Bilag

Indhold

Herunder fremgår alle bilagene i projektet. Hvert bilag vil have en kort forklaring som beskriver hvad den består af.

Bilag A – Kodebog

Kodebogen forneden består af koder observeret i de empiriske tekster: Digitaliseringsstrategien, Nationalstrategien, KI akten og Rapporten af HLEG. Første kolonne, Kode, indeholder en kode som fungerer som overkategori for andre koder. Anden kolonne giver en beskrivelse af koden i første kolonne. Tredje kolonne giver et eksempel fra en af de empiriske tekster på et tekststykke, hvor koden er anvendt. Sidste kolonne giver et overblik af de andre koder fra teksterne som hører under koden i første kolonne.

Kode	Beskrivelse	Eksempel	Underkategori
Lov	Overholdelse af loven på baggrund af juridiske virkemidler	"Sådanne teknologier kan indebære muligheder og risici, som bør imødegås og reguleres af en omfattende retlig ramme på EU-plan, osm afspejer etiske principper, der skal overholdes fra tidspunktet for udvikling og udbredelse af sådanne teknologier til anvendelsen heraf." Europa-Parlamentet, 2020, s.37	Juridisk, Lovlig og naturlig person
Etik	Principper og normer for handling som af den ene eller anden årsag anses som moralsk acceptable eller korrekte i en eller flere kontekster	Derudover indebærer kunstig intelligens en række etiske spørgsmål om forholdet mellem på den ene side fordelene ved anvendelse af ny teknologi og på den anden side hensynet til borgernes grundlæggende rettigheder, retssikkerhed og de grundlæggende samfundsmæssige	Gavn, Etiske principper, Moralsk status

		værdier, der ligeledes skal tages hånd om." Finansministeriet og Erhvervsministeriet, 2019,s. 17	
Robusthed	Faktor for troværdig og sikker teknologi i forhold til tekniske og ikke-tekniske aspekter af et stykke teknologi	"Det er vigtigt, at dataetik, sikkerhed og ansvarlighed er forankret dybt i den digitale udvikling, også når nye teknologier og digitale platforme vinder indpas." Digitaliserings- og Ligestillingsministeriet, 2023, s.22	Cybersikkerhed, Modstandsdygtighed, Sikkerhed, Rustet, Vedholdent, Nøjagtighed, Beskyttelse, Integritet, Whistleblowing, Evaluering, Test, Validering, Dokumentation, Revision
Ramme	Regulering som anvender forskellige virkemidler for at opnå en ønskede effekt	"A Union legal framework laying down harmonised rules on artificial intelligence is therefore needed to foster the development, use and uptake of artificial intelligence in the internal market that at the same time meets a high level of protection of public interests, such as health and safety and the protection of fundamental rights, including democracy, rule of law and environmental protection as recognised	Økonomisk ramme, Retningslinjer, Regulering, Standarder, Krav

		and protected by Union law." Europa-Kommissionen, 2024, s. 14	
Sociotekniske systemer	Netværk bestående af sociale og teknologiske aktanter	"Tillid til udviklingen, udbredelsen og anvendelsen af AI-systemer vedrører ikke kun teknologiens indbyggede egenskaber, men også kvaliteten af de sociotekniske systemer, der benytter AI-anvendelser." Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 5	Grundlag, Infrastruktur, Sprogområde, Samfund
Interessenter	Direkte og indirekte involverede aktører i ethvert projekt. "Alle, der forsker i, udvikler, designer, udbreder eller anvender kunstig intelligens, og alle, der (direkte eller indirekte) påvirkes af kunstig intelligens" Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 42	"Disse retningslinjer er rettet til alle AI-interessenter, der designer, udvikler, udbreder, implementerer, anvender eller berøres af kunstig intelligens, herunder bl.a. virksomheder, organisationer, forskere offentlige tjenester, offentlige instanser, institutioner, civilsamfundsorganisationer, enkeltpersoner, arbejdstagere og forbrugere." Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 5-6	Virksomheder, Udvikler, Udbreder, Bruger, Offentligheden, Erhvervslivet, Agentur, Civilsamfundet, Tilsynsmyndigheder, Skyldige, Ofre, Aktant, Autoritet
Domæne specificitet	Teknologi som kun er udviklet med henblik	"Der etableres en fælles dansk sprogressource, der	Sektorspecifik

	på anvendelighed i en bestemt kontekst eller med en bestemt type data	skal understøtte og accelerere udviklingen af sprogteknologiske løsninger på dansk." Finansministeriet og Erhvervsministeriet, 2019, s.21	
Værdier	Noget pålægges en form for værd og ønskes udbred	"This Regulation should be applied in conformity with the values of the Union enshrined in the Charter facilitating the protection of individuals, companies, democracy and rule of law and the environment while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI." Europa-Kommissionen, 2024, s. 12	Interesser, Offentlige interesser, Private interesser, Erhvervsinteresser, Demokrati, Lighed, Pluralisme, Respekt, Principper
Viden	Forståelse for et område der giver en dybere indblik i et fænomens underlæggende funktionalitet eller baggrund	"Danske børn og unde skal opbygge digitale kompetencer og digital dannelse, så de tidligt i livet får forståelse for den digitale virkelighed, de vokser op i. Fremover skal der desuden uddannes flere it-specialister for at imødekomme den store efterspørgel fra erhvervslivet."	Informeret, KI kendskab, Teknologiforståelse, Oplysning, Forskning, Uddannelse, Kompetencer,

		Digitaliserings- og Ligestillingsministeriet, 2023, s. 10	
Autonomi	Retten og muligheden for at styre og bestemme over ens egne handlinger og person, Kan både anvendes i forhold til naturlige og lovlige personer	"Menneskets autonomi prioriteres i udvikling og anvendelse af kunstig intelligens. Mennesket skal som i dag kunne træffe oplyste og selvstændige valg, uden af kunstig intelligens fjerner menneskets selvbestemmelse." Finansministeriet og Erhvervsministeriet, 2019, s. 28	Brugervenlighed, Menneskecentreret, Rettidig, Kontrol, Frihed, Frisættelse, Selvbestemmelse, Værdighed, Værdighed
Udvikling	Skabelse af materielle og immaterielle goder baseret på markedseffekter som produktion, innovation, effektivisering, mv.	"Den (KI) kan eksempelvis skabe nye muligheder for at nytænke velfærdsopgaver og øge produktivitet og vækst i erhvervslivet." Digitaliserings- og Ligestillingsministeriet, 2023, s.16	Konkurrence, Moderisering, Konkurrenceevne, Foregangsland, Førerposition, Eksport, Vækst, Økonomi, Innovation, Det indre marked, Arbejde, Marked, Værdikæde, SMV'er, Potentiale, Udnytte, Effektivitet, Teknologisk fix, Ressource, Drivhusgasser, Klimamål, Klimaaftryk, Klimabelastning, Miljø,

			Fremtidssikre, Fremtid, Bæredygtighed, Velfærd, Velvær, Sundhed, Levevilkår, Velstand
Holisme	Helheden af et fænomen med hensyn til dets livscyklus (før, under og efter), værdi (materiel og immateriel) og virke	"Indsatsen for at opnå pålidelig kunstig intelligens omfatter derfor ikke kun selve AI-systemets pålidelighed, men kræver en holistisk og systemisk tilgang, der omfatter pålideligheden af alle aktører og processer (...)" Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 5	Livscyklus, Implementering, Anvendelse, Udvikling, Implementering, Udbredelse
Risiko	Det mulige potentiale for at negative konsekvenser af et fænomen forekommer. "Risks are defined as the probabilities of physical harm due to given technological or other processes. Hence technical experts are given pole position to define agendas and impose bounding premises a	"Samtidig medfører teknologien nye udfordringer og trusler, eksempelvis for vores demokratiske samtale i forhold til misinformation og manipulation af billeder og lyd." Digitaliserings- og Ligestillingsministeriet, 2023, s. 16	Udfordringer, Manipulation, Misinformation, Frygt, Risikohåndtering, Fare, Fejl, Risikoanalyse, Trusler, Udsatte grupper, Høj-risiko, Kvæstelse eller skade, Forskelsbehandling, Forudindtagethed, Misbrug, Blackbox, Kategorisering, Konsekvensanalyse

	priori on risk discourses." Beck, 1992		
Gennemsigtighed	Tilgængeligheden og forklarligheden af et fænomens i dets helhed, komponenter og virke	"(...) udvikles, udbredes, og anvendes på en let forklarlig måde for at sikre, at der kan foretages en gennemgang af teknologiernes tekniske processer." Europa-Parlamentet, 2020, s. 53	Transparens, Klarhed, Forklarlighed, Troværdighed, Tillid
Retfærdighed	En juridisk, etisk og værdibaseret vurdering af korrekt handling	"(...) at sikre en ligelig og retfærdig fordeling af både fordele og omkostning og at sikre at personer og grupper ikke udsættes for urimelig skævhed, diskrimination og stigmatisering." Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 13	Ligestilling, Ligebehandling, Ikke-diskrimination, Lighed, Solidaritet, Ikke-forudindtagethed, Ikke-forskelsbehandling, Nøjagtighed, Objektivitet, Upartisk
Proportionalitet	Afbalancering af en handling baseret på et fænomens kontekst	"In accordance with the principle of proportionality as set out in that Article, this Regulation does not go beyond what is necessary in order to achieve that objective." Europa-Kommissionen, 2024, s. 92	Balance, Afbalanceret
Indflydelse	Evnen til at præge processer ved at ændre	"Dansk interessevaretagelse og implementering af EU's digitale dagsorden"	Præge, Inddrage

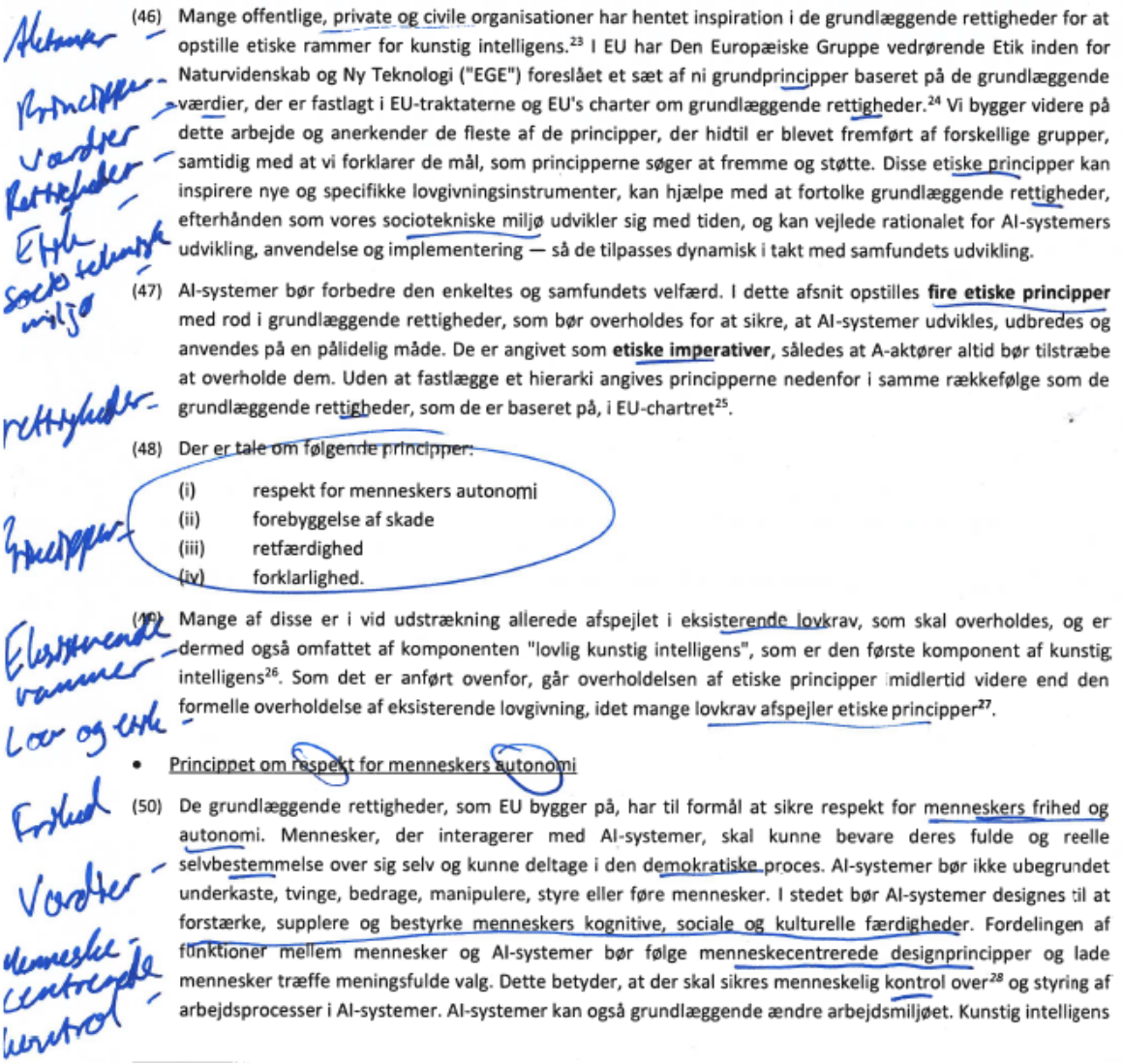
	deres forløb eller udfald	Digitaliserings- og Ligestillingsministeriet, 2023, s. 27	
Samarbejde	Udførelse af et eller flere projekter som indebærer en eller flere overensstemmende aspekter af flere forskellige aktanter og/eller interessenter med henblik på at opnå et fælles mål	"De enkelte nationale tilsynsmyndigheder bidrager til ensartet anvendelse af denne forordning i hele Unionen. Med henblik herpå samarbejder tilsynsmyndighederne i de enkelte medlemsstater med hinanden, med Kommissionen og/eller med eventuelle relevante EU-institutioner, -organer, -kontorer og -agenturer, som kan udpeges til dette formål." Europa-Parlamentet, 2020, s. 58	International, Inklusion, Inddragelse, Solidaritet, Fleksibilitet, Administration, Landegrænser, Engagement, Kompetencer, Åben adgang, Harmonisering, Koordinering
Ansvarlighed	Passende og opfyldende efterlevelse af etiske og/eller juridiske forventninger	"Alle led skal være ansvarlige for konsekvenserne af deres udvikling og anvendelse af kunstig intelligens, dvs. blandt andet udviklere, samarbejdspartnere, anvendere, myndigheder og virksomheder." Finansministeriet og Erhvervsministeriet, 2019, s. 28	Ansvar, Magt, Indflydelse, Socialt ansvar, Forpligtelser, Forvaltning, Forvaltningsstandarder, Erstatning, Forvaltningsansvar, Understøtte, Certificering
KI model	Den metodiske tilgang til og teknologiske udformning af et KI system	"pålidelig kunstig intelligens" Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 8	Pålidelig KI, Forklarelig KI, Troværdig KI, Sikker KI, Åben KI, Design, Metode, Tilgang, Praksis, KI

			styring, KI værdi-kæde, Definerede formål, Implicitte formål
Datastyring	Håndtering, opbevaring og bearbejdelse af data	"Regeringen etablerer et Forsyningsdigitaliseringsprogram, som skal skabe ensartet og nem adgang til data fra forsyningssektoren. Programmet skal understøtte bedre udnyttelse af ressourcer og infrastruktur på tværs af værdikæder og forsyningsarter." Digitaliserings- og Ligestillingsministeriet, 2023, s.18	Data, Dataøkonomi, Datastyring, Dataetik, Datadeling, Information, Åben adgang, Autensitet, Kvalitet, Big data, Personoplysninger, Ressource, Forsyning

Bilag B – Udsnit fra første og anden kodningsomgang

Udsnittene fra første og anden kodningsomgang er skanninger af de udprintet sider som er blevet læst og skrevet i under kodningsomgangene. Anvendeligheden af eksemplerne er at anvendelsen af koder bliver mindre sporadisk og mere målrettet i overgangen fra første til anden kodningsomgang.

Udsnit fra første kodningsomgang



(46) Mange offentlige, private og civile organisationer har hentet inspiration i de grundlæggende rettigheder for at opstille etiske rammer for kunstig intelligens.[23] I EU har Den Europæiske Gruppe vedrørende Etik inden for Naturvidenskab og Ny Teknologi ("EGE") foreslået et sæt af ni grundprincipper baseret på de grundlæggende værdier, der er fastlagt i EU-traktaterne og EU's charter om grundlæggende rettigheder.[24] Vi bygger videre på dette arbejde og anerkender de fleste af de principper, der hidtil er blevet fremført af forskellige grupper, samtidig med at vi forklarer de mål, som principperne søger at fremme og støtte. Disse etiske principper kan inspirere nye og specifikke lovgivningsinstrumenter, kan hjælpe med at fortolke grundlæggende rettigheder, efterhånden som vores socioetekniske miljø udvikler sig med tiden, og kan vejlede rationalet for AI-systemers udvikling, anvendelse og implementering — så de tilpasses dynamisk i takt med samfundets udvikling.

(47) AI-systemer bør forbedre den enkeltes og samfundets velfærd. I dette afsnit opstilles **fire etiske principper** med rod i grundlæggende rettigheder, som bør overholdes for at sikre, at AI-systemer udvikles, udbredes og anvendes på en pålidelig måde. De er angivet som **etiske imperativer**, således at A-aktører altid bør tilstræbe at overholde dem. Uden at fastlægge et hierarki angives principperne nedenfor i samme rækkefølge som de grundlæggende rettigheder, som de er baseret på, i EU-chartret[25].

(48) Der er tale om følgende principper:

(i) respekt for menneskers autonomi
(ii) forebyggelse af skade
(iii) retfærdighed
(iv) forklarlighed.

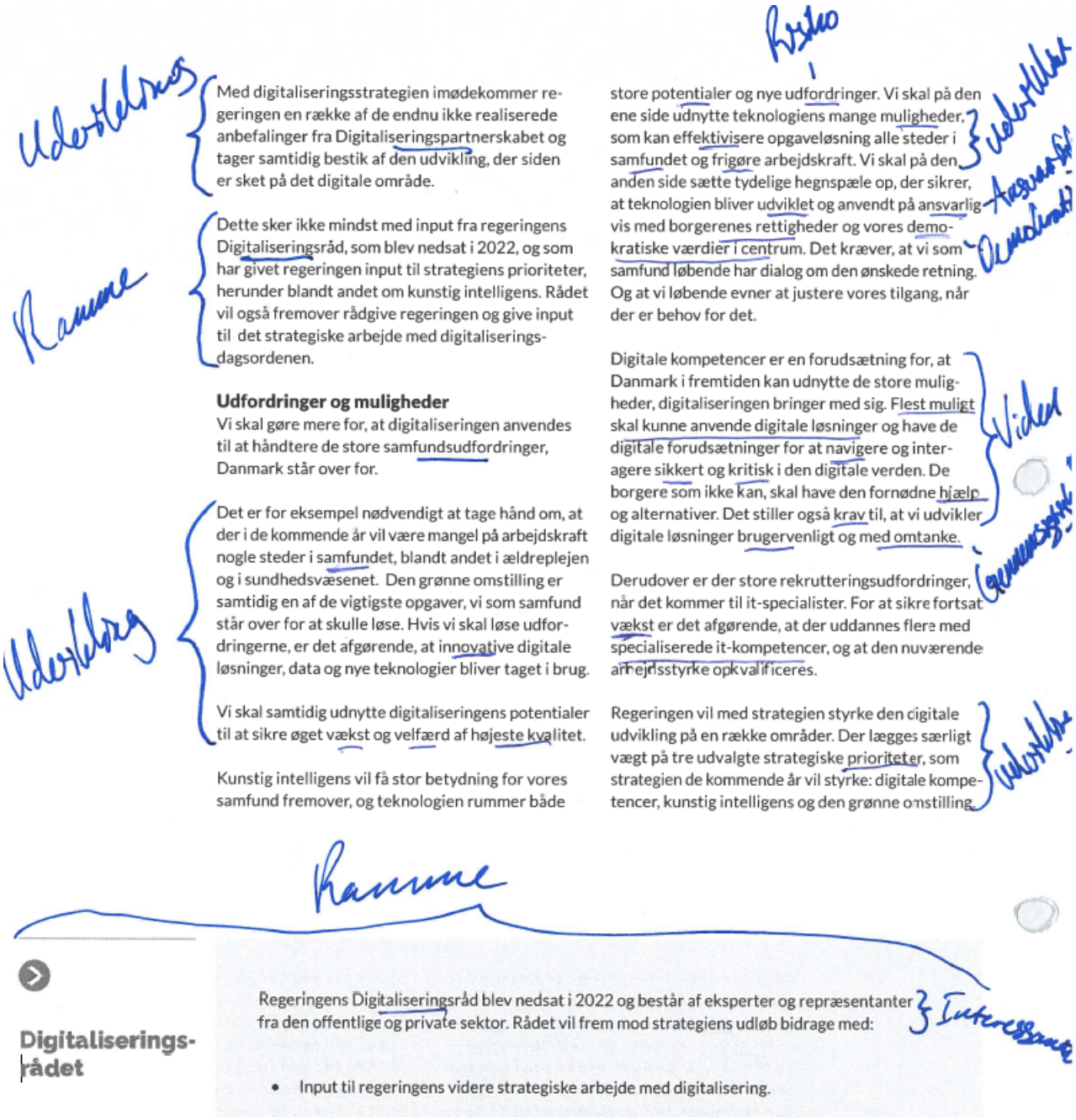
(49) Mange af disse er i vid udstrækning allerede afspejlet i eksisterende lovkrav, som skal overholdes, og er dermed også omfattet af komponenten "lovlig kunstig intelligens", som er den første komponent af kunstig intelligens[26]. Som det er anført ovenfor, går overholdelsen af etiske principper midlertid videre end den formelle overholdelse af eksisterende lovgivning, idet mange lovkrav afspejler etiske principper[27].

- Princippet om respekt for menneskers autonomi

(50) De grundlæggende rettigheder, som EU bygger på, har til formål at sikre respekt for menneskers frihed og autonomi. Mennesker, der interagerer med AI-systemer, skal kunne bevare deres fulde og reelle selvbestemmelse over sig selv og kunne deltage i den demokratiske proces. AI-systemer bør ikke ubegrundet underkaste, tvinge, bedrage, manipulere, styre eller føre mennesker. I stedet bør AI-systemer designes til at forstærke, supplere og bestyrke menneskers kognitive, sociale og kulturelle færdigheder. Fordelingen af funktioner mellem mennesker og AI-systemer bør følge menneskecentrerede designprincipper og lade mennesker træffe meningsfulde valg. Dette betyder, at der skal sikres menneskelig kontrol over[28] og styring af arbejdsprocesser i AI-systemer. AI-systemer kan også grundlæggende ændre arbejdsmiljøet. Kunstig intelligens

(Ekspertgruppen på Højt Niveau vedrørende Kunstig Intelligens, 2019, s. 12)

Udsnit fra anden kodningsomgang



Med digitaliseringsstrategien imødekommer regeringen en række af de endnu ikke realiserede anbefalinger fra Digitaliseringspartnerskabet og tager samtidig bestik af den udvikling, der siden er sket på det digitale område.

Dette sker ikke mindst med input fra regeringens Digitaliseringsråd, som blev nedsat i 2022, og som har givet regeringen input til strategiens prioriteter, herunder blandt andet om kunstig intelligens. Rådet vil også fremover rådgive regeringen og give input til det strategiske arbejde med digitaliseringsdagsordenen.

Udfordringer og muligheder

Vi skal gøre mere for, at digitaliseringen anvendes til at håndtere de store samfundsudfordringer, Danmark står over for.

Det er for eksempel nødvendigt at tage hånd om, at der i de kommende år vil være mangel på arbejdskraft nogle steder i samfundet, blandt andet i ældreplejen og i sundhedsvæsenet. Den grønne omstilling er samtidig en af de vigtigste opgaver, vi som samfund står over for at skulle løse. Hvis vi skal løse udfordringerne, er det afgørende, at innovative digitale løsninger, data og nye teknologier bliver taget i brug.

Vi skal samtidig udnytte digitaliseringens potentialer til at sikre øget vækst og velfærd af højeste kvalitet.

Kunstig intelligens vil få stor betydning for vores samfund fremover, og teknologien rummer både store potentialer og nye udfordringer. Vi skal på den ene side udnytte teknologiens mange muligheder, som kan effektivisere opgaveløsning alle steder i samfundet og frigøre arbejdskraft. Vi skal på den anden side sætte tydelige hegnspæle op, der sikrer, at teknologien bliver udviklet og anvendt på ansvarlig vis med borgerenes rettigheder og vores demokratiske værdier i centrum. Det kræver, at vi som samfund løbende har dialog om den ønskede retning. Og at vi løbende evner at justere vores tilgang, når der er behov for det.

Digitale kompetencer er en forudsætning for, at Danmark i fremtiden kan udnytte de store muligheder, digitaliseringen bringer med sig. Flest muligt skal kunne anvende digitale løsninger og have de digitale forudsætninger for at navigere og interagere sikkert og kritisk i den digitale verden. De borgere som ikke kan, skal have den fornødne hjælp og alternativer. Det stiller også krav til, at vi udvikler digitale løsninger brugervenligt og med omtanke.

Derudover er der store rekrutteringsudfordringer, når det kommer til it-specialister. For at sikre fortsat vækst er det afgørende, at der uddannes flere med specialiserede it-kompetencer, og at den nuværende arbejdsstyrke opkvalificeres.

Regeringen vil med strategien styrke den digitale udvikling på en række områder. Der lægges særligt vægt på tre udvalgte strategiske prioriteter, som strategien de kommende år vil styrke: digitale kompetencer, kunstig intelligens og den grønne omstilling.

Digitaliseringsrådet

Regeringens Digitaliseringsråd blev nedsat i 2022 og består af eksperter og repræsentanter fra den offentlige og private sektor. Rådet vil frem mod strategiens udløb bidrage med:

- Input til regeringens videre strategiske arbejde med digitalisering.

(Digitaliserings- og Ligestillingsministeriet, 2023a, s. 8)

Bilag C – Data fra sprogteknologi.dk

Et permanent hyperlink til et github webside hvorfra projektets data vedrørende metadata på sprogteknologi.dk kan ses og hentes ned som en csv-fil: https://github.com/TokeJoMu/sprogteknologi_dk_metadata/blob/3dd7ba98bacddd039197e1bf569cae693ab3773b/sprogteknologi_data.csv

Bilag D – R projekt

Forneden ligger det fulde R-markdown dokument, som også er tilgængeligt på github websiden: https://github.com/TokeJoMu/sprogteknologi_dk_metadata/blob/bfa1c20ff2f8ab8132ad018397720e284b23829c/analysedokument.Rmd

En kvantitativ analyse af sprogressourcer på sprogteknologi.dk

Toke Jøns Mulvad

2024-03-05

Herunder fremgår en analyse af metadata på sprogressourcer på sprogteknologi.dk. I overensstemmelse med FAIR-principperne vil hvert kodestykke være forklartet forud med henblik på at tilgængeliggøre de programmatiske tiltag. Den indsamlede data er indhentet ved at gennemgå de enkelte sprogressourcer på 'sprogteknologi.dk/group/corpora'. Indsamlingen tog sted fra den

Forberedende arbejde - Indlæsning af R-pakker

R har som programmeringssprog en række grundlæggende funktioner og metoder til manipulering og visualisering. Men i kraft af at teknologien er åben og tilgængelig, findes der en stor brugerbase som videreudvikler på de indboende funktioner, disse brugerudviklede funktionspakker kaldes 'packages'. Funktionspakkerne bliver distribuerede, opbevarede og vedligeholdt ved hjælp af CRAN, som er et netværk af servere og protokoller (The Comprehensive R Archive Network, 2024).

For at installere en funktionspakke benyttes funktionen 'install.packages()'. En funktionspakke skrives da i citationstegn ("funktionspakke"). Eftersom dette projket benytter sig af flere funktionspakkker skrives pakkerne i en liste, som gøres ved at skrive 'c()'. I denne liste

skrives hver funktionspakke og sepereres med ‘,’. Denne funktion skal kun køres en gang, eftersom funktionspakkerne bliver installeret på den enhed som anvendes til undersøgelsen.

```
install.packages(c("tidyverse", "wordcloud", "RColorBrewer", "tidytext", "ggplot2"), repos = "http://cran.us.r-project.org")
```

```
## Installing packages into 'C:/Users/tokej/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
## package 'wordcloud' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'wordcloud'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\tokej\AppData\Local\R\win-library\4.2\00LOCK\wordcloud\libs\x64\wordcloud.dll
## to
## C:\Users\tokej\AppData\Local\R\win-library\4.2\wordcloud\libs\x64\wordcloud.dll:
## Permission denied
```

```
## Warning: restored 'wordcloud'
```

```
## package 'RColorBrewer' successfully unpacked and MD5 sums checked
## package 'tidytext' successfully unpacked and MD5 sums checked
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\tokej\AppData\Local\Temp\Rtmpi4wRrq\downloaded_packages
```

Pakken ‘tidyverse’ og ‘tidytext’ indeholder en lang række funktioner med henblik på at rense data (data cleaning), manipulering og visualisering. ‘RColorBrewer’, ‘ggplot2’ og ‘wordcloud’ indeholder funktioner til at visualisere data.

```
library(tidyverse)
library(tidytext)
```

```
library(RColorBrewer)
library(wordcloud)
library(ggplot2)
```

Indhentning af data

Dataen er indhentet fra sprogteknologi.dk i perioden 14-04-2024 til 20-04-2024 (https://sprogteknologi.dk/group/corpora). Herunder indhentes den indsamlede data fra filen 'sprogteknologi_data.csv', med funktionen 'read_delim()', som anvendes til at indhente data i R. Inde i funktionen angives filnavnet mellem citationstegn ("filnavn"), og efterfølges af hvilket tegn der adskiller værdierne i filen. I dette tilfælde er den anvendte værdi semikolon (';'), hvilket angives mellem citationstegn ("tegn"). For at den indlæste data indhentes og gemmes i R projektet, skal det gemmes under en variabel, som er et navn (her 'data') for noget information. Den venstre pegende pil ('<-'), er en operator som fortæller at alt på højre side af pilen skal gemmes under den angivet variabel.

```
data <- read_delim("sprogteknologi_data.csv", delim = ";")
```

```
## Rows: 83 Columns: 12
## ── Column specification ────────────────────────────────────────────────
────
## Delimiter: ";"
## chr (10): Navn, Organisation, Type, Sprog, Filtype, Tags, Emne, url, Licens,
...
## dbl  (2): Periode (Start), Periode (Slut)
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this mes
sage.
```

Herunder vises det hvordan den indhentede data ser ud i R, ved at bruge funktionen 'print()', som viser den data en givet variabel indholder.

```
print(data)
```

```
## # A tibble: 83 × 12
##    Navn      Organisation Type  Sprog Filtype `Periode (Start)` `Periode (Sl
ut)`
##    <chr>     <chr>        <chr> <chr> <chr>               <dbl>            <
dbl>
##  1 Coral te… Alexandra I… <NA>  dan   parquet               NA
NA
##  2 Georg Br… Det Danske … Korp… dan   html                 1872
1890
##  3 PAROLE-D… Det Danske … Korp… dan:… xml, t…              1998
2015
##  4 DK-CLARI… Det Danske … Korp… dan   xml                   NA
NA
##  5 Fuldform… Det Danske … Korp… dan   csv                   NA
NA
##  6 KorpusDK  Det Danske … Korp… dan   txt                   NA
NA
##  7 Public A… Det Kongeli… Korp… dan   xml                   NA
NA
##  8 Referate… Folketinget  Korp… dan   xml, h…               NA
NA
##  9 Lyd fra … Folketinget  Korp… dan   mp3                   NA
NA
## 10 TV fra F… Folketinget  Korp… dan   mp4, h…               NA
NA
## # ℹ 73 more rows
## # ℹ 5 more variables: Tags <chr>, Emne <chr>, url <chr>, Licens <chr>,
## #   Dokumentation <chr>
```

Den indsamlede data fra sprogteknologi.dk består af: 'Navn' på ressourcen Den 'Organisation' som står bag ressourcen; Hvilken 'Type' ressourcen er; Det naturlige 'Sprog' ressourcen indeholder; Hvilke(n) 'Filtype' som ressourcen kan hentes i; Hvilken 'Periode' ressourcen stammer fra og til; Nøgleord ('Tags') om hvad ressourcen består af; Emneord ('Emne') om

hvad ressourcen kan anvendes til; Et 'url'-link til ressourcens hjemmeside; Ressourcens 'licens'; 'Dokumentation' til at vejlede i anvendelse af ressourcen.

Nu efter at data er blevet indhentet og præsenteret som det ser ud uden nogen datamanipulering er vi klar til at påbegynde den kvantitative undersøgelse af materialet.

Analyse

Analyse - Manglende metadata

Herunder gives et overblik af hvor mange værdier der mangler i hver kolonne i datasættet. Dette anvendes til at styre hvilke efterfølgende undersøgelser der bliver fortaget og giver et umiddelbart indblik i hvor fyldestgørende den angivet metadata er for sprogressourcerne. Først angives det hvilken data vi ønsker at bearbejde ved at skrive navnet på variablen (her 'data'). Derefter anvendes operatoren '%>%', som også kaldes et rør (fra det engelske "pipe"). En metafor for hvordan denne operator fungere er at den angivet variabel hældes ned gennem alle de efterfølgende manipuleringer. Her hældes vores variabel 'data' altså først ned i funktionen 'select()', som vælger de værdier vi ønsker at bearbejde i 'data', i select-funktionen er der angivet endnu en funktion, 'everything()', som vælger alle værdier i 'data'. Derfor har vi nu valgt (med 'select()') alle værdier (med 'everything()'). Derefter hældes 'data' videre over i den næste funktion, gennem endnu et rør. Her bruges funktionen 'summarise_all()' til at opsummere de værdier den gives. Inde i funktionen angives endnu en funktion, 'funs()', som laver en liste over de navngivet variabler som vi valgte med 'select(everything())'. Med funktionen, 'is.na(.)', i 'funs()', siger man at R skal fokusere på de manglende værdier, som fremgår i vores data som 'NA'.

```
data %>%
  select(everything()) %>%
  summarise_all(funs(sum(is.na(.))))
```

```
## # A tibble: 1 × 12
##    Navn Organisation  Type Sprog Filtype `Periode (Start)` `Periode (Slut)`
##   <int>        <int> <int> <int>   <int>             <int>            <int>
## 1     0            1    40    25       1                71               71
```

```
## # ℹ 5 more variables: Tags <int>, Emne <int>, url <int>, Licens <int>,
## #   Dokumentation <int>
```

På baggrund af denne oversigt er det tydeligt at fire typer metadata er stærkt overset eftersom omkring halvdelen af de 83 sprogressourcer mangler metadata om 'Type', 'Periode (Start)', 'Periode (Slut)' og 'Dokumentation'. Derfor fokuseres den kvantitative undersøgelse andet mere repræsentativt metadata for sprogressourcer på sprogteknologi.dk, som 'Navn', 'Organisation', 'Sprog', 'Tags' og 'Emne'.

Analyse - Organisationer

Herunder analyseres organisationerne bag sprogressourcerne som opgivet i deres metadata. Først optælles organisationerne bag sprogressourcerne, derefter visualiseres optællingen for at gøre dataen mere tilgængelig for undersøgelse.

Først angives den data man ønsker at undersøge ved at give navnet på variablen det er gemt under. Igen anvendes et rør for at gennemføre manipuleringer af dataen. Her optælles de hvor mange sprogressourcer de forskellige organisationer i dataen har. Med funktionen, 'count()', kan man tælle hvor de antal gange en værdi optræder i et datasæt ved at angive navnet på den kolonne man ønsker at tælle på. Man kan også fortælle funktionen at den skal angive de optalte værdier i en ordnet liste ved at skrive ', sort = TRUE' efter navnet på kolonnen man ønsker at tælle.

```
data %>%
  count(Organisation, sort = TRUE)
```

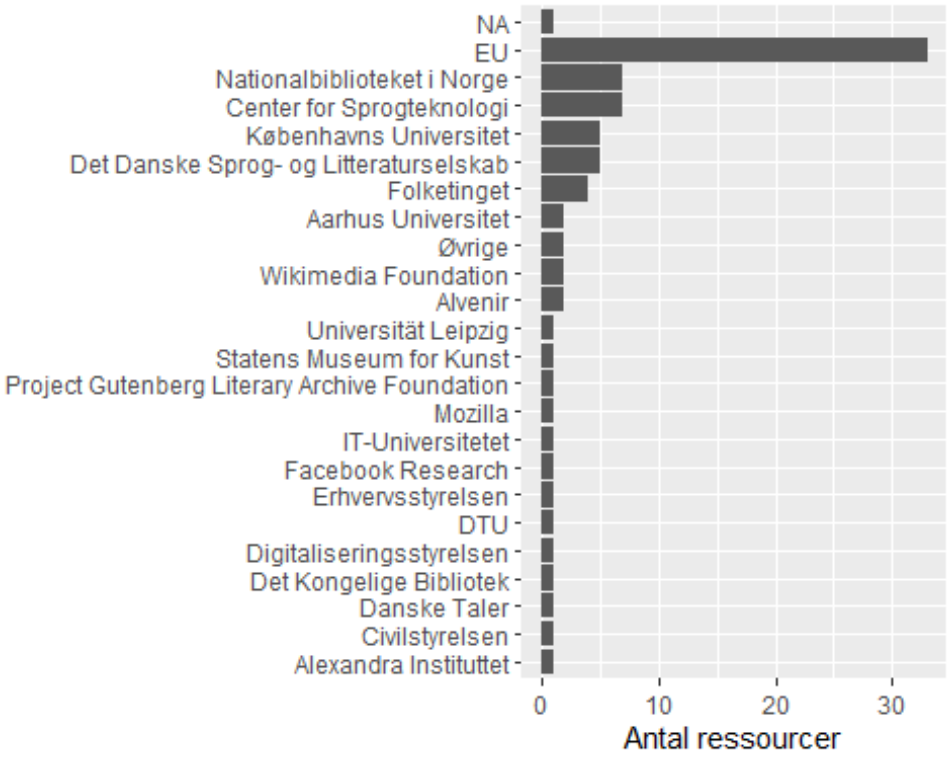
```
## # A tibble: 24 × 2
##    Organisation                              n
##    <chr>                                 <int>
##  1 EU                                       33
##  2 Center for Sprogteknologi                 7
##  3 Nationalbiblioteket i Norge               7
##  4 Det Danske Sprog- og Litteraturselskab    5
##  5 Københavns Universitet                    5
##  6 Folketinget                               4
```

```
##  7 Aarhus Universitet                          2
##  8 Alvenir                                     2
##  9 Wikimedia Foundation                        2
## 10 Øvrige                                      2
## # ℹ 14 more rows
```

Det er tydeligt EU har flest sprogressourcer ud af de 24 organisationer der er i datasættet, men for at fremhæve forskellene mellem hvor mange sprogressourcer hver organisation i visualiseres dette nedenfor.

Her gentages den forrige kode for at optælle organisationerne, derefter isoleres den data vi ønsker at bearbejde med funktionen, 'mutate()', hvor det angives hvilke værdier i variablen der skal isoleres. Så anvendes, '= reorder()' til at forberede dataen på visualisering og gør at dataen bliver sorteret i visualiseringen, ved at dataen 'Organisation' oplistes efter hvor mange gange de fremgår, 'n'. Dette føres videre med et rør til den egentlige visualisering med funktionen, 'ggplot()', inde i funktionen gentages det hvilken data der skal visualiseres, 'aes()', med antallet af observationer, 'n', i kolonnen, 'Organisation'. I funktionspakken, 'ggplot', har '+' den samme funktion som '%>%'. Derefter angives det hvilken type graf der skal visualiseres, her bruges, 'geom_col()', da visualiseringen skal være et horisontalt søjlediagram. Endeligt benyttes 'labs()' til at angive hvilket navn y- og x-aksen skal hedde ved at skrive det ønskede navn. X-aksen gives navnet 'Antal ressourcer' efter 'x='. For at organisationernes navne er i fokus ønskes der ikke noget navn på y-aksen, derfor angives 'y=' værdien 'NULL', som betyder at y-aksens navn ikke skal være noget. Hvis akserne ikke aktivt gives navne med 'labs()' får de automatisk navnene i 'aes()'.

```
data %>%
  count(Organisation, sort = TRUE) %>%
  mutate(Organisation = reorder(Organisation, n)) %>%
  ggplot(aes(n, Organisation))+
    geom_col()+
    labs(x="Antal ressourcer", y=NULL)
```



Analyse - Textmining og Ordsky

Herunder splittes navnene på sprogressourcerne ad, så de forekommer som enkelte ord. Funktionen, 'unnest_tokens()', anvendes til at dele variabler bestående af flere ord op i enkelte ord, men beholde alt den omliggende data i datasættet. For at gøre dette angives det inde i funktionen hvilken kolonne funktionen skal udføres på 'Navn'. Da funktionen 'unnest_tokens()' skaber en ny kolonne i datasættet med de enkelte ord, skal det også angives hvad den nye kolonne skal hedde, her anvendes ordet 'word'. Grunden til anvendelsen af det engelske ord for 'ord' benyttes forklares senere, da det har betydning for en senere datamanipulering. Endeligt gemmes disse indgreb i dataen under en ny variabel, ved at benytte pil operatoren '->', som gemmer alt på venstre side af pilen i variablen til højre for pilen, her under navnet 'navne_data'.

```
data %>%
  unnest_tokens(word, Navn) -> navne_data
```

Vi kan da se hvordan dette manipulere det originale datasæt med 'print()'. Nu er hvert ord i ethvert navn blevet til sin egen række i datasættet, og indeholder stadig alle de tilhørende data.

```
print(navne_data)
```

```
## # A tibble: 558 × 12
##    Organisation     Type  Sprog Filtype `Periode (Start)` `Periode (Slut)` T
ags
##    <chr>            <chr> <chr> <chr>               <dbl>            <dbl> <
chr>
##  1 Alexandra Insti… <NA>  dan   parquet                NA               NA c
ora…
##  2 Alexandra Insti… <NA>  dan   parquet                NA               NA c
ora…
##  3 Alexandra Insti… <NA>  dan   parquet                NA               NA c
ora…
##  4 Alexandra Insti… <NA>  dan   parquet                NA               NA c
ora…
##  5 Alexandra Insti… <NA>  dan   parquet                NA               NA c
ora…
##  6 Det Danske Spro… Korp… dan   html                 1872             1890 t
ekst
##  7 Det Danske Spro… Korp… dan   html                 1872             1890 t
ekst
##  8 Det Danske Spro… Korp… dan   html                 1872             1890 t
ekst
##  9 Det Danske Spro… Korp… dan   html                 1872             1890 t
ekst
## 10 Det Danske Spro… Korp… dan   html                 1872             1890 t
ekst
## # ℹ 548 more rows
## # ℹ 5 more variables: Emne <chr>, url <chr>, Licens <chr>, Dokumentation <ch
r>,
## #   word <chr>
```

Optæl ordene med ligende kode som forklaret under analysen af organisationer. I stedet for organisationer tælles det hvilke ord der fremgår.

```
navne_data %>%
  count(word, sort=TRUE)
```

```
## # A tibble: 245 × 2
##    word          n
##    <chr>     <int>
##  1 danish       39
##  2 corpus       29
##  3 bilingual    25
##  4 from         22
##  5 parallel     22
##  6 english      21
##  7 the          18
##  8 website      18
##  9 of           14
## 10 da           12
## # ℹ 235 more rows
```

Ud fra optællingen kan det ses at nogle af de mest hyppigt forekommende ord er såkaldte fyldeord, som 'from', 'the' og 'fra', hvilket ikke siger særlig meget om det egentlige indhold, derfor ønskes det at frasortere disse fyldeord.

Dette opnås med stopordslister. Stopordsliser indeholder en lang række fyldeord. Herunder indhentes en stopordsliste over danske stopord, som er blevet udarbejdet af Bertel Torp og gjort tilgængelig på github. Som udgangspunkt indhentes stopordslisten på samme måde som den indsamlede data fra sprogteknologi.dk blev indhentet, her anvendes funktionen, 'read_csv()'. For at understøtte det efterfølgende kodestykke sættes navnet på kolonnen i dataen til navnet 'word', ved at skrive det nye navn mellem citationstegn efter metoden, 'col_names ='.

```
stopord <- read_csv("https://gist.githubusercontent.com/berteltorp/0cf8a0c7afea7f25ed754f24cfc2467b/raw/fa34ef448aff6adbb4b6bab9bda62a8b0f1ee597/stopord.txt", col_names = "word")
```

```
## Rows: 351 Columns: 1
## ── Column specification ────────────────────────────────────────────────
```

```
──────
## Delimiter: ","
## chr (1): word
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this mes
sage.
```

Herunder vises det hvordan stopordlisten er opbygget af enkelte fylde ord per række.

```
print(stopord)

## # A tibble: 351 × 1
##    word
##    <chr>
##  1 ad
##  2 af
##  3 akkurat
##  4 al
##  5 aldrig
##  6 alene
##  7 alle
##  8 allerede
##  9 alligevel
## 10 alt
## # ℹ 341 more rows
```

Herunder hældes variablen 'navne_data' gennem funktionen, 'anti_join()', som går ind og fjerner alle overensstemmende værdier, først i 'navne_data' og 'stopord' der fjerner alle de danske stopord, og efterfølgende i 'navne_data' og 'stop_words', som er en indbygget stopordsliste i R som indeholder engelske fyldeord. Fordi 'stop_words' kun fungerer på kolonner navngivet "word", har vi anvendt samme navngivning i datasættet og til den danske stopordsliste. Til sidst gemmes datasættet uden fyldeord under en ny variabel, 'navne_data_u_stopord'.

```
navne_data %>%
  anti_join(stopord) %>%
  anti_join(stop_words) -> navne_data_u_stopord

## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

For forklaring henvises der til optælling af organisationer

```
navne_data_u_stopord %>%
  count(word, sort = TRUE)

## # A tibble: 218 × 2
##    word          n
##    <chr>     <int>
##  1 danish       39
##  2 corpus       29
##  3 bilingual    25
##  4 parallel     22
##  5 english      21
##  6 website      18
##  7 19            6
##  8 dataset       6
##  9 dk            6
## 10 nst           6
## # ℹ 208 more rows
```

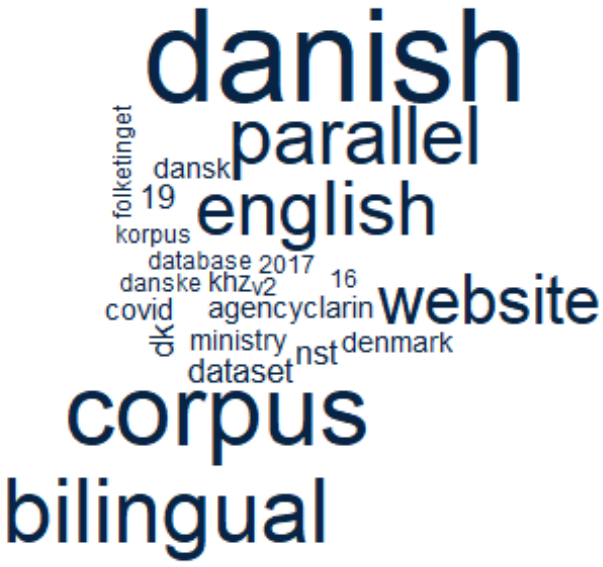
Nu da dataen over navne blevet manipuleret og renset, så den er klar til at blive visualiseret.

Ordsky

For at lave en ordsky gentages først optællingskoden for oven, derefter fremgår koden som visualiserer den rensede data. For at gøre dette anvendes funktionen, 'with()', som transformere dataen til at kunne blive udtrykt som en ordsky, altså er 'with()'-funktionen et forberedende trin inden funktionen 'wordcloud'. Ligesom visualiseringen af organisationer anvises det først hvilke værdier der skal visualiseres, 'word', er ordene og 'n', er hvor mange gange et ord forekommer. Derefter angives en nedre grænse for hvor få forekomster et ord skal have

for at blive inkluderet med argumentet, 'min.freq=', som her sættes lig tre. Altså ekskluderes alle de ord som kun forekommer to eller en gang(e). Så er det muligt at angive hvilken farve ordene i ordskyen skal visualiseres i med, 'colors =', værdien kan da blive angivet i en hex-kode, som er en kategorisering af farver i koder. Hver gang kodeblokken køres dannes visualiseres ordene i en ny formation.

```
navne_data_u_stopord %>%
  count(word, sort=TRUE) %>%
  with(wordcloud(word,n,min.freq=3, colors = "#082444"))
```



Analyse - Sprog

For at analysere hvilke sprog sprogressourcerne indeholder bliver dataen først manipuleret, hvorved metadataen om sprog udfoldes og isoleres.

Først benyttes funktionen, 'mutate()', til at manipulere kolonnen, 'Sprog'. Her opdeles de sprogressourcer som indeholder mere en ét sprog, ved at benytte funktionen, 'str_split()', som fungere ved at opdele ord og sætninger ved at dele dem på et bestemt tegn eller mønster. Her defineres stedet som værdierne i 'Sprog' skal deles på som kolon, der er blevet

anvendt i datasættet til at opdele værdierne når flere værdier var angivet under et punkt i metadataen. Efter dette benyttes funktionen, 'unnest()', som gør at sprog med flere sprog bliver gentaget lige så mange gang som de har antal af sprog, men kun med et sprog per række. Til sidst gemmes den manipulerede data i variablen, 'sprog_data'.

```
data %>%
  mutate(Sprog = str_split(Sprog, ":")) %>%
  unnest(Sprog) -> sprog_data
```

Vi kan da se hvordan koden har manipulere det originale datasæt med 'print()'. Nu er hvert sprog for hver ressource blevet til sin egen række i datasættet, og indeholder stadig alle de tilhørende data.

```
print(sprog_data)
```

```
## # A tibble: 148 × 12
##    Navn      Organisation Type  Sprog Filtype `Periode (Start)` `Periode (Sl
ut)`
##    <chr>     <chr>        <chr> <chr> <chr>                <dbl>            <
dbl>
##  1 Coral te… Alexandra I… <NA>  dan   parquet                NA
NA
##  2 Georg Br… Det Danske … Korp… dan   html                  1872
1890
##  3 PAROLE-D… Det Danske … Korp… dan   xml, t…              1998
2015
##  4 PAROLE-D… Det Danske … Korp… eng   xml, t…              1998
2015
##  5 PAROLE-D… Det Danske … Korp… fin   xml, t…              1998
2015
##  6 PAROLE-D… Det Danske … Korp… vls   xml, t…              1998
2015
##  7 PAROLE-D… Det Danske … Korp… fre   xml, t…              1998
2015
##  8 PAROLE-D… Det Danske … Korp… gre   xml, t…              1998
2015
```

```
##  9 PAROLE-D… Det Danske … Korp… dut   xml, t…              1998
2015
## 10 PAROLE-D… Det Danske … Korp… gle   xml, t…              1998
2015
## # ℹ 138 more rows
## # ℹ 5 more variables: Tags <chr>, Emne <chr>, url <chr>, Licens <chr>,
## #   Dokumentation <chr>
```

Herunder udføres en ligende kodning som forklares under analysen af organisationer, som optæller de enkelte sprog i sprogressourcerne på sprogteknologi.dk.

```
sprog_data %>%
  count(Sprog, sort = TRUE)
```

```
## # A tibble: 28 × 2
##    Sprog     n
##    <chr> <int>
##  1 dan      55
##  2 eng      35
##  3 <NA>     25
##  4 dut       2
##  5 fin       2
##  6 fre       2
##  7 ger       2
##  8 gle       2
##  9 ita       2
## 10 por       2
## # ℹ 18 more rows
```

Herunder udføres en ligende kodning som forklares under analysen af organisationer. Til forskel for den forrige optælling laver det understående kodestykke en optælling af sprogsammensætningerne i sprogressourcerne.

```
data %>%
  count(Sprog, sort = TRUE)
```

```
## # A tibble: 6 × 2
##   Sprog
n
##   <chr>                                                                <
int>
## 1 dan:eng
30
## 2 <NA>
25
## 3 dan
23
## 4 eng
3
## 5 dan:eng:fin:vls:fre:gre:dut:gle:ita:cat:nor:por:swe:ger
1
## 6 hrv:lav:slv:slo:ita:est:por:swe:dut:dan:spa:bul:eng:rum:gle:cze:mlt:ger…
1
```

Analyse - Tags

For forklaring af dette kodestykke se analysen af sprog.

```
data %>%
  mutate(Tags = str_split(Tags, ":")) %>%
  unnest(Tags) -> tag_data
```

Herunder udføres ligende kodning som forklares under analysen af organisationer.

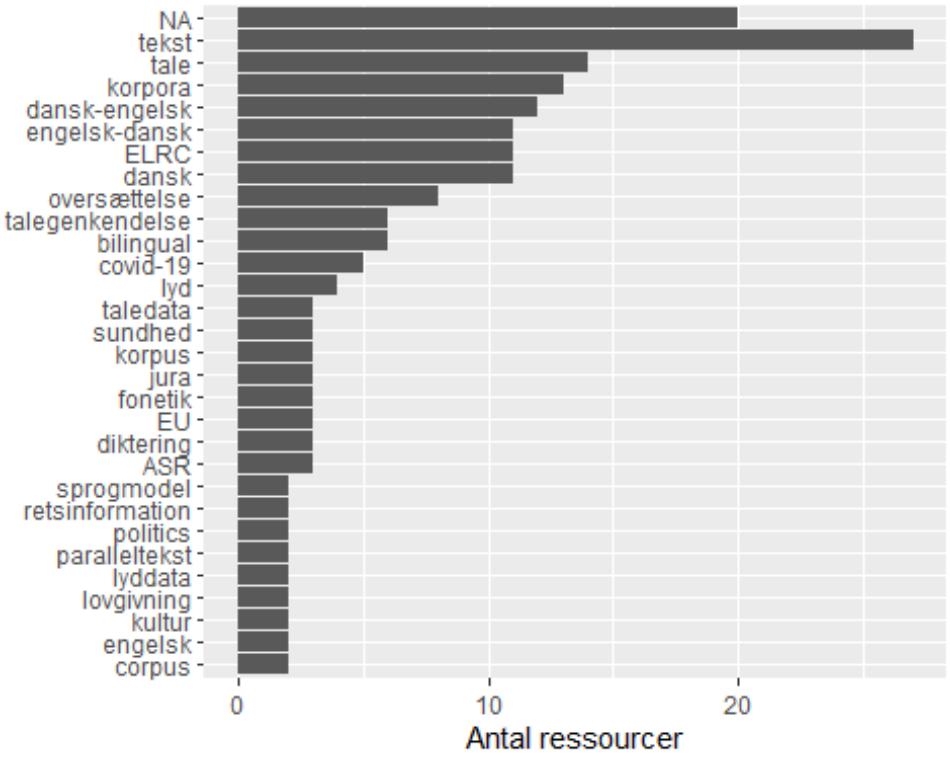
```
tag_data %>%
  count(Tags, sort=TRUE)
```

```
## # A tibble: 113 × 2
##    Tags              n
##    <chr>         <int>
##  1 tekst            27
##  2 <NA>             20
##  3 tale             14
```

```
##  4 korpora          13
##  5 dansk-engelsk    12
##  6 ELRC             11
##  7 dansk            11
##  8 engelsk-dansk    11
##  9 oversættelse      8
## 10 bilingual         6
## # ℹ 103 more rows
```

Denne visualisering følger samme tilgang som forklaret under analysen af organisationer. Dog anvendes funktionen, 'filter()' som filtrere værdierne den gives, ved at skrive, 'n>1', angives det at værdier som fremgår mindre end 2 gange skal frasorteres fra visualiseringen.

```
tag_data %>%
  count(Tags, sort = TRUE) %>%
  filter(n>1) %>%
  mutate(Tags = reorder(Tags, n)) %>%
  ggplot(aes(n, Tags))+
    geom_col()+
    labs(x="Antal ressourcer", y=NULL)
```



Analyse - Emner

For forklaring se analysen af sprog.

```
data %>%
  mutate(Emne = str_split(Emne, ":")) %>%
  unnest(Emne) -> emne_data
```

For forklaring se analysen af organisationer.

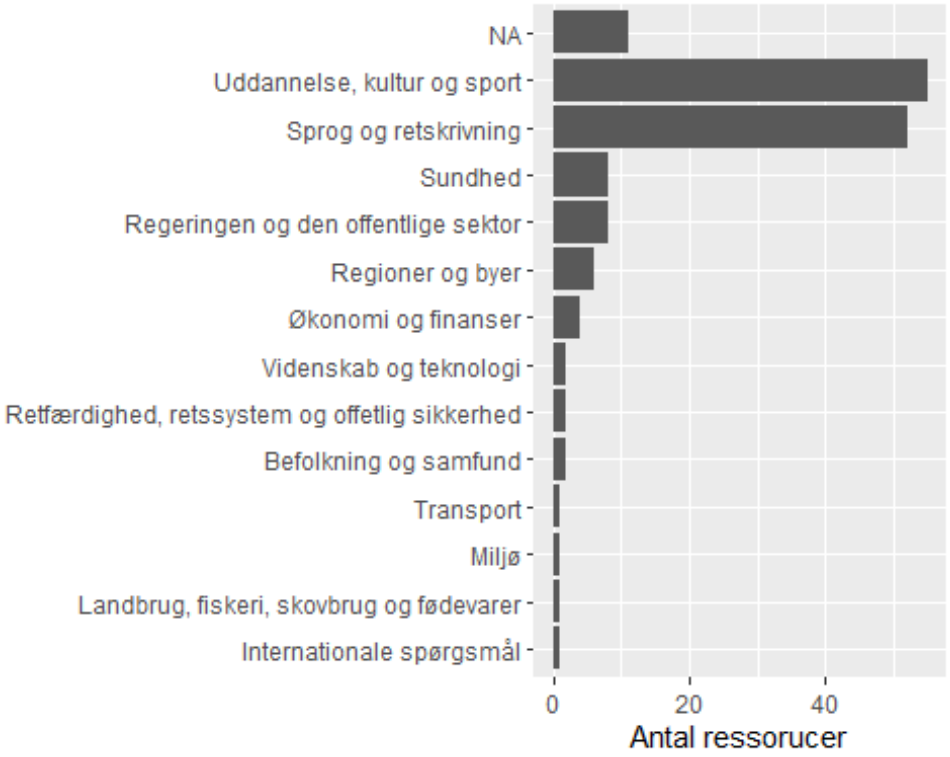
```
emne_data %>%
  count(Emne, sort=TRUE)
```

```
## # A tibble: 14 × 2
##    Emne                                          n
##    <chr>                                     <int>
##  1 Uddannelse, kultur og sport                  55
##  2 Sprog og retskrivning                        52
##  3 <NA>                                         11
##  4 Regeringen og den offentlige sektor           8
##  5 Sundhed                                       8
```

```
##  6 Regioner og byer                                  6
##  7 Økonomi og finanser                               4
##  8 Befolkning og samfund                             2
##  9 Retfærdighed, retssystem og offetlig sikkerhed    2
## 10 Videnskab og teknologi                            2
## 11 Internationale spørgsmål                          1
## 12 Landbrug, fiskeri, skovbrug og fødevarer          1
## 13 Miljø                                             1
## 14 Transport                                         1
```

For forklaring se visualisering under analysen af organisationer.

```
emne_data %>%
  count(Emne, sort = TRUE) %>%
  mutate(Emne = reorder(Emne, n)) %>%
  ggplot(aes(n, Emne))+
    geom_col()+
    labs(x="Antal ressorucer", y=NULL)
```



Anvendte resourcer

R for Data Science: https://r4ds.hadley.nz/data-import.html

Sprogteknologi.dk: https://sprogteknologi.dk/group/corpora

The R Graph Gallery: https://r-graph-gallery.com/barplot.html

Text Mining with R, A Tidy Approach: https://www.tidytextmining.com/