AN ODD COUPLE – CORPUS FREQUENCY AND LOOK-UP FREQUENCY: WHAT RELATIONSHIP?[1]

Lars TRAP-JENSEN, Henrik LORENTZEN, Nicolai H. SØRENSEN
Society for Danish Language and Literature

Trap-Jensen, L., Lorentzen, H., Sørensen, N. (2014): An odd couple – Corpus frequency and look-up frequency: what relationship? Slovenščina 2.0, 2 (2): 94–113.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_07.pdf.

In this paper, we investigate the relationship between log file records and corpus frequency. The study was motivated by practical considerations of how best to keep an already existing corpus-based dictionary updated. Should the next word in the dictionary be the one that follows next on a list of declining corpus frequency? Or the one that users most frequently look up but don't find? In order to establish manageable criteria, we analysed log files for The Danish Dictionary from 2009 to 2012 and compared the list of most popular words looked up by the users with the frequency of the same words in the corpus underlying The Danish Dictionary. The users' actual search behaviour was analysed in order to find answers to questions such as these: Are there words which are never looked up? If so, can we say something meaningful about their corpus frequency patterns – do they belong to particular parts of speech, are they particularly frequent or infrequent, could it even be that the pattern is cumulative, in such a way that a particular threshold can be identified? Ultimately, the question is whether it makes sense to use corpus frequency as a criterion for lemma selection.

Keywords: corpus frequency, lemma selection, look-up behaviour, updating dictionaries, log files

[1] This article is based on an investigation carried out in the autumn 2013. The investigation and its findings were presented at the conference eLex 2013, electronic lexicography in the 21st century: thinking outside the paper (cf. http://eki.ee/elex2013). The text is a revised version of a Danish article: Trap-Jensen (2014).

1 BACKGROUND

One significant consequence of the digital revolution in lexicography is the possibility to follow the users' actual behaviour in a digital dictionary. In this article, we address the question whether lexicographers should take advantage of this knowledge and use it as a guideline in the lemma selection procedure.

More specifically, our starting point was the maintenance of an existing dictionary, *The Danish Dictionary* ('Den Danske Ordbog', DDO), and the task of updating it with new entries. In this context, 'new entry' should be understood in a broad sense: any entry not previously included. It may be but does not have to be a neologism, it can also be an existing word or expression that has so far been neglected or opted out.

Being a corpus-based dictionary, DDO shares with similar dictionaries the assumption that corpus frequency reflects, or is at least closely connected with, the usage of a given word in the language. It follows that the case for including a word in the dictionary is strengthened as the word's corpus frequency increases. But we don't really know if frequent words are also the ones that users actually look up, for the simple reason that we so far haven't had sufficient empirical evidence to assess the relationship between corpus frequency and search behaviour. Previous studies have typically had a moderate empirical basis: de Schryver and Joffe (2004) was based on 21,337 look-ups, and Bergenholtz and Johnsen (2005) on 1,016,960 look-ups. However, this situation is gradually changing as online dictionaries grow older. DDO has been online since November 2009 and during the entire period all queries have been registered and stored in a query log. The only study we know of with a comparable empirical basis is Bergenholtz and Norddahl (2012), which investigated log files from *Den Danske Netordbog* ('The Danish Online Dictionary') for a nearly identical period of time, 31½ months, and with an average monthly number of successful look-ups of 568,062, equivalent to around 80 percent of the corresponding look-ups in DDO in this study. (A

look-up is successful if the item searched for is found in the dictionary.)

We decided to investigate the words that users actually look up and compare the results with the corpus representation of the same words, that is their frequencies. In particular, we wanted to examine the existing vocabulary of DDO to discover the patterns of search behaviour. By comparing the query log with corpus representation we hoped to find answers to questions such as the following:

(1) What DDO words are frequently looked up and how can they be characterised? Are they frequent or non-frequent words, and do the words belong to certain parts of speech?

(2) What other words, not attested in the dictionary, are users looking up and how can they be characterised?

If the answers to questions (1) and (2) differ markedly, it would be a motivation for us to change the principles of lemma selection for the future. However, before we address these questions, a few words about the design of the study are appropriate.

2 LOG FILE INVESTIGATION

2.1 Choice of method and reservations

We decided to use the original DDO corpus for the study (cf. Norling-Christensen and Asmussen 1998). This corpus is characterised by a well-balanced distribution with respect to genres, media and the sociological parameters of the authors, thus ensuring good overall validity of the findings. On the other hand, there are also obvious shortcomings. First, with c. 40 million running words, the corpus is fairly small by modern standards, and second, it is not up to date. The corpus covers the period 1983–1992, which means that, obviously, there are no examples of neologisms that have come into the language later than that (cf. also Lorentzen and Nimb 2011). In spite of these

shortcomings, we considered quality and balance to be more important than the quantitative and temporal considerations.

Some reservations should also be mentioned about the query log. It only records what the users have entered into the search box, and you can never be entirely sure what the user's intention was at the time of consultation. Consequently, if a string matches more than one lemma, for example in the case of homography or coinciding inflectional forms, it has been counted as a match for both words. Furthermore, queries may be performed by humans or robots. Robot crawling can be useful for various purposes but not in this case. We were only interested in look-ups by real humans and we have therefore tried to eliminate Google crawlers, malicious hacker attacks and our own test queries from the log. However, it is not easy to remove all and only these instances and nothing else. Sometimes too much may have been removed, sometimes too little. Meta queries are another case in point. When users write 'synonym' or 'dictionary' into Google's search box and are referred to DDO's entries for these words, it is more likely that they were in fact looking for the resources rather than the entry articles. But you can never be 100 percent sure. An example of the query log records is given in Figure 1.

id	query	normalized	caller	ip_address	found	created
228751	rekvisition	rekvisition	query	*A2E6DE5155C2E3E243B692EC03A794C1E26C6700	yes	2009-12-01 01:56:05
228759	forælder	forælder	query	*144363857FF3C34DEBB2C2AECEC0CC05C39D515B	yes	2009-12-01 02:05:21
228763	trapeze	trapeze	query	*0A73B32E298808D0FFA3FD8201B7CF8EF170A7B7	no	2009-12-01 02:07:13
228772	forstillelse	forstillelse	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:12:27
228777	vanskelig	vanskelig	query	*9716036C15A5DB086E6ECB345D72A6FF1F11CE9B	yes	2009-12-01 02:14:45
228784	skrømt	skrømt	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:17:08
228785	kareen	kareen	query	*606D60B3C30169187097 59B2B44AE7BCCEF68588	no	2009-12-01 02:17:47
228787	kapere	kapere	query	*606D60B3C30169187097 59B2B44AE7BCCEF68588	yes	2009-12-01 02:17:54
228788	karreen	karreen	query	*606D60B3C3016918709759B2B44AE7BCCEF68588	no	2009-12-01 02:18:07
228793	værge sig	værge sig	query	*4FDD3D9192DD1DE1F2C831D9971B9BC96347E394	yes	2009-12-01 02:19:13

Figure 1: Example from the query log.

2.2 Extent of the study and general results

The search behaviour was examined for a period of three years, from December 2009 until December 2012. In this period, the users executed a total of 29,551,938 look-ups in DDO, 2,208,872 of which were different. Out of the total number of look-ups, 24,478,138 were successful in the sense that they had a direct match in the base, whereas 5,073,800 look-ups were unsuccessful.

A simple comparison between the stock of lemmas in DDO and the query log is interesting in itself. It may be able to confirm or dispel some of the myths that have developed among lexicographers and dictionary users during the time when we did not know which words the users looked up. One such claim is that many words in a dictionary are never looked up. Taken literally, this claim can be dismissed quite easily. Altogether, there are only 202 entries for which the query log has no record, corresponding to 0.2 percent of the headwords, the total number of which being just under 100,000. However, a reasonable objection would be that a frequency of 0 is a very strict condition since some queries may be the result of undetected robot or hacker activity. If we instead set the threshold to 3 look-ups, the number of entries which are never (or almost never) looked up increases to about 5,000, and if we set the threshold at 10, this number increases to 16,000 entries. If this exercise is continued, with increasing threshold values and corresponding look-up frequencies, the correlation can be shown graphically. This is what we have done in Figure 2. Here you can see that many words have a fairly low look-up frequency, whereas only 10-20,000 words are looked up frequently. This is encouraging news for makers of small to medium-sized dictionaries. It seems that they are in fact able to cover the basic needs of most users.

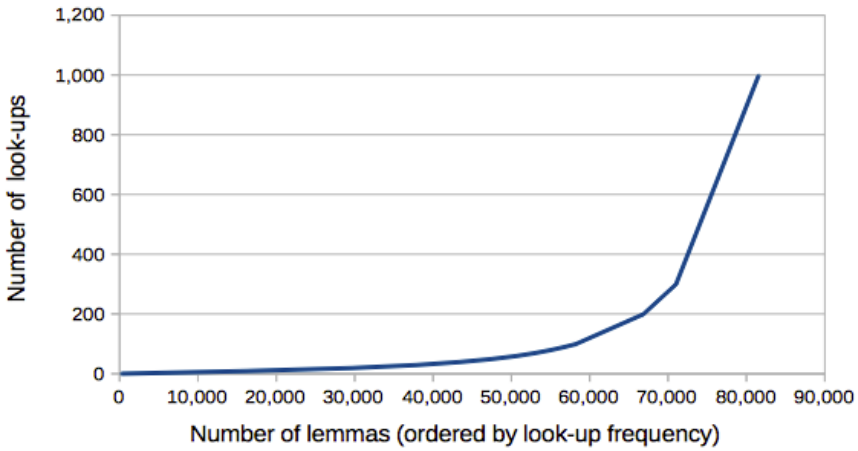

Figure 2: Correlation between look-up frequency and number of lemmas.

It may be a myth that many words are never looked up, but there is nevertheless some truth in it as many words are only looked up rarely. At the other end of the scale, it is the case that a limited number of words, around 20,000, have a very high look-up frequency.

The finding is noteworthy because it differs significantly from Bergenholtz and Norddahl (2012), who in their investigation found that 66.6 percent of the entries in *The Danish Online Dictionary* had a look-up frequency of 0. The difference is so pronounced that it cannot be coincidental. Among the possible explanations could be that: 1) the stocks of lemmas in the two dictionaries are basically different, 2) the dictionaries have very different user groups with different needs, or 3) the access paths to the contents of the dictionaries are so different that they lead to different search patterns. Intuitively, the last explanation seems most convincing: Whereas DDO is offered as a free service and has most of its users coming via Google searches, the look-ups in *The Danish Online Dictionary* (beyond two free daily look-ups) come primarily via queries at the pay service *ordbogen.com*. However, the reason for this difference may also be a combination of the factors mentioned or involve others not mentioned. It deserves further investigation.

One thing that needs to be taken into account when comparing the results of the two studies is the temporal aspect. Obviously, words that were added to the dictionary late in the period of study have less chance of being looked up than words that were accessible during the entire period. For example, the word *sprogteknolog* ('language technologist') is among the words that have a search frequency of 0, but as this entry was included in an update as late as 29 November 2012 (only a month before the end of the period being studied), the statistics do not give a true picture. Fortunately, the problem is not very grave as the number of entries added during the period is only 1,426.

Another claim which is often heard is that the common, basic words of a language are rarely, if ever, looked up. They are the words that we all know and use, and therefore we don't need to look them up, the argument goes. The records of the query log tell a different story and the claim can also be readily rejected as a myth. Among the most frequently looked-up words are many function words (the closed classes of prepositions, pronouns, conjunctions etc.): 64 function words among the 500 most looked-up words overall. The same is true in general of the most common words of the language: of the 1,000 most looked-up words, more than a quarter are also among the 1,000 most frequent words in the corpus, and nearly 60 percent belong to the 10,000 most frequent words in the corpus. The reverse tendency holds equally: among the words looked up 0, 1 or 2 times not a single one is a function word. The group of most rarely looked-up words consists largely of nouns and adjectives, with no simplex words among them. Typical examples of rare queries are compounds or multiword expressions such as *broderie-anglaise-flæse*, *middagsradioavis* and *rosenkålssuppe* ('broderie anglaise flounce', 'noon news bulletin', 'Brussels sprout soup').

The statistics might even be able to tell us more about who our users are. But in this study we have not investigated the distribution of queries on IP addresses and cannot say if the tendency remains when we take into account that some users are more diligent than others. Learners, for example, are typically busy

users with a need to look up function words and other basic vocabulary items. It would be interesting to see if the query log figures corroborate such a correlation but that was not possible in this study (for a qualitative approach, see Lorentzen and Theilgaard 2012).

3 QUERY LOG AND CORPUS REPRESENTATION

If corpus-based lemma selection is a healthy principle, it can be assumed that frequently looked-up words are also well represented in the corpus – bearing in mind the previously mentioned reservations about neologisms etc. If this is the case, they should either be in DDO already or be in the pipeline as part of the regular update routine.

One way of examining this is to look at the word forms in the corpus and their ability to accommodate actual queries. With this in mind, we ranked all the word forms in the corpus according to falling frequency and measured how large a percentage of the look-ups they were able to cover. The correlation is illustrated by Figure 3.

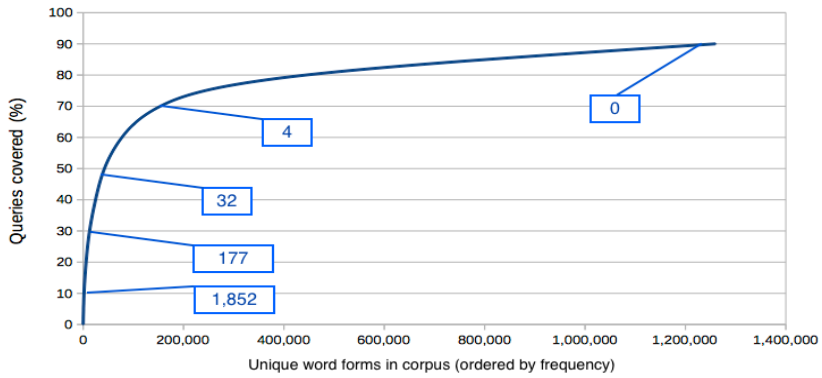

Figure 3: Correlation between corpus frequency and look-up coverage.

The figure shows that it takes 12,774 word forms to return 30 percent of the look-ups. Since the forms are ranked according to falling frequency, the absolute frequency of a given word form can easily be retrieved, in this case the

word form ranked 12,774 has an absolute frequency in the corpus of 177. Similarly, the absolute frequencies of selected percentages are shown in small boxes in Figure 3. For example, it appears that it takes 42,380 word forms to cover half the look-ups and the box pointing to the intersection between the curve and the 50 percent line indicates that the relevant word form occurs 32 times in the corpus, whereas the 70 percent limit corresponds to 4 forms in the corpus.

Continuing in this way, it would in principle be possible to return all the look-ups, but in reality it proves impossible for two reasons: first, it requires more word forms than the corpus contains. To reach 90 percent, 1,3 million forms are needed but the corpus only has about 600,000. Second, some search strings contain various kinds of 'noise', that is nonsense words and misspellings that are unlikely to occur in a corpus no matter how big it is. For that reason, it will probably never be possible to reach 100 percent even if the corpus size was multiplied.

What can be done instead, if only as an experiment, is to imagine that the corpus is a one-to-one reflection of the entire language, and only consider those look-ups that in fact return a match, i.e. getting rid of the noise by omitting the no-matches. The result of such an experiment is shown in Figure 4. The overall correlation is no different from the corresponding correlation found in Figure 3 but it differs in two important respects. If you look at the x-axis, you will see that considerably fewer occurrences are needed to cover the look-ups, now only about 600,000, which happens to equal the number of forms present in the corpus. And secondly, it is now possible to reach 100 percent – even if this should not be surprising as only successful look-ups were taken into consideration.

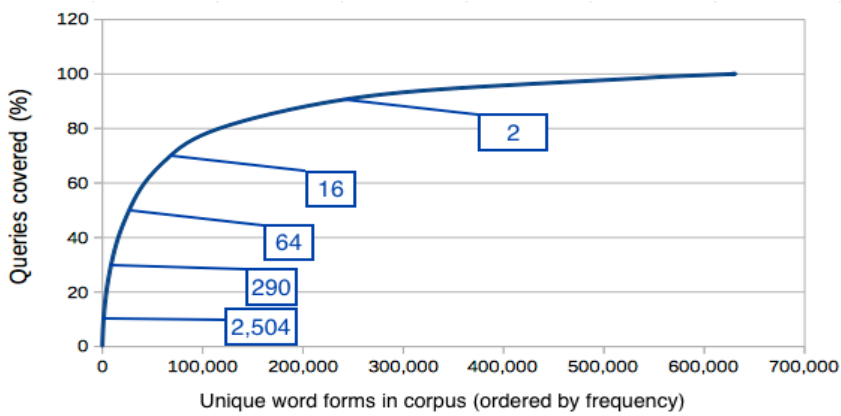

Figure 4: Correlation between corpus frequency and look-up coverage without no-matches.

In either case, the most important thing to note is the shape of the curve itself. We can see that it takes a large number of word forms to get from 80 to 100 percent. Moreover, the absolute frequencies of the forms are very low after the curve exceeds c. 100,000 occurrences. The curve flattens, and after that it takes a long time to reach the last part towards 100 percent. Our interpretation is that the usefulness of corpus data in lemma selection decreases after the point where the curve starts to level off. When this part of the curve has been reached, the next word on the list will have a frequency of only 1 or 2, which in practice means that one word is as good a candidate as another. As DDO has already reached this level, this is an important insight for us.

Our first conclusion of the study is that the corpus may be appropriate as a tool in lemma selection but it is not equally well suited in all phases of a project. It is most valuable for the first c. 100,000 word forms, diminishes gradually after that and after 200,000 forms the forms which follow have just 1 or 2 occurrences, and in these cases the corpus is not a particularly reliable basis for deciding what should be selected. On the other hand, it provides the editors with a good starting point, a pool of words from which to select lemma candidates. In this connection, it is of course important to distinguish between

a word form and a headword. For some headwords, the indeclinable ones, word form and headword coincide, whereas other headwords subsume several word forms, inflectional forms and spelling variants. As a rough estimate, there are 3.5 word forms per headword for Danish words in DDO.

One reservation that must be made about the conclusion is to do with corpus size. It is possible that the finding of this study is contingent on or otherwise related to the size of the underlying corpus. We have not been able to find clear evidence for the nature of a possible connection, so when it comes to generalizing to other projects, the reader must bear in mind that corpus size may have a role to play that really should be identified first.

4 CONTENTS OF THE QUERY LOG

Let us now take a look at the other condition: what the users actually look up. We were in particular curious to uncover the nature of the unsuccessful look-ups. In this connection, an unsuccessful look-up is defined as a look-up that does not have a direct match in the dictionary database. This may seem straightforward but even so it is useful to distinguish between different types of no-matches:

1. Look-ups that do not have a direct match but where the user gets a result after all. This is the case if the users use wildcards or, often, if they write multiple words in the search box.
2. Look-ups that have no direct match in the database and for which no result is shown, but where the user is guided to the right entry, either by means of alternative suggestions from the function 'Did you mean' or through a message informing the user that the word is available in another dictionary, in our case most often the historical dictionary *Ordbog over det danske Sprog* ('Dictionary of the Danish Language').
3. Look-ups for nonsense words or words so misspelled that the 'Did you mean' function cannot suggest appropriate alternatives.

4. Look-ups for words that are not in the dictionary.

The first type is probably not experienced as futile by the users. Even if the string does not have a direct match in the database and thus falls under the definition, the users will usually find what they were looking for. Similarly, the second type will not always be viewed as unsuccessful as the users are assisted in finding the right entry. The last two types are genuinely unsuccessful, technically as well as for the user experience. Here, the challenge is to distinguish what is noise and what are genuine words and potential lemma candidates.

To get an impression of the distribution of look-ups among the different types, 100 random no-matches were selected and their distribution analysed. The result is shown in Figure 5. In the figure, type 1) is classified as a direct match because the user does get a result from the look-up. 3) and 4) were merged because of the difficulty in deciding when a search string should count as a nonsense word or a misspelling and when as a real word. More on this later.

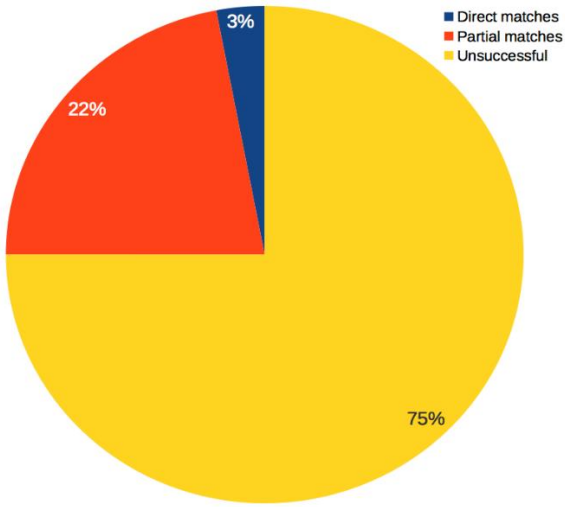

Figure 5: Distribution of no-matches by types among 100 randomly selected records.

It appears from Figure 5 that about a quarter of the no-matches are successful

or partially successful whereas three quarters are truly unsuccessful. Instead of looking at a random selection one could also consider the most frequent no-matches. This has been done in Figure 6.

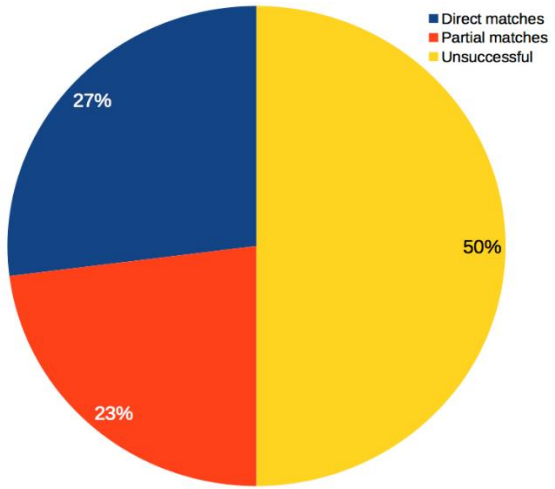

Figure 6: Distribution of no-matches by types among the 100 most frequent records.

Among the most frequent no-matches, half either give a result or offer guidance to an alternative, while the truly unsuccessful ones have been reduced to 50 percent, when compared with the random selection. The difference between the most frequent records and the random selection is mainly due to a higher proportion of successful look-ups among the frequent ones, either look-ups involving wildcards (*z*, *z, c*, x**), multiple words, previously official orthographic forms (for example *linie* 'line', *krem* 'cream', *bolche* 'sweet', *elvte* 'eleventh'), forms reproducing spoken language (*osse* 'also', *pive* 'squeak'), or abbreviations without an obligatory period (*su, ac, pt*). With our search algorithm these look-ups all lead to a result even if they have no direct match. On the other hand, the large number of rarely looked-up words make up a much larger proportion of the random look-ups, as explained below.

If we take a closer look at the unsuccessful look-ups and first consider the 100 most frequent ones, there are 8 proper nouns among the 50 true no-matches, including personal, geographical as well as company names such as *Jørgen, Danmark* and *TDC*, while the remaining no-matches are made up of nonsense lookups and possible lemma candidates. It is difficult to draw the line between nonsense and genuine look-ups, and in Figure 6 they are not discriminated. Some of the look-ups that seem meaningless could for example derive from participants in various word games checking the existence of words like *cu, po, zi* or *xa* or whatever annoying letters they may be confronted with in the game. They could also be queries performed automatically by hackers; it is difficult to know. The point is that more knowledge of the user's intention is required in order to classify a string correctly, and our division is no more than an estimate with the possibility of an occasional misinterpretation. We have estimated that some 35 of the look-ups are nonsense look-ups (examples are *bibhld, nyopmb, cu* (= 'see you'?)*, xa, config, ce, tilfr, tr, tac, npop, nbceq1visse, za, ci, zo, ma, xe*), while at most 6 qualify as true lemma candidates: *værv, %, tf, lol, malacostraca, tac*. The symbol *%* may also be a wildcard, *værv* could be a misspelling of *hverv*, 'task'; it is difficult to say when the string is all we have at our disposal. What the users had in mind when they wrote the string, we will never know. However that may be, probably no one would consider any of the possible lemma candidates crucial entries in a dictionary, and it is encouraging for the editorial staff to learn that there are very few true lemma candidates among the common no-matches, none of which are obvious candidates.

If we look at the random no-matches, the picture is somewhat different. Again, there are some proper nouns but also a substantially higher number of potential lemma candidates, for example *uvirkelighedsfølelser* 'feeling of unreality'*, brancheafdeling* 'business field'*, faghøjskoledanmark* 'folk high school Denmark'*, enzymgigantens* 'the enzyme giant's'*, pcbank* 'online banking'*, forskningskronerne* 'the research kroner'*, sluddertante* 'chatterbox'*, lydforhør* 'sound interrogation'*, udførelsesmetoder* 'execution methods'*, evg,*

jubilæumsmiddag 'anniversary dinner'*, lavadal* 'lava valley'*, differentiatorer* 'differentiators'*, hvilestue* 'resting room'*, vigilante*. Most of the words are compounds, quite a few are ad hoc compounds, many are semantically transparent, and there are also some neologisms and foreign words among them (*differentiatorer, vigilante*). Typically, they belong to that part of the lexicon where the curves of Figures 3 and 4 level off: they are by no means central words of the language but they are at least potentially words which according to the corpus-based principle could be selected for inclusion in the dictionary.

Finally, there is a residual group of misspellings (*fenimonal, illustraioner, tilrekkligt, forrørt*) and other words that may be difficult to interpret, typically just occurring once in the corpus (*femtinul, kulinaristisk, chemotolgy, caldérons, turos*).

So even if there are a few more genuine lemma candidates among the random look-ups, they are still peripheral, low-frequency words. An explanation why this is so can be found in the fact that we have already taken action on the basis of the no-match look-ups. In 2010 (cf. Lorentzen and Theilgaard 2012), we observed that the no-match lists contained inflected forms that were absent in the database, which led us to add these forms, most importantly present participle forms of verbs and genitive for nouns. Consequently, they no longer show up on the no-match lists. We have also added a number of common misspellings to the database, in particular words incorrectly written as either one or two words. Invisible to the user, these forms serve to improve the "Did you mean" function by increasing the chance of offering a correct alternative. The "Did you mean" function is adjusted regularly as new typos and misspellings are registered and added.

Among the real lemma candidates we have included some during the period of study, especially neologisms such as *swag* and *hipster*[2]. Among the remaining words there are some loanwords that are looked up fairly frequently and should be considered candidates for inclusion. They may not be frequent in the corpus, but that may have to do with the size of the corpus, the fact that the corpus does not include texts from the most recent period, or the representation of academic, specialist, technical or professional texts where these words are found.

The fairly high proportions of proper nouns among unsuccessful look-ups could also be a reason to rethink the traditions and editorial principles with an open mind. The most important argument against proper nouns in dictionaries has traditionally been considerations of space and the difficulties in delimiting the material but that does not really count in a digital world. It may be difficult for users to appreciate that they find derivations like *parisisk, cubaner* and *keynesiansk* ('Parisian', 'Cuban', 'Keynesian') in the dictionary but not *Paris, Cuba* and *Keynes*, and on the other hand, the orthography and etymology of proper names cause at least the same difficulty as the rest of the vocabulary. It makes an impression that proper nouns are so frequent among the words that users try to look up.

Finally, our dictionary contains a function where users are invited to send in reports of words they miss in the dictionary. The reports are stored in a database which is administered jointly by the Society for Danish Language and Literature and the Danish Language Council. Not surprisingly, there is a certain amount of overlap between these reports and the no-matches; examples include *egalitær* ('egalitarian'), *webinar, spilleliste* ('play list'), *stomp, zumba*, *app*,

[2] In theory, it poses a problem if there are many manual additions from the no-match list during the period of investigation because the quality of the remaining lemma candidates is thereby underplayed. However, the scope of the problem should not be overstated as it applies only to a limited number of instances. Besides, the examples mentioned would also have been selected on the basis of their frequency in the modern corpus.

skype (verb), *blingbling*, all examples of words that have been added to the dictionary subsequently. The value of the reported words is mixed, many proposals are hardly sincere, others have clear ad hoc status, but reports sent in independently by different users deserve to be taken seriously. At the moment, this database contains more than 6,000 posts.

5 CONCLUSIONS

The study has looked at lemma selection from different sources. The analysis of word form distribution in the corpus showed that a corpus is a good source but the value is highest for the first c. 20,000 headwords, after which point it gradually decreases. After that it is a good idea to take account of other sources. The query log is one possibility, especially if combined with user suggestions. No matter what method is used, it should be combined with the professional judgment of a skilled lexicographer, not least when it comes to the vocabulary beyond the stock of common words. We have seen that words belonging to the peripheral vocabulary typically have a low corpus frequency and for these words the best lemma selection principle is achieved by combining different methods. One could call this principle computer-aided introspection. Because one word is more or less as good a candidate as another, seen from a corpus linguistic point of view, the role of the editor becomes crucial. In spite of greater automation and the all-embracing use of computers in the editorial process, lexicographers and their competence are still much needed.

In itself, it is an interesting fact that so many of the words with a frequency of 1 or 2 are in fact looked up in our dictionary. If the ambition is to minimize the number of unsuccessful look-ups, it follows that the stock of headwords must be increased substantially.

REFERENCES

Bergenholtz, H., and Johnsen, M. (2005): Log Files as a Tool for Improving Internet Dictionaries. *Hermes* 34: 117-141.

Bergenholtz, H., and Norddahl, B. (2012): Ordbogsartikler, som ingen læser. *LexicoNordica* 19: 207-223.

DDO = *Den Danske Ordbog* ('The Danish Dictionary'). Copenhagen: Society for Danish Language and Literature. Available online at: http://ordnet.dk/ddo (27th October 2014).

Den Danske Netordbog ('The Danish Online Dictionary'). Available online at: http://www.ordbogen.com/ (27th October 2014).

de Schryver, G.-M., and Joffe, D. (2004): On How Electronic Dictionaries are Really Used. In G. Williams and S. Vessier (eds.): *Proceedings of the Eleventh EURALEX International Congress, Euralex 2004. Lorient, France. July 6-10. Volume I*: 187-196. Lorient: Université de Bretagne.

Lorentzen, H., and Nimb, S. (2011): Fra krydderkage til running sushi – hvordan nye ord kommer ind i Den Danske Ordbog. In M. H. Andersen and J. N. Jensen (eds.): *Nye ord, Sprognævnets Konferenceserie 1*: 69-85. Copenhagen: Dansk Sprognævn.

Lorentzen, H., and Theilgaard, L. (2012): Online dictionaries – how do users find them and what do they do once they have? In R. V. Fjeld and J. M. Torjusen (eds.): *Proceedings of the 15th EURALEX International Congress*: 654-660. Department of Linguistics and Scandinavian Studies, University of Oslo.

Norling-Christensen, O., and Asmussen, J. (1998): The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series* 8: 223-242.

Ordbog over det danske Sprog ('Dictionary of the Danish Language') 1-28 (1918-56) with Supplement 1-5 (1992-2005). Copenhagen: Society for

Danish Language and Literature. Available online at: http://ordnet.dk/ods (27th October 2014).

Trap-Jensen, L. (2014): Korpus eller brugerne – hvem får det sidste ord? In M. H. Andersen, J. N. Jensen and P. Jarvad (eds.): *Neologismer. Dansk Sprognævns 2. seminar om nye ord. København 5.-6. november 2013*. Sprognævnets konferenceserie 3: 129-144. Copenhagen: Dansk Sprognævn.

NENAVADEN PAR: POGOSTOST BESEDE V KORPUSU IN PRI UPORABNIŠKIH POIZVEDBAH

Prispevek se osredotoča na preučitev razmerja med dnevniki iskanj uporabnikov po spletnem slovarju in korpusno pogostostjo besed. Študijo so spodbudila razmišljanja, ki so se porajala pri rednem slovarskem delu in jih lahko strnemo v vprašanje: kako ohranjati na korpusu temelječ slovar aktualen? Bi morala biti naslednja beseda, ki jo uvrstimo v slovar, tista, ki sledi zadnji uslovarjeni besedi na frekvenčnem seznamu besed iz korpusa? Ali bi morala biti to beseda, ki jo uporabniki najpogosteje neuspešno iščejo v slovarju? Da bi prišli do ustreznih kriterijev, so avtorji analizirali dnevnike iskanj uporabnikov danskega slovarja v obdobju od 2009 do 2012 in seznam najpogosteje iskanih besed primerjali z njihovo pogostostjo v korpusu. S proučitvijo iskalnih navad uporabnikov so avtorji želeli priti do odgovorov na sledeča vprašanja: Ali so v slovarju besede, ki jih uporabniki nikoli ne iščejo? Če je odgovor da, ali lahko na podlagi njihove pogostosti v korpusu opazimo kakšne smiselne vzorce – gre za besede iste besedne vrste, so besede zelo pogoste ali zelo redke, se pojavljajo v določenem frekvenčnem območju? Ugotovitev prispevka je, da je pogostost v korpusu dober kriterij za 20.000 najpogostejših iztočnic, medtem ko je treba pri manj pogostih besedah dodati še druge metode, med katerimi je tudi pregled iskanj uporabnikov, nadvse pomembna pa je tudi presoja leksikografov.

Ključne besede: pogostost v korpusu, izbira lem, načini iskanj po slovarju, posodabljanje slovarjev, dnevniki iskanj uporabnikov