# Innsamling av språkdata i Norge for ELRC – utfordringer og muligheter

Jon Arild Olsen
seniorrådgiver
Nasjonalbiblioteket
jon.olsen@nb.no

# Oversikt

1. Nasjonalbibliotekets og Språkrådets roller i ELRC
2. Status for norsk språk i eTranslation
3. Oversettelse i offentlig sektor
4. Innsamling av språkressurser til eTranslation
5. Tiltak for nynorsk
6. Fremtidsplaner

# 1. Nasjonalbibliotekets og Språkrådets roller i ELRC

Nasjonalbiblioteket og Språkrådet fikk i 2017 i oppdrag å koordinere arbeidet med innsamling av språkdata til eTranslation.

- Språkrådet: ansvarlig for offentlig sektor ved Kristine Eide.
- Nasjonalbiblioteket: ansvarlig for språkteknologi ved Jon Arild Olsen.

Deltar på årlige styremøter i ELRC-konsortiet, samt månedlige videomøter med representanter fra nordiske og baltiske land.

# 2. Status for norsk språk i eTranslation

En utfordring:

- Norge er ikke EU-medlem – ingen oversettelsesminner fra Generaldirektoratet for oversettelse.
- Alle ressurser må skaffes lokalt.

Ca. 750 000 oversettelsesminner, hovedsakelig bokmål – engelsk.

- 732000 minner fra Utenriksdepartementets seksjon for EØS-oversettelse.
- 2000 oversettelsesminner fra Statens vegvesen.
- 9 000 oversettelsesminner fra Forbruker Europa.

Mindre enn 50 000 oversettelsesminner med nynorsk.

Trenger minst 300 000 minner for å bygge en oversettelsesmotor for nynorsk.

Heller ikke nok data mellom norsk bokmål og andre EU-språk til å bygge «direktemotorer».

Engelsk fungerer som «pivot language» for oversettelse mellom norsk bokmål og andre EU språk.

To mål:

- Øke det totale antallet oversettelsesminner bokmål – engelsk.
- Tilby nynorsk i eTranslation.

# 3. Oversettelse i offentlig sektor

Få offentlige virksomheter har interne oversettingstjenester som bruker profesjonelle oversettelsesverktøy.

Den klart viktigste er Utenriksdepartementet:

- Seksjon for oversettelse av EU-lover (ca. 19 stillinger)
- Seksjon for generelle oversettelser (ca. 6,5 stillinger).

I tillegg kommer:

- Sjøfartsdirektoratet (ca. 2 stillinger)
- Statens vegvesen (ca. 2,5 stillinger)
- EFTA-sekretariatet i Brussel (ca. 2,5 stillinger)

Noen offentlige virksomheter har oversettere, men anvender ikke profesjonelle oversettelsesverktøy.

De fleste offentlige virksomheter kjøper tjenestene av private oversettelsesbyråer.

Det som oversettes profesjonelt av statlige virksomheter er i stor grad tekster av juridisk og/eller teknisk karakter.

Tekstene som settes ut til private oversettelsesselskap, er mer varierte.

Offentlige virksomheter som kjøper oversettelsestjenester, kjenner ikke til eksistensen og/eller verdien av oversettelsesminner.

Nynorsk er lite brukt i oversettelser mellom norsk og andre språk utført av eller for offentlig sektor.

# 4. Innsamling av språkressurser til eTranslation

a) **Avtale med Amesto Translations, nå Semantix, om overdragelse av oversettelsesminner fra alle oppdrag for offentlige virksomheter.**

Ca. 1,2 millioner oversettelsesminner.

Nasjonalbiblioteket har påtatt seg juridisk ansvar for all videre bruk, inkludert håndtering av personopplysninger, konfidensialitet og intellektuelle rettigheter.

Nasjonalbiblioteket har gjennomgått materialet.

7% inneholdt så mange personopplysninger at det ble slettet.

Resten vil bli oversendt til eTranslation, men ikke for videre distribusjon.

**b) Avtale med Standard Norge om mottak av oversettelsesminner fra perioden 2011-17.**

Ca. 300 000 oversettelsesminner.

Anvendelsen skal ikke skade Standard Norges kommersielle interesser.

Løsning: «Scramble» oversettelsesminnene.

c) **Uni Computing har produsert bokmål/nynorsk – engelsk korpora basert på offentlige internettsider (NAV, Skatteetaten og Ny i Norge).**

Ca. 100 000 elementer.

**d) Tilde har lastet ned norske nettsider med paralleltekst:**

- ca. 800 000 elementer bokmål-engelsk
- ca. 8 000 elementer nynorsk-engelsk

Nasjonalbiblioteket har kontaktet ansvarlige institusjoner for å inngå avtaler om viderebruk av materialet.

De viktigste bidragsyterne er:

- Regjeringen (340 000)
- Equinor (110 000)
- Finanstilsynet (87 000)
- Norges bank (86 000)
- Norges musikkhøgskole (12 000)
- Nofima (10 000)
- Koro (10 000)

- Så langt er det totalt samlet inn 2 300 000 oversettelsesminner.
- Noe av det innsamlede materialet vil overlappe med de 750 000 minnene som allerede er levert til eTranslation.
- Uansett en betydelig økning, sannsynligvis nærmere en tredobling.
- Vi ligger fremdeles langt etter dansk, svensk og finsk som hver har ca. 5 250 000 oversettelsesminner fra DGT + lokale bidrag.

# 4. Tiltak for nynorsk

Nasjonalbiblioteket har så langt ikke identifisert store tilgjengelige kilder til paralleltekst nynorsk – engelsk.

Forslag fra eTranslation: gjøre bokmål til «pivot language» for nynorsk.

Behøver i så fall et stort parallelkorpus for bokmål og nynorsk.

To viktige kilder til parallelkorpus nynorsk-bokmål:

a) Nynorsk pressekontor

- oversetter nyhetssaker fra NTB til nynorsk.
- har inngått avtale med Nasjonalbiblioteket om overdragelse av 70 000 nyhetssaker.
- vil bli en viktig kilde til et variert og kontinuerlig oppdatert korpus.
- For å øke volumet trenger Nynorsk pressekontor ressurser til å videreutvikle den automatiske oversettelsestjenesten Apertium.

b) Lærebøker for norsk grunn- og videregående skole:

- Forskrift til opplæringsloven krever parallellutgaver av alle læremidler.
- Nasjonalbiblioteket mottar og digitaliserer alle publiserte læremidler som følge av pliktavleveringsloven.
- Vil utarbeide parallelkorpus basert på digitaliserte lærebøker for norsk skole.
- Har søkt CEF-programmet om midler.
- Foreløpig svar forventes innen nyttår.

# 6. Fremtidsplaner

a) Sammen med Irland, Island og Krotatien har Norge søkt CEF-programmet om støtte til tre prosjekter:

- Fjerne konfidensielt materiale fra minner fra Utenriksdepartementets seksjon for generelle oversettelser. Tre arbeidsmåneder. Ca. 750 000 oversettelsesminner.
- Kvalitetssikre oversettingsminner fra EFTA-sekretariatet i Brussel. En arbeidsmåned. Ca. 60 000 oversettelsesminner.
- Etablere et bokmål-nynorsk parallelkorpus basert på lærebøker fra grunn- og videregående skole. Ca. to millioner elementer.

- Inngå avtale med Sjøfartsdirektoratet om levering av oversettelsesminner.

- Inngå langsiktige avtaler med Nynorsk pressekontor, Semantix, Standard Norge og andre private aktører omlevering av språkressurser.

Forventet resultat innen 2021:

- Ca. 3 200 000 oversettelsesminner bokmål – engelsk
- Ca. 3 200 000 oversettelsesminner nynorsk – bokmål

# Takk for oppmerksomheten!