Freie Wortlisten und Trennmuster für die deutsche Sprache

Eine kurze Projektbeschreibung

Die deutschsprachige Trennmustermannschaft

15. Juni 2009

Inhaltsverzeichnis

1 Ziele

Dieses Projekt beabsichtigt, hochqualitative Wortlisten, Trennmuster [Lia83] und Ausnahmelisten für die deutsche Sprache zu schaffen, die auch österreichische und (deutsch)schweizerische Besonderheiten abdecken. Grundlage sind die verbindlichen Regeln des Dudens in der Fassung von 1991 [Wis91] und die amtlichen Regeln für die Rechtschreibung der deutschen Sprache in der Fassung von 2006 [Rato6, Wiso6]. Teil des Projekts ist eine Infrastruktur, die die Kontrolle der Listen in verteilter Arbeit ermöglicht. Wortlisten, Trennmuster und Ausnahmelisten sollen unter freien Lizenzen jedermann zugänglich gemacht werden.

2 Wer wir sind

Anstoß für dieses Projekt war eine Reihe von Artikeln in der *TEXnischen Komödie*[1], die sich mit den herkömmlichen Trennmustern für die deutsche Sprache auseinandersetzen [Lem03, Lem05, Hen08].

Die deutschsprachige Trennmustermannschaft ist eine offene Gruppe und setzt sich zur Zeit aus Mitgliedern des DANTE e. V. sowie Anwendern der Programme TEX und OpenOffice[2] zusammen. Die Kommunikation läuft derzeit über die Gruppe Trennmuster-Opensource bei Google[3], das Projekt soll jedoch bei einem anderen Internetdienst zur Verwaltung von Softwareprojekten angemeldet werden. Ersatzweise kann für die Diskussion auch die Newsgruppe `de.comp.text.tex` genutzt werden.

Wir benötigen dringend weitere Mithilfe. Informationen dazu können der Dokumentation des TEX-Pakets `dehyph-exptl` oder der folgenden Aufgabenliste entnommen werden. Bei Interesse oder mit Fragen wenden Sie sich bitte an die Projektliste.

3 Aufgabenliste

Dieser Abschnitt enthält eine Zusammenstellung von anstehenden Aufgaben. Die Liste ist weder sortiert noch vollständig. Die meisten Punkte bestehen nur als Idee und werden zur Zeit nicht aktiv bearbeitet.

3.1 Wortlisten

- Spezifikation und Implementierung der Datenbankanbindung (Wortliste v2.0, Backend)
- Spezifikation und Gestaltung der HTML-Maske zur verteilten Kontrolle der Wortliste (Frontend)
- systematische Erweiterung der Hauptwortliste:
  - um Wörter aus einer nach Häufigkeiten sortierten Liste (Quellen sind vorhanden, siehe Abschnitt 5),
  - um Wörter aus geläufigen Sprachbereichen (Quellen werden benötigt):

[1] Vereinsschrift der *Deutschsprachigen Anwendervereinigung TEX* (DANTE e. V.), `http://www.dante.de/`

[2] `http://de.openoffice.org/`

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

    * Küche und Haushalt,
    * Flora und Fauna,
    * Handwerk, Politik, Sport,
    * (alte) Literatur: Märchen, Liedtexte,
    * deutsche Vor- und Nachnamen,
    * Städte- und Flussnamen,
    * ~~Zahlwörter und römische Zahlen~~,
    * ...

- Erstellen von Wortlisten für fachspezifische Trennmuster (Quellen werden benötigt):
  - Naturwissenschaft und Technik,
  - Sprachwissenschaften,
  - Geschichte und Religion,
  - Medizin,
  - Jura,
  - Kunst,
  - ...
- Gewichtung der Trennstellen
- Markierung von Nottrennungen
- Behandlung von falschen Ligaturen
- Unterscheidung von Lang- und Rund-s für den Satz mit gebrochenen Schriften
- verteilte Kontrolle von Rechtschreibung und Trennung des Wortbestands

3.2 *Trennmuster*

- Können Strukturen der deutschen Sprache durch Startmuster in PATGEN berücksichtigt werden, zum Beispiel häufige Vor- und Nachsilben wie *auf-*, *aus-*, *ver-*, *-keit*, *-lich*, *-lein*, etc.?

3.3 *TEX*

Sprachanbindung Die Trennmuster für Schweizer Standarddeutsch (traditionelle Rechtschreibung) benötigen eine Sprachanbindung für `Babel` und `Polyglossia`.

Versionierung der Trennmuster In Dokumenten muss trotz weiterentwickelter Trennmuster langfristig umbruchtreuer Textsatz garantiert werden können. Daher benötigt TEXs Sprachenschnittstelle (Pakete `Babel` bzw. `Polyglossia`) einen Versionsmechanismus für die Trennmusteraktivierung. ⇒ Paket `hyphsubst`

mehrere Trennmuster laden Für einzelne Dokumente ist es wünschenswert, fachspezifische Trennmuster verwenden zu können. Allerdings können Trennmuster nur während der Formaterstellung geladen werden. Einmal eingebundene Trennmuster sind somit immer im Format vorhanden. Es wäre sinnvoll, Trennmuster auch dynamisch während der Ausführung laden zu können. Wie wird mit kollidierenden Mustern umgegangen? ⇒ LuaTEX

Trennstellengewichte Trennstellen sollen gewichtet werden können (z. B. Haupt-, Neben-, Nottrennungen).

Verallgemeinerung des Trennalgorithmus Derzeit benötigen TEX-Quelltexte in deutscher Sprache verschiedene Auszeichnungen:

`\/ "|`	zur Korrektur falscher Ligaturen (`Auf"|lage`)
`s:` *u. a.*	zur Unterscheidung von langem und rundem *s* in gebrochenen Schriften (`aus:setzen`)
`"ck`	für die richtige Trennung von *ck* in der traditionellen Rechtschreibung (`dru"cken`)
`"ff "ll` *u. a.*	für die richtige Trennung von Dreifachkonsonanten mit folgendem Vokal in der traditionellen Rechtschreibung (`Meta"llegierung`)

Für die Dokumentenerstellung mit TEX ergeben sich dadurch die folgenden Nachteile:

1. Alle genannten Eingriffe beeinflussen die Worttrennung in den übrigen Wortteilen.
2. Das Anbringen der Auszeichnung von Hand erfordert zusätzlichen Aufwand und ist fehleranfällig.
3. Physische Auszeichnung erschwert den Wechsel zwischen gebrochener und runder Schrift.
4. Physische Auszeichnung erschwert den Wechsel zwischen traditioneller und reformierter Rechtschreibung.

Idealerweise sollte ein Quelltext möglichst wenige physische Auszeichnungen enthalten.

Es handelt sich hier um ein Mustererkennungsproblem ähnlich der Worttrennung. Denkbar ist eine Verallgemeinerung des Silbentrennalgorithmus, mit der diese Fälle ohne Eingriffe im Quelltext erkannt und richtig behandelt werden können: Mit entsprechenden zusätzlichen Mustern werden während des Absatzumbruchs *verschiedene* Tries durchsucht und die notwendigen Ersetzungen vorgenommen. ⇒ Vorschlag für LuaTEX-Entwickler ausarbeiten

3.4 Sonstige

- Umzug auf einen anderen Host (BerliOS, Sourceforge o. ä.)
- Reservierung einer eigenen Domain

4 Zeitplan

Kurzfristig steht die Verbesserung der in TEX und OpenOffice verwendeten Trennmuster im Vordergrund.

Die Pflege der Wortlisten und Trennmuster ist jedoch ein Langzeitprojekt. Die verwendeten Werkzeuge und Listen sind daher gut zu dokumentieren.

5 Ressourcen

5.1 Auf welche Wortlisten können wir zurückgreifen?

Lembergs Liste

Urheber	Werner Lemberg
Rechte	(GPL angestrebt)
Wortformen	430 000
Sortierkriterium	alphabetisch
Rechtschreibung	gut
Bemerkung	manuell gepflegt
Zugriff	GIT-Repositorium:
	`$git clone git://repo.or.cz/wortliste.git`
Stand	15. 6. 2009

Leipziger Liste

Urheber	Liste des Wortschatzprojekts der Universität Leipzig[4]
Rechte	GPL
Wortformen	2 000 000
Sortierkriterium	Häufigkeit
Rechtschreibung	mangelhaft
Bemerkung	automatische Internetsuche (Datenbanken, Zeitungsarchive usw.)
Zugriff	...
Stand	28. 3. 2008

Mannheimer Liste

Urheber	Korpus des Instituts für Deutsche Sprache (IDS)[5]
Rechte	»Darf in ihrer Gesamtheit – wie vereinbart – nicht veröffentlicht oder an Dritte weitergegeben werden.« Abgeleitete Werke können nach unserer Wahl behandelt werden.
Wortformen	4 000 000
Sortierkriterium	Häufigkeitsklassen
Rechtschreibung	mittel
Zugriff	nicht öffentlich
Stand	9. 10. 2007

5.2 Welche Listen haben wir in Aussicht?

Berliner Liste

Urheber	Kernkorpus des Projekts Digitales Wörterbuch der Deutschen Sprache (DWDS)[6]
Rechte	...
Wortformen	2 000 000
Sortierkriterium	...
Rechtschreibung	...
Bemerkung	repräsentativer Wortschatz der deutschen Sprache
Zugriff	derzeit nicht
Stand	Juni 2009

[4] http://wortschatz.uni-leipzig.de/
[5] http://www.ids-mannheim.de/kl/
[6] http://www.dwdscorpus.de/

6 Bisherige Ergebnisse

6.1 *Trennmuster*

- Das LaTeX-Paket `dehyph-exptl` enthält die aktuellen experimentellen Trennmuster, die aus Werner Lembergs Liste generiert wurden. Diese können zusammen mit den Paketen `Babel` und `hyphsubst` verwendet werden. Das Paket `dehyph-exptl` ist im CTAN verfügbar.[7]
- Die bisherigen Trennmusterdateien für die traditionelle und reformierte Rechtschreibung, die Dateien `dehypht.tex` und `dehyphn.tex`, können im CTAN bezogen werden. Sie sind *nicht* im Zuge dieses Projekts entstanden.
- Die Ausnahmedatei `dehyphtex.tex` für die traditionellen deutschen Trennmuster ist im CTAN verfügbar.

6.2 *Wortlisten*

Die von Werner Lemberg erstellte und kontrollierte Liste steht im öffentlich zugänglichen Entwicklerrepositorium.[8] Eine Kopie kann mit

```
$git clone git://repo.or.cz/wortliste.git
```

bezogen werden.[9]

Für den Zugriff auf dieses Repositorium sollte Git 1.5.6.5 oder jünger verwendet werden. Ältere Versionen enthalten einen Fehler im Netzwerkzugriff, der dazu führt, dass statt eines Satzes von Änderungen häufig die komplette Wortliste übertragen wird.

Außerdem sollte für einen reibungslosen Betrieb nach dem Klonen dieses Repositoriums die Konfigurationsvariable `core.DeltaBaseCacheLimit` vom Standardwert 16 MB auf mindestens 64 MB erhöht werden. Mit älteren Git-Versionen können auch deutlich höhere Werte sinnvoll sein. Die Konfiguration kann beispielhaft so erfolgen:

```
$git config core.DeltaBaseCacheLimit 64m
$git config --get core.DeltaBaseCacheLimit
```

Das letzte Kommando sollte nun den Wert `64m` zurückliefern.

Das Format der Datei `wortliste` ist in der Datei `dateikopf` beschrieben. Im Repositorium sind auch einige Skripten zur Bearbeitung der Wortliste enthalten. Aktualisiert wird die Wortliste (das gesamte lokale Repositorium) mit

[7] Comprehensive TeX Archive Network, http://ctan.tug.org/
[8] http://repo.or.cz/w/wortliste.git
[9] http://repo.or.cz/

```
$git pull
```

6.3 HTML-*Prüfmaske*

Es existiert ein Entwurf für eine interaktive Maske zur Kontrolle der Rechtschreibung von Wortlisten und der Eingabe korrekter Trennungen.[10]

Literatur

[Heno8] Hennig, Stephan: *Einige Fragen zum Beitrag Hyphenation Exception Log für deutsche Trennmuster, Version 1*. Die TEXnische Komödie, 20(1):7–17, Januar 2008.

[Lemo3] Lemberg, Werner: *Hyphenation Exception Log für deutsche Trennmuster*. Die TEXnische Komödie, 15(2):28–31, Mai 2003.

[Lemo5] Lemberg, Werner: *Hyphenation Exception Log für deutsche Trennmuster, Version 1*. Die TEXnische Komödie, 17(2):24–51, Mai 2005.

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

[Rato6] 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.

[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.

[10] http://www.mnn.ch/opendehyph/index.php