
The *LICHEN* Project: The Linguistic and Cultural Heritage Electronic Network

Lisa Lena Opas-Hanninen

(*lisa.lena.opas-hanninen@oulu.fi*)

University of Oulu

Jean Anderson (*j.anderson@arts.gla.ac.uk*)

University of Glasgow

Ilkka Juuso (*ilkka.juuso@ee.oulu.fi*)

University of Oulu

Tapio Seppänen (*tapio@ee.oulu.fi*)

University of Oulu

The international, interdisciplinary and multilingual LICHEN project focuses on the languages and cultures of the northern circumpolar region, that is the region north of the 55th parallel. Its underlying assumption is that language and culture are as important to the survival and well-being of populations as more obvious ecological, social and health issues. We believe that the creation of a digital portal giving access to written and spoken texts in the languages of the region will further its well-being.

Faced with minority languages, governments in the recent past have pursued policies of assimilation. This has applied to indigenous languages in Canada, to Gaelic and Scots in Scotland, and to Finnic minority languages in the Circumpolar region: Meänkieli and Swedish Finnish in Sweden; the Kven language in Norway; Viena Karelian, Olonets Karelian and Vepsian in Russia; the Võro and Seto languages in Estonia; and Livonian in Latvia.

LICHEN aims to collect and disseminate information about the languages spoken in the circumpolar region in order to promote the linguistic confidence and self-image of their speakers. It will promote cultural awareness among the peoples of the North, facilitating cross-cultural communication between them in an age of rapid global change. LICHEN will create communication between research units in order to promote discussion on the common needs of research on the minority languages of the North. We are doing this by:

- creating an electronic framework for the collection, management, online display, and exploitation of corpora of the languages of the circumpolar regions;
- creating a website with information on these languages and the peoples speaking them (the LICHEN website will be launched in January 2005);
- creating a virtual learning environment for teaching the linguistic and cultural heritage of the circumpolar region;
- carrying out a pilot project on the Meänkieli and Kven languages;
- identifying research on topics of immediate importance and common interest;
- setting up an inter-institutional doctoral research programme.

LICHEN has existing resources and work has begun. We have Meänkieli tapes totalling about 150 hours and Kven language tapes of about 100 hours for our immediate use. These tapes are now being digitized. We have access to both the structure and contents of the *Scottish Corpus of Texts and Speech* (SCOTS) at the University of Glasgow, currently totalling 0.5 million words of spoken and written Scots and Scottish English. We have the nucleus of a research team based on the English and Finnish Departments and the Department of Electrical Engineering at Oulu and the English Language Department and SCOTS project at Glasgow. This team has considerable linguistic and computing expertise.

Our first aim in year 1 is to complete the technical specifications for the electronic framework through consultations between language and computing staff team members. We are also consulting other people working in the field of corpus building at meetings and conferences, and by email. In addition to housing the data, the system will accommodate management, administration and programs for concordancing and searching the data. The ultimate goal of the development of the computing tools is a shell which can be adapted to any language. For many languages at risk there is a need both to preserve existing materials and develop new ones. During this year, we will design and implement a prototype of this shell. A longer term goal is to provide an interface to the tools which allows the end user to define or rename all functions in their own language.

We will continue work on the Meänkieli and Kven recorded language material. We will work generally on the problem of languages without standard written forms, starting with Kven and Scots (a worldwide problem as people endeavour to record languages before they vanish). As an initial solution to the problem of written forms, it is proposed that several Kven speakers should be asked to transcribe a short passage and the results compared. Scottish Language Dictionaries will be consulted here. We will investigate the feasibility of working through community groups in minority language areas. In addition to harnessing local knowledge, we hope that a policy

of local workshops will stimulate skills development and job creation.

A poster presentation at ACH/ALLC 2005 will enable us to publicise the project to other minority language scholars and to enlist the considerable expertise of the conference participants in the discussion of the design of the online corpus tools. It is our intention that the functions of the corpus shell should include all the basic requirements for a corpus builder and for a corpus user in an easy-to-use environment. We know the end users will include many who are not technically sophisticated and would not have other avenues for finding advice on digitization or access to an Internet platform to share their materials. Our idea of 'basic' requirements for online use include corpus browsing, word and phrase searches, wildcard searches, concordancing; for corpus building we will include guidelines on recording, digitization, copyright and data protection. We would welcome this opportunity to discuss our proposed designs with the expert community of ACH/ALLC.

Bibliography

Anderson, J., et al. "The SCOTS Corpus." *Models and Methods in the Handling of Unconventional Digital Corpora*. Houndsmills: Palgrave Macmillan, Forthcoming.

Institute for the Languages of Scotland. <<http://www.arts.ed.ac.uk/celtscot/institutelanguagesscotland/>>

Kven bibliography. <<http://www.ub.uit.no/baser/kvensk/>>

Linguistic Atlas Projects. <<http://us.english.uga.edu/>>

The Linguistic Data Consortium. <<http://www.ldc.upenn.edu/>>

MediaTeam Oulu research group. <<http://www.mediateam oulu.fi/brief/?lang=en/>>

Meänkieli information. <<http://modersmal.skolutveckling.se/meankieli/index.html>>

The Newcastle Electronic Corpus of Tyneside English. <<http://www.ncl.ac.uk/necte/>>

Opas, L.L., and F.J. Tweedie. "Review of Michael P. Oakes, Statistics for Corpus Linguistics." *Literary and Linguistic Computing* 14.4 (1999): 541-543.

Palander, M., L.L. Opas-Hänninen, and F.J. Tweedie. "Neighbours or enemies? Competing variants causing differences in transitional dialects." *Computers and the Humanities* 37.4 (2003): .

Ruija kvenmuseum. <<http://museumsnett.no/alias/HJEMMESIDE/vadsomuseet/kven/>>

Ruijan Kaiku newspaper. <<http://www.ruijan-kaiku.no/>>

Scottish Corpus of Texts and Speech. <<http://www.scottishcorpus.ac.uk/>>

TAPoR tools. <<http://tapor.humanities.mcmaster.ca/home.html>>

The Text Encoding Initiative. <<http://www.tei-c.org.uk/>>

Thule Institute. <<http://thule oulu.fi/>>

University Centre for Computer Corpus Research on Language. <<http://www.comp.lancs.ac.uk/computing/research/ucrel/>>

Winsa, Birger. "Language attitudes and social identity. Oppression and revival of a minority language in Sweden." *Applied Linguistics Association of Australia Occasional paper* 17 (1998).