**Transkripsjonsveiledning for NORINT-korpuset**

14.06.2016

Av Kristin Hagen, Live Håberg og Annely Tomson[1]

## Innhold



[1] Transkripsjonsveiledningen er et resultat av samarbeid mellom mange ansatte i flere prosjekter, blant annet LIA, Nordiasyn og NoTa-Oslo

## 1 Om Norint-korpuset

NORINT-miljøet har i flere år gjort mindre forskningsarbeid som gjelder språket til internasjonale studenter. Vi har ønsket et materiale som kunne gi både ansatte ved ILN og ILNs bachelor- og masterstudenter mulighet til å forske i og lære mer om språket til studenter med et annet morsmål enn norsk. I juni-juli 2014 og 2015 ble det samlet inn et lyd- og videomateriale av 58 utenlandske studenter som gikk på trinn 3 på Den internasjonale sommerskole ved UiO. Det ble gjort tre opptak av hver student på tilsammen ca. 45 minutter. Hvert opptak består av ett intervju om morsmål, bakgrunn, arbeid, fremtidsplaner og en samtale mellom to informanter om valgfrie temaer, for eksempel kultur, fritid, reiser, Norge. Til slutt ble det gjort lydopptak av en opplesning av en gitt tekst og 60 setninger (de samme som ved Språkmøterprosjektet ved NTNU).

### *1.1 Ortografisk transkripsjon*

For å kunne søke i et talespråkskorpus må materialet transkriberes. Vi har valgt å bruke en rent ortografisk transkripsjon fordi den vil være lett å utføre, den vil gjøre det lett å søke i korpuset, den vil være lett å tagge automatisk, og lett å lese. Ortografisk transkripsjon skal gjengi informantens tale så godt det lar seg gjøre innenfor det skriftnormalen tillater. Den skal med andre ord være et best mulig kompromiss mellom krav om nøyaktig gjengivelse av det talte, og skriftmålets normer.

En uttalenær, fonetisk transkripsjon kan være nyttig for noen fonologiske studier, men det vil alltid være en mulighet for at akkurat de fenomenene som en bestemt fonolog er interessert i, likevel ikke er markert. Det vil også gjøre korpuset vanskelig å søke i og det er vanskelig å gjennomføre konsekvent, med forskjellige transkribører med ulik faglig og teoretisk bakgrunn.

## 2 Transkripsjonsprogrammet ELAN

*ELAN* er et gratis transkripsjonsprogram fra Max Planck Institute og The Language Archive i Nederland. Vil du lese mer om programmet, kan du gå inn på nettsiden http://tla.mpi.nl/tools/tla-tools/elan/. På nettsiden finner du også en manual for programmet.

ELAN har tre ulike moduser å arbeide i:

- **Segmentation Mode**
- **Transcription Mode**
- **Annotation Mode**

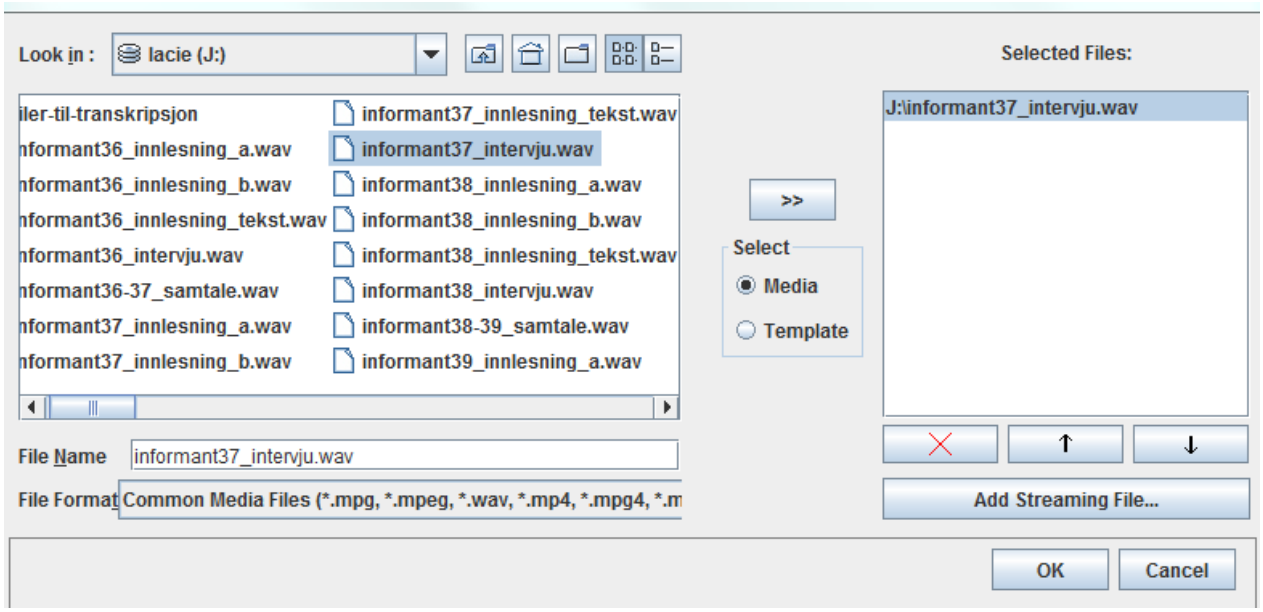
I **Segmentation Mode** deler vi transkripsjonen opp i tidsbolker. Når segmenteringen er unnagjort, kan du gjøre selve transkripsjonen i **Transcription Mode**. I **Annotation Mode** er det mulig både å segmentere og transkribere, men vi opplever at det er best å bare bruke denne modusen når man skal se over transkripsjonen og gjøre mindre endringer.

Det vil sannsynligvis oppleves enklest å veksle mellom perioder med segmentering og perioder med transkripsjon for å få litt variasjon i arbeidet.

### *2.1 Begynne på en ny transkripsjon og definere talere*

**Åpne en lydfil i ELAN:**

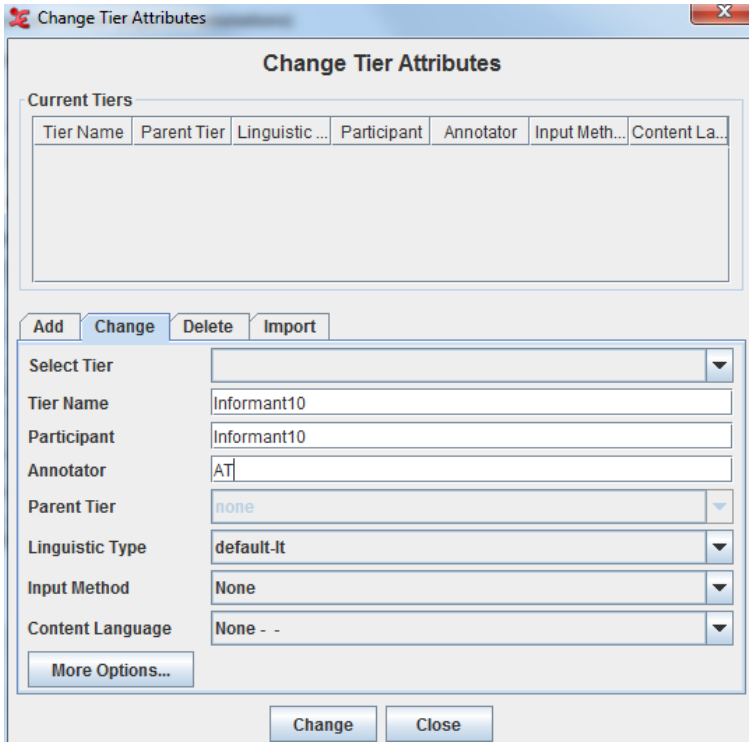
- Velg **File → New**
- I vinduet til venstre velger du riktig lydfil




**Definere talere:**

Hver taler skal ha sin **tier** (lag) under spektogrammet («bølgeformen») du ser i **Annotation Mode** (som er den forhåndsvalgte modusen i ELAN). Rekkefølgen på talerne har ingenting å si. Når du åpner programmet, har ELAN laget en taler som heter **default**. Du setter navn på talerne slik:

- Velg **Tier → Change Tier Attributes** for å endre navnet på **default**. I dialogboksen skal Tier Name og Participant være det samme.
  **Annotator**: sett initialene dine her.
  **Parent Tier** skal være **none**.
  **Lingustic Type** er **default-lt**.
  **Input Method** er **None**.
  **Content Language** er **None**.




- Er det flere talere i lydfilen du skal transkribere, klikker du på **Tier → Add New Tier** og fyller ut dialogboksen på samme måte. Du kan definere flere talere når som helst under transkripsjonen.

### *2.2 Fortsette med en påbegynt transkripsjon*

Dobbeltklikk på transkripsjonsfilen (filnavn.eaf) du vil jobbe med, og du vil automatisk få opp både transkripsjon og lyd-/videofil. Du kan også åpne programmet ELAN og deretter velge **Open** og riktig transkripsjonsfil. Lyd-/videofilen åpnes automatisk.

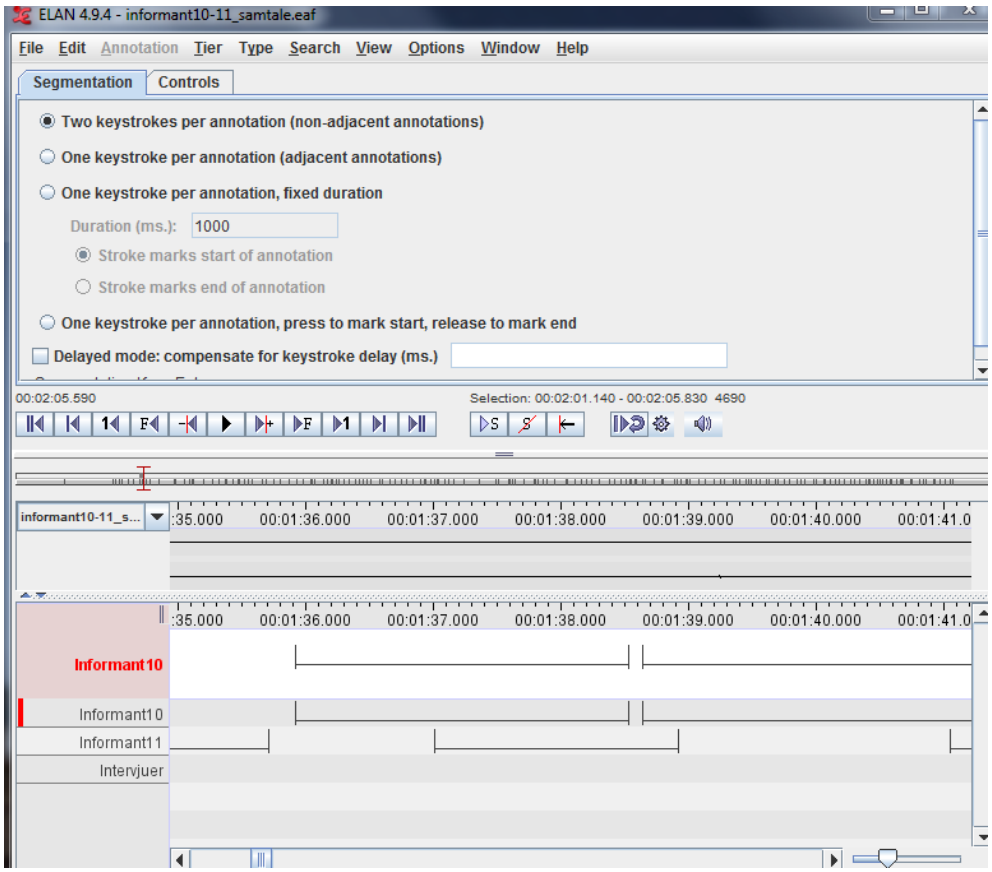
### *2.3 Segmentering*

Når vi segmenterer, deler vi lyden opp i mindre biter. For hver gang vi starter eller avslutter et segment, blir det satt inn en tidskode i transkripsjonen. Det er disse tidskodene som gjør at transkripsjonen blir koblet sammen med lyden.

#### 2.3.1 Segmentering i Segmentation Mode

Velg **Options →** **Segmentation Mode**.

I denne modusen er det ikke mulig å transkribere, men det går raskt å segmentere, det vil si å dele opp lydfilen i tidssegmenter for hver taler.

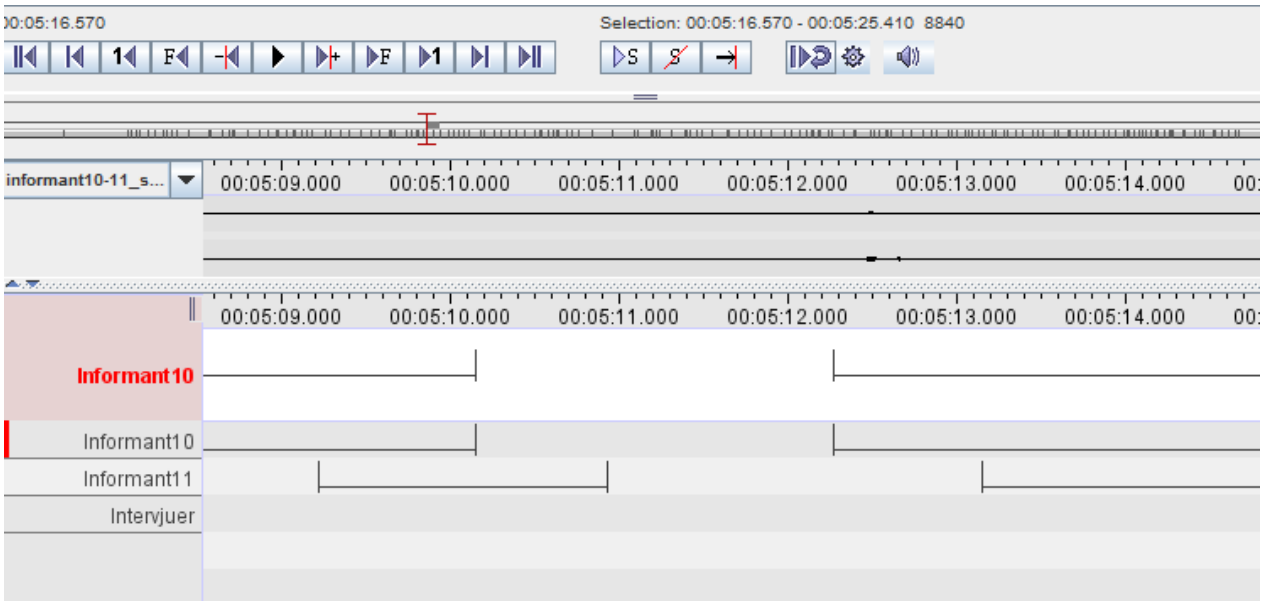



Du spiller av lyden ved å trykke på **play**-knappen på spilleren i programmet eller enklere, trykke **Ctrl+Mellomrom**. Du stopper lyden på samme måte.

Du kan veksle mellom aktive **tiers** med pil opp og pil ned. Den aktive tieren kan vi dele inn i segmenter (**annotations**) ved hjelp av markøren (**crosshair**) og **enter-**tasten. Sett markøren der segmentet skal starte i spektogrammet og marker starten på segmentet med **enter**-tasten. Sett så markøren der segmentet skal slutte, og marker slutten med et nytt trykk på **enter**-tasten.  Nå har du laget et segment mellom de to markerte stedene. Hvis du transkriberer en taler som snakker lenge, og du vil dele opp i flere segmenter uten pause mellom, trykk **enter** to ganger i slutten av et segment, slik at slutten på dette segmentet automatisk blir markert som starten på neste.

Du kan flytte markøren i spektogrammet med musa eller ved hjelp av snarveier (se under). Segment kan flyttes og finjusteres ved å klikke og dra, og de kan slås sammen eller deles opp ved hjelp av snarveier. Pass på at taleren du vil endre på, er aktiv (taleren med navn i rødt). Du bytter mellom talerne med piltastene (opp eller ned).

I en samtale eller et intervju med to deltakere kan det være greit å konsentrere seg om å segmentere en taler om gangen, særlig når det er mye overlappende tale. Segmentene vil ofte overlappe hverandre helt eller delvis, for eksempel slik:




### 2.3.2 Nyttige snarveier, Segmentation Mode

Merk at noen av snarveiene er ulike på PC og Mac (siden funksjonstastene er ulike).

| **Funksjon** | **PC og Mac** |
|---|---|
| | |
| Slå sammen (uthevet segment) med neste segment | Ctrl+A |
| Slå sammen (uthevet segment) med forrige segment | Ctrl+B |
| Del opp (uthevet segment) | Ctrl+Enter[2] |
| | |
| Gjør tieren over aktiv | Pil opp |
| Gjør tieren under aktiv | Pil ned |
| | |
| Spill/Pause | Ctrl+Mellomrom |
| Spill markert område | Shift+Mellomrom |

Bruk musa til å flytte markøren, eller bruk snarveiene:

| **Funksjon** | **PC og Mac** |
|---|---|
| | |
| Flytt markøren ett sekund til venstre | Shift+Pil venstre |
| Flytt markøren ett sekund til høyre | Shift+Pil høyre |

Ctrl/Cmd+Shift+Piltastene kan brukes for å flytte markøren et lite hakk til høyre eller venstre.

[2] Merk at denne snarveien ikke er forhåndsdefinert. I kapittel 5 kan du lese om hvordan du lager egne snarveier.

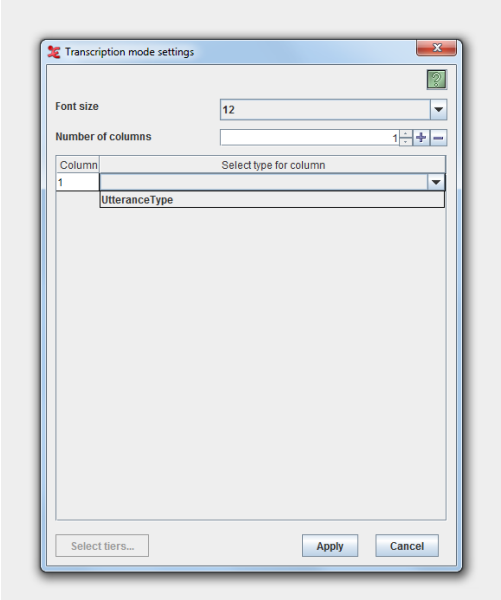
### *2.4 Transkripsjon*

I **Transcription Mode** går det raskt å transkribere den delen av opptaket som allerede er segmentert i **Segmentation Mode**. Du kan ikke redigere segmenteringen i denne modusen.

#### 2.4.1 Transkripsjon (og korrekturlesing) i Transcription Mode

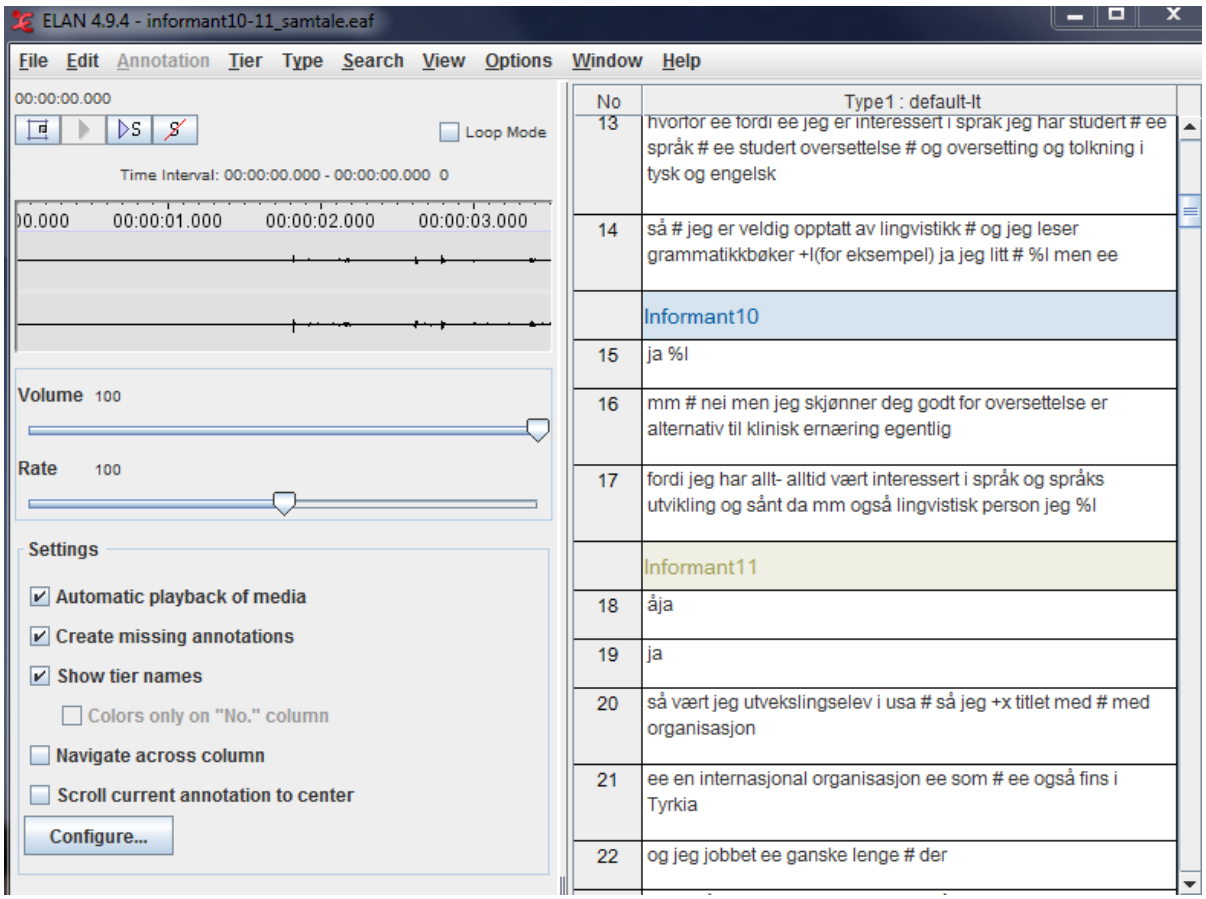
Vel **Options** → **Transcription Mode**.

Når du bytter til **Transcription Mode** første gang, må du konfigurere oppsettet. Klikk på det hvite feltet etter **Column 1** og velg **default-lt**.




12 punkt kan være en passende skriftstørrelse.

Nå ser du segmentene du laget i **Segmentation Mode** under hverandre med navnet på de ulike talerne i ulike farger.




Det er flere ulike funksjoner for å spille av lyden:

- Du kan klikke i det hvite feltet der transkripsjonen skal stå.
- Tab fungerer som en **play/pause**-knapp.
- Shift+Tab spiller av lyden fra starten av segmentet.
- **Enter** går til neste felt og spiller av lyden.
- Alt+Pil opp/ned skifter felt og spiller av lyden.

Krysser du av for **Loop Mode** (øverst til høyre over spektogrammet), blir lyden i segmentet spilt om igjen og om igjen til du stopper den med Tab.

Hastigheten på opptaket kan du endre ved å justere **Rate** (linjen under **Volume**). Som regel bør **Rate** være 100, men om noe er veldig utydelig eller taleren snakker veldig fort, kan du sette ned hastigheten litt.

Du transkriberer rett i de hvite feltene. Om du vil endre på segmenteringen, må du gå tilbake til **Segmentation Mode**.

### 2.4.2 Nyttige snarveier, Transcription Mode

| **Funksjon** | **PC og Mac** |
|---|---|
| | |
| Hopp ned til neste segment | Enter/Alt+Pil ned |
| Hopp opp til forrige segment | Alt+Pil opp |
| | |
| Spill/Pause | Tab |
| Spill segmentet om igjen fra start | Shift+Tab |

De samme snarveiene kan brukes til å flytte markøren som i **Segmentation Mode**.

## 3 Generelt om transkripsjon

Vi følger to hovedregler for transkripsjon:

1. Skriv bare former som finnes i Bokmålsordboka.
2. Ordene skal skrives ned nøyaktig i den rekkefølgen de blir uttalt, det vil si at vi ikke skal ta hensyn til syntaktiske regler for skriftlig bokmål.

Siden vi transkriberer ortografisk, bruker vi altså bare former som står i Bokmålsordboka. Slå opp i Bokmålsordboka på nett om du er i tvil: http://bokmålsordboka.uio.no/.

### *3.1 Tegnsetting og stor forbokstav*

Stor forbokstav brukes ikke i begynnelsen av en setning. Semikolon, kolon, komma, utropstegn og punktum brukes ikke. Spørsmålstegn brukes som vanlig ved spørsmål, også dersom spørsmålet blir avbrutt. Ved retoriske spørsmål kan spørsmålstegn sløyfes. Vi må ha mellomrom foran spørsmåltegn.

### *3.2 Ord som ikke står i Bokmålsordboka*

Dersom informanten bruker ord og uttrykk fra andre språk eller dialekter, markerer vi ordet/ordene med **+x**, se nedenfor. Dersom vi kjenner den godkjente skrivemåten for ordet (for eksempel *all right* på engelsk), normaliserer vi til det.

### *3.3 Grammatiske feil*

Dersom informanten bøyer et ord feil (sier «stolet» i stedenfor «stolen» eller «håpa» i stedet for «håpet», for eksempel), skriver vi korrekt form, men markerer ordet med **+g**:

+g stolen

+g håpet

Ved feil bruk av preposisjon skriver vi bare den preposisjonen som blir sagt, uten tagg.

### *3.4 Sammentrekking*

Sammentrekking markeres ikke. Om informanten sier *jei kan'ke gå*, skriver vi *jeg kan ikke gå*.

### *3.5 Forkortelser*

Forkortelser skrives som forkortelser. Forkortelser som *dvd*, *cd* og *pc* skrives med små bokstaver.

### *3.6 Sammensetninger*

Mange sammensetninger står ikke i Bokmålsordboka, men sammensetninger skal likevel normaliseres på vanlig måte:

førsteklasse

maskiningeniør

### *3.7 Tall*

Tall skrives med bokstaver. Slik får vi med om for eksempel *1600* uttales *ett tusen seks hundre* eller *seksten hundre*. Tall under hundre skrives som ett ord, alle tall over hundre skrives i flere ord:

tjuefire

trettisju

sekstifire tusen

ni millioner

tre tusen to hundre og tjuefem

Legg merke til at tallordet 1 skrives uten aksent:

en

Unntaket er årstall som skrives som ett ord:

nittenhundreogtrettiåtte

Ordenstall skrives også i ett ord med bokstaver:

tjuefjerde

Ordenstall over hundre skrives i flere ord:

to hundre tjuefjerde

Brøker skrives også med bokstaver:

to tredjedel

### *3.8 Navn*

Vi skriver vanligvis ikke inn navn på personer. I stedet skriver vi koder for hvert enkelt navn, **F** for jentenavn (fornavn eller fornavn pluss etternavn), **M** for guttenavn (fornavn eller fornavn pluss etternavn), og **E** for etternavn (uten fornavn). Navnene får fortløpende nummer etter som de dukker opp i teksten. Vi skriver informantkode og ikke navnet på informanten.

Informanten sier: «Jeg heter Kalina, og jeg kommer fra et lite sted som heter Elenite i Bulgaria.»

Vi skriver: jeg heter informant_32 og jeg kommer fra et lite sted som heter N1 i Bulgaria

Navn på offentlige personer behøver som regel ikke anonymiseres. Fotballag, stedsnavn, navn på kjæledyr og organisasjonsnavn mv. skrives normalt inn dersom du ikke vurderer at det er sensitive personopplysninger involvert. Bør et slikt navn anonymiseres, skriv **N1, N2** osv. Gatenavn kan

normalt stå, men dersom informanten oppgir hele adressen med gatenavn og nummer, erstattes adressen med **N**.

Navn skrives normalt med stor bokstav, også navn på sanger, filmer og bøker. Navn på verk osv. skrives i hermetegn når det består av flere ord:

«Seeta Aur Geeta»

Dersom et navn skal skrives som navn og ikke som kode, skal det skrives slik ”eieren” vanligvis skriver det:

TV2

CatoSenteret

Ringenes Herre

Dersom det er vanskelig å oppfatte navnet, skriver vi bare **N1, N2** osv. NB! Ikke bruk mye tid på navn!

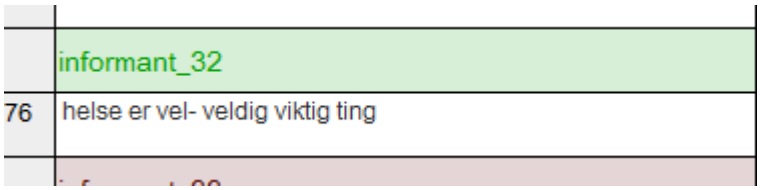
### *3.9 Uavsluttede ytringer*

Dersom en ytring ikke blir skikkelig avsluttet, og en annen informant overtar, for eksempel ved avbrytelse o.a., skal den avbrutte ytringen avsluttes med tre punktum med mellomrom foran:

*høres ut som sånn her ...*

### *3.10 Avbrutte ord*

Når et ord er avbrutt, skal de delene som er uttalt, likevel gjengis i ortografi. Dette markeres ved hjelp av bindestrek som skrives uten mellomrom rett etter den siste bokstaven:




### *3.11 Tagger til markering av ekstrainformasjon og ikke-språklige lyder*

Intervjuene og samtalene inneholder en del ikke-språklige elementer som latter, hosting osv. Lydopptakene kan også inneholde informasjon som ikke bør transkriberes (som informasjon om når lydopptakeren/kameraet er slått på eller skal slås av, innledende kommentarer eller sensitiv informasjon), og enkelte ganger kan det være nødvendig å markere ett eller flere ord som utydelige eller uforståelige. ELAN har ikke støtte for tagger eller hurtigkommandoer inne i selve transkripsjonen, og vi har derfor laget et eget system som bør være raskt og enkelt å bruke for å tagge teksten fortløpende.

### 3.11.1 Generelle prinsipper for tagger

**+** Plusstegn og bokstaven etter karakteriserer ordet eller ordgruppen som kommer etter. Dette kalles uavhengige tagger.

+u også
→ transkribøren mener «også» er uklart

**( )** Vanlige parenteser setter man rundt en ordgruppe som skal karakteriseres med en +tagg. Det skal ikke være mellomrom mellom taggen og parentesen.

+u(jeg dro til Norge)
→ transkribøren mener «jeg dro til Norge» er uklart

**{ }** Kommentarer blir markert med krøllparenteser.
{støy}

**%** Uavhengige tagger markeres med %, og de representerer en selvstendig sekvens i talestrømmen. En slik sekvens kan ikke uttrykkes med skrift og kan være latter, gjesping osv.

tror jeg %l
→ %l betyr at informanten ler etter at hun har sagt «tror jeg»

Flere uavhengige tagger kan stå til samme ordet eller samme uttrykket innenfor parenteser, rekkefølgen på taggene er uvesentlig.

+l +u (også jeg)
→ informanten ler mens hun sier «også jeg». «også jeg» er i tillegg uklart

Parentesuttrykk innenfor parenteser er ikke mulig. Dersom flere ord innenfor en parentes skal merkes, må hvert ord merkes separat (a), eller eventuelt må hele uttrykket merkes som (b).

(a) +l (også +u jeg)
(b) +l +u (også jeg)

### 3.11.2 Liste over avhengige og uavhengige tagger med betydning

| **Avhengige** | | **Uavhengige** | |
|---|---|---|---|
| +x | ord fra andre språk, norske dialektord, multietnolekt, | | |
| +g | grammatisk feil | | |
| | | %o | onomatopoetikon |
| +u | uklart | %u | uforståelig |
| +l | leende | %l | latter |
| +j | gjespende | %j | gjesp |
| +v | hviskende | | |
| | | %k | kremting |
| +s | stønnende/sukkende | %s | stønn/sukk |

### *3.12 Liste over interjeksjoner*

| | |
|---|---|
| ee | uansett lengde, nøling |
| eh | avstandsindikerende |
| ehe | «jeg skjønner», to stavelser |
| em | nøling |
| heh | imponert |
| hm | spørrende, undrende |
| m | nøling, ta til etterretning, nam |
| m-m | benektende |
| mhm | «jeg skjønner», to stavelser |
| mm | bekreftende |
| åh | utrop |
| aha | overrasket |
| hæ | spørrende |

### *3.13 Ellers*

Pause inntil ett sekund # og alle de andre pausene ##

**4 Opplest tekst**

NORINT-materialet består også av lydopptak av en opplest tekst og 60 setninger. De samme setningene og teksten ble brukt i språkmøterprosjektetet ved NTNU.

Disse opptakene skal segmenteres i ELAN, setning for setning. Etterpå skal de transkriberes, men siden vi kjenner teksten og setningene informantene har lest inn, kopieres disse bare og klippes inn i segmentene på rett sted.

Selv om informantene leser opp setninger fra et ark, hender det at de ikke sier nøyaktig det de skal.

> Setningen er: *den ærlige kjæresten fortalte alt i boka*
>
> Informanten sier: *den ærlig kjæreste fortalt alt i boken*
>
> Vi skriver: *den ærlige kjæresten fortalte alt i boka*

Av og til hopper informanten over et ord i teksten. Dette merkes med ***+mangler*** foran ordet. Dersom informanten derimot legger til et ord som ikke står i teksten, skriver vi inn ordet med en ***+ekstra***-tagg foran.

> Setningen er: *den ærlige kjæresten fortalte alt i boka*
>
> Informanten sier: *ærlige kjæresten fortalte alt i boka*
>
> Vi skriver: ***+mangler*** *den ærlige kjæresten fortalte alt i boka*

Eller

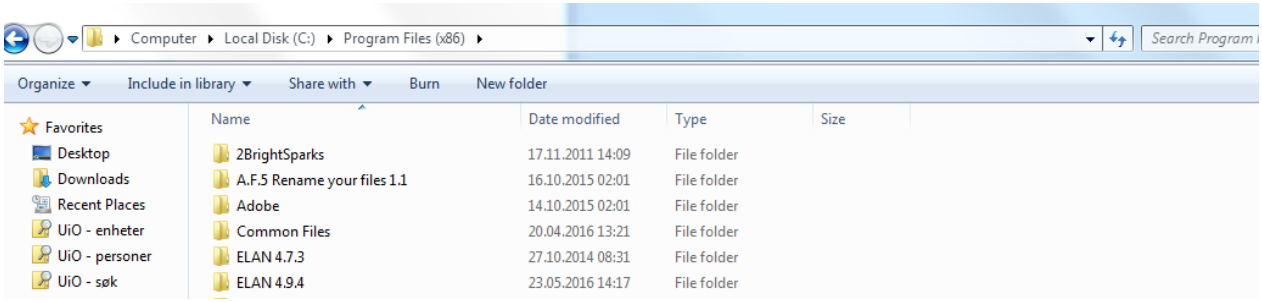
> Setningen er: *den ærlige kjæresten fortalte alt i boka*
>
> Informanten sier: *den den ærlige kjæresten fortalte alt i boka*
>
> Vi skriver: ***+ekstra*** *den den ærlige kjæresten fortalte alt i boka*

Pauser og nølelyder markerer vi ikke i den oppleste talen.

## 5 Når du begynner å jobbe på en ny maskin

### *5.1 Lage snarvei til ELAN på "taskbaren"*

Gå til **Computer** → **Local Disk (C:)** → **Program Files,** og åpne mappen som heter ELAN (velg den nyeste versjonen om det er flere å velge i):




Dra fila som heter **ELAN.exe** til taskbaren.

### *5.2 Fjern skjermbildesnuing i Windows:*

Ctrl + Alt + F12.

Velg **Advanced Mode** → **Options and Support** → **ikke** kryss av for **Hot Key Functionality**

### *5.3 Dersom du opplever at språket endrer seg fra norsk til engelsk (eller andre språk) mens du bruker ELAN:*

Gå til kontrollpanelet.

Klikk på **Change keyboards or other input methods**.

Klikk på **Change keyboards**. Under **General** kan du fjerne de språkene du ikke trenger. Dersom bare norsk blir stående igjen, blir problemet med språkskifte borte.

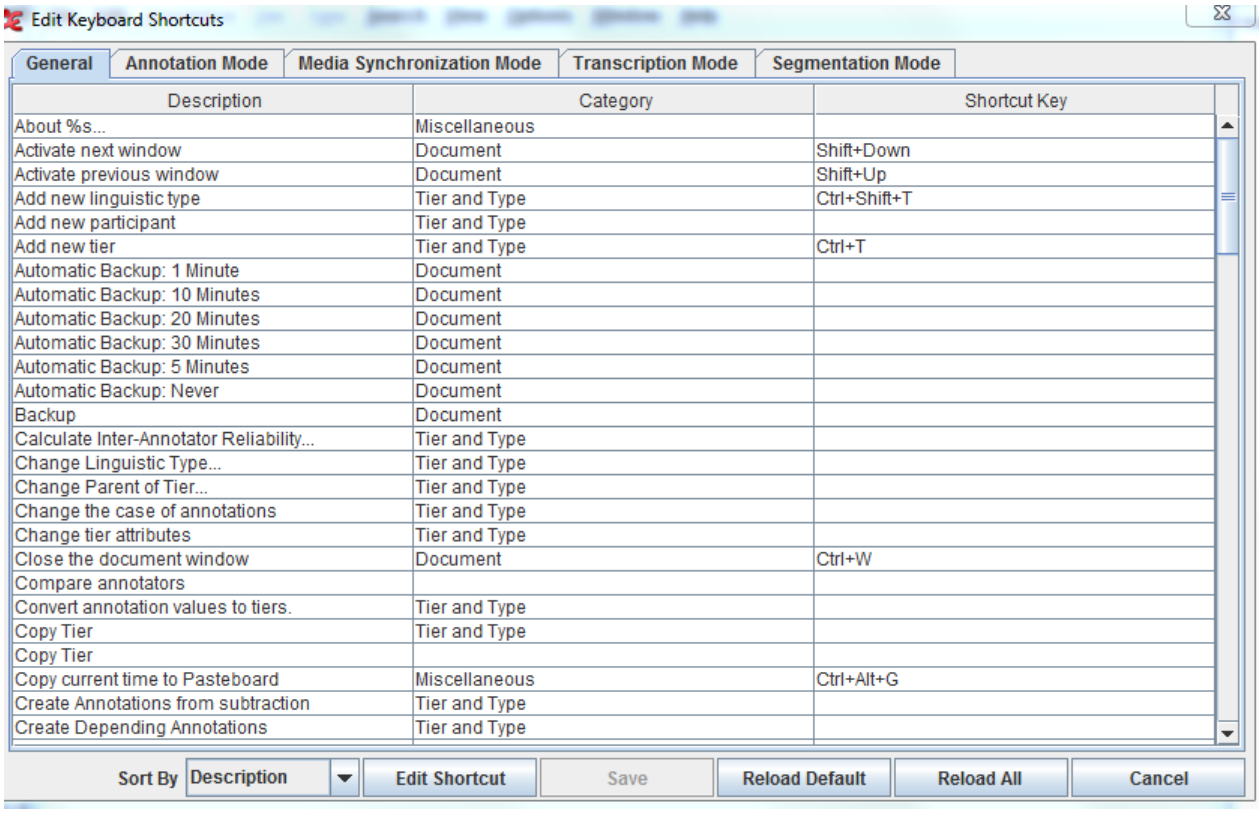
Eller:

Du kan også gå til **Advanced Key settings** og fjerne hurtigtasten (Left Alt+Shift) som endrer språket.

### *5.4 Definere egne snarveier:*

I ELAN er det mulig å definere egne tastatursnarveier for gitte funksjoner. Går du inn på

**Edit** → **Preferences** → **Edit Shortcuts**, får du opp en liste med snarveiene du kan velge egne hurtigtaster til:




Her kan du sette Ctrl + Enter til å dele opp uthevet segment (**Split Annotation**), slik som det står i listen over snarveiene i transkripsjonsveiledningen.