Willkommen im Trennmuster-Wiki

Ziele

In diesem Projekt werden Wortlisten für die Generierung von Trennmustern für die deutsche Sprache unter Einbeziehung der traditionellen und der (1996) reformierten Rechtschreibung wie auch österreichischer und schweizerischer Besonderheiten erstellt und gepflegt. Ziel ist die Verbesserung der Worttrennung in *TeX & Co*. Die Wortlisten und Trennmuster können auch in andere Anwendungen eingebunden werden. So verwenden z.B. der *Chromium-Browser* sowie *LibreOffice* und seine Verwandten TeXnische Trennmuster für ihre Textdarstellung. Die Wortlisten und Trennmuster sind für jedermann und unter freien Lizenzen zugänglich.

Wer wir sind

Wir sind ein Haufen *Nerds*, *Freaks* und Tunichtgute mit einem exklusiven gemeinsamen Interesse an Wortlisten und ungewöhnlichen, ja erstaunlichen Trenneffekten. Wir übernehmen den sau-ersten Teil der Planer-füllung bei der Altbauer-neuerung, rätseln dement-sprechend, was eine Salonal-bumserie wohl bein-halte, und diskutieren die Feinheiten der Regeln im Gebrauch von Lang-ſ und Rund-s, alles unter dem Vorwand, damit die Silbentrennung für Wirtschaftsdoktoranden und Otto-Ausnahme-Studenten zu verbessern. Einige von uns sind Mitglieder des DANTE e.V., der auch dieses Wiki und unsere Mailing-Liste betreibt.

Was TeX (bislang) nicht kann

Wir *TeXies* prahlen gern damit, wir hätten den besten Trennalgorithmus der Welt. Das bezieht sich darauf, daß *TeX* bei der Berechnung der Zeilenumbrüche nicht nur die jeweilige Zeile, sondern den gesamten Absatz betrachtet. Es berücksichtigt die entstehende Dichte und „Losigkeit“ der einzelnen Zeilen, die Anzahl der aufeinanderfolgenden Zeilen mit Wortumbruch am Ende, vermeidet unerwünschte Furchenbildungen und das Aufeinandertreffen von Unter- und Oberlängen im Text. Das ist alles in der Tat schon ziemlich faszinierend, allein es sagt noch nichts über die Güte der Trennstellenfindung, also über den Anteil der gefundenen möglichen Trennstellen an den tatsächlich vorhandenen, über den Anteil der falschen Treffer und nichts über die Gewichtung der verschiedenen möglichen Trennstellen.

Was TeX bislang nicht kann, sind

- die Unterscheidung von Haupt- und Ne-ben=trenn=stel-len,
- die Darstellung von „Spezialtrennungen“ in der traditionellen Rechtschreibung, z. B.
  Brücke → Brük-ke,
  Schiffahrt → Schiff-fahrt,
- die automatische Erkennung zulässiger und unzulässiger Ligaturen.

Unsere Wortlisten verbessern nicht nur signifikant die Treffsicherheit der Trennstellenfindung, sondern bereiten auch die Implementierung dieser Fähigkeiten vor. Sie markieren nicht nur die möglichen Trennstellen, sondern unterscheiden durch eine spezielle Semantik auch Haupt- von

Nebentrennstellen und zeichnen Spezialtrennungen aus.

Unser Gründungsmythos

Bereits zu Anfang der 90er Jahre entwickelte das *Institut für Praktische Informatik der Technischen Universität Wien* einen neuen Algorithmus für *sichere sinnentsprechende Silbentrennung* (SiSiSi), der besonders für die deutsche Sprache geeignet ist BS92], überführte ihn aber später in eine kommerzielle Nutzung. Den letzten, frei zugänglichen Stand findet man auf CTAN. 1995 beschrieb *Petr Sojka* einen Mechanismus, Haupt- von Nebentrennstellen zu unterscheiden, der aber nie implementiert wurde Soj95]. 1998 entwickelte *Matthias Clasen* Erweiterungen für den originalen Trennalgorithmus von *TeX*. Ebenfalls 1998 paßte *Walter Schmidt* die herkömmlichen Trennmuster für die „neue deutsche Rechtschreibung“ an und resümierte, daß auf lange Sicht die Trennmuster grundsätzlich neu erzeugt werden sollten Sch98]. 2003 rief *Werner Lemberg* in der DTK dazu auf, eine Liste der Trennmusterausnahmen anzulegen und diese, sobald eine hinreichend große Zahl zusammengekommen sei, in eine neue Version der deutschen Trennmuster aufzunehmen Lem03]. Aus dem Publikum erhielt er darauf keine Reaktion, aber *Walter Schmidt* meldete sich bei ihm und lieferte ihm für sein Vorhaben umfangreiches Datenmaterial. 2005 faßte *Werner Lemberg* die Ergebnisse seiner Tests zusammen und analysierte die Schwächen der vorliegenden Trennmuster:

> Eine überraschend hohe Anzahl von einfachen und zusammengesetzten Worten werden von den (alten) deutschen Trennmustern falsch getrennt. Ich glaube, dass es überlegenswert ist, neue Trennmuster für die alte Rechtschreibung zu generieren, sobald die Wortliste einen gewissen Umfang erreicht hat. … Meiner Meinung nach sollte [das] Augenmerk auf wirklich fehlerfreie Trennung einer möglichst großen Anzahl von einfachen und zusammengesetzten Wörtern gelegt werden. Lem05]

2008 antwortete ihm *Stephan Hennig* und merkte an:

> … für das Erstellen neuer Trennmuster [wird] weit mehr benötigt als eine Liste einiger derzeit falsch getrennter Wörter. Nämlich eine qualitativ möglichst hochwertige Liste getrennter deutscher Wörter. Hen08].

Also ward *dehyph-exptl*.

Mitarbeit

Wenn Sie diesen Punkt gefunden haben, stehen Sie an der Schwelle einer weiteren Stufe Ihres Pfads zur Erkenntnis.

Sie können

- weiter studieren,
- hier oder da stöbern,
- unser git-Repositorium „aus-checken“,
- sich bei der Mailingliste anmelden,
- sich im Wiki registrieren, wobei, wenn Sie eine kurze Mail an die Trennmuster-Mailingliste schreiben und dabei den gewählten Wikinamen angeben, die bestehenden Mitglieder den

Wikinamen der Trennmustergruppe hinzufügen und Ihnen so Schreibrecht erteilen können,
- DANTE beitreten.

Wenn Sie herausgefunden haben, wie Sie etwas beitragen können, dann haben Sie die nächste Stufe erreicht.

Ressourcen

- Literatur
- Teilprojekte
- Kommunikation
- Dateibereich
- Trennmustler
- Korpora
- Lizenzen
- Referenzen
- patgen selbst kompilieren