Ny forskning i grammatik, vol. 31 (2024), 53-70.
DOI 10.7146/nfg.v1i31.145292

Tekstordet som grammatisk domæne

Peter Juel Henrichsen

Abstract
In this paper, we study the grammar of the individual tokens of ordinary Danish text. We use the formal framework CLINK, developed at the Danish Language Council (DSN) for use in language technology. Within the CLINK environment, all morphs (text segments with individual semantics) are lexicalized, including not only traditional word forms, but affixes, multi-word expressions, punctuation marks, and more. The grammar rules involved are purely abstract, viz. those of the Lambek calculus (categorial grammar). The topics discussed in this paper are spinoff of DSN's work with COR (det Centrale Ordregister).

Nøgleord
morfologi, kategorialgrammatik, tekstanalyse, Det Centrale Ordregister, CLINK

1. Indledning

Denne artikel beskriver DSN's nye værktøj til automatisk tekstannotation, applikationen CLINK. Skønt CLINK især er rettet mod sprogteknologiske anvendelser, har udviklingsprojektet også kastet ny viden af sig om dansk ortografi og morfologi. Denne artikel fokuserer på projektets lingvistiske sider, især inden for morfologisk analyse, og her spiller det Centrale Ordregister (COR) hovedrollen som leksikalsk basis. Artiklen begynder med en kort introduktion til COR og kategorialgrammatik (parserens formelle grundlag) fulgt af hovedafsnittene om CLINK, og til slut nogle bemærkninger om sprogteknologien i sparring med ortografien. Fodnoter kan springes over uden skade for sammenhængen.

I de sproglige eksempler man møder i lingvistiske artikler, fremstår ordene som regel i deres leksikalske nøgenhed, uden apostroffer, forkortelser, parenteser, citationstegn, streger og prikker, uden versal på første ord, og

uden punktum efter sidste. Hverdagens tekster, derimod, er et konglomerat af alle tastaturets tegn. For at kunne analysere tekst maskinelt må tekstordene først transformeres til leksemer, i en proces som sprogteknologien kalder 'tokenizing'. Dette kan gøres mere eller mindre principfast, men mange tokenizere lider under en mangel på præcis orddefinition.

Vi har valgt en simplistisk definition af 'tekstordet' som en følge af teksttegn afgrænset af mellemrumstegn. For eksempel består teksten *ingen røg* uden brand af tekstordene [ingen], [røg], [uden] og [brand] mens *Ingen (af- og) pålæsning!* består af tekstordene [Ingen], [(af-], [og)] og [pålæsning!]. Denne definition er valgt især fordi den er enkel at implementere; i programmeringssproget Perl (populært blandt datalingvister) kan man for eksempel dele et tekstdokument op i tekstord ved hjælp af bare én kommando:

(1) `@tekstord = $tekstdokument =~ /(\S+)/g`

I øvrigt er definitionen ikke uden kognitiv relevans. Også vi læsere skanderer jo i første omgang en tekst i tegnblokke adskilt af mellemrum.

2. Det Centrale Ordregister

Retskrivningsordbogen (RO) har en særlig betydning for dansk skriftsprog. Den definerer vores ortografi, ikke kun for de leksemer den dækker eksplicit, men også for tegnsætning og for regelmæssigt dannede komposita. Den er en nødvendig ressource for skolevæsen, offentlig administration, forfattere, forlag og, ikke mindst, sprogteknologi. Sidstnævnte er afhængig af RO som input til stave- og grammatikkontrol, automatisk oversættelse, men også – måske overraskende – til taleteknologi. Mere herom senere.

På opfordring fra de sprogteknologiske brancher (Kirchmeier et al. 2019) har Sprognævnet udarbejdet en maskinlæsbar version af RO kaldet Det Centrale Ordregister (COR). COR-registeret har en procedure til at forsyne ethvert dansk lemma med et unikt id der, ligesom CPR-nummeret, fungerer som en eksakt henvisning til en entitet med et kompleks af egenskaber, som ellers kan være vanskelige at referere til under ét. COR-registerets leksikalske grundressource COR_1 omfatter alle Retskrivningsordbogens lemmaer som fulde ordformer (godt 530.000 former). Hvert leksem har et COR-id i denne form:

(2) **COR.** <*lemmaindeks*> **.** <*bøjningsindeks*> **.** <*variantindeks*>

I denne artikel anvendes forkortelserne LIX, BIX og VIX for de tre indekser (bestående af hhv. 5, 3 og 2 cifre).[1] Databasen COR_1 og dens BIX-tabel (afbildning af BIX-værdier på ordklasse og bøjning) kan downloades fra ordregister.dk. De tre tekstcitater (3) er fra korpus PAROLE (Keson 1999, Henrichsen 2023), her annoteret med leksikalske henvisninger ('COR-links').

(3a)	(3b)	(3c)
retten / `COR.43153.111`	**retten** / `COR.43157.111`	han / `COR.01880.991`
frikassé / `COR.53868.110`	til / `COR.00947.880`	**samlede** / `COR.30151.206`
af / `COR.00014.880`	den / `COR.00267.910`	et / `COR.00831.916`
gedde / `COR.69689.110`	**samlede** / `COR.30151.214`	års / `COR.30151.206`
og / `COR.00099.970`	flåde / `COR.60753.110`	**junkmail** / `?`
ørred / `COR.64233.110`		

Bemærk at de to forekomster af *retten* har forskellig LIX, men samme BIX (forskellige lemmaer, samme bøjning) mens det omvendte gælder *samlede*. Links til tekstord der ikke forekommer i COR_1, kræver særlig opmærksomhed (fx *junkmail*); mere om det herunder. Som det fremgår, er tekster annoteret med COR-links fri for homografi (modulo COR_1), hvad der i mange sammenhænge gør dem lettere at databehandle både for mennesker og maskiner.

COR_1 er, som allerede nævnt, Det Centrale Ordregister's leksikalske grundniveau; der findes også andre COR-ordbøger med fx ordsemantik (Nimb et al 2022), udtaleinformation (Henrichsen 2024), historisk ortografi (Widmann 2023) og – særlig relevant for denne artikel – specialordbøger med interpunktionstegn, affikser, fuger, numeraler m.v.

3. Kategorialgrammatik i morfologiens tjeneste

I dette afsnit gennemgås kategorialgrammatikkens grundtræk, med særligt fokus på den såkaldte 'sekventanalyse' som spiller en central rolle i CLINK. I (4a) og (4b) ses to komposita, *antiwoke* og *udrikkelighed*, analyseret som sekventer.

1 VIX bruges bl.a. til ortografiske varianter såsom *ressource/ressurse* (hhv. **`COR.95344.110.01`**, **`COR.95344.110.02`**) og *af sted/afsted* (**`COR.13533.900.01`**, **`COR.13533.900.02`**). VIX udelades ofte i denne artikel.

(4a)

A/A	A	==>	A
[anti]	[woke]		[antiwoke]

(4b)

A/A	V	V\A	A\N	==>	N
[u]	[drikke]	[lig]	[hed]		[udrikkelighed]

En 'sekvent' består af to dele, 'antecedenten' til venstre for 'sekventpilen' (==>), 'konsekventen' til højre. Stammen [*woke*] har kategorien A (for adjektiv) mens affikset [*anti*] har en sammensat kategori (A/A). Skråstregen / bruges til funktioner der søger argument til højre, bagstregen \ når argumentet søges til venstre. Funktioner af type *x**y* og *y*/x* siges at afbilde et *x* på et *y* (*mapping x to y*). Sekventanalyse består i at bevise at kategorien i konsekventen kan dannes af kategorierne i antecedenten gennem lutter lovlige 'applikationer' (*applications*), altså skridt hvor en funktion får det argument den beder om. I (4b) må *drikkelig* komponeres før *u-* kan affigeres, og *-hed* til sidst.

Morfologiske flertydigheder får hver sit bevis. Det illustrerer vi her med en lille anekdote. For nylig fortalte forfatterens datter om sin nystiftede studenterkontorklub, og faderen troede det drejede sig om en kontorklub (kun) for studenter. Men nej, klubbens navn skyldes at den holder til i studenterkontoret, og den er således også åben for seniorer[2]. Misforståelsen kan forklares med sekventeringen i (5) af navnet *studenterkontorklub*.

(5)

N	N\N/N	N	N\N/N	N	==>	N
[student]	[er]	[kontor]	[]	[klub]		[studenterkontorklub]

Hvis fugen [er] tages som udgangspunkt for delsekventen N N\N/N N ==> N, får man først komponeret *studenterkontor* og derefter tilføjet *klub*. Med fugen [] som basis komponeres *kontorklub* først.

Konstituenter med særlig grammatisk affinitet vil man ofte vælge at leksikalisere en bloc. For eksempel kan [student] og [er] komponeres vha. delsekventen N N\N/N ==> N/N og leksikaliseres som en sammensætningsform (*studenter* i kategorien N/N). Denne teknik kaldes 'partial deduction'. Retskrivningsordbogen har ca. 8000 sammensætningsformer af denne type. I tilfældet *studenter-* er det en fordel at den sjældne fuge [er] på denne måde bliver knyttet rent leksikalsk til de leksemer der skal bruge den, så den ikke er åben for orddannelse generelt. Bemærk

2 Faderen bades holde sig væk alligevel.

i (6) at den strukturelle tvetydighed stadig er bevaret efter den partielle deduktion.

(6)

N/N	N	N\N/N	N	==>	N
[studenter]	[kontor]	[]	[klub]		[studenterkontorklub]

Hvert morfem i et sammensat ord bidrager til ordets sekventering med en eller flere kategorier. En stamme bidrager med én grundlæggende kategori mens alle andre morfemer er funktioner: præfikser, suffikser, infikser, fuger, osv. Man kan udtrykke det sådan at hvert morfems kategori deklarerer morfemets bidrag til hele ordets struktur. Selve bevisførelsen hviler på såkaldte 'sekventregler'; det fører for vidt at forklare dem her, men der findes gode introduktioner til kategorialgrammatik (fx Wood 1993 for almene lingvister; Morill 1994 for formelle lingvister).[3]

I et kategorialgrammatisk perspektiv er de traditionelle termer præfiks og suffiks aliaser for henholdsvis *y*/*x* og *x**y* (hvor *x* og *y* er kategorier hver for sig) mens fugen, uanset sit materiale, er alias for *y**x*/*x*. I dette afsnit bruger vi {V,N,A} som grundlæggende kategorier og har altså ni forskellige fugekategorier til rådighed for orddannelse. Alle ni er rigt repræsenteret i moderne dansk (Henrichsen 2021), med V\V/V som den sjældneste (*løbetræne*, *tørretumble*, *sultestrejke*); se tabel 1.

Morfem	**Kategorial funktion**	**Eksempler** leksemer og sammensætninger	
stamme	grundlæggende kategori (svarende til en ordklasse)	[gul] : A [drikke] : V [ven] : N	-
suffiks	*x**y* (funktion fra kategori *x* til *y*)	[skab] : N\N [lig] : V\A	[ven][skab] [drikke][lig]
præfiks	*y*/*x* (funktion fra *x* til *y*)	[over] : A/A [for] : V/V	[over][klog] [for][blænde]
fuge	*x**y*/*y* (funktion fra *x* og *y* til *y*)	[] : A\V/V [e] : N\N/N [s] : N\A/A	[fin][][pudse] [barn][e][vogn] [mand][s][høj]

Tabel 1. Udvalg af danske funktionsmorfemer (tabel tillempet fra Henrichsen 2021)

3 Lambek-kalkylen har én aksiomregel: **A ==> A**, samt fire bevisregler for operatorerne **/** og **** (kaldet 'divisorer'), **L$_0$ A/B M$_1$ N$_0$ ==> C IF M$_1$ ==> B AND L$_0$ A N$_0$ ==> C** og **L$_1$ ==> A/B IF L$_1$ B ==> A**, samt **L$_0$ M$_1$ B\A N$_0$ ==> C IF M$_1$ ==> B AND L$_0$ A N$_0$ ==> C** og **L$_1$ ==> B\A IF B L$_1$ ==> A**, hvor **A, B, C** er vilkårlige kategorier, og **L$_n$, M$_n$, N$_n$** er lister af **n+** vilkårlige kategorier. Se detaljer i Wood (1993).

En fordel ved kategorialgrammatikkens koncise definitioner er at man i høj grad undgår dilemmaet mellem at udvikle en grammatisk nomenklatur afledt af sit eget sprog og at leve med one size fits all-termer fra et dominerende oversprog. I COR-sammenhæng er et sproguafhængigt regelsystem særlig vigtigt, alene fordi mange af COR_1's kernebrugere (sprogteknologerne) hverken har lingvistisk baggrund eller dansk som førstesprog.

4. COR møder kategorialgrammatikken

Vi bruger termen 'morf' for et tekstelement som i en given analytisk sammenhæng er atomisk, altså ikke kan opdeles uden at miste sin leksikalske afbildning. Eksempler på morfer: [hest], [km/t.], [7], [!], [(..)], men ikke [he], [st], [km/t] og [(..]. Herfra erstattes termerne substantiv, verbum og adjektiv (S, V og A) med BIX-indekser. Således er **123** kategori for sb.itk.pl.best (substantiv, intetkøn, pluralis, bestemt form). Også delvist instantierede kategorier forekommer, med punktum som vikar for et ciffer; fx står **1..** for en vilkårlig substantivform (sb), **12.** for en neutrumsform (sb.itk) og **1.3** for en bestemt flertalsform (sb.pl.best). Den helt uspecificerede kategori noteres med et variabelsymbol (***X***, ***Y***, ***Z***, **...**), som illustreret i (7).

(7)

SEGMENTER:	[studenter]	[kontor]	[]	[klubben]	
SEKVENT:	***Z/Z***	**129**	***Y\X/X***	**111**	**==> 111**

Valget af variabelsymboler er arbitrært. Fx kan ***X*** frit erstattes med ***W***, forudsat at ***W*** ikke forekommer i forvejen (proceduren kaldes 'alfakonversion'). Sekventen i (7) finder to beviser, nemlig for hhv. ***Z*=129** (*studenterkontor*|*klubben*)[4] og ***Z*=111** (*studenter*|*kontorklubben*). Bemærk at præfikskategorien er blevet generaliseret i forhold til kategorien N/N i afsnit 2. Dermed finder fx adjektivet *studenterløs* nu også en analyse (for ***Z*=300**, der er BIX for adjektiver som *løs*).

4.1. Segmentering af sammensatte tekstord

RO rummer en stor mængde hyppigt forekommende komposita, hvoraf langt hovedparten er regelmæssigt dannede (både morfologisk og se-

4 Kategorier med formen **1.9** svarer til RO's sammensætningsformer (sb.sms)

mantisk) og dermed i en vis forstand redundante. Men som bekendt har de mest benyttede komposita en tendens til at drive væk fra deres udgangspunkt mht. betydning og udtale (*hulvej*, *livmoder*), og det gør en vis overrepræsentation nødvendig. Som konsekvens står den praktisk virkende morfolog ofte over for et valg mellem flere segmenteringer. Tag for eksempel tekstordet *antifødselsdagssang-revolution*, der forekom i Politiken's stribe Ting Jeg Gjorde af Maren Uthaug (28/2 2024). I sin kontekst var ordets hensigt klar, nemlig at afsyngning af fødselsdagssange til voksne er en uskik og bør bekæmpes ved revolution. Segmentet [revolution] har altså skopus over resten af ordet, mens [anti] har skopus over [fødselsdagssang].

(8)

SEG:	[anti]	[fødselsdagssang]	[-]	[revolution]	
SEK:	***Z/Z***	**119**	***Y\X/X***	**110**	**==> 110**

Det viser sig imidlertid at COR_1 ikke har et opslag for *fødselsdagssang*. Hvis sekventanalysen skal dækkes ind af COR_1-ordbogen, må det lange token deles i mindst seks segmenter.

(9)

SEG:	[anti]	[fødselsdag]	[s]	[sang]	[-]	[revolution]	
SEK:	***Z/Z***	**119**	***V\W/W***	**119**	***Y\X/X***	**110**	**==> 110**

Underdelingen kan naturligvis fortsættes.

(10a)

SEG:	[fødsel]	[s]	[dag]	
SEK:	**119**	***Y\X/X***	**119**	**==> 119**

(10b)

SEG:	[fød]	[sel]	[s]	[dag]	
SEK:	**200**	**200\119**	***Y\X/X***	**119**	**==> 119**

Somme tider kan man argumentere for én segmentering frem for en anden med henvisning til den almindelige sprogfornemmelse eller til de mest udbredte ordbøger; men oftest afhænger svaret mere af det sprog-

teknologiske formål[5]. I alle tilfælde vokser det morfologiske analyserum med antallet af segmenter (se tabel 2).

Segmenter	**Analyser**
[anti][fødselsdagssang][-][revolution]	2
[anti][fødselsdag][s][sang][-][revolution]	5
[anti][fødsel][s][dag][s][sang][-][revolution]	14

Tabel 2. Sammenhæng mellem segmentering og analyserum

Jo større dele af et kompositum der genkendes, des mindre flertydighed står tilbage. Et fyldigt kompositum som *medarbejdertilfredshedsundersøgelse* er nemt at tolke på grund af sine velkendte komponenter, modsat det ligedannede *udarbejderafpasningsoverfindelse*. Leksikalsk information spiller en afgørende rolle for disambiguering, både for mennesker og computere, men den overordnede diskurs spiller også en vigtig rolle. Vi vender tilbage til CLINK's parsestrategi og til Uthaugs stribe i afsnit 6.

5. Den leksikalske kategori som nøgle til semantisk komposition

Det er et karakteristisk træk ved kategorialgrammatikken at konstituentanalyse (morfologi, syntaks) og semantisk analyse går hånd i hånd. Denne isomorfi er også et hovedprincip i CLINK.

For leksemer med et COR_1-id bruges lemmaindekset som stedfortræder for lemmaets semantik. Adjektivet *grøn* har fx lemmaindekset **15974**, så i CLINK-sammenhæng er egenskaben 'grøn' repræsenteret med symbolet **15974**. Skulle nogen engang i fremtiden ønske at indføre en maskinlæsbar forklaring på farvekvaliteten *grøn* (hvordan den end kommer til at se ud), kan denne forklaring så træde i stedet for sin proxy. Indtil da nøjes vi med at ræsonnere med LIX-værdier som stedfortrædere for betydning.
Morfer der ikke har en selvstændig betydning, men i stedet er funktioner af andre morfers betydning, kaldes somme tider 'synkategorematiske' i den semantiske litteratur. I CLINK kendes en funktionsmorf (fx affiks, fuge, bindestreg, interpunktionstegn) på at dens kategori har mindst én skråstreg. En sammensat kategori modsvares af en semantisk komposition, en såkaldt 'lambdafunktion'.

5 For eksempel behøver en maskinoversættelse en finere segmentering end en talesyntese. Ordet *antifødselsdagssangrevolution* kan læses op på samme måde uafhængigt af sin tolkning.

(11)

SEG:	[anti]	[woke]		
SEK:	**X/X**	**300**	**==>**	**300**
SEM:	**λa.a(02854)**	**27102**		**27102(02854)**

En lambdafunktion består af et hoved med et antal variabler (fx **λabc**) og en krop som de samme variabler indgår i. Funktionens argumenter indsættes i kroppen i den rækkefølge hovedet angiver. Når variablerne er brugt op, er formlen ikke længere en funktion, men et afsluttet semantisk udtryk der viser delenes relationer. Til *antiwoke* (funktionen **λa.a(02854)** for [anti], med argumentet **27102** for [woke]) svarer således udtrykket **27102(02854)**, der kan parafraseres som: adjektivet *woke* modificeret af præfikset *anti*.

RO's præfikser (fx *anti*, *ærke*, *skide*) og sammensætningsformer (fx *rigs*, *lamme*, *lande*, *lands*, *land*) er i CLINK kategoriseret som **X/X**, med lambdaudtrykket **λa.a(*LIX*)**, hvor ***LIX*** er lemmaindekset i COR_1 (fx **02854** for *anti*).

5.1. Det sammensatte farvenavn – præfiks versus fuge

De fleste sammensatte farvetermer følger et fast semantisk mønster, nemlig at det sidste led er hovedfarven og det første led en modifikation. Eksempler fra DDO ses i (12).

(12a) lysegul "af en lys gul farve"
(12b) karrygul "gul eller gulbrun som karry"
(12c) rødgul "af en gul farve med en orange tone"
(12d) orangerød "af en kraftig rød farve med en orange tone"
(12e) blegrød "af en svagt rød eller lyserød farve"
(12f) blårød "af rød farve med en blålig tone"

DDO forklarer desuden at formen *lys*+FARVE (*lysgul*, *lysrød*) også forekommer, dog "især fagligt". I RO finder man *lysegul/-blå/-rød/-brun/-grøn* og *-grå*, og dertil *lyslilla* og *lysviolet*; desuden sammensætningsformerne *lys-*, *lyse-* og *mørke-*, og dermed kan også *lysgult*, *mørkeviolet* osv. analyseres som præfiks+adjektiv.

(13)

SEG:	[lys]	[gult]		[lysgult]
SEK:	**X/X**	**301**	**==>**	**301**
SEM:	**λa.a(40871)**	**15978**		**15978(40871)**

Sammensatte farvenavne hvis førsteled ikke er et præfiks, men fx et adjektiv eller et substantiv, samles i stedet med en fuge (oftest [], [e] eller [s]).[6]

(14)

SEG:	[blå]	[]	[rødt]		[blårødt]
SEK:	**300**	**Y\(X/X)**	**301**	**==>**	**301**
SEM:	**15302**	**λab.b(a)**	**15892**		**15892(15302)**

Som nævnt findes sammensætningsformerne *lys-*, *lyse-* og *mørke-* i COR_1 – men ikke *mørk-*. Med en nulfuge kan fx *mørkviolet* komponeres af [mørk]:**COR.15581** og [violet]:**COR.24305.301**. Bemærk at BIX er irrelevant i alle andre kompositumled end det sidste; den danske morfologi bøjer jo komposita efter kun det sidste led. Det er afspejlet i fugens kategori ved at venstreargumentet er uspecificeret, dvs. at de to argumenter ikke behøver at have samme kategori (både **X=Y** og **X≠Y** er tilladt). Nu kan vi forsyne de to sekventeringer af *studenterkontorklubben* fra afsn. 3 med semantisk annotation.

(15a)

SEG:	[student]	[er]	[kontor]	[]	[klubben]	
SEK:	**119**	**V\(W/W)**	**129**	**Y\(X/X)**	**111**	**==> 111**
SEM:	**96483**	**λab.b(a)**	**48871**	**λcd.d(c)**	**47998**	

(15b)

SEG:	[studenter]	[kontor]	[]	[klubben]	
SEK:	**Z/Z**	**129**	**Y\(X/X)**	**111**	**==> 111**
SEM:	**λa.a(96483)**		**48871**	**λcd.d(c)**	**47998**

5.2. Rødhvide farver, vi skal ud at slås

Visse sammensatte farvetermer er, modsat *karrygul* og *blårød*, ikke seman-

6 Bemærk at der er føjet parenteser til fugens kategori **Y\(X/X)** for at vise at venstreleddet skal appliceres først da det leverer lambdafunktionens første argument (**15892 → a**).

tisk velbeskrevne som et møde mellem en grundbetydning og en moderator.

(16a) rødhvid ”ofte som symbol på den danske nation, fx i sportsbeklædning”
(16b) blågul ”blå og gul; fx om sportsbeklædning”
(16c) rødgrøn ”rød og grøn; som vedrører rød og grøn”

Ordforklaringerne i (16) er fra DDO. Der er lignende eksempler på komposita med ligestillede dele i mange andre ordklasser.

(17a) sursød ”Hjemmetlavet gammeldags sursød sovs”
(17b) majjuni ”I Danmark får den to kuld i majjuni og august”
(17c) fredaglørdage ”at score kassen på fire fredaglørdage med underholdning”
(17d) totre ”hvid skjorte, sorte bukser og totre dage gamle skægstubbe”
(17e) f3 ”hvid springer g1 til f3 kaldes Rétiåbningen”
(17f) spaghettikødsovs ”danskerne har fortjent deres spaghettikødsovs”

I den klassiske stilistik kaldes formerne i (17) for ‘dvandvasammensætninger’ (fx Albeck 1968: 79-80); de er karakteristiske ved at bestå af dele der er semantisk sideordnede, enten konjugerede (sursød = sur og sød), disjugerede (majjuni = maj eller juni) eller ubestemte mellem de to (fumlertumler = barn der fumler og/eller tumler).[7] Dvandvakomposita har brug for en ny semantisk operator.

(18)

SEG:	[sur]	[]	[sød]		
SEK:	**300**	***X*\(*X*/*X*)**	**300**	**==>**	**300**
SEM:	**27204**	**λab.(a,b)**	**15282**		**(27204,15282)**

De to sideordnede led noteres som et par, **(a,b)**.

5.3. Fugen og de tre hængsler

Dansk morfologi har tre produktive fuger, [], [e], [s] (i ortografien suppleret af [-] og enkelte andre) mens de øvrige sjældent forekommer i fri form. I den klassiske morfologi har fugen [] som bekendt det mistænkelige navn

7 Der er en tendens til at dvandvaformer skrives med bindestreg; fx har RO og DDO bindestreger i (16)'s leksemer. Dette er dog ikke et gennemført ortografisk princip, og formerne i (16) og (17) er også hyppigt forekommende. Eksemplerne i 17 er googlet.

'nulfuge'. Det er en uheldig term. Nok er fugen usynlig i skriftbilledet, men ikke uhørlig i udtalen og altså ikke *nul* i lingvistisk forstand. I kompositummet *flagbajer* med segmenterne [flag][][bajer] fungerer det midterste segment som en fonetisk operator der dels selekterer for udtalen /flAw/ (frem for den leksikalske udtale /fla:?/), dels styrer tryktabet i segmentet [bajer]. Regelmæssige udtaleændringer af denne type er vores ortografi dog for fattig til at repræsentere. Alle morfologiske fuger, inklusive [], kan derfor anses for materielle (i betydningen 'sansbare').

Vi bruger termen 'hængsel' for fugens semantik. Som omtalt i sidste afsnit kan to forskellige fuger godt have det samme hængsel (fx i *lysgul* og *mørkelilla*), og tilsvarende kan to forskellige hængsler være knyttet til den samme fuge (fx i *rødgul* og *rødhvid*). Hængslet λ**ab.b(a)** kaldes 'prædikationshængslet'; det samler sine argumenter i en underordningsrelation. Det andet hængsel λ**ab.(a,b)** sideordner argumenterne og kaldes 'dvandvahængslet'. Der er også brug for et tredje hængsel, nemlig til sammensatte tekstord hvis segmenter slet ikke indgår en morfologisk relation, som i for eksempel teksten *(højt skum)* mellem segmenterne [(] og [højt], samt [skum] og [)]. Da vi har defineret tekstordet som en gruppe af grafemer uden at skelne mellem bogstaver, cifre og andre tegn, er funktionstegnene nu kommet inden for den morfologiske analyses horisont – hvad der afslører en ny type af flertydighed i dansk ortografi.

(19a) Det var d. 19. feb. ph.d. T.H. Andersen betalte 1.200 kr. inkl. moms.

(19b) DSB overvejer Alstoms letbane-platform Citadis til diesel el. el.

(19c) Hun er 5 år, han er 6. Hun bor på 7. sal, han bor på 8.

De fleste danske funktionstegn, især punktum, komma og bindestreg, har mere end én ortografisk funktion, normalt uden at det giver anledning til misforståelser (jf. 19a). I (19b) forekommer tekstordet *el.* i to strukturelt forskellige varianter, dels som forkortelse (et leksem) og dels som to samskrevne morfer (et leksem plus et slutpunktum med sætningsskopus). I sidstnævnte tilfælde er der ingen morfologisk forbindelse mellem segmenterne, og det kalder på et nyt hængsel som ikke relaterer sine argumenter til hinanden, men bare sætter dem på listeform [*x*,*y*,*z*, ..]. Dette tredje hængsel, 'listehængslet', har kategorien **Y\((Y*X)/X)** med den semantiske form λab.**[a,b]**. Bemærk at dets kategori har samme ydre struktur

som de andre hængsler, med et venstre- og et højreargument (hhv. ***Y*** og ***X***), mens den indre funktor (***Y*X***) afspejler at de to kategorier ikke kombinerer.[8]

I (19c) forekommer tekstordene *6.*, *7.* og *8.*. Af konteksten fremgår det at de tre numeraler skal læses som hhv. seks, syvende og ottende.

(20a)

SEG:	[6]	[]	[.]		
SEK:	**601**	***Y*\((*Y**X*)/*X*)**	**x1**	**==>**	**601*x1**
SEM:	**NUM(6)**	**λab.[a,b]**	**f1**		**[NUM(6),f1]**

(20b)

SEG:	[7]	[.]		
SEK:	**601**	**601\602**	**==>**	**602**
SEM:	**NUM(7)**	**λa.ORDINAL(a)**		**ORDINAL(NUM(7))**

I (20b) tager punktumtegnet rollen som ordinalsuffiks, dvs. funktionen fra **601** (kardinaltal) til **602** (ordinaltal). I modsætning hertil er (20a) ikke en morfologisk komposition, men to uafhængige morfer og derfor forbundet med listehængslet. Man kan nu spørge om listehængslet kun er relevant for tekstord med funktionstegn, altså kun for rent ortografiske fænomener? Ikke nødvendigvis. Der findes som bekendt komposita med orddele der egentlig ikke indgår en morfologisk relation, men snarere minder om sætningskonstituenter, ord som *omsiggribende*, *italesætte*, *gudskelov*, *ovenikøbet* og *uladsiggørlig*. Hvis CLINK møder et sådant kompositum (og ikke har det leksikaliseret), kan en listehængselanalyse komme på tale. Dette emne, som også har bredere leksikologisk interesse, behandles dog bedst som del af en generel undersøgelse af grænsefladen mellem morfologi og sætningssyntaks. Herom i en senere publikation.

Med tekstordet *8.* indtræffer en ortografisk krise. Ordet læses tydeligvis *ottende* i (19c), altså med punktummet som ordinalsuffiks; men ortografien forlanger et slutpunktum. Med andre ord, punktumtegnet er 'overlastet' med to funktioner hvoraf ingen kan undværes. For den almindelige læser er det måske en skønhedsplet, men for grammatikeren (og CLINK)

8 Operatoren ***** betegnes som Lambek-kalkylens 'produkt'. Der er to bevisregler for *****: $\mathbf{L}_0\ \boldsymbol{A}\mathbf{*}\boldsymbol{B}\ \mathbf{M}_0\ \mathbf{==>}\ \boldsymbol{C}$ IF $\mathbf{L}_0\ \boldsymbol{A}\ \boldsymbol{B}\ \mathbf{M}_0\ \mathbf{==>}\ \boldsymbol{C}$ og $\mathbf{L}_1\ \mathbf{M}_1\ \mathbf{==>}\ \boldsymbol{A}\mathbf{*}\boldsymbol{B}$ IF $\mathbf{L}_1\ \mathbf{==>}\ \boldsymbol{A}$ AND $\mathbf{M}_1\ \mathbf{==>}\ \boldsymbol{B}$, hvor ***A, B, C*** er vilkårlige kategorier, og $\mathbf{L}_n$, $\mathbf{M}_n$ er lister af ***n+*** vilkårlige kategorier.

bryder overlastningen med det grundlæggende princip om én morf, ét semantisk bidrag. Den danske ortografi har mindst et andet eksempel på overlastning, nemlig i sammenhænge som *hunden der hyler, katten og hønen*, hvor kommategnet nødvendigvis fungerer som både opremsnings- og slutkomma (Togeby 2000 har et lignende eksempel).

6. Automatisk tekstanalyse

Programmet CLINK ('COR-linker') kan beskrives som en input-output-automat der læser en tekst og annoterer hvert tekstord for kategori (BIX) og for semantik (λ-udtryk). Tabel 3 viser CLINK's output for teksten "Gud ske Lov for Sofahjørnets Fløjl!" (fra Johs.V. Jensens digt "Ved Frokosten"). Bemærk at kategorier og semantiske udtryk udskrives med et løbenummer (0, 1, 2, ..) når de er genereret fra en flerordsforbindelse i COR_1 (i dette tilfælde *gud ske lov* med id **`COR.14223.900.01`**). Udråbstegnets kategori og semantik er let forenklet i tabel 3 (hhv. **`x8`** og **`f8`**).

Input (tekstord)	**Segmenter**	**Kategori**	**Semantik**
"Gud"	[gud]	`900:0`	`14223:0`
"ske"	[ske]	`900:1`	`14223:1`
"Lov"	[lov]	`900:2`	`14223:2`
"for"	[for]	`880`	`00093`
"Sofahjørnets"	[sofa][][hjørnets]	`125`	`41222(77215)`
"Fløjl!"	[fløjl][][!]	`110*x8`	`[48705,f8]`

Tabel 3. CLINK's annotation af input "Gud ske Lov for Sofahjørnets Fløjl!"

For tekstord der findes i COR_1, er annotationen direkte kopieret fra leksemets COR-id; i øvrige tilfælde er annotationen genereret med kategorial analyse (som omtalt i afsn. 4 og 5).

6.1. CLINK's algoritme – et overblik

CLINK's grundlæggende datastruktur er en 'template', en liste af leksikalske oplysninger af forskellig art der karakteriserer en morf.

- overflade (grafemisk form m.m.)
- kategori (ordklasse, bøjning, morfologisk funktion)
- betydning (proxy eller λ-formel)
- udtale (behandles i en senere artikel)

Templates for alle de morfer der er leksikalsk tilgængelige (i COR_1 og evt. andre COR-ordbøger), er samlet i CLINK's 'templatebase'. Basens tekniske specifikation er fremlagt andetsteds (Henrichsen 2024, ordregister.dk). Når CLINK-programmet kaldes med en inputtekst eller et tekstkorpus, hentes først alle relevante leksikalske oplysninger i templatebasen, og derefter begynder tekstordsanalysen. Algoritmen kan beskrives som en iteration i fem niveauer.

1. Teksten deles op i tekstord (tegnfølger adskilt af blanktegn)
2. Hvert tekstord deles op i segmenter (morfer med hjemmel i COR_1 m.fl.)
3. Hver segmentering omsættes til sekventer (med kategorier fra templatebasen)
4. Hver sekvent forsøges bevist (med Lambek-kalkylens bevisregler)
5. Hvert bevis lambdareduceres (med λ-udtrykkene fra templatebasen)

For de enkleste tekstord (*hesten*, *er*, *gammel*) giver denne procedure kun én analyse. De fleste tekstord er dog flertydige, for det første hvis ordet kan segmenteres på flere måder (*5.*, *trækvinden*), for det andet hvis et segment er homograf i COR_1 (*sky*, *får*) og for det tredje hvis en sekvent har flere beviser (*studenterkontorklubben*). I tilfælde af flertydighed bruger CLINK en lang række forskellige heuristikker til at udpege en foretrukken analyse baseret på klassiske grammatikregler, kontekstanalyse og statistik (Henrichsen 2022). Output af analysen giver et godt udgangspunkt for fx stave- og grammatikkontrol, taleteknologi og maskinoversættelse – en computer finder BIX-værdier og lambdaformler lettere at læse end ord.

Og så er tiden kommet til at udsætte CLINK for Uthaugs *antifødselsdagssang-revolution*. CLINK finder fem forskellige betydninger (se tabel 4).

λ-udtryk	**Parafrase**
`¬( SR(F) )`	det modsatte af: sang-revolution ang. fødselsdag
`¬( R(FS) )`	det modsatte af: revolution ang. fødselsdagssang
`SR( ¬F )`	sang-revolution ang. det modsatte af fødselsdag
`R( ¬FS )`	revolution ang. det modsatte af fødselsdagssang
`R( S( ¬F ) )`	revolution ang. sang om det modsatte af fødselsdag

Tabel 4. Semantiske analyser af antifødselsdagssang-revolution. λ-udtrykkene er let forenklet (`F`, `S`, `R` *og* ¬ *erstatter LIX for hhv. fødselsdag, sang, revolution og anti)*

Som før nævnt er den rette tolkning utvivlsomt **`R(¬FS)`**. Det er tankevækkende at verdens førende oversættelsesmaskine, Google Translate, selv med adgang til hele stribens tekst, konsekvent vælger en forkert morfologisk analyse blandt disse fem, uanset om målsproget er sat til engelsk, tysk, hollandsk, fransk, spansk, italiensk, svensk eller norsk.

6.2. CLINK i offentlighedens tjeneste

Programmet CLINK er her præsenteret i sin standardkonfiguration, men kan også indstilles til at aflevere andre former for output efter brugerens ønsker. Man kan fx vælge kun at få tekstord annoteret med en kategori (svarende til en almindelig PoS-tagger) eller at få samtlige analyser udskrevet (i CLINK's prioriterede rækkefølge). Endelig kan programmet kaldes med enhver COR-ordbog som er formelt velformet (dvs. holder COR's regler for indeksering m.m., se ordregister.dk). I begyndelsen af 2025 udgiver DSN ordbogen COR.TALE med udtaleinformation for alle morfer i COR_1, særligt rettet mod talesyntese. COR.TALE vil gøre det muligt at anvende CLINK til lydskrivning af tekst.

Denne korte artikel kan kun give et første indtryk af CLINK, og vi efterlader uden tvivl en mængde spørgsmål. Hvordan leksikaliseres RO's flerordsforbindelser, som jo går udover tekstordets domæne (*af sted*, *pro et contra*)? Hvordan kategoriseres et bindeord (*og*, *men*, *til*) for at afspejle kravet om kongruens mellem argumenterne? Hvordan leksikaliseres opremsningskommaet? Slutkommaet? Startkommaet??? Hvordan undgår CLINK at se nulfuger overalt? Og sidst men ikke mindst: Hvilke selektionskriterier benytter CLINK-applikationen til flertydige tekstord (*5.*, *el.*, *sky*, *rødgul*, *antifødselsdagssang-revolution* osv.)? Disse og andre spørgsmål bliver diskuteret i DSN's løbende publikationer. I øvrigt kan man altid kontakte Sprognævnet på pjh@dsn.dk (om CLINK), tw@dsn.dk (om COR) og tha@dsn.dk (om rettigheder til Sprognævnets leksikalske ressourcer). Program CLINK er p.t. i version 0.9 og frigives til offentligheden som 'SaaS' (Software as a Service) fra og med version 1.0, efter planen i perioden 2024-25. Besøg COR-projektets site ordregister.dk, hvor der bringes nyheder om projektet, og hvor sprogressourcer omtalt i denne artikel kan tilgås.

7. Afsluttende bemærkning

Sprognævnets CLINK-arbejde har blotlagt flere træk ved den danske ortografi som hverken er beskrevet i Retskrivningsordbogen eller andre

steder. Den rige informationsstruktur som CLINK-templaten knytter til enhver morf (ikke kun ordbogens almindelige lemmaer, men også affikser, fuger, funktionstegn m.m.), gør det lettere at beskrive hvordan de forskellige sproglige lag, grammatik, betydning og udtale, interagerer med ortografien. For eksempel tror vi at RO's paragraffer – a fortiori kommareglerne – kan formuleres mere præcist i CLINK-termer end på eksempelbasis. Skulle det være rigtigt, går vi måske en fremtid i møde med pålidelig kommakorrektur i tekstbehandlingen. Den dag vil Sprognævnets spørgetelefon få frigjort en fjerdedel af sin kapacitet til mere fornuftige emner.

Om forfatteren

Peter Juel Henrichsen, ph.d., seniorforsker, Dansk Sprognævn.

Litteratur

Albeck, Ulla (1968): *Dansk Stilistik*. 6. udg. København: Gyldendal

Henrichsen, Peter Juel (2021): Glemte Ord. En undersøgelse af H.C. Ørsteds nyord og deres plads i nudansk. *NyS – Nydanske Sprogstudier* 60, 7-36.

Henrichsen, Peter Juel (2022): Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi. I: *Nordiske Studier i Leksikografi.* Louise Holmer et al. (red.). Lund: Lund Universitet, 113-126.

Henrichsen, Peter Juel (2023): Diktaroriske Befølelser: Om ord og uord i Det Centrale Ordregister. I: *19. Møde om Udforskningen af Dansk Sprog*. Kirstine Boas et al. (red.). Aarhus: Aarhus Universitet, 133-148.

Henrichsen, Peter Juel (2024): Make each morf count. A new approach to computational lexicography for text processing. I: Kristina Despot et al. (red.). *Proceedings of EURALEX-21*. Cavtat: Institute of Croatian Language and Linguistics (accepteret).

Keson, Britt (1999): Vejledning til det Danske Morfosyntaktisk Taggede PAROLE-korpus. I: *Parole Dokumentation*. Jørg Asmussen (red.). København: Det Danske Sprog- og Litteraturselskab, 1-67.

Kirchmeier, Sabine, Peter Juel Henrichsen & Philip Diderichsen (2019): *Dansk Sprogteknologi i Verdensklasse. Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet*. København: Dansk Sprognævn

Morill, Glyn Verden (1994): *Type logical grammar. Categorial logic of signs*. Dordrecht: Kluwer Academic Publishers.

Nimb, Sanni et al (2022): COR-S – den semantiske del af Det Centrale OrdRegister. *LexicoNordica* 29, 75-97.

Togeby, Ole (2000): *'Henret ikke benådet!'*. Kronik i dagbladet Information, 28. februar 2000.

Widmann, Thomas (2023): Det Centrale Ordregister og dets leksikografiske anvendelser. I: *Nordiske Studier i Leksikografi.* Louise Holmer et al. (red.). Lund: Lund Universitet, 415-430.

Wood, Mary McGee (1993): *Categorial Grammars*. London: Routledge.

Internetlinks

https://ordregister.dk