dehyph-exptl*

Experimentelle Trennmuster für die deutsche Sprache

Die deutschsprachige Trennmustermannschaft

8. Juli 2008

Abstract

This package provides new, experimental hyphenation patterns for the German language, covering traditional and reformed orthography. The patterns can be used with packages `Babel` and `hyphsubst` from the OBERDIEK BUNDLE. Project-URL is http://groups.google.de/group/trennmuster-opensource.

Zusammenfassung

Dieses Paket enthält experimentelle Trennmuster für die traditionelle und reformierte deutsche Rechtschreibung. Die Trennmuster können mit den Paketen `Babel` und `hyphsubst` aus dem OBERDIEK-BÜNDEL verwendet werden.

Inhaltsverzeichnis

*This document describes the `dehyph-exptl` package v0.13.

1. Einleitung

Der in TEX implementierte Trennalgorithmus arbeitet musterbasiert [Lia83]. Prinzipiell können mit einem solchen Algorithmus nicht alle möglichen Wörter korrekt getrennt werden. Die Qualität der Worttrennung einer Sprache wird jedoch maßgeblich von der Qualität der Wortliste beeinflusst, aus der die verwendeten Trennmuster berechnet wurden.

Leider ist die Wortliste, die den herkömmlichen Trennmustern für die traditionelle deutsche Rechtschreibung zugrundeliegt, verschollen. Dies hat mehrere Konsequenzen:

- Die Trennmuster lassen sich nicht reproduzieren. Die Pflege der herkömmlichen Trennmuster ist daher schwierig bis unmöglich. Für freie Software ist dies kein zufriedenstellender Zustand.
- Die Qualität der ursprünglichen Wortliste und die der Trennmuster kann nur schlecht eingeschätzt werden. Für die traditionelle Rechtschreibung existiert jedoch inzwischen eine Ausnahmeliste mit über 3500 korrigierten Trennungen (Datei `dehyphtex.tex`).
- Für die Berechnung der Trennmuster für die reformierte deutsche Rechtschreibung stand keine Wortliste zur Verfügung. Diese Trennmuster entstanden durch manuelle Anpassung der Trennmuster für die traditionelle Rechtschreibung an die reformierten Regeln. Aus diesem Grund ist die Qualität der Trennmuster für die reformierte Rechtschreibung noch etwas schlechter als die der Trennmuster für die traditionelle Rechtschreibung.
- Eine Besonderheit der (deutsch)schweizerischen Rechtschreibung, der konsequente Ersatz des »ß« durch »ss«, wird mit den herkömmlichen Trennmustern nicht berücksichtigt. Dies beeinflusst auch den Versalsatz nach deutscher Rechtschreibung, wenngleich hier auf Trennungen möglichst verzichtet werden sollte.

Das Projekt *Freie Wortlisten und Trennmuster für die deutsche Sprache* hat sich deshalb das Ziel gesetzt, neue, hochqualitative Trennmuster für die Benutzung in TEX und OpenOffice zu schaffen.

Den experimentellen Trennmustern dieses Pakets liegt eine Wortliste mit den etwa fünfhunderttausend häufigsten deutschen Wörtern in deutscher und (deutsch)schweizerischer Schreibweise zugrunde. Diese Liste ist vermutlich

erheblich umfangreicher als die ursprüngliche Wortliste. Außerdem wurden Worthäufigkeiten in der ursprünglichen Wortliste wahrscheinlich überhaupt nicht berücksichtigt.

Mit den vorliegenden Trennmustern sollte für nicht-fachsprachliche Wörter eine sehr gute Trennqualität erreicht werden. Insbesondere sollte sich die Trennung häufig auftretender zusammengesetzter Wörter verbessern.

Aktuelle Trennmuster sind im Dateibereich unter der Projekt-URL[1] oder im CTAN erhältlich. Weitere Informationen sowie eine Aufgabenliste können der Projektbeschreibung entnommen werden.

Dieses Projekt benötigt Deine Hilfe!

2. Verwenden der Trennmuster

Die Installation der experimentellen Trennmuster ist in der Datei `INSTALL` beschrieben. Sie können mit den Paketen `Babel` und `hyphsubst` aus dem OBERDIEK-BÜNDEL aktiviert werden.

Das folgende Beispiel zeigt eine LaTeX-Präambel für die Aktivierung der experimentellen Trennmuster für die reformierte Rechtschreibung. Beachte, `<datum>` ist durch das bei der Installation angegebene Datum in ISO-Notation (`JJJJ-MM-TT`) oder die Zeichenkette `latest` zu ersetzen!

```
\RequirePackage[ngerman=ngerman-x-<datum>]{hyphsubst}
% \RequirePackage[ngerman=ngerman-x-latest]{hyphsubst}
\documentclass{article}
\usepackage[T1]{fontenc}
\usepackage[ngerman]{babel}
```

Die folgende Variante erleichtert das schnelle Umschalten zwischen verschiedenen Trennmustern im Editor. Weitere Hinweise können der Dokumentation des Pakets `hyphsubst` entnommen werden.

```
\RequirePackage{hyphsubst}
\documentclass{article}
\usepackage[T1]{fontenc}
% \HyphSubstLet{german}{german-x-<datum>}
% \usepackage[german]{babel}
\HyphSubstLet{ngerman}{ngerman-x-<datum>}
\usepackage[ngerman]{babel}
```

[1] http://groups.google.de/group/trennmuster-opensource?hl=de

traditionelle Rechtschreibung		*reformierte Rechtschreibung*	
herkömmlich	experimentell	herkömmlich	experimentell
lös-te	lö-ste	lös-te	lös-te
Fas-sa-de	Fas-sa-de	Fassa-de	Fas-sa-de
mo-d-ern-ste	mo-dern-ste	mo-d-erns-te	mo-derns-te
Abend-stern	Abend-stern	Abends-tern	Abend-stern
Mor-dop-fer	Mord-op-fer	Mor-dop-fer	Mord-op-fer

Tabelle 1: Trennvarianten

Ob die experimentellen Trennmuster korrekt aktiviert werden, kann mit dem folgenden Beispiel getestet werden. Die Ausgabe für die traditionelle und reformierte Rechtschreibung mit herkömmlichen und experimentellen Trennmustern ist in Tabelle 1 zusammengefasst.

```
\begin{document}
\showhyphens{löste Fassade modernste Abendstern Mordopfer}
```

Diese Trennmuster befinden sich im experimentellen Status. Sie können jederzeit durch umbruchinkompatible Versionen ersetzt und vom CTAN *oder aus TEX-Verteilungen entfernt werden. Sie sind daher nicht für Anwendungen geeignet, die einen dauerhaft stabilen Umbruch erfordern. Ausgiebige Tests sind erwünscht!*

3. Trennregeln und Konventionen

Es werden zwei Trennmuster bereitgestellt. Die Trennmuster für die traditionelle Rechtschreibung orientieren sich an den verbindlichen Regeln des Dudens in der Fassung von 1991. [Wis91] Die Trennmuster für die reformierte Rechtschreibung orientieren sich an den amtlichen Regeln für die Rechtschreibung der deutschen Sprache in der Fassung von 2006. [Rato6, Wiso6]

Die Regeln lassen gewisse Freiheiten bei der Schreibung und Trennung von Wörtern zu. Da sich solche Freiheiten nicht ohne weiteres auf die maschinelle Worttrennung übertragen lassen, wurden die folgenden Konventionen getroffen. Hauptsächlich betreffen diese die reformierte Rechtschreibung, die zusätzliche Freiheiten eingeführt hat. Die linke (grüne) Spalte zeigt jeweils die Trennung mit den experimentellen Trennmustern, die rechten (roten) Spalten zeigen alternative oder unerwünschte Trennungen.

Beachte, die folgenden Abschnitte enthalten keine vollständige Aufstellung der Silbentrennregeln. Diese sind den entsprechenden Regelwerken zu entnehmen.

3.1. *Traditionelle Rechtschreibung*

T1 In Ableitungen von Namen auf *-ow* wird die Nottrennung der Ableitungssilben *-er*, *-ern*, *-ers* unterdrückt [Wis91, R 180]:

Tel-tower	Tel-tow-er
Trep-towern	Trep-tow-ern
Pan-kowers	Pan-kow-ers

T2 Sinnentstellende und irreführende Trennungen werden möglichst vermieden [Wis91, R 181]:

An-alpha-bet	Anal-phabet
Kaf-ka-kenner	Kafkaken-ner
Tal-entwäs-se-rung	Talent-wässerung

3.2. *Reformierte Rechtschreibung*

R1 Sinnentstellende und irreführende Trennungen werden möglichst vermieden [Rato6, Wiso6, § 107]:

An-alpha-bet	Anal-phabet
Kaf-ka-kenner	Kafkaken-ner
Tal-entwäs-se-rung	Talent-wässerung

R2 In Fremdwörtern bleiben die Buchstabengruppen *bl, pl, fl, gl, cl, kl, phl; br, pr, dr, tr, fr, vr, gr, cr, kr, phr, thr; chth; gn, kn* im allgemeinen ungetrennt, nicht jedoch *str* [Rato6, Wiso6, § 112] i. V. m. [Wis91, R 179]:

Ar-thri-tis	Arth-ri-tis	*aber:*	In-dus-trie	In-du-strie	In-dust-rie
Di-plom	Dip-lom		de-struk-tiv		
igno-rie-re	ig-no-rie-re		sub-lim		
In-te-gral	In-teg-ral				

R3 Falls die Trennung nach Sprechsilben und die etymologische Trennung (nach Wortherkunft) kollidieren, wurde weitgehend die etymologische

Trennung gewählt [Rato6, Wiso6, § 113]:

in-ter-es-sant	in-te-res-sant
Lin-ole-um	Li-no-le-um
Päd-ago-ge	Pä-da-go-ge

R4 In Ableitungen von Namen auf *-ow* bleibt *-ow* ungetrennt, wenn es den Laut [o:] bezeichnet. Die Nottrennung der Ableitungssilben *-er*, *-ern*, *-ers* wird unterdrückt [Rato6, Wiso6, § 113] i. V. m. [Wis91, R 180]:

Tel-tower	Tel-to-wer	Tel-tow-er
Trep-towern	Trep-to-wern	Trep-tow-ern
Pan-kowers	Pan-ko-wers	Pan-kow-ers

4. Fehltrennungen

Für fehlerhafte (falsche, ausgelassene oder unerwünschte) Trennungen gibt es zwei mögliche Ursachen:

1. Die zugrundeliegende Wortliste enthält einen Fehler.
2. Das betreffende Wort ist in der zugrundeliegenden Wortliste nicht enthalten.

Da der Umfang der Wortliste nicht beliebig erweitert werden kann, sollten Fehltrennungen nur dann gemeldet werden, wenn eines der folgenden Kriterien erfüllt ist:

A. Das betreffende Wort wird mit den herkömmlichen Trennmustern für die traditionelle oder reformierte Rechtschreibung korrekt getrennt. Korrekt bedeutet hier: Nicht alle möglichen Trennstellen müssen erkannt werden; es werden jedoch in keinem Fall falsche Trennstellen ermittelt.

   Zum Testen kann der folgende Aufruf verwendet werden (die Ausgabe erfolgt in der LOG-Datei):

   ```
   \showhyphens{durch Leerzeichen getrennte Wörter}
   ```

B. Es handelt sich um eine sinnentstellende oder irreführende Trennung eines Wortes, das nicht aus mehr als zwei prä- und suffigierten Wörtern zusammengesetzt ist, zum Beispiel »Talent-wässerung«. Nicht berücksichtigt wird hingegen die »Talent-wässerungsanlage«.

C. Das Wort ist bereits in der Wortliste enthalten (siehe Anhang A) und Punkt 1 trifft zu.

Einige bekannte Fehler in den Trennmustern sind in der Datei `BUGS` verzeichnet. Noch nicht bekannte falsche, fehlende und unerwünschte Worttrennungen können an die folgenden E-Mail-Adressen gerichtet werden:

- trennmuster-opensource@googlegroups.com (Anmeldung erforderlich),
- wl@gnu.org (Werner Lemberg).

Fehltrennungen, die in den Trennmustern nicht korrigiert werden können, können mit Hilfe einer privaten Ausnahmeliste behandelt werden:

```
\hyphenation{Tal-entwäs-se-rungs-an-la-ge Kaf-ka-kenner-klub}
```

Happy TEXing!
Die deutschsprachige Trennmustermannschaft

Literatur

[Lia83] Liang, Franklin Mark: *Word Hy-phen-a-tion by Com-put-er*. Dissertation, Stanford University, 1983. http://www.tug.org/docs/liang/.

[Rat06] Rat für deutsche Rechtschreibung: *Deutsche Rechtschreibung*. http://rechtschreibrat.ids-mannheim.de/download/regeln2006.pdf, München, 2006.

[Wis91] Wissenschaftlicher Rat der Dudenredaktion (Herausgeber): *Duden : Rechtschreibung der deutschen Sprache*, Band 1 der Reihe *Der Duden in 12 Bänden*. Dudenverlag, Mannheim, 20. Auflage, 1991.

[Wis06] Wissenschaftlicher Rat der Dudenredaktion (Herausgeber): *Duden : Die deutsche Rechtschreibung auf der Grundlage der neuen amtlichen Rechtschreibregeln*, Band 1 der Reihe *Der Duden in 12 Bänden*, Seiten 1161–1216. Dudenverlag, Mannheim, 24. Auflage, 2006.

A. Die Wortliste

Die Wortliste ist über das öffentliche Entwicklerrepositorium[2] des Projekts[3] erhältlich. Eine Kopie kann mit

```
git clone git://repo.or.cz/wortliste.git      oder
git clone http://repo.or.cz/r/wortliste.git
```

bezogen werden.[4] Das Format der Wortliste wird in der Datei `dateikopf` beschrieben.

Der SHA1-Commit-Hash der Repositoriumversion, die den Dateien

```
dehyphn-x-<datum>.pat
dehypht-x-<datum>.pat
```

zugrundeliegt, sowie eine URL zum direkten Herunterladen der Wortliste (ca. 15 MB) kann dem Kopf beider Dateien entnommen werden.

[2] http://repo.or.cz/w/wortliste.git
[3] http://groups.google.de/group/trennmuster-opensource?hl=de
[4] http://repo.or.cz/