

Freie Wortlisten und Trennmuster für die deutsche Sprache

Eine kurze Projektbeschreibung

Die deutschsprachige Trennmustermannschaft

19. Juni 2008

Inhaltsverzeichnis

| | | |
|------------------------------------------------------------------|-------------------------------------------------------|----------|
| | 4.2 Welche Listen haben wir in Aussicht? | 4 |
| 1 Ziele | 1 | |
| 2 Wer wir sind | 2 | |
| 3 Aufgaben | 2 | |
| 4 Ressourcen | 3 | |
| 4.1 Auf welche Wortlisten können wir zurückgreifen? | 3 | |
| | 5 Bisherige Ergebnisse | 4 |
| | 5.1 Wortlisten | 4 |
| | 5.2 HTML-Frontend | 5 |
| | 5.3 Trennmusterdateien | 5 |
| | 6 Zeitplan | 5 |

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 amtlichen Regeln der deutschen Rechtschreibung in der traditionellen sowie der reformierten 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

Die deutschsprachige Trennmustermannschaft setzt sich zur Zeit aus Leuten der OpenOffice-Gemeinde¹ und Mitgliedern des Dante e. V.² zusammen.

Die Kommunikation läuft derzeit über die Gruppe TRENNMUSTER-OPENSOURCE bei Google³, das Projekt soll jedoch bei einem anderen Internetdienst zur Verwaltung von Softwareprojekten angemeldet werden.

3 Aufgaben

Dieser Abschnitt enthält eine Zusammenstellung von anstehenden Aufgaben. Die Liste ist weder sortiert, noch vollständig.

Arbeit an Trennmustern

- Spezifikation und Gestaltung der Eingabemaske (Frontend)
- Spezifikation und Implementierung der Datenbankbindung (Backend)
- Erweitern des Wortbestandes um Wörter aus geläufigen Sprachbereichen (Küche und Haushalt, Handwerk, Floristik, Märchen, deutsche Vor- und Nachnamen, Städte- und Flussnamen, römische Ziffern, etc.)
- Kontrolle von Rechtschreibung und Trennung des Wortbestands
- Prüfen, ob linguistische Besonderheiten der deutschen Sprache durch Startmuster in PATGEN berücksichtigt werden können (häufige Vor- und Nachsilben wie -lich, -keit, -lein, etc.)

Änderungen an T_EX

- T_EX sollte Trennmuster dynamisch laden können, beispielsweise für fachspezifische Begriffe (⇒ LuaT_EX).
- Babel sollte einen Versionsmechanismus für die Trennmusteraktivierung anbieten, um in einem Dokument trotz weiterentwickelter Trennmuster umbruchtreuen Textsatz garantieren zu können (⇒ Paket hyphsubst).
- Haupt- und Nebentrennstellen sollten im Trennalgorithmus berücksichtigt werden.

¹<http://de.openoffice.org/>

²Deutschsprachige Anwendervereinigung T_EX e. V., <http://www.dante.de/>

³<http://groups.google.de/group/trennmuster-opensource?hl=de>

Sonstiges

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

Wer Interesse oder Ideen hat, wende sich bitte an die Gruppe.

Dieses Projekt benötigt Deine Hilfe!

4 Ressourcen

4.1 Auf welche Wortlisten können wir zurückgreifen?

Lembergs Liste

| | |
|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Urheber | Werner Lemberg |
| Rechte | ... |
| Wortformen | 420.000 |
| Sortierkriterium | alphabetisch |
| Rechtschreibung | gut |
| Bemerkung | manuell gepflegt |
| Zugriff | GIT-Repository: <ul style="list-style-type: none">• <code>git clone git://repo.or.cz/wortliste.git</code>• <code>git clone http://repo.or.cz/r/wortliste.git</code> |
| Stand | 29. 3. 2008 |

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 |

⁴<http://wortschatz.uni-leipzig.de/>

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 |

4.2 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 | voraussichtlich 2008 |

5 Bisherige Ergebnisse

5.1 Wortlisten

Die von Werner Lemberg erstellte und kontrollierte Liste steht im öffentlich zugänglichen Entwicklerrepositorium.⁷ Eine Kopie kann mit

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

⁵<http://www.ids-mannheim.de/kl/>

⁶<http://www.dwdscorepus.de/>

⁷<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: 30. 5. 2008).

bezogen werden.⁸ Im Repository sind auch einige Skripten zur Bearbeitung der Wortliste enthalten. Aktualisiert wird die Wortliste (das gesamte lokale Repository) mit

```
git pull
```

5.2 HTML-Frontend

Es existiert ein Entwurf für eine interaktive Maske zur Kontrolle der Rechtschreibung von Wortlisten und der Eingabe korrekter Trennungen.⁹

5.3 Trennmusterdateien

Die bisherigen Trennmusterdateien für die traditionelle und neue Rechtschreibung, die Dateien `dehyph.t` und `dehyphn.t`, können am CTAN¹⁰ bezogen werden. Sie sind *nicht* im Zuge dieses Projekts entstanden.

Im Dateibereich der Google-Gruppe TRENNMUSTER-OPENSOURCE sind Trennmuster verfügbar, die aus Werner Lembergs Liste generiert wurden.¹¹

6 Zeitplan

Es ist kein Ende abzusehen.

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.

[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. Aufl. Auflage, 2006.

⁸Neben dem Protokoll unterscheiden sich auch die Adressen!

⁹<http://www.mnn.ch/opendehyph/index.php>

¹⁰Comprehensive T_EX Archive Network, <http://ctan.tug.org/>

¹¹<http://groups.google.de/group/trennmuster-opensource?hl=de>