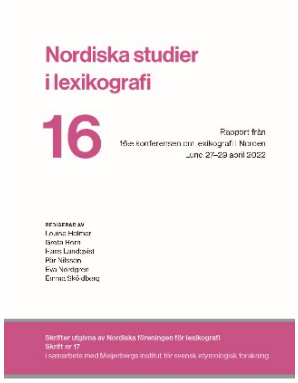
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi
Forfatter:	Peter Juel Henrichsen
Kilde:	Nordiska studier i lexikografi 16, 2023, s. 113–126
URL:	https://tidsskrift.dk/nsil/issue/archive



Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi

Peter Juel Henrichsen

Det Centrale Ordregister (COR) is a newly published Danish lexicographic register. Each Danish lemma and each Danish word form has (or can have) a unique COR index associated with it. By design, COR thus provides a stable and authoritative reference to the Danish vocabulary at large. COR includes (*i*) morphological information for 64,000 Danish lexemes (viz. all word forms covered by the official Danish orthographic norm), and (*ii*) a linking structure making a wide range of Danish linguistic resources inter-operable (corpora, dictionaries, term banks etc.). COR is especially intended for use in the language technological sectors. COR is published open-source (under licence CC0).

NØGLEORD: sprogteknologi, dansk sprog, automatisk tekstanalyse, leksikografi til NLP, CLINK

1. Det Centrale Ordregister – bedre input til sprogteknologerne

COR (det Centrale OrdRegister) er et register over det danske ordforråd særligt udviklet til brug i sprogteknologi. Hvert dansk lemma og hver dansk ordform har – eller kan få – tilknyttet et unikt COR-indeks.

Lemma

dansk	adj	`COR.15006`

Ordform

danske	adj.sg.best	`COR.15006.302.01`
danske	adj.plur	`COR.15006.303.01`

Tekst annoteret med COR-indekser er naturligt disambigueret for homografi og er dermed egnet som input til mange slags sprogteknologi som fx stave- og grammatikkontrol, tekstanalyse, talesyntese, maskinoversæt-

telse og natursprogsgrænseflader ('natural language interfaces'). Denne artikels hovedformål er at introducere COR som leksikologisk projekt, som leksikografisk grundressource og som værktøj for udvikleren af NLP ('natural language processing'). I artiklens sidste afsnit omtales projektets organisation og tidsplan.

Betegnelsen 'COR' er en parallel til 'CPR' (Det Centrale Personregister). Hver dansker får ved fødslen et CPR-nummer som følger personen hele livet. Med sit CPR-nummer har man let adgang til at søge i samfundets databaser – også baser der ikke er relateret indbyrdes – med information om adresse, sundhed, uddannelse, skatteforhold, og så videre. I samme ånd tildeler COR et unikt og uforanderligt indeks til hvert dansk ord.

Nu er det jo ikke fornuftigt – endsige muligt! – at leksikalisere det samlede danske ordforråd, alene fordi man ikke kan trække en klar grænse mellem etablerede leksemer og kortlivede neologismer. Ikke desto mindre *kan* ethvert dansk ord (lemma og ordform) få tilknyttet et COR-indeks. Vi har søgt at opnå denne 'praktiske fuldstændighed' ved at inddele Det Centrale Ordregister i tre niveauer, hvoraf niveau 1 dækker det centrale ordforråd og udgør det leksikalske grundlag som lemmaer og ordformer i de øvrige niveauer refererer til, mens COR's niveau 2 og 3 er åbne for tilgang af nye leksemer. Denne artikel fokuserer på COR's niveau 1 (herefter 'COR_1'), altså den leksikalske grundressource som hele Det Centrale Ordregister refererer til. Denne del af COR er udviklet af Dansk Sprognævn (dsn.dk) mens Det Danske Sprog- og Litteraturselskab (dsl.dk) og Center for Sprogteknologi (cst.dk) udvikler en ordsemantisk database i COR's niv. 2 (Pedersen et al. 2022). En indføring i hvordan eksisterende ordbøger gøres COR-kompatible gives i Widmann (2023) og mere forenklet i Dideriksen et al. (2022).

1.1. COR niveau 1 – det leksikalske fundament

COR_1 er en database med omkring 64.000 danske lemmaer (500.000+ ordformer) med oplysning om ortografi, bøjningsformer og sammensætningspotentiale. Den leksikalske dækning er identisk med Retskrivningsordbogens (Schack et al. 2012; se også Link1 1997), og oplysningerne i COR_1 er officielt ratificeret (Lov om Dansk Retskrivning, Link2 1997). COR_1 dækker det almindelige danske ordforråd forstået som de lemmaer der (*i*) benyttes alment (dvs. ikke er typiske for bestemte alders- eller sam-

fundsgrupper), (*ii*) forekommer stabilt over tid (efter redaktionens vurdering) og (*iii*) generelt ikke refererer til specifikke organisationer, sagsforhold, produkter og personer. COR_1 er – ifølge sine selektionskriterier, informationstyper og status som ortografisk norm – relevant for næsten alle typer af NLP-applikationer og er dermed det naturlige fundament for COR som helhed.

TABEL 1. Et udvalg af ordformer fra COR_1: bøjning, indeks, status i RO (se afsn. 1.1) og information om lemma.

Item	*Ordform*	*Bøjning*	*COR1-indeks*	*#RO*	*Lemma*
a.	*se*	vb.inf.akt	COR.30600.200.01	1	Leksem: {se}
b.	*ses*	vb.inf.pass	COR.30600.201.01	1	Klasse: verbum
c.	*ser*	vb.præs.akt	COR.30600.203.01	1	
d.	*så*	vb.præt.akt	COR.30600.206.01	1	*(udvalgte former)*
e.	*sås*	vb.præt.pass	COR.30600.207.01	1	
f.	*så*	vb.inf.akt	COR.30901.200.01	1	Leksem: {så}
g.	*sås*	vb.inf.pass	COR.30901.201.01	1	Klasse: verbum
h.	*sår*	vb.præs.akt	COR.30901.203.01	1	*(udvalgte former)*
i.	*så*	adv	COR.10147.900.01	1	Leksem: {så} Klasse: adverbium
j.	*så*	konj	COR.00364.970.01	1	Leksem: {så} Klasse: konjunktion
k.	*dansk*	adj.sg.ubest.fk	COR.15006.300.01	1	Leksem: {dansk}
l.	*dansk*	adj.sg.ubest.itk	COR.15006.301.01	1	Klasse: adjektiv
m.	*danske*	adj.sg.best	COR.15006.302.01	1	
n.	*danske*	adj.pl	COR.15006.303.01	1	*(udvalgte former)*
o.	*danskere*	adj.kompar	COR.15006.304.01	0	
p.	*danskest*	adj.superl.sg.ubest	COR.15006.305.01	0	
q.	*danskeste*	adj.superl.sg.best	COR.15006.306.01	0	
r.	*danskeste*	adj.superl.pl	COR.15006.307.01	0	
s.	*imedens*	konj	COR.00367.970.01	1	Leksem: {imedens}
t.	*imens*	konj	COR.00367.970.02	1	Klasse: konjunktion

Tabel 1 giver eksempler på COR_1-indeksering. Bemærk i række *d.*, *f.*, *i.* og *j.* at homografen *så* indekseres forskelligt efter sin klassifikation som vb.præt.akt, vb.inf.akt, adv og konj. Række *s.* og *t.* viser hvordan det tredje numeriske felt bruges til at adskille rent ortografiske varianter (`01` og `02`), altså ordformer der ikke afviger fra hinanden i betydning, bøjning, udtale osv. (i disse tilfælde vil der ofte være en ældre og en nyere stavemåde hvoraf den ene er på vej ud af retskrivningen; dette vil dog ikke påvirke det bagvedliggende lemmas COR-indeks). I kolonnen #RO vises ordformernes RO-status: 1 betyder at staveformen er eksplicit normeret i Retskrivningsordbogen, 0 bruges til former der kun er indirekte normeret (afledt af generelle staveregler). De fleste 0-former forekommer sjældent i hverdagens tekstarter, men træffes nu og da (ofte spøgende, 'tongue in cheek'). Adjektivet *danskeste* (se tabel 1) har fx 13 forekomster i korpuset DAGW:DANAVIS svarende til 0,46 ppm (DANAVIS er en del af DAGW, et stort dansk referencekorpus; omtales nærmere i afsn. 3.5).

Ordformerne i COR_1 dækker typisk 96-99 % af teksterne (undtaget proprier, numeralier og tekniske symboler) i populære tekstgenrer såsom aviser, magasiner, wiki, studietekster på lavere niveau, skønlitteratur, altså de genrer som korpuslingvister typisk udvælger til *balanced text corpora*. COR_1 er gratis, frit tilgængeligt og åbent for enhver anvendelse. Det kan downloades som fuldformsliste fra https://ordregister.dk, hvor man også finder relevante manualer, programmer, oplysning om COR-kompatible orddatabaser osv.

I det følgende afsnit giver vi eksempler på hvordan COR-opmærkning kan føre til forbedret sprogteknologi ved at udnytte COR's grundlæggende princip om entydig leksikalsk reference. Afsnit 3 introducerer CLINK – en applikation som annoterer hvert token i en tekst med dets relevante COR-indeks. Artiklen slutter med en perspektivering og en opfordring. Dansk bør ikke være det eneste sprog med et centralt ordregister.

2. Nøgen tekst er dårligt input

Ortografien er – på trods af sin dominerende rolle som sproglig repræsentation i alle moderne samfund – på mange måder en upålidelig afbildning af et ordforråd. Dette gælder i alle sprog med eksempler på homo-

grafi, men a fortiori i dansk, som er berygtet for sine uigennemskuelige skrift-til-lyd-principper. Som alle nordboer ved, er det ret umuligt at forudse hvordan et skrevet dansk ord lyder eller hvordan et udtalt ord staves (medmindre man ved det i forvejen). Tag som eksempel tekstordet *for*, her lydskrevet i den såkaldte SAMPA-formalisme (Link3 1995).

Skriftform	Udtale	Brug (eksempel)	Ordklasse
for	[fC]	*for fanden*	*for*/præp
for	[f"C]	*for og imod*	*for*/adv
for	[f"O:]	*for og bag*	*for_og_bag*/flerordsudtryk
for	[f"oR?]	*hun for afsted*	*for*/vb
for	[f"o:?R]	*frakkens for er slidt*	*for*/sb

Det er altså ikke muligt for fx en talesyntese at realisere en (dansk) tekst som oplæst tale uden at tekstordene er blevet leksikalsk disambigueret. På samme måde vil en automatisk oversætter gå i knæ over et input fuld af homografer.

Maj så Find så alle frøene, så hun bestilte nye så snart de slap op

Afprøv denne tekst som input til Google Translate, og få et underholdende svar.

2.1. PoS-annoteret tekst

En klassisk og velafprøvet måde at disambiguere en tekst på er ved at PoS-tagge den, altså give hvert ord et ordklassemærke (et *tag*).

(*a'*)	*Maj*/prop	*så*/vb	*Find*/prop	*så*/vb	*frø*/sb
(*b'*)	*Maj*/fornavn	*så*/vb.præt.akt	*Find*/fornavn	*så*/vb.inf.akt	*frø*/sb.itk.pl.ubest

PoS-tagging kan gøres manuelt eller overlades til en automatisk PoS-tagger (programmer af denne type kaldes 'classifiers'). Mange PoS-taggere er kun trænet til at fordele input i de overordende ordklasser, verbum, substantiv, adjektiv, konjunktion osv. Denne analysedybde er tilstrækkelig til en del formål. Den fanger ikke den lemmatiske forskel på de to instanser af *så* i eksempel *a'*, men dette vil ikke påvirke kvaliteten af

en talesyntese da de to instanser udtales ens; tilsvarende for homografen *frø* der både kan være en neutrums- og en utrumsform, men med samme udtale. En oversætter har naturligvis brug for den finere klassifikation i *b'*.

I andre situationer er disambiguering med PoS-tagging imidlertid utilstrækkelig eller irrelevant uanset den morfosyntaktiske finhedsgrad. Et nominalsyntagme som *flot fyr* kan ikke disambigueres leksikalsk ved PoS-analyse alene.

	flot	*fyr*	
(*c'*)	adj.sg.ubest.itk	sb.itk.sg.ubest	≈ NICE STOVE
(*d'*)	adj.sg.ubest.fk	sb.fk.sg.ubest	≈ HANDSOME FELLOW
(*e'*)	adj.sg.ubest.fk	sb.fk.sg.ubest	≈ BEAUTIFUL PINE TREE

Som det ses, er de to eksempler *d'* og *e'* formelt uskelnelige. Det er dårligt nyt for oversætteren, men her også for talesyntesen idet *fyr fellow* og *fyr pine* udtales forskelligt. For at redde situationen kunne vi overveje at tilføje nogle træksemantiske tags til PoS-inventaret. Men hvilke?

Det har vist sig vanskeligt at designe et versatilt tagsæt som er lige relevant til alle (sprogteknologiske) formål. Jagten på den universelle tagger er endt uden resultat og har – trods fem årtiers datalingvistisk indsats – efterladt os med et pletora af tekstkorpora der ikke kan lægges sammen på grund af forskelle i annotationsart og annotationsdybde. I vores undersøgelse (Kirchmeier et al. 2019, Kirchmeier et al. 2020) kunne vi konkludere at, ud af 100+ store danske korpora skabt for offentlige midler gennem de seneste 30 år, kun cirka fem har relevans i dag. Resten er endt på datakirkegården, ikke mindst på grund af inkompatible annotationsformater.

2.2. Leksikalsk disambigueret tekst

Oprettelsen af Det Centrale Ordregister giver mulighed for at skifte perspektiv. Vi foreslår at man erstatter ideen om en tagger med ideen om en linker. Linkeren klassificerer ikke tekstens elementer i forhold til et forud givet (og projektrelateret) inventar af kategorier, men annoterer i stedet hvert tekstord med en reference til et centralt ordregister (in casu COR_1) som i sig selv er applikationsneutralt. Sammenlign den lek-

sikalsk disambiguerede tekst i *c"-e"* herunder med den morfosyntaktisk disambiguerede tekst i *c'-e'* (bemærk især adskillelsen af *fyr*-instanserne i *d"* og *e"*).

	flot	fyr	
(*c"*)	COR.17043.301.01	COR.77168.120.01	≈ NICE STOVE
(*d"*)	COR.17043.300.01	COR.53883.110.01	≈ HANDSOME FELLOW
(*e"*)	COR.17043.300.01	COR.84448.110.01	≈ BEAUTIFUL PINE TREE

Princippet om COR-linking kalder naturligvis på to spørgsmål: Hvad gør man med tekstord som ikke forekommer i COR_1? Og hvad med de applikationer som behøver leksikalsk information der ikke findes i COR_1 (udtale, semantik, terminologiske relationer osv.)? Disse spørgsmål bliver belyst i det følgende.

3. CLINK ver. 1.0

Dansk Sprognævn udgav d. 1/10 2022 en COR-linker, kaldet CLINK, som frit kan downloades, bruges og videreudvikles (under licens CC0). CLINK er en input-output-applikation; den læser en tokenstreng (*plain text*) og udskriver den samme streng med hvert token annoteret med COR_1-indeks. Ord som ikke forekommer i COR_1, tagges med <OOV> (Out Of Vocabulary).

Algoritmen i CLINK 1.0 bygger på tre forskellige strategier; de præsenteres i 3.1-3.3 herunder, og interaktionen imellem dem i 3.4. Læsere uden interesse for datalingvistik kan springe til afsnit 3.5 uden skade for sammenhængen.

3.1. Strategi 1: Window-one

Den simpleste linker-strategi ser kun på ét token ad gangen. I parserteori omtales denne strategi somme tider som 'window-1-decisions', altså de beslutninger der kan træffes om et token uden at skele til dets omgivelser. I nogle situationer er denne strategi optimal, nemlig for former med nul eller ét indeks i COR_1 (her har man ingen gavn af konteksten). Men når et token er homografisk, er strategi 1 (S1) naturligvis dårlig (her afhænger den gode beslutning netop af de syntaktiske omgivelser).

Formelt betragtet er S1 komplet, forstået sådan at den altid kan annotere alle tokens i input (enten med et COR_1-indeks eller <OOV>). Ingen af de andre strategier er komplette i denne forstand, og S1 er derfor nødvendig som fallback. I classifier-termer har S1 en *recall* på 1.0. For de tokens som ikke er homografer i COR_1, er *precision* også 1.0, men for de homografe tokens er præcisionen dårlig: Her linker S1 jo bare til den mest frekvente form i COR_1 (hvis den har adgang til en frekvenstabel) eller også må den træffe et tilfældigt valg – hvad skulle den ellers gøre?

3.2. Strategi 2: De lokale omgivelser

Strategi 2 (S2) træffer beslutninger baseret på de morfosyntaktiske omgivelser omkring et token. Det simpleste eksempel er flerordsforbindelsen (MWE, 'multi word expression'), hvor et token indgår i et *n*-gram som er leksikaliseret. Der er ikke stor forskel, hvad den syntaktiske funktion angår, på en MWE og et enkelt leksem, og ofte er MWE'er leksikaliseret i COR_1 som rene ortografiske varianter.

ingen ting	pron	`COR.00561.991.01`
ingenting	pron	`COR.00561.991.02`
gud hjælpe mig	adv	`COR.12501.900.01`
gudhjælpemig	adv	`COR.12501.900.02`

Da CLINK arbejder token-for-token, er hvert token i en MWE i linkerens output annoteret med et link.

Der hvor S2 er mest effektiv, er til analyse af nominalsyntagmer hvor morfologisk kongruens spiller en central rolle (især konstruktioner som DET ADJ* CN MOD*, hvor DET er et determinativled, ADJ* en gruppe adjektiver, CN et appellativ og MOD* pladsen for præpositionsforbindelser og andre postmodifiers). Her gemmer sig den største (og interessanteste) udfordring for linkerudvikleren; af pladsgrunde må vi dog her nøjes med at illustrere S2 med nogle eksempler på leksikalsk disambiguering ved hjælp af unifikation af morfologiske træk.

S2 benytter den såkaldte PAROLE-formalisme til morfologisk trækstruktur (Keson 1999, Henrichsen 2002). En trækstruktur som *sb.itk. pl.ubest.gen* i COR_1 (svarende til ordformer som *bogstavers* og *æblers*)

ser i PAROLE sådan ud: NCNPG==I. Det første tegn er hovedordklassen, og hvert efterfølgende tegn koder for et bestemt morfologisk træk. Tagget NCNPG==I svarer således til Noun-Commonnoun-Neuter-Plural-Genitive-*void*-*void*-Indefinite. En trækværdi der er uspecificeret, markeres med et punktum, mens et træk der er udefineret markeres med = eller –. PAROLE-formalismen er udviklet specielt til brug i taggere (og linkere).

et stort fyr	P**I**–NSU---	ANPNSU=**I**U	NCNSU==**I**	≈ A BIG STOVE
et flot fyr	P**I**–NSU---	ANP.SU=**I**U	NCNSU==**I**	≈ A NICE STOVE
fyrre orange fyr	ACP--U=--	ANP...=.U	NCNPU==I	≈ FORTY ORANGE STOVES
fyrre flotte fyrre	ACP--U=--	ANP.PU=.U	NCCPU==I	≈ FORTY BEAUTIFUL PINES

I eksemplerne er kongruenser i trækket *bestemthed* markeret med **fed**, trækket *numerus* med understregning og trækket *genus* med dobbeltunderstregning. CLINK's opgave er altså at udvælge de COR_1-links der får hele kongruensregnestykket til at gå op. I datalogien kaldes denne øvelse 'unification'. Læsere med programmeringserfaring bliver næppe overraskede over at koden til S2 er skrevet i PROLOG, et sprog som er særligt egnet til netop unifikation.

3.3. Strategi 3: Den videre kontekst

S1 og S2 supplerer hinanden fint. S1 garanterer at linkeren som sådan har en *recall* på 1.0 mens S2 bidrager til en stærkt forbedret *precision*. Men der er stadig 'mørke områder' i en almindelig tekststreng med homografi der er uden for begge strategiers rækkevidde.

en flot fyr	*fyr* / sb.fk.sg.ubest / COR.53883.110.01
en flot fyr	*fyr* / sb.fk.sg.ubest / COR.84448.110.01

Godt nok kunne vi slå fast i afsnit 2.2 at COR-links (i modsætning til PoS-tags) kan adskille eksemplets to instanser af *fyr* leksikalsk. Men hvordan skal CLINK vælge det relevante link? Hvis input til linkeren kun består af *en flot fyr*, er afgørelsen umulig (selv for et menneske); men hvis input er en længere tekst, og *fyr* altså har en kontekst der går forud, kan denne afsøges for leksikalske pejlemærker. Dette er Strategi

3's funktion (S3). Den har adgang til en associationstabel som fx knytter ordformen

fyr	COR.84448.110.01	PINE TREE

til en række af semantisk associerede lemmaer som

træ	sb	COR.44241
nåletræ	sb	COR.91480
brænde	sb	COR.91213
brænde	vb	COR.30160

(og tilsvarende for *fyr* FELLOW). Da CLINK jo, for hvert token i inputstrengen, har adgang til sin egen analyse af de forudgående tokens, kan den (med lidt held) bruge de associerede lemmaer i venstrekonteksten som indicier. CLINK version 1.0 har kun omkring 80 leksikalske indgange i sin associationstabel – men vi forventer en markant udvikling på denne front når COR-projektet i slutningen af 2023 offentliggør sin omfattende ordsemantiske database (se Pedersen et al. 2022).

3.4. Parallel versus seriel kobling

Der er (grundlæggende) to måder at koble de tre strategier på, enten parallelt eller i serie. Ved parallelkobling, også kaldet 'voting', prøves alle tre strategier på ethvert token – og den endelige beslutning om inputordets link træffes ved afvejning af de tre bud. Hvis S1 er meget sikker i sin vurdering, mens S2 slet intet bud har, og S3 er usikker, går beslutningen til S1. Generelt træffes valget altså i forhandling mellem strategierne, og hvis de ikke kan blive enige, sørger en dommerstrategi for at lægge valget hos den strategi som viser størst sikkerhed ('confidence'). Parallelstrategier kan nå meget høje succesrater, men er vanskelige at optimere på grund af de mange variabler som programmøren skal justere.

Den serielle kobling fungerer som et samlebånd hvor hver strategi anvendes efter tur. Den har færre frihedsgrader og er lettere et optimere, og vi har derfor foretrukket en seriekoblet algoritme til CLINK ver. 1.0.

TOKEN → **S1**$^{-}$ → **S2** → **S3** → **S1**$^{+}$ → LINK

Bemærk at S1 i første omgang kun får myndighed til at beslutte links for ikke-homografe inputtokens (hvor *precision* er 1.0). Kun hvis hverken S2 eller S3 kan beslutte sig for et link, får S1 det sidste ord (men med lav *precision*).

Vi deler koden til S1, S2 og S3 med alle interesserede, og vores håb er at andre vil tage udfordringen op så vi i fællesskab får skabt en hurtigere, mere præcis COR-linker til glæde for alle.

3.5. Nyeste CLINK-udvikling

Dansk Sprognævn arbejder i disse måneder på at COR-linke korpus DAGW (Danish Gigaword Corpus, se www.sprogteknologi.dk). DAGW består af uannoterede tekster fri for copyright. Korpusteksterne er valgt i genrer som de fleste danskere er i jævnlig kontakt med, nærmere motiveret i Derczynski et al. (2021). Teksterne er uannoterede og foreligger som 'plain text' (i formatet UTF-8). DAGW har, trods sin unge alder, vundet stor udbredelse som dansk referencekorpus til både forskning og udvikling. Dansk Sprognævn sigter mod at have en komplet COR-linket DAGW klar inden jul 2023. Først på dét tidspunkt bliver det muligt at udmåle CLINK's performans kvantitativt. I skrivende stund har vi foretaget kvalitative målinger (stikprøvebaserede, manuelt evaluerede); efter vores bedste skøn har CLINK 1.0 en fejlrate (målt på blandede tekstgenrer a la DAGW) på 2.5-4.5 %. Læs mere om COR-linking i Henrichsen (2023).

4. Videre perspektiver

4.1. Kontrollerede udvidelser af Det Centrale Ordregister

Der er naturligvis mange danske ord som ikke findes i COR_1: fagord, neologismer, ældre ord, proprier og en ubegrænset mængde komposita. Desuden har mange NLP-applikationer brug for leksikalsk information som COR_1 ikke rummer.

Talesyntese	*fonetiske og prosodiske data*
Maskinoversættelse	*semantiske ækvivalenter i målsproget*
Terminologisystemer	*begreber og begrebsrelationer*
Betydningsordbøger	*ordbetydninger*
AI-baserede dialogsystemer	*sætnings- og diskurssemantik, verdensviden, ...*

Både nye leksemer og supplerende leksikalske data til eksisterende leksemer kan, på systematisk måde, gøres tilgængelige via Det Centrale Ordregister. Dertil har vi COR's niveauer 2 og 3 med eksplicit leksikalsk linking til COR_1 (Widmann 2023, Dideriksen et at 2022). Sprogressourcer som overholder COR-formatet, kaldes *COR-kompatible*.

Vi opfordrer alle danske korpus- og ordbogsredaktioner til at gøre deres ressourcer COR-kompatible (manualer findes i www.ordregister.dk). Øvelsen består i (*i*) at COR-indeksere hver leksikalsk indgang og/eller hvert tekstord, (*ii*) at udarbejde en afbildningstabel mellem de nye indekser og COR_1. Man kan vælge enten at publicere sin COR-kompatible ressource eller at nøjes med at udgive afbildningstabellen og så bevare sit indhold bag fx en betalingsmur. I begge tilfælde øger man sin ressources værdi som data til sprogteknologiske anvendelser.

4.2. Projekt COR – nu og i fremtiden

Det Centrale Ordregister er, i skrivende stund, stadig under udvikling. Projektgruppen består af Dansk Sprognævn (design af det formelle rammeværk, udvikling af COR_1), Det Danske Sprog- og Litteraturselskab (leksikografiske udvidelser og ordsemantisk annotation) og Center for Sprogteknologi (maskinlæring i forbindelse med semantisk annotation). Projektarbejdet er støttet af en treårig bevilling fra Innovationsministeriet, administreret af Digitaliseringsstyrelsen (digst.dk); læs mere om den danske sprogteknologiske satsning i Kirchmeier et al. (2019) – og besøg også www.sprogteknologi.dk, hvor den danske sprogteknologiske satsning er omtalt i detaljer.

Når COR-bevillingen rinder ud med udgangen af 2023, er det vigtigt at de nyudviklede sprogressourcer fortsat vedligeholdes og udvikles. Dansk Sprognævn, som har en naturlig forpligtelse til at støtte alle aspekter af dansk sprogbrug – inklusive de sprogteknologiske – har derfor indvilget i at stå for den fremtidige administration af Det Centrale Ordregister. Opgaven består ikke kun i at vedligeholde COR_1-ressourcens tilgængelighed og aktualitet, men også i at vejlede om forberedelsen af nye COR-kompatible ressourcer.

4.3. Nordisk samarbejde – et CALL for COR

Færøerne har nu udviklet sin egen parallel til COR kaldet OTAL (Simonsen et al. 2022). Det er en imponerende bedrift, og Færøernes eksempel viser at hvor der er vilje, er der vej. Vi vil hermed opfordre alle de nordiske sprogsamfund, større såvel som mindre, til at oprette deres egne centraliserede ordregistre til gavn for sprogteknologien i hele Norden.

Referencer

Litteratur

Derczynski, Leon et al. 2021. The Danish Gigaword Corpus. *Proceed. of NODALIDA-23*. Linköping Electronic Conference Proceedings 178 (2021).

Dideriksen, Christina, Peter Juel Henrichsen & Thomas Widmann 2022. Det Centrale Ordregister. I: *Nyt Fra Sprognævnet*. Oktober 2022. ISSN 2446-3124.

Henrichsen, Peter Juel 2002. *Sidste Års Aviser*. I: Institut For Datalingvistik (KU): LAMBDA 27.

Henrichsen, Peter Juel 2023. Diktaroriske Beføelser. Om Ord og Uord i Det Centrale Ordregister. I: *Proceedings of MUDS19* (Møde om Udforskningen af Dansk Sprog). Aarhus Universitet.

Keson, Britt 1999. *Vejledning til det Danske Morfosyntaktisk Taggede PAROLE-korpus*. DSL Press.

Kirchmeier, Sabine, Peter Juel Henrichsen & Philip Diderichsen 2019. *Dansk Sprogteknologi i Verdensklasse*. Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet. ISBN 978-87-89410-77-7.

Kirchmeier, Sabine, Bolette Sandford Petersen, Peter Juel Henrichsen; Sanni Nimb & Philip Diderichsen 2020. World Class Language Technology – Developing a Language Technology Strategy for Danish. *Proceedings of LREC 2020*.

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. *Proceed. of LREC2022*.

Schack, Jørgen et al. (red.) 2012. *Retskrivningsordbogen*. 4. udgave. Dansk Sprognævn.

Simonsen, Annika, Sandra Saxov Lamhauge, Iben Nyholm Debess & Peter Juel Henrichsen 2022. Creating a basic language resource kit for Faroese. *Proceed. of LREC2022*.

Widmann, Thomas 2023. Det Centrale Ordregister og dets leksikografiske anvendelser. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 415–430.

Links (verificeret maj 2023)

Link1 (1997) Lov om Dansk Sprognævn (LOV nr 320 af 14/05/1997)
https://www.retsinformation.dk/eli/lta/1997/320
https://kum.dk/ministeriet/organisation-og-institutioner/bestyrelser-raad-naevn-og-udvalg/dansk-sprognaevn-repraesentantskab

Link2 (1997) Lov om dansk retskrivning (LOV nr 332 af 14/05/1997)
https://www.retsinformation.dk/eli/lta/1997/332

Link3 (1995) Dansk SAMPA, fonetisk alfabet til computerbrug (se også DSN's modificerede SAMPA i www.dsn.dk)
https://www.phon.ucl.ac.uk/home/sampa/danish.htm