Deutsche Silbentrennung für TEX 3.1

Wilhelm Barth
Helmut Steiner

Zusammenfassung

TEX- und LATEX-Benutzer können jetzt das speziell für deutschsprachige Texte entwickelte *SiSiSi*

SICHERE
SINNENTSPRECHENDE
SILBENTRENNUNG

benutzen. Es gibt für VAX/VMS ein change-file und alle anderen notwendigen Files auf dem FTP-Server `eichow.tuwien.ac.at`[1]. Dort findet man auch eine Anleitung, mit der man das System leicht installieren kann. Diese Modifikation führt nur eine andere Hyphenation ein, alle anderen TEX-Funktionen bleiben unverändert[2].

Motivation

In der deutschen Sprache verwendet man gern lange, zusammengesetzte Wörter, z.B. Text=verarbeitungs=system oder Silben=trennungs=verfahren. Wenn solche riesigen Gebilde nicht am Ende einer Zeile abgeteilt werden können, entsteht meist ein sehr unschönes Schriftbild. Das Hauptproblem bei der Silbentrennung ist daher das Finden der Nahtstellen (durch = angezeigt) zwischen den Einzelwörtern. *SiSiSi* erreicht das mit Hilfe einer Worttabelle, genauer einer Tabelle aller Wortbestandteile: Vorsilben, Stämme, Endungen. Weniger als 8000 Eintragungen reichen aus, um fast alle deutschen Wörter und die gängigen Fremdwörter zu erfassen. Durch die Abstützung des Verfahrens auf eine Worttabelle ist es natürlich möglich, auch mit unvorhersehbaren, ungewöhnlichen Wortungetümen fertig zu werden. Dieses Zerspalten zusammengesetzter Wörter mit einer einfachen, überschaubaren Methode bringt wesentliche Vorteile gegenüber dem aus dem Amerikanischen übernommenen „pattern“-Verfahren.

SiSiSi sucht nach allen möglichen Zerlegungen. Findet es mehr als eine und kann dadurch die Zerlegung nicht eindeutig feststellen, z.B. Bau=mast/Baum=ast oder Stau=becken/Staub=ecken, nutzt es die zweifelhaften Trennstellen nicht aus. Dadurch ist *SiSiSi* sicher, d.h. es erzeugt keine falschen Trennungen.

[1] Anmerkung der Redaktion: Die numerische Adresse lautet `128.130.165.5`. Vorsicht: Es handelt sich um einen VMS-Rechner!

[2] Im vorliegenden Dokument wurde *SiSiSi* nicht verwendet.

Falls *SiSiSi* für ein Wort keine Zerlegung findet, dann handelt es sich dabei entweder um ein sehr ausgefallenes Wort (z.B. Eigennamen oder ähnliches) oder das Wort wurde falsch geschrieben. Auch in diesen Fällen wird nicht getrennt, *SiSiSi* bleibt sicher!

SiSiSi bevorzugt die Haupttrennstellen an den Nahtstellen zusammengesetzter Wörter (sie erhalten kleine „penalties“) gegenüber den Ne=bentrennstel=len in den Einzelwörtern. Dadurch unterstützt es eine sinnentsprechende Trennung.

Kurzbeschreibung des Verfahrens

Jedes Wort ist eine Folge von einem oder mehreren Einzelwörtern. Jedes Einzelwort besteht aus beliebig vielen Vorsilben (eventuell auch keinen), gefolgt von genau einem Stamm, abgeschlossen durch beliebig viele Endungen (ersatzweise Fugenzeichen). Eine Worttabelle enthält alle diese erwähnten Wortbestandteile (Morpheme).

Der Algorithmus sucht für jedes Wort, das eventuell getrennt werden soll, systematisch nach allen Zerlegungen, die nach der angegebenen Grammatik möglich sind. Dabei erkennt er durch Nachschauen in der Worttabelle, ob ein betrachtetes Teilstück des Wortes ein Wortbestandteil im erwähnten Sinn ist und gegebenenfalls von welcher Art es ist.

Auf diese Art findet der Algorithmus für jede Zerlegung alle Haupttrennstellen, nämlich zwischen der letzten Endung eines Teilwortes und der folgenden Vorsilbe bzw. dem folgenden Stamm. Außerdem erkennt es auch sofort alle Nebentrennstellen hinter den Vorsilben. Für den Rest jedes Einzelworts, bestehend aus Stamm und Endungen, muß man die Duden-Regeln für die Silbentrennung anwenden. Das sind solche Regeln wie „In einer Folge von Konsonanten ist vor dem letzten zu trennen“, z.B. Fül-lungen, kämp-fen. Diese Regeln sind mit all ihren Ausnahmen, z.B. Sonderbehandlung von „st“ und „ck“, vollständig in *SiSiSi* eingearbeitet. Ebenso wird schon bei der Zerlegung in Einzelwörter die 3-Konsonanten-Regel berücksichtigt.

Insbesondere bei Fremdwörtern gibt es Wörter, die nicht nach den Duden-Regeln getrennt werden, z.B. Pro-gramm oder Pan-orama. In diesen Fällen ist in der Worttabelle der Stamm als „Ausnahme“ deklariert und die möglichen Trennstellen sind angegeben. Der Algorithmus berücksichtigt natürlich solche Angaben.

Eine ausführliche Beschreibung des Verfahrens findet sich in: W. Barth, H. Nirschl — Sichere sinnentsprechende Silbentrennung für die deutsche Spra-

che, Angewandte Informatik 1985, S. 152–159 (oder Institutsbericht Nr. 26, Institut für Computergrafik der TU Wien).

Verwendung in TeX

Nach der Installation von *SiSiSi* kann das System wie ein normales TeX verwendet werden, die Silbentrennung erfolgt aber nach dem neuen Verfahren.

Man kann auch weiterhin den TeX-Befehl `\-` benutzen. Das ist sinnvoll, wenn man solche unsicheren Wörter wie Baumast oder Staubecken hat. Dadurch kann man für die augenblicklich gewünschte Alternative das Ausnutzen aller Trennstellen erreichen.

Zusätzlich besteht aber noch die Möglichkeit, die Worttabelle zu erweitern. Dadurch kann man die erwähnten Mehrdeutigkeiten schöner behandeln. Trägt man z.B. Stau=bek-ken mit den angezeigten Trennstellen in die Worttabelle ein, so werden diese Trennungen bei Bedarf bei jedem Vorkommen des Wortes verwendet. Darüberhinaus auch bei allen Ableitungen und Zusammensetzungen, die dieses Wort enthalten. Sie werden also auch in „des Stau=beckens“ oder in „Stau=becken=sicherheits=komission“ ausgenutzt. Aber man muß beachten, daß es jetzt keine „Staub=ecken“ mehr gibt! Natürlich ist diese Entscheidung für eine Alternative nur wirksam für Dokumente, bei denen man die erweiterte Worttabelle benutzt.

Man wird die Worttabelle auch dann erweitern, wenn man ein spezielles Vokabular verwendet, z.B. Eigennamen, Kunstwörter, geographische Begriffe, usw. Es gilt sinngemäß das im vorigen Absatz Gesagte.

Deutsches und internationales BibTeXing

Martin Wallmeier

Eigentlich ist BibTeX ein sehr schönes Programm. Es sortiert Literaturverzeichnisse automatisch und man muß nur einen neuen BibTeX-Style wählen, um die Literaturangaben so gesetzt zu bekommen, wie es vielleicht die Zeitschrift vorschreibt, in der man seinen Artikel veröffentlichen möchte.

Trotzdem haben die BibTeX-Standard-Styles auch einen entscheidenden Nachteil: sie können nur englische Literatur vernünftig setzen. Eine deutsche Literaturangabe würde so einfach furchtbar aussehen: