UPPSALA UNIVERSITY

Reports from the SCARRIE project
Editor: Anna Sågvall Hein

The SCARRIE Swedish Newspaper Corpus

Papers by:

Bengt Dahlqvist

WP CL&LE 6

Preface by the editor

SCARRIE is short for *Scandinavian Proof-reading Tools.* It is the name of a project within the EU TELEMATICS APPLICATIONS Programme. The project run for 30 months starting in November 1996.

"The SCARRIE project aims at the development of a high-quality proof-reading tool for the Scandinavian publishing industry. A concrete concrete result of the project will be a carefully evaluated demonstrator, the SCARRIE pilot, designed to meet the needs formulated by a user group consisting of representatives for Danish, Norwegian and Swedish newspapers and publishing houses. " (from LE3-4239 SCARRIE Project Programme. Annex I.)

The SCARRIE demonstrator for Swedish was developed by the Department of Linguistics at Uppsala university. Fundamental user input was provided by two Swedish newspapers, Svenska Dagbladet and Upsala Nya Tidning. We gratefully acknowledge their contributions. See further http://www.scarrie.com/ for general information about the project and http://stp.ling.uu.se/~ljo/scarrie-pub/ for a test version of the resulting Swedish SCARRIE pilot.

The achievements made in the SCARRIE project were continuously delivered to the European Commission via the project officer. When the project was still running the availability of most of the deliverables was restricted. Now the project is closed, and we make some of them available to a larger community. They all concern the development of the Swedish SCARRIE pilot.

An important part of a language checker is its dictionary. The SCARRIE dictionary for Swedish was based on large scale corpus studies of newspaper text provided by the two newspaper users. In this issue the primary investigations of the newspaper corpus underlying the dictionary is presented. The delivery of the report was made in May 1998. This issue also comprises a paper on the distribution and frequencies of single characters in the corpus as well as sequences of characters, so-called n-grams.

For reports on other aspects of the SCARRIE project, see Working Papers in Computational Linguistics & Language Engineering No. 3 - 5 and No. 7 -13.

Contents