Vorabversion. Erschienen in: Albert Bremerich-Voss, Cordula Löffler, Karl-Ludwig Herné (Hrsg.): Neue Beiträge zu Rechtschreibtheorie und –didaktik.Freiburg: Fillibach, 2004. S. 65-84.

Möglichkeiten und Grenzen der automatischen Analyse orthographischer Leistungen

Tobias Thelen

1 Einleitung

Eine Analyse orthographischer Leistungen beinhaltet nicht nur die Beurteilung „falsch“ oder „richtig“, sondern schließt Hypothesen über das Wie und Warum der Leistungen mit ein. Im Folgenden soll untersucht und an zwei Beispielen verdeutlicht werden, inwiefern sich solche Analysen automatisieren, d.h. der algorithmischen Verarbeitung durch ein Computerprogramm zugänglich machen lassen.

Als Ausgangspunkt ist festzuhalten, dass es ein vollständiges algorithmisches Verfahren für die Analyse orthographischer Leistungen nicht geben kann. Wenn es in allen Fällen das Zustandekommen orthographischer Leistungen korrekt beschreiben könnte, müsste das Verfahren über das gleiche Wissen und die gleichen kognitiven Fähigkeiten wie ein Mensch verfügen. Um eine Erklärung abzugeben wie *„<Hefte> wurde als *<Y> verschriftet, weil die Schreiberin die abgebildeten Hefte für Bücher gehalten hat und anschließend mithilfe einer Anlauttabelle den Buchstaben für den Anlaut gesucht hat. Sie fand etwas, dass sie für einen Büffel hielt, das aber eigentlich ein Yak darstellen sollte und schrieb deshalb *<Y>.“*[1] ist sehr viel Alltagswissen, Zugang zu weiteren Beobachtungen und Kenntnis der näheren Umstände des Schreibvorgangs notwendig. Aber auch ein Mensch kann keine letztlich sicheren Aussagen über die Hintergründe einer Schreibung treffen, da z.B. Kompetenz- und Performanzfehler nicht immer unterscheidbar sind, emotionale Ursachen wie Unlust, Trauer oder Wut eine Rolle spielen, Fehler abgeschrieben sein können etc.

Daraus folgt, dass die hier betrachteten Analysen immer nur auf einen Ausschnitt der möglichen Erklärungen, auf einen genau definierten Rahmen bezogen sind. Für die algorithmische Beherrschbarkeit muss dieser Rahmen formal spezifizierbar sein. Aus dem Methodeninventar der Informatik, Computerlinguistik und Künstlichen Intelligenz bieten sich mindestens drei Herangehensweisen an:

1. Die Informatik bietet zahlreiche Verfahren zum Vergleich von einfachen, linguistisch uninterpretierten Strukturen. Mithilfe von Alignment- oder String-Matching-Algorithmen werden Unterschiede zwischen Zeichenketten auf Ketten von Operationen wie „Kopieren“, „Auslassen“ und „Einfügen“ abgebildet. Diese Verfahren liefern nur Erklärungen auf einer oberflächennahen Ebene, sind dafür aber effizient und gut erforscht. (vgl. Stephan 1994).
2. Aus der Computerlinguistik stammen eine große Menge von Formalismen zur Beschreibung und Analyse von sprachlichen Formen. Besonders hervorzuheben sind hier formale Grammatiken, die durch die Verwendung in Generatoren und Parsern sowohl für die Erzeugung als auch die Analyse geeignet sind (vgl. Carstensen et al. 2001). Solche Grammatiken sind nicht auf einzelne Ebenen beschränkt, so dass auch Abbildungen z.B. zwischen phonologischer und orthographischer Ebene modelliert werden können (s. Sproat 2000, Thelen/Gust 2002).
3. Die Künstliche Intelligenz hat Verfahren des Maschinellen Lernen entwickelt (vgl. Mitchell 1997), die genutzt werden können, um aus gegebenen Analysebeispielen

[1] Dieses Beispiel verdanke ich einem mündlichen Bericht Swantje Weinholds.

implizite oder explizite Regeln zu extrahieren. In diesen Fällen ist es nicht notwendig, die zur Analyse genutzte Regelmenge manuell zu konstruieren. Diese Verfahren setzen aber trotzdem ein genaues Verständnis der Anwendungsdomäne voraus. Sie bieten keine vollständig sicheren Ergebnisse und sind oft nur schwer zu steuern.

Im Folgenden werden Verfahren aus allen drei Bereichen genutzt, um unterschiedliche Arten orthographischer Leistungen automatisch analysierbar zu machen.

2 Szenarien

Notwendige Voraussetzung für eine automatische Analyse orthographischer Leistungen ist es, dass die zu analysierenden Schreibungen in maschinenlesbarer Form vorliegen. Dies ist im Wesentlichen in zwei Szenarien der Fall: Die Schreibungen entstehen entweder direkt am Rechner, z.B. im Rahmen freier Textproduktion mit einem Textverarbeitungssystem oder durch die stärker übungsorientierte Arbeit mit Rechtschreiblehr- und -lernsoftware. Oder die Schreibungen werden nachträglich am Rechner erfasst, z.B. um eine maschinelle Analyse vornehmen zu lassen. Die beiden Szenarien ermöglichen unterschiedliche Zielrichtungen bei der automatisierten Analyse, die sich grundlegend auch auf das Design und die Optimierung der Algorithmen niederschlagen können.

2.1 Nutzung als Grundlage „intelligenter“ Rechtschreiblehr- und –lernsoftware

Auf dem Markt existieren sehr viele als „Lehr-/Lernsoftware“ deklarierte Programme für den Schrifterwerb. Diese Programme stehen isoliert nebeneinander, sind nicht miteinander kombinierbar. Deshalb heißt Einsatz von Lehr-/Lernsoftware immer: Fertige Produkte auswählen und nebeneinander benutzen. Nach einer kurzen Charakterisierung verschiedener Grundtypen von Lehr-/Lernsoftware wird ein Szenario entworfen, in dem die Trennung und Nicht-Kombinierbarkeit nicht weiter bestehen muss.

Die einfachste und daher auch größte Klasse verfügbarer Lehr-/Lernsoftware lässt sich als „behavioristisch“ beschreiben[2]. Die Software präsentiert Fakten und fragt Wissen auf einfache Weise, etwa durch Multiple-Choice-Wahl, ab. Im Fall von Software für den Schrifterwerb können das Auswahlübungen sein, bei denen die korrekte Schreibung identifiziert, ein korrekter oder falscher Buchstabe markiert oder eine Zergliederung eines Wortes in Silben, die Kategorisierung von Vokalen als Lang- oder Kurzvokal etc. vorgenommen wird. Die Auswertung der auf diese Weise erbrachten Leistungen kann auf einfache Weise erfolgen, da alle Kombinationsmöglichkeiten bereits bei der Erstellung des Programms vollständig berücksichtigt werden können. Auf diese Weise können einfache Schleifen, Wiederholungen und Verzweigungen in den Programmablauf eingebaut werden. Damit folgt derartige Software den Prinzipen des Programmierten Unterrichts und zielt auf den Aufbau und die Verfestigung von Faktenwissen ab.

Die zweite Klasse bietet dem Lerner mehr Freiheitsgrade. „Kognitivistische“ Lehr-/Lernsoftware präsentiert Regeln und Zusammenhänge als Modellierungen oder Simulationen. Sie stellt dem Lerner damit eine Menge von Werkzeugen zur Verfügung, mit denen er mehr oder weniger authentisches Übungsmaterial bearbeiten kann. Für den Schrifterwerb sind hier verschiedenste Werkzeuge wie „Wortbaukästen“, das Häusermodell[3]

---

[2] Die Darstellung der drei Typen von Software orientiert sich an Kerres (2001).

[3] Wie z.B. in der Software MoPs (Thelen 2001a) eingesetzt. Zwar enthält auch MoPs zu einem großen Teil behaviouristische Bestandteile, das zentrale Werkzeug „Haus“ und die damit möglichen Operationen lassen sich aber als kongitivistisch bezeichnen.

oder „Maschinen“, die bestimmte Aspekte von Schreibungen überprüfen können, denkbar. In Kontext solcher Software sind freie Eingaben einzelner Wörter, aber auch ganzer Sätze denkbar. Als Folge dieser größeren Freiheit muss die Software in der Lage sein, Leistungen und Fehler in weit größerem und freierem Rahmen zu analysieren, als dies für behavioristische Lehr-/Lernsoftware notwendig ist. Kognitivistische Lehr-/Lernsoftware folgt der Tradition der „Intelligenten tutoriellen Systeme“ (für einen Überlick s. Peylo 2001) und verfolgt als Ziel den Aufbau und die Überprüfung von Regelwissen.

Die letzte Gruppe bildet „konstruktivistische“ Lehr-/Lernsoftware. Sie ist am ehesten als nichtlineare Materialsammlung zu verstehen, die Kommunikation über das Gelernte erlaubt und dem Lerner große Freiheit bei der Auswahl zu bearbeitender Inhalte, das Tempo der Bearbeitung und der Anpassung der Software an eigene Vorlieben oder besondere Fähigkeiten lässt. Zum Teil ist es wünschenswert, diese Anpassungen von der Software selbst vornehmen zu lassen, dabei soll der Lerner aber jederzeit die volle Kontrolle über seine Arbeitsumgebung behalten. Im vorliegenden Kontext ist eine Umgebung vorstellbar, in der freie Texte verfasst und an andere verschickt werden, in der z.B. Lexika auf verschiedene Weise angebunden und durchsuchbar sind. Konstruktivistische Lehr-/Lernsoftware beruht auf Arbeiten zum CSCL (Computer Supported Collaborative Learning) und zu Adaptiven Systemen. Ihr Ziel ist die Vermittlung von Handlungskompetenz.

Die Reihenfolge und der beschriebene unterschiedliche Komplexitätsgrad soll keine Wertung der drei Typen beinhalten. Nicht in allen Fällen ist eine konstruktivistisch motivierte Lernumgebung optimal, nicht in allen Fällen sind behavioristisch aufgebaute Übungen sinnlos. Damit ist nun die Frage möglich: „Wie sieht die ideale elektronische Lernumgebung für den Schrifterwerb aus?“ Eine beispielhafte Antwort darauf könnte eine Umgebung beschreiben, in der es möglich ist, freie Texte zu schreiben oder auf angebotene Schreibanlässe zurückzugreifen. Die Umgebung würde Werkzeuge zum Analysieren von Wörtern und Sätzen sowie zum Konstruieren von Schreibungen anbieten, aber auch selbstständig tätig werden und dem Lerner an bestimmten Stellen Hinweise auf mögliche Problemstellen geben oder Übungen anbieten. Gleichzeitig sollte eine solche Umgebung offen sein, d.h. die Kombination unterschiedlicher Programmteile, Modellierungen und Wissensquellen erlauben.

2.2 Nutzung als diagnostisches Werkzeug

Die Einsatz von Analysealgorithmen in Lehr-/Lernsoftware ist sehr anspruchsvoll, da die Auswirkungen direkt und ungefiltert in die Interaktion des Lerners mit der Software einfließen und der Lerner selbst nicht in der Lage ist, die Validität der Analysen zu überprüfen. Etwas geringere Anforderungen an die Verlässlichkeit und Ausgereiftheit der Verfahren ist zu stellen, wenn die „Nutzer“ der Analyse Menschen sind, die ihrerseits selbst eine Analyse von Schreibungen Dritter vornehmen wollen. Dies sind z.B. Lehrkräfte, die für einzelne Schüler oder die gesamte Klasse den Stand der Rechtschreibfähigkeiten beurteilen, oder Wissenschaftler, die Korpora von Schreibungen auf spezifische Fragestellungen hin untersuchen.

Für die erste Gruppe liegen eine Reihe von standardisierten, in der Regel nicht automatisierten Analyseverfahren vor, wie die Hamburger Schreibprobe (May 2002) oder die Diagnostischen Rechtschreibtests (s. z.B. Grund/Haug/Naumann 2003). Diese Verfahren erlauben es, standardisiertes Wortmaterial nach jeweils eigenen Kategorien zu analysieren und daraus einen Wert zu gewinnen, der mit anderen Schreibern, die sich dem gleichen Test an anderen Orten und zu anderen Zeitpunkten unterzogen haben, zu vergleichen. Gleichzeitig sind die Tests so optimiert, dass sie manuell möglichst einfach und schnell auswertbar sind. In manchen Situationen ist es aber auch wünschenswert, freie Schreibungen zu analysieren. Das Ergebnis ist dann kein vergleichbares Testresultat, sondern eine Menge von gut beherrschten

und problematischen Bereichen der Orthographie. Für die Kategorien der Hamburger Schreibprobe existiert ein dezidierter Vorschlag, das Analyseschema auch auf freie Schreibungen anzuwenden (May 1999).

Die Vorteile einer (teil-)automatisierten Analyse von Schreibungen sind vor allem:

- Konsistenz: Versehen bei der Auswertung sowie unterschiedliche Interpretationen von Analysekategorien, unterschiedliche Beurteilungen von Zweifelsfällen etc. entfallen, da das Verfahren deterministisch ist und das Ergebnis nur von den eingegebenen Daten abhängt.
- Wiederholbarkeit: Aufwand fällt nur bei der Erfassung der Daten an, danach kann das Verfahren beliebig oft auf die Gesamtdaten oder verschiedene Ausschnitte der Daten angewendet werden. Das ermöglicht z.B. eine ständige Erweiterung der Daten, die dann jederzeit nach unterschiedlichen Fragestellungen ausgewertet werden können.
- Austauschbarkeit: Verschiedene Varianten des Verfahrens, z.B. solche, die unterschiedliche Hypothesen über die deutsche Orthographie implementieren, können parallel auf die Daten angesetzt werden. Damit ist es ohne Mehraufwand möglich, Daten aus verschiedenen Perspektiven und unter verschiedenen Annahmen auszuwerten.

Der größte Nachtteil der automatischen Analyse ist, dass bei großen Datenmengen den Ergebnissen zu leicht „blind“ vertraut wird. So können z.B. kleine Missverständnisse der Definitionen oder Fehler in den Algorithmen unüberschaubare Konsequenzen haben.

3 Wissenskommunikation

Die notwendige Grundlage für solche Software ist Wissenskommunikation. Die Bestandteile oder Module des Programms bzw. der miteinander zu kombinierenden Programme müssen in der Lage sein, sich über analysierte Leistungen, über vermittelte Regeln oder übergeprüftes Wissen zu verständigen. Dazu bedarf es einer gemeinsamen Kommunikationssprache, die genau genug ist, um die geforderten Aspekte zu berücksichtigen, aber auch allgemein genug, um verschiedene Theorien über die deutsche Orthographie und verschiedene didaktische Möglichkeiten abzudecken. In der Forschung zur Künstlichen Intelligenz werden für solche Zwecke Ontologie bzw. Ontologien eingesetzt, die wie folgt definiert werden können:

Definition (Ontology): The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the type of things that are assumed to exist in the domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D . (Sowa 2000)

Übertragen auf die vorliegende Domäne „orthographische Leistungen“ wären folgende Anforderungen an solch eine Ontologie zu stellen:

- Die in Erklärungen orthographischer Leistungen vorkommenden Begriffe müssen enthalten sein (unter der oben getroffenen Einschränkung, dass nur ein Ausschnitt der Menge der denkbaren Erklärungen abgedeckt werden soll).
- Diese Begriffe müssen als „Typen von Dingen“, d.h. Einzelfall übergreifenden orthographischen Kategorien oder Phänomene verstanden werden können.
- Die Existenz oder mögliche Existenz der Phänomene sollte durch sinnvolle Definitionen sichergestellt sein. Die Ontologie definiert die Begriffe nicht vollständig, sondern stellt nur Beziehungen zwischen ihnen dar.

Im Folgenden wird eine spezifische Einschränkung des Erklärungsvokabulars vorgenommen. Betrachtet werden nur Erklärungen, die sich im weitesten Sinne auf (linguistische) „Regeln“

beziehen. Eine orthographische Leistung lässt sich damit als Menge befolgter und nicht befolgter Regeln beschreiben. Diese Einschränkung ist kleiner, als sie zunächst erscheinen mag: Unter den Begriff „Regel“ lassen sich von sehr allgemeinen Aussagen wie „Am Satzanfang wird groß geschrieben“ bis hin zu speziellen Aussagen wie „Das Wort /taU.b@/ wird <Taube>[4] geschrieben“ weite Bereiche abdecken. Aus den beiden Beispiele wird aber auch schon deutlich, dass es mehrere alternative Erklärungen einer Leistung geben kann: Der Schreiber hat <Taube> geschrieben, weil er wusste, dass das Wort /taU.b@/ als <Taube> geschrieben wird. Aber auch: Der Schreiber hat <Taube> geschrieben, weil er die Regeln befolgt hat, dass der Anfangsrand /b/ in unbetonter Silbe als <b> und der Silbenreim /@/ als <e> verschriftet wird, usw. In einer Ontologie können also einander überlappende oder gar widersprechende Regeln enthalten sein. Erst damit ist es möglich, den angestrebten Zweck – die Kommunikation über orthographische Leistungen über Programmteile und Programme hinweg – zu erreichen.

Abb. 1 zeigt einen Ausschnitt einer möglichen Ontologie für die Beschreibung orthographischer Leistungen. Der Ausschnitt beschränkt sich auf phonographische Phänomene und geht von drei parallelen Zweigen aus: Einzelwortschreibungen sind demnach silbisch (silbenpositionsbezogene Laut-Buchstaben-Zuordnung und prosodische Markierungen), morphologisch (Konstantschreibung, Affixschreibungen) und lexikalisch (direkt angegebene Schreibung einzelner Wörter) begründbar. Für eine detaillierte Begründung dieser Sicht s. Thelen (2001b).

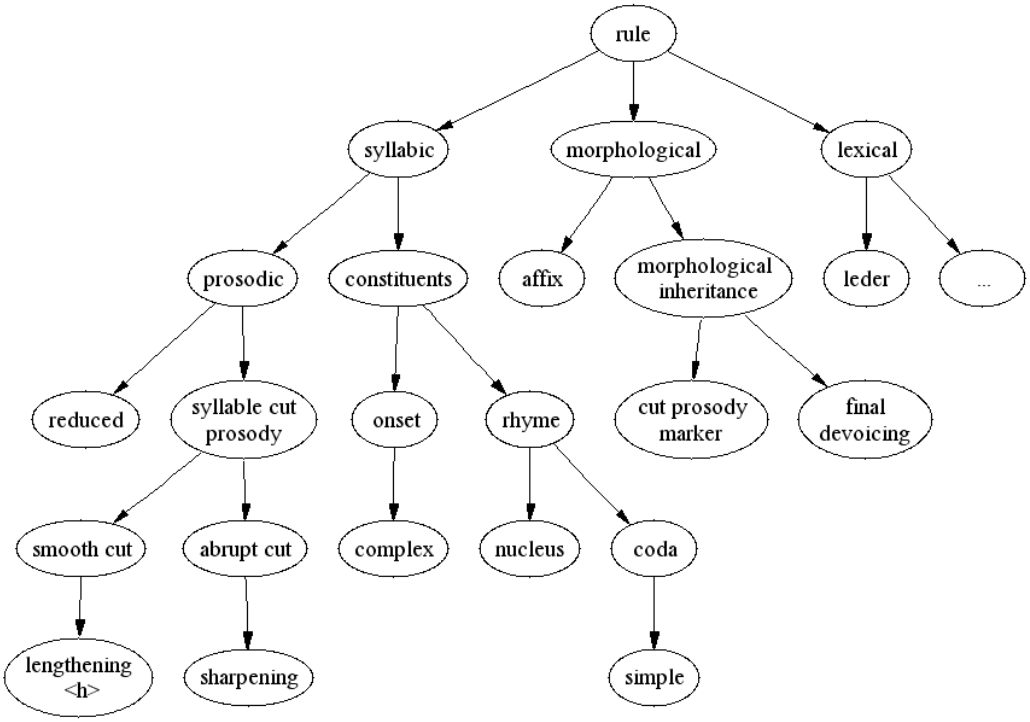

Abbildung 1: Ausschnitt einer Ontologie zur Beschreibung orthographischer Phänomene

[4] Diese Regel erfüllt dennoch die Forderung nach Einzelfall übergreifender Gültigkeit. Ein Einzelfall wäre eine Regel der Form „Das Wort W wird von Person P zu Zeitpunkt Z an Ort O als <W’> geschrieben.“

Vorabversion. Erschienen in: Albert Bremerich-Voss, Cordula Löffler, Karl-Ludwig Herné (Hrsg.): Neue Beiträge zu Rechtschreibtheorie und –didaktik.Freiburg: Fillibach, 2004. S. 65-84.

4 Beispiele

4.1 Beispiel 1: Analyse von Einzelwortschreibungen

In den oben entworfenen Szenarien sollen freie Texte automatisiert analysiert werden. Dieses Problem lässt sich in zwei Teilprobleme aufgliedern, die jeweils unterschiedliche Lösungsansätze bedingen:

1. Erkennung des „eigentlich gemeinten“ Textes. Hierunter sind Verfahren zu verstehen, die den geschriebenen Text in eine Repräsentation überführen, aus der sich der orthographisch korrekte Text sowie weitere Informationen über die einzelnen Bestandteile ableiten lassen.
2. Analyse der orthographischen Leistungen durch Vergleich von korrektem und tatsächlich geschriebenen Text. In diesem zweiten Schritt werden dann die vorliegenden Schreibungen mit den korrekten Schreibungen bzw. abgeleiteten Repräsentationen verglichen.

Für den ersten Fall wurden eine Reihe von Lösungsansätzen entwickelt, die z.B. in Textverarbeitungssystemen Anwendung finden. Hier besteht die Aufgabe nicht in einer erklärenden Analyse der Schreibungen, sondern in der Identifikation inkorrekter Schreibungen und der Generierung von Korrekturvorschlägen. In aller Regel vergleichen die verwendeten Algorithmen die Wörter des Textes mit einer gespeicherten Liste bekannter Wörter. Ist ein Wort nicht gespeichert, wird es als potenziell fehlerhaft markiert. Aus den gespeicherten Wörtern werden dann die „ähnlichsten“ herausgesucht und dem Benutzer als Korrekturvorschläge präsentiert (für Details der Verfahren s. Mitton 1996). Diese Verfahren schlagen allerdings fehl, wenn ein Wort auf solche Weise falsch geschrieben ist, dass es mit einem (anderen) gespeicherten Wort übereinstimmt. Das können Verwechslungen von Wörtern sein: „Ich *wahr in Hamburg.“, aber auch zufällige Übereinstimmungen mit gespeicherten Wörtern sein: „Ich war im Walt.“[5] Zur Lösung solcher Probleme, sowie zur Erkennung von nicht wortbezogenen Fehlern wie Groß- und Kleinschreibungs- oder Interpunktionsfehlern sind Verfahren notwendig, die die morphologische und syntaktische Struktur der Texte analysieren. Dies ist nach dem derzeitigen Stand der Technik zwar in weiten Teilen – auch mit effizienten Algorithmen – möglich, allerdings besteht noch erheblicher Forschungsbedarf.

Im Folgenden wird dieser erste Problembereich ausgeklammert und für alle Verfahren angenommen, dass Texte vorliegen, bei denen jeweils bekannt ist, welches Wort bzw. welche andere sprachliche Einheit geschrieben werden sollte. Für das Szenario „Diagnosesoftware“ lässt sich diese Situation auch häufig manuell herstellen, etwa durch Annotation der Korpora oder Nachfragen der Software beim Benutzer.

Für die Analyse von Einzelwortschreibungen sind mehrere Komplexitätsstufen der verwendeten Verfahren denkbar, die von der Art und Menge der zur Verfügung stehenden Informationen abhängen.

Im einfachsten Fall werden die beiden Zeichenketten der korrekten und der intendierten Schreibung verglichen und Abweichungen auf Regeln aus der Ontologie abgebildet. Durch Zusammenfassung von Buchstaben zu komplexen Graphemen[6] und einigen Heuristiken sind

---

[5] „Walt“ wie in „Walt Disney“. Wenn die gespeicherte Liste existierender Wörter zu groß wird, kommen solche Probleme häufig vor.

[6] Hier wird ein sehr weiter Graphembegriff zu Grunde gelegt, der Einheiten zusammenfasst, die als ganzes für bestimmte Phänomene stehen bzw. mit anderen Einheiten verwechselt werden können.

so bereits relativ weit gehende Erklärungen möglich. Für die Schreibung *<Hüte> für <Hütte> kann mit einem Levenshtein-Algorithmus[7] berechnet werden, dass anstatt <tt> ein <t> geschrieben wurde. Zusammen mit der Kontextinformation „steht vor <e>“ ist z.B. ein Verstoß gegen die Regel „silbisch -> prosodisch -> Silbenschnitt -> scharfer Schnitt -> Schärfung“[8] analysierbar. Durch eine zusätzliche, ebenfalls oberflächenbasierte Analyse der korrekten Form kann zudem eine Liste aller für die korrekte Schreibung geltenden Regeln erzeugt werden. Auf diese Weise ist das Ergebnis der Analyse eine Liste aller befolgten und aller nicht befolgten Regeln, wobei auch hier wieder gilt, dass sich diese Angaben überlappen oder auch widersprechen können.

Ein Problem dieses ersten, sehr einfachen Verfahrens ist, dass es heuristische Annahmen enthält, die nicht korrekt sein müssen. Dies betrifft z.B. die Betonungsverhältnisse in einem Wort, die Vokalquantitäten, aber auch bereits die Zusammenfassung von Buchstaben zu komplexen Graphemen. So enthält z.B. das Wort <annehmen> kein durch Schärfungsmarkierung entstandenes Graphem <nn>, das Wort <Kamel> ist nicht trochäisch betont und ein Wörtern wie <rasten> kann je nach morphosyntaktischem Zusammenhang einen Lang- oder Kurzvokal enthalten. Deshalb setzt ein zweites Verfahren auf zusätzliche phonologische Informationen, die entweder über ein ebenfalls automatisiertes Präanalyse-Verfahren (vgl. Thelen, Gust 2002), manuelle Annotation oder Lexikonzugriff gewonnen werden. Das Verfahren kann auf das vorhergehende aufsetzen, indem die graphemische Struktur des korrekten Wortes mit phonologischen Informationen angereichert wird. Ein Beispiel verdeutlicht dieses Vorgehen:

H - /h/, Anfangsrand betonte Silbe

Ü - /Y/, Nukleus prominente Silbe, Kurzvokal

TT - /t/, fester Anschluss, Endrand prominente Silbe, Anfangsrand Reduktionssilbe[9]

E - /@/, Nukleus Reduktionssilbe

Für den Einsatz als Diagnoseinstrumentarium für Lehrkräfte wurde ein tabellarisches Auswertungsschema erstellt, das die Ontologie aufgreift, aber nur Ausschnitte abbildet und Alternativen z.T. zusammenfasst. Zudem wurden für die Schreibung von Silbenkonstituenten vier Stufen der Regelübereinstimmung definiert:

1. Orthographisch korrekt – die vorgenommene Schreibung für das betrachtete Phänomen ist korrekt.
2. Phonetisch plausibel – die Schreibung für das betrachtete Phänomen ist nicht korrekt, aber die Abweichung ist phonetisch erklärbar.
3. Repräsentiert – die Schreibung ist weder korrekt noch phonetisch plausibel, es ist aber eine graphische Repräsentation der Konstituente vorhanden.
4. Nicht repräsentiert – die Konstituente ist graphisch überhaupt nicht repräsentiert.

Abb. 2 zeigt einen Teil der gesamten Tabelle zusammen mit der Gesamtanalyse eines vollständigen Textes aus dem Osnabrücker Bildergeschichtenkorpus (Thelen 2000).

Ein armer Hund

---

[7] Auch bekannt als „Minimal Edit Distance“, vgl. Stephan (1997:40ff.).

[8] Eine Regel wird immer durch die vollständige Angabe ihres Pfades in der Ontologie gekennzeichnet. Im Folgenden werden abkürzende und sprechendere Namen gewählt.

[9] Die Frage, ob der Endrand der betonten Silbe als „leer“, oder ein ambisyllabischer Konsonant angenommen wird, ist hier nicht von Belang.

Vorabversion. Erschienen in: Albert Bremerich-Voss, Cordula Löffler, Karl-Ludwig Herné (Hrsg.): Neue Beiträge zu Rechtschreibtheorie und –didaktik.Freiburg: Fillibach, 2004. S. 65-84.

Herr Jakob sas im Haus und kukt zum Fenstar raus . Da sit er ein Hund . Herr Jakob dengt ! *Was wil der Hund hirr ?* Dann gett er raus . Und sid (,) das er mit seiner Fußmate dafonrent . Dann folgt Herr Jakob den Spuren . Und dann siet Herr Jakob (,) des der Hund vrirt und dann denkt Herr Jakob *ich llas im die matte da (.) Ich kauf mir eine neue* .[10]

Phänomen	**max**[11]	**repräsentiert**	**Phonetisch plausibel**	**Orthographisch korrekt**
Silbenkerne	89	88 (98%)	86 (96%)	79 (88%)
S'	76	76 (100%)	74 (97%)	68 (89%)
S' (fester Anschluss)	36	36 (100%)	35 (97%)	35 (97%)
S' (loser Anschluss)	40	40 (100%)	39 (97%)	33 (82%)
Monophthong	22	22 (100%)	22 (100%)	16 (72%)
Diphthong	9	9 (100%)	9 (100%)	9 (100%)
Fallender Diphth.	9	9 (100%)	8 (88%)	8 (88%)
S°	8	7 (87%)	7 (87%)	6 (75%)
Schwa	4	4 (100%)	4 (100%)	4 (100%)
Vokalis. R	3	3 (100%)	3 (100%)	2 (66%)
Silbischer Sonorant	1	0 (0%)	0 (0%)	0 (0%)
S	5	5 (100%)	5 (100%)	5 (100%)
Anfangsränder	72	71 (98%)	70 (97%)	67 (93%)
Einfach	68	67 (98%)	66 (97%)	64 (94%)
Komplex	4	4 (100%)	4 (100%)	3 (75%)
Endränder	73	72 (98%)	70 (95%)	63 (86%)
Einfach	40	40 (100%)	38 (95%)	33 (82%)
Komplex	33	32 (96%)	32 (96%)	30 (90%)
Phonologische Markierungen	2			1 (50%)
Schärfung	2			1 (50%)
Konstantschreibung	22			15 (68%)
Auslautverhärtung	5			5 (100%)
Schärfung	13			10 (76%)
Silbentrennendes <h>	4			0 (0%)
Sonstiges				
Einfügungen	1			
Überfl. Schärfung	3			

Sowohl sehr hohe Werte (nahe an 100%), als auch nach unten abweichende können hohe Aussagekraft haben. Für den vorliegenden Text lassen sich aus der Tabelle folgende Schlussfolgerungen ableiten:

- Die Schreibung der Silbenränder wird gut beherrscht.
- Es gibt Probleme mit Silbenkernen, insbesondere Monophthongen mit losem Anschluss. Eine detailliertere Analyse zeigt, dass das Problem hauptsächlich auf falsche <ie>-Schreibungen zurückzuführen ist.

[10] Kursiv gesetzte Teile kennzeichnen (angenommene) direkte Rede, Satzzeichen in Klammern sind ausgelassen.

[11] Die Spalte „max“ enthält die Angabe, wie häufig das jeweilige Phänomen im analysierten Matrial vorkommt, also die maximale Anzahl korrekter Schreibungen

- Der gesamte Bereich „Schärfung“ ist noch problematisch, allerdings sind die Fallzahlen recht gering, so dass weiteres Material mit Schärfungsmarkierungen analysiert werden müsste.
- Das silbentrennende <h> ist nicht vorhanden; trotz geringer Fallzahlen ist diese Analyse relativ deutlich.
- Es liegen Einzelfehler bei vokalisiertem /r/ und silbischen Sonoranten vor.

Solche Zusammenfassungen der Analysen können auch automatisiert erfolgen, dennoch bleibt die Notwendigkeit einer Interpretation der Ergebnisse.

Es sind weitere Stufen der Anreicherung mit Informationen denkbar, die noch bestehende Unsicherheiten und Heuristiken in der Analyse ausräumen. Zum einen sind dies phonetische Verfahren, die Fast-Speech-Phänomene oder dialektale Phänomene mit in den Blick nehmen. Diese Phänomene sowie ihre Auswirkungen auf geschriebene Sprache sind gut beschrieben (vgl. Kohler 1995:201ff., Naumann 1989) und in großen Teilen gut formalisierbar. Auf diese Weise können unter Zugrundelegung zusätzlicher Annahmen, z.B. über dialektale Einflüsse, genauere Analysen vorgenommen werden. Vergleichsmaßstab ist dann nicht mehr die mit der Explizitlautung annotierte Schreibung, sondern evtl. verändert hergeleitete Schreibungen mit entsprechenden Markierungen. Die phonetischen Prozesse, die für die Analyse angenommen wurden, schlagen sich allerdings nicht in der Ontologie nieder, sondern sind zusätzliche Anmerkungen zu den Analyseergebnissen.

In den bislang skizzierten Verfahren wurden morphologische Informationen, wie z.B. die Klassifikation einer (korrekten) Schreibung wie <Wald> als Fall von Auslautverhärtung, nur heuristisch angenommen. Demnach wäre die Form <und> ebenfalls als Auslautverhärtungsfall betrachtet und folglich eine Schreibung wie *<unt> als Verstoß gegen die Auslautverhärtungsregel analysiert worden. Erst durch eine leistungsfähige morphologische Zerlegung sowie darauf basierenden Lexikonzugriff sind solche Informationen mit größerer Sicherheit herleitbar.

4.2 Beispiel 2: Analyse von Groß- und Kleinschreibleistungen

Neben den phonographischen Phänomen der deutschen Orthographie stehen solche, die nicht anhand isolierter einzelner Wortformen entscheidbar sind. Insbesondere sind dies die Interpunktion, die Getrennt- und Zusammenschreibung und die Groß- und Kleinschreibung. Für die automatisierte Analyse solcher Phänomene ist eine Gemeinsamkeit wichtig: Für alle drei Phänomenbereiche kann der Text in endlich viele gleichartige Einheiten unterteilt werden, deren Binnenstruktur für die Analyse unerheblich ist. Nur an den Grenzen zwischen diesen Einheiten treten die Phänomene auf und sind jeweils als binär betrachtbar: Komma oder kein Komma, Zwischenraum oder kein Zwischenraum, Groß- oder Kleinbuchstabe. Diese Ebenen sind im Einzelnen in Tabelle 1 aufgeführt.

Phänomenbereich	**Einheiten**	**Zu klärende Frage**
Getrennt- und Zusammenschreibung	Grapheme	Zwischenraum?
Groß- und Kleinschreibung	Orthographische Wörter[12]	Groß- oder Kleinbuchstabe?
Interpunktion	Orthographische Wörter	Komma? Punkt? …

Tabelle 1:Wortübergreifende Phänomenbereiche und ihre kleinsten Einheiten

[12] Orthographische Wörter sind Folgen von Graphemen, die nicht durch Wortzwischenräume unterbrochen werden.

Ein Algorithmus zur Analyse von Schreibleistungen kann auf der jeweiligen Ebene dann grundsätzlich wie folgt vorgehen:

```
1. Zerlege den Text in die relevanten Einheiten

2. Besorge relevante Zusatzinformationen für die Einheiten

3. Betrachte alle Grenzen zwischen den Einheiten:

      Wenn das betrachtete Phänomen gesetzt ist:

            Sammle alle Merkmale der benachbarten Einheiten
            in einer „Positivliste“

      Wenn das betrachtete Phänomen nicht gesetzt ist:

            Sammle alle Merkmale der benachbarten Einheiten
            in einer „Negativliste“
```

Es können damit nicht nur lokale Kriterien (links von X, rechts von X) abgebildet werden, sondern auch globalere. Dazu werden die Merkmale von Gruppen von Einheiten auf alle einzelnen Mitglieder dieser Gruppe projiziert und mitnotiert. Alle Elemente, die Teil einer Nominalphrase sind, bekommen so die Markierung „Teil einer Nominalphrase“. Sollte es als sinnvoll erachtet werden, z.B. Adjazenzbeziehungen zwischen solchen übergeordneten Einheiten mit zu berücksichtigen, könnte eine zusätzliche Markierung „Teil einer Nominalphrase, die nach einer Verbalphrase steht“ eingefügt werden.

Für die Groß- und Kleinschreibung wurde für das Projekt „Entwicklung eines linguistisch orientierten Rechtschreibkonzepts für alemannisch sprechende HauptschülerInnen“[13] ein Analyseverfahren entwickelt, das vor allem die manuelle Auswertung eines Diktattextes erleichtern sollte.

Die linguistische Analyse der Groß- und Kleinschreibung unterscheidet sich von der „naiven“ Begründung z.B. des Amtlichen Regelwerks. Dort heißt es grundsätzlich: „§ 55: Substantive schreibt man groß. “, worauf eine komplexe Auflistung von Regeln, Ausnahmen und deren Ausnahmen folgt. Die Erklärung „Nomen und Nominalisierungen werden groß geschrieben“ ist außerordentlich verbreitet und findet sich auch in didaktischen Darstellungen. Die linguistisch sauberere Erklärung „Expandierbare Kerne von Nominalphrasen werden durch Großschreibung markiert“ (Maas 1992:156ff.) wird mit der Begründung abgelehnt, dass die verwendeten Begriffe wie „expandierbar“, „Kern“ und „Nominalphrase“ für eine unterrichtliche Vermittlung nicht geeignet seien (s. aber Röber-Siekmeyer 1999 für eine Gegenposition). Die übliche Vermittlung der Groß- und Kleinschreibung impliziert eine Schwierigkeitsabstufung für Großschreibungen:

1. Konkreta („Alles, was man anfassen kann.“)

2. Abstrakta („Auch andere Nomen.“)

3. Nominalisierungen („Nominalisierte Verben, Adjektive, …“)

Eine Analyse vorliegender Schreibungen kann nach „Strategien“ forschen, die die Schreibenden verfolgt haben könnten. Die Frage, die dann untersucht werden kann, ist die, ob erfolgreiche Rechtschreiber sich tatsächlich an diesen Regeln orientieren oder ob ihre „inneren Regeln“ ganz anderer Natur sind, und z.B. mit der linguistisch begründeten Regel zur Kennzeichnung von Kernen in Nominalphrasen kongruieren.

---

[13] Durchgeführt an der PH Freiburg unter Leitung von Prof. Dr. Christa Röber-Siekmeyer (s. Röber-Siekmeyer et al. 2002).

Eine erfolgreiche Analyse besteht darin, eine Strategie oder eine Menge von einander nicht widersprechender Strategien zu finden, die die beobachtete Verwendung der Groß- und Kleinschreibung möglichst optimal erklären. Eine perfekte Erklärung durch eine einzelne Strategie wird nur in den seltensten Fällen möglich sein, da die Schreibungen immer unterschiedlichen Störfaktoren unterliegen. Die Art der zu betrachtenden Strategien ist grundsätzlich nicht beschränkt. Es kann sich um einfache positive Aussagen wie „Konkreta werden groß geschrieben“, einfache negative Aussagen wie „Abstrakta werden klein geschrieben“, oder zusammengesetzte Aussagen wie „Konkreta und Wörter am Satzanfang werden groß geschrieben“ handeln. Im Folgenden wird unter einer „atomaren Strategie“ die Korrelation einer Eigenschaft der zu schreibenden Wörter mit der tatsächlich vorgenommenen Majuskelmarkierung verstanden. Dazu werden im Originaltext alle Wörter mit Feature-Wert-Paaren annotiert, die all die Eigenschaften enthalten, die in die Untersuchung einbezogen werden sollen. Damit ist das Vokabular vorgegeben, aus dem Strategieerklärungen gebildet werden. Grundsätzlich können hier beliebige Eigenschaften gewählt werden. Denkbar sind morphosyntaktische Informationen („lexikalische Wortart“), syntaktische Informationen wie die Zugehörigkeit und die Position bzw. Rolle innerhalb syntaktischer Konstituenten, die Position im Satz, der Kontext des Wortes, semantische Angaben wie „abstrakt“ oder „konkret“ oder Frequenzinformationen. Wichtig ist, dass die automatisiert erstellten Erklärungen nur aus der Menge dieser Eigenschaften generiert werden können. Werden in einem Text z.B. alle mit <a> beginnenden Wörter groß geschrieben und alle anderen klein (eine solche Strategie ist denkbar, wenn auch nicht sehr wahrscheinlich), dann ist diese Erklärung nur erreichbar, wenn die Eigenschaft "`Anfangsbuchstabe'" oder bei binären Features „Beginnt mit <a>“ annotiert wird. Im vorliegenden Fall werden ausschließlich binäre Features verwendet. Damit sind keine offenen Klassen von Eigenschaften ausdrückbar, wie die absolute Position im Satz, der Anfangsbuchstabe, die Wortlänge etc. Solche Eigenschaften müssten durch eine Reihe binärer Features ausgedrückt werden, wie das hier bei der „Lexikonwortart“ durchgeführt ist. Üblicherweise wird jedem Lexikoneintrag genau eine „Lexikonwortart“ (part of speech) zugeordnet, mehrfache Zuordnungen werden durch mehrfache Lexikoneinträge aufgelöst. Die Feature-Annotation ist hier eine Projektion des Lexikons auf den zu schreibenden Text, die all die Kriterien enthalten soll, die ein Schreiber möglicherweise annimmt. Diese Liste kann kombinatorisch aus den linguistisch möglichen Alternativen gewonnen werden, aber auch aus der Auswertung von Interviews und Beobachtungen stammen (s. Röber-Siekmeyer 1999). So kann ein Wort mehrere Lexikonwortarten zugewiesen bekommen, wenn anzunehmen ist, dass ein Schreiber unterschiedliche Möglichkeiten wählen könnte. Im Beispiel unten ist die Form <gewohnten> mit [+adj +v] annotiert, da Partizipialformen dieser Art häufig adjektivisch verwendet werden können und ein Schreiber sie auf die gleiche Weise wie z.B. „schönen“ betrachten könnte. Ein anderes Beispiel ist die immer wieder problematische Form „morgen“, die adverbial oder nominal verwendet werden kann. Für die Konzeption einer Feature-Annotation ist damit zu beachten, dass „zu viele“ Alternativen erst einmal nicht schaden, sondern vorrangig sichergestellt werden soll, dass möglichst alle plausiblen Erklärungsmöglichkeiten enthalten sind.

Für die hier beschriebenen Untersuchungen wurde ein modifizierter Text aus einem „Harry Potter“-Band verwendet. Tabelle 2 gibt den ersten Satz des Textes mit einigen Features wieder.[14] Die Annotation mit Features kann teilautomatisiert vorgenommen werden, es bleibt

[14] Es sind nur positive Merkmalsausprägungen angegeben, die erste Zeile bezeichnet die lexikalische Wortart, danach sind Informationen zum Enthaltensein X_in, Beginn X_start oder Ende X_end verschiedener Einheiten X (Satz s, Nominalphrase np) gegeben, das Feature det_+ bezeichnet Wörter, die direkt nach einem Artikel stehen.

aber ein gewisser manueller Aufwand, wenn nicht komplexe syntaktische computerlinguistische Analyseverfahren eingesetzt werden.

Harry	**Potter**	**war**	**erst**	**seit**	**wenigen**	**Tagen**	**in**	**der**	**Zauberschule**
+n	+n	+v	+adv	+adv	+quant	+n	+präp	+art	+n
+name	+name				+np_start	+in_np		+np_start	+s_end
+s_start	+np_end				+in_np	+np_head		+in_np	+np_end
+np_start	+in_np					+abstrakt			+in_np
+in_np	+np_head								+np_head
+np_head									+det+_

Tabelle 2: Beispiel für einen annotatierten Textausschnitt

Da nur herausgefunden werden soll, welche Strategie der Schreiber verfolgt hat, ist die tatsächliche Groß- und Kleinschreibung der Wörter im Originaltext irrelevant. Es werden lediglich die Features des Originaltextes auf den geschriebenen Text abgebildet und dann mit der tatsächlich vorgenommenen Majuskelmarkierung verglichen. Zusätzlich kann natürlich berechnet werden, wie „gut“ eine Strategie in Hinblick auf die korrekte Markierung ist. Ergebnis der Analyse ist dann für jede atomare Strategie ein Wert zwischen -1 und +1, der die Korrelation der Strategie mit der tatsächlich vorgenommenen Majuskelmarkierung darstellt. Zusätzlich spielt die relative Fallzahl eine Rolle, für die eine betrachtete Strategie im Text relevant ist. Die atomaren Strategien können dann sortiert und mit Verfahren aus dem Maschinellen Lernen, den so genannten Entscheidungsbaumverfahren zu der komplexen Strategie kombiniert werden, die die beobachteten Schreibungen am besten erklärt. Es bleiben Restunsicherheiten, da die Schreibungen Abweichungen unterliegen und die Strategien nicht zu 100% befolgt werden. Komplexe Interaktionen zwischen Strategien können auf die beschriebene Weise nicht erkannt werden, wohl aber Präferenzordnungen.

Ein konkretes Beispiel soll die Art der Analyseergebnisse verdeutlichen. Der folgende Text ist ein relevanter Ausschnitt aus dem analysierten Gesamttext, er wurde ohne jegliche orthographische Korrekturen übernommen, berücksichtigt wurde hier aber nur die Groß- und Kleinschreibung.

… Er raste über vile Treppen in dem verwinkligen Gebeude enge kurze krumme wacklige. Manche fürten ***Freitags*** *nicht zu dem* ***gewonnten.*** *Manche hatten auf halber* ***höhe*** *eine Stufe die ganz plötzlich verschwand und man durfte nicht vergessen dieses unvorhersehbare* ***nichts*** *zu überspringen. Es gab auch Türen die sich nur öfneten wenn man sie höflich bat oder an der* ***Richtigen*** *Stelle kitzelte. Es war auch schwierig sich daran zu erinnern wo etwas* ***bestimmtes*** *war den alles schien morgens ziemlich oft die angestammten Plätze zu wegseln. Harri Potter musst noch viel lernen um das* ***geheimnisvolle*** *zu erreichen …*

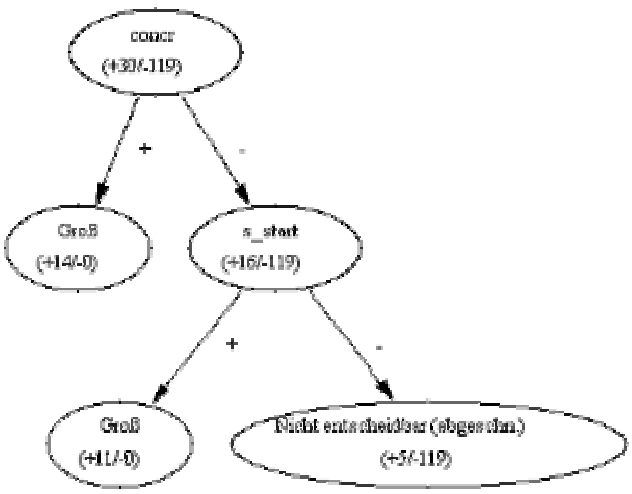

Abbildung 2: Entscheidungsbaum für einen analysierten Text

Abb. 2 zeigt nun eine automatisch gebildete komplexe Strategie in Form eines Entscheidungsbaumes. Demzufolge entscheidet sich der Schreiber zunächst anhand der Frage „Liegt ein Konkretum vor?“ für Großschreibung, wenn die Frage positiv (linker Zweig, +) beantwortet werden kann. Fällt die Antwort negativ aus, wird als nächstes die Frage gestellt: „Steht das Wort am Satzanfang?“ Falls ja, wird groß geschrieben, falls nein folgen weitere Fragen, die der Algorithmus hier abgeschnitten bzw. zusammengefasst hat, weil keine der noch folgenden Knoten relevant genug ist, d.h. eine Entscheidung trifft, die ausreichend hohe Fallzahlen berücksichtigt. Damit lässt sich als komplexe Strategie für diesen Schreiber ablesen: „Er schreibt groß, wenn ein Konkretum vorliegt oder ein Wort am Satzanfang steht. Die restlichen beobachteten Großschreibungen konnten nicht weiter systematisiert werden.“ Für andere Schreiber ergeben sich deutlich unterschiedliche Diagramme, so dass das Verfahren tatsächlich geeignet ist, Unterschiede in den Groß- und Kleinschreibstrategien einzelner Schreiber zu verdeutlichen. Die „Intelligenz“ bzw. das Wissen über sprachliche Strukturen liegt hier aber nicht im Verfahren selbst, sondern in der vorab zu leistenden Annotation des Textes.

5 Schlussfolgerungen

An zwei Beispielen wurde gezeigt, wie Leistungen in unterschiedlichen Phänomenbereichen der deutschen Orthographie mit unterschiedlichen Methoden automatisiert analysiert werden können. In beiden Fällen wurde deutlich, dass ein wichtiger Teil des Einsatzes in der sorgfältigen Aufbereitung und teilweise manuellen Annotation der zu analysierenden Texte besteht, sich dann aber große Vorteile durch Geschwindigkeit, Flexibilität, Modifizierbarkeit und Konsistenz der Analysen ergeben. Bezogen auf die beiden skizzierten Szenarien ergibt sich ein differenziertes Bild. Die dargestellten Algorithmen sind in der Lage, sowohl für den Einsatz in Lehr-/Lernsoftware als auch für diagnostische Zwecke eine wichtige Grundlage zu bilden. Allerdings sind für den vollautomatischen Einsatz in Lehr-/Lernsoftware entweder weitere komplexere Algorithmen zur Vorverarbeitung und –analyse der Texte notwendig, oder es müssen didaktische Einschränkungen in Kauf genommen werden, die nicht mehr beliebige Texte, sondern nur derart eingeschränkte zulassen, dass für die Software der „intendierte“ Text leicht identifizierbar ist. Damit eignen sich die Algorithmen für den Einsatz in „behavioristischer“ und „kognitivistischer“, nur bedingt aber in „konstruktivistischer“ Software und damit auch der skizzierten „idealen“ Lernumgebung für den Schrifterwerb. Eine Weiterentwicklung der Algorithmen in diese Richtung bzw. ihre Integration in größere Sprachanalyse-Frameworks bleibt eine wichtige Forschungsaufgabe. Diese Einschränkungen gelten nicht für den Einsatz zu diagnostischen Zwecken. Eine bessere und umfangreichere Aufarbeitung der Korpora verbessert die Analyseergebnisse, die zudem stets interpretiert und nicht unmittelbar umgesetzt werden. Daher können die dargestellten Algorithmen als wertvolle Hilfe bei der Analyse orthographischer Leistungen dienen.

Referenzen

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen: Computerlinguistik und Sprachtechnologie. Heidelberg, Berlin: Spektrum Akademischer Verlag.

Grund, Martin; Haug, Gerhard; Naumann, Carl-Ludwig (2003): Diagnostischer Rechtschreibtest für 5. Klassen. 2., aktualisierte Auflage. Göttingen: Beltz Deutsche Schultests.

Kerres, Michael (2001): Multimediale und telemediale Lernumgebungen. München: Oldenbourg.

Kohler, Klaus (1995): Einführung in die Phonetik des Deutschen. 2. Auflage. Berlin: Erich Schmidt Verlag.

Maas, Utz (1992): Grundzüge der deutschen Orthographie. Tübingen: Niemeyer.

May, Peter (1999): Strategiebezogene Rechtschreibdiagnose - mit und ohne Test: Analyse von freien Schreibungen mit Hilfe der HSP-Kategorien. In: Balhorn, Heiko; Bartnitzky, H., Büchner, I.; Speck-Hamdan, A. (Hrsg.): Schatzkiste Sprache I: Lesen und Schreiben von Anfang an. AKG-Band 103. Frankfurt a.M.: Arbeitskreis Grundschule.

May, Peter (2002): HSP 1-9. Diagnose orthografischer Kompetenz zur Erfassung der grundlegenden Rechschreibstrategien. 6. Auflage. Hamburg: Verlag für pädagogische Medien.

Mitchell, Tom (1997): Machine Learning. New York, St. Louis, San Francisco u.a.: McGraw-Hill.

Mitton, Roger (1996): English Spelling and the Computer. London (u.a.): Longman.

Naumann, Carl Ludwig (1989): Gesprochenes Deutsch und Orthographie : linguistische und didaktische Studien zur Rolle der gesprochenen Sprache in System und Erwerb der Rechtschreibung . Frankfurt am Main (u.a.): Lang.

Peylo, Christoph (2002): Wissen und Wissensvermittlung im Kontext von internetbasierten intelligenten Lehr- und Lernumgebungen. Berlin: Akademische Verlagsgesellschaft.

Röber-Siekmeyer, Christa (1999): Ein anderer Weg zur Groß- und Kleinschreibung. Stuttgart: Klett.

Röber-Siekmeyer, Christa; Noack, Christina; Dongus, Mareike; Eckert, Thomas (2002): Zwischenbericht zum Projekt „Entwicklung eines linguistisch orientierten Rechtschreibkonzepts für alemannisch sprechende HauptschülerInnen. Ms. PH Freiburg.

Stephan, Graham: String searching algorithms. Singapur, New Jersey, London u.a.: World Scientific.

Sowa, John (2000): Knowledge representation : logical, philosophical, and computational foundations. Pacific Grove (u.a.): Brooks/Cole, 2000.

Sproat, Richard (2000): A Computational Theory of Writing Systems. Cambridge: Cambridge University Press.

Thelen, Tobias (2000): Osnabrück Bildergeschichtenkorpus. Ms. Osnabrück. http://korpus.tobiasthelen.de

Thelen, Tobias (2001a): „Wie passt das Wort BETTEN in das Haus?“ Grundlagen und Ergebnisse des Computerprogramms MoPs zur Vermittlung der Schärfungsschreibung. In: Tophinke, Doris; Röber-Siekmeyer, Christa (Hrsg.): Schärfungsschreibung im Fokus. Baltmannsweiler: Schneider Verlag Hohengehren. S. 144-169.

Thelen, Tobias (2001b): Schrift ist berechenbar. Zur Systematik der Orthographie. In: Röber-Siekmeyer, Christa; Tophinke, Doris (Hrsg.): Schrifterwerbskonzepte zwischen Sprachwissenschaft und Pädagogik. Baltmannsweiler: Schneider. S. 66-82.

Thelen Tobias; Gust. Helmar (2002): Theorien auf dem Prüfstand – Evaluation phonologischer und orthograpischer Hypothesen durch computerlinguistische Simulation. In: Bommes, Michael; Tophinke, Doris; Noack, Christina: Sprache als Form. Opladen, Westdeutscher Verlag. S. 161-169.