NTNU
Innovation and Creativity

Sharing speech data
- notes from a small language

CLARIN and FLaReNet workshop
KTH, Stockholm, November 25 and 26, 2009

Ingunn Amdal
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

Main issues in sharing data

The user of the data (researcher, developer, ...) must

- Be aware of that the data exists
- Have access to the data (legally and economically)
- Be able to use the data (formats)

CLARIN (and FLaReNet) the solution — hopefully

What's special with speech?

- Large amounts of data
  - Multimodal corpora - even more data
    Should always include video for face-to-face dialogues
- Expensive to collect and annotate
- (Most) available speech data collected for speech technology purposes
- Existing organization for sharing data:
  - LDC
  - ELRA/ELDA
  - National language banks
- Informants
- Spontaneity

Speech issues (1)

- Companies and universities: different types of data and different interests - may still benefit from sharing data

- Heterogeneous types of data: same infrastructure?
  - Large speech technology databases
    e.g. Broadcast News data 100+ hours with
    complicated annotation and several parties (legally complicated)
  - Large linguistic research databases
    e.g. Scandinavian Dialect Corpus
  - Smaller linguistic research databases
    e.g. word tone - still interesting to share
  - Companies collecting data for speech technology applications
    may want to share if the infrastructure is there

Speech issues (2)

- Informants and spontaneity:
  - right to privacy and intellectual property
  - manuscript read speech easier - agreement with the informant
  - want spontaneous speech - more difficult
  - in-between: interviews and moderated dialogues
- Who should bear the "burden" of organizing, settling of legal issues, and maintenance needed for sharing?
  - Guidelines to help organizing, including legal issues
- Requirements to share to get funding?
  1. Minimum template to be registered in CLARIN
  2. Medium template to get public funding
  3. Maximum template to provide a structure for those who want to share information

Norwegian language bank

- BLARK reports in 1999, 2002, and 2008
- Formally started in 2008
- Now a part of the national library
  (but no separate budget or board …)
- NST-data secured and made available
- No money to collect more data, yet

Numbers from 2002 on Norwegian BLARK:

- Total of 100 mill NOK
- About 50 mill for speech data

Need for Norwegian speech data

Norwegian language bank report 2002 identifying needs:

- Telephone speech
  - Manuscript read (also for mobile phones): 400 hours
  - Spontaneous: 200 hours
- Studio quality
  - Manuscript read: 5-600 hours
  - Spontaneous: 500 hours
- Audio
  - Spontaneous: 100 hours
- Multimodal data: for later
- Multilingual data: for later

Existing Norwegian speech data

- Some data already available through ELRA:
  - SpeechDat (II), 65 hours manuscript read telephone speech over fixed telephone network, 1016 speakers
- NST data available through Norsk Språkbank:
  - Dictation data: app 500+50 hours (1000+100 speakers)
  - Telephone speech: 1000 hours (8000 speakers) of Norwegian
  - Speech synthesis database: 15 hours (1 male speaker)
- Data at the universities:
  - Varying availability
- Data at companies:
  - Varying availability

Small language issues

- Reports on “problems” on how to select from too much data at Interspeech this year
- We are NOT anyway near such problems
- Still lack of data – therefore motivated to share
- ”Same” interest as Humanities and Social sciences

- CLARIN must deal with less interest from larger languages:
  ”what’s in it for me”

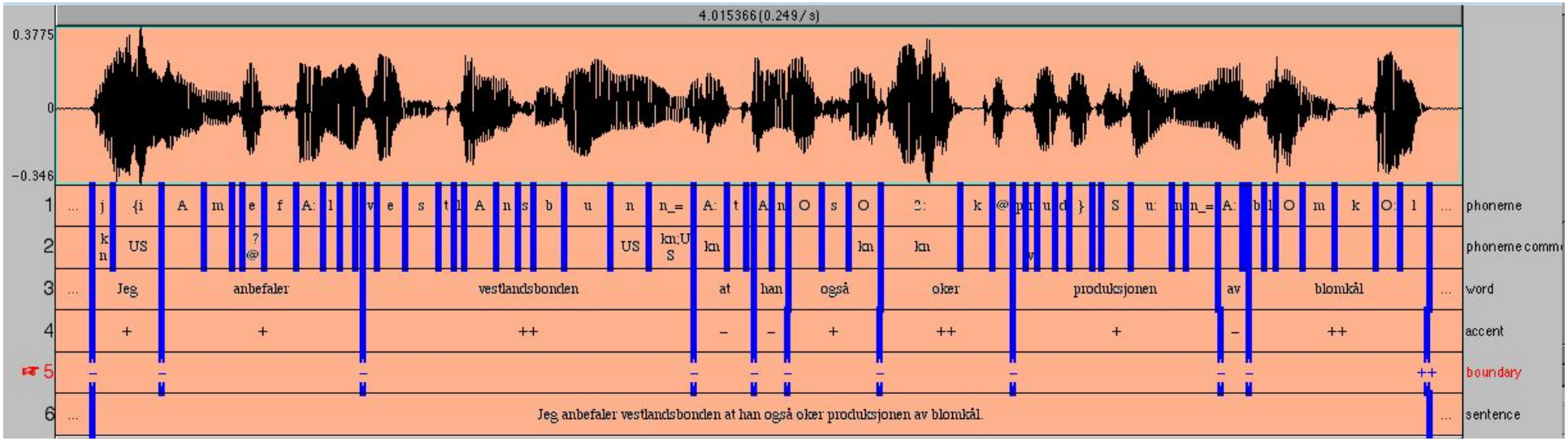
Suggestions

- Fortunately: Speech technology databases in ELRA will be available in CLARIN through ELRA (I hope)
- Suggestion: Target smaller databases at universities first. When CLARIN has become ”THE” repository others like companies will join
- Focus on future data collections
  Willingness in at least Norwegian national funding to have sharing demands for funding

First FONEMA database: FonDat1

- Development of automatic annotation tools (for TTS data) in the Fonema project requires evaluation experiments
- The database **FonDat1** was collected to serve as a test bed for the annotation tool development.
- In addition, the is used to conduct experiments with unit selection synthesis
- Contents:
  - 2 professional speakers (actors): one male, one female
  - To channel recording: Speech and EGG
  - 2000 sentences read newspaper texts (app 3 hours)
  - Manuscript: Greedy search using diphone coverage criteria
  - 10% manually annotated (phonemically and prosodically)

Manual phonemic and prosodic labeling of FonDat1



- Later TTS-databases:
  - Xenophones included
  - Automatic annotation: realistic manus=high OOV rate
  - POS-tags in text for pronunciation disambiguation (\PRT="v"\berget)
  - Speaker specific pronunciation lexicon

Rundkast

http://www.iet.ntnu.no/projects/rundkast/

Database of 70-80 hours radio broadcast news from the Norwegian Broadcasting Corporation (NRK):

- Read and spontaneous speech, as well as spontaneous dialogs and multipart discussions
- There is large variation between speakers, speaking styles and topics
- Speaker turns may be rapid and several speakers may talk simultaneously
- The quality of the recordings include studio and telephone (mobile, satellite etc)
- Frequent occurrences of background noise, jingles, music and audio illustrations

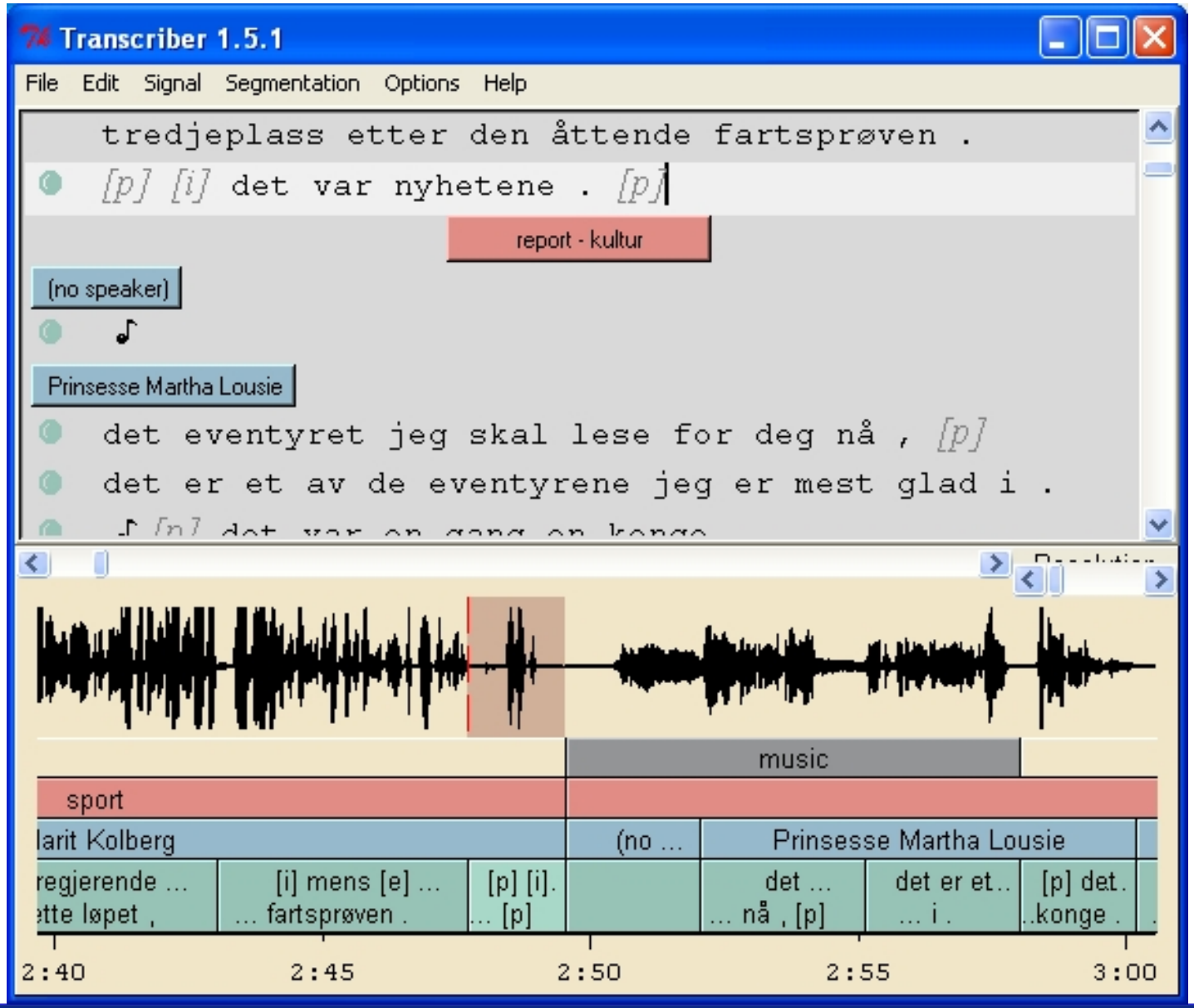
Funded by NTNU

Example speech metadata

Broadcast news database Rundkast

- Hierarchically organized and orthographically annotated:
  - Name of programme, type and date
  - Name of speaker (if known) and dialect (5 regions)
  - Type of speech: spontaneity, channel, recording quality
  - Segmented in speaker turns of app. 2-5 seconds
  - Orthographic transcription (standard Norwegian)
  - Labels for noise (speaker noise, background noise etc.)
  - Labels for pronunciation mistakes, foreign words, unintelligible speech etc.
- Transcriber:
  ”standard” tool and therefore annotation standard ...

Example speech annotation in Transcriber



Rundkast data for phonemic annotation

- 10 speakers (5 male and 5 female)
- Amount of speech per speaker:
  - app 5 min ”planned” speech and 1 min spontaneous speech
  - discard noisy parts (as far as possible)
  - from more than one programme
  - use turn segmentation from orthographic annotation
- All in all 1 hour of speech
- Approximately 1000 hours of work

Phonemic annotation

- Use existing standards
  – with necessary adjustments
- TIMIT: US English
  Again de facto standard, this time for phonetically based ASR experiments
- Norwegian speech synthesis: Specifications and experiences from several databases

- ”Suitable” level of detail: Acoustic boundaries should be labeled, but more phonemic than phonetic
- Consistence of utmost importance!

Main principles for phonemic annotation in Rundkast

- The annotation will be mainly phonemic using the phoneme symbols closest to the perceived sound
- Acoustic boundaries should be marked, include some acoustically motivated symbols
- A transcription as close as possible to the citation form is preferred
- Norwegian standard SAMPA is preferred
  - Some English phonemes included as well as dialect variants
  - Example: 3 variants of the /r/-sound
    /r/ (tap/trill)
    /R/ (uvular fricative)
    /r\/ (approximant)

Comments on deviations

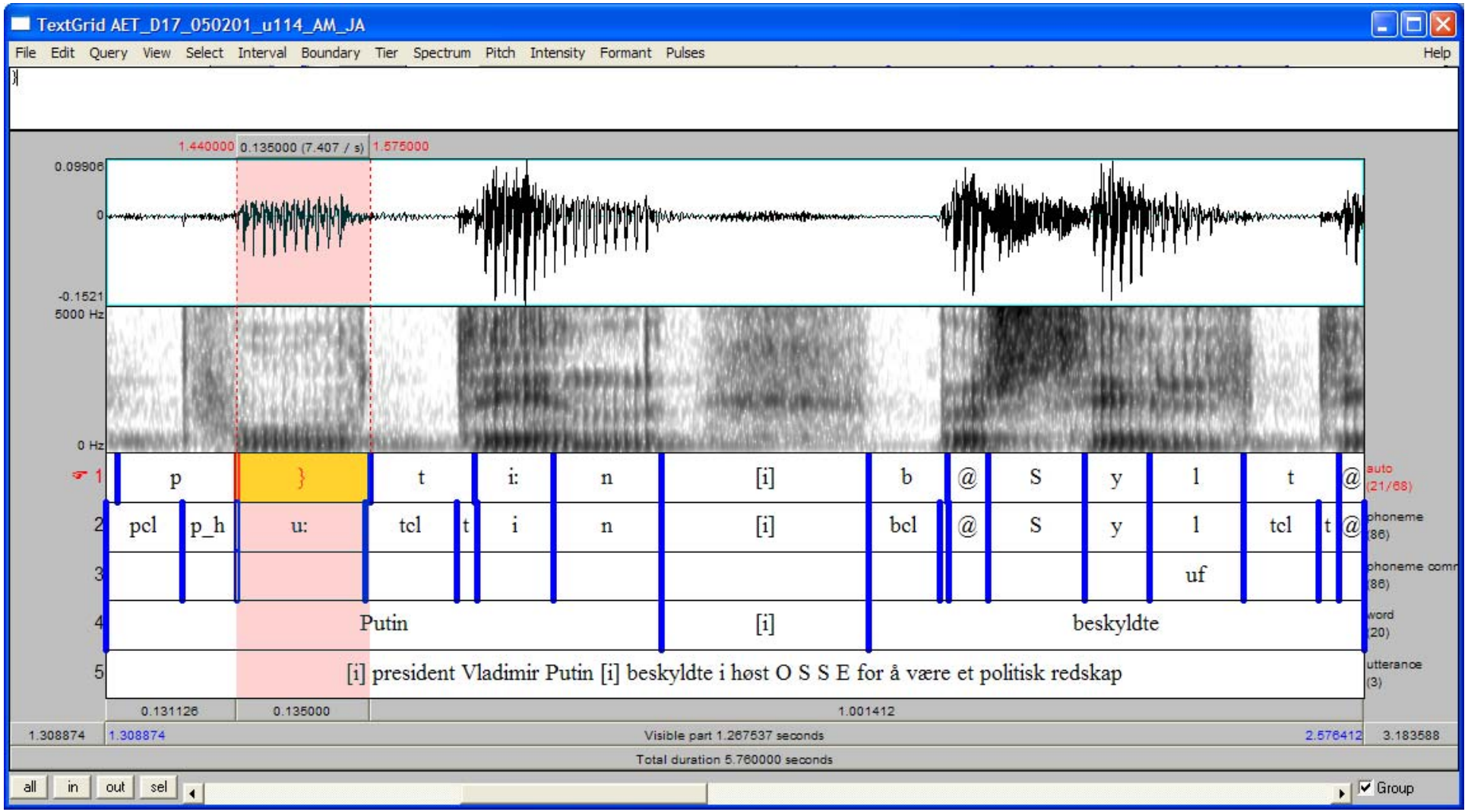
Always cases of uncertainty, need a log for these.

Problem: will the log be read?

Solution: Codes for deviations!

- Synchronous with the phoneme tier (use Praat)
- Easy to utilize automatically
- Examples:
  - creaky voice
  - unexpected voiced/unvoiced
  - uncertain boundary or symbol
- ... in addition a log file with whatever deviations left

Example annotation in Praat



University of Oslo

Tekstlaboratoriet

http://www.hf.uio.no/tekstlab/

- Mostly text (as given by the name)
- Data already registered in CLARIN

BigBrother-corpus: 150k words (500k planned)

Scandinavian Dialect Corpus

- Norwegian part: 400 speakers from 100 loacations (2012)
- UiO, UiT (Tromsø) and NTNU

NoTa-Oslo

- 166 speakers, 900k words
- Multimodal: video, audio and transcription
- Planned to be expanded (Bergen, Trondheim)

L1-L2 word tone database

- Non-natives repeating Norwegian minimal pair words with different word tones
- 22 speakers from 2 L1s (+ 2 teacher voices)
- Partly segmented
- Assessment from several judges included (May also apply for TTS voices)

- Too small for ELRA? and for CLARIN?
- Too much overhead to make available?

NorKompLeks

- Not speech
- Wait with sharing till it is completed?
  That will be never!
- Keep track of what is changed, how?
  Version control!

- What about changes/additions from
  a third party made available?
- Important for scientific work to know
  which version one refers to

“Almost” existing databases

Norwegian TIMIT

- Needed, but not existing (and not in the BLARK)
- First proposal rejected
- Annotation planned in Rundkast style

Customer care human-machine dialog data

- Large amounts
- Willingness to share was there
  but not to do the legal work

Conclusions

- Data important both for humanities and speech technology research
  – sharing is the solution!
- Standards important
  – must be well designed to be used!
- Difficult to decide level of detail

- Start the CLARIN work with the ones interested and hope the speech community will gain interest when they see first results