Log Files Can and Should Be Prepared for a Functionalistic Approach

Henning Bergenholtz, *Department of Afrikaans and Dutch, University of Stellenbosch, Stellenbosch, Republic of South Africa and Centre for Lexicography, Aarhus School of Business, Aarhus, Denmark (hb@asb.dk)*
and
Mia Johnsen, *Centre for Lexicography, Aarhus School of Business, Aarhus, Denmark (miajohnsen@gmail.com*)

Abstract: User surveys of printed dictionaries may be characterised as non-representative and non-realistic laboratory tests, often with retrospective questions based on memory. Log file analyses concerning the use of Internet dictionaries, on the other hand, are based on large numbers of users and look-ups. However, log file analyses have also been characterised by a juggling of numbers based on data calculations of limited direct relevance to practical and theoretical lexicography. This article proposes the development of lexicographically relevant log files for the use in log file analyses in order to give a true picture of how and why different dictionaries are employed for different purposes.

Keywords: LEXICOGRAPHY, LOG FILES, DICTIONARY, INTERNET DICTIONARY, SEARCH OPTIONS, DICTIONARY FUNCTION, RECEPTION, TEXT PRODUCTION, TRANSLATION, COMMUNICATIVE FUNCTION, COGNITIVE FUNCTION, DICTIONARY USE, USER SURVEY, LEMMA, DICTIONARY ITEM, USER NEEDS

Opsomming: Loglêers kan en behoort voorberei te word vir 'n funksionalistiese benadering. Gebruikersopnames van gedrukte woordeboeke kan gekarakteriseer word as nieverteenwoordigende en nierealistiese laboratoriumtoetse, dikwels met retrospektiewe vrae gebaseer op geheue. Loglêerontledings betreffende die gebruik van Internetwoordeboeke, aan die ander kant, is gebaseer op groot getalle gebruikers en raadplegings. Loglêerontledings word egter ook gekenmerk deur 'n gegoël met getalle gebaseer op databerekeninge van beperkte direkte tersaaklikheid vir die praktiese en teoretiese leksikografie. Hierdie artikel stel die ontwikkeling van leksikografies tersaaklike loglêers voor vir gebruik in loglêerontledings om 'n ware beeld te gee van hoe en waarom verskillende woordeboeke vir veskillende doeleindes aangewend word.

Sleutelwoorde: LEKSIKOGRAFIE, LOGLÊERS, WOORDEBOEK, INTERNETWOORDEBOEK, SOEKKEUSES, WOORDEBOEKFUNKSIE, ONTVANGS, TEKSPRODUKSIE, VERTALING, KOMMUNIKATIEWE FUNKSIE, KOGNITIEWE FUNKSIE, WOORDEBOEKGEBRUIK, GEBRUIKERSOPNAME, LEMMA, WOORDEBOEKITEM, GEBRUIKERSBEHOEFTES

1. Better Dictionaries through the Use of Log Files

The last decade has seen an explosive growth in the number of available online dictionaries. Significant technological improvements have led not only to an increase in the use of the Internet through high-speed connections, but also to new opportunities in the field of lexicography. In recent years, the focus has shifted from being mainly on the dictionary itself and the compilation thereof to the dictionary user — what does the user expect from the dictionary, and how do lexicographers best cater for the users' needs? This is one of the pivotal issues in the lexicographic debate and one that may be addressed through the use of the new technological possibilities.

The advent of online dictionaries has given rise to new ways of studying user habits and needs in order to improve a given dictionary in accordance with its users' wishes. Previously, lexicographers were forced to resort to such methods as user surveys, tests or protocols, which, due to their inherently subjective nature, do not necessarily provide a realistic picture. With online dictionaries, two new approaches have become accepted: log files and direct feedback. These methods may be used concurrently, and as De Schryver and Joffe (2004) describe in their article on the SeDiPro project, the results gained from the two methods will often supplement each other. In other instances, however, as was the case in direct feedback from users of the *Danish Internet Dictionary*, users request rarely used words which are not evident from the log files to be included in the dictionary, and the methods, rather than the results, may thus supplement each other.

One contribution on the use of log files for improving Internet dictionaries is Bergenholtz and Johnsen (2005). The article demonstrates that log files may be used to this end in various ways, most importantly to discover so-called lemma lacunae (i.e. words that could/should have been included in the dictionary), but also to discover other problematic issues. The aim is not to give a detailed account of this article, but merely to sum up the main points being made about this type of log file analysis.

As for lemma lacunae, the implications of log file analyses are obvious. Lexicographers can periodically analyse the log files and add words users have searched for without finding them, thus increasing the hit rate and the usability of the dictionary. This practice was also used in the SeDiPro project, resulting in an increase in the hit rate from 67% to 75% (De Schryver and Joffe 2004). Which words to include depends on the dictionary's functions, intended user group and genuine purpose and whether it is composed as a minimising or a maximising dictionary. Log file analyses may also reveal frequent searches for 'missing' orthographic forms. In Bergenholtz and Johnsen (2005), the passive and the imperative are mentioned as examples that may subsequently be added to the dictionary, thus making it possible for the users to search for these forms. This has, in fact, been done in the *Danish Internet Dictionary* as a result of log file analyses and feedback from users. Frequent misspellings or searches for non-existing words are also revealed in the log files and provide lexicographers with various options in terms of helping the user. The incorrect word may be added to the dictionary with a reference to the correct term, or the user may automatically be redirected to the correct form of the word. These strategies are, of course, only feasible in the case of frequently occurring mistakes.

There is no doubt that log file analysis is an extremely useful tool in discovering and providing a solution to such issues as those described above. The greatest advantage of the method, however, its objectivity, also constitutes a limitation as the hard data do not reveal the motivation for a search, e.g. whether the user in question uses the dictionary in connection with reception, production or translation, or whether the user actually found the answer to his/her question. This issue will be discussed later in this article.

2. Log Files, Corpus-based Lemma Selection and Search Options

In a recent article by De Schryver et al. (2006), a Swahili–English dictionary project is described and the use of corpus-based lemma selection is criticised on a number of counts. The intention was to investigate whether users really look up the most frequent words in a corpus on the basis of an analysis of the log files for the dictionary, the authors concluding that this is not the case. The same conclusion was reached by Johnsen (2005) in an analysis of the log files for the *Danish Internet Dictionary*. This relation between corpus-based lexicography and log files is very interesting as it challenges a generally accepted and widely used approach to lemmatisation. In De Schryver et al. (2006), it is concluded that corpus-based lemma selection is a valid strategy in connection with minimising dictionaries, but not in the case of maximising dictionaries in which users also expect to find less common words. The authors suggest that corpora may be used as guidance, but that additional software modules should be applied to help users, e.g. a module for redirecting frequent misspellings as also suggested by Bergenholtz and Johnsen (2005) and a module that deals with multi-word units (MWUs). The use of this latter module does indeed seem to be a valuable tool, and the problem posed by MWUs gives rise to a discussion of the definition of a lemma and which units should be awarded lemma status in a given dictionary.

In our consideration, log files are a useful supplement to corpus-based lemma selection as they may be used to reveal lemma lacunae, frequent misspellings, frequent searches for MWUs, etc. Whether the shortcomings of corpus-based lemma selection should be remedied on the basis of log file analysis, through the use of software modules or a combination of these is for the lexicographers to decide, but it does indeed stand to reason to employ the new technological possibilities offered by the Internet. Spelling and typing mistakes may, for example, be rectified through an integrated automatic spell checker as in common text editing programmes such as Microsoft Word (cf. Bergenholtz 2005 on the status of Word's spell checker as a dictionary in its own right), or the user may be presented with alternative suggestions if the search string is not found as is the case in the search engine Google. The latter strategy has also been implemented in the *Danish Internet Dictionary* where the user is presented with a list of 10 alternative suggestions ranked according to match frequency in the case of unsuccessful searches. The percentages shown after each suggestion indicate the probability of the suggestion being the correct alternative to the search string entered by the user. If the user searches for a word with incorrect spelling and the spelling mistake is a minor one with just one incorrect letter, the correct spelling of the word will usually occur as the first, and thus most likely, suggestion. This for example applies if the user searches for the incorrect spelling *hiraki* (hierarchy, the correct spelling being *hierarki*), in which case the following result appears (translations by HB and MJ):

hierarki (86%) (hierarchy)
hierarkisk (75%) (hierarchic)
harakiri (71%) (hara-kiri)
hierarkisere (67%) (sort hierarchically)
hik (67%) (hiccup)
hårlak (67%) (hair spray)
hiking (67%) (hiking)
hijacking (67%) (hijacking)
hak (67%) (notch, dent)
hagiografi (63%) (hagiography)

These options are in line with De Schryver et al. (2006) who, as mentioned above, suggest the use of various software modules to help the user and thus increase the hit rate.

De Schryver et al. (2006) mention another issue concerning lemmatisation which is specific to Swahili and other inflecting languages: the question of whether to include only word stems as is the common practice in Western lexicography, or whether to include also full word forms on the basis of a corpus. In this connection, Bergenholtz and Johnsen (2005) are criticised for being naïve in conducting their log file analysis on the basis of lemma strings only and not taking this issue into account. It should be noted, however, that the *Danish Internet Dictionary* also allows searches for morphemes, inflected forms of a lemma and random parts of words, and the various search options occur from the log files. Seeing that most users search for 'the lemma is' rather than 'the lemma begins with', 'the lemma ends with' or 'the lemma contains' (Bergenholtz and Johnsen 2005), it seems that this strategy is feasible for an agglutinative language like Danish, where the rules of word formation and grammar differ, for example, from Swahili. The lemma selection process and, consequently, the resulting log files, will necessarily to a certain extent be language-dependent as different languages require different considerations as far as dictionary compilation is concerned.

The focus of Bergenholtz and Johnsen (2005) was on improving the lemma selection on the basis of log file analysis, but other possibilities are also taken into account. It is suggested that improved search options, e.g. the possibility of searching directly in every field of the dictionary rather than just the lemma field, would result in far more detailed log files, but that the use of such detailed log files has yet to be described. This article aims to elaborate on the issue and present suggestions concerning the practical use and lexicographic relevance of such log files.

As regards the compilation of detailed log files, advanced search options have been implemented in a Danish online dictionary of music terms, the *Danish Music Dictionary*. In this dictionary, a search for a specific term will be conducted not only in the lemma field, but also in for example the definition field. The *Danish Music Dictionary* has been available on the Internet since August 2006 and was therefore not yet accessible at the time when Bergenholtz and Johnsen (2005) was written. The addition of these enhanced search options may be regarded as a further development of the log file theory. It serves a dual purpose — on the one hand, it increases the usability of the dictionary for its users, and on the other hand, it is part of a strategy for compiling more detailed, and thus more useful, log files.

The issue of MWUs was briefly touched on above, and the question of whether or not it should be possible to search for MWUs depends to some extent on the dictionary in question. It is less relevant to include MWUs in a dictionary like the *Danish Internet Dictionary* where it would be more useful to add advanced search options to enable users to search, for example, in collocations and examples, but it is highly relevant in a recently released online dictionary of Danish idiomatic expressions, the *Danish Phraseological Dictionary*. As idiomatic expressions inherently consist of more than one word, the dictionary compilers are testing a completely new approach to search options in this dictionary. The log files for the *Danish Phraseological Dictionary* will be different from lemma-oriented log files, and the initial results and implications of the strategy are described in sections 5 and 6 of this article.

3. Comparative Surveys and the Status of the Lemma

Log files are used ever more widely in the field of lexicography, not only in Denmark and South Africa, but also internationally. In Johnsen (2005), a survey of five Internet dictionaries representing different languages is carried out with the aim of establishing two facts: (1) the extent to which Internet dictionaries are used and whether the use of Internet dictionaries is increasing, and (2) whether the log files of these dictionaries show similarities in terms of user behaviour. The dictionaries in question are the above-mentioned *Danish Internet Dictionary*, *Eurodicautom* (a polylingual database compiled by translators in the European Union), *Wortschatz Deutsch* (a German and German–English dictionary), *Cambridge Dictionaries Online* (various monolingual English dictionaries) and *Bokmålsordboka* (a monolingual Norwegian dictionary).

The bare figures show that all five dictionaries are widely used with an average of 6 000–6 500 daily searches in the *Danish Internet Dictionary*, 162 074 in *Eurodicautom*, 329 657 in *Wortschatz Deutsch*, 205 480 in *Cambridge Dictionaries Online* and 7 932 in *Bokmålsordboka*. The survey also reveals that the number of queries have increased over time, particularly in the first years of a given dictionary's life. In order to evaluate the user behaviour, the top 20 queries of the dictionaries as established by the log files were compared (with the exception of *Eurodicautom* as no data were available). The top 20 lists are included as Appendix 1 with English translations by HB and MJ. Interestingly, the log files for the Danish and the Norwegian dictionaries show significant similarities as both mainly contain ordinary everyday words from practically all word classes. This is, however, not the case for the English and the German dictionaries. The log files from *Wortschatz Deutsch* show that seven of the top 20 words are the names of actual persons, and a number of words related to lexicography and linguistics also appear in the list. With the exception of one adjective, the top 20 for *Cambridge Dictionaries Online* contains only nouns and verbs. The log files from *Wortschatz Deutsch* thus differ most significantly from the log files of the other three dictionaries. This may be due to the fact that users with different needs search for different items, and differences in log files from various dictionaries may thus be linked with the intended function(s) of a given dictionary. It might be imagined that the log files of a dictionary compiled for translation purposes differ from those of a dictionary intended for text production or reception. As a clear function is only stated for the *Danish Internet Dictionary*, i.e. text production, it is difficult to say whether this has any effect on the differences and similarities in the log files described above. It is, however, relevant for lexicographers to pay attention to whether dictionary users search for words that should be included in the dictionary in accordance with its function(s) or whether they might have misconceived the purpose. If the log files for the *Danish Internet Dictionary* contained as many queries for the names of actual persons as those for *Wortschatz Deutsch*, it would clearly indicate that the users in question do not understand the purpose of the dictionary. In other cases, it may be that the lexicographer has misjudged the users' needs or has been unable to communicate the dictionary's purpose clearly.

Comparative surveys and log file analyses as those described above may be interesting, and they do provide lexicographers with new knowledge, but they still do not reveal the answers to the real questions: Who are the users, why did they use the dictionary, and did they find what they were looking for? In a dictionary allowing only searches for lemmata, log file analysis has certain limitations; cf. section 1 of this article. Restricting searches to the lemma field poses two problems: what is a lemma, and what did the users do with the information they found? In order to achieve the full benefits of log file analysis, changes in the basic conception of Internet dictionaries are thus required. For one thing, it is necessary to implement better, more advanced search options, but, more importantly, it is crucial to utilise the full potential of the electronic medium and move away from the traditional, lemma-oriented way of thinking. A structured, lemma-oriented approach is undoubtedly necessary in printed dictionaries, but not so in electronic dictionaries. Ideally, each building block of the dictionary should be entered into the database only once, and different users may choose to view the information in different ways, thus eliminating the need for a macrostructure (Bergenholtz 2005). This raises the question, then, of whether the lemma itself has also been made superfluous as every item in a dictionary may move into lemma position. In the *Danish Phraseological Dictionary* mentioned above, whatever the user chooses to enter into the search field is the lemma. The user may for instance search for the expression *stå med aben* (equivalent to the English expression: *be left holding the baby*). The dictionary then produces a list of expressions containing or relating to this idiom as it searches the entire database rather than just fields designated 'lemma'. The user may also choose to search for the exact word or phrase entered in the search field, entries containing the word(s) entered here, entries beginning with the word(s) or entries ending with the word(s). Furthermore, the user has three options regarding the purpose of the dictionary use: *hjælp til at forstå en tekst* (help in connection with understanding a text), *hjælp til at skrive en tekst* (help in connection with writing a text) and *hvis du vil vide mere* (further information). In other words, it is possible to specify or narrow down a search to exactly find the information required.

4. User Surveys

In this article, it was previously mentioned that log files may be used to collect data leading to the improvement of existing online dictionaries. This is not the only possible use of log file analyses. If one wishes to achieve new and general knowledge, it is advisable first of all to consider any kind of dictionary use, including the use of printed dictionaries. In other words, we suggest a holistic approach to user surveys according to which the purpose of conducting a user survey may be one or more of the following:

(1) In metalexicographic research: To present new, empirically substantiated results showing the dictionary user's need for, use of and benefits from one dictionary or a number of dictionaries.

(2) In metalexicographic research or a commercial survey: To test the quality of one dictionary or a number of dictionaries, possibly as part of a dictionary review or as a survey paid for by a publishing firm for the purpose of advertising and/or improving the dictionaries.

(3) In metalexicographic research: To collect, analyse and use data from user surveys in order to suggest concepts for new and better dictionaries.

Purposes (1) and (2) may be referred to as contemplative surveys (according to Tarp 2002). The starting point is existing dictionaries whose access methods, contents and structure are evaluated, and improvements may be suggested on the basis of the user survey. Almost all existing contributions on dictionary use of printed as well as electronic dictionaries, including log file analyses, fall under this type of survey. Purpose (3) may be referred to as a transformative survey as the purpose is not so much to repair existing dictionaries, but rather, on the basis of the analyses, to obtain arguments for future, possibly entirely new, dictionary concepts, which may not have any resemblance to the dictionaries analysed. This type of survey has not yet been carried out on a large scale. Laufer and Levitzky-Aviad (2006) have taken the first step, and this article aims to further develop the theory.

Giving priority to purpose (3) does not imply that purposes (1) and (2) should be neglected. Specifically, it is not true that user surveys conducted on the basis of printed dictionaries reveal nothing at all about dictionary use. Admittedly, compared to log file analyses, they involve only very few subjects who have carried out only a very limited number of look-ups, but unlike most existing contributions on log file analyses, they take the dictionary's various information types into account and do not focus almost solely on the lemma. In the early 1980s, the dictionary user was referred to as 'the known unknown' in several papers (Schaeder 1981: 62). This was undoubtedly the case at that time, but the question is whether the situation has changed fundamentally. In the 1980s, a large number of retrospective and introspective user surveys were carried out along the lines of Béjoint (1981) or Benbow et al. (1990); cf. the outlines in Ripfel and Wiegand (1988) and Nesi (2002: 277): *Which types of information do you need when you use a dictionary?* or *Which types of information do you consider to be important in a dictionary?* The surveys usually involved relatively few subjects, most often students, as in Benbow et al. (1990: 199):

	Daily	Weekly	Monthly	Never
Headword (e.g. for spelling)	23%	33%	25%	19%
Pronunciation	4%	16%	38%	43%
Phrases and idioms	13%	24%	37%	26%
Senses (definitions)	22%	38%	28%	12%
Illustrative quotations	12%	19%	39%	30%

Slightly, but only slightly, more realistic is the use of dictionary protocols where the user is asked to use specific dictionaries in connection with text production (Wiegand 1985), reception (Nesi 2002) or translation (Nielsen 1994: 20-32). These surveys involve very few users, and the results are thus heavily influenced by the users' individual actions. Furthermore, it is unrealistic to include bilingual dictionaries in surveys involving foreign students (as in Wiegand 1985).

Purpose (2) comprises surveys involving printed dictionaries, e.g. Benbow et al. (1990), as well as surveys involving Internet dictionaries, e.g. Ling et al. (2002), Bergenholtz and Johnsen (2005) and De Schryver et al. (2006). Surveys involving printed dictionaries have the same advantage (taking into account the various dictionary functions) and disadvantages (too few subjects, too few look-ups and completely unrealistic retrospective and introspective questions and answers) as was the case under purpose (1). Even with a relatively low number of look-ups, as in De Schryver and Joffe (2004), log files contain a much larger number of look-ups than practically all currently available surveys involving printed dictionaries. Log files of this type do not reveal any certain facts about the words, word parts or sentence combinations searched for, probably because the user often did not search for the word itself, but rather for information about the word, e.g. grammar, collocations, word formation etc., depending on the user's information needs. This may also explain the large differences between the most frequent queries in dictionaries of different languages; cf. the results in the Appendix where almost no similarities exist between the frequent queries in Danish, German, Norwegian and English, as certain words form part of numerous idiomatic expressions or involve complex grammar in some languages, but not in others. However, we have yet to find a reasonable explanation for the fact that only about half of the 128 000 entries in the *Danish Internet Dictionary* have been looked up after more than 8 million queries (Bergenholtz and Johnsen 2005). These data are entirely different from the data from three other online dictionaries of technical terms. In the *Danish Music Dictionary*, the *Danish–English Accounting Dictionary* and the *English–Danish Accounting Dictionary* (with approximately 4 000, 5 000 and 6 000 entries, respectively), almost all entries have been looked up after a much lower number of queries. This does not mean that existing log file analyses are of no value — they may be used to reveal lemma lacunae or improve the search options.

Purpose (2) also comprises more sophisticated log file analyses such as those presented by Ling et al. (2001 and 2002). These log files contain information not only about searches for lemmata and parts thereof, but also information based on search strings, e.g. full-text searches and searches using synonyms and conceptually associated words:

> A query pattern represents a set of user queries with the same or similar intention, and thus is associated with an article as the answer. It is difficult to capture the "similar intention" in natural language, and therefore, we use both syntactic constraint (such as stemming of keywords [...] as well as semantic constraints (such as generalized concepts and synonyms [...]) in our definition of queries with "similar intentions". (Ling et al. 2002: 1103)

The purpose of Ling et al. (2002) is to optimise the user's search options in the *Microsoft Encarta Encyclopaedia* by establishing patterns of associative search words for *Encarta*'s 42 000 entries. Furthermore, the aim is to find particularly popular themes in order to add more entries related to these themes. These log file analyses cannot rightly be characterised as a true description of the dictionary users' actions. The collected data are so general that they may at best be used to improve the search modalities, but not the lexicographic content. This is for example the case when, in their analysis of 4.8 million queries, Ling et al. (2002: 1103) conclude that 52.5% of the queries are carried out using one search word, 32.5% using two search words, 10% using three search words and 5% using four search words or more, which, in our opinion, is of no direct relevance to the lexicographic concept. Moreover, when analysing a much smaller number of queries (271 803), Ling et al. (2001) reach the same conclusion, i.e. that exactly 52% of the queries are carried out using one search word. These analyses show but one fact: that an online dictionary allowing the users to search by means of one search word only does not correspond to the users' habits or needs.

An interesting example of purpose (2) is Laufer and Hill (2000), who allow the user to compose an individual search profile:

> If, for example, the learner is interested in a quick L2–L1-translation, the option should be available. If, on the other hand, s/he is interested in examples of usage, or in grammatical information, or in a definition, each type of information should be accessible via another lookup option. Log files would record which of these options were selected for which words. (Laufer and Hill 2000: 59)

This suggestion has possibilities, but its consequences have not been given full consideration. If a user wants to understand a word or a collocation in a text, the user is only 'interested' in the equivalent and nothing else. Information on grammar or collocations, on the other hand, will help the user solve a problem in connection with L2 text production. Finally, the user may be 'interested' in all types of information if he/she wants to learn as much as possible in a cognitive process independent of text-related communicative problems. This test, as well as Laufer and Levitzky (2006), involves too few subjects to provide generally reliable indications of dictionary use. 75 students participated in a constructed reception test, not a test in a real life setting. In each case, the students were asked to select a search profile (Laufer and Levitzky-Aviad 2006: 150f) and given eight possible combinations to choose from. For text production in a foreign language, the students had the following preferences (listed according to the overall selection of all the students):

Translation + Definitions + Examples
English translations only
Translation + Definitions
Translation + Examples
Translation + Definitions + Examples + Thesaurus

The last three combinations were used only to a very limited extent:

Translation + Examples + Thesaurus
Translation + Examples + Thesaurus
Translation + Definitions + Thesaurus

We believe that these log files provide an unreliable and uncertain picture, as only the very experienced user (experienced in the use of the dictionary in question) will be able to choose among the many combinations correctly and in accordance with the required function(s). As we see it, a controlled process for inexperienced dictionary users and extensive freedom of choice for experienced users would enable lexicographers to document dictionary use in a way that would really be useful to transformative lexicography.

5. Suggested Structure of Log Files for Internet Dictionaries

If log files really are to reflect how and for which purpose dictionaries are used, a function-oriented set of search and link options is required. Surely, this is not to say that all users should be presented with as many lexicographic data as possible, or even with more data than they desire:

> I will be shamelessly selfish and ask for the impossible. I will advocate for a dictionary that will always adapt to my needs and always be ready to provide me with exactly the answer that I need and will also agree with. I also expect the dictionary to be able to give me satisfactory answers to those questions that I forget to ask. (Varantola 2002: 31)

On the contrary, all present experience with searches on the Internet suggests that the main problem usually involves receiving too many hits or texts that are too long. Users of Internet dictionaries may also be inundated by too much information. Another term used in connection with dictionary searches is 'over generalization' (Ling et al. 2002: 1105). In the mind of the user, the perfect dictionary provides just the amount of information required to fulfil the user's need. If a user for example needs assistance in translating a collocation, this collocation will be the search string entered by the user, and the translation will be exactly the short and exhaustive answer the user needs. Translated into the terminology used in connection with printed dictionaries, the user's search string is for instance the lemma **break a gene**, and this lemma and its corresponding equivalent constitute the dictionary entry, i.e. in an English–Danish dictionary **break a gene** *skære et gen over*. In the ideal situation, a one-to-one relation occurs, but most often a one-to-many relation exists between search strings and the entries shown. This type of dictionary does not contain a fixed number of entries, but rather all potential entries permitted by the dictionary database on the basis of the search criteria. Few, if any, Internet dictionaries function like this yet (but they will become reality within the next two or three years), and the user is merely offered ready-made entries like those found in printed dictionaries.

It is not, however, futuristic speculation to say that the current dictionary typologies of monolingual, bilingual and polylingual dictionaries or language dictionaries and subject dictionaries are relevant classifications when it comes to designing lexicographically relevant log files. In fact, the various user surveys conducted so far, and also the majority of log file analyses, are based on the assumption that a dictionary is a dictionary without any kind of distinction between dictionary types. They have, to a certain degree, distinguished between monolingual and bilingual dictionaries, but this is surely not the only suitable distinction to be made, as users very much need a combination of dictionary types taking into account the various needs of the users; cf. Mugdan (1992) and Laufer and Levitzky-Aviad (2006). This line of thought is not a new one:

> More original typologies are undoubtedly imaginable, for instance one that would be based on the functions of dictionaries and/or on the different types of organization of addresses (that is, types of organization of access to information). (Hausmann et al. 1989: xix)

The idea has been transformed into suggestions for a concrete dictionary typology in Bergenholtz and Kaufmann (1997: 100), who distinguish between reference books for needs relating to text-dependent and text-independent problems. Today, the use of the terms 'dictionaries for communicative functions' and 'dictionaries for cognitive functions' are preferred to refer to dictionaries designed to assist the user in resolving issues relating to text reception, text production and translation on the one hand and dictionaries designed to assist the user in resolving issues relating to the acquisition of knowledge on the other. In the first scenario, the user has a problem related to a text that he/she does not understand, has difficulties formulating or has difficulties translating. In the second scenario, the user has a general need for new knowledge, be it specific knowledge about a given word or expression or general knowledge about a particular subject. With this theory as a starting point, lexicographers may begin to comprehend how a dictionary is really used. Existing contributions on the subject, such as De Schryver et al. (2006), may present many figures, but their interpretational value to lexicography is limited.

These considerations lead to the following suggestions for the collection of log file data[†]:

(1) the user's IP address,
(2) the user's browser type,
(3) the user situation,
(4) which fields the user searches in if he/she chooses to define his/her needs according to the user situation as such,
(5) the user's search string,
(6) the search criteria,
(7) the date or time of the search,
(8) the result of the search, including all the entries that were found,

(9) which links the user clicks on in the entry found (entry being what is shown on the screen),
(10) which fields the user copies from, and
(11) which outside matter the user consults.

Not all data are of lexicographic relevance. The activities of search robots, for instance, are not relevant, yet they account for 30% of all searches in the *Danish Music Dictionary* over the first five months. They have IP addresses starting with 65 or 66 and should be filtered out. Log file item (10) cannot be logged in dictionaries using HTML, which is a so-called stateless language and therefore only allows for the logging of input commands initiated by the user. The most important items among the log file data listed above are (3)-(6) and (9)-(11). Item (5) is the search method used by the computer (the search engine). It is not just a question of defining the search string as 'is', 'contains', 'begins with' or 'ends with', but also of the algorithms used in the programming. Of particular interest is log file item (4), which will show the fields the user believes to be necessary in the basic user situation. We believe that the user will most often choose the field containing the lexicographic definition for reception problems, and in the case of text production problems, we assume that the user will also choose fields containing synonyms and antonyms, grammatical information, collocations and examples. In a translation situation, the user's choice in the case of L1 problems will most likely be the same as for reception problems, and in the case of L2 problems, it will probably be the same as for text production problems with the addition of the equivalent field. Finally, users who are on a quest for knowledge in the widest sense of the word are likely to require all information, including information on etymology and Internet references. Log file item (11) is also interesting. A recurring argument in dictionary related discussions is that a dictionary's outside matter, particularly its user instructions, are hardly ever consulted. This argument is put forward despite the fact that the only major survey known to us (Wolf 1994) shows that over 50% of all users have read the user instructions within the first two years of purchasing a dictionary. Our first log file data prove that Wolf's results concerning the use of outside matter in printed dictionaries are transferable to the use of outside matter in online dictionaries. The user clicks on the outside matter in more than 5% of all searches. Since every user on average conducts two searches a day, it means that 10% of the users consult the outside matter. The following figures (indicating the number of look-ups) are log file data from the *Danish Phraseological Dictionary* from the period 13 March to 26 March 2007:

om ordbogen (a kind of preface)	73
søgetips (short user instructions)	51
brugervejledning (more detailed user instructions)	46
litteratur (relevant literature)	31
kontakt (information on how to contact the editors)	28
copyright (information on dictionary rights)	22

These figures are to be viewed in relation to the number of dictionary searches conducted in the same period of time (13 March to 26 March 2007):

dictionary searches	4 010	94.11%
outside matter	251	5.89%
total	4 261	100.00%

The user's basic need for help, i.e. log file item (3), is also important. We suggest that this part be made mandatory, which means that the user will not be able to continue the search unless he/she makes a choice. Alternatively, instead of forcing the user to actively choose, one of the functions may be used as the default option in case the user does not make a choice. This would, however, produce figures that are not entirely reliable as the default button may lead to a biased result, seeing that many users might not take the trouble to make an active choice. Nonetheless, we have opted for a default button for the time being in our initial experiments with log files of this type for the purpose of user friendliness.

6. First Experiments with Function-oriented Log Files

We have used two dictionaries to conduct our initial experiments with the logging of dictionary functions. It seems that using a default option is of no major consequence. In the case of the *Danish Music Dictionary*, the default option was initially *help in connection with understanding a text*, but this was later changed to *further information*. Yet, the differences in the log file data before and after this change are relatively small:

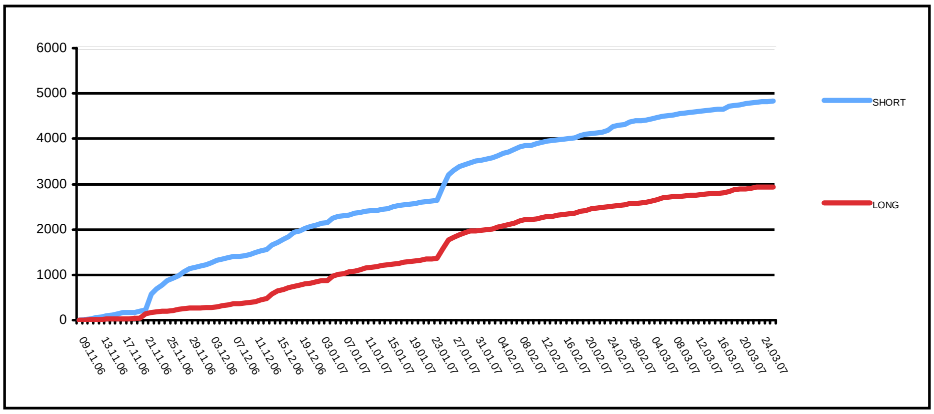
Searches from 9 November 2006 to 4 December 2006:

understanding a text	1 979	54.38%
further information	1 660	45.62%
total	3 639	100.00%

Searches from 5 December 2006 to 26 March 2007:

understanding a text	8 733	53.77%
further information	7 507	46.23%
total	16 240	100.00%

In other words, these figures indicate that the use of a default option does not measurably influence the result. However, it is interesting to observe that experienced users apparently to an increasing degree just 'settle for' the information they need, i.e. assistance in relation to text reception problems. Furthermore, the curves illustrating the two search options move ever further apart over time. This leads to the conclusion that an experienced user requiring help in a reception situation has learned that he/she will get a clear answer to his/her question by actively choosing the reception button rather than making use of the default option:



The *Danish Music Dictionary* is, as can be deduced from the above description of its log files, a polyfunctional dictionary with two functions:

(1) help in solving reception problems, which is indicated by the button *help in connection with understanding a text*, and

(2) help in the acquisition or expansion of knowledge, which is indicated by the button *further information.*

When using the first button, the user is presented with a short lexicographic definition, whereas clicking the other button provides the user with a detailed definition including illustrations, note examples and links to a short, integrated presentation of music theory. It was our assumption that most users would choose the detailed version. As can be seen from the log file data for both default options and a total of 19 881 searches, this is not the case, neither when the short definition, nor when the long definition for the acquisition of knowledge is used as default option:

understanding a text	10 714	53.89%
further information	9 167	46.11%
total	19 881	100.00%

Further statistics of the users who move from the reception entry to the knowledge acquisition entry, i.e. from the short entry to the long entry, or vice versa by clicking the internal link may be deduced from the figures. The remaining searches would reveal in how many instances the user's need for information was fulfilled by the first search. Such statistics are not available for the experimental dictionaries that we have compiled so far. Still, the log files reveal a crucial point. The initial log file results alone disprove the theories of Wierzbicka (1985) with her proposals for excessively long lexicographic definitions. Such detailed definitions are required when cognitive functions are involved whereas short definitions are more appropriate in relation to communicative functions.

Similar results appear from the log files of the *Danish Phraseological Dictionary* where a log file system has been in operation since the end of February 2007. This dictionary has three functions, two communicative functions and one cognitive function, which are indicated by the following buttons: (1) help in connection with understanding a text, (2) help in connection with writing a text, and (3) further information, i.e. knowledge acquisition. The first log files for this dictionary show that a large majority of its users only want information on (1) the meaning of a phrase, whereas (2) help in connection with text production, which includes further grammatical information, synonyms, collocations and examples, and (3) as much information about an idiom or a proverb as possible, which, in addition to the information provided under (2), includes etymology and Internet references to relevant contributions on phraseology, are less popular:

understanding a text	9 483	58.01%
writing a text	4 320	26.43%
further information	2 543	15.56%
total	16 346	100.00%

As was the case with the *Danish Music Dictionary*, we expected that a clear majority of the users would choose option (3), but we have had to adjust our expectations. The default option in this dictionary is help in connection with reception, and the figures should therefore be interpreted with a certain amount of caution. Further log file data will enable us to determine how many users move from the default search option to one of the other two options, but the differences between the three functions are considerable with more than half of the searches being conducted for reception purposes (which may indicate that the default option influences the result to some degree). Yet, it is interesting that more users choose option (2) than option (3), which leads us to conclude that more is not always better in the mind of the dictionary user. Removing the default option completely and thus forcing the users to make an active choice in order to conduct a search in the dictionary would therefore provide a more realistic picture.

When conducting these searches, the user cannot choose which types of information he/she wants to see. We believe that such a controlled search process is the most advantageous option for users who (just) want an answer to a specific question. For experienced users or users who are adept at experimenting and wish to do so, the possibility of choosing one of the three buttons and defining which fields to be shown should be implemented. There is no way of knowing for sure how the above figures would change if this possibility is realised. Our guess is that less than 10% of the users would bother. Intelligent log files of this type have yet to be analysed on a large scale. It is our belief that analyses of this kind would bring about new and constructive metalexicographic knowledge, particularly if used in conjunction with analyses of the large number of e-mails received by compilers of Internet dictionaries providing an e-mail address for user inquiries.

7. Moral

User surveys will never provide a 'true' picture without any limitations and conditions. Compared to the present guessing competitions and to lemma-oriented log file analyses, analyses of function-oriented log files are much closer to the truth, especially if those searches where the user has moved from one function to another are filtered out. *Theory without empirical data is empty.* This maxim is applicable to many, if not most, surveys concerning printed dictionaries in which the empirical basis is, at best, rather weak. *Empirical data without theory are empty.* This maxim may, in part, be applied to existing log file analyses concerning the use of Internet dictionaries. Many figures are presented, but their lexicographic relevance is limited. The aim of this article is to contribute to such a theory as it is our conviction that function oriented log file analyses are, for the time being, the best and most reliable way to look over the dictionary user's shoulder.

Endnote

† Some of these suggestions were put forward by Richard Almind, Aarhus School of Business, Aarhus, Denmark. We would like to thank him for several useful and constructive discussions.

References

Béjoint, Henri. 1981**.** The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* 2: 207-222.

Benbow, Timothy, Peter Carrington, Gayle Johannesen, Frank Tompa and Edmund Weiner. 1990. Report on the *NEW Oxford English Dictionary* User Research. *International Journal of Lexicography* 3(3): 155-203.

Bergenholtz, Henning. 2005. Den usynlige elektroniske productions- og korrekturordbog. *Lexico-Nordica* 12: 19-38.

Bergenholtz, Henning and Mia Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34: 117-141.

Bergenholtz, Henning and Uwe Kaufmann. 1997. Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. *Hermes* 18: 91-125.

Bokmålsordboka = Wangensteen, Boye (Ed.). *Bokmålsordboka — definisjons- og rettskrivningsordbok.* http://www.dokpro.uio.no/ordboksoek.html.

Cambridge Dictionaries Online. http://dictionary.cambridge.org.

Danish–English Accounting Dictionary = Sandro Nielsen, Lise Mourier and Henning Bergenholtz in cooperation with Mads Melgaard, Trine Middelboe, Brit Sørensen, Mia Johnsen, Rie Bobjerg Nielsen, Jóna Ellendersen, Amalie Kofoed Stender and Vibeke Vrang. 2004-2007. *Den Engelsk–Danske Regnskabsordbog/English–Danish Dictionary of Accounting.* Database and design: Richard Almind. Implementation and encoding of web pages: Caspar Thomsen. http://www.regnskabsordbogen.dk/regn/dkgb/dkgbregn.aspx.

Danish Internet Dictionary = Henning Bergenholtz and Vibeke Vrang in cooperation with Lena Lund, Helle Grønborg, Maria Bruun Jensen, Signe Rixen Larsen, Rikke Refslund, Mia Johnsen, Katja Å. Laursen, Sophie Leegaard and Maj H. Bukhave. 2002-2007. *Den Danske Netordbog*. Database and design: Richard Almind. http://www.ordbogen.com/ordboger/ddno/.

Danish Music Dictionary = Inger Bergenholtz in cooperation with Richard Almind and Henning Bergenholtz. 2006. *Musikordbogen*. http://www.musikordbogen.dk.

Danish Phraseological Dictionary = Henning Bergenholtz and Vibeke Vrang. 2007. *Ordbogen over Faste Vendinger*. Database and design: Richard Almind. http://www.idiomordbogen.dk.

De Schryver, Gilles-Maurice and David Joffe. 2004. On How Electronic Dictionaries are Really Used. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 187-196. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe and Sarah Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

English–Danish Accounting Dictionary = Sandro Nielsen, Lise Mourier and Henning Bergenholtz in cooperation with Mia Johnsen, Rie Bobjerg Nielsen, Jóna Ellendersen, Amalie Kofoed Stender and Vibeke Vrang. 2006–2007. *Den Engelsk–Danske Regnskabsordbog/English–Danish Dictionary of Accounting.* Database and design: Richard Almind. Implementation and encoding of web pages: Caspar Thomsen. http://www.regnskabsordbogen.dk/iasgbdk.

Eurodicautom. http://europa.eu.int/eurodicautom/Controller.

Hausmann, Franz Josef, Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta. 1989. Preface. Hausmann, Franz Josef, Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta (Eds.) 1989. *Wörterbücher. Ein internationales Handbuch zur Lexikographie/Dictionaries. An International Encyclopedia of Lexicography/Dictionnaires. Encyclopédie internationale de lexicographie*. First Volume: xvi-xxiv. Berlin/New York: De Gruyter.

Johnsen, Mia Steen. 2005. Logfiler som leksikografisk analyseinstrument og hjælpeværktøj. https://merkur2.asb.dk/F/C5NCNGIFXTRQS62JNBMTJJ8X22522SV62KHICN6YVYJQQXBEGM-00166?func=service&doc_library=HBA01&doc_number=000139835&line_number=0001&service_type=MEDIA.

Laufer, Batia and Monica Hill. 2000. What Lexical Information Do L2 Learners Select in a Call Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3(2): 58-76.

Laufer, Batia and Tamar Levitzky-Aviad. 2006. Examining the Effectiveness of 'Bilingual Dictionary Plus' a Dictionary for Production in a Foreign Language. *International Journal of Lexicography* 19(2): 135-155.

Ling, Charles X., Jianfeng Gao, Huajie Zhang, Weining Qian and Hongjiang Zhang. 2001. Mining Generalized Query Patterns from Web logs. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34) Volume 5.* Washington, DC: IEEE Computer Society Press 5020. http://research.microsoft.com/~jfgao/paper/hicss01.pdf (February 2007).

Ling, Charles X., Jianfeng Gao, Huajie Zhang, Weining Qian and Hongjiang Zhang. 2002. Improving Encarta Search Engine Performance by Mining User Logs. *International Journal of Pattern Recognition and Artificial Intelligence* 16(8): 1101-1116.

Mugdan, Joachim. 1992. On the Typology of Bilingual Dictionaries. Hyldgaard-Jensen, Karl and Arne Zettersten (Eds.). 1992. *Symposium on Lexicography V. Proceedings of the Fifth International Symposium on Lexicography May 3–5, 1990, at the University of Copenhagen*: 17-24. Tübingen: Niemeyer.

Nesi, Hilary. 2002. A Study of Dictionary Use by International Students at a British University. *International Journal of Lexicography* 15(4): 277-305.

Nielsen, Sandro. 1994. *The Bilingual LSP Dictionary. Principles and Practice for Legal Language.* Tübingen: Narr.

Ripfel, Martha and Herbert Ernst Wiegand. 1988. Wörterbuchbenutzungsforschung. Ein kritischer Bericht. Wiegand, Herbert Ernst (Ed.). 1988. *Studien zur neuhochdeutschen Lexikographie VI, 2. Teilband.* Germanistische Linguistik 87-90: 491-520. Hildesheim: Olms Verlag.

Schaeder, Burkhard. 1981. *Lexikographie als Praxis und Theorie.* Tübingen: Niemeyer.

Tarp, Sven. 2002. Translation Dictionaries and Bilingual Dictionaires — Two Different Concepts. *Journal of Translation Studies* 7: 59-84.

Varantola, Krista. 2002. Use and Usability of Dictionaries: Common Sense and Context Sensibility. Corréad, M.-H. (Ed.). *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*: 30-44. Göteborg: EURALEX.

Wiegand, Herbert Ernst. 1985. Fragen zur Grammatik in Wörterbuchbenutzungsprotokollen. Ein Beitrag zur empirischen Erforschung der Benutzer einsprachiger Wörterbücher. Bergenholtz, Henning and Joachim Mugdan (Eds.). *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.–30.6.1984*: 20-98. Tübingen: Niemeyer.

Wierzbicka, Anna. 1985. *Lexicography and Conceptual Analysis.* Ann Arbor: Karoma.

Wolf, Birgit. 1994. Wörterbuch und Benutzer — Versuch einer empirischen Untersuchung. *ed. by* Braußе, Ursula and Dieter Viehweger (Eds.). *Lexikontheorie und Wörterbuch. Wege der Verbindung von lexikologischer Forschung und lexikographischer Praxis*: 295-389. Tübingen: Niemeyer.

Wortschatz Deutsch = Quasthoff, Uwe, Stefan Bordag, Christian Biemann and Matthias Richter in cooperation with Christian Wolff, Fabian Schmidt, Karsten Böhm, Martin Läuter, Timo Böhme, Jens Drawehn, Frank Fischer, Sandra Liebold, Katja Mannekens and Martin Quested. 1998–2005. *Wortschatz Lexicon*. http://wortschatz.uni-leipzig.de.

Appendix: Top 20 queries in four different Internet dictionaries

The *Danish Internet Dictionary* (2003–2004)

1	gå (walk)
2	a (a, each)
3	hest (horse)
4	for (to, for)
5	ad (by, along)
6	arbejde (work)
7	kompetence (competence)
8	hus (house)
9	finke (finch)
10	tage (take)
11	bil (car)
12	empati (empathy)
13	tid (time)
14	kognitiv (cognitive)
15	indenfor (in, inside)
16	i (in)
17	vand (water)
18	godt (good, well)
19	linie (line)
20	hund (dog)

***Wortschatz Deutsch* (February 2005)**

1	synonyme (synonyms)
2	lexikon (lexicon)
3	wortschatz (vocabulary)
4	deutsch (German)
5	wortschatz deutsch (the name of the dictionary)
6	serena williams
7	michael jackson
8	britney spears
9	deutsches wörterbuch (German dictionary)
10	synonymwörterbuch (thesaurus)
11	lance armstrong
12	nicole kidman
13	wortschatz leipzig (vocabulary leipzig)
14	george bush
15	thesaurus
16	fremdwörter (loanwords)
17	wortschatz uni leipzig (vocabulary university leipzig)
18	duden (the name of a German reference work)
19	saddam hussein
20	deutsch wörterbuch (German dictionary)

***Bokmålsordboka* (2000–2005)**

1	feil (mistake, error)
2	dessverre (unfortunately)
3	verken (neither)
4	fitte (pussy)
5	hverken (neither)
6	interessant (interesting)
7	desverre (unfortunately)
8	internett (Internet)
9	engelsk (English)
10	interesse (interest)
11	kognitiv (cognitive)
12	hei (hi)
13	interessert (interested)
14	pragmatisk (pragmatic)
15	empati (empathy)
16	patetisk (pathetic)
17	tunnel (tunnel)
17	enda (even)
18	ordbok (dictionary)
20	nysgjerrig (curious)

***Cambridge Dictionaries Online* (2004)**

1	advice
2	liaise
3	effect
4	regard
5	comply
6	appreciate
7	commit
8	assess
9	endeavour
10	acquire
11	paradigm
12	information
13	analyse
14	intend
15	affect
16	provide
17	idiom
18	propose
19	emphasize
20	ubiquitous