Benedict: an EU Project for an Intelligent Dictionary

Mika Herpiö

Benedict is a large development project for an intelligent dictionary, which started as part of the second last call of the EU's 5th Framework IST (Information Society Technologies) research program. The driving force behind the project is the language technology firm Kielikone, also known for the new MOT GlobalDix multilingual dictionary that is based on the Kernerman Semi-Bilingual Dictionaries series. Kielikone is responsible for the software development and coordination of Benedict. Also involved are the University of Tampere, Gummerus Publishers and Nokia from Finland, and the University of Lancaster and HarperCollins Publishers from the UK.

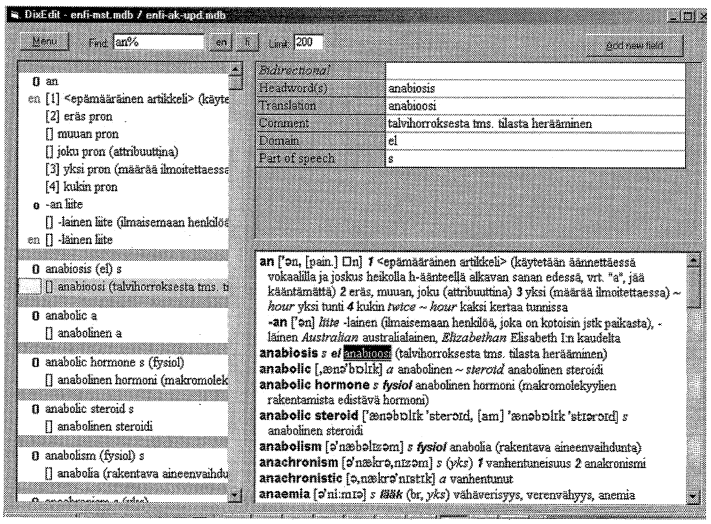
Benedict attempts to break the mental barriers caused by the tradition of the dictionary as a printed product, which still restrict its development in the electronic era. One such obstacle has been the notion that dictionary entries can appear only in one form: the one in which they are printed. The new Benedict dictionary will no longer look the same for each user – it will adapt to different users, even to different texts the users work with. How can such an adaptation be obtained so as to truly benefit the user? This is the challenge about to be solved during the three years of the project.

The Benedict product will provide an interactive user-specified access interface that tailors the dictionary content to user specifications, multi-layered entry structure, links to corpus data, and syntactically- and semantically-based corpus search tools in the dictionary database. Benedict is particularly aimed to cater for the demands of the multilingual corporate world. It will also transform the approach to economic space consumption, which has restricted the amount of data that can be accessed through a dictionary.

The new Benedict dictionary will not be created from scratch. The basis for its software development is MOT 3.0 – Kielikone's current dictionary engine that has been on the market (Windows, intranet/Internet, mobile) for more than 10 years. The new features will be integrated gradually in future editions of MOT. Since Kielikone and the University of Lancaster have considerable experience in language technology applications, the Benedict project will be able to employ state-of-the-art HLT (Human Language Technology) components in creating the final product. One of these components is the semantic tagger developed in Lancaster, which might have a vital role in directing the user to the relevant information in the dictionary entry. Kielikone has already developed also parsers, lemmatizers and other HLT solutions for the Finnish language, which will be applied in this project as well.

Significant sub-projects of Benedict include dictionary projects by Harper-Collins and Gummerus, in which content will be developed especially, but not entirely, for electronic use. The University of Lancaster has another big project to further develop their English semantic tagger and to develop a Finnish semantic tagger in collaboration with the University of Tampere – as an experiment for developing semantic taggers for more languages.

Kielikone plans to develop by-products in the project, like a system for updating the dictionary content, which applies user logs of the web dictionary. Perhaps more interesting for the dictionary community will be DixEdit, a structured content editor especially developed for lexicographic content. This will enable XML output of the dictionary, while the user doesn't have to input a single tag herself/himself. DixEdit shows the content in WYSIWYG view during the editing process, and the validity of the data is automatically surveyd all the time. The first version of DixEdit is available for Windows 98/2000/XP (a sample screen appears below).



Mika Herpiö graduated from the Helsinki University of Technology. He is the director of business development and a partner in Kielikone Oy.
mika@kielikone.fi

Kielikone is the leading Finnish language engineering company. It develops generic linguistic modules as a basis for language technology products such as dictionary software, machine translation software, terminology management software, and spell and grammar checkers for Finnish, on Windows, Internet/intranet, and mobile platforms.

http://mot.kielikone.fi/benedict