Automatisk orddeling

Hasse Hansson
Datalogisk Institut, Københavns Universitet
Sigurdsgade 41, DK-2200 København N, Danmark

Indledning

Når man skriver på et stykke papir, er man tvunget til at tage en beslutning, når man nærmer sig papirets højre kant: Skal hele det følgende ord flyttes ned på næste linie, eller skal det deles? Vilkårligheden af dette valg understreger, at en orddeling er et nødvendigt onde og ikke i sig selv rummer nogen information i modsætning til fx en indrykket linie, der markerer starten af et nyt afsnit. Det er derfor et rimeligt krav, at informationsindholdet i det delte ord bevares, således at orddelinger, der medfører midlertidig forvirring eller permanente misforståelser, undgås.

På trods af dette har den stadigt stigende anvendelse af automatisk tekstbehandlingsudstyr medført, at megen trykt tekst indeholder et stort antal forkerte orddelinger. Udviklingen er især tydelig inden for avisproduktion, hvor korrekturlæsning ofte foregår før ombrydning, og dermed før de automatiske orddelinger er udført. Nedenstående typiske eksempler på forkerte orddelinger er således fundet i de danske aviser Politiken og Ekstra Bladet.

middelhav-segnene	øjeb-likket
efte-rårskollektion	salg-sapparat
whiskyf-laske	rengøring-spersonalet
hvermand-seje	unds-lap
bes-laglagt	sanitet-stropperne

Da orddelinger tydeligvis er en væsentlig kilde til fejl, bør det undersøges, om orddeling overhovedet er nødvendig. Nedenfor vises to spalter med tilfældig tekst; den første indeholder ingen orddelinger, hvorimod der i den anden er foretaget orddelinger, hvor der er mulighed for det.

I de senere år har edb-baseret udstyr til tekstbehandling vundet en stadig større udbredelse. Indenfor forsknings- og undervisningssektoren kan især fremstillingen af dokumenter som rapporter og manualer lettes betydeligt ved anvendelse af tekstbehandlingsudstyr. Som oftest er sådant udstyr baseret på mini- eller mikrodatamater, men for brugere med terminaler tilsluttet et regnecenter som RECKU er det et naturligt ønske, at terminalerne også kan anvendes til tekstbehandlingsformål.

I de senere år har edb-baseret udstyr til tekstbehandling vundet en stadig større udbredelse. Indenfor forsknings- og undervisningssektoren kan især fremstillingen af dokumenter som rapporter og

Automatisk orddeling
Hasse Hansson
Proceedings of NODALIDA 1979, pages 33-42

manualer lettes betydeligt ved anvendelse af tekstbehandlingsudstyr. Som oftest er sådant udstyr baseret på mini- eller mikrodatamater, men for brugere med terminaler tilsluttet et regnecenter som RECKU er det et naturligt ønske, at terminalerne også kan anvendes til tekstbehandlingsformål.

Det ses, at linie 4 og 7 i øverste spalte indeholder uacceptabelt store mellemrum, og at den nederste spalte er 1 linie kortere end den øverste. Konklusionen må derfor være, at orddeling er nødvendig, dels for at få rimeligt udfyldte linier, dels for at opnå en papirbesparelse.

I perioden februar - september 1979 har jeg derfor i samarbejde med lektor Bente Maegaard fra Institut for anvendt og matematisk lingvistik, Københavns Universitet, udviklet en algoritme til automatisk deling af danske ord. Denne artikel beskriver dels de metoder, som vi har udviklet til brug for algoritmefremstillingen, dels den færdige algoritme. Det forudsættes, at læseren har et godt kendskab til det danske sprog; endvidere vil kendskab til programmeringssproget Pascal eller et andet ALGOL-lignende sprog være ønskeligt.

Regler for dansk orddeling

Målet for algoritmeudviklingen er at fremstille en algoritme, der i så mange tilfælde som muligt deler ordene i overensstemmelse med de regler for dansk orddeling, der er angivet i Retskrivningsordbogen [1]. Formålet med dette afsnit er at analysere reglerne for at klargøre i hvor stort omfang, de lader sig implementere på en datamat.

i) "Sammensatte ord deles efter deres bestanddele, når disse er let kendelige."
På dansk kan sammensatte ord frit dannes ved at skrive de to ord, der danner sammensætningen, uden et adskillende mellemrum, hvorimod man fx på engelsk i stor udstrækning beholder mellemrummet eller anvender en bindestreg. På trods af reglens klarhed er der derfor store vanskeligheder forbundet med maskinelt at afgøre, om et ord er sammensat.

ii) "Afledninger deles ligesom sammensætninger efter deres bestanddele, når disse er let kendelige."
Reglen er en parallel til i); men den er mindre konsekvent, idet det er usikkert, hvornår en afledning er "let kendelig". Antallet af afledningselementer, affikser,[1] er dog ret begrænset, hvorfor denne regel kan implementeres, hvis det er muligt at opstille fuldstændige lister over de affikser, der accepteres som let kendelige.

iii) "For usammensatte ord gælder følgende regler:"
   1) "En medlyd mellem to selvlyd skrives sammen med den sidste selvlyd."
   Denne regel kan umiddelbart implementeres; men der er dog, som bemærket af Spang Hanssen [2], en tradition for at skrive konsonanten *x* sammen med den første vokal.
   2) "Af to medlyd mellem selvlyd går en til hver linie; ..."
   Implementation af denne simple regel vanskeliggøres af følgende undtagelser (især b) og c)):
      a) "*sk, sp, st* kan gå sammen til ny linie."
      Der er tradition for at udnytte denne mulighed, der uden vanskelighed kan implementeres.
      b) "I ord, der ikke har tryk på første stavelse, går begge medlyd til den trykstærke

1. Affikser er en fællesbetegnelse for præfikser, infikser og suffikser; men kun præfikser (forstavelser) og suffikser (afledningsendelser) har interesse i denne sammenhæng

stavelse, hvis de danner en lydforbindelse, der kan forekomme i begyndelsen af et ord."
Reglen lader sig ikke implementere, idet en datamat ikke ud fra ordets stavemåde kan afgøre, hvor trykket ligger.

c) "Efter lang selvlyd kan man skrive begge medlyd sammen med den følgende selvlyd, hvis de danner en lydforbindelse, der kan forekomme i begyndelsen af et ord."
Heller ikke denne regel kan implementeres, idet det ikke er muligt på baggrund af ordets stavemåde at afgøre, om en vokal er lang.

3) "Tre eller flere medlyd behandles sådan, at der til den ny linie kun går så mange, som der kan forekomme i begyndelsen af et ord."
Denne regel kan - måske lidt overraskende - implementeres, idet det er muligt at opstille fuldstændige lister over de konsonantforbindelser, der kan forekomme i begyndelsen af et ord.

Regler for orddeling på andre sprog

Efter analysen af reglerne for deling af danske ord er det interessant at undersøge, i hvor stort omfang reglerne er i overensstemmelse for andre europæiske sprog. Nedenfor er derfor anført en simplificeret version af de danske regler:

a) Sammensatte ord deles i deres bestanddele.
b) Afledninger deles efter bestanddele.
For usammensatte ord:
c) Ved 1 konsonant mellem to vokaler skrives konsonanten sammen med den sidste vokal.
d) Ved 2 konsonanter mellem to vokaler går en konsonant til hver stavelse.
e) Ved 3 eller flere konsonanter skrives højst så mange konsonanter sammen med sidste vokal, som der kan forekomme i begyndelsen af et ord.

Det er muligt ved hjælp af reference [1] samt [3] til [7] at undersøge, i hvor høj grad ovenstående fem regler er gældende for dansk, engelsk, fransk, tysk og svensk. Resultatet er angivet i nedenstående skema, hvor + betyder, at reglen gælder, (+) at reglen gælder men ikke følges konsekvent eller at der er undtagelser, og - at reglen ikke gælder.

	dansk	engelsk	fransk	tysk	svensk
a	+	+	+	+	+
b	(+)	+	(+)	-	+
c	+	+	+	+	+
d	(+)	+	(+)	(+)	(+)
e	+	-	+	-	-

Som det fremgår af skemaet, er der ret store uoverensstemmelser - især for de usammensatte ords vedkommende. Man kan derfor ikke gøre sig håb om at udvikle én algoritme, der kan dele ord på to eller flere af ovennævnte sprog.

Metoder til automatisk orddeling

I dette afsnit gennemgås fem forskellige metoder, som vi har udviklet til automatisk orddeling. Grunden til at der er udviklet ikke mindre end fem forskellige metoder er, at problemet på naturlig måde kan opdeles i underproblemer, der hver kræver sin egen metode for at blive løst tilfredsstillende. Ligegyldigt hvilken metode, der anvendes i et givet tilfælde, vil vi kræve, at den skal give en orddeling, der i så høj grad som muligt er i overensstemmelse med de tidligere omtalte regler for dansk orddeling. Metoderne vil i stor udstrækning blive illustreret med

eksempler, i hvilke tegnet - betegner en korrekt orddeling og tegnet ≠ en fejlagtig orddeling.

Affiksgenkendelse

Regel i) kræver, at sammensatte ord deles i deres bestanddele. Som et eksempel kan nævnes ordet *afledningsendelse*, der deles *aflednings-endelse* og ikke *afledning≠sendelse,* som ville være en korrekt orddeling, hvis ordet ikke var sammensat. Det er imidlertid vanskeligt at afgøre, om et ord er sammensat og i givet fald, hvor det er sammensat.

En metode hertil er at genkende et affiks midt i ordet. I førnævnte eksempel kan man således genkende suffikset *nings* og opnå en korrekt deling. Nedenfor vises eksempler på suffikser, som det empirisk har vist sig bør genkendes. Venstre spalte indeholder affikser, som der kan deles efter, medens højre spalte indeholder de affikser, som der kan deles før.

nings-	-agtig
tets-	-bar
ments-	-dom
lig-	-mæssig
som-	-skab

Man skal dog ikke ukritisk forsøge at genkende samtlige affikser, som er nævnt i fx Hansen [8] og Skautrup [9]. Et eksempel er *af,* som korrekt genkendes i ordet *flertals-afgørelse,* medens genkendelsen fører til fejl i ordet *bortsk≠affe.* Ved at afprøve de opstillede affikslister med et omfattende inddatamateriale vil det hurtigt blive klart, hvilke affikser, som med godt resultat kan genkendes.

Regel ii) kræver, at afledninger deles efter deres bestanddele, når disse er let kendelige. Det er ikke muligt at give en præcis definition af, hvornår en afledning er "let kendelig"; men det er dog muligt at opstille en liste med et ret begrænset antal elementer, der accepteres som "let kendelige" afledninger af vide kredse. Det er derfor muligt ved hjælp af affikslister at implementere denne regel således, at der opstår meget få fejl.

Statistiske metoder

Hvis det ikke er muligt at genkende et affiks i ordet, må det deles på baggrund af statistiske resultater. Reglerne for deling af usammensatte ord deler op i tre tilfælde: 1 konsonant, 2 konsonanter og 3 eller flere konsonanter mellem to vokaler. Det er derfor naturligt, at lade programmet foretage den samme opdeling.

I tilfældet 1 konsonant skal der ifølge regel iii) pkt. 1 deles før denne konsonant, og dette påbud kan umiddelbart følges. Den eneste undtagelse er konsonanten *x*, som traditionelt skrives sammen med første vokal.

Matriks for tokonsonantkombinationer

Hvis der optræder to konsonanter mellem to vokaler, skal der ifølge regel iii) pkt. 2 deles mellem dem; men på grund af undtagelse b) og c) kan reglen ikke umiddelbart implementeres. Vi har derfor valgt at analysere et omfattende ordmateriale for empirisk at finde det bedste delepunkt for hver af de 400 mulige konsonantkombinationer. Da analysens kvalitet naturligvis afhænger af det valgte analysemateriale, er det vigtigt, at dette vælges med omhu. Et umiddelbart valg er en ordbog som fx Nudansk Ordbog [10], der findes på maskinlæsbar form. Imidlertid er et ordbogsmateriale uegnet af to grunde: Antallet af sammensatte ord i fx Nudansk Ordbog er forsvindende i forhold til antallet af mulige sammensætninger, og som tidligere nævnt er det netop sammensætninger, der virkelig volder problemer. Den anden grund er, at frekvensen i naturlig tekst af et givet ord fuldstændig ignoreres i ordbogsmaterialet. Dette betyder, at man kan risikere at vælge en orddeling ud fra et antal ord, som forekommer sjældent i naturlig tekst. Ved i stedet at vælge orddelingen ud fra færre, men ofte forekommende ord vil man opnå færre

fejl ved deling af ord i en naturlig tekst.

Vi har derfor valgt at samle et analysemateriale, som indeholder ord fra et stort antal forskellige tekster som vist i nedenstående skema:

Dansk prosa	250.000 ord
Avisartikler	20.000 ord
RECKU-Nyt[2]	50.000 ord
Økonomiske Råds rapport	22.000 ord
Matematisk fagtekst	11.000 ord
Geologisk fagtekst	5.000 ord
I alt	358.000 ord

"Dansk prosa" dækker over tekstprøver fra mange forskellige danske forfattere, hvis bøger er blevet udgivet i de senere år. Materialet indeholder i alt 358.000 ord, hvoraf de 35.000 er forskellige.[3] Til sammenligning kan nævnes, at Nudansk Ordbog indeholder ca. 51.000 ord.

Ved analysen af tokonsonantkombinationer viser det sig, at man kan finde sikre delepunkter for langt de fleste kombinationer, fx:

-bl: *øje-blik*
r-n: *tor-ne*
zz-: *jazz-orkester*

Der er dog enkelte kombinationer, der er meget usikre, som fx *dr:*

d-r: *yd-re*
-dr: *a-dresse*

Endelig er der en række kombinationer, som er umulige både som initial- og som finalkombinationer. Forekommer en sådan kombination midt i et ord, kan man med sikkerhed dele mellem de to konsonanter, fordi ordet da må være sammensat. Som et eksempel kan nævnes kombinationen *df:*

d-f: *flod-forurening*

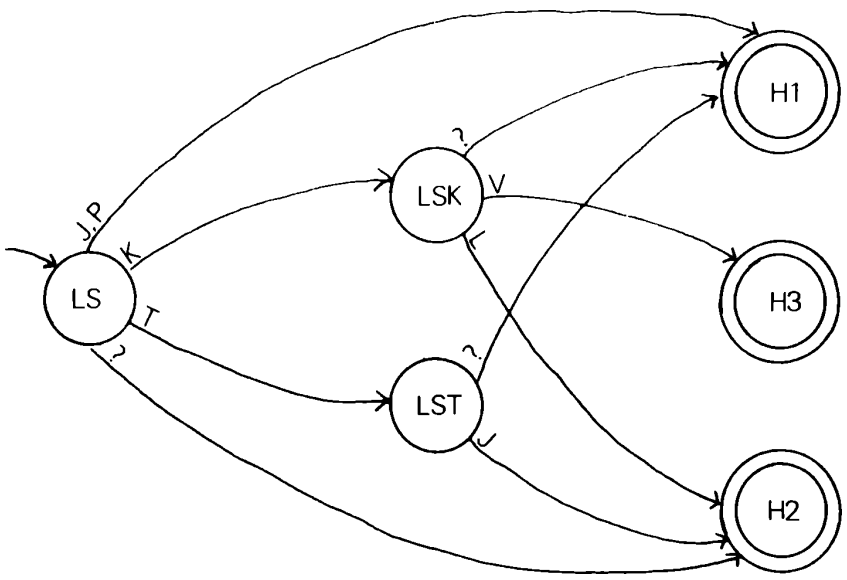
Resultatet af analysen er en 20 × 20 matriks, der for hver konsonantkombination angiver, om der skal deles til venstre, mellem eller til højre for de to konsonanter. Endelig indeholder matricen oplysninger om de kombinationer, der er usikre; anvendelsen af denne oplysning vil blive forklaret senere.

Tilstandsautomat

Ved kombinationer af tre og flere konsonanter er mulighederne så talrige, at det ikke er hensigtsmæssigt på tilsvarende måde at opstille en matriks i tre eller flere dimensioner. Dette skyldes dels sådanne matricers mange elementer, dels at de fleste konsonantkombinationer er utænkelige i praksis og derfor uinteressante. I stedet har vi ud fra analysen af materialet opstillet en endelig tilstandsautomat, hvor man - populært sagt - kun registrerer de kombinationer, som man er interesseret i. For at begrænse automatens størrelse indeholder den ikke de konsonantkombinationer, som man kan dele korrekt ved at anvende den tidligere omtalte tokonsonantmatriks på de to konsonanter umiddelbart foran den anden vokal. Et eksempel er ordet *overblik,* hvor matricen giver delingen *-bl* og dermed *over-blik.*

Der er dog stadig mange kombinationer, som må klares ved en tre- eller endda firkonsonantanalyse. Mange kombinationer forekommer kun i sammensætninger, hvorfor mange sammensatte ord bliver delt korrekt netop på grund af automaten. Nedenfor vises den lille del af automaten, der finder en orddeling i alle konsonantkombinationer, der indledes med konsonanterne *ls.*

2. RECKU-Nyt er en brugerorientering, der udsendes ca. 10 gange årligt af RECKU, Det Regionale Edb-center ved Københavns Universitet.
3. Forskellige bøjningsformer af samme ordstamme regnes i denne forbindelse for forskellige ord.



Tilstandene med dobbeltring - på tegningen H1, H2 og H3 - kaldes *sluttilstande*. Når en sådan nås, er analysen færdig, idet H1 betegner en deling efter den første konsonant, H2 en deling efter den anden konsonant etc. Ved at følge ord som ***hals-klud, bol-sje, detail-specifikation, elsk-værdig, formåls-tjenlig*** og ***galskab*** gennem automaten ses, hvorledes vidt forskellige ord - såvel sammensatte som usammensatte - analyseres og deles korrekt.

Fonetisk kontrol

Dansk orddeling følger ligesom fx fransk et fonetisk princip, hvorfor den konsonantkombination, der forekommer efter delestregen, skal kunne forekomme i begyndelsen af et ord. Når man derfor ved en af de ovenfor nævnte metoder har fundet en orddeling, undersøges om kombinationen er lovlig. Er dette ikke tilfældet, flyttes bindestregen mod højre, indtil kravet er opfyldt. Denne *fonetiske kontrol* forhindrer især de fejl, der opstår ved forsøg på genkendelse af et affiks, hvor den genkendte tegnfølge i det givne ord viser sig at indgå i en ganske anden sammenhæng. Et eksempel er ordet *angsten,* hvor præfikset *an* genkendes og ordet derfor deles *an≠gsten.* Den fonetiske kontrol sikrer, at *gs* ikke forekommer efter delestregen, og delingen rettes derfor til *ang-sten.*

Undtagelsesordbog

Vi vil i det følgende sammenfatte de hidtil omtalte metoder under betegnelsen de *algoritmiske metoder.* Selvom disse metoder forfines, vil der stadig findes ord, der ikke deles korrekt. For alligevel at kunne dele sådanne ord rigtigt opbygges en *undtagelsesordbog,* der indeholder alle de ord, som erfaringsmæssigt giver anledning til fejl. På grund af det store antal opslag i en sådan ordbog er det vigtigt, at den organiseres således, at søgning er effektiv; specielt skal søgning efter ord, der ikke findes i ordbogen, være effektiv. Dette opnås ved at benytte en såkaldt hashteknik. For en nærmere beskrivelse af teknikken henvises til Filsystemer og databaser [11].

Som et kuriosum kan nævnes, at selv med anvendelse af en undtagelsesordbog vil homografer[4)] kunne føre til forkerte orddelinger. Et eksempel er *vandrende,* som deles *van-drende,* hvis det er præsens participium af verbet *vandre,* medens det deles *vand-rende,*

4. Homografer er forskellige ord, der har fælles stavemåde.

hvis det er et substantiv. Vi har hidtil kun beskæftiget os med analyse af enkelte ord; men dette problem kan kun løses ved en syntaktisk analyse af hele den sætning, hvori ordet indgår.

Den komplette algoritme

Formålet med dette afsnit er at vise, hvorledes de fem metoder, som er omtalt i forrige afsnit, kan kombineres til en komplet algoritme. Desuden omtales nogle detaljer, som ikke er af principiel art og derfor ikke har været omtalt tidligere. Algoritmen, der i denne udformning læser ord fra en datakilde og finder samtlige mulige delepunkter i disse, er skitseret nedenfor i et Pascal-lignende sprog.

```
 1 while more_words( source ) do
 2 begin
 3    safe_hyphen( word );
 4    dictionary_lookup( word );
 5    search_prefix( word );
 6    while more_syllables( word ) do
 7    begin
 8       search_affixes( syllable, found );
 9       if not found then
10       begin
11          case consonants of
12             0: if divisible( syllable ) then save_hyphen;
13             1: save_hyphen;
14             2: matrix_hyphen( syllable );
15             otherwise
16             begin
17                finite_state_automata( syllable, found )
18                if not found then matrix_hyphen( syllable );
19             end
20          end
21       end
22    end
23 end.
```

I linie 3 findes de *sikre delesteder,* hvorved vi forstår forekomster af tegnene '-', ',' eller '/'. Det forekommer os indlysende, at et ord som *dansk-tysk* skal deles efter bindestregen; men på trods heraf genkender mange orddelingsalgoritmer ikke bindestreger, hvilket naturligvis fører til besynderligt udseende delinger. Grunden til at der deles efter et komma er, at der i visse tekster ikke er indsat et blanktegn efter kommaet. Ligeledes anvendes skråstregen undertiden til at danne sammensatte ord som fx *input/output.*

I linie 4 slås ordet op i undtagelsesordbogen, og hvis det findes, registreres de delepunkter, som ordbogen angiver.

I linie 5 søges efter et præfiks i en særlig præfiksliste. Dette skyldes, at det ved analysen af affikser viste sig, at der findes mange præfikser, som med fordel kan genkendes først i et ord men ikke midt i et ord. For ikke at miste disse ret sikre orddelinger samledes sådanne præfikser i præfikslisten.

Algoritmen har i linie 3 - 5 arbejdet på hele ordet; men i linie 6 - 22 findes en løkke, som anvendes til at finde delepunkter mellem de enkelte stavelser. Betingelsen for at der findes et delepunkt mellem to stavelser i denne løkke er, at der ikke allerede i linie 3 - 5 er registreret et delepunkt mellem de pågældende stavelser.

I linie 8 gennemløbes to affikslister, hvor den ene indeholder de affikser, som der skal deles efter, medens den anden består af de affikser, som der skal deles før. Hvis et sådant affiks ikke

genkendes, antages stavelsen at være en del af et usammensat ord, og der foregår derfor en opsplitning efter antallet af konsonanter ved hjælp af CASE-sætningen i linie 11.

Det tilfælde, hvor der ikke findes nogen konsonanter mellem de to vokaler, der afgrænser stavelsen, behandles i linie 12. Et sådant *vokalsammenstød* kan altid deles i oprindeligt danske ord; men i mange fremmedord betegner de to vokaler en diftong, som ikke kan deles. For eksempel kan hverken *ea* eller *au* deles i ordet *niveau.* Der er derfor opbygget en lille tabel, der angiver mellem hvilke vokaler, en deling er tilladelig.

Tilfældet 1 konsonant klares umiddelbart i linie 13, medens tilfældet 2 konsonanter løses ved opslag i tokonsonantmatricen. Ved tre eller flere konsonanter anvendes tilstandsautomaten; indeholder den ingen løsning, anvendes tokonsonantmatricen på de to konsonanter umiddelbart før den anden vokal.

Nøjagtighed

Den konstruerede algoritme er mere kompleks end tilsvarende algoritmer, som vi har fået kendskab til. Det er derfor interessant at undersøge, om resultatet er så meget bedre, at den forøgede programkompleksitet kan betale sig. Nedenfor vises de fejlprocenter, som opnås med den færdige algoritme.

Uden ordbog	1.8 % fejl
Med ordbog	1.2 % fejl

Fejlprocenten uden ordbog er målt på hele analysematerialet (358.000 løbende ord), medens fejlprocenten med ordbog er målt på tilfældige avisartikler, da det oprindelige analysemateriale er blevet brugt ved opbygning af undtagelsesordbogen.

Tallene virker måske ikke umiddelbart imponerende, men der er to faktorer, som skal med i vurderingen. For det første er resultatet klart bedre end tilsvarende, kommercielt udviklede programmer. Politiken hævder således, at deres program giver en fejlprocent på 3, og leverandører af tekstbehandlingsprogrammel er stolte af at markedsføre algoritmer, der har en fejlprocent på 2; men vores algoritme er med 1.2 % næsten dobbelt så god. For det andet indeholder undtagelsesordbogen på nuværende tidspunkt kun ca. 1000 ord, og en fuldt udbygget ordbog vil bevirke, at fejlprocenten falder til et niveau, hvor fejlene hovedsagelig skyldes nye sammensatte ord.

Anvendelse i tekstbehandlingsprogrammel

Jeg har i nogle år beskæftiget mig med udvikling af et tekstbehandlingssystem kaldet PHOTODOC, der anvendes ved fremstilling af sats på RECKU's[5)] fotosætter. Ved dette arbejde viste der sig et stort behov for en algoritme til at dele danske ord, og det er grunden til, at jeg begyndte at udvikle orddelingsalgoritmen. Imidlertid er en fleksibel grænseflade mellem orddelingsprogrammet og det øvrige tekstbehandlingssystem en forudsætning for, at man får det fulde udbytte af orddelingsalgoritmen. I det følgende beskrives derfor, hvorledes algoritmen kan indbygges i et tekstbehandlingssystem, således at en høj grad af fleksibilitet opnås.

Parametre til orddelingsalgoritmen

Da automatisk orddeling er en kilde til fejl, bør orddelingsalgoritmen ikke aktiveres, hvis man uden orddeling kan opnå en linie af typografisk acceptabelt udseende. Man kan ved hjælp af følgende fire parametre styre dels antallet af delinger, dels udseendet af disse:

1) Aktiv eller passiv
2) Relativ spildfaktor

5. RECKU er en forkortelse for Det Regionale Edb-center ved Københavns Universitet.

3) Sikkerhedsniveau
4) Minimum antal tegn før og efter delestreg

Første parameter giver mulighed for at hindre, at orddelingsalgoritmen aktiveres. Dette har især betydning, hvis kildesproget ikke er dansk, da algoritmen er uanvendelig til andre sprog end dansk.

Anden parameter, den *relative spildfaktor,* angiver den procentdel af linlen, der skal være ubrugt før orddelingsalgoritmen aktiveres. Angives en stor spildfaktor, fås få orddelinger men til gengæld også en "løs" og dermed uøkonomisk sats; omvendt vil man ved en lille spildfaktor få en "tæt" sats med mange delinger. Den bedste spildfaktor afhænger af personlig smag; men en værdi mellem 5 og 10 procent giver sædvanligvis gode resultater. Det er væsentligt, at spildfaktoren er relativ: En lang linie giver bedre mulighed for at fordele en vis uudnyttet plads end en kort linie.

Ved omtalen af tokonsonantmatricen blev det nævnt, at visse kombinationer bliver registreret som usikre; men også nogle få konsonanter i enkonsonantforbindelser samt enkelte kombinationer i tilstandsautomaten er registreret som usikre. Ved at angive et *sikkerhedsniveau* for orddelinger kan man undgå, at ord deles i usikre delepunkter, og man kan ligeledes vælge kun at dele ord i *sikre* delepunkter, hvortil regnes forekomster af bindestreg, komma eller skråstreg i ordet. Sidste mulighed har især betydning for engelsk, hvor man på denne måde har mulighed for at dele alle de sammensætninger, der er dannet ved hjælp af en bindestreg.

Sidste parameter angiver det minimale antal bogstaver, der skal findes såvel før som efter delepunktet. Der er almindelig enighed om, at to bogstaver må være minimum; men et minimum på tre tegn foretrækkes undertiden.

Interaktiv kontrol

Da orddelingsalgoritmen som nævnt har en fejlprocent på ca. 1.2, må man ved store tekster regne med at skulle rette nogle forkerte orddelinger i den ombrudte tekst. Imidlertid har man ved hjælp af en *interaktiv kontrol* mulighed for at rette forkerte orddelinger i det øjeblik, de opstår. Således kan man ved kun én programudførelse opnå en ombrudt tekst, der i hvert fald med hensyn til orddelinger er helt korrekt. Nedenfor vises et eksempel på en sådan interaktiv kontrol. Den kursiverede tekst er brugerens inddata, medens alt andet udskrives af datamaten.

@pho★to.doc file.elt
PHOTODOC 4R1 RLIB 73R1 Wednesday, 1979 November 14, at 12:27:46
@hyphen
★ Hyphen control is active ★
TEKSTBE H★ANDLING ->
UD ★STYR ->
TRYKFÆRDIG ★DIG ->*tryk færdig*
TY ★POGRAFISKE ->*typo*
Hyphen beyond maximum point - try again!
TY ★POGRAFISKE -> *t*
FACILITE T★ER ->*favili*
Misspelling - try again!
FACILITE T★ER ->*facili*
MU L★IGHED ->
0 ERRORS AND 0 WARNINGS
PROOF LIST HAS BEEN SENT TO PR2
TIME USED: 9.318 SECONDS.
END PHOTODOC

PHOTODOC præsenterer alle orddelinger for brugeren, idet der indsættes et blanktegn på det sted i ordet, hvor datamaten mener, at ordet skal deles. Stjernen indsættes for at markere den del af ordet, der maksimalt kan stå før delestregen. Det første delte ord i eksemplet, *tekstbehandling,* vil datamaten således dele *tekstbe-handling,* og delestregen kan ikke indsættes senere end efter *h.* Accepteres orddelingen, trykker brugeren blot på vognreturknappen, i modsat fald skrives ordet med et blanktegn indsat på det sted, hvor ordet ønskes delt. Et eksempel er vist ved ordet *trykfærdig,* som brugeren ønsker delt *tryk-færdig.* Af effektivitetsgrunde kan den del af ordet, der skrives efter blanktegnet, dog udelades.

Endelig vises eksempler på de to fejl, som brugeren kan begå. I det første tilfælde forsøger man at indsætte en delestreg til højre for maksimumspunktet, og i det andet tilfælde staves ordet forkert ved rettelsen. I begge tilfælde udskrives en fejlmeddelse, hvorefter brugeren har mulighed for at rette fejlen.

Konklusion

Det er lykkedes at udvikle en orddelingsalgoritme, som er klart bedre end de hidtil kendte algoritmer til deling af danske ord. Algoritmen har vist sin anvendelighed i praksis og giver ifølge brugerudsagn en signifikant besparelse i forhold til manuel orddeling.

Efter min mening er en væsentlig årsag til de gode resultater samarbejdet med en lingvist, der bevirkede, at vi under algoritmeudviklingen betragtede både lingvistiske og datalogiske aspekter. Anvendelsen af lingvistik sikrede dels, at vi fx gennem affiksgenkendelse ikke ignorerede ordenes sproglige struktur, dels at vi opbyggede et dækkende analysemateriale. Anvendelsen af datalogi førte til en programstruktur, der nøje afspejler de regler for dansk orddeling, som har været vort arbejdsgrundlag. Under udviklingen har det derfor været muligt at omskrive eller forfine et enkelt modul uden at berøre programmets øvrige moduler, hvorfor det har været tidsmæssigt overkommeligt at eksperimentere med forskellige løsningsmuligheder.

Dette arbejde er således et eksempel på de gode resultater, der kan opnås gennem et konstruktivt, tværfagligt samarbejde.

Referencer

[1] Retskrivningsordbogen. Udgivet af Dansk Sprognævn, 1955.

[2] Henning Spang-Hanssen: Orddeling ved linieskifte - er reglerne tidssvarende? Artikel i SAML, nr. 5 1979, p. 73 - 93.

[3] Poul Steller og Knud Sørensen: Engelsk grammatik, København 1966.

[4] Der Sprach-Brockhaus, Wiesbaden 1966.

[5] Poul Høybye: Fransk grammatik, København 1966.

[6] Skrivregler. Tekniske Nomenklaturcentralens publikationer nr. 37, Stockholm 1976.

[7] C. Hansen-Chrisensen og Sven Brüel: Oversigt over tysk grammatik, København 1979.

[8] Åge Hansen: Moderne dansk, København 1967.

[9] Peter Skautrup: Det danske sprogs historie, bd. 3, København 1968.

[10] Nudansk Ordbog, København 1969.

[11] Bratsbergsengen, Høfstad og Wibe: Filsystemer og databaser, Trondheim 1974.