Semantik og sprogforståelse i COR

(COR-S)

Semantik, sprogforståelse og AI

Danske virksomheder arbejder nu i stor stil med danske sprogdata med henblik på at skabe **sprogcentreret AI**, derfor: stigende efterspørgsel på danske ordbogsdata med basale oplysninger om semantik, fx om ord er positive eller negative

COR-projektet: en ressource der gør alle basisoplysninger om danske ord tilgængelige på en standardiseret og internationalt kompatibel måde – også hvad angår semantik til sprogforståelse

Forankret i viden om dansk sprog og samfund

COR skal facilitere sprogcentrerede AI-systemer for dansk og tager derfor afsæt i lokalt forankret viden om dansk sprog og samfund i stedet for blot at tilpasse fra engelsk

CORs semantiske del bygger derfor på eksisterende danske ordbøger af høj kvalitet

Leksikalske ressourcer til sprogforståelse

Fremgangsmåde i COR:

Gennemgå eksisterende leksikalske ressourcer til sprogteknologi udviklet igennem de senere år, alle med reference til *Den Danske Ordbog* med henblik på:

- **Validering**
- **Forenkling** af struktur og betydningsinventar
- **Udvidelse af dækningsgrad**

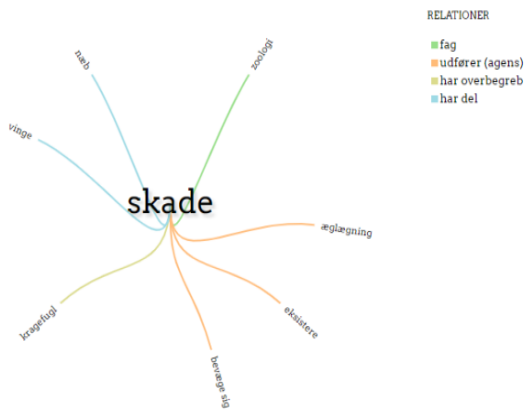
Sprogteknologiske kilder genbruges i COR

Først og fremmest det danske wordnet: DanNet
(70.000 ord)

Indhold: Betydningsstruktur, overbegreber, ontologiske typer og semantiske relationer, positiv-negativ på visse ord

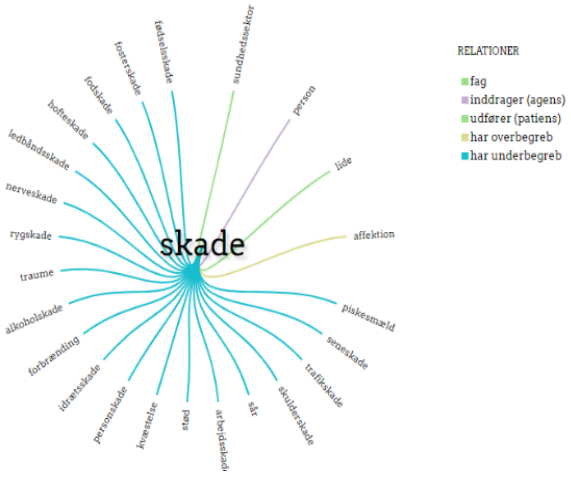
Fugl (ANIMAL)

ca. 46 cm lang kragefugl med hvid og blåsort fjerd ...



Defekt (PROPERTY)

(midlertidig) fysisk defekt fremkommet ved ydre på ...



Sprogteknologiske kilder genbruges i COR

Det Danske FrameNet-leksikon (11.500 verber og verbasubstantiver)

Semantiske rammer og roller for handlinger og hændelser

Fx

Sansning	*lugte, se, føle..*
Bevægelse	*løbe, gå, svømme*..
Kommunikation	*tale, argumentere, drøfte*..

Det Danske Sentimentleksikon (14.000 ord)

Ords negative/positive konnotation til brug for sentimentanalyse

Fx *En overbevisende* vs. *en fejlbehæftet argumentation*

positiv negativ

Mål 1: Forenkling af struktur og betydningsinventar

Udgangspunkt: Den Danske Ordbog (DDO)
En stor udfordring af finde det rette niveau

Ikke for finkortnet til at sprogmodellerne kan skelne,
men finkornet nok til at centrale betydningsforskelle beskrives

COR udvikler principielle retningslinjer for:

- sammenlægning af betydninger fra DDO
- sletning af forældede eller for specialiserede betydninger

Fremgangsmåde:

- Opbygge en guldstandard for forenkling ud fra en række principper (ca. 5000 ord)
- Herudfra maskinlære på resten

Mål 1: Forenkling af struktur og betydningsinventar

Guldstandard for sammenlægning (håndkodet)

Machine prediction

Betyd 1

Betyd 1

Betyd 2

Betyd 1

Betyd 1

Betyd 1

Betyd 2

Betyd 1

'*Hær*' i Den Danske Ordbog

Mål 2: Udvidelse af dækningsgrad

- Ikke alle 'centrale' betydninger er med i sprogteknologiske kilder (DanNet)

Mål:

- lukke 'huller'
- sikre at centrale ord er fuldt dækket mht. betydninger

Metode:

1. DanNet's links til centrale begreber i Princeton WordNet
2. Nøgleord i Den Danske Begrebsordbog
3. Validering: DanNet i opmærkning (ELEXIS)

Mål 2: Udvidelse af dækningsgrad i

DET DANSKE SPROG- OG LITTERATURSELSKAB

DanNets links til centrale begreber i Princeton WordNet

- ~5000 centrale engelske begreber
- → manuelt linket til DanNet og til opslagsordet (med øvrige betydninger) i DDO

→ **~ 4600 ord**

(*abe, acceptere, adgang, ekspert, elegant, knække, spise osv.*)

- 5 % af opslagsord i DDO, men
- 17 % of betydningerne i DDO

ekspert har 4 overbegreber (person → væsen → dyr → organisme)

Alignering

Det Engelske Wordnet har begreber der svarer til det Danske:

→ *(Engelsk)* expert: *a person with special knowledge or ability who performs skillfully*

Underbegreber

- ernæringsekspert: *(ingen definition)*
- finanseksper: *(ingen definition)*
- fingeraftryksekspert: *(ingen definition)*
- militæreksper: *(ingen definition)*
- miljøeksper: *(ingen definition)*
- rationaliseringsekspert: *(ingen definition)*
- sprængningsekspert: *(ingen definition)*
- sprængstofeksper: *(ingen definition)*
- valgekspert: *(ingen definition)*
- våbeneksper: *(ingen definition)*

DET DANSKE SPROG- OG LITTERATURSELSKAB

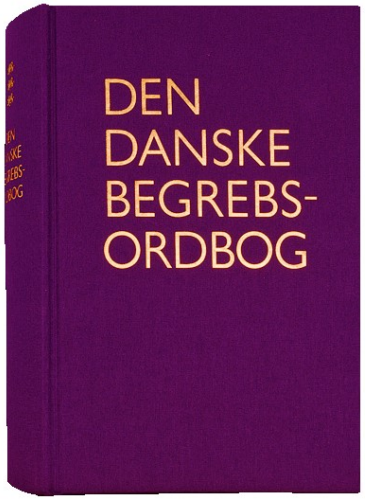
Mål 2: Udvidelse af dækningsgrad

Nøgleord i Den Danske Begrebsordbog

- Ca. 11.500 ord er nøgleord
  fx
  it, computer, pc, bærbar, terminal,.. , datakraft, funktionalitet

- Mange gengangere fra 1., men yderligere
  8.400 ord

Afsnittet 'computer'



B • **it**, edb, informationsteknologi • inform
datamatik, datalogi, kybernetik, datateknik,
teknik, computerteknik, datateknologi, com
teknologi • informatik, informationsteori, k
• sprogteknologi, datalingvistik, informatic
kommunikationsteknologi, ikt, informations
intelligens, AI, KI • robotteknologi • **com**
datamaskine, datamat, edb-maskine • langsom computer, hurtig computer • elektronhjerne, kalkulator, robot • stationær, stationær computer • ***pc***, personlig computer • bambusmaskine sl., desktop • hjemmecomputer, hjemme-pc, hjemmedatamat • tekstbehandlingsanlæg • ***bærbar***, laptop • notebook • PDA, palmtop, lommecomputer, lomme-pc • ***tablet***, tabletcomputer, tavlecomputer, tavle-pc • e-læser • mikrodatamat, minidatamat • minicomputer, mikrocomputer, supercomputer • ***terminal***, dataterminal, edb-terminal • skærmterminal, arbejdsstation • client/server, client/server-system, client/server-miljø, client/server-løsning • edb-anlæg, edb-system, computersystem • ***simulator***, flysimulator • ***server***, mainframe, webserver, webhotel • ***computernetværk***, net, netværk, datanet, intranet, ekstranet, familienetværk, intranet • internet • **datakraft**, maskinkraft, regnekraft • performance, nedetid, oppetid • ram-størrelse, rom-størrelse • klokfrekvens • konfiguration, opsætning • understøttelse • **funktionalitet**, funktion,

DET DANSKE SPROG- OG LITTERATURSELSKAB

Mål 2: Udvidelse af dækningsgrad

Validering: DanNet i opmærkning (ELEXIS)

2000 sætninger

- Sikre at de frekvente manglende betydninger er dækket af 1 og 2

DET DANSKE SPROG- OG LITTERATURSELSKAB

Sprogteknologiske kilder genbruges i COR

Øvrige ord i DanNet

- ~ 50.000 ord
- alle artefakter (ca. 8000 ord)
- alle personer?
- alle ord i DanNet med kun én betydning i DDO?
- sammensætninger der er underbegreber til centrale ord?
- osv.

Møbel, Artefakt og Objekt
4 overbegreber (stol → siddemøbel → møbel → genstand)

Efter en lang dag på den velpolstrede kontorstol er der behov for en gang motion, der kan mærkes

COR-S, dec 2023: Ordforråd med formel betydningsbeskrivelse

DET DANSKE SPROG- OG LITTERATURSELSKAB

~11.000 centrale ord med
~20.000 COR-betydninger
(efter reduktion)