Kirstine Boas, Inger Schoonderbeek Hansen, Tina Thode Hougaard
& Ea Lindhardt Overgaard (udg.):
19. Møde om Udforskningen af Dansk Sprog, Århus 2022

Diktaroriske befølelser
Om ord og uord i Det Centrale Ordregister

Peter Juel Henrichsen
Dansk Sprognævn

To kendte russiske historikere Andronik Mirganjan og Igor Klamkin tror ikke, at Rusland kan udvikles uden en "jernnæve". De hævder, at Ruslands vej til demokrati går gennem diktatur. I en af deres artikler hedder det: "I et autoritært regime lagdel samfundet og forskellige interesser modnes. Og når deres repræsentanter er parate til at gå i struben på hinanden, så stopper en "jernnæve" det. På den måde skabes hele tiden betingelserne for en harmonisering af interesser og følgelig for demokratiske reformer".

Og de tilføjede: "Netop fordi vi end ikke har et kim af civilt samfund, var vi imod Folkekongressen. Bortset fra illusionen om demokrati kunne den ikke give noget reelt. Det kunne den ikke helt objektivt. Det var slet ikke på grund af apparatets lænker. Illusioner er farlige. De føder skuffelser og leder til sidst til en destabilisering af samfundet, hvad vi allerede har set og desværre vil komme til at se mere af." De mener, at Folkekongressen skal give præsidenten diktaroriske befølelser.

Det danske PAROLE-korpus (to første afsnit)[1]

1. Det Centrale Ordregister – en introduktion

Det Centrale Ordregister (COR) er et register for det danske ordforråd. Hvert dansk ord har (eller kan få) et COR-nummer der fungerer som ordets leksikalske holdepunkt.

[1] PAROLE-korpusset består af små, ukomplette tekststykker og kan derfor citeres frit iflg. citatretten (se Link3 2023). Tekststykket er citeret fra en artikel i Det Fri Aktuelt d. 1/12 1992 (nærmere info herunder).

’To’ / `COR.01528.600.01`

’kendte’ / `COR.18159.303.01`

’russiske’ / `COR.22261.303.01`

’historikere’ / `COR.58774.112.01`

Her ses de første fire ord i det danske PAROLE-korpus annoteret med COR-numre (herfra kaldet *links*). Et COR-link består af præfikset "COR" efterfulgt af tre numeriske indekser.

`COR` . <*lemmaindeks*> . <*bøjningsindeks*> . <*variantindeks*>[2]

Lemmaet *russisk* har altså lemmaindekset `22261` mens *historiker* har `58774`. Bøjningsindekset viser ordets morfologiske klasse og bøjning, fx står `112` for bøjningen sb.fk.pl.ubest (i Retskrivningsordbogens notation) mens `303` står for adj.pl. Sidst i artiklen findes en tabel over de bøjningsindekser der forekommer her i artiklen (den fulde definition ses i ordregister.dk, både som tabel og som opslagsfunktion).

Såvel lemmaindekser som bøjningsindekser er entydige, dvs. to forskellige lemmaer vil altid have to forskellige lemmaindekser, og tilsvarende har forskellige bøjningsformer (fx ’russiske’$_{sg.best}$ og ’russiske’$_{pl}$) altid forskellige bøjningsindekser (hhv. `302` og `303`). Tekst der er annoteret med COR-links, er altså fri for homografi og er dermed lettere at databehandle (for både mennesker og computere).

COR-annoterede tekster er gode korpusressourcer – bedre end de velkendte Part-of-Speech-taggede korpora. PoS-tagging er, som bekendt, en af de hyppigst anvendte metoder til forberedelse af inputtekst til maskinoversættelse, talesyntese, tekstbehandling og mange andre slags NLP (*Natural Language Processing*). Ulempen ved PoS-tagging er at selve tagdefinitionen, altså de tags som man klassificerer tekstens tokens med, varierer fra det ene projekt til det næste. En grov ordklassebestemmelse er fx nok til at støtte en talesyntese mens et tekstbehandlingsprogram med grammatikkontrol behøver en meget finere morfologisk klassifikation. To korpora der er annoteret med forskellig tagdefinition, er besværlige at lægge sammen, i praksis ofte umulige (Kirchmeier et al. 2020). En af fordelene ved COR’s standardiserede bøjnings- og

[2] Variantindekset bruges i de tilfælde hvor ortografien tillader flere forskellige stavninger af den samme ordform (fx ”ressource” eller ”resurse”, ”allevegne” eller ”alle vegne”). Lemmaindekser har 5 cifre, bøjningsindekser har 3, variantindekser har 2. Det faste format gør indekserne lette at genkende i artiklens eksempler. Formatet gælder i grundregisteret, emnet for denne artikel, men ikke for alle COR-ressourcer (omtales herunder).

lemmainformation er at COR-annoterede sprogressourcer lavet til projekt A og projekt B umiddelbart kan lægges sammen og bruges som korpus i projekt C, uanset hvilke forskelle der måtte være mellem A, B og C's formål og metoder.

COR-links er mere informationsrige end PoS-tags. En COR-annoteret tekst afbildes på en PoS-tagget tekst simpelthen ved at fjerne lemmaindekserne. Fjerner man dertil de to sidste cifre i bøjningsindekserne, står hovedordklasserne tilbage (`1` for substantiver, `2` for verber, `3` for adjektiver osv.). Har man brug for en lemmatiseret tekst, fjernes i stedet bøjningsindekserne. På den måde erstatter COR-linking såvel PoS-tagging som lemmatizing.

Variantindekset spiller en mindre rolle, og vi udelader det i resten af artiklen for læselighedens skyld; men faktisk kan også dette indeks gøre nytte for eksempel som input til et tekstbehandlingsprogram med stilkontrol (forfatteren er fx tilbøjelig til at flakke mellem varianterne 'resurse' og 'ressource', og her ville en løftet pegefinger være rart).

I artiklens første del præsenterer vi COR's grundregister (COR1), som alle andre COR-ressourcer refererer til. Man kan frit downloade COR1 og andre COR-relaterede sprogressourcer (ordbøger, korpora, værktøj etc.). En af de nyeste tilføjelser er PARC, det danske PAROLE-korpus[3] med COR-annotation, der har leveret teksteksemplerne til denne artikel. I artiklens sidste del giver vi et overblik over de vigtigste COR-ressourcer og -metoder, med omtale af bl.a. automatisk COR-linking og praktiske COR-anvendelser i sprogteknologi. Vi må erkende at det bliver et hastigt flyby, men vi håber at læseren får lyst til at følge referencerne til de uddybende publikationer.

COR-ressourcer kan downloades fra www.ordregister.dk, som også giver adgang til forskellige online-tjenester såsom leksikalske opslag i COR1.

2. Nøgen tekst er dårligt input

Computere har svært ved at arbejde med naturligt sprog. Sproget er fuldt af flertydigheder, som vi mennesker navigerer sikkert imellem takket være vores forståelse af andre menneskers intentioner; computerprogrammer, derimod, opfatter ethvert inputelement som et entydigt datapunkt uanset om det drejer sig om et ord, et ciffer, en lydsample eller en pixel. Når et NLP-program får galt fat på et ord, giver det tit anledning til følgefejl af en karakter som mennesker

[3] Den oprindelige projektbeskrivelse for PAROLE indebar at alle tokens skulle bevares præcis som de forekom i kildeteksterne. Korpusset er rigt på stavefejl og andre anomalier og derfor en god kilde til 'økologiske' tekstprøver. Læs om PAROLE i Bilgram et al. (1993, 1998), Keson (1999) og Henrichsen (2002).

aldrig ville begå, fejl uden en gnist af fornuft.

Når MS-Word's stavekontrol læser ordet 'hundehvalpenne', bliver det godkendt selv i konteksten "kattekillingerne og hundehvalpenne drak af samme vandskål". Ordet kan analyseres som et kompositum 'hunde-hval-penne' og er dermed et velstavet ord, pedantisk betragtet. En sætning som "killingerene og hvalpenne drag ad same skol" giver ikke anledning til en eneste kommentar fra funktionen Stave- og Grammatikkontrol. Alle ordene er jo velstavet (med 'skol' som imperativ). Der er ikke megen hjælp at hente for hverken dyslektikere eller andre.

Men stavetjekkerne er ikke værre end resten af NLP-familien. Hvis man siger til Siri: "Pæren i lampen er sprunget, hvor kan jeg købe en ny pære?", svarer hun "Jeg ved ikke hvor der er en grønthandler". Og hvis man serverer denne tekst til Google Translate:

"Jeg har lige været til to studenterfester. Lærkes var fed. Majs var dødssyg."

og beder om at få den oversat til engelsk, så vender Translate tilbage med denne tekstperle

"I just went to two student parties. Larks were fat. Maize was deathly ill."[4]

Translate har åbenbart forvekslet den intenderede form 'Majs'$_{\text{prop.gen}}$ med dens homograf 'majs'$_{\text{sb.fk.sg.ubest}}$ mens 'Lærkes'$_{\text{prop.gen}}$ er forvekslet med 'lærkes'$_{\text{sb.fk.pl.ubest}}$ (som ret beset er et uord).

Skal man pege på én grundlæggende udfordring som næsten alle slags sprogprogrammer lider under, uanset om de behandler tekst eller tale, så er det at de danske ordformer kun sjældent er leksikalsk entydige (ganske særligt de kortere ord). Det problem løser NLP-udvikleren, som før nævnt, ved at tagge hvert teksttoken, fx morfologisk, prosodisk eller terminologisk, ud fra sit eget formål. Der er mange annotationsstandarder i brug i de danske NLP-værksteder i dag (Kirchmeier et al. 2019). Alt for mange. Det kan alle blive enige om. Derfor har den fælles og versatile annotationsstandard – lige relevant til alle NLP-formål – været et mål for datalingvistikken gennem mindst fire årtier. Desværre er vi ingen vegne kommet, og derfor står vi i dag med et pletora af inkompatible tekstkorpora. En nylig undersøgelse (Kirchmeier et al. 2020) konkluderede at, ud af 100+ store danske korpora skabt for offentlige midler gennem de seneste 30 år, har kun omkring fem relevans i dag. Resten er endt på datakirke-

[4] Afprøvningen af MS-Word (aktuel version distribueret af Statens-IT), Google Translate (i Chrome) og Apple's Siri (afprøvet på ipad, ny model 2022) blev foretaget d. 5/1 2023.

gården, ikke mindst på grund af inkompatible annotationsformater.

COR-linking blev udviklet som et versatilt alternativ. Vi vender tilbage til hvordan COR-baseret tekstbehandling kan bruges mod homografblindhed (som i eksemplerne herover).

3. COR's grundregister

COR's grundregister, kaldet COR1, er afledt af Retskrivningsordbogen og tildeler hver af RO's godt 530.000 ordformer et unikt COR-indeks. RO dækker, ifølge sine redaktionsprincipper, det almene danske ordforråd forstået som de ord der (*i*) benyttes i hele samfundet (dvs. ikke er typiske for bestemte alders- eller samfundsgrupper), (*ii*) forekommer stabilt over tid og (*iii*) generelt ikke refererer til specifikke organisationer, sagsforhold, produkter og personer (Jervelund et al. 2012). Den offentlige sektor er forpligtet til at følge RO's ortografi i skriftligt arbejde, fra ministerierne til folkeskolen (iflg. Lov om Dansk Retskrivning, se ref. Link2 1997). RO 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 Det Centrale Ordregister.

Her er nogle eksempler (fra PARC) der viser hvordan homografer kan adskilles med COR1-links, selv i tilfælde hvor der ikke er forskel på deres stavning og bøjningsparadigme. Hvis COR-linking skal slå an for alvor, er der imidlertid to spørgsmål som må besvares: Hvordan linker man et token som ikke forekommer i COR1? Og hvorfra skal NLP-applikationen hente de arter af leksikalsk information som ikke findes i RO (stavning, bøjning og sammensætningspotentiale)? Det ser vi nærmere på i de næste afsnit.

4. COR-linking af det ukendte token

Når man afprøver en almindelig PoS-tagger på en tekst, får hvert eneste token et tag, selv de særeste stavefejl. Det ukendte token er ikke et principielt problem for en tagger, den skal bare klassificere det ud fra dets materielle lighed med kendte tokens. Linkeren, derimod, er forpligtet til at finde et leksikalsk mål til hvert eneste token. Lad os tage nogle konkrete eksempler. I PAROLE-citatet først i artiklen er der tre tokens (foruden proprierne) som ikke findes i COR1: 'jernnæve', 'diktaroriske' og 'befølelser', og desuden også formen 'lagdel' der ikke synes at passe i sin syntaktiske sammenhæng. Disse fire anomale tokens kan annoteres på forskellig måde i COR, afhængigt af hvilken status man ønsker at tillægge teksten; er den fx kanonisk, et 'værk', så alle tokens *skal* tages bogstaveligt, eller er den et løbende input til et tekstbehandlingssystem som skal granskes og rettes?

Der er i alt fald seks måder at COR-linke det ukendte token på, som alle bidrager med forskellig teksttolkning og dermed tilgodeser forskellige typer af NLP. Tabel 1 kobler de forskellige varianter og NLP-typer til hinanden og giver eksempler på de enkelte varianters relevans.

Variant	'jernnæve'	'lagdel'	'diktaroriske'	'befølelser'	***NLP-relevans***
afvisning	`00000.000`	`00000.000`	`00000.000`	`00000.000`	_
PoS-tagging	`00000.110`	`00000.204`	`00000.303`	`00000.112`	tegnsætning
korrektion	_	`37419.204`	`29293.303`	`77788.112`	stavekontrol oversættelse
abstraktion	_	`37419`	_	_	emnesøgning resumering
analyse	`[97002][]` `[98252.110]`	_	_	`[be-][30029]` `[-else][-r]`	talesyntese læremidler
accept	`PARC.`*`p`*`.110`	`PARC.`*`q`*`.204`	`PARC.`*`r`*`.303`	`PARC.`*`s`*`.112`	tekstkanon

Tabel 1. COR-linking af det ukendte token – seks forskellige tilgange. Id'er af typen `PARC.p.110` er forkortelser af `COR.PARC.p.110` (osv.). Symbolerne `p`, `q`, `r` og `s` er lemmaindekser i ordbogen COR.PARC. Den højre kolonne giver eksempler på relevante NLP-formål.

De seks tilgange til annotation af det ukendte token har forskellige fordele og ulemper. Her må vi nøjes med at give nogle stikord, men reelt afspejler de en grundlæggende holdningsforskel til teksten.

Afvisning. Blot at konstatere at et ukendt token er *out of vocabulary* (ofte kaldet OOV) er det mindst risikable, men også det mindst værdiskabende valg.

PoS-tagging. Hvis formålet er at forberede en tekst til automatisk tegnsætning (kommatering), er en klassisk PoS-tag et naturligt valg. I COR-termer svarer det til kun at instantiere bøjningsindekset og lade lemmaet forblive ukendt.

Korrektion. En automatisk oversætter er afhængig af en egentlig korrektur, altså en erstatning af det ukendte (eller fejlbøjede) token med et kendt alternativ.

Abstraktion. Denne tilgang er komplementær til PoS-tagging: at identificere lemmaet, men

ikke bøjningen. Automatisk tekstresumering og emnesøgning refererer typisk kun til lemmaer, ikke bøjningsformer.

Analyse. En talesyntese omsætter tekst til lyd, med lydskrift som mellemled. Udtaleordbogen er ofte et 'morfemikon' med udtaleregler for morfemer snarere end fuldformer. Det ukendte token skal derfor analyseres i morfemer, så langt som muligt.

Accept. I de fem første tilfælde er det ukendte token tolket som en tekstfejl. Men hvad hvis inputteksten er 'kanonisk', altså fejlfri a priori? I det tilfælde skal der oprettes en ny COR-ordbog til supplement af COR1 (in casu ordbogen COR.PARC), og det ukendte token linkes til et helt nyt leksem.

5. Nye COR-ordbøger

COR's grundregister dækker naturligvis ikke det samlede danske ordforråd (hvordan det end bestemmes), og derfor er det nødvendigt også at kunne linke *ud af* COR1. Oplagte eksempler på leksemer der er svagt repræsenteret i COR1, er proprierne og numeralierne, som RO kun dækker nødtørftigt. COR1 mangler også fagord, fremmedord, neologismer, samt en ubegrænset mængde af komposita. Dertil kommer interpunktionstegnene, som sprogfolk ganske vist ikke regner for rigtige sprogtegn (med Victor Borge som en undtagelse), men som ikke desto mindre bærer syntaktisk information der er uhyre relevant for NLP. Hvis COR-linking skal være sit salt værd, må ethvert teksttoken derfor kunne få et link. I dette afsnit diskuterer vi hvordan nye ordbøger kan føjes til COR-systemet og give NLP-programmerne mulighed for at importere leksikalsk information fra mange sider.

Vi har allerede nævnt den nye ordbog COR.PARC (se tabel 1), der dækker de ordformer i PAROLE-korpusset som ikke findes i COR1. Korpus PARC kan downloades fra ordregister.dk, ligesom ordbøgerne COR.PARC og COR.XP, sidstnævnte med COR-indekser for alle de almindelige interpunktionstegn, fx

```
!/COR.XP.1.03
,/COR.XP.2.01
(/COR.XP.8.02
```

For at gøre en eksisterende ordbog COR-kompatibel skal hver enkelt leksikalsk indgang blot

forsynes med et unikt COR-indeks efter visse formelle regler[5]. Når alle indgange er indekseret, kan man sikre at ordbogen i sin helhed er formelt COR-kompatibel ved hjælp af Sprognævnets COR-værktøj. Når en ny COR-kompatibel ordbog er klar, kan man vælge at distribuere den frit; men alternativt kan man nøjes med at udstille dens opslagsord og holde den leksikalske information skjult bag en licens- eller betalingsmur (se tabel 2). På den måde kan fx en salgsvirksomhed gøre opmærksom på sit leksikalske produkt, dets omfang og dækning, uden at offentliggøre indholdet.

Artikler	**BETALINGSMUR**	**COR-indeks**	**COR1-afbildning**
		`COR.OPEN.SMOVS.70250`	`COR.70250` [bouillon] ← `COR.96878` [kapital]
		`COR.OPEN.SMOVS.60058`	`COR.22093` [*adj.* genert] `COR.32735` [*vb.* undgå] `COR.60058` [*sb.* gelé] ← `COR.73562` [*sb.* vanddamp]

Tabel 2. Firmaet Smovs-i-Sovs har udgivet en COR-kompatibel ordbog. Tabellen viser to af ordbogens artikler (”fond” og ”sky”) med deres COR-indekser. Hvor det giver mening, kan indekser i ordbogen afbildes på COR1 (fx `COR.OPEN.SMOVS.70250` ← `COR.70250`). Dermed kan firmaets sovse-app importere leksikalsk info fra RO uden at blive vildledt af de mange homografer til ’fond’ og ’sky’. Bemærk at Smovs-i-Sovs (kun) har offentliggjort oplysningerne til højre for betalingsmuren; de fungerer som reklame for den nye ordbog. COR-annotation kan således også være en forretningsmodel.

Der findes, i skrivende stund, omkring 10 COR-kompatible ordbøger og korpora der er offentligt tilgængelige (heriblandt COR1 og PARC). Mange flere er på vej. Sprognævnet arbejder på at COR-annotere det største frit tilgængelige tekstkorpus i Danmark, nemlig DAGW (DAnish GigaWord-corpus; Derczynski et al. 2021). Dette korpus, der tæller omkring 1 mia.

[5] En ordbog er COR-kompatibel når hver indgang har et COR-indeks af formen COR.OPEN.*NAVN.ix*, hvor *NAVN* er navnet på ordbogen, og *ix* indekserne for de informationstyper der er defineret for ordbogen. Denne type indeksering kaldes niveau 3, og den er åben for alle. Professionelle sprogredaktioner (DSN’s repræsentantskab iflg. Lov om Dansk Sprognævn, se ref. LINK1) har også mulighed for at indeksere i niveau 2 (med linkformat COR.*NAVN.ix*). COR-ordbøger i niv. 2 forventes at være kvalitetssikret i henhold til alm. leksikografiske principper (Widmann 2022; Dideriksen et al. 2022).

tokens, er hurtigt blevet et referencekorpus for dansk korpuslingvistik takket være sin størrelse og tilgængelighed. En COR-annoteret version vil opgradere DAGW betydeligt som sprogressource for både forskning og udvikling.

Også på den taleteknologiske front er der nyt i vente. Alexandra-instituttet er netop gået i gang med at udvikle Danmarks største frit tilgængelige talesprogskorpus (projekt CoRaL). Over 1000 timers indtalinger vil, efter planen, blive transskriberet og tidskodet i de kommende år og gjort tilgængelige for træning af akustiske modeller til bl.a. talegenkendelse. Alle transskriptioner bliver COR-annoteret.

Og sidst, men ikke mindst, er en afgørende vigtig COR-ordbog ved at være færdigudviklet, nemlig en omfattende betydningsordbog udviklet af DSL (Det Danske Sprog- og Litteraturselskab) og CST (Center for Sprogteknologi, KU), se Pedersen et al. (2022). Når ordbogen står klar i slutningen af 2023, vil den sætte nye standarder for hvordan dansk tekst kan annoteres semantisk – og dermed for udviklingen af NLP i retning ad kunstig intelligens.

6. COR-linket tekst som udgangspunkt for NLP

COR-linking giver et forbedret grundlag for de fleste typer af automatisk tekstprocessering. Det gælder ikke mindst tekstbehandling. Med COR-links kan man give et tekstbehandlingsprogram med stave- og grammatikkontrol stærkere briller på, og det giver vi her et eksempel på.

”I et autoritært regime lagdel samfundet og forskellige interesser modnes.”

Set fra et grammatisk synspunkt kan denne tekststreng ikke umiddelbart analyseres som en velformet sætning på grund af det skæve 5. token, ’lagdel’. En rimelig hypotese, understøttet af en rundspørge blandt kolleger, er at ordet egentlig skulle have været ’lagdeles’. En ordentlig stavekontrol bør altså foreslå korrekturen ’lagdel’→’lagdeles’, med det bagvedliggende argument at sætningen dermed bliver normaliseret både syntaktisk og semantisk. Med den COR-linkede tekst som udgangspunkt kan man opstille det følgende formelle ræsonnement (som kan overlades til en computer).

1 Teksten kan ikke, som den står, analyseres som en sætning. Det antages at årsagen er en stavefejl. Der gives derfor licens til at ændre ét token; korrekturen skal dog være realistisk (begrænset til nogle få bogstaver og også på anden måde minimal). Denne *licence to change* kalder vi *LiC*.
2 Der er to sætningsled (forstået som kongruensdomæner) som kan identificeres

sikkert vha. COR-annotationen: [’et’ ’autoritært’ ’regime’] og [’forskellige’ ’interesser’]

3 Tekstens 7. token er ’og’$_{970}$. En `970`’er (sideordningskonjunktion) knytter to kongruerende led sammen.

4 Konjunktionsfrasen [*V* ’og’ *H*] har tre mulige højreled: *H*=[’forskellige’]$_{AP}$, *H*=[’forskellige’ ’interesser’]$_{NP}$ og *H*=[’forskellige’ ’interesser’ ’modnes’]$_{S}$. Den første mulighed kan ikke finde et kongruerende venstreled *V* (modulo *LiC*, altså selv efter én korrektur); den anden kan godt kongruere med et venstreled, *V*=[’samfundet’]$_{NP}$, men det åbner stadig ikke for en sætningsanalyse (mod. *LiC*). Det tredje *H*-led tillader derimod en fuldstændig sætningsanalyse (mod. *LiC*). Med korrekturen ’lagdel’$_{37419}$→’lagdeles’$_{37419.204}$ kan teksten analyseres som [[’I’..’regime’ ’lagdeles’ ’samfundet’]$_{S}$ ’og’ [’forskellige’,’interesser’,’modnes’]$_{S}$]$_{S}$.

5 Der udskrives et forslag til brugeren: korrekturen ’lagdel’→’lagdeles’, evt. med en motiverende note.

Ræsonnementet i 1-4 viser sig altså at føre til et velmotiveret korrekturforslag. I modsat fald kunne man have løsnet *LiC* og givet flere frihedgrader. Arten og graden af korrektur kunne endda præsenteres som et valg for brugeren af tekstbehandlingsprogrammet.

Også korrekturerne ’diktaroriske’→’diktatoriske’ og ’befølelser’→’beføjelser’ kan understøttes med COR-baserede ræsonnementer; den sidstnævnte vil dog forudsætte en COR-ordbog med ret sofistikeret betydningsinformation. Vi venter spændt på DSL og CST’s magnum opus.

7. Automatisk COR-linking

Indtil nu har vi diskuteret *nytten* af COR-links, men ikke fremstillingen. Hvordan COR-linker man en tekst i praksis? Mindre tekststykker til grammatisk analyse og undervisningsmaterialer kan man naturligvis håndlinke; men til korpusannotation og NLP-udvikling er automatiske metoder et must. Sprognævnet har derfor udviklet et program til tekstannotation i stor skala, en såkaldt *linker*. Linkeren fungerer som en input-output-automat; den tager en tekststreng som input og udskriver teksten med hvert token annoteret med et COR-link.

Rent teknisk sker det ved at hvert teksttoken, som en indledende øvelse, annoteres med alle de links som er mulige i COR1. For eksempel har ’sky’ ikke mindre end ni alternative links (tjek selv i ordregister.dk). Når ’sky’ forekommer som token i en sammenhængende tekst, er

otte af de mulige links altså forkerte og skal fjernes. Algoritmen bag COR-linkeren kan beskrives som et loop, en stadigt gentagen analyse af teksten in toto. For hvert gennemløb fjernes ét link, og dermed ændres også beslutningsgrundlaget for den følgende iteration. Bliver man ved længe nok, står teksten til sidst tilbage, færdiglinket.

Sprognævnets COR-linker har fået navnet CLINK (af oplagte grunde), og programmet er p.t. i version 1.0. Denne version bruger tre forskellige strategier, kaldet LSYN (lokalsyntaktisk analyse), CTXT (kontekstanalyse) og FREQ (frekvens). Hver af de tre strategier er implementeret som et selvstændigt modul. LSYN ræsonnerer efter klassiske syntaktiske principper (som i afsnit 6) hvorimod CTXT benytter semantiske associationer mellem lemmaer uden at skele til bøjningen. FREQ foretager bare simple opslag i en frekvensordbog. De tre moduler er seriekoblet, sådan at det ene moduls output er input til det næste. Man kan eksperimentere med at ændre på deres rækkefølge, og desuden kan man skyde nye moduler ind. For tiden arbejder en af Sprognævnets tekniske assistenter på at træne et neuralt net (baseret på *TensorFlow*) som også kan indgå som modul. CLINK-moduler er hyggelige at programmere, og vi opfordrer alle til at bidrage. Med hvert nyt, velskrevet modul bliver CLINK stærkere, til gavn for os alle. Læs mere om CLINK i ordregister.dk og referencer dér.

8. Det nye økosystem af danske sprogressourcer

Det Centrale Ordregister har været under udarbejdelse siden april 2020 (på en statslig bevilling administreret af Digitaliseringsstyrelsen); men selve projektideen er af langt ældre dato. Gennem de seneste 20 år er det blevet stadigt tydeligere at Danmark er et svært sted at overleve som NLP-udvikler (Pedersen et al. 2011, Kirchmeier et al. 2019). Vores sprogsamfund er lille, og vores talesprog er vanskeligt at processere som data – men det kan vi ikke gøre noget ved. Hvad vi *kan* gøre noget ved, er tilgængeligheden af grundlæggende sprogressourcer for det danske sprog. Staten har, ind til 2020, stort set ikke bidraget til at sørge for gode og frit tilgængelige sprogressourcer, sådan som det er sket i vores nabolande. Hver NLP-virksomhed har måttet skabe sine egne sprogdata – maskinlæsbare ordbøger, tekstkorpora, talematerialer, transskriptioner og sprogmodeller. Det kan ikke undre at et firma som har ofret tusinder af arbejdstimer på at skabe et talekorpus for at kunne træne en talegenkender, ikke føler trang til at dele sit korpus med andre. Derfor har danske taledata til brug for NLP stort set været låst inde i firmaerne i årtier. Hvilke chancer giver det et nyt lille firma for at teste en god ide til et sprogprodukt og føre det til marked? Ikke mange. Det er for dyrt at komme gang når man begynder uden data. Som følge heraf har udviklingen af ny sprogteknologi til det danske sprog været begrænset til mellemstore danske virksomheder og internationale kæmper. Sverige har i

modsætning til Danmark i snesevis af mikrofirmaer med egen NLP-udvikling der drager nytte af en række offentlige sprogressourcer, ligesom Norge, Finland, Island og Færøerne (Kirchmeier et al. 2019; Simonsen et al. 2022).

Men nu er der grøde i luften herhjemme. Staten bevilgede i 2020 omkring 40 mio. kroner til udvikling af danske sprogressourcer. Nye materialer til NLP er under udarbejdelse og bliver løbende frigivet open-source. Hold øje med Digitaliseringsstyrelsens hjemmeside www.sprogteknologi.dk, hvor historien om Statens inititativ bliver fortalt og de nye sprogressourcer præsenteret.

Vi i COR-projektet (DSN, DSL og CST) mener at man som dansker har ret til sit sprog, også som materiale til NLP. Vi ser Det Centrale Ordregister som vores bidrag til den nye demokratisering og opfordrer alle korpusejere og ordbogsredaktioner til at gøre deres data frit tilgængelige – cum COR.

Litteratur

Bilgram, Thomas; Hans Arndt (1993) Automatisk morfologisk analyse af danske tekster. I Mette Kunøe og Erik Vive Larsen (red.). *4. Møde om Udforskning af Dansk Sprog*, Aarhus Universitet.

Bilgram, Thomas; Britt Keson (1998) The Construction of a Danish Tagged Corpus. I *NODALIDA '98 Proceedings*.

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

Diderichsen, Christina; Peter Juel Henrichsen; Thomas Widmann. (2022) Det Centrale Ordregister. I *Nyt Fra Sprognævnet*. Oktober 2022.

Henrichsen, Peter Juel (2023) Det Centrale Ordregister – et indeks for det danske ordforråd – en gave til dansk sprogteknologi. *Proceedings NFL-2022*, Lund Universitet.

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

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*.

Kirchmeier, Sabine; Bolette Sandford Petersen; Peter Juel Henrichsen; Sanni Nimb; Philip Diderichsen (2020) World Class Language Technology – Developing a Language Technology Strategy for Danish. *Proceed. 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*.

Pedersen, Bolette Sandford; Wedekind, J.; Bøhm-Andersen, S.; Hoffensets-Andersen, S.; Juel Henrichsen, P.; Kirchmeier-Andersen, S.; Kjærum, J. O.; Larsen, L. B., Nimb, S., Rasmussen, J., Revsbech, P. & Thomsen, H. (2011) Languages in the European Information Society – Danish. *META-NET White Paper Series*.

Budapest, Ungarn. (37 sider).

Jervelund, Anita, Schack, Jørgen, Jensen, Jørgen Nørby & Andersen, Margrethe Heidemann (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. *Proceedings of LREC2022*.

Widmann, Thomas (2022) Det Centrale Ordregister og dets leksikografiske anvendelser. *Proceedings of NLF 2022*.

Link

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 (2023) Det Danske Sprog og Litteraturselskabs' præsentation af PAROLE-projektet (såvel de pan-europæiske som de specifikt danske aktiviteter): https://korpus.dsl.dk/resources/details/parole.html.

Appendiks

1. Tabel for bøjningsindeks og PoS (komplet tabel i ordregister.dk)

Bøjningsindeks (COR1)	**Part-of-speech (Retskrivningsordbogen)**
100	sb.sg.ubest
102	sb.pl.ubest
110	sb.fk.sg.ubest
111	sb.fk.sg.best
112	sb.fk.pl.ubest
114	sb.fk.sg.ubest.gen
120	sb.itk.sg.ubest
122	sb.itk.pl.ubest
201	vb.inf.pass
203	vb.præs.akt
204	vb.præs.pass
214	vb.perf.part.sg.best

300	adj.sg.ubest.fk
302	adj.sg.best.
303	adj.pl
500	prop
600	talord
880	præp
900	adv
910	art.best.sg.fk
915	art.ubest.sg.fk
970	konj

2. PAROLE med COR-links (første 22 tokens)

To/COR.01528.600.01

kendte/COR.18159.303.01

russiske/COR.22261.303.01

historikere/COR.58774.112.01

Andronik/COR.PARC.110001.500

Mirganjan/COR.PARC.110002.500

og/COR.00099.970.01

Igor/COR.PARC.110003.500

Klamkin/COR.PARC.110004.500

tror/COR.30425.203.01

ikke/COR.10161.900.01

,/COR.XP.2.01

at/COR.00145.970.01

Rusland/COR.05764.500.01

kan/COR.30580.203.01

udvikles/COR.30794.201.01

uden/COR.00053.880.01

en/COR.00831.915.01

“/COR.XP.8.04

jernnæve/COR.PARC.100002.110

”/COR.XP.9.04

./COR.XP.1.01


COR
DET CENTRALE ORDREGISTER