Yvonne Cederholm, Språkdata, Dept. of Swedish, Göteborg University

A Historical Lexical Database of Swedish. The O.S.A Project

Abstract

Large historical dictionaries have sometimes been called information graves because of the difficulty to perform systematic searches in the material. Recently, there have been efforts to make these dictionaries machine tractable. The O.S.A project is carrying out the computerization of the largest historical dictionary of Swedish, *Svenska Akademiens ordbok* (SAOB). This paper describes the main work, which is currently being done on the project, namely the semi-automatic SGML encoding of the dictionary. It also gives a brief account of the work on some interface prototypes and the optical scanning acquisition. Some future developments and the publishing plans are outlined.

1. Introduction

In the 19th century some monumental historical dictionaries were initiated, for example the *Oxford English Dictionary* (OED), *Woordenboek der Nederlandsche Taal* (WNT) and *Deutsches Wörter-buch*. The dictionaries contain an enormous amount of information, but because of their size and complexity they are also difficult to use. Lots of interesting facts are buried in these masses of information. *Svenska Akademiens ordbok* (SAOB),[1] which is edited under the auspices of the Swedish Academy, is the largest Swedish historical dictionary. It comprises 31 volumes and is not yet completed. SAOB is currently being transferred into a lexical database by the O.S.A project, which is funded by the Swedish Academy and is headed by Professor Sture Allén. The project can be described as a Swedish parallel to the New OED Project at the University of Waterloo. It was started a year earlier, however, and thus independently of the latter. The primary reason for the computerization of SAOB is to make the information more readily accessible for linguistic research. Since the dictionary is not completed yet, it will be an important aid for the editors. The preparation of a second edition will also be facilitated. The current work of the O.S.A project is mainly SGML encoding of the dictionary. We are also building interface prototypes in order to try out various presentation formats.

2. Svenska Akademiens ordbok (SAOB)

The purpose of SAOB is to record the vocabulary of written standard Swedish from 1521 to the present day, a period normally identified as New Swedish. Its focus is mainly to describe the semantic development in that period. Medieval Swedish is covered by two other dictionaries, *Ordbok till Samlingen av Sweriges Gamla Lagar* by P.C.J. Schlyter, which is exclusively concerned with the legal language of that period, and *Ordbok öfver svenska medeltidsspråket* by K F. Söderwall.

The first fascicle of SAOB was printed in 1893. The most recent volume, which was printed in 1993, records the words *stod–stå*. Approximately 500 000 words are described in the 31 volumes.

The dictionary format is well documented, principally in the manuals used by the editors and also in several books and articles (e.g. Ekbo & Loman 1965; Lundbladh 1992). Considering that the work has been progressing for more than a hundred years and that there has been a succession of chief editors, it must be claimed that the original format has been followed very closely over the years. The first parts that were printed before the 1920's differ in many respects from the parts that were printed later, which reflects a major reorganization of the editing routines that took place in these years. Some structural variation can be found in the later parts as well, but abbreviations and typography are used quite consistently in SAOB.

3. Acquisition of a machine-readable dictionary

The acquisition of the printed dictionary is now completed. The method used was scanning and optical character recognition (OCR). The text was then proofread once. It can be argued that manual typing is a preferable acquisition method, considering that the text is typographically very complex and that scanning errors occur frequently. We have found that an advantage of OCR is that the scanning errors are of another character than the errors made by humans. They are therefore more striking and easier to discover during proofreading. There is also the risk that by manually typing the text, the typists will incorrectly normalize the historical examples.

In 1994 the editors started to use desktop publishing. The process of transferring the desktop publishing format to SGML is fairly simple compared to the encoding of the scanned text.

4. The markup process

It is of course a difficult, as well as a labour intensive, undertaking to mark up a dictionary of this dimension. Most of the entries span several columns, but the size of the entries varies from a few lines up to 80–90 columns. The variation in entry size implies a great variation in the lexical entry structure. As mentioned above there is also some variation due to the fact that the work has proceeded for more than a hundred years. Most of the elements of the entry structure are optional and some elements can appear almost everywhere in the structure. This is the case with, for instance, all kinds of bibliographical references. The macrostructure of SAOB is hierarchical. Most compounds and derivations which are considered lexical entries in their own right, are placed under a headword. The maximum depth of the hierarchy of entries is three levels. Compared to the OED entries, the structure of the SAOB entries is often much more complex. It is partially due to linguistic differences between the two languages, for example the numerous compounds in Swedish, but also the fact that the sense descriptions are very elaborate in SAOB. The depth of the sense hierarchies is usually four or five levels deep, but the maximum depth is seven levels.

4.1 The markup strategy

The structural complexity and variation of SAOB causes a combinatorial explosion. To overcome some of the difficulties of variation in the markup process we used the following strategy. The text was not parsed in sequence, but we first tried to identify the lexical entities, from which we could build a kind of skeleton structure. The structure consists of the headwords; the labels of the sense hierarchies; compounds; derivations; verbal constructions and cross references. More than 800 000 entities are identified so far. The establishment of the skeleton structure was made solely on the basis of the typographical information. By looking for patterns of a tabulator followed by characters in upper case, we marked up approximately 95% of the headwords. Some of the other lexical entities were, however, more difficult to locate.

Once the main structure – the skeleton – was established, the next step was to identify certain well-structured isolated elements, for example usage information and bibliographical references. The isolated elements will later be combined into more complex elements and the final structure will be validated using a SGML parser. We used the following method for the markup of the isolated elements. We began by sorting the right contexts of certain lexical entities in the skeleton structure in order to find the most frequent regular patterns that followed the entities. The patterns did not only consist of typographical information, but of combinations of the following types of information:

- Style information. The OCR program recognized bold, italics and plain text.
- Information on upper versus lower case. This is of great importance as the words treated are all in upper case.
- Punctuation information. This information is quite unreliable because of frequent scanning errors. The most frequent errors occur for the most common punctuation marks – full stops and commas are confused and so are colons and semicolons.
- Lists of abbreviated labels. The labels represent, for example, gender information (*m., f., r., n.*, etc.) and part of speech information (*sbst., v., adj.*, etc.). The usage information also has a very strict form and lists of more than 350 usage labels have been used.
- Bibliographical information. We extracted information from a bibliographical database based on the source registers, containing approximately 19 000 records. The bibliographical database was created by Lars Svensson, chief editor of SAOB. It covers all the sources excerpted until 1990.

The patterns were then converted into SGML elements using common UNIX tools (*sed, awk* etc.) and *Perl.* We then used the encoded elements as a starting-point for further analysis of the right contexts. Most of the usage information was identified in this way. The identification of the citations and the bibliographical references demanded another supporting skeleton structure. We first encoded the date information, which was quite easy to identify, and then sorted the left context of the dates in order to find the authors and the titles of the bibliographical references.

The manual work has mainly been a control of the automatic markup. In more difficult cases the material is also preprocessed manually. This was especially helpful in the first stages when no context information was available and the markup was more of a hazard. The greatest manual effort so far was the markup of the divisions between the definition and first citation that follows the definition. It was not possible to do it automatically without a complete analysis of the definitions and that is too big a task for this project.

The following information categories are marked up: headwords; pronunciation; part of speech; gender; alternative forms; etymology; usage; definitions; quotations; bibliographical references (author, title, date, page reference); cross-references (to other entries); compounds; derivations; verb combinations; and general comments. There is no further analysis of the inner structure of most of these elements.

The present tag set has developed rather organically and it also includes lots of technical place holders. We do not, however, see any reason to make comprehensive adjustments to some standard or recommended tag set at this stage of the project. The final tag set will most probably be based on the tag set for print dictionaries recommended by the Text Encoding Initiative (Speerberg-McQueen & Burnard 1994: 321–370).

The bibliographical database will not only be used for the identification of the bibliographical references. Most of the references are identified solely on the basis of the typographical information and we now intend to automatically compare all the references in the dictionary to the information in the database. We will of course find some remaining scanning errors, but there is also a great variation in the bibliographical references that will be recognized. For instance, the titles can be abbreviated in many ways: August Strindberg's *Bland franska bönder*, is abbreviated both *FrBönd.* and *FranskaBönd.* We will add all these variants to the database and it may later be used to normalize the bibliographical references in the dictionary.

4.2 Database format and presentation format

The SGML format was decided on some years ago, and we then planned to use it as a temporary format which could later be transferred into a relational or object-oriented database. The SGML format seemed to be the best way to keep the text properties of the dictionary during the identification phase. As for the transfer into a traditional database model, it now seems possible to keep SGML as the base format, as it seems to have a potential beyond that of an interchange format. There are several ongoing research projects aimed at the development of a query language that can handle the SGML format, for example Blake et al. (1984:267–280) who have proposed extensions to SQL that handles SGML documents as well.

We are currently trying out different presentation formats. We decided, for instance, to make SAOB available in unstructured format, in spite of remaining scanning errors. It was incorporated in a concordance system, where the contexts of the KWIC-concordances can be expanded to entire columns. This system is used regularly by the SAOB editors via Internet. We have published information on World-Wide Web (WWW)[2]. Lists of all the headwords and derivations were sorted in reverse order and can be used, for instance, to find certain endings. There are also lists of compounds sorted by the last element and verb combinations sorted by the particle. We have developed a prototype hypertext dictionary which covers approximately 100 columns, *a–advokat*, and also includes a rudimentary search routine. The HTML documents are generated on the fly from the richer SGML dictionary format. The bibliographical references in the hypertext dictionary are linked directly to the bibliographical database.

5. Future Work

The encoded dictionary will probably be published on CD ROM, but the flexibility of the Internet and World-Wide Web seem to bring about new possibilities. SAOB could be connected to other material and in this way it is possible to create a large historical language database of Swedish, which could grow dynamically. There are many interesting projects that can be foreseen. The obvious connection is to link the bibliographical references to the bibliographical database. As the database contains complementary information on the sources and the authors, for example genre and sex, it will be possible to perform searches like "female authors from the period 1750-1800". A very interesting project would be to link to the two dictionaries of medieval Swedish to SAOB and in this way create a lexical database which covers eight centuries. Söderwall was also one of the early chief editors of SAOB, and there are obvious similarities in the structure of SAOB and the dictionary by Söderwall. It would also be interesting to link SAOB to some of the actual source texts from which the examples are derived, and in this way to form a large language database of texts and dictionaries. Linking to present-day synchronic dictionaries is also a possibility, but it is of course not trivial to bring together the synchronic and diachronic approaches.

It would of course be very intriguing to trace changes of meaning in a systematic way in SAOB, for example by analysing the definitions. This is however a very difficult task, as the definitions really consist of combinations of several definitions and synonyms. They often span more than ten lines.

6. Conclusions

In the past it was difficult to perform systematic searches in large historical dictionaries.

Today, computers are powerful enough to handle such large quantities of text, and it is now possible to transfer them into lexical databases. The standards and recommendations for text encoding are also making the task easier. The SGML encoding of SAOB is making fast progress and approximately 50% of the text is now encoded. The method used is a semi-automatic encoding. As the variation in entry structure is very large, we decided not to parse the text in sequence. First we identified the main structure and then separate elements that were easy to find. These separate elements were then combined and finally attached to the main structure. A machine readable version of SAOB is already being used regularly by the editors. The encoded version will probably be published on CD ROM and Internet/WWW.

Notes

1. The correct title, which is seldom used, is *Ordbok över svenska språket.*
2. The address of O.S.A project home page is: http://svenska.gu.se/saob/saobusers.html

References

Allén. S., Loman., B. & Sigurd, B. 1986. *Svenska Akademien och svenska språket.* Norstedts, Stockholm.

Berg, D.L., Gonnet, G.H. & Tompa, F.W. 1988. "The New Oxford English Dictionary Project at the University of Waterloo". University of Waterloo, UW Centre for the New Oxford English Dictionary OED-88-01.

Blake, G.E., Consens, M.P., Kilpeläinen, P., Larson P.-Å., Snider, T., & Tompa, F.W. 1994. "Text/Relational Database Management Systems: Harmonizing SQL and SGML". In *Applications of Databases* (ADB-94). Springer Verlag, Berlin Heidelberg. pp. 267–280.

Ekbo, S. & Loman, B. 1965. *Vägledning till Svenska Akademiens ordbok.* Norstedts, Stockholm.

Kruyt, J. G. & Van der Voort van der Kleij, J.J. 1992–93. "Towards a Computerized Historical Dictionary of Dutch". In *Acta Linguistica Hungarica*, Vol. 41 (1–4), pp. 159–174.

Lundbladh, C.-E. 1992. *Handledning till Svenska Akademiens ordbok.* Norstedts, Stockholm.

Malmgren, S.-G. 1988. "The O.S.A project: Computerisation of the Dictionary of the Swedish Academy". In *Literary and Linguistic Computing* 3: 166–168.

Rydstedt. R. 1988. "Creating a Lexical Database from a Dictionary". In *Studies in Computer-Aided Lexicology.* Almqvist & Wiksell, Stockholm, pp. 228–267.

Schlyter, P.C.J., 1877. *Ordbok till Samlingen av Sweriges Gamla Lagar.* Gleerups, Lund.

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

Söderwall, K.F., *Ordbok öfver svenska medeltidsspråket.* 1884–1918. Berlingska, Lund.