

Det Centrale Ordregister (COR): Hvad, hvor, hvorfor og hvordan?

Thomas Widmann
Dansk Sprognævn

Hvad er COR?

Det Centrale Ordregister (COR) er en standardiseret og centraliseret database over alle lemmaer og ordformer på dansk, med tilhørende unikke og permanente ID-numre. COR er et samarbejde mellem Dansk Sprognævn, Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi (Københavns Universitet), og projektet er støttet af Digitaliseringsstyrelsen. Opdelingen i lemmaer i COR's grundregister er baseret på Retskrivningsordbogen fra Dansk Sprognævn, og den første version af denne grundlæggende del af COR blev offentliggjort den 30. september 2022.

Hvordan er COR designet?

Fælles for alle COR-resurser er et lemmanummer. De enkelte resurser kan have deres egen struktur, altså egne underopdelinger og datafelter.

Som et eksempel kan vi se på grundregisteret. Her består en indgang af præfikset/domænet 'COR' efterfulgt af et 5-cifret indeksnummer der angiver det specifikke lemma.

Fx COR.29645 samsisk/samsk

Derefter kommer der et 3-cifret nummer som angiver ordklasse og bøjning for ordformen:

Fx COR.29645.302 samsiske/samske (302 er adj.sg.best)

Slutteligt har et COR-indeks i grundregisteret altid et 2-cifret løbenummer som angiver ortografisk variation.

Fx COR.29645.302.02 samske (samsiske er COR.29645.302.01)

Den 'rå' databaseindgang for *COR.29645.302* ser således ud:

```
COR.29645.302.01    samsisk adj.sg.best samsiske    1
COR.29645.302.02    samsisk adj.sg.best samske      1
```

Man ser her at grundregisteret tillige rummer et normeringsfelt ("1" sidst på linjen). Det er 0 for enkelte former (fx visse komparativer og superlativer) som ikke er blevet tjekket manuelt endnu og derfor ikke er omfattet af Retskrivningsordbogens status som normativ for det danske skriftsprog.

Hvad er et lemma?

Hvad angår definitionen af et lemma, følger vi Retskrivningsordbogens definition af opslagsord:

Opdelingen i opslagsord er principielt uafhængig af ordenes betydning. Det bevirker at ord med forskellig betydning er slået sammen i ét opslagsord hvis de i øvrigt har samme stavemåde, udtale, ordklasse og bøjning, og hvis de indgår i sammensætninger på samme måde.

Det betyder fx at fisken *skade* og fuglen *skade* har samme COR-id da forskellen mellem dem udelukkende er semantisk.

Tre niveauer

COR består af tre niveauer, hvilket giver mulighed for at inkludere data fra forskelligartede kilder.

- Niveau 1 svarer til Retskrivningsordbogen (grundregisteret, også kaldet COR1).
- Niveau 2 rummer andre resurser fra de professionelle sprogmiljøer i Danmark (specifikt dem som er medlem af Dansk Sprognævns repræsentantskab). Allerede nu ligger her bl.a. en del ekstra lemmaer fra DDO: **COR.EXT**.
- Niveau 3 vil rumme alle andre resurser, og her er der ingen begrænsninger – alle og enhver kan få allokeret et præfiks/domæne og et ID-interval ved at henvende sig til Dansk Sprognævn.

Relationer

I Det Centrale Ordregister (COR) er relationer en måde at forbinde lemmaer og ordformer på, så man kan indikere hvordan de er forbundet til andre. Relationer kan kun “pege indad”, således at Niveau 3-resurser kan pege på alle resurser, Niveau 2-resurser kun på sig selv, hinanden og grundregisteret, og grundregisteret på Niveau 1 kun på sig selv; det hviler således i sig selv og refererer aldrig til andre resurser.

Disse relationer gør det nemmere at organisere og søge efter data i COR da man kan følge relationerne mellem lemmaer og ordformer for at finde relaterede sproglige elementer.

Derudover er relationerne også nyttige for udviklere af sprogteknologiske værktøjer; de kan fx bruge relationerne til at forstå hvordan lemmaer og ordformer er forbundne, og de kan dermed forbedre deres værktøjs funktionalitet.

Der er mange slags mulige relationer, fx:

forkortelse	betydning
fus	fusion af to eller flere COR-indekser
rep	erstatning for et eller flere COR-indekser
spl	opsplitning i to eller flere COR-indekser
sms	sammensætning af to COR-indekser (til sammensatte ord)
hyr	hyperonym (overbegreb) til to eller flere COR-indekser
hyp	hyponym (underbegreb) til et andet COR-indeks
rim	rim som involverer et eller flere COR-indekser (til rimordbøger)

Ovenstående er blot eksempler. Hver resurse kan definere sine egne relationer, og grundregisteret bruger p.t. ingen relationer.

Hvordan kan man få adgang til COR?

COR's grundregister og enkelte andre resurser kan tilgås på to måder:

1. Det hele kan downloades som en CSV-fil fra *ordregister.dk*. Dette giver mulighed for at arbejde med bl.a. grundregisteret offline og integrere det i egne systemer.
2. Der er også en onlinegrænseflade der kan bruges til at søge i grundregisterets data og tilgå information om lemmaer og ordformer med. Disse informationer kan enten vises i HTML eller tilgås fra et program i CSV- eller JSON-format. Denne grænseflade findes på samme adresse: *ordregister.dk*.

Andre resurser vil kun have lister over definerede lemmaer og former liggende på *ordregister.dk*. Der vil så være et link til den URL som resursen kan hentes fra (hvis den er offentligt tilgængelig).

COR og Retskrivningsordbogen

Hvordan afviger Det Centrale Ordregister fra Retskrivningsordbogen?

Det Centrale Ordregisters grundregister er baseret på Retskrivningsordbogen, men der er nogle væsentlige forskelle mellem de to resurser:

- COR er lavet så et computerprogram nemt kan bruge det; RO er designet til mennesker.
- COR rummer de almindeligt brugte fuldformer for alle lemmaer; RO har en del huller i dækningen (og de vises ikke).
- COR kan bruges uden nogen restriktioner; RO må ikke bruges til at lave ordbøger med.
- COR har ingen eksempler; det har RO.
- COR har ingen regelhenvisninger; det har RO.
- COR har færre og kortere glosser end RO.

Opdatering til nyeste retskrivning

Det Centrale Ordregister (COR) er baseret på Retskrivningsordbogen (RO) og afspejler derfor den nyeste retskrivning. Når der sker ændringer i retskrivningen, vil COR også blive opdateret for at reflektere disse ændringer.

Helt konkret vil der ske flg. når den næste udgave af Retskrivningsordbogen offentliggøres:

1. Den hidtidige udgave omdøbes (sandsynligvis til *COR.RO2012*) og vil fortsat kunne tilgås på *ordregister.dk*.
2. Ord som udgår af Retskrivningsordbogen, vil fortsat beholde deres unikke ID. Hvis vi fx hypotetisk set forestiller os at *COR.62542 prokrustesseng* udgår i den næste udgave, vil *COR.62542* stadigvæk være tilknyttet lemmaet *prokrustesseng*, men lemmaet vil ikke have nogen former i den gældende retskrivning, og der vil være en note om at lemmaet er udgået. I *COR.RO2012* vil *prokrustesseng* stadig være defineret (altså som *COR.RO2012.62542*).
3. Ord som ændrer stavemåde i den nye retskrivning, vil ikke få noget nyt ID-nummer. Hvis man igen rent hypotetisk forestiller sig at *Odense* mister sit *d* (som jo er stumt på fynske dialekter), vil *COR.05162* stadig være det relevante COR-ID – det vil blot rumme stave-

måden *Oense*, og *Odense* vil nu være forvist til *COR.RO2012.05162*. Det samme gør sig gældende hvis der indføres ortografiske alternativer uden betydningsforskel (altså hvis både *Odense* og *Oense* tillades).

4. Nye ord vil få ID-numre i intervallet fra 0 til 99999, og disse vil naturligvis ikke defineres i *COR.RO2012*.
5. Hvis et lemma splittes op, vil de nye lemmaer få nye ID-numre, og vi laver relationer som viser sammenhængen mellem det gamle og de nye numre. Se eksemplet med *fjeder/fjer* i afsnittet om *RO^{hist}* nedenfor.
6. Hvis to lemmaer slås sammen, vil det fusionerede lemma også få et nyt ID-nummer, og igen vil vi lave en relation som beskriver sammenhængen.

Deling af egne resurser

Oprettelse af resursen

For at oprette sin egen COR-resurse skal man kontakte Dansk Sprognævn, som forvalter uddeelingen af nummerserier, og oplyse følgende:

- Et navneforslag (fx *Ordbog over dansk udtale i 1922*).
- Et forslag til et COR-domænenavn (fx *COR.UDT1922*).
- En oversigt over hvilke oplysninger man vil angive om hvert enkelt COR-indeks (fx glosse, udtale i Dania, udtale i IPA).
- Antal indeksnumre man regner med at få brug for (fx 10.000 – man skal ikke medregne de id-numre fra andre resurser som man regner med at genbruge).
- Licensbetingelser.
- Ejerforhold.
- Kontaktperson.

Når resursen er klar, bedes man sende en liste over de lemmaer som er defineret i den (men ikke indholdet af datafelterne), til Dansk Sprognævn. Det vil tillade os for et givet COR-ID at oplyse hvilke resurser der har data for dette lemma.

Tildeling af COR-ID-numre til resursen

For hvert lemma man har, er der basalt set fire scenarier:

1. Perfekt match – brug det eksisterende ID: *COR.OPEN.MUDS.0003919 skål!*

2. Blot ortografisk variation – brug eksisterende ID (evt. med underinddeling):
COR.OPEN.MUDS.0003919.02 skål!
3. Intet match – brug nyt ID fra egen nummerserie: *COR.OPEN.MUDS.2220002 cheers!*
4. Dårligt match – brug nyt ID med relationsopmærkning:
COR.OPEN.MUDS.2220022 whiskey (irsk); i grundregisteret er der kun *COR.94863 whisky*, så lav en relation mellem dem i *COR.OPEN.MUDS*-resursen.

Hvis man er i tvivl, er man naturligvis velkommen til at kontakte Dansk Sprognævn.

Eksempler på anvendelse af COR

Opmærkning af tekst

I COR's grundregister skelnes der som sagt mellem enslydende ordformer med forskellig grammatisk funktion. Fx kan formen *skål* svare til fire forskellige COR-ID-numre:

- ”“Her er en *skål* med øl i”: Det drejer sig om *COR.82511.110.01*, sb.fk.sg.ubest af **skål** (*beholder*).
- ”Lad os udbringe en *skål*”: Her er det *COR.55947.110.01*, sb.fk.sg.ubest af **skål** (*det at skåle*).
- ”Kom nu, *skål* med mig!”: I denne sætning er det *COR.32888.209.01*, vb.imp af **skåle**.
- ”*Skål!*”: Sidst, men ikke mindst, har vi her *COR.03919.870.01*, udråbsordet **skål**.

Hvis man – manuelt eller automatisk – tilføjer links til de korrekte COR-numre til hvert ord i en løbende tekst, kaldes det COR-linking. Se Henrichsen (dette bind) for mere information.

Opmærkning af ordbog

Man kan opmærke en ordbog med COR-ID-numre på mange måder. Et typisk eksempel kan fx se således ud:

```
<entry>
  <orth COR="55947">skål</orth>
  <pos>sb</pos>
  <tran>toast</tran>
  <example>De udbragte en <hw/> for MUDS.</example>
  <tran>They toasted the MUDS conference.</tran>
</entry>
```

Det vil gøre det muligt at tjekke at man staver opslagsordene i overensstemmelse med nyeste retskrivning, og man kan tilføje bøjningsoplysninger automatisk. Man vil også kunne opdatere ordbogen automatisk når den næste udgave af Retskrivningsordbogen udkommer.

Hvis nogen på et tidspunkt publicerer en COR-opmærket udtaleordbog, vil udtale-angivelser herfra også automatisk kunne tilføjes til ens ordbog (hvis man i øvrigt har rettighederne hertil).

RO^{hist}

Dansk Sprognævns RO^{hist}-projekt (rohist.dk) er en søgemaskine (baseret på det svenske SAOLhist-system) hvormed man kan sammenligne retskrivningerne fra 1872 til 2012:

Tidslinje

Seneste opslagsform	DHO 1872	SRO 1892	SRO 1918	DRO 1923	DRO 1946	RO 1955	RO 1986	RO 1996	RO 2001	RO 2012
fråde, sb.	Fraade		Fraade			fråde ¹	1. fråde	1. fråde	1. fråde	1. fråde
fråde, vb.	fraade		fraade			fråde ²	2. fråde	2. fråde	2. fråde	2. fråde
Antal forekomster	DHO 1872: 2	SRO 1892: 0	SRO 1918: 2	DRO 1923: 0	DRO 1946: 0	RO 1955: 2	RO 1986: 2	RO 1996: 2	RO 2001: 2	RO 2012: 2

Et eksempel på brug af RO^{hist}.

Den langsigtede plan er at udvide RO^{hist} med alle historiske retskrivningsordbøger og andre ortografiske ”rettesnore”. Fx er vi for tiden ved at omdanne Ove Mallings *Store og gode Handlinger af Danske, Norske og Holstenere* fra 1777 til en ordbog som kan tilføjes til RO^{hist} (se Hartling og Widmann 2020).

Vi vil også bruge den information som ligger i RO^{hist}, til at tildele COR-ID-numre til de historiske ordbøger. Dette vil tillade os at simplificere og forbedre implementeringen, og det vil gøre det muligt at lægge de historiske ordbøger ind som COR-resurser på ordregister.dk.

Det vil foregå således at det samme ID-nummer bruges hvis lemmaet er det samme – også hvis stavemåden har ændret sig. Et ord som *fråse* (RO2012) vil altså få samme COR-nummer som *frådse* i RO1996 og som *fraadse* i Dansk Haandordbog (1872):

ID-nummer	lemma
<i>COR.RO2001.59309</i>	fråse
<i>COR.RO1996.59309</i>	frådse
<i>COR.DHO1872.59309</i>	fraadse

Hvis man søger efter *fråse* i den fremtidige RO^{hist}, vil man basalt set blot skulle finde COR-nummeret (her 59309) og derefter slå dette op i de andre ordbøger.

Det er ikke nogen forhindring hvis et ord ikke er med i RO2012 da hver ordbog vil få tildelt sit eget nummerområde til ekstra lemmaer. Ordet *backfisch* var fx sidste gang med i 2001; hvis vi forestiller os at RO2001's ekstra nummerområde er 4002000-4004999, kan man så ved COR'ificeringen af denne ordbog tildele dette lemma ID-nummeret *COR.RO2001.4002123*.

Andre ordbøger kan så genbruge dette – *Backfisch* i DHO1872 vil altså kunne hedde *COR.DHO1872.4002123*. Nummeret viser at det oprindeligt blev defineret af RO2001 ved at ligge i intervallet 4002000–4004999.

Det vil være muligt at indikere sammenfald eller opdeling af stavemåder ved brug af relationer. Fx var der i DRO1923 fri variation mellem formerne *Fjeder* og *Fjer* uafhængigt af betydningen, hvorimod vi i dag har to separate lemmaer:

ID-nummer	lemma	glosse	relation
<i>COR.RO2012.79703</i>	fjeder		
<i>COR.RO2012.83551</i>	fjer		
<i>COR.DRO1923.4008020</i>	Fjeder/Fjer		fus:79703+83551

Det samme gør sig gældende hvis to lemmaer i historiske ordbøger svar til ét i dag. Fx skelnedes der i RO1955 mellem **skade** (*en fisk*) og **skade** (*en fugl*), mens de i dag som sagt er slået sammen i ét opslagsord da de kun adskiller sig semantisk:

ID-nummer	lemma	glosse	relation
<i>COR.RO2012.03094.110.01</i>	skade	en fugl; en fisk	
<i>COR.RO1955.4011080.110.01</i>	skade	en fugl	spl:03094
<i>COR.RO1955.4011081.110.01</i>	skade	en fisk	spl:03094

RO^{hist} vil altså også skulle tjekke relationsfeltet for at kunne præsentere fyldestgørende resultater.

Vi kan således opmærke alle de historiske ordbøger i rækkefølge, og vores liste over kendte lemmaer vil gradvist vokse i processen. Vi véd endnu ikke hvor meget af COR-opmærkingen af de historiske ordbøger vi kan foretage automatisk, men vi håber at de eksisterende links mellem ordbøgerne i den eksisterende RO^{hist}-database vil hjælpe os i processen.

Sådan laver man et lille sprogpil

COR er nemt at bruge fra et program da man ikke behøver at downloade noget som helst på forhånd. Se fx flg. Python-program:

```
from urllib.request import urlopen # hent data fra en URL
from urllib.parse import quote     # formatér URL-parametre
import json                        # håndtér JSON-data
from random import randint         # generér tilfældige tal

def beregnsvar(ord, gaet):
    """
    Funktionen tager to parametre: ordet der skal gættes, og det
    gættede ord. Den returnerer et svar hvor bogstaverne i det
    gættede ord er enten store (hvis de er korrekte på korrekt plads),
    små (hvis de er korrekte på forkert plads) eller bindestreger
    (hvis de er forkerte).
    """
    svar = ""
    for pos in range(len(gaet)):
        if len(ord) > pos and gaet[pos] == ord[pos]:
            svar = svar + gaet[pos].upper() # korrekt bogstav på korrekt plads
        elif ord.find(gaet[pos]) != -1:
            svar = svar + gaet[pos].lower() # korrekt bogstav på forkert plads
        else:
            svar = svar + "-" # forkert bogstav
    return svar

def tjeksvar(ord, gaet):
    """
    Funktionen tager to parametre: ordet der skal gættes, og brugerens gæt.
    Den tjekker om det gættede ord findes i ordregisteret, og hvis det gør
    det, returnerer den svaret fra beregnsvar-funktionen.
    """
    url = "https://ordregister.dk/fuldform/COR/json/" + quote(gaet) # lav URL
    data = json.loads(urlopen(url).read())
    status = data['status']
```

```

    if status == 'ok':
        return beregnsvar(ord, gaet)
    else:
        return "Det ord findes ikke!"

# Vælg et tilfældigt lemma fra COR med en længde på 5 bogstaver
found = False
while not found:
    try: # hvis noget går galt, havner vi i except og begynder forfra
        r = randint(1, 99999)
        url = "https://ordregister.dk/id/COR." + str(r) + ".json"
        data = json.loads(urlopen(url).read()) # går galt hvis COR.r ikke
        findes
        ord = data['lemma']
        if len(ord) == 5:
            found = True # COR-ID findes, og lemmaet er på fem bogstaver -
            hurra!
    except:
        continue

# Beder brugeren om at gætte på ordet og tjekker om det er korrekt.
# Prøver igen hvis det er forkert.
while True:
    gaet = input("Jeg tænker på et ord på fem bogstaver ... Hvad er det? ")

    if gaet == ord:
        break
    else:
        print(tjekksvar(ord, gaet))

input("Tillykke, det var rigtigt!")

```

Programmet laver simple JSON-kald til *ordregister.dk* for at tilgå lemmaer og fuldformer. Man kan sagtens forbedre programmet – fx fungerer *beregnsvar*-funktionen ikke på samme måde som i Wordle, og der tjekkes ikke om det fundne lemma indholder ikke-alfabetiske tegn (såsom bindestreger) – men det overlades her som en øvelse til læseren.

Konklusion

Vi har i denne artikel vist hvad Det Centrale Ordregister er, hvor man kan tilgå det henne, hvorfor man bør begynde at anvende det, og hvordan man rent praktisk gør det. Vi har også set eksempler på konkrete anvendelser af COR, inklusive et spil skrevet i Python. Forhåbningen er at læseren nu føler sig inspireret til at begynde at bruge COR – og gerne på måder som vi slet ikke har forestillet os!

Litteratur

- DHO1872 = Grundtvig, Sven (1872) *Dansk Haandordbog med den af Kultusministeriet anbefalede Retskrivning*. 2. udgave. København: C. A. Reitzel.
- DRO1923 = Glahder, Jørgen (1923) *Dansk Retskrivningsordbog*. Udgivet af Undervisningsministeriets Retskrivningsudvalg. København: Gyldendal.

- Dideriksen, Christina, Peter Juel Henriksen og Thomas Widmann (2022) Det Centrale Ordregister. I Nyt fra Sprognævnet 2022/3.
- Hartling, Anna Sofie & Thomas Widmann (2020) Den første ortografiske rettesnor for dansk – fra læsebog til ordbog: Malling (1777) på <http://rohist.dsn.dk>. I Goldshtein, Yonatan, Inger Schoonderbeek Hansen & Tina Thode Hougaard (red.). 18. Møde om Udforskningen af Dansk Sprog. Aarhus: NORDISK, Institut for Kommunikation og Kultur, Aarhus Universitet:213-230.
- Henriksen, Peter Juel (2023a) Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi. Nordisk Forening for Leksikografi. (Endnu ikke udkommet).
- Henriksen, Peter Juel (2023b) Diktatoriske befølelser. I denne konferencerapport.
- RO1955 = *Retskrivningsordbog*. Dansk Sprognævn (1955) København: Gyldendalske Boghandel.
- RO1996 = *Retskrivningsordbogen. 2. udgave*. Dansk Sprognævn (1996) København: Aschehoug.
- RO2001 = *Retskrivningsordbogen. 3. udgave*. Dansk Sprognævn (2001) København: Alinea – Aschehoug Dansk Forlag.
- RO2012 = *Retskrivningsordbogen. 4. udgave*. Dansk Sprognævn (2012) København: Alinea.
- Widmann, Thomas (2023) Det Centrale Ordregister og dets leksikografiske anvendelser. Nordisk Forening for Leksikografi. *Endnu ikke udkommet*.