The Project of Korpus 2000 Going Public

Mette Skovgaard Andersen
Helle Asmussen
Jørg Asmussen
The Society for Danish Language and Literature
Christians Brygge 1, 1.
1219 Copenhagen K
{msa, hea, ja}@dsl.dk

Abstract:

Among experts, corpora have become widely accepted and appreciated as an indispensable resource for lexicographic and NLP purposes. Laymen (or non-experts), however, seem to know very little about publicly available corpora and the advantages of using these in conjunction with dictionaries and as a means of linguistic inspiration. Thus in Denmark, the use of corpora till now has been limited to a small group of people with specific linguistic interests.

The Society for Danish Language and Literature, DSL, has a long tradition of creating and using corpora for lexicographic purposes for instance for the creation of The Danish Dictionary which will be published 2002-2003. The present paper discusses some of the aspects of a corpus project at DSL called Korpus 2000. The project aims at creating a relatively balanced corpus of general text from the years 1998-2002 documenting Danish around the turn of the millennium. Korpus 2000 will be made publicly available on the internet and one of the main purposes of the project is to increase laymen's awareness of the advantages of corpora. This paper focuses on aspects of designing a corpus, planning a corpus layout and presenting the project keeping this target group of non-experts in mind.

Introduction

At the 'New Trends in Reference Sciences' conference, Rundell [1996] predicted changes in the following three areas of corpus-related activities:

1. changes in the uses and users of corpora
2. changes in the types of corpora
3. changes in the statistical and analytical corpus tools

All of these areas have indeed developed enormously during the last five years: Corpora have become a natural resource for most linguists, the internet has facilitated the creation of user-defined ad hoc corpora, new corpus tools have been developed and existing ones have been improved. Still, some aspects of corpus-based activities seem to be neglected. The increasing use of computers has made corpus resources available to people outside the lexicographic and NLP communities. In Denmark, according to Danmarks Statistik an estimated 75 per cent of the population has access to the internet but until very recently the situation with respect to corpora and public knowledge of these left a lot to be desired. Various LSP corpora existed at different universities but they were not publicly available and corpora in general were hardly known outside the linguistic community. This was partly due to the size of existing corpora as for instance the indexed version of the corpus of The Danish Dictionary took up as much disk space as 1 GB which back in 1993 when the corpus was completed constituted an unmanageable size for an ordinary pc. Thus this corpus was only available on a server[1].

In 2000, The Society for Danish Language and Literature, DSL, obtained funding for the creation of the first major national and publicly available Danish corpus called Korpus 2000. The aim of the Korpus 2000 project has been to create a relatively balanced corpus of general texts from the years 1998-2002 documenting and reflecting Danish around the turn of the millennium. In the funding conditions it is explicitly stated that Korpus 2000 is to be **a linguistic resource of public interest** i.e. not only a resource for experts but also for laymen interested in language use. In the following, we describe our considerations with respect to designing and creating corpus tools and the corpus itself for such target groups. In particular, we focus on our considerations regarding the design of the corpus and the corpus interface on the internet as well as aspects of our effort to make Korpus 2000 publicly known.

User groups

As Kruyt & Dulith point out [1997], corpus users have different attitudes towards corpus design. Lexicographers need balanced corpora whereas computational linguists need large corpora. Thus before designing the corpus and the corpus interface, our main prospective target groups had to be defined. Two distinct user groups with different needs and wishes were identified: a layman user group and an expert user group. In our terms, a layman is anybody who holds just a slight interest in the functions and use of language including journalists, teachers, and students. The latter user group includes the above-mentioned lexicographers and other linguists.

Some might ask if the use of corpora is at all interesting for and relevant to laymen. Is it worth the effort to try to draw laymen's attention to corpus use and are laymen able to use a corpus in the "correct" way i.e. as a descriptive, not a normative language tool? As we have already defined laymen as one of our user groups, we obviously do believe that corpora have great potential for laymen. Whereas dictionaries often give a rather narrow definition of the potential meaning of words and a limited number of examples of the word's use in a certain context, corpora provide no definitions but a vast number of authentic examples always in context. A native speaker will thus more often find his own intuitions about language mirrored in a corpus as he is likely to find examples of a word in contexts that correspond to his own use and understanding of that word. Furthermore students' use of corpora for instance can increase their creativity and learning responsibilities as they are forced to participate in the whole process of learning i.e. asking questions, finding answers and drawing conclusions.

On the other hand, when giving laymen with no in-depth knowledge of the nature of a corpus access to such a huge collection of authentic language examples it is also crucial to educate them about what kind of language tool a descriptive corpus is. For someone who is used to using a dictionary as his main tool and guide to normative language it might be difficult to get used to the idea that what you see in a corpus is not necessarily "correct" language use in the sense that it might not correspond to the language norm. Thus it is very important to emphasise that the user cannot take for granted that corpus examples are all normative examples. Assuming that most laymen are unfamiliar with the use of corpora, one of the main tasks of Korpus 2000 has been to educate these non-experts about advantages as well as disadvantages of corpora.

Corpus Size

In an effort to combine and fulfil the above-mentioned needs of the experts we opted for a relatively large and balanced corpus of 25 million words from at least 20,000 different texts. This number was chosen for two reasons: firstly because of the limited project time and secondly because we assumed that this size would also be a workable size for laymen returning a number of query results that would not be completely overwhelming to a non-expert.

As our aim was to create a dynamic linguistic tool, we chose to construct Korpus 2000 on the basis of a so-called text bank. By contacting more than 2,000 potential text suppliers from very different areas (newspapers, publishers, ordinary people, companies etc.), we obtained a vast amount of text material - several hundreds of million words. Each text was supplied with a header primarily containing text external information about the text supplier, the author(s), publishing data etc. but also text internal information for instance about the topics of the text. All documents were saved in the text bank in a specific format inspired by the TEI guidelines [Sperberg-McQueen & Burnard 1994] and on the basis of the header information we were able to choose from the text bank the text needed to create a balanced corpus of general text.

Design, Interface and Facilities

As opposed to the expert user group, the layman user group can be expected neither to hold any prior knowledge about corpora nor to acquire this knowledge on their own. This is an important aspect to bear in mind when designing a corpus interface aimed at non-expert corpus users. Traditionally, user interfaces for corpora have not received much attention. As Johannesen et al. point out:

> "It is a rather surprising fact that while user interfaces tend to be simple and self explanatory in most areas of life represented electronically, corpus interfaces are still extremely user unfriendly." [Johannesen et al. 2000]

In our experience, interfaces of many publicly available corpora possess an overload of facilities confusing the user and reducing the immediate accessibility. Making any feature of the corpus searchable via such facilities might be a reasonable scientific wish but in our opinion this effort is wasted on any other than expert corpus users. For laymen, simplicity and fast access are much more important [Hackos & Stevens 1997]. In the Korpus 2000 project, the problem of satisfying the needs of both user groups has been solved by designing two interfaces. The expert interface consists of a search box and by means of the CQP query formalism [Schulze 1994] it is possible to search for any feature in the corpus provided that the user familiarises himself with this particular but well-documented query language. Search results are presented as KWIC concordances. Main features of the layman user interface are simplicity and familiarity. Thus it consists of a single search box very similar to the interfaces of well-known internet search engines.

Query Scenarios

In the following, we describe different query scenarios and our ways of dealing with the different problems involved. Basically, two query scenarios are possible:

- singleword queries

- multiword queries

From the point of view of the corpus designer, the main difference between these scenarios is that the number of possible different singleword queries is finite and known in advance (i.e. the number of types in the corpus). This means that word-related statistical corpus information can be computed in advance making it possible to return this kind of information to the user immediately. Multiword queries, however, are much more unpredictable: we do not know in advance, how many words are involved or whether they are immediate neighbours (n-grams) or not. This makes it rather difficult to preprocess supplementary statistical information for this query type.

From the point of view of the corpus user, one of the major problems is that corpus queries may yield too few or too many examples of the word or phrase in question - the phenomenon often referred to as the problem of scarcity and abundancy. If the visible result in both cases is a KWIC concordance, the result very likely turns out to be useless for a layman and discourage him from making further queries. In order to avoid such discouragement we decided to provide the user with alternative information which would hopefully help him obtain a more useful result.
The above mentioned two query types cause different challenges for the design of the user interface of the query system. A singleword query is more likely than a multiword query to result in an abundant number of occurrences, thus demanding certain means to shorten and structure the resulting concordance. On the other hand, multiword queries require a certain easy-to-learn query syntax. Scarcity problems may be common to both singleword and multiword queries. In the following, we look at methods of dealing with abundancy and scarcity problems in the case of singleword queries. Later, the idea behind the query syntax for multiword queries is discussed.

In Korpus 2000, the user is not presented with the classical KWIC concordance if a singleword query results in too many corpus instances (more than 100-200). Instead he is provided with a statistically generated list of typical or frequent collocates as well as some additional statistical information such as distributional reports on the relative word frequency in different types of text. From these lists he may choose to see a concordance based on a certain collocate or text type. The full concordance can be shown on demand, a feature that may be useful if the user wants to download the concordance for further processing. One of the main design principles in connection with such a search has been to enable the user to decide on his own what he wants to see and what he does not want to see.

On the other hand, if a singleword query results in only a few instances or none at all, in Korpus 2000 the user is given other supplementary information on the word searched for or at least a possible explanation of why he did not get a more copious result. In this way the user still gets some use out of his query. Other supplementary information on words may be related words such as synonyms, antonyms, semantically related words, major terms (hyperonyms), minor terms (hyponyms), compounds or derivatives. All this information is derived from The Danish Dictionary which is in its last phase and to which we have electronic access. Furthermore, it is stated whether the word searched for can be found in The Danish Dictionary or not. For the time being, however, it is not possible to show the

dictionary entry itself. This feature will be implemented when an official electronic version of The Danish Dictionary is launched in a couple of years.

The query system registers each query in a log file together with other information e.g. the number of occurrences of the word searched for. By analysing this log, we can retrieve information on how many users we have, what they search for and how often and the results they obtain. Searches for words that do not occur in the corpus are particularly interesting. By analysing the log file, we discovered that these words fall into three characteristic categories:

1. The word is seldom used or used mainly within LSP, whereas Korpus 2000 represents LGP
2. The word is too new to be in the corpus
3. The word is misspelled

We use these observations to guide the user by telling him that he might be searching for an infrequently used word, which had better be looked up in a dictionary. Or that the word might be too new - if possible, we provide him with a list of new words by comparing our newest text material with what is in the corpus. And in case the word is misspelled, we try to inform him of alternative spellings based on typical phonetic and orthographic misunderstandings. These alternative spellings can all be found in the corpus. We believe that this kind of guidance is considerably more useful for the user than a simple "no matches found" message, which might lead the user to dismiss the corpus as useless all together.

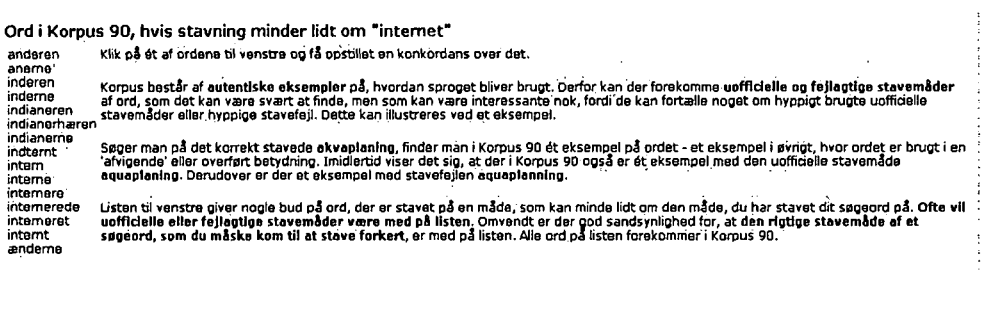

Figure 1: Example of alternative spellings of the query word *internet*.

A special case of a singleword query is a query containing wildcards, which may be interesting to use for experts as well as non-experts. Available wildcards are ? (exactly one character) and * (zero, one or more characters). Instead of generating an unmanageable KWIC concordance containing all matches, the user is presented with a list of matching words and their number of occurrences in the corpus. From this list he can pick the words he wants to see in a KWIC format. It is, however, also an option is to see all matches in one concordance.

As mentioned above, multiword queries require a query syntax that is easy to learn for our users. One way of facilitating this kind of queries is to employ a graphical user interface, where one can specify a couple of words, their possible positions relative to a fixed keyword, and a couple of attributes for each word such as part of speech or inflection. Though such interfaces may be very dynamic and make almost any kind of query possible they easily

grow very complex making it difficult to use for simple queries. Our way of solving this problem is to use the same interface, merely just a query box, for all types of searches. The difference is that if the user types in several words instead of just one, the system will ask the user whether the query has to be interpreted

1. literally, e.g. as a fixed string of words in exactly that order
2. as words contained in the same orthographic sentence in any order
3. as a syntactic pattern

The first two cases are quite trivial and are well known from search engines on the web. The third case is somewhat more unusual and will be explained in more detail. Assume that the user types in the following phrase

han taler om sin fremtid
(he speaks about his future)

If the user wants the query to be treated as a syntactic pattern he will be presented with a checkbox containing the query words. In the box he can check the words which should be treated as constants in the pattern. The rest of the words will be treated as variables: in their place any word with the same part of speech may occur. Assuming that the user wants to have the words *taler* and *om* treated as constants, then the above sentence will be treated as
[pronoun] taler om [pronoun] [noun]

and result in examples like

han taler om hendes forældre
(he talks about her parents)

de taler om deres arbejde
(they talk about their work)

hun taler om sin fremtid
(she talks about her future)

The strength of syntactic patterns can be augmented by the use of the wildcards ? and * where ? means exactly one arbitrary word in that position, and * means zero, one or any number of arbitrary words in that position (interpreted in a non-greedy manner, corresponding to the sequence .*? in regular expressions). An example of a syntax pattern query with wildcards is

*han taler ? om * fremtid*
*(he talks ? about * future)*
interpreted as
[pronoun] taler [exactly one word] om [zero, one or many words] [noun]

Matching examples for this query are

han taler ikke om deres fremtid

(he does not talk about their future)

han taler gerne om sin lysende fremtid
(he likes to talk about his bright future)
han taler aldrig om andet end deres fælles fremtid
(he never talks about anything but their common future)

Another intricate query problem from the point of view of corpus designers has to be mentioned here: part of speech and inflection. Should query words be treated literally or as instances of certain lemmas? If the user types in *speaks*, does he then mean, literally, *speaks*, or any form of the lemma *to speak*, e.g. *speaking*, *spoke*, *spoken*? In singleword queries we assume that the user means literally what he has typed in, thus *taler* gives a concordance with occurrences of *taler* as immediate result. The resulting concordance contains, though, an inflectional scheme of the typed in word, with all its possible forms and pos's - by the way, the lemma *tale* both is the noun *a speech* and the verb *to speak*. The user can then click on one of these alternative forms to get another concordance. In the case of syntactic patterns the fixed, constant words are treated in the same manner.

A future augmentation of the user interface could be an implementation of facilities dealing with syntactic functions, thus making queries like

[subject] taler om [prepositional object]

possible as the corpus already is marked-up with this kind of syntactic information. The morpho-syntactic mark-up of the corpus has been made with an constraint-grammar based tagger at the VISL project at Syddansk Universitet [http://visl.hum.sdu.dk/visl/].

Project Homepage

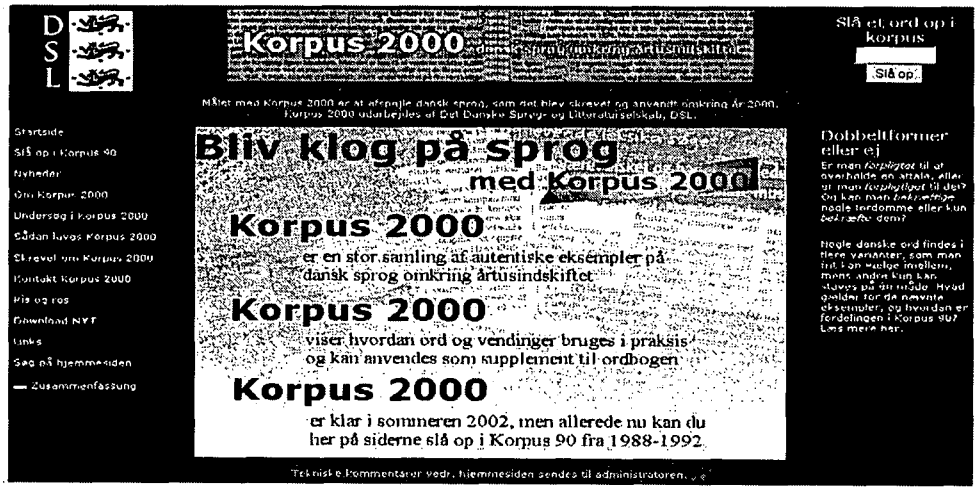

Figure 2: The Korpus 2000 homepage

We have attempted to design the corpus query interface on the basis of simplicity and good layout. As the query interface is part of the project homepage it has been our goal to transfer these principles to the homepage as well. Popularisation is a major component of the homepage. The advantages of corpus use are exemplified through pre-generated queries and the project itself is explained in a popular manner emphasising how corpora can be used by everybody for inspiration and in conjunction with dictionaries. This might not be of great interest to the expert user group but in order to appeal to laymen and educate them about the virtues of corpora we feel that popularising the information is a necessary means to cater for the needs of these non-experts.

The homepage contains a variety of examples of what is searchable in a corpus. Some of our searches exemplify how a corpus can be put to very practical use when solving crossword puzzles because of the wildcard facility. Others are frequency lists of somehow related words like weekdays, months, kinship terms etc. Some of these lists are inspired by the lists made by Leech et al. [2001], others by holiday seasons etc. At Christmas time, a list was displayed showing the most frequent Christmas compounds to mention just one example. This kind of corpus use may be very trivial to experts but we feel it necessary to provide our non-expert users with such simple examples of corpus use to inspire them to make their own - maybe more complex - searches.

Making Korpus 2000 Known to the Public

Having no corpus tradition in Denmark, a major task for us has been to inform potential users about the Korpus 2000 project. Our homepage has played an important part but we have also felt it necessary to take further steps to spread the word of Korpus 2000.

1. At the very beginning of the project, we contacted most Danish universities informing them about the project and urging them to contribute with ideas, wishes or subcorpora they might possess. The result was, however, disappointing. Only few of the universities responded and none of these made their own corpora available to the project.
2. We have made a special effort to encourage so-called ordinary people to contribute with private texts. This step was facilitated by making use of The Danish Dictionary's corps of 'word watchers' (at DSL called 'spORDhunde'). As a result of contacting these, we obtained a relatively large amount of non-professional writings.
3. Several newspapers and magazines have been contacted, seven of which (March 2002) have printed or written articles or announcements about the project.
4. At the end of the project, we plan to contact Danish schools and high schools in order to promote the use of corpora in language teaching. A future augmentation in this respect could be to make corpus courses for teachers as we are convinced that the use of corpora has great potential in the Danish school system
5. At the end of the project, press announcements about Korpus 2000 will be made.

At the moment, it is too early to evaluate the results of our efforts to make Korpus 2000 publicly known. We do, however, keep records of particular user data and are thus able to evaluate user behaviour as the project develops. Month by month the number of queries has grown by app. 50 percent and our hope is that Danes eventually will use corpora to the same extent that they use dictionaries. For a small language like Danish, the availability and accessibility of corpora are of great importance and will hopefully result in more competent and conscious language users. If so, Korpus 2000 has more than fulfilled its purpose.

Endnotes

1) Today, the space required is found on most pc's which is why it was decided to make the corpus of The Danish Dictionary available for download from our homepage along with a corpus query tool developed at the Society for Danish Language and Literature. The current version of the corpus takes up app. 750 MB of disk space.

References

Danmarks Statistik: http://www.dst.dk

[Hackos & Stevens 1997] Hackos, J.T. & Stevens, D.M., 1997. *Standards for Online Communication: Publishing Information for The Internet/WorldWide Web/Help Systems/Corporate Intranets*. John Wiley & Sons, Inc., New York.

[Johannesen et al. 2000] Johannesen, J., Nøklestad, A. & Hagen, K., 2000. A Web-Based Advanced and User Friendly System; The Oslo Corpus of Tagged Norwegian Texts. In *Proceeding, Second International Conference on Language Resources and Evaluation.* Athen. 1725-1729.

[Kruyt & Dulith 1997] Kruyt, J.G. &. Dulith, M.W.F, 1997. A 38 Million Words Dutch Text Corpus and its Users. In *Lexikos 7.* 229-244.

[Leech et al. 2001] Leech G., Rayson P. & Wilson A., 2001. Word Frequencies in Written and Spoken English based on the British National Corpus. Pearson Education Limited, Harlow.

[Rundell 1996] Rundell M., 1996. *The Corpus of the Future and the Future of the Corpus*. Talk at Exeter, special conference on 'New Trends in Reference Science' (29/3/96). [http://www.ruf.rice.edu/~barlow/futcrp.html]

[Schulze 1994] Schulze, B.M., 1994: *Entwurf und Implementierung eines Anfragesystems für Textcorpora.* Diplomarbeit, IMS, Universität Stuttgart.

[Sperberg-McQueen & Burnard 1994] Sperberg-McQueen, C.M. & Burnard, L. (eds.), 1994. *Guidelines for Electronic Text Encoding and Interchange*. TEI P3 Text Encoding Initiative. Chicago, Oxford.

VISL: http://visl.hum.sdu.dk/visl/