The Danish Dictionary at large: Presentation, Problems and Perspectives

Henrik Lorentzen
Det Danske Sprog- og Litteraturselskab
The Society for Danish Language and Literature
Christians Brygge 1
1219 COPENHAGEN K
DENMARK
hl@dsl.dk

Abstract
This paper is both a presentation of the first, completely corpus-based dictionary of Danish and a discussion of some of the general difficulties involved in long, large-scale dictionary projects. The first part of the paper concentrates on features that reflect the editors' descriptivist point of view, i.e. non-standard usage, statistically based collocations and authentic examples. In the second part problems such as finding a common practice in definition writing and sense discrimination, the grouping of tasks and the eternal question of time are discussed. At the end of the paper the outlines of future projects are presented.

1. Introduction

Last year, in November 2003, a long-expected dictionary came out in Denmark: the first volume of a six-volume, monolingual dictionary of modern Danish was published after more than 12 years' work, The Danish Dictionary. This paper will deal with three issues: firstly I am going to present the printed dictionary and give examples of the most remarkable features; then I intend to talk about the project, the making of the dictionary, the difficulties we encountered in the process and to discuss the lessons we have learnt. To conclude, I should like to suggest some of the future work that we would like to do if sufficient funding can be found.

2. Presentation – the dictionary

The Danish Dictionary, the first, entirely corpus-based dictionary of Danish, is a printed dictionary in six volumes. It contains about 100 000 headwords of which 60 000 have their own entry and another 40 000 are mentioned as derived and composed words. It conveys information on spelling, morphology, pronunciation, meaning, collocation, fixed phrases, sentence-construction, usage, word formation and etymology. The data in the dictionary were compiled from The Corpus of the Danish Dictionary, a general text corpus of about 40 million words (cf. Asmussen and Norling-Christensen 1998), in conjunction with other sources of information such as newspaper corpora, informants and the Internet. The main purpose was to give as true as possible a picture of the Danish language as it was written and spoken towards the end of the 20th century. This descriptive point of view is reflected in the

dictionary in different ways, among others by the features that I am going to explain more in detail, viz.

- spoken language
- non-standard usage
- statistically based collocations
- authentic examples

2.1 Spoken language

When The Corpus of The Danish Dictionary was established, the editors had the opportunity to include a lot of spoken language, in fact just under 20 per cent of the corpus, which means almost 8 million words. It is even today the largest corpus of spoken Danish, and it gave the editors a unique possibility to give a lexicographical description of the spoken language on an empirical basis. For instance, it allowed us to include **interjections** that were registered in few, if any, dictionaries, like *ad* ('yuck', expressing disgust), *arh/ahr* (expressing reservation, disagreement or doubt; close to 'well'); **pragmatic phrases** like *og alt det der, og alt sådan noget* (lit. 'and all that, and all that kind of thing', equivalent to 'and so on, and so forth'), *det er jo det* ('that's it'); **informal phrases** like *det kan du bande på* (lit. 'you can swear on that', meaning 'you bet'), *være dum til* (lit. 'be stupid at', meaning 'be bad at something, e.g. swimming or driving').

For further information on spoken language in The Danish Dictionary, see Trap-Jensen 2004.

2.2 Non-standard usage

2.2.1 Spelling. When you work with a corpus of authentic, genuine language, you discover very quickly that it quite often does not conform to the rules set out by authorities such as teachers, language councils and spelling dictionaries. The divergences from the standard can be found in various areas: spelling, inflexion, meaning, construction. How will you handle this in the dictionary compiled from the corpus? The policy adopted in The Danish Dictionary is both descriptive and normative. Let's take spelling first. Whenever an entry word is in bold typeface, it is acknowledged as official orthography by the Danish spelling dictionary, but we also have entry words that are not official orthography. They are printed in ordinary typeface and direct the user to the correct form. In many cases the unofficial forms are also mentioned in the entries for the correctly spelled forms. Below you will see a few examples of this.

The official Danish spelling of the word *basar* is with an *s*. In Danish texts, however, a lot of occurrences with a *z* are found, probably inspired by the spelling in many other languages, for instance French, from which the word was borrowed. In The Danish Dictionary this fact is reflected in the entry *bazar*, that has a different typography than other headwords and tells the user that this spelling is unofficial, but frequent, and that he/she has to look up the correct form *basar*.

bazar sb. *uofficiel, men alm. stavemåde af* →basar.

basar sb. fk. *uofficiel, men alm. stavemåde*: bazar.
...

Another example is the word *beautyboks* ('vanity case'). Many Danes believe that this is a word borrowed from English, and consequently they spell it with an *x* and often also with an unauthorized hyphen. In this case the unofficial forms have no entry of their own because they would be alphabetical neighbours to the officially spelled form.

beautyboks sb. fk. *uofficielle, men meget alm. former*:
beautybox, beauty-box
...

The purpose of this is not only a practical and pedagogical one (leading the user to the right entry), but also a descriptive one (describing the language as it is, not what it 'ought' to be).

2.2.2 Inflexion. As far as inflexion is concerned, the official Danish spelling dictionary also gives information about this. In real usage quite a number of unofficial inflexional forms can be found by means of a corpus. The plural morpheme *-s*, for instance, is not a genuine Danish morpheme and is often avoided in the spelling dictionary, but Danes tend to use it nevertheless in words borrowed from other languages. A few examples: *aficionados* (Spanish) – the correct plural is *aficionadoer*; *auteurs* (French) – the correct form is *auteurer*; *bachelors* (English) – correct Danish is *bachelorer*; *divertimenti* (Italian) – correct Danish *divertimentoer*. But also in old Danish words, there can be dispute about inflexional forms: the verb *bede* ('pray, ask for') is traditionally inflected *bad* in the past tense, but young people and children tend to inflect it *bedte* as a parallel to other verbs like *lede* ('lead') and *sprede* ('spread') that have *ledte* and *spredte* in the past tense. This is accounted for in the dictionary: the 'new' past tense is given after the 'old' one, with the remark 'unofficial':

bede[3] vb.
-r, bad *el. (uofficielt)* bedte, bedt; ...

In this way, the dictionary is at the same time normative and descriptive: the user is never in doubt whether a given way of spelling or inflecting a word is official or not.

2.2.3 Meaning and construction. When it comes to meaning and construction, it is much more difficult to agree on correctness: the Danish Language Council, for instance, has no authority in this area. The policy in The Danish Dictionary is quite clear: whenever a meaning or construction is frequent, it is reported in the dictionary – even if it is not in accordance with traditional, accepted usage. The question, of course, is: what exactly does 'frequent' mean? You cannot give a mathematical answer to this, but a practical answer is that when a phenomenon is salient in a concordance and when the human lexicographer can

confirm it, either in his/her own usage or by consulting the linguistic environment (often: colleagues next door!), there are good reasons to describe it in your dictionary. A few examples from the dictionary are the words *crescendo* and *chance* ('chance', noun).

Crescendo is an international, originally Italian, musical term meaning 'a gradual increase in volume'. It is, however, quite frequent to use it in the sense of 'the most exciting, important or intense point', close to *climax*, and then the focus is on the point where the crescendo ends, so to speak. Of course, it is hard for people who are familiar with the musical sense to accept the broader sense, but in a descriptive dictionary it is natural to account for it. In The Danish Dictionary we thus have a sense number 2 with the remark that this use is disapproved of by some people (this is in fact very close to the way the issue is treated in Longman 1984 and American Heritage 1992).

crescendo adv.
1 *(mus.)* med gradvis stigende lydstyrke; **ANT** decrescendo, diminuendo ...
• *(som sb. itk.)* gradvis stigende lydstyrke ...
2 *(som sb. itk.)* højdepunkt i et forløb *(denne brug regnes af nogle for ukorrekt)*; **SYN** kulmination, klimaks ...

The noun *chance* ('chance') is traditionally followed by the preposition *for*: *have en chance for* ('have a chance to'), but in the corpus the editors found a number of instances of the preposition *til* ('to'). This usage is rather new, possibly influenced by English *a chance to*, and needs to be described in the dictionary. The construction pattern is given in the entry together with the traditional pattern, and it is pointed out that some people consider the pattern incorrect. To be true to the descriptivist approach, the citations in the entry reflect both the traditional and the new usage.

chance[1] sb. fk.
-n, -r, -rne
mulighed for et positivt resultat el. en gunstig udvikling · fx sejr, fremgang el. succes; **ANT** risiko; [en chance for at ./NGT. en chance til at+INF (konstruktionen med *til* regnes af nogle for ukorrekt)] □ *store chancer, sidste ~, gode chancer, større ~, ingen ~, en enestående ~, en rimelig ~, en reel ~, vores eneste ~; (ikke) have en ~, få chancen, gribe chancen, give en ~*□ *Landboforeningerne øjner en historisk chance for at få afskaffet jordskatten* JyP92, *Leonards eneste chance for sejr lå i en knockout* BT91
• mulighed for at nogen kan vise sine evner, el. for at noget kan lykkes □ *få chancen, give en ~* □ *Så fik jeg en chance til at vise alle, hvor dygtig og erfaren jeg er* Hjemm92
...

2.3 Statistically based collocations

Collocations are far from being an innovation in dictionaries, but the way they are selected has changed considerably. Before the age of computer-driven corpus lexicography, collocations would derive from the lexicographer's intuition and introspection combined

with existing material from other dictionaries. This was of course of great value, but when large corpora appeared, there was suddenly the possibility of finding the most frequent and typical word combinations instead of the ones the lexicographer could think of. During the period when The Danish Dictionary was being compiled, a lot of research in automatic collocation extraction was being undertaken, ranging from mutual information and T-score to word sketches (Kilgariff and Tugwell 2002). In our software the only relevant facility was mutual information, which is far from being ideal since it often favours the unusual cases and does not discard the noisy elements. Nevertheless, we found that it gave us a lot of material for presenting collocations in the dictionary. In figure 1 there is a raw mutual information list for the adjective *absolut* ('absolute') with an interval of 1 word to the right and a co-occurrence of 10 or more, in other words the list is supposed to show us the nouns that *absolut* typically modifies.

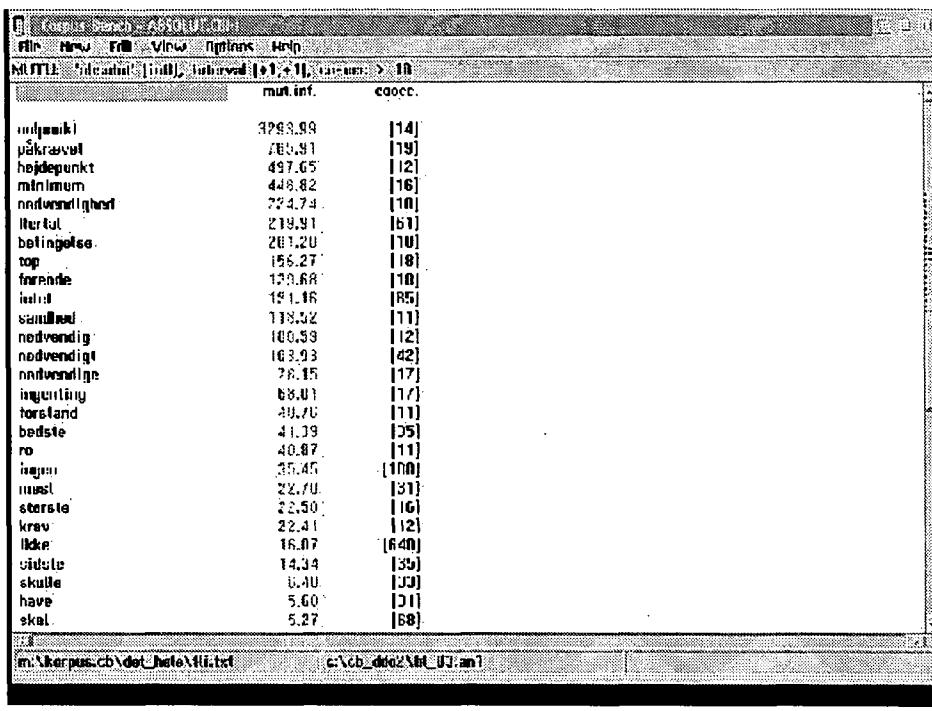

Figure 1: Mutual information for *absolut*.

The result is not too bad, but needs processing by humans: some of the collocations are fixed phrases that need an explanation such as *det absolutte nulpunkt* ('absolute zero') and *absolut flertal* ('absolute majority') (see figure 2). Some of the other nouns occurring at the top of the list are genuine collocators of *absolut* in the main sense ('without restriction') such as *højdepunkt* ('summit, climax'), *minimum* ('minimum'), and *betingelse* ('condition'). The lexicographer selected those three nouns for inclusion in the entry, but did not take

account of *nødvendighed* ('necessity'), *top* ('top'), and *ro* ('silence') which seem just as typical as the other ones. This is in fact a typical situation: the lexicographer has more material than can be included in a normal dictionary entry, and a selection has to be made. In a printed dictionary the question of space is always important, as well as the desire not to burden the user with too much information, whereas in an electronic version these questions can be dealt with differently.

Other items from the list are exploited in the dictionary entry, for instance the frequent occurrence of negative particles in the neighbourhood of *absolut* when used as an adverb ('absolutely'): *intet, ingenting* ('nothing'), *ingen* ('nobody'), *ikke* ('not') and *nej* ('no'). This is accounted for, not by mentioning the collocations, but by giving a general remark that in this sense the word is often combined with 'not' and other negatives (see figure 2).

absolut[2] adj.
-, -te
1 som er uden indskrænkning, forbehold el. konkurrence; fuldstændig, total; □ *~ højdepunkt, ~ minimum, ~ betingelse*
...
• *(som adv.)* i meget høj grad; uden vaklen el. tvivl; helt bestemt; [bruges ofte sammen med *ikke* el. anden nægtelse]
...
2 som er ubetinget og uafhængig af andre (lignende) forhold · bruges især inden for videnskab og filosofi; **ANT** relativ
...
absolut flertal *se* →flertal
absolut gehør *se* →gehør
det absolutte nulpunkt *se* →nulpunkt
...

Figure 2: The entry *absolut*[2].

2.4 Authentic examples

Examples play a very important role in dictionaries, whether they are for production or reception. A remark that is often heard from dictionary users is that it is only after reading the examples that they understand the definition or the explanation. An important issue here is whether the examples should be made by the lexicographer or taken from genuine texts (Lorentzen 2001). Again, the point of view adopted in The Danish Dictionary is descriptive: we want to describe the language as it is. Therefore, all examples in the dictionary are authentic citations from corpus texts, one part being the collocations mentioned above, another part being full sentences cited with an indication of the source. This means that on a random page in the dictionary you can find citations from such different sources as

- newspapers
- magazines
- radio and television broadcasts

- novels and short stories
- books on various subjects such as politics, literature and geography
- private diaries

This broad spectrum of examples ensures that the picture of the language that we give in the dictionary is sufficiently comprehensive to cover most of the general language used in everyday life, and as a positive side-effect: the user may want to browse the dictionary just to read the citations because it is instructive and fun. In future electronic versions of the dictionary, it might be interesting to look at the corpus formed by the citations, maybe to reveal biases in the editors' choice of examples!

3. The project – the making of the dictionary

After this presentation of some of the main features of The Danish Dictionary, I should like to turn to the project itself and to the difficulties we encountered in the process. Of course the problems that arise during 12 years are numerous, some of them banal and boring, others more intriguing; I intend however to focus on three aspects: difficulties in finding a common practice, the grouping of tasks, and finally the question of time.

3.1 Finding a common practice

The team of lexicographers behind The Danish Dictionary comprises 16 editors, 6 etymologists, and 15 student assistants. Not all of these people worked on the project at the same time, but it is obvious that when so many persons contribute to a work, there are bound to be differences in their practice. These differences may occur in different situations, among others definition style and sense discrimination.

3.1.1 Definition style. The definition style in the dictionary is not the well-known Cobuild format that appeared in the 1980's, but rather more the traditional Aristotelian genus-differentiae type or – occasionally – synonyms. The very long and detailed definitions that used to be common in large monolingual dictionaries like *The Oxford English Dictionary* or *Ordbog over det danske Sprog* (a 33-volume Danish dictionary that describes Danish from 1700 to 1950) were discarded at the beginning of the project: a more straightforward way of defining words was supposed to be adopted even if there was the risk of being imprecise and incomplete. However, when many entries had been written and they began to reach the stage of proof reading, it became clear that different editors would emphasize different features in their definitions. Let me give an example. An editor who wrote a lot of zoological entries fancied including information about the minimum and maximum size of the animals in his definitions. This type of information may be very useful, but runs the risk of becoming slightly ridiculous if the focus is always on the extreme cases (which may be interesting to a zoologist, but not to the layman). For instance, the original definition for the word *reje* ('shrimp') described the animal as a crustacean of up to 35 centimetres' length. I think we all agree that the prototypical shrimp is much smaller than that! The solution adopted by the senior editor was to describe the average shrimp and talk about 'a small crustacean' instead.

Another difficulty in defining words is the wish to foresee all imaginable possibilities, which may lead to an overuse of formulae such as *or, and so on, and the like.*

The purpose, of course, is to find an irreproachable wording that your colleagues cannot criticize, but that may be unintelligible to the average user if too many reservations are made. Syntax in definitions is another problem: it is not all lexicographers who have the gift of writing elegant definitions that mention exactly the necessary features – and nothing more – in each case, and it is not all senior editors who have the inspiration or the time to make bad definitions better. In our case, the policy at the outset was clear: complicated definitions were to be avoided, but I must say that too little attention was paid to this aspect when the heavy production began and the dreadful deadline loomed on the horizon.

3.1.2 Sense discrimination. Just how many senses does a word have? Many words can be described by most people as having one, two or three distinct senses, but when a whole team of lexicographers are confronted with a lot of corpus evidence, disagreement ensues very quickly. A well-known sense development is that of 'container-content': words denoting containers like glasses, cups and plates may also mean the content: *have a cup of coffee.* Words denoting institutions may also have the senses 'the persons who are connected with the institution' and 'the building that houses the institution': *school* is an example of this. In The Danish Dictionary these subsenses of the word are given in the entry, but what about other words denoting schools, such as *gymnasium* ('grammar school') and *kostskole* ('boarding school')? In these entries the derived senses are neglected, but that does not mean that they cannot be found in the corpus.

Another recurrent case is adjectives that indicate qualities in human beings such as *aggressive, active, ambitious* and *arrogant.* Very often such adjectives can also be used about activities, behaviour and other things connected with humans: you can play an active role, have ambitious plans or make an arrogant remark. In the case of these four adjectives the entries in the dictionary have separate sections for the subsenses, but in lots of other cases the possible subsense is either not accounted for or 'built' into the main sense by means of *or*-constructions, parentheses and the like. That is for example the case of the entries *assertiv* ('assertive') and *asocial* ('anti-social') where wordings like 'being, showing or characterized by x' are used in order to lump together humans and non-humans in the same definition. Again, a common line can be difficult to establish, and a lot of the responsibility falls on the lexicographer who is writing the entry in question, but the governing principle was corpus salience all the way, even though the concept of salience may be subject to individual interpretations.

3.2 Grouping of tasks

Making whole dictionary entries involves information types that go beyond the describing of senses, for instance morphology, pronunciation and etymology. These three tasks were taken care of in separate rounds by in-house staff or by external contributors. The advantages of this are several: the team working on the semantic description are free to concentrate on that, and the other groups can pay attention to their particular part of the entry. Consistency within fields like morphology or pronunciation is much easier to reach if the whole dictionary is processed by few people in a relatively short time. On the other hand, it is our experience that etymology is hard to separate from semantics, and it was only towards the end of the editorial process that this became clear. The fact is that the entries – and the etymologies –

had not necessarily been written in alphabetical order, but according to subject fields or specific semantic domains. When the entries passed the final proof reading, we realized that there were major differences in the way words with similar etymology had been treated, and unfortunately there was not enough time to rectify this in all cases.

In spite of this, we found that abandoning the alphabetical straitjacket and focussing on the content when writing the entries, is an advantage: the semantic description of words with similar meaning or related to the same subject appears to be far more consistent in The Danish Dictionary than in other dictionaries of Danish.

3.3 Time

It is a well-known fact that dictionary projects may go on forever if there are no limitations in time. When the plan for The Danish Dictionary was made and the funding granted, one of the conditions was that the project should be finished within 8 years. This, however, turned out to be impossible, but luckily additional funds were granted so that the dictionary could be finished within 12 years. At any rate, the dictionary has been made within a specific time-scale, the obvious advantage being that the dictionary would never have been published if there had not been a time-limit. On the other hand, one of the disavantages is that the dictionary is not as good as we wanted it to be because the constant demand for production did not always allow us to explore the semantics of a word in depth; on the contrary it often made us work too fast with many formal and stupid errors as a result, errors that it took time to correct. An important factor that appeared during the process was the arrival of other corpora. In the beginning the only corpus source was the corpus made for this particular dictionary, but gradually several other corpora made their appearance, not to speak of the Internet, that provided the editors with a lot of raw, interesting and different material. This was of course not foreseen in the project plan and funding, and it no doubt prolonged the writing process for some of the editors, the dilemma being whether to make the expected number of entries and skip interesting and even necessary information or to lag behind the production rate and include linguistic phenomena that were not present in the original corpus.

4. The future – new projects

Having discussed some of the general problems involved in the dictionary-making process, I should like to end this paper by suggesting some of the present and future work undertaken by some of the editors behind The Danish Dictionary. As mentioned in the introduction it is a printed dictionary, but of course it would be unthinkable not to have an electronic version of it. We are currently working on a prototype for an online version of the dictionary linked to corpora, meaning that the user will be able to click on words and expressions in the entries and see matching concordance lines from corpus texts. Similarly, words and expressions in the corpus should be clickable so that they can be looked up in the dictionary. In this work much of the know-how learnt during the Korpus 2000 project is of great use (Andersen et al. 2002). The plan is to develop a concept for digital dictionaries that involves searching for linguistic information in completely new ways, for instance searching for a specific content instead of a specific word form ('give me all words denoting musical instruments with strings') (Asmussen 2004).

This system will ultimately integrate all the dictionaries, corpora and texts that belong to the Society for Danish Language and Literature, thus covering a span of more than thousand years. This may sound ambitious, and luckily we managed to convince politicians and funding bodies that the Danish population would benefit from such a web-based system; thus, the Danish Ministry of Culture and the Carlsberg Foundation recently announced their support of the project. The first tasks for the staff will be the digitization of the great Danish dictionary in 33 volumes, the development of a joint web interface for the two dictionaries and the integration of corpora.

References

The American Heritage Dictionary of the English Language, Third Edition. 1992. Boston: Houghton Mifflin Company.

Andersen, M. S., Asmussen, H., Asmussen, J. 2002. 'The Project of Korpus 2000 going public' in Braasch, A. and Povlsen, C. (eds.), *Proceedings for the Tenth EURALEX International Congress*. Copenhagen. Pp. 291-299.

Asmussen, J. 2004. 'Feature Detection – a Tool for Unifying Dictionary Definitions' in Williams, G. and Vessier, S. (eds.), *Proceedings for the 11th EURALEX International Congress*. Lorient.

Asmussen, J. and Norling-Christensen, O. 1998. 'The Corpus of the Danish Dictionary' in *Lexikos 8, AFRILEX Series 8:1998*. Stellenbosch: Buro van die WAT. Pp. 223-242.

Kilgariff, A. and Tugwell, D. 2002. 'Sketching Words' in Corréard, M.-H. (ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Euralex. Pp. 125-137.

Longman Dictionary of the English Language, First Edition. 1984. Harlow: Longman.

Lorentzen, H. 2001. 'Jagten på det gode citat. Om vanskelighederne ved at finde egnede ordbogseksempler i et korpus' in Gellerstam, M. et al. (eds.), *Nordiska studier i lexikografi 5. Rapport från Konferens om Lexikografi i Norden, Göteborg 26.-29. maj 1999*. Göteborg. Pp. 202-216.

Trap-Jensen, L. 2004. 'Spoken Language in Dictionaries: Does It Really Matter?' in Williams, G. and Vessier, S. (eds.), *Proceedings for the 11th EURALEX International Congress*. Lorient.