Hvorfor forstår ikke mobilen dialekta mi? Om utviklinga av grunnlagsressurser for norsk språkteknologi ved Nasjonalbiblioteket

Per Erik Solberg
Språkbanken ved
Nasjonalbiblioteket

Nasjonalbiblioteket

Språkbanken ved Nasjonalbiblioteket

- Oppretta i 2010 som et språkpolitisk tiltak
- Utvikler og tilgjengeliggjør grunnlagsressurser for norsk språkteknologi
- **Åpen lisens:** ressursene kan fritt brukes til forskning og kommersiell utvikling
- https://www.nb.no/sprakbanken/

Nasjonalbiblioteket

Hvorfor forstår ikke mobilen dialekta mi?

- Språkteknologi bruker ML-algoritmer trent på store datasett med tekst og/eller tale
- Datasettene bør inneholde variasjon (dialekt, alder, kjønn) og være tilpassa bruksdomenet
- En utfordring for norsk
  - Et lite språksamfunn og et lite marked
  - Stor dialektvariasjon, og dialektene brukes overalt
  - To skriftspråk

Hvorfor forstår ikke mobilen dialekta mi?

- Språkbanken deler datasett for språkteknologi:
  - laga av lingvister med kompetanse på norsk
  - med dialektal variasjon
  - på bokmål og nynorsk
  - tilpassa ulike formål

- Selvlagde datasett og datasett arva fra andre, bl.a. konkursboet til Nordisk språkteknologi

- Vi skal se på noen utvalgte ressurser

Plan

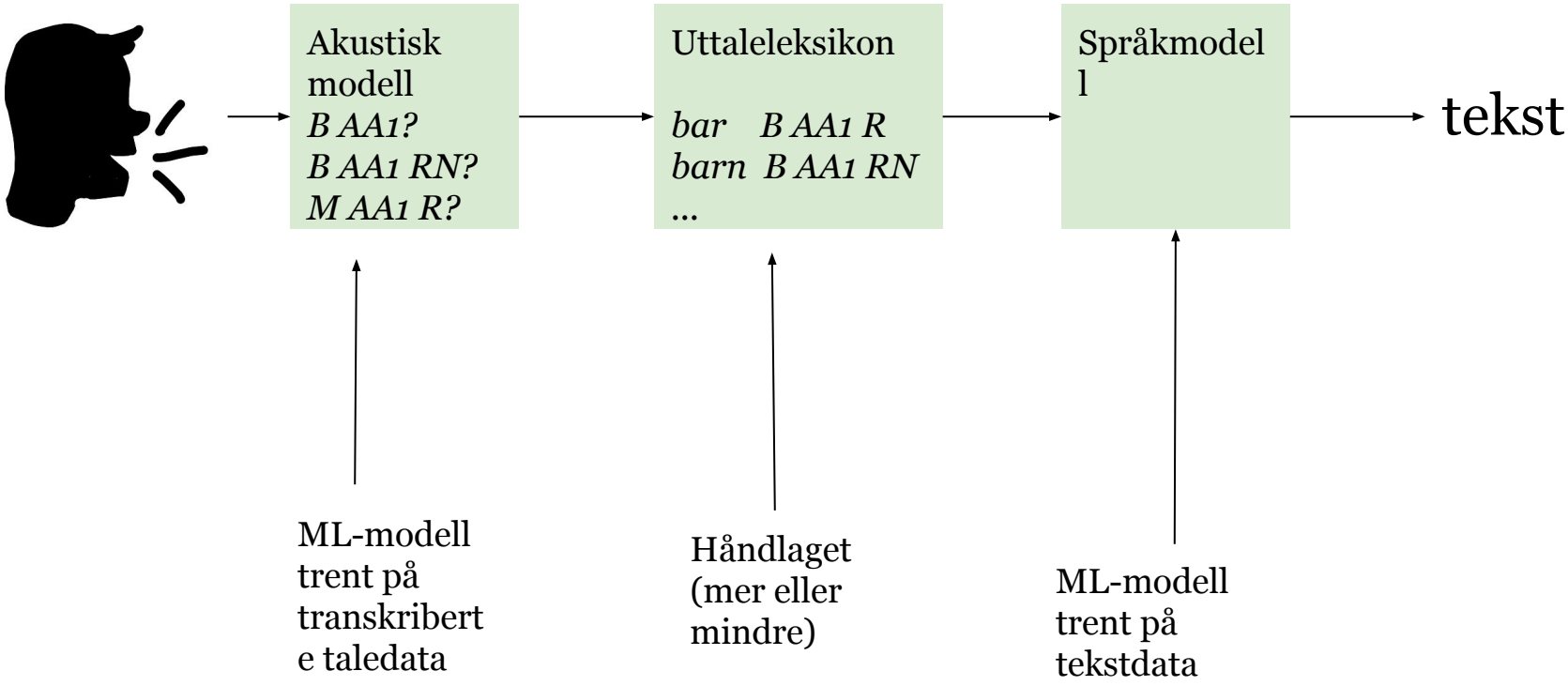
1. Talespråksressurser
2. Tekstressurser
3. Konklusjon

1. Talespråksressurser

- Taleteknologi:
  - talegjenkjenning
  - talesyntese
  - talerindentifikasjon
  - taleassistenter (Google-assistenten, Siri etc.)

- Case: Talegjenkjenning

Talegjenkjenningsalgoritmer



Talekorpus

- Treningsdata for akustiske modeller
- Opptak av tale + ortografisk transkripsjon + metadata om talerne
- Variasjon er viktig for god talegjenkjenning (kjønn, alder, dialekt)
- Man trenger både generelle og domenespesifikke taledata (diktering, foredrag, taleassistenter...)

Eksisterende talekorpus i Språkbanken

- **Talekorpus for talegjenkjenning fra Nordisk språkteknologi**
  - 982 talere fra forskjellige deler av landet
  - 540 timer taleopptak
  - Oppleste setninger
  - God ressurs for grunnleggende talegjenkjenning

- **Spesialiserte talekorpus:** diktering, telefonkvalitet, talesyntese, korpus med fonetisk transkripsjon

Stortingstranskripsjonene

- ortografisk transkripsjon av stortingsmøter fra 2017 og 2018
- Fritt tilgjengelige taledata, detaljerte referat, mye metadata om talerne, stor dialektvariasjon
- Velegna for talegjenkjenning av foredrag o.l.
- Stortinget skal utvikle talegjenkjenning

Uttaleleksikon

- Essensiell ressurs, men dyr å utvikle
- Utviklinga kan i noen grad automatiseres, men krever mye håndsøm fra lingvister
- Språkbankens uttaleleksikon:
  - ca. 800 000 ord
  - utvikla på slutten av 90-tallet
  - bare østlandsk
- Prosjekt i 2021: Utvide med 4 nye dialekter + nyord

2. Tekstressurser

- **Natural Language Understanding (NLU)**
  - Automatisk gjenkjenning av meningsinnholdet i tekst
  - prateroboter, konversjon fra løpende tekst til strukturerte data, sentimentanalyse, emneklassifisering

- **Ustrukturerte treningsdata:** samling med tekst uten oppmerking
- **Strukturerte treningsdata:** tekst med forskjellige former for oppmerking (annotasjon)

Norsk dependenstrebank

- Annotert tekstkorpus laga av Språkbanken i 2011-2013
- *Dependenstrebank:* tekstkorpus med grammatiske relasjoner mellom ord
- 600 000 ord (50/50 bokmål/nynorsk)
- Flere lag med grammatisk annotasjon
- Manuelt annotert
- https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-10/

Ordklasser

Dette er en setning med ordklassetagger
pron verb art subst prep subst

- Ordklassetaggere (POS-taggere): trent på manuelt POS-tagga tekstkorpus
- Ordklassetagging (*POS-tagging*): utgangspunkt for syntaksparsing, navnegjenkjenning, informasjonsekstraksjon etc.
- All tekst i NDT er manuelt POS-tagga

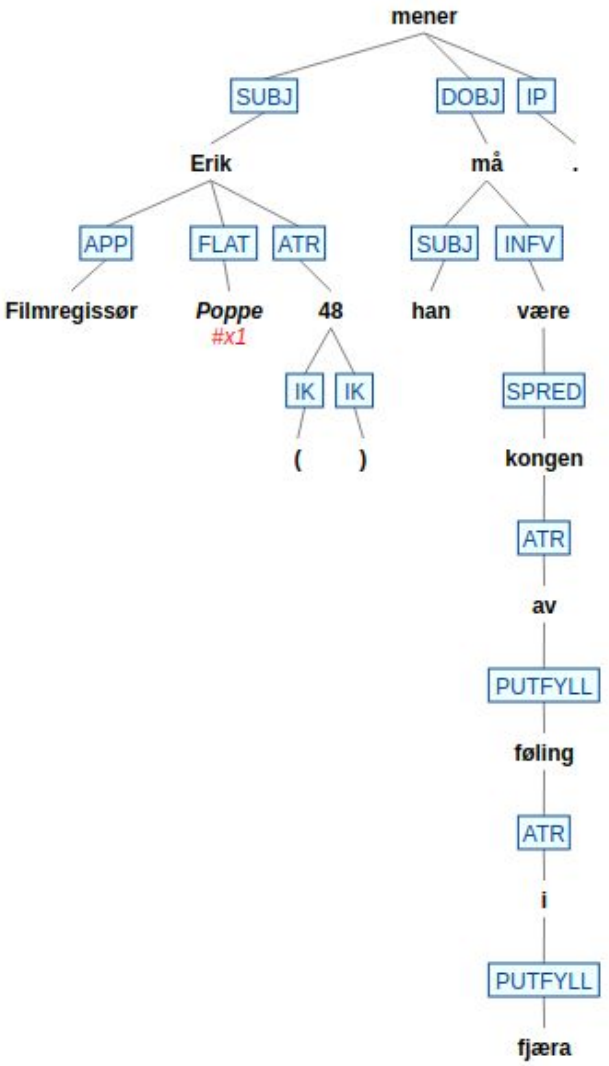
Syntaktiske relasjoner

Filmregissør Erik Poppe (48) mener han må være kongen av føling i fjæra.

- NDT: treningssett for syntaktisk parsing
- Relasjoner er viktig informasjon for forståelse:

BMW imponerer, men Tesla innfrir ikke

- Chunking: *Filmregissør Erik Poppe (48)*



Norwegian Named Entities

- **NorNE:** Lag med navneannotasjon bygd oppå NDT
- Samarbeidsprosjekt mellom Språkbanken og Schibsted og språkteknologigruppa ved UiO
- Alle navn i NDT er merket opp og klassifisert

Norske Gunnar Kolås som leier i Torrevieja mener tiden for å kjøpe nærmer seg
person
sted

Norwegian Named Entities

- NorNE: treningsmateriale for navnegjenkjenning (Named Entity Recognition)
- NER-modellen i NLP-pakka Spacy er trent på NorNE
- NER: sentral NLU-oppgave
  - Spill Lady Gaga på Spotify!
  - Hvem er statsminister i Mongolia?
- https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-49/
- https://github.com/ltgoslo/norne

Konklusjon

- God språkteknologi krever gode språklige datasett med variasjon
- Kostnadskrevende å utvikle for norsk
- Språkbanken utvikler og deler slike datasett med åpen lisens
- Vi er veldig glade for spørsmål, tilbakemeldinger og forslag til nye ressurser

sprakbanken@nb.no


https://www.nb.no/sprakbanken/
Nasjonalbiblioteket