Dansk Sprognævn

Udvikling af sprogteknologi i Danmark

Hvilken rolle spiller de offentlige institutioner?

Sabine Kirchmeier

Oversigt

- Hvorfor har maskiner så svært ved at håndtere dansk?
- Hvordan fungerer sprogteknologi?
- Hvad kan offentlige institutioner bidrage med?
- Det sprogteknologiske udvalg.

Dansk Sprognævn

Hvorfor har maskiner så svært ved dansk?

Sprog er svært

- **Sprog er flertydigt på mange måder**
  - Ord der betyder det samme (*hest – krikke*)
  - Ord der staves ens og betyder noget forskelligt (*skimmel*)
  - Ord der skifter betydning alt efter situationen (*din gamle buk - den gamle buk - springe buk - buk hjørnet om - buk dig*)
  - Sætninger er flertydige (*den gamle elsker piber*).
- **Sprog er tilsyneladende ustrukturerede data** – men der er en struktur i sproget.
- **Sprog er styret af konventioner som kan være forskellige fra tekst til tekst og fra situation til situation** – fx e-mail vs. opslag på Twitter vs. Facebook**.**
- **Sprog er knyttet til vores identitet** – fx er der ikke to mennesker der taler og skriver på samme måde.
- **Sprog er uendeligt** – både når det gælder antallet af elementer og deres kombinationsmuligheder.
- **Sproget ændrer sig hele tiden** – nye ord (*disruption*) – nye betydninger (*adressere*).
- **Sprog er en resurse der vokser i takt med at den bliver brugt.**

Hvorfor er dansk så svært?

- Mange vokaler og uigennemskuelig udtale af dem (*fisk*– 'fesk', *hest* – 'hæst').
- Tryk signalerer fx processer: *læse a'vis - læse avisen.*
- Stød signalerer forskellige betydninger: *stien* (+stød) – *stigen* (-stød).

- Sammensatte ord kan dannes i det uendelige: *borgmesterkædedans.*
- Mange små partikler med forskellige betydninger: *skrive af/afskrive - tale ud/udtale.*

Ordstilling mere fleksibel end på engelsk, men mindre fleksibel end på andre sprog:

F	v	s	a	V	S	A
Jeg	*har*		*ikke*	*hentet*	*bogen*	*i dag*
I dag	*har*	*jeg*	*ikke*	*hentet*	*bogen*	
Bogen	*har*	*jeg*	*ikke*	*hentet*		*i dag*

Rekursive konstruktioner: *Denne bog går der rygter om at du har læst _.*

Dansk Sprognævn

Hvordan fungerer sprogteknologi?

Sprogteknologi er en forudsætning for kunstig intelligens.

Intelligens

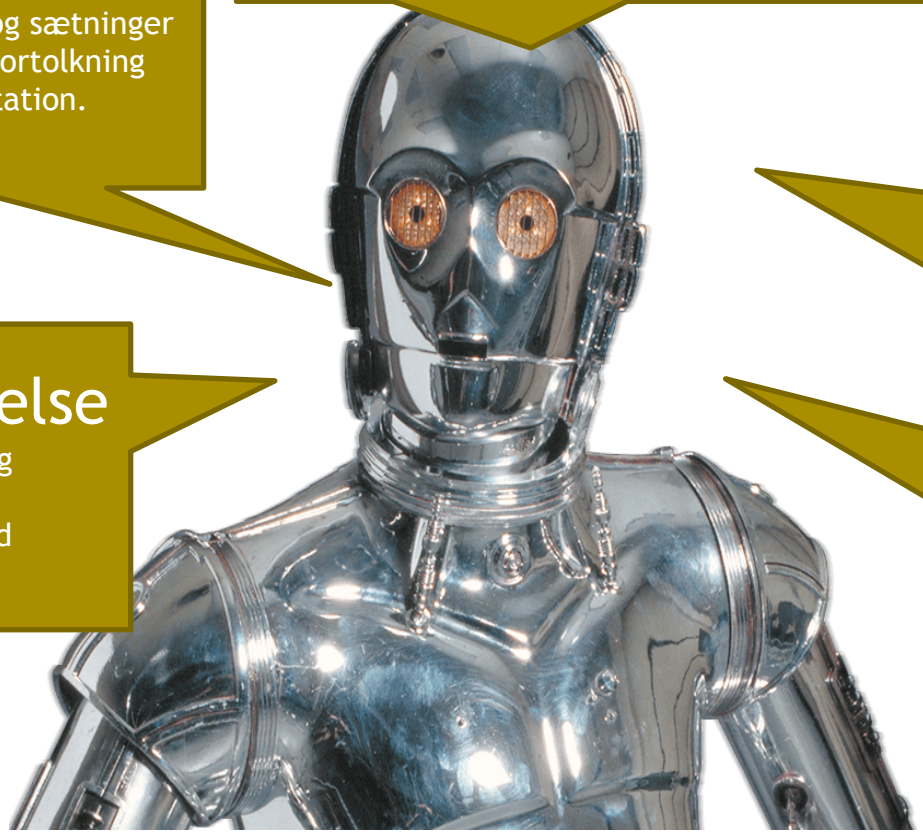
- Indholdsanalyse baseret på viden om genre, fakta, logik, følelser, ironi og humor
- Ræsonnementer og konklusioner baseret indholdsanalyse og betydningsrepræsentationer.

Sprogforståelse

- Genkendelse af ord og sætninger
- Sætningsanalyse og fortolkning
- Betydningsrepræsentation.

Oversættelse

- Transformation af betydning til målsprogets struktur
- Generering af sætninger.



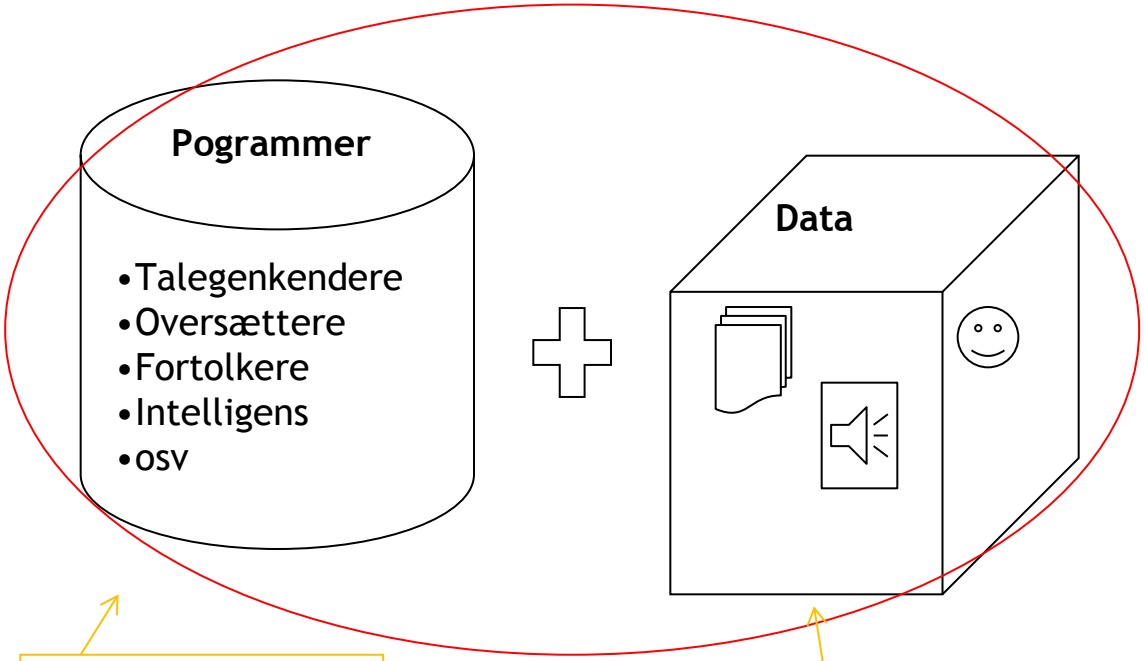
Talegenkendelse

- Identifikation af sprog
- Genkendelse af lyd
- Repræsentation af lyd
- Sprogmodeller.

Talesyntese

- Transformation af sætninger og ord til lyd
- Genskabelse af meningsfuldt tryk og prosodi.

Sprogteknologiens basisbyggesten



Produkter:
Siri
Google Translate
Alexa
Dialogsystemer
Oversættelse
Stavekontroller
Søgemaskiner
osv.

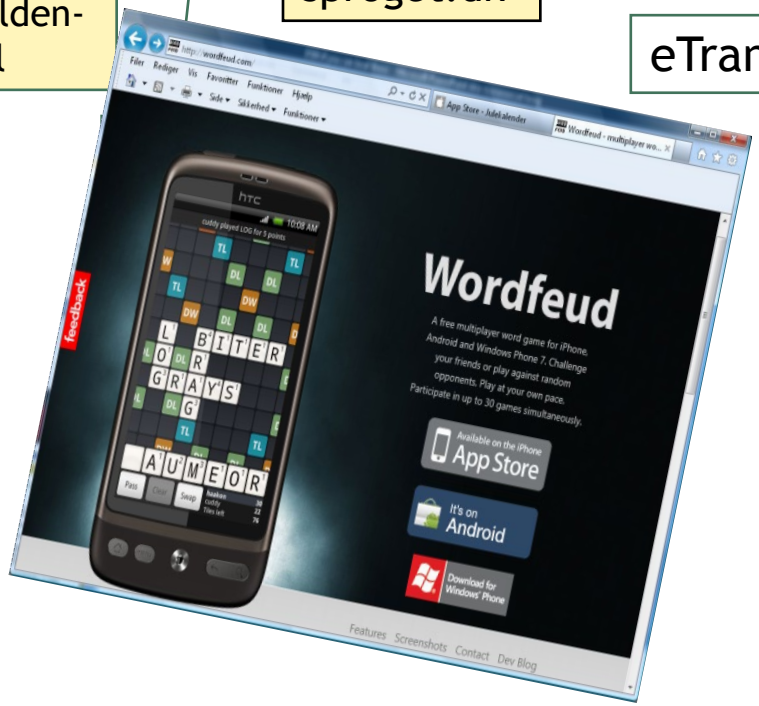
Udvikles af forskere og producenter.

Indsamles af forskere og producenter i virksomheder og i offentlige institutioner.

Retskrivningsordbogen er også en dataresurse

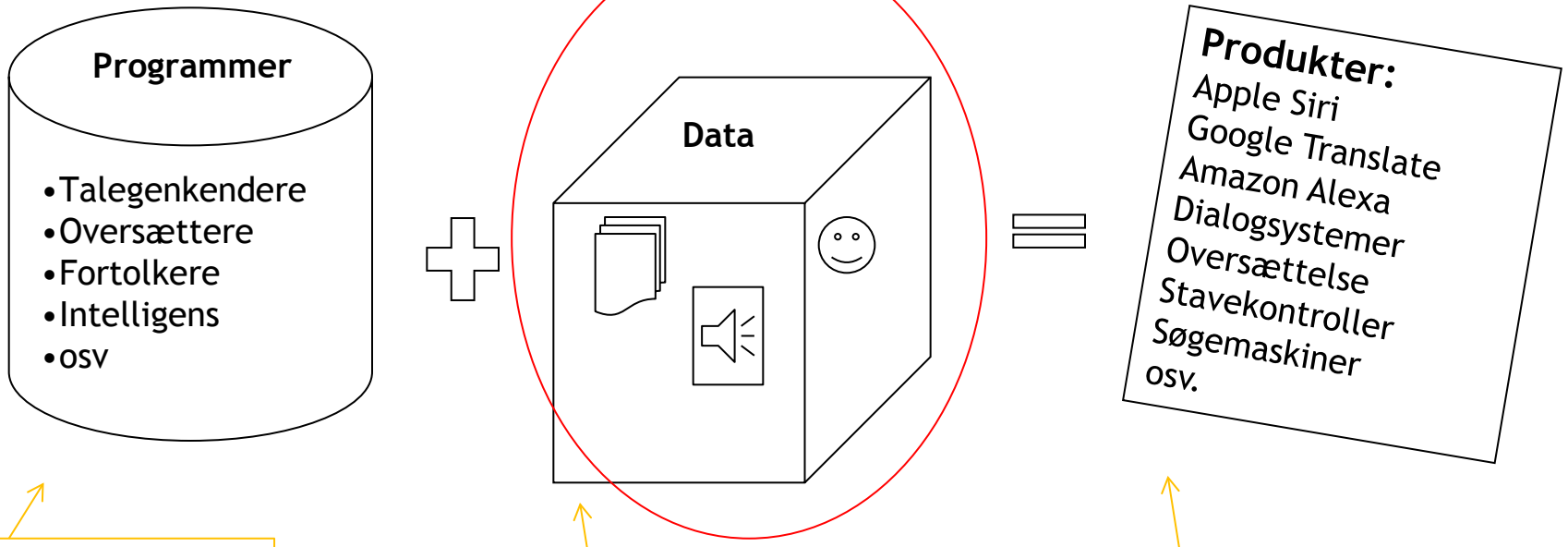
Ordbogsbasen

- Bog
- Web
  - Ordb. com
  - Gylden-dal
  - sproget.dk
- App
- Ordliste
  - Spiludviklere
  - Aviser/tv
  - KMD
  - eTranslation
  - Microsoft
  - Sundheds-platformen



Stavekontrol, søgemaskiner, højresider i tosprogsordbøger, spil, kontrol af taleinput, orddeling, undervisningsprogrammer, generering af brugernavne, osv.

Hvem ejer data?

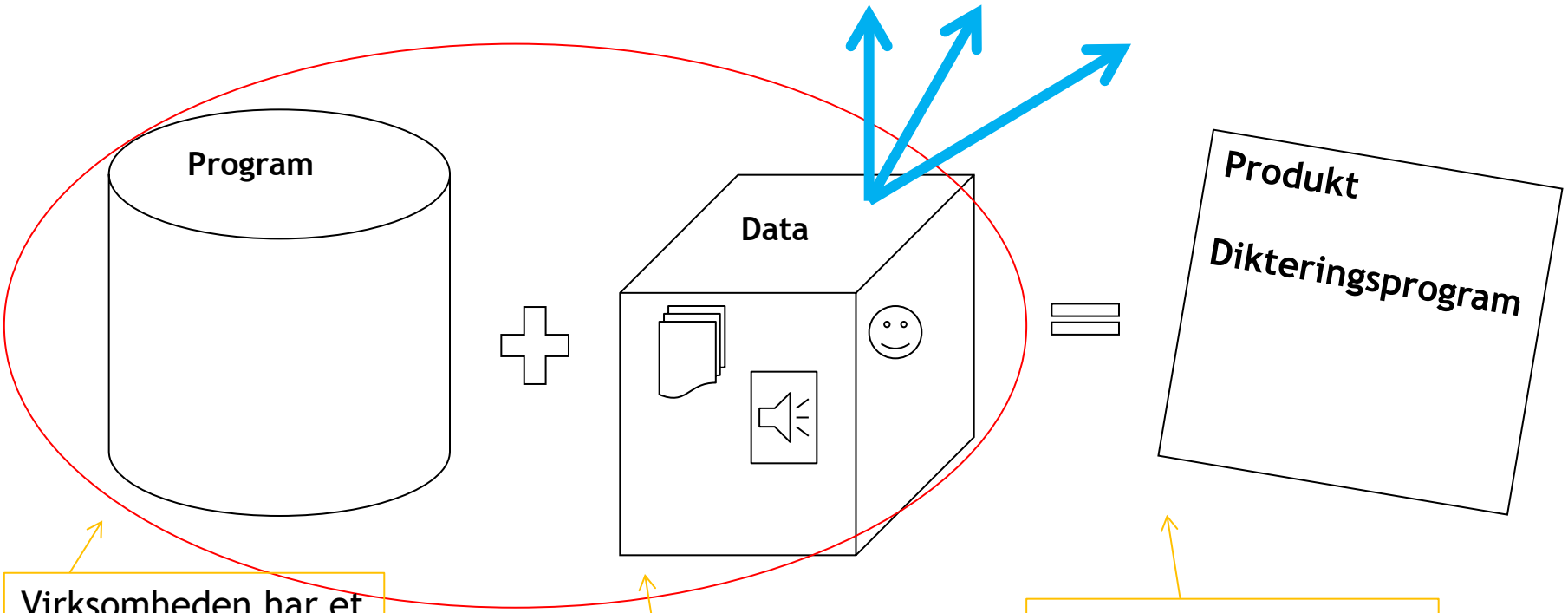


Ofte virksomheder, men kernen er tit open source.

Oftest virksomheder. Få data er open source.

Monopolisering.
Ringe muligheder for konkurrenceudsættelse.
Ingen genbrug af data.

Mangel på konkurrence – et eksempel.

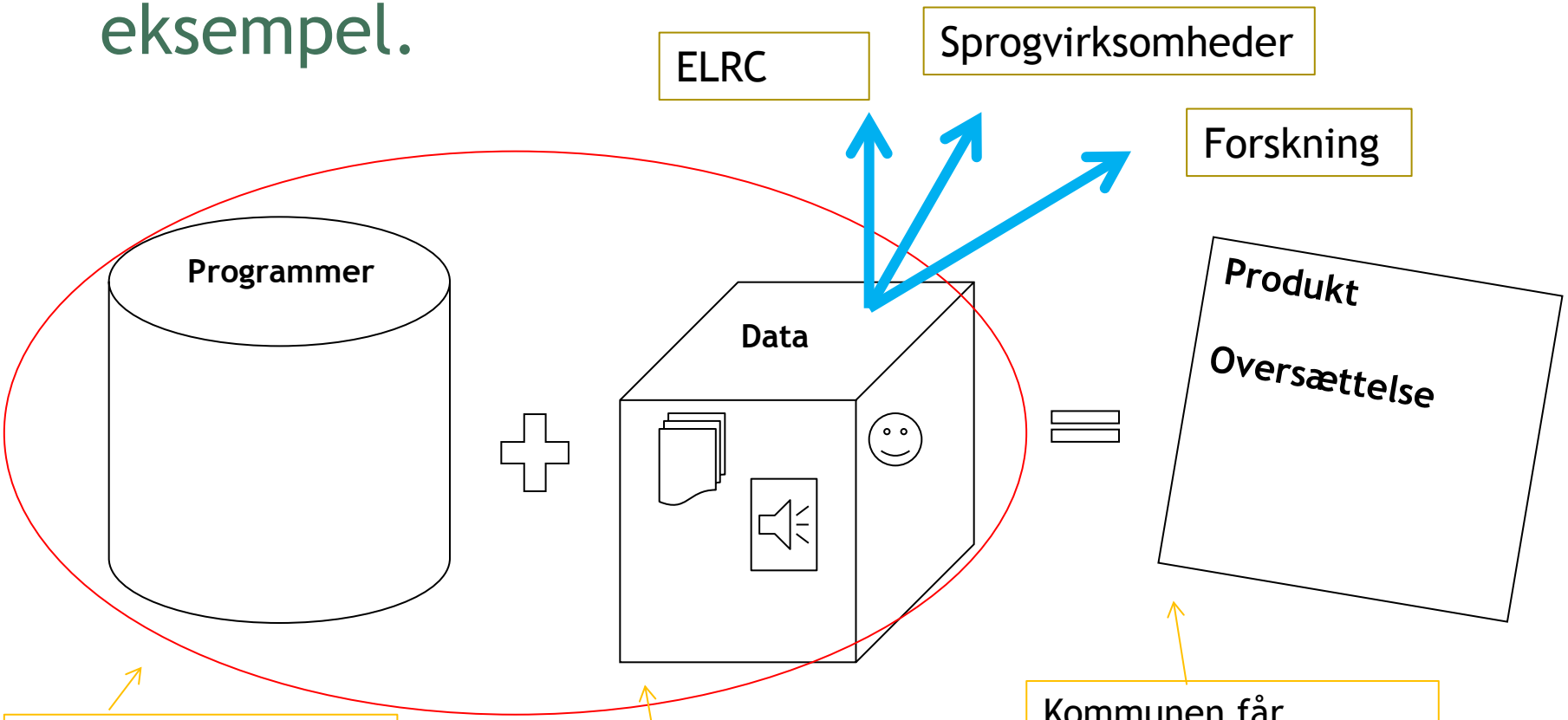


Virksomheden har et talegenkend-elsesprogram for engelsk som tilpasses dansk.

Virksomheden får data fra en dansk kommune, fx tekster og lydoptagelser.

Virksomheden bearbejder data og sælger produktet til kommunen.

Mangel på konkurrence – et andet eksempel.



Oversættelses-bureauet bruger oversættelses-hukommelse.

Kommunen giver virksomheden en del tekster til oversættelse.

Kommunen får oversættelsen, men ikke den værdifulde oversættelses-hukommelse og holder ikke snor i de oversatte tekster.

Sprogresurser (data) kan genbruges uendeligt mange gange

- Derfor giver det mening at oprette en offentlig sprogbank hvorfra virksomheder og forskere kan hente data.
- Derfor er det vigtigt for offentlige institutioner at sikre sig ejerskab til data så de kan konkurrenceudsætte sproglige produkter og stimulere produktudvikling og forskning.

- Jo flere data der bliver offentligt tilgængelige, jo bedre sprogteknologi kan der udvikles.
- Dataindsamling og bearbejdning er dyrt. Der er mange penge at spare ved at dele.

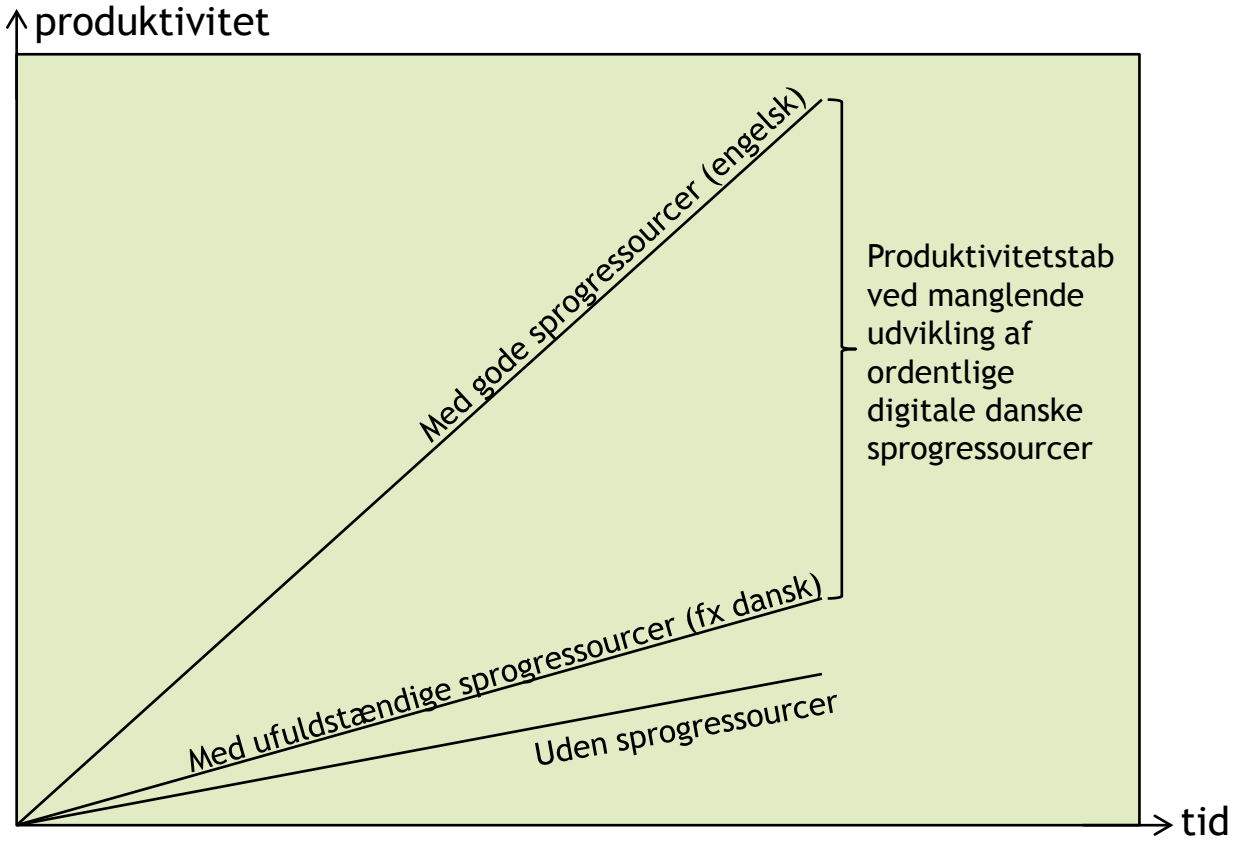
Men det sker ikke.

Dansk halter bagefter

Det skyldes:

- At dansk er et vanskeligt sprog for systemer der i første omgang er udviklet til engelsk.
- At dansk er et lille marked med få aktører og stærke monopoltendenser.
- At der ikke har været satset tilstrækkeligt på forskning og udvikling af sprogteknologi for dansk.
- At der ikke forefindes tilstrækkeligt med sprogresurser som forskere og virksomheder kan bruge.

Manglende og dårlige digitale sprogressourcer på dansk betyder produktivitetstab



Dansk Sprognævn

Kulturministerens sprogteknologiske udvalg

Kulturministerens sprogteknologiske udvalg

- Placering i Dansk Sprognævn
- Varighed 1.1.2018 – 1.8.2019.

Formål

- at udrede perspektiver og udfordringer for sprogteknologi i en dansk kontekst
- at komme med forslag til hvordan Danmark bedst sikrer brugen af dansk og andre sprog i digitale tjenester
- at afklare behovet og perspektiverne for en national termbank.

- CTO Klaus Akselsen, MIRSK
- Linguist, Partner, Head of Research and Innovation Esben Alfort, Ankiro ApS
- Udviklingschef Lars Fremerey, GTS-foreningen
- Computational Linguist Anna Katrine Jørgensen, Google
- Sekretariatschef Jens Kellerup, Ballerup Kommune/OS2 – (Offentligt digitaliseringsfællesskab)
- Direktør Sabine Kirchmeier, Dansk Sprognævn (formand for udvalget)
- Direktør Jens Otto Kjærum, Dictus
- Professor Bodil Nistrup Madsen, CBS – Copenhagen Business School
- Seniorredaktør Sanni Nimb, Det Danske Sprog- og Litteraturselskab
- Professor Bolette Sandford Petersen, Center for Sprogteknologi, Københavns Universitet
- Forsknings- og Innovationsdirektør Anders Quitzau, IBM Research – Watson Advocate
- Founder, Chief Visionary Officer Mads Rydahl, Unsilo
- Kontorchef Jens Krieger Røyen, Digitaliseringsstyrelsen
- Chefkonsulent Carl Østergaard, Odense Kommune

Kommissorium

- Med udgangspunkt i et brugerorienteret perspektiv kortlægge de nuværende og fremtidige behov for at benytte dansk og andre sprog samt sproglig viden i forhold til digitale tjenester og applikationer baseret på kunstig intelligens i centrale sektorer af samfundet.
- Vurdere i hvilket omfang det vil være muligt at imødekomme disse behov under inddragelse af de relevante spillere i erhvervslivet, den offentlige sektor samt uddannelses- og forskningssektoren.
- Afklare behovet og perspektiverne for en national termbank (”sprogtermbank”) og inddrage resultater fra arbejdet med dansk terminologi og danske vidensbaser samt med begrebs- og datamodellering i det fællesoffentlige digitale arkitektursamarbejde.
- Inddrage relevante resultater fra arbejdet med sprogteknologi og terminologi i andre lande, herunder EU og Norden,
- Pege på måder hvorpå en styrkelse af dansk sprogteknologi vil kunne gavne den enkelte borger og bidrage til at skabe vækst og effektivisering i samfundet.

Kommissorium

- Levere en rapport der udreder behovet for sprogteknologi inden for centrale sektorer.
- Bidrage til oplysning og offentlig debat om sprogets rolle i kunstig intelligens og ny teknologi.
- Inddrage offentlige institutioner, virksomheder, brancheforeninger, fagforeninger, fageksperter og borgere med henblik på at sikre at så mange aspekter som muligt bliver belyst.

Organisering af udvalgsarbejdet

- Udvalgsmøder
- Vidensindsamling
  - **Workshops**
  - **Spørgeskemaundersøgelser**
  - Interviews med/besøg hos udvalgte grupper/personer
  - Indsamling af kommentarer via bloggen sprogtek2018.dk
  - Gennemgang af rapporter fra EU og Norden
- **Seminar**
- **Afsluttende konference**

ELRC-initiativet indgår i udvalgets arbejde

- Vi håber at offentlige institutioner bliver mere opmærksomme på at deres sproglige data i form af tekstdokumenter og oversatte dokumenter repræsenterer en stor værdi.
- Vi håber at offentlige institutioner vil begynde at identificere data som kan deles, og at de vil dele dem med EU-kommissionen og evt. også med andre udviklere af sprogteknologi.
- ELRC vil gerne hjælpe med at rense og pakke data og udrede fx juridiske og GDPR-relaterede spørgsmål.
- EU gør en stor indsats for at sikre at alle EU-sprog kan bruges i den nye teknologi.

- Men vi i Danmark bliver også nødt til at bidrage – først og fremmest med vores data.
- Offentlige institutioner kan blive en vigtig drivkraft i denne proces.


retskrivning sb., -en, -er, i
sms. retskrivnings-, fx ret-
skrivningssystem.
retslig (el. retlig) adj., -t.
retslæge sb., -n, -r.
retslægeråd sb., -et, rets-
lægeråd, bf. pl. -ene.
retslærd adj., itk. d.s.
retsløs adj., -t.