Datamanagementplan for COR.SEM – den semantiske komponent til det Centrale Ordregister for dansk

Endelig version, 27-08-2024

Om projektet

Centralt OrdRegister for dansk (COR) er et samarbejdsprojekt mellem Det Danske Sprog- og Litteraturselskab (DSL), Dansk Sprognævn, Center for Sprogteknologi (CST), NorS, KU og Digitaliseringsstyrelsen om at udvikle en fælles dansk sprogressource til AI-formål. Formålet med projektet er dels at støtte en effektiv deling af danske sprogressourcer ved at udarbejde et nummersystem for alle danske ord, dels at stille et betydningsinventar for danske ord (lemmaer) til rådighed for virksomheder og forskere der arbejder med AI. Denne semantiske komponent kaldes COR.SEM, og den indarbejder og opgraderer bl.a. ressourcer der tidligere er udviklet ved CST og DSL såsom det danske wordnet DanNet, Dansk FrameNet-leksikon og Dansk Sentimentleksikon.

Denne datamanagementplan gælder kun for COR.SEM.

Internt på Center for Sprogteknologi, NorS, KU, fungerer

- Bolette S. Pedersen (bspedersen@hum.ku.dk) som PI.
- Sussi Olsen (saolsen@hum.ku.dk) som data-ansvarlig (back-up).

Beskrivelse af data:

COR indeholder ikke personfølsomme data, men data der indgår i enten Retskrivningsordbogen eller Den Danske Ordbog.

Rettighederne til disse data ligger hos de to ordbogsudbydere, Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab.

Derudover indeholder COR.SEM oplysninger udtrukket og bearbejdet fra

- Den Danske Begrebsordbog (rettighedshaver: DSL),
- DanNet (https://wordnet.dk/dannet, licens CC-BY-SA 4.0)
- Det Danske FrameNet-leksikon (DSL, ret åben licens, se https://github.com/dsldk/dansk-frame-net/blob/master/LICENSE)
- Dansk Sentiment-leksikon (https://github.com/dsldk/danish-sentiment-lexicon, licens CC-BY-SA 4.0)

Datamanagement i projektforløbet

CST, NorS, KU deltager i arbejdspakkerne 2: COR-S (som arbejdspakkeleder) og og 3: COR-SX sammen med Det Danske Sprog- og Litteraturselskab (DSL).
For begge arbejdspakker gælder det at data udtrækkes til fungerende arbejdsfiler af DSL.

For at flere ad gangen og fra begge institutioner kan arbejde i datafilerne samtidig, arbejdes der hovedsagelig i Google docs/sheets. DSL har backup-plan for alle data, og CST laver desuden backup af alle nyligt ændrede filer, ca. en gang om ugen. Disse back up-filer ligger på KU's N-drev (N:\HUM-NORS-cst-projekter\COR\back-up_filer).

Alle færdige data kopieres løbende til DSL's iLex-database, som der også foretages regelmæssig backup af.

Se Bilag 1 for en erklæring fra Sanni Nimb, DSL, arbejdspakkeleder af arbejdspakke 3, om opbevaring af data i løbet af projektet.

Efter projektets afslutning

Anvendelse og udstilling

I projektbeskrivelsen (pkt. 6) står der følgende om distribution af data efter projektets afslutning:

"Anvendelse og udstilling

COR bliver tilgængeligt via sprogteknologi.dk, dsn.dk, clarin.dk, CST's github-repositorie samt DSL's hjemmeside + github-repositorie. COR leveres som et API designet efter aktuelle best practices. Der udformes manualer og arrangeres seminarer til målgruppen om mulighederne ved at anvende COR i takt med at de forskellige elementer af COR lanceres (jf. arbejdspakkerne).

Der gives licens til at COR frit kan anvendes til de fleste formål, også kommercielt. Der må frit vælges ud og føjes til, og COR må videredistribueres i et hvilket som helst medie eller format. COR og de sproglige ressourcer heri må dog ikke bruges til at udvikle selvstændigt ordbogsprodukter som er i konkurrence med Den Danske Ordbog, Den Danske Begrebsordbog og Retskrivningsordbogen hverken i trykt eller elektronisk form, men udelukkende integreret i programmer til andre formål, fx stavekontrol, ordspil, undervisning, informationsøgning mv. Ligeledes må der ikke laves ordbøger henvendt til mennesker ud fra en delmængde af COR-data, fx inden for et bestemt domæne.

COR skal krediteres i publikationer og digitale produkter der helt eller delvist gør brug af COR, og enhver brug skal registreres i API'et.

Ud fra den beskrevne betragtning om anvendelse og udstilling vil Digitaliseringsstyrelsen parallelt med udviklingsprojektet undersøge og fastlægge en nærmere licensbeskrivelse til brug for COR. "

Pga. licensforhold er COR.SEM opdelt i to resurser:

- COR.SEM med de fleste oplysningstyper under en **CC0**-licens.
- COR.SEM.EXT, som indeholder IPR-beskyttet materiale fra DDO, under licensen **CC BY-NC-ND**.

Hele COR-resursen (inkl. COR.SEM) ligger på adressen

- https://ordregister.dk (hostet af Dansk Sprognævn).

Derudover ligger COR.SEM i

- DSL's iLex-database (ikke offentlig), som backes up regelmæssigt.
- CST's og DSL's github-sider:
  - https://github.com/kuhumcst
  - https://github.com/dsldk/
- CLARIN-DK:

- https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/50 (COR.SEM)
- https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/51 (COR.SEM.EXT)

DSL hoster en hjemmeside, hvor man kan søge i COR.SEM-data: https://corsem.dsl.dk/

Drift og vedligehold

Driften af hjemmesiden https://ordregister.dk foretages af Dansk Sprognævn der også administrerer den del COR som ikke er COR.SEM og COR.SEMEXT.

Driften af de øvrige repositorier, hvor COR.SEM og COR.SEM.EXT findes, hører under disse.

Jf. projektbeskrivelsen for COR skal COR.SEM (og COR.SEMEXT) videreudvikles som beskrevet nedenfor

- *"DSL vil, under forudsætning af fortsatte bevillinger til at videreudvikle Den Danske Ordbog, udarbejde semantiske oplysninger for den delmængde af Retskrivningsordbogens nye opslagsord, som også er beskrevet i Den Danske Ordbog, dog maksimalt 100 opslagsord.*
- *CST forpligter sig til i tre år efter projektets afslutning i samarbejde med DSL at vedligeholde COR's semantikkomponent, dvs. udvikle semantiske oplysninger for nye ord i dansk, dog maksimalt 100 opslagsord."*

Arbejdet med nye ord planlægges udført på samme måde som det hidtidige udviklingsarbejde via delte google sheets som backes up både af CST og DSL. Data vil løbende bliver uploadet til DSL's iLex-database. Én gang årligt vil en ny udvidet version af COR.SEM (og COR.SEMEXT) blive udgivet og uploadet til alle de repositorier, hvor den første version ligger (se ovenfor).

From: Sanni Nimb
To: Sussi Olsen
Cc: Bolette Sandford Pedersen
Subject: Re: Datamanagement for COR-data
Date: 2. juni 2023 12:30:18
Attachments: signature.png

Kære Sussi
vi har ikke særlige aftaler vedr. adgang til data når projektet er slut (udover at de skal lægges på ordregister.dk) - dvs. det står allerede i din tekst. Men vi vil formentlig også lægge dem på vores egen hjemmeside. Det vil jeg bare ikke love direkte nu.
Alle data som Thomas udarbejder til brug for projektet, ligger også på DSL's egne servere, er det blot det vi skal erklære?
Og de færdige data ligger også i en kopi på DSl's egen server.
Og i enndu en kopi via iLex.
Det hele backes op systematisk på DSL.
Og vi laver backup af data/arbejdsfiler på egne bærbare hele tiden.
Kh. Sanni

Den fre. 2. jun. 2023 kl. 12.17 skrev Sussi Olsen <saolsen@hum.ku.dk>:

> Kære Sanni
>
> Bolette har modtaget en forespørgsel fra KU ang. data management for COR-dataene. Hun har svaret at dem har I ansvaret for, men de vil gerne have en beskrivelse af hvordan vi behandler data - så de har dokumentation for at det ikke er KU's ansvar.
>
> Se et uddrag af mailen:
>
> "Hvis der er indgået en aftale om at du/NorS har adgang til data efter projektafslutning, så bør den aftale (både hvis det er en mail eller et selvstændigt dokument) journaliseres, når projektet ophører.
>
> Men der skal udarbejdes – og journaliseres - **en kort beskrivelse af at det er DSL, der har haft al datahåndtering i projektet**. Denne korte beskrivelse kan man i dette tilfælde godt kalde en datamanagementplan.
>
> Så helt lavpraktisk kan du eller en anden der har arbejdet på projektet svare på denne mail med den info.
>
> Samt vedlægge hvad I måtte have af aftaler om adgang til data efter projektet er slut."
>
> Jeg har forfattet et lille stykke som forhåbentlig indeholder den info de vil have.
>
> Gider du læse det igennem og evt. rette hvis jeg skriver noget du ikke mener passer?
>
> Og så skrive i en mail tilbage at det er korrekt at I opbevarer data. Så kan jeg sende dem dokumentet og vedlægge din mail. Det skulle være nok til at tilfredsstille dem.
> Kh. Sussi
> --

Sussi Olsen
Videnskabelig medarbejder/Research Associate

Københavns Universitet/University of Copenhagen
Institut for Nordiske Studier og Sprogvidenskab
Center for Sprogteknologi
Emil Holms Kanal 2
2300 København S

DIR 35 32 90 64
MOB 30 26 70 04
saolsen@hum.ku.dk
www.cst.ku.dk

--

Sanni Nimb
Ledende redaktør, ph.d.

DET DANSKE SPROG- OG LITTERATURSELSKAB, DSL
Society for Danish Language and Literature

Christians Brygge 1
DK-1219 København K
Tel. +45 27265640
Fax +45 33140608

http://www.dsl.dk