



Sprogteknologi.dk – Visioner, udfordringer og fremgang

November 2021



DIGITALISERINGSSTYRELSEN

Agenda



Sprogteknologi.dk



Fremadrettet arbejde for sprogteknologi



Samarbejde med EU



Hvordan løfter vi dansk sprogteknologi

» Initiativet sprogteknologi.dk



- Fællesoffentligt samarbejde
- **Formål – at styrke udviklingen af løsninger inden for dansksproget kunstig intelligens**
- Målgrupper – sprogteknologiske virksomheder samt slutbrugerne i form af myndigheder, forskere, virksomheder og borgere
- Samarbejde med private og offentlige aktører

Sproressourcer



Dansk sprogteknologi er i udvikling

Analysebureau: Anerkendelse af angreb på Facebook

SPROGMODELLEN AELÆCTRA VIL FORBEDRE DANSK SPROGTEKNOLOGI PÅ EN KLIMAVENLIG MÅDE

Superalgoritme kortlæst og afslører yndlingsopførelser
Kommune: Dorte gennemgår samtaler og træner AI'en manuelt



rt mere end

Her er robotten Bridget. Hun skal læse tv-nyheder op for bornholmerne
Danish Gigaword

Danske had book

Ny skandinavisk NER model med S₀



»» Hvor ser vi udfordringerne for dansk sprogteknologi?



Dansk vil altid være et mindre udbredt sprog på verdensplan



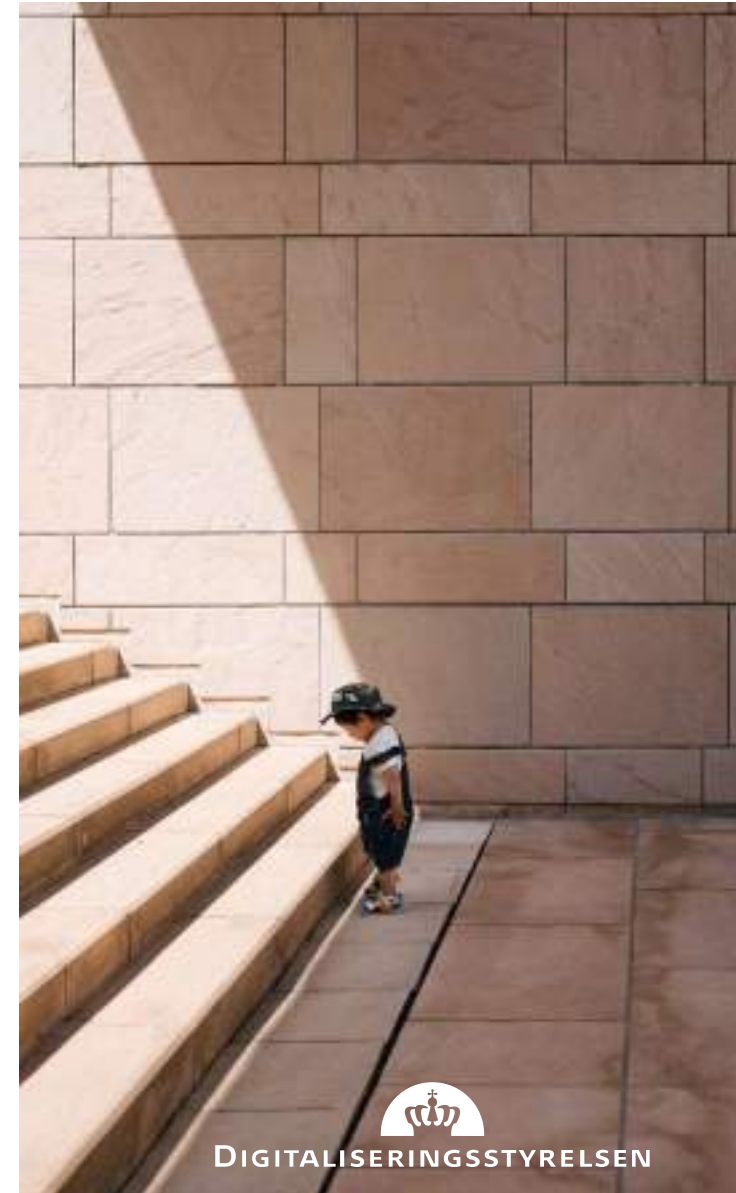
Behov for flere sproressourcer, særligt på lyd



Ressourcekrævende at udvikle eller ibrugtage dansk sprogteknologi



Vedvarende og løbende kvalitetssikring af eksisterende løsninger



» Arbejdsprogram 3.0



Sprogteknologi.dk

Portal for den fælles danske sprogressource

Over 120 ressourcer på portalen

- Datasæt
- NLP værktøjer
- NLP modeller



Centralt Ord Register (COR)

Samarbejde med Dansk Sprognævn, Center for Sprogteknologi ved KU, Det Danske Sprog- og Litteraturselskab



Taleteknologi

Et større datasæt (talekorpus med lyd, tekst og annotationer i høj kvalitet og frit tilgængeligt.

Analyse af mulighederne for samarbejde med interessenter om at udarbejde datasæt baseret på lydbogsmateriale

» Arbejdsprogram 3.0



Standarder

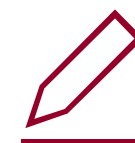
- For licensvalg
- Opmærkning
- Strukturering af data
 - Metadata
 - Annotation



Forbinde miljøet

Flere events og begivenheder

Deltagelse i følgegrupper på sprogteknologiske projekter



Guide til samtykke

Gøre det nemmere for udviklere og anvendere at indsamle den fornødne data in-house

Europæiske arbejde

Deltagelse i forskellige initiativer

- European Language Equality (ELE)
- European Language Resource Coordination (ELRC)
- European Language Grid (ELG)

Arbejder for at dansk bliver tilgodeset i EU

- Digital Europe Programme 2022 - 2023 → Language Data Spaces
- Løbende inputs til CEF platformen, særligt eTranslation
- Følger med i EU projekter, der indeholder AI relateret sprogteknologi



Hvordan løfter vi dansk sprogteknologi?

Del jeres
sprogressourcer

Tænk over
fremtidig
indsamling af
sprogressourcer

Open-source
så meget som
muligt



DIGITALISERINGSSTYRELSEN



Christian Plaschke, kontorchef
info@sprogteknologi.dk



DIGITALISERINGSSTYRELSEN

Vores scope for sprogressourcer

Sprogressourcer

I scope af sprogteknologi.dk

Sprogdata

Talekorpora
Tekstkorpora
Ordbøger
Termbaser

Sprogteknologisk supportværktøj

Dataindsamlingsværktøjer
Anonymiseringsværktøjer
Værktøjer til automatisk opmærkning
Værktøjer til manuel opmærkning
Modeltræningsværktøjer
(...)

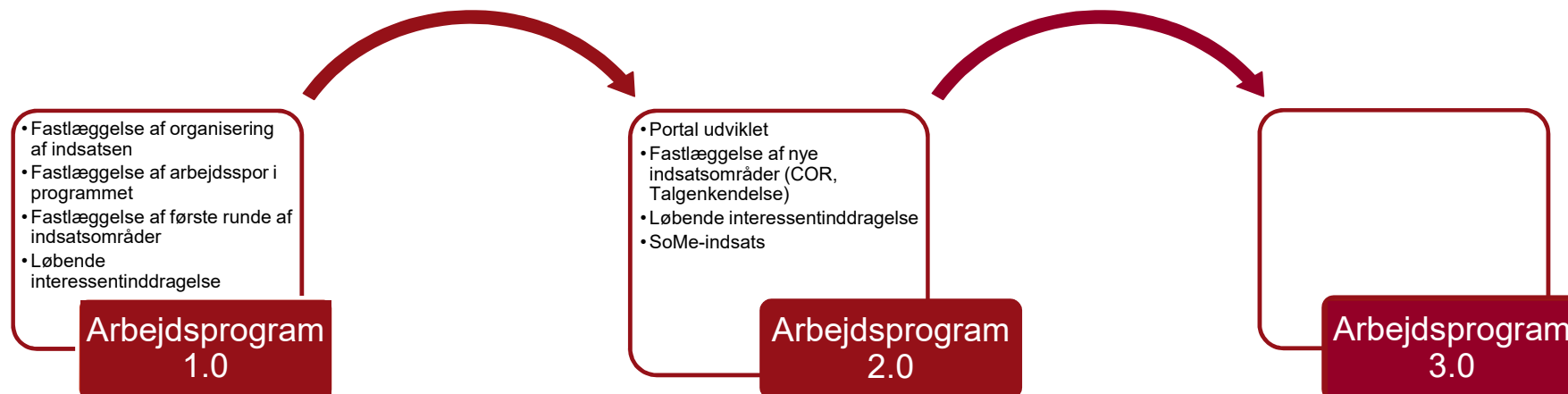
Sprogteknologisk infrastrukturkomponent

Talegenkendelseskomponent
Talesyntesekomponent
Sprogforståelseskomponent
Maskinoversættelseskomponent
(...)

Slutbruger softwareløsning

Tale-til-tekst
Tekst-til-tale
Dialogsystem
Maskinoversættelsessystem
Læse-/skriveværktøj
(...)

Organiseringen af arbejdet



» Initiativet sprogteknologi.dk



- Fællesoffentligt samarbejde aftalt i Digitaliseringspagten og ØA20
- **Formål – at styrke udviklingen af løsninger inden for dansksproget kunstig intelligens**
- Målgrupper – sprogteknologiske virksomheder samt slutbrugerne i form af myndigheder, forskere, virksomheder og borgere
- Samarbejde med både private og offentlige aktører

Bidrag til dansk sprogteknologi

- Flere sprogressourcer også gerne tekst
- Scrapes af hjemmesider, hvis der forelægger oversat udgave
- Hav en open-source tilgang til sprogmodeller, værktøjer og sprogressourcer → Del så meget som muligt
 - Sprogressourcer har oftest stor værdi, hvor vi mindst venter det

