Dansk Sprognævn

Status for sprogteknologi i forskning, udvikling og uddannelse i Danmark

Sabine Kirchmeier
sabine@dsn.dk

Oversigt

- Hvad er status for udvikling af sprogteknologi på dansk?
- Hvad er status for uddannelse og forskning?
- Hvordan har man hidtil arbejdet strategisk med sprogteknologi?
- Hvad har man gjort i andre lande?
- Anbefalinger til et sprogteknologisk udvalgsarbejde

Hvorfor interesserer Dansk Sprognævn sig for sprogteknologi, kunstig intelligens og big data?

- Sprogteknologi på dansk skal følge de regler som gælder for dansk.
- Det dansk som vi møder når vi bruger fx kunstig intelligens, skal være korrekt.

- Sprognævnet bruger selv sprogteknologi og arbejder med store datamængder.
- Sprognævnets repræsentantskab har nedsat et fagråd for fagsprog og sprogteknologi.

- Vi har tre bekymringer:
  - At de nye produkter der udnytter big data og kunstig intelligens, ikke udvikles for dansk - eller ikke hurtigt nok.
  - At den sproglige kvalitet i de danske produkter bliver for ringe.
  - At den negative udvikling accelererer - især når vi begynder at bruge kunstig intelligens i større stil.

Robotteknologien og kunstig intelligens er her allerede.

Robotter i samspil med børn, ældre og handicappede

Kunstig intelligens og tingenes internet i hjemmet

Automatisk oversættelse bruges allerede i offentlige sektor, fx i EU's klageportal.

Klageportalen ODR – Online Dispute Resolution

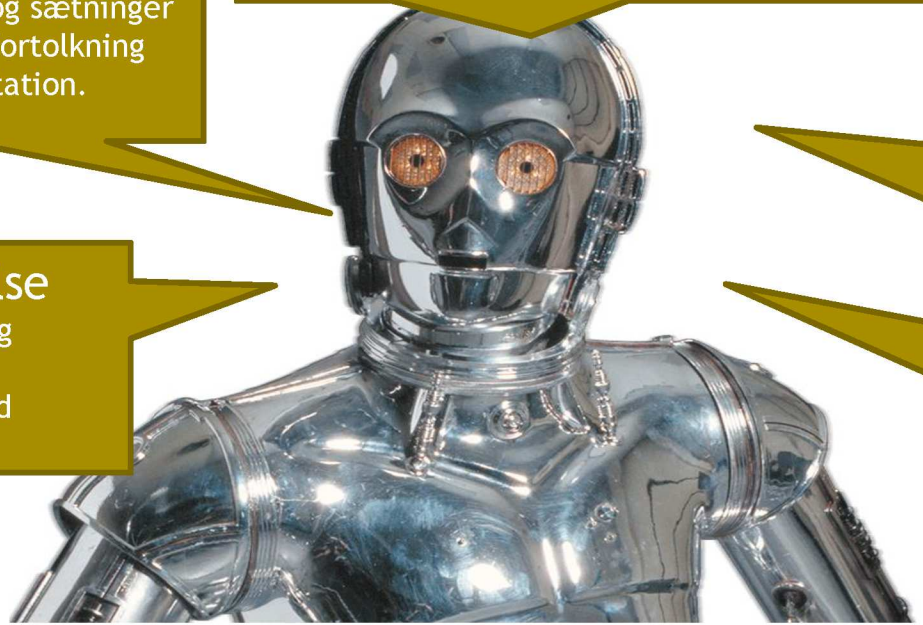
- Servicerer borgere på deres eget sprog – med EU's eget oversættelsessystem.
- Kan forbedres hvis den fodres med flere data.
- EU arbejder med ELRC-programmet på at forbedre den sproglige infrastruktur for alle sprog.

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.

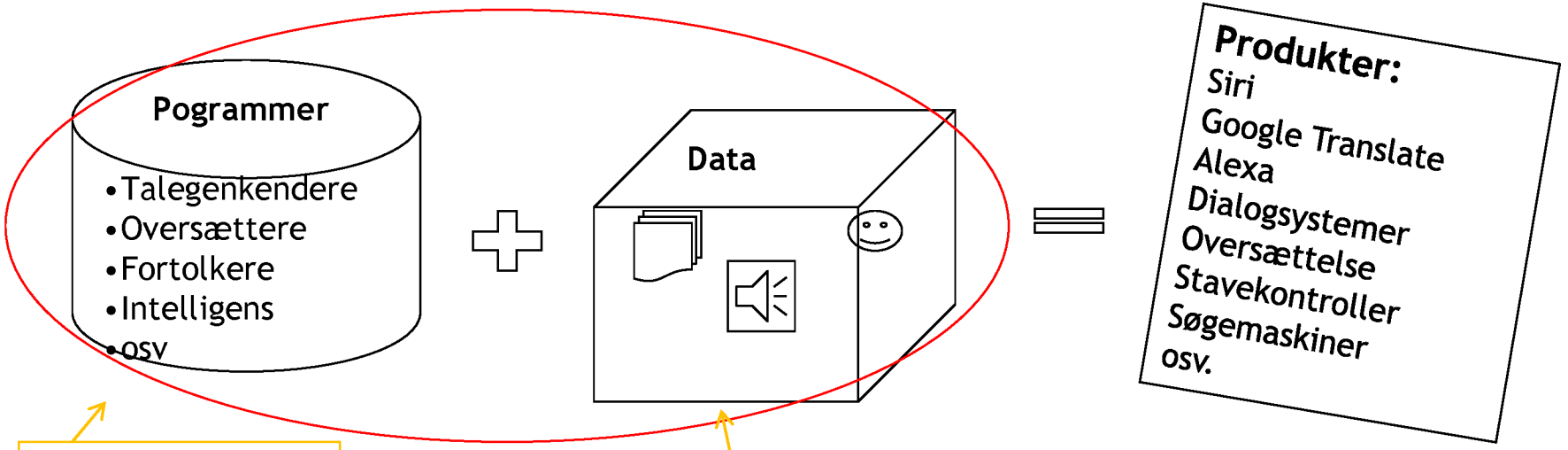
Oversættelse
- Transformation af betydning til målsprogets struktur
- Generering af sætninger.



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



Udvikles af forskere og producenter.

Indsamles af forskere, producenterne, i virksomheder og i offentlige institutioner.

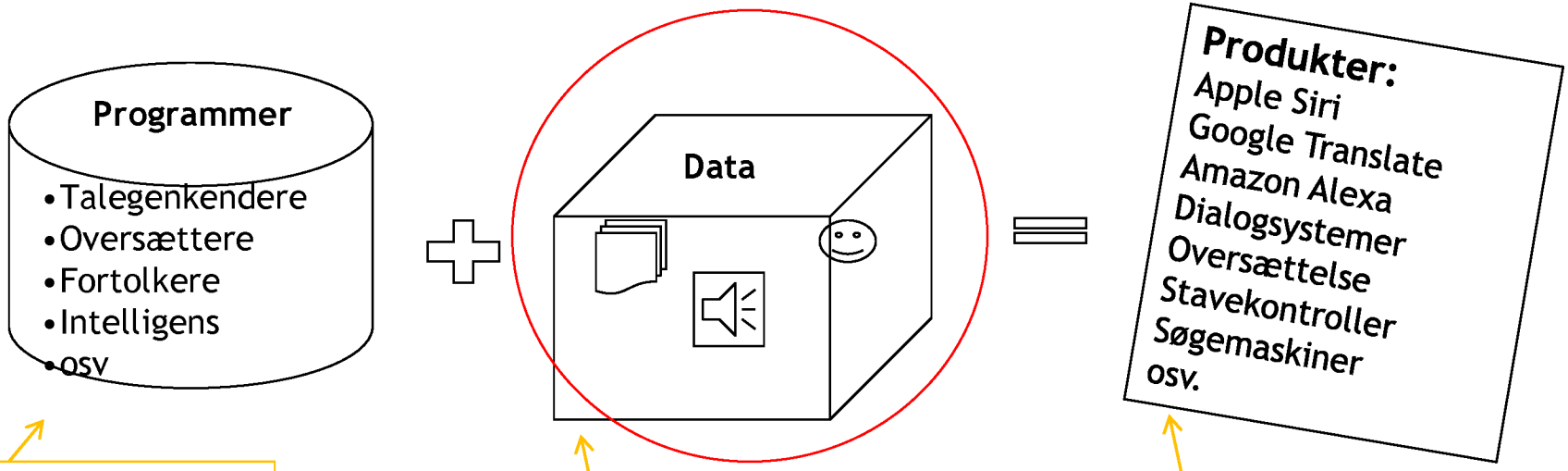
Retskrivningsordbogen er også en dataresurse

Ordbogsbasen

- Bog
- Web
  - Ordb.com
  - Gylden-dal
  - sproget.dk
- App
- Ordliste
  - Spiludviklere
  - Aviser/tv
  - KMD/Uni-C
  - CEF.at
  - Mircosoft

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

Hvem ejer data?

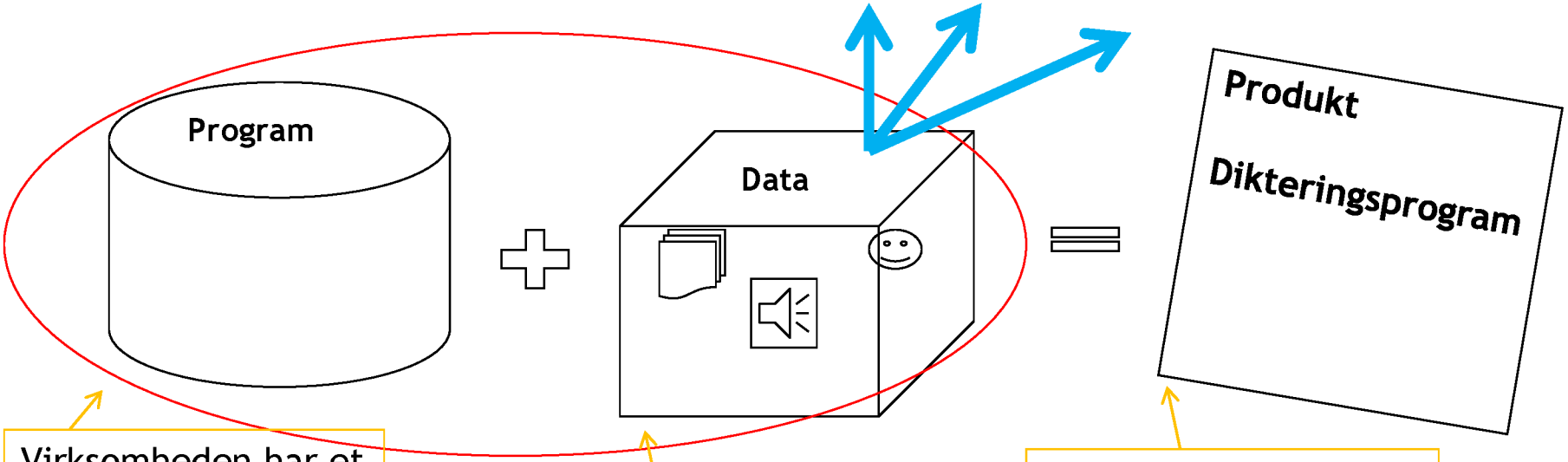


Ofte virksomheder, men kernen er tit open source.

Oftest virksomheder. Få data er open source.

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

Mangel på konkurrence – et eksempel.

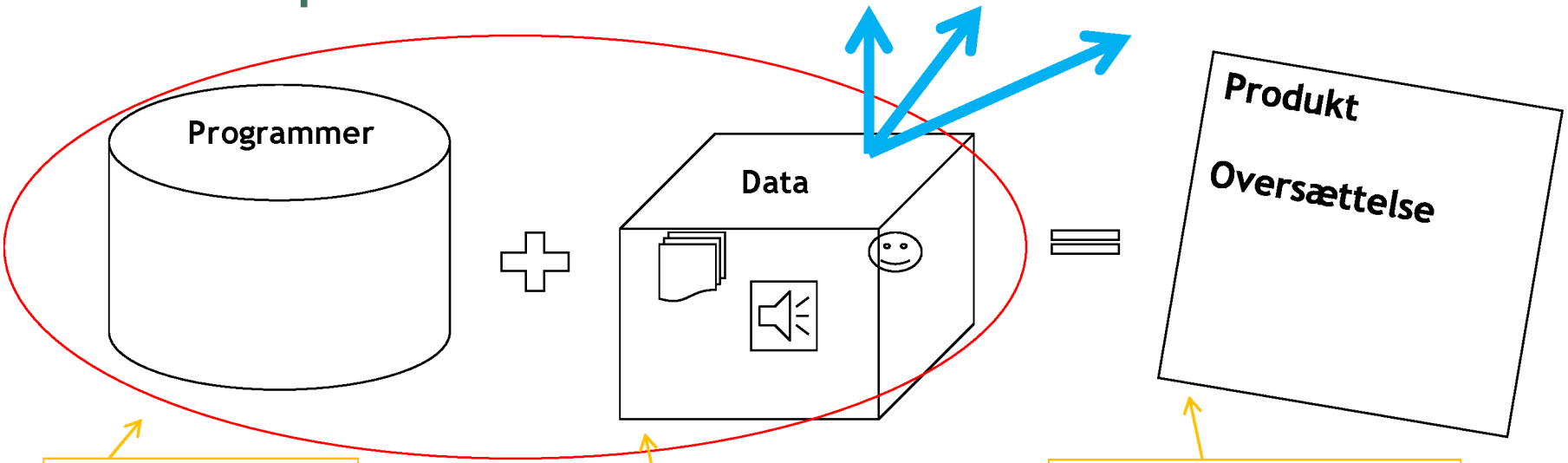


Virksomheden har et talegenkend-elsesprogram for engelsk som tilpasses dansk.

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

Virksomheden bearbejder data og sælger produktet til kommunen. Kommunen er låst.

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.

Sprogresurser 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 Sprogteknologi

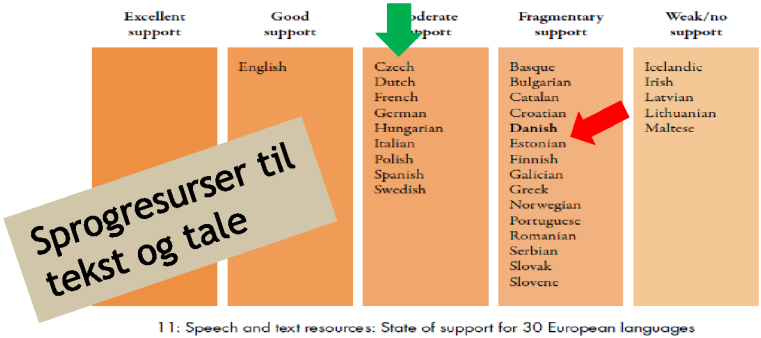
Kvaliteten for dansk er ringe.

Det skyldes:

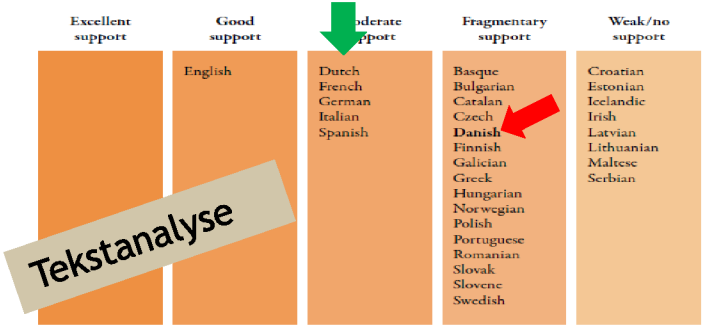
- At dansk er et vanskeligt sprog for systemer der i første omgang er udviklet til engelsk.
- At dansk er et lille marked med få aktører og stærke monopoltendenser.
- At der ikke har været satset tilstrækkeligt på forskning og udvikling af sprogteknologi for dansk.
- At der ikke forefindes tilstrækkeligt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

Dansk sprogteknologi ifølge META-net-rapporten (2012)

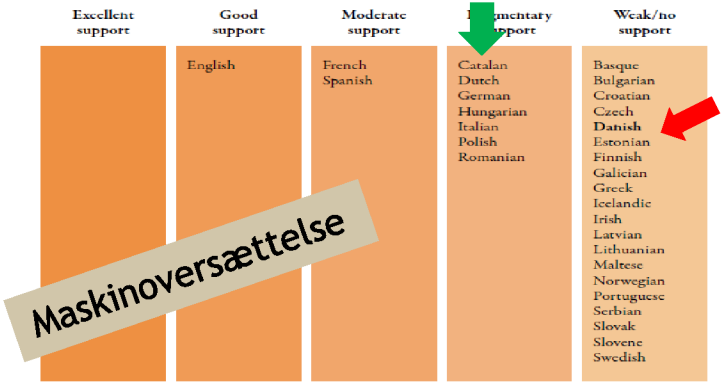
8: Speech processing: state of language technology support for 30 European languages



11: Speech and text resources: State of support for 30 European languages



10: Text analysis: state of language technology support for 30 European languages



9: Machine translation: state of language technology support for 30 European languages

Dansk er undereksponeret.

- Det meste sprogteknologi er udviklet på datasæt (samlinger af talt og skrevet sprog).
- Hvis datasæt ikke er tilgængelige for et givet sprog, bliver nye produkter ikke tilgængelige for sprogbrugerne.
- Hvis et produkt ikke trænes på et relevant datasæt for et givet sprog og domæne, kan produktet ikke afspejle de sproglige og kulturelle træk som er nødvendige for en vellykket kommunikation.

Undereksponeringens negative spiral

- Selvom der findes store mængder data tilgængelige for andre sprog, er de fleste produkter som kan håndtere sprog og kunstig intelligens, orienteret mod engelsk.
- Der er flere kommercielle incitamenter til at overeksponere engelsk frem for andre sprog.
- Der findes langt flere sprogteknologiske hyldevarer for engelsk. Det gør det nemmere at afprøve nye ideer. Startomkostningerne er langt lavere, hvilket betyder at fx danske forskere hellere vil arbejde med engelsk end med deres eget sprog.

Hovy & Spruit (2016)

Dansk Sprognævn

Hvordan har man hidtil arbejdet strategisk forskningspolitisk med sprogteknologi?

Forskning og undervisning i Danmark

- Forskningsmiljøer: KU, CBS, SDU (AAU, DTU, ITU)
- Undervisning: KU (IT og kognition), CBS og SDU (Fagsprog og vidensmodellering)
- Forskningsråd- og forskningsprogrammer
    - Danske
    - Nordiske
    - EU (Nyt: Human Language Project)

Sprogteknologi er et højt specialiseret fag der kræver god indsigt i både sprog og programmering. Der er typisk kun få studerende på et hold, og det giver problemer i forhold til dimensioneringen.

Forskning i sprogteknologi i Danmark

- EU
  - Eurotra (1984-1982), PAROLE, SIMPLE
- SHF/FKK
  - Constraint grammar for dansk, engelsk, portugisisk (VISL 1995-2000 (?))
  - UDOG (1993-1997) korpusbaseret udforskning af ordforråd og grammatik
  - CMOL (Computational modelling of language 2003-2007) statistiske metoder
  - DANNET (2005-2008) (bygger på EU-projekter (PAROLE og SIMPLE)
  - Semantic processing across domains (CST/DSL 2013-2016)
- STVF (1998-2002) fokus på udvikling af komponenter til taleteknologi
- Tværvidenskabelige IT-forskningsprogrammer under forskningsrådene
  - Spoken language dialogue systems (1991-1995)
  - Ontoquery (1999-2004)
- Private fonde
  - Carlsberg
  - Velux: DANTERMbank (2010-2015)

Andre sprogpolitiske aktører

Kommuner

- KL. Offentligt digitaliseringsfællesskab. Fælleskommunalt samarbejde om talegenkendelse (OS2 –KOMBIT –CBS).

Staten

- Digitalisering i det offentlige - PSI-direktivet
- Forskningsinfrastruktur
- Digitaliseringen af kulturarven.

Faglige Interessegrupper

- Terminologigruppen
- FORVIR - Forum for Vidensmodellering i Offentligt Regi
- Nordiske Sprognævn: Arbejdsgruppen for sprogteknologi i Norden (ASTIN)
- Sprognævnets fagråd for sprogteknologi.

Problem: Manglende kommunikation og samarbejde om sprogteknologi.

Regeringsinitiativer

- Udvikling af talesyntese 1998-2001.
- Forskning i multimedia (1997-2002) Staging: dialog med tale, håndbevægelser mm.
- It for alle (2003) talesyntese til blinde, svagtsynede og læsesvage.
- IT-infrastruktur
  - Clarin-DK (2008-2011) indsamling og tilgængeliggørelse af tekst, tale og billeder til humanistisk forskning (EU)
  - DigHumLab (2012- ?) herunder videreførelse af CLARIN-DK
  - Deic (Dansk e-infrastruktur 2012- ?) skal sikre den bedst mulige nationale resurseudnyttelse på e-infrastrukturområdet.

Problem: Kun få midler til sprog. Ingen samlet strategi hvor sprog indgår. Kun lidt sammenhæng mellem initiativerne. For lidt opfølgning på projekterne.

Sprogteknologi og terminologi i Folketinget

2008 Sprog til tiden (Regeringens sprogudvalg).

2009 Beretning i Folketingets Kulturudvalg (S, DF, SF og EL).

Flertallet i udvalget ønsker et terminologicenter oprettet under Dansk Sprognævn som skal drive en offentligt tilgængelig flersproglig termbank.

Terminologicenteret skal understøtte arbejdet med fagsproget i de faglige miljøer og sikre videndeling indbyrdes mellem miljøerne og det omgivende samfund.

Nedstemt. Der blev iværksat en sprogkampagne (Gang i sproget 2009-2011).

Sprogteknologi og terminologi i Folketinget

2012 Dansk Sprogs Status. Dansk Sprognævn.

2012 Beslutningsforslag nr. **B 13** Folketinget 2012-13. Dansk Folkeparti.

"VII. Sprogteknologiske tiltag

Der satses fuldt og helt på udvikling af dansk sprogteknologi i form af erhvervsstøtteordninger og øremærkede midler til sprogteknologisk forskning og infrastruktur.

VIII. Sprogtermbank

Offentligheden skal have adgang til en national flersproget termbank, hvilket var en af de vigtige anbefalinger i »Sprog til tiden« fra 2008. Det vil gøre det lettere at undervise og formidle på dansk, så det helt undgås at udbyde uddannelser på engelsk for danske studerende."

Nedstemt. Det blev besluttet at der skal udarbejdes jævnlige statusrapporter for brugen af engelsk på universiteterne.

Dansk Sprognævn

Hvad har man gjort i andre lande?

Hvad har man gjort i andre lande?

- Nederlandsk (Holland og Belgien)
  - Stevin (2004-2009) => **BLARK (Basic language kit).**
- Letland (og andre nyere EU-medlemslande)
  - Statsstøttede projekter fra 2005-2009 – systematisk opbygning af sprogteknologikomponenter.
- Sverige, Norge, Finland
  - Termbanker (nationale)
  - Store infrastrukturprojekter
  - Flere statslige initiativer (udarbejdelse af **BLARK** for norsk og svensk).
- EU
  - Oversættelse som e-infrastrutur i det digitale indre marked
  - Debat om Human Language Project i EU-Parlamentet.

BLARK – sprogteknologiske byggesten – nationale prioriteringer for nederlandsk.

For language technology:

Modules:

- Robust modul
  (tokenisation
- Morphological
  disambiguatio
- Syntactic anal
- Semantic anal

Data:

- Monolingual le
- Annotated cor
  syntactic, mor
  structures)
- Benchmarks for evaluation

> Only by dedicated programmes such as the STEVIN programme was it possible to create language resources and basic tools to be able to carry out research on language technology for the Dutch language, and to make it more attractive to companies to develop and offer products and services in the Dutch language.
> Meta-net, NL.

For speech technology:

Modules:

- eech recognition (including tools
  eech recognition, recognition of
  adaptation, and prosody
- esis (including tools for unit
- culating confidence measures
- ntification (speaker identification as
  age and dialect identification)
- ora for specific applications, such as
  ory assistance, etc.

Data:

- Multi-modal speech corpora
- Multi-media speech corpora
- Multi-lingual speech corpora
- Benchmarks for evaluation

Dansk Sprognævn

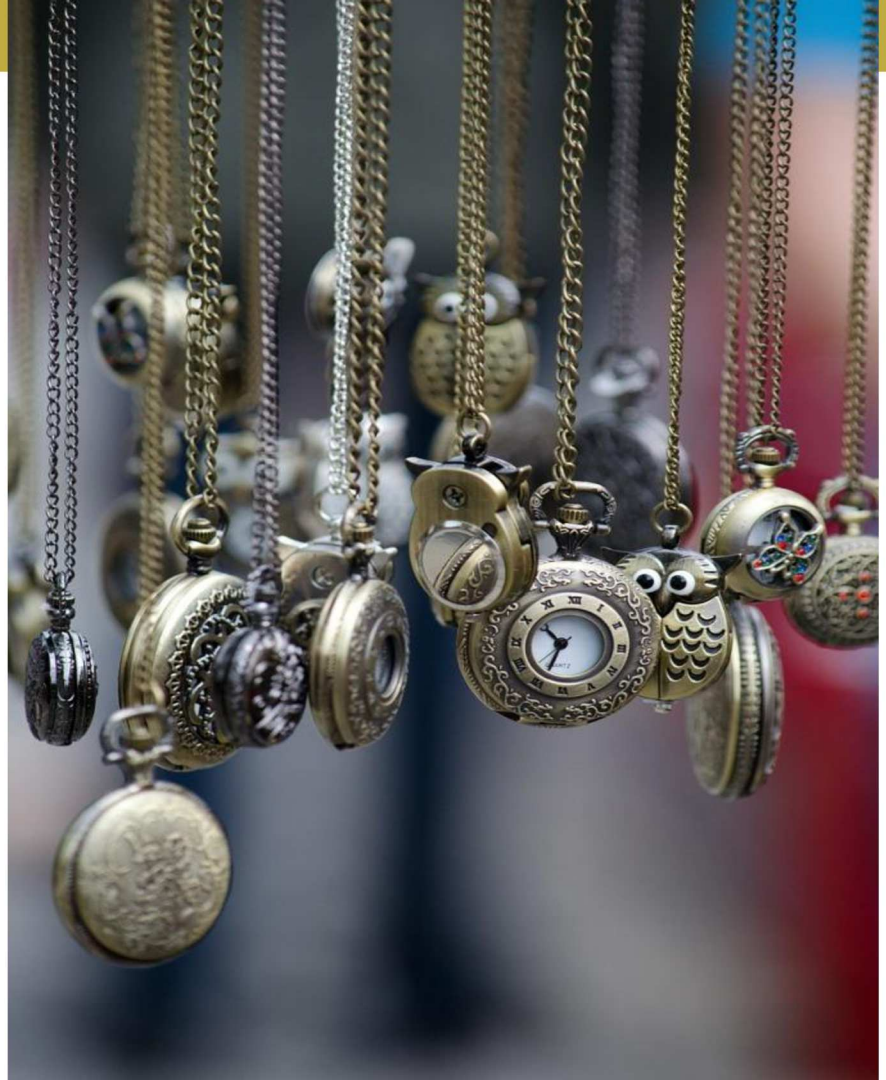
Hvordan styrker vi dansk sprogteknologi?

Betragt sprogteknologi som infrastruktur. Del viden og data!

- Centraliser indsamling af data (basisprogrammer, tekst og lyd), og gør dem tilgængelige for offentlige og private aktører.
- Styrk forskningen gennem et strategisk program.
- Inddrag private aktører løbende i processen.
- Sørg for at anskaffelse af nye produkter kan konkurrenceudsættes ved at sikre det offentlige ejerskab til egne data.
- Sørg for at udbrede viden om sprogteknologi.
- Vær opmærksom på at samfundet og dermed sproget hele tiden er under forandring. Indsatsen skal være kontinuerlig.

Konklusion

- Sprogteknologi er infrastruktur.
- Der er brug for en klar strategi.
- Der bør løbende indsamles og deles data for udvikle produkter og holde dem ajour.
- Det kan ikke overlades til markedskræfterne, men en delestrategi kan stimulere markedet.
- Staten og de offentlige institutioner kan spille en afgørende rolle.
- Tiden er knap.



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

Tak!