NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Det Centrale Ordregister og dets leksikografiske anvendelser
Forfatter:	Thomas Widmann
Kilde:	Nordiska studier i lexikografi 16, 2023, s. 415–430
URL:	https://tidsskrift.dk/nsil/issue/archive

Det Centrale Ordregister og dets leksikografiske anvendelser

Thomas Widmann

In this article, we introduce the new Danish language resource called *Det Centrale Ordregister* 'the Central Word Registry' (abbreviated *COR*). It assigns unique and permanent ID numbers to lemmas and word forms in Danish. It consists of three levels; the first one corresponds to the official Danish orthographical dictionary (*Retskrivningsordbogen*), published by the Danish Language Council. The second level will contain resources produced by various professional linguistic organisations, and the third one will be open to everybody who is interested.

We start out by looking at the structure of the COR in some detail. Following on from this, we examine how various historical orthographical dictionaries can be encoded in the COR, and how this can be used to improve the implementation of the Danish Language Council's historical dictionary comparison site, RO^{hist}.

Finally we explore other lexicographic uses of the COR, and we examine how this could be expanded to create links to closely related languages, in particular Norwegian.

NØGLEORD: sprogteknologi, leksikografisk database, retskrivning, sproghistorie

1. Introduktion

Der findes mange elektroniske resurser for dansk – ordbøger i maskinlæsbare formater, korpusser, taggere, etc. – men de kan være svære at genbruge da de ikke er baseret på de samme grundresurser og ikke deler databasenøgler og tilsvarende, og det gør det sværere end nødvendigt at lave sprogteknologi på dansk. Vi har derfor lavet en ny resurse som forsøger at løse dette problem: Det Centrale Ordregister.

Det Centrale Ordregister (COR) tildeler unikke id-numre til alle lemmaer og ordformer på dansk. Det er et projekt under Digitaliseringsstyrelsen, med deltagelse af Dansk Sprognævn, Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi ved Københavns Universitet. I Dansk Sprognævn er vi ansvarlige for grundregisteret: ortografi og morfologi for det ordforråd som dækkes af Retskrivningsordbogen. Dette grundregister blev lanceret i september 2022 og er tilgængeligt på ordregister.dk

Vi vil i denne artikel først præsentere COR's opbygning, beskrive grundresursens struktur og demonstrere hvordan nye COR-resurser kan tilføjes. Derefter vil vi se på forskellige leksikografiske anvendelser med særligt fokus på RO^{hist}, Dansk Sprognævns hjemmeside som tillader sammenligning af forskellige historiske retskrivningsordbøger. Vi vil dernæst diskutere COR-linkere (programmer der automatisk tildeler COR-id-numre til alle ord i en løbende tekst), og til sidst vil vi som perspektivering se på hvordan man kan forestille sig at det danske COR kan sammenknyttes med et hypotetisk norsk parallelprojekt. Vi håber herved at give læseren både lyst til at begynde at bruge COR og de praktiske færdigheder til at gøre det.

2. Motivation

I Danmark har vi Det Centrale Personregister (CPR) som tildeler CPR-numre eller personnumre (svarende til *fødselsnummer* i Norge, *personnummer* i Sverige, *henkilötunnukset/personbeteckningar* i Finland og *kennitölur* i Island). Det centrale element er disse unikke numre som alle indbyggere har. Det er smart fordi det er en nøgle som tillader forskellige databaser at tale sammen.

Uden en sådan nøgle er databasesammenkøring en besværlig og tidskrævende proces da man typisk skal matche personerne på navn, fødselsdato og adresse; ingen af dem er unikke, og både navn og adresse kan ændre sig over tid.

På det sproglige område har en tilsvarende offentlig databasenøgle manglet. Mange leksikografiske og datalingvistiske projekter har måske tilknyttet unikke id-numre til deres lemmaer, men hvis disse ikke deles af andre projekter, vil det fortsat være besværligt og tidskrævende at køre forskellige databaser sammen så de kan bruges i andre projekter.

Hvis man fx har to ordbøger, en med udtaleangivelser og en anden med betydningsangivelser, kan man ikke umiddelbart flette dem sammen automatisk. Den største del af et sådant projekt er ikke vanskeligt – de fleste lemmaer har kun én udtale og ét bøjningsparadigme – men en automatisk sammenfletning kommer typisk til kort når den møder homografer og homonymer; på dansk fx ord som *kost* (/kɔsd/ "fejeredskab" eller /kʌsd/ "føde"), *tag* ("øverste del af et hus" eller det engelske låneord) eller *frø* (*en* frø "padde" eller *et* frø "plantedel"). Det gør det besværligt og tidskrævende at genbruge sproglige resurser.

Vi prøver derfor med COR-projektet at støtte sprogteknologi på dansk ved at lancere de manglende offentlige databasenøgler i håbet om at mange projekter vil begynde at bruge dem.

Der er paralleller mellem COR og den norske *Metaordboka* (se fx Grønvik & Ore 2018 og Ore et al. 2023); den primære forskel er nok at hovedmotivationen for COR er at dele data og lave resurser som kan arbejde sammen, hvorimod *Metaordboka* i udgangspunktet er et internt værktøj.

3. Hvordan fungerer det?

COR's opdeling af grundordforrådet i lemmaer er baseret på *Retskrivningsordbogen* fra Dansk Sprognævn. Derfor følger det samme princip for hvad et lemma er (RO2012:13f):

> 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.

Et COR-nummer vil derfor svare til et lemma (med ordklasse), og alle almindelige bøjningsformer anføres derunder.

Vi kender ikke til noget projekt, hverken i Danmark, Norden eller resten af verden, som er fuldstændigt sammenligneligt med COR. Der findes naturligvis talrige leksikalske databaser med morfologisk information, men det særlige ved COR er den åbne og fleksible struktur med en fælles nøgle som muliggør at man tilføjer alskens resurser (herunder historiske ortografiske ordbøger), sammen med de såkaldte *relationer* (se nedenfor) som tillader at disse mangfoldige resurser peger på hinanden.

COR's id-numre er principielt arbitrære. Grundresursens id-numre ligger mellem 0 og 99.999, og de er ikke tildelt alfabetisk. Af praktiske årsager har vi opdelt dette interval efter ordklasse, fx ligger adjektiverne mellem 15.000 og 29.999, og substantiverne mellem 40.000 og 99.999, men dette er ikke et formelt krav, og andre COR-resurser forventes ikke at følge dette mønster. Som et eksempel på id-numrenes arbitraritet er her de første 15 adjektiver: COR.15006 *dansk*, COR.15021 *travl*, COR.15026 *lille*, COR.15027 *vidunderlig*, COR.15049 *smart*, COR.15052 *alvorlig*, COR.15053 *lækker*, COR.15064 *ny*, COR.15066 *flest*, COR.15067

yderlig, COR.15073 *politisk*, COR.15075 *gylden*, COR.15081 *dyr*, COR.15082 *nær*, COR.15083 *kongelig*.

4. COR's struktur: eksempler

Lad os nu se på et konkret eksempel: *bark*. Det kan på dansk betyde både "det yderste lag på et træ" og "et skib", men det kan kun staves på denne måde, ordklassen kan kun være et substantiv, kønnet er fælleskøn, og udtalen kan kun være /ba:g/. Det er derfor tæt på kun at kræve ét COR-id (da opdelingen i opslagsord som sagt er uafhængig af ordenes betydning). Men der er to bøjningsmønstre: ét med pluralis og ét uden. Det betragtes derfor som to lemmaer, og de tildeles to COR-id-numre, COR.69850 til lemmaet uden pluralis, og COR.36198 til det med.

Da grundformen staves ens ("bark"), tildeler vi også en disambiguerende glosse til de to indgange. Vi har således:

- COR.58005 *sb* (*yderste lag på et træ*)
- COR.59594 *sb* (*sejlskib*)

Til bøjningsformerne bruges der to udvidelser: grammatisk kode og ortografisk variation. Til et substantiv af fælleskøn i singularis ubestemt er den grammatiske kode 110, i singularis bestemt er den 111, og så videre. Ordklassen, som altid er det første element i den grammatiske kode, er helt baseret på *Retskrivningsbogen*. I tabel 1 nedenfor er den ortografiske variation altid 01:

TABEL 1. lemmaet 'bark' i COR.

COR-id	lemma	Glosse	gram. funk.	form
COR.58005.110.01	bark	yderste lag på et træ	sb.fk.sg.ubest	bark
COR.58005.111.01	bark	yderste lag på et træ	sb.fk.sg.best	barken
COR.58005.114.01	bark	yderste lag på et træ	sb.fk.sg.ubest.gen	barks
COR.58005.115.01	bark	yderste lag på et træ	sb.fk.sg.best.gen	barkens
COR.59594.110.01	bark	sejlskib	sb.fk.sg.ubest	bark
COR.59594.111.01	bark	sejlskib	sb.fk.sg.best	barken
COR.59594.112.01	bark	sejlskib	sb.fk.pl.ubest	barker
COR.59594.113.01	bark	sejlskib	sb.fk.pl.best	barkerne
COR.59594.114.01	bark	sejlskib	sb.fk.sg.ubest.gen	barks
COR.59594.115.01	bark	sejlskib	sb.fk.sg.best.gen	barkens
COR.59594.116.01	bark	sejlskib	sb.fk.pl.ubest.gen	barkers
COR.59594.117.01	bark	sejlskib	sb.fk.pl.best.gen	barkernes

Som et eksempel på den ortografiske variation kan vi se på COR-lemmaet *coronavirus*; det kan i ubestemt singularis staves både *coronavirus* og *koronavirus*; de tildeles hhv. 01 og 02 (se tabel 2).

TABEL 2. lemmaet 'coronavirus' i ubestemt singularis.

COR-id	lemma	glosse	gram. funk.	form
COR.53473.110.01	coronavirus		sb.fk.sg.ubest	coronavirus
COR.53473.110.02	coronavirus		sb.fk.sg.ubest	koronavirus

Hver af disse to stavemåder har tre forskellige flertalsformer, se tabel 3 nedenfor:

TABEL 3. lemmaet 'coronavirus' i ubestemt pluralis.

COR-id	lemma	glosse	gram. funk.	Form
COR.53473.112.01	coronavirus		sb.fk.pl.ubest	coronavira
COR.53473.112.02	coronavirus		sb.fk.pl.ubest	coronavirus
COR.53473.112.03	coronavirus		sb.fk.pl.ubest	coronavirusser
COR.53473.112.04	coronavirus		sb.fk.pl.ubest	koronavira
COR.53473.112.05	coronavirus		sb.fk.pl.ubest	koronavirus
COR.53473.112.06	coronavirus		sb.fk.pl.ubest	koronavirusser

Vi forbinder altså ikke de enkelte ortografiske variationer til hinanden, men nummererer dem sekventielt for hver grammatisk funktion.

Hvis man henter COR, vil man også lægge mærke til et normeringsfelt (som ikke er vist i eksemplerne i denne artikel): Det er normalt 1 (= *normeret*), men er nogle gange 0 (= *ikke normeret*); det drejer sig om enkelte bøjningsformer som er blevet autogenereret og derfor ikke bør bruges til fx stavekontrol.

5. Fra form til lemma

Informationerne i COR kan præsenteres på mange måder. På ordregister.dk kan man søge på COR-id, lemma og fuldform. Hvis man bruger den sidste mulighed, kan COR bruges til at finde ud af hvilke lemmaer og grammatiske funktioner en ordform kan referere til. Som et eksempel kan vi søge på ordformen *los*; resultatet heraf kan ses i tabel 4.

TABEL 4. ordformen 'los' i COR.

COR-id	lemma	glosse	gram. funk.	form
COR.22838.300.01	los		adj.sg.ubest.fk	los
COR.37712.209.01	losse		vb.imp	los
COR.55736.110.01	los	et dyr	sb.fk.sg.ubest	los
COR.70821.120.01	los	et spark	sb.itk.sg.ubest	los
COR.70821.122.01	los	et spark	sb.itk.pl.ubest	los
COR.78188.114.01	lo		sb.fk.sg.ubest.gen	los

6. Tre niveauer

Det Centrale Ordregister er opdelt i tre niveauer. Niveau 1 er det samme som grundregisteret og rummer altså de samme lemmaer som *Retskrivningsordbogen*.

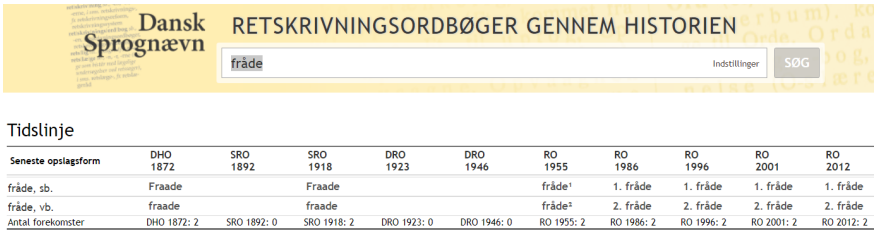
Niveau 2 rummer andre resurser fra de professionelle sprogmiljøer i Danmark (specifik dem som er medlem af Dansk Sprognævns repræsentantskab), og her vil mange andre nyttige resurser tilføjes med tiden. Allerede nu ligger her bl.a. en del ekstra lemmaer fra Den Danske Ordbog (udgivet af Det Danske Sprog- og Litteraturselskab [DSL]); denne resurse kaldes *COR.EXT*. Niveau 2 vil også komme til at rumme en semantisk udvidelse til grundregisteret produceret af DSL og CST (Center for Sprogteknologi ved Københavns Universitet). Se fx Nimb et al. (2022) for mere information om deres arbejde med at udvikle denne semantiske komponent.

Niveau 3 vil rumme alle andre resurser, og her er der ingen begrænsninger – ethvert relevant projekt kan få allokeret et præfiks og et id-interval hvis man henvender sig til Dansk Sprognævn.

Resurser kan relateres til andre resurser på samme eller lavere niveau. Alle resurser kan derfor markere relationer til grundresursen, men grundresursen hviler i sig selv og refererer aldrig til andre.

7. RO^{hist}

Dansk Sprognævns RO^{hist}-projekt (rohist.dk) er en søgemaskine hvormed man kan sammenligne de danske retskrivningsordbøger fra 1872 til 2012:



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

FIGUR 1. Et eksempel på en søgning i RO^{hist}.

Det er en videreudvikling af det svenske SAOLhist-system (som tillader parallelle opslag i *Svenska Akademiens ordlista över svenska språ-*

ket (1874, 1889-), *Ordbok öfver svenska språket* af A. F. Dalin (1850–1853) og Svenska Akademiens *Svensk ordbok* (2009). Se Diderichsen et al. (2015) for flere detaljer om sammenhængen mellem SAOLhist og RO^{hist}.

Vi arbejder løbende på 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} (jf. Hartling & Widmann 2020).

Id-numrene i COR svarer til den nyeste udgave af *Retskrivningsordbogen* (p.t. 4. udgave fra 2012), men vi håber på at vi på sigt også kan tildele COR-numre til de historiske ordbøger i RO^{hist}. Disse ordbøger vil komme til at være niveau 2-resurser, og de vil derfor få deres eget præfiks og id-nummer-interval.

Man vil dog genbruge det samme id-nummer 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 (*Retskrivningsordbogen* fra 1996) og som *fraadse* i DHO1872 (*Dansk Haandordbog* fra 1872):

COR-id	lemma	gram. funk.	form
COR.37337.200.01	fråse	vb.inf.akt	fråse
COR.RO2001.37337.200.01	fråse	vb.inf.akt	fråse
COR.RO1996.37337.200.01	frådse	vb.inf.akt	frådse
COR.DHO1872.37337.200.01	fraadse	vb.inf.akt	fraadse

Dette vil simplificere implementeringen af RO^{hist} betydeligt. Hvis man søger efter *fråse* i den fremtidige RO^{hist}, vil man blot skulle finde COR-nummeret (her 37337) og så ved et simpelt opslag konstatere hvorvidt dette er defineret i de historiske ordbøger.

De eksisterende links mellem ordbøgerne i RO^{hist} vil danne basis for dette arbejde. Vi vil altså tage den relationelle database som ligger til grund for RO^{hist}, analysere data og herudfra tildele historiske COR-idnumre. Det betyder også at hvis man finder en fejl i RO^{hist} – fx at en historisk stavemåde er blevet knyttet til det forkerte lemma – vil man blot skulle rette COR-id-nummeret i den historiske ordbog hvor fejlen er.

8. Uoverensstemmelser

Retskrivningsordbogen, og dermed COR's grundregister, rummer kun grundordforrådet, og når man indkoder en historisk ordbog, vil man derfor nogle gange stå med et lemma som ikke findes i den ældre ordbog. Det er dog ikke noget problem da alle ordbøger (og alle andre resurser som tilføjes til COR) vil få tildelt deres egne nummerområder 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 *backfisch* id-nummeret `COR.RO2001.4002123`.

Andre ordbøger kan så genbruge dette – *Backfisch* i DRO1946 (*Dansk Retskrivningsordbog* fra 1946) vil altså kunne hedde `COR.DRO1946.4002123`. Nummeret viser at det oprindeligt blev defineret af RO2001 ved at ligge i intervallet 4002000–4004999.

Det er heller ikke noget problem hvis stavemåder er blevet slået sammen eller splittet op i forhold til grundresursen eller de andre historiske ordbøger. COR gør det på niveau 2 og 3 muligt at bruge et såkaldt relationsfelt til dette formål. Dette er et ekstra felt i COR som indeholder et link til en eller flere andre COR-indgange, samt en type. Da grundresursen aldrig peger på andre resurser, bruges relationsfeltet aldrig her, men andre resurser kan bruge det til at vise hvordan afvigelser skal forstås. Forskellige resurser kan definere deres egne relationstyper afhængigt af deres behov; i arbejdet med de historiske retskrivningsordbøger regner vi med at få brug for flg.:

forkortelse	betydning
fus	fusion af to eller flere COR-indekser
rep	erstattet af et eller flere COR-indekser
spl	splittet op i to eller flere COR-indekser
sms	sammensætning af (til sammensatte ord)

Eksempelvis var der i DRO1923 (*Dansk Retskrivningsordbog* fra 1923) fri variation mellem formerne *Fjeder* og *Fjer* uafhængigt af betydningen, hvorimod vi i dag har to separate lemmaer. Det løser vi ved at lave et nyt id-nummer (her 4008020) som rummer information om at det har en relation til de to moderne indgange; se tabel 5.

TABEL 5. 'fjer' og 'fjeder'.

COR-id	lemma	glosse	gram. funk.	form	relation
COR.70131.110.01	fjeder		sb.fk.sg.ubest	fjeder	
COR.70759.110.01	fjer		sb.fk.sg.ubest	fjer	
COR.DRO1923.4008020.110.01	Fjeder		sb.fk.sg.ubest	Fjeder	fus:70759+70131
COR.DRO1923.4008020.110.02	Fjeder		sb.fk.sg.ubest	Fjer	fus:70759+70131
COR.DRO1923.70759					rep:4008020
COR.DRO1923.70131					rep:4008020

Det samme gør sig gældende hvis to historiske lemmaer er smeltet sammen til ét moderne – som et eksempel ses i tabel 6 lemmaet/lemmaerne *skade*; bemærk at de to lemmaer i RO1955 har samme indhold i relationsfeltet.

TABEL 6. 'skade'.

COR-id	lemma	glosse	gram. funk.	form	relation
COR.45662.110.01	skade	en fugl; en fisk	sb.fk.sg.ubest	skade	
COR.RO1955.4011080.110.01	skade	en fugl	sb.fk.sg.ubest	skade	rep:45662
COR.RO1955.4011081.110.01	skade	en fisk	sb.fk.sg.ubest	skade	rep:45662
COR.RO1955.45662					spl:4011080+4011081

(Fuglen *skade* har det latinske navn *Pica pica*; fisken *Dipturus batis*.)

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

Planen er at opmærke alle de historiske ordbøger i omvendt kronologisk rækkefølge, altså i første omgang RO2001, og vores liste over kendte lemmaer vil derved gradvist vokse.

9. Næste udgave af *Retskrivningsordbogen*

Vi regner med at den næste udgave af *Retskrivningsordbogen* vil udkomme i 2024. I den forbindelse vil vi dels opdatere COR-grundserien så den fort-

sat afspejler nyeste retskrivning, dels publicere ændringerne mellem den gamle og den nye udgave i et maskinlæsbart format. Helt konkret vil den nuværende udgave få tildelt et nyt præfiks, COR.RO2012, og COR vil blive forbeholdt den nye udgave. Lad os rent hypotetisk forestille os at *sprog* ændres til *språg* i overensstemmelse med udtalen (men dette kommer helt sikkert ikke til at ske i virkeligheden!). COR-nummeret ændres ikke – det vil forsat være COR.40015.

Det er de konkrete former som vil blive ændret. De former vi har i dag, kan ses i tabel 7; efter ændringen vil de se ud som i tabel 8 nedenfor.

TABEL 7. lemmaet 'sprog' i COR.

COR-id	lemma	gram. funk.	form
COR.40015.120.01	sprog	sb.itk.sg.ubest	sprog
COR.40015.121.01	sprog	sb.itk.sg.best	sproget
COR.40015.122.01	sprog	sb.itk.pl.ubest	sprog
COR.40015.123.01	sprog	sb.itk.pl.best	sprogene
COR.40015.124.01	sprog	sb.itk.sg.ubest.gen	sprogs
COR.40015.125.01	sprog	sb.itk.sg.best.gen	sprogets
COR.40015.126.01	sprog	sb.itk.pl.ubest.gen	sprogs
COR.40015.127.01	sprog	sb.itk.pl.best.gen	sprogenes
COR.40015.129.01	sprog	sb.itk.sms	sprog-

TABEL 8. lemmaet 'sprog' i en hypotetisk fremtidig udgave af COR.

COR-id	lemma	gram. funk.	form
COR.40015.120.01	språg	sb.itk.sg.ubest	språg
COR.40015.121.01	språg	sb.itk.sg.best	språget
COR.40015.122.01	språg	sb.itk.pl.ubest	språg
COR.40015.123.01	språg	sb.itk.pl.best	språgene
COR.40015.124.01	språg	sb.itk.sg.ubest.gen	språgs
COR.40015.125.01	språg	sb.itk.sg.best.gen	språgets
COR.40015.126.01	språg	sb.itk.pl.ubest.gen	språgs
COR.40015.127.01	språg	sb.itk.pl.best.gen	språgenes
COR.40015.129.01	språg	sb.itk.sms	språg-

Og de gamle former vil nu få et RO2012-præfiks, så RO2012 dermed nærmest automatisk bliver tilføjet til de historiske ordbøger som kan tilgås i RO^{hist}, se tabel 9.

TABEL 9. lemmaet 'sprog' fra 2012-udgaven af Retskrivningsordbogen som det vil se ud når retskrivningen har ændret sig.

COR-id	lemma	gram. funk.	Form
COR.RO2012.40015.120.01	sprog	sb.itk.sg.ubest	sprog
COR.RO2012.40015.121.01	sprog	sb.itk.sg.best	sproget
COR.RO2012.40015.122.01	sprog	sb.itk.pl.ubest	sprog
COR.RO2012.40015.123.01	sprog	sb.itk.pl.best	sprogene
COR.RO2012.40015.124.01	sprog	sb.itk.sg.ubest.gen	sprogs
COR.RO2012.40015.125.01	sprog	sb.itk.sg.best.gen	sprogets
COR.RO2012.40015.126.01	sprog	sb.itk.pl.ubest.gen	sprogs
COR.RO2012.40015.127.01	sprog	sb.itk.pl.best.gen	sprogenes
COR.RO2012.40015.129.01	sprog	sb.itk.sms	sprog-

Denne information kan man så bruge til at generere en ændringsliste med; denne viser hvordan RO2012 forholder sig til denne hypotetiske fremtidige retskrivning i et format som kan bruges til fx at ændre en tekst med:

```
COR.RO2012.40015.120.01: sprog < språg
COR.RO2012.40015.121.01: sproget < språget
COR.RO2012.40015.122.01: sprog < språg
COR.RO2012.40015.123.01: sprogene < språgene
COR.RO2012.40015.124.01: sprogs < språgs
COR.RO2012.40015.125.01: sprogets < språgets
COR.RO2012.40015.126.01: sprogs < språgs
COR.RO2012.40015.127.01: sprogenes < språgenes
COR.RO2012.40015.129.01: sprog- < språg-
```

10. Opdatering af tekster og ordbøger til en anden udgave af retskrivningen

COR-opmærkning af historiske ordbøger har også mange andre leksikografiske anvendelser.

Man vil fx kunne ændre stavemåden af en COR-opmærket tekst fra en ortografi til en tidligere eller senere. Her er fx et fragment fra Mallings lærebog (1777):

… gandske forskiellige i Sprog, Sæder og Levemaade …

Hvis vi nu har en passende ordbog over Mallings sprog, gerne med fuldformer, kan vi nu tildele ordene `COR.MALLING`-id-numre. Disse kan vi så slå op i det moderne COR-register og derved opnå de moderne staveformer:

… ganske forskellige i sprog, sæder og levemåde …

I tabelform:

Malling	COR	RO2012
gandske	`COR.MALLING.10069.900`	ganske
forskiellige	`COR.MALLING.15189.303`	forskellige
i	`COR.MALLING.00852.880`	i
Sprog	`COR.MALLING.40015.122`	sprog
Sæder	`COR.MALLING.92121.112`	sæder
og	`COR.MALLING.00099.970`	og
Levemaade	`COR.MALLING.84179.110`	levemåde

Denne proces kan naturligvis bruges i begge retninger – man kunne lige så godt bruge den til at gøre en moderne tekst kunstigt gammeldags med.

En tilsvarende procedure kan anvendes på bl.a. ordbøger. Lad os eksempelvis antage at vi har en dansk-engelsk ordbog med COR-opmærkning af opslagsordene:

```
<entry>
  <orth COR="37337">frådse</orth>
  <pos>vb</pos>
  <tran>gorge</tran>
</entry>
```

Det vil nu være ganske let at slå op i COR og se at den gældende stavemåde er *fråse*. Man vil endda kunne gøre det næsten fuldautomatisk – der vil kun være behov for menneskelig assistance hvis en stavemåde splittes op i to, afhængigt af betydningen. Man kunne også bruge COR til at tilføje bøjningsoplysninger med.

11. COR-linkere

Til brug for korpuslingvistik og andre datalingvistiske anvendelser vil der også blive udviklet COR-linkere, dvs. programmer som tildeler det korrekte COR-id (med under-id, altså grammatisk kode) til alle ord i en løbende tekst. Når en tekst på denne måde er blevet COR-linket, vil det være trivielt at generere en ordklasseopmærket tekst (da alle de nødvendige informationer ligger i under-id'et), og hvis man har en COR-opmærket udtaleordbog, vil man også kunne tilknytte udtaleangivelser til alle ord (hvor også homografer får tilknyttet den korrekte udtale). Se Peter Juel Henrichsen (2023) for flere oplysninger om hvordan en sådan COR-linker kan se ud.

12. Norsk

Lad os til sidst til perspektivering se på muligheden for engang at forbinde dansk og norsk bokmål i værktøjer som RO^{hist}.

De to sprog deler som bekendt et ortografisk udgangspunkt, så når vi opmærker de historiske ordbøger med COR-id-numre, vil det blot kræve et norsk parallelprojekt for at kunne forbinde de to sprogs retskrivninger. Dette parallelprojekt kunne fx være en udvidelse af Metaordboka (Grønvik & Ore 2018; Ore et al. 2023).

Man kunne fx tage et dansk lemma som *sprog* og slå det op i COR-registeret; dets id-nummer er COR.40015. Vi kan så tage en resurse som er gammel nok til at den er en del af både dansk og norsk ortografihistorie, fx Ove Mallings læsebog *Store og gode Handlinger af Danske, Norske og Holstenere* (Malling 1777), som er ved at blive lavet om til en ordbog med COR-id'er (se Hartling og Widmann 2020), og tjekke at lemmaet også er defineret der: COR.MALLING.40015 *Sprog*. Hvis vi nu i fremtiden er så heldige at det hypotetiske norske parallelprojekt også har indekseret Malling-ordbogen, kunne det måske have id-nummeret `SOR.MALLING.123456` dér. Dette kan så bruges til at finde den moderne norske form med, her altså SOR.123456 *språk*.

Resultatet bliver ikke en tosproget ordbog, men forbindelser mellem de ord som har samme ortografiske ophav. Norsk *kveld* vil altså knyttes sammen med dansk *kvæld*, ikke med *aften*.

I teorien kunne man også lave forbindelser til andre sprog, men det kræver at nogen skaber disse links manuelt når man ikke kan gå direkte tilbage til sidste fælles retskrivning.

13. Konklusion

I denne artikel har vi givet en introduktion til Det Centrale Ordregister (COR) som tildeler unikke id-numre til alle lemmaer og ordformer på dansk, og vi har beskrevet grundregisterets struktur demonstreret hvordan nye resurser kan tilføjes. Vi har også set på forskellige leksikografiske anvendelser af COR, herunder RO^{hist}, som tillader sammenligning af forskellige historiske retskrivningsordbøger, og COR-linkere, som kan tildele det korrekte COR-id til alle ord i en løbende tekst. Endelig har vi diskuteret muligheden for at sammenknytte det danske COR med et hypotetisk norsk parallelprojekt.

Det Centrale Ordregister er frit tilgængeligt i dag på ordregister.dk. Vi håber at mange vil tilføje deres egne resurser til COR, så dansk dermed bliver et af de sprog som har de bedste resurser til sprogteknologiske og leksikografiske projekter.

Litteratur

DHO1872 = Grundtvig, Sven 1872. *Dansk Haandordbog med den af Kultusministeriet anbefalede Retskrivning*. 2. udgave. København: C. A. Reitzel.

Diderichsen, Philip, Anna Sofie Hartling, Anne Kjærgaard & Anna Kristiansen 2015. Retskrivningsordbøger gennem historien – følg retskrivningens udvikling ord for ord på http://rohist.dsn.dk. I Hansen, Inger Schoonderbeek & Tina Thode Hougaard (red.). *15. Møde om Udforskningen af Dansk Sprog*. Aarhus: Aarhus Universitet, 117-126.

DRO1923 = Glahder, Jørgen 1923. *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg*. København: Gyldendal.

DRO1946 = Glahder, Jørgen 1946. *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg. 5. Optryk*. København: Gyldendal.

Grønvik, O., & Ore, C.-E. S. 2018. Bokmål og nynorsk samindeksert – Metaordboka som verktøy for jamføring og utforsking av ordtilfang. *Nordiske studier i leksikografi 14*. Reykjavík, 87-95.

Hartling, Anna Sofie & Thomas Widmann 2020. Den første ortogra-

fiske 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.

Henrichsen, Peter Juel 2023. Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 113–126.

Malling, Ove 1777. Store og gode Handlinger af Danske, Norske og Holstenere.

Nimb, Sanni, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård 2022. COR-S – Den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica* 29, 73-95.

Ore, Christian-Emil Smith, Oddrun Grønvik & Trond Minde 2023. Et fullformsystem for analyse av eldre tekst på tidlig nynorsk, bygd på Aasen-normalen. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 267–279.

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.