, the outcome of this may also influence the dictionary-making process (as described in de Schryver/Prinsloo 2000a and b). From this, it follows to define a computer-lexicographical process for an online dictionary under construction not as simply linear, but as circular. While other lexicographic processes lead to an end (i.e. the publication of the dictionary), theoretically, working on an online dictionary under construction could go on forever. An online dictionary under construction is an open system (see Schröder 1997: 16). This also implies that organizing (calculating, segmenting, regulating) and controlling this process (see Wiegand 1999: 134) is of the utmost importance. Training the staff to work simultaneously on many different things in different stages is also helpful (as discussed by Landau 1984: 264). It is also very important to render this process transparent for users (cf. 4.).

4. ONLINE DICTIONARIES AND THEIR USERS

Defining modules and deciding on the order of working on them in an online dictionary project is often guided by practical conditions instead of lexicographic considerations. If, for example, an online dictionary project computer linguistic competence is strong, then modules to be compiled automatically will be defined and published first, not necessarily respecting the users' needs. Another example is the decision on the publication rate (daily, quarterly, or only when a whole module is finished), which projects make without asking users about their preference. In order to answer these questions and to ensure high quality, usage research is indispensable, thus, leaving the experimental stage of online lexicography behind (see Klosa 2007: 85; other de Schryver/Prinsloo 2000b: 199).

This is especially important because online dictionaries under construction may leave users unsatisfied. Many generally tend to expect that dictionaries cover all the words from A to Z in a language and that a dictionary is a reliable source of information. It is hard to say how users will adapt to online dictionaries under construction in the future, but as Internet users become more and more accustomed to collaborative information sites such as WIKI, it is to be expected that they will also learn what to expect from dictionaries under construction. This process can be advanced when online dictionaries stay in contact with their users. Feedback concerning mistakes or missing words and contributions from persons outside the dictionary staff (e.g. specialists) may be asked for

through the Internet's communication facilities as wikis or e-mail in a user-friendly way, thus, also promoting the user's linking with the dictionary.

Looking up a word in a dictionary under construction may also frustrate users when they go back to the dictionary and no longer find the information they had found before or if the information has been changed. An online dictionary that changes or delets entries in an unmarked way cannot be used when reliability and quotability are necessary. Any dictionary under construction should, therefore, only offer reliable and quotable information (Storrer 2001: 65). This contrasts with the Internet's general features of rapid growth and constant change, but it can be achieved when changes, additions, or deletions are recorded on the dictionary's site (e.g. the list of quarterly updates of the OED). An archive with older versions of entries or the list of headwords may be useful, thus, also providing data sustainably and establishing the dictionary as a reliable source of information. In general, the question of versioning and archiving an online dictionary under construction is still open (see Klosa 2007: 83).

Further research should not only investigate how users cope with online dictionaries under construction but should also reflect more closely the role of the lexicographer in a computer-lexicographical process for online dictionaries in general.

5. SELECT BIBLIOGRAPHY

5.1 DICTIONARIES

DDN = Den Danske Netordbog. http://www.ordbogen.com/ordboger/ddno/index.php?dict=ddno (17.06.2009).

DRW = Deutsches Rechtswörterbuch Online. http://drw-www.adw.uni-heidelberg.de/drw/ (17.06.2009).

DWDS = Digitales Wörterbuch der deutschen Sprache. http://www.dwds.de (19.06.2009).

ELEXIKO = elexiko – Online-Wörterbuch zur deutschen Gegenwartssprache. http://www.owid.de/elexiko_/index.html (17.06.2009).

OED = The Oxford English Dictionary Online. http://www.oed.com/ (17.06.2009).

WIKI = Wikipedia. The free encyclopedia. www.wikipedia.org (17.06.2009).

5.2 OTHER REFERENCES

Derouin, M.-J./Le Meur, A. (2008): ISO-Standards for Lexicography and Dictionary Publishing. In: Bernal, E./De Cesaris, J. (eds.): Proceedings of the Thirteenth EURALEX International Congress. Barcelona, 663-669.

Dubois, C. (1990): Considérations générales sur l'organisation du travail lexicographique. In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. (= HSK 5.2), Berlin/New York, 1574-1588.

Engelberg, S./Lemnitzer, L. (forthcoming): Lexikographie und Wörterbuchbenutzung. 4th, revised edition, Tübingen.

Haß, U. (2005): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York.

Klein, Wolfgang (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: Scharnhorst, J. (ed.): Sprachkultur und Lexikographie. Frankfurt a. M., 281-308.

Klosa, A. (ed.) (2007): Lexikografische Portale im Internet. Mannheim.

Knowles, F. E. (1990): The Computer in Lexicography. In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. (= HSK 5.2), Berlin/New York, 1645-1672.

Landau, S. L. (1984): Dictionaries. The Art and Craft of Lexicography. New York.

Lemberg, I. (2001): Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In: Lemberg, I./Schröder, B./Storrer, A. (eds.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen, 71-91.

Müller-Spitzer, C. (2003): Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen. In: Lexicographica 19, 140-168.

Prinsloo, D. J. (2001): The Compilation of Electronic Dictionaries for the African Languages. In: Lexikos. Journal of the African Association for Lexicography 11, 139-159.

Riedel, H./Wille, M. (1979): Über die Erarbeitung von Lexika. Leipzig.

Schaeder, B. (1987): Germanistische Lexikographie. Tübingen

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

de Schryver, G.-M. (2003): Lexicographers' Dreams in the Electronic-Dictionary Age. In: International Journal of Lexicography 16, 2, 143-199.

de Schryver, G.-M./Prinsloo, D .J. (2000a): The Concept of "Simultaneous Feedback": Towards a New Methodology for Compiling Dictionaries. In: Lexikos. Journal of the African Association for Lexicography 10, 1-31.

de Schryver, G.-M./Prinsloo, D. J. (2000b): Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In: Heid, U./Evert, S./Lehmann, E./Rohrer, C. (eds.): Proceedings of the Ninth EURALEX International Congress. Stuttgart, 197-209.

Storrer, A. (1998): Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In: Wiegand, H. E. (ed.): Wörterbücher in der Diskussion III. Tübingen, 106-131.

Storrer, A. (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, I./Schröder, B./Storrer, A. (eds.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen, 54-69.

Storrer, A./Freese, K. (1996): Wörterbücher im Internet. In: Deutsche Sprache 24, 1996, 97-136.

Wiegand, H. E. (1999): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York.

Zgusta, L. (1971): Manual of Lexicography. The Hague.