The Dictionary of the Danish Language Online: From Book to Screen – and Beyond

LEXICOGRAPHIE ET INFORMATIQUE : BILAN ET PERSPECTIVES

Colloque international à l'occasion du 50e anniversaire du lancement du projet du *Trésor de la Langue Française*

Dates : 23-25 janvier 2008

Lieu : Nancy, Campus Lettres et Sciences Humaines

Organisation : UMR ATILF (CNRS/Nancy-Université)

Henrik Lorentzen (1) hl@dsl.dk

Lars Trap-Jensen (2) ltj@dsl.dk

Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature), Christians Brygge 1, DK-1219 Copenhagen (1)

Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature), Christians Brygge 1, DK-1219 Copenhagen (2)

Keywords: historical dictionary, digitisation, XML mark-up, supplementary volumes, dating information, advanced search, corpus, WordNet

Abstract : The Dictionary of the Danish Language was digitised and published online for the first time in 2005. The paper first describes the digitisation method (the so-called double-keying) and then presents and discusses some of the challenges for the future, e.g. the integration of supplementary volumes, a refined mark-up and better dating information. Finally, the perspectives involved in the integration of corpora, dictionaries and other language resources are briefly introduced.

Introduction

This paper is concerned with a large-scale monolingual dictionary, the Danish counterpart to le Trésor de la Langue Française, to the Oxford English Dictionary, to the Woordenboek der Nederlandsche Taal and others, i.e. historical dictionaries in many volumes that aspire to be comprehensive. The Dictionary of the Danish Language (Ordbog over det danske Sprog, henceforth ODS) was published from 1919 to 1956 in 28 volumes. It covers the Danish language from about 1700 until about 1950. During the compilation period, editorial guidelines changed considerably, and as the staff continued to collect material, the inevitable consequence was that the last part of the alphabet was covered more comprehensively than the first. Soon after the last volume was published it was decided to start gathering material for a supplement. With only a small staff, it took some decades to complete, but from 1992 to 2005 five supplementary volumes were published, meaning that the complete work comprises 33 volumes.

The original dictionary was conceived by the Danish linguist Verner Dahlerup as a much smaller project, originally designed to be carried out by one man, himself, but even before the first volume was published, he realised that the job far exceeded his working capability, and it was decided that the project should come under the auspices of the Society for Danish Language and Literature (Det Danske Sprog- og Litteraturselskab, DSL). It is also within the framework of this institution that the work with the online version of the dictionary takes place.

The goals of the current project, which runs from 2004 until 2010, are threefold:

1. to provide public online access to the dictionary
2. to integrate the supplementary volumes with the original dictionary

3. to integrate the new supplemented dictionary with other linguistic resources developed by DSL, in particular The Danish Dictionary (Den Danske Ordbog, DDO, a medium-size dictionary of modern Danish) and corpora of modern Danish

Digitisation

Since the original was only available as a printed book, the first step was raw digitisation. Different methods have been explored by other projects:

1. keying and proofreading (OED)
2. scanning and proofreading (SAOB, the large Swedish dictionary)
3. double-keying without proofreading (Grimm's Deutsches Wörterbuch)

In this project it was decided to adopt the model used for the German dictionary founded by the Grimm brothers, Deutsches Wörterbuch, the so-called double-keying method. The printed dictionary data is keyed twice, by two independent typing teams, and afterwards the two versions are compared electronically. In this way, the number of typing errors is reduced to almost nothing, and thus the labour-intensive process of proofreading that is normally applied can be avoided.

This part of the project was carried out in collaboration with the University of Trier, and monitored by the same department which was responsible for the Grimm project (http://germazope.uni-trier.de/Projects/KoZe2/). The actual typing of the dictionary took place in Nanjing, China, by the company TQY DoubleKey which has specialised in this type of assignment and can deal with different sorts of typefaces including black-letters and even hand-written manuscripts. After the keying process, the two versions were automatically compared in Trier and a list of discrepancies generated. The list was subsequently processed semi-automatically and manually. All the difficult and dubious cases had been specially marked by the keyboarders, and about 2,000 instances had to be solved by the editors in Copenhagen because native-speaker competence was required. In many cases the answer was straightforward for a native speaker, but in other cases it was necessary to go back to the original slip to resolve the issue. The percentage of genuine errors has not yet been calculated, but spot checks carried out by the Grimm project yielded a result as low as 1 error in 33,000 characters – hardly surprising when you think of it, as it is highly unlikely that two keyboarders would make identical mistakes at exactly the same place. In the ODS, the rate may be expected to be even lower, an estimated 1 in 100,000 characters, because of a clearer structure and typography in the printed text.

The ultimate aim is to establish a fine-grained XML mark-up of the dictionary text, but it cannot be done in a single round. As a first step, a crude mark-up has been implemented in which only the headword, homograph number (if any) and the part-of-speech are identified. The rest of the entry is treated as one chunk. This mark-up allowed us to release a first preliminary version in November 2005 where the only possible search was for headwords. A second version was launched in April 2006 where wildcards and parts-of-speech were introduced as search criteria. The new features have improved the online version, no doubt, but there is still a lot of work to be done.

The files from Trier are in a format close to the TUSTEP[1] standard where every typographic detail from the book is rendered by means of codes: font size, bold, italics, spacing; special characters like the Danish *æ*, *ø* and *å*, and symbols used as labels (e.g. anchor = nautical language; book = literary; note = music). The files also give exact information about the page and the line in the dictionary, information that will prove useful when cross-references are to going be processed.

In the current online version the dictionary contains more than 180,000 headwords to which number may be added about 70,000 sub-headwords (words that do not have their own entry, but are nested within another entry) and supplementary articles which will take the total number to 250,000. The number of definitions, citations and multi-word units is still unknown but the improved mark-up will reveal it in a couple of years.

[1]TUSTEP (= Tübinger System von Textverarbeitungsprogrammen) is a text processing and layout system especially used for philological text editing.

Future work

1.1 Integrating the supplement

An important task is to integrate the five supplementary volumes into the original dictionary, which is far from being a trivial issue. In the online version of the OED (the so-called third version), the editors have included the Additions to the Second Edition (printed in three volumes 1993-1997) as well as additions that have never been published in print. The additions are always presented in a section of their own, after the original entry or as an entirely new entry if that is the case. At least as a preliminary measure for a work in progress, this seems a reasonable approach and it is likely that a similar procedure will be adopted for the integration of the ODS and its supplementary volumes.

A major challenge lies in the fact that the supplementary volumes were meant to be used as printed books in connection with the original printed volumes; therefore a thorough knowledge of the structure of the original dictionary is required in order to benefit from the additions and corrections proposed by the supplement entries. An elaborate system of markings is used to indicate how a supplement entry should be interpreted.

Additions are marked by the plus sign (+). A plus sign before a headword means that it is a completely new entry. This is a rather straightforward situation since the new entry can be treated on a par with the existing entries and the headword can be added to the general list of headwords. Examples of this are for instance loan words that were excluded from the original dictionary due to a somewhat purist approach on the part of the editorial staff: *caddie, café au lait, cafeteria, cancer* (cf. [Hjorth, 1990]). Another category of new headwords is compounds that were not included in the original dictionary, for instance due to low or no frequency in the collection of dictionary slips; examples of compounds with the word *dag* 'day' are *dagsaktuel* 'topical', *dagsdosis* 'daily dosis', *dagsprogram* 'programme or plan for the day'. A third major category is new words that have entered the language during the period covered, but after publication of the relevant volume, e.g. *a-bombe* 'A-bomb'*, bilradio* 'car radio', *dobbeltmoral* 'double standard'.

Things turn more complicated when the plus sign occurs in front of other information types. In principle any part of the microstructure can be affected by the additions. Thus the plus sign may precede a new main sense, a new sub-sense, a new citation including a new date and a new author, a new cross-reference etc. In contrast to the added headwords this type of addition is hard to handle automatically and requires a substantial amount of manual effort because the editors of the supplementary volumes put confidence in the human user's capability to interpret the information correctly and locate the right place in the original entry.

1.2 Refining the mark-up

Another important and necessary task is to refine the mark-up of the entries. The information that is encoded in the typographical setting as well as in the serial ordering of information reveals to a high degree which microstructural element is involved.

The headword, the homograph number and the part-of-speech have already been identified. The next information categories we want to identify are the definitions, the citations and the citation sources. The printed book has some typographical pointers as to the interpretation of the data. The definition for instance is in italics and occurs after the etymology which is in turn placed within bold brackets. In contrast to many other dictionaries the citations are in ordinary typeface, not in italics, whereas the sources for the citations are in italics and relatively well documented in a list of sources in volume 28. Multi-word units are another subtype which is fairly easy to identify as many of them are in spaced typeface and often preceded by formulae like "in the expression", "in the phrase" etc.

It is possible to identify the etymological information by means of the bold brackets, but the actual contents are often pretty hard to interpret and structure. However, we hope to benefit from the experience of the modern Danish dictionary, DDO, which is stored in XML, and possibly also from the work on the etymologies of the TLF conducted by Salmon-Alt [Salmon-Alt, 2006].

1.3 Dating information

Unfortunately, the ODS does not provide explicit information about the first occurrence of a word form in the language, but the information can to a certain extent be deduced from either the earlier word forms given in the etymological section or from the dates assigned to the citations in the source list. As a rule, the ODS brings the oldest occurrence as the first citation, meaning that the first edition of the source text could be used as basis for dating. However, this fact is obscured by an editorial practice of using collections of texts as a source rather than first editions, e.g. complete editions of a writer's work. The year given for a particular citation is therefore

potentially misleading for dating purposes if the citation is taken from a collection of works. In that case, the year would refer to the publication of the collection and not to the first edition of the text. By way of example, the fairy tales of Hans Christian Andersen are cited from an edition that was published in 1919, 44 years after Andersen's death. This may perhaps lead some users to conclude that The Ugly Duckling was published in 1919 and not in 1843 which is the actual year of first publication. At the moment work is being undertaken to improve the dating information by assigning the first year of publication to all texts being cited from collected works. Thereby we will, on the one hand, be able to provide explicit information about the first year of publication for all citations and, on the other, always be in a position to use the first edition of a text when using a citation as evidence for dating a word.

1.4 Cross-references

As with all large dictionaries, the ODS is full of cross-references, and there is a lot of work to be done in detecting and structuring them. Two major distinctions can be made:

1. between internal references (the target is in the same entry as the source) and external references (the target is outside the entry, i.e. in another entry, in the source list or even outside the dictionary)
2. between complete references (all information is to be found in the target) and references that only provide supplementary information (some information is already given in the source entry)

The original dictionary text provides many clues as to the interpretation of cross-references, e.g. the complete references are generally marked "see <target>" and the supplementary references are marked "cf. <target>" or "as opposed to <target>". In principle, all target entries should be marked as hyperlinks and thus made clickable in order to facilitate the user's navigation within the dictionary.

Perspectives

The ultimate goal is to create a state-of-the-art dictionary base, thereby reviving a project that is not only the largest one undertaken in Danish lexicography, but also one of the country's finest academic achievements in the field, and granting it the public attention it deserves. A fine-grained mark-up in combination with advanced search facilities will enable its users to broaden the scope of their inquiries as they will be able to make queries and find answers to questions such as: "Which words were borrowed from Arabic in the 18th century?", "Which adjectives occur in citations from authors like Hans Christian Andersen or Søren Kierkegaard?" or "How do the entries and the citations reflect the Danes' view of the Jews during the first half of the 20th century?" This means that the dictionary may be used in new ways and also for new purposes, for example cultural or literary studies, because the great wealth of information becomes accessible in ways which were not immediately available in the printed medium.

As a simple example of this you can compare the following entry for *Jødeskole* ('Jewish school') which is presented in full in the screen version (figure 1). Figure 2 shows a possible XML tagging of a relevant extract of the same entry. The citation in figure 2 runs like this in English translation: 'I command everybody else to be quiet; this place will soon be like a Jewish school'; apparently a Jewish school was seen as a place full of noise and shouting.

Jødeskole, en. *egl.: synagoge; ogs.: skole for jødiske børn. Moth.J109. VSO. MO. StSprO.Nr.113.11.* || *nu især (dagl.) i udtr. for stærk støjen af mennesker, raaben i munden paa hinanden olgn.* (*jf.* -kirke). Her er en Støi som i en Jødeskole. *VSO. Mau.I.496.* Uden at fornærme nogen kunde man jo tro, at man var i en Jødeskole. *Sven Clausen. Forensiske Skuespil.(1920).25. (jeg)* kommanderer . . alle andre til at tie stille imens; her er jo snart som i en Jødeskole! *Borregaard.VL.III.363.*

Figure 1. Screen version of an entry.

```
<Lemma>Jødeskole</Lemma>
    <POS>en</POS>
    <Citation>(jeg) kommanderer . . alle andre til at tie stille imens;
    her er jo snart som i en Jødeskole!</Citation>
        <Source>Borregaard.VL.III.363.</Source>
        <Author>Einar Borregaard</Author>
        <Title>Viktor Løwe, I-III</Title>
        <PublYear>1924-26</PublYear>
```

Figure 2. Extract of the entry in XML format.

The XML format allows queries for particular words in citations from a particular period, in this case for instance compound words with *jøde* ('Jew') and the period 1900 to 1950. Such a query leads to a number of words, not all equally flattering, such as *jødepris* ('Jew's price, exorbitant price'), *jøderente* ('Jew's interest, usury interest'), *jødesmovs* ('yid') and *jødesnabel* ('Jew's conk', literally 'Jew's trunk').

Another goal is to integrate the ODS with other linguistic resources, both dictionaries and corpora. The ODS covers the period from about 1700 to about 1950, the DDO covers the period from 1950 until today, so together the two dictionary resources provide coverage of the last three centuries of the Danish language. A dictionary of the Old Danish language, i.e. from 1100 to 1500, is being compiled at the DSL, unfortunately by a very small staff so the number of actual dictionary entries is not very high. There are plans, however, to digitise the dictionary slips and publish them on the Internet so that the raw material can be made available to scholars and other people who take an interest in the earlier stages of the language.

The DSL already has corpora of modern Danish which were used for the compilation of the DDO and have been made publicly available [Andersen et al., 2002]. The idea is to link the corpora and the dictionary so that the user can navigate freely between the two resources. It is not very difficult to do since both resources are marked up in XML. A corpus of the language of the 18th and 19th centuries, however, is not yet available but it can be developed on the basis of the texts gathered in the Archive of Danish Literature (www.adl.dk) which covers the essential parts of the canonical Danish literature. A major obstacle is the large number of spelling variants but fortunately the ODS accounts for many of them. The long-term goal is to create the same integration and easy navigation in the older dictionary and corpora as in the modern ones.

A third perspective is the possible integration of wordnet resources. A wordnet for contemporary Danish following the format of Princeton WordNet, GermaNet, EuroWordNet and others is being prepared on the basis of several resources, among others the DDO [Pedersen et al., 2006]. It would be a useful improvement of the wordnet if it could link to the more comprehensive older dictionary, but it would also be a quite labour-intensive task to adapt the material to the wordnet standard and no concrete plans exist at the moment.

References

[Andersen et al., 2002] Andersen, Mette Skovgaard; Asmussen, Helle; Asmussen, Jørg (2002): The Project of Korpus 2000 Going Public. In: Braasch & Povlsen (eds.): *Proceedings of the Tenth EURALEX International Congress.* Copenhagen, 291-299.

The Dictionary of the Danish Language online: http://ordnet.dk/ods

[Hjorth, 1990] Hjorth, Poul Lindegård (1990): Danish Lexicography. In: *Wörterbücher, Dictionaries, Dictionnaires – Handbücher zur Sprach- und Kommunikationswissenschaft,* Band 5.2. Berlin, New York: Walter de Gruyter, 1913-1922.

[Pedersen et al., 2006] Pedersen, Nimb, Asmussen, Sørensen, Trap-Jensen, Lorentzen (2006): DanNet – a wordnet for Danish. In: *Proceedings from Third International Conference on Global Wordnets*. Jeju, South Korea.

[Salmon-Alt, 2006] Salmon-Alt, Susanne (2006): Data Structures for Etymology: towards an Etymological Lexical Network. In: Corino, Marello & Oensti (eds.): *Proceedings XII EURALEX International Congress.* Torino, 79-87.