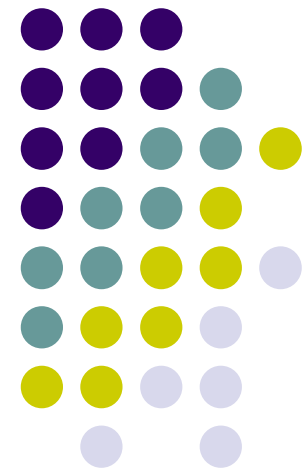


Automatische Sil-ben-tren-nung in TeX

Steve Reich
12.06.06





Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- einige Befehle
- Probleme und Lösungen



Quellen

- „Einführung in TeX“
Norbert Schwarz
- „TeX für Fortgeschrittene“
Wolfgang Appelt
- „The TeXbook“
Donald E. Knuth



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Einige Befehle
- Probleme und Lösungen

Einführung



- Die meisten Menschen trennen nach Gefühl!
- Wenn man sich unsicher ist oder es genau wissen will, hat man drei Möglichkeiten:
 - Ein anderes Wort mit gleicher Bedeutung suchen.
 - Das Wort nicht trennen. (auf die neue Zeile schreiben oder quetschen)
 - In einen Wörterbuch nachschauen.



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Befehle
- Probleme und Lösungen

Vorgehensweise von TeX



- TeX versucht Silbentrennung zu vermeiden. Das bedeutet solange es die gesetzten Grenzen zulassen, werden Worte nicht getrennt.
- Intern wird das Trennen sowie andere Unschönheiten über den Zeilenumbruch geregelt. Es werden für alle Unschönheiten Minuspunkte aufsummiert und die Lösung mit den wenigsten Minuspunkten gewählt.



Vorgehensweise von TeX

- `\hyphenpenalty = 10 000`
 - Es wird nicht mehr getrennt.
- `\doublehyphendemerits = 10 000`
 - Zwei aufeinander folgende Trennungen kommen nicht vor.
- `\finalhyphendemerits = 5000`
 - In der vorletzte Zeile eines Absatzes wird nicht getrennt.
- `\brokenpenalty = 100`
 - In der letzte Zeile der Seite wird nicht getrennt.



Vorgehensweise von TeX

- Jedes `penalty` das 10 000 oder größer ist, ist so groß das TeX niemals dort trennen wird.
- Jedes `penalty` das -10 000 oder kleiner ist, ist so klein das TeX immer dort trennen wird.
- Das `\nobreak` Makro in plain TeX ist ein `'\penalty10000'`.



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Befehle
- Probleme und Lösungen



Der Algorithmus

- von Frank M. Liang 1980-1982
 - schnell
 - findet fast alle erlaubten Trennungen
 - findet fast keine falschen Trennungen
 - beansprucht relativ wenig Speicher
 - ist sehr flexibel (z.B. auf andere Sprachen übertragbar)



Der Algorithmus

- Der Algorithmus benutzt eine Liste von „hyphenation patterns“.
- Diese Muster bestehen aus einer Buchstabenfolge mit Zahlenwerten zwischen 0 und 5.

0 h 0 y 3 p 0 h 0

- Die Zahlen geben die Trennbarkeit an:
Gerade bedeutet Trennung verboten, ungerade bedeutet Trennung erlaubt.



Der Algorithmus

.hyphenation.

- Daraus werden jetzt alle Buchstabenfolgen bis zu einer bestimmten Länge ermittelt.

.h hy yp ph he en na at ti io on n.

.hy hyp yph phe hen ena nat ati tio ion on.

usw.

- Es wird überprüft welche der entstehenden Muster in der Liste (hyphenation patterns) vorkommen.



Der Algorithmus

- Nur gefundene Muster werden benutzt:
 - 1. Muster 0 h 0 y 3 p 0 h 0
 - 2. Muster 0 h 0 e 2 n 0
 - 3. Muster 0 h 0 e 0 n 0 a 4
 - 4. Muster 0 h 0 e 0 n 5 a 0 t 0
 - 5. Muster 1 n 0 a 0
 - 6. Muster 0 n 2 a 0 t 0
 - 7. Muster 1 t 0 i 0 o 0
 - 8. Muster 2 i 0 o 0
 - 9. Muster 0 o 2 n 0
 - max. Wert: 0 h 0 y 3 p 0 h 0 e 2 n 5 a 4 t 2 i 0 o 2 n 0



Der Algorithmus

0 h 0 y 3 p 0 h 0 e 2 n 5 a 4 t 2 i 0 o 2 n 0

- Trennen an ungeraden Stellen ergibt:
hy-phen-ation
- Die Trennmuster für Englisch wurden auf der Basis eines Wörterbuchs mit 50 000 Wörtern erstellt. Daraus ergaben sich 4447 Muster mit einer max. Länge von 8 Zeichen.
- Damit lassen sich 89% der Trennung aus dem Wörterbuch reproduzieren.



Der Algorithmus

- Die max. Länge der Zeichenketten von 8 Zeichen ist Ergebnis vieler Versuche.
- Längere Zeichenketten ergeben mehr Muster und verlangsamen so den Trennalgorithmus.
- Kurze Zeichenketten (2-3 Zeichen) haben zur Folge, dass nicht mehr alle wichtigen Trennstellen gefunden werden.



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Befehle
- Probleme und Lösungen

Das PATGEN- Programm



- Dieses Programm dient zur Erstellung der Liste mit Trennmuster.
- Als Eingabe dient ein Wörterbuch mit Trennstellen (auch gebeugte Formen).
- z.B. „hy-phen-a-tion“
- Nun ermittelt das Programm alle Buchstabenfolgen bis zu einer bestimmten Länge und stellt für alle, durch Vergleich, die Trennbarkeit fest.
- Am Ende gibt es eine Liste der Trennmuster aus.

Das PATGEN- Programm



- Das PATGEN- Programm ist in der Standardinstallation dabei.
- Die Eingabe erfordert jedoch sehr viel Arbeit und Fachkenntnisse.
- Die Kontrolle der Qualität der erstellten Pattern Liste erfordert umfangreiche Tests.
- Man sollte die Erstellung von Pattern Listen mit dem PATGEN- Programm Profis überlassen.



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- **Befehle**
- Probleme und Lösungen



Befehle

- Ab Version 3 trennt TeX auch nach verschiedenen Sprachen. Die Sprache wird durch das Register `\language` festgelegt.
(i.d.R. 0 = Englisch, theoretisch 256 Sprachen)
- Es kann auch innerhalb eines Dokuments nach mehreren Sprachen getrennt werden.
- Wird eine Sprache gesetzt zu der keine Trennregeln geladen wurden, wird nicht getrennt.

Befehle



- Trennvorgaben:
 - Ein Wort kann explizit durch den Vortrenner ,\-' getrennt werden. Es wird dann aber nur an diesen Stellen getrennt.
- Ausnahmelexikon:
 - Ein Eintrag wird durch folgenden Befehl realisiert.
`\hyphenation {Ur-instinkt}`
 - Dadurch kann z.B. kein Urin - stinkt entstehen, da diese Trennung Vorrang hat.
 - Die Größe ist jedoch begrenzt.



Befehle

- Sich Trennungen ausgeben lassen:
 - `\showhyphens {Testworte...}`
- Ein bestimmtes Trennschema festlegen:
 - `\discretionary { <Text vor der Trennung> }
 { <Text nach der Trennung> }
 { <Text ohne Trennung> }`



Gliederung

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Befehle
- Probleme und Lösungen



Probleme und Lösungen

- Deutsche Umlaute
(\ "a, \ "o, \ "u, \ "A, \ "O, \ "U)
sind ein Problem, da die Anführungsstriche nicht als Buchstabe eines Wortes erkannt werden, können Wörter mit Umlauten nach diesen nicht mehr getrennt werden.
- Dadurch ergibt sich ein Problem für Wörter wie:
Öffentlichkeitsarbeitreferat
- Lösung auf TeX-Ebene:
umdefinieren des \ "- Befehls



Probleme und Lösungen

- `\def"#1{{\accent"7F #1\penalty10000\hskip 0pt plus 0pt}}`

Dieser Befehl bewirkt, dass ein Wort nach dem Umlaut zu Ende geht und ein neues Wort beginnt, das auch wieder getrennt wird. Dabei wird durch `\penalty10000` ein Zeilenumbruch zwischen den so geschaffenen Teilwörtern verhindert.

- Auf LaTeX-Ebene können die Umlaute mittels eines `includepackage` als Buchstaben zur Verfügung gestellt werden.
- z.B. `\includepackage [T1] {fontenc}`



Probleme und Lösungen

- Problem „ß“:
 - Im Englischen gibt es kein ß, deshalb erkennt TeX es nicht als Buchstaben.
- Lösung:
 - dem ß einen Wert zuordnen: `\lccode'31='31`
 - oder fonts mit Umlauten benutzen: z.B. DC-fonts
- Kontextabhängige Trennung:
 - z.B. erb-lich, er-blich, Stau-becken, Staub-ecken
 - kann nicht automatisch getrennt werden



Probleme und Lösungen

- richtige aber unschöne Trennungen:
z.B. Anal – phabet oder Urin – stinkt
Lösung: Eintrag ins Ausnahmelexikon
- Wörter verändern beim Trennen ihre Schreibweise:
z.B. backen, bak-ken, Brennessel, Brenn-nessel
Lösung: `\def\ck{\discretionary{k-}{k}{ck}}`
alte Rechtschreibung!
Das discretionary findet jedoch immer noch Verwendung.

Ende

- Quellen
- Einführung
- Vorgehensweise von TeX
 - Der Algorithmus
 - Das PATGEN Programm
- Befehle
- Probleme und Lösungen

