Dissertation

zur Erlangung des Doktorgrades der Philosophie
im Fachgebiet Computerlinguistik und Künstliche Intelligenz
im Fachbereich Sprach- und Literaturwissenschaft
der Universität Osnabrück

zum Thema

Automatische Analyse orthographischer Leistungen von Schreibanfängern

vorgelegt von

Tobias Thelen

aus Haselünne

Osnabrück, 2010

Inhaltsverzeichnis

1. Einleitung

1.1. Analyse orthographischer Leistungen

Eine aussagekräftige Analyse orthographischer Leistungen beinhaltet nicht nur die Beurteilung „falsch“ oder „richtig“, sondern schließt Hypothesen über das Wie und Warum der Leistungen mit ein. Zunächst soll an zwei Beispielen verdeutlicht werden, inwiefern sich solche Analysen überhaupt 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 ein solches Verfahren in allen Fällen das Zustandekommen orthographischer Leistungen korrekt beschreiben könnte, müsste es über das gleiche Wissen und die gleichen kognitiven Fähigkeiten wie ein Mensch verfügen. Um eine Erklärung abzugeben, wie:

> <Hefte>[1] 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>. [2]

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. Stephen, 1994).

[1] Der linguistischen Konvention gemäß, werden orthographische Repräsentationen in <> eingefasst, phonologische (d.h. eine in Bezug auf eine Einzelsprache idealisierte Lautung) in // und schließlich phonetische (d.h. an der tatsächlich auftretenden Lautung orientierte) in []. Ein vorgestellter * kennzeichnet (beabsichtigt) falsche bzw. ungrammatische Formen.

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

2. Aus der Computerlinguistik stammt 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 (vgl. Sproat, 2000).

3. Die Künstliche Intelligenz hat Verfahren des Maschinellen Lernens entwickelt (vgl. Mitchell, 1997), die genutzt werden können, um aus gegebenen Analysebeispielen 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.

In der vorliegenden Arbeit werden Verfahren aus allen drei Bereichen genutzt, um unterschiedliche Arten orthographischer Leistungen automatisch analysierbar zu machen und die Möglichkeiten und Grenzen dieser Analysen aufzuzeigen.

1.2. Einsatzszenarien

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.

1.2.1. 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 und 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[3]. 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

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

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[4] 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, 2002) 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.

1.2.2. 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“

[4] Wie z.B. in der Software MoPs (Thelen, 2002) eingesetzt. Zwar enthält auch MoPs zu einem großen Teil behaviouristische Bestandteile, das zentrale Werkzeug „Haus“ und die damit möglichen Operationen lassen sich aber als kognitivistisch bezeichnen.

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 et al., 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, 1998).

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

1. Konsistenz: Versehen und Fehler 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.
2. 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.
3. 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 bezüglich der Definitionen oder Fehler in den Algorithmen unüberschaubare Konsequenzen haben.

1.3. Gliederung der Arbeit

In Kapitel 2 wird eine linguistische Analyse eines Teilbereichs der deutschen Orthographie, der Wortschreibung, in Hinblick auf a) Formalisierbarkeit und b) didaktische Vermittelbarkeit vorgenommen. Dabei werden zwei Sichtweisen gegenübergestellt und hinsichtlich ihrer Erklärungsmächtigkeit und Implementierbarkeit verglichen.

Kapitel 3 nimmt orthographische Leistungen, also sowohl Fehlschreibungen als auch korrekte Schreibungen in den Blick und definiert einen methodischen Rahmen für die Messung orthographischer Leistungen. Innerhalb dieses Rahmens spielen Fehler- bzw. Leistungstypologien eine herausragende Rolle. Zwei solcher Typologien, die zu förderdiagnostischen Zwecken weit verbreitet sind, werden

beschrieben und sowohl linguistisch als auch hinsichtlich des definierten methodischen Rahmens untersucht.

Aus festgestellten Defiziten der untersuchten Leistungstypologien heraus wird in Kapitel 4 ein neues, insbesondere linguistisch fundiertes Auswertungsschema für Wortschreibungen definiert und anhand eines praktischen Beispiels ausführlich mit den beiden in Kapitel 3 vorgestellten Verfahren verglichen.

Kapitel 5 beschreibt die Erhebung und Repräsentation von Texten, die als Testkorpus für die nachfolgend zu entwickelnden automatisierten Analyseverfahren dienen. Es wird ein komplexes Repräsentationsschema für Korpora frei verfasster Texte von Schreibanfängern vorgeschlagen und dafür notwendige Verarbeitungssoftware vorgestellt.

Den ersten Teilbereich zu entwickelnder und zu beurteilender Analyseverfahren bilden solche, die ohne Informationen über die korrekte, d.h. die vom Schreiber intendierte Schreibung auskommen. Ihnen steht nur der tatsächlich geschriebene Text zur Verfügung, in dem es potenziell fehlerhafte Formen zu erkennen gilt. In Kapitel 6 werden dazu regel-, muster- und lexikonbasierte Verfahren vorgestellt und anhand des Testcorpus bezüglich ihrer Leistungsfähigkeit verglichen.

Qualitativ deutlich weitergehende Analyseaussagen ermöglichen Verfahren, die vertiefte linguistische Analysen zur Annotation von korrekten Wortformen sowie zum Vergleich korrekter und fehlerhafter Formen verwenden. Kapitel 7 definiert ein Verfahren, das unter Kenntnis der intendierten Schreibungen in der Lage ist, das in Kapitel 4 vorgestellte Auswertungsschema auf große Corpora anzuwenden. Da es nicht Ziel der Arbeit ist, verallgemeinernde Aussagen aus dem erhebungsmethodisch nicht dafür geeigneten Testcorpus zu gewinnen, werden lediglich exemplarische Gesamtauswertungen präsentiert, die die grundsätzliche Leistungsfähigkeit des Verfahrens demonstrieren sollen.

Kapitel 8 verlässt den Bereich der Wortschreibung und definiert ein Analyseverfahren, mit dessen Hilfe Strategiehypothesen zur Groß- und Kleinschreibung generiert werden können. Anhand korrekter Texte, die manuell mit Features annotiert wurden, werden Abweichungen in Groß- und Kleinschreibung auf einzelne Merkmale zurückgeführt, die nach Auswertung zu schreiberspezifischen Entscheidungsbäumen zusammengeführt werden.

Ein Ziel dieser Arbeit ist es, praxistaugliche Analyseverfahren grundlegend zu entwerfen. In Kapitel 9 werden drei Implementierungen beschrieben, in denen die Verfahren bereits praktisch eingesetzt werden. Für das Lernportal www.ich-will-schreiben-lernen.de werden Teile der entwickelten Algorithmen verwendet, um individuelle Arbeitspläne für Lerner zu erstellen; in einem Forschungsprojekt der Universität Lüneburg kommt eine Online-Umsetzung der Verfahren zum Einsatz, um große Längsschnitterhebungen von Schülerschreibungen zu analysieren. Die für die Korpusrepräsentation und -aufbereitung genutzten Verfahren werden in Kombination mit den Analysealgorithmen außerdem genutzt, um eine E-Learning-Umgebung für die Deutschlehrerausbildung zu gestalten.

Ein zusammenfassendes Fazit wird in Kapitel 10 gezogen und weitere mögliche, sowie notwendige Entwicklungsschritte werden skizziert.

Der Arbeit ist eine CD-Rom mit ausführlichem Anhang beigefügt. Dort finden sich die XML-Daten des Testcorpus, die Quellen der entwickelten Analysesoftware und die Ergebnisse vollständiger Analyseläufe.

2. Orthographie und Linguistik

Die immer wieder aufflackernde Diskussion um die Rechtschreibreform zeigt großes Interesse, aber auch große Unsicherheiten in bezug auf den Gegenstand „Rechtschreibung“. Hauptanliegen der Reform war eine Vereinfachung der deutschen Orthographie mit Blick auf die Hauptschwierigkeiten von Schreibanfängern, d.h. die Ausschaltung häufiger Fehler. Dabei wurde davon ausgegangen, dass ein häufiger Fehler gleichzeitig mit einer „Unregelmäßigkeit“ der Orthographie zusammen fallen müsse. Nicht berücksichtigt wurde die Möglichkeit, dass ein anderer Blick auf die deutsche Rechtschreibung als System und eine damit einhergehende veränderte didaktische Herangehensweise solche „Unregelmäßigkeiten“ als sehr wohl begründet und „berechenbar“ erscheinen lassen könnte. Im Folgenden werden für einen Teilbereich der deutschen Orthographie, die Wortschreibung, zwei solcher Blickwinkel aus sprachwissenschaftlicher Sicht vorgestellt und diskutiert.

Die Aussage „Schrift ist berechenbar“ ist durchaus doppeldeutig zu lesen. Zum einen als Versicherung, dass die Schrift - genauer hier: die Rechtschreibung des Deutschen - nicht „falsch spielt“, sondern verlässlich und vorhersagbar ist. Zum anderen in einem strengeren, mathematischen Sinne, dass Schreibungen durch logische Schritte herleitbar seien. Damit stellt sich die Frage nach der korrekten Schreibung als ein Abbildungsproblem dar: Gegeben eine Menge sprachlicher Informationen und eine Menge von Handlungsanweisungen (im folgenden als „Regeln“ bezeichnet) ist bei vollständiger Berechenbarkeit der Schrift eine eindeutige Abbildung möglich: Aus gesprochener oder gedachter Sprache wird Schrift. Diese Abbildung soll nicht nur in eine Richtung möglich sein, sondern auch umgekehrt gelten: Aus einer gegebenen schriftlichen Repräsentation soll mithilfe klar angebbarer Regeln die Rücküberführung in gesprochene Sprache möglich sein.

Übertragen in die Terminologie der Informatik hat das Problem zwei Aspekte:

- Datenstrukturen: Eine Spezifikation der zur Verfügung stehenden Informationen. Hier betrifft dies zwei Ebenen: Die schriftliche und die „interne“ (d.h. z.B. lautliche, grammatische usw.). Datenstrukturen sind Repräsentationen von Informationen, die hier zu betrachtenden Datenstrukturen sind schriftliche Repräsentationen einerseits und nicht-schritliche andererseits.
- Algorithmus: Ein Algorithmus ist eine endliche Menge von Handlungsanweisungen, die mit hinreichender Genauigkeit angegeben werden müssen. In der Regel werden zur Beschreibung eines Algorithmus elementare Operationen verwendet, die einem Formalismus entstammen, für den bewiesen ist, dass mit ihm alle berechenbaren Funktionen ausgedrückt werden können. Eine Klasse solcher Formalismen sind Programmiersprachen.

Mit der Beschreibung geeigneter Datenstrukturen und Algorithmen ist die Formalisierbarkeit und deterministische Bearbeitbarkeit eines Problems bewiesen, d.h. die Leistungsfähigkeit einer dadurch ausgedrückten Theorie ist exakt und sicher bestimmbar. Gleichzeitig ist damit eine Rückführung des Problems auf eine mathematisch fassbare Basis geleistet: Jede andere Beschreibung bzw. Formalisierung des Problems, die die gleichen Resultate liefert, ist mathematisch äquivalent. Daraus folgen zwei Aussagen:

- Um die Frage zu klären, ob das Problem lösbar ist, ist es ausreichend, (nur) einen Weg aufzuzeigen und exakt zu operationalisieren.
- Keine Lösung ist im Sinne der Berechenbarkeitstheorie „richtiger“ als eine andere. Dabei ist das übliche Problem von Wahrheitsaussagen über Modelle und Theorien zu berücksichtigen:

> Physikalische Begriffe sind freie Schaffungen des Geistes und ergeben sich nicht etwa, wie man sehr leicht zu glauben geneigt ist, zwangsläufig aus den Verhältnissen in der Außenwelt. Bei unseren Bemühungen, die Wirklichkeit zu begreifen machen wir es manchmal wie ein Mann, der versucht, hinter den Mechanismus einer geschlossenen Taschenuhr zu kommen. Er sieht das Zifferblatt, sieht, wie sich die Zeiger bewegen, hört sogar das Ticken, doch er er hat keine Möglichkeit, das Gehäuse aufzumachen. Wenn er scharfsinnig ist, denkt er sich vielleicht irgendeinen Mechanismus aus, dem er all das zuschreiben kann, was er sieht, doch er ist sich wohl niemals sicher, daß seine Idee die einzige ist, mit der sich seine Beobachtungen erklären lassen. Er ist niemals in der Lage, seine Ideen an Hand des wirklichen Mechanismus zu überprüfen. (A. Einstein, zitiert nach von Glasersfeld (1996))

Für sprachliche Daten und somit auch das hier untersuchte Problem ist Eindeutigkeit und vollständige Berechenbarkeit nicht zu erwarten, wie die Beispiele im nächsten Abschnitt schnell zeigen werden. Damit verkompliziert sich die Situation, denn der Grad der Berechenbarkeit des Gegenstandes ist eine Unbekannte, gleichwohl eine Obergrenze für mögliche Theorien. Der Grad der Berechenbarkeit ist eine Funktion der zur Verfügung stehenden Informationen und der verwendeten Algorithmen.

2.1. Erste Betrachtung

Ein neutraler Betrachter kann beim Betrachten des amtlichen Regelwerks der deutschen Orthographie leicht zu der Auffassung gelangen, dass es zwar einen regelhaften Teil gibt, der aber durch viele Einzelfälle und Ausnahmen verwässert wird. Als Beispiele seien hier Teile der Formulierungen zur Schärfungs- und Dehnungsmarkierung zitiert:

§4: In acht Fallgruppen verdoppelt man den Buchstaben für den einzelnen Konsonanten nicht, obwohl dieser einem betonten kurzen Vokal folgt.

§12: In Einzelfällen kennzeichnet man die Länge des Vokals [i:] zusätzlich mit dem Buchstaben h und schreibt ih oder ieh.

Beides sind Regeln der „zweiten Ebene“, d.h. sie betreffen Ausnahmen zu einer vorher gegebenen Hauptregel. Zusätzlich gibt es noch ein Verzeichnis von Einzelfällen - das eigentliche Wörterbuch. Ein unsicherer Schreiber, der die Schreibung eines Wortes mithilfe des amtlichen Regelwerks herausfinden möchte, müsste folgendermaßen vorgehen:

1. In der Wortliste nachschlagen, ob das gesuchte Wort dort als „Ausnahme“ verzeichnet ist.
2. Wenn nicht: Unter der einschlägigen Regel nachschlagen, ob die Bedingungen für einen der untergeordneten Fälle erfüllt sind.
3. Wenn nicht: Die einschlägigen Hauptregeln anwenden.

Dieses Vorgehen ist nicht praktikabel, wenn die Formulierung der Regeln viele Ausnahmen, d.h. nicht regelhafte Fälle bedingt und das Wörterverzeichnis so umfangreich wird, dass die Anwendung einer Regel zumindest vom Vorgehen her umständlicher ist und selbst zur Ausnahme wird. Daher ist es nachvollziehbar, dass das Wörterverzeichnis, z.B. im Duden, mit dem Anspruch auf Vollständigkeit den größten Raum einnimmt und das Regelwerk vom durchschnittlichen Benutzer nicht verwendet wird. Der Blick auf die Rechtschreibung als ein systematisches Gebilde ist damit verstellt.

2.2. Wortschreibung

Der Bereich der Wortschreibung betrifft die Schreibung einzelner Wörter, wie sie ohne Kenntnis des Kontextes entscheidbar ist. Ausgenommen sind demnach die Bereiche Groß- und Kleinschreibung - der syntaktische Informationen benötigt (vgl. z.B. Maas, 1992, 156ff., Eisenberg, 1998, 326ff.) - und Getrennt- und Zusammenschreibung, ebenso die Fragen der Zeichensetzung und der Wortbrechung am Zeilenende. Für geübte Schreiber stellt die Wortschreibung oft nur ein geringes Problem dar, hingegen ist sie für Schreibanfänger und weniger geübte Schreiber eine große Hürde.

(1) /valt/ <Wald> <wallt>
/kantə/ <Kante> <kannte>
/vaɪzə/ <Weise> <Waise>
/hoɪtə/ <heute> <Häute>
/reːdə/ <Rede> <Reede>

(2) /hɛlt/ <Held> <hält> <hellt>
/fɛlt/ <Feld> <fällt>
/vɛlt/ <Welt> <wellt>

(3) /mɛlt/ <melt> <meld> <mellt> *<melld> <mält> ?<mäld> <mällt> *<mälld>

Die Vorstellung, dass Wörter im Deutschen im Wesentlichen so geschrieben werden, „wie man sie spricht“ erweist sich schnell als hinfällig, wenn Homophone, d.h. gleichlautende Wörter mit unterschiedlicher Schreibung, wie in (1), betrachtet werden. Für einige der Fälle in (1) oder (2) können die meisten Schreiber des Deutschen Gründe für die unterschiedliche Schreibung nennen, wie z.B. „<Häute> kommt von <Haut> - deshalb wird es mit <äu> geschrieben.“ oder „Für <Wald> gibt es <Wälder>, deshalb wird ein <d> geschrieben.“ Allein die Kenntnis der Lautung eines Wortes reicht also nicht aus, um sicher auf die Schreibung zu schließen. Die Schreibungen sind aber auch nicht unabhängig von der Lautung, wie die relativ gleichmäßige Variation bei den drei Beispielen in (2) verdeutlicht: <l> oder <ll>, <e> oder <ä>, <d> oder <t> am Ende. Kombinatorisch ergeben sich für ein Pseudowort wie /mɛlt/ in (3) acht Möglichkeiten, von denen allerdings die beiden mit * markierten Schreibvarianten sofort auffallen und ausgeschlossen werden können. Zumindest zweifelhaft erscheint die Möglichkeit <mäld>.

(4) <lache> /laxə/ /laːxə/
<rasten> /rastən/ (von Rast) /raːstən/ (von rasen)
<knie> /kniː/ (Singular) /kniːə/ (Plural)
<käschen> /kɛsçən/ (kl. Käse) /kɛʃən/ (mit dem Käscher fangen)

(5) a. <Kabel> /kaː.bəl/
<Kamel> /ka.meːl/
<Gamel> /gaː.məl/ /ga.meːl/

b. <Befel> /beː.fəl/ /bə.feːl/
<Befehl> /bə.feːl/

In umgekehrter Richtung ist es aber auch nicht mit Sicherheit möglich, vom geschriebenen Wort auf die korrekte Lautung zu schließen. In (4) sind - Groß- und Kleinschreibung ignorierend - drei Fälle aufgeführt, die für eine Schreibung verschiedene Lautungen zulassen. Einige der Beispiele wirken exotisch, was darauf zurückzuführen ist, dass solche Fälle, die zu zwei verschiedenen tatsächlich möglichen Wörtern führen, selten sind. Die Beispiele in (5) zeigen aber, dass es auch für vermeintlich eindeutige Wörter verschiedene Lesarten gibt. <Kabel> und <Kamel> unterscheiden sich in ihrer Betonung, dem Wortakzent: <Kabel> ist anfangsbetont, <Kamel> endbetont. Das Pseudowort <Gamel> könnte nach diesen beiden Mustern auf zwei Arten betont werden, allerdings wird ein Leser, wenn er nicht direkt zuvor mit <Kamel> konfrontiert wurde, automatisch eine Betonung wie in <Kabel> vornehmen. Das Paar in (5-b) zeigt, dass es auch Fälle gibt, in denen die Betonung absolut eindeutig ist: <Befehl> kann nicht anfangsbetont gelesen werden.

2.3. Die Standardsicht auf die deutsche Wortschreibung

Die Standardsicht auf die Wortschreibung im Deutschen ist die einer zugrundeliegenden Laut-Buchstaben-Zuordnung:

- „Als grundlegend im Sinne dieser orthographischen Regelung gelten die folgenden Laut-Buchstaben-Zuordnungen.“ (Amtliches Regelwerk, §1)
- „Die graphische Ebene steht in einem Wechselverhältnis mit der phonologischen Ebene, wobei zwischen den Einheiten der phonologischen und der graphischen Ebene eine mehr oder weniger ausgeprägte Parallelität besteht.“ (Nerius und Autorenkollektiv, 1989, 64)
- „Das Deutsche hat eine Alphabetschrift, und damit ist klar, daß die Beziehung zwischen Lauten und Buchstaben grundlegend für die Schreibung ist [...]“ (Augst und Stock, 1997, 115)
- „Hauptregel: Unter Berücksichtigung des Wortaufbaus gelten [...] folgende Zuordnungen zwischen abstrakten Lauten (Phonemen) und Buchstaben als Basis [...]“ (Augst und Dehn, 1998, 95)

Diese „traditionelle Sicht“ ist in der Deutschdidaktik weitgehend vorherrschend. Schuster (1995, S. 184) schreibt Mitte der Neunziger Jahre: „Da sich die deutsche Sprache über Jahrhunderte hinweg entwickelt hat, ist sie nicht widerspruchsfrei. Es gibt zwar Prinzipien, die aber untereinander konkurrieren, und immer wieder durchbrechen Ausnahmen die Regel.“

Demnach sind zwei Ketten diskreter Einheiten zu betrachten, auf der einen Seite eine Kette von Lauten (Phonemen) und auf der anderen eine Kette von schriftlichen Einheiten (Buchstaben bzw. Graphemen). Die zugrundeliegende Annahme einer Laut-Buchstaben-Zuordnung oder Graphem-Phonem-Korrespondenz (GPK) ist, dass Grapheme Phoneme repräsentieren und so aus einer Folge von Graphemen - einem geschriebenen Wort - eine Folge von Phonemen ableitbar ist. Das Wort wird durch sukzessive Anwendung der Korrespondenzregeln in eine Kette von Lauten überführt, es wird erlesen. Die Beispiele in (1)-(5) haben bereits gezeigt, dass dieser Vorgang, wie auch seine Umkehrung - aus einer Kette von Lauten die korrekte Schreibung eines Wortes abzuleiten - nicht durch 1:1-Zuordnungen von Phonemen und Graphemen zu korrekten Ergebnissen führen kann. Die Beispiele in (2) zeigen sehr deutlich eine Variation bei den Phonemen /ɛ/, das als <e> oder <ä>, und /t/, das als <d> oder <t> geschrieben wird. Auch für die andere Richtung lassen sich keine eindeutigen Regeln angeben, so kann das Graphem <d> dem Phonem /d/ (wie in <Feder>) oder dem Phonem /t/ (wie in <Wald>) entsprechen.

Viele Autoren nehmen vor dem Hintergrund der Nichteindeutigkeit der Laut-Buchstaben-Zuordnung an, dass es für jedes Phonem eine Standard- oder Basisentsprechung auf graphemischer Ebene gibt. In (6) sind Standardentsprechungen für einige Phoneme angegeben, wobei zu beachten ist, dass nicht nur Einzelphoneme eine Entsprechung haben können, sondern auch Phonemgruppen.

(6)
/ɛ/ → <e>
/a/ → <a>
/iː/ → <ie>
/l/ → <l>
/m/ → <m>
/n/ → <n>
/kv/ → <qu>
/ʃ/ → <sch>
/ŋ/ → <ng>

(7)
/r/ → <r> (99%)
/l/ → <l> (85%)
/f/ → <f> (59%)
/ʃ/ → <sch> (55%)
/iː/ → <ie> (78%)
/ɛ/ → <e> (88%)

Ermittelt werden können diese Standardzuordnungen grundsätzlich auf zwei verschiedenen Wegen. Zum einen über Häufigkeiten, d.h. für eine repräsentative Menge deutscher Wörter werden von Hand oder maschinell (vgl. dazu Maas et al., 1999, Kap. 3) Laut-Buchstaben-Zuordnungen ermittelt und ausgezählt. Naumann (1989, 88ff.) legt eine auf ausführlichen Untersuchungen beruhende Liste von Standardentsprechungen samt Häufigkeitsangaben vor, die in (7) ausschnittsweise wiedergegeben ist.

Thomé (1998, 72) bezeichnet die Standardzuordnungen als „Basisgrapheme“, die so interpretiert werden können, dass sie zu wählen sind, wenn keine besonderen Voraussetzungen für die Wahl eines anderen, so genannten „Orthographems“ vorliegen. Die gleiche Kategorisierung einzelner Laut-Buchstaben-Zuordnungen verwendet die „Hamburger Schreibprobe“ (s. z.B. May, 1999), die aus der Analyse gewählter Zuordnungen bei Schülerschreibungen Aussagen über vorherrschende Strategien (wie dementsprechend die alphabetische oder orthographische Strategie) ableitet.

Standardzuordnungen oder Basisgrapheme begründen die Redeweise von „lautgetreuer Schreibung“ (vgl. Naumann, 1989, 82). Die Orthographie eines Wortes ist dann „lautgetreu“, wenn sich seine korrekte Schreibung vollständig aus der Übersetzung der Phonemkette mithilfe der Standardzuordnungen ergibt. Besondere Probleme bereitet der Begriff vor allem dann, wenn die Häufigkeitszählung kein klares Bild ergeben hat und sich keine eindeutigen Regeln für die Entscheidung angeben lassen (s.u.). Im Falle von /iː/ → <ie> vs. /iː/ → <i> ist eine solche Situation gegeben. Welche der beiden möglichen Schreibungen für /miːnə/ sollte als „lautgetreu“ angesehen werden: <Mine> oder <Miene>? Eine Aufweichung des Begriffs hilft nur partiell, weil Eindeutigkeit dann nur noch für eine Richtung gewährleistet ist.

(8) /ɛ/ → <e>, <ä> <Held>, <hält>
/aː/ → <a>, <ah>, <aa> <Rat>, <Naht>, <Saat>
/iː/ → <ie>, <i>, <ieh> <Miene>, <Mine>, <sieht>
/l/ → <l>, <ll> <Wal>, <Wall>
/k/ → <k>, <ck>, <g>, <gg> <kalt>, <dick>, <Zug>, <Brigg>
/s/ → <s>, <ss>, <ß>, <> <bis>, <Biss>, <Blöße>, <Nation>
/ŋ/ → <ng>, <n> <lang>, <Schrank>

In (8) sind eine Reihe von vollständigen Laut-Buchstaben-Zuordnungen angegeben, wobei die Standardzuordnung jeweils an erster Stelle steht. Wie oben bereits erwähnt, sind viele dieser Schreibungen mittels Regeln herleit- und erklärbar. Erfolgreiche Schreiber des Deutschen sind offensichtlich in der Lage, diese Regeln anzuwenden, jedoch häufig nicht, sie zu explizieren, bzw. sie geben falsche Begründungen, die die intuitiven Regeln aber nicht stören (vgl. Weingarten, 2000). Damit ist ein Argument für die prinzipielle Lernbarkeit bzw. „intuitive Berechenbarkeit“ der deutschen Wortschreibung gegeben.

In der sprachwissenschaftlichen Literatur werden zur Erklärung der von den Standardzuordnungen abweichenden Zuordnungen eine Reihe von „Prinzipien“ eingeführt und diskutiert (vgl. Nerius, 1986; Nerius und Autorenkollektiv, 1989; Naumann, 1990), von denen einige im folgenden überblicksartig zusammengefasst werden:

Phonematisches Prinzip. Damit ist das soeben skizzierte Modell von Laut-Buchstaben-Standardzuordnungen gemeint.

Syllabisches Prinzip. Nerius und Autorenkollektiv (1989, 76) nehmen die Einheit „Silbe“ nur für die graphische Worttrennung als relevant an. In den vergangenen Jahren sind silbenorientierte Erklärungsansätze für die Dehnungs- und Schärfungsmarkierung stark in den Vordergrund der Diskussion gerückt (vgl. Maas, 1992, 278ff.; Eisenberg, 1998, 295ff.; Augst und Dehn, 1998, 113; Primus, 2000). Im nächsten Abschnitt wird eine Sicht auf die deutsche Wortschreibung vorgestellt, die prosodische, d.h. auch syllabische Eigenschaften von Wörtern in den Mittelpunkt rückt.

Morphematisches Prinzip. Viele der von den Standardzuordnungen abweichenden Schreibungen lassen sich durch feste oder „vererbte“ Schreibungen des Stammmorphems begründen. Ein <ä> für /ɛ/ ist demnach über eine verwandte Wortform zu erklären, die ein /a/ enthält, wie z.B. <hält> wg. <halten>. Die oben zitierte Schärfungsregel aus dem amtlichen Regelwerk (§2) ist ebenfalls stammorientiert. Das morphematische Prinzip erklärt ebenfalls die Markierung von Auslautverhärtung: <Hund> wg. <Hunde> und je nach betrachteter Theorie die Vererbung von Dehnungs-, Schärfungs- und Silbentrennendes-<h>-Markierungen auf Fälle, in denen sie nicht nötig wären (s. dazu den nächsten Abschnitt).

Lexikalisches Prinzip. Es gibt eine Reihe von Schreibungen, die sich mit den eben skizzierten Prinzip nicht erklären lassen. Das morphematische Prinzip kann z.B. nicht das <d> in <und> begründen, je nach angenommener Schärfungsregel sind entweder <in> und <ab> oder aber <wenn> und <dann> nicht erklärbar (s. dazu Maas et al., 1999, 105), usw. Hier ist anzunehmen, dass eine ganze Reihe von Schreibungen, bzw. nur deren „Besonderheiten“ im Lexikon abgelegt, d.h. auswendig gelernt werden müssen und nicht weiter begründet werden können.

Graphisches/ästhetisches Prinzip. Es gibt einige Regularitäten der Wortschreibung, die nur durch graphische Filter erklärt werden können. Vor der Neuregelung der Rechtschreibung wurde ein Aufeinandertreffen von drei gleichen Konsonantenbuchstaben vor einem weiteren Konsonantengraphem verhindert, in dem einer der Buchstaben gestrichen wurde, bei der Schärfungsmarkierung werden nur einbuchstabige Grapheme verdoppelt, also: <Fläche> statt *<Flächche>. Desweiteren wird von verschiedenen Autoren ein Zusammenhang zwischen „Schwere“ eines Wortes und der Dehnungsmarkierung angenommen, etwa der Art, dass Wörter mit einer weniger komplexen Repräsentation des Anfangsrandes eher ein Dehnungs-<h> aufweisen, als andere (vgl. Eisenberg, 1998, 301; Primus, 2000, für eine Auszählung der Fälle Maas et al., 1999, 121).

Etymologisches/historisches Prinzip. Bei vielen Wörtern ist Wissen über die Herkunft die einzige Möglichkeit, ihre Schreibung herzuleiten. Die Frage, ob /iː/ als <ie> oder <i> verschriftet wird, wird laut amtlichem Regelwerk (§1, Absatz 2) von der Unterscheidung „einheimisches Wort“ oder „Fremdwort“ bestimmt (s. dazu auch Maas et al., 1999, 106ff.). Einige Schreibungen sind nur über verwandte Formen bzw. Wortbildungsmuster erklärbar, die synchron nicht mehr anzutreffen sind: <ll> in <Kellner>, <Ä> in <Ähre> usw.

Das Zusammenwirken dieser Prinzipien ergibt ein komplexes Bild, das den Vorteil einer weitestmöglichen Begründung jeder einzelnen Schreibung hat. Eine Implementation als Computerprogramm, dass Schreibungen mithilfe von Regeln aus zugrundeliegenden Informationen ableiten soll, ist mit diesem Ansatz gut möglich, wenn die Regeln selbst hinreichend gut formalisiert werden können. Im Projekt „Computerbasierte Modellierung orthographischer Prozesse“ wurde ein solches Programm zur Überprüfung verschiedener Regelhypothesen entwickelt (Maas et al., 1999), die von Sproat (2000) entworfene sprach- und schriftsystemübergreifende Architektur zur Abbildung orthographischer auf lautliche Daten geht ebenfalls wesentlich von einer segmentalen Entsprechung verschiedener relevanter Ebenen aus.

2.4. Eine alternative Sicht: Schrift ist für den Leser da

Es ist offensichtlich, dass Schreiben aufwendiger als Lesen ist. Sinn und Zweck der Schrift ist die möglichst einfache und orts- sowie leserunabhängige Rekonstruktion des Ausgedrückten. Orthographie als Normierung von Schrift stellt die Austauschbarkeit und Dauerhaftigkeit sicher. Das Vorbild für die Schrift ist die gesprochene Sprache. Eine zu einfache, sich nur auf einzelne Teilaspekte der gesprochenen Sprache beschränkende Kodierung von Sprache in Schrift würde diesem Ziel widerstreben, weil der Rekonstruktionsprozess dann auf große Probleme und Mehrdeutigkeiten stoßen würde. Allerdings darf das Schreiben auch nicht zu kompliziert sein, weil die Rechtschreibung dann nicht mehr erlernbar wäre. Die Lösung besteht in einer Abbildung der Strukturen gesprochener Sprache auf Strukturen der Schrift, so dass „Schreiben lernen“ sich auf schon gelernte Strukturen beziehen kann. Die Orthographie wird demnach von einer grundlegenden Maxime geprägt, die verschiedenen Randbedingungen unterliegt (Maas, 2000, 44f.):

M[axime]1 Schreib, wie du gelesen werden willst.

R[andbedingung]1 Die Orthographie soll durch ihre Fundierung in der Struktur der gesprochenen (deutschen) Sprache optimal lernbar sein.

Die Struktur von Wörtern kann unter zwei Aspekten betrachtet werden: Auf der einen Seite die Lautung, die Formseite, die „erste Artikulation“ (Maas, 1999, 20), auf der anderen die morphologische Struktur, die Inhaltsseite, die „zweite Artikulation“ (ebd.). Die Dekodierung einer Äußerung aus der Schrift ist um so einfacher, je mehr Informationen in der Schrift enthalten sind. Wie im folgenden gezeigt wird, ist es möglich, die deutsche Wortschreibung dadurch zu charakterisieren, dass beide Aspekte mit denselben Mitteln repräsentiert werden, ohne sich dabei gegenseitig zu behindern. Dies ist kein Ergebnis einer notwendigen Entwicklung, sondern komplexer Adaptions- und Umdeutungsprozesse der für das Lateinische entstandenen Schrift (vgl. Maas, 2000, 51f.; als Überblick zu Klassifikationsansätzen von Schriftsystemen s. z.B. Sproat, 2000, 131ff.).

Im letzten Abschnitt wurde ganz selbstverständlich der Begriff des Lautes bzw. des Phonems verwendet, der für die Schreibung in Form von Laut-Buchstaben-Zuordnungen grundlegend sein soll. Zum erfolgreichen Erlernen der deutschen Orthographie wäre es demnach notwendig, Äußerungen bzw. Wörter in Phoneme zerlegen zu können. Ein häufig vorgebrachter Einwand ist, dass verwendete Phonembegriff selbst schriftbasiert sei (vgl. Eisenberg, 1998, 295). Demnach wäre es eine Voraussetzung zum Erlernen der Schrift, schon eine Vorstellung davon zu haben, was Schrift ist, bzw. welche Abbildungsmittel Schrift verwendet. D.h. wenn Laut und Buchstabe (bzw. Graphem) aufeinander bezogen werden, beide Konzepte aber nur verschiedene Seiten einer Medaille sind, dann kann nicht das eine aus dem anderen erklärt werden. Für den Leseanfänger ist es wenig nützlich, eine Kette von Buchstaben als eine Kette von Lauten, d.h. Buchstabennamen „lesen“ zu können - dabei ist es unerheblich, ob der Buchstabe <b> als /beː/ oder /bə/ bezeichnet wird (vgl. Röber-Siekmeyer und Pfisterer, 1998). In der Schrift steckt mehr als eine Kette von Lauten, nämlich Hinweise auf die weitere sprachliche Struktur. Die Grundoperationen der lautlichen Produktion sind nicht solche auf Einzelphonemen, sondern prosodische. Das grammatische „Rohmaterial“ einer Äußerung wird nicht durch Übertragung in eine Phonemkette zu lautlichem Material, sondern durch Anpassung an metrische, d.h. rhythmische Muster.

Abbildung 2.1 zeigt eine Hierarchie von prosodischen Ebenen. Diese Ebenen werden definiert als Domänen bestimmter Phänomene, von denen einige exemplarisch erläutert werden. Die Intonationsphrase ist der Teil einer Äußerung, in dem die Grundfrequenz der Stimme fällt und am Ende

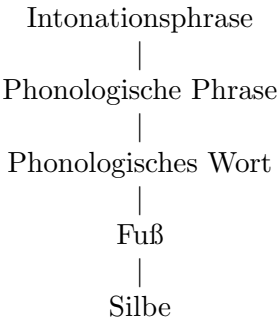

Abbildung 2.1.: Prosodische Hiearchie (aus: Wiese, 1996, 83)

	zweielementig	dreielementig
fallend	Trochäus `x -`	Daktylus `x - -`
steigend	Jambus `- x`	Anapest `- - x`

(`x` bezeichnet das betonte Element/den Kopf, `-` unbetonte Elemente)

Abbildung 2.2.: Häufige Fußtypen des Deutschen

mit einer Grenzmarkierung (Steigen, Fallen oder Halten der Stimmhöhe) den Typ einer Äußerung anzeigen kann, wie z.B. Frage, Aussage, oder Weiterleitung. Die Ebenen sind strikt geordnet, d.h. eine Intonationsphrase besteht genau aus einer oder mehreren phonologischen Phrasen, der Beginn einer Intonationsphrase fällt mit dem Beginn einer phonologischen Phrase zusammen, das Ende mit einem Ende und es gibt keine Lücken zwischen den phonologischen Phrasen. Das gleiche gilt analog für die anderen Ebenen. Die phonologische Phrase und das phonologisches Wort können durch das Phänomen der Koordinationsreduktion verdeutlich werden. Gleiche Teile zweier zusammen auftretender Wörter können bei einem von beiden weggelassen werden: „Feuerwehrmänner und -frauen“, „An- und Abfahrt“, aber nicht *„verlieren und -gessen“, *„Ho- und Hebel“. Das ausgelassene Teilstück ist immer ein phonologisches Wort (oder mehrere) und beide Teile stehen innerhalb einer phonologischen Phrase. Das phonologische Wort ist auch die Schnittstelle zur Morphologie: Der Beginn eines phonologischen Wortes fällt immer mit einer Morphemgrenze zusammen, genauso wie das Ende. Ein phonologisches Wort besteht aus einem oder mehrere Füßen, der Begriff ist analog zu dem in der klassischen Metrik verwendeten zu sehen. Ein Fuß besteht aus mehreren Elementen (hier: Silben), die Träger von Betonung sein können - konstitutiv für einen Fuß ist es, dass er genau ein betontes Element enthält, den Kopf. Alle anderen sind unbetont. In Abbildung 2.2 sind häufige Fußtypen klassifiziert.

Nur im Kontext von Fußstrukturen (und damit der gesamten prosodischen Analyse einer Äußerung) sind Silben sinnvoll betrachtbar. So ergeben sich drei Typen von Silben (vgl. Maas, 1999):

Betonbare (prominente) Silben. Jede prominente Silbe ist Kopf eines Fußes und umgekehrt. Die Ränder der Silbe können komplex sein, den Kern bildet immer ein Vollvokal, der zusätzlich durch die Art des Anschlusses an den folgenden Konsonanten charakterisiert wird. Ein loser Anschluss - d.h. der Vokal kann zu Ende artikuliert werden, „austrudeln“ - entspricht einem „Langvokal“, ein fester Anschluss - d.h. die Artikulation des Vokals wird vorzeitig abgebrochen,

der Vokal wird vom folgenden Konsonanten „abgeschnitten“ - einem „Kurzvokal“ (vgl. Kap. 8.2 Maas, 1999; Ramers, 1988, 106ff.; zum akustischen Korrelat des Silbenschnitts Spiekermann, 2000).

Nicht betonbare reduzierte Silben. Reduktionssilben enthalten entweder einen Schwa-Vokal: /ə/ wie in /'reːdə/ und /ɐ/ wie in /'vaːtɐ/, häufig aber auch einen silbischen Sonoraten (vgl. /'reːdn̩/, /'neːbl̩/ als Kern, darüber hinaus sind eine Reihe weiterer phonotaktischer Beschränkungen zu beachten (s. Maas, 1999, Kap. 9).

Nicht betonbare, nicht reduzierte Silben enthalten einen Vollvokal als Kern, der aber keine Anschlussopposition wie bei der prominenten Silbe, also auch keine Längenunterschiede, aufweisen kann (s. Maas, 1999; Ramers, 1988, Kap. 7,83ff.).

Das metrische Grundmuster des Deutschen ist der Trochäus (Maas, 1999, 105), wobei das unbetonte Element in der Regel eine Reduktionssilbe ist. Trochäische Formen können verlängert oder verkürzt werden, im ersten Fall können daktylische oder noch komplexere Muster entstehen, im letzteren spricht man von degenerierten Fußstrukturen, die zwar nur noch einsilbig sind, aber das Potenzial zum vollständigen Fuß in sich tragen. Die prosodische Grundform eines Wortes ist nicht immer identisch mit der im Wörterbuch aufgeführten „Zitierform“, d.h. die prosodische Grundform der Zitierform <Bett> ist <Betten>. Eine fallende Fußstruktur beginnt per Definition mit einem betonten Element, jedoch kann ein Auftakt in Form einer unbetonten Silbe vor den Fuß treten. Zunächst verwirrend erscheint der Fall eines degenerierten trochäischen Fußes mit Auftakt (z.B. <Gebell>), der wie ein echter steigender Fuß aussieht. Solche Jamben kommen aber als einheimische deutsche Wörter nicht vor, wohl aber in aus anderen Sprachen übernommenen oder entlehnten Wörtern. In (9) a.-c. sind Wörter mit fallender Fußstruktur, in (9) d. solche mit steigender aufgeführt.

(9)
a. fallend:
<Leben>, <Löffel>, <gebe>, <saures>, <Vater>, <sandig>, <Atem>, <Arbeiter>, <blauere>, <spannendere>
b. degeneriert fallend:
<Bett>. <Sand>, <Tag>, <läuft>, <bringst>, <blau>
c. fallend mit Auftakt:
<Gebell>, <Verluste>, <Forelle>, <erwünscht>
d. steigend:
<Balkon>, <Natur>, <Kamel>, <salopp>, <Büro>

Betrachtet man nun die Mittel der Schrift, um Fußstrukturen - also Betonungsmuster - zu repräsentieren und leicht erschließbar zu machen, wird eine sehr systematische Kennzeichnung des Anschlusstyps durch die Schrift sichtbar. Reduktionssilben enthalten immer ein <e>, das nach dieser Sichtweise nicht vorrangig das Phonem /ə/ repräsentiert, sondern Erkennungsmerkmal einer reduzierten Silbe ist. Auch in Fällen, in denen phonetisch kein Schwa vorhanden ist, steht ein <e>: <Vater>, <Löffel>, <geben>. Damit sind aus Sicht des Lesers Reduktionssilben leicht erkennbar, wie die Beispiele in (5) schon angedeutet haben. Probleme bereiten dann allerdings Formen, die wie trochäische Wörter aussehen, aber keine sind, wie <Kamel>. Solche Formen sind selten und als lexikalisiert anzunehmen.

<...VKe...>	loser Anschluss	<Rede>, <schlafen>
<...VVKe...>	loser Anschluss	<Raute>, <Paare>
<...Vhe...>	loser Anschluss	<Rehe>, <gehen>
<...VhKKe...>	loser Anschluss	<wohnte>, <Hühnchen>
<...VKKe...>	fester Anschluss	<Reste><Retter>

Abbildung 2.3.: Kennzeichnung des Anschlusstyps

Nicht nur die Betonungskontur des Wortes, sondern auch der Anschlusstyp - lose oder fest, d.h. in üblicher Terminologie: Kurz- oder Langvokal in der betonten Silbe - ist für den Leser einfach erschließbar, wie die „Leseregeln“ in Abbildung 2.3 zeigen.

Anders ausgedrückt: Stehen zwischen dem Vokalgraphem, das den Kern der prominenten Silbe repräsentiert und dem <e>, das den Kern der Reduktionssilbe repräsentiert, zwei Konsonantengrapheme, dann liegt ein fester Anschluss (Kurzvokal) vor, ansonsten ein loser (Langvokal). Das Dehnungs-<h> kann die Lesart als festen Anschluss „blockieren“ und losen Anschluss erzwingen. Damit erscheinen Dehnung, Schärfung und silbentrennendes <h> nicht als Störungen einer grundlegenden Laut-Buchstaben-Zuordnung, sondern als zentraler Bestandteil der deutschen Wortschreibung, der sicher stellt, dass die Lautung eines Wortes mit den Aspekten Betonungsmuster und Anschlussverhältnis beim Lesen leicht erkennbar ist. Im Einzelnen stellen sich die Regeln wie folgt dar:

Schärfung: In einem Wort wie <Hütte> ist nur ein /t/-Laut hörbar, jedoch fordert die Leseregel für den festen Anschluss das Vorhandensein von zwei Konsonantengraphemen. Eine Möglichkeit, dies zu erreichen, ist die Verdopplung des entsprechenden Graphems eine Möglichkeit - die Fälle <ck> und <tz> statt <kk> und <zz> zeigen, dass es nicht die einzige ist.

Silbentrennendes-<h>: Um die Eigenständigkeit des die Reduktionssilbe repräsentierenden <e> zu sichern, wird ein <h> gesetzt, wenn kein konsonantisches Material an der Schnittstelle zwischen prominenter und reduzierter Silbe vorhanden ist. Die Markierung erfolgt im Falle von Monophthongen sehr regelhaft, bei Diphthongen allerdings nicht.

Dehnung: Aus Sicht des Lesers ist die Funktion des Dehnungs-<h> sehr klar. Die Fälle, in denen es nach der dritten Leseregel zwingend stehen müsste, sind allerdings selten, insbesondere wenn gefordert wird, dass beide Konsonantengrapheme zum Wortstamm gehören sollen.

Wiese (1996, 48) setzt für einsilbige Formen wie <Mond> oder <Obst> extrasyllabische Positionen außerhalb der Silbenstruktur an, so dass diese Formen als markiert anzusehen und entsprechend selten sind. Daher wird auch die relative Komplexität der Dehnungsmarkierung aus Sicht des Schreibers deutlich, es ergibt sich insgesamt ein Bild, dass nach Anwendung einiger Subregularitäten (wie der, dass ein Dehnungs-<h> nur vor <l>,<m>,<n> und <r> stehen kann und dem oben erwähnten tendenziellen Zusammenhang zwischen Komplexität des Anfangsrands und Wahrscheinlichkeit einer Dehnungsmarkierung) immer noch einige Unsicherheiten zurücklässt. Es bleibt aber festzuhalten, dass das Dehnungs-<h> in den Fällen, in denen es steht, eine Unterstützung beim Dekodieren der Schrift bietet.

Zu allen genannten Regeln gibt es eine Reihe von Fällen, die von ihnen nicht erklärt werden können. Dabei sind drei Bereiche hervorzuheben:

1. Mehrdeutigkeit des <e> und nicht-trochäische Formen: <Kamel>, <Beleg>, <gelesen>.
2. Mehrbuchstabige Konsonantengrapheme, die im Falle eines festen Anschlusses nicht verdoppelt werden: <Lusche>, <Dusche>, <Lache> - <Lache>, <waschen>, <wuschen>
3. Nicht markierter loser Anschluss und mehrere Konsonantengrapheme: <Monde>, <duschten>, <schulte> - <Schulter>

Experimente mit Pseudowörtern zeigen aber regelmäßig, dass die Leseregeln in Abbildung 3 als Standardannahmen verwendet werden.

Zu der Möglichkeit, prosodische Verhältnisse aus der Schreibung eines Wortes zu gewinnen, tritt die Kodierung morphologischer Informationen, die sich in der Forderung: „Stämme werden orthographisch maximal konstant repräsentiert." (Maas, 2000, 331) niederschlägt. Dabei ist es wichtig, dass die Leseregeln nicht verletzt werden. Gleiche Stämme werden durch die Verwendung gleicher Markierungen gekennzeichnet, wenn nicht eine Änderung der Anschlussverhältnisse (und damit eine potenzielle Verletzung der Leseregeln) dagegen spricht. Durch die Forderung nach Stammkonstanz stehen die Sondermarkierungen also auch in Fällen, in denen sie nach den Leseregeln nicht notwendig wären. (10-a) zeigt Beispiele für Dehnung, Schärfung und silbentrennendes-<h>, in denen eine konstante Schreibung des Stammes mit den Leseregeln vereinbar ist, (10-b) solche, bei denen das nicht der Fall wäre.

(10) a. <Ball> wg. <Bälle>
<geht> wg. <gehen>
<stellst> wg. <stellen>
<sieht> wg. <sehen>
<fällt> wg. <fallen>
<wohnen> wg. <wohnte>
b. <kam> trotz <kommen>
<ging> trotz <gehen>
<nimmt> trotz <nehmen>
<traf> trotz <treffen>

Die konstante Repräsentation von Stämmen ist auch in anderen Bereichen der Wortschreibung anzutreffen, z.B. der Markierung von Auslautverhärtung, der Begründung von <ä> und <äu>-Schreibungen sowie von <r>-Schreibungen wie in <klar> für die norddeutsche Varietät des Deutschen, in der erst eine Form wie <klare> ein /r/ erkennbar werden lässt.

(11) a. <Attacke>, <arrogant>, <Kommode>
b. <City>, <Psychologie>, <Bureau>/<Büro>

Die vom Leser ausgehende Betrachtungsweise der Wortschreibung kann dazu dienen, aus anderen Sprachen übernommene Wörter und Schreibungen zu identifizieren, weil sie häufig Muster beinhalten, die im Deutschen nicht vorkommen. Die Leseregeln können in Fällen wie in (11-a) nicht sinnvoll angewendet werden, in (11-b) werden Grapheme oder Graphemkombinationen verwendet, die in „normalen“ deutschen Wörtern nicht vorkommen: Beides kann als „Warnhinweis“ gesehen werden, sich bei der Dekodierung nicht auf das übliche Regelsystem zu stützen.

Unter dem Aspekt der Berechenbarkeit ist das Modell noch nicht vollständig. Es fehlt eine Komponente, die (beim Lesen) zu tatsächlichem phonetischen Material gelangt - das können Phoneme sein, müssen es aber nicht. Als Modell des Schreibens sind mehr Informationen notwendig als eine reine Phonemkette, aber diese Information muss in irgendeiner Form vorliegen. Eine Variante von Graphem-Phonem-Kombinationen kann auch hier angenommen werden, nicht als Ausgangspunkt, sondern als Modul, dass unter Berücksichtigung aller vorliegenden Informationen und Analysen positionsabhängig (in Bezug auf prosodische Ebenen wie die Silbe) entscheidet, welches Graphem/Phonem bzw. welche Kombination davon zu wählen ist.

2.5. Zusammenfassung

Es wurden zwei vom Ansatz her unterschiedliche Sichtweisen auf die deutsche Wortschreibung vorgestellt. Dabei hat sich die von einer Standardzuordnung von Lauten und Buchstaben ausgehende Sicht, die vom Standard abweichende Schreibungen als Auswahl eines anderen Graphems unter dem Einfluss verschiedener Prinzipien erklärt, als für viele Fälle brauchbares System erwiesen, dem allerdings eine grundsätzliche Erklärungsfähigkeit fehlt. Die zweite Sicht kann aus einer deutlich bezogenen Position - der eines Lesers - viele unzusammenhängend erscheinende Phänomene in einen gemeinsamen Erklärungsrahmen einbinden. Sie ist nicht in der gleichen Weise vom problematischen Phonembegriff abhängig wie die erste Sicht, sondern stützt sich auf insbesondere unsicheren Schreiblernern einfacher zugängliche metrische Verhältnisse.

Die beiden Modelle schließen sich nicht aus, sondern zeigen das Problem von verschiedenen Seiten. Die Eigenschaften in bezug auf „Berechenbarkeit“ sind unterschiedlich: Wenn beliebig komplexe Regeln zugelassen werden können, die auf die vorhandenen Informationen auf unterschiedliche Weise zugreifen, bietet das GPK+„Prinzipien“-Modell den Vorteil, alle verfügbaren Subregularitäten und Einzelfälle miteinzubeziehen. Wenn es allein um eine möglichst große Abdeckung der beobachtbaren Fälle geht, ist dieses Modell vorzuziehen. In Fällen allerdings, in denen ein vorwiegend systematischer Blick auf die deutsche Wortschreibung notwendig ist, kann das leserorientierte Modell entscheidende Zusammenhänge erhellen.

3. Orthographische Leistungen

3.1. Messung von Leistungen

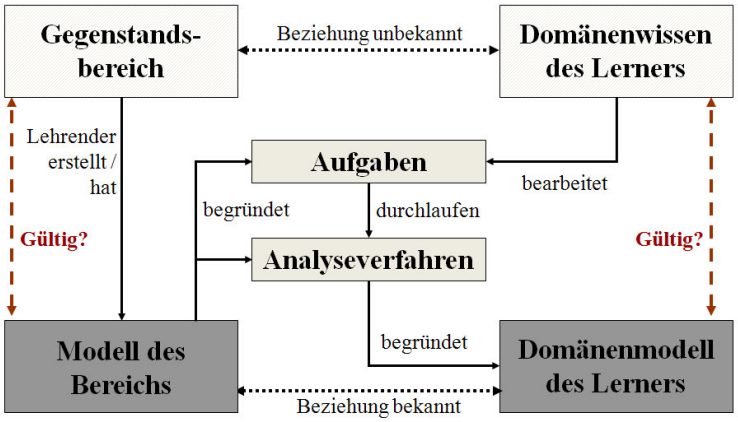

Abbildung 3.1.: Grundmodell der Messung orthographischer Leistungen

In Abbildung 3.1 ist ein Grundschema für Analyseinstrumente dargestellt. Das Domänenwissen des Lerners (hier: Fähigkeiten im Umgang mit Schrift und Hypothesen über Schrift) ist einer direkten Beobachtung entzogen. Ebenso ist der Gegenstandsbereich (hier: Schrift, bzw. die deutsche Orthographie) nicht durch unmittelbare Beobachtung erfassbar. Aussagen über Kenntnisse und Fähigkeiten von Lernern stellen dennoch eine Beziehung zwischen Gegenstandsbereich und Domänenwissen her. Das kann nur unter Zuhilfenahme zweier konstruierter Größen geschehen: Einem Modell des Gegenstandsbereichs und einem Modell des Lernerwissens. Ersteres definiert Aufgaben und Bewertungsschemata, die wiederum zu einer Beurteilung, d.h. Bildung eines Lernermodells führen. Da die Beziehung zwischen diesen beiden Modellen bekannt ist, können Gegenstand und Wissen auf dieser Ebene modellhaft in Beziehung gesetzt werden. Entscheidend für die Validität von Analyseinstrumenten ist allerdings die Frage, inwiefern die Modelle den zu modellierenden Größen angemessen sind.

Es wird hier davon ausgegangen, dass sich das Modell des Gegenstandes dann als gültig erweisen kann, wenn es sich an linguistischen Erkenntnissen über Strukturen und Prinzipien der Schrift bzw. der Orthographie orientiert. Aus den Analysen müssen dann didaktische Handlungen ableitbar sein, die auf die wahrnehmbaren Strukturen aufsetzt.

3.2. Leistungstypologien

Auf den ersten Blick erscheinen Spontanschreibungen von Schreibanfängern chaotisch und – vor allem – falsch. Aus Angst vor möglichen negativen Einflüssen von Fehlern auf die Lernentwicklung, wie etwa der Befürchtung, dass sich unkorrigierte Schreibungen oder an die Tafel angeschriebene falsche Wortformen dauerhaft einprägen, herrschten lange Zeit unterrichtliche Methoden vor, die vor allem auf Vermeidung von Fehlerquellen ausgerichtet waren. Insbesondere sind hier Fibellehrgänge zu nennen, die den Umgang mit Schrift im ersten Schuljahr im Wesentlichen auf vorgegebenes Wortmaterial beschränken. Als Form der Leistungskontrolle ist dabei und auch später das „geübte Diktat“ vorsehen – die Lernleistung wird vorrangig im Reproduzieren bekannter und auswendig gelernter Muster gesehen. In diesem Zusammenhang stehen auch Grundwortschätze, die eine Menge von ca. 1500 Wortformen für die gesamte Grundschulzeit als zu beherrschenden Grundstock definieren, der es den Lernern ermöglichen soll, korrekte Schreibungen anderer Wörter per Analogieschluss aus dem bekannten Material herzuleiten.

Bedenken und Einwände gegen dieses Vorgehen enstanden aus zwei Richtungen:

Zum einen pädagogisch motiviert aus der Auffassung, dass mit der Beschränkung auf Vorgegebenes kreatives Potenzial zurückgedrängt und die für Lernmotivation und -erfolg wichtige Berücksichtigung der spezifischen Situation und eigenen Interessen des Kindes unterbleibt.

Zum anderen aus der Auffassung, dass Fehler wichtige und notwendige „Zwischenschritte“ auf dem Weg zur vollständigen Beherrschung der orthographischen Norm sind und darüber hinaus Einblicke in den Stand und die Entwicklung des Schrifterwerbsprozesses ermöglichen. Neben lernpsychologischen Begründungen für diese Auffassung, die in der Entwicklung von „Stufenmodellen“ (s.u.) gipfeln, sind vor allem sprachsystematische Erwägungen zu nennen, die die deutsche Orthographie als System ansehen, das aus lautlichen und grammatischen Eigenschaften der deutschen Sprache erklärt und gelernt werden kann (vgl. Maas, 1992; Röber-Siekmeyer, 1993).

3.2.1. Standardisierte Testverfahren

Seit den achtziger Jahren werden eine Reihe standardisierter Tests angeboten und weiterentwickelt, die es dem Lehrer ermöglichen sollen, mittels einfach durchzuführender Schreibproben und anschließender Auswertung objektive Aussagen über den individuellen Stand des untersuchten Schülers zu gewinnen. Daraus abgeleitet sollen „Störungen im Aneignungsprozess differenziert erfassbar“ (May, 1999, 40) werden.

Im Wesentlichen sind hier der *Diagnostische Rechtschreibtest* (DRT), die *Hamburger Schreibprobe* (HSP), und die *Aachener förderdiagnostische Rechtschreibanalyse* (AFRa, s. Herné und Naumann, 2002) zu nennen. Die folgenden Ausführungen konzentrieren sich auf die *Hamburger Schreibprobe*, die wegen ihrer augenscheinlichen Einfachheit besonders beliebt ist. Die Tests funktionieren sämtlich nach einem vergleichbaren Schema:

- Eine für eine bestimmte Klassenstufe ausgewählte feste Menge an Wortformen bzw. Sätzen wird vom Probanden geschrieben. Es wird dabei vorausgesetzt, dass die Wortformen den Probanden nicht geläufig sind, bzw. der Einfluss bekannter Wortformen für alle Schüler gleich ist.

- Die Auswertungskategorien sind so operationalisiert, dass a) die Auswertung vom Lehrer selbst manuell in kurzer Zeit (3-10 Minuten) durchgeführt werden kann und b) die Leistungen auf eine oder mehrere (in der Regel numerische) Skalen abgebildet werden. Das können sehr einfache Kategorien, wie die Anzahl der korrekt geschriebenen Wortformen sein, etwas differenzierter ist das unten diskutierte Konzept der „Graphemtreffer“, bis hin zur Zählung spezifischer Phänomene, wie z.B. „Auslautverhärtung nicht markiert“. Gemeinsames Ziel dieser Skalen ist die Reduktion der konkret vorliegenden Schreibungen auf eine Reihe zählbarer Merkmale.
- Für jede Klassenstufe liegen „geeichte“ Normwerte vor: Für die ausgewählten Wortformen werden repräsentative Ergebnisse erhoben, für die ein Bezug zwischen Skalenwert und klassenstufenspezifischer „Norm“ geschaffen wird. Das Testdesign besteht also aus den beiden Schritten Auswahl der Testwortmenge und Ermittlung von Ergebnissen für diese Menge (Eichung).

  Durch Kombination der Einzelergebnisse ergibt sich letztendlich ein einzelner skalarer Wert, der den Stand des Probanden relativ zur „normalen“ Entwicklung wiederspiegeln soll. Bei der *Hamburger Schreibprobe* (HSP) ist dies ein „Prozentrang“, der angibt, welcher Anteil der Schüler derselben Klassenstufe eine schlechtere Leistung zeigt. Ein Prozentrang von 50 kennzeichnet demnach eine durchschnittliche Leistung, ein Prozentrang von 90 eine sehr überdurchschnittliche Leistung.

Hinter diesen Verfahren steht die Motivation, die komplexen und auf den ersten Blick schwer durchschaubaren Leistungen auf einen einzelnen, vorgeblich objektiven Wert zu reduzieren. In Analogie zum „Intelligenzquotienten“, der ebenfalls auf einer Eichung der Testauswertung in Bezug auf eine repräsentative Menge an Rohergebnissen basiert, sind Abweichungen nach oben oder unten einfach ablesbar und dienen als objektive Rechtfertigung für die Einteilung in schwache und gute Schüler. Letztendlich liefern sie im Zweifelsfall auch eine Begründung für das Versagen des schulischen Schreiblehrgangs.

Das Vorgehen standardisierter Tests ist aus mehreren Gründen zu kritisieren:

1. Die Tests behaupten, lehrgangsunabhängig zu sein. Wenn sie tatsächlich für jeden Schreiblehrgang objektive Ergebnisse lieferten, müsste dies entweder bedeuten,
   - dass der Lehrgang keinen Einfluss auf die Schreibentwicklung der Schüler hat. Diese Auffassung wird unterstützt durch die Annahme relativ starrer „Stufenmodelle“ des Schrifterwerbs (für eine umfassende Übersicht s. Thomé, 1998). Oder
   - dass entgegen der behaupteten Unabhängigkeit doch eine bestimmte Art von Lehrgang angenommen wird, der die vom Test bzw. der vorgenommenen Eichung unterstellte Entwicklung zur Folge hat. Dies würde die Frage aufwerfen, ob nicht eine andere Art von Lehrgang andere Ergebnisse zeigen könnte und der Test somit gegen die Forderung nach Validität (s. May und Malitzky, 1999), d.h. der Erfassung zentraler Bereiche der orthographischen Kompetenz, der Übereinstimmung mit Lehrerbeurteilungen und der Voraussage künftiger Entwicklungen, verstieße.

2. Die Testwortmenge, an der der Test geeicht wurde und die deshalb nicht verändert werden darf, muss die tatsächlichen Problemfelder des Schrifterwerbs repräsentativ abdecken. Daher sind die für die konkrete Auswahl der Wörter angewendeten Kriterien von besonderer – gerade auch theoretischer – Bedeutung. Die Autoren des HSP nennen insbesondere:

   - „Bei der Wortauswahl wurde darauf geachtet neben der Repräsentation der wichtigsten Phänomene der deutschen Orthographie darauf geachtet, dass die zu schreibenden Begriffe der Erfahrungswelt der Schüler nahekommen und inhaltlich bedeutsam sind." (May und Malitzky, 1999)
   - „Daher wurde bei der Wortauswahl der Hamburger Schreibprobe auf eine geschlechtstypische Ausgewogenheit der Schreibwörter geachtet." (May und Malitzky, 1999)

   Damit gehen neben Annahmen über das „System Orthographie" auch solche über allgemeinere pädagogische Lernvoraussetzungen ein, deren Bezug zum Schrifterwerb vor allem durch lernpsychologische, kaum aber sprachwissenschaftliche, Argumente begründet wird.[1]

3. Die angesetzten Auswertungskategorien folgen nur in Ansätzen sprachwissenschaftlichen Erkenntnissen über die deutsche Orthographie. So wird in Erklärungen zur HSP (z.B. May, 1999) angenommen, dass Schärfung, sibentrennendes <h> und Auslautverhärtung regelhaft sind („Elemente, deren Verwendung hergeleitet werden kann"), jedoch wird in der Auswertung nicht zwischen der Beherrschung regelhafter und als nicht regelhaft angenommener Bereiche (Dehnung, /ks/-Schreibung, <f>/<v>-Schreibung) unterschieden. Insbesondere werden keine Hypothesen über die Regelbildung des Schreibenden ablesbar, eine Grundvoraussetzung für unterrichtliches Handeln, das auf Vermittlung bzw. „Entdecken" von Regeln im Sinne der Möglichkeit der Herleitung normgerechter Schreibungen aus der eigenen Artikulation und dem erworbenen Wissen über Sprache basiert.

4. Die HSP, aber auch vergleichbare Verfahren, betonen sehr stark den interindividuellen Aspekt der Untersuchung, die Leistungen des Schülers werden primär auf Leistungen Anderer, bzw. abstrakter auf eine „idealisierte Normentwicklung" bezogen, nicht auf seine eigene Entwicklung. Die oben skizzierte Motivation für den Entwurf solcher Verfahren als objektiver, valider Bewertungsmaßstab begründet ein solches Vorgehen zwingend, allerdings zeigen die hier skizzierten Problemfelder die extreme Schwierigkeit, wenn nicht Unmöglichkeit des Entwurfs eines solchen Maßstabs.

[1] Zur Bedeutung der „geschlechtstypischen Ausgewogenheit der Schreibwörter" s. z.B. May (1994). May argumentiert dort, dass Jungen „persönlich bedeutende Wörter" sorgfältiger, ergo häufiger korrekt schreiben, Mädchen hingegen auch bei persönlich weniger bedeutenden Wörter größere Sorgfalt an den Tag legen, weil sie die Autorität der Schule eher anerkennen.

3.2.2. Lernbeobachtungen

Insbesondere in Hinblick auf den vierten Kritikpunkt schlägt Mechthild Dehn (Dehn, 1994, S. 210ff.) eine systematische *Lernbeobachtung* als Beobachtung des Schrifterwerbsprozesses während des ersten Schuljahres vor. Ziel der Lernbeobachtung soll es sein, die Voraussetzung für das frühe Anbieten von Lernhilfen zu gewährleisten.

Die Bewertung der individuellen Entwicklung steht im Vordergrund, indem zu drei Zeitpunkten während des ersten Schuljahrs eine jeweils erweiterte, aber im Kern konstante Menge von Wörtern geschrieben wird. Zusätzlich sind zu erlesende Sätze vorgesehen, auf die im Folgenden nicht weiter eingegangen wird.[2]

Die Auswertung der Schreibungen geschieht nach einem dreiteiligen Raster, nach dem Schreibungen als „diffus“, „rudimentär“ oder „besser“ klassifiziert werden. Hinter diesem Raster steht der Grad der „Regelgeleitetheit“ als wachsende Größe:

- „Diffuse“ Schreibungen sind nicht regelgeleitet, der Schreibende benutzt Buchstaben nicht in einer Weise, die erkennen ließe, dass Buchstaben spezifische Funktionen haben, sondern lediglich aus der Erkenntnis heraus, dass Wörter überhaupt aus Buchstaben bestehen. Dehn nennt Schreibungen wie *<LBED> für <wo> oder *<ltL> für <Lampe> als Beispiele für diffuse Schreibungen (s. Dehn, 1994, S. 41).
- „Rudimentäre“ Schreibungen enthalten „erste wichtige Bausteine“. Im Gegensatz zu diffusen Schreibungen hat der ganz überwiegende Teil der Zeichen einen lautlichen Gegenwert, *<HDL> für <Hundeleine> ist ein Beispiel für rudimentäre Schreibungen (a.a.O. S. 42). An dieser Stelle werden auch silbische Kriterien genannt, wie die Beobachtung, dass die rudimentäre Schreibung *<LMND> für <Limonade> genau die Anfangsränder aller Silben repräsentiert. Es erfolgt allerdings keine Systematisierung oder gar eine prinzipielle Nutzung prosodischer Analysen.
- „Bessere“ Schreibungen sind solche, die einen großen Teil der für die korrekte Schreibung notwendigen Leistungen beinhalten. Der Übergang von „rudimentär“ zu „besser“ wird auch quantitativ definiert: „Wenn mehr als zwei Drittel der Buchstaben als regelgeleitet erkennbar sind, gilt die Schreibung als besser“ (Dehn, 1994, S. 229). „Besser“ umfasst also das breite Spektrum aller Schreibungen von „überwiegend korrekt“ bis „vollständig korrekt“. Dehn bietet eine genauere Differenzierung der „besseren“ Schreibungen in 5 Unterkategorien an, die durchaus auch in Analogie zu übliche Stufenmodellen der Schreibentwicklung (s.o.) gesehen werden können (vgl. Dehn, 1994, S. 230, 272):

    1. Unvollständig: Schreibungen, die nicht mehr als rudimentär anzusehen sind, denen aber trotzdem noch Elemente fehlen. Beispiel: *<PNSL> für <PINSEL>.

[2] Einen Vorschlag, wie die im folgenden herauszuarbeitenden Grundlagen für sinnvolle Analysen auch für die Verbesserung der Lesefahigkeit eingesetzt werden können, machen Chr. Röber-Siekmeyer und K. Pfisterer in Röber-Siekmeyer und Pfisterer (1998).

2. An der Artikulation orientiert: Schreibungen, die phonetisch korrekt wahrgenommene, aber für die Schrift nicht relevante Phänomene widerspiegeln (vgl. aber die differenziertere Analyse Dehns im Hauptteil (Dehn, 1994, S. 57), die z.B. auch dialektal bedingte Phänomene mit einbezieht).
   Beispiel: Aspiration in [hˈɛf.t^{h}ə] → *<Hefthe>, aber auch solche, die phonetische Ähnlichkeiten nicht im Sinne der für die deutsche Orthographie zu erlernenden Zusammenfassungen differenzieren: *<Hond> für <Hund>. Daneben gilt das Nichtbeherrschen orthographischer Konventionen für komplexe Grapheme (v.a. Diphthonge) und die graphische Repräsention von Reduktionssilben (*<FEDA> für <FEDER>) als „an der Artikulation orientiert".

3. Einsichten aus der Auseinandersetzung mit der Schrift falsch verallgemeindernd: Hierunter sind sämtliche falschen Verwendungen spezieller Markierungsmöglichkeiten der deutschen Orthographie zu fassen: *<HUNND>, *<WELD>, *<FRIEHDE>.

Mechthild Dehn greift zur Klassifikation der Schreibleistungen an einigen Stellen auf phonetische und prosodische Eigenschaften der tatsächlich gesprochenen Sprache der Kinder zurück, sieht darin aber nicht primäre Grundlagen für Untersuchung und unterrichtliches Aufgreifen von Leistungen und Defiziten. Vielmehr sieht sie den Schriftspracherwerb als einen gerichteten Lernprozess, und den Sinn der Analyse im Feststellen des Standorts auf einer linearen Achse.

3.2.3. Voraussetzungen für sinnvolle Analysen

Das Ziel, Lernhilfen anbieten zu können, setzt gemäß Kritikpunkt 3 (s.o.) die Möglichkeit voraus, aus den Testergebnissen Hypothesen über die Regelbildung des Schülers ableiten zu können und mit Annahmen über die Orthographie als systematisches, d.h. regelhaftes, erlernbares Gebilde zu kombinieren. Lernhilfen bestehen dann darin, den Schüler auf Verstöße gegen Regelmäßigkeiten in der Orthographie aufmerksam zu machen und ihm Wege anzubieten, diese Regelmäßigkeiten zukünftig zu berücksichtigen. Der Erfolg dieser Wege hängt entscheidend davon ab, ob dem Schüler damit ein Verfahren an die Hand gegeben wird, mit dem er sicher und nachvollziehbar zur korrekten Schreibung gelangen kann.

Die Forderung „hör genau hin, wie du das Wort sprichst" ist kein erfolgversprechender Weg, weil sie zirkulär ist. Es wird das Vorhandensein einer Fähigkeit vorausgesetzt, nämlich die Artikulation einer Wortform in einzelne Laute zu segmentieren, die wiederum einzelnen Graphemen entsprechen sollen. Schrifterfahrene Erwachsene sehen diese Fähigkeit als „natürlich" an, tatsächlich ist sie aber erst aus dem Erlernen der Schrift entstanden, d.h. die Lautung eines Wortes wird unwillkürlich durch die Brille der Schrift gesehen.

Beispiel: Aus wievielen Lauten besteht das Wort <Otto>?

Die Antwort ist nicht eindeutig, weil der Begriff „Laut" nicht eindeutig ist. Auf jeden Fall enthält das Wort zwei Vokale[3], und zwar zwei verschiedene: [ɔ] und [o], die aber beide durch dasselbe Schriftzeichen, das Vokalgraphem <o> repräsentiert werden. Diese Abbildung ist nicht selbstverständlich, sondern muss gelernt werden.

[3] Wichtig ist hier die genaue Unterscheidung zwischen *Vokalen* und *Vokalbuchstaben* bzw. *Vokalgraphemen*. Vokale sind Laute, also Elemente der gesprochenen Sprache, Vokalgrapheme sind Schriftzeichen, also Elemente der geschriebenen Sprache.

Schwieriger ist die Frage im Falle der Konsonanten (auch hier ist wieder streng zwischen *Konsonanten* als lautlichen und *Konsonantengraphemen* als schriftlichen Elementen zu unterscheiden). Keinesfalls haltbar ist die Aussage, das Wort enthielte zwei „t"-Laute – diese Annahme ist allein durch die Schrift hervorgerufen. Insbesondere ist es daher nicht möglich, das Kind durch den Hinweis auf „deutliches" und „langsames" Sprechen zur Schreibung der zwei Konsonanten*grapheme* <tt> zu bringen; es ist dadurch nicht in der Lage zu erschließen, dass die Lautung „ot – to" „korrekt" sei, „kant – te" oder „vat – ter" hingegen aber nicht. „Otto" enthält also nur einen Konsonanten [t], dagegen aber noch zwei weitere Konsonanten, die in der Schrift nicht repräsentiert werden. Zum einen ist dies der Glottisverschluss am Wortanfang, zum anderen ein [h]-Laut, der bei der Lösung des [t] entsteht. Eine phonetisch akzeptable Darstellung des Wortes wäre demnach z.B. [ʔɔt^{h}o]. Dieses (noch recht einfache) Beispiel zeigt schon, dass die Annahme „Ein Laut wird durch ein Graphem repräsentiert" nicht haltbar ist, sondern auf einem Konzept von „Laut" basiert, das nur solche Elemente gesprochener Sprache als „Laute" sieht, die auch in der Schrift Niederschlag finden – ein Zirkelschluss.

Sowohl die HSP als auch M. Dehns Lernbeobachtung benutzen zur Analyse von Schreibungen das Konzept der „Graphemtreffer". Hervorgegangen aus der Beobachtung, dass das bloße Auszählen vollständig korrekt geschriebener Wörter in Verhältnis zur Gesamtmenge zu ungenau ist, liegt hier ein einfaches Instrument vor, mit dem graduelle Aussagen über die Korrektheit einer Schreibung möglich werden. Der Algorithmus ist einfach:

1. Zerlege das Zielwort in eine Liste von Graphemen.
2. Zerlege das geschriebene Wort in eine Liste von Graphemen.
3. Richte die Graphemketten aneinander aus und zähle die Übereinstimmungen. Nichtübereinstimmungen können sein: Wahl eines falschen Graphems oder Auslassen eines Graphems. Das Einfügen überflüssiger Grapheme muss auf andere Weise behandelt werden.

Die Frage, was als Gesamtmenge der Grapheme anzunehmen ist, ist nicht trivial. In verschiedenen theoretischen Darstellungen wird das Graphemsystem zweigliedrig gesehen, Thomé (1998, S. 71f.) etwa unterscheidet Basis- und Orthographeme. *Basisgrapheme* sind die für ein Phonem „per default" stehenden, also bei Abwesenheit besonderer Gründe, ein anderes Graphem zu wählen. Zugleich werden sie über Häufigkeitszählungen von Phonem-Graphem-Entsprechungen begründet. *Orthographeme* sind dagegen zum einen seltenere Varianten (wie <v> für /f/ statt <f> als Basisgraphem), zum anderen durch Regeln (<d> für /t/ in <Hund>, <tt> in <Bett>) oder ausnahmshafte Einzelfälle (<ai> statt <ei> in <Mai>) begründbare Schreibungen.

Dieser Aufteilung folgt im Grunde die HSP, wenn sie drei „Strategien" ansetzt, die mit Hilfe der Graphemtrefferzählung analysiert werden sollen:

- Alphabetische Strategie: Hierunter ist die Fähigkeit zu verstehen, Phoneme, für die ihr jeweiliges Basisgraphem zu wählen ist, korrekt zu verschriften. Der für die „alphabetische Strategie" ermittelte Prozentwert soll angeben, in welchen Maße das Kind die „Grundlage" der deutschen Orthographie, nämlich die Defaultzuordnung bestimmter Laute zu ihren Basisgraphemen, verstanden hat und sicher anwenden kann.

- Orthographische Strategie: Die Fähigkeit, denjenigen Teil der Orthographeme, der nicht durch Abgleich mit verwandten Wortformen zu begründen ist, korrekt zu schreiben, wird als „orthographische Strategie“ bezeichnet. Ein Kind, dass diese Strategie beherrscht, soll in der Lage sein, Bedingungen zu erkennen, die die Wahl des Basisgraphems verbieten – seien sie regelhaft oder einfach als „der seltenere Fall“ (wie <v> statt <f> für /f/) begründet. Wichtig ist hier anzumerken, dass die Schärfungsmarkierung nicht, wie in der Literatur (vgl. Maas, 1992 und Eisenberg, 1998, davon beeinflusst auch Augst und Dehn, 1998) angenommen, durch silbenstrukturelle Bedingungen (fester Anschluss an heterosyllabischen Konsonanten (vgl. Maas, 1992; Maas, 1999) bzw. „Silbengelenk“ (vgl. Eisenberg, 1998)) begründet und per „Vererbung“ auch in damit verwandten Formen gesetzt werden muss, sondern entsprechend der „Duden-Regel“ (Duden, 1991, §2) gesehen wird, die keinen morphologischen Abgleich benötigt und somit jede Schärfungsschreibung an der betrachteten Wortform selbst erklären kann. Allerdings bezieht sich diese Regel auf die Kategorie „Wortstamm“ – eine morphologische Kategorie: Somit gehörten die Schärfungsschreibungen doch zur „morphematischen Strategie“.[4] Dehnungsschreibung und silbentrennendes <h> werden höchstens als teilweise regelhaft angenommen und zählen ebenfalls vollständig zur „orthographischen Strategie“.
- Morphematische Strategie: Für die korrekte Schreibung von Orthographemen, die nur über den Rückgriff auf verwandte Wortformen erklärt werden können, ist nach May die Beherrschung einer „morphematischen Strategie“ notwendig. Hierzu zählen vor allem die Phänomene *Auslautverhärtung* (<Hund> wg. <Hunde>, <Flug> wg. <Flüge> . . .) und *<ä>-Schreibung* (<Bälle> wg. <Ball>, <Räuber> wg. <Raub> . . .). Gleichzeitig zählt noch ein gänzlich anderes Phänomen zur „morphematischen Strategie“: Das Auftreten ansonsten nicht vorkommender Graphemfolgen an Kompositionsgrenzen wie <pfpf> in <Topfpflanze> oder <stst> in <Miststall>. Die Frage, inwiefern diese Kategorie phonetisch begründet ist (wie z.B. die Geminatenreduktion in [ʔan.ˈneː.mən] → [ʔa.ˈneː.mən] (vgl. Kohler, 1995)) ließ sich in den Arbeiten Mays nicht feststellen, wie generell anzumerken ist, dass keine detaillierten Ausführungen zur Begründung der Zuordnungen von orthographischen Phänomenen zu Strategien gemacht werden.

Daneben enthält die HSP noch die Kategorien „überflüssige Elemente“ und „Oberzeichenfehler“. Erstere meint überflüssige Grapheme, aber auch die Verwendung orthographischer Markierungen an solchen Stellen, wo sie nicht nötig ist; letzteres bezeichnet eine Reihe von Performanzfehlern beim Schreiben: gedrehte oder nicht existierende Buchstaben, vertauschte Buchstaben usw.

Für all diese Kategorien werden Graphemtreffer gezählt, d.h. es scheint immer möglich, alle relevanten Bereiche direkt auf ein einzelnes Graphem zu beziehen, die Annahme übergeordneter Strukturen ist nicht nötig.

Dieser Sicht ist zu widersprechen:

- Aus linguistischer Sicht ist die Fundierung der Orthographie in prosodischen Strukturen sehr fruchtbar gewesen. Sie ist in der Lage, eine Reihe von Phänomenen einfach und regelhaft zu beschreiben, wohingegen ein rein segmental basierter oder an Wortstämmen ausgerichteter Ansatz mit einer sehr viel größeren Menge an Regeln und Ausnahmen operieren muss und dadurch den Blick auf das System der deutschen Orthographie verstellt.

[4] Für einen Vergleich der beiden Regelhypothesen zur Schärfungsschreibung s. Maas et al. (1999).

- Aus Sicht der Schriftspracherwerbsforschung ist festzustellen, dass die Fähigkeit, Wörter (oder allgemeiner: sprechsprachliche Einheiten) in Phoneme zu zergliedern, nicht von vornherein vorhanden, sondern eine Folge des Schriftspracherwerbs ist – ergo ist es nicht möglich, Kindern mit Problemen beim Zugang zur Schrift genau dieses Zergliedern als „Hilfe“ anzubieten. Einen sehr viel intuitiveren und direkteren Zugang haben Kinder im Einschulungsalter zu prosodischen Domänen, insbesondere zu metrischen. Aus der dominierenden metrischen Kategorie, dem Fuß, lässt sich recht einfach der Begriff „Silbe“ begründen. Schon hier ist aber zu beachten, dass Silben immer im Zusammenhang mit größeren metrischen Strukturen zu sehen sind und deshalb eine qualitative Unterscheidung von Silbentypen[5] nicht Ziel, sondern Ausgangspunkt der Überlegungen ist.

[5] Maas (1992, 1999) unterscheidet drei Silbentypen zum Ersten nach ihrer Position im Fuß und damit ihrer Akzentuierbarkeit (prominent oder nicht-prominent) und zum Anderen nach ihrer Belegung im Kern: /ə/ oder silbischer Sonorant für die reduzierte, Vollvokale für die nicht-reduziuerte Silbe.

4. Entwurf eines Auswertungsschemas

4.1. Motivation

Aus den oben diskutierten Unzulänglichkeiten etablierter Auswertungsschemata für orthographische Leistungen erwächst die Notwendigkeit, ein weiteres, neues Schema zu konstruieren. In Abschnitt 4.2 werden zunächst die Ansprüche definiert, die das Schema und die darauf aufbauenden Verfahren erfüllen sollen. Abschnitt 4.3 stellt dann im Einzelnen die verwendeten Kategorien sowie ihre linguistische Fundierung vor und legt den Rahmen für die nachfolgend zu entwerfenden automatisierten Verfahren fest. Zur vertiefenden Veranschaulichung und Darstellung der grundsätzlichen Leistungsfähigkeit werden in Abschnitt 4.6 einige Beispiele angeführt und einer manuellen Auswertung unterzogen.

4.2. Zielsetzungen

Das Verfahren sollte folgende Kriterien erfülen:

- Klare und objektive Definition der Auswertungskategorien. Dies ist durch eine linguistische Fundierung möglich.
- Ableitbarkeit differenzierter qualitativer Analysen orthographischer Leistungen im Einzelfall, d.h. sowohl von Fehlschreibungen als auch korrekten Schreibungen.
- Aggregationsmöglichkeit sowohl für Gruppen von Schreibenden als auch für zeitlich verteilte Leistungen Einzelner.

4.3. Auswertungskategorien

Mit einem Blick auf die Systematik der deutschen Orthographie (s. Kapitel 2) ergeben sich drei Bereiche:

- Silben und Grapheme: Ein grundlegender Bereich, der auf das erwähnte Konzept der Basisgrapheme abgebildet werden kann. Dabei ist hier allerdings die Position eines Phonems bezogen auf die Silbe entscheidend, d.h. das <s> für /ʃ/ (statt <sch>) an bestimmten Positionen, nämlich vor Plosiven im Anfangsrand, als „normale“ Entsprechung gesehen werden kann.
- Phonologische Markierungen: Ein über die Silbe bzw. einzelne Konstituenten der Silbe hinausgehender Bereich, der sich auf den Anschluss zwischen Vokal in prominenter Silbe und nachfolgendem Konsonanten, sei er in derselben Silbe (tautosyllabisch) oder nicht (heterosyllabisch), bezieht. Dieser Bereich entspricht in Teilen dem, was z.B. in der HSP als „orthographische Strategie“ bezeichnet wird, beschränkt sich jedoch auf eine exakt definierte Menge von Phänomenen, die eine linguistisch begründete Gemeinsamkeit aufweisen und im Zweifelsfalle nicht nur statistisch motiviert werden können.
- Morphologische Konstantschreibung: Darüber hinaus ist ein Bereich notwendig, der die möglichst konstante Schreibung von Morphemen über die gesamte Wortfamilie betrifft. Damit ist ein Teil der HSP-Strategie „morphematisch“ abgedeckt, allerdings in systematischerer Weise: Durch die Verknüpfung der Schärfungs- und Dehnungsschreibung und des silbentrennenden <h> an silbenstrukturelle Bedingungen ist in Fällen, die die spezielle Bedingung nicht aufweisen, wohl aber eine verwandte Form, auch eine Markierung vorzunehmen – ein Fall von Konstantschreibung.

4.3.1. Aufbau eines tabellarischen Schemas

Es soll nun ein Auswertungsschema entworfen werden, das Schreibleistungen hinsichtlich dieser drei Bereiche analysierbar werden lässt. Jeder der Bereiche subsummiert eine Menge von Regeln, die Abbildungen zwischen der orthographischen Ebene und anderen sprachlichen Ebenen beschreiben. Diesen Regeln wird hier zunächst kein kognitiver Status zugeschrieben – das Ergebnis der Analyse ist also nicht eine direkte Aussage über Intentionen und Modellbildungen der Schreiber. Diese ergeben sich aus der linguistischen Analyse der deutschen Orthographie.

Die allgemeine Form einer solchen Regel ist:

> Wenn Vorbedingungen $v_1 \dots v_n$ gelten, dann gelten Abbildungen $a_1 \dots a_m$.

Phänomenbereich	max	korrekt
Silben		
Phonologische Markierungen		
Morphologische Konstantschreibung		

Tabelle 4.1.: Zweidimensionale Tabelle als Grundlage des Auswertungsverfahrens

Für die korrekte Schreibung eines Wortes sind nach dieser Vorgehensweise eine Reihe von Regeln *relevant*, die jeweils *korrekt angewendet* werden müssen. Die Relevanz einer Regel ergibt sich aus dem Vorkommen oder Nichtvorkommen der Phänomene, die in den Vorbedingungen enthalten sind. Eine Regel, die sich z.B. auf die orthographische Repräsentation von silbischen Sonoranten in Reduktionssilben bezieht, ist nur für Wortformen relevant, die eine Reduktionssilbe mit silbischen Sonoranten aufweisen.

Gruppen von Regeln mit gleichen Vorbedingungen oder solchen Vorbedingungen, die gegenseitig als Verallgemeinerungen oder Spezialisierungen betrachtet werden können, entsprechen damit Regeln, die sich auf den gleichen Phänomenbereich beziehen. Eine erste Strukturierung der Auswertung lässt sich damit nach den oben genannten Phänomenbereichen vornehmen:

1. Silben und Grapheme
2. Phonologische Markierungen
3. Morphologische Konstantschreibung

Resultat des Auswertungsprozesses sollen klare Aussagen über Beherrschung und Nichtbeherrschung einzelner Regeln – und in der Zusammenfassung – Phänomenbereichen sein. Einfache Aussagen über die Häufigkeit von korrekten, falschen oder nicht erfolgten Regelanwendungen sind nicht ausreichend, weil sie nicht in Relation gesetzt werden können zu anderen Regeln und Leistungen. Deshalb ist es vorteilhaft, die Beherrschung einer Regel als Verhältnis der korrekten Anwendungen zu allen Fällen zu sehen, in denen die Vorbedingungen der Regel zutreffen. In der Zusammenfassung zu Phänomenbereichen gibt das Verfahren dann an, welcher Anteil der jeweils zugehörigen Phänomene mit den dafür angenommenen Regeln korrekt verschriftet wurden.

Es ergibt sich wie in Tabelle 4.1 dargestellt ein zweidimensionales Schema, in dem als Zeilen die Phänomenbereiche und als Spalten die Anzahl der Vorkommen (bezeichnet mit `max`) und die Anzahl der korrekten Regelanwendungen aufgetragen sind.

Eine weitere Ausdifferenzierung der Tabelle erfolgt durch eine Spezialisierung der Phänomenbereiche. Dadurch ergibt sich eine Hierarchie von Auswertungskategorien, die von oben nach unten betrachtet als sich erweiternde Differenzierung von Phänomenblöcken gesehen werden kann. Damit ist es möglich, bei der Interpretation des Auswertungsergebnisses zunächst auf oberen Ebenen besondere Problembereiche zu identifizieren und dann den Unterkategorien detaillierte Informationen über Beherrschung und Nichtbeherrschung von Teilbereichen zu entnehmen. Umgekehrt bedeutet dies für die Erstellung der Auswertung, dass sich die Resultate in Oberkategorien als Summe der Einzelergebnisse in den Unterkategorien ergeben.

4.3.2. Phänomenbereich „Silben“

Für die Untergliederung des Phänomenbereiches „Silben“ bieten sich zwei Dimensionen an. Zunächst wäre es möglich, die Schreibleistungen nach der korrekten Wiedergabe der drei Silbentypen zu unterscheiden. Dies ist insofern sinnvoll, als es für jeden der drei Typen spezifische Phänomene oder Belegungsbeschränkungen gibt, die eine eindeutige Unterscheidung möglich machen. So ist es notwendig, bei der Repräsentation von Silbenkernen nach den drei Typen zu unterscheiden, da sich hier die größten Unterschiede zeigen: Ein nicht repräsentierter Kern einer Reduktionssilbe ist ein Fehler deutlich anderen Typs als die Nichtrepräsentation eines Kerns einer prominenten Silbe.

Hinsichtlich der Anfangs- und Endränder ergeben sich aber geringere Unterschiede, wenn als Vereinfachung für dieses Verfahren angenommen wird, dass alle vokalischen Elemente Teil des Nukleus sind und Anfangs- sowie Endränder ausschließlich aus konsonantischem Material bestehen. Diese Konstruktion ist unter dem Begriff des „verzweigenden Nukleus“ verbreitet. Hier unterscheiden sich die Silbentypen zwar deutlich in den zulässigen bzw. auftretenden Belegungen der Ränder, diese sind aber nicht prinzipiell anderer Art.

Deshalb wird hier der Phänomenbereich „Silbe“ zunächst nach den drei Silbenkonstituenten Anfangsrand, Nukleus und Endrand unterteilt. Innerhalb dieser Unterkategorien wird dann – soweit notwendig – nach Silbentypen unterschieden.

Silbenkerne

Wie beschrieben ist für Silbenkerne der Silbentyp von entscheidender Bedeutung. Der Phänomenbereich „Silbenkern“ wird demnach in drei Bereiche unterteilt: S’ (prominente Silben), S^0 (Reduktionssilben) und S (nicht-prominente, nicht-reduzierte Silben).

Kerne prominenter Silben In prominenten Silben ist der Anschlusstyp zwischen vokalischem Element und nachfolgendem konsonantischen Material ein wesentliches Unterscheidungsmerkmal. Deshalb liegt eine erste Aufteilung des Bereichs in festen und losen Anschluss nahe. Fälle festen Anschlusses sind zunächst nicht weiter zu unterteilen, da die Fälle, die zu unterschiedlichen orthographischen Markierungen führen, hier nicht als Eigenschaft des Silbenkernes, sondern unter der Sonderrubrik „phonographische Markierungen“ geführt werden. Bei losem Anschluss hingegen ist eine Unterscheidung in vier Unterfälle sinnvoll:

1. Monophthonge. Hier sind zunächst als „unproblematische Fälle“ die Monophthonge /aː/, /eː/, /oː/, /uː/, /øː/, /yː/ zusammengefasst. Eine spezielle Behandlung benötigen evtl. die Phoneme /iː/ und /ɛː/. Der lose angeschlossene Monophthong /iː/ hat orthographisch zwei häufige Entsprechungen, <ie> und <i>. Auch wenn <ie> als Standaradentsprechung angesehen werden kann, ist die Entsprechung <i> so häufig, dass der Gesamtkomplex eine wichtige Fehlerquelle für Schreibanfänger darstellt. Das Phonem /ɛː/ wird in Teilen des deutschen Sprachraumes als [eː] realisiert, so dass lexikalisches Wissen oder ein morphologischer Abgleich wie bei <ä> als Entsprechung eines Silbenkernes mit festen Anschluss notwendig wird.
2. Diphthonge, das sind /aʊ/, /aɪ/ und /ɔʏ/. Die Einsprechung /au/ → <au> ist unproblematisch, eine genauere Unterscheidung kann für /aɪ/ → <ai> / <ei> getroffen werden. Der Fall /ɔʏ/ → <eu> / <äu> wird in der Realisierung <äu> durch eine Unterkategorie des Bereiches „Morphologische Konstantschreibung“ abgedeckt.

3. Öffnende Diphthonge, also Kombinationen aus Vokal und vokalisiertem /r/. Ob öffnende Diphthonge als eigene Kategorie im Auswertungsschema Berücksichtigung finden sollten, hängt von der Art der zugrundegelegten phonologischen bzw. phonetischen Repräsentation ab. Eine phonologische und damit schriftnähere Repräsentation könnte /ɐ/-Phoneme grundsätzlich als Konsonanten analysieren und dem Endrand zuschlagen. Eine stärker phonetisch orientierte Repräsentation würde die Vokalisierung mitberücksichtigen und daher das vokalisierte /r/ mit zum Kern rechnen. Für die Erklärung orthographischer Leistungen ist ein Rückgriff auf phonetische Realisierungen statt auf eher abstrakte phonologische Repräsentationen häufig von Vorteil, allerdings ist dann zu berücksichtigen, in welcher Art und Weise der untersuchte Schreiber die angenommene phonetische Realisierung tatsächlich vornimmt.
4. /aː/ → <ar>. Wird mit öffnenden Diphthongen von einer Repräsentation des vokalisierten /r/ als vokalisches Element ausgegangen, sollten in einem nächsten Schritt Formen mit /ar/ getrennt behandelt werden, da ausgehend von [a] keine weitere Öffnung mehr möglich ist und es zu einer Realisierung als [aːɐ] kommt.

Kerne reduzierter Silben Reduktionssilben zeichnen sich orthographisch durch das Auftreten eines <e> als Kernmarkierung aus. Diese Markierung ist aber nicht als alleinige oder direkte Entsprechung eines Schwa-Vokals im Kern der Reduktionssilbe zu sehen. Stattdessen sind drei Fälle reduzierter Silben zu unterscheiden. Eine getrennte Betrachtung von Nukleus und Endrand ist für Reduktionssilben nicht sinnvoll, so dass die nachfolgenden Fälle als Belegungen des Reimes gesehen werden müssen.

- /ə/ – In diesen Fällen ist die orthographische Repräsentation des Reimes <e>.
- /ɐ/ – In diesen Fällen ist die orthographische Repräsentation des Reimes <er>, eine Differenzierung gegenüber /ə/ ist sinnvoll, da sich in beiden Fallgruppen bei Schreibanfängern unterschiedliche Leistungen zeigen, bzw. sich beide Fallgruppen unterschiedlich entwickeln.
- Silbischer Sonorant - ein silbischer Sonorant als Reim einer Reduktionssilbe erhält in seiner orthographischen Repräsentation dennoch ein <e>. Damit ergibt sich ein von den beiden vorhergehenden klar unterscheidbarer Phänomenbereich. Die Annahme silbischer Sonoranten als Reim einer Reduktionssilbe setzt wiederum eine eher phonetische Analyse der Wortform voraus.

Kerne nicht-prominenter, nicht-reduzierter Silben Die Kerne nicht-prominenter, nicht-reduzierter Silben unterliegen keiner Alternation in den Anschlussverhältnissen. Eine detailliertere Unterscheidung ist deshalb nicht notwendig, auch Subregularitären wie /iː/ → <ie>/<i> sind nicht zu beobachten.

Für Silbenkerne ergibt sich somit insgesamt folgende Strukturierung:

S’
- fester Anschluss
- loser Anschluss
  - Monophthong
  - Diphthong
  - Öffnender Diphthong
  - /aː/ → <ar>

S^0
- /ə/
- /ɐ/
- silbischer Sonorant

S

Anfangsränder

Anfangs- und Endränder bieten sich aus verschiedenen Gründen für eine getrennte Behandlung und Analyse an. Für sie gelten grundsätzlich unterschiedliche phonotaktische Beschränkungen und sie sind Domänen unterschiedlicher phonologischer Prozesse wie z.B. der Auslautverhärtung. Für den Aufbau des Analyseschemas stellt sich nun die Frage, ob es sinnvoll ist, Anfangs- bzw. Endränder zuerst (oder überhaupt) nach Silbentypen zu unterscheiden.

Für Anfangsränder ist folgendes zu beobachten:

- In prominenten Silben sind komplexe Anfangsränder möglich, der Anfangsrand muss belegt sein, erhält aber im Defautlfall (Glottisverschluss) eine leere orthographische Repräsentation, die nicht von der Repräsentation leerer Anfangsränder zu unterscheiden ist.
- In reduzierten Silben ist der Anfangsrand maximal einfach belegt, er kann auch leer bleiben. In diesem Fall können die Bedingungen für eine Markierung durch ein silbentrennendes <h> gegeben sein, die jedoch unter dem Phänomenbereich „phonologische Markierungen“ behandelt wird.
- In nicht-prominenten, nicht-reduzierten Silben sind komplexe Anfangsränder möglich, aber in geringerer Variationsbreite und seltener als bei prominenten Silben. Der Anfangsrand kann auch leer bleiben, allerdings dann oft in Alternation mit einer Default-Belegung durch Glottisverschluss.

Für alle drei Typen bleibt zusammenfassend festzuhalten: Leere Anfangsränder unterscheiden sich in ihrer (korrekten) orthographischen Repräsentation nicht von Anfangsrändern mit Default-Belegung durch einen Glottisverschluss. Aus Sicht einer Analyse, die die Repräsentation oder Nichtrepräsentation von Silbenkonstituenten in den Vordergrund stellt, müssen daher keine verschiedenen Typen von Aussagen – abhängig davon, ob ein Silbentyp leere Ränder zulässt oder nicht – möglich sein.

Alle drei Typen lassen einfache Anfangsränder zu, zwei der drei Typen auch komplexe. Grundsätzlich zwischen einfachen und komplexen Rändern zu unterscheiden ist sinnvoll, weil für komplexere Anfangsränder z.T. spezielle orthographische Subregeln gelten und zudem andere Fehlerhäufigkeiten und Verteilungen zu beobachten sind.

Endränder

Das Phänomen des losen oder festen Anschlusses wird hier nicht als Eigenschaft des Endrandes betrachtet, da die Fälle, die zu einer orthographischen Sondermarkierung führen, unter dem Phänomenbereich „phonologische Markierungen“ behandelt werden. Die Hauptunterschiede in der orthographischen Repräsentation ergeben sich abgesehen davon bei der Repräsentation des vokalischen Materials (s.o.). Ähnliches gilt für die Auslautverhärtung, die sich ebenfalls in der Orthographie niederschlägt, aber hier auch als gesonderter Bereich unter „morphologische Konstantschreibung“ behandelt wird.

Es bleibt die Unterscheidung zwischen einfachen und komplexen Endrändern, die analog zu der Zusammenfassung der drei Silbentypen bei den Anfangsrändern begründet werden kann.

4.3.3. Phänomenbereich „Phonologische Markierungen“

Phonologische Markierungen – hier sind nur die direkt silbenstrukturell begründbaren Phänomene zu erfassen:

- Schärfung
- Dehnung
- Silbentrennendes h

4.3.4. Phänomenbereich „Morphologische Konstantschreibung“

Morphologische Konstantschreibung – hier erfolgt zunächst keine Differenzierung nach der „Entfernung“ des zu findenden Abgleichs (Flexion oder Wortbildung):

- Auslautverhärtung
- Umlautschreibung – hier sind sämtliche <ä>-Schreibungen zu verzeichnen
- Schärfung
- Dehnung
- Silbentrennendes h

4.3.5. Sonderbereiche

Sonstiges – hier werden alle Normabweichungen gesammelt, für die keine klare Hypothesenbildung möglich ist:

- Vertauschungen – von den anderen Kategorien nicht abgedeckte Verwechslungen einzelner Grapheme, aber auch Vertauschungen der Reihenfolge.
- Einfügungen – überflüssige Grapheme, für die keine plausible phontische Entsprechungvorhanden ist und die nicht durch Übergeneralisierungen erklärt werden können.
- Regelwidrige Markierungen – Übergeneralisierungen der regelhaften phonologischen oder morphologischen Markierungen, aber auch willkürliche Verwendung solcher Markierungen an nicht möglichen Positionen.

4.4. Grade von Korrektheit

Für eine differenzierte Auswertung ist es sinnvoll, die bislang eingeführten Tabellenspalten `max` – die gesamten in der Testwortmenge enthaltenen Fälle jedes einzelnen Phänomens – und korrekt – der Anzahl der tatsächlich orthographisch korrekt repräsentierten Fälle – zu erweitern. Neben der vollständigen orthographischen Korrektheit werden noch zwei weitere Abstufungen von `korrekt` notiert.

Der Abstufungsgrad `repräsentiert` klärt die Frage, ob eine prosodische Einheit oder ein orthographisches Phänomen überhaupt in der Schreibung Niederschlag findet, egal ob orthographisch korrekt oder nicht, ob phonetisch plausibel oder nicht.

Der Grad `phonetisch plausibel` gibt an, ob eine Einheit oder ein Phänomen in einer solchen Weise in der Schreibung vorhanden ist, dass ihre/seine Repräsentation als phonetisch plausibel angesehen werden kann. Phonetische Plausibilität in diesem Sinne beinhaltet fünf Problemfelder:

1. Vokalismus: ungespannte Vokale ([ɛ], [ɪ], [ɔ], [ʊ], [ʏ]) und Reduktionsvokale ([ə], [ɐ]) werden bei abweichender Akzentuierung der Wortform häufig mit einem Graphem für phonetisch ähnliche (gespannte) Vollvokale repräsentiert. Das geschieht insbesondere bei einer Artikulation, in der jede Silbe als prominente Silbe mit gespanntem Vokal realisiert wird („Pilotsprache“, „Dehnsprechen“). Beispiele:

   [hʊnt] . . . [hoːnt] → <h**o**nt>

   [heɪ̯.fə] . . . [heɪ̯.fɛː] → <hefä> oder [heɪ̯.fə] . . . [heɪ̯.fœː] → <hefö>

2. Konsonantismus: Insbesondere das Merkmal [±stimmhaft] ist für Konsonanten in einigen Silbenpositionen nicht bestimmbar, sondern wird in der Wahrnehmung durch das Merkmal [±aspiriert] ersetzt. Dadurch werden nicht-aspirierte Plosive an bestimmten Positionen als stimmhaft wahrgenommen. Dazu kommen orthographische Sonderfälle wie die Repräsentation von silbeninitialem /ʃt/ oder /ʃp/ durch <st> und <sp> statt *<scht> bzw. *<schp>, /f/–/v/-Schreibung oder /s/-Schreibung im Silbenendrand.

   Beispiele:

   [hɛf.t^{h}ə] ... [hɛf.tə] ... [hɛf.t̥ə] ... [hɛf.də] → <Hefde>

   [ʃt^{h}aɪn] ... [ʃtaɪn] ... [ʃt̥aɪn] ... [ʃdaɪn] ⟶ <Schdein>

3. Morphologische Konstantschreibung: Die orthographische Markierung von Auslautverhärtung und Umlautschreibung (<ä> bzw. <äu> für /ɛ/, /eː/ bzw. /ɔʏ/, wenn eine verwandte Wortform an korrespondierender Stelle /a/ bzw. /aʊ/ aufweist) ist phonetisch nicht zu motivieren. Daher sind die Schreibungen /ɛ/,/eː/ → <e> und /ɔʏ/ → <eu> sowie nicht markierte Auslautverhärtung in allen Fällen als „phonetisch plausibel“ zu klassifizieren.

4. Dialektale Variationen: Für den Entwurf des Auswertungssschemas wurde zunächst einmal von norddeutschen Sprechern ausgegangen, die Probleme verschieben sich jedoch, bzw. es können andere entstehen, wenn Sprecher aus anderen Dialektregionen oder Sprecher nichtdeutscher Familiensprachen untersucht werden. Eine genaue Untersuchung und Einbeziehung solcher Phänomene steht noch aus.

5. Sprechregister: Die oft postulierte „Phonem-Graphem-Korrespondenz“ orientiert sich an einer idealisierten, in tatsächlich gesprochener Sprache nicht auftretenden Explizitlautung. Abhängig vom Sprechregister (vgl. Maas, 1999) treten eine Reihe von Reduktionsmechanismen auf, die durch Assimilation und Elision die lautliche und prosodische Struktur einer Äußerung verändern (für eine Zusammenfassung dieser Mechanismen s. Kohler, 1995). Eine Schreibung wie *<Lebm> für /lˈeː.bən/ → /lˈeː.bn̩/ ist als Folge solcher Reduktionen gut zu erklären.

4.5. Auswertungsschema

In Tabelle 4.2 ist die (fast) vollständige Menge aller Auswertungskategorien aufgeführt, es fehlen lediglich einige Sonderprobleme, die sich auf Einzellaute beziehen (/s/ im Endrand, /f/-/v/-Probleme, /iː/-Schreibung) und lexikalisierte Schreibungen für „Funktionswörter“ (<und>, <ab>, <ihn>, <denn> usw.), die je nach Untersuchung hinzugefügt werden müssen.

Phänomen	**max**	**repräsentiert**	**phonetisch plausibel**	**orthographisch korrekt**
Silben				
Vollsilben				
Reduktionssilben				
Silbenkerne				
S’				
S’ (fester Anschluss)				
S’ (loser Anschluss)				
Langvokal				
Diphthong				
Öffnender Diphthong				
/aː/ → <ar>				
S^0				
S^0 (/ə/)				
S^0 (/ɐ/)				
S^0 (silbischer Sonorant)				
S				
Anfangsränder				
einfach				
komplex				
Endränder				
einfach				
komplex				
Phonologische Markierungen			—	
Schärfung			—	
Dehnung			—	
Sibentrennendes <h>			—	
Konstantschreibung				
Auslautverhärtung				
Schärfung				
Umlautgraphie				
Sibentrennendes <h>				
Sonstiges				
Vertauschungen	—		—	—
Einfügungen	—			—

Tabelle 4.2.: Vollständige Übersicht der Auswertungskategorien

4.6. Anwendungsbeispiel

Im folgenden Abschnitt soll das hier entworfene Verfahren an zwei Beispielen mit Auswertungen nach der HSP und dem Ansatz von Dehn (s.o.) verglichen werden. Dabei erkannte Stärken und Schwächen werden in Abschnitt 4.7 abschließend zusammengefasst. Es wird kein Beispiel aus dem Testkorpus verwendet, das in Kapitel 5 vorgestellt wird. Sowohl die HSP als auch das Auswertungsverfahren nach Dehn beziehen sich eher auf isolierte Schreibungen weniger einzelner Wörter. In ihren Darstellungen wird demonstriert, wie sich aus gleichartigen Erhebungen zu verschiedenen Zeitpunkten Entwicklungsverläufe analysieren lassen. Für das hier vorgeschlagene Verfahren soll zunächst im Vergleich an ähnlich geartetem Testmaterial demonstriert werden, dass es für vergleichende Analysen von Einzelwörtern geeignet ist, bevor es auf das gesamte Testkorpus angewendet wird.

4.6.1. Durchführung der Untersuchung

Wort	**2. Dezember**	**2. Februar**
Hund	HONT	HONT
Feder	FEDA	FEDER
Brillen	BRLÄN	BRLEN
Pinsel	PÖNSEL	PNSEL
Hefte	HFTÄ	HEFTE
Griaffe	GAFÄ	GERAFE
Reiter	—	REITER
Gürtel	—	GÜRTEL
Kirby	KABI	—
Bomberman	—	BOMBENHEN

Tabelle 4.3.: Jeromes Schreibungen der Testwörter

Wort	**2. Dezember**	**2. Februar**
Hund	HUNT	HUNT
Feder	FDR	FEDA
Brillen	BNL	BLN
Pinsel	PSU	PESL
Hefte	ED	HFD
Griaffe	GRF	GRAV
Reiter	—	RATA
Gürtel	—	GÜTL
Pizza	PTZZA	—

Tabelle 4.4.: Nadines Schreibungen der Testwörter

Jerome und Nadine, zwei Erstklässlern aus Nordwestdeutschland, wurden Anfang Dezember und Anfang Februar der ersten Klasse jeweils eine Liste von Wörtern vorgegeben, die sie während der

Unterrichtszeit in einem abgetrennten Raum niederschreiben sollten. Ihnen wurde zuvor erklärt, dass es nicht darum ginge, alles „richtig“ zu machen, sondern darum, festzustellen, wie sie für sich diese Wörter schreiben würden.[1]

Die Wörter wurden durch Schwarz-Weiß-Zeichnungen symbolisiert und der einzige Eingriff der den Test durchführenden Person bestand darin, bei Unklarheiten auf das „gemeinte“ Wort hinzuführen.

Tabelle 4.3 führt die zu schreibenden Wörter und die von Jerome vorgenommenen Schreibungen auf, analog Tabelle 4.4 für Nadine.

Die zweite Untersuchungsreihe umfasste zwei zusätzliche Wörter (Reiter, Gürtel), außerdem wurden Jerome und Nadine jeweils aufgefordert, ihr derzeitiges „Lieblingswort“ aufzuschreiben. Jerome wählte beide Male Begriffe aus seiner außerschulischen Erfahrungswelt (Computerspiele), <Kirby> (['kœɐ.bi]) und <Bomberman> ([ˌbom.bɐ.'mɛːn], die Transkriptionen geben seine eigene Artikulation wieder). Nadine hat nur beim ersten Mal ein Lieblingswort geschrieben, nämlich *<PTZZA> für <Pizza>.

4.6.2. Subjektiver Eindruck

Jerome schreibt bei beiden Erhebungen sehr sicher – sicher in dem Sinne, dass er sehr zielgerichtet und selbstbewusst an die Aufgaben herangeht. Er ist jeweils nach sehr kurzer Zeit (weit weniger als 5 Minuten) mit allen Wörtern fertig, lässt sich während des Schreibens nicht ablenken, gibt aber nur sehr zögerlich und wenig Auskunft über die Motive seiner Schreibungen.

Zuerst fällt auf, das er beim zweiten Mal alle Reduktionssilben richtig niedergeschrieben hat, beim ersten Mal war es nur eine. Ansonsten fällt auf, dass die Wörter vollständiger sind, und dass alle drei Silben des Wortes Giraffe beim zweiten mal repräsentiert waren, bei der ersten nur zwei Silben.

Nadine hat an beiden Terminen etwa 10 Minuten für die Schreibungen gebraucht. Dabei hat sie häufig laut gedacht und sich die Wörter laut vorgesprochen. Beim Vorsprechen der Wörter hat sie diese sehr gedehnt, teilweise auf einzelne Silben reduziert und den Anlaut stark hervorgehoben. Auch kam es vor, das ihr ein Buchstabe, der in die Mitte eines Wortes gehörte, erst zum Schluss einfiel, so dass sie ihn hinten angehängt hat.

Die Schreibleistung betreffend fällt bei Nadine auf, dass die Wörter tendenziell länger geworden sind. Zwar ist noch keines vollständig korrekt geschrieben, allerdings schreibt sie bei der zweiten Erhebung mehr Vokalbuchstaben.

Schreibung	Zielwort	Wort	Grapheme	A	O	M	Überfl.	Oz-Fehler
HONT	Hund	0	2/4	2/3		0/1		
FEDA	Feder	0	3/5	3/5				
BRLÄN	Brillen	0	3/6	3/5	0/1			
PÖNSL	Pinsel	0	5/6	5/6				
GAFÄ	Giraffe	0	2/6	2/5	0/1			
HFTÄ	Hefte	0	3/5	3/5				
	Summe	0/6	18/32	18/29	0/2	0/1		
		0%	56%	62%	0%	0%		

Tabelle 4.5.: HSP-Analyse für Jerome, Dezember

Schreibung	Zielwort	Wort	Grapheme	A	O	M	Überfl.	Oz-Fehler
HONT	Hund	0	2/4	2/3		0/1		
FEDER	Feder	1	5/5	5/5				
BRLEN	Brillen	0	4/6	4/5	0/1			
PNSEL	Pinsel	0	5/6	5/6				
GERAFE	Giraffe	0	4/6	4/5	0/1			
HEFTE	Hefte	1	5/5	5/5				
REITER	Reiter	1	5/5	5/5				
GÜRTEL	Gürtel	1	6/6	6/6				
	Summe	4/8	36/42	36/40	0/2	0/1		
		50%	86%	90%	0%	0%		

Tabelle 4.6.: HSP-Analyse für Jerome, Februar

Schreibung	Zielwort	Wort	Grapheme	A	O	M	Überfl.	Oz-Fehler
HUNT	Hund	0	3/4	3/3		0/1		
FEDR	Feder	0	4/5	4/5				
BLN	Brillen	0	2/6	2/5	0/1			
PSU	Pinsel	0	2/6	2/6				
GRF	Giraffe	0	2/6	2/5	0/1			
ED	Hefte	0	1/5	1/5				
	Summe	0/6	14/32	14/29	0/2	0/1		
		0%	44%	48%	0%	0%		

Tabelle 4.7.: HSP-Analyse für Nadine, Dezember

Schreibung	Zielwort	Wort	Grapheme	A	O	M	Überfl.	Oz-Fehler
HUNT	Hund	0	3/4	3/3		0/1		
FEDA	Feder	0	3/5	3/5				
BLN	Brillen	0	2/6	2/5	0/1			
PESL	Pinsel	0	3/6	3/6				
GRAV	Giraffe	0	3/6	3/5	0/1			
HFD	Hefte	0	2/5	2/5				
RATA	Reiter	0	2/5	2/5				
GÜTL	Gürtel	0	4/6	4/6				
	Summe	0/8	22/42	22/40	0/2	0/1		
		0%	52%	55%	0%	0%		

Tabelle 4.8.: HSP-Analyse für Nadine, Februar

4.6.3. Auswertung nach der HSP

Da May selbst vorschlägt, auch freie Schreibungen nach HSP-Kategorien zu analysieren (vgl. May, 1998), werden die vorliegenden Schreibungen auch mit dem HSP-Schema analysiert. In diesem Verfahren werden die Graphemtreffer des Kindes gezählt und dann noch einmal in drei Strategien (alphabetisch, orthographisch und morphematisch) unterteilt. Zum Schluss ergeben sich fünf Prozentzahlen, die die Leistung des Kindes beschreiben sollen.

Jerome hat sich insgesamt um 50% (vier von acht Wörtern) verbessert. Ansonsten gibt es eine große Verbesserung bei den Graphemtreffern und bei den alphabetischen Treffern. Orthographisch und morphologisch hat er sich nicht verbessert (0%). Nach dieser Diagnosemethode dominiert bei Jerome zu beiden Zeitpunkten klar die alphabetische Strategie (s. Tabelle 4.5 und 4.6). Der im Februar erreichte Wert einer 90%igen Beherrschung der alphabetischen Strategie zeigt, dass er in diesem Bereich eine große Sicherheit erlangt hat (dieses Ergebnis wird auch von den oben beschrieben Beobachtungen gestützt). Ein Stufenmodell der Schreibentwicklung zugrunde gelegt, wäre Jerome im Februar in der Situation, in absehbarer Zeit die ersten Schritte auf der nächsten Stufe zu versuchen – das ist in diesem Fall die „orthographische Strategie“. Ob er überhaupt schon „orthographische Elemente“ in seinen Spontanschreibungen verwendet, ließ sich aufgrund der Auswahl der Testwörter nicht feststellen.

Tabelle 4.7 und 4.8 zeigen die Auswertung von Nadines Schreibungen nach dem HSP-Schema. Sie schreibt bei beiden Untersuchungen kein Wort absolut richtig, und liegt sowohl bei Graphemtreffern als auch der alphabetischen Strategie um 50%. Die beobachtbare Verbesserung ist gering, aber feststellbar. Ein Strategiewert von nur 50% besagt, dass sie diese Strategie noch eine Weile beibehalten und ausbauen wird, bevor sie die ersten Schritte in der „orthographischen Strategie“ unternimmt. Aufgrund dieses Diagnoseverfahrens ist es nicht möglich zu sagen, welche 50% sie schon beherrscht und wie die fehlenden 50% erklärbar sind. So ist es nicht ersichtlich, dass im Bereich der Vokalgrapheme besondere Schwierigkeiten bestanden, die sich aber bei der zweiten Untersuchung schon gemildert haben – dieses Ergebnis liefert das HSP-Schema nicht.

[1] Die Untersuchung fand im Rahmen des Seminars „Erstlesen, Erstschreiben“ (Dozentin: Karin Winkler) im Wintersemester 2001/2002 an der Universität Osnabrück statt.

4.6.4. Auswertung nach Dehns Auswertungsverfahren

Zielwort	2.12.	BT	Klass.	2.2.	BT	Klass.
Hund	HONT	2/4	besser / A	HONT	2/4	besser / A
Feder	FEDA	3/5	besser / A	FEDER	5/5	besser / OK
Brillen	BRLÄN	3/7	besser / U	BRLEN	4/7	besser / U
Pinsel	PÖNSEL	5/6	besser / A	PNSEL	5/6	besser / A
Hefte	HFTÄ	3/5	besser / U	HEFTE	5/5	besser / OK
Giraffe	GAFÄ	3/7	rudimentär	GERAFE	4/7	besser / U
Reiter	—	—	—	REITER	6/6	besser / OK
Gürtel	—	—	—	GÜRTEL	6/6	besser / OK

BT: Buchstabentreffer, Klass.: Klassifikation

U: „regelgeleitet, aber unvollständig“, A: „An der eigenen Artikulation orientiert“

Tabelle 4.9.: Auswertung von Jeromes Schreibungen nach Dehns Auswertungsstrategie

Zielwort	2.12.	BT	Klass.	2.2.	BT	Klass.
Hund	HUNT	3/4	besser / A	HUNT	3/4	besser / A
Feder	FDR	3/5	rudimentär	FEDA	3/5	besser / A
Brillen	BNL	1/7	rudimentär	BLN	3/7	rudimentär
Pinsel	PSU	2/6	rudimentär	PESL	3/6	besser / U
Hefte	ED	1/5	rudimentär	HFD	2/5	rudimentär
Giraffe	GRF	3/7	rudimentär	GRAV	3/7	rudimentär
Reiter	—	—	—	RATA	2/6	rudimentär
Gürtel	—	—	—	GÜTL	4/6	besser / U

BT: Buchstabentreffer, Klass.: Klassifikation

U: „regelgeleitet, aber unvollständig“, A: „An der eigenen Artikulation orientiert“

Tabelle 4.10.: Auswertung von Nadines Schreibungen nach Dehns Auswertungsstrategie

Dehn (1994, S.218ff.) beurteilt die Schreibungen nach der Anzahl der richtigen Buchstaben, sowie der richtigen Stellung im Wort, nach vier selbst definierten Kategorien und nach einem Stufenmodel. Die vier Kategorien lauten: „nichts geschrieben“, „diffus“, „rudimentär“ und „besser“.

Jeromes Schreibungen (Tabelle 4.9) weisen, wie auch die HSP-Analyse ergeben hat, beim zweiten Mal mehr Buchstabentreffer auf als bei der ersten Erhebung. Die genauere Klassifikation der Schreibungen zeigt einen Fortschritt bei der Anzahl der vollständigen Schreibungen. Die anderen Kategorien lassen keine deutlichen Schlüsse auf Stärken oder Schwächen zu.

Nadines Schreibungen (Tabelle 4.10) sind in der ersten Untersuchung fast alle als „rudimentär“ einzustufen, bei der zweiten Untersuchung hat sich der Anteil der „besseren“ Schreibungen auf 50% erhöht. Die Ausdifferenzierung bei der Kategorie „besser“ zeigt noch einmal mehrere Stufen, im Gegensatz zum HSP-Schema, das beide Kinder der gleichen Stufe zugeordnet hat. Nadines „bessere“ Schreibungen sind in die Bereiche A (an der Artikulation orientiert) und U (unvollständig) einzuordnen. Ihre Schreibungen sind nicht mehr „diffus“ (also nicht regelgeleitet) und in zunehmenden

	Phänomen	max	repräsentiert	phonetisch korrekt	orthogr. korrekt
1	**Silbenanzahl**	12	**12 → 12**	08 → 10	**02 → 07**
2	Vollsilben	07	07 → 07	03 → 05	01 → 02
3	Reduktionssilben	05	05 → 05	05 → 05	**01 → 05**
4	**Silbenkerne**	12	09 → 10	09 → 10	**02 → 08**
5	S’	06	04 → 04	04 → 04	01 → 03
6	S’ (fester Anschluss)	05	**03 → 03**	03 → 03	01 → 02
7	S’ (loser Anschluss)	01	01 → 01	01 → 01	01 → 01
8	S^0	05	05 → 05	05 → 05	**01 → 05**
9	S^0 (/ə/)	02	02 → 02	02 → 02	**00 → 02**
10	S^0 (/ɐ/)	01	01 → 01	01 → 01	**00 → 01**
11	S^0 (silbischer Sonorant)	02	02 → 02	02 → 02	01 → 02
12	S	01	00 → 01	00 → 01	00 → 00
13	**Anfangsränder**	12	11 → 12	11 → 12	11 → 12
14	einfach	11	10 → 11	10 → 11	10 → 11
15	komplex	01	01 → 01	01 → 01	01 → 01
16	**Endränder**	03	03 → 03	03 → 03	02 → 02
17	einfach	02	02 → 02	02 → 02	02 → 02
18	komplex	01	01 → 01	01 → 01	00 → 00
19	**Phonologische Markierungen**	02	00 → 00	—	00 → 00
20	Schärfung	02	00 → 00	—	00 → 00
21	**Konstantschreibung**	01	00 → 00	01 → 01	00 → 00
22	Auslautverhärtung	01	00 → 00	01 → 01	00 → 00

Tabelle 4.11.: Analyse der Testwörter Jeromes, die in beiden Untersuchungen vorkamen.

Maße nicht mehr „rudimentär“ (also nur teilweise regelgeleitet), sondern überwiegend regelgeleitet. Allerdings steht sie noch am Anfang der „besseren“ Schreibentwicklung – ihr gelingt es noch nicht, alle zu verschriftenden Phoneme zu identifizieren und normgerecht niederzuschreiben.

4.6.5. Auswertung nach dem eigenen Auswertungsverfahren

Die in Abschnitt 4.5 entwickelte Auswertungstabelle bietet eine Möglichkeit, die Schreibentwicklung näher zu analysieren. Dazu wurden zunächst nur die Analysen der in beiden Untersuchungen vorhandenen Wörter in Tabelle 4.11 und 4.12 eingetragen, also <Hund>, <Feder>, <Brillen>, <Pinsel>, <Hefte> und <Giraffe>. Besondere Entwicklungen, auffällige Leistungen oder Defizite werden fett gedruckt.

Auffällig bei Jerome sind folgende Punkte:

- Schon bei der ersten Untersuchung hat er alle Silben in seinen Schreibungen repräsentiert, einen Großteil davon auch phonetisch korrekt (1. Zeile).
- Die Anzahl der orthographisch korrekt geschriebenen Silben hat sich mehr als verdreifacht und liegt schließlich bei über 50% (1. Zeile).

- Bei Betrachtung der unterschiedlichen Silbentypen wird sofort deutlich, dass Jerome zwischen den beiden Untersuchungen gelernt hat, Reduktionssilben orthographisch vollständig korrekt zu repräsentieren (3. Zeile).
- Die Silbenkerne sind entgegen der Repräsentation aller Silben in beiden Untersuchung nicht komplett vorhanden. Einen großen Sprung gibt es wieder bei der orthographisch korrekten Schreibung. Diese Steigerung ist aber größtenteils durch den Fortschritt bei den Reduktionssilben zu erklären. Um diese Frage genauer zu klären, sind die Silbenkerne in den folgenden Zeilen detailliert analysiert.
- Vokale mit festem Anschluss notiert Jerome nur zu 60%, diese allerdings phonetisch korrekt. Es ist hier auch kein Fortschritt erkennbar – die Erkennung von (Kurz-)Vokalen mit festem Anschluss als zu verschriftende Silbenkerne ist offensichtlich auch im Februar eine besondere Schwierigkeit für Jerome (Zeile 6).
- Auffällig ist, dass die Anfangsränder schon bei der ersten Untersuchung fast vollständig selbst orthographisch korrekt geschrieben wurden (Zeile 13). Gleiches gilt für die Endränder (Zeile 16).
- Die in der Testwortmenge vorkommenden phonologischen Markierungen kann Jerome nicht als besonders zu markierende Phänomene erkennen (Zeile 19). Gleiches gilt für morphologische Konstantschreibung (Zeile 21). Allerdings sind beide Phänomene in der Testwortmenge nur selten vorhanden.

Bei der Betrachtung von Nadines Auswertungsergebnissen (Tabelle 4.12) fallen folgende Punkte auf:

- Sie hat in beiden Untersuchungen alle Silben repräsentiert, wobei sowohl bei der phonetischen als auch der orthographischen Korrektheit noch Schwächen bestehen (Zeile 1).
- Ihr gelingt es auch im Februar nicht, auch nur eine einzige Reduktionssilbe korrekt zu schreiben.
- Bei der Repräsentation von Silbenkernen hat sich die Anzahl der phonetisch plausibel verschrifteten stark erhöht, liegt aber dennoch nur bei 60%. Die Gründe sind nicht auf ein einzelnes Phänomen zurückzuführen, sondern betreffen den gesamten Vokalismus (Zeile 4).
- Anfangränder werden sicher identifiziert und phonetisch plausibel verschriftet – orthographisch korrekt allerdings nur zu 75% (Zeile 13). Endränder werden ebenfalls erkannt und dann auch plausibel verschriftet.

	Phänomen	**max**	**repräsentiert**	**phonetisch korrekt**	**orthogr. korrekt**
1	**Silbenanzahl**	12	**12 → 12**	**01 → 06**	00 → 02
2	Vollsilben	07	07 → 07	01 → 03	00 → 02
3	Reduktionssilben	05	05 → 05	**00 → 03**	**00 → 00**
4	**Silbenkerne**	12	05 → 07	03 → 07	02 → 03
5	S’	06	02 → 04	02 → 04	02 → 03
6	S’ (fester Anschluss)	05	02 → 03	02 → 03	02 → 02
7	S’ (loser Anschluss)	01	01 → 01	01 → 01	01 → 01
8	S^0	05	03 → 03	01 → 03	00 → 00
9	S^0 (/ə/)	02	00 → 00	00 → 00	00 → 00
10	S^0 (/ɐ/)	01	01 → 01	00 → 01	00 → 01
11	S^0 (silbischer Sonorant)	02	02 → 02	01 → 02	00 → 00
12	S	01	00 → 00	00 → 00	00 → 00
13	**Anfangsränder**	12	11 → 12	10 → 11	09 → 09
14	einfach	11	10 → 11	10 → 11	09 → 09
15	komplex	01	01 → 01	01 → 00	00 → 00
16	**Endränder**	03	01 → 02	01 → 02	00 → 01
17	einfach	02	02 → 02	00 → 01	00 → 01
18	komplex	01	01 → 01	01 → 01	00 → 00
19	**Phonologische Markierungen**	02	00 → 00	—	00 → 00
20	Schärfung	02	00 → 00	—	00 → 00
21	**Konstantschreibung**	01	00 → 00	01 → 01	00 → 00
22	Auslautverhärtung	01	00 → 00	01 → 01	00 → 00

Tabelle 4.12.: Analyse der Testwörter Nadines, die in beiden Untersuchungen vorkamen.

	HSP	Dehn	eigenes Verfahren
Jerome	alphabetische Strategie gut beherrscht	sämtliche Schreibungen sind „besser“	alle Silben sind repräsentiert, meist phonetisch plausibel
	keine orthographische Strategie	noch vorhandene Abweichungen sind entweder unvollständig oder an der Artikulation orientiert	Reduktionssilben werden gut beherrscht
			Probleme bestehen bei Vokalen mit festem Anschluss
			Schärfungsschreibung wird noch nicht beherrscht
Nadine	alphabetische Strategie noch mit Schwächen	Übergang zwischen rudimentären und „besseren“ Schreibungen	alle Silben sind repräsentiert, die Hälfte davon phonetisch plausibel
	keine orthographische Strategie	die „besseren“ Schreibungen sind unvollständig oder an der Artikulation orientiert	Probleme bestehen bei der korrekten Schreibungen von Reduktionssilben
			Probleme bestehen im Vokalismus, insbesondere der korrekten orth. Repräsentation

Tabelle 4.13.: Vergleich der Ergebnisse

4.6.6. Konsequenzen für den Unterricht

Die Ergebnisse der Analyse sollten genutzt werden können, um aus ihnen konkrete Schritte abzuleiten. D.h. eine Analyse sollte nicht nur dazu dienen, einen Standort zu bestimmen, sondern den Standort so genau zu bestimmen, dass es möglich ist, dem Kind Hilfen anzubieten, die vorhandenen Fähigkeiten zu sichern und erkannte Schwächen zu beseitigen. Eine kurze Zusammenstellung der Analyseergebnisse aller drei Verfahren in Tabelle 4.13 zeigt die Unterschiedlichkeit der Ergebnisse für die zweite Untersuchung.

Der unterschiedliche Detailgrad wird in solch einer Zusammenfassung sofort ersichtlich. Das sollte nicht verwundern, schließlich ist die Granularität der Auswertungskategorien bei dem hier vorgestellten Verfahren weit höher als bei den anderen beiden. Es wird aber deutlich, dass die mit der HSP und Dehns Auswertungsschema erzielten Ergebnisse sich nicht für die Bestimmung konkreter unterrichtlicher Handlungen eignen, sondern als grobe Standortbestimmung aufzufassen sind, die aber u.U. wichtige Lernfortschritte oder -defizite ganz unerkannt lassen kann.

Aus den Ergebnissen des eigenen Auswertungsverfahrens lassen sich direkt einige Vorschläge zur weiteren Arbeit mit den beiden Kindern ableiten.

Jerome:

- Überprüfung der Reduktionssilben anhand „schwieriger“ Fälle mit silbischen Sonoranten oder mehreren Reduktionssilben.
- Wörter mit Kurzvokal in Häuser einordnen(vgl. Röber-Siekmeyer und Pfisterer, 1998; Saure et al., 1997). Die dort angebotene Regel lautet: „Es muss immer ein Vokalgraphem da sein“, das gilt auch, wenn der Vokal phonetisch weniger „prominent“ ist als Vokale mit losem Anschluss.
- Wenn die Anschlussopposition zwischen fest und lose sicher erkannt wird, können Jerome Wörter mit Notwendigkeit einer Schärfungsmarkierung zur Einordnung in Häuser präsentiert werden.
- Der Bereich der morphologischen Konstantschreibung ist in der Testwortmenge unterrepräsentiert gewesen, so dass keine genauen Aussagen möglich sind. Zur Überprüfung können Wörter mit unterschiedlich „weiten“ Abgleichen (Hund - Hunde, Staub - staubig) präsentiert werden.

Nadine:

- Bestehende Unsicherheiten bei der Repräsentation von Silbenkernen können durch verstärkte Übungen mit den Häusern beseitigt werden. Die Aussage, dass „im zweiten Zimmer immer ein Vokalbuchstabe stehen muss“ kann sie noch nicht bei freien Schreibungen umsetzen.
- Genauso kann bestehenden Unsicherheiten bei der Reduktionssilbe durch die analoge Erklärung, dass im zweiten Zimmer der Garage immer ein <e> vorhanden sein muss, begegnet werden.
- Der Bereich der morphologischen Konstantschreibung ist in der Testwortmenge unterrepräsentiert gewesen, so dass keine genauen Aussagen möglich sind. Zur Überprüfung können Wörter mit unterschiedlich „weiten“ Abgleichen (Hund - Hunde, Staub - staubig) präsentiert werden.

4.7. Zusammenfassung

Das vorgestellte eigene Auswertungsschema konnte sehr konkret Entwicklungen, Fähigkeiten und Probleme der beiden untersuchten Kinder aufzeigen. Die erzielten Ergebnisse decken sich gut mit der intuitiven Einschätzung der Schreibungen, die für einen im Umgang mit Schriftspracherwerbsproblemen erfahrenen Auswerter durch keine noch so detaillierte Analyse übertroffen werden dürften. Der beabsichtigte Nutzen aller drei vorgestellter Verfahren liegt demnach zum einen in der Auswertung großer Mengen von Daten und einer Zusammenfassung der festgestellten Leistungen, die ohne eine aufgezeichnete Auswertung sonst intuitiv bleiben muss. Desweiteren sollen die Verfahren Sicherheit geben, keinen wichtigen Bereich ausgelassen zu haben.

Gerade den letzten Punkt können Dehns Auswertungsschema und stärker noch die Hamburger Schreibprobe nicht erfüllen. Durch die ausschließliche (oder zumindest grundlegende) Konzentration auf Graphemtreffer bzw. eine nicht weiter differenzierte Phonem-Graphem-Korrespondenz können sie wichtige Zugänge der Kinder zur Schrift – nämlich die über ihre eigene gesprochene Sprache über prosodische Kategorien – nicht erfassen. Beide Verfahren benutzen ein wenig geeignetes Instrumentarium, um Symptome zu untersuchen und können daher die Ursachen nicht benennen.

Das vorgestellte eigene Auswertungsschema ist sehr detailliert und in der Anwendung recht aufwendig. Die beiden Beispiele haben aber gezeigt, dass eine sehr genaue Betrachtung der Schreibungen notwendig ist, um tatsächlich aussagekräftige Ergebnisse zu erzielen. Die HSP und auch Mechthild Dehn reklamieren Lehrgangsunabhängigkeit für ihre Verfahren. Dass zumindest Zweifel an dieser Aussage angebracht sind, wurde auf S. 26 gezeigt. Für das hier entwickelte Verfahren gilt dieser Anspruch nur zum Teil: Auf der reinen Auswertungsseite werden keine Vorannahmen über eine bestimmte Art von Unterricht gemacht, wohl aber bei der Formulierung von Konsequenzen. Die Orientierung an etablierten und überprüften sprachwissenschaftlichen Erkenntnissen (vgl. Maas, 1992; Maas et al., 1999) sorgt für Validität der Ergebnisse: Die klare Definition der einzelnen Kategorien macht außerdem Verwechslungen und Unsicherheiten bei der Einordnung, die Untersuchungsergebnisse verfälschen könnten, unwahrscheinlicher – z.B. im Gegensatz zu den HSP-Kategorien, wenn nicht auf vorklassifizierte Testwörter zurückgegriffen wird.

5. Erhebung und Repräsentation von Testdaten

Im Rahmen dieser Arbeit werden eine Reihe von Verfahren entworfen und evaluiert, die auf verschiedenen Ebenen orthographische Analysen korrekter und fehlerhafter Schreibungen vornehmen können. Der Grund, nicht ein einzelnes, möglichst mächtiges Verfahren zu entwickeln, liegt darin, dass in Anwendungssituationen wie den in der Einleitung vorgestellten Szenarien nicht alle Informationen verfügbar sind bzw. mit der gleichen Sicherheit erschlossen werden können. Es ist daher sinnvoll, Verfahren mit unterschiedlichem Bedarf an Informationen zu definieren, um auch in unsicheren Situation Analysen liefern zu können. Die Frage, welche Art und Menge von Informationen welche Art und welchen Detailgrad von Analysen ermöglicht, ist darüber hinaus von theoretischem Interesse.

Die hier entwickelten Verfahren lassen sich grob in zwei Klassen einteilen: Solche, die Analysen allein aus der Ausgangsschreibung abzuleiten versuchen und solche, die einen Abgleich zwischen Ausgangs- und Zielschreibung vornehmen. In Kapitel 6 werden zunächst Verfahren entworfen, die keine Informationen über die Zielschreibung zur Verfügung haben. Kapitel 7 definiert Verfahren, die auf die Zielschreibung und ggf. weitere linguistische Informationen zurückgreifen. Ein vollständiges, maximal mächtiges Verfahren würde diese Schritte miteinander verbinden und aus einer Analyse der Ausgangsschreibung heraus auf die Zielschreibung schließen sowie schließlich die Abweichungen analysieren. Hier wird die Auffassung vertreten, dass es möglich und sinnvoll ist, diese Schritte zu trennen und ihre Eigenschaften und Erfolgsbedingungen unabhängig voneinander zu überprüfen.

5.1. Maschinenlesbare Kodierung

Natürlichsprachliche Texte haben in ihrer graphischen, d.h. visuellen Manifestation zunächst eine zweidimensionale Struktur. Sie sind auf Flächen, wie z.B. einem Blatt Papier notiert. Sie fungieren als Abbildung gesprochener Sprache, die als zeitlich verlaufende Äußerung eine eindimensionale Struktur hat. Daher muss es möglich sein, zweidimensional notierte Texte in eine eindimensionale Repräsentation zu überführen. Das für das Deutsche verwendete Schriftsystem benutzt als kleinste Bestandteile Buchstaben. Das sind graphischen Zeichen, die zwar für sich komplex sind, deren einzelne Formmerkmale jedoch nur zur Unterscheidung bzw. Erkennung von Buchstaben dienen.[1] Für das Deutsche und alle anderen bekannten Schriften kann eine endliche Menge von grundlegenden Zeichen angenommen werden, die miteinander kombiniert werden können, um sprachliche Aussagen zu repräsentieren. Für die Kombination dieser Zeichen auf einer Fläche sind grundsätzlich

[1] Es ist sehr schwierig, genau definieren, was z.B. ein `A` ausmacht (s. Hofstadter, 1991, 246ff.). Jedes einzelne Merkmal, wie eine zulaufende, geschlossene Spitze oder ein waagerechter Balken in der Mitte, kann fehlen oder verändert sein, jedoch zeichnet alle Formen von `A` eine Familienähnlichkeit (Wittgenstein, 1984, § 67, S. 278) aus. Für einen sinnvoll definierten Font ist es notwendig, hinreichend große Unterschiede zwischen einzelnen Buchstaben aufzuweisen, so dass auch ähnliche Buchstaben (wie `A`/`H` oder `O`/`Q`) unterscheidbar sind. Für die Entzifferung handschriftlicher Texte ist es u.U. notwendig, die verschriftete Sprache zu kennen, um Wörter mittels Erwartungen, die von der sprachlichen Umgebung gestützt werden, entziffern zu können.

komplexe Möglichkeiten denkbar,[2] die auf dem lateinischen System basierenden Alphabetschriften werden jedoch per Konvention zeilenweise von links nach rechts geschrieben, wobei das auf das letzte Zeichen am Ende einer Zeile folgende Zeichen am Beginn der darunter liegenden Zeile steht (für andere Möglichkeiten s. Sproat, 2000, 59ff.). Daraus lässt sich für einen deutschen Text eine lineare Struktur ableiten, die aus einer Kette von Buchstaben (und Zahlen), Satzzeichen und Leerstellen besteht.

Eine maschinenlesbare Kodierung dieser Kette ist leicht möglich, indem Leerstellen und Zeilenwechsel durch ein spezielles Zeichen (Leerzeichen oder Blank bzw. Carriage Return/Line Feed) kodiert werden und jedes Zeichen eine eindeutige Repräsentation zugeordnet bekommt. Für den Austausch von Texten ist es notwendig, ein gemeinsames Zeicheninventar festzulegen und die Zuordnung zu standardisieren. Der derzeit am weitesten verbreitete Standard ist die ASCII-Kodierung[3], die allerdings in ihrer Ursprungsform neben einer Menge von Steuercodes nur Zeichen enthält, die für die Repräsentation des Englischen notwendig sind. Erweiterungen für andere Sprachen, wie z.B. die Kodierung von Umlauten und ß für das Deutsche, waren lange Zeit nicht standardisiert und betriebssystemabhängig. Heute gibt es eine Reihe von ISO-ASCII-Standards, wie z.B. ISO-LATIN-1, der für die meisten westeuropäischen Sprachen ausreicht. Die Zukunft gehört UNICODE The Unicode Consortium (2003), einer 16-Bit-Repräsentation, die ca. 65.000 Zeichen direkt und weitere 1 Million Zeichen durch indirekte Kodierung darstellbar macht. Unicode enthält den Latin-1-Code als echte Untermenge (UTF-8), so dass in ASCII/Latin-1 vorliegende Texte auch in Zukunft weiter verwendbar sein werden. Gleichwohl wird ASCII/Latin-1 dort eine wichtige Rolle behalten, wo es um die Darstellung formaler Repräsentationen geht, die wie Programmquelltexte vorrangig für die maschinelle Verarbeitung gedacht sind.

Für die im Folgenden definierten Verfahren wird eine ASCII-ISO-LATIN-1-Kodierung der Texte verwendet. Quelltexte von Programmen müssen in der Regel in 7-Bit-ASCII vorliegen, zumindest sind Schlüsselwörter, Variablennamen etc. auf diesen Zeichensatz beschränkt. Der Datentyp `string` (Zeichenkette) kann überlicherweise auch ASCII-Erweiterungen bzw. Unicode aufnehmen, jedoch ist es im speziellen Fall von Prolog notwendig, ein-eindeutige Ersatzdarstellungen für `ä,ö,ü` und `ß` zu wählen, wenn der Datentyp `atom` zur Repräsentation von Wörtern oder Graphemen gewählt wird.

5.2. Trainingskorpora

Es werden im Folgenden zwei Klassen von Verfahren zur automatischen Analyse orthographischer Leistungen vorgestellt. Die erste Klasse beschränkt sich darauf, (möglicherweise) fehlerhafte Formen in einem Text zu identifizieren, die zweite leitet qualitative Analysen in Form von phänomenbezogenen Erklärungen aus den Schreibungen ab. Die Verfahren der ersten Gruppe stützen sich auf große, möglichst umfassende Wortlisten des Deutschen, die als Trainings- oder Abgleichmenge verwendet werden.

Die Leistung eines solchen Analysealgorithmus wird durch das Verhältnis zweier Werte bestimmt, die im Information Retrieval und in der statistischen Sprachverarbeitung als *Precision* und *Recall*

[2] Sproat (2000, 34ff.) demonstriert eine Modellierung komplexerer Kombinationen für das Chinesische und Koreanische.

[3] American Standard Code for International Interchange. Der reine ASCII-Code ist ein 7-Bit-Code, d.h. er kann maximal 128 verschiedene Zeichen repräsentieren.

bezeichnet werden (Manning und Schuetze, 1999, 267ff.). Die Menge der zu untersuchenden Wortformen lässt sich für ein Verfahren, das erkennen soll, ob es sich jweils um eine korrekt oder fehlerhaft geschriebene Wortform handelt, in zwei disjunkte Mengen teilen: Die Menge der tatsächlich korrekt geschriebenen Wortform (k) und die Menge der tatsächlich fehlerhaft geschriebenen Wortformen (f). Der Algorithmus wird auf die Vereinigung beider Mengen angewendet und liefert ebenfalls zwei disjunkte Mengen, nämlich die Menge der als korrekt klassifizierten Formen (markiert mit $^+$) und die Menge der als inkorrekt klassifizierten Wortformen (markiert mit $^-$). Bei einem idealen Verfahren sind die beiden Aufteilungen der Gesamtmenge identisch, jedes suboptimale Verfahren liefert vier Teilmengen:

- Korrekt geschriebene Formen, die als korrekt klassifiziert wurden (k^+).
- Falsch geschriebene Formen, die als falsch klassifiziert wurden (f^-).
- Korrekt geschriebene Formen, die als falsch klassifiziert wurden (k^-).
- Falsch geschriebene Formen, die als korrekt klassifiziert wurden (f^+).

Die ersten beiden Fälle stellen korrekte Leistungen des Verfahrens dar, die letzten beiden Fehler. Für ein Verfahren, dass Rechtschreibfehler erkennen soll, sind insbesondere die letzten drei Mengen interessant. Die Zahl der als korrekt klassifizierten korrekten Wortformen ist üblicherweise sehr hoch und besitzt daher die geringste Aussagekraft. Die *Precision* wird hier definiert als das Verhältnis der korrekt klassifizierten Fehler zur Menge aller als Fehler klassifizierten Formen, sie sagt aus, wie hoch die „Trefferquote“ des Verfahrens ist:

Definition 1 *Precision:* $P = \frac{f^-}{f^- + k^-}$

Unter *Recall* wird das Verhältnis der korrekt klassifizierten Fehlschreibungen zur Menge aller tatsächlichen Fehler gefasst, d.h. wie vollständig die Erkennungsleistung des Verfahrens ist:

Definition 2 *Recall:* $R = \frac{f^-}{f^- + f^+}$

Als kombiniertes Maß für die Leistung eines Fehlererkennungsverfahrens kann das *F-Maß* dienen, das definiert ist als:

Definition 3 $F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$

α ist ein Faktor, der die Gewichtung zwischen *Precision* und *Recall* steuert. Ein häufig verwendeter Wert ist 0,5, der zu einer Gleichgewichtung führt. In diesem Fall ergibt sich $F = 2PR/(R + P)$.

Abbildung 5.1.: Bildergeschichte: Der arme Hund

5.3. Korpuserfassung

Um aussagekräftige Testdaten für die zu entwickelnden Verfahren zur Verfügung zu haben, wurde ein eigenes Korpus erhoben und maschinenlesbar kodiert. Mit Unterstützung von Prof. Dr. Christa Röber-Siekmeyer (PH Freiburg) ist eine Bildergeschichte ausgewählt worden, die direkt und indirekt an potenziell kooperationswillige Lehrkräfte zweiter Klassen im gesamten deutschen Sprachraum geschickt wurde.

Die ausgewählte Bildergeschichte aus der Reihe „Der kleine Herr Jakob“ von Hans-Jürgen Press (Press, 1987) ist in Abbildung 5.1 wiedergegeben. Die Verwendung solcher Aufsatzformen in der zweiten Klasse ist nicht unbedingt üblich (“Bildergeschichten werden etwa von der vierten Klasse an geschrieben“ Widmann, 2001, S. 1), allerdings kam es bei der Konzeption der Erhebung nicht auf die Untersuchung narrativer Kompetenz an. Es sollte vielmehr ein Schreibanlass geschaffen werden, der – zumindest prinzipiell – strukturell vergleichbare Texte hervorbringt, die jedoch trotzdem als *freie* Texte der Kinder aufzufassen sind: *Frei* in der Wortwahl, *frei* in der Ausgestaltung und Auschmückung der Erzählung, jedoch gebunden an ein Thema, das auch bestimmte Wortformen induzieren sollte.

Den Schülerinnen und Schülern wurde die Aufgabe gestellt, ohne weitere Anleitung, insbesondere der Vorgabe einzelner Wörter oder Textteile, den Inhalt der Bildergeschichte in eigenen Worten wiederzugeben. Daneben wurde je Klasse ein Fragebogen zu Unterrichtsform, Methodik der Schrifteinführung, Art und Umfang der bislang behandelten orthographischen Regeln usw. ausgefüllt.

Die Texte wurden manuell erfasst und annotiert, d.h. für jede fehlerhafte Wortform wurde notiert, was die korrekte Schreibung gewesen wäre, und ob es sich um einen Grammatik- oder Orthographiefehler handelt. Orthographiefehler wurden auch wortübergreifend (z.B. Getrennt- und Zusammenschreibungsfehler wie *<geter> für <geht er>), Grammatikfehler nur wortbezogen notiert. Im

Falle von Grammatikfehlern sind dies vor allem Kongruenzfehler, Flexionsfehler wie *<schleichte> für <schlich>, nicht aber Wortstellungs- oder -auslassungsfehler. Zusätzlich wurden Zeichensetzungsfehler erfasst.

Namen wurden in der Regel nicht korrigiert, außer bei eindeutig unorthographischen Formen (*<Spaeig> für <Spike>, *<Frtz> für <Fritz>). Auf einheitliche Verwendung im Text wurde jedoch geachtet. Beispiele für nicht markierte Namensschreibungen sind: <Maksl>, <Jackob>, <Belo>, <Bällo>, <Oska> und <Ranternplan>. Falls ein Name innerhalb eines Textes sehr inkonsistent geschrieben wurde, wurde die häufigste Schreibung als die intendierte angenommen (*<Wili>, *<Wieli>, *<Wilie>, <Willi>).

Um optimale Maschinenlesbarkeit des Korpus zu gewährleisten, wurde eine XML-Kodierung gewählt (s.u.), die eine Verarbeitung mit Standardtools oder Standardbibliotheken verbreiteter Programmiersprachen (z.B. Java, Perl oder Python) erlaubt. Auf diese Weise ist die Wiederverwertbarkeit des Korpus und die flexible Verarbeitbarkeit mit unterschiedlichen Analysetools mit geringem Aufwand gewährleistbar.

5.4. Korpusrepräsentation

Wenn für jedes Korpus bzw. jedes Analysevorhaben ein eigenes Format entwickelt und gewählt wird, ist es praktisch ausgeschlossen, Ergebnisse wiederzuverwenden und damit zu vergleichen. Es gilt also, ein allgemeines, standardisiertes Format zur maschinenlesbaren Notation von potenziell orthographisch fehlerhaften Texten zu finden, das die Integration verschiedener Informationen ermöglicht und die Verarbeitung mit verschiedenen automatisierten und halbautomatisierten Analyseverfahren erlaubt.

Mitte der 80er Jahre wurde der SGML-Standard (Standard Generalized Markup Language) zur Textauszeichnung festgelegt, der es ermöglicht, dem Text eine inhaltliche Struktur zuzuweisen, d.h. einzelne Bereiche des Textes werden mit einer Bezeichnung ihrer strukturellen Bedeutung markiert (z.B. „dieser Textbereich ist eine Überschrift 1. Ordnung“, „dieser Textbereich ist eine Anmerkung“ usw.). In Textverarbeitungssystemen wie Microsoft Word finden sich solche Verfahren ebenfalls, nämlich in der Verwendung von Absatzvorlagen: Einzelne Absätze werden mit der Information „Überschrift 1“, „Standard“ usw. belegt. Die Angabe darüber, wie eine Überschrift darzustellen ist (Schriftart, Schriftgröße, Einrückung), findet sich an anderer Stelle, getrennt vom eigentlichen Text. Dieses Vorgehen hat gegenüber der direkten Formatierung den Vorteil, dass sichergestellt ist, dass alle Überschriften im Text gleich dargestellt werden, die Darstellung für alle Überschriften gleichzeitig schnell geändert und das gleiche Dokument durch Verwendung einer anderen Dokumentvorlage schnell umformatiert werden kann. Der Benutzer ist jedoch nicht gezwungen, diese Strukturierungsmöglichkeiten zu nutzen. Zudem sind die Auszeichnungen in hohem Maße herstellerspezifisch, d.h. bei Weiterbearbeitung eines Textes mit einem anderen Programm (oder auch nur einer anderen Version desselben Programms) gehen die Informationen verloren.

Der oben erwähnte SGML-Standard vermeidet solche Inkompatibilitäten und definiert einen Weg, spezielle Beschreibungssprachen (d.h. eine Menge verwendeter Strukturbezeichungen) festzulegen, so dass sie von allen SGML-fähigen Programmen (SGML-Parsern) verarbeitet werden können. SGML selbst legt also keine Strukturbezeichnungen (Tags) fest, sondern ermöglicht eine Menge von

SGML-Anwendungen. Die bisher bekannteste SGML-Anwendung ist HTML (Hypertext-Markup-Language), das als Format für WWW-Seiten verwendet wird.

Eine typische HTML-Datei sieht wie folgt aus:

```
<html>
  <head>
    <title> Der arme Hund </title>
  </head>

  <body>

  <h1>Der arme Hund</h1>

  <p>Es schneit.
  Und der kleine Herr Jakob kokt aus dem Fenster.
  und er sit Einen Hund der die Fusmate klaut.
  </p>

  </body>
</html>
```

Jede Bereichskennzeichnung besteht aus einem Starttag, z.B. `<html>` und dazu passendem Endtag, z.B. `</html>`. Die obige Datei besteht also auf höchster Ebene aus einem Bereich, der von den Tags `<html>` und `</html>` eingeschlossen ist. Die Datei besteht aus einem `html`-Block. Dieser Block besteht aus zwei weiteren Blöcke, einem `head` und einem `body` Block. Der `head`-Block beinhaltet Meta-Informationen, wie z.B. den Titel oder den Namen des Autors, und der `body`-Block den eigentlichen Text.

Die SGML-Definition von HTML beschreibt, welche Tags innerhalb welcher Blöcke vorkommen können, welche zusätzlichen Argumente für die Tags zulässig sind, nicht aber, wie der Text letzendlich dargestellt oder ausgedruckt wird. Ein Nachteil von HTML ist, dass keine neuen Tags benutzt werden können, um Informationen außerhalb der Dokumentgliederung darzustellen. Damit disqualifiziert sich HTML als Beschreibungsformalismus für Vorhaben wie die Repräsentation von Texten als Grundlage orthographischer Analysen, denn dort sollen z.B. spezifische Informationen über den Schreibenden integriert und als solche gekennzeichnet werden. Dafür wird eine andere – vom Aussehen ähnliche, aber mit anderen Tags ausgestattete Sprache benötigt.

Eine vereinfachte Version von SGML, XML (Bray et al., 2000, Extended Markup Language) hat HTML auf vielen Gebieten ersetzt und ist ein allgemein anerkanntes Format für den Austausch von Daten aller Art geworden, weil es die flexible Definition eigener Tags (also die Beschreibung einer eigenen Markierungssprache) erlaubt. XML ist ideal geeignet, um die oben aufgestellten Kriterien an eine Formatierung von Texten als Grundlage für eine orthographische Analyse zu erfüllen.

SGML und XML werden zur Informationsstrukturierung verwendet. Dabei liegt ihre Stärke darin, dass sie nicht, wie z.B. eine relationale Datenbank mit zu Grunde liegender Entity-Relationship-Modellierung eine vollständige Strukturierung und totale Zuordnung aller Informationen zu kleinsten Informationstypen vornehmen müssen, sondern auf halbstrukturierten Daten operieren. Gleichzeitig ist die Tiefe der Strukturierung nicht beschärnkt und kann für unterschiedliche zu repräsentierende Phänomene bzw. Informationen unterschiedlich weit gehen. Eine häufige Anwendung von SGML/XML-Kodierungen ist die Annotation laufenden natürlichsprachlichen Textes, der in einer (relationalen) Datenbankstruktur nur schwer mit der gleichen Flexibilität untergebracht werden kann. Art und Umfang der Strukturierung werden durch eine *Document Type Definition* (DTD) festgelegt, die als kontextfreie Grammatik die gültigen Verschachtelungen von Blöcken beschreibt. Damit ist allerdings nur die Syntax eines SGML/XML-Dokumentes festgelegt, die semantische Interpretation der Daten nimmt das verarbeitende Programm anhand der vorgefundenen Strukturierung vor und hält sich dabei an Konventionen evtl. standardisierter DTDs.

Lobin (1999, 9ff.) unterscheidet zwei Ebenen der Informationssturkturierung:

- Eine konkrete Ebene der Daten-Elemente und
- eine abstrakte Ebene der Container-Elemente, die Daten Funktionen zuordnen und sie gruppieren.

Bei der Kodierung natürlichersprachlicher Texte besteht die konkrete Ebene aus dem reinen (druckbaren) Text, die abstrakte Ebene strukturiert den Text auf logischer (Absätze, Kapitel, Überschriften etc.) und konzeptueller (Fließtext, Tabellen, Abbildungen, Fußnoten etc.) Ebene und weist weitere Annotationen aus. Solche Annotationen können Layoutanweisungen, linguistische Informationen, Metainformationen z.B. über Erstellungsdatum, Autor usw. sein. Obwohl ein Text derart auf verschiedenen Ebenen strukturiert werden kann, werden die abstrakten Elemente in einer Hierarchieebene platziert, so dass sich grundsätzlich eine Baumstruktur ergibt. Damit ist es nicht möglich, überlappende Strukturierungen vorzunehmen, wie z.B. eine syntaktische Strukturierung in Sätze und eine damit nicht kongruente semantische Strukturierung, die Beginn und Ende einer Einheit innerhalb von Sätzen und satzgrenzenübergreifende Einheiten zulässt. Solche Strukturierungen sind gleichwohl möglich, allerdings nur auf Attribut- und nicht auf Elementebene (vgl. Lobin, 1999, 61ff. zur Repräsentation gerichteter Graphen anstelle von Bäumen).

In einer Baumstruktur kann die direkte Beziehung zweier Informationseinheiten durch zwei verschiedene Relationen ausgedrückt werden: sequenziell geordnet oder hierarchisch geordnet. Im ersten Fall sind beide Einheiten Tochterknoten eines übergeordneten Knotens, im zweiten Fall ist eine Einheit in einem Tocherknoten der anderen repräsentiert. Für den Entwurf einer geeigneten Repräsentationssprache für Korpora potenziell orthographisch fehlerhafter Texte ist also eine Menge sinnvoller Strukturierungseinheiten auf abstrakter Ebene zu finden und gleichzeitig in sequenzielle oder hierarchische Beziehung zueinander zu setzen.

Zunächst wird für das vorliegende Korpus zwischen zwei Strukturierungsebenen unterschieden:

1. Text übergreifende Ebene: Die Einzeltexte des Korpus stehen miteinander in Beziehung und ihre Gruppierung ergibt sich aus den verfügbaren Metainformationen. Jeder Text verfügt über Metainformationen wie Entstehungszeitpunkt, Art der Niederschrift (handschriftlich oder am Computer) oder Bezug zu einer Bildergeschichte bzw. Aufgabe. Für jeden Schüler sind individuelle Informationen über Alter, Herkunft, Muttersprache, besondere Förderung etc. verfügbar, so dass es sinnvoll ist, alle Texte eines Schreibers zu gruppieren und die Metainformationen nur ein einziges Mal zu kodieren. Gleiches gilt für Schulklassen, für die Informationen über Klassenstufe, Ort, verwendete Unterrichtsmethoden etc. vorhanden sind. Es ergibt sich also folgende hierarchische Gliederung: Korpus – Schulklasse – Schüler – Einzeltext. Einen Überblick über die erhobenen Metadaten und ihre Codierung gibt Abschnitt 5.6.
2. Innertextliche Ebene: Die interne Strukturierung eines Texte verwendet zunächst das konkrete Datenmaterial, d.h. den tatsächlich geschriebenen Text. Neben der Trennung von Titel und Textkörper werden Informationen über Zeilenenden kodiert, um die ursprüngliche Gestalt des Textes in wesentlichen Punkten rekonstruieren zu können. Schließlich wird jeder Text um linguistische Informationen angereichert, die im folgenden Abschnitt beschrieben werden.

5.5. Kodierung linguistischer Informationen

Das Korpus wurde erhoben, um die zu entwickelnden Analyseverfahren in ihrer Leistungsfähigkeit einschätzen zu können. Eine solche Analyse muss sich, um absolute Aussagen herleiten zu können, an eine Vorgabe, d.h. eine Vorklassifikation der zu analysierenden Daten halten. Für die Analyse orthographgischer Leistungen sind solche Vorgaben nur durch manuelle Annotation zu erhalten, die sich im vorliegenden Fall auf verschiedene Ebenen linguistischer Information bezieht. Gleichzeitig ist es durch die Vorgabe verschiedener Ebenen möglich, Teilverfahren zu testen, also solche, die davon ausgehen, dass Informationen anderer Ebenen, z.B. durch vorhergehende Analyseschritte, bereits verfügbar sind.

5.5.1. Satzgrenzen und Interpunktion

Ein Text gliedert sich syntaktisch in Sätze, die durch Punkte am Satzende voneinander abgegrenzt werden. Mit der Einheit Satz hängt ebenfalls ein Teilbereich der Groß- und Kleinschreibung, nämlich die Großschreibung des Satzanfangs zusammen. In potenziell fehlerhaften Texten können weder Satzendpunkte noch die Großschreibung am Satzanfang grundsätzlich als korrekt angenommen werden, so dass sich verschiedene Situationen ergeben:

1. Ein Satz wurde korrekt durch einen Satzendpunkt beendet.
2. Ein Satz wurde nicht durch einen Satzendpunkt beendet, d.h. ein Satzendpunkt wurde ausgelassen.
3. Ein Satzendpunkt wurde an einer Stelle eingefügt, an der kein Satz endet.

Zu Abbildung dieser Informationen wurde das Tag `punkt` als leeres Element (vgl. Lobin, 1999, 11f.) eingeführt, das über die Attribute `vorhanden` und `noetig` verfügt und beiden jeweils die Werte `ja` oder `nein` zuweist.[4] Die Attribute sind mit `"ja"` als Default belegt, so dass die Erfassung weniger arbeitsintensiv ist. Als Repräsentation der drei genannten Fälle ergeben sich also:

1. `<punkt vorhanden="ja" noetig="ja"/>`, Abkürzung: `<punkt/>`
2. `<punkt vorhanden="nein" noetig="ja"/>`, Abkürzung: `<punkt vorhanden="nein">`
3. `<punkt vorhanden="ja" noetig="nein"/>`

Aus dieser Kodierung lassen sich auch alle notwendigen Informationen zur Behandlung der Großschreibung am Satzanfang ableiten. Zusätzlich wurden allerdings Großschreibungen, die auf ihre Position am Satzanfang zurückgeführt werden können, mit dem Element `<sanf>...</sanf>` markiert. Dies ist dann von Vorteil, wenn Wortlisten, z.B. als Auflistung aller Schreibungen eines Wortes, extrahiert werden sollen, die nicht nach Großschreibung am Satzanfang differenzieren. Neben Punkten werden auch Doppelpunkte, Folgen von drei Punkten (als Auslassungs- oder Fortführungsmarkierung) sowie Ausrufe- und Fragezeichen als Satzendmarkierung verwendet, zusätzlich kann es zu Fehlverwendungen dieser Möglichkeiten kommen. Das Element `punkt` verfügt daher über zwei weitere Attribute, die das an dieser Stelle korrekte und das tatsächlich verwendete Zeichen angeben. So wird z.B. die Verwendung eines Punktes anstelle eines Fragezeichens als `<punkt vorhanden="ja" noetig="ja" zeichen="frage" fzeichen="punkt"/>` kodiert. Durch Verwendung des Wertes `"punkt"` als Default kann aber der Kodierungsaufwand auf diese seltenen Fälle beschränkt werden. Abschließend kann noch die Anzahl der verwendeten Zeichen kodiert werden. Das Attribut `anzahl` hat die Zahlen 1 – 6 als gültige Werte, wodurch z.B. die Verwendung von drei Ausrufezeichen erfasst werden kann. Von der Möglichkeit, auch Zeichenfolgen wie „?!?!?!“ kodieren zu können, wurde wegen des zu hohen Aufwandes und des zu geringen Nutzens Abstand genommen. Kommata werden analog zur Markierung der Satzendpunkte durch das Tag `<komma>` markiert.

Die vollständige Definition der Elemente `punkt` und `komma` stellt sich damit wie folgt dar:

```
<!ELEMENT punkt EMPTY>
     <!ATTLIST punkt vorhanden (ja | nein) "ja">
   <!ATTLIST noetig (ja | nein) "ja">
   <!ATTLIST zeichen (punkt | ausruf | frage | doppel | 3punkt) "punkt">
   <!ATTLIST fzeichen (punkt | ausruf | frage | doppel | komma | 3punkt) "punkt">
   <!ATTLIST anzahl (1 | 2 | 3 | 4 | 5 | 6) "1">

<!ELEMENT komma    EMPTY>
   <!ATTLIST komma vorhanden (ja | nein) "ja">
   <!ATTLIST komma noetig    (ja | nein) "ja">
   <!ATTLIST komma fzeichen  (punkt | ausruf |frage | doppel) "komma">
```

[4] Dadurch ergibt sich noch ein vierter Fall: An einer Stelle, an der kein Satz endet, wurde kein Satzendpunkt eingefügt. Dieser Fall gilt als Default-Fall und bleibt unmarkiert.

Als weiteres Interpunktionsphänomen wird schließlich die Markierung wörtlicher Rede repräsentiert. Wörtliche Rede ist in den Texten des Korpus ein sehr häufiges Phänomen, das allerdings in vielen Fällen nicht durch die Verwendung von Anführungszeichen gekennzeichnet wird. Um diese Leistungen zumindest ansatzweise erfassen zu können, wurde das Tag `<rede>` eingeführt, das den als wörtliche Rede zu interpretierenden Text umschließt. Als Attribut des Tags wird angegeben, ob die wörtliche Rede vollständig oder teilweise markiert wurde. Die genaue Art der Markierung wird dabei ebenso wenig erfasst wie das genaue Erfassen von wechselnden Sprechern in Dialogwiedergaben usw. Als einzige Differenzierung wird zwischen teilweiser Markierung: nur Beginn und teilweiser Markierung: nur Ende unterschieden. Die vollständige Defintion des Tags `<rede>` ist im Folgenden wiedergegeben:

```
<!ELEMENT rede     (#PCDATA | f | versuch | punkt | komma | sanf)*>
   <!ATTLIST rede  markiert  (ja|nein|anfang|ende) "nein">
```

5.5.2. Orthographische und grammatische Fehler

Das Korpus wurde als Testdatensatz für die zu entwickelnden automatischen Analysealgorithmen konzipiert. Um differenzierte Aussagen über die Qualität der Algorithmen und unterschiedliche Parametrisierungen über das Gesamtkorpus ableiten zu können, ist es notwendig, die Testläufe selbst ebenfalls zu automatisieren. Automatisierte Tests bestehen aus zwei Schritten; im ersten Schritt werden die Analysedaten durch das zu testende Verfahren generiert und in einem zweiten Schritt werden diese Ergebnisse mit gespeicherten Referenzanalysen verglichen. Referenzanalysen können entweder manuell unter Anwendung von Expertenwissen oder automatisiert durch Verfahren mit bekannten Leistungen erstellt werden. Da für das vorliegende Vorhaben keine Referenzalgorithmen zur Verfügung standen, wurde eine manuelle Voranalyse notwendig. Hier ist zwischen Eigenschaften des Textes und Eigenschaften unterschiedlicher Analysemodelle zu unterscheiden. Eine Eigenschaft des Textes ist insbesondere die „intendierte“ korrekte Schreibung des Textes, eine Eigenschaft eines Analysemodells z.B. die Annahme bestimmter Analyse- oder Fehlerkategorien. Es ist sinnvoll, die fest stehenden Eigenschaften des Textes fest mit der Repräsentation des Textes zu verbinden und spezielle verfahrensabhängige Analysen davon zu trennen.

Als feststehende Eigenschaften des Textes, die in der XML-Repräsentation des Korpus mitzukodieren sind, werden folgende Punkte definiert:

- Die „intendierte“ Schreibung. Dahinter verbirgt sich eine Interpretationsleistung bei der Korpuserfassung und -kodierung, da für jeden angenommenen Fehler nach der subjektiv wahrscheinlichsten Textform gesucht wird, die mit den Intentionen des Schreibers übereinstimmt. Da bei der Kodierung nur die fertigen handschriftlichen Texte vorlagen, erhöht sich die Interpretationsunsicherheit noch. Es war nicht möglich, die Schreiber noch einmal zu befragen und Zweifelsfälle auf diesem Wege auszuräumen.
- Eine Grobkategorisierung von Fehlern in Wortschreibungs-, Groß-/Kleinschreibungs-, Getrennt-/Zusammenschreibungs-, Grammatik- und Zeichensetzungsfehler. Diese Unterteilung ist im strengen Sinne eine Eigenschaft der Analyseverfahren, allerdings ließe sie sich mit einem sehr einfachen Algorithmus automatisieren und ist außerdem Grundlage der hier zu Grunde gelegten linguistischen Fundierung des Gegenstandsbereichs. Die Grobkategorisierung dient nur der Orientierung und kann von konkreten Analyseverfahren ignoriert werden.

Eine manuelle Analyse birgt immer die Fehlerquelle, Kategorisierungen inkonsistent vorzunehmen und Interpretationsleistungen zu unterschiedlichen Zeitpunkten und in unterschiedlichen Kontexten voneinander abweichend zu erbringen. Es gibt verschiedene Möglichkeiten, diese Fehlerquellen zu minimieren:

- Mehrmalige Analyse der Texte durch eine Person zu unterschiedlichen Zeitpunkten.
- Mehrmalige Analyse der Texte durch unterschiedliche Personen.
- Durchführung der Analyse anhand strenger Kriterien.

Da es Ziel der Untersuchung war, eine möglichst große Datenmenge als Testdaten zu gewinnen und nicht, möglichst fehlerarm kodiertes Material für empirische Untersuchungen von Schreibleistungen vorlegen zu können, wurde auf ein aufwendiges Kodierungsverfahren verzichtet. Die Hauptkodierung wurde von einer einzelnen Person innerhalb von 6 Wochen vorgenommen, Zweifelsfälle wurden mit einer zweiten Person besprochen und einzelne Texte stichprobenartig von einer zweiten Person überprüft. Grundsätzlich wurden die Analysen nach dem Prinzip vorgenommen, dass so wenige Fehler wie nötig angenommen wurden, d.h. bei mehreren Interpretationsmöglichkeiten stets diejenige gewählt wurde, die keinen oder die geringste Anzahl an angenommenen Fehlern zur Folge hatte.

Die Behandlung von Zeichensetzungsfehlern wurde bereits im letzten Abschnitt erläutert, für die anderen Fehlertypen wurde das Tag `<f>` eingeführt, das genau den fehlerbehafteten Text umschließt und als Attribute Informationen über die „intendierte“ Schreibung (im Folgenden: „Zielschreibung“) und die Grobklassifizierung des Fehlers enthält. Die Fehlerklassen sind wie folgt definiert:

Groß- und Kleinschreibungsfehler: Die beobachtete Schreibung und die Zielschreibung eines Einzelwortes unterscheiden sich lediglich in der Groß- und Kleinschreibung des initialen Buchstabens.

Getrennt- und Zusammenschreibungsfehler: Werden sowohl bei der beobachteten Schreibung als auch bei der davon abweichenden Zielschreibung alle Leerräume entfernt, ergeben sich identische Zeichenketten. Dieser Fehlertyp kann sich sowohl aufseiten der beobachteten Schreibung als auch bei der Zielschreibung auf Wortgruppen beziehen.

Grammatikfehler: Entspricht die beobachtete Schreibung einer korrekt bzw. regelkonform geschriebenen Wortform und ist anzunehmen, dass der Schreiber diese Wortform intendiert hat und führt die Annahme dieser Wortform zu einer morphologisch, morphosyntaktisch oder syntaktischen ungrammatischen Konstruktion, dann wird ein Grammatikfehler angenommen. Beispiele: Bildung einer falschen Wortform (*<gehte> für <ging>), Kongruenzfehler (*<die Mann ging> für <der Mann ging>).

Wortschreibungsfehler: Führt die Annahme der intendierten Form nicht zu einem Grammatikfehler und weicht die beobachtete Schreibung von der Zielschreibung in einer anderen als den in Punkt 1 und 2 beschriebenen Form ab oder treten zu den dort beschriebenen Abweichungen noch weitere hinzu, wird ein Wortschreibungsfehler angenommen.

Kombinierter Grammatik- und Wortschreibungsfehler: Führt die Annahme der intendierten Form zu einem Grammatikfehler und stimmt die beobachtete Schreibung nicht mit der intendierten Schreibung überein, wird der Fehler als kombinierter Grammatik- und Wortschreibungsfehler klassifiziert. Zudem gibt es zwei Zielschreibungen: Zum einen die grammatisch und orthographisch korrekte Form, zum anderen die orthographisch korrekte[5] aber ungrammatische Form. Beispiel: *<gete> für *<gehte> für <ging>.

Die formale Definition des `<f>`-Tags ergibt sich entsprechend:

```
<!ELEMENT f     (#PCDATA | punkt | komma | sanf)*>
   <!ATTLIST f  ortho  CDATA #IMPLIED>
   <!ATTLIST f  fortho CDATA #IMPLIED>
   <!ATTLIST f  typ (o|g|gks|og) "o">
```

Die Attribute bedeuten im Einzelnen:

`ortho="..."`	orthographisch und grammatisch korrekte Zielschreibung
`fortho="..."`	grammatisch inkorrekte aber orthographisch „korrekte“ Zielschreibung
`typ="o"`	Wortschreibungsfehler
`typ="g"`	Grammatikfehler
`typ="gsk"`	Groß- und Kleinschreibungsfehler
`typ="og"`	Kombinierter Wortschreibungs- und Grammatikfehler

Bei der Großschreibung unterscheidet sich die Großschreibung am Satzanfang deutlich von anderen Reguläritäten und wird daher in den meisten Darstellungen getrennt behandelt. Um Zählungen auf oberer Ebene zu erleichtern, wurden im Korpus Satzanfänge speziell markiert. Prinzipiell lässt sich diese Information auch aus den Angaben über Satzendzeichen erschließen, dazu ist allerdings eine kontextabhängige Verarbeitung bzw. Vorverarbeitung notwendig. Um solche zusätzlichen Schritte einzusparen bzw. nur einmal durchführen zu müssen, wurde das Tag `sanf` eingeführt, das einzelne Wortformen als Satzanfang markieren kann. Die Satzanfangsmarkierung ist dabei das innerste Markierungsmerkmal und steht z.B. innerhalb von `f`-Tags.

5.6. Metadaten

Das Korpus ist als Sammlung von Texten von Schreibern organisiert, die wiederum in (Schul-) Klassen zusammengefasst sind. Damit ist es möglich, mehrere Texte desselben Schreibers und Klassen als Gruppen von Schreibern zu untersuchen. Diese Aufteilung ermöglicht es auch, Metadaten auf jeder dieser Ebenen zu repräsentieren, um Redundanzen in der Codierung zu vermeiden und auf die für eine Ebene relevanten Daten leicht zugreifen zu können.

[5] Im Falle von nicht existierenden Wortformen wie *<gehte> ist hier wiederum ein Interpretationsspielraum gegeben. Es wird jeweils angenommen, dass die nicht existente aber orthographisch mögliche Form soweit wie möglich den orthographischen Regularitäten folgt. Im Falle von *<gehte> würde also die Notwendigkeit angenommen, ein „vererbtes“ silbentrennendes <h> zu markieren.

5.6.1. Klassenebene

Die oberste Gliederungsebene ist aufgrund der besonderen Situation der Korspuserhebung die Ebene der Schulklassen. Ein Teil der auf dieser Ebene repräsentierten Metadaten sind tatsächlich nur für Schulklassen als relativ homogene Gruppen von Schülern sinnvoll, ein anderer Teil ist auch für andere Gruppen geeignet. Innerhalb des Korpus gibt es eine Gruppe, die keine Schulklasse darstellt, nämlich eine außerschulische Fördergruppe gemischten Alters. Für diesen Einzelfall ist die Repräsentation als Schulklasse vertretbar, für eine breitere Datenbasis mit einem größeren Anteil von Gruppen, die keine Schulklassen sind, müsste eine andere Darstellung gefunden werden, die z.B. die schulklassenspezifischen Merkmale ausnahmsweise auch als Eigenschaften einzelner Schüler zulässt.

Jeder Klasse wurde eine eindeutige Kennziffer (ID) der Form `k001` bis `k038` zugewiesen, die Reihenfolge entspricht der Reihenfolge des Eingangs der erhobenen Daten. Den Daten erhebenden Lehrerinnen und Lehrern wurde ein klassenspezifischer Fragebogen zur Verfügung gestellt, der die Grundlage für die nachfolgend dargestellte Codierung bildet. Nicht alle Informationen, die in den ausgefüllten Fragebögen enthalten waren und z.T. durch die Fragen direkt abgefragt wurden, sind im Korpus repräsentiert. So wurden die meisten der Fragen als offene Fragen formuliert und aus den vorhandenen Antworten nachträglich Skalen oder Kategorien gebildet. Zusätzlich Informationen wurden in einem freien Textfeld codiert, so dass sie zwar im Korpus vorhanden, aber einer maschinellen Auswertung nicht unmittelbar zugänglich sind. Generell ist festzuhalten, dass die Informationen vor allem erhoben wurden, um informelle Hintergrundinformationen über die Datensätze zu gewinnen und nicht, um umfangreiche statistische Analysen unter Einbeziehung der Informationen durchzuführen.

Grunddaten

Jede Klasse bekommt eine eindeutige Kennzeichnung (`id`) innerhalb des Korpus. Der Name des Schulortes (`ort`) und die ersten beiden Ziffern der Postleitzahl des Schulortes (`plz`) werden kodiert. Diese Kennzeichnung kann als grobe Zuordnung einer Klasse zu einer Dialektregion dienen. Um mehrere Klassen einer Schule miteinander in Beziehung setzen zu können, bekommt jede Schule ebenfalls eine eindeutige Kennzeichnung (`schule`) zugewiesen. Diese Kennzeichnung hat die Form `s001` - `s036`. Die Klassenstufe (`stufe`)wird als 0 (unbekannt/gemischt) oder als Zahl (1 = 1. Klasse, 2 = 2. Klasse, ...) gespeichert.

Unterrichtsform

Der versandte Fragebogen erhebt die in der Klasse vorherrschende Unterrichtsform in Form einer Freitextangabe. Die von den Lehrkräften gemachten Angaben wurden gesichtet und zusammengefasst. Hierbei geht es nur um eine grobe Einschätzung der üblichen Unterrichtsformen, so dass insgesamt sechs verschiedene Formen als ausreichend angesehen werden können: Offener Unterricht, Frontalunterricht, Werkstattunterricht, Unterricht mit Montessori-Materialien und Wochenplanunterricht. In den meisten Fällen wurden mehrere Formen angegeben, häufig anzutreffen war auch die Nennung ein oder zwei konkreter Formen und der Angabe „gemischt“. Um eine möglichst umfassende Repräsentation zu gewährleisten und nicht im Einzelfall vereinfachende und evtl. verfälschende Entscheidungen treffen zu müssen, wurde eine Codierung gewählt, die beliebige Kombinationen der

fünf genannten Formen und der Angabe „gemischt“ ermöglicht. Dazu wurden numerische Codierungen gewählt, die addiert werden können: 0 = unbekannt, 1 = offen, 2 = frontal, 4 = Werkstatt, 8 = Montessori-Materialien, 16 = Wochenplanarbeit, 32 = gemischt. Aus einer solcherart gebildeten Summe lassen sich die Einzelformen rekonstruieren, der Wert 37 z.B. lässt sich eindeutig auf 32 + 4 + 1 = offen + Werkstatt + gemischt zurückführen.

Schrifteinführung

Die vorrangig verwendete Methode der Schrifteinführung wurde in Form einer zweistufigen Frage abgefragt. Die erste Frage lautete „Wurde eine Fibel verwendet? (Welche?)“, die zweite „Wenn nein: Wie haben Sie die Schrift eingeführt?“ Aus den Antworten auf beide Fragen wurden fünf Kategorien gebildet: Fibel (Codierung: 1), Anlauttabelle (2), „Phonetisches Schreiben“ (4), „Häusermodell“ (8) und Grundwortschatz (16). Dazu kommt die Codierung 0 = unbekannt, es sind ebenfalls wiederum eindeutige Kombinationen der Kategorien als Zahlenwert repräsentiertbar. Konkrete Angaben zur verwendeten Fibel wurden unter „Bemerkungen“ (s.u.) kodiert.

Freie Texte

Um abschätzen zu können, wie vertraut die Schülerinnen und Schüler mit der gestellten Aufgabe waren, eine Bildergeschichte frei nachzuerzählen und zu verschriftlichen, wurde zum Einen nach der grundsätzlichen Häufigkeit freien Schreibens gefragt. Die Antworten sind an dieser Stelle besonders mit Vorsicht zu genießen, da der verwendete Begriff „freier Text“ unterschiedlich interpretiert wurde. In einigen Fällen umfasste die Interpretation die vorliegende Bildergeschichten-Aufgabe, in anderen Fällen nicht. Ursache ist eine unterschiedliche Auslegung des „Freiheitsbegriffs“ in der Textproduktion: Das eine Extrem ist das völlige Fehlen äußerer Vorgaben wie z.B. einer nachzuerzählenden Geschichte, das andere jede Form von Textproduktion, die nicht auf einer exakt zu reproduzierenden Vorgabe beruht. Die offenen Antworten auf die Frage „Seit wann verfassen die Kinder freie Texte, wie häufig?“ wurden gemäß einer fünfteiligen Skala kategorisiert, die folgende Werte umfasst: 0 = unbekannt, 1 = noch nie, 2 = selten, 3 = unregelmäßig, 4 = häufig, 5 =sehr häufig.

Bildergeschichten

Als Spezialisierung der Fragen nach freien Texten wurde gefragt, ob bereits ähnliche Bildergeschichten verschriftet wurden. Die Antworten wurden in die gleiche Skala eingeordnet, die auch für die vorige Frage verwendet wurde.

Code	Bedeutung
0	unbekannt / nicht angegeben
1	nicht behandelt
2	in Einzelfällen / Satzanfang (GKS)
3	wiederholt in Einzelfällen, am Grundwortschatz / Punkt (Zeichens.); Nomen, Verben (GKS)
4	systematisch / Punkt, Frage, Ausruf (Zeichens.)
5	sehr ausführlich /+Komma (Zeichens.)

Tabelle 5.1.: Codierung der Angaben zur Einführung orthographischer Regeln

Sprachbuchgebrauch

Die Frage, ob und wie intensiv Sprachbücher im Deutschunterricht verwendet werden, hängt mit der vorherrschenden Unterrichtsform zusammen, ist aber dennoch als relativ unabhängige Frage zu betrachten und kann zu einer weiteren Klärung der freien Angaben zur Unterrichtsform dienen. Aus diesem Grund wurden zwei Fragen nach dem verwendeten Sprachbuch und der Häufigkeit des Sprachbuchgebrauchs aufgenommen. Die freien Antworten auf die erste Frage wurden als Bemerkung (s.u.) aufgenommen. Für die zweite Frage wurden drei Antwortalternativen vorgegeben: unter 50%, 50-75% und 100% des Sprachunterrichts. Für die Kodierung wurden die Alternativen um zwei Werte ergänzt: 0 = unbekannt, 1 = gar nicht, 2 = unter 50%, 3 = 50-75%, 4= 100%.

Einführung von orthographischen Regeln

Neben den eher allgemeinen Fragen zum Deutschunterricht sollte erfasst werden, wie einzelne orthographische Phänomene im Unterricht behandelt wurden. Unter der Frage „Wie detailliert wurden Regeln zu den folgenden orthographischen Phänomenen im Unterricht behandelt? (Und welche?)“ wurden neun Phänomenbereiche aufgeführt, die jeweils mit einer freien Antwort versehen werden konnten. Bei der Sichtung der Antworten hat sich herausgestellt, dass einige Phänomene den Lehrkräften nicht geläufig waren bzw. es wiederum zu unterschiedlichen Interpretationen gekommen ist. Insbesondere konnten die Antworten nicht eindeutig auf linguistisch fundierte Regeln und Regelalternativen abgebildet werden. Die Antworten waren auch ansonsten nicht sehr detailliert, so dass sich eine Codierung auf einer Skala von 1 (nicht behandelt) bis 5 (sehr ausführlich behandelt) als die sinnvollste Lösung herausgestellt hat. Lediglich bei den Komplexen Groß- und Kleinschreibung sowie Zeichensetzung konnten klare Regeltypen identifiziert werden, die sich aber ebenfalls auf die Skala abbilden ließen. Für die Groß- und Kleinschreibung waren dies die Großschreibung am Satzanfang (codiert als 2) und die wortartenbezogene Groß- und Kleinschreibung von Nomen und Verben (codiert als 3), für die Zeichensetzung die Setzung von Punkten (codiert als 3), die Verwendung von Punkten, Ausrufe- und Fragezeichen (codiert als 4) und die Einführung von Kommaregeln (codiert als 5). Tabelle 5.1 fasst die Codierung zusammen.

Die neun Phänomenbereiche wurden ohne detaillierte Erläuterungen aufgeführt, in einigen Fällen wurden jedoch Beispiele angegeben, um evtl. unbekannte Begriffe zu klären. In der folgenden Auflistung sind die gegebenen Beispiele enthalten.

1. Laut-Buchstaben-Zuordnung

2. Schärfung (z.B. <kommen>, <Hütte>)
3. Dehnung (z.B. <fahren>, <Lehrer>)
4. Silbentrennendes <h> (z.B. <sehen>, <Ruhe>)
5. <s>/<ss>/<ß>-Schreibung
6. Morphologische Konstantschreibung (z.B. <Hund> wg. <Hunde>, <kommt> wg. <kommen>, <fahren> wg. <fährt>)
7. Groß-/Kleinschreibung
8. Getrennt-/Zusammenschreibung
9. Zeichensetzung

Computereinsatz im Unterricht

Der Einsatz von Rechnern im Deutschunterricht variiert stark zwischen einzelnen Schulen und Klassen, sowohl die grundsätzliche Rechnerausstattung als auch den didaktischen Einsatz von Software betreffend. Um einen groben Eindruck von der Intensität des Computereinsatzes in den untersuchten Klassen zu gewinnen, wurde die Frage „Setzen Sie Computer im Unterricht ein? (Hard- und Softwareausstattung)“ aufgenommen. Nur in neun der 36 Klassen wurde diese Frage positiv beantwortet, detaillierte Angaben wurden nur in fünf Fällen gemacht, so dass auf eine genauere Auswertung verzichtet wurde.

Entstehung der Texte

Das Begleitschreiben zur Korpuserhebung enthielt keine exakten Anweisungen zur Erhebung der Texte, lediglich die Aufforderung, „freie Texte zu der beiligenden Bildergeschichte verfassen zu lassen“. Vor diesem Hintergrund kann es besonders wichtig sein, die genaueren Begleitumstände der Textentstehung zu erfassen. Der Fragebogen ermöglichte Anmerkungen zu Vorbesprechung und Erläuterungen, zur Art der gegebenen Hilfestellung und zu der Frage, ob die Texte handschriftlich oder am Computer erstellt wurden.

Der genaue Zeitraum der Textentstehung wurde ebenfalls abgefragt, vor allem, um eine Einordnung in das laufende Schuljahr zu ermöglichen. In den allermeisten Fällen lag der Entstehungszeitraum zwischen April und Mai 1999. Im Korpus wurden Angaben zum Entstehungszeitraum frei unter dem Tag `datum` erfasst. Die Angaben zu näheren Umständen der Entstehung wurden aufgeteilt in `hilfestellung` und `erlaeuterungen` und unter einem Tag `entstehung` zusammengefasst.

Kinder nichtdeutscher Familiensprache

Die Lehrkräfte wurden gebeten, unter der Frage „Sind in ihrer Klasse Kinder nicht-deutscher Familiensprache? (Welche?)“ einen Überblick über die sprachliche Situation in der Gesamtklasse zu geben. Für Detailangaben gab es zusätzliche Möglichkeiten bei den einzelnen Schülerinnen und Schülern. Die Angaben wurden im Korpus frei unter dem Tag `nichtdeutsch` angegeben. Auf eine genaue Aufschlüsselung der Angaben wird an dieser Stelle verzichtet, da ein Großteil der Lehrkräfte auf Angaben für die Gesamtklasse verzichtet hat.

Weitere Bemerkungen

Alle weiteren Anmerkungen, die sich auf die Gesamtklasse bezogen, wurden in einem Tag `bemerkungen` gesammelt. Das waren zum einen direkte Antworten auf die Frage nach weiteren, besonderen Bemerkungen, zum anderen Zusatzangaben zu den anderen Fragen wie z.B. Titel der verwendeten Fibel oder des verwendeten Sprachbuchs, genauere Erläuterungen zur Unterrichtsform, orthographischen Regeln etc.

DTD

Alle klassenbezogenen Informationen werden entweder als Attribute des Tag `klasse` oder unter dem Tag `info` innerhalb von `klasse` codiert. Der vollständige DTD-Ausschnitt für die Repräsentation von Metadaten auf Klassenebene ist nochfolgend wiedergegeben.

```
<!ELEMENT klasse   (info, person+)>
   <!ATTLIST klasse id CDATA #REQUIRED>
   <!ATTLIST klasse ort CDATA "unbekannt">
   <!ATTLIST klasse plz CDATA "0">
   <!ATTLIST klasse schule CDATA "0">
   <!ATTLIST klasse uform CDATA "0">
   <!ATTLIST klasse stufe (0|1|2|3|4) "0">
   <!ATTLIST klasse schrift CDATA "0">
   <!ATTLIST klasse freie_texte (0|1|2|3|4|5) "0">
   <!ATTLIST klasse bildergeschichten (0|1|2|3|4|5) "0">
   <!ATTLIST klasse sprachbuch (0|1|2|3|4) "0">
   <!ATTLIST klasse regeln CDATA "0,0,0,0,0,0,0,0,0">
   <!ATTLIST klasse computer CDATA "nein">

<!ELEMENT info     (datum?, entstehung, nichtdeutsch?, besonderheiten?)>

   <!ELEMENT datum                (#PCDATA)>
   <!ELEMENT entstehung           (hilfestellung, erlaeuterungen)>
       <!ATTLIST entstehung typ (hand|computer) "hand">
   <!ELEMENT nichtdeutsch         (#PCDATA)>
   <!ELEMENT besonderheiten       (#PCDATA)>
```

```
<!ELEMENT hilfestellung       (#PCDATA)>
<!ELEMENT erlaeuterungen      (#PCDATA)>
```

5.6.2. Schreiberebene

Grunddaten

Ebenso wie Klassen und Schulen werden einzelne Schreiber durch eine ID gekennzeichnet. Diese ID ist nicht global, sondern nur innerhalb einer Klasse eindeutig und setzt sich nach dem Muster `s00` - `s99` zusammen. Zusammen mit der ID kann der Name des Schülers als freier Text gespeichert werden. Aus Datenschutzgründen sind im vorliegenden Korpus die Namen entfernt und die Datensätze somit anonymisiert worden. Als weitere Grunddaten können Geschlecht und Alter gespeichert werden.

Im Fragebogen wurde Gelegenheit gegeben, weitere Informationen über einzelne Schüler anzugeben. Diese Informationen sollten hauptsächlich dazu dienen, den sprachlichen Hintergrund und sprachliche Besonderheiten zu erheben. Der Fragebogen enthielt daher mehrere Abschnitte, die wie folgt aussahen:

Schüler(in):
1. Stammt nicht aus dieser Region, sondern ...
2. Sprachliche Auffälligkeiten (z.B. starker Dialekt) ...
3. Erhält besondere Förderung: ...
4. Weitere Anmerkungen: ...

Diese Angaben wurden nicht für alle Schüler vorgenommen, sondern nur für diejenigen, bei denen nach Auffassung der Lehrkraft Besonderheiten vorlagen. Für alle anderen Schülerinnen werden Default-Werte angenommen (s.u.). In den allermeisten Fällen war jeder einzelne Text mit dem Namen des Schreibers versehen, Informationen über Geschlecht und Alter wurden nicht explizit erhoben. Teilweise wurden die Texte nur mit Initialen oder Kürzeln versehen, so dass eine Codierung des Geschlechts hier nicht möglich war. Insgesamt ist das Alter im Korpus gar nicht, das Geschlecht nur dort repräsentiert, wo es sich eindeutig aus dem Vornamen ableiten ließ.

Code	Bedeutung
0	unbekannt
1	wie Klasse
freier Text	andere Herkunft

Tabelle 5.2.: Kodierung der Herkunft einzelner Schreiber

Code	Bedeutung
0	unbekannt
1	deutsch
2	deutsch, leichter örtl. Dialekt
3	deutsch, starker örtl. Dialekt
8	bilingual deutsch/gem. Herkunft
9	gemäß Herkunft
10	gemäß Herkunft, gute Deutschkenntnisse
freier Text	andere Angaben

Tabelle 5.3.: Kodierung von Muttersprache und Dialekt einzelner Schreiber

Herkunft und Muttersprache

Zu jeder Klasse wurden bereits Informationen über den Schulort und Schüler nichtdeutscher Familiensprache gespeichert. Genauere bzw. abweichende Informationen über Herkunft und Muttersprache und Dialekt können für jeden einzelnen Schreiber erfasst werden. An dieser Stelle wurde eine teilweise numerische Kodierung gewählt, die in den Tabellen 5.2 und 5.3 dargestellt ist.

Die Quellen der Informationen über Muttersprache und Dialekt waren die Fragen nach Herkunft und sprachlichen Auffälligkeiten aus dem Fragebogen. Die Codes wurden aus einer Durchsicht der freien Angaben gewonnen und sinnvoll zusammengefasst. Aus der Auflistung wird deutlich, dass es sich lediglich um eine wenig systematische, grobe Kategorisierung handelt, die zudem von unterschiedlichen Interpretationen der Lehrkräfte beeinträchtigt wird. Insbesondere hinsichtlich der Dialektfrage sind starke Unterschiede in der Auslegung anzunehmen, z.B. stellte eine Lehrerin aus dem nordwestdeutschen Raum fest, dass ihre Schüler keinen Dialekt sprächen. Die Angabe „gemäß Herkunft“ bezieht sich auf alle Abweichungen der Herkunft vom Klassenort. In den meisten Fällen ist hier ein Land angegeben, aus dem die Muttersprache erschlossen wurde, in einigen Fällen aber auch eine andere Region des deutschen Sprachraumes, was begrenzte Rückschlüsse auf abweichende dialektale Bedingungen zulässt.

Besondere Förderung

Die Angaben zu einer besonderen Förderung einelner Schüler wurde wie in Tabelle 5.4 aufgeschlüsselt codiert. Wie bei den vorangegangenen Codierungen wurde auch hier eine Liste häufiger Antworten erstellt und zu groben Kategorien zusammengefasst.

Code	Bedeutung
0	unbekannt
1	keine
2	Deutsch als Fremdsprache/Zweitsprache
3	Lesen, Rechtschreiben
4	Logopädie
5	im Rahmen der Binnendifferenzierung
freier Text	andere Angabe

Tabelle 5.4.: Kodierung der besonderen Förderung einzelner Schreiber

Inhalt	Anzahl	Anteil
Sprachliche Informationen/Beurteilung	63	44,1%
Sonstige schulische Informationen/Beurteilung	24	16,8%
Familiäre Informationen	30	21,8%
Entstehungsinformationen/Hintergrund	21	13,9%
Metakommentar	21	13,9%

Tabelle 5.5.: Grobkategorisierung freier Anmerkungen zu einzelnen Schreibern

Anmerkungen

Von der Möglichkeit, weitere Anmerkungen zu einzelnen Schülern als freien Text anzugeben, wurde sehr unterschiedlich Gebrauch gemacht. Daher wurden die Angaben für das Korpus nicht systematisiert, sondern direkt übernommen. In einigen Fällen wurden Metakommentare zur Codierung mit aufgenommen, wie Anmerkungen über offensichtlich abgeschriebene Texte. Eine grobe inhatliche Aufschlüsselung ist in Tabelle 5.5 aufgeführt, in Tabelle 5.6 finden sich beispielhafte Angaben. Soweit zur Erklärung sinnvoll, werden die Angaben im Folgenden bei konkreten Auswertungsbeispielen herangezogen.

DTD

Der vollständige DTD-Ausschnitt zu den Metadaten für einzelne Schreiber stellt sich wie folgt dar:

```
<!ELEMENT person (anmerkungen?, text+)>
   <!ATTLIST person id CDATA #REQUIRED>
   <!ATTLIST person name CDATA "unbekannt">
   <!ATTLIST person alter CDATA "0">
   <!ATTLIST person geschlecht (m|w|u) "u">
   <!ATTLIST person herkunft CDATA "1">
   <!ATTLIST person muttersprache CDATA "1">
   <!ATTLIST person foerderung CDATA "1">
```

Sprachliche Informationen/Beurteilung
- „Bei dem Schüler wurde eine Lese-Rechtschreib-Schwäche diagnostiziert.“
- „Linkshänder, verwechselt immer wieder b,d“
- „Hat erst mit drei Jahren sprechen gelernt“
- „keine häusliche Übung, sehr viele Abschreibfehler, harte Aussprache“
- „Sprach überhaupt nicht bis vor einigen Wochen, hatte große Probleme, Schriftliches von sich zu geben. Dies ist die erste eigene Geschichte.“
Sonstige schulische Informationen/Beurteilung
- „private Förderung seit Jahren, graphomotorischen und Wahrnehmungsstörungen“
- „Eltern sehr interessiert an schulischer Leistung“
- „Sehr schwacher Schüler, wurde schon mit Bedenken eingeschult. Wird wiederholen.“
- „sonderpädagogischer Förderbedarf“
- „kommuniziert nicht im Unterricht“
Familiäre Informationen
- „Asylanten, lernt seit 2 Jahren deutsch, hat außerhalb der Schule kaum Kontakt zu dt. Kindern“
- „Mutter spricht nur fehlerhaftes Deutsch. Leistungsstarke Schülerin, aber ohne häusliche Förderung“
- „deutscher Vater, amrikanische Mutter, deutsch+englisch zu Hause“
- „Elterntrennung belastet sie sehr“
Entstehungsinformationen/Hintergrund
- „Hat die Geschichte überwiegend von der Nachbarin abgeschrieben“
- „Klasse 2, Text in außerschulischer LRS-Förderung entstanden, ohne Hilfestellung.“
- „Gipsarm, geschrieben von der Lehrerin, daher keine Fehler“
- „Hatte heute keine Lust, er wollte die Geschichte nicht mehr zu Ende schreiben“
Metakommentar
- „Text (modulo Fehler) fast identisch mit s02“
- weitere ähnliche Kommentare zu auffällig ähnlichen Texten

Tabelle 5.6.: Beispiele freier Anmerkungen zu einzelnen Schreiber

Bezug	Anzahl	Anteil
Bildergeschichte 1	520	72,8%
Bildergeschichte 2	35	4,9%
Bildergeschichte 3	57	8,0%
Bildergeschichte 4	93	13,0%
Erlebnisbericht	9	% 1,3

Tabelle 5.7.: Bezüge der Texte

```
<!ELEMENT anmerkungen       (#PCDATA)>
```

5.6.3. Textebene

Für jeden Schreiber ist es möglich, mehrere Texte zu erfassen. Im vorliegenden Fall sind alle Texte Bildergeschichtenverschriftungen oder freie Erzählungen, so dass auf eine tiefer gehende Klassifizierung der Texte verzichtet wurde. So bestehen die Metadaten zu einzelnen Texten lediglich aus einer ID, der Angabe eines Bezugs, wie z.B. einer bestimmten Bildergeschichte und der Trennung von Überschrift und eigentlichem Text. Überschrift und Text selbst sind wie in 5.5 beschrieben kodiert. Tabelle 5.7 gibt eine Übersicht über die Bezüge der Texte.

Der vollständige DTD-Ausschnitt zu den Metadaten für einzelne Texte stellt sich wie folgt dar:

```
<!ELEMENT text     (titel?, body)>
   <!ATTLIST text     bezug CDATA #REQUIRED>
   <!ATTLIST text     id CDATA "t01">
```

5.7. Beispiel

Ein Beispiel soll das Format der Daten und die Art der zusätzlich vermerkten Informationen verdeutlichen:

```
<klasse id="k001" ort="Freiburg" plz="79" stufe="2">

<info>
  <datum>2/1999</datum>
  <entstehung>
     <hilfestellung></hilfestellung>
     <erlaeuterungen></erlaeuterungen>
  </entstehung>
  <nichtdeutsch></nichtdeutsch>
  <besonderheiten>
  </besonderheiten>
```

```
</info>

<person id="s01" name="Alex" geschlecht="m">
  <text bezug="bildergeschichte01">
    <titel><sanf>Ein</sanf> armer Hund</titel>
    <body>
      <sanf>Ein</sanf> alter <f ortho="Mann">man</f>
      <f ortho="guckt">ckugt</f> aus dem Haus
      <punkt vorhanden="nein"/>
      er <f ortho="sieht">sieht</f> <f typ="g" ortho="einen">ein</f> &ze;
      Hund
      <punkt vorhanden="nein"/>
      er <f ortho="geht">gehd</f> zur <f typ="gks">tür</f>
      <punkt vorhanden="nein"/>
      der Hund <f ortho="rennt">rend</f> <f ortho="weg">veg</f>
      <punkt vorhanden="nein"/>
      der <f ortho="Mann">man</f> &ze; <f ortho="folgt">volgt</f>
      den <f ortho="Fußspuren">vusspuren</f>
      <punkt vorhanden="nein"/>
      <f ortho="dann">dan</f> kommt er zu einer
      <f ortho="Hundehütte">Hundehüte</f>
      <punkt vorhanden="nein"/> &ze;
      da ist der Hund <punkt/>
    </body>
  </text>
</person>
</klasse>
```

5.8. Verarbeitungssoftware

Durch die Verwendung von XML zur Korpusrepräsentation ist es möglich, standardisierte Routinen zur Verarbeitung der Daten zu nutzen. Ein XML-Parser stellt allerdings nur die syntaktische Verarbeitungslogik zur Verfügung, die semantische Interpretation der Daten muss jedes Anwendungsprogramm DTD-spezifisch implementieren. XML-Parser können generell validierend oder nicht-validierend sein, ein validierender Parser überprüft außerhalb der anwendungsspezifischen Logik, ob die XML-Datei der angegebenen DTD entspricht, ein nicht-validierender Parser überprüft lediglich, ob die Datei grundsätzlich XML-konform ist, d.h. öffnende und schließende Tags ausbalanciert sind und spezielle Kodierungen und Verweise (Entitäten) korrekt verwendet werden.

Grundsätzlich gibt es zwei verbreitet Typen von XML-Parsern, SAX-Parser und DOM-Parser. Ein SAX-Parser arbeitet nach dem Callback-Delegate-Prinzip, der Parser liest die XML-Datei(en) sukzessive ein und benachrichtigt die umgebende Applikation über jeden neu beginnenden bzw. endenden XML-Tag und dazwischen liegende Datenblöcke. Die Anwendung implementiert typischerweise eine Subklasse eines abstrakten SAX-Parsers und enthält switch-case-Anweisungen, die öffnende bzw. schließende Tags verarbeiten. Dabei ist die Anwendung selbst für die Buchführung zuständig, d.h. sie merkt sich, typischerweise mithilfe von Instanzvariablen, in welchem Zustand

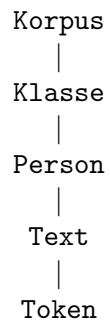

Abbildung 5.2.: Inklusionsbeziehung der Korpusverarbeitungsklassen

sich der Parser bezüglich der XML-Struktur befindet. Auf diese Weise kann die Anwendung weitestgehend selbst entscheiden, welche Daten berücksichtigt und in welcher Weise sie verarbeitet und gespeichert werden sollen. SAX-Parser sind insbesondere dann sinnvoll, wenn nicht alle Inhalte einer XML-Datei verarbeitet werden sollen, wenn die Datei zu groß ist, um vollständig in den Speicher geladen werden zu können und wenn einfache Aufgaben besonders effizient erledigt werden sollen.

Ein DOM-Parser baut zunächst eine vollständige interne baumartige Repräsentation des XML-Dokumentes auf und reicht diese im Ganzen an die umgebende Anwendung weiter. Der Parse-Tree wird üblicherweise als Instanz eines DOM-Objektes geliefert und bietet eine Reihe verschiedenartiger Methoden zur Baum-Traversierung und zur Manipulation des Baumes. DOM-Parser eignen sich für abstraktere Werkzeuge, da die Anwendung weniger Kenntnis der konkreten DTD-Struktur haben muss, als bei einem SAX-Parser. DOM-Parser vereinfachen zudem die Umstrukturierung von Informationen und Aufgaben, bei denen die Daten nicht sequenziell verarbeitet werden können.

Die Programmiersprache Python bietet integrierten XML-Support und stellt sowohl SAX- als auch DOM-Parser zur Verfügung. Da sie zudem über umfangreiche Möglichkeiten zur Stringverarbeitung verfügt und sich gut eignet, Algorithmen auf einer relativ hohen Abstraktionsebene zu implementieren, wurde sie für die Implementation der meisten hier vorgestellten Algorithmen verwendet.

Grundlage der Implementationen bildet eine objektorientierte Klassenbibliothek, die die aus dem Korpus stammenden Testdaten einlesen und auf verschiedene Weise aufbereiten kann. Die Klassen folgen eng der Schichtung von Ebenen des Korpus, wie die Abbildung 5.2 verdeutlicht. Ein Objekt der Klasse `KorpusReader` stellt die Methode `parse` zur Verfügung, mit der mehrere Korpusdateien eingelesen und in Ptyhon-Objekte verkapselt werden können. Das so entstehende `korpus`-Objekt kann von den nachfolgend definierten Algorithmen genutzt werden, um auf alle enthaltenen Informationen zuzugreifen und über verschiedene Ebenen des Korpus (Klassen, Personen, Texte) zu iterieren.

6. Verfahren ohne Informationen über die Zielschreibung

Die Analyse von Rechtschreibleistungen wird nachfolgend grundsätzlich in zwei Fälle unterschieden: (a) Solche, in denen die „intendierte Schreibung“ oder „Zielschreibung“, d.h. die korrekte Form des Wortes oder des Textes, das bzw. den der Schreiber schreiben wollte, unbekannt ist und (b) solche, in denen die Zielschreibung dem Analyseverfahren zugänglich ist.

6.1. Verfahren ohne Zusatzinformationen

Das einfachste denkbare Analyseverfahren müsste sich ausschließlich auf den zu analysierenden Text stützen. Es dürften weder Informationen über korrekte Schreibungen noch lexikalische Ressourcen oder Vorannahmen irgendeiner Art verwendet werden. Die Daten, auf deren Grundlage die Analyse vorgenommen werden soll, wären in diesem Falle ausschließlich die Ketten von Buchstaben, Leer- und Satzzeichen, die die Ausgangsschreibung bilden.

Ein solches Verfahren ist nicht implementierbar, weil jede Aussage der Art „Die Schreibung `ckugt` enthält Fehler.“ oder „Die Schreibung `guckt` ist korrekt.“ bereits A-priori-Wissen über die deutsche Rechtschreibung verwendet. Dass die Zeichenkette `ckugt` kein deutsches Wort repräsentieren kann, ist keine inhärente Eigenschaft der Zeichenkette, sondern eine Eigenschaft des orthographischen Systems des Deutschen. Daraus folgt: Ein orthographisches Analyseverfahren muss in irgendeiner Form auf Wissen über das den Schreibungen zu Grunde liegende Schriftsystem und seine Orthographie zurückgreifen können.

Aufgrund der Zusammensetzung des Zeichensatzes aus Buchstaben, Zahlen, Satz- und einigen Sonderzeichen sowie Leerstellen ist eine Segmentierung geschriebenen Textes möglich, die auf der Eigenschaft der deutschen Orthographie beruht, verschiedene syntaktische und morphosyntaktische Merkmale der deutschen Sprache zu kodieren. Satzgrenzen sind nicht so einfach bestimmbar, wie es die Setzung von Punkten an Satzenden zunächst vermuten ließe. Der Punkt hat auch andere Funktionen: Abkürzungszeichen, Kennzeichnung von Ordnungszahlen und Datumsangaben oder Auslassungszeichen. Mit einigen heuristischen Zusatzannahmen sind allerdings hohe Raten der Satzgrenzenerkennung möglich (Evert und Fitschen, 2001, 372).

Eine erste Strukturierung von Zeichenketten kann dadurch vorgenommen werden, dass das Konzept *Wort* operationalisiert wird. Definition 4 gibt eine solche Operationalisierung an.

Definition 4 *Wort: Die jeweils längste Kette von Buchstaben, die nicht durch Leerzeichen, Satz- oder Sonderzeichen unterbrochen wird.*

6.2. Verwendung von Vollformenlexika

Spell-Checker basieren im Kern auf einer Liste „korrekter“ Wortformen, mit der der zu überprüfende Text abgeglichen wird (vgl. Mitton, 1996; Fliedner, 2001). Bei diesem Verfahren wird durch explizite Auflistung korrekter Wortformen in großem Umfang Wissen über die deutsche Orthographie zu Grunde gelegt. Allerdings ist dieses Wissen in Hinblick auf eine theoretische Modellierung „flach“, da über Einzelfälle hinaus keine weiteren Informationen ableitbar sind. Vorteile des Verfahrens sind zum einen die leichte Verfügbarkeit der notwendigen Daten, zum anderen ist ein sehr einfacher Algorithmus verwendbar. Nachteil ist die geringe Erklärungsmächtigkeit des Verfahrens. Außer der Aussage, dass das betreffende Wort nicht gefunden wurde, und deshalb Rechtschreibfehler enthalten *könnte*, sind keine weiteren Analysen möglich. In interaktiven Textverarbeitungssystemen kann diese Aussage verwendet werden, fragliche Wörter während der Eingabe zu markieren und somit dem Nutzer zu signalisieren, an welchen Stellen eventueller Korrekturbedarf besteht. In aktuellen Textverarbeitungssystemen wie Microsoft Word oder Open Office ist diese Funktion standardmäßig aktiviert, was darauf schließen lässt, dass sie von den meisten Nutzern akzeptiert und als stabil und sicher genug erachtet wird. Die darüber hinausgehende Funktion, eine Menge von Korrekturvorschlägen anzuzeigen und zur Auswahl anzubieten, wird hingegen erst nach expliziter Aufforderung des Nutzers aktiv.[1]

In ihrer einfachsten Form vergleichen Algorithmen zur Markierung möglicherweise fehlerhaft geschriebener Wortformen die Liste der Wörter im Text mit einer gespeicherten geschlossenen Liste von Wortformen in einer Lexikondatei. Darüber hinaus gehende Verfahren verwenden zusätzlich morphologische Analysealgorithmen, um einerseits nicht komplette Flexionsparadigmen speichern zu müssen und andererseits ad hoc gebildete Wortformen wie z.B. seltene Komposita im Deutschen erkennen zu können.

Zur sicheren Erkennung potenzieller Fehler sind Vollformenlexika problematisch. Durch produktive Kompositabildung, Namen etc. ist die Wortmenge nicht abgeschlossen, sondern offen. Damit ist es prinzipiell nicht möglich, eine vollständige Abgleichliste zu unterhalten. Zusätzlich besteht das Problem fehlerhafter Formen, die mit existierenden Wortformen übereinstimmen.

Zur Überprüfung dieser Aussage wurden die im Corpus auftretenden korrekten und fehlerhaften Types (unterschiedliche Formen) mit der CELEX-Datei verglichen. Es wurde gezählt, wie viele der korrekten Wortformen aus dem Bildergeschichtencorpus in der Datenbank vorkommen und wie viele nicht, sowie für wie viele der fehlerhaften Wortformen das analog gilt. Tabelle 6.1 fasst die Ergebnisse zusammen.

Im Ergebnis ist festzuhalten, dass die Fehlerquote des Verfahrens mit einem F-Wert von 0,23[2] sehr hoch ausfällt. Nur 70% der korrekten Schreibungen wurden als korrekt erkannt, 30% aller verwendeten (korrekten) Wortformen kommen in der ausgewählten Wortliste also nicht vor. Zugleich werden nur 20% der tatsächlichen Fehlschreibungen auch als solche erkannt. 80% der Fehlschreibungen entsprechen also (zufällig) Wortformen, die auch als korrekte Schreibungen in der Wortliste vorkommen. Als Beispiele können hier *<den> für <denn> oder *<man> für <Mann> dienen.

Lässt sich die Fehlerquote bei als falsch klassifizierten korrekten Schreibungen (k^-) durch Erweiterung des Vollformenlexikons senken, besteht diese Möglichkeit für als korrekt klassifizierte Fehlschreibungen (f^+) nicht.

[1] Detailliert zur Nutzung vorhandener Korrekturfunktionen s. Berndt (2002).
[2] S. Definition 3 auf S. 56

k^+	1334
k^-	930
f^+	2857
f^-	564
Precision	0,378
Recall	0,165
F-Maß (α=0.5)	0,23

Tabelle 6.1.: Ergebnisse Verfahren Vollformenlexikon

6.3. Verfahren zur Erkennung möglicher deutscher Wortformen

Im Folgenden werden zwei nicht lexikonbasierte Verfahren vorgeschlagen, mit denen potenzielle Rechtschreibfehler aus der bloßen Betrachtung einer vorliegenden Schreibung erkannt werden sollen. Wie oben ausgeführt muss ein solches Verfahren Wissen über die deutsche Orthographie enthalten. Ein Vollformenlexikon enthält explizite, unverdichtete Informationen, die eine abgeschlossene Menge als korrekt erkennbarer Wortformen definiert. Die beiden hier vorgestellten Verfahren enthalten Wissen über die deutsche Rechtschreibung in verdichteter Form und weisen zusätzlich Produktionsregeln auf, die eine potenziell unendlich große Menge von Formen als korrekt akzeptieren können.

Die Entscheidung über das Vorliegen eines potenziellen Rechtschreibfehlers wird für die nachfolgende beschriebenen Verfahren auf die Frage nach der Wohlgeformtheit von Zeichenketten hinsichtlich einer formalen Sprache abgebildet. Die Beschreibung der formalen Sprache kann explizit oder implizit vorliegen.

Kondensierte Informationen über Wohlgeformtheitsbedingungen von Zeichenketten, d.h. kompakte Beschreibungen formaler Sprachen, können grundsätzlich auf zwei Wegen gewonnen und kodiert werden:

1. Durch Formulierung von symbolischen Regeln, die in möglichst kompakter und allgemeiner Weise beschreiben, welche Wohlgeformtheitsbedingungen für korrekte deutsche Schreibungen gelten. Solche Regeln werden manuell durch einen Experten formuliert, der umfassende Erfahrungen im Umgang mit dem Phänomenbereich hat und dem ein theoretisches Beschreibungsvokabular zur Verfügung steht.
2. Durch automatische Extraktion von Mustern wird aus einer großen Menge von Daten eine möglichst minimale Menge von charakteristischen Mustern gebildet, die eine Entscheidung über Wohlgeformtheit zulassen. Die verwendeten Trainingsdaten müssen alle abzudeckenden Phänomene in einer Weise und Zusammensetzung enthalten, die dem späteren Anwendungszwecke entspricht.

Auf beide Weisen werden Verfahren definiert, die analog zu den in Kapitel 6.2 beschriebenen Verfahren die Frage beantworten, ob eine vorliegende Schreibung evtl. fehlerhaft ist. Es ist abhängig von den Verfahren im Einzelnen zu untersuchen, ob sich aus den Ergebnissen auch Aussagen über die Art des Fehlers ableiten lassen. Grundsätzlich ist damit zu rechnen, dass solche Aussagen möglich sind, weil das Wissen über korrekte Schreibungen in Form von Regeln oder Mustern vorliegt und festgestellte Abweichungen prinzipiell auf einzelne Regeln oder bestimmte Muster zurückgeführt werden können, die qualitative Aussagen über die Fehler ermöglichen könnten.

6.3.1. Regelbasierte Verfahren

Ein regelbasiertes Verfahren zur Erkennung potenziell fehlerhafter Schreibungen deutscher Wörter besteht aus einer formalen Grammatik, die den Aufbau als korrekt eingestufter Formen beschreibt und einem Parser, der die Grammatik auf zu untersuchende Schreibungen anwendet. Als Parsingalgorithmen werden im Folgenden Standardalgorithmen verwendet, die nicht auf eine besonders effiziente Abarbeitung von Analyseaufgaben hin optimiert wurden, da es hier um die Klärung von Fragen der grundsätzlichen Leistungsfähigkeit einer Klasse von Verfahren geht.

Eine formale Grammatik enthält im Wesentlichen eine Menge von elementaren Symbolen und Regeln, wie diese Symbole zu wohlgeformten Ausdrücken der durch die Grammatik beschriebenen Sprache kombiniert werden dürfen (vgl. Klabunde, 2001, 60ff.). Zu den sichtbaren, an der Oberfläche erscheinenden Zeichen ist es sinnvoll, nicht sichtbare Tiefenstrukturen anzunehmen, die Regeln einfacher, d.h. knapper formulierbar werden lassen und theoretische Annahmen über Ähnlichkeiten und strukturelle Verwandtschaft zwischen Formen enthalten. Eine formale Grammatik, die ein Modell für mögliche deutsche Wörter enthält, kann drei verschiedene Arten elementarer Symbole annehmen:

1. Aufbau von Wörtern aus Graphemen,
2. Aufbau von Wörtern aus graphischen Repräsentationen prosodischer Einheiten,
3. Aufbau von Wörtern aus Morphemen.

In allen drei Fällen sind jeweils zwei Beschreibungen notwendig. Zunächst ist zu definieren, wie die grundlegenden Einheiten aussehen bzw. aufgebaut sind und dann, wie diese zu gültigen Wörtern zusammengesetzt werden.

Aufbau von Wörtern aus Graphemen

Die Menge der deutschen Grapheme ist – wie in Kapitel 2.3 ausgeführt – definitionsabhängig. Zunächst wird hier von einer einfachen Liste ausgegangen, die so wenige komplexe Grapheme wie möglich enthält:

Σ = {<a>, <b>, <d>, <e>, <f>, <g>, <h>, <i> , <j>, <k>, <l>, <m>, <n>, <o>, <p>, <r>, <s>, <t>, <u>, <v>, <w>, <x>, <y>, <z>, <ä>, <ö>, <ü>, <ß>, <ck>, <ch>, <sch>, <qu>}

Das einfachste Verfahren überprüft lediglich, ob eine Schreibung genau aus einer Kette zulässiger Grapheme besteht. Dabei werden Kombinationsmöglichkeiten, d.h. die Position und Umgebung eines Graphems im Wort nicht berücksichtigt.

Algorithmus 6.1 definiert das Verfahren:

Im Ergebnis zeigt sich, dass das Verfahren nicht geeignet ist, potenzielle Fehlschreibungen zu identifizieren. Von den 3039 unterschiedlichen Fehlschreibungen des Osnabrücker Bildergeschichtenkorpus[3] werden lediglich 15 als potenziell falsch klassifiziert, zusätzlich wurden 8 korrekte Schreibungen als potenziell falsch eingeordnet.

[3] Groß- und Kleinschreibung sowie Getrennt- und Zusammenschreibung wurden dabei normalisiert, so dass nur phonographische Fehler berücksichtigt wurden.

```
S ← 0
grapheme ← [a, b, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, v, w, x, y, z, ä, ö, ü, ß, ck, ch, sch, qu]
while schreibung do
  found ← FALSE
  for all g ∈ grapheme do
    if schreibung.startswith(g) then
      schreibung ← schreibung − g
      found ← TRUE
    end if
  end for
  if ¬found then
    return FALSE {Abbrechen, wenn kein passendes Graphem gefunden}
  end if
end while
return TRUE
```

Algorithmus 6.1: Algorithmus 0

k^+	2481
k^-	8
f^+	3405
f^-	15
Precision	0,6521
Recall	0,0047
F-Maß (α=0.5)	0,0092

Tabelle 6.2.: Ergebnisse Verfahren 0

Dieses Ergebnis lässt sich leicht begründen. Bis auf zwei Ausnahmen sind alle in deutschen Wörtern vorkommenden Buchstaben auch als einfache Grapheme in der Liste enthalten. Daher kann das Verfahren überhaupt nur Schreibungen als potenziell fehlerhaft markieren, die ein <c> oder <q> außerhalb der Verbindungen <ch>, <ck> und <qu> enthalten.

Eine Verbesserung des Recalls, also der Erkennungsleistung tatsächlicher Fehlschreibungen, ließe sich ohne Einbeziehung des Kontextes nur dann erreichen, wenn die Menge einfacher Grapheme verkleinert werden könnte. In Kapitel 2.3 werden weitere komplexe Grapheme wie <tt>, <ah>, <er> etc. vorgeschlagen, die jedoch sämtlich als Ergänzung zu den vorhandenen Einzelgraphemen zu betrachten sind. Es lässt sich kein einziges der einfachen Grapheme

<a,b,d,e,f,g,h,i,j,k,l,m,n,o,p,r,s,t,u,v,w,x,y,z,ä,ö,ü,ß>

wie <c,q> so in feste Kontexte einbetten, dass ein alleiniges Auftreten nicht mehr angenommen werden muss.

Das einfache Graphemlistenverfahren eignet sich also nicht für die Erkennung potenziell fehlerhafter oder potenziell korrekter Schreibungen und bietet keine Optimierungsmöglichkeiten durch veränderte linguistisch fundierte Annahmen über die Grapheme des Deutschen.

Eine nicht primär linguistisch begründete Erweiterung des Verfahrens auf Buchstabenkontexte findet sich in Abschnitt 6.3.2 unter der Überschrift „Musterbasierte Verfahren“.

Aufbau von Wörtern aus graphemischen Repräsentationen prosodischer Einheiten

In einem prosodischen Modell können Wörter als Ketten von Füßen bzw. Silben aufgefasst werden, die intern aus konsonantischem Anfangsrand, vokalischem Nukleus und konsonantischem Endrand bestehen (vgl. Kapitel 4.3.2). Wenn angenommen wird, dass sich die Schreibung eines Wortes als graphische Repräsentation dieser Ebenen und ihrer Wechselwirkungen auffassen lässt, ist es möglich, geschriebene Wortformen mit einem phonologischem Beschreibungsvokabular zu analysieren. Dieses Vorgehen wird bei den nachfolgend definierten Algorithmen zu Grunde gelegt und deshalb in einem kurzen Exkurs genauer begründet.

Exkurs: Phonologische Interpretation graphemischer Ketten Es stellt sich die Frage, warum und unter welchen Umständen es zulässig ist, geschriebene Wortformen, d.h. Ketten von Graphemen, mit phonologischen Kategorien wie Silben, Silbenkonstituenten, Füßen, Betonungsmustern etc. zu analysieren. Zunächst sind phonologische und graphemische Beschreibung auf zwei materiell unterschiedlichen Ebenen angesiedelt. Eine Gleichsetzung von Elementen der einen Ebene mit Elementen der anderen ist daher nicht ohne weiteres möglich.

Der orthographische Wortbegriff ist zunächst als syntaktisch motiviert anzusehen, wenn genau das als ein Wort betrachtet wird, was durch Zwischenräume und/oder Satzzeichen abgegrenzt wird. Für die Verbindung der orthographischen und der phonologischen Ebene ist also nun das orthographische Wort als syntaktisch atomare Einheit mit phonologischen Einheiten eindeutig in Verbindung zu bringen. Atomare Einheiten einer Ebene sind nicht mit Begriffen der betreffenden Ebene zu beschreiben, sondern mit Begriffen „niedrigerer“ Ebenen. Im Falle der Syntax bietet sich hier die morphologische Ebene an. Jedes syntaktische Wort ist genau als Folge von Morphemen zu beschreiben, d.h. ein syntaktisches Wort kann aus einem oder mehreren Morphemen bestehen, Beginn und Ende eines solchen Wortes fallen aber immer mit Beginn bzw. Ende eines Morphems zusammen. Gemäß Wieses prosodischer Hierarchie (Wiese, 1996) besteht die Schnittstelle zwischen Phonologie und Morphologie in der Ebene des prosodischen Wortes. Prosodische Wörter können im Falle satzphonetischer Reduktionen mehr als ein syntaktisches Wort umfassen. Da hier die Betrachtungsweise aber von bereits explizit isolierten syntaktischen Wortformen ausgeht, kann dieser Aspekt zunächst vernachlässigt werden. Jedes syntaktische Wort ist auch phonologisch isoliert betrachtbar und besteht dann genau aus einer Kette phonologischer Wörter. Phonologische Wörter setzen sich genau aus Füßen und Füße genau aus Silben zusammen. In diesem Sinne können syntaktische Wörter genau als Folge von Silben betrachtet werden, d.h. jedes syntaktische (und damit orthographische) Wort beginnt mit dem Beginn (also einem vollständigen Anfangsrand) einer Silbe und endet mit dem Ende (also einem vollständigen Endrand) einer Silbe. Damit sind orthographische Wörter als ganzes phonologisch verankert, indem sie genau als Folge von vollständigen prosodischen Einheiten beschrieben werden können.

Jede orthographische Wortform besitzt eine oder mehrere phonetische Realisationen, d.h. die Transformation graphemischen in phonetisches Material ist für jede korrekte geschriebene Wortform möglich – so wie es z.B. in „Aussprachewörterbüchern“ angegeben ist. Phonetische Realisationen werden hier als Manifestationen eines phonologisch codierbaren „Planes“ verstanden, der lautliche Invarianten beinhaltet. Es ist nun zu untersuchen, ob und wie sich die graphemische und phonologische Binnenstruktur von Wortformen aufeinander abbilden lassen.

Die Unterscheidung von Silbenkernen und Silbenrändern geht einher mit der Unterscheidung von Vokalen und Konsonanten. Der Zusammenhang zwischen Vokalen und Vokalgraphemen bzw. Konsonanten und Konsonantengraphemen muss nicht einfach stipuliert, sondern kann auf qualitative sowie strukturelle Weise begründet werden. Im Folgenden werden beide Argumentationsrichtungen angedeutet, ohne sie miteinander in ihrer Leistungsfähigkeit vergleichen zu wollen. Für die hier verfolgten Zwecke reicht es aus, begründen zu können, dass und wie sich graphemische Ketten phonologisch interpretieren lassen.

Verschiedene phonologische Theorien gehen davon aus, dass sämtliches vokalische Material einer Wortform dem Kern, sämtliches konsonantische Material den Rändern zuzuordnen ist. Somit lässt sich unter Vernachlässigung der Frage nach Silbengrenzen ein einfacher Algorithmus zur Identifikation von Silbenkernen implementieren.

Auf graphemischer Ebene andererseits können Vokal- und Konsonantengrapheme unterschieden werden. Diese Unterscheidung muss nicht zirkulär über die Heranziehung von Graphem-Phonem-Korrespondenzen getroffen werden, sondern kann über eine rein graphemische Distributionsanalyse geschehen. In dem Projekt „Computerbasierte Modellierung orthographischer Prozesse“ (Maas et al., 1999) wurde eine solche Analyse mit Hilfe selbstorganisierender Karten (Kohonen, 1995) für einfache Grapheme vorgenommen. Die Ergebnisse rechtfertigen eine klare Trennung von zwei Graphemklassen aufgrund der graphemischen Kontexte ihres Auftretens. Die beiden Klassen entsprechen exakt den üblicherweise als Vokalgrapheme und Konsonantengrapheme bezeichneten Graphemmengen. Betrachtet man nun Vokalgrapheme als Repräsentationen von Silbenkernen, wird deutlich, dass sie sich in ihrem graphemischen Kontext von Konsonantengraphemen, die dann Repräsentationen von Silbenrändern bilden, deutlich unterscheiden. Der oben skizzierte einfache Algorithmus zur Identifikation von Silbenkernen ist damit erweiterbar auf die Identifikation graphemischer Repräsentationen von Silbenkernen.

Silbengrenzen sind an Beginn und Ende einer Wortform unmittelbar gegeben, damit dort ebenfalls die Abgrenzung von erstem Anfangsrand und letztem Endrand. Orthographische Repräsentationen einsilbiger Wortformen sind damit vollständig silbenstrukturellen Einheiten zuordenbar, d.h. es werden Silbenkonstituenten (=Ketten von Phonemen) mit Ketten von Graphemen in Verbindung gebracht. Eine unmittelbare Zuordnung von Graphemen und Phonemen ist damit allerdings nicht geleistet. Es ist jedoch nicht immer sinnvoll und notwendig, eine direkte Zuordnung auf atomarer Ebene, d.h. von Graphemen und Phonemen vorzunehmen.

Zur vollständigen Lösung des Zuordnungproblems verbleiben nach der Identifikation von Silbenkernrepräsentationen und Anfangs- und Endrändern an den Worträndern noch Verbindungen von je einem End- und einem Anfangsrand und Ketten von Graphemen. Das Problem ist dann gelöst, wenn eine Grenze innerhalb der Graphemkette angegeben werden kann, so dass alles graphemische Material vor dieser Grenze als Repräsentation des Endrandes und alles Material nach dieser Grenze als Repräsentation des Anfangsrandes anzusehen ist. Eine solche Grenze ist schon phonologisch nicht immer eindeutig bestimmbar, so dass in Zweifelsfällen eine Zuordnung einer Graphemkette zu einer Kombination aus End- und folgendem Anfrangsrand ausreichend sein kann.

Grammatik Die Menge möglicher graphemischer Repräsentationen von Silben wird in erster Näherung über eine kontextfreie Grammatik definiert, die Wortformen zunächst als beliebige Folge möglicher Silben zulässt. Eine mögliche Silbe besteht dabei aus einer Verkettung eines möglichen Anfangsrandes, eines Nukleus und eines möglichen Endrandes. In dieser einfachen Grammatik bleiben eine Reihe von Ebenen und Gegebenheiten unberücksichtigt: Fußstrukturen, Silbentypen, kontextuelle Auftretensbeschränkungen etc. Die Menge der möglichen Silbenkonstituenten ist maximal, d.h. jede in deutschen Wörtern vorkommende Kombination ist aufgenommen. Sie ist explizit, d.h. Binnenstrukturen graphemischer Repräsentationen von Silbenkonstituenten werden nicht über Regeln sondern durch Aufzählung abgebildet.

Die Liste der graphemischen Konstituentenrepräsentationen lässt sich grundsätzlich automatisiert gewinnen, in dem zunächst alle einsilbigen Formen (das sind solche, die genau ein Vokalgraphem enthalten) identifiziert werden und für diese Formen Anfangs- und Endränder gesammelt werden.

Für den nachfolgenden dargestellten Algorithmus wurden die möglichen Silbenkonstituenten manuell ermittelt und durch sukzessive Verfeinerung nach Testläufen anhand des Osnabrücker Korpus

k^+	2434
k^-	45
f^+	3105
f^-	315
Precision	0,8750
Recall	0,0921
F-Maß (α=0.5)	0,1666

Tabelle 6.3.: Ergebnisse Verfahren 1

optimiert. Die Entscheidung, welche Graphien noch als dem Kernbereich der deutschen Orthographie zugehörig klassifiziert werden, wurde heuristisch getroffen. Denkbar ist hier eine Absicherung über Häufigkeitszählungen, die aber stark von der Zusammensetzung des verwendeten Korpus abhängen. So tritt die Anfangsrandrepräsentation <wr> im Osnabrücker Korpus 8 mal, die Anfangsrandrepräsentation <ps> 69 mal am Wortanfang auf. Eine sinnvolle Verwendung von Häufigkeiten ließe sich nur problemabhängig erreichen, im vorliegenden Fall also durch Auswertung eines Korpus von Kinderschreibungen mit einer dem untersuchten Korpus ähnlichen Leistungsstruktur. Ein solches Korpus lag bei der Konstruktion der Grammatik nicht vor, eine Verwendung des Korpus selbst verbietet sich aus methodischen Gründen. Weder die Anfangsrandrepräsentation <wr>, noch <ps> treten hier bei korrekten Schreibungen am Wortanfang auf, allerdings ist <wr> in 4 Fällen als Anfangsrandrepräsentation fehlerhafter Schreibungen zu beobachten. Eine aus Häufigkeitszählungen hergeleitete Nichtaufnahme von <wr> würde also eine verfälschende Verbesserung des Recall-Wertes zur Folge haben. Aus diesen Gründen wurden bei der Erstellung der Konstituentenlisten heuristische Annahmen und Vorwissen, z.B. über die Herkunft bzw. Beeinflussung von Schreibungen aus anderen Sprachen einbezogen.

Die nichtterminalen Symbole der in Algorithmus 6.2 aufgeführten Grammatik begründen wie erwähnt eine relativ flache Analysestruktur für graphemische Wortformen. Es wird vereinfachend ein einheitlicher Silbentyp angenommen, Wörter bilden sich aus einer beliebig langen Verkettung solcher Silben. Tabelle 6.3 stellt die Ergebnisse für das Bildergeschichtenkorpus zusammen.

Auch hier zeigt sich, dass das Verfahren nicht als generelles Verfahren zur Erkennung orthographischer Fehler geeignet ist. Zwar werden nur 1,8% der korrekten Formen fälschlicherweise als Fehler analysiert, allerdings entgehen dem Verfahren damit auch 89,9% der tatsächlichen Fehler. Immerhin weist das Verfahren, wenn es einen Fehler erkannt hat, nur eine Fehlerkennungsquote von 13% auf (Verhältnis k^- zu f^-). Es könnte somit als zusätzliches Erkennungsverfahren neben anderen gute Dienste leisten.

6.3.2. Musterbasierte Verfahren

Bei musterbasierten Verfahren werden die zu untersuchenden Wortformen mit einer Menge automatisiert gewonnener Muster verglichen. Lässt sich eine Wortform vollständig als aus bekannten Mustern bestehend analysieren, so gilt sie als korrekt.

Ein ebenso einfaches wie grundlegendes Verfahren, um automatisiert Muster zu gewinnen, sind n-Gramme (vgl. Frötschl und Lindstrot, 2001, 117). Die n-Gramm-Berechnung für Graphemketten

```
w --> silbe.
w --> silbe, w.
silbe --> anf,vok,end.
silbe --> vok,end.
silbe --> anf,vok,endkurz.
silbe --> vok,endkurz.
silbe --> anf,vok.
silbe --> anf,langvok.
silbe --> anf,langvok,endlang.
silbe --> langvok,endlang

anf --> <b> | <d> | <f> | <g> | <h> | <j> | <k> | <l> | <m> | <n> | <p> | <r> |
        <s> | <t> | <v> | <w> | <x> | <y> | <z> | <st> | <sp> | <qu> | <ch> |
        <bl> | <br> | <dr> | <fl> | <fr> | <gl> | <gr> | <kl> | <kn> | <kr> |
        <ph> | <pl> | <pr> | <pf> | <tr> | <wr> | <zw> | <sch> | <spr> |
        <str> | <pfr> | <schw> | <schm> | <schn> | <schl>.

vok --> <a> | <e> | <i> | <o> | <u> | <ä> | <ö> | <ü> | <y>.

langvok --> <au> | <ai> | <aa> | <ah> | <ei> | <eu> | <ee> | <eh> | <ie> |
        <uh> | <oo> | <oh> | <öh> | <üh> | <äu> | <äh> | <eih> | <ieh>.

end --> <b> | <bs> | <bst> | <bsts> | <ch> | <chs> | <chst> | <d> | <ds> |
        <dst> | <f> | <fs> | <fst> | <g> | <gs> | <gst> | <k> | <ks> | <kst> |
        <l> | <ls> | <lst> | <lg> | <lgs> | <lgst> | <lk> | <lks> | <lkst> |
        <ln> | <lns> | <lnst> | <lm> | <lms> | <lmst> | <ld> | <lds> | <ldst> |
        <lt> | <lts> | <ltst> | <lb> | <lbs> | <lbst> | <lp> | <lps> | <lpst> |
        <lf> | <lfs> | <lfst> | <lz> | <lzt> | <m> | <ms> | <mst> |<md> |
        <mds> | <mdst> | <mt> | <mts> | <mtst> | <mb> | <mbs> | <mbst> | <mp> |
        <mps> | <mpst> | <mf> | <mfs> | <mfst> | <mpf> | <mpfs> | <mpfst> |
        <mz> | <mzt> | <n> | <ns> | <nst> | <nd> | <nds> | <ndst> | <nt> |
        <nts> | <ntst> | <nf> | <nfs> | <nfst> | <nz> | <nzt> | <p> | <ps> |
        <pst> | <pt> | <pts> | <pf> | <pfs> | <pfst> | <r> | <rs> | <rst> |
        <rg> | <rgs> | <rgst> | <rk> | <rks> | <rkst> | <rn> | <rns> | <rnst> |
        <rm> | <rms> | <rmst> | <rd> | <rds> | <rdst> | <rt> | <rts> | <rtst> |
        <rb> | <rbs> | <rbst> | <rbsts> | <rp> | <rps> | <rpst> | <rf> |
        <rfs> | <rfst> | <rl> | <rls> | <rlst> | <rz> | <rzt> | <s> | <st> |
        <sts> | <sk> | <t> | <ts> | <v> | <vs> | <w> | <x> | <z> | <zt> | <zts>.

endlang --> <ß> | <ßt>.

endkurz --> <bb> | <bbs> | <bbst> | <ck> | <cks> | <ckst> | <dd> | <ff> |
        <ffs> | <ffst> | <gg> | <ggs> | <ggst> | <ll> | <lls> | <llst> | <mm> |
        <mms> | <mmst> | <nn> | <nns> | <nnst> | <ng> | <ngs> | <ngst> | <nk> |
        <nks> | <nkst> | <pp> | <pps> | <ppst> | <rr> | <rrs> | <rrst> | <ss> |
        <sst> | <tt> | <tts> | <tz> | <tzt>.
```

Algorithmus 6.2: Kontextfreie Grammatik für Verfahren 1

n	Zerlegungen
1	k, o, m, m, e, n
2	_k, ko, om, mm, me, n_
3	__k, _ko, kom, omm, mme, men, en_, n__
4	___k, __ko, _kom, komm, omme, mmen, men_, en__, n___
5	____k, ___ko, __kom, _komm, komme, ommen, mmen_, men__, en___, n____

Tabelle 6.4.: n-Gramm-Zerlegungen für $n = 1 \ldots 5$

Kursform	Einträge	Erklärung
de100	100	100 häufigste Wortformen (Projekt dt. Wortschatz)
de1000	1000	1000 häufigste Wortformen (Projekt dt. Wortschatz)
de10000	10000	10000 häufigste Wortformen (Projekt dt. Wortschatz)
gws	1150	Nieders. Grundwortschatz für Grundschulen
kaeding	5795	Aufbereitete Form des Keading-Orthmann-Corpus (s. Maas et al., 1999)
ids	30000	30000 häufigste Wortformen aus den IDS-Corpora
osc	89595	Osnabrücker Corpus (s. Maas et al., 1999)
celex	143442	Aufbereitete Form der Celex-Corpus (s. Maas et al., 1999)

Tabelle 6.5.: Verwendete Trainingscorpora für Verfahren ohne Zusatzinformationen

zerlegt eine Menge von Wortformen in jeweils gleich große, d.h. gleich lange Teilketten. Tabelle 6.4 zeigt die Zerlegungen der Wortform <kommen> für $n = 1 \ldots 5$. _ steht dabei als Metazeichen für den Wortrand.

Mit zunehmender n-Gramm-Größe nimmt also die Zahl der pro Wortform generierten Muster zu. Gleichzeitigt steigt auch die Gesamtzahl der Muster, da größere Muster wie z.B. <ommen> mit geringerer Wahrscheinlichkeit auch in anderen Wortformen gefunden werden als kleinere Muster wie z.B. <om>.

Größere Muster bieten grundsätzlich höhere Sicherheiten bei der Erkennung. Als Beispiel diene die zu untersuchende Schreibung *<qu>. Um diese Schreibung als falsch zu erkennen, sind – eine geeignete Trainingswortliste vorausgestzt – mindestens Muster der n-Gramm-Größe 3 notwendig. Bei Verwendung der Mustergröße 2 muss das Trainingscorpus lediglich die Muster <_q>, <qu> und <u_> enthalten, was z.B. bereits durch die Wörter <Quark> und <du> gewährleistet ist. Bei Mustern der Länge drei müsste das Trainingscorpus auch das Muster <qu_> enthalten. Für eine solche Wortform lässt im Deutschen kein Beleg finden, so dass die Fehlschreibung *<qu> bereits mit Mustern der Länge 3 als nicht mögliche Wortform klassifiziert werden kann. Für die folgenden Untersuchungen wurden acht Wortlisten des Deutschen verwendet, die in Tabelle 6.5 aufgeführt sind. Bei diesen Korpora handelt es sich um aufbereitete Wortlisten, die hinsichtlich Kodierungsinformationen, Groß- und Kleinschreibung und bekannter Fehlern (vor allem bei der CELEX-Datenbank) bereinigt und vereinheitlicht wurden.

Die Gesamtergebnisse für alle Korpora und n-Gramm-Größen mit $n = 2 \ldots 6$ sind in Tabelle 6.6 aufgeführt. Die F-Maß-Werte rangieren dabei zwischen 0.02 für einige 2-Gramm-Musterlisten und ca. 0.75 bei einigen 5- und 6-Gramm-Musterlisten. Tendenziell sind große Wortlisten bei größeren Musterlängen sehr viel leistungsfähiger als kleine. Kleine, nicht speziell ausgewählte Trainingskorpora sind deutlich weniger erfolgreich. Würde man nur die korrekten Formen des Osnabrücker

Korpus	n	k^+	k^-	f^+	f^-	P	R	F
de100	2	13.53%	86.47%	13.64%	86.36%	0.50	0.86	0.63
	3	4.44%	95.56%	2.18%	97.82%	0.51	0.98	0.67
	4	3.84%	96.16%	1.24%	98.76%	0.51	0.99	0.67
	5	3.63%	96.37%	1.17%	98.83%	0.51	0.99	0.67
	6	3.42%	96.58%	1.14%	98.86%	0.51	0.99	0.67
de1000	2	84.53%	15.47%	77.08%	22.92%	0.60	0.23	0.33
	3	37.00%	63.00%	24.32%	75.68%	0.55	0.76	0.63
	4	21.89%	78.11%	5.99%	94.01%	0.55	0.94	0.69
	5	18.25%	81.75%	3.42%	96.58%	0.54	0.97	0.69
	6	17.12%	82.88%	2.93%	97.07%	0.54	0.97	0.69
de10000	2	98.06%	1.94%	95.77%	4.23%	0.69	0.04	0.08
	3	83.16%	16.84%	64.84%	35.16%	0.68	0.35	0.46
	4	60.22%	39.78%	25.42%	74.58%	0.65	0.75	0.70
	5	49.19%	50.81%	11.20%	88.80%	0.64	0.89	0.74
	6	45.07%	54.93%	8.01%	91.99%	0.63	0.92	0.75
gws	2	84.65%	15.35%	76.07%	23.93%	0.61	0.24	0.34
	3	49.07%	50.93%	29.07%	70.93%	0.58	0.71	0.64
	4	31.14%	68.86%	7.81%	92.19%	0.57	0.92	0.71
	5	23.99%	76.01%	3.91%	96.09%	0.56	0.96	0.71
	6	22.62%	77.38%	3.39%	96.61%	0.56	0.97	0.71
kaeding	2	93.94%	6.06%	88.93%	11.07%	0.65	0.11	0.19
	3	69.14%	30.86%	49.22%	50.78%	0.62	0.51	0.56
	4	48.87%	51.13%	18.94%	81.06%	0.61	0.81	0.70
	5	41.68%	58.32%	9.38%	90.62%	0.61	0.91	0.73
	6	38.93%	61.07%	7.49%	92.51%	0.60	0.93	0.73
osc	2	99.76%	0.24%	99.15%	0.85%	0.78	0.01	0.02
	3	94.22%	5.78%	83.33%	16.67%	0.74	0.17	0.27
	4	78.51%	21.49%	48.27%	51.73%	0.71	0.52	0.60
	5	62.11%	37.89%	20.83%	79.17%	0.68	0.79	0.73
	6	52.95%	47.05%	12.79%	87.21%	0.65	0.87	0.74
celex	2	99.65%	0.35%	98.63%	1.37%	0.80	0.01	0.03
	3	90.23%	9.77%	75.42%	24.58%	0.72	0.25	0.37
	4	73.34%	26.66%	40.55%	59.45%	0.69	0.59	0.64
	5	61.91%	38.09%	18.72%	81.28%	0.68	0.81	0.74
	6	57.60%	42.40%	13.47%	86.53%	0.67	0.87	0.76
ids	2	99.72%	0.28%	99.19%	0.81%	0.74	0.01	0.02
	3	94.06%	5.94%	79.98%	20.02%	0.77	0.20	0.32
	4	77.58%	22.42%	41.80%	58.20%	0.72	0.58	0.64
	5	66.03%	33.97%	19.27%	80.73%	0.70	0.81	0.75
	6	61.19%	38.81%	12.60%	87.40%	0.69	0.87	0.77

Tabelle 6.6.: Erkennungsraten, Precision, Recall und Fehlermaß für verschiedene Trainingskorpora und n-Gramm-Größen

Bildergeschichtenkorpus als Trainingsmenge zugrundelegen, ergäbe sich zwar auch ein relativ kleines Trainingskorpus. Es würde k^+-Werte von 100% erreichen, da alle in den korrekten Formen des Testkorpus enthaltenen Muster auch im Trainingskorpus enthalten waren. Allerdings ist diese Form des Trainings methodisch unzulässig und erlaubt keine Beurteilung der Leistungsfähigkeit des allgemeinen Verfahrens (vgl. Rojas, 1993, 222ff.).

Das beste Fehlermaß von 0.77 ist vorsichtig positiv zu beurteilen. Ein triviales Verfahren, das ohne jegliche Analyse entweder alle beobachteten Formen als korrekt oder falsch klassifiziert, kann je nach Verteilung falscher und korrekter Formen in untersuchten realitätsnahen Korpora durchaus Fehlermaß-Werte von ca. 0.5 erreichen. Das bestee hier untersuchte Trainingskorpus erkennt für 6-Gramm-Musterlängen 12,6% der Fehler nicht und klassifiziert gleichzeitig immerhin 38,81% der korrekten Schreibungen als falsch.

6.4. Möglichkeiten und Grenzen von Analyseverfahren ohne Informationen über die Zielschreibung

Grundsätzliches Ziel einer Analyse von Schreibungen ohne Kenntnis der „intendierten“ Schreibung ist, sichere Aussagen darüber zu treffen, ob eine orthographisch korrekte oder normabweichende Schreibung vorliegt. Allein aufgrund der Problematik, dass die isolierte Betrachtung von Wortformen für diese Entscheidung nicht ausreicht, wäre ein vollständig verlässliches Verfahren nur unter Einbeziehung von Kontextinformationen und einer semantisch und pragmatisch fundierten Analyse des Gesamttextzusammenhanges möglich.

In diesem Kapitel wurden drei Verfahren entwickelt und an einem Korpus tatsächlich beobachteter Schreibungen von Schreibanfängern erprobt. Als Gesamtergebnis ist festzuhalten, dass keines der Verfahren in verlässliche Bereiche vorstoßen kann. Am ehesten erfolgversprechend sind musterbasierte Verfahren, die anhand möglichst großer und damit für den Gesamtsprachgebrauch repräsentativer Wortlisten trainiert wurden. Hier sind Fehlermaße (s. Kapitel 3, S. 56) von maximal 0.77 erreichbar. Somit scheidet ein solches Verfahren als alleiniges Entscheidungskriterium bei der Analyse unbekannter Schreibungen aus.

Ein ähnliches Bild ergeben manuell konstruierte Beurteilungsverfahren wie das einer silbenorientierten Grammatik, die mögliche Wortformen mit Produktionsregeln beschreibt, die Grapheme als Repräsentation segmentaler und suprasegmentaler phonologischer Gegebenheiten verstehen. Die vorgestellte Grammatik konnte zwar nur relativ wenige fehlerhafte Formen tatsächlich als fehlerhaft erkennen (ca. 10%), dafür bietet es aber eine besonders niedrige Quote an fälschlicherweise als fehlerhaft markierten korrekten Formen.

Das untersuchte Korpus unterscheidet sich von Schreibungen, wie sie im Büroalltag auftreten dadurch, dass es eine besonders hohe Grundfehlerquote von ca. 25% aufweist. Die Rechtschreibkorrektur-Algorithmen von Spell-Checkern, wie sie Microsoft Office und andere für den Büroalltag konzipierte Pakete aufweisen, sind auf andere Gesamtumgebungen hin optimiert und wenig geeignet, typische Fehler von Schreibanfängern zu erkennen. Insgesamt ist es wahrscheinlich, dass die in diesem Kapitel vorgestellten Verfahren in Kombination eine gute Erkennungsrate bei solchen Schreibungen aufweisen. Als Grundlage für verlässliche Aussagen für eine qualitative Fehleranalyse taugen sie aber nur sehr bedingt. Deshalb wird im Folgenden davon ausgegangen, dass die „intendierten Schreibungen“ bekannt sind und ein Verfahren entwickelt, dass aufgrund des Abgleichs von Ausgangs- und Zeilschreibung zu differenzierten Aussagen gelangen kann.

7. Verfahren zum Abgleich von Ausgangs- und Zielschreibung

7.1. Grundlegendes Verfahren

Verfahren, die Informationen über die Zielschreibung zur Verfügung haben, sind grundsätzlich anders aufgebaut als die in Kapitel 6 vorgestellten Verfahren. Eine bekannte Zielschreibung ermöglicht es, die vorliegende Schreibung nicht isoliert betrachten und analysieren zu müssen. Kern der Untersuchung ist vielmehr die Betrachtung einer Differenz zwischen beobachteter und intendierter Schreibung.

Die Differenzbetrachtung geschieht immer auf Buchstaben- bzw. Graphemebene, d.h. zwei Zeichenketten gleicher Materialität werden miteinander verglichen. Der einfachste Fall eines solchen Verfahrens, der lediglich feststellt, ob die Formen übereinstimmen oder nicht, ist für weitergehende Analysen ungeeignet. Es ist also notwendig, ein abgestuftes Ähnlichkeitsmaß für Schreibungen zu verwenden, aus dem weitergehende Aussagen abgeleitet werden können. Dieses Ähnlichkeitsmaß darf nicht einfach numerisch sein, da daraus ebenfalls keine qualitativen Analysen abgeleitet werden könnten. Erforderlich ist vielmehr eine Liste von Abweichungen, die für die weitere Analyse ausgewertet werden kann. Eventuell kann es auch sinnvoll sein, mehrere solcher Listen als Differenzhypothesen zu betrachten.

Für solche weitergehenden Aussagen kann es notwendig sein, weitere Informationen, die nicht allein auf der graphematischen Oberfläche beobachtbar sind, einzubeziehen. In Hinblick auf das in Kapitel 4 entworfene Auswertungsschema sind dies z.B. Informationen über die Rolle als Repräsentant von Silbenkonstituenten, die ein Graphem oder Graphemcluster annehmen kann. Für die Zielschreibungen wird also eine mehr oder weniger detaillierte linguistische Analyse vorausgesetzt. Diese kann auf zweierlei Arten gewonnen werden: Zum Einen durch Zugriff auf ein Lexikon, das die Informationen für alle vorkommenden Zielschreibungen enthält und zum Anderen durch einen Analysealgorithmus, der die graphemische Form analysiert und die notwendigen Informationen erzeugt. Ein vollständiges Lexikon ist bei geschlossenen Anwendungen wie Diktaten oder Lückentexten prinzipiell vorstellbar, jedoch zeigen selbst hier die Erfahrungen, dass mit abweichenden intendierten Schreibungen gerechnet werden muss. Der vielversprechendste Weg ist daher eine eigenständige Anlaysekomponente, die für nicht korrekt analysierbare Fälle auf ein Ausnahmelexikon zurückgreifen kann.

Die Betrachtung der annotierten Zielschreibung und der Liste vorgefundener Differenzen kann dann schließlich zu einem Analyseergebnis führen, das das Analyseschema aus Abschnitt 4.5 füllt. Somit ergibt sich für das hier entwickelte Analyseverfahren ein dreischrittiges Vorgehen:

1. Annotation der Zielschreibung
2. Stringvergleich von Zielschreibung und beobachteter Schreibung
3. Generierung der Analysetabelle

Phänomen	**max**	**repräsentiert**	**phonetisch plausibel**	**orthographisch korrekt**
Silbenkerne				
Anfangsränder				
Endränder				
Phonologische Markierungen			—	
Konstantschreibung				
Sonstiges				
Vertauschungen	—		—	—
Einfügungen	—			—

Tabelle 7.1.: Kurzübersicht der Auswertungskategorien

7.2. Annotation der Zielschreibung

Die möglichst automatisierte Annotation der Zielschreibung soll alle Informationen generieren, die für Analysetabellen relevant sein können. Die Zeilen der Tabelle gliedern sich in vier große Bereiche: silbenpositionsbezogene Aussagen, phonologische Markierungen, Konstantschreibung und Sonstiges (s. Tabelle 7.1).

Die Spalte **max** weist für eine vollständige Analyse die Vorkommen des jeweiligen Phänomens für die Zielschreibung aus, ist also eine positionsunabhängige Aggregation der hier zu definierenden Annotation. Die Annotationskategorien sind damit definiert. Die Fälle unter Sonstiges müssen nicht berücksichtigt werden, da sie nur für den Vergleich eine Rolle spielen. Eine korrekte Schreibung kann keine Einfügung oder Vertauschung enthalten.

Um aus dem späteren Stringvergleich aussagekräftige Informationen gewinnen zu können, ist eine postionsabhängige Zuordnung der Analyseannotationen notwendig. Das Annotationsverfahren muss somit jedem Graphem der Zielschreibung eine Menge von Attributen aus der Analysetabelle zuordnen. Graphemübergreifende Annotationen kommen ebenfalls vor, können aber problemlos durch eine Folge von Einzelgraphemannotationen abgebildet werden, da gleichartige graphemübergreifende Annotationen wie z.B. „(Repräsentation von) komplexer Endrand“ sich weder überlappen noch direkt aneinandergrenzen können.

Das Annotationsverfahren für Zielschreibungen arbeitet fünfschrittig:

1. Zugriff auf Vollanalyselexikon
2. Zerlegung komplexer Formen durch Lexikonzugriff
3. Wortparser
4. Silbentyp-Zuweisung
5. Feature-Zuweisung

7.2.1. Zugriff auf Vollanalyselexikon

Formen, die von den nachfolgend beschriebenen Schritten nicht korrekt analysiert werden, können manuell korrigiert bzw. erfasst und in einem Vollanalyse-Lexikon abgelegt werden. Dieses Lexikon ist in der vorliegenden Implementierung direkt als Python-Datei umgesetzt, die die erforderlichen Datenstrukturen als Python-Code enthält. Plausibilitätsprüfungen finden nicht statt und die dort vorgefundenen Analyse haben immer Vorrang, d.h. beenden den Analyseprozess sofort erfolgreich mit einem Ergebnis.

Das abzulegende Format enthält alle in Abschnitt 7.2.6 aufgeführten Informationen in Python-Array- bzw. Dictionary-Syntax.

7.2.2. Zerlegung komplexer Formen durch Lexikonzugriff

Morphologisch komplexe Wortformen können bei der folgenden Token-Zerlegung Probleme bereiten. Es ist deshalb vorgesehen, dem Zerlegungsalgorithmus Hinweise zu geben, an welchen Stellen auf jeden Fall Grenzen von Silbenrepräsentationen anzunehmen sind. Das Lexikon für die Vorab-Zerlegung ist äußerst simpel aufgebaut, es besteht aus einer Textdatei, in der zu zerlegende Wortformen aufgeführt sind, die Zerlegungsgrenzen sind dabei durch ein `#` gekennzeichnet, wie im folgenden Beispiel:

```
bade#hose
bett#decke
bitter#kalt
```

Es ist nicht notwendig, alle derartigen Grenzen in der Datei zu markieren, die Grenzmarkierungen werden lediglich als gesetzt angenommen und evrhindern die Analyse der restlichen Bestandteile nicht.

Für die vorliegende Untersuchung wurde diese Datei manuell erstellt. D.h. alle im Testkorpus vorkommenden Formen, die zerlegt werden müssen, um korrekt analysiert werden zu können, wurden entsprechend markiert und in die Datei aufgenommen. Anstelle des Lexikons könnte auch eine automatische Analysekomponente verwendet werden, die entsprechende Ergebnisse liefert. Der Lexikonzugriff ist in der Klasse `morphlex` gekapselt, die durch eine automatisierte Variante ersetzt werden könnte. Im Rahmen dieser Arbeit wurde allerdings keine entsprechenden Versuche unternommen.

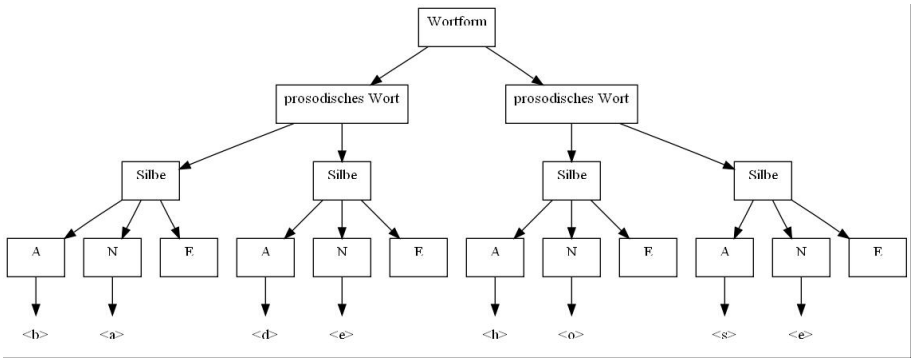

Abbildung 7.1.: Resultat eines Parserdurchlaufs

Zustand	Zeichen	Folgezustand	Aktion
A	V	N	Neue Silbe gefunden, Stack aufteilen, h- und Schärfungsmarker setzen
N	V	N	Mehrere Vokalbuchstaben in Folge. Prüfen, ob zulässiger Cluster, dann in gleicher Silbe bleiben, sonst neue Silbe beginnen.
E	V	N	Zeichen auf Stack.
A	K	A	Zeichen auf Stack.
N	K	A	Zeichen auf Stack. Auf h-Marker prüfen.
E	K	E	Zeichen auf Stack. Auf h-Marker prüfen.
A	#	E	Silbe abschließen.
N	#	N	Silbe abschließen.
E	#	E	Silbe abschließen.

Tabelle 7.2.: Zustandsübergänge des Wordparser-Automaten

7.2.3. Wortparser

Die gegebenenfalls vorzerlegte Zielschreibung wird von einer Parser-Komponente als Repräsentation einer prosodischen Grundstruktur aus phonologischen Wörtern, Silben und Silbenkonstitutenten interpretiert. Abbildung 7.1 zeigt ein beispielhaftes Ergebnis des Gesamtdurchlaufs. Die Zuweisung von Silbentypen geschieht erst in einem nachgelagerten Schritt.

Der Wortparser ist im Kern als endlicher Automat mit drei Zuständen für Anfangsrandrepräsentation, Nukleusrepräsentation und Endrandrepräsentation aufgebaut, dem die zu analysierende Schreibung in rückläufiger Form zur Analyse gegeben wird. Beim Durchlaufen des Automaten wird für jede erkannte Silbenrepräsentation eine Featurestruktur mit den Elementen `{A:, N:, E:, marker:}` aufgebaut. Als Hilfskonstrukt wird ein Stack für noch nicht zugeordnete, aber schon abgearbeitete Zeichen verwendet, da die genaue Position einer Silbenrepräsentationsgrenze erst festgelegt werden kann, wenn ein folgendes (bzw. durch die Umkehrung: vorangehendes) Vokalgraphem gefunden wurde.

Tabelle 7.2 listet die möglichen Zustandsübergänge auf, wobei zwischen drei Eingabezeichenklassen unterschieden wird: Vokalbuchstaben (V), Konsonantenbuchstaben (K) und eindeutigen Grenzmarkierungen (#). Das grundsätzliche Vorgehen des Parsers lässt sich wie folgt zusammenfassen:

- Durchlaufe die Wortform rückwärts.
- Auftretende Vokalbuchstaben sind Zeichen für Silbenkerne. Mehrere aufeinanderfolgende Vokalbuchstaben können als komplexes Vokalgraphem einen Kern bilden.
- Die vor einem Vokalbuchstaben auftretenden Konsonantenbuchstaben werden aufgesammelt und dann auf Anfangs- und Endrandrepräsentation verteilt, wenn ein neuer Vokalbuchstabe auftaucht. Die Aufteilung behandelt genau einen Buchstaben als Anfangsrandrepräsentation, alle anderen als Endrandrepräsentation. Ausnahme sind die komplexen Grapheme <ch> und <sch>, die vollständig als Anfangsrandrepräsentation analysiert werden.
- Tritt ein <h> in Endrandposition auf, wird anhand des vorausgehenden Buchstaben markiert, um was für einen Typ es sich handelt:
  1. Dehnungs <h>: Wenn vor dem <h> ein <l>, <m>, <n>, <r> aufgetreten ist. Als Marker für die analysierte Silbe wird `Pdehn` gesetzt.
  2. Silbentrennendes <h>, phonologische Bedingung erfüllt: Wenn vor dem <h> ein Vokalbuchstabe aufgetreten ist. Als Marker für die analysierte Silbe wird `Msilbh` gesetzt.
  3. Silbentrennendes <h>, phonologische Bedingung nicht erfüllt (Konstantschreibung angenommen): Wenn vor dem <h> kein Vokalbuchstabe aufgetreten ist. Als Marker für die analysierte Silbe wird `Psilbh` gesetzt.
- Bei Schärfungsschreibungen werden die beiden Konsonantenbuchstaben zu einem komplexen Graphem zusammengefasst und im Fall erfüllter phonologischer Bedingung (fester Anschluss in offener Silbe) vollständig als Anfangsrandrepräsentation gespeichert. Die Endrandrepräsentation erhält als Spur eine Markierung (`!`).

Insgesamt implementiert der Wortparser eine relativ einfache Heuristik, die sich mit anderen Silbenparser-Ansätzen vergleichen lässt (vgl. Maas et al., 1999).

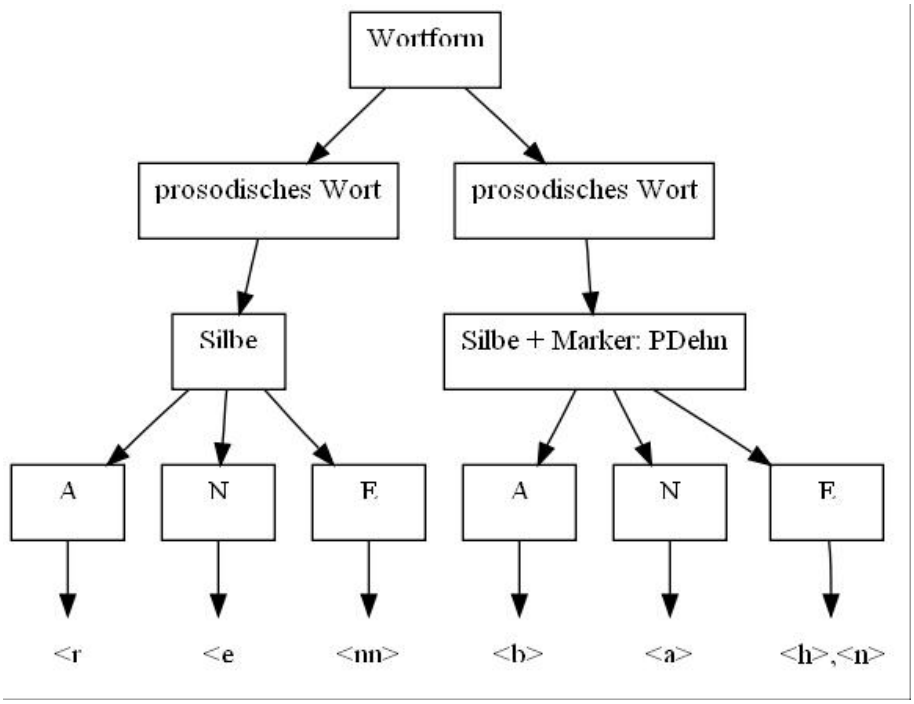

Abbildung 7.2.: Resultat eines Parserdurchlaufs

7.2.4. Silbentyp-Zuweisung

Der vom Silbenrepräsentationsparser erzeugte Baum wird von der Komponente *Silbentyp-Zuweisung* analysiert. Ihr Sinn ist es, jeder analysierten Silbe einen der drei Typen prominent (S+), reduziert (S^0) oder nicht-prominent, nicht-reduziert (S) zuzuweisen.

Ausgangspunkt ist eine vom Wortparser generierte Struktur wie in Abbildung 7.1. Ein etwas komplexeres Beispiel zeigt Abbildung 7.2. Hier wurde die zweite Silbe (als ganzes) mit dem Marker `Pdehn` versehen. Die Schärfungsschreibung in der ersten Silbe wurde zwar zu einem Graphem zusammengefasst, aber bislang nicht gesondert markiert.

Der Silbentyp-Zuweisungs-Algorithmus behandelt jedes prosodische Wort isoliert. Da in jedem prosodischen Wort genau eine prominente Silbe vorkommt und es keine Abhängigkeiten bezüglich zulässiger Silbentypen und Silbentypkombinationen zwischen prosodischen Wörtern gibt, ist dieses Verfahren sinnvoll und zulässig. Das Verfahren kann damit allerdings keine Differenzierung nach Haupt- und Nebenakzent in Wortformen mit mehreren prosodischen Wörtern leisten. Im Folgenden wird eine solche Differenzierung auch nicht benötigt, anderenfalls könnte aber eine nachgelagerte Komponente anhand einer Heuristik die Akzentzuweisung vornehmen.

Für jedes prosodische Wort geht der Silbentyp-Zuweiser dreischrittig vor:

1. Nach sicheren Hinweisen für prominente Silben suchen und markieren.
2. Falls keine sichere prominente Silbe gefunden wurde, wird eine Heuristik angewendet.
3. Den nicht-prominenten Silben wird anhand einer Heuristik der Typ S^0 oder S zugewiesen.

Erkennung sicherer Hinweise auf prominente Silben

Prominente Silben sind sicher erkennbar, wenn Dehnung, Schärfung oder silbentrennendes <h> vorliegen. Alle drei Fälle wurden vom Wortparser so aufbereitet, dass sie leicht erkannt werden können:

- Dehnungsmarkierung durch den Silbenmarker `Pdehn`
- Silbentrennendes <h> durch die Silbenmarker `Msilbh` und `Psilbh`
- Schärfung durch die Markierung ! oder einen Schärfungskonsonantenbuchstaben-Cluster im Endrand

Absolute Sicherheit bietet dieses Vorgehen nicht, da es sich auf die korrekte Zuordnung durch den Wortparser verlassen muss. Insbesondere Fälle scheinbarer Schärfungsschreibung wie z.B. <illegal> werden falsch klassifiziert.

Heuristik für die Identifizierung prominenter Silben

Falls keiner der drei genannten Fälle zutrifft, greift eine Heuristik ein:

- Bei Einsilbern ist die einzige vorhandene Silbe prominent
- Bei Zweisilbern wird die erste Silbe als prominent markiert
- Bei drei- und mehrsilbigen Formen:
  - Wenn die n>=1 letzten Silben einen durch <e> repräsentierten Kern aufweisen, wird die von hinten betrachtet erste Silbe als prominent markiert, deren Kern nicht durch ein <e> repräsentiert wird. Die Suche wird bei der viertletzten Silbe abgebrochen, da mehr als drei Reduktionssilben am Wortende sehr selten sind. Das viertletzte oder, falls die Wortform kürzer ist, erste <e> im Wort wird dann als Kernrepräsentation einer prominenten Silbe angenommen.
  - Wenn die Wortform nicht auf eine Silbe endet, deren Kern durch <e> repräsentiert wird, wird die vorletzte Silbe als prominent angenommen.

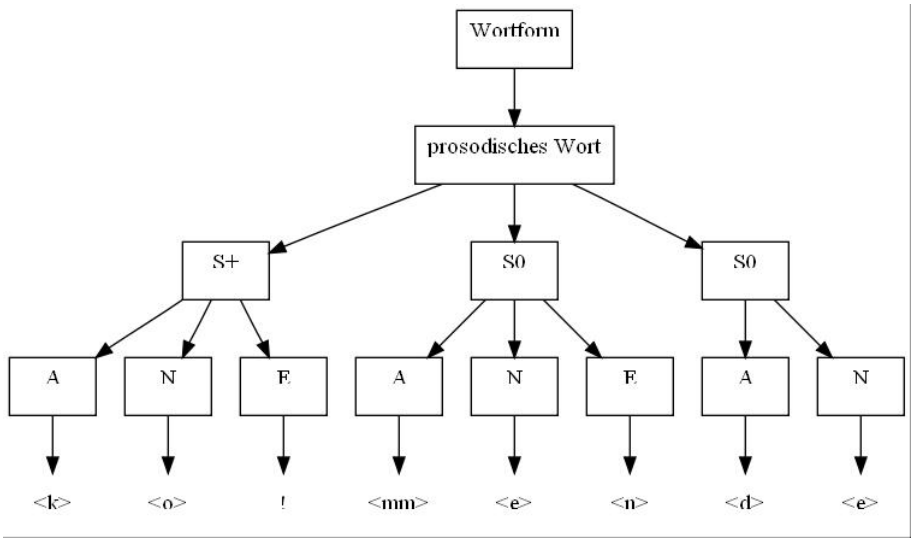

Abbildung 7.3.: Resultat eines Parserdurchlaufs

Klassifizierung nicht-prominenter Silben

Alle noch nicht klassifizierten Silben werden nun als Reduktionssilben (S^0) oder nicht-prominente, nicht-reduzierte Silben klassifiziert. Als Reduktionssilben werden nur Silben klassifiziert, deren Kern durch <e> repräsentiert wird, eine Erkennung von Endsilben <ig> oder <ich> wurde nicht implementiert. Für alle Nicht-<e>-Silben wird der Typ S angenommen, bei <e>-Silben greift folgende Heuristik:

- Vom Wortende her betrachtet, werden alle <e>-Silben als Reduktionssilben markiert, bis eine S+ oder S-Silbe auftritt. Alle vorangehenden (d.h. aus Sicht des rückläufigen Algorithmus: folgenden) <e>-Silben werden als S klassifiziert.
- Ausnahme bildet die erste Silbe eines prosodischen Wortes, die immer, wenn ihr Kern durch <e> repräsentiert wird, als Reduktionssilbe markiert wird. Damit werden Präfixe wie <ge> und <be> korrekt analysiert, die insgesamt selteneren Fälle wie <entscheiden> bekommen durch diese Heuristik allerdings einen falschen Silbentyp zugewiesen. Eine genauere Differenzierung wäre möglich, wurde aber für die vorliegende Untersuchung nicht implementiert.

Mit Abschluss dieser Schritte sind alle Silben eines prosodischen Wortes mit einem der drei Silbentypen markiert. Es gibt exakt eine prominente Silbe, die vorzugsweise aufgrund sicher hinweisender orthographischer Markierungen erkannt wird. Einige zusätzliche Heuristiken liefern keine absolut sichereren Zuordnungen, aber insgesamt gute Resultate.

7.2.5. Feature-Zuweisung

Nach Abschluss der Silbentyp-Zuweisung liegen Parsebäume wie in Abbildung 7.3 vor. Die abschließende Feature-Zuweisung propagiert nun übergreifende Features als Markierungen in die Blätter und erkennt einige zusätzliche orthographische Besonderheiten.

Die Tokenisierung bzw. Graphemzerlegung des Wordparsers ist nicht vollständig, da nur bis zur Ebene vollständiger Konstituentenrepräsentationen analysiert wurde. Konsonantengraphemcluster wie z.B. <schr> in <schrift> liegen also noch unaufgelöst vor. Zunächst wird der `OrthoAnalyse-Tokenizer` (s. 7.3) verwendet, um alle Graphemcluster zu tokenisieren.

Für jedes Graphem wird im Folgenden ein Featureset aufgebaut. Es enthält neben dem Graphem selbst (Feature `{gr:}`) weitere Markierungen, die Blatt-Kategorien in der später zu füllenden Analysetabelle entsprechen. Im Einzelnen sind dies die in Tabelle 7.3 aufgeführten Fälle. Besonders beachtenswert sind auch hier wieder einige Heuristiken.

Für die Erkennung loser und fester Anschlüsse bzw. von Repräsentationen von Kurz- und Langvokalen in prominenten Silben wird angenommen, dass offene Silben, in deren Endrandanalyse keine Spur `!` für eine phonologisch begründete Schärfungsmarkierung gefunden wird, immer auf einen losen Anschluss/Langvokal hindeuten. Diese Annahme ist unproblematisch. Schwieriger ist der Umkehrschluss, dass eine geschlossene Silbe, in der sich keine explizite Dehnungsmarkierung findet auf einen festen Anschluss hinweist. Formen wie <fragt> oder <Mond> werden somit falsch klassifiziert. Dennoch ist diese Annahme systematisch und statistisch gesehen als gut zu bewerten.

Sämtliche Markierungen, die auf Phänomene morphologischer Konstantschreibung hinweisen, sind lediglich Hypothesen, da der Algorithmus nicht überprüft, ob tatsächlich entsprechende Formen existieren. In der Praxis würde das eine Reihe komplexer Probleme aufwerfen, die z.B. auf die „Nähe“ der abzugleichenden Form abzielen und ein vollständiges Lexikon mit paradigmatischen Informationen und Wortbildungszusammenhängen erfordern. Mit dem vorliegenden Algorithmus werden also Formen wie <und>, <Magd> oder <ob> als morphologisch herleitbare Fälle von Auslautverhärtung klassifiziert.

7.2.6. Beispiele

Einige Beispiele sollen die Resultate der mehrschrittigen Analysekette noch einmal verdeutlichen.

<kommende>

```
(s+)
A: [{'feat': ['Sanf_einf'], 'gr': 'k'}]
N: [{'feat': ['Ss+_fest'], 'gr': 'o'}]
E: []
(s0)
A: [{'feat': ['Sanf_einf', 'Pschrf'], 'gr': 'mm'}]
N: [{'feat': ['Sred_silbson'], 'gr': 'e'}, {'feat': ['Sred_silbson'], 'gr': 'n'}]
E: []
(s0)
A: [{'feat': ['Sanf_einf'], 'gr': 'd'}]
N: [{'feat': ['Sred_schwa'], 'gr': 'e'}]
E: []
```

Feature	Konstit.	Bedingungen	Beispiel
`Sanf_einf`	A	einzelnes Graphem im Anfangsrand	Haus
`Sanf_komplex`	A	Bestandteil eines komplexen Anfangsrandes	Schrank
`Pschrf`	A	Schärfungsrepräsentation für phonologisch begründeten Schärfungsfall	Bett̲en
`Ss+_fest`	N	Nukleus einbuchstabig, Endrand belegt, aber nicht mit <h> oder <r>: fester Anschluss	Band, Bett, bester
`Ss+_lang`	N	Nukleus einbuchstabig, Endrand leer oder beginnt mit <h>, aber nicht <hr>	baden, Bahn
`Ss+_lang`	N	Nukleus <aa>, <ee>, <ie> oder <oo>, Endrand beginnt nicht mit <r>	Zoo, Dieb
`Ss+_oeffdiphth`	N	Nukleus einbuchstabig, Endrand belegt und beginnt mit <hr> oder <r>: Öffnender Diphthong	Fahrt, her
`Ss+_diphth`	N	Nukleus mehrbuchstabig, keiner der vorherigen Fälle	Haus
`Ss`	N	Kern(bestandteil) einer nicht-reduzierten, nicht-prominenten Silbe	Auto
`Sred_schwa`	N	Kern einer Reduktionssilbe, Endrand leer	Rede
`Sred_vokr`	N / E	Kern einer Reduktionssilbe und folgendes <r> im Endrand. Beide Grapheme bekommen das Feature und werden dem Nukleus zugewiesen (aber weiterhin als zwei getrennte Grapheme, Begründung s. 7.4)	Vater
`Sred_silbson`	N / E	Kern einer Reduktionssilbe und folgendes <l>, <m> oder <n> im Endrand. Beide Grapheme bekommen das Feature und werden dem Nukleus zugewiesen (aber weiterhin als zwei getrennte Grapheme, Begründung s. 7.4)	Löffel, gehen
`Send_einf`	E	einzelnes Graphem im Endrand	Haus
`Send_komplex`	E	Bestandteil eines komplexen Endrandes	fragt
`Pdehn`	E	<h> am Beginn eines Endrandes in `Pdehn`-markierter Silbe	Zahn
`Psilbh`	E	<h> am Beginn eines Endrandes in `Psilbh`-markierter Silbe	gehen
`Msilbh`	E	<h> am Beginn eines Endrandes in `Msilbh`-markierter Silbe	geht
`Mauslv`	E	<d>, <b>, <g> im Endrand: morphologisch begründbare Auslautverhärtung angenommen	Band, fragt
`Mschrf`	E	Schärfungsgraphem am Beginn des Endrandes: morphologisch begründbare Schärfung angenommen	bannt, Bett

Tabelle 7.3.: Bei der Feature-Zuweisung vergebene Features

<**Fahrtziel**>

Eingabe ist `fahrt#ziel`.

```
(s+)
A: [{'feat': ['Sanf_einf'], 'gr': 'f'}]
N: [{'feat': ['Ss+_oeffdiphth'], 'gr': 'a'}]
E: [{'feat': ['Send_komplex', 'Pdehn'], 'gr': 'h'},
    {'feat': ['Send_komplex'], 'gr': 'r'},
    {'feat': ['Send_komplex'], 'gr': 't'}]
#
(s+)
A: [{'feat': ['Sanf_einf'], 'gr': 'z'}]
N: [{'feat': ['Ss+_lang'], 'gr': 'ie'}]
E: [{'feat': ['Send_einf'], 'gr': 'l'}]
```

<**Wald**>

```
(s+)
A: [{'feat': ['Sanf_einf'], 'gr': 'W'}]
N: [{'feat': ['Ss+_fest'], 'gr': 'a'}]
E: [{'feat': ['Send_komplex'], 'gr': 'l'},
    {'feat': ['Send_komplex', 'Mauslv'], 'gr': 'd'}]
```

7.2.7. Bewertung

Das vorgestellte Verfahren arbeitet vergleichsweise einfach mit einer Reihe von Heuristiken. Es ist sehr effizient, weist aber eine Reihe von Einschränkungen in der Erkennungssicherheit auf, die zum Teil durch aufwendigerere Verfahren umgangen werden könnten. Die Verwendung von Ausnahmelexika ermöglicht es aber, nicht korrekt analysierbare Fälle entweder manuell zu erfassen oder mit anderen Verfahren vorab zu analysieren und transparent in die Gesamtanalyse einzubinden.

7.3. Stringvergleich von Zielschreibung und beobachteter Schreibung

Im zweiten Schritt des Gesamtanalyseverfahrens wird die annotierte Zielschreibung mit der beobachteten Schreibung verglichen. Hierzu findet als Standard-Algorithmus der Levenshtein-Distance-Algorithmus (s. Stephen, 1994) Verwendung. Die im Paket `Comparer` implementierte Levenshtein-Algorithmus-Klasse `ld` ist dabei generisch gehalten und verwendet ein übergebenes Objekt vom Typ `Tokenizer` zur Berechnung der Kosten. Damit ist die Grundimplementation von unterschiedlich definierten Kostenfunktionen unabhängig. `ld`-Objekte sind in der Lage, nicht nur den Editierabstand als numerischen Wert, sondern auch gefundene Alignments zu liefern. Die Methode `oneAlignment` liefert dabei eines der möglicherweise mehreren besten Alignments, die Methode `allAlignments` alle. Für die hier beschriebene Implementation der Gesamtanalyse wurde `oneAlignemt` verwendet. Ein Vergleich der Analysen unterschiedlicher Alignments wäre einfach zu implementieren, hat aber Effienzverluste zur Folge.

	+	-	*	n.v.
+	0	1	0	5
-	1	0	0	5
	0	0	0	5
n.v.	5	5	5	5

Tabelle 7.4.: Ersetzungskosten für die vier Featurebelegungen

Der Vergleich zwischen annotierter Zielschreibung und beobachteter Schreibung könnte direkt auf Buchstabenebene erfolgen. Die dafür als Kostenfunktion verwendbare Klasse `FastTokenizer` ist vorhanden, liefert aber im Hinblick auf das zu füllende Analyseschema unbefriedigende Ergebnisse. Zum einen sind Misalignments aufgrund unterschiedlicher Zeichenzahl bei eigentlich gleicher Graphemzahl unvermeidlich, zum anderen ist die Kategorien „phonetisch plausibel“ mit den Ergebnissen nicht zu unterstützen, weil Buchstaben entweder gleich oder verschieden sind, und keine graduelleren Ähnlichkeitswerte erhalten können.

Die beobachtete Schreibung durchläuft daher auch einen einfachen Tokenisierungs- und Annotationsprozess mit dem `OrthoAnalyseTokenizer`, der bereits für die Graphemzerlegung von Einzelkonstituenten bei der Zielschreibungsannotation verwendet wurde. Dieser Tokenizer verwendet eine relativ große Menge an Graphemen, die mit einigen wenigen phonologischen und orthographischen Features versehen werden. Hier ergibt sich wieder das Materialitätsproblem: Graphemen sind nicht ohne weiteres phonologische oder phonetische Features zuzuweisen, da sie in sehr unterschiedlichen Realisierungskontexten auftreten können. Diesem Problem wird auf zweierlei Weise durch Unterspezifikation begegnet: Durch Wahl einer recht kleinen Menge an Features, die nicht alle Eigenschaften der Grapheme oder damit in Verbindung gebrachten Phoneme abbilden und durch eine dreiwertige (implizit vierwertige) Belegung der Features. Die vier Belegungsvarianten sind *+, -, *, nicht vorhanden (implizit, n.v.)* und werden beim Matching relevant. Tabelle 7.4 gibt an, welche Belegungen im Matching zu welchen Kosten führen. Die Werte 1 und 5 sind dabei heuristisch gewählt. Der hohe Wert 5 sorgt dafür, dass in der Praxis keine inkonsistenten Featuresets so gematcht werden, dass sie zu einem besten Alignment gehören – die Kosten für Auslassung bzw. Einfügung sind jeweils mit 3 gewählt (s.u.). Ein Beispiel für die Unterspezifizierung durch * ist der Eintrag für <v>, der das Feature *stimmhaft* mit * belegt.

Zu den phonologisch motivierten Merkmalen zählt die Klassifikation als Vokalgraphem (`vok`) oder Konsonantengraphem (`kons`), einige artikulatorisch motivierte Features wie (`oben`)/(`unten`) (für Vokalgrapheme), nasal (`nas`), plosiv (`plos`) oder stimmhaft (`sth`). Orthographisch zu motivieren sind Kennzeichnungen wie (`dehn`), (`schrf`) oder (`kombionly`) für Buchstaben, die nicht alleinstehend als korrekte Grapheme auftreten.

Das Feature (`kombionly`) für die Buchstaben <c> und <q> ist dem Umstand geschuldet, dass auch offensichtlich fehlerhafte Formen wie *<cukt> tokenisiert werden können müssen. Schließlich wird der Tokenizer nicht nur auf die orthographisch korrekten Zielformen, sondern auch auf die potenziell fehlerhaften beobachteten Formen angewandt. Ein weiteres Beispiel ist das Graphem *<ngng>, das in orthographisch korrekten Schreibungen nicht vorkommt aber hier als Schärfungsmarkierung aufgenommen ist.

Nachfolgend sind alle Grapheme mit ihren Feature-Sets aufgeführt.

```
['a',   {'vok':'+', 'unten':'+', 'dehn':'-'}],
['aa',  {'vok':'+', 'unten':'+', 'dehn':'+'}],
['e' ,  {'vok':'+', 'unten':'-', 'dehn':'-'}],
['ee',  {'vok':'+', 'unten':'-', 'dehn':'+'}],
['o' ,  {'vok':'+', 'unten':'+', 'dehn':'-'}],
['oo',  {'vok':'+', 'unten':'+', 'dehn':'+'}],
['i' ,  {'vok':'+', 'unten':'-', 'dehn':'-'}],
['ie',  {'vok':'+', 'unten':'-', 'dehn':'+'}],
['u',   {'vok':'+', 'unten':'+', 'dehn':'+'}],
['ö',   {'vok':'+', 'unten':'+', 'dehn':'+'}],
['ü',   {'vok':'+', 'unten':'+', 'dehn':'+'}],
['ä',   {'vok':'+', 'unten':'+', 'dehn':'+'}],
['ei',  {'vok':'+', 'diphth':'+', 'dehn':'-'}],
['eu',  {'vok':'+', 'diphth':'+', 'dehn':'-'}],
['äu',  {'vok':'+', 'diphth':'+', 'dehn':'-'}],
['au',  {'vok':'+', 'diphth':'+', 'dehn':'-'}],
['y', {'vok':'+', 'dehn':'-'}],
['m' ,  {'kons':'+', 'nas':'+', 'lab':'+', 'schrf':'-'}],
['n' ,  {'kons':'+', 'nas':'+', 'dent':'+', 'schrf':'-'}],
['ng',  {'kons':'+', 'nas':'+', 'vel':'+', 'schrf':'-'}],
['mm' , {'kons':'+', 'nas':'+', 'lab':'+', 'schrf':'+'}],
['nn' , {'kons':'+', 'nas':'+', 'dent':'+', 'schrf':'+'}],
['ngng', {'kons':'+', 'nas':'+', 'vel':'+', 'schrf':'+'}],
['l' , {'kons':'+', 'lat':'+', 'schrf':'-'}],
['ll' , {'kons':'+', 'lat':'+', 'schrf':'+'}],
['r' , {'kons':'+', 'vel':'+', 'schrf':'-'}],
['rr' , {'kons':'+', 'vel':'+', 'schrf':'+'}],
['b' , {'kons':'+', 'plos':'+', 'sth':'+', 'lab':'+', 'schrf':'-'}],
['p' , {'kons':'+', 'plos':'+', 'sth':'-', 'lab':'+', 'schrf':'-'}],
['bb', {'kons':'+', 'plos':'+', 'sth':'+', 'lab':'+', 'schrf':'+'}],
['pp', {'kons':'+', 'plos':'+', 'sth':'-', 'lab':'+', 'schrf':'+'}],
['d' , {'kons':'+', 'plos':'+', 'sth':'+', 'dent':'+', 'schrf':'-'}],
['t' , {'kons':'+', 'plos':'+', 'sth':'-', 'dent':'+', 'schrf':'-'}],
['dd', {'kons':'+', 'plos':'+', 'sth':'+', 'dent':'+', 'schrf':'+'}],
['tt', {'kons':'+', 'plos':'+', 'sth':'-', 'dent':'+', 'schrf':'+'}],
['g' , {'kons':'+', 'plos':'+', 'sth':'+', 'palat':'+', 'schrf':'-'}],
['k' , {'kons':'+', 'plos':'+', 'sth':'-', 'palat':'+', 'schrf':'-'}],
['gg', {'kons':'+', 'plos':'+', 'sth':'+', 'palat':'+', 'schrf':'+'}],
['ck', {'kons':'+', 'plos':'+', 'sth':'-', 'palat':'+', 'schrf':'+'}],
['f',  {'kons':'+', 'frik':'+', 'sth':'-', 'lab':'+', 'schrf':'-'}],
['ff', {'kons':'+', 'frik':'+', 'sth':'-', 'lab':'+', 'schrf':'+'}],
['v',  {'kons':'+', 'frik':'+', 'sth':'*', 'lab':'+', 'schrf':'-'}],
['vv', {'kons':'+', 'frik':'+', 'sth':'*', 'lab':'+', 'schrf':'+'}],
['w',  {'kons':'+', 'frik':'+', 'sth':'+', 'lab':'+', 'schrf':'-'}],
['ww', {'kons':'+', 'frik':'+', 'sth':'+', 'lab':'+', 'schrf':'+'}],
['s',  {'kons':'+', 'frik':'+', 'sth':'*', 'dent':'+', 'schrf':'-'}],
['ss',  {'kons':'+', 'frik':'+', 'sth':'+', 'dent':'+', 'schrf':'+'}],
['ß',  {'kons':'+', 'frik':'+', 'sth':'-', 'dent':'+', 'schrf':'-'}],
['sch', {'kons':'+', 'frik':'+','sth':'-', 'schib':'+', 'schrf':'-'}],
['ch', {'kons':'+', 'frik':'+','sth':'-', 'vel':'+', 'schrf':'-'}],
['h', {'kons':'+', 'glott':'+', 'schrf':'-'}],
['x', {'kons':'+', 'afrik':'+', 'schrf':'-'}],
['q', {'kons':'+', 'kombionly':'+', 'schrf':'-'}],
['c', {'kons':'+', 'kombionly':'+', 'schrf':'-'}],
['j', {'kons':'+', 'glide':'+', 'schrf':'-'}],
['z', {'kons':'+', 'afrik':'+', 'schrf':'-'}],
['tz', {'kons':'+', 'afrik':'+', 'schrf':'+'}]]
```

	<a>	<b>	<d>	<tt>	<ie>
<a>	0	15	15	15	2
<b>	25	0	5	7	25
<d>	25	5	0	2	25
<tt>	25	7	2	0	25
<ie>	2	15	15	15	0

Tabelle 7.5.: Ersetzungskosten für die vier Featurebelegungen

Der Levenshtein-Algorithmus versucht die beiden vom Tokenizer gelieferten Graphemketten aufeinander abzubilden und verwendet folgende Kostenfunktion des Tokenizers:

- Graphem ausgelassen: 3
- Graphem eingefügt: 3
- Identisches Graphem: 0
- Vertauschung nichtidentischer Grapheme: Ergebnis des Featurematchings

Das Featurematching durchläuft alle Features eines Graphems der Zielschreibung und versucht sie mit den Features den potenziell zuzuordnenden Graphems der beobachteten Schreibung abzugleichen. Gemäß Tabelle 7.4 schlagen dabei inkonsistente Feature-Sets (z.B. zwischen Vokalgraphemen und Konsonantengraphemen) mit besonders hohen Kosten zu Buche. Durch die unterspezifizierenden Merkmale können auch solche Grapheme mit Ersetzungskosten von 0 belegt werden, die nicht identisch, aber entsprechend der Definition phonetischer Plausibilität im Analsyseschema „phonetisch ähnlich“ realisiert werden. Tabelle 7.5 zeigt einige exemplarische Graphem-Ersetzungskosten.

7.4. Generierung der Analysetabelle

Ergebnis des Stringvergleichs ist eine Zuordnungsliste von annotierten Graphemen der Zielschreibung und nicht-annotierten Graphemen der beobachteten Schreibung. Damit stehen alle Informationen zur Verfügung, um die Analysetabelle zu füllen.

Der `TableExplainer` iteriert über die Liste der Zuordnungen, die als Tupel <Zielschreibungsgraphem - Graphem der beobachteten Schreibung> vorliegen und trifft folgende Fallunterscheidung:

1. Leeres Zielschreibungsgraphem: Eingefügtes Graphem ohne Entsprechung in der Zielschreibung, also Wert für „Einfügung“ in der Analysetabelle erhöhen.
2. Leeres Graphem der beobachteten Schreibung: Für jedes Feature des Zielschreibungsgraphems den Wert für `max` um 1 erhöhen, die Werte für „repräsentiert“ `rep` (und folglich auch „phonetisch plausibel“ `phon` und „orthographisch korrekt“ `ortho`) unberührt lassen. Ein ausgelassenes Graphem wird also dahingehend behandelt, dass die mit ihm verbundenen Eigenschaften des Wortes als nicht repräsentiert gezählt werden.
3. Grapheme gleich. Für alle Features des Zielschreibungsgraphems werden die Werte für `max`, `rep`, `phon` und `ortho` um 1 erhöht.

4. Grapheme nicht gleich. Hier werden mehrere Aktionen angestoßen:

   a) Trägt das Graphem der beobachteten Schreibung das Merkmal `{schrf:+}`, das Graphem der Zielschreibung aber nicht, wird eine übergeneralisierende Schärfungsmarkierung angenommen und der Wert für `uebgen_schrf` inkrementiert.

   b) Trägt das Graphem der beobachteten Schreibung das Merkmal `{dehn:+}`, das Graphem der Zielschreibung aber nicht, wird eine übergeneralisierende Dehnungsmarkierung angenommen und der Wert für `uebgen_dehn` inkrementiert.

   c) Unabhängig davon, wie ähnlich die Graphemfeaturesets der zugeordneten Grapheme sind, wird für jedes Analysefeature des Zielschreibungsgraphems der Wert für `max` und `repr` erhöht. Das heißt, jede Realisierung eines Zielschreibungsgraphems wird als „repräsentiert“ gezählt.

   d) Sind die beiden Grapheme phonetisch ähnlich, wird auch der Wert für „phonetisch plausibel“ `phon` erhöht. Die Ähnlichkeit wird ebenfalls in der Tokenizer-Klasse definiert. In der vorliegenden Implementation werden Vokalgrapheme als ähnlich gewertet, wenn ihre Featurewerte für `unten` matchen, Konsonantengrapheme dann, wenn sie einen positiven Wert entweder für Nasal, Plosiv oder Frikativ aufweisen.

Diese Fallunterscheidung wird für jede Graphemzuordnung jeder zu untersuchenden Schreibung vorgenommen. In einem letzten Schritt werden die zusammenfassenden Kategorien der Analysetabelle durch Addition der untergeordneten Werte gebildet.

Die resultierende Tabelle kann in verschiedenen Formaten ausgegeben werden, implementiert sind HTML, LaTeX und SQL.

7.5. Beispiele für Gesamtanalysen

7.5.1. Beispiel: *<gomen> für <kommen>

1. Annotation der Zielschreibung <kommen>

```
Pros.Woerter:
[[{'a': 'k', 'marker': '', 'e': '!', 'n': 'o'},
  {'a': 'mm', 'marker': '', 'e': 'n', 'n': 'e'}]]
2 Silben.

kommen ->
(s+)
A: [{'feat': ['Sanf_einf'], 'gr': 'k'}]
N: [{'feat': ['Ss+_fest'], 'gr': 'o'}]
E: []
(s0)
A: [{'feat': ['Sanf_einf', 'Pschrf'], 'gr': 'mm'}]
N: [{'feat': ['Sred_silbson'], 'gr': 'e'}, {'feat': ['Sred_silbson'], 'gr': 'n'}]
E: []
```

2. Tokenisierung der Schreibung *<gomen>

```
Tokenisierung: [{'gr': 'g'}, {'gr': 'o'}, {'gr': 'm'}, {'gr': 'e'}, {'gr': 'n'}]
```

3. Stringvergleich

```
#2 Distance:
[({'feat': ['Sanf_einf'], 'gr': 'k', 'pos': 'a'}, {'gr': 'g'}),
 ({'feat': ['Ss+_fest'], 'gr': 'o', 'pos': 'n'}, {'gr': 'o'}),
 ({'feat': ['Sanf_einf', 'Pschrf'], 'gr': 'mm', 'pos': 'a'}, {'gr': 'm'}),
 ({'feat': ['Sred_silbson'], 'gr': 'e', 'pos': 'n'}, {'gr': 'e'}),
 ({'feat': ['Sred_silbson'], 'gr': 'n', 'pos': 'n'}, {'gr': 'n'})]
```

4. Berechnete Tabelle

Phänomen	max	repr.	%	plaus.	%	orth.korr.	%
Silbenkerne	3	3	100%	3	100%	3	100%
S'	1	1	100%	1	100%	1	100%
S' (fester Anschluss)	1	1	100%	1	100%	1	100%
S' (loser Anschluss)	0	0	0	0	0	0	0
Monophthong	0	0	0	0	0	0	0
Diphthong	0	0	0	0	0	0	0
Öffnender Diphthong	0	0	0	0	0	0	0
a: -> <ar>	0	0	0	0	0	0	0
S°	2	2	100%	2	100%	2	100%
Schwa	0	0	0	0	0	0	0
<er>	0	0	0	0	0	0	0
Silbischer Sonorant	2	2	100%	2	100%	2	100%
S	0	0	0	0	0	0	0
Anfangsränder	2	2	100%	2	100%	0	0%
einfach	2	2	100%	2	100%	0	0%
komplex	0	0	0	0	0	0	0
Endränder	0	0	0	0	0	0	0
einfach	0	0	0	0	0	0	0
komplex	0	0	0	0	0	0	0
Phonologische Markierungen	1	1	100%	0	0%	0	0%
Schärfung	1	1	100%	0	0%	0	0%
Dehnung	0	0	0	0	0	0	0
Silbentrennendes <h>	0	0	0	0	0	0	0
Konstantschreibung	0	0	0	0	0	0	0
Auslautverhärtung	0	0	0	0	0	0	0
Schärfung	0	0	0	0	0	0	0
ä/äu-Schreibung	0	0	0	0	0	0	0
Silbentrennendes <h>	0	0	0	0	0	0	0
Sonstiges							
Vertauschungen	0	0	0	0	0	0	0
Einfügungen	0	0	0	0	0	0	0
überfl. Schärfung	0	0	0	0	0	0	0
überfl. Dehnung	0	0	0	0	0	0	0
überfl. ä/äu-Graphie	0	0	0	0	0	0	0

7.6. Ergebnisse

7.6.1. Analyse des Gesamtcorpus

Das Gesamtcorpus besteht aus 37 Klassensätzen mit 750 Texten mit insgesamt 43685 Token und 6431 Types (unterschiedliche Token). Eine Gesamtanalyse ergibt folgende Ergebnisse:

Token gesamt/korrekt/falsch:	43685 / 32165 / 11520	Types gesamt/korrekt/falsch:	6431 / 2282 / 4149
Fehler Orth./Gramm./Or.+Gr.:	8931 / 1169 / 211	Fehler Orth./Gramm./Or.+Gr.:	3679 / 156 / 90
Fehler GKS:	1631	Fehler GKS:	404

Phänomen	max	repr.	%	plaus.	%	orth.korr.	%
Silbenkerne	66153	65381	98%	64626	97%	63253	95%
S’	47739	47553	99%	46894	98%	45703	95%
S’ (fester Anschluss)	24543	24469	99%	24322	99%	23926	97%
S’ (loser Anschluss)	23196	23084	99%	22572	97%	21777	93%
Monophthong	9184	9103	99%	9008	98%	8293	90%
Diphthong	6352	6332	99%	5960	93%	5960	93%
Öffnender Diphthong	7660	7649	99%	7604	99%	7524	98%
S0	16343	15775	96%	15700	96%	15625	95%
Schwa	5273	5125	97%	5120	97%	5111	96%
<er>	4676	4569	97%	4507	96%	4506	96%
Silbischer Sonorant	6394	6081	95%	6073	94%	6008	93%
S	2071	2053	99%	2032	98%	1925	92%
Anfangsränder	52524	51984	98%	51794	98%	49272	93%
einfach	45551	45153	99%	44974	98%	42884	94%
komplex	6973	6831	97%	6820	97%	6388	91%
Endränder	47058	45122	95%	44578	94%	41290	87%
einfach	25272	24525	97%	24137	95%	21817	86%
komplex	21786	20597	94%	20441	93%	19473	89%
Phonologische Markierungen	3645	3047	83%	1567	42%	1567	42%
Schärfung	2678	2666	99%	1190	44%	1190	44%
Dehnung	832	283	34%	280	33%	280	33%
Silbentrennendes <h>	135	98	72%	97	71%	97	71%
Konstantschreibung	11942	11360	95%	9032	75%	9032	75%
Auslautverhärtung	6619	6533	98%	6041	91%	6041	91%
Schärfung	4252	4238	99%	2441	57%	2441	57%
Silbentrennendes <h>	1071	589	54%	550	51%	550	51%
Sonstiges							
Einfügungen	3298	0	0%	0	0%	0	0%
überfl. Schärfung	457	0	0%	0	0%	0	0%
überfl. Dehnung	322	0	0%	0	0%	0	0%

Die Tabelle gibt die kumulierten Analysewerte für alle Token des Gesamtcorpus wieder. Angesichts der insgesamt relativ geringen Fehleranzahl (knapp 25% aller Schreibungen enthalten überhaupt Fehler), sind vor allem Werte unter 90% als auffällig zu betrachten. Exemplarisch seien einige auffällige Analyseregebnisse genannt:

- S’ (loser Anschluss) - Monophthong: orth.korrekt 90% aber phonetisch plausibel 98% - hauptsächlich zurückzuführen auf Schreibung von <ie> als <i>.
- Phonologische Markierung - Schärfung: orth. korrekt 44% - weniger als die Hälfte aller notwendigen Schärfungsmarkierungen wird bei anzuwendendem phonologischen Kriterium gesetzt.

- Konstantschreibung - Schärfung: orth. korrekt 57% - auch bei morphologisch zu begründender Schärfungsmarkierung werden nur ungefähr die Hälfte der notwendigen Markierungen korrekt vorgenommen. Der Wert ist höher als bei den phonologisch begründeten Fällen. Eine allgemeine Aussage über die Hintergründe dieser Beobachtung lässt sich daraus aber nicht ableiten. Zum einen ist das Corpus nicht repräsentativ für alle Schreibanfänger im deutschsprachigen Raum, so dass sich verallgemeinernde Aussagen bereits aus methodischen Gründen verbieten. Zum anderen müsste das verwendete Wortmaterial genauer untersucht werden. Es ist z.B. denkbar, dass die morphologisch begründeten Schärfungsmarkierungen tendenziell bei Wörtern mit höherer Frequenz vorkommen (z.B. <Bett> im Gegensatz zu <Hütte>) und die Schreibungen daher tendenziell eher bekannt sind, als hergeleitet werden müssen.
- Überflüssige Schärfung: 457 Fälle. Bei insgesamt 3631 korrekt vorgenommen Schärfungsmarkierungen liegt die Übergeneralisierungsquote bei ca. 12,6%.

Diese Schlussfolgerungen sind nur Beispiele für mögliche Analysen, die sich aus Gesamtbetrachtungen zu untersuchender Corpora ergeben können. Die Validität dieser Schlussfolgerungen hängt wesentlich von Faktoren der Corpuserstellung ab, die hier nicht im Vordergrund standen. Das Corpus wurde nur insofern als repräsentativ angesehen, als dass es möglichst viele der bei Schreibanfängern zu beobachtenden Schreibvarianten und Fehlertypen enthält und somit eine gute Testumgebung zur Optimierung der Algorithmen darstellt.

7.6.2. Analysen für Teilcorpora

Der Analyseprozess kann über Attribut-Filter gesteuert werden. Alle Attribute, die in den XML-Dateien verzeichnet sind, können zur Auswahl zu analysierender bzw. nicht zu analysierender Daten verwendet werden. Kombinationen von Filtern sind ebenfalls möglich.

Analyse aller Texte von männlichen Schreibern

Phänomen	max	repr.	%	plaus.	%	orth.korr.	%
Silbenkerne	27621	27239	98%	26930	97%	26359	95%
S’	19860	19762	99%	19502	98%	19003	95%
S’ (fester Anschluss)	10238	10193	99%	10134	98%	9961	97%
S’ (loser Anschluss)	9622	9569	99%	9368	97%	9042	93%
Monophthong	3746	3708	98%	3675	98%	3382	90%
Diphthong	2685	2678	99%	2529	94%	2529	94%
Öffnender Diphthong	3191	3183	99%	3164	99%	3131	98%
S0	6966	6691	96%	6656	95%	6616	94%
Schwa	2212	2141	96%	2139	96%	2134	96%
<er>	1982	1935	97%	1907	96%	1906	96%
Silbischer Sonorant	2772	2615	94%	2610	94%	2576	92%
S	795	786	98%	772	97%	740	93%
Anfangsränder	21762	21521	98%	21430	98%	20338	93%
einfach	18904	18723	99%	18634	98%	17735	93%
komplex	2858	2798	97%	2796	97%	2603	91%
Endränder	19600	18806	95%	18556	94%	17064	87%
einfach	10649	10332	97%	10157	95%	9069	85%
komplex	8951	8474	94%	8399	93%	7995	89%
Phonologische Markierungen	1493	1236	82%	616	41%	616	41%
Schaerfung	1084	1079	99%	461	42%	461	42%
Dehnung	349	118	33%	116	33%	116	33%
Silbentrennendes <h>	60	39	65%	39	65%	39	65%
Konstantschreibung	4981	4749	95%	3694	74%	3694	74%
Auslautverhärtung	2748	2711	98%	2488	90%	2488	90%
Schärfung	1811	1805	99%	991	54%	991	54%
Silbentrennendes <h>	422	233	55%	215	50%	215	50%
Sonstiges							
Einfügungen	1316	0	0%	0	0%	0	0%
überfl. Schärfung	177	0	0%	0	0%	0	0%
überfl. Dehnung	138	0	0%	0	0%	0	0%

Gegenüber den Werten der Gesamtanalyse sind hier beinahe durchgängig geringere Erfolgsquoten zu beobachten. Einzelne Phänomene fallen dabei aber nicht besonders ins Auge, so dass für das vorliegende Corpus für männliche Schreiber insgesamt leicht unterdurchschnittliche Rechtschreibleistungen konstatiert werden können.

Analyse aller Texte von weiblichen Schreiberinnen

Phänomen	max	repr.	%	plaus.	%	orth.korr.	%
Silbenkerne	27092	26862	99%	26589	98%	26073	96%
S’	19546	19490	99%	19242	98%	18796	96%
S’ (fester Anschluss)	9976	9957	99%	9902	99%	9764	97%
S’ (loser Anschluss)	9570	9533	99%	9340	97%	9032	94%
Monophthong	3925	3900	99%	3859	98%	3573	91%
Diphthong	2547	2536	99%	2400	94%	2400	94%
Öffnender Diphthong	3098	3097	99%	3081	99%	3059	98%
S0	6628	6459	97%	6439	97%	6418	96%
Schwa	2188	2146	98%	2145	98%	2142	97%
<er>	1776	1737	97%	1720	96%	1720	96%
Silbischer Sonorant	2664	2576	96%	2574	96%	2556	95%
S	918	913	99%	908	98%	859	93%
Anfangsränder	21582	21369	99%	21315	98%	20374	94%
einfach	18675	18526	99%	18479	98%	17696	94%
komplex	2907	2843	97%	2836	97%	2678	92%
Endränder	19145	18422	96%	18229	95%	17056	89%
einfach	10154	9897	97%	9760	96%	8948	88%
komplex	8991	8525	94%	8469	94%	8108	90%
Phonologische Markierungen	1519	1282	84%	722	47%	722	47%
Schaerfung	1118	1112	99%	554	49%	554	49%
Dehnung	352	130	36%	129	36%	129	36%
Silbentrennendes <h>	49	40	81%	39	79%	39	79%
Konstantschreibung	4724	4512	95%	3730	78%	3730	78%
Auslautverhärtung	2631	2602	98%	2441	92%	2441	92%
Schärfung	1663	1658	99%	1053	63%	1053	63%
Silbentrennendes <h>	430	252	58%	236	54%	236	54%
Sonstiges							
Einfügungen	1308	0	0%	0	0%	0	0%
überfl. Schärfung	208	0	0%	0	0%	0	0%
überfl. Dehnung	112	0	0%	0	0%	0	0%

Gegenüber den Werten der Gesamtanalyse und insbesondere der männlichen Vergleichsgruppe sind hier beinahe durchgängig höhere Erfolgsquoten zu beobachten. Einzelne Phänomene fallen dabei aber nicht besonders ins Auge, so dass für das vorliegende Corpus für weibliche Schreiberinnen insgesamt leicht überdurchschnittliche Rechtschreibleistungen konstatiert werden können.

8. Analyse von Groß- und Kleinschreibungsleistungen

Die in den vorangegangenen Abschnitten beschriebenen Verfahren beziehen sich sämtlich auf den phonographischen Teil der deutschen Orthographie. Für das zu Grunde liegende Szenario ist es jedoch wünschenswert, auch die anderen Bereiche miteinzubeziehen. Für die Groß- und Kleinschreibung wurde für das Projekt „Entwicklung eines linguistisch orientierten Rechtschreibkonzepts für alemannische sprechende HauptschülerInnen“[1] ein Analyseverfahren entwickelt, das vor allem die manuelle Auswertung eines Diktattextes erleichtern sollte.

8.1. Grundsätzliches Vorgehen

Die Erklärung „Nomen und Nominalisierungen werden groß geschrieben“ ist außerordentlich weit verbreitet und findet sich auch in den didaktischen Darstellungen. Die linguistisch sauberere Erklärung „Expandierbare Kerne von Nominalphrasen werden duch Großschreibung markiert“ (vgl. Maas, 1992, S. 161ff.) wird mit der Begründung abgelehnt, dass die verwendeten Begriffe wie „expandierbar“, „Kern“ und „Nominalphrase“ für eine unterrichtliche Vermittlung nicht geeignet seien[2]. 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, z.B. mit der linguistisch begründeten Regel zur Kennzeichnung von Kernen in Nominalphrasen kongruieren.

Eine solche Analyse kann prinzipiell von Hand durchgeführt werden, indem gezählt wird, in welche der genannten Kategorien die beobachteten Fehler fallen. Für einen Schreiber, der nach der „Nomen“-Regeln handelt, wären nach obiger Aufzählung am wenigsten Fehler bei Wörtern der 1. Stufe, am meisten Fehler bei Wörtern der 3. Stufe zu erwarten. Zusätzlich muss berücksichtigt werden, wie häufig die zu zählenden Fallklassen überhaupt in dem zu schreibenden Text aufgetreten sind.

[1] Durchgeführt 2000-2002 an der PH Freiburg unter Leitung von Prof. Dr. Christa Röber-Siekmeyer (s.a. Noack, 2002a).

[2] S. aber Röber-Siekmeyer (1999) für eine Gegenposition

Eine erfolgreiche Analyse besteht darin, eine Strategie oder eine Menge von einander nicht widersprechenden 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.

Der Nachteil einer manuellen Analyse ist die Notwendigkeit, für jede neue Strategie den gesamten Auswertungsprozess zu wiederholen. Bei größeren Mengen von Texten wird der Aufwand schnell so groß, dass nur wenige Strategien in Betracht gezogen werden. Ein maschinelles Verfahren kann hingegen, wenn die Strategien und Testverfahren implementiert sind, beliebig viele Untersuchungen ohne weiteren Aufwand durchführen. Wird ein einheitliches Format zur Formalisierung der Strategien verwendet, ist es sogar möglich, automatisiert Zusammensetzungen von „atomaren“ Strategien zu testen.

8.2. Featureannotation von Texten

Im Folgenden wird unter einer „atomare 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 morphologische Informationen (“lexikalische Wortart“), syntaktische Informationen wie die Zugehörigkeit und die Position oder Rollen 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.

Tabelle 8.1 führt die für das hier beschriebene Experiment verwendeten Features zusammen mit einer weiter unten erläuterten Kodierungskonvention auf. Im vorliegenden Fall werden aussschließlich 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 Featureannotation 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 (vgl. 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>

Attribut	Erklärung	Kennzeichnung
„Lexikonwortart“ (POS)		
v	Verb	[v]
n	Nomen	[n]
adj	Adjektiv	[j]
adv	Adverb	[d]
p	Präposition	[p]
conj	Konjunktion	[c]
detminierer	Artikel	[a]
pro	Pronomen	[o]
quant	Quantifizierer	[q]
Position im Satz		
s_start	Start eines Satzes	[automatisch]
s_end	Ende eines Satzes	[automatisch durch '.']
:_vollst	Vollständiger Satz nach Doppelpunkt	[automatisch erkannt]
:_unvollst	Unvollständiger Satz nach Doppelpunkt	[automatisch erkannt]
Position und Rolle innerhalb einer NP		
in_np	Teil einer NP	[automatisch durch NP-Klammerung]
np_start	Start einer NP	[automatisch durch NP-Klammerung]
np_end	Ende einer NP	[automatisch durch NP-Klammerung]
np_head	Kopf einer NP	[H] wenn expandierbar, sonst [h]
np_det	NP mit Artikel	[automatisch]
sing_head	Kopf einer NP und Singular	[s]
expandable	Expandierbarer Kopf einer NP	[H]
expanded	Expandierter Kopf einer NP	[e]
Kontext		
det+_	Steht direkt nach det	[automatisch erkannt]
quant+_	Steht direkt nach quant	[automatisch erkannt]
“Nominalisierungs-Eigenschaften“		
nomin	„Nominalisierung“	[automatisch erkannt]
denomin	„Entnominalisierung“	[automatisch erkannt]
Semantische Kriterien		
abstr	Abstraktum (nur bei POS=n)	[A]
concr	Konkretum (nur bei POS=n)	[automatisch erkannt bei +n,-A]
name	Eigenname	[N]

Tabelle 8.1.: Für die Groß-/Kleinschreibungsanalyse verwendete Features

mit $\begin{bmatrix} +adj \\ +v \end{bmatrix}$ 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“, das adverbial oder nominal verwendet werden kann. Für die Konzeption einer Featureannotation ist damit zu beachten, dass „zu viele“ Alternativen erst einmal nicht schaden, sondern vorrangig sichergestellt werden soll, dass möglichst alle plausiblen Erklärungsmöglichkeiten enthalten sind.

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.

Im vorliegenden Beispiel sind keine Frequenzinformationen eingeflossen, so dass Hypothesen der Art „Die Wahrscheinlichkeit, Wort x (oder Wortklasse X) groß zu schreiben ist gleich dem Anteil,

zu dem Wort x (oder Wortklasse X) im Deutschen groß geschrieben vorkommt.“ nicht möglich sind. So wünschenswert eine solche Möglichkeit wäre, so schwierig ist sie zu realisieren, da dazu ein ausgewogenes Korpus von Texten notwendig ist. Die Ausgewogenheit ist hier um so kritischer, da nur solche Texte infrage kommen sollten, mit denen die Kinder bereits Kontakt hatten. Allerdings wäre es falsch anzunehmen, die in der amtlichen Regelung als Ausnahme bezeichnete „Nominalisierung“ komme in von Kindern gelesenen Texten nicht vor. Eine manuelle Auszählung auf den ersten zehn Seiten des Bandes „Harry Potter und der Feuerkelch“ (Rowling, 2000) ergibt folgende „Nominalisierungen“: <etwas Merkwürdiges>, <etwas Entsetzliches>, <die Älteren>, <im *Gehängten Mann*>, <dem Trockenen>, <Genaueres>, <von Neugierigen>, <etwas Merkwürdiges>, <auf etwas Hartem>, <das dumpfe Kratzen>, <trat Schweigen ein>, <das Knistern>, <ein Zischen>, <ohne Aufsehen>.

Die vorliegende Untersuchung wurde mit einem abgewandelten „Harry-Potter“-Text durchgeführt, der um einige nach dem „Nominalisierungs-Konzept“ kritische Formen angereichert wurde. Die vollständige Annotation mit positiven Featureausprägungen eines Satzes ist in (1) angegeben. Allerdings ist es sehr mühsam und fehleranfällig, einen vollständigen Text derartig zu annotieren. Einige der Features sind nicht voneinander unabhängig, [*+exandiert*] impliziert auch [*+expandierbar*], [*+abstrakt*] hat [*−konkret*]zur Folge usw. In Tabelle 8.1 gibt die dritte Spalte eine Notationskonvention an, die es mit Hilfe eines entsprechend arbeitenden Aufbreitungsprogramms ermöglicht, den manuellen Aufwand drastisch zu reduzieren. Der vollständige Text ist in Abbildung 8.1 dargestellt.

(1)
$$\begin{array}{llllllll}
\text{Harry} & \text{Potter} & \text{war} & \text{erst} & \text{seit} & \text{wenigen} & \text{Tagen} & \\
\left[\begin{array}{l}+n\\+name\\+s_start\\+in_np\\+np_start\\+np_head\end{array}\right] & \left[\begin{array}{l}+n\\+name\\+in_np\\+np_end\\+np_head\end{array}\right] & [\ +v\ ] & [\ +d\ ] & [\ +d\ ] & \left[\begin{array}{l}+q\\+in_np\\+np_start\end{array}\right] & \left[\begin{array}{l}+n\\+in_np\\+np_head\\+expandierbar\\+expandiert\\+abstrakt\end{array}\right] & \\
\text{in} & \text{der} & \text{Zauberschule.} & & & & & \\
[\ +p\ ] & \left[\begin{array}{l}+a\\+np_start\\+in_np\end{array}\right] & \left[\begin{array}{l}+n\\+s_end\\+np_end\\+in_np\\+np_head\\+expandierbar\\+det+_\\+sing_head\\+konkret\end{array}\right] & & & & &
\end{array}$$

8.3. Auswertung

Ein Originaltext $T = \{w_1, \ldots, w_n\}$ besteht aus Wörtern $w_i = \left[\begin{array}{l} gks : gks_i \\ f_1 : v_{i_1} \\ \ldots \\ f_m : v_{i_m} \end{array}\right]$, $f_i \in F, v_{i_j} \in \{+, -\}$.

Die Ermittlung der Featurewerte kann wie oben beschrieben manuell oder teilautomatisiert geschehen, im Folgenden wird angenommen, dass die Werte bekannt sind. Eine beobachtete Schreibung

```
In/p (der/a Zauberschule/nHs).
(Harry/nNHs Potter/nNHs) war/v erst/d seit/d
(wenigen/q Tagen/nAHe) in/p (der/a Zauberschule/nHs).
(Er/oh) fand/v (es/oh) (jeden/q Morgen/dnAHes) schwer/j,
(den/a Weg/nHs) (ins/pa Klassenzimmer/nHs) zu/p finden/v.
(Er/oh) raste/v über/p (viele/q Treppen/nHe) in/p
(dem/a verwinkelten/j Gebäude/nHes):
(enge/j, kurze/j, krumme/j, wacklige/j).
(Manche/q) führten/v freitags/dnA nicht/d zu/p (dem/a Gewohnten/Hjvs).
(Manche/q) hatten/v auf/p (halber/qj Höhe/nsAHe) (eine/a Stufe/nsH),
die/oa ganz/d plötzlich/d verschwand/v, und/c (man/ho) durfte/v
nicht/d vergessen/v, (dieses/ao unvorhersehbare/j Nichts/dHes)
zu/p überspringen/v. (Es/oh) gab/v auch/d (Türen/nH), die/oa
(sich/oh) nur/d öffneten/v, wenn/c (man/oh) (sie/oh)
höflich/dj bat/v oder/c (sie/oh) an/p (der/a richtigen/j
Stelle/nsHe) kitzelte/v.
(Es/oh) war/v auch/d schwierig/dj, (sich/oh) (daran/oh) zu/p erinnern/v,
wo/p (etwas/q Bestimmtes/vdHs) war/v,
denn/c (alles/qh) schien/v morgens/dnA ziemlich/d oft/d
(die/a angestammten/j Plätze/nHe) zu/p wechseln/v.
(Harry/nsNH Potter/nsNH) musste/v noch/d viel/q lernen/v,
um/c (das/a Geheimnisvolle/jsH) zu/p erreichen/v.
(Er/oh) ließ/v (sich/oh) von/p (vielem/qh) reizen/v.
Aber/c (es/oh) fiel/v (ihm/oh) anfangs/dnA sehr/d schwer/dj,
immer/d an/p (alles/qh) zu/p denken/v.
Immer/d wieder/d vergaß/v (er/oh) (etwas/qh).
Aber/c (er/oh) gab/v (das/a Hoffen/vsH) nicht/d auf/p.
```

Abbildung 8.1.: Annotierter Diktattext für die Groß-/Kleinschreibungsanalyse

(Diktat des Textes) $D^T = \{d_1^T \dots d_n^T\}$, enthält für jedes Wort d_i^T Informationen über die Majuskelverwendung: $d_i^T \in \{+, -, unbekannt\}$. Damit ist jede Schreibung des Textes als ein Vektor aufzufassen, der keine weiteren Informationen als die Groß- oder Kleinschreibung der einzelnen Wörter enthält. Dieser Vektor wird mit den w_i aus dem Originaltext abgeglichen, so dass die Indizes im Originaltext den Indizes der beobachteten Schreibung vollständig entsprechen. Falls Wörter nicht als einzelne Wörter abgetrennt sind, Wörter ausgelassen wurden oder die Verwendung eines Groß- oder Kleinbuchstabens, etwa bei Handschriften, nicht eindeutig feststellbar ist, wird der Wert „unbekannt“ angenommen. Daraufhin fallen diese Wörter genauso wie Einfügungen aus der weiteren Analyse heraus.

$T \leftarrow$ Originaltext mit Featureannotationen ($array(|T|, |F|)$)
$D \leftarrow$ Beobachtete Schreibung des Originaltextes (bereits aligned)
$F \leftarrow$ Liste der Strategien
$result \leftarrow array(|F|, 3)$

for all $f \in F$ **do**
  $result[f][0] \leftarrow 0$ {Anzahl Übereinstimmungen für Feature f}
  $result[f][1] \leftarrow 0$ {Anzahl der Nichtübereinstimmungen bei Großschreibung}
  $result[f][2] \leftarrow 0$ {Anzahl der Nichtübereinstimmungen bei Kleinschreibung}
  $unknown \leftarrow 0$ {Anzahl Wörter mit Klassifikation $'unbekannt'$}
  for $i \leftarrow 0$ to $|T|$ **do**
    if $T[i][f] = D[i]$ **then**
      $result[f][0] \leftarrow result[f][0] + 1$
    else if $D[i] = +$ **then**
      $result[f][1] \leftarrow result[f][1] + 1$
    else if $D[i] = -$ **then**
      $result[f][2] \leftarrow result[f][2] + 1$
    else
      $unknown \leftarrow unknown + 1$ {$D[i]$ = 'unbekannt'}
    end if
  end for
  $result[f][3] \leftarrow \frac{result[f][0]}{(|T| - unknown)}$ {Anteil Übereinstimmungen}
end for

Algorithmus 8.1: Bestimmung der Strategieleistungen

Für jedes Feature, d.h. jede einfache Strategie, werden, wie in Algorithmus 8.1 dargestellt, vier Werte errechnet. Die Anzahl der Übereinstimmungen gibt an, wie oft Merkmalsausprägung (+ oder -) und gewählte Großschreibung („groß“ = +, „klein“ = -) für das jeweilige Merkmal übereinstimmen. Die Nichtübereinstimmungen werden nach „Unterabdeckung“ und „Überabdeckung“ getrennt gezählt. Eine „Unterabdeckung“ ist dann gegeben, wenn bei negativer Merkmalsausprägung eine Großschreibung erfolgte, eine „Überabdeckung“ dann, wenn bei positiver Merkmalsausprägung klein geschrieben wurde. Schließlich wird noch ein Abdeckungsgrad zwischen 0 und 1 als Verhältnis von Übereinstimmungen und gesamter Wortmenge des geschriebenen Textes berechnet, der als „Erklärungsleistung“ der Strategie betrachtet werden kann. Für diese einfache Auswertung sind die Werte für Über- und Unterabdeckung nicht relevant, das Ergebnis kann nach Abschluss des Verfahren als nach Abdeckungsgrad sortierte Liste der Merkmale ausgegeben werden.

Aus Signifikanz und Häufigkeit (s. Tabelle 8.2) lässt sich der so genannte Entropie-Wert bilden, der

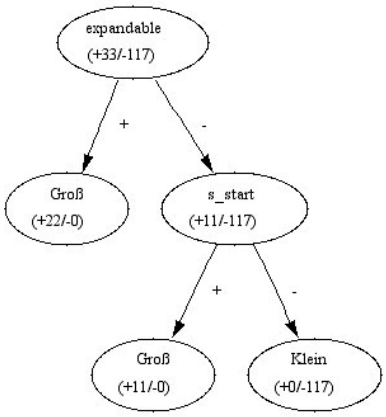

Abbildung 8.2.: ID3-Entscheidungsbaum für den orthographisch korrekten Text

ausdrückt, wie hoch die Erklärungsmächtigkeit eines ausgewählten Features ist. Werden nach und nach immer wieder die Attribute mit dem geringsten Entropiewert ausgesucht und zur Konstruktion eines Baumes verwendet, so ergibt sich daraus ein Entscheidungsbaum, der eine anschauliche Strategiehypothese für den untersuchten Schreiber darstellt. Berechnung der Werte und Konstruktion des Baumes wurden hier mit dem ID3-Verfahren (Quinlan, 1986) durchgeführt.

Der Algorithmus umfasst 6 Schritte:

1. Berechne für jedes Attribut seinen Entropie-Wert
2. Wähle das Element A mit der geringsten Entropie
3. Unterteile die Daten in separate Mengen, die jeweils einen bestimmten Wert für A aufweisen (z.B. +name / -name)
4. Konstruiere für jede dieser Mengen einen Ast des Baumes
5. Wiederhole den Prozess ab Schritt 1 für jeden der Äste
6. In jeder Iteration wird ein Attribut aus der Attributmenge entfernt. Der Prozess endet, wenn keine Attribute mehr verbleiben, oder wenn alle verbleibenden Daten den gleichen Wert (groß oder klein) haben.

Abbildung 8.2 zeigt den resultierenden Entscheidungsbaum für den vollständig korrekt geschriebenen Originaltext. Hier sind nur zwei Attribute relevant: expandierbar und s_start. Die Reihenfolge der verwendeten Attribute (expandierbar zuerst) ist darin begründet, dass der Informationsgewinn aufgrund der höheren Anzahl damit zu entscheidender Fälle größer ist.

8.4. Beispiel

Als Beispiel dient folgender Text, die wie in Abschnitt 8.2 beschrieben, mit dem featureannotierten Originaltext in Verbindung gebracht wurde:

> In der Zauberschule
> Harry Potter war erst seit wenigen Tagen in der Zauberschule. Er fand es jeden Morgen schwer den Weg ins Klassenzimmer zu finden. Er raste über viele Trepen in dem verwinkelten Gebäude enge kurze grumme waklige. Manche führten **Freitags** nicht zu dem Gewohnten. Manche hatten auf halber **höhe** eine Stufe die ganz plötzlich verschwand und man durfte nicht vergessen dieses unforherbares Nichts zu überspringen. Es gab auch Türe die sich nur öffneten wenn man sie höflich bat oder sie an der richtigen Stelle kitzelte. Es war auch schwierig sich daran zu erinern wo etwas **bestimtes** war den alles schien **Morgens** zimlich oft die angestanden Plätze zu wechseln. Harry Potter musste noch viel lernen um das **geheimnisvolle** zu erreichen. Er lies sich von vielem reizen. Aber es viel ihm **Anfangs** sehr schwer immer an alles zu denken. Aber er gab das **hoffen** nicht auf.[3]

In Tabelle 8.2 sind die Abdeckung und Signifikanz der einzelnen Attribute für den Text aufgeführt, Abbildung 8.3 zeigt den vollständigen, vom ID3-Verfahren generierten Entscheidungsbaum. Hier ist zu erkennen, dass das Feature `n` den größten Informationsgewinn bietet. Im linken Teilbaum findet sich nur noch ein einziger Fall von Kleinschreibung, der durch eine Kaskade von drei weiteren Entscheidungen herausgearbeitet werden muss. Im rechten Teilbaum bietet zunächst das Feature `s_start` einen großen Informationsgewinn, der Restbaum ist dafür zuständig, die zwei verbliebenen Großschreibungsfälle zu erklären. Die gemäß ID-Verfahren gefundene Erklärung entzieht sich einer kompakten sprachlichen Beschreibung, da sowohl Satzende (`s_end`) als auch die Unterscheidung zwischen `expandierbar` und `expandiert` eine Rolle spielt. Aus solchen Entscheidungsregeln, die letztlich nur Einzelfälle isolieren, lassen sich keine verallgemeinernden Rückschlüsse ziehen, da die Datenbasis dazu nicht ausreicht. Aus diesem Grunde werden Entscheidungsbäume häufig gekappt (pruning), wenn unterhalb eines Knotens nur noch geringe Informationsgewinne zu erzielen sind. Abbildung 8.4 zeigt den abgeschnittenen Entscheidungsbaum für den betrachteten Text. Hier spielen nur noch die Attribute `n` und `s_start` eine Rolle, in den abgeschnittenen Knoten ist angegeben, wie viele Fälle von Groß- und Kleinschreibung jeweils darunter zu fassen sind. So verbleiben nach der Entscheidung `+n` noch 19 Groß- und 1 Kleinschreibungsfall. Als Hypothese kann also formuliert werden, dass Regel `+n` für den Schreiber eine große, aber keine vollkommen sichere Rolle zu spielen scheint.

[3] Alle nicht auf die Groß- und Kleinschreibung bezogenen Fehler werden ignoriert. Groß- und Kleinschreibungsfehler sind hier der besseren Übersichtlichkeit wegen fett gesetzt. Der Text ist im Anhang im Abschnitt Groß- und Kleinschreibung unter der Bezeichnung KG11 zu finden.

Merkmal	Abdeckung	Signifikanz
+name	1.00	1.00
+s_start	1.00	1.00
+denomin	1.00	1.00
+concr	1.00	1.00
+n	0.95	0.90
+expanded	0.875	0.75
+abstr	0.8333	0.67
+expandable	0.8182	0.64
+sing_head	0.7778	0.56
+np_head	0.5476	0.10
+np_end	0.5349	0.07
+det+_	0.4286	-0.14
+nomin	0.4	-0.2
+in_np	0.3623	-0.28
+np_det	0.3571	-0.29
+quant+_	0.3333	-0.33
+conj	0.2857	-0.43
+pro	0.25	-0.50
+s_end	0.25	-0.50
+np_start	0.2326	-0.53
+adv	0.2083	-0.58
+quant	0.1818	-0.64
+adj	0.0667	-0.87
+p	0.0556	-0.89
+v	0.333	-0.93
+det	0.00	-1.00
+:_unvollst	0.00	-1.00

Tabelle 8.2.: Beispiel für errechnete Abdeckungsraten von Einzelstrategien

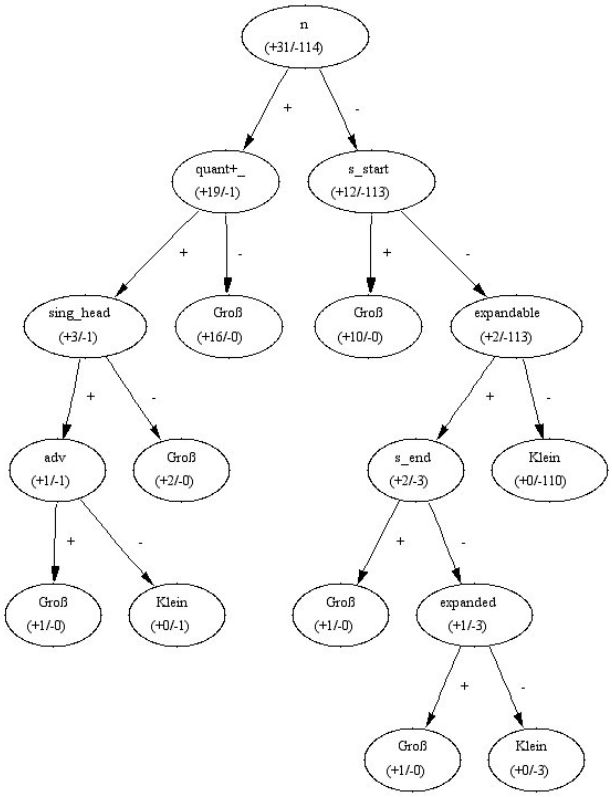

Abbildung 8.3.: Vollständiger ID3-Entscheidungsbaum für den fehlerhaften Text

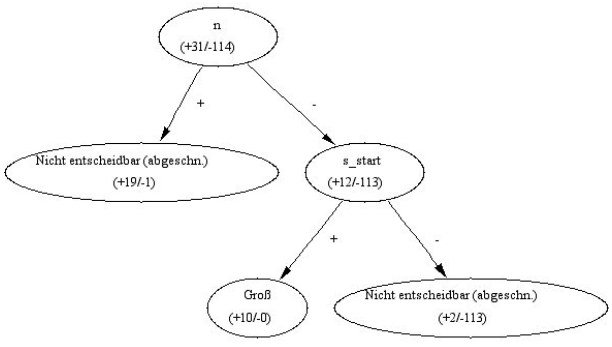

Abbildung 8.4.: Gekappter ID3-Entscheidungsbaum für den fehlerhaften Text

9. Anwendungen

9.1. Lehr-/Lernsoftware

Das in Kapitel 7 entwickelte Verfahren wird seit September 2004 für die Lernplattform *www.ich-will-schreiben-lernen.de*[1] des Deutschen Volkshochschul-Verbandes und des Bundesverbandes Alphabetisierung eingesetzt. Ziel dieser Lernumgebung ist es, den Lerner in die Lage zu versetzen, Wissen und Fähigkeiten im Umgang mit dem Gegenstandbereich „Schrift“ zu erwerben oder zu vertiefen.

Abbildung 9.1 zeigt die Einstiegsseite des Lernportals. Hauptziel ist es für den Gegenstandsbereich „Schreiben“, funktionalen Analphabeten, für die der Gang zur Behörde oder auch schon der alltägliche Einkauf mit großen Hindernissen verbunden ist, eine niedrigschwellige Möglichkeit anzubieten, selbstständige erste Schritte hin zu strukturiertem Lernen zu ermöglichen. Anspruch der Plattform ist es nicht, als alleiniges Hilfsmittel zum Lernerfolg zu führen. Stattdessen stehen Tutoren bereit, die Lerner zu begleiten und zu ermutigen, den für sie häufig schwierigen Schritt zu wagen, an einem Alphebetisierungskurs der Volkshochschulen teilzunehmen. Das Programm bietet eine Einstufung in sechs verschiedene Niveaustufen. Der Lernende erhält anschließend maßgeschneiderte Übungspakete. Zusätzlich bietet das Portal aktuelle Nachrichten, Hörtexte, Lesetexte und die Möglichkeit, über Voice-Mail mit anderen Lernenden in Kontakt zu kommen. Aktuelle und schwierige Texte können gelesen, aber auch gehört werden. Das in Kapitel 7 vorgestellte Verfahren wird dabei in modifizierter Form genutzt, um sowohl Lernenden wie dem Trainer Vorschläge zur Arbeit mit der Software zu machen und dem Trainer detaillierte Diagnosen zu präsentieren bzw. für den Lernenden automatisiert auf seine Fortschritte zugeschnittene Wochenarbeitspläne anzubieten.

Nach einer Selbsteinschätzung werden dem Lerner verschiedene Themenkomplexe („Lebenswelten“) angeboten, die die inhaltlichen Schwerpunkte seiner Arbeit mit der Umgebung festlegen. Hierzu zählen z.B. die Wohnungssuche, die Bewältigung von Bankgeschäften oder das Lesen von Hobby-Zeitschriften. Alle nachfolgend angebotenen Übungen, wie z.B. die in Abbildung 9.2, weisen in ihrer situativen Einbettung und beim Wortmaterial Bezüge zu den gewählten Lebenswelten auf.

Da die didaktische Konzeption auf der Grundlage bestehender Alphabetisierungskurse entwickelt wurde, konnte das in Kapitel 4 entwickelte Auswertungsschema nicht unverändert übernommen werden. Stattdessen wurden die Analysen auf ein vorgegebenes groberes Analyseschema übertragen, das in Tabelle 9.1 aufgeführt ist. Im Einzelfall sind die Kategorien bei genauerer linguistischer Analyse unangemessen bzw. weisen begriffliche Probleme auf. In fast allen Fällen konnten aber Abbildungen auf den in dieser Arbeit vorgeschlagenen Analyseansatz gefunden werden. Die anderen Fälle, wie z.B. die Verwechslung formähnlicher Buchstaben konnten durch einfache Listen bzw. eine Erweiterung des Stringabgleichsalgorithmus (s. Kapitel 7.3, S. 101) abgedeckt werden.

[1] Das Lernportal wurdem mittlerweile auch auf andere Gegenstandsbereiche der Grundbildung ausgeweitet und ist nun unter http://www.ich-will-lernen.de zu finden.

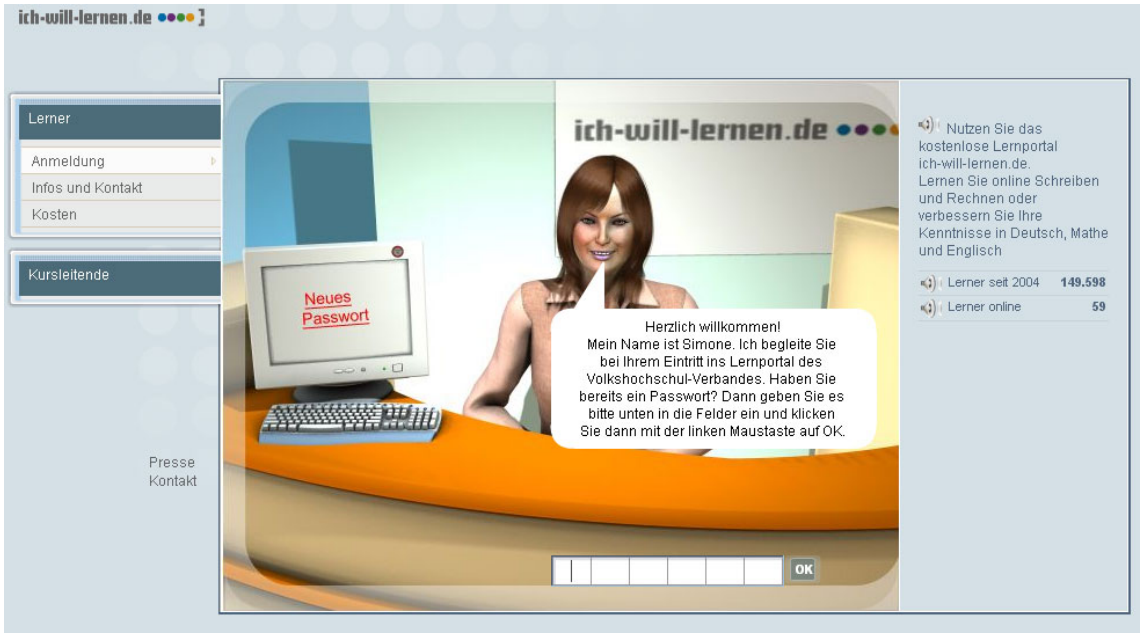

Abbildung 9.1.: Einstiegsseite der Lernumgebung www.ich-will-lernen.de

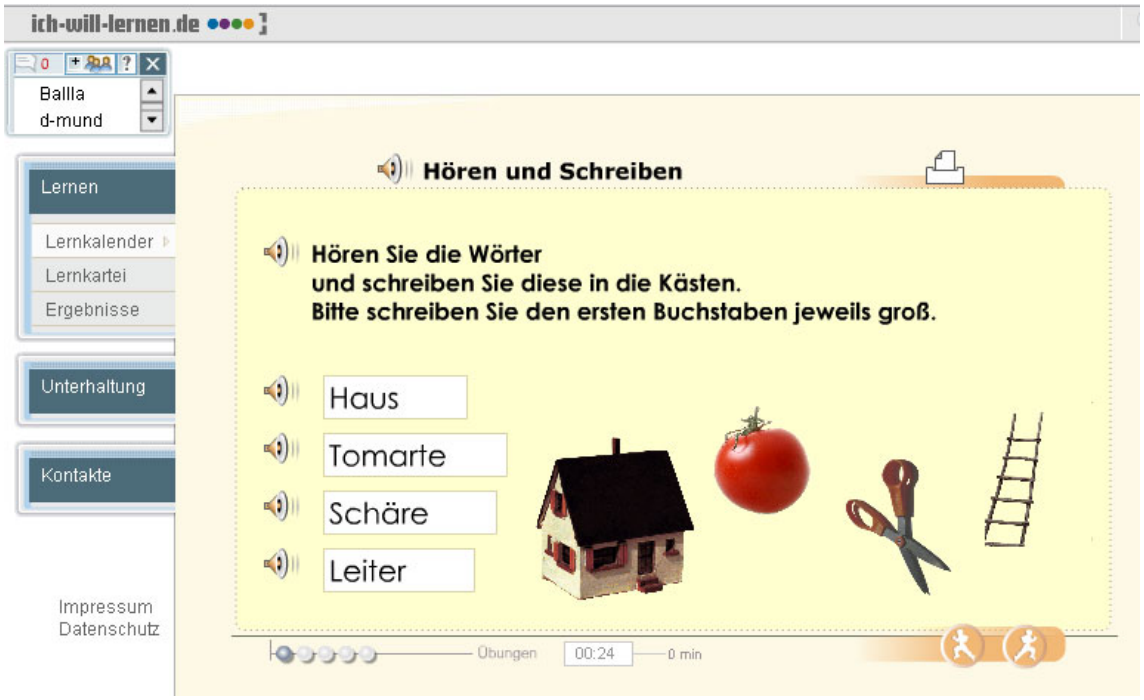

Abbildung 9.2.: Übung zu Einzelwortschreibungen

Nr	Sprachliches Thema	Beschreibung
1	Anlaute	Fehler in der Schreibung von Silbenanlauten
2	Auslaute	Fehler in der Schreibung von Silbenauslauten
3	Buchstabenreihenfolge	Fehler in der Buchstabenreihenfolge
4	Buchstabenverbindungen	Fehler bzgl. <sch>, <ch>, <st>, <sp>, <pf>
5	Lautgetreue Wörter	Fehler bei der Schreibung lautgetreuer Wörter
6	Selbstlaute	Fehler bei der Scheibung von Selbstlauten
6.1	Lange Selbstlaute	Fehler bei der Schreibung langer Selbstlaute
6.2	Kurze Selbstlaute	Fehler bei der Schreibung kurzer Selbstlaute
6.3	Ähnlich klingende Selbstlaute	Fehler bei der Schreibung von e/ä, ei/ai, eu/äu/oi, o/u, e/i (in best. Umgebungen)
7	Konsonanten	Fehler bei der Schreibung von Konsonanten
7.1	Mitlauthäufung	Fehler bei der Schreibung aufeinander folgender Konsonanten
8	Klangähnliche Buchstaben	Fehler durch Verwechslung von Buchstaben, deren mögliche Lautwerte Verwechslungsgefahr bergen (b/p, d/t, g/k, f/v/w, m/n)
9	Formähnliche Buchstaben	Fehler durch Verwechslung von Buchstaben, deren graphische Form ähnlich ist
10	Schärfung	Fehler bei der Schärfungsmarkierung
10.1	Schärfung	Dopplung von einem Mitlaut
10.2	Sonderfälle	Sonderfälle von Dopplung: ck/tz
11	S-Laute	Fehler bei der Schreibung von s-Lauten
11.1	<s>-Fehler	Fehler bei der Schreibung von <s>
11.2	<ß>-Fehler	Fehler bei der Schreibung von <ß>
11.3	<ss>-Fehler	Fehler bei der Schreibung von <ss>
12	Dehnung/ie	Fehler bei der markierten Schreibung von Langvokalen
12.1	Dehnungs-h	Fehler bei der Schreibung von Dehnungs-h (*<faren> statt <fahren>, *<Schuhle> statt <Schule>)
12.2	Doppelter Selbstlaut	Fehler bei der Schreibung doppelter Selbstlaute
13	Wortfamilien	Fehler bei der Schreibung, der Hinweis auf Wortfamilienmitglieder erlaubt (*<färt> statt <fahren>)

Tabelle 9.1.: Vorgegebene Analysekategorien des Volkshochschulverbandes

Abbildung 9.3.: Generierter Wochenplan

Das Interface zwischen der Lernportalsoftware und den Analysealgorithmen beschränkt sich auf zwei einfache Python-Funktionen:

`einschlaegige_kategorien(wort)`: Liefert Liste aller Kategorien, die evtl. als Fehler bei der Schreibung von `wort` auftreten könnten.

`fehlerkategorien(wort,schreibung,nurkategorien)`: Liefert für ein Paar aus (Fehl-) Schreibung und korrekter Schreibung eine Liste der gefundenen Fehlerkategorien. Mit `nurkategorien` kann eine eingeschränkte Liste übergeben werden, die die Analyse evtl. beschleunigt.

Die Funktion `einschlaegige_kategorien` wird für die Autorenumgebung verwendet. Die Autoren der Übungen und Übungstexte können aus den inhaltlichen definierten Wortschätzen („Lebenswelten“) damit gezielt solche Wortformen auswählen, die den in der jeweiligen Übung zu behandelnden sprachlichen Themen entsprechen. Die Funktion `fehlerkategorien` wird schließlich bei der Auswertung der Schreibleistungen verwendet, um ein einfaches Benutzermodell aufzubauen. Dieses Benutzermodell enthält für die 24 Kategorien Aussagen darüber, wie intensiv sie schon geübt wurden und wie hoch der jeweilige Fehleranteil (auch im Verlauf) war. Mithilfe dieser Informationen werden dann aus einem Übungspool die als nächstes zu bearbeitenden Aufgaben ausgewählt (vgl. Abbildung 9.3).

Das Portal ist als erfolgreich zu bewerten. Seit Beginn des Projektes haben mehr als 100.000 Nutzer das Portal genutzt, pro Woche sind derzeit mehr als 1.500 Lerner regelmäßig aktiv. Das Projekt wurde 2005 mit der Comenius-Medaille und 2006 mit dem Europäischen E-Learning Award eureleA und dem Deutschem Bildungssoftware-Preis digita ausgezeichnet (Deutscher Volkshochschulverband, 2006).

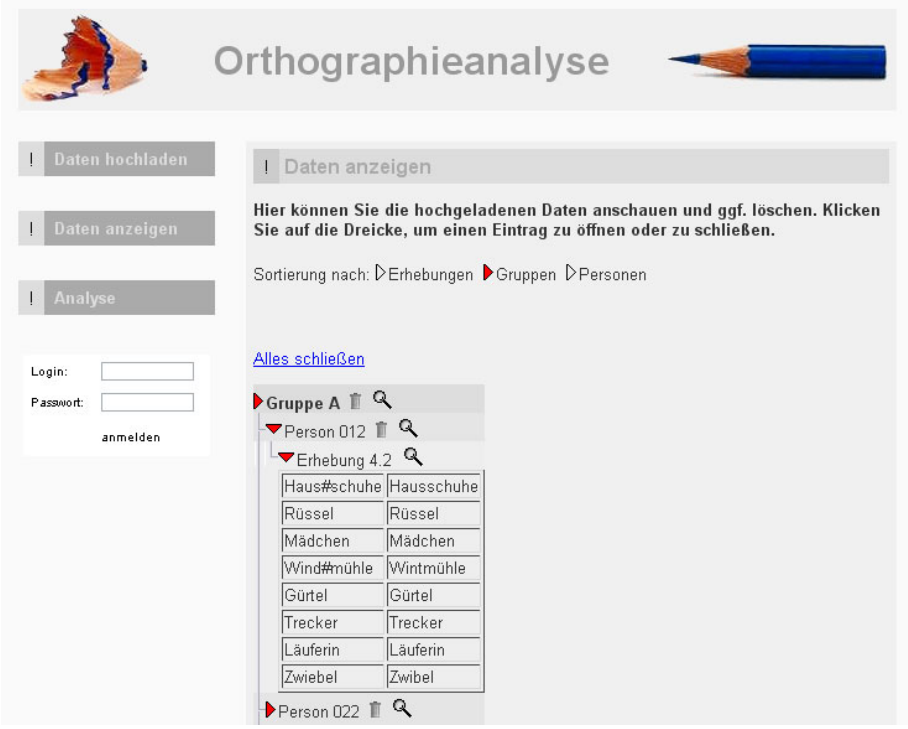

Abbildung 9.4.: Online-Analyse: Korpusansicht

9.2. Forschungsunterstützung

Vorläufer der hier entwickleten Analyseverfahren wurden bereits zur Unterstützung linguistischer und sprachdidaktischer Untersuchungen genutzt (s. Noack, 2000, 2002b; Mehlem, 2002). Es wurden aus dem Projekt „Computerbasierte Modellierung orthographischer Konzepte“ (Maas et al., 1999) entstandene Verfahren, eigene Entwicklungen für eingeschränkte Einzelwortanalysen (Thelen, 1998) und vorläufige Teile der hier vorgestellten Verfahren verwendet und anhand der gewonnen Ergebnisse und erkannten Probleme und Fehler verbessert.

Für das Projekt „Entwicklungsverläufe im Lesen- und Schreibenlernen in Abhängigkeit verschiedener didaktischer Konzepte (ELSAK). Eine empirische Längsschnittuntersuchung in Klasse 1-4“ von Prof. Dr. Swantje Weinhold an der Universität Lüneburg (s. Weinhold, 2006) wird das in Kapitel 7 beschriebene Verfahren mit projektspezifischen Erweiterungen (s. Thelen, 2006) eingesetzt:

> „Untersucht werden drei verschiedene didaktisch-methodische Konzepte zum Lesen- und Schreibenlernen in der Grundschule: die Arbeit mit Fibeln, die Methode „Lesen durch Schreiben“ und die „Silbenanalytische Methode“, die bisher noch gar nicht empirisch erforscht wurde. Die Konzepte unterscheiden sich in ihrer linguistischen Fundierung ebenso wie in ihrem Grad an Strukturiert- bzw. Offenheit. Die Lese- und Schreibdaten, die mithilfe standardisierter und nicht standardisierter Verfahren über zwei bzw. vier Schuljahre hinweg in 13 Klassen erhoben werden (n=156), werden mit einem speziell für das Projekt entwickelten Computerprogramm Konzept-kontrastiv analysiert. Das Programm geht von einer phonologischen Fundierung der deutschen Orthographie aus

	A	B	C	D	E	F	G
1	Code	Pflanze	Rüssel	Hähnchen	Wind#mühle	Gürtel	Trecker
2	A0123.16aW	Pflanze	Rüssel	Hähnchen	Wintmüle	Gürtel	Träker
3	A0223.16aW	%	%	%	%	%	%
4	A0323.16aW	Pflanze	Rüssel	Hähnchen	Windmühle	Gürtel	Träcker
5	A0423.16aW	Pflanze	Rüssel	Hänchen	Windmüle	Gürtel	Trecker
6	A0523.16aW	Pflanze	Rüssel	Hähnchen	Windmühle	Gürtel	Trecker
7	A0623.16aW	%	%	%	%	%	%
8	A0713.16aW	Pflanze	Rüssel	Hünchen [Hühnchen]	Windmühle	Gürtel	Träcker
9	A0813.16aW	Pflanze	Rüssel	Hähnchen	Windmüle	Gurtel	Trecker
10	A0913.16aW	Pflanze	Rüssel	Hänchen	Windmühle	Gürtel	Treker
11	A1013.16aW	Pflanze	Rüssel	Hänchen	Mühle [Müh	Gürtel	Träcker
12	A1113.16aW	%	%	%	%	%	%
13	A1213.16aW	Pflanze	Rüssel	Hühnchen [Hühnchen]	Mühle [Müh	Gürtel	Trecker
14	B0123.16aW	Pflanze	Rüssel	Hänchen	Windbühle	Gürtel	Träcker

Abbildung 9.5.: Online-Analyse: Datenerfassung mit EXCEL

> und wird durch standardisierte Analyseverfahren ergänzt, denen ein Phonem-Graphem-Korrespondenz Konzept zugrunde liegt. Zur Kontrolle der beiden zentralen Variablen Unterricht und Migrationshintergrund werden Unterrichtsbeobachtungen durchgeführt, Arbeitsblätter analysiert und regelmäßige Lehrerbefragungen durchgeführt.“ (Weinhold, 2009)

Im Zentrum der projektspezifischen Erweiterungen stand die Entwicklung einer webbasierten Oberfläche (s. Abbildung 9.4). So konnten Korpuserfassung und Analyseläufe von den Projektmitarbeitern eigenständig durchgeführt werden, ohne dass lokale (und fehleranfällige) Installationen der Analysesoftware notwendig waren. Die Umsetzung von Erweiterungswünschen und Behebung von Fehlern in den Analysen konnte somit ebenfalls direkt in der Serverinstallation erfolgen und hat sowohl Verbesserung der Software als auch Projektablauf erheblich beschleunigt.

Die Datenerhebung im Projekt bestand in mehreren Erhebungswellen, bei denen die Schreiber jeweils aufgefordert wurden, eine Menge von Wortformen, die durch Bilder symbolisiert wurden, isoliert niederzuschreiben. In späteren Erhebungen kamen auch einfache, kurze Sätze hinzu. Die Eingabe der erhobenen Daten geschieht in EXCEL-Tabellen, in denen neben der vereinbarten Codierung beliebige Zusatzangaben gemacht werden können und die damit auch für andere Arbeiten innerhalb des Projektes genutzt werden können. Abbildung 9.5 zeigt eine solche EXCEL-Tabelle. In der ersten Zeile stehen die Vorgaben, d.h. welche Wörter geschrieben werden sollten. Damit die automatische Analyse zuverlässig arbeitet, müssen Komposita an geeigneter Stelle ein # enthalten - das muss aber pro Wort nur einmal geschehen (s. Wind#mühle).

In der linke Spalte steht ein Code, der für das Lüneburger Projekt mehrere Informationen zusammenfasst: Klasse, männlich oder weiblich, Schülernummer, Erhebungsnummer. Diese Codes können frei gewählt werden. Wenn sie in die Strukturierung der Online-Untersuchung mit einfließen sollen, muss die Auswertungssoftware entsprechend etwas angepasst werden. Wichtig ist jedenfalls: Das, was nachher in der Auswertung differenziert werden soll, muss auch hier differenziert abgelegt werden. Jede Spalte entspricht damit den Schreibungen eines Schülers in einer Erhebung. % steht für: Schüler war krank bzw. hat nicht teilgenommen. Falls ein Proband zwar an der Erhebung teilgenommen hat, aber bei einem einzelnen Wort nichts geschrieben hat, steht `/`, ansonsten die jeweilige

! Daten analysieren

Gruppe A, alle Erhebungen

Zusammenfassung:

Schreibungen berücksichtigt: 72
Fehlende Schreibungen: 0
Fehlend wg. Krankheit: 24
Diffuse Schreibungen: 0

Kategorie	max	repräsentiert	plausibel	korrekt
Silbenkerne	171	170 (99%)	168 (98%)	164 (95%)
S'	90	90 (100%)	90 (100%)	86 (95%)
S' (fester Anschluss)	35	35 (100%)	35 (100%)	35 (100%)
S' (loser Anschluss)	55	55 (100%)	55 (100%)	51 (92%)
Monophthong	36	36 (100%)	36 (100%)	32 (88%)
Doppelvokal	0	/	/	/
ie	9	9 (100%)	9 (100%)	6 (66%)
Diphthong	9	9 (100%)	9 (100%)	9 (100%)
Fallender Diphthong	10	10 (100%)	10 (100%)	10 (100%)
S°	62	61 (98%)	60 (96%)	60 (96%)
Schwa	18	17 (94%)	17 (94%)	17 (94%)
<er>	8	8 (100%)	7 (87%)	7 (87%)
Silbischer Sonorant	36	36 (100%)	36 (100%)	36 (100%)
S	19	19 (100%)	18 (94%)	18 (94%)
Anfangsränder	180	180 (100%)	177 (98%)	166 (92%)
einfach	144	144 (100%)	141 (97%)	130 (90%)
komplex	36	36 (100%)	36 (100%)	36 (100%)
Endränder	118	118 (100%)	117 (99%)	106 (89%)
einfach	100	100 (100%)	99 (99%)	89 (89%)
komplex	18	18 (100%)	18 (100%)	17 (94%)
Phonologische Markierungen	35	35 (100%)	32 (91%)	32 (91%)
Schärfung	17	17 (100%)	15 (88%)	15 (88%)
Dehnung	9	9 (100%)	9 (100%)	9 (100%)
Silbentrennendes <h>	9	9 (100%)	8 (88%)	8 (88%)
Konstantschreibung	18	18 (100%)	17 (94%)	17 (94%)
Auslautverhärtung	18	18 (100%)	17 (94%)	17 (94%)
Schärfung	0	/	/	/
Silbentrennendes <h>	0	/	/	/
Einfügungen	1	0 (0%)	0 (0%)	0 (0%)
Fugen	0	/	/	/

Download der Analyseergebnisse als EXCEL-Datei (mit rechter Maustaste anklicken -> Ziel speichern unter...)

Abbildung 9.6.: Online-Analyse: Analyseansicht

Schreibung. Falls ein anderes als das angegebene Wort verschriftet werden sollte, ist das in eckigen Klammern angegeben (s. Hünchen [Hühnchen] für Hähnchen in Zelle D:13).

Die EXCEL-Dateien können über die Weboberfläche der Analyseumgebung hochgeladen und wie in Abbildung 9.4 gezeigt strukturiert betrachtet, unterschiedlich sortiert und durchsucht werden. Jedes Lupensymbol führt zu einer Analyse der jeweils untergeordneten Datenmenge. In Abbildung 9.4 führt ein Klick auf die Lupe rechts neben der Überschrift „Gruppe A“ zu Abbildung 9.6, einer Analyse aller hochgeladenen Schreibungen für Gruppe A. Die Ergebnisse werden online angezeigt, können aber auch in einer EXCEL-kompatiblen Form heruntergeladen werden, um anschließend weiter statistisch analysiert und mit anderen Ergebnissen zusammengeführt werden zu können. Je nach Stabilität der projektspezifisch vorgenommenen Änderungen kommen noch eine Reihe von Debug-Ausgaben hinzu, die es den Projektbeteiligten erlauben, die Qualität der Analyse zu beurteilen und nach Ursachen für fehlerhafte Analyse zu suchen. So kommt es z.B. bei der Annotation der Zielschreibungen wie in Kapitel 7.2.7 ausgeführt immer wieder zu Fällen, die nicht vollautomatisch entschieden werden können. Hier sind dann zusätzliche Markierungen in der EXCEL-Datei vorzunehmen oder das Ausnahmelexikon entsprechend zu erweitern.

Die Online-Analyse-Umgebung ermöglich es Forschenden, ihre Daten in einem leicht zu vermittelnden Format (EXCEL mit einigen Codierungskonventionen) zu erfassen bzw. umzuwandeln. Anschließend können sie ohne technische Kenntnisse der beteiligten Unterprogramme und ohne die Notwendigkeit, Software lokal auf ihrem Rechner zu installieren, die hochgeladenen Daten nach selbst gewählten Strukturierungs- und Zusammenfassungskriterien kontrollieren und zur Analyse bringen. Die Analyseergebnisse können unmittelbar kontrolliert und ggf. anhand von Detailinformationen über den Herleitungsprozess nachvollzogen werden. Gleichzeitig können sie in einem für die weitere Untersuchung geeigneten Standardformat (EXCEL, csv) heruntergeladen werden. In dem Lüneburger Projekt konnten auf diese Weise große Datenmengen effizient und transparent analysiert werden. Der Einsatz der Online-Analyse-Umgebung in weiteren Projekten ist geplant.

9.3. Exploration

Die sichere Beurteilung von Rechtschreibleistungen und die sprachwissenschaftlich fundierte Einordnung orthographischer Phänomene des Deutschen sind Basiskompetenzen für alle Lehrkräfte, die Rechtschreibkenntnisse vermitteln und bewerten müssen. In der deutschdidaktischen und orthographietheoretischen Forschung hat die Untersuchung empirischer Daten in den letzten Jahren erheblich an Bedeutung gewonnen, für die fachwissenschaftliche Ausbildung im Fach Deutsch ist daher ebenfalls eine stärkere empirische Orientierung notwendig. Darüber hinaus fördert die Beschäftigung mit empirisch erhobenen Rechtschreibleistungen als Vorbereitung auf den Beruf praktische Kompetenzen im Umgang mit Fehlern und Leistungen.

Auf Basis der hier vorgestellten Arbeiten ist eine internetbasierte Experimentierumgebung „Ortholab“ entworfen und in Grundzügen implementiert worden, mit der als E-Learning-Anwendung Wissen und Fertigkeiten über

- die Analyse von Rechtschreibleistungen und
- die Untersuchung und Veranschaulichung orthographischer Phänomene des Deutschen

Abbildung 9.7.: Einstiegsseite der Experimentierumgebung Ortholab

aufgebaut und eingeführt werden könen.

Lehramtsstudierende und Teilnehmer von Weiterbildungsprogrammen sollten in Blended-Learning-Szenarien in Form von Übungsaufgaben und Arbeitsaufträgen Zugriff erhalten auf

1. empirische Daten, d.h. Sammlungen von Rechtschreibleistungen/-fehlern unterschiedlicher Schülergruppen und
2. linguistisch aufbereitete Wortlisten und typische Beispiele deutscher Orthographie, sowie
3. einführende und vertiefende Materialien, Anleitungen und Beispiele.

Die Umgebung sollte verschiedene Werkzeuge bereitstellen, mit denen sich die Daten strukturiert durchsuchen, gruppieren, gegenüberstellen, analysieren und weiterverarbeiten lassen. Such- und Arbeitsergebnisse können in persönlichen oder kursbezogenen Bereichen gespeichert, dokumentiert und ausgetauscht und somit für verschiedenste Arbeitsaufträge und Projekte verwendet werden.

Ein grundlegende Implementierung der Umgebung ist mithilfe der in Kapitel 5.8 beschriebenen Verarbeitungssoftware umgesetzt worden (s. Abbildung 9.7). Die Korpusverarbeitungsroutinen sind mit zusätzlichen Ausgabetemplates für SQL-Kommandos erweitert worden, so dass die Ausgaben der Korpusverarbeitung direkt in eine SQL-Datenbank exportiert werden können. Die Experimentierumegbung greift über SQL-Abfragen auf die so entstandene Datenbasis zu und stellt Werkzeuge zur Untersuchung und Exploration der Korpora bereit.

Die skizzierte Lernumgebung zielt nicht darauf ab, vorstrukturiertes Wissen über die deutsche Rechtschreibung und typische Rechtschreibleistungen und -fehler zu vermitteln. Vielmehr versteht

es sich als Unterstützung für reguläre Präsenzlehrveranstaltungen, die Übungsaufgaben, Arbeitsaufträge sowie Projektarbeiten definieren. Die Umgebung ist offen gestaltet, so dass je nach Szenario sowohl individuelle Zugänge als auch interaktive Nutzungen möglich sind.

Im Vordergrund des didaktischen Konzeptes stehen neben den aufbereiten Daten die Such- und Darstellungsroutinen sowie unterschiedliche Analysewerkzeuge, die frei, d.h. unabhängig von konkreten Aufgabenstellungen, genutzt werden können. Lernende können Arbeitsergebnisse speichern und später weiter analysieren oder mit anderen austauschen.

Damit unterstützt die Umgebung auch offenere Arbeitsformen explorativen Lernens. Ergänzend werden in einer „Bibliothek“ auch Materialien bereitgehalten, die Basiswissen wiederholen, unterschiedliche Analyse- und Beurteilungsansätze beschreiben und Hinweise für die Arbeit mit empirischen Daten in der Deutschdidaktik geben. In diesem Umfeld werden auch Rundgänge durch die Umgebung angelegt, die kleinschrittig den Umgang mit den Werkzeugen und die korrekte Interpretation der Ergebnisse vermitteln. In die Materialien sind abschnittsweise Tests zur Selbstkontrolle eingebunden.

Die Umgebung stellt sich den Lernenden als Werkzeug dar, mit dem sie vor dem Hintergrund ihrer Lehrveranstaltungen vor Ort effizient und flexibel auf große Datenmengen zugreifen können. Lehrende haben die Möglichkeit, die Umgebung kursspezifisch zu konfektionieren. Die Experimentierumgebung ermöglicht grundsätzlich zwei Nutzungsformen:

- freie Nutzung der allgemein zugänglichen Daten und Werkzeuge
- kursspezifisch gestaltete Nutzung mit ausgewählten, zusätzlichen oder abweichenden Inhalten, Zugriff auf geschützte Datensammlungen und spezifisch angepasste Werkzeuge.

Die erste Form bietet sich für den punktuellen Einsatz in Lehrveranstaltungen an, die zweite Form kann bei speziell daraus ausgerichteten Veranstaltungen zum Einsatz kommen. Typischerweise werden die Studierenden in einer Reihe von Aufgaben und Arbeitsaufträgen schrittweise an die Experimentierumgebung, ihre Daten und Werkzeuge herangeführt und dabei tutoriell betreut. Eine exemplarische Abfolge mit einem Workload von ca. 30 Stunden (stark abhängig von Vorerfahrungen, Art der begleiteten Lehrveranstaltung etc.) kann wie folgt aussehen, Arbeitsergebnisse sind dabei jeweils im persönlichen Bereich zu speichern und zu kommentieren:

1. Heranführung an die Umgebung und einfache Suchaufträge
    - Machen Sie sich mithilfe der „Rundgang“-Funktion mit der Umgebung vertraut.
    - Durchsuchen Sie unterschiedliche Wortlisten nach deutschen Wörtern, die auf die Buchstabenfolge „-ere“ enden. Klassifizieren Sie die Ergebnisse und finden Sie heraus, welche Eigenschaften von Wortlisten die gefundenen Ergebnisse erheblich beeinflussen (s. Abbildung 9.8).
    - Finden Sie heraus, welche 10 Wortformen in den vorhandenen Corpora am häufigsten vorkommen (s. Abbildung 9.9). Wie verändert sich die Liste, wenn Sie nur Schreibungen von Zweitklässlern berücksichtigen?
    - Finden Sie alle Fehlschreibungen der Wortform <Hund> (s. Abbildung 9.10).
2. Einbeziehung linguistischer Kategorien

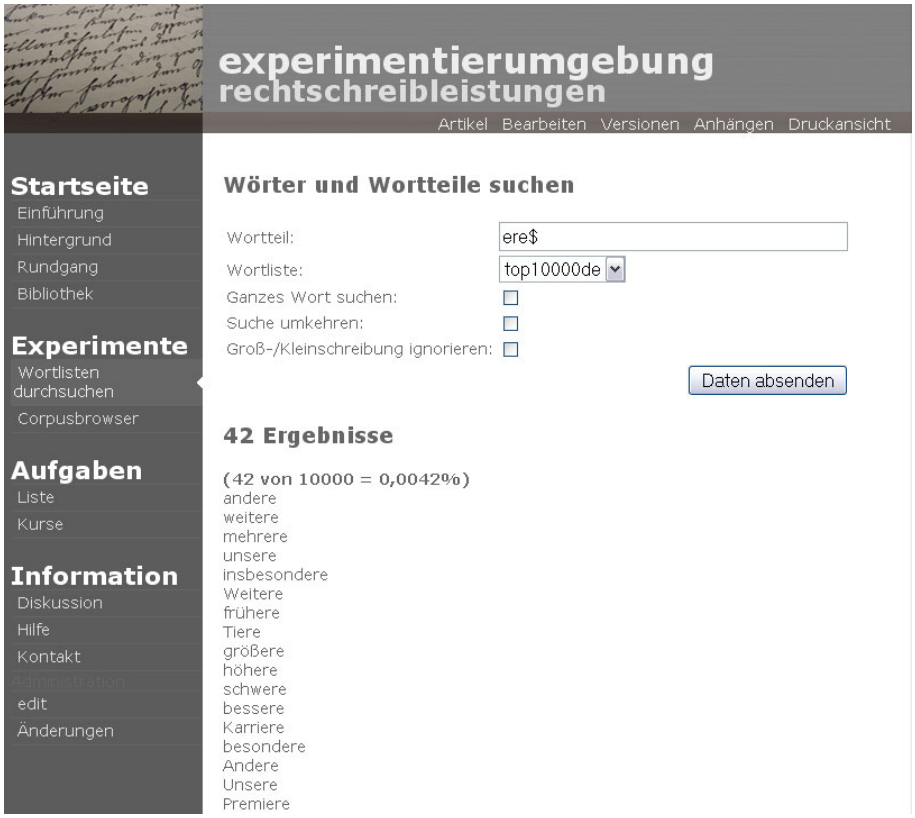

Abbildung 9.8.: Suchergebnis: Wörter, die auf die Zeichfolge ere enden

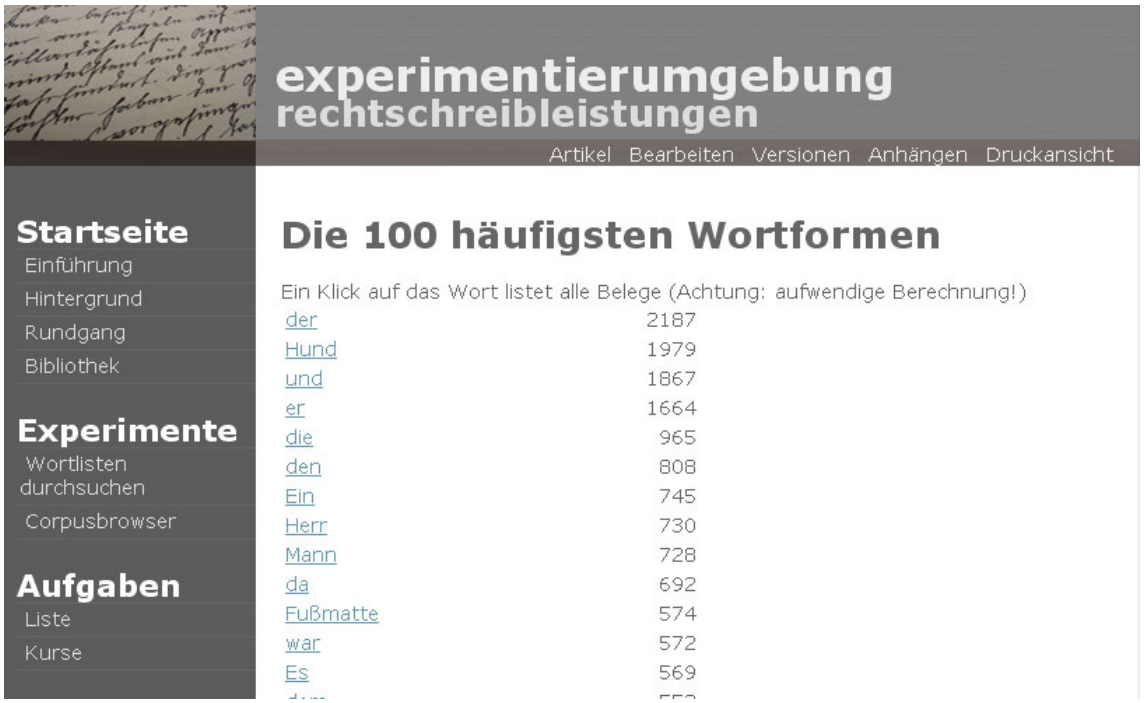

Abbildung 9.9.: Auflistung der häufigsten Wortformen im Korpus

Abbildung 9.10.: Korpussuche nach allen Fehlschreibungen für <Hund>

- Erstellen Sie eine Tabelle mit Häufigkeiten von Betonungsmustern zwei- und dreisilbiger Wortformen. Verschaffen Sie sich einen Überblick über die jeweils selteneren Muster und versuchen Sie eine Grobklassifizierung.
- Listen Sie Fälle von fälschlicher <ie>-Schreibung in unbetonten Silben auf.
- Finden Sie drei Wortformen, deren Schreibung bei nord- und süddeutschen Schreibern deutliche Unterschiede aufweist.

3. Analyse von Leistungen und Fehlern

- Belegen Sie, dass die Häufigkeit von fälschlich klein geschriebenen Abstrakta vom 2. zum 4. Schuljahr abnimmt. Vergleichen Sie die Abnahme mit der Abnahme aller Groß- und Kleinschreibungsfehler in diesem Zeitraum.

4. Projektartige Arbeitsaufträge

- Überprüfen Sie einige Faustregeln zu Subregularitäten der Dehnungsschreibung, die Sie in Sprachbüchern finden.

Bei der Bearbeitung der Aufgaben haben die Studierenden Zugriff auf ein Diskussionsforum und andere E-Learning-Kommunikationswerkzeuge, um Fragen untereinander oder mithilfe von Tutoren und Dozenten zu klären. Diese Werkzeuge können in einfacher Form entweder in der Umgebung selbst, oder in einer einbettenden Lernplattform wie Stud.IP (s. Schulze und Meeh, 2005) genutzt werden.

Für die erfolgreiche Nutzung der Lernumgebung sind vier Typen von Materialien in aufbereiter Form notwendig:

1. Empirische Daten (Korpora)

Sowohl orthographisch korrekte Daten, die für Wortlisten und die Untersuchung grundlegender orthographischer Phänomene des Deutschen verwendet werden, als auch Korpora von Lernertexten, die für die Untersuchung von Rechtschreibleistungen und -fehlern genutzt werden, liegen aus verschiedenen Forschungsprojekten in ausreichender Zahl zugänglich vor. Umfangreiche Neuerhebungen für die skizzierte Umbegung sind nicht notwendig. Die Daten müssen allerdings mit corpuslinguistischen Verfahren einheitlich aufbereitet werden, das umfasst die Annotation mit Metadaten und die teilautomatisierte Anreicherung mit linguistischen Informationen sowie die maschinenlesbare Codierung in dem in Kapitel 5 vorgestellten Format.

2. Erläuternde Begleitmaterialien

Hier könnte vor allem auf existierende Vorlesungsskripten, wissenschaftliche Artikel, besonders geeignete studentische Arbeiten und Projektbereichte zurückgegriffen werden, die teilautomatisiert in das für die Plattform notwendige Format transferiert werden. Die Verwendung von Audio- und Videomaterial ist punktuell ebenfalls vorzusehen.

3. Aufgabensammlungen

Neben Selbstüberprüfungsfragen, die in den erläuternden Begleitmaterialien Anwendung finden, müssten kursspezifische Aufgaben für die Arbeit mit der Experimentierumgebung entwickelt und erprobt werden, die allen Anwendern in Form eines strukturierten Aufgabenpools zur Verfügung gestellt werden.

4. Projektdokumentationen

Insbesondere umfangreichere Arbeitsaufträge führen zu aufwendigen Projekt- und Experimentierdokumentationen, die als Trainings- und Anschauungsmaterial für die Arbeit mit der Umgebung oder als Ausgangspunkt für neue Fragestellungen verwendet werden. Projektergebnisse in diesem Sinne sind auch speziell zusammengestellte Teil- und Querschnittskorpora, die Einzelphänomene in den Vordergrund stellen und als Material für weitere Aufgaben Verwendung finden.

Technische Plattform für die Experimentierumgebung ist eine kollaborative Textproduktionsplattform, die mit spezifischen Erweiterungen (Plugins) für die Arbeit mit Rechtschreibcorpora und Wortlisten erweitert wird. Die Wahl ist hier auf PmWiki[2] gefallen, da umfangreiche positive Vorerfahrungen mit dieser Plattform als Autoren- und Arbeitsumgebung, dem Einsatz von Wikis in E-Learning-Szenarien, der Erweiterung von PmWiki mit Plugins und der Verknüpfung von PmWiki-Inhalten mit anderen eingesetzten E-Learning-Werkzeugen vorliegen (Thelen, 2008). PmWiki-Inhalte (so genannte Wiki-Fields) sind über die E-Learning-Content-Schnittstelle mit dem Lernmanagement-System Stud.IP kombinierbar, außerdem ist die separate Nutzung über andere Authenfizierungsmechanismen (z.B. LDAP) möglich.

Die Werkzeuge, Schemata und Algorithmen, die für Aufbereitung, linguistische Voranalyse der Daten und Online-Analyse von Wortformen und Rechtschreibleistungen genutzt werden, liegen in einer allgemeinen Grundform wie in dieser Arbeit beschrieben vor. Sie sind auf vermittlungsspezifische Fragestellungen hin anzupassen und mit weiteren Ausgabe- und Darstellungsformen zu ergänzen.

[2] http://www.pmwiki.org

10. Fazit

Ziel der vorliegenden Arbeit war es, Möglichkeiten und Grenzen der automatisierten Analyse orthographischer Leistungen von Schreibanfängern zu erforschen und zu bewerten. Ausgangspunkt der Betrachtungen war ein linguistisch orientiertes Modell der deutschen Orthographie, für das zudem erprobte didaktische Umsetzungen für den Anfangsunterricht im Lesen und Schreiben existieren. Dieses Modell diente, vor allem auf den Bereich der Wortschreibung fokussiert, als Beurteilungskriterium etablierter standardisierter und nicht-standardisierter diagnostischer Rechtschreibtests.

Bei den betrachteten Tests wurden Defizite auf mehreren Ebenen festgestellt. Zum einen operieren sie mit linguistisch nicht sauber begründbaren Begriffen und Kategorieabgrenzungen. Zum anderen und zum Teil daraus folgend erlauben sie nur wenige diagnostischen Schlüsse, die zielgerichtetes didaktisches Handeln begründen können. Aus diesen Gründen wurde ein eigenes, vor allem auf phonologische Aspekte der Wortschreibung abzielendes Auswertungsschema definiert und anhand eines ausführlichen Beispiels mit den etablierten Tests verglichen. Ergebnis dieses Vergleichs war, dass das eigene Verfahren zwar sehr viel aufwendiger in der Handhabung ist, dafür aber eine weit größere Differenzierung aufweisen kann. Diese ist notwendig, um Leistungen und Defizite zu identifizieren, denen dann durch wohlbegründete Regeln und entsprechende didaktische Maßnahmen begegnet werden kann.

Auf dieser linguistisch gesicherten und für den Praxiseinsatz vielversprechenden Grundlage sollten unterschiedliche algorithmische Verfahren entworfen, umgesetzt und erprobt werden, die die notwendigen Analyseschritte automatisieren oder zumindest teilautomatisieren können. Für die Erprobung der Verfahren wurde ein Testkorpus von ca. 750 Bildergeschichtenverschriftungen aus verschiedenen Regionen Deutschlands erhoben, das „Osnabrücker Bildergeschichtenkorpus“. Die Texte, die vorwiegend von Zweitklässlern verfasst wurden, umfassen insgesamt ca. 45.000 Wortformen (Tokens), von denen ca. 11.500 orthographisch fehlerhaft sind. Auf Type-Ebene wurden 6.431 unterschiedliche Wortschreibungen erfasst, von denen 4.149 auf unterschiedliche Fehlschreibungen entfielen. Um dieses große Korpus authentischer Rechtschreibleistungen von Schreibanfängern umfassend nutzbar zu machen, wurde ein umfangreiches XML-Repräsentationsschema entworfen, das neben den erfassten Texten auch unterschiedliche Ebenen von Meta-Informationen abbildet. Es war niemals Ziel der Korpuserhebung, daraus statistisch belastbare Aussagen über Rechtschreibkompetenzen in Abhängigkeit miterfasster Parameter wie Herkunft, Muttersprache, Dialektregion oder Art des Schreiblehrganges zu gewinnen. Vielmehr stellt die Definition eines Datenformats für solche Informationen zusammen mit den darauf operierenden Analyserverfahren eine eigenständige Leistung dieser Arbeit dar. Innerhalb der Arbeit wurden nur exemplarische Auswertungen vorgenommen, um die Bandbreite der denkbaren Analysen mit methodisch sauber erhobenen Korpora aufzuzeigen. Das Korpus wurde im Rahmen dieser Arbeit vor allem zur Kontrolle der Algorithmen und zur Aufdeckung von Fehlern genutzt. Durch die Größe und heterogene Zusammensetzung des Korpus sollte zudem sichergestellt werden, dass möglichst viele typische Schreibungen von Schreibanfängern berücksichtigt werden können, auch wenn dadurch keine repräsentative Häufigkeitsverteilung angenommen werden kann.

Die entworfenen Analyseverfahren decken drei Bereiche ab: Die Analyse von Wortschreibungen ohne Kenntnis der korrekten Schreibungen, den Vergleich korrekter und beobachteter Wortschreibungen und die Analyse von Groß- und Kleinschreibungsleistungen.

Verfahren, die keine Kenntnis über die korrekte Schreibung haben, wurden auf Basis von Vollformenlexika, manuell erstellten Regeln und automatisiert gewonnen Mustermengen entwickelt. Dabei stellte sich heraus, dass solche Verfahren keine qualitativen Analysen leisten, sondern höchstens Hinweise auf potenziell fehlerhafte Formen geben können. Alle Verfahren wurden anhand des Osnabrücker Bildergeschichtenkorpus getestet und mithilfe von Leistungsmaßen aus dem Information Retrieval verglichen. Das beste Verfahren, das musterbasiert mithilfe eines großen Standardkorpus deutscher Wortformen trainiert wurde, konnte dabei ein F-Maß von 0.77 erreichen. Konkret bedeutet dies, dass das Verfahren 12,6% der aufgetretenen Fehler nicht erkannt hat und zusätzlich 38,81% der korrekten Schreibungen fälschlicherweise als fehlerhaft markiert hat. Damit ist – erwartungsgemäß – festzustellen, dass automatisierte Verfahren zur Fehlererkennung keine allzu verlässlichen Aussagen liefern und höchstens in Kombination verschiedener Verfahren zur weiter nutzbaren Ergebnissen führen.

Für den Abgleich zwischen beobachteter Schreibung und bekannter korrekter Schreibung wurde ein Verfahren entwickelt, dass den Anforderungen des linguistisch und didaktisch begründeten Analyseverfahren voll genügt. Das Verfahren versucht zunächst, die korrekten Schreibungen linguistisch zu analysieren und mit verschiedenen phonologischen und morphologischen Zusatzinformationen zu versehen. Dieser Automatismus kann, wie gezeigt, nicht in allen Fällen fehlerfrei arbeiten, so dass ein Ausnahmelexikon bzw. manuell hinzugefügte morphologische Hinweismarkierungen notwendig sind, um korrekte Analysen zu gewährleisten. Die so angereicherten korrekten Formen werden mittels etablierter String-Matching-Algorithmen mit den beobachteten Schreibungen vergleichen. Aus den gelungenen bzw. fehlgeschlagenen Abgleichversuchen werden die Informationen gewonnen, die zum Füllen der Analysetabellen notwendig sind. Das Verfahren hat sich insgesamt als robust und leistungsfähig erwiesen und liegt als Python-Implementation eines allgemeinen Korpusverarbeitungs- und Analyseframeworks vor.

Über den Bereich der Wortschreibung geht die Analyse von Groß- und Kleinschreibungsleistungen hinaus. Es wurde ein Ansatz vorgestellt, der auf bekannten korrekten und manuell merkmalsannotierten Texten arbeitet. Die vorgestellten Merkmale sind sehr unterschiedlicher Natur und sind geeignet, unterschiedlichste Hypothesen über die Gründe zu unterstützen, warum ein Schüler Groß- und Kleinschreibungen vorgenommen hat. So sind sowohl linguistische Kategorien wie expandierbare Kerne von Nominalphrasen, als auch Kategorien des amtlichen Regelwerks wie Wortarten, Substantivierung und Desubstantivierung und schließlich „naive“ didaktisch begründete Kategorien in die Beispielanalysen aufgenommen worden. Über eine Analyse der potenziellen Relevanz einzelner Kategorien für die beobachteten Schreibungen werden letztendlich individuelle Strategiehypothesen gebildet. Grundlegend ist hier das Konzept des Informationsgewinns aus dem Maschinellen Lernen. Die Strategiehypothesen werden als Entscheidungsbäume dargestellt, die diejenigen Kategorien beinhalten, die – nach informationstheoretischer Analyse – für den Schreiber besonders bedeutsam waren.

Die in dieser Arbeit entworfenen und implementierten Verfahren liefern keine unmittelbaren Erkenntnisse über Wissen und kognitive Zustände der Schreiber. Vielmehr liegt ihnen ein Modell zur Validität von Analyseergebnissen zugrunde, das sich wesentlich auf die Gültigkeit zweier Übertragungen beruft: Zum Einen die Übertragung des Gegenstandsbereichs „Orthographie“ in ein linguistisch fundiertes Modell der deutschen Orthographie und daraus begründete Analysekategorien.

Und zum Anderen die Übertragung der beobachteten Leistungen in ein Wissensmodell des Lerners, das wiederum linguistisch fundiert ist und mithilfe linguistisch begründeter didaktischer Modelle in unterrichtliches Handeln umgesetzt werden kann.

Die vorgestellten Verfahren konnten so weit entwickelt und umgesetzt werden, dass sie für den tatsächlichen Praxiseinsatz geeignet sind. Für die in der Zielsetzung definierten Einsatzszenarien – die Nutzung in Lehr-/Lernsoftware und die Verwendung als forschungsunterstützendes Werkzeug – konnten erfreulicherweise Projekte gefunden werden, die bereit waren, die Verfahren produktiv einzusetzen. Das erfolgreiche Lernportal für funktionale Analphabeten „www.ich-will-schreiben-lernen.de“ verwendet ein vereinfachtes Wortschreibungsanalyseverfahren, um Übungsautoren zu unterstützen und den Lernern individuell zugeschnittene Aufgabenbündel zu präsentieren. Für das Projekt „Entwicklungsverläufe im Lesen- und Schreibenlernen in Abhängigkeit verschiedener didaktischer Konzepte (ELSAK). Eine empirische Längsschnittuntersuchung in Klasse 1-4“ wird das in dieser Arbeit entwickelte Analyseschema mitsamt der Auswertungsalgorithmen verwendet, um große Mengen an Wortschreibungen effizient zu analysieren. Für das Projekt wurde eine webbasierte Oberfläche für die Analyseverfahren entwickelt, die es Proektmitarbeitern erlaubt, ohne technische Detailkenntnisse Daten zu erfassen und nach unterschiedlichsten Gruppierungskriterien zu analysieren. Die Automatisierung der Auswertung hat es dem Projekt überdies ermöglicht, unterschiedliche Varianten der Auswertung und eine fortlaufende, iterative Verfeinerung der Analysekonzepte umzusetzen. Bei üblicher manueller Analyse der Daten hätte der notwendige Zeit- und Personalaufwand ein solches Vorgehen unmöglich gemacht. Der Online-Zugang zu Korpora und Analyseverfahren eröffnet zudem noch ein drittes Anwendungsfeld. Die in dieser Arbeit entwickelten Verarbeitungs- und Analyseverfahren ermöglichen eine hochschuldidaktisch ausgerichtete E-Learning-Plattform für die Deutschlehrer-Ausbildung, die den Lernern einen explorativen Zugang zu authentischen Daten und unterstützenden Auswertungsalgorithmen gibt. Diese als Werkzeuge begriffenen Möglichkeiten lassen eine Lernumgebung realisierbar werden, die unterschiedlich ausgerichtete Aufgabenstellungen als Ergänzung, Vertiefung und praktisches Anwendungsfeld für im Studium erworbenes theoretisches Wissen beinhalten. Eine solche Lernumgebung ist im Rahmen dieser Arbeit grundlegend technisch entwickelt und mit dem Osnabrücker Bildergeschichtenkorpus als Beispielmaterial befüllt worden. Die notwendigen Projektmittel für die vollständige Ausgestaltung der Umgebung und ihre Implementierung in der Deutschlehrer-Ausbildung werden derzeit beantragt.

Als Ergebnis dieser Arbeit ist deutlich geworden, dass verlässliche und aussagekräftige automatisierte Analysen orthographischer Leistungen nur für wohldefinierte Teilbereiche der Orthographie und wohldefinierte Datenmengen entwickelt werden können. Ein vollständig autonom agierendes Analyseverfahren für beliebige freie Texte ist nach derzeitiger linguistischer Modellierung der deutschen Orthographie und den derzeitigen Möglichkeiten computerlinguistischer Verfahren unmöglich. Dennoch ist es möglich und – wie gezeigt – gewinnbringend, gut beherrschbare Teilbereiche zu identifizieren und praktisch nutzbare Analyseschemata- und -verfahren für sie zu definieren. Die in dieser Arbeit beschrittenen Wege bieten viel Raum für Verbesserungen und Erweiterungen. Neben der Ausdehnung auf weitere Teilbereiche der Orthographie, wie der Getrennt- und Zusammenschreibung und der Zeichensetzung wären vor allem eine Verfeinerung und weitere empirische Absicherung der Analysekategorien wünschenswert. Auch die automatisierte Behandlung speziellerer Fragestellungen, wie dem Einfluss dialektaler und fremdsprachlicher Einflüsse auf Rechtschreibleistungen konnte im Rahmen dieser Arbeit nicht angegangen werden.

Es besteht aber die Hoffnung, dass die konzeptionellen Leistungen dieser Arbeit – der Entwurf eines Repräsentationsformates für Korpora grundschulischer Rechtschreibleistungen und der Entwurf

linguistisch fundierter Analyseschemata für Wortschreibungen und Groß- und Kleinschreibungsleistungen – der weiteren Forschung dienen. Alle weiteren Beiträge dieser Arbeit, das Osnabrücker Rechtschreibkorpus und die implementierten Analysealgorithmen werden mit der Veröffentlichung der Arbeit unter freien Lizenzen veröffentlicht und stehen damit der wissenschaftlichen Öffentlichkeit auch für eigene Untersuchungen und abgeleitete oder modifizierte Verfahren zur Verfügung.

Literaturverzeichnis

Augst, Gerhard und Dehn, Mechtild (1998): *Rechtschreibung und Rechtschreibunterricht.* Stuttgart: Klett.

Augst, Gerhard und Stock, Elisabeth (1997): "Laut-Buchstaben-Zuordnung". In: *Zur Neuregelung der deutschen Orthographie: Begründung und Kritik*, herausgegeben von Augst, Gerhard, Tübingen: Niemeyer.

Berndt, Elin-Birgit (2002): *Interaktion mit digitalen Rechtschreibhilfen.* Universität Bremen.

Bray, Tim; ean Paoli; Sperberg-McQueen, C.M. und Maler, Eve (Herausgeber) (2000): *Extensible Markup Language (XML) 1.0 (Second Edition).* World Wide Web Consortium.

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf und Langer, Hagen (Herausgeber) (2001): *Computerlinguistik und Sprachtechnologie.* Heidelberg, Berlin: Spektrum Akademischer Verlag.

Dehn, Mechthild (1994): *Zeit für die Schrift.* Bochum: Kamp, vierte Auflage.

Deutscher Volkshochschulverband (2006): *Portal Zweite Chance Online: Awards.* Deutscher Volkshochschulverband.

Duden (1991): *Der Duden. Band 1: Die Rechtschreibung.* Mannheim, Leipzig, Wien, Zürich: Dudenverlag, zwanzigste Auflage.

Eisenberg, Peter (1998): *Grundriss der deutschen Grammatik. Band 1: Das Wort.* Stuttgart: Metzler.

Evert, Stefan und Fitschen, Arne (2001): "Textkorpora". In: Carstensen et al. (2001), S. 369–376.

Fliedner, Gerhard (2001): "Korrekturprogramme". In: Carstensen et al. (2001), S. 411–417.

Frötschl, Bernhard und Lindstrot, Wolf (2001): "Wahrscheinlichkeitstheorie und Hidden-Markov-Modelle". In: Carstensen et al. (2001), S. 107–133.

Grund, Martin; Haug, Gerhard und Naumann, Carl-Ludwig (2003): *Diagnostischer Rechtschreibtest für 5. Klassen.* Beltz Deutsche Schultests.

Herné, Karl-Ludwig und Naumann, Carl Ludwig (2002): *Aachener Förderdiagnostische Rechtschreibfehler-Analyse.* Alfa Zentaurus.

Hofstadter, Douglas (1991): *Metamagicum.* Stuttgart: Klett-Cotta.

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

Klabunde, Ralf (2001): "Automatentheorie und Formale Sprachen". In: Carstensen et al. (2001), S. 59–86.

Kohler, Klaus J. (1995): *Einführung in die Phonetik des Deutschen.* Berlin: Erich Schmidt, zweite Auflage.

Kohonen, Teuvo (1995): *Self-organizing maps.* Springer.

Lobin, Henning (1999): *Informationsmodellierung in XML und SGML.* Berlin, Heidelberg, New York u.a.: Springer.

Maas, Utz (1992): *Grundzüge der Orthographie des Deutschen.* Tübingen: Niemeyer.

Maas, Utz (1999): *Phonologie – Eine Einführung in die funktionale Phonetik des Deutschen.* Opladen: Westdeutscher Verlag.

Maas, Utz (2000): "Materialien zu einem erklärenden Handbuch der Orthographie des Deutschen". Unveröffentlichtes Manuskript, Osnabrück.

Maas, Utz; Gust, Helmar; Albes, Christian; Noack, Christina und TobiasThelen (1999): "Abschlußbericht des Projektes Computerbasierte Modellierung orthographischerProzesse". Universität Osnabrück.

Manning, Christopher und Schuetze, Hinrich (1999): *Foundations of Statistical Natural Language Processing.* Cambridge, London: MIT Press.

May, Peter (1994): "Regeln sind für Mädchen - Jungen brauchen Sensationen !?" In: *Mädchen lernen* anders *lernen Jungen, Geschlechtsspezifische Unterschiedebeim Schriftspracherwerb*, herausgegeben von Richter, S. und Brügelmann, H., Bottinghofen am Bodensee: Libelle, S. 83–89 u. 110–120.

May, Peter (1998): "Strategiebezogene Rechtschreibdiagnose – mit und ohne Test – Analyse vonfreien Schreibungen mit Hilfe der HSP-Kategorien". In: *Schatzkiste Sprache 1. Von den Wegen der Kinder in die Schrift*, herausgegeben von Balhorn, Heiko; Bartnitzky, Horst; Büchner, Inge und Speck-Hamdan, Angelika, Frankfurt/Main, Hamburg: Arbeitskreis Grundschule, Deutsche Gesellschaft für Lesen und Schreiben, S. 279–293.

May, Peter (1999): "HSP - Was ist das?" *lernchancen* 11: S. 38–40.

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

May, Peter und Malitzky, V. (1999): "Erfassung der Rechtschreibkompetenz in der Sekundarstufe mit der HamburgerSchreibprobe". In: *Konkrete Handlungsanleitungen für erfolgreiche Beratungsarbeit mit Schülern,Eltern und Lehrern*, herausgegeben von Lade, E. und Kowalczyk, W., Kissing: WEKA Fachverlag.

Mehlem, Ulrich (2002): "Die Nutzung orthographischer Strukturen des Deutschen in der Verschriftungfamiliensprachlicher Texte durch marokkanische Migrantenkinder". In: *Schrifterwerbskonzepte zwischen Pädagogik und Sprachwissenschaft*, herausgegeben von Röber-Siekmeyer, Christa und Tophinke, Doris, Baltmannsweiler: Schneider Verlag Hohengehren.

Mitchell, Tom M. (1997): *Machine Learning.* Boston, Burr Ridge usw.: McGraw-Hill.

Mitton, Roger (1996): *English spelling and the computer.* New York: Longman.

Naumann, Carl-Ludwig (1989): *Gesprochenes Deutsch und Orthographie.* Frankfurt: Lang.

Naumann, Carl-Ludwig (1990): "Nochmals zu den Prinzipien der deutschen Orthographie". In: *Zu einer Theorie der Orthographie*, herausgegeben von Stetter, Christian, Tübingen: Niemeyer.

Nerius, Dieter (1986): "Zur Bestimmung und Differenzierung der Prinzipien der Orthographie". In: *New trends in graphemics and orthography*, herausgegeben von Augst, Gerhard, Berlin, New York: de Gruyter, S. 11–24.

Nerius, Dieter und Autorenkollektiv (1989): *Deutsche Orthographie.* Leipzig: Bibliographisches Institut, 2. Auflage.

Noack, Christina (2000): *Regularitäten der deutschen Orthographie und ihre Deregulierung.* Dissertation, Universität Osnabrück.

Noack, Christina (2002a): "Mundart und Schrifterwerb am Beispiel des Alemannischen. Aktueller Forschungsstand und Perspektiven." In: Tophinke und Röber-Siekmeyer (2002), S. 222—247.

Noack, Christina (2002b): "Regularities in German Orthography: A Computer-Based Comparison of DifferentApproaches to Sharpening". In: *The Relation of Writing to Spoken Language*, herausgegeben von Neef, Martin; Neijt, Anneke und Sproat, Richard, Tübingen: Niemeyer, Linguistische Arbeiten 460, S. 149–168.

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

Press, Hans-Jürgen (1987): *Der kleine Herr Jakob.* Ravensburg: Ravensburger Verlag.

Primus, Beatrice (2000): "Suprasegmentale Graphematik und Phonologie: Die Dehnungszeichen im Deutschen". *Linguistische Berichte* 181: S. 9–34.

Quinlan, Ross (1986): "Induction of Decision Trees". *Machine Learning* 1: S. 81–106.

Ramers, Karl Heinz (1988): *Vokalquantität und -qualität im Deutschen.* Tübingen: Niemeyer.

Röber-Siekmeyer, Christa (1993): *Die Schriftsprache entdecken.* Weinheim: Beltz.

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

Röber-Siekmeyer, Christa und Pfisterer, Katja (1998): "Silbenorientiertes Arbeiten mit einem leseschwachen Zweitklässler". In: *Schriftspracherwerb*, herausgegeben von Weingarten, Rüdiger und Günther, Hartmut, Baltmannsweiler: Schneider Verlag Hohengehren, S. 36–61.

Rojas, Raúl (1993): *Theorie der Neuronalen Netze.* Springer.

Rowling, Joanne K. (2000): *Harry Potter und der Feuerkelch.* Hamburg: Carlsen. Übersetzung Klaus Fritz.

Saure, Michael; Thelen, Tobias und Trommer, Jochen (1997): "Modellierung orthographischer Prozesse – Abschlußbericht und Ausblick". Institut für Semantische Informationsverarbeitung.

Schulze, Annette und Meeh, Holger (2005): "Erfahrungen mit Stud.IP in der Hochschullehre". In: *Virtuelle Lernumgebungen im Deutschunterricht*, herausgegeben von Möbius, Thomas und Ulrich, Stefan, Schneider Verlag Hohengehren, S. 138–147.

Schuster, Karl (1995): *Einführung in die Fachdidaktik Deutsch.* Baltmannsweiler: Schneider Verlag Hohengehren, 5. Auflage.

Spiekermann, Helmut (2000): *Silbenschnitt in deutschen Dialekten.* Tübingen: Niemeyer.

Sproat, Richard (2000): *Computational theory of writing systems.* Cambridge: Cambridge University Press.

Stephen, Graham (1994): *String Searching Algorithms.* Singapur: World Scientific.

The Unicode Consortium (2003): *The Unicode Standard, Version 4.0.* Reading: Addison-Wesley.

Thelen, Tobias (1998): *Automatische Analyse orthographischer Fehler bei Einzelwortschreibungen.* Magisterarbeit, Universität Osnabrück.

Thelen, Tobias (2002): "Wie passt das Wort BETTEN in das Haus? Grundlagen und Ergebnisse eines Computerprogramms zur Vermittlung der Schärfungsschreibung". In: Tophinke und Röber-Siekmeyer (2002).

Thelen, Tobias (2006): "Praktische Möglichkeiten computergestützter Rechtschreibanalyse". In: *Schriftspracherwerb empirisch. Konzepte, Diagnostik, Entwicklung.*, herausgegeben von Weinhold, Swantje, Schneider Verlag Hohengehren, S. 178–198.

Thelen, Tobias (2008): "Wikis als flexible E-Learning-Werkzeuge". In: *Lernen Organisation Gesellschaft. Das eCampus-Symposium der Osnabrücker Hochschulen. Tagungsband 2008.*, herausgegeben von Knaden, Andreas; Hoppe, Uwe; Bergs, Alexander; Andersson, Robby und Hübner, Ursula, Osnabrück: epOs-media, S. 39–46.

Thomé, Günther (1998): *Orthographieerwerb.* Frankfurt: Lang.

Tophinke, Doris und Röber-Siekmeyer, Christa (2002): *Schärfungsschreibung im Fokus. Zur schriftlichen Repräsentation sprachlicher Strukturen im Spannungsfeld von Sprachwissenschaft und Pädagogik.* Baltmannsweiler: Schneider Verlag Hohengehren.

von Glasersfeld, Ernst (1996): "Welt als Black-Box". In: *Die Natur ist unser Modell von ihr*, herausgegeben von Braitenberg, Valentin und Hosp, Inga, Hamburg: Rowohlt.

Weingarten, Rüdiger (2000): "Orthographisch-grammatisches Wissen". In: *Wissenstransfer zwischen Experten und Laien*, herausgegeben von Wichter, S.; Antos, G. und Schieholz, St., Frankfurt/Main: Lang.

Weinhold, Swantje (2006): "Entwicklungsverläufe im Lesen- und Schreibenlernen in Abhängigkeit verschiedener didaktischer Konzepte". In: *Schriftspracherwerb empirisch. Konzepte, Diagnostik, Entwicklung.*, herausgegeben von Weinhold, Swantje, Schneider Verlag Hohengehren, S. 120–151.

Weinhold, Swantje (2009): "Arbeitsschwerpunkte in Lehre und Forschung". Online: http://www.fb1.uni-lueneburg.de/fb1/deutsch/mitarbeiter/Weinhold/schwerpunkte.php [1.2.2009].

Widmann, Gerhard (2001): *Aufsatz Bildergeschichte. Übungsprogramm mit Lösungen für die 4. bis 6.Klasse.* München: Adolf Hauschka Verlag.

Wiese, Richard (1996): *The Phonology of German.* Oxford, New York: Oxford University Press.

Wittgenstein, Ludwig (1984): "Philosophische Untersuchungen". In: *Werkausgabe, Bd. 1*, Frankfurt a.M.: Suhrkamp.

A. Anhang auf CD-ROM

Auf der beiligenden CD-ROM sind Daten, entwickelte Software und vollständige Analyseläufe enthalten.

`/ergebnisse`	Vollständiger Analyselauf für das Osnabrücker Bildergeschichtenkorpus (PDF)
`/gks`	Analysesoftware (Python), Ausgangsdaten und Ergebnisse (HTML) für die Groß- und Kleinschreibungsanalyse (s. Kapitel 8)
`/ortholab`	Explorative E-Learning-Umgebung für die Deutschlehrerausbildung (s. Kapitel 9.3, benötigt Webserver mit PHP- und Python-Unterstützung sowie mySQL-Datenbank-Server)
`/pycoa`	Korpusverarbeitungs- und Analyseframework als Python-Implementation
`/pycoa/analyse`	Analyseroutinen
`/pycoa/data`	Verwendete Wortlisten und aufbereitete Korpora
`/pycoa/data/grammar`	Grammatik für das regelbasierte Analyseverfahren (Kapitel 6.3.1)
`/pycoa/data/lists`	Wortlisten für die musterbasierten Analyseverfahren (Kapitel 6.3.2)
`/pycoa/data/morphlex`	Type-Liste des Osnabrücker Bildergeschichtenkorpus mit manuell eingefügten Morphemgrenzen (s. Kapitel 7.2.2)
`/pycoa/data/osbigk`	XML-Repräsentation des Osnabrücker Bildergeschichtenkorpus
`/pycoa/helper`	Hilfsscripte für die Datenaufbereitung
`/pycoa/templates`	LaTeX-, HTML- und SQL-Templates für die Analyseausgabe