Möglichkeiten zur Integration der sicheren sinnentsprechenden Silbentrennung in TEX

Mag. Gabriele Kodydek

Technische Universität Wien
Institut für Computergraphik und Algorithmen

e-mail: kodydek@ads.tuwien.ac.at

http://www.ads.tuwien.ac.at/research/SiSiSi.html

1. März 2001

Übersicht

1. Einleitung
2. Sichere sinnentsprechende Silbentrennung (SiSiSi)
3. Silbentrennung in TEX
4. SiSiSi und TEX
5. Zusammenfassung

Motivation

- *Schriftbild:*
  qualitativ hochwertige Dokumente erfordern Silbentrennung

  ◇ ohne Silbentrennung: große Wortzwischenräume
  (wegen langer Wortzusammensetzungen)

  ◇ mit Silbentrennung: ruhigeres Schriftbild

- *häufige Trennfehler:*
  Grund: Wortzusammensetzungen werden nicht erkannt

  → falsche Trennung: z.B. *Grammati-kregeln*, *Wor-tende*

Silbentrennung durch Wortzerlegung

1. Schritt: Zerlegung des Wortes

Vormerken von Haupttrennstellen (=) an Wortfugen und
von Nebentrennstellen (-) nach Vorsilben

z.B. `Wort=zerlegungs=ver-fahren`

2. Schritt: Auffinden weiterer Trennstellen

Silbentrennung auf Basis der Vokal-Konsonanten-Folgen
Vormerken von Nebentrennstellen innerhalb von
Stamm mit Endung(en) sowie in mehrsilbigen Vorsilben

z.B. `Wort=zer-le-gungs=ver-fah-ren`

Sichere **Si**nnentsprechende **Si**lbentrennung

sinnentsprechend:

Vergabe unterschiedlicher Prioritäten ermöglicht:

- Bevorzugung von Haupttrenstellen (die den Lesefluss fördern)
- Zurückdrängung von Nebentrennstellen (die eher stören)

sicher:

alle legalen Zerlegungen und ihre Trennstellen werden ermittelt

nicht in allen Zerlegungen vorkommende Trennstellen:
unsichere Trennstellen

z.B. `Wach=stu-be, Wachs=tu-be`

Bestandteile

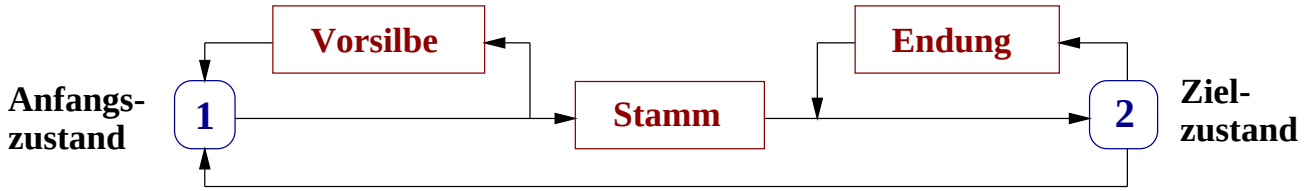
- **Atomtabelle**: enthält ca. 6000 atomare Bestandteile deutscher Wörter zusammen mit ihren Attributen
- **Grammatikregeln**: legen fest, wie die Atome sinnvoll kombiniert werden dürfen
- **Rekursiver Zerlegungsalgorithmus**: zerlegt die Wörter in ihre atomaren Bestandteile gemäß der Atomtabelle und den Grammatikregeln, merkt Trennstellen an Wortfugen vor
- **Dudentrennung**: findet Trennstellen in Einzelwörtern auf Basis der Vokal-Konsonaten-Folgen

Prinzipien der Wortanalyse

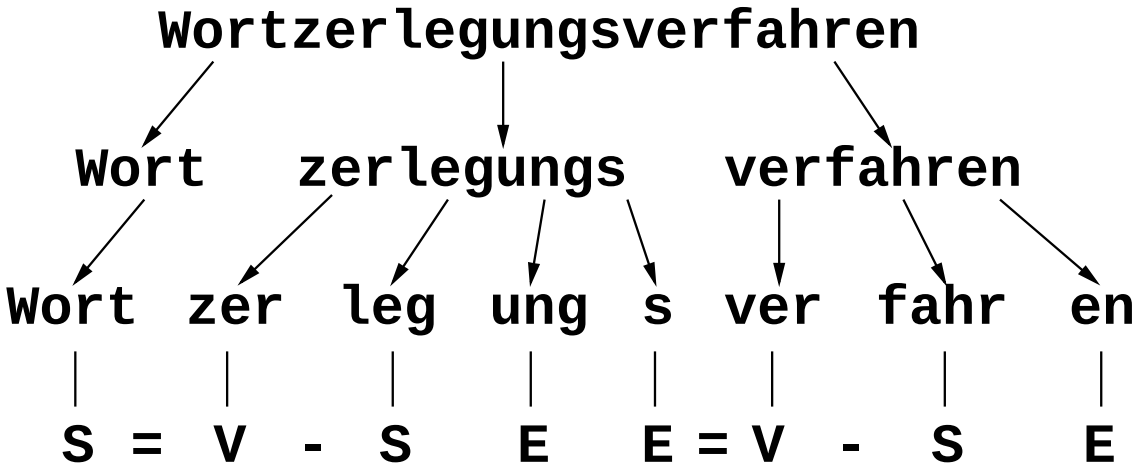
Einfache Klassifizierung:

- ◇ Stämme (S): `text, arbeit`
- ◇ Vorsilben (V): `ver`
- ◇ Endungen (E): `en, ung`

Einfache Regeln zur Wortbildung:



Beispiel



Folgt auf Stamm oder Endung wieder Vorsilbe oder Stamm
→ Haupttrennstelle

Folgt auf Vorsilbe wieder Vorsilbe oder Stamm → Nebentrennstelle

Zerlegungsalgorithmus im Pseudocode

```
PROCEDURE zerlegen(zustand, wortrest)
BEGIN
   IF (wort ist Leerstring) AND (zustand ist Endzustand) THEN
      Trennvektor an Dudentrennung übergeben
   ELSE
      FOR i ← 1 TO Länge von wortrest DO
         IF wortrest[1..i] ist Atom THEN
            FOR atomklasse IN Atomklassenmenge dieses Atoms DO
               IF Übergang(zustand, atomklasse, neuzustand) THEN
                  eventuell Trennstelle in Trennvektor eintragen
                  zerlegen(neuzustand, wortrest[i+1..Länge])
END
```

Dudentrennung

Trennregeln für *Stämme mit Endungen* bzw. *mehrsilbige Vorsilben*

Wesentliche Trennregeln:

- Trennung vor letztem Mitlaut (`Bü-cher`)
- Trennung zwischen selbständigen Selbstlauten (`Befrei-ung`)

Vermerk von *Ausnahmen* in der Atomtabelle (vorwiegend Fremdwörter)

z.B. `helikopter as_6` → `He-li-ko-pter`

Ergebnis der Silbentrennung

Trennvektor

- *Haupttrennstellen* (=): zwischen Wortgrenzen, bevorzugt
- *Nebentrennstellen 1. Ord.* (-): innerhalb der Einzelwörter
- *Nebentrennstellen 2. Ord.* (_): „unerwünschte“ Nebentrennstellen

z.B.: `Wortzerlegungsverfahren`
`...=.._.-....=.._..-...`

Mehrdeutige und unbekannte Wörter

- **eindeutig** → eindeutiger Trennvektor
- **mehrdeutig**:

  unsichere Trennungen werden unterdrückt:
  `Spiel=en-de, Spie-len-de` → `Spielen-de`

  nur vom Autor zu beseitigen: `Wachs=tube/Wach=stube`

  ungewöhnliche Zerlegungen: `Messer=attentat/Messe=ratten=tat`

- **unbekannt:**
  - ◇ Atomtabelle unvollständig
  - ◇ Rechtschreibfehler
  - ◇ Eigenname/Fremdwort/Abkürzung

Einfache Wortgrammatik

Einfache Wortgrammatik → erlaubt zuviele unsinnige Wörter

z.B.: `Generalintendant`

```
1. Zerlegung:  General  +  intendant
                  S           S
2. Zerlegung:  General  +  in  +  t  +  end  +  ant
                  S        E     E      E       E
```

Berücksichtigung der deutschen Wortbildungsgrammatik
zur Verhinderung von unsinnigen Zerlegungen

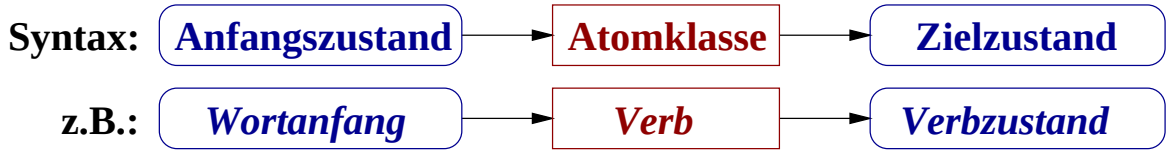
Verbesserte Wortgrammatik

Atomtabelle: Klassifizierung nach Wortarten (ca. 200 Klassen), z.B.

```
teil ... Substantiv, Verb
alt  ... Adjektiv
s    ... Fugenzeichen, Substantivendung, Adverbableitung
```

Wortgrammatik: ca. 3500 Regeln für Wortbildung

- ◇ Menge von Atomklassen und Zuständen
- ◇ Menge von Zustandsübergängen



Rechtschreibreform

Neue Rechtschreibung seit 1998 (alte bis 2005 gültig)

Neuschreibung mancher Stämme: z.B. *daß* → *dass*

Neue Trennregeln:

- ◇ *ck* bleibt ungetrennt: z.B. *Bä-cker* (statt *Bäk-ker*)
- ◇ *st* wird getrennt: z.B. *kos-ten* (statt *ko-sten*)
- ◇ zwei Trennvarianten für manche Wörter:
  1. nach der Herkunft: z.B. *He-li-ko-pter*
  2. nach Sprechsilben: z.B. *He-li-kop-ter*

SiSiSi anwendbar auf alte und neue Rechtschreibung

Silbentrennung in TEX

- Automatisch: *pattern*-Methode von Liang
- Manuelle Eingabe von Trennstellen
- Penalty-Werte für Trennstellen
- Trennung von Zusammensetzungen

Pattern-Methode von Liang

ursprünglich für Englisch entwickelt, für Deutsch adaptiert

Trennung eines Wortes wird aus *patterns* abgeleitet

z.B.: `algorithm` → `l1g4 lgo3 1go 2ith 4hm`
→ `al1g4o3r2it4hm` → `al-go-rithm`

automatische Erzeugung der *pattern*-Tabelle
(aus Liste aller Wortformen mit allen möglichen Trennstellen)

Eigene *pattern*-Tabellen für alte und neue Rechtschreibung
(neue Trennregeln selektiv umgesetzt)

Manuelle Eingabe von Trennstellen

Im Deutschen häufig Fehler bei neuen Wortzusammensetzungen

z.B. *Grammati-kregeln*

→ Der Benutzer kann selbst geeignete Trennstellen vorgeben

- direkt im Text: mit `\-` oder `\discretionary`, z.B.

  `Grammatik\-regel`, `Bä\discretionary{k-}{k}{ck}er`

  Nachteil: alle Vorkommnisse einzeln eintragen

- als Liste im Vorspann: mit `\hyphenation{...}`

  Nachteile: alle Wortformen eintragen, begrenzter Speicher

Penalty-Werte für Trennstellen

Optimale Absatzgestaltung:
für jede Trennstelle ein *penalty*-Wert (`hyphenpenalty`) vergeben

- negativ: fördert Trennung
- positiv: hindert Trennung

pro Absatz nur *ein penalty*-Wert für Trennstellen definierbar
→ keine Bevorzugung von Haupttrennstellen möglich

Trennung von Zusammensetzungen

Problematik der Trennung von Zusammensetzungen → Forderung nach sprachspezifischem Modul

Inzwischen – Ansatz von Petr Sojka:

◇ Umfangreichere *pattern*-Tabelle auf Basis einer Wortliste, in der Wortfugen in Zusammensetzungen markiert sind

◇ kommt mit kleinen Eingriffen in TeX aus

◇ Einführung einer `\compoundwordhyphenpenalty`

Anwendung von SiSiSi für TEX

Frühere SiSiSi-Versionen

- SiTEX: Erweiterung von TEXund LATEX
- iSiTEX: interaktive Version
  (erlaubt Editieren bei unbekannten/mehrdeutigen Wörtern)

◇ Trennalgorithmus von Liang wurde komplett durch SiSiSi ersetzt

◇ Einführung eines zweiten Strafwertes für Trennstellen
(`\nebenhyphenpenalty`)

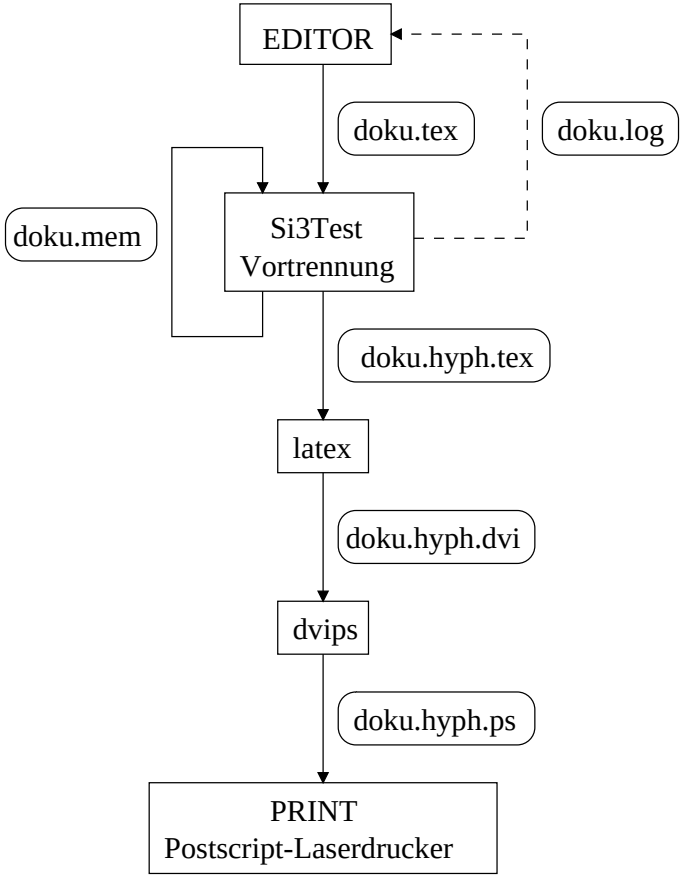
Vortrennung für neue SiSiSi-Versionen

Mit verbesserter Wortgrammatik und neuer Rechtschreibung

Möglichkeiten zur Anwendung von SiSiSi in TEX:

- direkte Implementierung (noch nicht realisiert)
- vorläufig: Vortrennung mittels Preprocessors
  - ◇ Kennzeichnung der Trennstellen im Text
  - ◇ Alternative: Trennliste im Vorspann

Vorgangsweise



Vortrennung von TeX-Dokumenten
mit SiSiSi mittels Preprocessors

Interaktive Vortrennung

Wort mit mehreren Trennvarianten:

überwiegende

Gewünschte Trennvariante auswählen oder eigene Trennvariante eingeben:

ü_ber-wiegen_de

ü_ber-wieg=en_de
ü_ber-wie-gen-de
ü_ber-wiegen_de

Übernehmen

☐ Im ganzen Dokument

Ursprünglicher Text:

(3) Der Hauptwohnsitz einer Person ist dort begründet, wo niedergelassen hat, hier den Mittelpunkt ihrer Lebensbezie beruflichen, wirtschaftlichen und gesellschaftlichen Leben bezeichnen, zu dem sie das überwiegende Naheverhältnis hat deklarative Landesbürgerschaft

Optionen für die Vortrennung

Version
- ○ ReSi
- ○ HelSi
- ○ UrSi

Trennvariante (bei ReSi)
- ○ nach Silben
- ○ nach Herkunft
- ○ alle

Mitprotokollieren
- ☐ mehrdeutige Wörter
- ☐ unbekannte Wörter

Trennvariante (bei ReSi)
- ○ wie im Trennvektor
- ○ wie in LaTeX
- ○ spezielle Zeichenfolge ▭

Trennprofil
- ○ optimale Trennstellen
- ○ nur Haupttrennstellen
- ○ Haupt- und Nebentrennstellen 1. Ord.
- ○ alle Trennstellen

Zusammenfassung

- *Pattern*-Methode ist für Zusammensetzungen nicht gut geeignet
- SiSiSi: sichere, sinnentsprechende Silbentrennung für die deutsche Sprache
- mittels eines Preprocessors kann SiSiSi auf TEX-Dokumente angewendet werden
- Unterscheidungsmöglichkeit für Haupt- und Nebentrennstellen in TEX wünschenswert
- Ziel: direkter Einbau von SiSiSi in TEX