*Jørg Asmussen**

The Text Corpus and Corpus Retrieval System of The Danish Dictionary

The project

A new dictionary of modern Danish

The Danish Dictionary (*Den Danske Ordbog*) is the newest dictionary project carried out by The Society for Danish Language and Literature. The Society is an institution under the Danish Ministry of Culture funded by the Danish Government. The aim of the Society is to make scholarly editions of Denmark's literature as well as bibliographies and dictionaries. Work on The Danish Dictionary began in 1991, and the dictionary will be published in six volumes by Gyldendal Publishers in 1998-99.

The Danish Dictionary aims at a comprehensive description of contemporary Danish from about 1950 until today, focusing its attention on the decade from 1983 to 1992. The dictionary will cover the written language and also pay attention to the spoken. It will be a broad common language dictionary including commonly used technical language. It will contain approximately 100,000 entries and give information on orthography, word class, inflection, pronunciation, meaning, phraseology, etymology. Authentic quotations will illustrate the use of the words. The dictionary will be as descriptive as possible — but still guide the user, no matter whether he is a native speaker or a learner of Danish. The dictionary will cater for both professional and general users of Danish.

Sources of the dictionary

The dictionary will mainly be based on a text corpus containing 40 million words of text drawn from a wide variety of sources. The lexico-

* *Jørg Asmussen*
The Danish Dictionary
University of Copenhagen
Njalsgade 80
2300 Copenhagen S (DK)

graphic information evidenced in the corpus is supplemented by

- inflectional information and semantic templates drawn from a couple of machine-readable dictionaries,
- material from The Board of the Danish Language (*Dansk Sprognævn*) — a national advisory body with a collection of close to one million slips with authentic examples of the use of Danish,
- and notes submitted by ‘word watchers’ (*spORDhunde*) — a group of more than 600 people collecting authentic language material for The Danish Dictionary.

The text corpus

Characteristics and design

The Text Corpus of The Danish Dictionary contains 40 million words of written and spoken Danish produced during the decade 1983-1992. Even though The Danish Dictionary aims at describing modern Danish from the fifties until today, the corpus doesn’t contain older texts. Older texts are normally not available in machine-readable form and typing or scanning is a rather time consuming process.

The corpus has been compiled by the editorial staff of The Danish Dictionary to get a reliable source for the dictionary. It consists of more than 43,000 annotated text samples from a wide variety of linguistic sources, e.g., common and technical, written and spoken, professional and non-professional, ‘young’ and ‘adult’ language. The text samples represent a variety of different media, genres, and topics. The corpus pays special attention to spoken language and contains 7 million words of private interviews, political debates, broadcasts, etc. Non-professional language amounts to a total of 4.5 million words of private diaries, letters, school exercises, etc. Main media are books, magazines, newspapers with 28 million tokens; radio and television with 3.8 million tokens; ephemera with 2 million tokens. The corpus thus aims at being as linguistically ‘mixed’ as possible.

Annotations

Each of the 43,000 text samples in the corpus consists of a header followed by the text proper. A header consists of a finite number of fields that have been filled in with appropriate information on the text during the compilation of the corpus. A number of fields are for statistical purposes — only a value from a finite set can be assigned to them. Furthermore these fields allow us to define special ‘filters’ in order to create special virtual sub-corpora. A header is, by means of SGML tags, structured like this (fields for statistical purposes are marked with a dot •):

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 not be used by others than The Danish Dictionary before [year]
- **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, .. , 1991, 1992}
  - **Sure**...................The year of publication is known exactly (-), or not (?)
- • **Location** ....E.g. book volume, newspaper section, page number
- • **LangType**................{general, LSP}
- • **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 partly media-dependent values
- • **GenreType**..............A reduced genre-classification with 17 values – for statistical purposes
- • **Topic** ....{philosophy, geography, computing, physics, ..} – 66 possible 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)
- **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}
•	**YearOfBirth**	{1880, 1881, .., 1989, 1990}
	Sure	The year of birth is known exactly (-), or not (?)
	PlaceOfBirth	Place of birth
	PlaceOfResid	Place of residence
•	**Region**	Dialectal region, derived from PlaceOfBirth/Resid – 11 possible values
	Education	Education of the language user
	Occupation	Occupation of the language user
•	**LangVariant**	{standard, regional}
	Role	Communicative role of the language user, e.g. teacher, pupil

The corpus retrieval system Corpus·Bench

Overview

Corpus·Bench (CB) is the computational tool that the team of lexicographers at The Danish Dictionary use to retrieve linguistic information from the corpus. The software has been developed by TEXTware A/S, Copenhagen, on the basis of requirement specifications from The Danish Dictionary and Longman. CB consists of two components — Corpus·Build and Corpus·View.

Corpus·Build allows you to design the overall structure of the corpus database, to define an alphabet, character mapping, and separators. It provides you with tools to build and maintain an optional inflectional dictionary that can be accessed by the retrieval system in all kinds of searches and enables you to lemmatize word forms. Corpus·Build can handle the indexing of an arbitrary amount of SGML-annotated corpora. Annotations may be any kind of information on text documents, e.g., headers, morpho-syntactical tags, etc. The Corpus·Build software works in a common DOS-environment, thus allowing us to store the indexed corpus on any DOS-machine.

Corpus·View gives you access to your corpus database. You can interactively generate concordances, word lists and statistical reports. Search criteria can be specified by using wild cards, a lemmatization dictionary, POS-tags, and they can be modified by filters based on other words in the context of the key word or on the contents of certain header-fields, thus enabling you to define virtual sub-corpora. The Corpus·View software runs under OS/2, but can access a corpus stored and built on a DOS-machine, e.g., a server in a PC-network. OS/2 allows you to run OS/2, Windows and DOS applications concurrently.

Concordances

It is possible to create concordances according to almost any search criteria. A very simple example could be the key word form *engelsk* ('English'). The generation of an *engelsk*-concordance based on the 40 million word corpus of The Danish Dictionary takes approximately 20 seconds (3058 occurrences). As Danish has a more complex inflectional system than English, a concordance should rather be based on a lemma than on a single word form. CB can derive inflectional information from an inflectional dictionary and use it in different kinds of queries. The generation of a concordance with the key lemma *engelsk* takes approximately 25 seconds and displays the 5672 context lines in a window on the screen. It is possible to scroll through the concordance listing, view contents of header fields together with the corresponding lines in the concordance, jump into the corresponding text document by mouse clicking on a concordance line, mark up lines with own, e.g., lexical, annotations, sort concordance listings by almost any criteria, print them and copy from both concordance listings and documents either to a file or to the Windows-OS/2 clipboard and thus paste them into any other document, e.g., a dictionary document in a dictionary compilation system.

Key word based search criteria can be combined with two types of filters: word filters and/or text type filters. Word filters define the absence or presence of certain additional words or lemmas in specific contextual positions or ranges around the key word. Any logical combination of several word filters can be defined. Text type filters are defined on contents of certain pre-defined header fields — in our case those marked with a dot in the header description given above. This enables you to specify a query as, e.g., *please display a concordance listing with the key lemma 'engelsk' and the word form 'mad' (which means 'food' in Danish) in context position +1 in either newspaper articles written by men born in the sixties or any text on the subject 'food' written by a woman*. In the corpus of The Danish Dictionary we find 1 occurrence that matches these conditions — a newspaper article on the subject British theater. The linguistic relevance of this example is probably somewhat restricted, but it might give an indication of the type of queries you can carry out with CB, mainly that by means of text type filters you can define any virtual sub-corpus you can think of. Filters can be defined for all types of queries — also word lists and statistics.

Word lists and statistics

Word lists show words according to certain search patterns (often containing some wild cards). As compound words are very common in Danish, a word list can give you an idea of the productivity of certain words, e.g., CB can list all words ending on ‘*engelsk*’ — the resulting list can be sorted alphabetically or by frequency and shows that *dansk-engelsk* (‘Danish-English’) is the most frequent compound word ending on ‘*engelsk*’. Weird examples on the list are *alpe-engelsk* (in a text on Arnold Schwarzenegger), *baby-engelsk*, *Dallas-engelsk*, *kolibri-engelsk*, *management-engelsk*, *pseudo-engelsk*. In word lists you can define an additional frequency filter.

Frequency lists simply list the absolute and relative frequency of certain word forms belonging to one lemma. By defining filters you can get an idea of the use of certain words in different sub-corpora. Or you can compare the frequency of words that might be related to each other in certain aspects: thus the overall relative frequency for the lemma *dansk* (‘Danish’) in our corpus turns out to be 1174 occurrences per million running words, whereas *engelsk* only occurs 141 times in a million — so the corpus of The Danish Dictionary in fact seems to be very Danish!

A word distribution report shows the use of certain words distributed according to the contents of a header element, e.g., year of birth, topic, publishing time etc. In Danish *starte* (‘to start’) and *begynde* (‘to begin’) are close synonyms. A word distribution report can show if a word has a significant high or low frequency within e.g. a certain age group, sex, region, medium etc. The distribution of *starte* over the year of birth shows that the overall average frequency is 301 occurrences per million tokens. The report shows that *starte* seems to be more popular among speakers born in the sixties (relative frequency: 332) and seventies (381), whereas those who are born in the decade 1910-19 only use *starte* 158 times in a million words.

A mutual information report displays a list of words that co-occur with a significantly high probability together with the key word in a certain contextual position or range around the key word and thus gives an indication of typical collocations. A mutual information report can simply tell you what, e.g., is typically English (from a Danish corpus point of view). A mutual information report on *engelsk* and the word in position +1 to it gives the following strong collocations at the very top of the list (descending order of collocability): *bookmakere*, *underhus* ('House of Commons'), *hooligans, dronning* (queen), *bullterrier*, *fodboldfans…*

T-score reports can be used to detect differences in the use of words that in some aspects are related to each other. A t-score report can be described as two mutual information reports compared to each other. If you want to detect words that are 'typically English' but at the same time 'untypically Danish' and vice versa, you could generate a t-score report on *engelsk* and *dansk*. T-score reports based on 'national' adjectives normally do not show any unexpected results. In lexicography, t-score reports are very useful to detect slight differences in the use of almost synonymous adjectives, e.g. *strong* vs. *powerful, big* vs. *large*, etc.

Who can use Corpus·Bench?

What makes Corpus·Bench different compared to most other commonly used corpus retrieval systems is the capability of handling a lot of extra-textual information. Queries are not only limited to the raw text in the corpus, but can be modified by any extra information on the text documents in the corpus — provided that every text document is marked up with such information (headers, POS-tags, etc.). Corpus·Bench is designed for very large corpora (100 million words and more), that contain a vast, but strictly organized amount of extra-textual information. Almost any kind of corpus database can be set up with Corpus·Bench, but it requires a lot of planning and testing. Once a corpus database is set up, CB handles any query in a fast and efficient way. So if one works with corpora of the above mentioned type, CB will be a good choice. If one's corpus merely is a collection of text documents without any annotations that can be used for filtering and statistical purposes, Corpus·Bench probably is sheer overkill.

References

Church et al.: *Using Statistics in Lexical Analysis* (in Zernik: *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, New Jersey, 1991) — exact definitions of mutual information and t-score.

TEXTware: *Corpus·Bench User’s Manual. Version 1.0. December, 1993*. (TEXTware A/S, Hørsholmsgade 20,2, DK-2200 Copenhagen N) — comprehensive description of all features of Corpus·Bench.