UNIVERSITETET
I OSLO

Challenges in transcribing spoken language

Janne Bondi Johannessen
(The Text Laboratory, UiO)

PhD training course:
Infrastructural tools for the study of linguistic variation
Fefor Høifjellshotell, Gudbrandsdalen, Norway, 2.-6. June, 2009

UNIVERSITETET I OSLO

Structure of this lecture

- What is a spoken language corpus?
  - The Big Brother Corpus
    - http://www.tekstlab.uio.no/talespraak/bigbrother/
  - NoTa
    - http://www.hf.uio.no/tekstlab/prosjekter/NoTa/NoTa.htm
- TAUS
- Why transcribe spoken data
- Purposes of a spoken language corpus
- What to transcribe?
  - Transcription in NoTa
- Informants - Who, where, when, how?
- Other spoken language corpora
  - Gothenburg Spoken Language Corpus
    - http://www.ling.gu.se/projekt/tal/
    - http://www.ling.gu.se/~leifg/tal/
  - Danish BySoc
    - http://www.id.cbs.dk/%7Epjuel/cgi-bin/BySoc_ID/index.cgi
  - Swedia
  - Scandiasyn
  - British National Corpus

UNIVERSITETET
I OSLO

The Big Brother Corpus

- Pros
  - Lots of spontaneous speech data
  - Lots of dialogue and polylogue
  - Lots of emotional speech in different dialogue situations
    - Conflict, argument, love, irritation etc.
- Cons
  - Not a dialect corpus
  - No representativity w.r.t. age, social class, education etc.
  - Not ”controlled” recording situations
  - Small number of informants

UNIVERSITETET I OSLO

NoTa (Norsk talemålskorpus)

- Goal
  - Record the speech in the Oslo area
  - Representative samples w.r.t. age, education, social status, geographical location
    - But not easy to do (clustering of properties)
    - How to find them
  - Main focus on spontaneous speech
    - Each informant
      - half an hour of dialogue with some other informant (family, friend, acquantance, unknown)
      - 10 minutes of interview
- Number of informants: 144

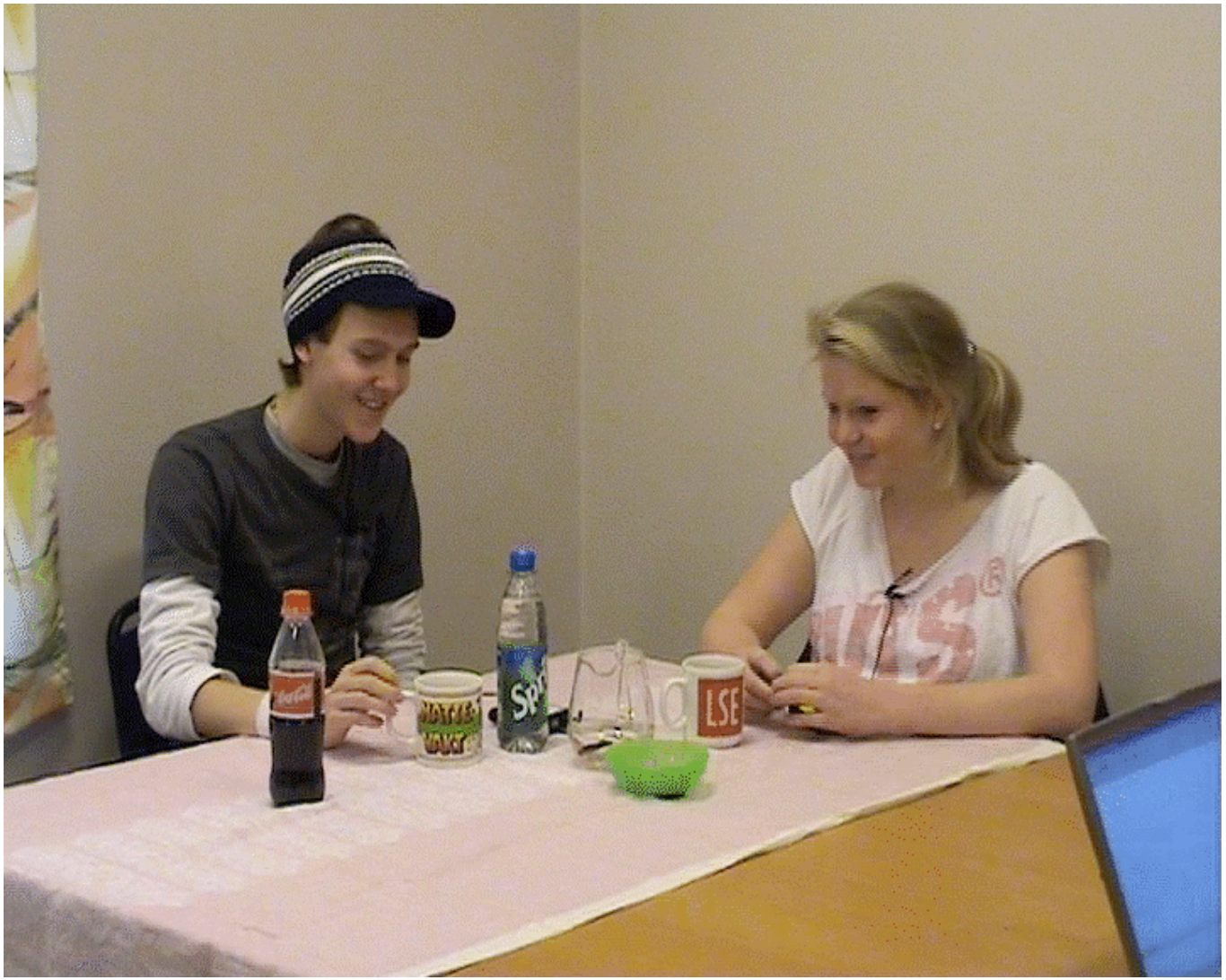
UNIVERSITETET
I OSLO

NoTa

- Pros
  - Representativity
  - Quantity
- Cons
  - Too controlled setting
    - Are the informants influenced by the situation?
      - Swearing, register w.r.t. vocabulary, inflections, pronunciation
    - Are informants influenced by interviewer?
  - Few emotions

UNIVERSITETET
I OSLO

Two young informants



UNIVERSITETET
I OSLO

Why transcribe spoken data?

- In the past
  - In order to make the data available for a wider audience (unpractical or even impossible to distribute tapes - no internet…)
  - Get a good overview of the data (read and browse)
- Now
  - Get a good overview of the data
  - Make data searchable (due to software not previously available)
  - Tag data grammatically and make more interesting searches
  - Less important: make data available to others

UNIVERSITETET
I OSLO

Purpose of transcribed corpus

- Pragmatic research
- Morphological research
- Syntactic research
- Semantic research
- Conversation analysis research
- Phonetic/phonological research
- Socio-linguistic
- Etc.

UNIVERSITETET
I OSLO

Does the purpose of the transcribed data determine what to transcribe?

- First answer: Yes
  - No detailed phonological transcription needed in syntactic research
  - Morphological variation perhaps not necessary in conversation analysis
  - Extra-linguistic information possibly not necessary for socio-linguistic studies
  - Etc.

UNIVERSITETET
I OSLO

Does the purpose of the transcribed data determine what to transcribe (continued)?

- No!
  - In order to make the corpus maximally searchable, orthographic transcription is necessary.

UNIVERSITETET
I OSLO

Example

- Search for all occurrences of the pronoun *jeg (’I’)*.
  - Alternative 1:
    - Search for each of the forms that occur in the dialects that constitute your corpus:
      - /æ:/, /je:/, /jæi/, /jæ/, /e:/, /i:/ etc.
  - Alternative 2:
    - Search for *jeg*, and get all occurences immediately - then listen to each with your favourite sound program or look at additional transcriptions that accompany the orthographic forms

=> Orthographic transcription is necessary

UNIVERSITETET
I OSLO

What's the "best" dialect data?

- Answers to questions posed by an interviewer with the same (or different) dialect?
- A monologue (e.g. a story) prompted by the interviewer?
- Dialogue produced by dialect speakers under controlled conditions?
- Dialogue/polylogue produced by dialect speakers under uncontrolled conditions?

UNIVERSITETET
I OSLO

UNIVERSITETET
I OSLO

If dialogue data, then many features have to be dealt with - even if dialogue is not the main interest of the study

- overlapping speech, interruptions
- punctuation
- pauses
- emphasis
- morphology
- phonology
- extralinguistic features (laughter, sighs, ...)
- sounds that are in the borderline betweeen extralinguistic and linguistic (interjections)

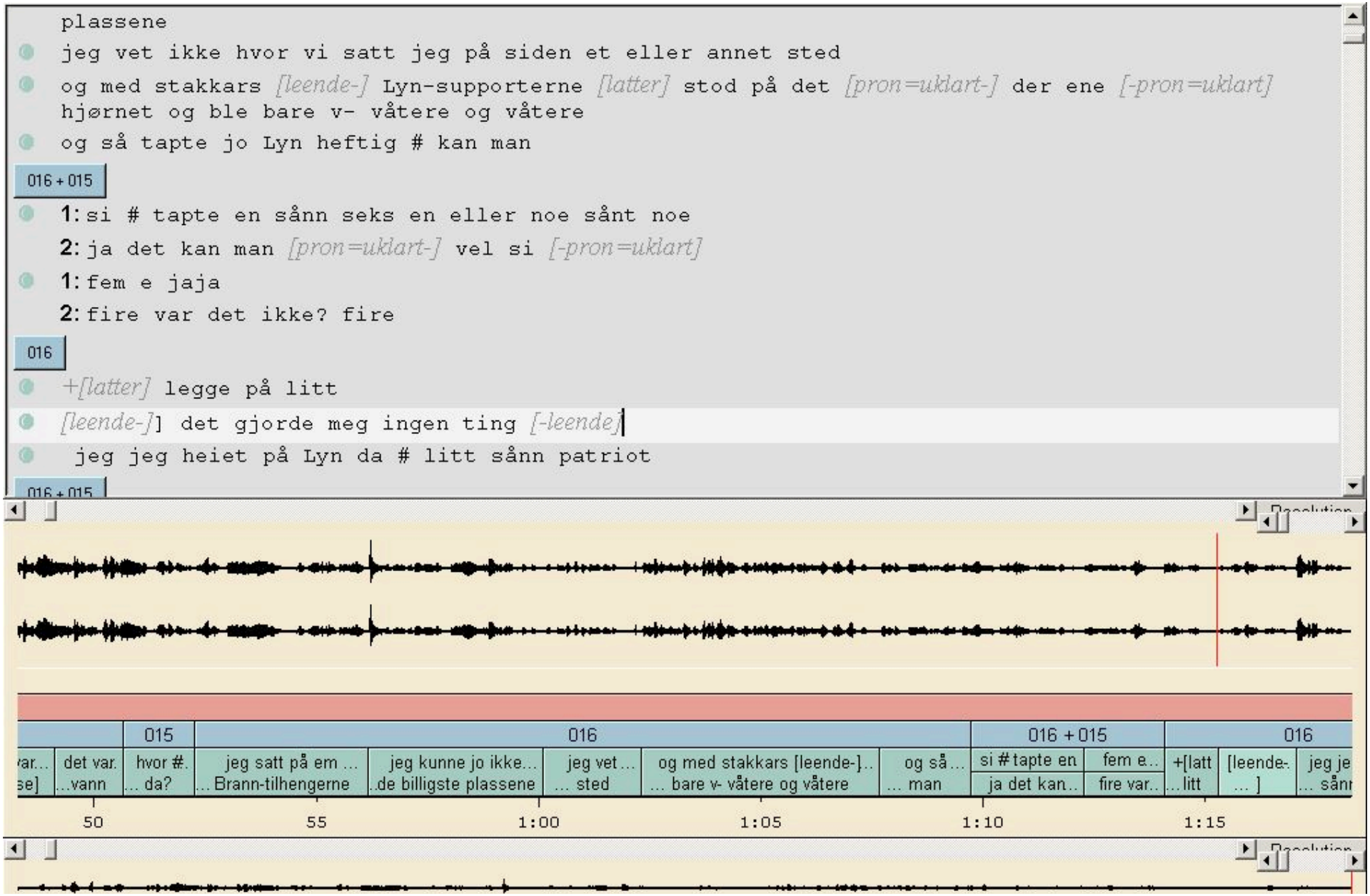
UNIVERSITETET I OSLO

NoTa

- Dialogue
  - http://www.hf.uio.no/tekstlab/prosjekter/NoTa/internt/AMB_samtale_003-004.wav.mp3
- Transcription
  - http://www.hf.uio.no/tekstlab/prosjekter/NoTa/samtale.html
- Transcription done in the Transcriber program
  - http://www.etca.fr/CTA/gip/Projets/Transcriber/Index.html

UNIVERSITETET I OSLO

The Transcriber program



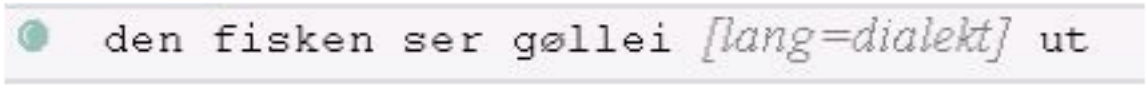
UNIVERSITETET
I OSLO

```
<Turn speaker="spk1 spk2" startTime="72.269" endTime="73.299">
<Sync time="72.269"/>
<Who nb="1"/>
fem e
<Who nb="2"/>
fire var det ikke?
</Turn>
<Turn speaker="spk2 spk1" startTime="73.299" endTime="74.065">
<Sync time="73.299"/>
<Who nb="1"/>
ja ja
<Who nb="2"/>
fire
</Turn>
<Turn speaker="spk2" startTime="74.065" endTime="80.114">
<Sync time="74.065"/>

<Event desc="latter" type="noise" extent="instantaneous"/>
 legge på litt
<Sync time="75.293"/>

<Event desc="leende" type="noise" extent="begin"/>
 det gjorde meg ingen ting
<Event desc="leende" type="noise" extent="end"/>

<Sync time="77.03"/>
 jeg jeg heiet på Lyn da # litt sånn patriot
</Turn>
```

UNIVERSITETET I OSLO

Transcription in NoTa - additional annotation

- **-** **pronounce**
- **noise**
- **-** **language**
  - » (for words not in the *Bokmålsordboka*, e.g. foreign or dialect words)
- **-** **lexical**
  - » (for specifying the pronunciation of certain words)
- **comment**
  - » (for comments on problems, sensitive person information etc. *)*



UNIVERSITETET
I OSLO

NoTa - orthographic transcription at word level - but keep gender and ”wrong” use of words

Informant says:	We transcribe:
je jikk på vægen	jeg gikk på vegen
henne jikk	henne gikk
vi snakka på det	vi snakka på det
jei ga det til de	jeg ga det til de
jei bruker ei maskin	jeg bruker ei maskin
da får døm si det sjøl	da får dem si det sjøl
jei mener det ass	jeg mener det altså

UNIVERSITETET
I OSLO

NoTa - when more than one variety is allowed, choose the one that is closest to the one used by the informant

Informant says:	We transcribe:
sne	snø
røyk	rauk
mjølken	mjølken
åssen	åssen
trur	trur
vart	vart
blei	blei
hu	ho {lex=hu}

UNIVERSITETET
I OSLO

NoTa - stick to the norm w.r.t. ”deviation” in stem and phonological variation in suffixes

Informant says:	**We transcribe:**
itte	**ikke**
søvi	**sovet**
hestær	**hester**
prate (present)	**prater**

UNIVERSITETET
I OSLO

NoTa - special treatment of pronouns

- Pronouns are written w.r.t. standard norms and as they are used by the informant

```
jeg tok boka deres +[lex=dems]
```

- Two pronouns have been added to the standard ones - because they are different
  - a
  - n

UNIVERSITETET
I OSLO

The pronouns *a* (3p.sg.f) and *n* (3.p.sg.m)

- These clitic pronouns differ in phonological form from the full pronouns , and it is not clear of which, if any, pronouns they are variants.
  - A
    - Hun (3p.sg.f.nom)
    - Henne (3p.sg.f.acc)
  - N
    - Han (3p.sg.m.nom)
    - Ham (3p.sg.m.acc)
- Since speakers differ w.r.t. how they use the full form-pronouns (nominative is not always used in subject position etc., it would be wrong to take syntactic function as a guideline for their transcription.
  - **der er a**
  - **jeg så a i går**
  - **jeg så n**
  - **jeg så n Lars**

UNIVERSITETET
I OSLO

Exceptions from the norm

- **Keep gender**
  - **ei maskin (norm: en maskin)**
  - **maskina (norm: maskinen)**
- **Keep lexical words that are not found in the main dictionary**
  - **Dette kuper**
  - **Det er illere**

UNIVERSITETET I OSLO

Other things

- Abbreviations
  - **de sa det på NRK**
- Compounds
  - **trafikksituasjon**
- Numbers
  - **sekstifire tusen**
- Names
  - **jeg så at F1 ga F2 dokumentene til E1**
  - **det foregikk på N1 # ikke sant 081**
- Dialect words and words from other languages
  - **Yes [lang=english] slik er det**
- Citations
  - **da kjørte jeg den "jeg? hæ?" da ljuger jeg**
  - **jeg sier ikke ″sne″ jeg jeg sier ″snø″**
- New words, swearing
- Spellings
  - **Kutt staves [pron=stavet-] c u t [-pron=stavet] på engelsk**
- Noises
- Emphasis
  - Is not marked (no criterion availble)

UNIVERSITETET
I OSLO

Interruptions, pauses and unclear passages

- Interrupted words
  - *hvo-*, *hvo-*, *hvordan*
- Self-interrupted utterances
  - **høres ut som sånn her ...**
  - **du har du har du har ikke gjort det**
- Pauses
  - **og # jeg tror ## det er slik at**
- Unclear passages
  - **Men takk for at du {uforståelig}**

UNIVERSITETET
I OSLO

Punctuation

- Since spoken language differs from written language, comma and full stop and capital letters utterance-initially are not used
- Capitals are used in names.
- Question mark and exclamation mark are used
- **`kommer du i morgen?`**
- **`kom hit!`**

UNIVERSITETET
I OSLO

Turns and segments

- Turns are marked
- Overlaps are marked
- Segments are marked
  - For time coding
  - For separating out ”natural” units (intonation)
  - For presumed ease of later grammatical tagging

Monica

- ja fordi at jeg går rundt og prøver være venner med alle
- og det er ikke noe jeg har bevisst prøvd å gjøre
- jeg har prøvd å være hyggelig mot folk
- for det er ingen her inne jeg direkte misliker så hvorfor skulle jeg gå rundt da å være ...
- *[snufsing]* ja, selvfølgelig er det noen jeg liker mer enn andre men det er # jeg kjenner ikke alle godt nok til å ...

UNIVERSITETET
I OSLO

Overlapping speech

001 + AMB

1: gamle hus som det ikke er # blitt pussa opp så veldig mye på fasadene i hvert fall
2:
1: så mm
2: ja
1: men jeg syns det er veldig fint der så
2:
1: trives godt
2: ja
1: mm
2:

UNIVERSITETET
I OSLO

The most common noises - predefined

- fremre klikkelyd
- bakre klikkelyd
- sugelyd
- labial frikativ
- labial vibrant
- sibilant
- latter
- gjespende
- gråt
- hosting
- knipsing
- kremting
- lattermild
- leende
- lydmalende ord
- pause
- pusting
- snufsing
- stønning
- sukking
- trekker pusten

UNIVERSITETET I OSLO

Many ”new” interjections - (interjection: a word with a constant meaning)

aha	**(overraskende) BMO**
e	**(nøling – uansett lengde på een)**
eh	**(avstandsindikerende)**
ehe	**(”Jeg forstår” – to stavelser)**
em	**(nøling)**
heh	**(imponert)**
hm	**(spørrende, undrende)BMO i betydningen kremting**
hæ	**(spørrende) BMO**
jaha	**(forsterkende ”ja”)BMO**
m	**(nøling, ta til etterretning, nam)**
m-m	**(benektende)**
mhm	**(”Jeg forstår” – to stavelser)**
mm	**(bekreftende)**
nja	**(tvilende) BMO**
næhei	**(forsterkende ”nei”)**
u	**(imponert)**
ææ	**(konstaterende – to stavelser)**
å-å	**(”oj”**
å ja	**(overraskende)**

UNIVERSITETET
I OSLO

Conclusion

- Developing a spoken language corpus is very different from a written corpus.
- This is important to know for future users.
- Many of the decisions made in NoTa might not be made in future spoken language corpora
  - Time is a decisive factor w.r.t. transcription, and every decision is time consuming.
  - Decisions without clear criteria for choice are even more time consuming (what is a turn, how long is a pause, which interjection do I hear…)
- But spoken language corpora are fun to use, and they will certainly reveal new information about language, and possibly gestures, interplay between modalities and many other things.