

Indeksering i COR - teknisk specifikation

(ver 1.1, 13/1 2025)

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-id. Registeret er gratis, frit tilgængeligt og åbent for enhver anvendelse. Registrering af COR-ressourcer varetages af Dansk Sprognævn (DSN).

2 COR-indeksering – formel specifikation

COR's niveau 1 udgøres af leksemerne i RO5 (Retskrivningsordbogen, 5.udg.). Nye leksikalske ressourcer kan tilføjes i COR's niveau 2¹ og niveau 3. Alle COR-ressourcer skal overholde de formelle syntaksregler for navngivning og COR-indeksering for at kalde sig **COR-kompatible**. COR-id'er kan fx se sådan ud:

COR.EXT.100002.300.01	(COR niv.2)
COR.TALE.02668.890.01	(COR niv.2)
COR.KEMI.065825.120	(COR niv.2)
COR.OPEN.MITFIRMA.009876543.543.123.0.0.0	(COR niv.3)
COR.OPEN.KANINAVLEREN.456	(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 Bekendtgørelse af Lov om Dansk Retskrivning, Retsinformation LBK nr 217 af 13/02/2015 (<https://www.retsinformation.dk/eli/lt/2015/217>); den aktuelle besætning ses i <https://kum.dk/ministeriet/organisation-og-institutioner/bestyrelser-raad-naevn-og-udvalg/dansk-sproгнаevn-repraesentantskab>

Syntaks for COR-id (EBNF-stil)

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

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

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

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

Syntaks for COR-id (regex i unix/python/perl-stil)

```
Grundlæggende serie (niv.1, lemma):    /^COR\.\d{5}$/
Grundlæggende serie (niv.1, ordform):   /^COR\.\d{5}\.\d{3}\.\d{2}$/
Restringeret serie (niv. 2):            /^COR\.[A-Z0-9_]+(\.\d+)$/i
Åben serie (niv. 3):                    /^COR\.OPEN\.[A-Z0-9_]+(\.\d+)$/i
```

Hertil kommer fem ekstra regler af praktisk natur:

- domænenavne skal have mindst 3 tegn
- COR, OPEN, AABEN, PRIVAT og PRIVATE 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 i navne
- 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.

3 Specifikation af den grundlæggende indeksserie

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

Lemmaer: "COR" "." Cifre₁

Ordformer: "COR" "." Cifre₁ "." Cifre₂ "." Cifre₃

Desuden gælder

- Lemmaindeks (Cifre₁)
 - Lemmaindekser i grundserien er 5-cifrede, om nødvendigt med foranstillede 0'er. Lemmaindekser i niv. 2 og 3 kan indeholde flere (eller færre) cifre.
 - Nummerområdet 00010-99999 er opdelt i underområder der afspejler RO5's ordklassetaksonomi. Kontakt DSN for nærmere detaljer.
- Bøjningsindeks (Cifre₂)
 - Bøjningsindekser er 3-cifrede, om nødvendigt med foranstillede 0'er.
 - Hver bøjningsform (fx de regelmæssigt bøjede substantivers 9 former (inkl. ejefald og sammensætningsform) har et unikt indeks.
 - Indekser 100-199 er forbeholdt substantiver; 200-299 verber; 300-399 adjektiver.
 - Indekser i 400+ samt øvrige detaljer ses i tabellen bix2bform_UTF8_01.tsv (i ordregister.dk)
 - Flere indeksserier kan evt. inddrages senere.
- Variantindeks (Cifre₃)
 - Variantindekser er 2-cifrede, om nødvendigt med foranstillede 0'er.
 - Hver variant (fx "mix", "miks" for lemmaet *miks*; "gud ske lov" og "gudskelov" for lemmaet *gud ske lov*) 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-id'er uden begrænsninger (jf. COR-id og Cifre i EBNF-syntaks), med andre ord: Ciffergruppernes længde og antal kan vælges frit.

Det tilrådes at benytte deskriptive præfiksnavne for nye COR-ressourcer (korpus, leksikon, termbase, etc.).

For lemmaer i nye COR-ordbøger som anses for ækvivalente med lemmaer i RO5 tilrådes det kraftigt at benytte det samme lemma- og bøjningsindeks som i niv.1 (evt. med foranstillede 0'er).

Formatet for COR-id'er i niv.2 og 3 kan frit tilføjes nye ciffergrupper hvis mere finstruktur er påkrævet (fx til disambiguering af lemmaer), jf. EBNF-syntaksen. Se eksempler i afsn. 2.

Er et COR-id (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 (COR niveau 2) administreres af DSN. Alle medlemmer af DSN's repræsentantskab er fødte medlemmer af gruppen af **registrerede COR-redaktioner** (bl.a. de danske universiteter, DR, TV2, DSL, Dansk Journalistforbund, Dansk Forfatterforening m.fl.) og kan altid oprette nye COR-ordbøger i niv. 2.

COR-serierne i niveau 3 (COR.OPEN.xyz) er åbne for enhver bruger. Kontakt DSN for detaljer.

COR-redaktioner *opfordres kraftigt* til at offentliggøre afbildningstabeller af deres egen COR-serie på andre COR-serier, a fortiori på den grundlæggende indeksserie (niv.1). Dette vil gøre det lettere at importere og eksportere data til og fra samfundets sprogpources (COR-annoterede ordbøger, korpora, termdbaser, sprogværktøjer etc., mange af dem frit tilgængelige over sprogteknologi.dk og ordregister.dk).

5 Vedligehold af Det Centrale Ordregister

DSN er ansvarlig for administrationen af COR-registeret, dvs. ajourføring af den grundlæggende indeksserie (niv.1), samt registrering af nye serier (niv.2 og 3).

Derudover er DSN 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.

Ordbøger med registrerede serienumre (COR niv.2 og 3) vil blive betragtet som COR-kompatible hvis (og kun hvis) de overholder ovenstående retningslinjer.

6 Referencer

6.a Artikler om Det Centrale Ordregister

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

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

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

Widmann, Thomas (2024) "The Central Word Register of the Danish Language". Proceed. of eLex 2023 (Brno); <https://www.youtube.com/watch?v=1xykLvLYgjo>.

6.b Artikler om anvendelser af Det Centrale Ordregister

Henriksen, P.J. (2024) Tekstordet som grammatisk domæne. In: Heegård, J. (Eds.), *Ny forskning i grammatik* 31 (pp. 53-70).

Henriksen, Peter Juel (2024) Tekstordet som grammatisk domæne. IN: *Ny Forskning I*

Grammatik, 1(31), 53–70. <https://doi.org/10.7146/nfg.v1i31.145292>

Henrichsen, Peter Juel (2024) Neology and Nomenclature. Concept clarification as a precursor to computational lexicography. IN: *Proceedings of GWLN-6 (Global Workshop for Lexicography and Neology)*. ISBN 978-0-7961-8474-0.

Henrichsen, Peter Juel (2024) Make Each Morph Count - A New Approach to Computational Lexicography for Text Processing. IN: Despot, Kristina Š. et al. (eds) *Proceedings of EuraLex 2024*, 349-357

Nguyen, Michael Nguyen & Peter Juel Henrichsen (2024) *The Locum Hyphen – A Formal Approach to the Lexicalization of Multiword Expressions With Rich Internal Semantics*; N: Despot, Kristina Š. et al. (eds) *Proceedings of EuraLex 2024*, 113-122

Nimb, S. et al. (2022) COR.SEM – den semantiske del af Det Centrale OrdRegister (COR). *Lexico Nordica* 29 (pp. 73-95).

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. IN: *Proceedings of LREC2022*.

Kontakt

Dansk Sprognævn
www.dsn.dk
3374 7400

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