# Bretter, die die Welt bedeuten

## *Hyphenation Exception Log* für deutsche Trennmuster

Werner Lemberg

Dem englischen Vorbild der *TEX Users Group* folgend soll dieser Artikel ein Aufruf zur Mitarbeit sein, die deutschen Trennmuster zu verbessern.

### Einleitung

Seit vielen Jahren werden in unserer Schwesterzeitschrift *TUGboat* in unregelmäßigen Abständen Ergänzungen zu den originalen US-englischen Trennmustern von TEX veröffentlicht. Barbara Beeton, die Verfasserin dieser *Hyphenation Exception Logs*, ruft dazu auf, von TEX falsch getrennte Wörter ihr zu melden. Die letzte Ausgabe erschien im Doppelheft 21/1-2 (2000), S. 50–51, und umfasst, wenn man Plural- und Konjugationsformen mitrechnet, rund 890 Einträge.

Ich plane, das Gleiche für die deutschen Trennmuster zu tun, und stelle mich hiermit als Verwalter dieser Trennmusterausnahmen zur Verfügung.

### Der Trennalgorithmus von TEX

Ohne genauer ins Detail gehen zu wollen, müssen doch einige Bemerkungen vorausgeschickt werden, damit interessierte Leser wissen, ob es Sinn macht, fehlende oder inkorrekte Trennungen zu melden.

TEX fügt in einem Wort keine Trennstellen für die ersten `\lefthyphenmin` und für die letzten `\righthyphenmin` Buchstaben ein. Für die deutschen Trennmuster ist `\lefthyphenmin=2` und `\righthyphenmin=2`. Wörter mit weniger als vier Buchstaben werden also gar nicht getrennt. Genaueres findet sich in *The TEXbook*, Seite 454.

Im Gegensatz zum US-Englischen gibt es im Deutschen in der Regel keine mehrdeutigen Trennstellen wie z. B. *rec-ord* und *re-cord*; gleichgeschriebene Komposita, die sich aus verschiedenen Wörtern zusammensetzen (Wachstube, Staubecken), werden in den deutschen Trennmustern zugunsten der tatsächlich benutzten Varianten aufgelöst, also Wach-stu-be und Stau-becken. Ich plane, solche Wörter im Katalog der Trennmusterausnahmen nur zu erwähnen. Die ersten mir bekannten echten Einträge sind „spielende“ und „Druckerzeugnis“, da ohne Kontextanalyse nicht entschieden werden kann, ob „Spiel-en-de“ oder „spie-len-de“ beziehungsweise „Druck-er-zeug-nis“ oder „Dru-cker-zeug-nis“ gemeint ist (die deutschen Trennmuster erzeugen „spielen-de“ und „Drucker-zeug-nis“) und beide Trennungsmöglichkeiten Sinn machen. Aus gegebenem Anlass kommt noch ein Wort hinzu, das falsch getrennt wird: „Trenn-al-go-rith-mus“ statt „Tren-nal-go-rith-mus“.

Zusätzlich zu beachten ist, dass bestimmte Trennstellen in manchen zusammengesetzten Wörtern unterdrückt werden müssen, um beim Lesen nicht falsche Assoziationen zu erzeugen. Das wohl bekannteste Beispiel ist das Wort „Ur-in-stinkt“, wo aus naheliegenden Gründen die Trennung „Urin-stinkt“ inakzeptabel ist.

## Haupt- und Nebentrennstellen

Im Deutschen kommt der Unterscheidung zwischen Haupt- und Nebentrennstellen aufgrund der großen Anzahl von zusammengesetzten Wörtern eine viel größere Bedeutung zu als im Englischen.[1] Leider ist es mit TEX nicht möglich, ohne manuelle Eingriffe in einem Dokument die Trennstellen zu gewichten. Das gleiche gilt (derzeit?) auch für ε-TEX und Omega. Es macht daher keinen Sinn, grammatikalisch korrekte, aber „unschöne“ Trennungen zu melden.

In solchen Fällen sollte `\-` verwendet werden, um ungewollte Trennungen zu unterbinden. Der in `german.sty` definierte Befehl `"-` ist dazu nicht geeignet, da er ja bekanntlich zusätzliche Trennstellen einfügt, ohne andere mögliche Trennstellen im Wort zu unterbinden.[2]

[1] Beispielsweise Ne-ben—trenn—stel-len; lange Bindestriche zeigen die bevorzugten Haupttrennstellen an.

[2] Das ist übrigens so nicht ganz richtig. In ungünstigen Fällen kann es durchaus vorkommen, dass durch Einfügen von `"-` falsche Trennungen erzeugt oder bereits gefundene Trennstellen nicht erkannt werden, da TEX jetzt Trennstellen für die Wortteile links und rechts von `"-` sucht anstatt für das gesamte Wort.

Es gibt mehrere Ansätze, um dieses Problem automatisiert zu lösen:

- Die *Sichere sinnentsprechende Silbentrennung* (*SisiSi*, CTAN:`/systems/unix/sisisi`). Vor rund zehn Jahren von W. Barth, H. Steiner und H. Herbeck am Institut für Praktische Informatik der Technischen Universität Wien entwickelt, implementiert *SisiSi* einen anderen Trennalgorithmus für TEX, der speziell für die deutsche Sprache geeignet ist. Ein Bericht dazu ist in „Die TEXnische Komödie“ 1/1992, erschienen.

- Matthias Clasen hat 1998 einige Erweiterungen für den originalen Trennalgorithmus von TEX geschrieben (das Paket ist erhältlich von CTAN:`/systems/tex-extensions/clasen`). In seiner Implementation werden u. a. Trennstellen gewichtet, wobei maximal zehn verschiedene Gewichtsklassen zur Verfügung stehen.

- Petr Sojka beschreibt in dem Artikel *Notes on Compound Word Hyphenation in TEX* (erschienen in *TUGboat* 16/3 (1995), S. 290–296) eine weitere interessante Möglichkeit, Haupttrennstellen zu bevorzugen, ohne sie allerdings zu implementieren.

  Bekannterweise versucht TEX bis zu dreimal, einen Absatz zu formatieren. Im ersten Durchgang wird versucht, ohne Wortabtrennungen auszukommen. Im zweiten Durchgang benützt TEX Trennmuster, falls der erste Durchgang nicht befriedigende Ergebnisse gebracht hat. Der dritte Durchgang wird nur ausgeführt, falls `\emergencystretch` einen positiven Wert hat. TEX versucht dann, *badness*-Werte zu verringern; das Ergebnis sind Zeilen mit übergroßen Wortzwischenräumen.

  Sein Vorschlag ist nun, vor dem zweiten Durchgang einen zusätzlichen TEX-Lauf auszuführen, wobei spezielle Trennmuster verwendet werden, welche nur Haupttrennstellen enthalten. Erst danach sollten zusätzlich Nebentrennstellen berücksichtigt werden (enthalten in einer *zweiten* Trennmusterdatei), wobei Trennstellen aus dem Extra-Durchgang erhalten bleiben und mit einem höheren Gewicht versehen werden.

  Des weiteren schlägt er vor, dass TEX bei den im Extra-Durchgang ermittelten Haupttrennstellen zusätzlich das Zeichen `\compoundwordchar` einfügt (die `T1`-Kodierung enthält es an Position 23). Dadurch würde der Befehl `"|` von `german.sty`, um z. B. „Auflage“ korrekt als „Auflage“ darzustellen, nur noch in Ausnahmefällen nötig sein, da TEX automatisch eine inkorrekte Ligatur unterbinden kann.

Jede der oben erwähnten Ideen hat Nachteile. *SisiSi* wird meines Wissens nicht mehr gepflegt und ist auch nicht weit verbreitet. Clasens Erweiterung benötigt gewichtete Trennmuster, die erst zu erstellen sind. Sojkas Vorschlag wurde bis jetzt gar nicht implementiert, und auch hier fehlen separate Trennmuster für Haupt- und Nebentrennstellen.

## Verbesserungen

Es gibt eine sehr einfache Möglichkeit, Trennungen in einem deutschen Text zu verbessern, indem man `\lefthyphenmin` und `\righthyphenmin` zu Beginn eines Dokuments auf den Wert 3 setzt. Die Ausnahme ist mehrspaltiger Text mit kurzen Zeilenlängen, wo selbst ungünstige Umbrüche immer noch besser als halbleere Zeilen sind.

Weiterhin kann man Umbrüche bei abgeteilten Wörtern erschweren; das interne TEX-Register dafür ist `\hyphenpenalty`. Standardmäßig setzen TEX und LATEX es auf 50, jedoch kann für längere Absätze ein Wert von 1000 oder mehr Sinn machen. Kürzere Absätze können dadurch normalerweise nicht beeinflusst werden, da TEX zu wenig Möglichkeiten hat, Absätze verschieden zu formatieren.

Eine kleine Randbemerkung: Um Wortumbrüche gänzlich zu unterdrücken, sollte nicht `\hyphenpenalty=10000`, sondern `\lefthyphenmin=65` benutzt werden, was den gleichen Effekt hat, aber deutlich schneller ist.

## Ausblick

Sobald eine genügend große Zahl von Trennmusterausnahmen zusammengekommen ist, wird ein Folgeartikel in „Die TEXnische Komödie“ erscheinen. Im Weiteren könnten die Ausnahmen in eine neue Version der deutschen Trennmuster aufgenommen werden – im Gegensatz zu Knuths `hyphen.tex` werden diese ja weiter gepflegt und bei Bedarf verbessert.