

# Theorien auf dem Prüfstand – Evaluation phonologischer und orthographischer Regelhypothesen durch computerlinguistische Simulation

Tobias Thelen, Helmar Gust

## 1. Einleitung

Die Beurteilung von Theorien und der Vergleich von Theorien bezüglich ihrer Erklärungsmächtigkeit in Hinblick auf ihren Phänomenbereich ist nicht einfach. Dies resultiert zum einen aus der generellen Resistenz empirischer Theorien gegenüber einer Falsifikation (vgl. Schnell 1999:111ff.) und zum anderen aus den häufig unklaren Vergleichsparametern und der Unvollständigkeit der Testdaten. Darüber hinaus ist der Formalisierungsgrad einer Theorie sowie die formale Aufbereitung des Phänomenbereichs mitentscheidend zumindest für eine umfassende computerunterstützte Überprüfung.

Für den Phänomenbereich Schrift ist die Aufarbeitung einfach zu handhaben, zumindest solange keine handschriftlichen Phänomene betrachtet werden, denn die Daten liegen entweder in großer Menge maschinenlesbar vor oder lassen sich leicht (d.h. in einer 1:1-Kodierung) erfassen.

Die Formalisierung der Theorien so weit zu treiben, dass sie keine impliziten oder unklaren Bestandteile mehr enthält, vollständig und widerspruchsfrei ist, und somit die Mindestvoraussetzungen erfüllt, um vermittels eines Computerprogramms evaluiert zu werden, bleibt aber ein nichttriviales Problem. Wer eine Theorie durch Implementation überprüfen will, darf keinen Spielraum haben, der es ihm – willentlich oder unbewusst – ermöglichen würde, die Theorie zu verfälschen, auszuhebeln oder ins Gegenteil zu verkehren.

Utz Maas hat in seinen Schriften zur Phonologie und Orthographie weitgehend auf stark formalisierte Teile verzichtet, jedenfalls soweit es sich um die durchgängige Darstellung der theoretischen Annahmen in *einem* Formalismus handelt. Er bezeichnet eine formale Modellierung als „eine besonders weit getriebene Form dessen, was wissenschaftliche Argumentation ohnehin auszeichnet: die Verallgemeinerung einer

Beobachtung hin zu einer symbolischen Fassung, die die Ableitung von Hypothesen erlaubt, die wiederum wissenschaftlich überprüfbar sind.“ (Maas 1999:312).

Liegen Theorien in implementierbarer Form vor, können Computersimulationen gute Daten liefern, um sie zu vergleichen. Der Vergleich selber bleibt aber ein intellektueller Akt, da die von den Simulationen gelieferten Daten häufig keine eindeutigen Aussagen liefern und daher interpretiert werden müssen. Nicht allein die Zahl der abgedeckten oder nicht abgedeckten Fälle ist entscheidend. Im Falle komplementärer Leistungen verschiedener Theorien muss z.B. über den Status einzelner Fälle – „Ausnahme“ oder „Fehlleistung der Theorie“ – entschieden werden.

Im Folgenden werden zwei computerlinguistische Experimente vorgestellt, die beide eine formale Modellierung des phonographischen Teils der deutschen Orthographie leisten und die Durchführung umfangreicher Untersuchungen über die empirische Leistungsfähigkeit von Regelhypothesen ermöglichen. Beide Experimente haben ihren Ausgangspunkt in den entsprechenden Arbeiten von Utz Maas (1991, 1992, 1999, 2000), so dass die hier vorgestellten Ergebnisse auch Aussagen über die Formalisierbarkeit seiner Theorien sind.

## ***2. Testen von Regeln zur Herleitung korrekter Schreibungen***

In der aktuellen Orthographiediskussion wird erneut die Frage aufgeworfen, ob die deutsche Orthographie ein eigenständiges oder abhängiges System sei (Ossner 2001, Primus/Neef 2001), d.h.: Sind korrekte Schreibungen allein aus anderen sprachlichen Informationen ableitbar, oder hat die Schrift eigene Gesetzmäßigkeiten? Hypothesen über diese Frage werden theoretisch durch den Entwurf eines Systems oder einer systematischen Regel aufgestellt. Die Überprüfung dieser Hypothese kann auf verschiedene Weisen geschehen:

- Durch Einordnen der Hypothese in einen systematischen Zusammenhang mit anderen Annahmen über die Orthographie.
- Durch Test der Hypothese an speziell ausgewählten oder möglichst umfassenden Wortlisten.

Diese beiden Möglichkeiten sind nicht getrennt voneinander zu sehen. Weder ist eine empirisch nicht abgesicherte Hypothese, noch eine unsystematische Regelformulierung wünschenswert. Es ergeben sich Maßstäbe für die Frage, welche Hypothese „besser“ ist.

Eine, die möglichst viele Fälle abdeckt, oder eine, die den „systematischen Kern“ der Orthographie besonders herausstellt. Die Formalisierung und Implementierung als Computerprogramm erleichtert die Klärung beider Fragen. Ob eine Einzelregel zu einem Gesamtsystem von Regeln passt, lässt sich u.a. daran feststellen, ob sie mit demselben Vokabular, denselben Annahmen über sprachliche Strukturen und ihren Einfluss auf die Schreibung formulierbar ist. Sproat (2000:19f.) bezeichnet diese Eigenschaft als „consistency“. Der empirische Abdeckungsgrad einer Regelhypothese lässt sich bei erfolgter Formalisierung überprüfen: Es ist eine regelinterpretierende Umgebung zu schaffen, die automatisiert eine Menge von Eingaben abarbeitet, Schreibungen generiert und die dabei aufgetretenen Fehler klassifiziert. Dies ist im DFG-Projekt „Computerbasierte Modellierung orthographischer Prozesse“ (s. Maas et al. 1999) geschehen.

Ziel des Projektes war es, ein System zu entwerfen, in dessen Rahmen möglichst viele verschiedene Theorien ausdrückbar sind. Diese Eigenschaft hängt von zwei Bestandteilen ab: Dem verwendeten Regelformat samt Verarbeitungsmechanismus und der Menge der in den Eingabedaten vorgesehenen Informationen, auf die die Regeln zugreifen können. Die Verarbeitung der Eingaben geschieht in drei hintereinandergeschalteten Stufen. Im ersten Regelblock werden Regeln verwendet, die auf verschiedene phonologische und morphologische Eigenschaften zugreifen, spezifische Konfigurationen erkennen und diese Informationen durch spezielle Markierungen kennzeichnen. Auf dieser Stufe sind Regeln angesiedelt, die keine Bedingungen an den graphemischen Kontext stellen und nicht direkt graphemische Strukturen erzeugen. Als Beispiel lässt sich eine Regel zur Auslautverhärtung nennen, die für stimmlose Phoneme überprüft, ob es verwandte Formen gibt, in denen das Phonem stimmhaft realisiert wird. Die Formen, die für einen Abgleich herangezogen werden können, werden primär durch das Lexikon und die darin definierten Ableitungszusammenhänge bestimmt. Darüber hinaus kann jede Regel Bedingungen an erlaubte Abgleichformen stellen. Das System stellt also eine regelunabhängige Gesamtmenge an verwandten Formen bereit, über die theoriespezifische Filter laufen. Im zweiten Regelblock werden die mit Markierungen angereicherten Phoneme durch reguläre Ausdrücke in Grapheme übersetzt. Dies geschieht in Bezug auf die graphemische Ebene kontextfrei, d.h. rein graphotaktische Constraints wie „Kein Wort

endet auf <bb>“ (Neef 2002) können auf dieser Ebene nicht formalisiert werden. Dadurch können Strukturen generiert werden, die im abschließenden Schritt noch modifiziert werden müssen. Rein graphemische Filterregeln bilden somit den Abschluss. Für das unten dargestellte Experiment zur Schärfungsschreibung gibt eine Regel der zweiten Ebene, die für das Phonem, das mit einem Schärfungsmarker versehen ist, das Standardgraphem verdoppelt. Im dritten Schritt werden Strukturen wie <schsch> oder <chch> wieder reduziert und <kk> oder <zz> auf <ck> und <tz> abgebildet.

Aus computerlinguistischer Sicht ist diese Architektur komplex und widerspricht üblichen Forderungen nach Effizienz durch genau einen Verarbeitungsformalismus (s. Kaplan / Kay 1994). Sproat (2000:16f.) stellt die These auf, die Ableitung orthographischer Strukturen aus anderen linguistischen Ebenen sei grundsätzlich durch reguläre Ausdrücke beschreibbar. Der Vorteil des hier beschriebenen Verfahrens liegt aber in der einfachen Abbildung von Theorievorschlägen auf Regeln im System. Durch ein Protokoll der Regelanwendungen können zudem genauere Aussagen über die Anwendung einzelner Regeln getroffen werden. Dies wäre mit einer Menge von Regeln, die zu einem einzigen endlichen Automaten komprimiert ist (vgl. z.B. Karttunen 1994), nicht möglich.

Am Beispiel der Schärfungsschreibung soll kurz die Leistungsfähigkeit des Systems demonstriert werden. Es wurden vier verschiedene Regelhypothesen formalisiert und an drei verschiedenen Wortlisten getestet:

1. Regel NIE: Es findet keine regelhafte Schärfungsmarkierung statt.
2. Regel IMMER: Eine Schärfungsmarkierung erfolgt nach jedem betonten Kurzvokal.
3. Regel STAMM: „Folgt auf einen betonten Kurzvokal im Wortstamm nur ein einzelner Konsonant, so wird der Buchstabe für ihn verdoppelt.“ (Deutsche Rechtschreibung 1996:§2)
4. Regel ANSCHLUSS: Der heterosyllabische feste Anschluss zwischen Vokal und Konsonant wird durch Verdopplung des Buchstabens für den Konsonanten markiert. Diese Markierung „vererbt“ sich auf verwandte Formen, die einen homosyllabischen festen Anschluss aufweisen. (vgl. Maas 1992, 285ff.)

Die Regel ANSCHLUSS ist äquivalent zu einer Regel, die mit der Kategorie Silbengelenk operiert (Eisenberg 1998: 297ff.). Die Darstellung eines festen Anschlusses über eine Silbengrenze hinweg lässt sich im System auch als Markierung von Ambisyllabizität interpretieren. Die beiden letzteren Regeln wurden mehrfach diskutiert, auch in sehr heftiger Form (s. Eisenberg 1997, Ramers 1999a, Eisenberg 1999, Ramers 1999b, Sternefeld 2000). Bei der Vielzahl der Argumente und Gegenargumente fehlt eine systematische Untersuchung der Leistungsfähigkeit der Regeln, wie sie mit dem hier beschriebenen System durchgeführt werden kann.

Für das im Folgenden beschriebene Experiment wurden drei Wortlisten verwendet. Die Liste GWS enthält den ca. 1.200 Wörter umfassenden Niedersächsischen Grundwortschatz für Grundschulen, die Liste KAEDING 5.600 Formen aus dem Kaeding-Ortmann-Korpus (Ortmann 1975). Beide Listen sind gegenüber den Ausgangskorpora um Komposita bereinigt und enthalten größtenteils hochfrequente Wörter des Kernwortschatzes. Sie unterscheiden sich damit von der Wortliste CELEX, die aus dem CELEX-Korpus (Baayen et al. 1995) gewonnen wurde und mit insgesamt 148.000 Wortformen (für eine detaillierte Beschreibung der Korpora s. Maas et al. 1999) vollständige Flexionsparadigmen und viele Wörter aus peripheren Bereichen des deutschen Wortschatzes enthält. Die Regeln NIE und IMMER dienen zunächst dazu, die Menge einschlägiger Formen abzustecken. Die Wörter, die selbst von der Regel IMMER nicht mit einer Schärfungsmarkierung versehen werden (Fehlertyp: Nicht geschärft), sind zum einen „Pseudoschärfungen“ wie <Batterie> oder <illegal>, bei denen sich die Doppelkonsonantenschreibung nur durch ihre Herkunft aus einer anderen Sprache erklären lässt, zum anderen „Ausnahmen“ wie <allein>, bei denen die Konsonantengraphemverdopplung nur unter Rückgriff auf diachrone Formen möglich ist. Gleichzeitig sind die einfachen Regeln als „naive“ Hypothesen zu betrachten, die verdeutlichen können, wie groß der Erklärungsgewinn durch eine aufwendigere linguistische Analyse des Phänomens ist.

<b>Regel / Fehlertyp</b>	<b>GWS</b>		<b>KAEDING</b>		<b>CELEX</b>	
<b>NIE</b>						
Nicht geschärft	221	100%	833	100%	27844	100%
<b>IMMER</b>						
Nicht geschärft	5	2,5%	23	2,8%	2878	10,3%
überfl. Schärfung	122	100%	1254	100%	27521	100%

falsche Markierung	0 0%	0 100%	18 100%
<b>Mögliche Fehler Gesamt</b>	<b>343 100%</b>	<b>2087 100%</b>	<b>55383 100%</b>
<b>ANSCHLUSS</b>			
nicht geschärft	19 8,6%	41 4,9%	3925 14,1%
überfl. Schärfung	7 5,8%	17 1,9%	1527 5,5%
<b>Gesamt</b>	<b>28 7,8%</b>	<b>58 2,8%</b>	<b>5480 9,9%</b>
<b>STAMM</b>			
nicht geschärft	13 5,9%	62 7,4%	3900 14,0%
überfl. Schärfung	22 18,0%	48 3,8%	2553 9,3%
<b>Gesamt</b>	<b>35 10,2%</b>	<b>110 5,3%</b>	<b>6471 11,7%</b>
Verhältnis	1:1,25	1:1,90	1:1,18

**Tabelle 1: Ergebnisse verschiedener Schärfungsregeln**

Tabelle 1 zeigt die Resultate der vier Regeln für die drei Korpora. Wie zu erwarten, sind die beiden naiven Regeln deutlich schlechter, als die beiden linguistischen Regelhypothesen. Deren Vergleich ergibt für alle drei Korpora eine bessere Leistung der Regel ANSCHLUSS. Dieses Ergebnis widerspricht der Schlussfolgerung von Ramers (1997a, 62), nach der die Regel STAMM „leichte Vorteile“ auf ihrer Seite habe.

Neben dieser rein quantitativen Analyse ist eine genauere Auswertung möglich. Detaillierte Listen von Formen, für die die beiden Regeln unterschiedliche Schreibungen liefern, sind in Maas et al. (1999) aufgeführt, der Versuch einer systematischen Aufzählung der Fälle, in denen beide Regeln notwendigerweise zu identischen Resultaten gelangen, wird in Thelen (2001) unternommen. Hier sollen nur Fallgruppen zusammengefasst werden:

- Wortformen aus geschlossenen Wortarten, die nicht flektieren. Gerade einsilbige Formen weisen häufig keine Markierung auf, obwohl sie nach der Regel STAMM notwendig wäre (<in>, <um>, <mit>) und in wenigen Fällen eine Markierung, die nach der Regel ANSCHLUSS unnötig wäre (<denn>, <wann>). In der Wortliste KAEDING machen diese Formen einen größeren Anteil als in den Listen GWS und CELEX aus, was das unterschiedliche Verhältnis der Regelleistungen zu einem Großteil erklärt.
- Nomina mit –s Plural, für die sich keine trochäische Abgleichform finden lässt, oder nur fragwürdige Dativformen wie ?<dem Tricke> werden nach der Regel STAMM markiert, nach der Regel ANSCHLUSS aber nicht. Hier kann keine der beiden Regeln einen klaren Vorteil verbuchen.

Als Ergebnis lässt sich festhalten, dass die Regel ANSCHLUSS durchgängig einen höheren Abdeckungsgrad erreicht. Eine genauere Analyse der relevanten Fälle ergibt klare Kriterien und definierbare Fälle, die als Zusatzannahmen, als „systematische Ausnahmen“ formuliert werden können. Unter diesen Umständen zeigen dann beide Regeln nahezu identische Leistungen. Die Frage nach der grundsätzlich „besseren“ ist also nicht zu beantworten, sondern hängt von den Anforderungen an eine orthographische Regel ab. In den systematisch orientierten Darstellungen von Maas (1991, 1992, 2000) und den sprachdidaktischen Arbeiten von Röber-Siekmeyer (s. z.B. Röber-Siekmeyer / Pfisterer 1997) wird der Regel ANSCHLUSS der Vorzug gegeben, da sie auf Grundannahmen beruht, die auch für andere Phänomene der deutschen Orthographie relevant sind, bzw. sich für eine unterrichtliche Vermittlung besser eignen, als die der Regel STAMM.

An einem konkreten Beispiel wurde gezeigt, dass es möglich ist, linguistische Hypothesen über die Regelmäßigkeit der Orthographie so zu formalisieren, dass empirische Untersuchungen anhand repräsentativer Wortlisten möglich sind. Durch die Methode der computerlinguistischen Simulation ist dies mit geringem manuellen Aufwand für verschiedene Regeln und verschiedene Korpora möglich.

### **3. Phonologische Grammatik zum Parsen geschriebener Wörtern**

Im vorigen Abschnitt wurde ein System vorgestellt, das geschriebene Formen aus phonologischen und grammatischen Informationen mithilfe von Regeln ableitet. In der angewandten Computerlinguistik ist die umgekehrte Richtung – die Herleitung phonologischer bzw. phonetischer Repräsentationen aus geschriebener Sprache – in „Text-to-Speech“-Systemen zur Sprachsynthese ein wichtiger Forschungsbereich (vgl. Sproat 2000). In praktisch eingesetzten Systemen basieren manuell erstellte Regeln meist auf generativen Ansätzen (Bird 1994: iv), oder es werden Regeln mit maschinellen Lernverfahren induktiv erzeugt (s. z.B. Daelemans / van den Bosch 1996). Im Folgenden soll gezeigt werden, dass eine phonologische Grammatik, wie sie von Maas (1999: Kap. 12.3) vorgeschlagen wird, geeignet ist, orthographische Formen zu parsen. Das Ergebnis ist ein Baum, dessen Knoten phonologische Kategorien darstellen, und dessen Blätter Phoneme sind, die anhand der von Maas (1992, 2000) entwickelten

Systematik aus den vorgefundenen Graphemen struktur- und positionsabhängig abgeleitet werden.

Auffällig an diesem Vorgehen ist zunächst der Wechsel der Materialität: Geschriebenes wird mit Grammatikregeln für Gesprochenes interpretiert. Damit wird nicht die Existenz von Strukturen auf graphischer Ebene postuliert, die als Korrelate phonologischer, suprasegmentaler Kategorien gelten, wie etwa „Schreibsilben“ (vgl. Butt / Eisenberg 1990). Sondern es liegt umgekehrt die Auffassung zu Grunde, dass die phonologischen Strukturen aus den graphemischen rekonstruierbar sind – d.h. geschriebene Wortformen enthalten Merkmale, die auf suprasegmentale Eigenschaften der phonologischen Form rückschließen lassen. Einfache Beispiele von Homographen wie <Montage>, <rasten> oder <Lachen> zeigen, dass diese Hypothese nicht vollständig zutrifft (vgl. Thelen 2002). Das hier beschriebene Experiment einer Implementation der phonologischen Grammatik zum Parsen orthographischer Wortformen soll zeigen, wie es mittels computerlinguistischer Simulation möglich ist, den Abdeckungsgrad dieser und ähnlicher Hypothesen zu überprüfen.

Die von Maas vorgeschlagene Grammatik lässt sich in zwei Blöcke aufteilen. Der erste Block regelt die Struktur von Wörtern als Folge von Silben bestimmten Typs, der zweite Block die interne Struktur von Silben. Der erste Block besteht aus folgenden Regeln (Maas 1999:324):

$$\begin{array}{ll}
 (1) \quad W \rightarrow W + W & W \rightarrow (S^\circ +) F'' \\
 F'' \rightarrow (S +) F'' & F'' \rightarrow F' \\
 F' \rightarrow F' + S^\circ & F' \rightarrow F \\
 F \rightarrow F + S & F \rightarrow S'
 \end{array}$$

Demnach bestehen Wörter aus Füßen (F), die je genau eine prominente Silbe S' aufweisen und zusätzlich reduzierte (S°) und nicht-prominente nicht-reduzierte Silben (S) enthalten. Zum Parsen von Wortformen ist diese Grammatik allerdings nicht unmittelbar geeignet, da sie fast jede beliebige Folge von Silben(typen) als wohlgeformtes Wort zulässt, also mehr umfasst als nur die möglichen Wortformen des Deutschen (Maas 1999: 328). Das Problem wird in der vorliegenden Implementation durch die Einführung von Constraints gelöst, die wie in der Optimalitätstheorie (Prince / Smolenksy 1993) behandelt werden. Es wird eine geordnete Menge von Forderungen an wohlgeformte Formen gestellt, die Präferenzen widerspiegeln. Wenn es nicht möglich



ist, eine Forderung zu erfüllen, so wird derjenige „Kandidat“ (d.h. diejenige Strukturbeschreibung) ausgewählt, der eine niederrangige Forderung erfüllt. Im vorliegenden Fall wurden folgende Constraints verwendet:

- Wörter weisen möglichst wenig Füße auf.
- Die prominente Silbe eines Fußes bildet den linken Rand.
- Silben vom Typ S werden vermieden.

Im zweiten Regelblock ist der interne Aufbau von Silben festgelegt und es sind je nach Silbentyp und –position Paare von Graphemen und Phonemen angegeben. Damit ist es möglich, prosodische Eigenschaften der Wörter aus der graphischen Repräsentation abzulesen. Neben erlaubten Phonem- bzw. Graphemkombinationen in den Rändern ist insbesondere die Frage des Silbentyps und des losen bzw. festen Anschlusses zu klären. Dafür finden sich folgende Hinweise:

- Sondermarkierungen wie Schärfung oder Dehnung treten nur in prominenten Silben, bzw. an der Schnittstelle zwischen prominenten und reduzierten Silben auf.
- Reduzierte Silben weisen immer ein Graphem <e> auf.

Einige kleine Beispiele in (2) verdeutlichen das Zusammenspiel der Informationen über die graphemische, phonematische und prosodische Ebene. Alle Silben bestehen aus Anfangsrand (A) und Reim, der wiederum in Nukleus (N) und Endrand (E) verzweigt. Die Argumente in Klammern sind Informationen, die zwischen den Regeln ausgetauscht werden und über die z.B. Übereinstimmung bzgl. des ersten Arguments (Silbentyp) erreicht werden kann.

- (2) A( \_, /b/) -> <b>. ( möglicher Anfangsrand für alle Silbentypen)  
 A(S' ,/Str/) -> <str>. (<str> nur in prominenten Silben)  
 A(S' ,/h/) -> <h>. (<h> steht nur für /h/ im S'-Anfangsrand)  
 N(S' ,lose, /e:/) -> <eh>. (<eh> impliziert losen Anschluss)  
 N(S' ,lose, /e:/) -> <e>. (<e> kann mitlosem oder festen Anschluss gelesen  
 werden, der folgende Rand kann Aufschluss geben.)  
 N(S' ,fest, /E/) -> <e>. (Fester Anschluss, nur in S')  
 E( \_, \_, /p/) -> <b>. (Auslautverhärtung, in allen Silbentypen möglich)  
 E(S' ,fest, /tt/) -> <tt>. (Fester Anschluss, nur in S')

Die Grammatik wurde auf die oben erwähnte Wortliste GWS angewendet und zeigte gute Resultate. Tabelle 2 listet die Ergebnisse mit einer teilautomatisch vorgenommenen Fehlerklassifikation auf. Die aufgetretenen Fehler beziehen sich zum größten Teil auf nicht aus der graphischen Struktur dekodierbare Anschlussverhältnisse, die durch eine

morphologische Analyse bzw. den Abgleich mit verwandten Formen erkennbar wären. Insgesamt sind 97% der Formen entweder korrekt erkannt worden oder konnten systematisch erklärt werden.

Korrekt geparkt	81,6%	
Falscher Anschlusstyp	11,3%	<sprach>, <gut>, <Tasche>
Falsche Betonung	4,1%	<bevor>, <fabrik>, <höchstens>, <Moment>
Sonstiger Fehler	1,5%	<v>-Fehler, Vokalcluster <feiern>, <ferien>
Nicht parsbar / Datenfehler	1,5%	

**Tabelle 2: Ergebnisse für die Wortliste Gws**

Das Experiment zeigt, dass die Hypothese, dass sowohl die prosodische als auch die phonematische Struktur von Wörtern aus ihrer Schreibung ableitbar ist, für den Kernbereich des deutschen Wortschatzes weitgehend zutrifft.

#### **4. Zusammenfassung**

An zwei Beispielen wurde gezeigt, wie computerlinguistische Methoden benutzt werden können, um sprachwissenschaftliche Hypothesen zu formalisieren, experimentell an Daten zu überprüfen und die Ergebnisse dieser Experimente auszuwerten. Die Modellierungen haben in beiden Fällen nicht den üblichen Effizienzüberlegungen der Computerlinguistik Rechnung getragen, sondern Formalismen und Vorgehensweisen gewählt, bei dem die ursprünglich aufgestellten Regeln möglichst isoliert und klar erkennbar geblieben sind. Dadurch ist es möglich, den Einfluss und die Leistungsfähigkeit einzelner Regeln nachzuvollziehen. Für den Bereich Orthographie wurden für ein abgeschlossenes und ein laufendes Projekt Verfahren vorgestellt, die über den engeren linguistischen Rahmen hinaus weitere Anwendungsmöglichkeiten, z.B. für die Erklärung von Rechtschreibfehlern, haben (s.a. Maas et al. 1999, Thelen 2001).

#### **Literatur**

Baayen, Harald / Piepenbrock, Richard / van Rijn, Hedderik (1995). The Celex Lexical Database (CDROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

- Bird, Steven (1994): Introduction to Computational Phonology. In: Computational Linguistics, 20 (3): iii-ix.
- Butt, Matthias / Eisenberg, Peter (1990). Schreibsilbe und Sprechsilbe. In: Stetter, Christian (Hrsg.). Zu einer Theorie der Orthographie. Interdisziplinäre Aspekte gegenwärtiger Schrift- und Orthographieforschung. Tübingen: Niemeyer, 33-64.
- Daelemans, Walter / van den Bosch, Antal (1996). "Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion." In: Van Santen, Jan / Sproat, Richard / Olive, Joseph / Hirschberg, Julia (Hrsg.). Progress in Speech Synthesis. New York: Springer Verlag, 77-90.
- Deutsche Rechtschreibung (1996). Deutsche Rechtschreibung. Regeln und Wörterverzeichnis. Text der amtlichen Regelung. Tübingen: Narr.
- Eisenberg, Peter (1997). Die besondere Kennzeichnung der kurzen Vokale – Vergleich und Bewertung der Neuregelung. In: Augst, Gerhard / Blüml, Karl / Nerius, Dieter / Sitta, Horst (Hrsg.). Zur Neuregelung der deutschen Orthographie. Begründung und Kritik. Tübingen: Niemeyer, 323-335.
- Eisenberg, Peter (1998). Grundriss der deutschen Grammatik. Band 1: Das Wort. Stuttgart: Metzler.
- Eisenberg, Peter (1999). Vokallängenbezeichnung als Problem. In: Linguistische Berichte 179, 343-349.
- Kaplan, Ronald / Kay, Martin (1994). Regular models of phonological rule systems. In: Computational Linguistics, 20 (3):331-378.
- Karttunen, Lauri (1998). The Proper Treatment of Optimality in Computational Phonology. In: Proceedings of FSMNLP '98. International Workshop on Finite-State Methods in Natural Language Processing. Bilkent University, Ankara.
- Maas, Utz (1991): Orthographie als wissensbasiertes System. In: Osnabrücker Beiträge zur Sprachtheorie 44, 13-39.
- Maas, Utz (1992). Grundzüge der deutschen Orthographie. Tübingen: Niemeyer.
- Maas, Utz (1999). Phonologie – Einführung in die funktionale Phonetik des Deutschen. Opladen: Westdeutscher Verlag.
- Maas, Utz (2000). Orthographie. Materialien zu einem erklärenden Handbuch der Rechtschreibung des Deutschen. Ms. Osnabrück.

- Maas, Utz / Gust, Helmar / Albes, Christian / Noack, Christina / Thelen, Tobias (1999).  
Abschlußbericht des Projektes Computerbasierte Modellierung orthographischer  
Prozesse. Universität Osnabrück.
- Neef, Martin (2002). Warum man keinen Job mit Doppel-<b> bekommen kann.  
Graphematische Beschränkungen und Fremdwortintegration im Deutschen.  
Vortrag DGfS-Jahrestagung Mannheim, 28.2.2002.
- Ossner, Jakob (2001): Das <h>-Graphem im Deutschen. In: Linguistische Berichte  
187, 325-351.
- Ortmann, Wolf Dieter (1975): 7995 Wortformen der KAEDING-Zählung,  
rechnersortiert in alphabetischer und rückläufiger Reihenfolge, nach Häufigkeit  
und Hauptwortarten. München: Goethe-Institut, Arbeitsstelle für  
wissenschaftliche Didaktik, Projekt Phonotheke.
- Prince, Alan / Smolensky, Paul (1993): Optimality Theory: Constraint Interaction in  
Generative Grammar. Rutgers University Center for Cognitive Science,  
Technical Report 2.
- Primus, Beatrice / Neef, Martin (2001). Stumme Zeugen der Autonomie – Eine Replik  
auf Ossner. In: Linguistische Berichte 187, 353-378.
- Ramers, Karl-Heinz (1999a). Vokalquantität als orthographisches Problem: Zur  
Funktion der Doppelkonsonanzschreibung im Deutschen. In: Linguistische  
Berichte 177, 52-64.
- Ramers, Karl-Heinz (1999b). Zur Doppelkonsonanzschreibung im Deutschen: Eine  
Replik. In: Linguistische Berichte 179, 350-360.
- Röber-Siekmeyer, Christa / Pfisterer, Katja (1997). Silbenorientiertes Arbeiten mit  
einem leseschwachen Zweitklässler. Begründung und Beschreibung einer nicht  
buchstabenorientierten Unterrichtsfolge zum Lesenlernen. In: Weingarten,  
Rüdiger / Günther, Hartmut (Hrsg.): Schriftspracherwerb. Baltmannsweiler:  
Schneider Verlag Hohengehren, 36-61.
- Sproat, Richard (2000). A Computational Theory of Writing Systems. Cambridge:  
Cambridge University Press.
- Thelen, Tobias (2001). Was bringen linguistische Regeln? Untersuchungen zur  
Regelmäßigkeit von Rechtschreibung und Rechtschreibfehlern. In: Andresen,  
Ute / Kluge, Wolfhard (Hrsg.): Wir reden miteinander. Theorien und Praxen des

Schriftspracherwerbs. München, Lich: Deutsche Gesellschaft für Lesen und Schreiben, 90-97.

Thelen, Tobias (2002). Schrift ist berechenbar. Zur Systematik der Orthographie. In: Röber-Siekmeyer, Christa / Tophinke, Doris (Hrsg.). Schrifterwerbskonzepte zwischen Sprachwissenschaft und Pädagogik. Baltmannsweiler: Schneider Verlag Hohengehren.