Die TUSTEP-Programme FORMATIERE und SATZ führen, wenn die Silbentrennung nicht über Parameterangaben oder Steueranweisungen unterbunden wird, eine automatische Silbentrennung durch, falls Zeileneinteilung und ggf. Randausgleich dies erfordern. Dabei werden die Regeln der Silbentrennung für die deutsche Sprache zugrunde gelegt.

Keine automatisch durchgeführte Silbentrennung arbeitet fehlerfrei. Selbst umfangreiche Ausnahmelisten würden dem nicht abhelfen, da es viele Wörter gibt, die je nach ihrem Kontext anders zu trennen sind (Beispiel: »spie-lende Kinder«, aber: »kurz vor Spiel-ende«; »Staubecken« und »Staub-ecken«). Vor allem zusammengesetzte Wörter bieten eine schier unerschöpfliche Quelle von Fehlermöglichkeiten.

Für deutsche Texte mit einem nicht ungewöhnlich hohen Anteil an zusammengesetzten Wörtern muß man bei TUSTEP erfahrungsgemäß mit 0,5 bis 1 Prozent fehlerhaften Trennungen (und einer höheren Rate sogenannter »unschöner« Trennungen wie »Versaufbau«) rechnen.

Obwohl das verwendete Trennprogramm die fürs Deutsche geltenden Trennregeln voraussetzt, ist es auch für romanische Sprachen noch brauchbar; die Fehlerrate liegt dann bei etwa 5 %. Für englischsprachige Texte sind diese Trennregeln ungeeignet; hier empfiehlt es sich, die Zahl der Trennungen insgesamt (und damit die Zahl der Falschtrennungen) durch geeignete Parameter zu reduzieren.

Viele Formatier- und Satzprogramme bieten die Möglichkeit, mit Hilfe von Trennwörterbüchern, die der Benutzer selbst erweitern kann, die Zahl der falschen oder unschönen Trennungen zu reduzieren. Aus Gründen des Durchsatzes ist diese Möglichkeit in den beiden genannten TUSTEP-Programmen nicht vorgesehen, sondern in eigene Prozeduren verlagert, die bei Bedarf vor dem Setzen oder dem Formatieren aufgerufen werden können.

Um falsche oder unschöne Trennungen zu vermeiden, muß in dem zu trennenden Wort hinter einem Buchstaben, hinter dem nicht getrennt werden soll, die Steueranweisung für Trennverbot \\ enthalten sein. Soll bevorzugt an bestimmten Stellen getrennt werden, so muß hinter den Buchstaben, hinter denen getrennt werden soll, die Steueranweisung für Kanntrennstelle \ stehen. Bei der automatischen Silbentrennung werden so markierte Stellen bevorzugt berücksichtigt. Ein Wort, das durch \ markierte Kanntrennstellen enthält, wird an anderen Stellen nur getrennt, wenn mindestens drei Buchstaben oder andere Zeichen zwischen dieser Trennung und der Kanntrennstelle liegen.

## Die Makros *SILMARKE und *SILMARKO

In der UNIX-Version (und der VMS-Version) von TUSTEP stehen zwei Standard-Makros zur Verfügung, die vor dem Setzen bzw. Formatieren das Markieren der Textwörter anhand von Ausnahmen-Lexica und die Pflege dieser Ausnahmen-Lexica selbst erleichtern. In der DOS-Version von TUSTEP sind diese Makros noch nicht enthalten.

Mit dem Standard-Makro *SILMARKE können in einem Text die Wörter markiert werden, die bei der automatischen Silbentrennung eine Sonderbehandlung erfahren sollen. Zu diesem Zweck wird der Text von der QUELL-Datei in die ZIEL-Datei kopiert; dabei werden die Textwörter, die in einem vom Benutzer angegebenen Ausnahmen-Lexikon enthalten sind und dort die genannten Markierungen aufweisen, im Text ausgetauscht. Dieses Ausnahmen-Lexikon muß in einer Datei stehen, die zur Spezifikation MARKIERUNGEN angegeben wird. In dieser Datei müssen die Wörter an den Kanntrennstellen durch \ und an den Stellen, an denen sie nicht getrennt werden dürfen, durch \\ markiert sein.

Dieses Verfahren ersetzt die Verwendung von Ausnahmelisten, wie sie bei anderen Silbentrennprogrammen üblich sind.

Es empfiehlt sich, nur die Wörter in das Ausnahmen-Lexikon aufzunehmen, die bei der automatischen Silbentrennung falsch oder unschön getrennt würden. Mit dem Makro *SILMARKO kann diese Datei entsprechend optimiert werden.

Das Standard-Makro *SILMARKO entfernt doppelte und überflüssige Wörter aus Ausnahmen-Lexika, die im Makro *SILMARKE benutzt werden sollen. Überflüssig sind in jedem Fall Wörter, die eine Markierung enthalten und bei automatischer Silbentrennung nur und genau an den Stellen getrennt würden, an denen eine Kann-Trennstelle markiert ist. Auf weitere Einzelheiten und Sonderfälle braucht hier nicht eingegangen zu werden; sie können der Beschreibung des Makros im Handbuch entnommen werden.

Die Optimierung solcher Wortlisten ist zweckmäßig, da sie vor allem für fremdsprachige Texte häufig sehr umfangreich sind, das Makro *SILMARKE aber nur wenige 1000

Einträge auf einmal verarbeiten kann (in der DOS-Version werden es sehr viel weniger sein).

Enthält eine solche Ausnahmenliste sehr viele Einträge, so muß die Verarbeitung in mehreren Schritten durchgeführt werden. Gegebenenfalls muß das Makro *SILMARKE also mehrfach hintereinander aufgerufen werden, wobei die jeweils zu verarbeitenden Einträge mit der Spezifikation BEREICH ausgewählt werden müssen. Bei jedem weiteren Aufruf wird die ZIEL-Datei des jeweils vorhergehenden Aufrufs zur QUELL-Datei.

Ausnahmen-Lexikon für das Englische

Herr Dr. Fritz Kemmler vom Seminar für Englische Philologie der Universität Tübingen hat in den vergangenen Jahren viel Arbeit in eine Wortliste investiert, die mehr als 35.000 Ausnahmen zur Silbentrennung in TUSTEP für das Englische enthält, und zwar nach den in Großbritannien geltenden Normen. Er stellt diese Daten allen TUSTEP-Nutzern zur Verfügung. Sie stehen in Tübingen auf dem Textserver in der Datei `zrlddv1*engsil` zur Benutzung bereit; beim Anmelden dieser Datei muß zur Spezifikation TRAEGER eine Systemvariable angegeben werden, die zuvor auf Betriebssystem-Ebene definiert werden muß. Dies kann z. B. mit folgendem UNIX-Befehl geschehen:

`setenv LDDV /home/textserv/zr`

Außerhalb Tübingens ist diese Datei am leichtesten über den ITUG-Informationsserver (FTP oder Gopher:

`wgex03.germanistik.uni-wuerzburg.de`

vgl. BI 94/7+8 S. 14) in der Rubrik `Daten` zugänglich.

Die Datei `engsil` enthält 35218 Wörter, die länger als 3 Buchstaben sind. Die Datei ist so organisiert, daß sie mit der Mindestzahl von aufeinanderfolgenden Aufrufen des Makros *SILMARKE benutzt werden kann.

Zu diesem Zweck sind die Einträge nach aufsteigender Wortlänge sortiert; die ganze Datei ist eine Segment-Datei und enthält nach einem Segment mit Namen *inhalt*, dem weitere Hinweise zu entnehmen sind, Segmente mit den Namen *teil_n* (wobei *n* eine laufende Nummer ist). Diese Segmente sind (außer dem Segment *teil_1*, siehe unten) so groß, daß der in der UNIX-Version von TUSTEP programmintern zur Verfügung stehende Platz optimal ausgenutzt wird. Die Namen dieser Segmente müssen jeweils bei den einzelnen Aufrufen des Makros *SILMARKE zur Spezifikation BEREICH angegeben werden.

Der Bereich *teil_1* enthält nur Wörter mit 4 bis 7 Buchstaben (8854 Einträge). Wenn also nur Wörter mit 8 oder mehr Buchstaben überhaupt getrennt werden sollen, so genügt es, die übrigen Bereiche außer *teil_1* bei *SILMARKE zu benutzen.

Ausnahmen-Lexikon für das Französische

Herr Prof. Dr. Kurt Kloocke vom Romanischen Seminar der Universität Tübingen hat für die von ihm betreute Edition der Werke von Benjamin Constant eine Liste von französischen Wörtern zusammengestellt und mit entsprechenden Markierungen versehen, die mit TUSTEP falsch getrennt wurden. Er stellt diese Daten ebenfalls allen TUSTEP-Benutzern zur Verfügung. Die entsprechende Datei heißt `zrlddv1*franzsil` und enthält derzeit knapp 3000 Einträge. Sie ist kurz genug, um mit einem einzigen Aufruf von *SILMARKE verarbeitet zu werden. Sie steht an den im vorigen Abschnitt genannten Stellen zur Benutzung bereit. Auch diese Datei ist eine Segmentdatei mit den beiden Segmenten *inhalt* und *alles*. Beim Aufruf von *SILMARKE muß zur Spezifikation *bereich* der Segmentname *alles* angegeben werden.

Beide Herren bitten zu berücksichtigen, daß die Dateien unvollständig sind (trotz der großen Zahl der Belege fehlen z. B. viele Flexionsformen) und bitten um Verständnis dafür, daß sie keine Gewähr für die Richtigkeit der Einträge übernehmen können.

*Wilhelm Ott*
*tustep@zdv.uni-tuebingen.de*