Erla Hallsteinsdóttir & Uwe Quasthoff & Sabine Fiedler – poster
Syddansk Universitet i Odense, Universität Leipzig
erla@sdu.dk; quasthoff@informatik.uni-leipzig.de; sfiedler@uni-leipzig.de

Frequency Dictionaries as a tool for analog and digital dictionary user

The Leipzig Corpora Collection at the University of Leipzig contains corpora in many different languages. The collection aims at producing text corpora with comparable empirical linguistic data for a large number of different languages, e.g. for use in linguistic research and language technology. Since 2011, the corpora have been used to compile frequency dictionaries and dictionaries for German, Icelandic, English, French, Hungarian and Esperanto have already been published in the series Frequency Dictionaries at Leipzig Universitätsverlag, more languages are in preparation.

The dictionaries are available in print and electronic version. The printed version provides the most frequent 1,000 word forms in the corpus ordered by frequency and the most frequent 10,000 word forms in alphabetical order. The enclosed CD-ROM contains a more comprehensive version of the dictionary as an e-book, with data on relative frequency of up to 1,000,000 word forms given in alphabetical order. The number of word forms for a particular language depends on the size and composition of the corpus used. This list of words, with frequency classes, is also available as a plain text file on the CD-ROM ordered alphabetically as well as ordered by frequency. The series Frequency Dictionaries aims at producing comparable linguistic information, e.g. frequency data on word forms, for a large number of different languages.

The dictionaries can be seen as a standalone tool, but they can also be used as an integrated part of the linguistic tools and resources that the NLP-group at the University of Leipzig has developed.

The potential users of the dictionaries are both analog and digital (human and machine), as an example, the electronic version gives access to data for use in language technology and research projects.

With examples from the Icelandic and the Danish frequency dictionaries, we will present the concept and the content of the series. Then we will discuss different usage of the dictionaries in the context of the thematic approach of the conference: the digital dictionary user.

http://corpora.informatik.uni-leipzig.de/
http://univerlag-leipzig.de/catalog/category/142-Frequency_Dictionaries_Haeufigkeitswoerterbuecher