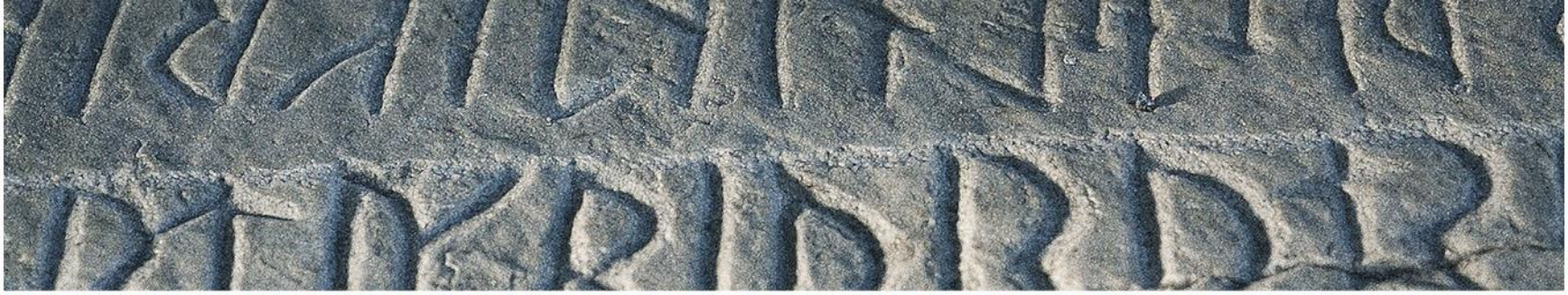

UiO : Department of Linguistics and Scandinavian Studies
University of Oslo

Discussion Workshop

Subjects: Making a speech corpus. Tools, computer software, field work, transcription. Installation of software on individual computers as required.

Kristin Hagen, Anders Nøklestad, Joel Priestley,
The Text Laboratory, University of Oslo
kristin.hagen, anders.noklestad, joel.priestley @iln.uio.no

tekstlab

Introduction of participants

- Name
- Project
- Language
- Main project aim
- Fieldwork
- Transcription

Corpora in the Norhed project *Linguistic Capacity Building*

- Promised in the application:
  - 4 text corpora
  - 5 speech corpora
- Corpora – especially speech corpora are expensive to develop = make the corpora:
  - Reusable for other researchers and projects
  - Useful for more than one goal (dictionaries, wordlists, grammars, syntactic research etc.

Building a speech corpus

- Free speech –
  - interviews ((semi)formal setting)
  - conversations (informal atmosphere)
- Audio and video if possible
- Various topics
- Representative
  - Both young and older
  - Both women and men
  - All social classes
- Size: as big as possible

Information to and from informants

- Information sheet (about the project and how the data will be used)
- Consent form
- Informant form (data about age, gender, parents, linguistic and/or multilingual situation etc.
- List of discussion topics
- Gifts

Transcription

- Use Elan
  - Transcription is fast
  - Connects audio/video and transcription automatically
  - Can be imported in to Glossa
- Two or three transcriptions (tiers)
  - One phonetic (NB not too fine-grained)
  - One orthographic or/and English

Elan transcriptions in the search tool Glossa:

- Important that word specific attributes in subtiers are aligned one to one (positionaly) to make all the information from the different tiers about the particular token available in Glossa
- (Exception: free translation in English)

Transcription: Budget

- The project has a generous transcription budget. Use it ☺

- (Part III,13c):
  - Transcription of speech data: 40 hours recordings x factor 20, 6 languages
    = 4800 hours
  - Translation of speech corpora: 4800 hours
  - = 9600 hours transcription and translation = 60 months

Planning a transcription project

- Hire one or two experienced students
- Teach them how to transcribe the recordings
- Make detailed transcription guidelines – alone or in cooperation with the transcribers
- Hire more students for transcription with the most experienced students as group leaders
- NB Proofread each others transcriptions to get consistent transcriptions and results

Glossa demonstration

Future work

- Record and transcribe as much as possible
- Create recording and transcription guidelines and consent forms
- Elan training workshop
- Make it as simple as possible first – work incrementally, adding more tiers later as required
  - Orthographic first and phonetic later? The other way around?
  - POS tagging/parsing later
- Agree on metadata schema and format (Addis and Oslo)
- Make mp4 and wav format of video and audio files
- Make corpus generating tool for Glossa (Oslo)
- Adding more features to Glossa (login, collocations, deleting and saving hits)
- One or more demo corpora ready for workshop in November?