Henning Bergenholtz & Mia Johnsen**

Log Files as a Tool for Improving Internet Dictionaries

Abstract

In their advertisements, dictionary publishers often praise their dictionaries for taking into account the exact needs of the users. Until the beginning of the 1980s, however, no theoretical contributions on dictionary use were available, neither in the form of purely theoretical considerations nor in the form of empirical research. Since then, the situation has changed completely. Such a large number of user surveys have been carried out that it is no longer possible to give a complete overview. Nevertheless, this has led to no significant improvement of the situation as the majority of these surveys are not related to concrete examples of dictionary use. The surveys, which have always been concerned with printed dictionaries, have therefore not contributed to substantial improvements of dictionary conception.

In the case of internet dictionaries, on the other hand, technical possibilities enable lexicographers to monitor user behaviour in a different and much more precise way. Analyses of log files reveal exactly which lemmas and which types of information have been requested, and, perhaps more significantly, which lemmas and which types of information have been requested but were not found in the dictionary. Furthermore, log files allow lexicographers to see the types of information which have not, or not yet, been searched for. All in all, log files may thus be used as a tool for improving internet dictionaries – and perhaps also printed dictionaries – quite considerably.

1. Internet Dictionaries only?

As new tools are invented, old tools become obsolete. This process has been described in relation to electronic dictionaries vs. paper dictionaries (Simonsen 2000 with references to other scholars). We are not convinced, however; on the contrary, we are certain that paper dictionaries will remain a popular tool for at least the next two or three

* *Henning Bergenholtz*
Centre for Lexicography
Aarhus School of Business
Fuglesangs Allé 4
DK – 8210 Aarhus V
hb@asb.dk

* *Mia Johnsen*
Centre for Lexicography
Aarhus School of Business
Fuglesangs Allé 4
DK – 8210 Aarhus V
miajohnsen_1@hotmail.com

centuries (Bergenholtz 1996). Already now, CD-ROM dictionaries and even DVD dictionaries have had their time and will only be known by the next generation as a lexicographical medium that is no longer used. We believe that internet dictionaries, on the other hand, will have a much longer life. We are convinced, too, that the next 20-30 years will see not only internet dictionaries free of charge, most of them of quite low quality, but also really high quality dictionaries for which the user can pay monthly, yearly or pay per view. But we will still have printed dictionaries, especially those that comprise only one or two volumes. It is more doubtful whether we will have multi-volume printed dictionaries and encyclopaedias as they are too expensive in relation to the limited number of years in which they will be up to date. Here, the internet dictionaries and encyclopaedias will offer a version that is updated daily, and the user is not compelled to search for answers in several different volumes if he has different questions and a limited amount of time at his disposal. But for the smaller dictionaries, there is no doubt that there is a demand for paper and internet versions of the same dictionary. We have recently made this experience with the free internet dictionary THE DANISH-ENGLISH DICTIONARY OF ACCOUNTING (http://www.regnskabsordbogen.dk/iasdk). This dictionary has been available since August 2003, but very soon after the release there were so many requests from the users for a paper version that a publishing company asked for the possibility of publishing such a paper version (DICTIONARY OF ACCOUNTING DANISH-ENGLISH 2004).

In this paper, we are dealing with the use of internet dictionaries, more specifically with THE DANISH INTERNET DICTIONARY (http://www.netordbog.asb.dk), which has been available on the internet since April 2002.

2. User Surveys in Theory and Practice

Dictionaries are utility products. They are tools designed to help a potential dictionary user solve problems with producing, comprehending or translating a text and to provide cultural, encyclopaedic or linguistic knowledge. The function of a given dictionary is to provide assistance to a specific user group with specific characteristics in order to meet the complex needs that arise in a specific type of user situation. A concrete dictionary can have one or more functions, i.e. it can be mono or multifunctional. As any other utility product, dictionaries also have a genuine purpose. This genuine purpose comprises the totality of functions of a given dictionary and the subject field(s) that it covers (Bergenholtz/Tarp 2003). References to users and their needs have been made in dictionary prefaces and other lexicographic contributions for centuries. There is nothing new in that. It is therefore rather a paradox that the German lexicographer, Wiegand (1977), was right in his conclusion that the dictionary user is the "known unknown". Similarly, 25 years later, the dictionary user was referred to as a yeti (Bergenholtz 2002). This does not mean that no research has been carried out on dictionary use yet – on the contrary. From 1985 until today, so many monographs, editions and papers in journals have been published that it is difficult or even impossible to get a complete overview. When Bergenholtz (2002) insists on the comparison between the dictionary user and a yeti, he does not mean that no one has made fatiguing expeditions. Rather, Bergenholtz (2002) argues that, in reality, the main part of the research on dictionary use has only found unclear tracks of the dictionary user. We can add that the big majority of the investigations have not made clear why they want to find out how users use a dictionary. Perhaps it is too evident for those scholars that knowledge about user habits leads to better dictionary conceptions, which, in the end, leads to better and more helpful dictionaries. There could be another, but naturally related, goal: It is indeed interesting to know about the dictionary users' habits and experiences with different dictionaries. It is a goal sui generic, and, at the same time, it could be a contribution to dictionary criticism of a single dictionary or a set of dictionaries.

A distinction exists between two kinds of investigation. The first one is the criticized type – and the most practiced kind – and is undertaken without a direct relation to a concrete dictionary use. In such **questionnaire surveys**, the same methods are employed as in other forms of market analysis: a number of standard questions are asked of a selected sample concerning a certain product or behaviour, e.g. Atkins (1998). However, the answers from the informants do not necessarily reflect a real genuine user situation. It cannot be ruled out that the problems, behaviour, etc. described by the informants differ from their real problems. The questions asked have to do with future activities, as in "Under which headword would you look for the following collocations?", or with past activities: "Which types of information do you look for most often?". There is no guarantee that the answers correspond to why and how the informants really have used or will use dictionaries. Such surveys are quite problematic because they presuppose that the informants remember exactly how they have used dictionaries in the past and that they are able to predict how they will do it in the future. And as far as we can see, none of the surveys meet the normal requirements of representativity, e.g. it is very often students only, and the informants are not selected in accordance with the principles applied in the social sciences.

More realistic are the so-called **dictionary protocols** written by selected dictionary users directly after each dictionary use. They are more realistic because they refer to authentic user situations. The problem leading to the dictionary use is still clearly remembered by the informants, they can describe the result of the looking up and the way they used the dictionary items, also how they found the wanted pieces of information or that they did not find an item which enabled them to solve the problem. In practice, however, such protocols are insufficient; compare the results from some of the investigations of this kind:

- Wiegand (1985) asked foreign students (in Germany) with another native language than German to translate a text from their mother tongue into German. They were allowed to use any bilingual dictionary. He did not ask the informants to write a protocol from this part of the translation. He asked the students to improve the German translation by using a monolingual German dictionary. Each time they encountered a problematic word or text part, the students were to describe the problem in the protocol, use a dictionary, correct the text and write down in the protocol what they had found or not found in the dictionary, and how they had used it. The results were quite interesting, but in the end not typical because the really interesting aspect of dictionary use, the translation phase, was ignored. Furthermore, it is doubtful whether a small number of language students is representative of other students or all other kinds of users.
- Another attempt was undertaken by Danish libraries. Next to the shelves with dictionaries, the dictionary user would find a questionnaire to be filled out after each use of a dictionary. The investigation was part of a ministerial report on the need for additional or different dictionaries in Denmark (Vilkår 1982). After a year, however, the ministry

had received only some fifty replies, most of them very imprecise.

All those investigations concern printed dictionaries, but in principle, questionnaire surveys are also valid for internet dictionaries or other electronic dictionaries. We have no knowledge of any dictionary protocols regarding internet dictionaries. Other possibilities exist for internet dictionaries, however. With a log file, you can track every single use of the dictionary, depending, of course, on the search possibilities. If it is only possible to search for the lemma, only data for the first access step in the dictionary will be available. Which lemmas have been looked up how often? Which lemmas have never been looked up at all? And which words have been used in the search field without result, i.e. how many and which lemma lacunas does the dictionary use indicate? It is this kind of user investigation that we will describe in the following section. It is possible, however – if there is direct access to every dictionary item class (or field) – to get exact data for the use of the semantic item, the grammar item, the collocation item, etc. As far as we know, the use of such exact log files has not yet been described. Obviously, such log files do not reveal exactly the kind of problem which the user had; they do not reveal whether the user did indeed find the information to fulfil his needs. This could only be done by using a kind of dictionary protocol, e.g. if some of the users were asked to or made to fill out a questionnaire after every use of the dictionary (this could function technically if the user is not allowed to use the dictionary for free unless he fills out such a questionnaire). This possibility has not yet been practiced either, at least not as far as we know.

Several other interesting investigations may be carried out, too: In which way does the use of internet dictionaries differ from the use of paper dictionaries? Do we have user groups who never used paper dictionaries but now use internet dictionaries? To which extent do internet dictionaries function as lexicotainment dictionaries, i.e. only for entertainment? Does the use of paper dictionaries decrease? Or – as we believe – does the total use of dictionaries as a helping aid in connection with communicative and knowledge-related questions increase? All this will not be discussed in this paper, but it is certainly a relevant topic for further contributions. In this paper, we will analyse log files from THE DANISH INTERNET DICTIONARY. It is a Danish monolingual dictionary with 108,000 dictionary entries and a total of 126,000 different "records", i.e. the dictionary contains 18,000 subentries for polysemy. The genuine purpose of the dictionary is to help users with Danish as their mother tongue or with a good knowledge of Danish when encountering problems in a text production process and looking for help in the dictionary. The results will be used for decisions in the ongoing (and never-ending) work of improving dictionaries – not only this single internet dictionary but also other monolingual and even bilingual dictionaries and printed dictionaries, too, at least to some extent.

There are only a few published scholarly descriptions of internet dictionary log files. The most interesting contribution from de Schryver/Joffe (2004) describes the log file for a South-African bilingual dictionary, a Sesotho sa Leboa–English dictionary. The number of visitors and the number of lookups is not very high: 21,337 lookups made by 2,530 different visitors. De Schryver/Joffe write that the dictionary is partly used as a lexicotainment dictionary with no less than 17 sexually related words in the top 100 list. The most frequently requested words in both languages are greeting routine formulas like *hello, good morning, goodbye*, resp. *dumêla, thôbêla* and *sepela*. The users also look for non-existing words, especially for misspellings, but not as often as would be expected, e.g. there is only one misspelling in the top 100 list. Furthermore, they describe emails from the users, most of them thanking the dictionary makers for the free dictionary. De Schryver/Joffe (2004) fail to mention one very interesting point: With 28,000 English lemmas and 25,000 Sesotho sa Leboa lemmas, the users cannot have looked up all lemmas (with only 21,337 lookups). It would be most interesting to know which types of words are not looked up: Is about 90% or 80% of the dictionary never used at all? The very limited number of lookups indicates that no more than 40-50% of the dictionary is actually being used. Will all lemmas in the dictionary be looked up in time when the dictionary has had many more users? Or are there some lemmas that will never be looked up? If future dictionary makers knew the answers to those questions, they would not have to waste time describing words of no interest to the users.

3. How Frequently Are Internet Dictionaries Used?

To determine how frequently internet dictionaries are used in practice, it may be useful to look at a number of specific internet dictionaries that carry statistics of the search frequency. Unfortunately, only a limited number of internet dictionaries make such information available on the web site, and it has therefore not been possible to carry out a systematic analysis of e.g. the 10 most commonly used online dictionaries according to the Danish telecommunications provider, TDC (see below).

An example of a dictionary that does allow the user to access information on the search frequency is the German-French dictionary, ALLGEMEINES WÖRTERBUCH DEUTSCH-FRANZÖSISCH (http://site.ifrance.com/allinfor/dico/index.htm). According to the counter on the web site, 262,768 searches have been performed since 20 October 2002, which results in an average of approximately 380 searches per day. A French-Swedish online dictionary (FRANSK-SVENSKT LEXIKON; http://www.azoria.com/lexikon/indexsw.shtml) contains statistics of the number of searches per month for the last year, a total of 1,199,122 searches. Thus, approximately 3,406 searches are performed daily. Interestingly, however, the number of searches is lower at the end of the period in question than at the beginning. This may be contrasted with the search frequency of two other internet dictionaries, the German-English-German dictionary QUICKDIC (http://quickdic.org/index_d.html) and the above-mentioned SESOTHO SA LEBOA (NORTHERN SOTHO) - ENGLISH DICTIONARY (http://africanlanguages.com/sdp/). The latter does not provide a counter on the web site, but an article on the dictionary in which statistics on the search frequency are included appeared in the EURALEX Proceedings 2004 (de Schryver/Joffe 2004). The statistics of QUICKDIC show an increase in the number of searches from less than 20,000 in 1997 to almost 100,000 in 2001, and the same trend appears from the Sesotho sa Leboa-English dictionary, which was launched in 2003. This dictionary had a frequency of 1,308 searches in the first month, and 6 months after the release, this number had grown to 3,673 with numbers varying from just over 2,000 to almost 6,000 in the months in between.

As the following sections of this article will focus on the most commonly used Danish online dictionary[1], THE DANISH INTERNET DICTIONARY (http://www.netordbog.asb.dk), it may be interesting to examine whether this dictionary also shows an increase in search frequency. According to

[1] According to statistics made by the Danish telecommunications company, TDC (http://links.tdconline.dk/top.php?cid=2502&linktype=2)

the statistics, an average of 1,631 searches were performed daily in the first period of 2003, whereas the following period of 2004 showed an increase to 2,520 searches per day. The following section will elaborate further on these figures, but it is clear that this dictionary, too, is used more frequently now than 6 months ago.

Although it is difficult to make any final conclusions on the basis of statistics from randomly selected dictionaries, the trend seems to be that internet dictionaries in general are used ever more frequently. As mentioned above, the Danish telecommunications provider, TDC, carries daily statistics of the 10 most used internet dictionaries on their web site. Previously, TDC also carried a list of the 50 most used internet dictionaries, but this list is no longer available. From the top 10 list, it appears that the Danish Internet Dictionary is number one in the vast majority of cases, e.g. on the very recent list from 29 September 2004 where it is followed by On-line Dictionaries, Retskrivningsordbog, Cambridge, Eurodicautom, Britannica.com, IT-leksikon, Svenska Akademiens ordbok, Ordbøger - Gyldendal and Wortschatz Deutsch. Unfortunately, the only one of these dictionaries for which statistical information is available is the Danish Internet Dictionary, and we therefore contacted TDC in order to determine what they base this top 10 list on and how many users are involved as regards the other dictionaries appearing on the list. However, no one at TDC knew anything about this top 10 list, and we have therefore not been able to determine how it is compiled. Nevertheless, it is still of interest to this article that the Danish Internet Dictionary appears so frequently at the top of the list as it indicates the wide use of this dictionary.

Specifically, the Danish Internet Dictionary had the following contents on 3 August 2004[2]:

total records:	126,416
dictionary articles*:*	108,016
records with polysemy:	18,395

We did not install a log file system when the Danish Internet Dictionary was first made available to the public in 2002. During the time in

[2] We wish to thank Richard Almind, who is responsible for the database used for the Danish Internet Dictionary, for his help in preparing this paper. He prepared all the statistical data relating to the log files. Without his help, we would not have been able to begin this research.

which a log file system was in operation from 1 January 2003 to 3 August 2004, there was a period from June to September 2003 in which the log file system was switched off for technical reasons. In total, we had 1,021,139 single searches on 456 logging days. Thus, someone looks something up in the dictionary 2,239 times on average each day. We can see an increase in the number of searches over time: In 2003, the average number of searches was 1,631; in 2004 it increased to an average of 2,520. On the first four days of the week, between 4,000 and 4,500 searches are normally performed, and on Fridays, the number is about 3,500 searches. During school holidays and in weekends, between 1,000 and 1,500 searches are carried out. By a single search we mean every new article searched for, either by looking for a new lemma or by linking from one article to another (all items that are identical with lemmas in THE DANISH INTERNET DICTIONARY are also functioning links). The log files also allow us to see if the user looked for a word and did not find it, a so-called lemma lacuna. In total, there are

searches:	**1,016,960**	
records ”found”:	818,613	(80.5%)
records ”not found”:	198,347	(19.5%)

Some of the searches may be termed “empty” because the user did not write anything in the search field before hitting the search button. They are not included in the statistics:

“empty” records:	4,179

In this article, we do not include the number of unique users as it is not possible to distinguish between a single user and a unique user. In most cases, the unique user will be a single dictionary user, but this is not always the case, e.g. if a whole school uses the same identity number for all computers connected in a network.

The search string can be used to look for the lemma or for an inflected form with the lexeme represented by the lemma, or for only a part of the lemma. Most users try the “traditional” way, i.e. looking directly for a lemma, but the other possibilities are used too:

the lemma is:	859,682
the lemma begins with:	78,126
the lemma ends with:	17,119
the lemma contains:	66,212

In total, the users have searched for 104,097 different orthographical forms found in the Danish Internet Dictionary. This figure comprises about 35,000 different lexemes represented by a lemma (a detailed discussion and explanation of this appears in section 4). In comparison to this, the number of searches for different orthographical forms not found in the dictionary is higher; more specifically, 116,066, a difference of 12,000. This result is quite thought-provoking and gives the dictionary makers behind the dictionary cause to a renewed lemma selection, especially in the case of real lemma lacunas and of unsuccessful searches due to misspellings (more about that in section 4). It is obvious that misspellings account for almost all words in the top 100 list of searched, but not found, search strings. In the top 100 list of found words, a mixture of words from all word classes appears, compare the 10 words that have been looked up most frequently, e.g. *gå* (walk) 956 times, *hest* (horse) 703 times:

gå (walk)
hest (horse)
for (for)
ad (to)
arbejde (work)
kompetence (competence)
hus (house)
finke (finch)
tage (take)
bil (car)

As was the case in the log files described by de Schryver/Joffe (2004), the number of searches for sexually related expressions on the top 100 list is quite high, e.g. *pik* (cock) is number 25, *fisse* (cunt) is number 29. Such words are mainly looked for in the evening and during the night. Furthermore, the log file enables us to follow individual users linking from one word to another, probably using the synonyms as links. Such use of a dictionary will hardly be due to communicative problems; the user knows the words and how to use them and is thus only interested in seeing how they are explained and what kind of exciting collocations they have. More closely related to the intended genuine purpose of the Danish Internet Dictionary, i.e. helping Danish users solve text production problems, is the search for synonymous or almost synonymous expressions, e.g.

præliminær → indledende → indledningsvis
(preliminary → initial → by way of introduction)

4. Concrete Searches

A study of the log file from THE DANISH INTERNET DICTIONARY also reveals a number of specific problems encountered by the users of the dictionary. These problems fall into different categories which will be discussed in turn below.

The Passive

At present, it is not possible to search for the passive form of verbs in THE DANISH INTERNET DICTIONARY. However, the log file reveals that quite a large number of users actually attempt to do this: a search in the entire log file (1,021,139 searches) for the error string *-es*, i.e. not-founds that end in the letters *-es* (one of the two most common passive endings in Danish, the other being *-s*), returns a total of 4,141 hits. The vast majority of these are passive forms, probably between 3,000 and 3,500 of these searches. The top 100 list of not-founds also contains 5 passive forms, i.e. *fås* (is available), *nås* (is reached), *fåes* (is available), *nåes* (is reached) and *gennemgåes* (is examined). In addition to this, the top 500 of not-founds contains a further 7 searches for passive forms, i.e. *opnås* (is achieved), *gåes* (is walked), *foreslås* (is suggested), *anses* (is considered), *forståes* (is understood) and *opnåes* (is achieved).

In practically all of the above cases, the users have subsequently conducted a search for the infinitive of the word, but the examples nonetheless show that many users are unsure of how to form the passive of certain words (with or without *e*, i.e. *-s* or *-es*). Thus, it may be considered whether it would be relevant to add the passive form in the dictionary to enable the users to search for it in cases of doubt.

The majority of searches for passive forms concerns the present tense of the passive, but there are also examples of the past tense, e.g.

hjælpedes (was helped)
gaves (was given)
agtedes (was intended)
afbilledes (was depicted)
tryggedes (was pushed, the correct spelling is trykkedes)
prøvedes (was tried)
standsedes (was stopped)
fundes (was found)
øgedes (was increased)
passeredes (was passed)
overdragedes (was handed over)
konstateredes (was ascertained)
anbefaledes (was recommended)
påstodes (was claimed)
nægtedes (was refused)
rekrutteredes (was recruited)
erindredes (was remembered)
udrensedes (was cleansed)
tegnedes (was drawn)
koncentreredes (was concentrated)
stakkedes (was stacked)
gennemluftedes (was ventilated)
dannedes (was formed)
samledes (was gathered)
noteredes (was noted)
akkompagneredes (was accompanied)
svaredes (was answered)
udkrystalliseredes (was crystallized)
forværredes (was made worse)
implementeredes (was implemented)
nåedes (was reached)
udelukkedes (was excluded)
ytredes (was said)
domesticeredes (was domesticated)
ynkedes (was pitied)
frasagdes (was renounced)
lykkedes (was successful)
udfærdigedes (was issued)
skingredes (was shrilled)
normaliseredes (was normalized)
fokuseredes (was focussed)
planlagdes (was planned)

Most of these examples are grammatically correct, although some are used less frequently than others (eg *gaves*, *påstodes*), whereas yet others do not exist at all, e.g. *hjælpedes* and *fundes*. This illustrates that users may also be unsure of how to form the past tense of the passive as the correct form is not always evident. Consequently, this is another argument for including the passive form in the dictionary.

The Imperative

As is the case with the passive, it is presently not possible to search for imperative forms in THE DANISH INTERNET DICTIONARY. Among the 100 most frequent not-founds, no searches for imperative forms appear, whereas 15 searches are registered among the 500 most frequent not-founds, ie

fortsæt (continue)
søg (search)
åbn (open)
annuller (cancel)
registrer (register)
send (send)
mob (bully)
ærger (annoy)
skrid (go away)
husk (remember)
gem (hide)
beslut (decide)
niv (pinch)
scan (scan)
order (order)

In some of these cases, however, it is impossible to determine for certain whether the user actually searched for the imperative, e.g. in the case of *registrer* (may also be a misspelling of the plural form of *register* (register), or *registre* (registers) and *ærger* (may also be a misspelling of *ærgre* (to annoy) or *ærgrer* (annoys).

If these two terms are viewed in the context of the entire log file, however, it seems most likely that the user did not search for the infinitive form as he subsequently searched for *register, registre* and *ærgre, ærgrer*. Even so, the examples mentioned indicate that there is a need among the users for being able to search for the imperative as the formation of it is not always as simple as it may seem.

A search for specific imperatives in the log file (other than those mentioned above) reveals that users have searched for *vent* (wait) at 17 occasions, 11 times for *find* (find) and *lyt* (listen), respectively, 9 times for *installer* (install), 7 times for *luk* (close) and *returner* (return), respectively, and 6 times for *aflever* (hand over), *angiv* (state) and *skriv* (write), respectively. This supports the assumption that it would be relevant to include the imperative form in the dictionary to enable users to search for it.

Some imperative forms, such as *spis* (eat), *hør* (hear), *drik* (drink), *sig* (say), *sabl* (bill), *saboter* (sabotage), *sagtn* (slacken), *saliggør* (save), *saluter* (salute), *samkør* (co-ordinate), *køb* (buy), *skub* (push), *træk* (pull), *kast* (throw) and *kør* (drive), do not appear from the log file at all, i.e. no searches have been performed for these words.

The reference book HANDBOOK OF CONTEMPORARY DANISH lists a number of imperative forms that may cause the user problems. A search for those words in the log file yields the following results:

- *affjedr* (spring): No hits, no hits on *affjeder* (alternative, though not correct, spelling) either
- *behandl* (treat)*:* 4 hits, 2 hits on *behandel* (alternative, though not correct, spelling)
- *hamstr* (hoard)*:* No hits
- *krydr* (season)*:* 3 hits, no hits on *krydder* (alternative, though not correct, spelling)
- *pensl* (paint): 2 hits
- *sagtn* (slacken): No hits, no hits on *sagten* (alternative, though not correct, spelling) either

- *saml* (collect): 2 hits, 1 hit on *sammel* (alternative, though not correct, spelling)
- *smuldr* (crumble): 2 hits, no hits on *smulder* (alternative, though not correct, spelling)
- *åbn* (open): 23 hits

Clearly, the problem of not being able to search for imperative forms is not as widespread as the problem of the ”missing” passive forms. However, quite a number of examples do appear from the log files, and it might therefore still be relevant to include this form in the dictionary for the reasons mentioned above.

Dyslexia

Among the not-founds, a number of misspellings appear in which the problem seems to be reversal of letters.

Neither the top 100 of most frequent not-founds, nor the top 500 register any examples of this problem. A study of 300 consecutive searches in a random place in the log file revealed only 3 examples of the phenomenon, i.e. *onanym* instead of *anonym* (anonymous), *akapolypse* instead of *apokalypse* (apocalypse) and *medmnidre* instead of *medmindre* (unless). Thus, the problem does not seem to be very common. Also, it is impossible to tell whether these misspellings are actually due to dyslexia or whether they are simply typing errors.

Spelling Mistakes Affected by Pronunciation

An extremely large proportion of the misspellings found in the log file can be ascribed to users spelling the word as it is pronounced. Among the 100 most frequent not-founds, 8 examples appear, and a further 48 examples can be found among the 500 most frequent not-founds:

Top 100:

kompetance instead of *kompetence* (competence)
seperat instead of *separat* (separate)
hieraki instead of *hierarki* (hierarchy)
reperation instead of *reparation* (repair)
ærgeligt instead of *ærgerligt* (unfortunate)
udemærket instead of *udmærket* (excellent)
ærgelig instead of *ærgerlig* (unfortunate)
hiraki instead of *hierarki* (hierarchy

Top 500:

statestik instead of *statistik* (statistics)

parantes instead of *parentes* (paranthesis)

osse instead of *også* (also)

senarie instead of *scenarie* (scenario)

irreterende instead of *irriterende* (annoying)

pavilion instead of *pavillon* (pavilion)

ekseptionel instead of *exceptionel* (exceptional)

ærge instead of *ærgre* (annoy)

hovedsaglig instead of *hovedsagelig* (mainly)

kadence instead of *kadance* (cadence)

disiplin instead of *disciplin* (discipline)

intresse instead of *interesse* (interest)

ærgelse instead of *ærgrelse* (annoyance)

brilliant instead of *brillant* (brilliant)

transperant instead of *transparent* (transparent)

kotyme instead of *kutyme* (custom)

ancinnitet instead of *anciennitet* (seniority)

desideret instead of *decideret* (pronounced)

ekvivalent instead of *ækvivalent* (equivalent)

sattelit instead of *satellit* (satellite)

bobbel instead of *boble* (bubble)

trals instead of *træls* (annoying)

ærger instead of *ærgre* (annoy)

flematisk instead of *flegmatisk* (phlegmatic)

misære instead of *misere* (misery)

pressening instead of *presenning* (tarpaulin)

præsis instead of *præcis* (exact)

matriale instead of *materiale* (material)

kambolage instead of *karambolage* (collision)

presentere instead of *præsentere* (present)

halvfjers instead of *halvfjerds* (seventy)

revance instead of *revanche* (revenge)

synagi instead of *synergi* (synergy)

estetisk instead of *æstetisk* (aesthetic)

rimlig instead of *rimelig* (reasonable)

intresseret instead of *interesseret* (interested)

intressant instead of *interessant* (interesting)

apperat instead of *apparat* (device)

referance instead of *reference* (reference)

resonnement instead of *ræsonnement* (argument)

dimmitend instead of *dimittend* (graduate)

farisær instead of *farisæer* (Pharisee)

gardarobe instead of *garderobe* (cloak room)

ingenør instead of *ingeniør* (engineer)

artikkel instead of *artikel* (article)

singel instead of *single* (single)

distret instead of *distræt* (absent-minded)

pavilion instead of *pavillon* (pavilion)

Consequently, these two lists alone show that the problem of misspellings affected by pronunciation is the reason for a very large part of the non-successful searches. A study of 200 consecutive searches in a random place in the log file provides the same picture. 19 examples were found:

fransbrød instead of *franskbrød* (white bread)
senarie instead of *scenarie* (scenario)
reperation instead of *reparation* (repair)
kottelet instead of *kotelet* (chop)
mæling instead of *mægling* (mediation)
vanilie and *vannilie* instead of *vanille* (vanilla)
premisser instead of *præmisser* (premisses)
premisserne instead of *præmisserne* (the premisses)
napolion instead of *napoleon* (Napoleon)
parentes instead of *parantes* (paranthesis)
pediatri instead of *pædiatri* (pediatrics)
rensage instead of *ransage* (search)
statestikker (twice) instead of *statistikker* (statistics)
statestik (4 times) instead of *statistik* (statistics)

One or Two Words?

Another common problem revealed by the log file is whether a particular term should be written in one or two words. The top 100 and top 500 lists of most common not-founds contain the following examples:

Top 100 (11 words):

udfra (from)
ud fra (from)
iøvrigt instead of *i øvrigt* (besides, by the way)
istedet instead of *i stedet* (instead)
imorgen instead of *i morgen* (tomorrow)
allesammen instead of *alle sammen* (everybody)
fornylig instead of *for nylig* (recently)
tilgengæld instead of *til gengæld* (in return)
ovenikøbet instead of *oven i købet* (in addition, even)
tilgode instead of *til gode* (owed, due)
pånær instead of *på nær* (except)

Top 500 (another 15 words):

sålænge instead of *så længe* (in the meantime)
såfald instead of *så fald* (that case)
henimod instead of *hen imod* (towards)
iorden instead of *i orden* (allright)
tildels instead of *til dels* (partly)
forsent instead of *for sent* (too late)
istand instead of *i stand* (in order)
ligemeget instead of *lige meget* (of no consequence)
hvadenten instead of *hvad enten* (whether)
så som instead of *såsom* (such as)
foreksempel instead of *for eksempel* (for example)
vedlige instead of *ved lige* (in good order)
forøvrigt instead of *for øvrigt* (apart from this)
her fra instead of *herfra* (from here)
ud af (out of)

The many examples among the 100 most common not-founds show that this issue is highly relevant. However, not quite as many occurrences are found among the 500 most common not-founds, and a study of 200 consecutive searches in a random place in the log file yields only 10 examples:

holdspil (team play) (this is actually the correct spelling, but the word is not included in the dictionary – see also the section “Lemma Lacuna” below)
hold-spil
hold spil
lamme koteletter (twice) instead of *lammekoteletter* (lamb chops)
fornylig instead of *for nylig* (recently)
så som instead of *såsom* (such as)
somregel instead of *som regel* (usually)
ligemeget instead of *lige meget* (of no consequence

The above-mentioned reference book Handbook of Contemporary Danish contains a section that explains the use of one or more words, both in regard to compounds and prepositions. Based on the search results from the log file, it may be relevant to include similar information on this issue in the Danish Internet Dictionary, e.g. under the heading “Sprognormer” (linguistic norms).

Non-existing Words

A study of the not-founds also reveals that a number of searches have been performed for words that are non-existing. The top 100 contains 4 such words, ie

bekræftigelse instead of *bekræftelse* (confirmation)
omstændig instead of *omstændelig* (laborious)
forespørgelse instead of *forespørgsel* (request)
bekræftige instead of *bekræfte* (confirm)

The top 500 contains a further 4 words, i.e.

forspørgelse instead of *forespørgsel* (request)
privilegie instead of *privilegium* (privilege)
implementation instead of *implementering* (implementation)
imidlertidig instead of *imidlertid* (however) or *midlertidig* or (temporary)

In some of these cases, however, e.g. *privilegie* and *implementation*, it is debatable whether the words exist or not. A search on Google shows that these forms are very widely used, although the correct forms are *privilegium* and *implementering.*

Other examples of non-existing words occurring in the log file include:

gravbil instead of *rustvogn* (hearse)
beskrivning instead of *beskrivelse* (description)
fåreunge instead of *lam* (lamb)
smukhed instead of *skønhed* (beauty) (interestingly, a search on Google actually returned 631 hits on smukhed)
stavningsfejl instead of *stavefejl* (spelling mistake)
udspørgelse instead of *udspørgning* (questioning)
spørgerier instead of *spørgsmål* (questions)
selvgroet instead of *hjemmedyrket* (home-grown)
forbedrelse instead of *forbedring* (improvement)
antiautoritetstro (unorthodox)
indlandssø instead of *indsø* (lake)
selvudvalgte instead of *selvvalgte* (self-elected)
grisling instead of *gris* (piglet)
sideeffekt instead of *bivirkning* (side-effect)
hævnbegærlig instead of *hævngerrig* (vindictive)
middelaldrende instead of *midaldrende* (middle-aged) (Again, a search on Google showed that this is a common mistake - 636 hits were returned on middelaldrende)
personlighedssplitning instead of *personlighedsspaltning* (split personality)
priviligisere instead of *privilegere* (privilege)
nervøshed instead of *nervøsitet* (nervousness)
kampduelig instead of *kampdygtig* (effective)
ignorantisk instead of *ignorant* (ignorant)
gæstevenlighed instead of *gæstfrihed* (hospitality)
forpligtning instead of *forpligtelse* (obligation)
urelevant instead of *irrelevant* (irrelevant)
brudgift instead of *medgift* (dowry)
hangris instead of *orne* (boar)
handue instead of *duerik* (male pigeon)

The above examples generally fall into the following categories:

1. Incorrect/unusual word formation:

bekræftigelse (bekræftelse – confirmation)
forespørgelse (forespørgsel – request)
bekræftige (bekræfte – confirm)
forspørgelse (forespørgsel – request)
privilegie (privilegium – privilege)
implementation (implementering – implementation)
beskrivning (beskrivelse – description)
udspørgelse (udspørgning – questioning)
spørgerier (spørgsmål – questions)
forbedrelse (forbedring – improvement)
nervøshed (nervøsitet – nervousness)
ignorantisk (ignorant – ignorant)
stavningsfejl (stavefejl – spelling mistake)
forpligtning (forpligtelse – obligation)
urelevant (irrelevant – irrelevant)

2. Another word exists:

imidlertidig (imidlertid – however or midlertidig – temporary)
gravbil (rustvogn – hearse)
fåreunge (lam – lamb)
selvgroet (hjemmedyrket – homegrown)
indlandssø (indsø – lake)
sideeffekt (bivirkning – side-effect)
hævnbegærlig (hævngerrig – vindictive)
middelaldrende (midaldrende – middle-aged)
personlighedssplitning (personlighedsspaltning – split personality)
gæstevenlighed (gæstfrihed – hospitality)
kampduelig (kampdygtig – effective)
brudgift (medgift – dowry)
hangris (orne – boar)
handue (duerik – male pigeon)

As mentioned above, it may be debatable whether all of the examples listed in group 1 may be defined as non-existing words. For example, a search on Google returns a number of hits for *implementation* although the correct form is *implementering.* This means that such words exist in common usage even though they are not officially authorized. Many of these incorrect word formations, particularly those concerning the endings *-else/-ing* and *-ing/-ion*, are very common, and it may therefore be a good idea to include them in the dictionary with a reference to the correct term. This is particularly relevant for the words occurring in the top 100 and top 500 of the most common not-founds.

It is much more difficult to make allowances for words like those listed in category 2 as very few users have searched for them. Thus, these words are not part of common usage, and none of them occur in the top 100 or top 500 either.

Linking Morphemes

Neither the top 100 nor the top 500 of most common not-founds contain any examples of problems related to linking morphemes.

A study of 200 consecutive searches in a random place in the list of not-founds from the log file yields only two examples, *tidalder* instead of *tidsalder* (age or era) and *vidensproget/videnssproget* (the language of knowledge). A search for *tidalder* reveals that two different users have searched for this word at two different occasions, whereas only the one search has been carried out for *vidensprog/videnssprog*.

Other words related to the issue of linking morphemes that can be found in the log file include *kontraktsforhold* (contractual relationship), which has also been searched for at two occasions, and *isterningmaskine/isterningemaskine/isterningsmaskine* (ice machine). According to the Danish Internet Dictionary, the correct form is *isterningsmaskine*, and the other two alternatives have references to this term.

In connection with compounds containing *kontrakt-/kontrakts-*, a search in the list of not-founds from the log file shows that users are often unsure of whether they should include the linking morpheme *-s*. A search for *kontrakts**, i.e. compounds starting with *kontrakts-*, reveals not only 2 hits on *kontraktsforhold*, but also 3 hits on *kontraktsmæssig* (conforming to the contract), 2 on *kontraktspart* (party to a contract), *kontraktsgrundlaget* (the contractual basis), *kontraktssum* (contractual amount) and *kontraktsret* (contract law) as well as 1 hit on *kontraktsbeløb* (contractual amount), *kontraktsperiode* (term of contract), *kontraktsindgåelse* (formation of contract), *kontraktsafdeling* (contract department), *kontraktsvilkår* (terms of a contract), *kontraktsgrundlag* (contractual basis), *kontraktslige* (contractual) and *kontraktsfaktura* (contract invoice). Similarly, a search for *kontrakt**, i.e. compounds starting with *kontrakt-*, returns 4 hits on *kontraktmodel* (model contract), 3 on *kontraktret* (contract law), 2 on *kontraktansættelse* (employment relationship on a contractual basis) and *kontraktgrundlag* (contractual basis) and 1 hit on *kontraktgaranti* (contractual guarantee), *kontraktindgående* (contracting), *kontraktpart* (party to a contract), *kontraktgæld* (contractual debt), *kontraktbinding* (binding effect of a contract), *kontraktperiode* (term of contract), *kontraktmodellen* (the model contract), *kontraktafdeling* (contract department), *kontraktfrihed* (freedom of contract), *kontraktvilkår* (terms of a contract) and *kontraktfaktura* (contract invoice).

Thus, it seems that only certain compounds cause problems for users, whereas problems relating to linking morphemes in general do not seem to be very common. In the case of compounds that occur more than once, e.g. *tidalder* and the compounds containing *kontrakt-/kontrakts-*, it may be considered whether it would be relevant to include these words in the dictionary with a reference to the correct form as it has been done with *isterningsmaskine*.

Lemma Lacuna

Undoubtedly, the most obvious way that log files can be used to improve internet dictionaries is as a tool to discover lemma lacuna. Particularly the lists of the 100 and 500 most common not-founds are interesting in this connection as a large number of users have searched for the terms included on these lists, i.e. the terms are commonly used. Consequently, it is very relevant to add these words to the dictionary. The top 100 list contains 6 examples of words that can be classified as lemma lacuna, whereas the top 500 contains a further 53 such words:

Top 100:

sparring (sparring)
site (site)
adsl (adsl)
suboptimering (suboptimization)
franchise (franchise)
volatilitet (volatility)

Top 500:

reliabilitet (reliability)
revitalisering (revitalisation)
pluralistisk (pluralistic)
revitalisere (revitalise)
prævalens (prevalence)
eksplorativ (explorative)
prospektiv (prospective)
invasiv (invasive)
appendix (appendix)
portfolie (portfolio)
synops (synopsis)
interdependens (interdependence)
inferens (inference)
respektiv (respective)
debit (debit)
mix (mix)
per se (per se)
præjudicerende (prejudicial)
efficiens (efficiency)
pønal (penal)
fasi (fasi)
tilretning (trimming)
vanhjemmel (defective title)
intertekstualitet (intertextuality)
akkomodation (accomodation)
metroseksuel (metrosexual)
alzheimer (Alzheimer)
integrator (integrator)
rucola (rocket)
kollegie (student hostel)
ankebegæring (notice of appeal)
endemisk (endemic)
inkremental (incremental)
fokal (focal)
submissiv (submissive)
pizzeria (pizzeria)
resektion (resection)
forforståelse (preconception)
avatar (avatar)
dysfunktionel (dysfunctional)
kritikalitet (criticality)
replication (replication)
e-learning (e-learning)
semitisme (semitism)
addendum (addendum)
programpakke (programme package)
kontraherende (contracting)
eradikation (eradication)
moderator (moderator)
metadata (metadata)
nomotetisk (nomothetic)
emergens (emergence)

A large proportion of these terms, more specifically 38, are technical terms that may be classified as follows:

Computer-related terms:

Top 100: *site* (site), *adsl* (adsl)

Top 500: *integrator* (integrator), *avatar* (avatar), *e-learning* (e-learning), *programpakke* (programme package)

Financial terms:

Top 100: *volatilitet* (volatility), *franchise* (franchise)

Top 500: *debit* (debit), *efficiens* (efficiency)

Legal terms:

Top 100: none

Top 500: *per se* (per se), *præjudicerende* (prejudicial), *pønal* (penal), *vanhjemmel* (defective title), *ankebegæring* (notice of appeal), *kontraherende* (contracting)

Medical terms:

Top 100: none

Top 500: *prævalens* (prevalence), *invasive* (invasive), *alzheimer* (Alzheimer), *fokal* (focal), *resektion* (resection), *replikation* (replication), *eradikation* (eradication)

Other:

Top 100: none

Top 500: *reliabilitet* (reliability), *eksplorativ* (explorative), *prospektiv* (prospective), *interdependens* (interdependence), *inferens* (inference), *intertekstualitet* (intertextuality), *akkomodation* (accomodation), *endemisk* (endemic), *inkremental* (incremental), *submissiv* (submissive), *dysfunktionel* (dysfunctional), *kritikalitet* (criticality), *moderator* (moderator), *nomotetisk* (nomothetic), *emergens* (emergence).

The main question arising from this analysis is whether such terms should be included in a dictionary like THE DANISH INTERNET DICTIONARY, or whether this dictionary should merely contain words from the common language. We suggest that at least the most common of the technical terms, e.g. *site* (site), *franchise* (franchise), *e-learning* (e-learning) etc., which are part of everyday usage, should be included in THE DANISH INTERNET DICTIONARY, whereas highly technical terms such as *nomotetisk* (nomothetic) or *fokal* (focal) might be left to specialised dictionaries.

Lemmas not searched for

Another interesting aspect to consider in connection with dictionary use is whether the dictionary is actually used to its full extent, i.e. whether all dictionary entries have been searched for, and if not, how many of the total number of entries have never been requested by the users.

According to the statistics for THE DANISH INTERNET DICTIONARY, a total of 104,097 orthographical words have been searched for. This figure constitutes approximately one third of all possible searches as it is possible to search for inflections of a particular lexeme, i.e. for the head word itself as well as for all grammatical forms listed in the field of grammatical inflections. As mentioned earlier, however, passive and imperative forms are not included, and neither is the genitive form of nouns. In other words, if only one third of all entries have been searched for, approximately two thirds of the entire dictionary are not used in practice. But this can only be true if the user searches for inflected forms as often as for the non-inflected form, the lemma. Normally, or at least more frequently, the user will use the lemma sign as a search string; therefore we assume that the 104,097 search strings represent more than one third of the lemmas.

In order to establish whether this assumption is true and whether it is possible to discern a pattern in the words that have not been searched for, we examined the first 100 entry words of THE DANISH INTERNET DICTIONARY starting with the letter *b*. The result was that 52 of these words had been searched for (eg *B-aktie* (B share), *babysitter* (babysitter) and *bacon* (bacon)), whereas 48 had not (eg *B-dur* (B major), *B-skål* (B cup) and *babysprog* (baby talk)). Clearly, this corresponds to the above estimate that more than one third of the lemmas in the dictionary are actually used, and we believe that this small test is representative of the entire dictionary. The really interesting question is: By how much will the number of used articles in the dictionary increase when we have got a bigger number of log-filed lookups in the dictionary? We do not believe that the dictionary will ever be used to its full extent, but this topic is indeed interesting. We intend to make further investigations into "lemmas not searched for".

Obviously, it is not possible to discern a distinct pattern on the basis of the above examination, e.g. that certain types of words, such as semantic or orthographical variants, are never requested. However, it is still unclear whether such an investigation would be of practical use to lexicographers. It depends on whether a systematic description of the requested words compared with the non-requested words is possible.

5. Perspectives

It is beyond the scope of interpreting log files to give a detailed report of the approximately 2,000 emails that the dictionary makers behind THE DANISH INTERNET DICTIONARY have received from users. About half of these users merely express their thanks to the dictionary makers for the dictionary and report that they use it often, whereas less appreciative emails were received during a period of technical problems with the server; the reason being that, in the case of a dictionary free of charge, the user expects to have access to his tool whenever he needs it. In many other cases, the users propose new lemmas or report spelling mistakes found in the dictionary. All this, however, is not the topic of this paper. We believe to have shown that log files can be used by dictionary makers for improving their dictionary, in this case for improving the lemma selection. The use of a better search system which includes the possibility of a direct search for synonyms, collocations, antonyms, word formations, grammar items etc. will provide a much more precise way of obtaining knowledge about real dictionary use. On the basis of such data, it will be possible to prepare much better internet dictionaries. However, this is only possible if the necessary funding to do this work is granted by national or private organisations in the future, or if the charging of a pay per view fee or a monthly or yearly payment for using quality internet dictionaries becomes more common.

6. Literature

Atkins, B.T. Sue (ed.) 1998: *Using Dictionaries. Studies of dictionary use by language learners and translators.* Tübingen: Niemeyer.

Allgemeines Wörterbuch Deutsch-Französisch http://site.ifrance.com/allinfor/dico/index.htm

Bergenholtz, Henning 1996: Håndbogens dage er talte. In *HHÅ info 5, No 3*, 1996, 2-3.

Bergenholtz, Henning 2002: Das de Gruyter Wörterbuch Deutsch als Fremdsprache und das neue DUDEN-Wörterbuch in zehn Bänden. Ein Vergleich im Hinblick auf die Grammatik. In *Perspektiven des pädagogischen Lexikographie des Deutschen II. Untersuchungen anhand des de Gruyter Wörterbuches Deutsch als Fremdsprache.* Hrsg. von Herbert Ernst Wiegand. Tübingen: Niemeyer, 36-56.

Bergenholtz, Henning/Sven Tarp 2003: Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. In *Hermes 31*, 2003, 171-196.

Büdenbender, Stefan: *QuickDic*. http://quickdic.org/index_d.html

de Schryver, Gilles-Maurice/David Joffe 2004: On How Electronic Dictionaries are Really Used. In Geoffrey Williams/Sandra Vessier (Eds.): *Proceedings of the Eleventh EURALEX International Congress, Euralex 2004, Lorient, France. July 6-10, 2004.* Volume I. Lorient: Université de Bretagne, 187-196.

Dictionary of Accounting Danish-English = Sandro Nielsen/Lise Mourier/Henning Bergenholtz 2004: *Regnskabsordbogen dansk-engelsk*. København: Thomson.

Grostabussiat, Pascal: *Fransk-Svenskt Lexikon*.
http://www.azoria.com/lexikon/indexsw.shtml

Handbook of Contemporary Danish 2002 = Galberg Jacobsen, Henrik/Peter Stray Jørgensen: *Politikens Håndbog i Nudansk*. København: Politikens Forlag.

Simonsen, Henrik Köhler 2000: Papirordbogen er død! Computerordbogen længe leve! In *MDTnyt nr. 3*, 19-35.

TDC Online http://links.tdconline.dk/top.php?cid=2502&linktype=2

the Danish Internet Dictionary = Henning Bergenholtz/Vibeke Vrang in cooperation with Lena Lund, Helle Grønborg, Maria Bruun Jensen, Signe Rixen Larsen, Rikke Refslund, Jette Pedersen 2002: *Den Danske Netordbog*. Database and layout: Richard Almind. http://www.netordbog.asb.dk 2002-2004.

the Danish-English Dictionary of Accounting = Sandro Nielsen/Lise Mourier/Henning Bergenholtz in cooperation with Mads Melgaard, Trine Middelboe, Brit Sørensen, Helle Grønborg 2003: *Den Dansk-Engelske Regnskabsordbog/the Danish-English Dictionary of Accounting*. Database and layout: Richard Almind.
http://www.regnskabsordbogen.dk/iasdk 2003-2004.

Vilkår 1982 = *Vilkår for ordbogsarbejde i Danmark. Betænkning afgivet af det af Ministeriet for kulturelle anliggender nedsatte ordbogsudvalg.* Betænkning nr. 967 København 1982.

Wiegand, Herbert Ernst 1977: Einige grundlegende semantisch-pragmatische Aspekte von Wörterbucheinträgen. Ein Beitrag zur praktischen Lexikologie. In *Kopenhagener Beiträge zur Germanistischen Linguistik 12*, 59-149.

Wiegand, Herbert Ernst 1985: Fragen zur grammatik in Wörterbuchbenutzungsprotokollen. Ein Beitrag zur empirischen Erforschung der Benutzung einsprachigen Wörterbuch. In *Lexikographie und Grammatik. Akten des Essener Kolloquiums 1984*, hrsg. von Henning Bergenholtz und Joachim Mugdan. Tübingen: Niemeyer, 20-98.