SCARRIE LEXICAL RESOURCE, LMF COMPLIANT VERSION

The original SCARRIE lexical resource was developed for use in automatic proofreading of Norwegian Bokmål (nob). The current LMF compliant version has been derived from the original lexicon by Koenraad De Smedt at the University of Bergen in the context of the the META-NORD project, with the aim of making the resource easier to share and reuse.

Norwegian Bokmål has a considerable number of alternative forms. The SCARRIE lexical resource is to our knowledge the first and so far only Norwegian lexical resource with information about the 'style' or 'subnorm' that each wordform belongs to. It was designed to help a proofreading system find the alternative which fits best in the chosen overall style or subnorm of a text. SCARRIE has not taken into account official changes to the spelling after 1998 and this resource has not been fully manually checked for correctness.

LICENSE

The authors and institutions to be credited in connection with the SCARRIE lexical resource are: Victoria Rosén and Koenraad De Smedt at Universitetet i Bergen, and Torbjørn Nordgård at Norges Teknisk-Naturvitenskapelige Universitet.

FORMAT

The SCARRIE lexical resource is basically a full form lexicon where forms belonging to the same lemma are grouped together, but the lemma itself does not carry any information.

This version has been converted from the original IDF files to a file compliant with the Lexical Markup Framework (LMF, see http://www.lexicalmarkupframework.org/). It has been validated against DTD_LMF_REV_16.dtd. It consists of one file with a single LexicalResource containing seven Lexicon elements with the following names:

- prefixes
- suffixes
- gramwords (grammatical words)
- gramwords2x (more grammatical words)
- abbrevwords (words occurring in abbreviations)
- idiomwords (words in multiword expressions)
- main (normal open class words)

- WordForm count in prefixes.....327
- Lemma count in prefixes 327
- WordForm count in suffixes 562
- Lemma count in suffixes 125
- WordForm count in gramwords 707
- Lemma count in gramwords 539
- WordForm count in gramwords2x.....39
- Lemma count in gramwords2x 8
- WordForm count in abbrevwords 23

- Lemma count in abbrevwords 23
- WordForm count in idiomwords 811
- Lemma count in idiomwords 811
- WordForm count in main.....359684
- Lemma count in main 72617
- WordForm count, total.....362153
- Lemma count, total.....74450

Each LexicalEntry has an empty Lemma and at least one WordForm. Normal WordForm elements have elements Feat with the following att: writtenForm, corrStyle, featureList, replacement and synCat. Special entries, in particular prefixes and word forms occurring in abbreviations or idioms neither have replacement nor synCat. The use of the attributes is explained in SCARRIE deliverable 3.3.1 Tagset.

ABOUT THE SCARRIE PROJECT

SCARRIE was an RTD project (LE3-4239) in the Language Engineering sector of the Telematics programme of the European Union. The project began on Dec. 1, 1996 and was concluded on Feb. 28, 1999. The coordinator of the project was WordFinder Software AB (Växjö, Sweden). The other main partners in the project were Universitetet i Bergen, Institutionen för lingvistik at Uppsala Universitet, Center for Sprogteknologi (København) and Svenska Dagbladet (Stockholm).

The aim of the project has been to build proofreading tools for Danish, Norwegian and Swedish. In order to achieve its goals, SCARRIE has researched effective error detection and correction mechanisms for the Scandinavian languages. Resources for these languages have been integrated in the CORRie platform, which was originally developed for Dutch by Cognitech. The prototype proofreading system provides several linguistically motivated error detection and correction mechanisms at both word level and sentence level.

The work for Norwegian was coordinated at the University of Bergen. The chief researcher on the project was Victoria Rosén. The scientific coordination was done by Prof. Koenraad de Smedt.

The Norwegian part of SCARRIE has been aimed at advanced spelling correction in Bokmål. It uses word form dictionaries in combination with special mechanisms for handling multi-word expressions and for recognizing newly seen compounds, proper names and other words not present in the dictionaries. In cooperation with NTNU, a suitable Norwegian word form dictionary has been built. The word forms in this list are tagged with information about lemma (basic form), standard, style or written norm, morphosyntactic characteristics and possibly replacement. Predictable misspellings are supplied with recommendations for corrections.

New compounds are detected by an analysis based on rules supplied by the University of Oslo. Words that are outside the scope of the dictionary and are likely errors are processed by the correction mechanisms including sound-based similarity. In addition, a robust grammar was developed for the detection and correction of certain classes of errors which cannot be handled at word level, i.e. agreement errors. Finally, suggestions for correction are chosen so as to fit in the written norm which the document is written in (on a range from conservative to radical Bokmål).

More information of the project has been archived at
http://ling.b.uib.no/projects/scarrie/