L2/01-143

ISO/IEC JTC 1/SC 2/WG 2 **N 2340**

Encoding of Swedish dialect transcripts

The aim of this project is to develop methods and standards for data entry and encoding of dialect transcripts. Though independent, the project is associated with other Swedish dialect documentation projects.

Background

The archives of Swedish dialect research institutions contain large collections of recordings and transcripts. Most of these transcripts are written with an unique phonetic alphabet, the Swedish dialect alphabet, specially designed in 1878 by Professor J. A. Lundell of Uppsala. Only a limited number of scholars are able to read these transcripts, which contain a large amount of information of both linguistic and ethnological interest. To make these dialect collections accessible to the general public, various kinds of software tools are needed for the encoding, conversion, search, and display of these texts.

The project

The project is divided into three subprojects: Encoding tools, Character codes, and Text conversion.

1. Encoding tools

This subproject involves the creation of detailed standards for dialect recording transcript formats and the development of software and other tools for entering dialect texts.

• Making an XML schema for dialect transcripts, containing tags for various types of metadata.

• Collecting information about existing software tools - fonts, keyboard modifiers, OCR software etc.

• Development of new software where existing software isn't sufficient.

• Creation of manuals and other types of training material for the persons carrying out the actual work of entering, importing, and cataloguing dialect transcripts.

2. Character codes

This subproject involves the documentation of the phonetic values of all the characters of the Swedish dialect alphabet and including these characters in the ISO 10646/Unicode character code standard.

• Investigating actual character usage in different dialects and various types of dialect texts. This phase should include cataloguing the usage of similar dialect alphabets in Denmark and Norway.

• Making translation tables between the Scandinavian dialect alphabets and the International phonetic alphabet (the IPA alphabet).

• Making a proposal for including the characters of the dialect alphabets in ISO 10646/Unicode, the international character code standard.

• Making Unicode-encoded fonts for free distribution, containing all the Scandinavian dialect characters.

3. Text conversion

This subproject contains the development of sets of rules for conversion between various orthographic systems, e.g. conversion from the Swedish dialect alphabet into IPA or 'phonetic spelling', i.e. ordinary alphabetic characters indicating the pronunciation. There will probably be a large number of rule sets, due to phonetic differencies between the dialects and the demand for different levels of accuracy.

ə áə, ȷa míns mın fár, dæ va mın fá‿ʂʉm hàde ə gáɖən dá. han sadə ǽʈèr ən fjòʈənə máʂ, ɑ ə dæ ȷık ȷu brá ... dı fık ȷu. vɑra ǽldrə sadə dı sàdə altı tìdıarə æn dı ȷø‿ɳú fɵ‿ʈín. ... sı dı hàdə san ʌèʂkap, ʂa dı va tvùgnə te ɑ ɑ bòrȷa nàgərlùnda, sa dı kʉnə hına mǽ. nu há‿ɖı san ʌèʂkap ʂa nú ə ... mæ traktórər ʂá sa dǽ dæ dæ gá‿ɖə mykə ʌàskarə ...

ə ɒːˊə, ɟɒ miˊnːs min fɒːˊr, dɛ vɒ min fɒːˊ ʂœm hɑˋdːe ə goːˊɖən doːˊ. han sodə ɛˋʈːeˌr ən fɟʊːˋʈənə mɑˊʂː, o ə dɛ ɟɪk ɟʉ brɒːˊ ... di fɪk ɟʉ. voːrɑ ɛˊlːdrə sodə di soˋdːə ɑlːti tiːˋdiɑrə ɛn di ɟøː ɳʉˊ fœ ʈiːˊn. ... si di hɑˋdːə son ʁeˑˋʂkɑp, ʂo di vɒ tvəˋŋːnə te o o bœˋrːɟɑ noːˋgərləˌnːdɑ, so di kʉnə hinːɑ mɛːˊ. nʉ hɒːˊ ɖi son ʁeˑˋʂkɑp ʂo nʉːˊ ə ... mɛ trɑktʊːˊrər ʂoːˊ so dɛːˊ dɛ dɛ goːˊ ɖə mykə ʁɑˋsːkarə ...

e ae, ja minns min far, dä va min far sôm hadde e gården då. han såde ärter en fjortene mars, å e dä jik ju bra ... di fik ju. våra älldre såde di sådde allti tidiare än di jör nu för tin. ... si di hadde sån rerskap, sjå di va tvungne te å å börrja någerlunnda, så di kune hinna mä. nu har di sån rerskap sjå nu e ... mä traktorer sjå så dä dä dä går de myke rasskare ...

Figure 1. At the top, a transcript written with the Swedish dialect alphabet; in the middle, the same text automatically converted to IPA characters; at the bottom, the same text automatically converted to Swedish 'phonetic spelling'.

Uppdaterad 2001-03-10