

dehyph-expt1*

Experimentelle Trennmuster für die deutsche Sprache

Die deutschsprachige Trennmustermannschaft

21. Juni 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	3. Trennregeln	4
1. Einleitung	4. Fehltrennungen	5
2. Verwenden der Trennmuster	A. Die Wortliste	6

1. Einleitung

Der in T_EX implementierte Trennalgorithmus arbeitet musterbasiert [Lia83]. Prinzipiell können mit einem solchen Algorithmus nicht alle möglichen Wörter

*This document describes the dehyph-expt1 package v0.1.

korrekt getrennt werden. Die Qualität der Trennmuster einer Sprache wird jedoch maßgeblich von der Qualität der Wortliste beeinflusst, aus der sie berechnet werden.

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.

Das Projekt *Freie Wortlisten und Trennmuster für die deutsche Sprache* hat sich deshalb das Ziel gesetzt neue, hochqualitative Trennmuster für die Benutzung in $\text{T}_{\text{E}}\text{X}$ und OpenOffice zu schaffen.

Den experimentellen Trennmustern dieses Pakets liegt eine Wortliste mit den etwa fünfhunderttausend häufigsten deutschen Wörtern in deutscher und (deutsch)schweizerischer¹ Schreibung zugrunde. Diese Liste ist vermutlich erheblich umfangreicher als die ursprüngliche Wortliste. Außerdem wurden Worthäufigkeiten in der ursprünglichen Wortliste wahrscheinlich überhaupt nicht berücksichtigt.

¹Die Berücksichtigung dieser Schreibweise kommt auch dem Versalsatz zugute. Wenngleich auf Trennungen dabei möglichst verzichtet werden sollte.

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-URL² oder am 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 README beschrieben. Sie können mit den Paketen Babel und hyphsubst aus dem OBERDIEK-BÜNDEL aktiviert werden.

Das folgende Beispiel zeigt eine L^AT_EX-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! Weitere Hinweise können der Dokumentation des Pakets hyphsubst entnommen werden.

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

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 sollten ausgiebig getestet werden! Sie sind jedoch noch nicht für Zwecke geeignet, die einen dauerhaft stabilen Umbruch erfordern. Diese Trennmuster können jederzeit durch umbruchin-kompatible Versionen ersetzt und vom CTAN oder aus T_EX-Verteilungen entfernt werden.

²<http://groups.google.de/group/trennmuster-opensource?hl=de>

<i>traditionelle Rechtschreibung</i>		<i>reformierte Rechtschreibung</i>	
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-ster[n]	Abend-ster[n]	Abends-tern	Abend-ster[n]
Mor-dop-fer	Mord-op-fer	Mor-dop-fer	Mord-op-fer

Tabelle 1: Trennvarianten

3. Trennregeln

Es werden zwei Trennmuster bereitgestellt, entsprechend den traditionellen und den reformierten³ amtlichen Regeln für die Rechtschreibung der deutschen Sprache.

Da sich Freiheiten bei der Schreibung und Trennung von Wörtern nicht ohne weiteres auf die maschinelle Worttrennung übertragen lassen, wurden die folgenden Konventionen getroffen. Hauptsächlich betreffen diese die reformierte Rechtschreibung. Die grünen Spalten zeigen die Trennung mit Trennmustern für die traditionelle und reformierte Rechtschreibung. Rot werden alternative oder unerwünschte Trennungen dargestellt:

1. Falls die Trennung nach Sprechsilben und die etymologische Trennung (nach Wortherkunft) kollidieren, wurde weitgehend die etymologische Trennung gewählt [Rato6, Wiso6, § 113]:

Heli-ko-pter	Heli-ko-pter	Heli-kop-ter
Päd-ago-ge	Päd-ago-ge	Pä-da-go-ge
poe-tisch	poe-tisch	po-e-tisch

2. In Fremdwörtern bleiben die Verbindungen aus Buchstaben für einen Konsonanten + *l*, *n* und *r* ungetrennt, einschließlich *phl*, *phr*, *thr* (Ausnahme: *str*) [Rato6, Wiso6, § 110, § 112]:

Ar-thri-tis	Ar-thri-tis	Arth-ri-tis
Co-gnac	Co-gnac	Cog-nac
Di-plom	Di-plom	Dip-lom
In-du-strie	In-dus-trie	In-du-strie

³in der Fassung von 2006 [Rato6, Wiso6]

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

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

4. Fehltrennungen

Für auftretende Trennfehler (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.

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 T_EXing!

Die deutschsprachige Trennmustermannschaft

Literatur

- [Lia83] Liang, Franklin Mark: *Word Hy-phen-a-tion by Com-put-er*. Dissertation, Universität Stanford, 1983. <http://www.tug.org/docs/liang/>.
- [Rato6] Rat für deutsche Rechtschreibung: *Deutsche Rechtschreibung*. <http://rechtschreibrat.ids-mannheim.de/download/regeln2006.pdf>, München, 2006.
- [Wiso6] 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 des Projekts⁴ erhältlich.⁵ Eine Kopie kann mit

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

bezogen werden.⁶

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

⁴<http://groups.google.de/group/trennmuster-opensource?hl=de>

⁵<http://repo.or.cz/w/wortliste.git>

⁶<http://repo.or.cz/> Eine Git-Version für Windows ist unter <http://code.google.com/p/msysgit/downloads/list> erhältlich, Datei `Git-1.5.5-preview20080413.exe` (Stand: 11. 6. 2008).

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.