Ein Wortanalysesystem für Silbentrennung, Volltextsuche und Rechtschreibprüfung unter Berücksichtigung der Rechtschreibreform

Gabriele Kodydek, Institut für Computergraphik, Abteilung für Algorithmen und Datenstrukturen,
Technische Universität Wien, Favoritenstraße 9-11/186, A-1040 Wien, Österreich
kodydek@apm.tuwien.ac.at, http://www.apm.tuwien.ac.at/

Kurzfassung

Deutsche Wörter erfordern in Textverarbeitungssystemen besondere Aufmerksamkeit aufgrund der Möglichkeit, jederzeit neue zusammengesetzte Wörter aus bereits existierenden zu erzeugen. Daher präsentieren wir ein Wortanalysesystem, welches die Analyse von deutschen Wörtern gemäß ihrer atomaren Bestandteile ermöglicht. Ein rekursiver Zerlegungsalgorithmus, der die Regeln für Flexion, Derivation und Bildung von Zusammensetzungen berücksichtigt, zerlegt die Wörter in ihre kleinsten relevanten Bestandteile (*Atome*), welche in einer Atomtabelle gespeichert sind. Das System basiert auf den in diesem Aufsatz beschriebenen Grundlagen und wird zur sicheren, sinnentsprechenden Silbentrennung (SiSiSi) sowie zur sinnentsprechenden Volltextsuche eingesetzt; in eingeschränkter Form kann es auch zur Rechtschreibprüfung herangezogen werden.

1 Einleitung

Ein besonderes Merkmal der deutschen Sprache ist die Möglichkeit, jederzeit neue Wörter als Zusammensetzungen aus bereits existierenden Wörtern zu erzeugen. Diese Besonderheit erfordert in Textsystemen eine spezielle Behandlung der deutschen Wörter. Dazu wird ein umfassendes Wortanalysesystem vorgestellt, das es ermöglicht, alle Wörter in deutschsprachigen Texten hinsichtlich ihrer atomaren Bestandteile zu analysieren. Die Wörter werden gemäß der deutschen Wortbildungsgrammatik durch einen rekursiven Zerlegungsalgorithmus in ihre kleinsten für die Wortzerlegung noch relevanten Bestandteile (*Atome*) zerlegt. Da für jedes Wort alle möglichen Zerlegungen bestimmt werden, können alle kritischen Fälle erkannt und sinnvoll bearbeitet werden. Etwa 6000 Atome, die in einer Atomtabelle zusammen mit ihren Attributen gespeichert werden, reichen bereits für die Analyse fast aller deutschen Wörter und der gängigen Fremdwörter aus.

In Abschnitt 2 werden andere Verfahren zur Analyse von Wortformen vorgestellt. Abschnitt 3 beschreibt die Grundlagen der Wortanalyse sowie ihre Anwendung auf den Gebieten der Silbentrennung, Volltextsuche und Rechtschreibprüfung. Abschnitt 4 erläutert die Anpassung des Wortanalysesystems an die reformierte Rechtschreibung. Abschnitt 5 fasst schließlich die Eigenschaften des Systems zusammen und gibt Auskunft über weitere Arbeiten im Zusammenhang mit dem Wortanalysesystem.

2 Andere Verfahren

Eine Reihe von Morphologiesystemen für die deutsche Sprache arbeitet ähnlich wie das hier präsentierte Wortanalysesystem, jedoch mit unterschiedlicher Zielsetzung und dadurch bedingt auch mit anderen Schwerpunkten bei der Realisierung. Als Beispiele seien hier das als Freeware erhältliche Morphologiesystem Morphy [7] und das auf dem Zwei-Ebenen-Modell TWOL [6] basierende Morphologiesystem für die deutsche Sprache GERTWOL [5] genannt, welche ebenfalls auf einem Stammlexikon (das entspricht der Atomtabelle), der Einteilung der Stämme nach ihren Wortklassen und bestimmten Wortbildungsregeln basieren. Im Vordergrund stehen bei diesen Systemen die Rückführung einer Wortform auf ihren Stamm (Lemmatisierung) und eine anschließende kontextfreie oder auch kontextbezogene Kategorisierung (Tagging), welche Informationen über die grammatikalische Funktion eines Wortes innerhalb des Satzes liefert. Oft enthalten solche Systeme auch eine Synthesefunktion zur Erzeugung von Wortformen. Meist wird auch die Anwendung der morphologischen Analyse zur Volltextsuche erwähnt. Unseres Wissens nach wird jedoch keines dieser Systeme zur sinnentsprechenden Silbentrennung verwendet.

Zusammensetzungen, die sich auf mehr als eine Art lesen lassen (z.B. *Wachstube*) werden in den Morphologiesystemen unterschiedlich gehandhabt. Einige, wie z.B. GERTWOL, suchen zwar alle möglichen Zerlegungsvarianten eines Wortes, nehmen aber bestimmte

Wörter, die häufig Mehrdeutigkeiten liefern (wie etwa *au*, *ende*) von der Kompositionsbildung aus und speichern Zusammensetzungen mit solchen Wörtern in einem Teillexikon ab. Das kann dazu führen, dass neue Wortschöpfungen mit diesen Wörtern nicht erkannt werden können. Andere dagegen, wie etwa Morphy, welches Zusammensetzungen nach einer *longest-matching-rule* analysiert, liefern für die auf unterschiedliche Weise zerlegbaren Wörter meist nur eine der möglichen Zerlegungsvarianten. In unserem Wortanalysesystem werden vor allem im Hinblick auf die Sicherheit bei der Silbentrennung alle grammatikalisch korrekten Zerlegungen ermittelt.

3 Das Wortanalysesystem und seine Anwendungen

3.1 Die Wortanalyse

In einer ursprünglichen Version [1] werden die Atome nach ihrer Funktion bei der Wortbildung lediglich in Vorsilben (V), Stämme (S) und Endungen (E) eingeteilt. Die entsprechenden primitiven Grammatikregeln für die Wortbildung lauten: Ein Einzelwort wird aus beliebig vielen Vorsilben, einem Stamm und beliebig vielen Endungen (einschließlich Fugenzeichen) gebildet; ein zusammengesetztes Wort besteht aus beliebig vielen Einzelwörtern. **Bild 1** zeigt die Wortanalyse am Beispiel des zusammengesetzten Wortes *Wortzerlegungsverfahren*.

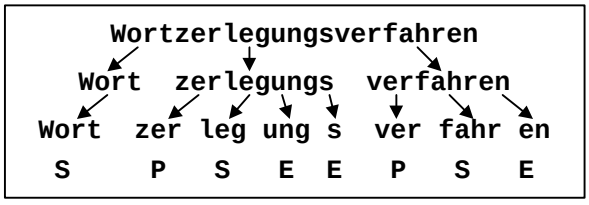

Bild 1 Beispiel für die Zerlegung eines zusammengesetzten Wortes

Jedes Atom ist in der Atomtabelle zusammen mit einer Menge von Attributen, die seiner Klassifikation entsprechen, gespeichert. Dabei ist es möglich, dass ein Atom aufgrund verschiedener Funktionen bei der Wortbildung in mehrere Klassen fällt: beispielsweise kann das Atom *end* sowohl als Stamm (*enden*) als auch als Endung für das 1. Partizip (*gehend*) verwendet werden.

Der in **Bild 2** dargestellte Zerlegungsalgorithmus besteht im Wesentlichen aus der Suche von Atomen in der Atomtabelle und dem Aneinanderfügen der gefundenen Atome gemäß bestimmter Grammatikregeln.

Die äußere *for*-Schleife testet für jeden möglichen Anfang des Wortrestes *wr*, ob er ein Atom ist. Wenn das der Fall ist und das Atom an dieser Stelle angefügt werden darf, wird durch einen rekursiven Aufruf mit dem um das gefundene Atom verkürzten Wortrest untersucht, ob es zu einer legalen Zerlegung kommt. Ob ein Atom an dieser Stelle angefügt werden darf, wird anhand seiner Atomklasse und des aktuellen Zustandes festgestellt. Eine Atomklasse *ak* ist dann erlaubt, wenn es eine Grammatikregel gibt, die für diese Atomklasse einen Übergang aus dem aktuellen Ausgangszustand *az* in einen Zielzustand *zz* ermöglicht. Aufgrund der mehrfachen Klassenzugehörigkeit müssen für ein in der Tabelle enthaltenes Atom *wr[1..i]* in der inneren *for*-Schleife alle seine Atomklassen der Reihe nach betrachtet werden. Jede Atomklasse, die an dieser Stelle erlaubt ist, verursacht dann einen rekursiven Aufruf mit dem entsprechenden Zustand *zz*. Eine gültige Zerlegung ist dann erreicht, wenn der Zustand nach der Abarbeitung des gesamten Wortes ein zulässiger Endzustand ist; andernfalls wird die gefundene Zerlegung verworfen. Es ist zu beachten, dass der Algorithmus nicht endet, wenn das Wort zum ersten Mal ganz abgearbeitet wurde, sondern es werden danach alle weiteren Zerlegungen gesucht.

```
procedure zerlege(az,wr,inf)
begin
 if (wr = Leerstring) and
    (az = Endzustand) then
  Nachbehandlung
 else
  n:=Länge von wr;
  for i:=n downto 1 do
   if wr[1..i] ist Atom then
    for ak in Atomklassenmenge
        dieses Atoms do
     if Übergang az->ak->zz then
      spezialinfo inf speichern;
      zerlege(zz,wr[i+1..n],inf)
end.
```

Bild 2 Zerlegungsalgorithmus

Die primitive Wortgrammatik von [1] lässt allerdings auch eine Vielzahl unsinniger Wortbildungen zu, wie z.B. Stämme mit beliebig vielen gleichen Endungen. Aufbauend auf die Arbeiten in [11, 12] wurde das Verfahren durch Einteilung der Atomklassen nach Wortarten und Schaffung entsprechender Grammatikregeln für die Zusammensetzung dieser Elemente deutlich verbessert. Die neue Klassifizierung sieht für die Stämme z.B. die Klassen Substantiv (*StSb*), Verb (*StVb*), Adjektiv (*StAd*) und Pronomen (*Pn*) vor. Bei den Endungen wird zwischen Flexionsendungen, Derivationsendungen und Fugenzeichen unterschieden. Die Flexionsendungen werden dabei weiter unterteilt in Deklinationsendungen für Substantive (*DkSb*), Konjugationsendungen für Verben (*KjVb*) und dergleichen. Daneben gibt es mehrere Klassen mit Deri-

vationsendungen, die etwa einen Verbstamm in ein Substantiv ableiten. Bei den Vorsilben wird unterschieden zwischen allgemeinen Vorsilben (*VoAllg*) und speziellen Vorsilben, wie etwa verbableitenden Vorsilben (*VoAbVb*). Eine weitere Verbesserung wird durch die Unterteilung der Stammklassen in Unterklassen entsprechend ihrer Flexionsendungen erzielt: beispielsweise fällt das Substantiv *Ding*/*Ding**e*** in die Klassen *StSbS1* (Substantiv mit Genetiv Singular auf *(e)s*) und *StSbP1* (Substantiv mit Nominativ Plural auf *e*), während dem Substantiv *Kind*/*Kind**er*** die Klassen *StSbS1* und *StSbP5* (Substantiv mit Nominativ Plural auf *er*) zugeordnet werden. In Kombination mit einer entsprechenden Klassifizierung der Flexionsendungen und mit passenden Grammatikregeln garantiert diese Einteilung, dass das System einen Stamm mit Flexionsendungen nur dann als gültig betrachtet, wenn die verwendete Endung seiner Unterklasse entspricht. Andernfalls wird das Wort als falsch zurückgewiesen. Auch die Verwendung von Endungen bei der Bildung von Komposita wurde durch die genauere Klassifizierung und entsprechende Grammatikregeln eingeschränkt: nur auf den letzten Stamm eines Kompositums darf eine (passende) Flexionsendung folgen, alle anderen Stämme können allerdings von der Wortklasse entsprechenden Fugenzeichen (z.B. Fugen-*s*) oder Derivationsendungen (z.B. *ung*) gefolgt sein.

Während in der ursprünglichen Version nur wenige Grammatikregeln existierten, die fix in den Zerlegungsalgorithmus implementiert werden konnten, gibt es in der verbesserten Version eine Vielzahl von Grammatikregeln. Die Beschreibung dieser Grammatik durch Anweisungen im Programmcode wäre hierbei zu kompliziert, daher werden die Atomklassen und die durch sie bedingten Zustände so zu Regeln verknüpft, dass ein Wortzerlegungsautomat entsteht. Die Regeln werden in der Form

Ausgangszustand ® *Atomklasse* ® *Zielzustand*

in einer separaten Datei gespeichert. Die Zustände sind im Wesentlichen nach den Atomklassen benannt, da es für die meisten Atomklassen einen eigenen korrespondierenden Zustand gibt, in den der Automat nach einem Atom der betreffenden Atomklasse übergeht. Als Ausgangszustand am Anfang eines Wortes dient der spezielle Zustand *Wortanfang*. Von den übrigen Zuständen werden jene als Endzustände gekennzeichnet, die das Ende einer gültigen Zerlegung andeuten. Der echte Endzustand *Wortende* darf jederzeit als Zielzustand benützt werden, wenn das Wort nach dem Auftreten einer bestimmten Atomklasse unbedingt zu Ende sein muss, z.B. nach Auftreten einer Flexionsendung.

Im Zuge dieser Grammatikverfeinerung wurden die Atome im Bereich der Vorsilben und Endungen bereits vollständig klassifiziert. Die Zuordnung der neuen Atomklassen zu den Stämmen unterblieb jedoch in dieser Phase weitgehend und wurde erst während der Testphase nachgeholt.

3.2 Sichere Sinnentsprechende Silbentrennung

Die automatische Silbentrennung ist in Textverarbeitungssystemen von großer Bedeutung. Um eine optimale Absatzausrichtung ohne stark gefranste Ränder oder (bei Randausgleich) ohne unansehnliche, große Wortzwischenräume zu erreichen, ist es günstig, lange Wörter zu trennen. Dadurch entsteht beim Drucken von hochwertigen Dokumenten wie Büchern und beim Setzen von schmalen Zeitungsspalten ein ruhiges Schriftbild, das den Lesefluss optimal unterstützt.

Die gebräuchlichen Silbentrennverfahren, wie etwa die für das Englische erfolgreich angewandte *pattern*-Methode von Liang [8], basieren auf vollständigen Wörterbüchern und versagen daher im Deutschen aufgrund der Möglichkeit, jederzeit neue Wörter als Zusammensetzungen von existierenden Wörtern erzeugen zu können. Die inhärente Unvollständigkeit der Wörterbücher führt dazu, dass Wortfugen in neuen Wortschöpfungen nicht erkannt werden und in der Folge Trennfehler entstehen können. Aus demselben Grund führen auch Silbentrennverfahren, die auf dem Nachschlagen in einem großen Wörterbuch beruhen, nicht zu befriedigenden Resultaten. Andere Methoden versuchen die Wörter mit Hilfe von Trennregeln, die aus Grammatiken [3, 4] abgeleitet wurden, zu analysieren und zu trennen. Diese Regeln trennen im Wesentlichen bestimmte Vokal-Konsonanten-Folgen, was sehr leicht zu implementieren ist. Für bessere Ergebnisse werden Vorsilben und Endsilben gesondert berücksichtigt. Diese Methode arbeitet allerdings nur für Einzelwörter zuverlässig. An Wortfugen versagt sie ebenso wie die *pattern*-Methode.

Die Silbentrennung von SiSiSi dagegen basiert auf der Zerlegung von zusammengesetzten Wörtern in ihre Einzelwörter (siehe Bild 1). So werden Wortfugen immer erkannt und bereits während des Zerlegungsprozesses als Haupttrennstellen (=) vorgemerkt, zum Beispiel wenn wie in *Wort=zer*... eine Vorsilbe auf einen Stamm folgt.

Trennstellen innerhalb eines Einzelwortes werden dagegen als Nebentrennstellen (–) gekennzeichnet. Diese werden auf unterschiedliche Weise aufgefunden. Die Trennstellen zwischen Vorsilben und Stämmen und vor bestimmten Endsilben wie *chen* oder *heit* werden ebenfalls bereits im Zuge der Wortzerlegung markiert. Die übrigen Nebentrennstellen werden mittels eines zusätzlichen Algorithmus zur Silbentrennung aufgefunden, der auf der Abfolge von Konsonanten und Vokalen beruht. Dabei wird auf der Buchstabenebene eine eindeutige Unterscheidung zwischen konsonanti-

schen und vokalischen Lauten getroffen, die neben den einzelnen Buchstaben des Alphabets auch alle Buchstabengruppen berücksichtigt, die im Sinne der Silbentrennung als ein Laut zu behandeln sind (siehe **Bild 3**).

Konsonanten: b, c, d, f, g, h, j, k, l, m, n, p q, r, s, t, v, w, x, z; ß; ch, ph, qu, sch, st (nur nach alter Silbentrennung), th;
Vokale: a, e, i, o, u; ä, ö, ü, y; ai, au, äu, ei, eu, ie;

Bild 3 Zuordnung von Buchstaben und Buchstabengruppen zu Konsonanten und Vokalen

Es gelten zwei wesentliche Trennregeln, die aus den im Duden [3, 4] angegebenen Regeln zur Silbentrennung abgeleitet sind (daher bezeichnen wir diesen Algorithmus auch als *Dudentrennung*):

(1) Getrennt wird vor einem einzelnen Konsonanten oder vor dem letzten Konsonanten einer Konsonantenfolge (z.B. *Bü–cher, Imp–fung*) oder
(2) zwischen zwei verschieden lautenden Vokalen (z.B. *Mau–er, bö–ig*).

Mit Hilfe der Dudentrennung können alle Nebentrennstellen in einem aus einem heimischen Stamm und beliebigen Endungen bestehenden Wortteil und in mehrsilbigen Vorsilben gefunden werden. Eine Sonderbehandlung erfordern lediglich einige Fremdwörter, für die es in Anlehnung an die Trennregeln in ihrer Herkunftssprache auch im Deutschen besondere Regeln gibt. Die richtige Trennung dieser Fremdwörter wird durch die Angabe sogenannter Ausnahmetrennstellen in der Atomtabelle erzielt. Diese unterdrücken für bestimmte Abschnitte des Atoms die Dudentrennung und setzen bei Bedarf eigene Trennstellen an den korrekten Positionen, z.B. im Wort *Ma–gnet*.

Die Vergabe von unterschiedlichen Bewertungen für Trennstellen ermöglicht es, die den Lesefluss eher fördernden Haupttrennstellen an den Fugen zusammengesetzter Wörter (z.B. *Wort=zerlegungs=verfahren*) gegenüber den eher störenden Nebentrennstellen (z.B. *Wortzer–le–gungsver–fah–ren*) innerhalb der Einzelwörter zu bevorzugen und dadurch die sinnentsprechende Trennung zu fördern. In manchen Fällen erscheint es sinnvoll, bei den Nebentrennstellen weitere Unterteilungen zu treffen, um die Ausnutzung von Trennstellen, die besonders sinnentstellende Trennungen erzeugen, wie z.B. *Spargel–der* oder *Ge–hörner–ven*, noch weiter zu unterdrücken. Es handelt sich dabei durchwegs um Nebentrennstellen, die dicht hinter einer Haupttrennstelle folgen. Sie sollten noch schlechter bewertet werden als andere Nebentrennstellen, damit sie vom Textverarbeitungsprogramm nur in Ausnahmefällen als Trennstellen eingesetzt werden.

Das Hauptaugenmerk von SiSiSi liegt jedoch auf der Sicherheit: Es werden alle möglichen Zerlegungen mit ihren Trennstellen ermittelt, z.B. *Per–son=alm=an–gel* und *Per–so–nal=man–gel*; nur jene Trennstellen, die in allen Zerlegungen vorkommen, sind sicher, d.h. niemals falsch; alle anderen sollten nur mit äußerster Zurückhaltung, z.B. nach Befragung des Benutzers, verwendet werden, weil es möglich ist, dass eine solche unsichere Trennstelle nur in einer nicht beabsichtigten Zerlegung vorkommt.

3.3 Sinnentsprechende Volltextsuche

Die Suche nach Dokumenten, die ein oder mehrere bestimmte Schlagwörter enthalten, erfolgt oft mittels *pattern-matching*-Methoden. Dies hat den Nachteil, dass manchmal Dokumente gefunden werden, die den Erwartungen des Benutzers nicht entsprechen (Schlagwort: *Autor* → gefundenes Dokument enthält *Autorennen*).

Flexiblere Möglichkeiten bietet dagegen die auf der Wortanalyse basierende sinnentsprechende Volltextsuche [2], bei der sowohl die Schlagwörter als auch die Wörter in den Textdokumenten in ihre sinngebenden Bestandteile zerlegt werden. Der Sinn eines zusammengesetzten Wortes wird durch seine Einzelwörter (z.B. *Textverarbeitungssystem* → *text, verarbeitung, system*) bestimmt; der Sinn eines Einzelwortes durch den Stamm in Verbindung mit einer eventuell vorhandenen Vorsilbe (*verarbeitung* → *verarbeit*); Endungen sind in der Regel unbedeutend.

Gewöhnlich werden flektierte Formen von Substantiven, Verben und Adjektiven durch Anhängen eines bestimmten Suffixes an den Stamm gebildet. Es gibt jedoch eine beträchtliche Anzahl von unregelmäßigen Stämmen, bei welchen sich die Schreibweise des Stammes in flektierten Formen verändert, so dass er als eigenes Atom gespeichert werden muss, wie etwa das Adjektiv *hoch* und seine Steigerungsform *höher*. In diesen Fällen muss darauf geachtet werden, dass die unterschiedlichen Schreibweisen des Stammes zueinander in Beziehung gesetzt werden, so dass eine Suche nach verwandten Wörtern mit demselben Sinn möglich ist. Dies wird durch das Konzept der Wortfamilien erreicht. Eine Wortfamilie umfasst alle unterschiedlichen Schreibweisen eines bestimmten Stammes. Beispielsweise repräsentiert die Wortfamilie <*gehen v*> (*v* steht dabei für Verb) die Stämme {*geh, ging, gang*}. Die Bezeichnung der Wortfamilien folgt dabei folgender Konvention: Substantive werden im Nominativ Singular angegeben, gefolgt vom Geschlecht (*m*/*f*/*n*) und dem Buchstaben *s*; Verben werden im Infinitiv angegeben, gefolgt vom Bezeichner *v*; Adjektive werden im Positiv ohne Deklinationsendungen angegeben, gefolgt vom Wortartenbezeichner *adj*. Die Wortfamilie wird als Attribut bei den betroffenen Stämmen in der Atomtabelle vermerkt. Als Ergebnis der Wortanalyse wird für unregelmäßige Stämme also

die zugehörige Wortfamilie (anstelle des Wortstammes) geliefert, was die Suche nach allen möglichen Wortformen eines eingegebenen Schlagwortes ermöglicht.

3.4 Rechtschreibprüfung

In eingeschränktem Maß ermöglicht die hier beschriebene Wortanalyse auch eine Rechtschreibprüfung. Falls nämlich keine Zerlegung gefunden wird, handelt es sich meistens um eine Konstruktion mit orthographischem oder grammatikalischem Fehler. Die einzig andere Möglichkeit besteht darin, dass die Atome des Wortes nicht in der Atomtabelle enthalten sind, wenn das zu analysierende Wort etwa ein Eigenname oder ein seltenes Fremdwort ist. In diesem Fall kann das betreffende Atom mit geeigneten Attributen mühelos zur Atomtabelle hinzugefügt werden; dadurch werden in Zukunft auch etwaige Zusammensetzungen mit diesem Atom erkannt.
Das System kann jedoch nicht alle Rechtschreibfehler finden, da in manchen Fällen ein Rechtschreibfehler ein Wort in ein anderes Wort verwandelt, für welches das System eine gültige Zerlegung findet, auch wenn diese mitunter unsinnig sein mag.

4 Umsetzung der Rechtschreibreform

Die Reform der deutschen Rechtschreibung [10] im Jahre 1998 erforderte eine sorgfältige Überarbeitung unseres Wortanalysesystems. Aufgrund der neuen Regeln zur Laut-Buchstaben-Zuordnung ist für bestimmte Wörter eine andere Orthographie vorgeschrieben. In manchen Fällen wird dabei die alte Schreibweise durch die neue ersetzt (z.B. *rau* statt *rauh*, *Fluss* statt *Fluß*), in anderen Fällen tritt eine zusätzliche Schreibweise neben die bisherige (neben *Delphin* wird auch *Delfin* erlaubt). Das System wurde dahingehend adaptiert, dass die Wortanalyse sowohl nach dem neuen als auch nach dem alten Regelwerk möglich ist, da die beiden Regelwerke noch bis 2005 nebeneinander gültig sind und ein Großteil der privaten Anwender sich noch an die alte Rechtschreibung hält. Da sich die Änderungen in der Schreibweise der Wörter nur auf einen geringen Prozentsatz aller Atome auswirken, wird weiterhin nur eine Atomtabelle geführt, die aber folgendermaßen modifiziert wurde: Für neue Schreibvarianten wurden neue Atome aufgenommen und als *neu* gekennzeichnet; ebenso wurden alle Atome, die nur nach dem alten Regelwerke gültig sind, als *alt* gekennzeichnet (z.B. neu: *rau*, alt: *rauh*; neu: *Delfin*).

Im Zuge der Rechtschreibreform wurde auch die Silbentrennung neu geregelt. Die offensichtlichste Änderung betrifft hierbei die Buchstabenfolgen *ck* und *st*: während *ck* neuerdings ungetrennt bleibt (früher *k–k*), ist die Trennung von *st* nun erlaubt. Die Lautzuordnung im Algorithmus zum Auffinden der Trennstellen in Einzelwörtern wurde entsprechend abgeändert: *ck* wurde zur Menge der Konsonanten hinzugefügt, *st* daraus entfernt. Ebenso wurde berücksichtigt, dass die neuen Regeln auch die Trennung eines einzelnen Vokals am Wortanfang zulassen, z.B. *A–der*. Da sich eine solche Trennstelle aber auf den Lesefluss äußerst störend auswirken kann (insbesondere in Zusammensetzungen wie *Schlaga–der*), wird sie als besonders nachrangig zu behandelnde Nebentrennstelle markiert.
SiSiSi kann Wörter sowohl nach den neuen als auch nach den alten Rechtschreibregeln trennen. So ermöglicht SiSiSi nach der neuen Rechtschreibung in jenen Wörtern, die neuerdings entweder nach Sprechsilben oder nach der Herkunft getrennt werden können, die Trennung nach beiden zulässigen Varianten: z.B. *He–li–kop–ter* (nach Sprechsilben) und *He–li–ko–pter* (nach seinen griechischen Bestandteilen *helix* und *pterón*). Davon betroffen sind neben Zusammensetzungen, die nicht mehr als solche empfunden werden (z.B. *Helikopter*, *darin*) auch Fremdwörter mit bisher untrennbaren Konsonantenverbindungen mit *r* oder *l*, sowie *gn* und *kn* (z.B. *Magnet*). Zur Trennung nach der Herkunft wird die Trennung gemäß der beim Atom angeführten Ausnahmetrennstellen durchgeführt. Zusätzlich erfolgt für die betroffenen Wortteile eine Trennung nach Silben durch die Anwendung des Trennalgorithmus für die Trennung nach Konsonanten-Vokal-Folgen. Die beiden resultierenden Trennvektoren werden danach zu einem gesamten Trennvektor kombiniert. Dabei wird jede Trennstelle, die nur in einer der beiden Trennvarianten auftritt, als solche markiert: Trennstellen nach Herkunft werden durch „ , “ gekennzeichnet, Trennstellen nach Silben durch „ ’ “. Trennstellen, die in beiden Varianten vorkommen, werden weiterhin als gewöhnliche Nebentrennstellen (–) betrachtet. Ein Beispiel dafür findet sich in **Bild 4**. Dasselbe gilt natürlich auch für Zusammensetzungen und abgeleitete Wörter, die aus den betroffenen Wörtern gebildet werden (z.B. *He–li–ko, p’ter=pi–lot, Ma, g’ne–tis–mus)*.

Trennung nach Herkunft:	`He-li-ko-pter`
Trennung nach Silben:	`He-li-kop-ter`
vollständiger Trennvektor:	`He-li-ko,p’ter`

Bild 4 Beispiel für mehrere zulässige Trennvarianten nach der neuen Rechtschreibung

Ebenso erkennt SiSiSi nach der alten Rechtschreibung Wörter, bei denen aufgrund der 3-Konsonanten-Regel einer von drei gleichlautenden Konsonanten an der

Wortfuge entfallen ist, weil danach ein Vokal folgt, und trennt diese richtig ab, indem der entfallene Buchstabe wieder eingefügt wird (*Schiffahrt* → *Schiff=fahrt*). Die neuen Regeln benötigen diese Spezialbehandlung nicht, da die 3-Konsonanten-Regel nicht mehr zur Anwendung kommt.

5 Schlussbemerkungen

Das vorgestellte System zur Silbentrennung und Volltextsuche beruht auf der Wortzerlegung.
Das System ist *sicher*: fast alle deutschen Wörter und die häufigsten eingedeutschten Fremdwörter werden vom System korrekt zerlegt. Für die Silbentrennung bedeutet das, dass mit wenigen Einschränkungen alle Wörter richtig getrennt werden, für die sinnentsprechende Volltextsuche, dass alle gewünschten Dokumente gefunden werden.
Das System ist sehr *robust*: Auch wenn ein Wort grammatikalisch auf mehrere Arten zerlegt werden kann, wird keine falsche Trennstelle erzeugt, Unsicherheiten werden als solche erkannt und können sinnvoll behandelt werden. Ebenso werden orthographisch oder grammatikalisch falsche Wörter erkannt. Unbekannte Wörter (z.B. Eigennamen) werden von der Silbentrennung ausgenommen, so dass hier keine Trennfehler entstehen können. Bei der Volltextsuche wird bei nicht zerlegbaren Wörtern nach dem gesamten Wort gesucht.
Das System ist überdies *leicht wartbar*: durch Aufnahme eines neuen Stammes (etwa eines seltenen Fremdwortes) in die Atomtabelle und seine entsprechende Attributierung werden künftig sämtliche Zusammensetzungen mit diesem Wort erkannt.
Das Wortanalysesystem wurde in einer speziell dafür entwickelten Testumgebung mit großen Textdateien eingehend getestet. Dabei kamen besondere Testmethoden zur Anwendung, um die wenigen problematischen Fälle aus der großen Menge der analysierten Wörter herausfiltern zu können, siehe dazu [9]. Aufgrund der Testergebnisse konnte die Atomtabelle um fehlende Wortstämme, zumeist Fremdwörter, ergänzt werden. In der vorliegenden Form kann SiSiSi zur Vortrennung von Texten verwendet werden.[1] Eine direkte Einbindung des Silbentrennalgorithmus in das Textsatzsystem TEX ist geplant. Diese ist jedoch nicht trivial, weil die Ansprüche an die Trennsicherheit in Zweifelsfällen eine Interaktion mit dem Anwender nötig machen, um den korrekten Sinn eines mehrdeutig zerlegbaren Wortes zu erkennen.

[1] Im vorliegenden Dokument wurde die Silbentrennung mittels SiSiSi durchgeführt.

6 Literatur

[1] Barth, W., Nirschl H.: Sichere sinnentsprechende Silbentrennung für die deutsche Sprache. Angewandte Informatik 4, S. 152–159, 1985

[2] Barth, W.: Volltextsuche mit sinnentsprechender Wortzerlegung. Wirtschaftsinformatik, 32. Jahrgang, Heft 5, S. 467–471, 1990

[3] Duden, Grammatik der deutschen Gegenwartssprache (Duden Band 4). 4. Aufl., Hrsg. G. Drosdowski, Bibliographisches Institut, Mannheim, 1984

[4] Duden, Rechtschreibung der deutschen Sprache (Duden Band 1). 20. Aufl., Hrsg. G. Drosdowski, Dudenverlag, Mannheim, 1991

[5] Haapalainen, M., Majorin, A.: GERTWOL und morphologische Disambiguierung für das Deutsche. In: Proceedings of the 10th Nordic Conference on Computational Linguistics, Ed.: K. Koskenniemi, Helsinki, Finland, 1995

[6] Koskenniemi, K.: Two-Level morphology: A General Computational Model for Word-Form Recognition and Production, Helsinki, 1983

[7] Lezius, W., Rapp, R., Wettler, M.: A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German. In: Proceedings of the COLING-ACL 1998, Montreal, Canada, 1998

[8] Liang, F. M.: Word Hy-phen-a-tion by Com-puter. Ph.D. Thesis, Dep. of Computer Science, Stanford University, Report No. STAN-CS-83-977, 1983

[9] Schönhacker, M., Kodydek, G.: Testmethoden für die sichere sinnentsprechende Silbentrennung und andere Anwendungen einer Wortanalyse. In: Tagungsband Konvens'2000, Ilmenau, Deutschland, 2000

[10] Sitta, H., Gallmann, P.: Duden, Informationen zur neuen deutschen Rechtschreibung, 2. aktualisierte Ausgabe, Hrsg. Dudenredaktion. Dudenverlag, Mannheim, 1996

[11] Steiner, H.: Automatische Silbentrennung durch Wortbildungsanalyse. Dissertation, Institut für Computergraphik, Technische Universität Wien, 1995

[12] Steiner, H., Barth, W.: Sichere sinnentsprechende Silbentrennung mit Berücksichtigung der deutschen Wortbildungsgrammatik. In: Tagungsband Konvens'94, Hrsg. H. Trost, Wien, S. 330–340, 1994

Dieses Projekt wurde zum Teil durch die freundliche Unterstützung der Hochschuljubiläumsstiftung der Stadt Wien unter Geschäftszahl H-75/99 ermöglicht, für die sich die Autorin herzlich bedankt.