SCRIBE: Machine transcription of Norwegian conversational speech

SCRIBE: Machine transcription of Norwegian conversational speech

- Collaborative project – partners are NTNU, NRK, Telenor and Nasjonalbiblioteket
  - RCN/IKTPluss project
  - 27MNOK total budget (+NTNU-financed PhD), 12MNOK RCN contribution
- Main goal: Develop transcription system for Norwegian speech, with particular focus on conversational speech (2 or more speakers)
- Secondary goals
  - Robust models for syntactic phenomena in natural speech
  - Handling of dialects and generation of orthography close to speech
  - Develop metrics for quality in transcription systems

SCRIBE – some challenges

- Large dialect variation
  - dialect words vs dialect pronunciation (accent)
  - scarcity of available speech data
- Two variants of Norwegian text – with high degree of legal variation in orthography and inflictions/conjugations
- Relatively small data sets available
  - text: English ~1 trillion words (Google 2006), Norwegian ~900 million, mainly *bokmål*
  - Speech: English ~125 000 hours (Google), Norwegian < 1000 hours
    - 125 000 hours = 14,2 years…
- Must develop data efficient learning algorithmsr

and some applications

- real-time subtitles for hearing impaired
  - meetings
  - streamed content
  - broadcast TV
- instantaneous meeting minutes
- data-assisted language learning able to handle dialects and accents
- meeting transcriptions – Parliament, court proceedings,++
- Efficient production of new language resources
- Transcription for information retrieval and generation of metadata
- Transcription and analysis of conversations in customer centers, support telephonee etc.

Some other projects

- **NordTrans** - Technology for automatic speech transcription in selected Nordic languages (EEA NO/CZ: NTNU/TU Liberec/Newton Tech)
  - Utvikle forbedret teknologi for norsk og svensk transkripsjon av strømmet audio.
    - Basert på eksisterende flerspråklig system
    - Forbedre ASR-teknologi, og tilpasse til skandinaviske språk
    - Støtte fra NRK
- **TEFLON** - Technology-enhanced foreign and second-language learning of Nordic languages (NordForsk: Aalto, Univ Helsinki, Karolinska, UiO, NTNU)
  - Tverrfaglig gruppe: Målgruppe fremmedspråklige barn som skal lære et nordisk språk
  - Dataspill-basert opplæring. **Taleteknologi for barn**, utvikling og hjerneforskning for språkopplæring, spill og taleterapi
- **eLadda** – Early Language Development in the Digital Age (EU MSCA, NTNU leder, 16 deltakere, 13 PhD-stipendiater)
  - Tverrfaglig prosjekt
  - Vårt fokus: Talegjenkjenning for barn