

# NB taledatabasen

– et innskudd i Språkbanken

*Lingit AS produserer en akustisk-fonetisk taledatabase som skal være en ressurs for forskning og utvikling av norsk talegjenkjenning. Prosjektet gjennomføres på oppdrag av Nasjonalbiblioteket. Sentrale stikkord for ressursen er dialektal/regional/idiolektal uttalevariasjon, fonetisk spredning og akustisk stabilitet.*

## **Modul 1 (ferdigstilles høsten 2013)**

- Totalt 240 innlesere
- 12 regionale grupper
- 20 innlesere pr. region
- Hver region er likt fordelt mht. alder (over/under 40 år) og kjønn

## **Modul 2 (ferdigstilles høsten 2014)**

- 12 grupper med totalt 140 innlesere:
  - 120 NO2-innlesere gruppert etter språklig bakgrunn
  - 20 NO1-innlesere
- Alle innlesere er over 18 år
- Alle grupper har minst 40% fra hvert kjønn

## **Tekstmateriale**

Alle innlesere leser 20 setninger. Setningene leses enkeltvis, og utgjør ikke en sammenhengende tekst. Manus pr. innleser er strukturert i tre deler:

- 3 a-setninger. Leses av alle 380 lesere
- 12 b-setninger. Leses av tre ulike innlesere (isolert pr. modul)
- 5 c-setninger. Leses av én innleser.

Tekstmaterialet i NB taledatabasen utgjør med dette 3423 unike setninger. Totalt vil NB taledatabasen ha 7600 setningsfiler.

Setningene er valgt ut fra et antall nyhetstekst-korpora, gjennom et såkalt grådig søk som baserer seg på et antall forventede fonemiske egenskaper.

## **Opptak og lydfiler**

Opptakene er med få unntak gjort i støydempet lydstudio. Alle opptak er i to separate kanaler, med kanal 1 til hodemikrofon og kanal 2 til studiomikrofon.

Lydfilene har 48 KHz punktprøvefrekvens, og 16 biters oppløsning.

## **Annotering**

Alle opptak er annotert på ord- og fonemnivå. Med annotering mener vi transkripsjon og tidsmerking. Det er brukt et tilpasset fonotypisk inventar med tegnbruk i henhold til X-SAMPA-standard.

Annoteringen er gjort på grunnlag av autogenererte merkefiler, i analyseprogrammet Praat.

## **Innleserdata**

Anonymisert informasjon om innleserne blir distribuert som en del av ressursen.