Indeksering i COR – teknisk specifikation

(ver 1.0 30/9 2022)

1 Det Centrale Ordregister (COR)

COR er et register over det danske sprogs ord og lemmaer særligt udviklet til support af sprogteknologiske udviklingsprojekter. Hvert dansk lemma og hver dansk ordform har (eller kan få) tilknyttet et unikt COR-indeks. Registeret er gratis, frit tilgængeligt og åbent for enhver anvendelse.

2 COR-indeksering – formel specifikation

Nye leksikalske ressourcer kan tilføjes til COR niveau 2[1] og 3. Nye COR-ressourcer skal overholde de formelle syntaksregler for navngivning (domænenavn og indeksering) for at kunne kalde sig **COR-kompatible**. COR-indekser kan fx se sådan ud:

`COR.KEMITERMER.012`	(COR niv.2)
`COR.FONETIK001.0012345.0002.03.06`	(COR niv.2)
`COR.DAN_ENG.00012345.002.01.02.03.0123`	(COR niv.2)
`COR.OPEN.KANINAVLEREN.456`	(COR niv.3)
`COR.OPEN.MITFIRMA.009876543.543.123.0.0.0`	(COR niv.3)

Domænenavne skal skrives med engelske bogstaver (`A-Z`) og cifre. Underscore (_) kan også bruges. Case er ikke signifikant (dvs. `FISK24`, `Fisk24` og `fisk24` er det samme domæne), men vi anbefaler at man kun bruger uppercase (`A-Z`) i domænenavne.

1 Kun medlemmer af Dansk Sprognævns repræsentantskab kan tilføje ressourcer til COR niveau 2; niveau 3 er frit for alle. Repræsentantskabet er udpeget af Kulturministeriet, se `https://kum.dk/ministeriet/organisation-og-institutioner/bestyrelser-raad-naevn-og-udvalg/dansk-sprognaevn-repraesentantskab`

EBNF-syntaks

```
COR-indeks    =  Præfiks  { "." Ciffergruppe }+
Præfiks       =  Niv1 | Niv2 | Niv3

Ciffergruppe =  Ciffer {Ciffer}+
Ciffer        =  "0".."9"

Niv1          =  "COR"
Niv2          =  "COR"  "."  Domænenavn
Niv3          =  "COR"  "."  "OPEN"  "."  Domænenavn

Domænenavn    =  Bogstav Alfanum+
Bogstav       =  "A"."Z"
Alfanum       =  Bogstav | Ciffer | "_"
```

Hertil kommer fem ekstra regler af praktisk natur:

- domænenavne skal have mindst 3 tegn
- `"COR"`, `"PRIVAT"`, `"PRIVATE"`, `"AABEN"`, `"OPEN"` er ikke lovlige domænenavne
- nye domænenavne skal være hidtil ubenyttede (dvs. må ikke være registreret hos DSN hverken i niv.2 eller niv.3)
- vi anbefaler at domænenavne er deskriptive, samt at `COR` ikke indgår
- Case er ikke signifikant (dvs. `FISK24`, `Fisk24` og `fisk24` er det samme domæne), men vi anbefaler at man kun bruger uppercase (`A-Z`) i domænenavne.

Syntaksen for COR-indekser udtrykt som regex

(unix/perl/python-stil)

Grundlæggende serie (lemma): `/^COR\.\d+$/`
Grundlæggende serie (ordform): `/^COR\.\d+\.\d+\.\d+$/`
Registreret serie: `/^COR\.[A-Z][A-Z0-9_][A-Z0-9_]+(\.[0-9]+)+$/`
Åben serie: `/^COR\.OPEN\.[A-Z][A-Z0-9_][A-Z0-9_]+(\.[0-9]+)+$/`

Ovenstående regexes beskriver de anbefalede serienavne. Tilføjelse af et `i` til sidst i udtrykkene vil gøre dem case-insensitive.
OBS. 'COR', 'PRIVAT', 'PRIVATE', 'AABEN' og 'OPEN' er ikke lovlige domænenavne

3 Specifikation af den grundlæggende indeksserie

Som nævnt ovenfor har COR-indekser i grundserien (også kaldet niv.1) formen

Lemmaer: "COR" "." $Cifre_1$
Ordformer: "COR" "." $Cifre_1$ "." $Cifre_2$ "." $Cifre_3$

Desuden gælder

- Lemmaer ($Cifre_1$)
  - Lemmaindekser i grundserien er altid 5 cifrede, om nødvendigt med foranstillede 0'er. Lemmaindekser fra niv. 2 og 3 kan indeholde flere cifre.
  - Nummerserien 00010-00999 er forbeholdt funktionsord (dvs. lemmaer i RO som hverken er bestemt som subst., verb., adj., proprier, talord eller adv.
  - Nummerserien 10000-14999 er forbeholdt de lemmaer der er bestemt som adverbier i RO.
  - Nummerserien 15000-29999 er forbeholdt de lemmaer der er bestemt som adjektiver i RO.
  - Nummerserien 30000-39999 er forbeholdt de lemmaer der er bestemt som verber i RO.
  - Nummerserien 40000-99999 er forbeholdt de lemmaer der er bestemt som substantiver i RO.
- Bøjningsindeks ($Cifre_2$)
  - Bøjningsindekser er altid 3-cifrede, om nødvendigt med foranstillede 0'er.
  - Hver bøjningsform (fx de regelmæssigt bøjede substantivers 8 former inkl. ejefald) har et unikt indeks.
  - Indekser 100-199 er forbeholdt verber; 200-299 adjektiver og adverbier (sidstnævnte bøjes dog kun undtagelsesvist), 300-399 substantiver.
  - Flere indeksserier kan evt. inddrages senere.
  - Den konkrete indeksering vil have en finhedsgrad sammenlignelig med PAROLE's morfosyntaktiske annotationstabel for dansk (den ikke-reducerede version, se ref.).
- Varianter ($Cifre_3$)
  - Variantindekser er altid 2-cifrede, om nødvendigt med foranstillet 0.
  - Hver variant (fx "mix", "miks" for lemmaet "miks") har et unikt indeks.

4 Retningslinjer for indeksserier uden for den grundlæggende serie

I de registrerede og åbne COR-serier (dvs. serier med sammensat præfiks) er ciffergrupper i COR-indekser uden begrænsninger (jf. COR-indeks og Cifre i EBNF-syntaks), med andre ord: Ciffergruppernes længde og antal kan vælges frit.

I de registrerede og åbne COR-serier (serier med sammensat præfiks) *tilrådes* det kraftigt at benytte et deskriptivt præfiksnavn som henviser til en konkret ordressource (korpus, leksikon, termbase, etc.).

I de registrerede og åbne COR-serier (serier med sammensat præfiks) *tilrådes* det kraftigt at benytte den samme indeksering, evt. med foranstillede 0'er, for lemmaer som kan anses for identiske med lemmaer i den grundlæggende serie (dvs. samme konkrete værdier for $Cifre_1$ omtalt herover).

I de registrerede og åbne COR-serier (med sammensat præfiks) *tilrådes* den samme brug af bøjningsindekser, evt. med foranstillede 0'er, som i den grundlæggende serie (samme konkrete ciffergruppe for $Cifre_2$ omtalt herover).

I de registrerede og åbne COR-serier kan der frit tilføjes nye ciffergrupper efter behov hvis mere finstruktur er påkrævet (fx til disambiguering af lemmaer) jf. EBNF-syntaksen. Eksempler: COR.DDO.012345.002.0034 (for niv.1-lemma 12345, bøjning 002, betydningsvariant 0034); COR.ODS.0012345.002.0034.02 (for samme lemma med finere underdelt betydningsvariant); COR.KLE.012345.9876.54.321 (for samme lemma, med indekseret klassifikation).

Er et COR-indeks (uanset serie) først publiceret, må det under ingen omstændigheder ændre denotation. Indekser kan derimod godt tilbagekaldes og annulleres (hvis de fx viser sig at være uhensigtsmæssige).

COR-præfikserne i de registrerede serier (se Niv2 i EBNF-syntaksen) administreres i projektperioden af den samlede projektgruppe; efter projektets slutning af DSN alene. Alle medlemmer af DSN's repræsentantskab er fødte medlemmer af gruppen af **registrerede COR-redaktioner** (bl.a. DSL, CST, DSN, de danske universiteter, DR, TV2, Forfatterforeningen, Oversætterforeningen m.fl.). Registrerede COR-redaktioner kan altid definere nye COR-præfikser til brug for konkrete ordressourcer så længe syntaksen for COR-præfikser overholdes.

COR-serierne COR.OPEN.*xyz* er åbne for enhver bruger. Man meddeler bare DSN sin åbne præfikskode (*xyz*) der kan vælges frit så længe (1) den overholder syntaksen for COR-præfikser og (2) ikke er registreret i forvejen hos DSN. Bemærk dog at de ovenstående anbefalinger også gælder for implementeringen af COR-serier i COR.OPEN.*xyz*

COR-redaktioner for registrerede og åbne COR-serier *opfordres kraftigt* til at offentliggøre afbildningstabeller af deres egen COR-serie på andre COR-serier, a forteriori på den grundlæggende indeksserie (Niv1). Dette vil gøre det lettere at importere og eksportere data til og fra samfundets sprogressourcer (COR-annoterede ordbøger, korpora, termbaser, sprogværktøjer etc., mange af dem frit tilgængelige over sprogteknologi.dk).

5 Fremtidig vedligehold af Det Centrale Danske Ordregister

Efter COR-projektets slutning er DSN ansvarlig for den fremtidige administration af COR-registeret, dvs. ajourføring af den grundlæggende indeksserie, samt registrering af nye serier.

Derudover vil DSN være ansvarlig for uddeling og allokering af nye serienumre via en ansøgningsformular. I den forbindelse bliver følgende registreret for alle serienumre: navn, præfiks, nummerserie, url, licens, beskrivelse og en liste over definerede lemmaer.

Serienumre vil blive betragtet som COR-kompatible, hvis de overholder ovenstående retningslinjer.

6 Referencer

Dideriksen, Christina; Peter Juel Henrichsen; Thomas Widmann (2022) *Det Centrale Ordregister*. Nyt Fra Sprognævnet. Oktober 2022. ISSN 2446-3124.

Henrichsen, Peter Juel (2022) *Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi*. Proceedings of NLF2022. Lund University Press.

Pedersen, Bolette; Nathalie Carmen Hau Sørensen; Sanni Nimb; Sussi Olsen; Ida Flørke; Thomas Troelsgård (2022) *Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes. The Open Source COR Lexicon*. Proceedings of LREC2022.

Widmann, Thomas (2022) *Det Centrale Ordregister og dets leksikografiske anvendelser.* Proceedings of NLF2022. Lund University Press.

Kontakt

Dansk Sprognævn
www.dsn.dk
3374 7400

Thomas Widmann, tw@dsn.dk
Peter Juel Henrichsen, pjh@dsn.dk