Bibliographischer Hinweis sowie Verlagsrechte bei den online-Versionen der DD-Beiträge:

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
http://www.didaktik-deutsch.de
16. Jahrgang 2011 – ISSN 1431-4355
Schneider Verlag Hohengehren GmbH

Albert Bremerich-Vos & Miriam Possmayer

ZUR RELIABILITÄT EINES MODELLS DER ENTWICKLUNG VON TEXTKOMPETENZ IM GRUNDSCHULALTER

In: Didaktik Deutsch. Jg. 16. H. 31. S. 30-49.

Albert Bremerich-Vos & Miriam Possmayer

Zur Reliabilität eines Modells der Entwicklung von Textkompetenz im Grundschulalter[1]

Der Beitrag bezieht sich auf ein von Gerhard Augst, Katrin Disselhoff, Alexandra Henrich, Thorsten Pohl und Paul-Ludwig Völzing entwickeltes und 2007 publiziertes Modell. Augst u. a. begleiteten 39 Kinder vom zweiten bis zum vierten Schuljahr; sie hatten in diesen drei Schuljahren jeweils identische Schreibaufgaben zu den Textsorten bzw. -mustern Erzählung, Bericht, Instruktion, Beschreibung und Argumentation zu bearbeiten. Von jedem Kind liegen also dreimal fünf Exemplare dieser Textsorten vor, d. h. 585 Texte. Für jede Textsorte wurde ein Modell mit vier Entwicklungsniveaus postuliert, darüber hinaus nahm man an, es handele sich um Varianten eines einzigen Kernmodells. Die Frage, ob die Zuordnung der Texte zu den Niveaus reliabel ist, wurde von Augst u. a. nicht im Einzelnen überprüft. Sie steht im Zentrum des vorliegenden Beitrags und kann, bedenkt man einige Einschränkungen, vorsichtig positiv beantwortet werden. Besondere Aufmerksamkeit gilt narrativen und argumentativen Texten, vor allem deshalb, weil geplant ist, die Modelle für diese Textsorten auf ein Korpus von Texten aus der Feder von Dritt- und Viertklässlern anzuwenden, das im Rahmen der Arbeiten des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) entstanden ist.

1 Zur Ausgangslage

1.1 Zur aktuellen Debatte

Bei der Diskussion darüber, wie die Schreibentwicklung angemessen zu verstehen ist, hat ein Modell von Carl Bereiter Referenzstatus (Bereiter 1980, Augst/ Faigel 1986, Bereiter/ Scardamalia 1987, Schneuwly 1988, Jechle 1992, Feilke 1996, Becker-Mrotzek 1997, Hug 2001, Bachmann 2002, Feilke 2003, Feilke/ Schmidlin 2005, Fix 2006). Demnach integrieren Experten im Rahmen entwickelten Schreibens folgende Teilfähigkeiten: Flüssigkeit der Produktion geschriebener Sprache, Flüssigkeit im Bereitstellen von Wissen, Beherrschung von Schreibkonventionen, Übernahme der Perspektive von anderen (d.h. Lesern), reflexive Bewertung des Geschriebenen auf lokaler und globaler Ebene.

Die Fähigkeiten, flüssig zu schreiben und Wissen bereitzustellen, konstituieren die basale „Strategie“ des assoziativen Schreibens (associative writing). Man schreibt hin, was einem gerade einfällt, Wiederholungen z. B. eingeschlossen. Es kommt al-

1 Das Bundesministerium für Bildung und Forschung unterstützt das Projekt, auf das hier Bezug genommen wird, unter dem Förderkennzeichen 01 GJ 0902.

lenfalls zu lokaler Kohärenzbildung. Ist diese „inhaltszentrierte“ Strategie beherrscht, kommt es zur allmählichen Integration eines weiteren Fähigkeitskomplexes. Man orientiert sich verstärkt an Schreibkonventionen (nicht nur im Hinblick auf Syntax einschließlich Orthografie, sondern auch mit Rücksicht auf Textsortennormen) – normorientiertes Schreiben (performative writing). Über „Inhalt“ und „Form“ hinaus wird dann auch die Größe „Leser“ systematisch berücksichtigt. Mit der Integration insbesondere der Fähigkeit zur Perspektivenübernahme korrespondiert die Fähigkeit, die eigene Perspektive zu dezentrieren. Das Schreiben wird kommunikativ (communicative writing). Die Fähigkeit zur Dezentrierung ist Voraussetzung für das unified writing, von Baurmann als „authentisches“, von Bachmann als „kritisches“ Schreiben etikettiert. Jetzt hat der Schreibende genügend Ressourcen für eine (selbst)kritische Prüfung der Prozesse und (bisherigen Vor-) Produkte zur Verfügung. „Der Text wird nun als etwas empfunden, das zu gestalten ist, d.h. Schreiben wird nicht mehr nur als instrumentelle Fähigkeit, etwas mitzuteilen, gesehen, sondern weit stärker als eine produktive Tätigkeit.“ (Eigler u. a. 1990, 17) Ist hier wieder das Produkt im Fokus, so ist auf einem letzten im Modell vorgesehenen Level wie schon beim assoziativen Schreiben die Prozesskategorie involviert. Das Schreiben wird nicht mehr als nachträgliche Fixierung von bereits Gewusstem begriffen, sondern der Schreibende gelangt zu neuem, „tieferem“ Verstehen von Sachverhalten: epistemisches Schreiben (epistemic writing).

An Bereiters Modell ist u. a. kritisiert worden, dass keine Theorie zugrunde liege, auf deren Basis plausibel wird, warum für dieses oder jenes Niveau u. a. einmal ein Prozess- (wie beim assoziativen und epistemischen Schreiben), einmal ein Produkt- (wie beim normorientierten und authentischen Schreiben) und einmal der Leseraspekt (beim kommunikativen Schreiben) leitend sein soll. Außerdem werde stufen- bzw. niveauintern zu wenig differenziert, was insbesondere beim normorientierten Schreiben auffalle (vgl. Feilke 2003, 181). Andererseits kann im Rahmen dieses Modells verständlich werden, „wie durch Automatisierung und Routinisierung von Tätigkeiten auf unteren Ebenen komplexere Fähigkeiten […] ins Spiel kommen können und so die jeweils begrenzte Verarbeitungskapazität ausgelastet, aber nicht überlastet wird […]“ (Eigler u. a. 1990, 18f).

Weitgehend einig ist man sich in der Beschreibung des Beginns sowie des Ziels der Schreibentwicklung. Novizen zeichnen sich durch assoziatives, wenig geplantes und kaum adressatenorientiertes Schreiben aus, was Bereiter/ Scardamalia (1987) als Resultat einer Strategie des „knowledge telling“ beschreiben. Experten sind demgegenüber in der Lage, die zum Teil widerstreitenden Anforderungen an einen Text auszubalancieren, indem sie ihr Wissen über Inhalte, Adressaten, Textsorten bzw. -muster nach Maßgabe der eigenen Schreibziele umstrukturieren und auf diese Weise das Schreiben auch in epistemischer Funktion nutzen, auch als Strategie des „knowledge transforming“ (ebd.) bezeichnet.

Didaktisch zentral ist nach wie vor die Modellierung textsortenspezifischer[2] Kompetenzmodelle. Unterscheidet man grob dominant narrative, informative (u. a. deskriptive, instruktive) und argumentative (persuasive) Texte (vgl. Loomis/ Bourque 2001), dann lässt sich im Hinblick auf die Curricula eine markante Abfolge erkennen: In der Grundschule dominiert das Schreiben narrativer Texte, in der Sekundarstufe sind es überwiegend deskriptiv-instruktive Texte und gegen Ende der Sekundarstufe I und vor allem in der Sekundarstufe II überwiegt interpretativ-argumentatives Schreiben. Hierbei handelt es sich um didaktisches Brauchtum, das durch Rekurs auf empirische Forschung kaum gestützt werden kann. So konstatierte Feilke noch 2003 (183): „Es gibt eine überschaubare Zahl textsortendifferenzierender Untersuchungen […]. Untersuchungen zur Interdependenz von Textsorten in der Entwicklung gibt es bis heute nicht."

Soweit es um das Grundschulalter geht, ist die Entwicklung der Kompetenz, schriftlich zu **erzählen**, am intensivsten untersucht (z. B. Bitter Bättig 1999, Schmidlin 1999, Hug 2001, Becker 2002). So nimmt z. B. Wolf (2000) in Übereinstimmung mit Resultaten von Studien zur Genese mündlicher Erzählfähigkeit (vgl. Boueke u. a. 1995) vier „Stufen" bzw. Niveaus an: In einem **enumerativen** Modus werden Aussagen (nur) lokal gereiht. Das Motiv ist inhaltlich assoziativ, wie von Bereiter angenommen. Im Rahmen eines zweiten, linear **sequenzierenden** Modus werden bereits Textmerkmale bedacht, insofern Sätze vor allem auf der Basis stereotyper temporaler Konnektoren wie „und dann" verbunden werden. Auf einem dritten Niveau wird zusätzlich **kontrastiert**. Dasjenige, was die Geschichte überhaupt erzählenswert macht, die Komplikation, der Erwartungsbruch, wird auch sprachlich markiert. Das höchste Niveau ist nach Wolf erreicht, wenn affektive Markierungen integriert sind, wenn also angenommen werden kann, dass der Leser, die Leserin bewusst in das Geschehen **involviert** wird – kommunikatives Schreiben nach Bereiter. Dieses Niveau ist, folgt man den Darstellungen von Boueke und Wolf, in der Regel zu Beginn der Sekundarstufe I erreicht.

Studien, in denen die Entwicklung der Fähigkeit thematisch wird, informative, insbesondere **instruierende Texte** zu schreiben, sind eher spärlich (Becker-Mrotzek 1997, Bachmann 2002). Nach Becker-Mrotzek ist ein großer Teil von Viertklässlern nicht in der Lage, Rudimente einer Bedienungsanleitung für eine digitale Stoppuhr

2 Hier wird nicht auf die Nuancen des Gebrauchs von Ausdrücken wie „Textmuster", „Textart", Textfunktion" und „Textsorte" eingegangen. Schulisch relevantes Erzählen, Beschreiben, Berichten, Instruieren und Argumentieren kann man als prototypische Handlungsmuster bzw. als Texttypen fassen. Ein Text als „Fall" einer Textsorte mag dann aus verschiedenen Handlungsmustern bzw. Texttypen bestehen. So kann in einem primär argumentativen Text eine narrative Passage vorkommen, die als Exempel zur Stützung einer These fungiert, und ein Exemplar der Textsorte Erzählung kann mehr enthalten als das, was zum Handlungsmuster Erzählen gehört, z. B. Teile des Musters Beschreiben. Auf diese ‚Komplikationen' wird im Unterricht oft nicht eingegangen; insofern kann man die in der Schule in der Regel vermittelten Textsortennormen mit einigem Recht als (zu) rigide ansehen.

zu formulieren. Bachmann (2002) fand u. a., dass (einige wenige) Schülerinnen und Schüler vierter Klassen im Rahmen von Anleitungsaufgaben (z. B. Erläuterung eines Strategiespiels) nur sehr wenige Kohäsionsmittel nutzten.

Studien zur Entwicklung der Fähigkeit von Kindern im Grundschulalter, **Berichte** zu verfassen, gab es bis vor kurzem – d.h. vor Augst u. a. (2007) – nach unserer Kenntnis nicht. Dies mag u. a. dem Umstand geschuldet sein, dass der Bericht immer noch vornehmlich als eine Textsorte angesehen wird, die erst in der Sekundarstufe I Lerngegenstand ist (Feilke 2006).

Auch Arbeiten zur Entwicklung der Kompetenz, schriftlich zu **beschreiben**, sind spärlich (vgl. vor allem Schneuwly/ Rosat 1986, Feilke 2004). Schneuwly und Rosat nehmen für die Entwicklung der Kompetenz, Räume zu beschreiben, ebenfalls vier „Stufen" bzw. „Strategien" an: So finden sich auf der ersten Stufe Listen, d.h. stereotype Aufzählungen in sprachlich mehr oder weniger identischer Form. Mehr als zwei Drittel der Zweitklässler und ein Viertel aller Viertklässler folgen, so ihr Befund, dieser Strategie. Auf der zweiten Stufe beschreiben Kinder zwar ihnen wichtige Objekte und fassen auch für eine Raumbeschreibung prototypische Objektgruppen zusammen. Es fehlt aber noch ein Referenzpunkt, der als Bezugsrahmen und Orientierungshilfe für den Leser dienen könnte. Knapp die Hälfte aller Viertklässler befindet sich auf dieser Stufe. Auf den folgenden Stufen werden Gegenstände zwar lokal im Raum verortet und auch aufeinander bezogen; die Konstruktion eines globalen Referenzrahmens gelingt, so die Autoren, Grundschulkindern aber noch nicht. Selbst viele Achtklässler hätten noch Schwierigkeiten, diese für die Orientierung der Leser zentrale Teilaufgabe zu meistern.

In Arbeiten zur **Entwicklung der literalen Argumentationskompetenz** wird zuweilen betont, Argumentationen seien deshalb vergleichsweise schwierig, weil die Textsorte wenig prägnant ist. Primär argumentative Texte können z. B. ausgeprägt narrative Partien in der Funktion von stützenden Beispielen enthalten und sind dann unter Umständen von Narrationen nur schwer zu unterscheiden. Es wurde angenommen, dass die Texte der Schülerinnen und Schüler in vierten Klassen primär als Reihen von isolierten Feststellungen, Einstellungsäußerungen usw. erscheinen, die in der ersten Person formuliert sind (Feilke 1995, zusammenfassend Feilke 2003, 187ff).

1.2 Zur Studie von Augst, Disselhoff, Henrich, Pohl und Völzing

Die erste längsschnittlich angelegte linguistisch-sprachdidaktische Untersuchung zur Entwicklung der Textsortenkompetenz im deutschsprachigen Raum haben Augst u. a. 2007 vorgelegt. Die Stichprobe bestand aus 39 Kindern zweier paralleler Klassen einer Schule in Hessen, die zu drei Zeitpunkten, am Ende des 2., 3. und 4. Schuljahrs, jeweils fünf Texte zu schreiben hatten, und zwar einen erzählenden, einen berichtenden, einen instruktiven, einen beschreibenden und einen argumentativen. Die Schreibaufgaben waren jeweils identisch, was u. a. in motivationaler Hinsicht einige Probleme bereitete. Die Mitteilungen über die Umstände, unter denen die Texte geschrieben wurden, sind recht spärlich (Augst u. a. 2007, 37f), sodass of-

fen bleiben muss, ob sie hinreichend standardisiert waren. Es wurden nur Texte von Kindern mit Deutsch als Muttersprache berücksichtigt. Von jedem Kind gibt es also 15 Texte, jeweils drei Bearbeitungen derselben Schreibaufgabe; insgesamt liegen 585 Texte vor.[3]

Bei der Aufgabe zum Erzählen bekamen die Schülerinnen und Schüler ein Bild, auf dem ein Kind oder auch Zwerg mit einer brennenden Kerze in der Hand zu sehen war, das bzw. der in eine Höhle, einen Tunnel oder dergleichen schaut bzw. geht. Die Instruktion lautete: „Denk dir zu diesem Bild eine interessante Geschichte aus und schreibe sie für das Geschichtenbuch auf." (Ebd., 46) Als Schreibauftrag zum Bericht wurde formuliert: „Manche Kinder, die neu nach Deutschland kommen, kennen das Weihnachtsfest gar nicht. Schreibe für sie doch einmal ganz genau auf, wie du in deiner Familie das Weihnachtsfest feierst." (Ebd., 97) Als Aufgabenstellung im Kontext der Instruktion wählte man: „Im Unterricht, z.B. im Sportunterricht, spielt ihr ganz unterschiedliche Spiele. Schreibe doch mal eine Spielanleitung zu deinem Lieblingsspiel auf, sodass Kinder, die neu nach Deutschland kommen und das Spiel nicht kennen, gleich mitspielen können." (Ebd., 123) Die Instruktion im Fall der Beschreibung: „Beschreibe für die Kinder, die neu nach Deutschland kommen, dein Zimmer/ deinen Klassenraum. Dann können sie sich genau vorstellen, wie es dort aussieht, wo du wohnst/ wo du in der Schule lernst." (Ebd., 168) Und schließlich der Schreibauftrag, der sich auf das Argumentieren bezieht: „Professor Augst von der Universität in Siegen ist auf eine Idee gekommen. Er meint, dass man Autos abschaffen soll. Was hältst du von diesem Vorschlag? Schreibe ihm einen Brief." (Ebd., 200)

Auf der Basis detaillierter Analysen der Schülertexte präsentierten Augst u. a. einen Vorschlag, wie die Schreibentwicklung zum einen **textsortenübergreifend und** zum anderen **textsortenspezifisch** zu modellieren ist.

In einem ersten „Stadium" bzw. auf einer ersten „Stufe"[4] liegt ein knowledge telling im Sinne Bereiters und Scardamalias (1987) vor. Es dominiert ein subjektiver, assoziativer Zugang zum Schreibgegenstand; von „Texten" kann eigentlich noch nicht die Rede sein. In einem zweiten Stadium werden – bereits textsortenspezifisch – Sachverhaltsbeziehungen selegiert, die im Fall des Erzählens z.B. in Form des häufig dargestellten „Freskostils" verschriftet werden, d.h. als „und dann"-Kette (s. Wolf 2000). Im dritten Stadium werden die stereotyp gestalteten Sequenzen „aufgebrochen" und es kommt zu einer, allerdings nur lokalen, Ausdifferenzierung verschiedener Textteile, z.B. zur Markierung einer Komplikation. Schließlich kann das Textganze (bzw. das Exemplar der Textsorte) nicht nur lokal und „regional", sondern global antizipiert werden. Der Fokus hat sich verlagert bzw. ausgedehnt. „Der Text in seiner Gegliedertheit wird jetzt von den Autoren von seinem Ende her bzw. von seinem funktionalen Ziel her wahrgenommen und gestaltet." (Augst u. a. 2007,

3 Das Korpus ist unter www.text-sorten-kompetenz.de abrufbar.

4 Die Autoren sprechen sowohl von „Stadien" als auch von „Stufen". Auf die Implikationen dieser Unschärfe soll hier nicht weiter eingegangen werden.

235) Es wird z.B. eine Pointe sprachlich realisiert oder im Fall eines argumentativen Textes eine Konklusion explizit formuliert.

Die vier „Stufen“[5] werden als „selektierte Assoziationen“, „sequenzierte Selektionen“, „perspektivierte Sequenzen“ und „synthetisierte Perspektiven“ bezeichnet. In dieser Abstraktheit können sie, so die Autoren, als textsortenübergreifend verstanden werden. Es sei aber mit textsortenspezifisch mehr oder weniger „schnellen“ Entwicklungen zu rechnen.

Zur Illustration des Modells zunächst auf „horizontaler“ Ebene die Charakterisierungen dessen, was textsortenspezifisch als „perspektivierte Sequenzen“, als dritte „Stufe“ also, firmiert: Beim Erzählen gelingt die Aufspaltung der linearen Ereignisreihung vornehmlich in Exposition und Planbruch. Beim Berichten werden Teilhandlungen unterschieden, etwa in vorbereitende und solche, die sich auf das eigentliche Fest beziehen. Beim Instruieren kommt es zu einer Gliederung nach Spielutensilien bzw. Vorbereitungsteil auf der einen und dem eigentlichen Spielablauf auf der anderen Seite. Beim Beschreiben gelingt der Ausgang von einem Fixpunkt (z. B. der Klassen- oder Zimmertür) bzw. die Bildung von zwei Raumachsen. Beim Argumentieren schließlich werden Pro- und Contra-Argumente unterschieden.[6]

---

5 Augst u. a. sind, wenn wir recht sehen, Verfechter der „starken“ Version einer Stufentheorie der Entwicklung. Sie weisen zwar darauf hin, dass im schulischen Kontext auch Instruktion, Kommunikation bzw. Kooperation und Rezeption eine Rolle spielen. So werde der komplexe Lerngegenstand via Instruktion zerlegt und man setze z. B. auf Kooperation in Form von Schreibkonferenzen. Darüber hinaus „imitierten“ die Schülerinnen und Schüler in wachsendem Maß „diejenigen sprachlichen Ausdrucks- und Struktureigenschaften, die sich in der Sprachgemeinschaft konventionell mit einer bestimmten Textsorte verbinden“ (Ebd., 26). Sprachliche Sozialisation, auf Seiten des Lernenden dergestalt als Imitation gefasst, wird dann als „Gegenbegriff“ (ebd., 26) zum Begriff der Aneignung verstanden. Aneignung sei im Wesentlichen durch die Merkmale des Gegenstands Schriftsprache bedingt. „Es sind die internen Konstitutions-, Abhängigkeits- und Restriktionsbedingungen des Aneignungsgegenstandes, die den Lerner zu einem bestimmten Vorgehen oder zu einer bestimmten Strategie zwingen. Sie prägen den Gang der Entwicklung in seiner internen Strukturiertheit vor. Aus dieser Konstellation erwächst eine integrative Aneigungsfolge, in der ein nächster Schritt notwendig die vorausgehenden voraussetzt […].“ (Ebd., 232) Wir halten diese Sichtweise aus verschiedenen Gründen für problematisch, u. a. deshalb, weil Aneignung hier primär als monologischer Prozess gedacht und kommunikativen Prozessen zu wenig Beachtung geschenkt wird (vgl. Bremerich-Vos 1996 und vor allem Miller 1986). Auf die damit angedeuteten Differenzen kommt es hier aber nicht an.

6 Augst u. a. fanden keine Evidenz dafür, „dass das Erzählen der Motor auch für die Entwicklung der anderen vier Textsorten ist, wie es Wolf annimmt (2000). Die Höhe der Texteinstufung in der 2., 3. und 4. Klasse lässt keine Vorhersage zu über die Höhe der Einstufung der anderen Textsorten, bezogen auf jedes einzelne Kind“ (Ebd., 356). Wolf (2000, 375) schreibt zwar, „daß dem Erzählerwerb eine über sich selbst hinausweisende paradigmatische Rolle […] zuzugestehen ist“, aber auch: „Die Versuche, die bereichsübergreifende Wirkung der narrativen Strukturelemente plausibel zu machen, haben niemals dem Zweck gedient, die Behauptung zu unterstützen, daß direkte Rückschlüsse von

Auf „vertikaler“ Ebene ist das Modell der „Stufen“ narrativer Kompetenz wie folgt bestimmt:

> 1. Stufe
> Unzusammenhängendes assoziatives erzählerisches Einzelnes; oft geschichtenübliches Repertoire; formelhafte sprachliche ‚geschichtenübliche‘ Wendungen; sehr oft Ich-Erzählungen als (außergewöhnliche) Alltagsgeschichten ohne *sprachlichen* Planbruch und ohne Pointe
>
> 2. Stufe
> Zusammenhängendes kohärentes erzählerisches Geschehen auf der Zeitachse ausgewählter notwendiger Ereignisse mit geschichtenüblichem Repertoire; formelhafte sprachliche, geschichtenübliche Wendungen, neben Ich-Erzählungen als außergewöhnliches Alltagsgeschehen auch Er-Erzählungen, beides mit beginnender Fiktionalität; meist noch ‚und-dann‘-Verknüpfung; gelegentlich nicht vollendete Erzählungen; in vielen Fällen inhaltliche Planbrüche, seltener Pointen, jedoch noch nicht sprachlich realisiert.
>
> 3. Stufe
> Zusammenhängendes erzählerisches Geschehen auf der Zeitachse mit einem sprachlich klar herausgearbeiteten Planbruch, jedoch mit keiner oder einer inhaltlich wie sprachlich sehr schwachen Pointe, die wenig Überraschendes enthält; es überwiegen Er-Erzählungen mit fiktionalem Charakter; die ‚und-dann‘-Verknüpfung wird durch temporale Adverbien und Konjunktionen abgelöst. Die Texte sind oft gerahmt (vor allem Einleitung – seltener Schluss) und enthalten teilweise eine Coda.
>
> 4. Stufe
> Zusammenhängendes erzählerisches Geschehen mit versprachlichtem Planbruch, Aufbau von Spannung, klarer inhaltlicher und sprachlicher Pointe, die meist ein Überraschungsmoment enthält; meist Er-Erzählungen mit fiktionaler Struktur; das Erzähltempus ist eindeutig Präteritum; die Texte sind gerahmt, oft mit Coda; durch Rede und Gegenrede wird eine szenische Dramaturgie […] aufgebaut; in manchen Texten wird ein durchgängiger Erzählton erreicht. (Augst u. a. 2007, 51f)

Zwei Textexempel, die Stufe 1 und 4 illustrieren sollen: [7]

> Das Bergwerk // es war einmal / ein Bergwerk / es wurde vor / 80 Jahren / nicht mehr ben / utzt es lebte / ein Kobold in / der Höhle. // sie leben ohne Essen und Trinken. / auf einmal kam ein Monster / und auf einmal kam ein Kobold / raus, und das Monster sagte: ‚Ich besch / ütze eure Höhle.‘ (Domenic, 3. Klasse)

> Eines Tages / ging ein / Junge in / eine Höhle / er verirrte / sich und / kam nie / wieder. Ein / anderer Junge / wollte der / Sache auf den Grund gehen. Er / ging in die Höhle. Als er eine halbe / Stunde ging hörte er eine Stimme / sagen: ‚Wer ist da, was willst du / hier?‘ Der Junge zitterte. Dann / fragte wieder jemand: ‚Was ist da, was willst / du hier.‘ Der Junge sagte mit zittriger / Stimme: ‚Ich woll wollte nu nu / nur der der Sache auf den Grund ge ge / gehen wegen dem ver ver / verschwundenen Jungen. Ich / bin der verschwundene Junge.‘ / sagte eine Stimme. Warum bist du denn verschwunden / fragte der andere Junge. Ich bin verschwunden weil, ich wissen / wollte was ihr in der Höhle ist‘ /

---

den Leistungen in einem Bereich auf Leistungen in einem anderen möglich sind.“ (Ebd., 403f.) Insofern dürfte der Dissens kleiner sein als zunächst angenommen.

7 Ein „ /“ markiert einen Zeilensprung, ein „ //“ eine Leerzeile.

> sagte der Verschwundene. Komm mit ich führe dich raus‘ sagte der andere Junge. / Sie gingen aus der Höhle und / alle waren froh das der verschwundene Junge wieder da / ist. Aber er ist jetzt kein / Junge mehr sondern ein erwachsener Mann. Er lebt mit seiner Familie jetzt glücklich und / gesund. (Cornelia, 4. Klasse) (Ebd., 52f)

Die Einstufung der Texte und die Entwicklung des Stufenmodells gingen Hand in Hand. Jeweils zwei Projektmitglieder beurteilten Texte eines Jahrgangs und einer Textsorte; Dissense wurden vermerkt und anschließend in der Autorengruppe diskutiert und entschieden (ebd., 40). Es wird allerdings nicht mitgeteilt, wie hoch die Übereinstimmungen bei den verschiedenen Textsorten waren, ob die Urteile deutlich oder nur gering differierten usw.

An diesem Punkt setzt unsere Studie an. Uns geht es nicht um detaillierte linguistisch-didaktische Analysen der Schülertexte, die im Übrigen bereits von Augst u. a. geleistet werden. Es geht auch nicht um Erörterungen des Verhältnisses von Analyse und Interpretation, z.B. im Kontext der Frage, inwiefern allein anhand von Merkmalen der „Textoberfläche“ zu entscheiden ist, ob ein inhaltlicher Planbruch klar, andeutungsweise oder gar nicht vorliegt. Es soll ebenfalls z.B. nicht gefragt werden, wie Urteile zur Kohärenz der Texte objektiv zu begründen sind, vor allem dann, wenn Konnektoren fehlen und zu entscheiden ist, ob Schlüsse auf Leserseite eher autorseitig intendiert sind oder aber elaborativen Charakter haben. Für uns ist entscheidend, dass das Modell, das Augst u. a. vorgelegt haben, jedenfalls dann, wenn seine Zuverlässigkeit überprüft werden soll, ein holistisches Rating erfordert. Die Schülertexte wurden ja jeweils **als ganze** einer Stufe zugeordnet. Ratings sind per definitionem nicht objektiv; das Ausmaß der Übereinstimmung verschiedener Beurteiler ist aber überprüfbar – und es kann erhöht werden.

2 Zur Anlage unserer Studie

2.1 Ein Vergleich von zwei holistischen Kompetenzniveaumodellen

Während die Reliabilität des Modells bei Augst u. a. allenfalls beiläufig zur Sprache kam, steht sie im Folgenden im Zentrum. Die Schülertexte sollten von mehreren Beurteilern unabhängig voneinander **holistisch** kodiert werden. Jeder Beurteiler produziert für jeden Text also eine der Zahlen 1 bis 4, womit er die seiner Ansicht nach jeweils erreichte Stufe markiert. Bei einem **analytischen** Kodieren dagegen hätte man Punkte vergeben können z.B. für die inhaltliche, die strukturelle und die sprachliche Dimension und letztere ließe sich z.B. noch differenzieren nach Aspekten des Wortschatzes, nach Arten und Anzahl orthographischer Fehler, nach syntaktischer Variabilität usw. Die auf diesem Wege gewonnenen Informationen wären für diagnostische Zwecke, z.B. für den Aufweis je individueller Profile, hilfreich. So mag es sein, dass ein Schüler in inhaltlicher und struktureller Hinsicht passabel verfährt, aber wenig Kontrolle über den Satzbau hat. Eine andere Schülerin wiederum mag syntaktisch versiert sein, aber Schwierigkeiten damit haben, ihren Text plausibel zu organisieren.

Es liegt auf der Hand, dass eine analytische Kodierung im Hinblick auf die Diagnostik der schriftsprachlichen Fähigkeiten **einzelner** Schülerinnen und Schüler mehr Informationen liefert als ein holistisches Verfahren. Für die individuelle Förderung ist die analytische Kodierung insofern das angemessene Mittel. Die hier gewählte holistische Variante bietet sich dagegen an, wenn es um die ökonomische Beurteilung einer großen Zahl von Texten geht. Das ist z.B. bei den Large-scale-Untersuchungen zur Normierung der Bildungsstandards der Fall. Hinreichend reliable holistische Kodierungen können auch für Zwecke eines Screenings genutzt werden: Man schaut sich dann weitere Texte von Schülerinnen und Schülern an, die eine bestimmte Stufe nicht erreicht haben, und überprüft genauer, ob Förderbedarf besteht.

In der deutschdidaktischen Forschung hat die holistische Kodierung der Texte von Grundschülern nach unserer Kenntnis keine Tradition – anders als in den USA. Hier wurde im Auftrag des National Center for Education Statistics und im Rahmen des National Assessment of Educational Progress (NAEP) zuletzt im Jahr 2002 eine „Writing Report Card" vorgestellt (Persky, Daane, Jin 2003). Sie bezog sich u. a. auf Ergebnisse einer Testung der Schreibfähigkeiten einer für die USA repräsentativen Stichprobe von Viertklässlern. Man unterschied zwischen narrativem, informativem und persuasivem (bzw. argumentativem) Schreiben und nahm jeweils sechs Kompetenzniveaus an. Auf jedem Level finden sich vier komplexe Deskriptoren, zum Teil in identischer Form, zum Teil textsortenspezifisch. Sie stehen für die Dimensionen der sprachlichen Richtigkeit, der sprachlichen Angemessenheit, der Struktur bzw. Organisation und die inhaltliche Dimension. Das unterste und das oberste Niveau beim informativen Schreiben, das – bezogen auf das Modell von Augst u. a. – in etwa die Textsorten Beschreiben, Berichten und Instruieren umfasst, sind z.B. wie folgt bestimmt:

1 Unsatisfactory Response (may be characterized by one or more of the following)
- Attempts a response, but may only paraphrase the task or be extremely brief.
- Exhibits no control over organization.
- Exhibits no control over sentence formation; word choice is inaccurate across the response.
- Characterized by misspellings, missing words, incorrect word order; errors in grammar, spelling, and mechanics severely impede understanding across the response. –

6 Excellent Response
- Develops ideas well and uses specific, relevant details across the response.
- Is well organized with clear transitions.
- Sustains varied sentence structure and exhibits specific word choices.
- Exhibits control over sentence boundaries; errors in grammar, spelling, and mechanics do not interfere with understanding.

(National Center for Education Statistics 2003, 87)

Vergleicht man das damit angedeutete Modell mit dem von Augst u. a., wie es oben anhand der Textsorte Erzählen exemplifiziert wurde, dann zeigt sich Folgendes: Die Dimension der sprachlichen Richtigkeit (orthographische Fehler, Grammatikfehler) spielt bei Augst u. a. erklärtermaßen kaum eine Rolle, allenfalls dann, wenn es um das Erzähltempus geht. Auch die stilistische Dimension ist nicht zentral. Der Ange-

messenheit des Wortschatzes und der Variabilität des Satzbaus wird keine besondere Aufmerksamkeit geschenkt. Insofern Erzählen gefragt ist, spielt allerdings die direkte Rede als Mittel, den Leser zu involvieren, eine Rolle. Sprachliche Aspekte kommen bei Augst u. a. in erster Linie dann ins Spiel, wenn die Dimension der Struktur (bzw. „organization") betrachtet wird: Sind ein Planbruch und eine Pointe nicht nur inhaltlich, sondern auch sprachlich zu erkennen? Gibt es nicht nur „und dann"-Verknüpfungen, sondern kommen auch andere Junktoren vor? Im Modell von Augst u. a. dominiert eindeutig der Aspekt der Struktur; was diese Dimension angeht, so wird im US-Modell deutlich weniger differenziert.[8] Im Hinblick auf die inhaltliche Dimension schließlich interessiert Augst u. a. vor allem, inwiefern strukturell relevante Inhalte vorkommen, während im amerikanischen Modell vornehmlich auf den Reichtum inhaltlicher Details abgehoben wird.

Der Vergleich erhellt, welche Aspekte im Modell von Augst dominieren und welche eher marginal sind. Es wäre reizvoll, die Texte aus dem Korpus von Augst u. a. mithilfe **beider** Modelle beurteilen zu lassen. Die Korrelationen müssten hoch ausfallen. Man spricht hier von konvergenter Validität. Auch Lehrerurteile zu den Schreibkompetenzen der Kinder könnten berücksichtigt werden – ein Weg, der im Rahmen der Validitätsprüfung häufig gewählt wird. Dies ist zwar didaktisch von Belang, bleibt hier aber außer Betracht.

2.2 Zur Bestimmung der Beurteilerübereinstimmung

Einfach zu berechnen ist die prozentuale Übereinstimmung, sei es für alle Rater insgesamt oder als gemittelte Übereinstimmung für alle Raterpaare. Bezugsgröße wäre jeweils der Modalwert, d.h. der am häufigsten vorkommende Messwert. Wählt man die prozentuale Übereinstimmung als Maß, nimmt man in Kauf, dass sie u. a. zufallsbedingt ist. Beurteiler können ja (buchstäblich) raten. Als zufallskorrigiertes Maß der Übereinstimmung kommt vor allem Cohens Kappa in Betracht. Es nimmt Werte zwischen -1 und +1 an, womit das Ausmaß bezeichnet ist, in dem die beobachtete Übereinstimmung von der Zufallserwartung abweicht, wobei gilt, „that a negative value indicates poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement." (Fleiss, Cohen 1973, 613) Die Formel lautet:

$$\frac{\text{Differenz der beobachteten prozentualen und der bei Zufall erwarteten prozentualen Übereinstimmung}}{\text{Differenz der maximal möglichen und der bei Zufall erwarteten prozentualen Übereinstimmung}}$$

8 Dies sei wiederum anhand der Hinweise zur Einstufung informierender Texte belegt. Für die Niveaus 2 bis 5 finden sich die folgenden Formulierungen: „Is very disorganized or too brief to detect organization." – „Is disorganized or provides a disjointed sequence of information." – „Provides a clear sequence of information; provides pieces of information that are generally related to each other. " – „Is clearly organized; information is presented in an orderly way, but response may lack transitions. " (Ebd., 87)

bzw.

$$\frac{P_0 - P_e}{1 - P_e}$$

Für Zwecke der Illustration dieser Formel sei angenommen, dass zwei Beurteiler unabhängig voneinander 100 Texte auf vier Levels einzustufen hatten. Hier die fiktiven, d.h. nicht auf das Korpus von Augst u. a. bezogenen Ergebnisse:

	Rater 1				
		1	**2**	**3**	**4**	**Summe**
Rater 2	**1**	6	3	0	0	9
	2	13	44	7	2	66
	3	4	3	9	2	18
	4	1	0	3	3	7
	Summe	24	50	19	7	100

Tab. 1: Fiktives Beispiel zur Berechnung von Beurteilerübereinstimmung

Die Ergebnisse des Kodierers 2 sind zeilen-, die des Kodierers 1 spaltenweise eingetragen. Die Zahl der identischen Urteile ist aus der Diagonalen ersichtlich. Es wurden 62 (6 + 44 + 9 + 3) von 100 Texten identisch beurteilt, d.h. 62 Prozent (= P_0). Die bei bloßem Raten zu erwartende Übereinstimmung wird wie folgt bestimmt: Man addiert die Produkte der jeweiligen Randsummen, also der spalten- und zeilenweisen Summen jeweils für die Kategorien 1, 2, 3 und 4, und dividiert diese Summe durch das Quadrat der Anzahl der Objekte. Dann erhält man im Zähler 24 x 9 + 50 x 66 + 19 x 18 + 7 x 7 (= 3907) und im Nenner 100 x 100 (= 10000). Als bei Zufall erwartete prozentuale Übereinstimmung resultiert demnach 0,39. Setzt man die Werte in die Formel ein, ergibt sich als Cohens Kappa (0,62 – 0,39) : (1 – 0,39) = 0,38. Cohens Kappa fällt also deutlich niedriger als die prozentuale Übereinstimmung aus.

Im Beispiel haben die Rater einzelne Kategorien unterschiedlich häufig vergeben. So wurde Kategorie 1 von Rater 2 neun Mal, von Rater 1 aber in 24 Fällen vergeben, bei Kategorie 2 ist das Verhältnis 66 zu 50. Die Grundwahrscheinlichkeiten, dass 1 bzw. 2 gewählt werden, sind also deutlich unterschiedlich. Rater 1 scheint im Hinblick auf die Vergabe der Kategorie 1 strenger zu sein als Rater 2, bei Kategorie 2 ist es umgekehrt. Nur bei Kategorie 4 sind die Randsummen für beide Rater identisch. Wenn die Randverteilungen wie hier zum Teil massiv voneinander abweichen, dann fällt der Wert von Kappa niedriger aus. Wären die Randverteilungen nicht deutlich voneinander verschieden und wäre der Kappa-Wert gleichwohl niedrig, dann läge das daran, dass die Rater nicht konsistent urteilen, also unsystematisch strenger bzw. milder urteilen als ihre jeweiligen Partner.

Die Daten, die im Rahmen der vierstufigen Beurteilung der Schülertexte anfallen, haben nicht nur nominales, sondern ordinales Niveau. Ein Text, der auf Stufe 3 ein-

gestuft wird, ist dem Anspruch nach nicht nur **anders** als ein Text auf Stufe 1, sondern er gilt auch als **besser**. Ist zusätzlich ein Vergleich der **Distanzen** zwischen den Werten 1 bis 4 möglich? Ist also der Abstand zwischen 1 und 2 so groß wie der zwischen 2 und 3? Wenn es sich so verhielte, dann läge ein intervallskaliertes Rating vor und man könnte die Reliabilität mithilfe der Intraklassenkorrelation (ICC) bestimmen. Um eine **Intra**klassenkorrelation handelt es sich deshalb, weil für **dasselbe** Objekt (hier: denselben Text) mehrere Messwerte, d.h. Urteile verschiedener Rater, vorliegen. Sie ist der Produkt-Moment-Korrelation ähnlich, die sich auf den linearen Zusammenhang zweier intervallskalierter Merkmale bezieht. Den Fragen, ob nicht nur Ordinal-, sondern sogar Intervallskalenniveau gegeben ist, bzw. ob man wie z.B. häufig im Umgang mit schulischen Noten „so tun kann, als ob" es sich um intervallskalierte Daten handle, soll hier nicht weiter nachgegangen werden. Wir halten dafür, dass die Daten „nur" ordinales Niveau haben.[9]

Wünschenswert wäre ein Ergebnis, bei dem einerseits die Unterschiede der Modalwerte für die einzelnen Texte möglichst groß sind. Das spräche dafür, dass die wahre Varianz der Texte groß ist und ihre Qualität über die vier Kategorien hinweg streut. Dies aber nur dann, wenn zugleich die Unterschiede der Messwerte für einen einzelnen Text möglichst klein sind. Cohens Kappa ist ein Maß für die **Gleichheit** der vergebenen Werte. Jeder Messunterschied bei ein und demselben Text wird also als Fehler aufgefasst. Man könnte angesichts der komplexen, mehrdimensionalen Größe „Textqualität" aber auch eine weniger strenge Version wählen, wonach nicht die absolute Übereinstimmung, sondern nur die **Ähnlichkeit** der Urteile angestrebt wird. Dabei gibt es Grade von Nichtübereinstimmungen, die sich entsprechend gewichten lassen: Eine Abweichung vom Modalwert um eine Stufe sollte weniger stark ins Gewicht fallen als eine Differenz um mehrere Kategorien.[10] Es gibt eine

9 Details zu Maßen der Reliabilität intervallskalierter Ratings findet man bei Wirtz & Caspar (2002, 157ff). Zum Thema Skalenniveau von Rating-Skalen schreiben Bortz, Döring (1995, 168): „Die Kontroverse zu diesem Thema hat eine lange Tradition und scheint bis heute noch kein Ende gefunden zu haben. Die messtheoretischen ‚Puristen' behaupten, Rating-Skalen seien nicht intervallskaliert; sie verbieten deshalb die statistische Analyse von Rating-Skalen mittels parametrischer Verfahren […]." Anders die „Pragmatiker". Entscheidend sei, zwischen messtheoretischen Interpretationsproblemen und mathematisch-statistischen Voraussetzungen zu unterscheiden. „Die Skalenqualität der Zahlen wird erst bedeutsam, wenn man die Ergebnisse interpretieren will. Es sind dann meßtheoretische Erwägungen, die dazu veranlassen, die Ergebnisse einer Varianzanalyse über Nominalzahlen für nichtssagend zu erklären, weil die Mittelwerte derartiger Zahlen, die in diesem Verfahren verglichen werden, keine inhaltliche Bedeutung haben […]." (Ebd.) Bezogen auf unsere Fragestellung: Welche Bedeutung hätte etwa ein mittlerer Wert der Urteile aller Rater über einen Text von 2,51, etwa im Vergleich mit einem Mittelwert von 2,49? Wäre dem ersten Text Niveau 3 zuzuordnen, dem zweiten Niveau 2? Die Annahme einer solchen inhaltlichen Differenz lässt sich unserer Meinung nach nicht rechtfertigen. Deshalb halten wir die ordinale Lesart für angemessen.

10 Die Formel zur Berechnung des gewichteten Kappa findet man bei Wirtz & Caspar (2002, 79).

Reihe von Möglichkeiten, diesem Postulat Rechnung zu tragen. So kann man die quadrierten Differenzen als Gewichte eintragen (vgl. Wirtz /Caspar 2002, 80f.) und erhielte für eine Abweichung um eine Stufe den Wert 1, bei einer Abweichung um zwei Stufen resultierte ein Wert 4 und bei einer Differenz von drei Stufen müsste die Zahl 9 in die Matrix eingetragen werden. Eine andere Variante besteht darin, die Gewichte jeweils zu verdoppeln, was z.B. so realisiert werden kann, dass man für eine Differenz von einer Stufe einen Gewichtungsfaktor von 0,25, für eine von zwei Stufen 0,5 und für eine von drei Stufen den Faktor 1 vorsieht.

2.3 Zum Ablauf und den Ergebnissen der ersten Kodierertrainings

Zunächst suchten und fanden wir per Aushang 12 Kodiererinnen und Kodierer, durchgängig Studierende der Germanistik. Sie wurden zunächst mit dem Modell von Augst u. a. vertraut gemacht, das ihnen auch in schriftlicher Form vorlag. Zu jeder der fünf Textsorten gab es zunächst Benchmark-Texte, welche die einzelnen Stufen in paradigmatischer Weise repräsentieren. Die Zuordnungen der Schülertexte zu den Stufen stammten aus Augst u. a. (2007). So war gesichert, dass die Urheber des Modells selbst als „Beurteilungs-Meister" fungierten. In der Folge hatten die Rater Texte zu beurteilen, die den Stufen in zufälliger Reihenfolge zugeordnet waren. Im Anschluss an die Probekodierungen wurden die Ergebnisse besprochen (vgl. zum Prozedere detaillierter Weigle 2002, 130f.). Zuhause kodierten die Studierenden zunächst jeweils 20 Exemplare der fünf Textsorten, also insgesamt 100 Texte. Die Resultate waren nicht zufriedenstellend. Betrachtet man für jedes Kodiererpaar die Übereinstimmungen im Hinblick auf den Modalwert, so ergaben sich bei insgesamt 66 paarweisen Vergleichen Übereinstimmungen von weniger als 60 Prozent: beim Erzählen in 43 (von 66), beim Berichten in 26, beim Beschreiben in 33, beim Instruieren in 30 und beim Argumentieren in 29 Fällen. Dabei ist die Grenze bei 60 Prozent bereits sehr „liberal" angesetzt. Es handelt sich hier nur um die prozentuale Übereinstimmung, die noch nicht im Hinblick auf das zufällig zu erreichende Agreement relativiert ist.[11] Untersuchungen zum Verhalten der einzelnen Rater ergaben u. a., dass mehrere Beurteiler inkonsistent urteilten.

Ausgehend von der Hypothese, dass der Umgang mit Schülerprodukten, die sich nicht weniger als fünf verschiedenen Textsorten zuordnen lassen, die Kodierer verwirrt haben könnte, wurde eine Nachschulung angesetzt, und zwar zu den Textsorten, für welche die Ergebnisse des ersten Trainings am besten und am schlechtesten ausgefallen waren. Diese Extreme bildeten die Berichte und die Erzählungen. Es

11 Der Fixierung eines Werts als kritischer Grenze haftet etwas Willkürliches an. Richtwerte sollten nicht am Schreibtisch, sondern unter Berufung auf Resultate anderer Studien festgelegt werden, die dieselbe Domäne betreffen. Solche Studien liegen nach unserer Kenntnis bislang nicht vor. Wirtz & Caspar (2002, 59) sprechen bei einem Kappa ab etwa .6 (bis .75) von guter Übereinstimmung. Hat eine Skala aber nur wenige Kategorien und bleiben mehrere Zellen der Matrix unbesetzt, fällt das Kappa niedriger aus (ebd.). Insofern lässt sich eine Grenze bei einer Übereinstimmung von 60 Prozent rechtfertigen.

sollte überprüft werden, ob sich die besonders schlechten Ergebnisse für die Erzählungen massiv und ob auch die relativ befriedigenden Resultate für die Berichte noch merklich verbessert werden könnten.

Die ersten Schritte der Nachschulung entsprachen dem Vorgehen bei der ersten Schulung; im Anschluss hatten die Studierenden jetzt aber **alle** Erzähltexte und Berichte zu kodieren, also jeweils 117 Texte, d.h. für jedes der 39 Kinder je drei Erzählungen und drei Berichte. Wieder war der Grad der Übereinstimmung vergleichsweise niedrig. In 38 von 55 Fällen beim Berichten und sogar in 52 von 55 Fällen beim Erzählen wurde die Grenze von 60 Prozent unterschritten.

Bedenkt man mögliche Quellen der Beurteilungsvarianz, dann kommen vor allem das Modell selbst, die „Qualität" der Rater und die Güte der Schulung in Betracht. Die erste Schulung war für die Rater womöglich überfordernd. Aber die zweite hatte bei verändertem Setting keine überzeugendere Wirkung. Was die Rater angeht, so gab es vereinzelt eine Tendenz zur Strenge; einige Rater urteilten wie angedeutet aber auch inkonsistent. Was die Debatte darüber angeht, inwiefern das Modell selbst als Varianzquelle in Frage kommt, so ergab sich, bezogen auf die Textsorten, die nur im **ersten** Kodierdurchgang eine Rolle gespielt hatten: zunächst Folgendes:

- Die Aufgabe zum **Instruieren** ist so formuliert, dass die Schülerinnen und Schüler benachteiligt sein mögen, die sich ein vergleichsweise komplexes Lieblingsspiel ausgesucht haben. Die Testung wird hier ungewollt unfair und für die Kodierer resultiert die nur schwer zu lösende Aufgabe, der Heterogenität der Spiele gerecht zu werden.
- Ähnlich verhält es sich mit der Aufgabe zum **Beschreiben**, jedenfalls dann, wenn das eigene Zimmer als Gegenstand gewählt wird. Es ist leichter, dem Leser eine Orientierungshilfe zu geben, wenn der Grundriss des Zimmers „einfach", die Zahl der als relevant erachteten Utensilien klein ist usw.
- Zu den **argumentieren Texten**: Hier erscheint nicht die Aufgabenstellung als solche als problematisch, sondern deren Beziehung zu zwei Indikatoren der Textgüte. Gefragt ist ja, was man von Professor Augsts Vorschlag hält, Autos abzuschaffen. Es ist nicht explizit verlangt, dass man dialektisch bzw. dialogisch argumentiert, also Pro **und** Contra berücksichtigt. Auch eine „lineare" Argumentation, in deren Rahmen nur ein Pro, ein Contra oder auch eine vermittelnde Position begründet wird, ist u.E. mit der Aufgabenstellung kompatibel. Diese lineare Version wird von den Verfassern des Modells aber als nicht hinreichend angesehen. Sie sind darüber hinaus der Auffassung, dass in einem auf der höchsten Stufe platzierten Text die Konklusion am Ende zu stehen hat. Auch diese Forderung erscheint als zu streng. Warum sollte sie (bzw. die Position) nicht bereits zu Beginn formuliert sein dürfen? Verführe man in diesen beiden Hinsichten weniger streng, so die Erwartung, würden mehr Schülertexte auf höheren Stufen platziert und die Varianz würde größer.

Im Hinblick auf die Textsorten, die im **zweiten** Kodierdurchgang im Zentrum standen, ergab sich:

- Die Aufgabe zum **Berichten** kann von den Schülerinnen und Schülern auf verschiedene Weisen verstanden werden, was den Konstrukteuren des Modells durchaus bewusst ist: „Schreibe ich nun, wie *man* Weihnachten (bei uns, in Deutschland) feiert oder wie *wir* es feiern oder wie *wir* es *das letzte Mal* gefeiert haben, so dürfte es in den Hinterköpfen – kaum bewusst – geklungen haben“ (Völzing 2007, 98). Diese Auslegungsbreite erschwert ersichtlich die Kodierung.
- Es ist nicht klar, ob bei der **narrativen** Aufgabe, die auf eine „Höhepunkterzählung“ zielt, bereits das **Betreten** der Höhle, des Kellers oder dergleichen als unerwartetes, vom „normal course of events“ abweichendes Ereignis, als Planbruch, aufzufassen ist (Augst u. a. 2007, 58). Verhielte es sich so, dann wäre der Planbruch bereits vorgegeben. Oder „darf“ erst in der Höhle Unerwartetes geschehen? Darüber hinaus haben Divergenzen der Beurteiler, folgt man ihren Auskünften, mit dem Konzept der Pointe zu tun. Für Augst et al ist, anders als etwa für Boueke u. a. (1995), der Planbruch nur „eine notwendige, aber keine hinreichende Bedingung der Erzählwürdigkeit. Erzählwürdig wird eine Geschichte durch eine Pointe, auf sie hin ist die ganze Geschichte angelegt. […] Dies ist der meist witzige, aber auf jeden Fall für den Leser unerwartete, daher überraschende Einfall, der sich gegen das bisher abgelaufene Geschehen plötzlich auftut.“ (Ebd., 50f) Augst halten dafür, dass das Geschehen erst nach der Pointe wieder „normal“ wird. Das erzählerische Gelingen der Pointe ist dem Modell zufolge zentrales Merkmal von Niveau 4. Die Kodierer stimmten öfter gerade im Hinblick auf die Pointe nicht überein. Das verwundert nicht, geht es hier doch nicht um ein dichotomes, sondern um ein graduierbares Merkmal. Was für die einen unerwartet bzw. überraschend **genug** war, war für die anderen **zu wenig** unerwartet oder überraschend. Unserer Auffassung nach ist dieses Kriterium, zumal im Grundschulkontext, unangemessen streng.

2.4 Modifizierte Modelle für Erzählen und Argumentieren und Resultate einer erneuten Kodierung

Die Modelle für die einzelnen Textsorten sind intrinsisch mit den jeweiligen Schreibaufgaben verknüpft. Fielen die Aufgaben anders aus, wären auch die Modelle zu modifizieren. Bei der Instruktion, der Beschreibung und beim Bericht sind die Einwände gegen die Aufgaben**formulierung** gewichtig: Bei den ersten beiden Textsorten kann sie, so die These, die systematische Benachteiligung derer begünstigen, die es mit einem komplexeren Spiel bzw. Zimmer zu tun haben. Beim Bericht sind erklärtermaßen mehrere Lesarten plausibel, was eine reliable Auswertung erschwert. Anders verhält es sich beim Erzählen und Argumentieren. Hier sind es unserer Auffassung nach nicht die Schreibaufgaben selbst, die eine reliable Kodierung erschweren, sondern nur die genannten Indikatoren, d.h. im Fall des Erzählens vor allem die geforderte Pointe und beim Argumentieren die Position der These bzw. Konklusion und die für das oberste Level allein zugelassene „dialektische“ Version.

Für einen **dritten** Kodierdurchgang wurden die Modelle für Erzählen und Argumentieren in der damit angedeuteten Weise modifiziert. Was das **Erzählen** angeht, so

wurden zwar diverse Details vorsichtig reformuliert und anders gruppiert. Inhaltlich gravierend sind aber wie erläutert nur Änderungen, die das Konzept der Pointe betreffen. Das wird besonders deutlich anhand des Vergleichs der Deskriptionen des vierten Levels. In der Version von Augst u. a. heißt es (vgl. S. 36 dieses Beitrags):

> *Stufe IV*
> Zusammenhängendes erzählerisches Geschehen mit versprachlichtem Planbruch, Aufbau von Spannung, klarer inhaltlicher und sprachlicher Pointe, die meist ein Überraschungsmoment enthält; meist Er-Erzählungen mit fiktionaler Struktur; das Erzähltempus ist eindeutig Präteritum; die Texte sind gerahmt, oft mit Coda; durch Rede und Gegenrede wird eine szenische Dramaturgie […] aufgebaut; in manchen Texten wird ein durchgängiger Erzählton erreicht. (Augst u. a. 2007, 51f)

Die modifizierte, weniger „strenge“ Version lautet:

> *Stufe IV*
> Es gibt einen inhaltlich und sprachlich klar herausgearbeiteten Planbruch und eine Auflösung, in der die Verhältnisse vor dem Planbruch wiederhergestellt werden oder sonst irgendeine Lösung zu finden ist. Zusätzlich kann es eine Coda geben. Der Text wirkt ‚von hinten her‘ geplant. Das Tempus ist fast durchgängig Präteritum (bzw. Plusquamperfekt).
>
> Es handelt sich fast durchgängig um Er-Erzählungen. Direkte Rede (auch Wechselrede) ist sehr häufig.
>
> Andere Mittel, den Leser in die Geschichte ‚hineinzuziehen‘ bzw. die Emotionen der Akteure und/oder des Erzählers darzustellen, tragen zum Eindruck eines durchgängigen ‚Erzähltons‘ bei. (Augst u. a. 2007, 51f)

Auch bei den argumentativen Texten ist in erster Linie die Modifikation, die sich auf die Beschreibung der vierten Stufe bezieht, inhaltlich relevant. Ursprünglich verlangt waren ein Diskussionsteil mit Pro- und Contra-Argumenten und eine darauf folgende Conclusio bzw. ein „Schlusssatz“ (Völzing 2007, 203).

Die revidierte Fassung dieser Stufenbeschreibung lautet:

> *Stufe IV*
> Es können Argumente für und gegen eine Abschaffung von Autos bzw. für eine vermittelnde Position unterschieden werden. Das muss nicht immer geordnet geschehen. Die argumentationstypischen Tätigkeiten des Abwägens und Einschränkens sind entweder klar erkennbar oder es wird stringent nur für eine Seite argumentiert. Es gibt ein klares Fazit, eine Festlegung auf ein Pro oder Contra oder eine vermittelnde Position. Dieses Fazit ist inhaltlich stringent auf das Folgende bzw. auf das bis dahin Geschriebene bezogen. Insofern wirkt der Text geschlossen bzw. vom Ende her geplant. (Augst u. a. 2007, 51f)

Sieben Studierende der Germanistik, die auch an der ersten Schulung teilgenommen hatten, wurden mit den beiden modifizierten Modellen vertraut gemacht. Sie schätzten alle narrativen und argumentativen Texte der Grundschüler ein, also jeweils 117 Produkte. Für die resultierenden paarweisen Vergleiche wurde das gewichtete Kappa berechnet. Dabei schlagen Abweichungen vom Modalwert, die maximal drei Niveaus betragen können, mit den Faktoren 0,25 (bei Abweichung um eine Stufe), 0,5 (bei zwei Stufen) und 1 (bei drei Stufen) zu Buche. Bei 21 Vergleichen ergaben sich für die Erzählungen ein Mittelwert von .59 und eine Standardabweichung von .05. Für die Argumentationen resultierten im Mittel Übereinstimmungen von .64; die

Standardabweichung betrug .03. Anders als bei der prozentualen Übereinstimmung handelt es sich hier, daran sei erinnert, um ein Maß, bei dem die durch Zufall erreichbare Übereinstimmung berücksichtigt ist. Betrachtet man zum Beispiel die paarweisen Vergleiche der Ratings zum Argumentieren, so unterschritten hier nur 4 von 21 die Grenze von .6 – also rund 20 Prozent. Inwiefern diese Grenze, berechnet auf der Basis der beschriebenen Version des gewichteten Kappas, mittelfristig als zufriedenstellend angesehen werden kann, wird sich erweisen, wenn weitere Studien vorliegen. Zieht man Arbeiten heran, die sich auf die Sekundarstufe I beziehen, dann zeigen sich dort für fünfstufige Skalen mittlere Korrelationen von etwa .45 bis .70 (Böhme, Bremerich-Vos & Robitzsch 2008, 299). Dabei ist zu bedenken, dass die Varianz der Schülerleistungen in der Sekundarstufe I deutlich größer ist als im Grundschulbereich.

Zurzeit wird geprüft, inwiefern sich die Modelle im Umgang mit einem großen Korpus von Schülertexten bewähren, die im Rahmen von Testungen des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin entstanden sind.

3 Kleines Fazit

Das von Augst, Disselhoff, Henrich, Pohl und Völzing vorgelegte, nach unserer Kenntnis bislang ambitionierteste und fünf Textsorten integrierende Modell wurde ansatzweise auf seine Reliabilität hin überprüft. Die Resultate sollten u.E. dazu ermutigen, weitere Untersuchungen von Stufenmodellen dieser Art, ergänzt um Analysen der Validität, in Angriff zu nehmen. Solche Arbeiten sollen detaillierte Analysen einzelner Schülertexte im Hinblick auf inhaltliche, strukturelle und sprachliche Merkmale nicht **ersetzen**. Für solche Detailstudien ist das Buch von Augst u. a. übrigens eine Fundgrube. Holistische Ratings von Schülertexten können aber eine **ergänzende** Funktion haben. Sie eignen sich für Large-scale-Untersuchungen, bei denen mit geschulten Testleitern gearbeitet wird, wohl aber auch für Projekte wie VERA 3 (Vergleichsarbeiten im Fach Deutsch in dritten Klassen), bei denen die Lehrkräfte selbst für die Auswertung der Schülertexte verantwortlich sind. Die Ergebnisse können dann mit den Annahmen zur Schreibkompetenz der Schülerinnen und Schüler verglichen werden, die im Verlauf des Unterrichts und in Kenntnis vieler weiterer Texte entstanden sind. So mögen sich Hinweise auf eine Bestätigung, aber auch auf eine Korrektur bisheriger diagnostischer Urteile ergeben, die für die Unterrichtsentwicklung genutzt werden können. Auch jenseits von VERA könnten sich Lehrkräfte darauf verständigen, Schülertexte aus verschiedenen Klassen ab und zu gemeinsam zu beurteilen. Ein auf seine Güte hin überprüftes Modell wie das von Augst u. a. wäre ein geeignetes, ökonomisch einsetzbares Instrument.

Potentielle Nachteile sollen aber nicht verschwiegen werden. Sollen möglichst viele Dimensionen der Schülertexte holistisch erfasst werden, z.B. auch Aspekte der Orthographie und der Grammatik, dann können sich zumal bei jüngeren Schülerinnen und Schülern unterschiedliche Profile ergeben – eine wichtige Information, die man im Rahmen holistischen Kodierens nicht bewahren kann. So mag ein Text zwar ko-

härent sein, aber zahlreiche Grammatikfehler enthalten; ein anderer wiederum kann syntaktisch elaboriert, inhaltlich aber belanglos sein. Sollen solche „Profile“ erfasst werden, bietet sich ein analytisches Rating an, das überdies den Vorteil hat, tendenziell zuverlässiger zu sein: „just as reliability tends to increase when additional items are added to a discrete-point test, so a scoring scheme in which multiple scores are given to each script tends to improve reliability […]“ (Weigle 2002, 120) Es kommt also auf die Zwecke an, die man verfolgt, und darauf, das Verhältnis von Aufwand und Ertrag zu bedenken. Immer sollte es aber auch um eine Antwort auf die Frage gehen, wie bei der Beurteilung von Schülertexten ein möglichst hohes Maß an Übereinstimmung gesichert werden kann.

Literatur

Augst, G., Disselhoff, K., Henrich, A., Pohl, T. & Völzing, P.-L. (2007): Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter. Frankfurt/M.

Augst, G. & Faigel, P. (1986): Von der Reihung zur Gestaltung. Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13 bis 23 Jahren. Frankfurt/M.

Bachmann, T. (2002): Kohäsion und Kohärenz: Indikatoren für Schreibentwicklung. Innsbruck.

Becker, T. (2001): Kinder lernen erzählen. Zur Entwicklung der narrativen Fähigkeiten von Kindern unter Berücksichtigung der Erzählform. Hohengehren.

Becker-Mrotzek, M. (1997): Schreibentwicklung und Textproduktion. Der Erwerb der Schreibtätigkeit am Beispiel der Bedienungsanleitung. Opladen.

Bereiter, C. (1980): Development in Writing. In: Gregg, L. W. & Steinberg, E.R. (Hrsg.): Cognitive Processes in Writing. Hillsdale, 73-93.

Bereiter, C. & Scardamalia, M. (1987): The Psychology of Written Composition. Hillsdale

Bitter Bättig, F. (1999): Die Entwicklung der schriftlichen Erzählfähigkeit vom 4. bis 6. Primarschuljahr. Bern.

Boueke, D., Schülein, F., Büscher, H., Terhorst, E. & Wolf, D. (1995): Wie Kinder erzählen. Untersuchungen zur Erzähltheorie und zur Entwicklung narrativer Fähigkeiten. München.

Böhme, K., Bremerich-Vos, A. & Robitzsch, A. (2008): Aspekte der Kodierung von Schreibaufgaben. In: Granzer, D., Köller, O., Bremerich-Vos, A., van den Heuvel-Panhuizen, M., Reiss, K. & Walther, G. (Hrsg.): Evaluation der Bildungsstandards Deutsch und Mathematik – Erste Ergebnisse. Weinheim/Basel, 290-329.

Bremerich-Vos, A. (1996): Aspekte des Schriftspracherwerbs – Stufentheorien, das ‚Neue‘ und die Lehrer-Schüler-Interaktion. In: Peyer, A. & Portmann, P. R. (Hrsg.): Norm, Moral und Didaktik – Die Linguistik und ihre Schmuddelkinder. Tübingen, 267-290.

Eigler, G., Jechle, T., Merziger, G. & Winter, A. (1990): Wissen und Textproduzieren. Tübingen.

Feilke, H. (1995): Auf dem Weg zum Text. Die Entwicklung der Textkompetenz im Grundschulalter. In: Augst, G. (Hrsg.): Frühes Schreiben. Studien zur Ontogenese der Literalität. Essen, 69-88.

Feilke, H. (1996): Die Entwicklung der Schreibfähigkeiten. In: Günther, H. & Ludwig, O. (Hrsg.): Schrift und Schriftlichkeit. Writing and its Use. Ein interdisziplinäres Handbuch internationaler Forschung, 2. Halbband. Berlin, 1178-1191.

Feilke, H.(2003): Entwicklung schriftlich-konzeptualer Fähigkeiten. In: Bredel, U., Günther, H., Klotz, P., Ossner, J. & Siebert-Ott, G. (2003) (Hrsg.): Didaktik der deutschen Sprache. Ein Handbuch, Bd.1. Paderborn, 178-192.

Feilke, H. (2003): Beschreiben und Beschreibungen. In: Praxis Deutsch, H. 182, 6-15.

Feilke, H. & Schmidlin, R. (2005) (Hrsg.): Literale Textentwicklung. Frankfurt/M.

Feilke, H. (2006): „Der Stand der Dinge“ - Berichten und Berichte. In: Praxis Deutsch, H. 1995, 6-15.

Fix, M. (2006): Texte schreiben. Schreibprozesse im Deutschunterricht. Paderborn.

Fleiss, J. L. & Cohen, J. (1973): The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. In: Educational and Psychological Measurement 33, 613-619.

Flower, L. S. & Hayes, J. R. (1981): A Cognitive Process Theory of Writing. In: College Composition and Communication 32, 1981, 365-387.

Grzesik, J. & Fischer, M. (1985): Was leisten Kriterien bei der Aufsatzbeurteilung? Opladen.

Hayes, J.R. & Flower, L. (1980): Identifying the Organization of Writing Processes. In: Gregg, L.W. & Steinberg, E.R. (Hrsg.): Cognitive Processes in Writing. Hillsdale, 3-30.

Hayes, J.R. (1996): A new Framework for Understanding Cognition and Affect in Writing. In: Levy, C.M. & Ransdell, S. (Hrsg.): The Science of Writing, Mahwah, 1-27.

Harsch, C., Lehmann, R., Neumann, A., Schröder, K. (2007): Schreibfähigkeit. In: Beck, B. & Klieme, E. (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI Ergebnisse, Bd. I. Weinheim, 42-62.

Hug, M. (2001): Aspekte zeitsprachlicher Entwicklung in Schülertexten. Eine Untersuchung im 3., 5. und 7. Schuljahr. Frankfurt/M.

Jechle, T. (1992): Kommunikatives Schreiben. Prozeß und Entwicklung aus der Sicht kognitiver Schreibforschung. Tübingen.

Köller, O. & Baumert, J. (2002): Entwicklung schulischer Leistungen. In: Oerter, R. & Montada, L. (Hrsg.): Entwicklungspsychologie. Weinheim, 756-786.

Langer, I. & Schulz von Thun, F. (2007): Messung komplexer Merkmale in Psychologie und Pädagogik [= Standardwerke aus Psychologie und Pädagogik – Reprints, hrsg. von D. H. Rost]. Münster.

Lehmann, R. H. (1990): Aufsatzbeurteilung – Forschungsstand und empirische Daten. In: Ingenkamp, K. & Jäger, R.S. (Hrsg.): Tests und Trends. Jahrbuch der Pädagogischen Diagnostik, Bd.8, 64-94.

Loomis, S. C. & Bourque, M. L. (2001) (Hrsg.): National Assessment of Educational Progress – Achievent Levels 1992 – 1998 for Writing. Washington, DC.

Neumann, A. & Lehmann, R.H. (2008): Schreiben Deutsch. In: DESI-Konsortium (Hrsg.): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie, Bd.II. Weinheim, 89-103.

Miller, M. (1986): Kollektive Lernprozesse. Studien zur Grundlegung einer soziologischen Lerntheorie. Frankfurt/ M.

National Center for Education Statistics (2003): The Nation´s Report Card: Writing 2002.Washington, DC.

Rehbein, J., (1984): Beschreiben, Berichten, Erzählen. In: Ehlich, K. (Hrsg.): Erzählen in der Schule. Tübingen, 67-124.

Schmidlin, R. (1999): Wie Deutschschweizer Kinder schreiben und erzählen lernen. Textstruktur und Lexik von Kindertexten aus der Deutschschweiz und aus Deutschland. Tübingen.

Schneuwly, B. & Rosat, M. (1995): »Ma chambre« ou comment linéarisier l`espace. Etude ontogénétique de textes écrits. Bulletin de linguistique appliqué 61, 83-100.

Schneuwly, B. (1988): Le langage écrit chez l'enfant. La production des textes informatifs et argumentatifs. Lausanne.

Sieber, P. (2003): Modelle des Schreibprozesses. In: Bredel, U., Günther, H., Klotz, P., Ossner, J. & Siebert-Ott, G. (Hrsg.): Didaktik der deutschen Sprache. Ein Handbuch, Bd.1. Paderborn, 208-223.

Thonke, F., Groß Ophoff, J., Hosenfeld, I. & Isaac, K. (2008). Kriteriengestütze Erfassung von Schreibleistungen im Projekt VERA). In: B. Hofmann & R. Valtin (Hrsg.): Checkpoint Literacy [Tagungsband 2 zum 15. Europäischen Lesekongress der Deutschen Gesellschaft für Lesen und Schreiben] Berlin, 28-35.

Weigle, S. C. (2002): Assessing Writing. Cambridge.

Wirtz, M. & Caspar, F. (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen.

Wolf, D. (2000): Modellbildung im Forschungsbereich sprachliche Sozialisation. Zur Systematik des Erwerbs narrativer, begrifflicher und literaler Fähigkeiten. Frankfurt/M.

Anschrift des Verfassers und der Verfasserin:

Prof. Dr. Albert Bremerich-Vos, Miriam Poßmayer, M.A., Germanistik/ Linguistik/ Sprachdidaktik, Universität Duisburg-Essen, Universitätsstr.12, 45117 Essen

albert.bremerich-vos@uni-due.de
miriam.possmayer@uni-due.de