The Corpus of the Danish Dictionary

Ole Norling-Christensen and Jørg Asmussen,
The Society for Danish Language and Literature,
Copenhagen, Denmark

Abstract: A Danish corpus, holding 40 million words of general language from the period 1983-92, was designed and compiled by DSL (The Society for Danish Language and Literature) in order to serve as a major source for a new six volume dictionary of contemporary Danish. The corpus includes written and spoken, private and professional, general and specialised language, and each of the 44 000 text samples is annotated with formalized information on these and other features of linguistic and sociological importance. The resulting multidimensional text type specification is useful for the extraction of (virtual or real) subcorpora and for statistical analyses. Specialized software has been developed for flexible interactive concordancing and analysis. The corpus is currently only accessible at the site of DSL; nevertheless, several scholars and students have been using it in their research. The experience gained by the staff of DSL is being reused in cooperative language engineering projects within the European Union, and in 1998 a publicly available corpus will be released as an outcome of the PAROLE project.

Keywords: CONCORDANCE, COPYRIGHT, CORPUS, DANISH, DICTIONARY, FREQUENCY, LANGUAGE ENGINEERING, MUTUAL INFORMATION, SGML, STATISTICS, SUBCORPUS, T-SCORE, TEXT TYPOLOGY, WORD DISTRIBUTION

Opsomming: Die korpus van die Deense Woordeboek. 'n Deense korpus wat 40 miljoen woorde uit die algemene taal van die periode 1983-92 bevat, is ontwerp en saamgestel deur die DSL (The Society for Danish Language and Literature) om te dien as 'n primêre bron vir die saamstel van 'n nuwe ses-volume woordeboek van hedendaagse Deens. Die korpus sluit geskrewe en gesproke, private en amptelike, algemene en gespesialiseerde taal in, en elk van die 44 000 teksvoorbeelde word voorsien van formele inligting oor hierdie en ander kenmerke van taalkundige en sosiologiese belang. Die geskepte multidimensionele tekstipe spesifikasie is nuttig vir die onttrekking van (virtuele of ware) subkorpora en vir statistiese ontledings. Gespesialiseerde programmatuur is ontwikkel vir veeldoelige interaktiewe konkordansiebou en ontleding. Alhoewel die korpus tans slegs toeganklik is by DSL, het verskeie leerlinge en studente dit al gebruik in hulle navorsing. Die ervaring wat opgedoen is deur die personeel van DSL word hergebruik in koöperatiewe taalmanipulasieprojekte binne die Europese Unie, en in 1998 sal 'n korpus wat beskikbaar sal wees aan die publiek, vrygestel word as 'n uitvloeisel van die PAROLE-projek.

Sleutelwoorde: KONKORDANSIE, KOPIEREG, KORPUS, DEENS, WOORDEBOEK, FREKWENSIE, TAALMANIPULERING, WEDERSYDSE INLIGTING, SGML, STATISTIEK, SUBKORPUS, T-TELLING, TEKSTIPOLOGIE, WOORDVERSPREIDING

Lexikos 8 (AFRILEX-reeks/series 8: 1998): 223-242

1. The Danish Dictionary

The DDO Corpus was built during the period 1991-93 in order to serve as a primary source for *The Danish Dictionary* (Den Danske Ordbog, DDO), a new dictionary of contemporary Danish being edited by The Society for Danish Language and Literature (Det Danske Sprog- og Litteraturselskab, DSL). This Society, which is a kind of academy, was founded in 1911 with the aim of providing scholarly editions of Danish works of linguistic or literary importance, as well as dictionaries of the Danish language. Legally it is a semipublic institution under the jurisdiction of the Danish Ministry of Culture, and its activities are financed in part by the Danish Government and in part by the Carlsberg Foundation[2] and various other public and private foundations.

DSL edited the 28 volume *Ordbog over det Danske Sprog*, which was published 1918-56. It is the authoritative dictionary of newer Danish (i.e. from after c. 1700). DSL is currently in the process of editing five supplementary volumes which extend the coverage of all the volumes to 1955. A dictionary of Old Danish (1100-1510) is also in progress, and among the recent text editions of the Society is *Dansk Nationallitterært Arkiv* (Archive of Danish National Literature) on CD-ROM (1992). During 1995-98 DSL took part in the European Union language engineering oriented project *PAROLE* (MLAP63-386/LRE-63368), the aim of which is the production of comparable, harmonized corpora and lexica for the languages of the Union.

The history of the Danish Dictionary project dates back to 1989, when the plans for changing the European Community into an Economic and Monetary Union were launched. A large minority of the Danish people was, and still is, sceptical of the Union. Among other things it is feared that Danish culture and language will slowly but surely disappear in the new Europe. In order to allay this fear, several initiatives were taken by the Government, and a think-tank set up by the prime minister advocated the idea of creating a Danish national encyclopaedia and a dictionary of modern Danish. Both projects were launched in 1991 with the support of private foundations and the Government. The dictionary work was entrusted to DSL, which had submitted a plan and a budget for it by the end of 1989. The funding is shared equally by the Government and the Carlsberg Foundation. An electronic manuscript ready for printing will be delivered to the publishing house Gyldendal in the course of 2002, and the six volumes will be published in 2002–03. The royalties are earmarked for future lexicographical work.

The dictionary will contain approximately 100 000 entries and provide information on spelling, word-class, inflection, valency, pronunciation, meaning, phraseology and etymology. Entries are supplemented with original quotations, illustrating the different usages. It aims to fulfil the needs of both professional and general users of Danish, whether native speakers or advanced learners. The dictionary is basically descriptive, but the description includes information on acceptability, i.e. the norm. In other words: it shows the language as it is, not as it should be, but at the same time it also guides the user. There was

therefore no doubt in the minds of the chief editors, Ebba Hjorth, Kjeld Kristensen and Ole Norling-Christensen, that the work should be largely corpus-based. Foreign experience in the field was eagerly studied, especially the English dictionary project *Collins Cobuild*, the implications of which gave much inspiration to the first phase of the work, the building of a corpus. Thanks to the authors, papers like Atkins et al. (1992) and Church et al. (1991) were available to the editors in manuscript during this period.

Some domestic experience was also available, including the theoretical considerations of the makers of the first Danish corpus *DANwORD*, 1,25 million words for frequency studies of five distinct text types from the period 1970-74 (Maegaard and Ruus 1987). Thanks to funding from the Danish Research Council for the Humanities, a few more corpora had been created around the end of the 1980s: a collection of Danish, English and French texts in the field of contract law, and a collection of Danish, Spanish and German texts about genetic engineering, each holding c. one million words for each language. The latter was of special interest to the dictionary project, as some of the texts were not technical language (LSP), but written by or for laymen. Furthermore, Prof. Henning Bergenholtz of the Aarhus Business School collected one million words of general language (newspapers, magazines and novels) for each of the years 1987-90. This corpus, *DK87-90*, is the reference corpus most widely distributed among researchers of Danish.

2. Design of the corpus

It is important to underline that DDO is a *dictionary* project having a fixed budget of around six million ECU and a fixed time frame of twelve years. The corpus was thus not an end in itself, but was primarily established in preparation for the dictionary, even though some thought was also given to other future needs. Consequently, time and costs had to be among the premises for many of the decisions made during the planning and compiling of the corpus, including the decision of limiting the corpus period to ten years with some overrepresentation of the most recent three years.

2.1 Size and structure

The corpus consists of samples of written and spoken Danish produced during the decade 1983-92. The samples were collected, standardized and annotated by the staff of the Danish Dictionary, with the assistance of several students and external typists.

The following three aspects were taken into consideration during the initial design of the corpus: how many running words should be included, what period should be covered, and what types of text should be included.

In view of the Cobuild experience, it was decided that the corpus should consist of 40 million running words and should cover the Danish general lan-

guage as comprehensively as possible. Setting the number of running words to be included was not a main criterion, as this number naturally depends on other important considerations, such as the breadth, variety and balance of the coverage. Even though the dictionary is meant to describe contemporary Danish from the 1950s until today, texts from before 1983 were not included in the corpus. The decade from 1983 onwards was mainly selected because most machine-readable texts available are from this period, and it was estimated to be too costly and time-consuming to extend the coverage with scanned and/or typed text dating back to the 1950s. Furthermore, supplementary sources would be available to cover the language from 1955 up to the start of the corpus. They include just under one million slips with excerpts made by the Board of the Danish Language (Dansk Sprognævn) since 1955, two newly updated comprehensive bilingual dictionaries (Danish-English Dict. 1990) and (Danish-French Dict. 1991), and a special dictionary of New Words in Danish (Riber Petersen 1984). For the time after 1992 no systematic investigations are made. However, observations made by the staff, as well as slips submitted by the *spORDhunde*, a group of c. 300 voluntary "word watchers" who collect original material for the project, are continuously considered for inclusion.

The decade 1983-92 was designated as the Dictionary's primary period, meaning i.a. that the quotations used to supplement the dictionary definitions are chosen mainly from this period. Furthermore, it was accepted that the later part of the primary period would receive special emphasis because the supplementary sources would partially cover the earlier part of the decade. However, the corpus is balanced in this respect to allow for diachronic studies. As can be seen from figure 1, subcorpora of up to 16 million words, equally distributed over the years in question, may be selected from the main corpus by taking up to 1,6 million words from each of the years 1983-92[3].

The aim for the broadest possible coverage meant that the corpus was designed to comprise of general and specialized language, written and spoken language, "public" and "private" language (technically a distinction is made between *reception* and *production*), "young" and "adult" language, as well as a variety of different media, genres and subject areas. Two kinds of text were intentionally excluded viz. translated text, which will notoriously be biased by the source language, and technical language, i.e. language produced by specialists for other specialists in the same field, which is outside the scope of a dictionary of general language. In this context, *specialized language* (which is included) therefore means nonfictional written (or spoken) language for non-specialists, for instance textbooks or magazines on specific topics. Only a single intensional exception was made to the exclusion of translated text: parts of a new translation of the Bible were included. However, even though news-agency stories and subtitles of foreign films and telecasts were avoided, the origin of, for instance, newspaper stories cannot always be known. Finally, in order to cover as much different text as possible, entire novels, textbooks etc. were not included, but only one or a few randomly selected chapters up to a maxi-

mum of 10 000 words from each.

Year	No. of samples	Pct.	No. of words	Pct.
1983	2 199	5,0	1 601 379	4,0
1984	2 069	4,7	1 978 855	4,9
1985	2 291	5,2	2 295 799	5,7
1986	2 234	5,1	2 812 292	7,0
1987	1 809	4,1	3 639 409	9,1
1988	1 442	3,3	2 918 484	7,3
1989	1 821	4,2	2 798 556	7,0
1990	7 160	16,3	5 734 530	14,3
1991	14 155	32,3	8 688 920	21,7
1992	8 611	19,7	7 309 353	18,2
1993	15	0,0	329 765	0,8
Total	43 806	99,9	40 107 342	100,0

Figure 1: Number of text samples and running words by the year they were produced/published

As it is difficult to use objective criteria to establish what makes a balanced corpus, a more common-sense approach was adopted. Three dichotomies were selected (written vs. spoken, reception vs. production, general vs. specialized), and on the basis of these the corpus was divided into eight distinct classes. For each class, the possible text sources were reviewed and a preliminary word-number target was set. In some cases, this was done very informally, such as for spoken language, where the target was "as much as possible, up to a maximum of 10 million words". The collection of text samples was thus an iterative process: after a part of the corpus had been collected, statistical information was used to investigate which classes were still underrepresented and the selectional criteria were adjusted accordingly. The statistics were calculated on the basis of the information contained in the annotations (the headers, see below) of each text sample.

2.2 Selection of the text samples

The main sources for data acquisition were (a) books, magazines and news-

papers (28 million running words), (b) radio and television broadcasts (3,8 million running words), and (c) leaflets, booklets, pamphlets etc. (2 million running words). Furthermore, the relevant parts of existing Danish corpora were included, viz. the 4 million words of *DK87-90* and those parts of the corpus of genetic engineering which were not technical language. Several publishers, as well as Danmarks Radio (the National Broadcasting Company), were extremely helpful in supplying us with machine-readable text, the biggest donation being three volumes of three (very different) newspapers from the newspaper publisher Berlingske, a total of c. 75 million words distributed over more than 200 000 separate pieces of text.

It should be noted that only a relatively small part of this newspaper text was included in the corpus. However, the large number of separate articles etc. was most useful for the final balancing and annotating of the corpus, as the text had been downloaded from an information retrieval system which also contained some information on the individual articles. Even though this information was rather informal and inconsistent, parts of it could be transformed by a computational analysis into the standard categories for genre and topic, after which a balancing selection could be made. The information on authors (mostly journalists) was collected in a database which meant that information on year of birth, sex, etc., only had to be looked up once for each language user. Moreover, the database counted the number of newspaper articles by each author, which helped to avoid overrepresentation of the most productive journalists.

One of the explicit aims of the Danish Dictionary is to account for the use of spoken as well as written language. However, while the Dictionary aims to *cover* written Danish, it settles for only *considering* spoken Danish. The reason for this is twofold: it is theoretically difficult to define and represent spoken language usage in a corpus, and it is not economically feasible to collect and transcribe a large body of spoken language samples. Special emphasis was still put on the inclusion of spoken language, and the corpus does in fact contain 7 million words from private interviews, political debates, radio and television broadcasts etc., which represent 17 pct. of the total corpus. Again, great willingness to help was encountered: transcribed sociolinguistic material and interviews made for sociological research were given by colleagues at universities, and the unedited transcriptions of several animated debates with improvised contributions from a large number of members were received from the parliament and the city council of Copenhagen.

Another explicit aim of the Danish Dictionary is to describe the Danish language as it is used "privately" by the majority of the population *(production)*, instead of concentrating solely on "public" language users, such as journalists, authors, and politicians *(reception)*. Great emphasis was therefore placed on incorporating such material as private letters, letters to the editor, diaries, and school essays, which represent a total of 11 pct. of the corpus.

3. Building of the corpus

During the early period of the dictionary project (September 1991 – December 1993) the text samples were scanned, typed in, or, if already in a machine-readable format, converted from various kinds of wordprocessing or typesetting formats. Information on author(s), text type etc. was attached manually or, to some extent, automatically to the respective text samples.

SGML, the international standard for generic description of textual structures and marking up texts, was used for annotating the corpus. An SGML document type called *CorpusEntry* was defined. It provides the means for registering extralinguistic information about the text and for unambiguously tagging some (socio)linguistic features of it. Each of the 43 806 text samples of the corpus is one CorpusEntry element which consists of a header followed by the text proper. In the language of an SGML document type definition this is formally expressed as:

```
<!DOCTYPE    CorpusEntry [
<!ELEMENT    CorpusEntry (Header, Text)>
-- followed by declarations of the Header and Text elements -- ]>
```

3.1 Coding of the header

The header is structured by means of SGML tags as shown in figure 2. It is made up of a number of fields which have been filled in with formalized information (attribute/value pairs) about the respective text samples during the compilation of the corpus. The fields typically specify the authors' age, sex and language variant (standard or regional), as well as medium, genre and subject area (topic) of the text. Some of the fields are of special importance in that only a value from a finite set can be assigned to them; they are marked by bars (‖) in the figure. These fields are used for corpus statistics, and they permit the use of special "filters" for creating virtual or real subcorpora according to a multidimensional text type specification; these can in turn be accessed separately or compared statistically, thus making the concept of "a balanced corpus" more flexible.

TextInfo	
TextID	Unambiguous identifier of the text sample — for citation purposes
Restrictions	
Anonymity	Proper names must be altered (A), or not (-), if cited
DD_Only	Text must only be used by The Danish Dictionary
TextTitle	Title of the text
VolTitle	Name of anthology, newspaper, magazine etc.
Publisher	Publishing house, broadcaster etc.

	PublTime	
	Day	{1, 2, ..., 30, 31}
	Month	{1, 2, ..., 11, 12}
‖	**Year**	{1983, 1984, ..., 1992, 1993}
	Certainty	The year of publishing is known exactly (-), or not (?)
	Location	E.g. book volume, newspaper section, page number
‖	**LangType**	{general, specialized}
‖	**Expression**	{written, spoken, and two intermediate types}
‖	**Aspect**	{reception, production}
‖	**AgeRelation**	{adult–adult, adult–juvenile, adult–child, ..., child–child}
‖	**Medium**	{book, journal, radio, diary, ...} — 13 possible values
	Genre	{novel, interview, essay, ...} — 131 possible values
‖	**GenreType**	A reduced classification for statistical use — 17 values
‖	**Topic**	{philosophy, geography, physics, ...} — 66 possible values
‖	**TopicType**	A reduced classification for statistical use — 12 values
	Group	Unambiguous identifier of a group of related text samples
	Number	Serial number within the text group
	Size	Number of tokens in the following text sample
	UserInfo+	(one or more language users: author(s)/speaker(s))
	UserID	Identifier referred to by speaker turns in the text
	Surname	Surname of the language user
	FirstName	First name of the language user
‖	**Sex**	{male, female, unknown}
‖	**Born**	{1880, 1881, ..., 1989, 1990}
	Certainty	The year of birth is known exactly (-), or not (?)
	BirthPl	Place of birth
	Residence	Place of residence
‖	**Region**	Dialectal region — 11 values
	Education	Education of the language user
	Occupation	Occupation of the language user
‖	**LangVar**	Language variant {standard, regional}
	Role	Communicative role of the language user, e.g. teacher, pupil

Figure 2: Structure of the header information which accompanies each text sample

3.2 Coding of the text

An important consideration when designing a corpus is how the printed and spoken text should be represented computationally. As a matter of course one specific character set must be used. Because work is done in a PC environment (operating system: OS/2), Code Page 850[4] was chosen. However, this is only the first decision to be made. One uniform and consistent annotation system is also needed. This must be suitable for future computational searches and ana-

lyses, and information of importance for these uses must be recorded. On the other hand, it may not prove feasible to spend resources (human as well as computational) on recording information which is regarded to be of less or no importance. Defining such a format is no trivial task. It implies a series of decisions on *which features* of the text one wants to depict in the corpus. Should there, for instance, be specific codes for the smell of the paper? — Probably not. The colour of the paper? — It might have some special meaning. The size of the letters? — Differences in size are likely to signify differences in text type, but the meaning of such differences will differ from one text to another. An obvious conclusion from these kinds of question is that the coding has to be generic and not just mirror how the printer chose to represent the different kinds of text: *business pages*, not pink paper; *headline*, not big bold type; *highlighted*, not italics, bold or small caps.

For the Danish corpus a very restricted set of textual features has been chosen to be marked up. The structure of the element Text depends on whether it consists of written language or of (transcribed) spoken language. Written language is divided up into paragraphs (the element *p*) which in turn are mostly nontagged strings of characters (the SGML category #PCDATA); these may, however, be interspersed with elements of special categories of text, like highlighted text or notes.

For spoken language the first level of subdivision normally is not paragraphs, but speaker turns. Most of the spoken text samples are conversations or interviews with more persons involved. Consequently, the header may contain two or more instances of the element *UserInfo*. Each of these contains a different three letter string in the subelement *UserID*, and each element *speaker_turn* contains an attribute *id* which refers to the *UserID*. The *speaker_turn* element consists of #PCDATA interspersed with entity references[5] like {hesitation} representing nonverbal sounds like "eh", "mmm"; {pause}; {uf} representing a passage that was incomprehensible to the transcriber; {laughter}; and with the elements *comment* (the transcriber's "stage directions" that are not part of the speech), and *uncertain* (a word or passage that the transcriber was not sure about). The full set of SGML tags used is defined and explained in figure 3.

```
<!ELEMENT Tekst   ( ekst.tekst ¦ ill.tekst ¦ lyd ¦ p ¦ regi ¦ replik ¦ skrift ¦ tanke ¦
                    vers )+ +(kommentar)>
   -- The tag for the Text element --
<!ELEMENT ekst.tekst   ( ill.tekst ¦ p )+ >
   -- external text: part of text which was typographically placed in e.g. margins or
      boxes and thus not part of the running text --
<!ELEMENT f   ( #PCDATA ) >
   -- highlighted: enhanced part of the running text (e.g. italics, bold, or small caps in
      the printed original) --
<!ELEMENT ill.tekst   ( p ¦ vers )+ >
   -- caption: underline of illustration, table etc.; text inside an illustration --
<!ELEMENT kommentar   ( ( #PCDATA ¦ usikker)+ ¦ ( p ¦ vers )+ ) -(kommentar) >
```

```
    -- comment: transcriber's or editor's comment; not part of the text --
<!ELEMENT lyd   ( p+ ) >
    -- sound: in comic strips etc: rendering of sounds like Riiinnggg, Bam! --
<!ELEMENT note   ( #PCDATA | f )+ >
    -- foot- and end-notes; this element which contains the text of the note, is placed at
       the point of the text where the note-reference of the original was written --
<!ELEMENT p   ( #PCDATA | f | note | usikker )+ >
    -- paragraph: One or more empty lines between paragraphs in the original are
       represented by one instance of the newline-entity {NL} --
<!ELEMENT regi   ( p | skrift )+ >
    -- stage direction, in comics: the narrative text, like "Copenhagen, Townhall Square.
       An evening in September. It is six o'clock" --
<!ELEMENT replik   ( p | vers )+ >
    -- speaker turn in spoken language; speech balloon in comics --
<!ELEMENT skrift   ( p+ ) >
    -- writing, in comics etc.: text in the picture which is neither sound, stage direction,
       speaker turn, nor thought, but for e.g. posters, notes, letters, documents, graf-
       fiti --
<!ELEMENT tanke   ( p+ ) >
    -- thought, in comics etc.: the thoughts of the characters, rendered in thought
       balloons --
<!ELEMENT usikker   ( #PCDATA ) >
    -- uncertain: part of spoken language which was not identified with certainty by the
       transcriber --
<!ELEMENT vers   ( p+ ) >
    -- verse: metrical text. Each stanza under this element is marked up as a paragraph;
       its verses are separated by slashes --
```

Figure 3: Structure of the text element as defined in the DTD

4. The lexical database

In parallel to the corpus building, methods for reuse of existing lexical sources were developed, and a database of 340 000 words (i.e. lemmas) was extracted/constructed from the machine-readable versions of some standard printed dictionaries, *viz.* the official spelling dictionary (Retskrivningsordbogen 1986), Danish-English Dict. (1990), Danish-French Dict. (1991), supplemented with word-lists from the Board of the Danish Language. The database holds formalized morphological information, as well as unformalized (except for subject field) semantic and contextual information extracted from the source dictionaries. Using the inflectional information given in the database, all possible inflected forms of the lemmas were generated and compared to the stock of word forms that are present in the corpus. The remaining forms, which were not identified during this run, have been further investigated and gradually added to the database as new words or as unofficial spelling variants of exist-

ing ones. The selection of lemmas for the Danish Dictionary, as well as a tentative assignment of dictionary entry size, was made by the help of the information kept in the lexical database, including the word frequencies found in the corpus and the relative size of entries in the printed dictionaries.

5. Using header information for making a subcorpus

As a simple example of the use of the feature/value pairs of the headers for the design and extraction of subcorpora, as well as for the evaluation and further balancing of the resulting subcorpus, brief consideration will be given to the case of a Danish research institute, active in the field of machine translation, which needed a specialized corpus of text covering a range of 10 different, but somehow related, subject fields. Summing up the numbers of text samples and running words of the entire corpus for the specified values of the text type feature *topic* rendered the result shown in figure 4, which is at the same time the composition of the largest possible subcorpus that will fit the demand.

Topic code and name		No. of samples	Pct.	No. of words	Pct.
191	science (general)	76	1,6	115 407	3,4
195	communication	178	3,7	108 491	3,2
30	society	1 122	23,1	925 329	27,4
331	business	1 428	29,4	815 653	24,1
38	social services	635	13,1	453 074	13,4
50	natural sciences (in general)	158	3,3	127 143	3,8
60	technology	144	3,0	147 253	4,4
627	transportation	731	15,1	389 582	11,5
66	industry	147	3,0	68 052	2,0
68	crafts	234	4,8	230 535	6,8
Total		**4 853**	100	**3 380 519**	100

Figure 4: The number of samples and running words (tokens) for 10 of the 66 topics (subject fields) which are distinguished in the corpus

The composition of this subcorpus according to other text type features can now be investigated and compared to that of the general (reference) corpus and

be used for further balancing, if needed. Figure 5 is just one short example to serve as demonstration.

	Entire corpus		Selected subcorpus	
Authors' sex	No. of words	Pct.	No. of words	Pct.
unknown	14 539 129	36,3	1 164 447	34,4
female	7 648 688	19,1	532 964	15,8
male	17 919 525	44,7	1 683 108	49,8

Figure 5: Composition of the subcorpus summed up by the text type feature *sex of the author*, compared to that of the entire corpus

It can be seen from the table that an author has been identified for a greater proportion of the selected text than for the source corpus, and that the over-representation of male authors is even more marked. If another balance is desired, the user must discard some text samples, thus making a smaller, but more balanced subcorpus.

6. Exploitation of the corpus

There are two versions of the corpus, a master copy, which is a collection of SGML-coded text files, and a compiled and indexed version which is available on-line; the latter version is used every day by the editorial staff for making concordances and statistical analyses as part of the work on the dictionary. The master copy is used for special examinations which cannot be made by the interactive tool. It is continually refined, and at intervals a new compiled version is made from it. The refinement of the corpus includes correction of (technical) errors, the disambiguation of certain characters, and the making of some additional annotations. Among the errors that were corrected, were multiple instances of the same text sample, wrong dating (the machine-readable version supplied by a publisher proved to be a later version than the known printed book) and a few data conversion errors.

As to disambiguation, a clear-cut definition of which characters are part of a word and which are not, is necessary for simple and efficient computational processing of text. Apostrophes may be part of (contracted) words, but they are also frequently used as quotation marks; a hyphen is part of a word, whereas a dash is a punctuation mark, but quite often the same character is used for both. These ambiguities were resolved automatically with a high degree of certainty.

6.1 Two problems: abundance and scarcity

The lexicographer working with corpora runs into two basic problems: the theoretical problem of the significance of infrequent or missing occurrences of

some linguistic phenomena, and the practical problem of being flooded with too many instances of others. The former problem can only be solved by making the corpus even larger, or by relying on sources external to the corpus. To cope with the latter, computational tools are needed in order to structure the flood; without such tools, large corpora will not be of much use.

Finding the sense in a large corpus can be seen as the repetitive process of making ever more specific queries. The first basic query is that all the instances of a certain lemma be given. The following queries include contextual restrictions which can be made more precisely the more annotated the corpus becomes. The querying is repeated until some characteristic behaviour of the lemma crystallizes. Once such behaviour (e.g. one meaning, one valency frame) has been recognized and described by the lexicographer, the instances of it may be discarded and the procedure repeated for the remaining instances.

There is, however, one class of important questions that cannot meaningfully be answered solely on the basis of the immediate context of the instances of a lemma. Computational exploration of the collocational behaviour of a word is not possible without some knowledge of the corpus as a whole. The mere observation that one word seems to be occurring frequently in the neighbourhood of another word does not in itself indicate an affinity between the two, neither does a seemingly infrequent occurrence indicate the absence of such an affinity. Only a statistical calculation that takes into account the total number of occurrences of the words in question can give a reliable indication. A useful survey of methods and tools for identifying collocations in corpora is given in Fontenelle et al. (1994).

Since the work of Church et al. (1991) three statistical methods for collocational studies have become more or less standard. These or similar methods should be part of any toolbox for the analysis of large corpora. *Mutual information* (or the cognate *Z-score statistics*) reveals positional interdependence between two words by comparing the observed frequency of a co-occurrence to the calculated frequency for co-occurrence by chance. *Scale statistics* calculates the mean and the standard deviation of the distance between such pairs, thus giving a measure of the fixedness of the collocation in question. The more sophisticated *T-score test* looks for significant differences between the immediate neighbourhoods of two different words, typically pairs of near synonyms like "strong"/"powerful" or "his"/her". The observed neighbouring words, e.g. words in the position immediately to the right of the two, are ranged on a scale spanning from those having greatest affinity to one of the synonyms, through those which are neutral, to those with greatest affinity to the other synonym.

6.2 An interactive corpus tool

For corpus search and interactive analysis, a tool called Corpus·Bench was developed by the Danish software house TEXTware A/S according to specifications made jointly by Longman Publishers (UK) and the Danish Dictionary. It is

commercially available and is also being used by a few other publishing houses and academic institutions.

Concordances can be built in real time according to complex search criteria. The concordance lines can be interactively tagged according to several user-defined criteria, and they can be sorted by almost any combination of criteria. Moreover, the statistically-based methods for collocational analysis mentioned above are available, and frequency information, including frequency distribution over e.g. text types, can be obtained.

For the use by Corpus·Bench, the corpus must be compiled and indexed by a separate software package called Corpus·Build. It allows the user to design the overall structure of the corpus database, such as the definition of the alphabet, character mappings and separators. It also provides a tool for building and maintaining an optional inflectional dictionary that can be accessed by the retrieval system and facilitate searching for lemmas rather than individual word forms. Corpus·Build can handle the indexing of large SGML-annotated corpora (at least 100 million words). The annotations may reflect any kind of information on the text document, e.g. headers, morphosyntactic tags etc.

6.3 Working with Corpus·Bench

Almost any search criterion can be used to create concordances from the corpus. As Danish has a more complex inflectional structure than e.g. English, a concordance normally should be based on a lemma rather than a single word form. An inflectional dictionary, based on the above-mentioned lexical database, was therefore added to the retrieval system.

One can scroll through a concordance listing, view the contents of header fields together with the corresponding lines in the concordance, jump into the corresponding document by clicking the mouse on a concordance line, mark up lines with one's own annotations, and sort the lines according to any combination of keyword, left context, right context, user-defined tags, and text type information. Concordances or parts of them can be printed out or copied either to a file or to the Windows-OS/2 clipboard in order to paste them into another document, such as a dictionary entry in the dictionary compilation system.

Search criteria based on keywords can be combined with two types of filters: word filters and/or text type filters. Word filters specify the absence or presence of additional words or lemmas in a given contextual position or range. Text type filters specify the contents of certain header fields (cf. 3.1 above). Any logical combination of up to eight word filters and text filters can be applied to a query, which allows the user to specify queries such as

> "display a concordance listing with the keyword 'typisk' *typical(ly)* AND the word 'dansk' *Danish* OR 'engelsk' *English* OR 'fransk' *French* OR 'tysk'

German in context position +1 in text by persons born outside Denmark (Region=X)".

What came out of this example query were two statements: *Hiding one's light under a bushel may be a typical Danish expression, but doing so is not a salient Danish feature* and *Something being typically French implies that the opposite, too, is typically French.* Both authors happen to be born in the former Soviet Union.

Filters can be defined for all types of queries, including word-lists and statistics.

Word-lists show words according to specified search patterns (which will normally contain wild cards). As compound words are very common in Danish, a word-list can be used to investigate the productivity of a given word. For example, Corpus-Bench can list all words with the string "engelsk" in them (search pattern: **engelsk**), and the resulting list can be sorted alphabetically or, like here, by frequency:

Word	**Abs. frequency**
engelsk *(English)*	3 081
engelske *(English)*	2 630
engelsksprogede *(English-language [adj.])*	53
engelsktalende *(English-speaking)*	38
engelsksproget *(English-language [adj.])*	35
engelsklærer *(English teacher)*	24
engelskundervisning *(teaching of English)*	22
engelskundervisningen *(the teaching of English)*	10
engelsktime *(English lesson)*	10
engelskfødte *(English born)*	9
engelskkundskaber *(knowledge of English)*	8
engelskgræs *(thrift [a plant])*	7
dansk-engelsk *(Danish-English)*	7
engelsklæreren *(the teacher of English)*	6
oldengelske *(Old English)*	5
engelsk-amerikanske *(Anglo-American)*	5

Frequency lists give the absolute and relative frequency of the word forms belonging to a given lemma. By defining filters, one can investigate the use of a given word in different subcorpora. It is also possible to compare the frequencies of words that are related to each other. For the word "virus", two genders and several inflectional variants are permitted; the frequency list, giving the number of instances and the number per million running words, shows that inflection is normally avoided, and that far from all of the inflected forms are used:

	Abs.no.	Per mil.		Abs.no.	Per mil.		Abs.no.	Per mil.
virus	813	20,35	virussets	1	0,03	viruserne	0	0,00
virus's	0	0,00	virusets	4	0,10	virussene	0	0,00
viruses	0	0,00	virusser	5	0,13	virusserne	0	0,00
virussen	11	0,28	viruser	6	0,15	viraene	1	0,03
virusen	20	0,50	vira	43	1,08	virusenes	0	0,00
virusset	5	0,13	virussers	0	0,00	virusernes	0	0,00
viruset	25	0,63	virusers	0	0,00	virussenes	0	0,00
virussens	1	0,03	viras	3	0,08	virussernes	0	0,00
virusens	4	0,10	virusene	0	0,00	viraenes	0	0,00
						Total	942	23,58

A **word distribution report** shows the use of words which are distributed according to the contents of a header element, e.g. the year of birth, subject area, or time of publication. The verb "start", borrowed from English, was originally only used in connection with motors, cars and the like. However, it is gradually also taking over the more general meaning of "begynde" (*begin*); this is mirrored by the word distribution report by age:

Birth	Abs.no.	Total	Per mil.	Dev.pct.
?	6 808	18 921 566	359,80	+19%
1910s	146	1 129 424	129,27	–57%
1920s	431	2 322 307	185,59	–39%
1930s	734	3 842 267	191,03	–37%
1940s	1 735	6 570 818	264,05	–13%
1950s	1 612	5 196 341	310,22	+2%
1960s	643	1 916 274	335,55	+11%
1970s	20	48 836	409,53	+35%
Total	12 129	39 947 833	303,62	

A **mutual information report** displays a list of words that occur with a significantly high probability together with the keyword in a given contextual position or range. The report thereby identifies typical collocations. Most of the following left-side collocators of "interesse" (*interest*) represent expressions of the type *in the interest of ....* The factor *mut.inf.* measures how many times more frequent than chance the co-occurrence is[6], and *coocc.* is the actual frequency of each co-occurrence:

	mut.inf.	coocc.
nyhedens (*of novelty*)	6 582,28	[15]
sandhedens (*of truth*)	1 513,92	[23]
alles (*of all, common*)	730,34	[34]
medlemmernes (*the members'*)	677,59	[10]
offentlighedens (*of the public*)	639,94	[10]
almen (*common*)	440,22	[15]
fornyet (*renewed*)	288,62	[14]
stigende (*growing*)	270,42	[78]

befolkningens *(of the population)*	197,33	[10]
størst *(greatest)*	187,52	[21]
manglende *(missing)*	181,31	[40]
voksende *(growing)*	167,07	[19]
speciel *(special)*	147,33	[20]
øget *(additional)*	111,03	[22]
stor *(great)*	107,82	[280]
betydelig *(considerable)*	98,78	[14]
offentlig *(public)*	96,42	[14]
historisk *(historical)*	95,49	[14]
samfundets *(of society)*	93,75	[10]
fælles *(common)*	84,83	[59]
særlig *(special)*	82,82	[61]

Finally, **T-score reports** are used for investigating differences in the use of words that are related to each other in some aspect. A T-score report can be thought of as two mutual information reports compared to each other. The report given below shows what is — to a Dane — typically German but at the same time untypically French and vice versa. While T-score reports normally do not show unexpected results when based on adjectives of nationality, they are very useful in lexicography for the investigation of slight differences in the use of almost synonymous adjectives, e.g. "strong" *vs.* "powerful" or "big" *vs.* "large".

German	T-score	French	T-score
genforening *(reunification)*	9,78	revolution *(revolution)*	–10,55
2 *(television channel)*	7,14	præsident *(president)*	–6,95
besættelse *(occupation)*	6,14	franc *(= the currency)*	–6,80
soldater *(soldiers)*	5,87	skole *(school)*	–5,70
forbundsbank *(federal bank)*	5,72	francs *(= the currency)*	–5,59
rente *((rate of) interest)*	5,27	kartofler (franske kartofler) *(potatoes (potato crisps))*	–4,50
1 *(television channel)*	5,03	koloni *(colony)*	–4,23
stater *(lands)*	4,98	visit *(f.v. = flying visit)*	–4,03
soldat *(soldier)*	4,82	konge *(king)*	–3,78
forbundskansler *(Federal Chancellor)*	4,67	polynesien *(Polynesia)*	–3,64
enhed *(unity)*	4,56	køkken *(kitchen, cooking)*	–3,62
fransk *(French)*	4,43	kunst *(art)*	–3,62
værnemagt *(Wehrmacht)*	4,10	off. *(= official (language))*	–3,48
rige *(State, Reich)*	4,10	ord *(word)*	–3,42
kansler *(Chancellor)*	4,10	revolutions *(revolution's)*	–3,36
besættelsesmagt *(occupying power)*	4,10	alper *(Alps)*	–3,36

What makes Corpus·Bench different compared to most other commonly-used corpus retrieval systems is its capability of handling extra-textual information. Queries are not limited to the raw text of the corpus, but may be modified by the information supplied in the headers, as well as by part-of-speech tags, if available.

7. Third parties' use of the corpus

The linguistic resources developed for the dictionary project, the corpus, as well as the lexical database have already been widely used by researchers and students of Danish. Among the topics for corpus based term papers and theses written by university students are "The concept *sand* (true)", "Topology and interpretations of the adverb *kun* (only, just)", and "Topology of some adverbials in spoken language". For a term paper on automatical identification of technical terms in professional text, a lemmatized list of frequent words in general language was produced. PhD theses and studies by senior researchers include work on prototypical sensory and speech act verbs; onomatopoeic words in written and spoken Danish; valency patterns of adjectives; the concept *politician*; stylistics; lexical semantics; and some derivational affixes. A corpus-based study of types of language errors was made as part of preparatory work on a syntax checker for Danish.

7.1 Criteria for access

The access to the corpus for external users is regulated by three kinds of considerations: copyright, resources available, and a wish for survival.

During the compilation of the corpus no formal copyright agreements were made; and it would in fact have been a major job to find the authors of 44 000 distinct pieces of text and get their permission. The publishers and others who supplied text, were promised that it would only be used for dictionary work and other research; furthermore, as far is known most of them did not ask permission from the actual copyright holders, namely the authors. Consequently, the corpus had to be handled like photocopies: it is permissible to make one copy for personal use, but illegal to duplicate and distribute copies. External users, therefore, normally do not receive (sub)corpora, but rather concordances or word-lists, or they are invited to query the corpus on the premises of the Dictionary, where a special subcorpus for guests is available. The "guest corpus" excludes a few million words on which special restrictions were laid by the suppliers. However, making concordances or word-lists, and instructing guests in the use of the corpus tools, encroaches upon time for working on the dictionary, and given the sparse resources available, help has to be somewhat limited. On the other hand: widespread use of the corpus for many different purposes would prove the need for its continuation after the end of the dictionary project. That is where the wish for survival comes in, and that is one reason why every instance of external use is carefully recorded. No charges have been made so far, partly because quite a few of the users of the corpus — or their institutions — were in fact also providers of textual material to the corpus.

8. The PAROLE project

A new dimension, and a new approach to the question of availability, was added to DSL's corpus work when the Society became a partner of the PAROLE project in 1994, the aim of which was to provide publicly accessible harmonized comparable corpora and lexica (i.e. dictionaries which can be accessed and used by computer programs) for the official languages of the European Union and for Catalan and Irish — a total of 14 languages. The corpora focus on written language, and their primary target group is the language industry. Consequently, the design criteria are not the same as for the corpus of the Danish Dictionary; among other things, childrens' language and other nonstandard variants have been left out. Three kinds of corpora should be made, viz. a 20 million word publicly accessible corpus, a 3 million word distributable corpus, and a 250 000 word morphosyntactically tagged, and manually checked, corpus. Producing the tagged corpus was by far the most labour-intensive part of the job, as no experience in this field, let alone an automatic tagger, was available for Danish. The next step will now be to use the 250 000 words for training some taggers which are known to have been successfully used with other languages.

9. Future development

As already mentioned, the immediate goal of the project is a manuscript for a six volume dictionary of contemporary Danish, which will be completed by 2002-03. Further objectives for the future of the corpus include a strengthening of the diachronic dimension of the corpus, as well as the integration of computational methods in the philological editorial work of the Society. Techniques used for corpus building and analysis may also prove useful for the preparation of scholarly text editions, as well as for the use of such editions, which are likely to be published electronically in the not too distant future. As to the future of the dictionary, an electronic version is likely to be the next step. It will be accessible not only by headwords (semasiologically) but also by concepts (onomasiologically). Preparations for such access are part of the ongoing work. Furthermore, it may provide far more examples of real language than the printed version.

Notes

1. The authors want to thank their colleagues at *The Danish Dictionary*, Henrik Andersson and Ebba Hjorth, for input to and comments on the manuscript.

2. The Carlsberg Foundation, the owner of most Danish and several foreign breweries, is among the most important sponsors of Danish science and scholarship.

3. The 15 texts from 1993 were a series of transcribed interviews planned for 1992. They happened to be delayed for a couple of months. It was, nevertheless, decided to include them.

4. The IBM specific character set "Code Page 850 (Latin 1)" holds an inventory of Western European letters which is close to that of the ISO character set 8859-1. Conversion between the two character sets is not a major problem.
5. In order to make the text more readable to humans braces {...} have been chosen for the delimiting of SGML-entity references, instead of the standard SGML-delimiters &...; (ampersand ... semicolon) which can therefore be used with their original meaning. The braces are reserved characters that are not used for other purposes. A newline-entity {NL} is inserted as section delimiter, i.e. in places where the original text had one or more empty lines between paragraphs.
6. For instance, the word "stor" *(big, great)* appears 107 times more frequently to the left of "interesse" than would be expected if the words were randomly distributed. Strictly speaking, in information theory *mutual information* is defined as the logarithm to the base 2 of the figures which are here called *mut.inf.*

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

Church, Kenneth, William Gale, Patrick Hanks and Donald Hindle. 1991. Using Statistics in Lexical Analysis. Zernik, Uri (Ed.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.* New Jersey: Erlbaum.

Danish-English Dict. 1990. Vinterberg, Hermann and C.A. Bodelsen: *Dansk-engelsk Ordbog*. Third edition edited by Viggo Hjørnager Pedersen. Copenhagen: Gyldendal.

Danish-French Dict. 1-2. 1991. Blinkenberg, Andreas and Poul Høybye. *Dictionnaire Danois-Français/Dansk-fransk Ordbog*. Fourth edition edited by Jens Rasmussen et al. Copenhagen: Arnold Busck.

Fontenelle, Th., W. Bruls, L. Thomas, T. Vanallemeersch and J. Jansen. 1994. *Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora.* Document D-1a of DECIDE (MLAP-Project 93-19), Luxembourg.

Maegaard, Bente and Hanne Ruus. 1987. The Composition and Use of a Text Corpus. Cappelli, A., L. Cignoni and C. Peters (Eds.). 1987. *Studies in Honour of Roberto Busa S.J.* Linguistica Computazionale IV-V: 103–21. Pisa: Giardini Editori.

Retskrivningsordbogen. 1986. Copenhagen: Dansk Sprognævn/Gyldendal.

Riber Petersen, Pia. 1984. *Nye Ord i Dansk 1955-75.* With contributions by Jørgen Eriksen. Copenhagen. Dansk Sprognævns skrifter 11. Copenhagen: Gyldendal.