Herziene afbreekpatronen voor het Nederlands

Piet Tutelaers

Technische Universiteit Eindhoven
Rekencentrum
rcpt@urc.tue.nl

1 Waarom nieuwe afbreekpatronen?

De CELEX patronen van Henk Penning (Rijksuniversiteit Utrecht, henkp@cs.ruu.nl) zijn de meest gebruikte voor het Nederlands. Ze zijn beschikbaar op alle goede TEX-fileservers zoals op ftp.cs.ruu.nl. Deze afbreekpatronen hebben echter de volgende bezwaren:

1. werken niet goed voor woorden met diacritische tekens (8-bits karakters)
2. breken niet af in de eerste en laatste twee letters van een woord
3. voldoen niet aan de nieuwere afbreekregels uit de 'Herziene Woordenlijst Nederlandse taal' (Groeneboekje boekje 1990, in het vervolg afgekort als GB90).

Feitelijk was punt 1 voor mij de aanleiding om de bestaande patronen te herzien. TEX3.0 biedt namelijk de mogelijkheid om te werken met 8-bits karakters en 8-bits fonts (Extended Computer Modern bijvoorbeeld). Wil je daar optimaal profijt van hebben dan heb je uiteraard ook afbreekpatronen nodig waarin deze 8-bits karakters voorkomen. Met de huidige CELEX patronen wordt Curaçaoënaar afgebroken als Cu-raça-oë-naar. Volgens het GB90 is Cu-ra-çao-e-naar de correcte spelling. We zullen later zien waarom Cu-ra-çaoë-naar het beste is wat we met TEX3.0 kunnen bereiken.

Bij het maken van de CELEX patronen zijn woorden waarin diacritische tekens voorkomen weggelaten omdat TEX toendertijd alleen ASCII codes kon verwerken. Je kunt je afvragen hoe slecht deze 7-bits patronen voor woorden met een diacritisch teken zijn. In het door mij gebruikte woordenboek komen 2090 woorden voor waarin een trema of accent aanwezig is. Deze woorden bevatten in totaal 4774 afbreekplaatsen. Hiervan worden er 54 in de verkeerde lettergreep geplaatst (1.2%) en 497 (10.4%) gemist. In het algemeen breken de CELEX patronen niet af voor een trema op 30 gevallen na. In de volgende voorbeelden zijn deze plaatsen aangegeven met een '.':

```
afge.ëist
associ.ëer
mede.ëter
mede.ïngezetene
shampoo.ën
```

De overige 24 fouten treden op in verschillende woordtypen zoals:

```
associ.és
Curaça.oënaar
ing.rediënt
procédee.tje
vol.tampère
```

Meestal wordt ontraden om woorden af te breken in de eerste twee en laatste drie letters van een woord. Er kunnen echter bijzondere omstandigheden zijn, bijvoorbeeld als je met erg smalle tekstkolommen werkt, waardoor je van deze hoofdregel wilt kunnen afwijken. Je kunt dan \lefthyphenmin en \righthyphenmin op de gewenste waarden zetten.

De afbreekregels in het GB90 zijn ten opzichte van de regels uit het Groene boekje 1954 op een aantal punten vereenvoudigd. Bastaardwoorden worden zoveel mogelijk via de Nederlandse regels afgebroken (hockey-en in plaats van hock-ey-en, cros-sen in plaats van cross-en). Er zijn echter ook een aantal veranderingen ingevoerd met het doel om de afbreekregels te vereenvoudigen:

GB54	GB90
ab-er-ra-tie	a-ber-ra-tie
ad-o-ra-tie	a-do-ra-tie
ad-e-quaat	a-de-quaat
il-lus-tra-tie	il-lu-stra-tie
leeuw-e-rik	leeu-we-rik
pres-crip-tief	pre-scrip-tief
pres-crip-tie-ve	pre-scrip-tie-ve
pros-pec-tief	pro-spec-tief
pros-pec-tie-ve	pros-pec-tie-ve
re-gis-ter	re-gis-ter
re-gis-tra-tie	re-gi-stra-tie
ver-nieuw-end	ver-nieu-wend

2 Afbreekregels in TEX

Wanneer TEX, in een poging om een mooie rechterkantlijn te maken, genoodzaakt is om een woord af te breken dan gaat dit met de volgende prioriteiten:

1. op de plek van een expliciete afbreekplaats (\discretionary)
2. volgens een uitzonderingsgeval (\hyphenation)
3. volgens de algemene afbreekregels (\patterns).

2.1 Expliciete afbreekplaats

Het Nederlands heeft een aantal situaties die netjes op te lossen zouden zijn met het `\discretionary` commando. Denk maar aan woorden als `menuutje` (`me-nu-tje`), `procédeetje` (`pro-cé-dé-tje`) en `Curaçaoënaar` (`Cu-ra-çao-e-naar`). De praktische bruikbaarheid van het discretionary commando in TEX is echter zeer beperkt omdat je het niet kunt gebruiken in uitzonderingsgevallen en patronen. Wil je de genoemde woorden volgens de Nederlandse taalregels afbreken dan rest er niets anders dan de woorden kompleet in te voeren:

```
menuutje
  me\-nu\discretionary{-}{}{u}tje
procédeetje
  pro\-cé\-d\discretionary{é-}{}{ee}tje
Curaçaoënaar
  Cu\-ra\-çao\discretionary{-}{e}{ë}naar
```

Nu kun je de pijn wel proberen te verzachten met macros als:

```
\def\uu{u\discretionary{-}{u}{u}}
\def\ee{\discretionary{é-}{}{ee}}
\catcode‘\"==\active % Babel30d methode!
\def"#1{\allowhyphens\discretionary{-}
   {#1}{\"{#1}}\allowhyphens}
```

Als je nu de woorden invoert als:

```
men\uu tje
procéd\ee tje
Curaçaoënaar
```

dan worden deze woorden alleen afbroken op de plaatsen waar de discretionaries voorkomen (aangegeven met ‘`*`’). En dus niet op alle andere plaatsen (aangegeven met ‘`-`’):

```
me-nu*tje
pro-cé-dé*tje
Cu*ra-çao*e-naar
```

De `\allowhyphens` in de `"`-macro zorgt ervoor dat een tremawoord wordt opgedeeld in drie onafhankelijke delen. Het deel vóór het trema, het trema zelf en het deel ná het trema. Hierdoor worden in het algemeen meerdere afbreekplaatsen gevonden. Er zijn echter ook gevallen waarin deze benadering tot fouten leidt, zoals in:

```
ar*chaï-see.rde
ar*chaï-see.r*den
be*doeïe-n.en*tent
be*doeïe-n.en*ten*ten
geë-ve-n.aard
on-geë-ve-n.aard
on-geë-ve-n.aarde
```

Normaal wordt niet afgebroken binnen de eerste `\lefthyphenmin` en de laatste `\righthyphenmin` karakters van een woord. Wanneer zich echter in een woord een `\discretionary`-commando bevindt dan wordt van deze regel afgeweken. Woorden als `België` en `aëroob` worden ten onrechte afgebroken voor het e-trema. In de `"`-macro wordt bovendien geen onderscheid gemaakt tussen een trema en een umlaut. Woorden als `fröbelen` en `maïs` gaan dan ook fatikaal fout.

Het Nederlands kent ook woorden waarin een apostroph wordt gebruikt (`baby'tje` en `juffertje-in-'t-groen`). Deze woorden worden niet altijd goed afgebroken (`baby.'.tje`).

2.2 Uitzonderingen

Woorden die met de afbreekpatronen foutief worden afgebroken kunnen afzonderlijk worden behandeld. Dit kan voor de duur van één document (preamble van LATEX) of voor alle LATEX-runs als je deze uitzonderingen met IniTEX meevertaalt.

De file `GB90.b+m` bevat de woorden uit de gebruikte woordenlijst die fout gaan met `GB90.8pat`. Plaatsen aangegeven met ‘`.`’ gaan fout (er komt ten onrechte een afbreekstreepje), plaatsen met ‘`*`’ aangegeven gaan goed en plaatsen gemarkeerd met ‘`-`’ worden gemist.

Uit `GB90.b+m` zijn de volgende ‘exceptions’ gedestilleerd:

```
% Exception list for GB90.8pat (July ’93)
\hyphenation{
aan-pers-te
acht-en-der
acht-en-ders
acht-en-der-tig
acht-en-der-tig-ste
ant-arc-tis
be-scherm-en-gel
be-scherm-en-ge-len
don-der-aal
drie-ster
gast-rol-len
ge-laats-trek-ken
han-dels-taal
ket-ting-ste-ken
lands-taal
lui-ste
mi-nis-ters-por-te-feuil-le
mi-nis-ters-por-te-feuil-les
moet-je
pa-ling-ste-ken
schel-linkje
spie-gel-ei
ti-chel-aar-de
vier-en-der-tig
vier-en-der-tig-ste
}
```

2.3 Afbreekpatronen

De patronen gebaseerd op het GB90 zijn gemaakt inclusief de diacritische tekens. Vanwege de problemen die optreden met het `\discretionary` commando worden woorden niet afgebroken voor tremas. Onregelmatige woorden (`menuutje`) en woorden met verschillende betekenissen (`buur-tje` en `buurt-je`, `wets-taal` en `wet-staal`) worden ook niet afbroken op plaatsen waar deze afbreking niet eenduidig is.

De patronen zijn gemaakt met `patgen2` (een 8-bits versie van `patgen`) uitgaande van een woordenlijst die is afgeleid van het Groeneboekje '54. Deze woorden zijn afgebroken met de CELEX patronen en vervolgens in een aantal slagen met de hand aangepast aan de spelling van GB90 (zeer tijdrovend!). Wanneer het mij lukt om alsnog de officiële woorden van het GB90 te lenen van het Instituut voor Lexicologie dan zal ik de patronen opnieuw berekenen. Tot dan blijven deze patronen van kracht.

De gebruikte woordenlijst bevat 163861 woorden met in totaal 361780 afbreekplaatsen. De meeste woorden stammen uit het GB54. Ook de woorden uit het CELEX bestand die fout werden afgebroken met de CELEX patronen of waarin bepaalde afbreekplaatsen gemist werden, zijn aan deze lijst toegevoegd (plus minus 2000). In tabel 1 zijn de patgen[1] resultaten bij elkaar geplaatst. Zie [7] voor de betekenis van de waarden uit de eerste drie kolommen.

level	lengte	parameters	patronen	goed	fout	mis
1	1(1)4	1 2 20	832	97.48	11.13	2.52
2	1(1)4	2 1 4	1196	92.29	0.55	7.71
3	1(1)5	1 1 1	3672	99.86	3.87	0.14
4	1(1)6	3 2 1	2396	87.87	0.01	2.13
5	1(1)8	1 1000 1	2167	100.00	0.01	0.00
					(24)	(7)

Tabel 1: *Patgen resultaten van GB90.8pat*

	GB90	CELEX (GB54)
patterns	11803	12698
trie ops	185	188
trie words	9985	10335
bytes	52113	51748

Tabel 2: *Geheugen beslag van CELEX en GB90 patronen*

De resultaten zijn zeer goed. Het aantal foutieve afbreekplaatsen is maar 24 terwijl slechts 7 plaatsen gemist worden. De CELEX patronen leverden 210 (0.02%) foutieve en 5877 (0.49%) gemiste afbreekplaatsen op. Daarbij moet worden opgemerkt dat het CELEX bestand fouten bevatte die niet zijn gekorrigeerd.

De afbreekpatronen zijn zo ontworpen dat ze een minimaal geheugenbeslag leggen op TEX. Als we de CELEX patronen en de nieuwe GB90 patronen vergelijken dan zien we uit tabel 2 dat dat onderling weinig verschilt.

3 Conclusie

Het model dat door TEX wordt gebruikt voor het afbreken van woorden levert voor het Nederlands zeer acceptabele resultaten op. Echter woorden waarin een trema voorkomt of woorden met een veranderende lettergreep kunnen niet goed worden behandeld vanwege de beperkingen die TEX oplegt aan het discretionary commando. Omdat de methode die nu gebruikt wordt in `dutch.sty` allerlei foute resultaten geeft, zijn bij het genereren van de herziene patronen afbreekplaatsen voor tremas, in veranderende lettergrepen en in woorden waarin de afbreking niet eenduidig is, weggelaten. Woorden met een koppelteken erin worden alleen op die plaats afgebroken omdat dit impliciet wordt afgebeeld op een discretionary. Wanneer zo'n koppelteken lange woorden verbindt dan worden die niet afgebroken zoals 'juffertje' in 'juffertje-in-'t-groen' en 'tolletje' in 'a-al-tolletje' En er nog geen goede oplossing voor woorden met een apostroph.

Samengevat:

- vóór een trema kan beter niet worden afgebroken
- woorden waarvan de afbreking onregelmatig is (veranderende lettergreep) kunnen beter niet worden af-

[1] Een run op een DEC/OSF1 alpha kost ongeveer 45 minuten rekentijd. Op mijn 486DX/33Mhz systeem thuis met 386BSD UNIX 1 uur en 17 minuten.

gebroken in die lettergreep
- woorden waarin koppeltekens '-' voorkomen worden alleen op die plaatsen afgebroken
- woorden met een apostroph worden niet altijd goed afgebroken.

4 De beschikbaarheid

Realiseer je dat deze patronen alleen zin hebben als je overstapt naar 8-bits fonts. Je kunt dan in je LaTeX invoer rechtstreeks een i-trema plaatsen zonder dat je daarvoor het 7-bits `\"{\i}` commando gebruikt. Diverse systemen ondersteunen ISOlatin1 zodat je deze tekens met je editor rechtstreeks kunt invoeren.

Wil je de patronen eens uit proberen? Ze zijn beschikbaar via ftp op `ftp.urc.tue.nl` in `tex/8bit/newhyph`. Ook als je ze niet onmiddellijk in LaTeX wilt gaan gebruiken, kun je ze uitproberen. Er is namelijk een programma `hyphenate.c` bij aanwezig waarmee je interaktief woorden kunt afbreken. Dit programma accepteert als argument de naam van een file waarin zich afbreekpatronen en uitzonderingen bevinden en bepaalt van de ingevoerde woorden hun afbreekplaatsen. Met de optie `-L`*left* en `-R`*right* kun je opgeven hoeveel letters aan het begin en eind van een woord niet mogen worden afgebroken. Deze waarden zijn standaard ingesteld op *left*=2 en *right*=3. Met de `-d` optie krijg je extra tussenresultaten zoals de patronen die zijn gebruikt voor het berekenen van de afbreekplaatsen (`-d1`). Hier is een kleine demonstratie van de mogelijkheden van `hyphenate` [2]:

```
hyphenate -d1 -L1 -R1 GB90.8pat
Curaçaoënaar
3cu
2ur
u1ra
1ç
a3o
4oë
4ë
ë3na
a4a4
4ar.
4r.
.3C2u1ra1ça4o4ë3na4a4r.
Cu-ra-çaoë-naar
```

In `tex/8bit/newhyph` tref je de volgende files aan:

```
(8) patronen.tex
      (deze file, voor de verwerking heb
       je NFSS en de DC-fonts nodig)
    patronen.ps.z
      (PostScript uitvoer van
       patronen.tex)
(8) GB90.8pat
      (de 8-bits patronen)
    GB90.7pat
      (de 7-bits gedaante van GB90.8pat)
(8) GB90.b+m
      (bad+missed woorden)
(8) WdNT.G-B.z
      (de gebruikte woor-den-lijst)
    patgen.out
      (de log van patgen)
    patgen.in
      (de invoer waarden van patgen)
(8) dutch.tra
      (de transition file voor patgen)
    dcdutch.sty
      (een 8-bits versie van dutch.sty;
       zorgt ervoor dat voor tremas niet
       wordt afgebroken)
(8) test.tex
      (een plain TeX test file; legt
       diverse problemen van het
       afbreken in het Nederlands bloot!
       Voor de verwerking heb je NFSS en
       de DC-fonts nodig)
    src
      (deze directory bevat de sources
       van een aantal hyphenation
       tools zoals hyphenate.c)
    src/README
      (bevat nadere gegevens hierover)
```

De files voorafgegaan door een (8) bevatten 8-bits karakters. En de files met achtervoegsel `.z` zijn gecomprimeerd met `gzip`.

References

[1] Peter Breitenlohner. *The eight bit patgen extensions.* Unpublished, patgen2.web sources

[2] *CELEX woordenbestand.* Centre for Lexical Information. Universiteit van Nijmegen

[3] Donald E. Knuth. *Computers and Typesetting Vol. A–E.* Addison–Wesley, Reading, MA, 1987–1991

[3a] Vol A: *The TeXbook*, 11. ed. 1991

[3b] Vol B: *TeX: The Program*, 4. ed. 1991

[4] Yannis Haralambous. *Syntax of translation file.* Unpublished, distributed via BITNET

[5] *Herziene Woordenlijst Nederlandse taal*, SDU, Den Haag, 1990, vijfde oplage

[6] Leslie Lamport. *LaTeX – A Document Preparation System* Addison–Wesley, Reading, MA, 1985

[7] Franklin Mark Liang. *Word Hy-phen-a-tion by Com-put-er*, Department of Computer Science, Stanford University, 1983

[8] Norbert Schwarz, *DC-fonts.* Unpublished DC METAFONT sources. Rechenzentrum Ruhr-Universitaet Bochum, Germany

[2] Zie appendix H van het TeXbook voor meer informatie over afbreekpatronen.