The Norwegian Human Language Technology Resource Collection

Torbjørg Breivik, The Language Council of Norway

Språkrådet

Background

- eNorway Action Plan, version 1.0, spring 2000

- The Norwegian Research Council: Action Plan for HLT Research, autumn 2000 – and the KUNSTI programme, 2001-2006

- The Norwegian Language Council: Action Plan for Norwegian Language and ICT, February 2001

- Prospect for a Norwegian HLT resource collection, October 2002

Språkrådet

The Consortium

- Nordisk Språkteknologi Holding AS went bankrupt in 2003. Among other resources the company had some high-quality acustic databases

- Three universities (Bergen, Oslo and Trondheim), one commercial company (IBM) and a governmental council (Språkrådet) bought the estate

- The language resources (especially the acoustic databases and the lexicons) constitute a starting point for a Norwegian HLT resource collection

Språkrådet

Distribution has started

- The consortium has started distribution of the copyright checked resources
- The resources has not been validated to see to what extend they meet todays quality demands (- must be done!)
- The resources has been distributed as is to researchers and commercial companies
- *Aksis* in Bergen has been engaged to carry out this assignement as this company has the essential expertise
- The Norwegian Language Council speaks for the consortium, handles the administration and is responsible for outlining standard agreements

Torbjørg Breivik, 15.-16.12.2008

The Norwegian HLT Resource Collection

- The white paper "Mål og meining" stated that The Norwegian HLT Resource Collection should be established from 1 January 2009

- Some money allocated in the Budget for 2009 (more indicated)

- A business plan has been made – still some open questions from the Ministry for Culture and Church Affairs

- A board has to be appointed by the Ministry (HLT-experts and -users)

- The board has to decide and give priorities

- Limited administration: identify materials, chech copyrights, sign licenses, pass on accepted material for quality control and validation, update available resources, national and international relations, etc.

Språkrådet

Torbjørg Breivik, 15.-16.12.2008

The Resource Collection

- Purpose: a resource collection for human language technology

- Users: researchers *and commercial companies*

- Owner: A Norwegian State property represented by The Ministry for Culture and Church Affairs

- Contents validated and produced according to international standards (The BLARK concept)

- Tasks to be outsourced to relevant existing organizations

AND: cleared copyright issues!

Språkrådet

Time frame and priorities

- Compilation within six years

- Listed demands for specific resources from the HLT industry and research

- All data to be incorporated must be validated according to international criteria and standards (the BLARK and CLARIN initiatives) by an independent institution

- The project plan specifies minimum and recommendable contents

- The board will adjust priorities and work schedule when the project is launched

- Price: NOK 90-100 millions

The HLT Resource Collection and CLARIN

- Goals: both intend to establish infrastructure of language resources and its technology, but the HLT resource collection for language technology purposes only

- Users: The HLT Resource Collection for research and industry - CLARIN for research only

- Copyrights: easier for research purpose than for both research and industry

- Important: Avoid doing the work twice: contents available in The Norwegian HLT Resource Collection are available to all kind of researchers – not only for those working in the HLT field