TillTal

Accessible Cultural Heritage for Speech Research

Johanna Berg, Rickard Domeij, *Jens Edlund*, Gunnar Eriksson, David House, Zofia Malisz, Susanne Nylund Skog, Jenny Öqvist

TillTal

- New project (proposed and funded 2016; starts 2017)
- Funded by Riksbankens Jubileumsfond
  - One of Sweden's largest funding agencies for the humanities
  - ~ 1 MEUR over 3 years
- Involves researchers from several disciplines, e.g. spoken language & conversation; linguistics & phonetics; human-machine interaction; anthropology; speech technology…
- Collaboration between 3 SWE-CLARIN members

SWE-CLARIN

The SWE-CLARIN working group for speech

- Informal special interest group for speech
  - Currently 3 members
- KTH Speech, Music & Hearing (KTH; CLARIN-SPEECH K-centre):
  - Research: Spoken language; spoken human(-human/machine) interaction
  - Research: Speech technology
- The Institute for Language and Folklore (ISOF)
  - Research: anthropology; linguistics
  - Research: archives, methodology
  - Archives: e.g. 13000 h of speech
- The Swedish National Archives (DIGISAM)
 - Digitalization, information, collaboration
  - (Huge) archive

SWE-CLARIN

Background I: "found speech"

- 2015 survey (by KTH, on behalf of the the government and the Swedish Post and Telecom Authority):
  - Vast quantities of speech data, but
  - intellectual property ownership,
  - privacy and integrity, and
  - legal issues relating to the (often unknown) content of the recordings, so:
  - virtually none of it is used in research (of any kind).

- Interspeech
  - 2015: few papers on "found speech"
 - 2016: about a dozen papers
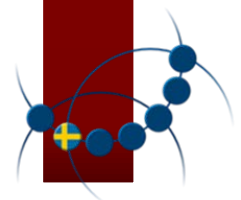
SWE-CLARIN

Background: CLARIN, speech and SSH

- Active effort to find and start new speech technology+SSH projects
- Swe-Clarin and the spoken language I – Research collaborations between *resource holders, speech technologists, and researchers within the human and social sciences* (2015)
- Triplet brain storming
  - Through a SSH researcher, a data holder and a speech technologist together
  - Task: come up with project focussing on the SSH researcher's research, using the data holder's data and the speech technologist's skills
- Successful: TillTal first to materialize.

SWE-CLARIN
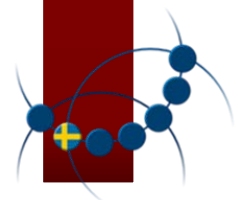
Opportunity calls!

- Riksbankens jubileumsfond
  - Funds humanities research
  - Early 2016: a second call over "Research and the collections"
  - Claims of low quality in the first call ☺

- Speech research + speech collections:
 - A perfect fit!

- Just add SSH research.
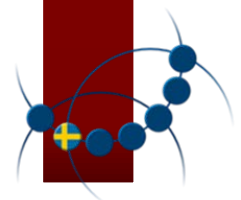
Project outline

- Top-level objective: evaluate the use of modern speech technology methods to access large "found" speech data
  - Quantify errors
  - Develop tweaks (and possibly some new methods)
- Test cases: 3 full-scale, research projects from different SSH disciplines
  - Avoid home-grown potential usage and go for real research projects!
  - Sacrifice generic solutions to avoid "nail syndrome" (thank's Sally!)
  - Spend out on custom-designed solutions that can be evaluated, quantified and verified on their specific task: gain validity.

SWE-CLARIN
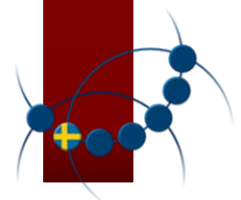
(1) From stories to cultural heritage

- Investigates a collection created by Karl Gösta Gilstring (1915-1986),
  - consisting of 8000 original letters,
  - and 250 hours of recorded speech (mainly interviews made by Gillstring).
  - (Largest Nordic collection by a single researcher in modern time.)
- Original SSH research.
- Also: will act as a use case for speech technology methods that provide new entry points to the speech data and relate these to the texts.

SWE-CLARIN
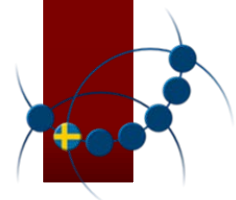
(2) Linguistic variation in time and space

- Investigates linguistic variation and change in speech materials.
- Currently often captured by painstaking annotation, e.g.
  - all occurrences of a specific vowel.
- Will test and evaluate different methods of finding these occurrences automatically.
  - Allows for larger data sets by orders of magnitude.
  - Frees up the researchers' labelling time for other, more challenging pursuits.

SWE-CLARIN

(3) Interaction patterns over time and type of conversation

- Relation between different contributions to the flow of conversation:
  - Who speaks when?
  - How are the speaker changes managed?
  - How is common ground achieved?
  - How are attitudes towards the spoken signalled?
- Interaction models based on chronograms – simple diagrams showing segments of speech and silence over time – have proven a successful approach to study these kind of questions.
- Generally, these models are built automatically, but how well the methods used work on archive data is unknown.

SWE-CLARIN
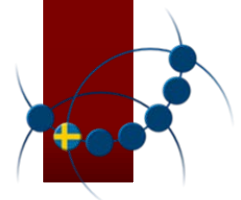
Thank you!

Qustions? (Several of us are here!)