



## Orddeling

### 1. Generelt

Forutsatt at verdien av `\hyphenchar` ligger innenfor intervallet  $[0-255]$ , vil  $\text{\TeX}$  normalt dele ord automatisk, hvis deling er nødvendig for å unngå overfylte eller underfylte bokser. Ord som begynner med stor forbokstav deles bare dersom `\uchyph`  $> 0$ . Antallet delte ord avhenger dels av satsspeilets bredde i forhold til størrelsen på den aktuelle font, dels av verdien av forskjellige parametre ( $\leftrightarrow$  **Linjefall 1.**).

$\text{\TeX}$  prøver først å sette et avsnitt uten orddelinger. Hvis *badness* ved dette første forsøket overstiger verdien av `\pretolerance`, gjør  $\text{\TeX}$  et nytt forsøk og prøver nå å dele ord på grunnlag av: (1) en unntaksliste, (2) en samling orddelingsmønstre (*patterns*) (se ekskursen nedenfor). Under dette andre forsøket erstattes `\pretolerance` med `\tolerance` som grenseverdi for hvilken *badness* som kan aksepteres. Forhåndsverdiene for `\pretolerance` og `\tolerance` er i plain  $\text{\TeX}$  satt til hhv. 100 og 200.

I hovedtrekk foregår orddelingsprosessen slik: Etter et  $\lim$ <sup>34</sup> som ikke opptrer i en matematisk formel, vil  $\text{\TeX}$  prøve å identifisere et «ord» (dvs. en tegnsekvens som det om nødvendig kan dele hvis orddelingsmønstrene tillater det).  $\text{\TeX}$  oppfatter det som om et slikt ord begynner når det påtreffer et tegn i font *f* med `\lccode`  $\neq 0$ , fulgt av ett eller flere tegn med samme egenskap. I ordet kan det, foruten ligaturer definert i `.tfm`-filen, også inngå kontrollsekvenser eller aktive tegn, såfremt disse ikke ekspanderer til noe annet enn  $\_{11}$ - og  $\_{12}$ -token med den riktige `\lccode`.  $\text{\TeX}$  betrakter ordet det i ferd med å lese, som avsluttet når det støter på noe som: (1) ikke er eller ekspanderer til et tegn med `\lccode`  $\neq 0$  (eller en ligatur dannet utelukkende av slike tegn); (2) ikke er en implisitt utligning. For at et ord skal være delelig, må det etterfølges av null eller flere tegn, ligaturer eller implisitte utligninger, som i sin tur følges av et  $\lim$  eller en eksplisitt utligning eller en *penalty* eller en *whatsit* eller vertikalt materiale fra `\mark`, `\insert` eller `\vadjust`. En boks, en strek, en matematisk formel og en `\discretionary`-kommando som opptrer like etter et ord, vil følgelig forhindre orddeling (jf. **Ligaturer 2.**).

<sup>34</sup> Et ord som følger umiddelbart etter en `\kern`-kommando, blir altså **ikke** forsøkt delt.

- Hvis man bruker CM-fonter, vil  $\TeX$  ikke prøve å dele ‘årsresultatsforbedring’, siden ‘å’ ekspanderer til `\accent 23a`, hvilket innebærer en eksplisitt utligning (jf. **Aksent** og **Utligning**). Når  $\TeX$  støter på en slik utligning, innstiller det forsøkene på å identifisere et delelig ord, og gjenopptar dem først etter neste forekomst av lim i inndatastrømmen. Derimot vil  $\TeX$ , når det leser sekvensen ‘regnskapsåret’, identifisere det som står foran ‘å’, dvs. ‘regnskaps’, som et delelig ord.

Ord med ‘æ’ og ‘ø’ vil imidlertid bli delt som normalt, såfremt ‘æ’ og ‘ø’ er utstyrt med `\lccode` (om de skrives inn som aktive tegn eller som kontrollsekvenser, er likegyldig).

Med introduksjonen av  $\TeX$  3.x ble håndteringen av orddeling vesentlig forbedret. Man kan nå veksle mellom forskjellige orddelingsmønstre ved å gi `\language` forskjellige verdier ( $\leftrightarrow$  **Nasjonale tilpasninger**). De nye parametrene `\lefthyphenmin` og `\righthyphenmin` lar en dessuten angi det minste antall bokstaver som kan stå til venstre og til høyre for delingstegnet når et ord deles. Standardverdiene for engelsk er hhv. 2 og 3, men for norsk kan man sette begge til 2. Verdiene kan være forskjellige for ulike delingsmønstre, og kan endres når som helst.

- Appendiks H i *The  $\TeX$ book* (ss. 449–455) gir en beskrivelse av  $\TeX$ s orddelingsalgoritme:

Når programmet skal prøve å dele et ord, f.eks. ‘imidlertid’, utvider det først ordet med en markør i hver ende (heretter betegnet med et punktum): `.imidlertid`. Det utvidede ordet omfatter delord av lengde 1: `. i m i d l e r t i d .`; av lengde 2: `.i im mi id dl le er rt ti id d.`; av lengde 3: `.imi mid idl dle ler ert rti tid id.` osv.

Hvert delord av lengde  $k$  er et mønster som definerer  $k + 1$  heltallsverdier i intervallet  $[0-9]$ , relatert til ønskeligheten av bindestrek foran eller bak delordet, eller mellom dets elementer. Disse verdiene er lik null for alle delord som ikke forekommer i den mønsterordlisten `INITEX` leser inn når det lager et format.

$\TeX$  finner den maksimale mellomverdi mellom ordets enkelte elementer på grunnlag av verdiene i de av ordets delord som opptrer i mønsterordlisten. For ‘imidlertid’ blir dette med norske delingsmønstre: `1mi d1l 1ler r1ti 3tid` (0 er utelatt), hvilket gir `i1mid1ler3tid` som resultat av maksimeringen.

Deling mellom to bokstaver anses som akseptabel hvis mellomverdien mellom dem er et ulike tall. Med `\lefthyphenmin=2` kan ‘imidlertid’ altså deles på steder: ‘imid-ler-tid’.<sup>35</sup>

Mønstrene genereres som regel maskinelt av programmet `PatGen`, på grunnlag av en ordliste som viser alle mulige delingspunkter i de ordene (bokstavsekvensene) som inngår i den. Brukeren angir den maksimale lengde av de delord som skal undersøkes ved beregningen av mønstrene. Størrelsen på denne parameteren er et kompromiss mellom to kryssende hensyn: behovet for å holde antall mønstre så lavt som mulig, og ønsket om å finne flest mulig av de tillatte orddelinger i en tekst.

I løpende engelsk tekst klarer  $\TeX$  på grunnlag av de 4447 mønstrene i `hyphen.tex` å finne ca. 95 % av alle mulige orddelinger. At ca. 5 % blir oversett, vil bare i de færreste tilfeller skape vanskeligheter.

I ord som ‘record’, som på grunn av ulik betoning deles forskjellig avhengig av ordklasse, (substantiv: ‘rec-ord’; verb ‘re-cord’) avskjæres enhver deling,

<sup>35</sup> Med amerikanske orddelingsmønstre vil de tillatte delinger bli: ‘imi-dler-tid’.

siden T<sub>E</sub>X ikke kan vite hvilken av de to muligheter som er den riktige. På tilsvarende måte hindrer delingsmønstre for tysk at T<sub>E</sub>X deler et ord som ‘erblich’ (riktig deling: ‘er-blich’ som preteritum av ‘erbleichen’, ‘erb-lich’ som adjektiv).

Spesielle problemer reiser seg i språk som norsk og tysk, hvor det ofte opptrer spontant dannede sammensetninger. Her vil det, selv med svært store mønsterfiler, neppe være praktisk mulig å forebygge enhver feilaktig deling.

## 2. Brukerstyrt deling av enkeltord

Gale orddelinger kan elimineres og riktige fremtvinges ved en av fremgangsmåtene nedenfor:

### A. Lokalt

Mulige orddelinger angis i løpende tekst med kommandoen `\-`, f.eks. `bidrags\lyter`. Vær oppmerksom på at **alle** tillatte bruddpunkter i så fall må markeres med `\-` hvis T<sub>E</sub>X skal kunne bruke dem. Når vi retter opp en tekst, er dette ofte den enkleste måten å fjerne gale orddelinger på.

- Primitiven `\-` er likeverdig med `\discretionary{-}{#1}{}`. En `\discretionary`-kommando angir hva som skal innføres ved et eventuelt skjønsmessig (*discretionary*) brudd til ny linje på vedkommende punkt, og forhindrer samtidig enhver deling av ordet som ikke skyldes en `\discretionary`-kommando. Den har syntaksen: `\discretionary{før brudd}{etter brudd}{uten brudd}`. Ved hjelp av `\discretionary` kan vi løse de spesielle problemene som ord av typen *oppakning*, *oppasser*, *toppunkt* reiser, siden siden de må deles *opp-pakning* osv. Vil vi slippe å korrigere feildelinger i slike tilfeller, kan vi f.eks. redefinere `\-`, som normalt gir oss `_`:

```
\def\#1{\discretionary{-}{#1}{}}
```

og så skrive `opp\_pakning`, `topp\_punkt` osv. På tilsvarende måte kan vi definere `\ck`, så vi på tysk kan skrive `ba\ck en`:

```
\def\ck{\discretionary{k-}{k}{ck}}
```

### B. Alment

Istedenfor å angi tillatte orddelinger lokalt kan vi lage en liste over særlig vanskelige ord, f.eks. `\hyphenation{fri-vil-lig ord-de-ling osv.}`, og plassere den først i teksten. Navn o.l. som **ikke** skal deles, kan tas med her uten bindestrek. Unntakslisten er kumulativ, dvs. vi kan gjenta `\hyphenation{...}` så ofte vi vil, uten at de tidligere unntakene forsvinner. Hvor mange unntak den har plass til, avhenger av den T<sub>E</sub>X-implementasjonen vi bruker.

- Dersom vi lurer på hvordan T<sub>E</sub>X vil dele f.eks. *orddeling* og *frivillig*, kan vi skrive `\showhyphens {orddeling frivillig}`.

Hvis vi ikke bare bruker 7-bits CM-fonter, bør vi erstatte `\tenrm` i definisjonen av `\showhyphens` (se *The T<sub>E</sub>Xbook*, s. 364) med Times eller en annen 8 bits skrift, ellers vil alle tegn med karakterkode fra og med 128 forsvinne, istedenfor å bli vist med notasjonen  $\langle \text{heksadesimalt tall} \rangle$  ( $\leftrightarrow$  **Karakterer 2**).

I enkelte T<sub>E</sub>X-implementasjoner (f.eks. OzT<sub>E</sub>X) kan brukeren nå isteden velge å la tegn med karakterkode over 127 bli skrevet til skjerm og fil på vanlig måte, uten bruk av notasjonen  $\langle \text{heksadesimalt tall} \rangle$ .

### 3. Orddeling ved eksplisitt bindestrek

Etter karakterer med samme karakterkode som verdien av `\hyphenchar` for den aktive font (normalt 45, dvs. bindestrek), skyter  $\TeX$  automatisk inn et `\discretionary{}{}{}{}` (jf. 2.A. ovenfor), slik at orddeling ved bindestreken blir mulig. Straffen for å bryte til ny linje på et slikt punkt er lik verdien av `\exhyphenpenalty`.

I bindestreksord som ‘engelsk-amerikansk’, ‘historisk-filosofisk’ vil  $\TeX$  altså bryte til ny linje bare umiddelbart etter bindestreken. Dersom vi ønsker å åpne for andre bruddpunkter, må vi angi dem med `\-`. Merk at siden en eksplisitt bindestrek setter inn et `\discretionary{}{}{}{}`, kan  $\TeX$  finne på å bryte til ny linje også rett etter en *innledende* bindestrek, slik at man får:

havebord og - stoler	istedenfor	havebord og -stoler
-------------------------	------------	------------------------

Se nedenfor om hvordan dette kan forhindres.

### 4. Omgåelse av $\TeX$ s begrensninger

#### A. Ord med eksplisitt utligning

For å hindre at  $\TeX$  slutter å dele et ord når det treffer på en eksplisitt *kerning* i forbindelse med `\accent`, kan vi definere en makro, `\allowhyphens`, som føyer inn lim uten utstrekning, men samtidig sperrer for deling akkurat der:

```
\def\allowhyphens{\penalty 10000\hskip 0pt plus0pt minus0pt }
```

og deretter, hvis ‘å’ ekspanderer til `\accent 23a`, gi ‘å’ følgende definisjon:

```
\def å{\allowhyphens\aa\allowhyphens}
```

Uten `\allowhyphens` vil f.eks. ‘studielånsrente’ bli delt ‘stu-di-elånsrente’; med `\allowhyphens` vil  $\TeX$  oppdage ytterligere to delingspunkter: ‘stu-di-elåns-ren-te’. I det første tilfellet finner  $\TeX$  ett «ord», som det kan forsøke å dele, nemlig ‘studiel’; i det andre tilfellet finner det to «ord»: ‘studiel’ og ‘nsrente’.

#### B. Bindestrek først i ord

Når bindestreken innleder et ord, som i ‘suffikset *-het* eller ‘*jordbruksmaskiner og -redskaper*’, bør den ikke ha samme karakterkode som verdien av `\hyphenchar`. En reservebindestrek til bruk i slik tilfeller finnes på plass 127 i en del 8-bits foner (se **Tillegg 2**), slik at man kan definere en erstatningskommando, f.eks. ‘`\def ≠{\char127 \kern0pt}`’, og deretter skrive: ‘*jordbruksmaskiner og ≠redskaper*’ (dersom man ikke vil sperre for deling av ordet etter den innledende bindestreken, må `\kern0pt` fjernes fra definisjonen).

#### C. Ut- og inngangsbindestrek

I enkelte språk (f.eks. polsk og tsjekkisk) skal en eksplisitt bindestrek alltid kunne skjernes fra en fakultativ. Ved deling av sammensatte ord som ‘list-koperta’, som uten deling skrives ‘list-koperta’, skal det altså være utgangsbindestrek på linje 1 og inngangsbindestrek på linje 2. Istedenfor å skrive inn ‘`\discretionary{}{}{}{-}`’ kan man definere en kommando, f.eks. `\≠`, som gjør dette.

**D. Alternative markerte bindestreker**

T<sub>E</sub>Xs normale bindestreker, enten de er fakultative (skrevet som \-) eller eksplisitte (skrevet som -), forhindrer alle andre, ikke uttrykkelig angitte delinger av ordet. Denne begrensningen kan man komme utenom ved å definere \÷ som et alternativ til \-, og ÷ som et alternativ til bindestrek:

```
\def\÷{\penalty\10000\-\allowhyphens}
\def÷{\penalty\10000-\hskip Opt plus Opt minus Opt}
```

**5. Forebygging av orddeling****A. Lokalt:**

Vil vi forebygge at ett enkelt ord blir delt på et bestemt sted i teksten, kan vi plassere det i en \hbox eller sette \kern Opt foran det.

**B. Alment:**

Deling av ord som begynner med stor bokstav, kan forhindres ved å sette parameteren \uchyph lik null eller et negativt tall. Standardverdien er 1, slik at T<sub>E</sub>X normalt vil prøve å dele slike ord.

All orddeling kan slås av ved å sette \hyphenpenalty (eventuelt også \exhyphenpenalty, som gjelder ord med bindestrek) lik 10 000. Sannsynligheten for orddeling mellom to sider kan minskes ved å øke verdien av \brokenpenalty, som normalt er 100.

Hvis vi vil gå mer radikalt til verks, kan vi sette \pretolerance=10000. Da vil T<sub>E</sub>X overhodet ikke gjøre noe forsøk på å dele ord, med mindre det skulle støte på potensielle bindestreker (*discretionary hyphens*), angitt med \- (se ovenfor under **2. Brukerstyrt deling**). Slike eksplisitte delingspunkter tar T<sub>E</sub>X nemlig hensyn til uansett verdien av \pretolerance.

Å anvende \pretolerance på denne måten er imidlertid ikke å anbefale, for resultatet kan bli utrolig stygt! En mer hensiktsmessig fremgangsmåte er å sperre for all orddeling, som angitt ovenfor i annet avsnitt (samme resultat gir også \lefthyphenmin=100 eller \language=255<sup>36</sup>), og så sette \tolerance=9999 og \emergencystretch=10pt (f.eks.). For å unngå en uendelig strøm av deprimerende og fremfor alt tidkrevende skjermmeldinger om utilstrekkelig fylte horisontale bokser bør \hbadness i så fall økes til 10 000.

Ønsker vi å forhindre orddeling når en bestemt font er aktiv, kan vi gi fontens \hyphenchar, som normalt er lik verdien av \defaultthyphenchar, dvs. 45, en verdi utenfor intervallet [0–255]. Med \hyphenchar\tentt=-1 vil vi f.eks. avskjære all deling av ord i fonten \tentt. Vær klar over at en slik tilordning a l d r i er lokal, selv ikke inne i en gruppe. Dersom vi ønsker å gå tilbake til den opprinnelige tilstanden, må vi altså gjenopprette den eksplisitt.

**Output-rutiner**

↔ Utdatarutiner

<sup>36</sup> Forutsatt at vi ikke bruker et T<sub>E</sub>X-format som har delingsmønstre for \language255, hvilket antagelig er lite sannsynlig. NB: Hvis \language gis en verdi mindre enn 0 eller større enn 255, setter T<sub>E</sub>X \language=0.