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

Abstract

Schlagwörter: Wortanalyse, Silbentrennung, sinnentsprechende Volltextsuche, Rechtschreibprüfung, Rechtschreibreform

Es 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. Etwa 6000 Atome, die in einer Atomtabelle gespeichert werden, reichen bereits für die Analyse fast aller deutschen Wörter und der gängigen Fremdwörter aus. Dieses System baut auf den im Folgenden beschriebenen Grundlagen auf und wird angewendet zur sicheren, sinnentsprechenden Silbentrennung (SiSiSi) und zur sinnentsprechenden Volltextsuche, in eingeschränkter Form auch zur Rechtschreibprüfung.

Wortanalyse In einer ursprünglichen Version [1] werden die Atome dabei nach ihrer Funktion bei der Wortbildung in Vorsilben, Stämme und Endungen eingeteilt. Die entsprechenden primitiven Grammatikregeln für die Wortbildung lauten: Ein Einzelwort wird aus beliebig vielen Vorsilben, einem Stamm und beliebig vielen Endungen gebildet; ein zusammengesetztes Wort besteht aus beliebig vielen Einzelwörtern. Diese primitive Wortgrammatik lässt allerdings auch eine Vielzahl unsinniger Wortbildungen (z.B. Stämme mit beliebig vielen gleichen Endungen) zu. Aufbauend auf die Arbeiten in [5] wurde das Verfahren durch Einteilung der Atomklassen nach Wortarten (Substantiv, Substantivendungen, Verb, Verbendungen etc.) und Schaffung entsprechender Grammatikregeln für die Zusammensetzung dieser Elemente verbessert.

Sichere sinnentsprechende Silbentrennung Silbentrennung unterstützt den Lesefluss durch Vermeidung von großen Wortzwischenräumen und ist daher wichtig für die Erstellung hochwertiger Textdokumente. Die gebräuchlichen Silbentrennverfahren, wie etwa die *pattern*-Methode von Liang [3], die auf vollständigen Wörterbüchern basieren, versagen im Deutschen aufgrund der Möglichkeit, jederzeit neue Wörter als Zusammensetzungen von existierenden Wörtern erzeugen zu können. Die Unvollständigkeit der Wörterbücher führt dazu, dass Wortfugen in Wortzusammensetzungen nicht erkannt werden und in der Folge Trennfehler entstehen können. Die Silbentrennung von SiSiSi dagegen basiert auf der Zerlegung von zusammengesetzten Wörtern in ihre Einzelwörter. So werden Wortfugen immer erkannt und als Haupttrennstellen (=) vorgemerkt. Trennstellen innerhalb eines Einzelwortes werden mittels eines zusätzlichen Algorithmus zur Silbentrennung, der auf der Abfolge von Konsonanten und Vokalen beruht, aufgefunden und als Nebentrennstellen (-) gekennzeichnet. Die bevorzugte Ausnutzung der Haupttrennstellen fördert die sinnentsprechende Trennung von zusammengesetzten Wörtern an den Wortfugen. Das Hauptaugenmerk von SiSiSi liegt jedoch auf der Sicherheit: es werden alle möglichen Zerlegungen ermittelt (z.B. *Per-son=alm=an-gel*, *Per-sonal=man-gel*); nur die Trennstellen, die in allen Zerlegungen vorkommen, sind sicher; alle anderen sollten nur mit äußerster Zurückhaltung, z.B. nach Befragung des Benutzers, verwendet werden.

Sinnentsprechende Volltextsuche Die Suche nach Dokumenten, die 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 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* → *ver+arbeit*), Endungen sind in der Regel unbedeutend. Bei unregelmäßigen Stämmen muss darauf geachtet werden, dass die unterschiedlichen Schreibweisen des Stammes zueinander in Beziehung gesetzt werden (z.B. *Haus - Häuser*), so dass eine Suche nach verwandten Wörtern mit demselben Sinn möglich ist. Dies wird durch das Konzept der Wortfamilien erreicht, das hier vorgestellt werden soll. Eine Wortfamilie umfasst alle unterschiedlichen Schreibweisen eines bestimmten Stammes. Für unregelmäßige Stämme wird also die Wortfamilie anstelle des Wortstammes als Ergebnis der Wortanalyse geliefert.

Rechtschreibprüfung In eingeschränktem Maß ermöglicht die Wortanalyse eine Rechtschreibprüfung. Falls nämlich keine Zerlegung gefunden wird, handelt es sich meistens um einen Eigennamen oder ähnlichen Begriff, der nicht in der Atomtabelle enthalten ist, oder um eine Konstruktion mit orthographischem oder grammatikalischem Fehler.

Umsetzung der Rechtschreibreform Die Reform der deutschen Rechtschreibung [4] im Jahre 1998 erforderte eine weitgreifende Überarbeitung unseres Wortanalysesystems. Aufgrund der neuen Regeln zur Laut-Buchstaben-Zuordnung ist für bestimmte Wörter eine andere Orthographie vorgeschrieben (z.B. *rau* statt *rauh*, *Fluss* statt *Fluß*). Der Zerlegungsalgorithmus und die Atomtabelle wurden dahingehend adaptiert, dass die Wortanalyse sowohl nach dem neuen als auch nach dem alten, noch bis 2005 gültigen Regelwerk möglich ist. Für neue Schreibvarianten wurden neue Atome aufgenommen; alle Atome, die nur nach einem der Regelwerke gültig sind, wurden speziell gekennzeichnet (z.B. neu: *rau*, alt: *rauh*). Im Zuge der Rechtschreibreform wurde auch die Silbentrennung neu geregelt. Der Algorithmus zum Auffinden der Trennstellen in Einzelwörtern wurde entsprechend abgeändert. 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 (z.B. *He-li-kop-ter* bzw. *He-li-ko-pter*), die Trennung nach beiden zulässigen Varianten. Ebenso erkennt SiSiSi nach der alten Rechtschreibung Wörter, bei denen aufgrund der 3-Konsonanten-Regel ein Konsonant an der Wortfuge entfallen ist und trennt diese richtig ab (*Schiffahrt* → *Schiff=fahrt*).

Schlußbemerkungen 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. 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 Textverarbeitungssystem TeX ist geplant. Weil allerdings die Ansprüche an die Trennsicherheit in Zweifelsfällen eine Interaktion mit dem Anwender nötig machen, um den korrekten Sinn eines mehrdeutig zerlegbaren Wortes zu erkennen, ist die Einbindung in TeX nicht trivial.

Zusammenfassung 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*: Der Fall, dass ein Wort grammatikalisch auf mehrere Arten zerlegt werden kann, wird entsprechend berücksichtigt, ohne Fehler zu produzieren. Ebenso werden grammatikalisch falsche Wörter erkannt. Das System ist außerdem *leicht wartbar*: durch Aufnahme eines neuen Stammes (z.B. eines seltenen Fremdwortes) in die Atomtabelle werden künftig sämtliche Wortzusammensetzungen mit diesem Wort erkannt.

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

[4] Sitta, H., Gallmann, P.: Duden, Informationen zur neuen deutschen Rechtschreibung, 2. aktualisierte Ausgabe, hrsg. von der Dudenredaktion. Dudenverlag, Mannheim, 1996.

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

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