Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
http://www.didaktik-deutsch.de
23. Jahrgang 2018 – ISSN 1431-4355
Schneider Verlag Hohengehren GmbH

Christopher Sappok / Johanna Fay

PROSODISCHE ASPEKTE VON LESEFLÜSSIGKEIT MESSEN

Evaluation einer Ratingprozedur mit Audioaufnahmen von DrittklässlerInnen.

In: Didaktik Deutsch. Jg. 23. H. 44. S. 61-83.

Christopher Sappok, Johanna Fay

Prosodische Apekte von Leseflüssigkeit messen

Evaluation einer Ratingprozedur mit Audioaufnahmen von DrittklässlerInnen

Zusammenfassung

Der Beitrag stellt ein Instrument zur Erhebung von Lernständen bzgl. prosodischer Aspekte von Leseflüssigkeit vor. Das Design orientiert sich an Methoden der perzeptiven Phonetik. Einbezogen werden Audioaufnahmen von zwölf lesestarken DrittklässlerInnen und eine Ratingprozedur mit 96 RaterInnen sowie akustische Messungen. Hauptanliegen ist die Evaluation der Ratingprozedur und die Erhebung „prosodischer Register" für die einzelnen Kinder. Diese Register ergeben sich aus den Ratingscores und den akustischen Messungen und ermöglichen Aufschluss zur individuellen Verwendung bestimmter Sprechstilmittel. Folgende Ergebnisse können festgehalten werden: Die Ratingprozedur liefert die präzise Quantifizierung einer einfachen, aber aussagekräftigen Facette von Kompetenz im Gebiet des prosodischen Lesens. Der Erfolg der Ratingprozedur wird auf die simple Hörinstruktion, die Verwendung von kurzen Hörstimuli (unter 10 Sekunden) und deren systematische Dekontextualisierung bei der Darbietung zurückgeführt. Der Erfolg der einzelnen Drittklässler wiederum wird auf den intensiven Einsatz prosodischer Stilmittel zurückgeführt, wobei diese durch ihre Heterogenität auffallen. Dieser diagnostische Befund bietet eine Reihe relevanter Anknüpfungspunkte für die Didaktik des lauten Lesens in der Grundschule.

Abstract

This study introduces an instrument for the assessment of skill levels concerning prosodic aspects of reading fluency. The design draws from methods of perceptual phonetics. It involves audiosamples of 12 strong third-grade readers and a rating procedure with 96 raters, as well as acoustic measures. The main purpose is the evaluation of the rating procedure and the acquisition of the children's „prosodic registers". These registers result from the rating scores and the acoustic measures and give insight into the individual use of certain stylistic devices. Main outcomes are: The rating procedure precisely quantifies a plain but relevant facet of competency in the field of prosodic reading. The success of the rating procedure is attributed to the simpleness of the listening instruction, the use of short stimuli (below 10s) and their systematic decontextualisation within the rating session. The success of the individual third-grader on the other hand can be attributed to the intensive use of stylistic devices with noticeable heterogeneity among those. This diagnostic evidence provides a number of reference points for the didactics of oral reading in primary school.

Einführung

Die Leseflüssigkeit von Kindern wird in der Deutschdidaktik immer intensiver und differenzierter erforscht (vgl. Holle 2006, Nix 2011, Rosebrock / Nix 2014, Lauer-Schmaltz et al. 2014). Ein Definitionsrahmen, der den Gegenstand aktuell zusammenfasst, findet sich bei den amerikanischen Autorinnen Kuhn, Schwanenflugel und Meisinger (2010):

> Fluency combines accuracy, automaticity, and oral reading prosody, which, taken together, facilitate the reader's construction of meaning. It is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation. It is a factor in both oral and silent reading that can limit or support comprehension (Kuhn et al. 2010: 240).

Zu unterscheiden ist grob zwischen Aspekten von Leseflüssigkeit beim Erstlesen (ca. 1. bis 3. Klasse) und beim weiterführenden Lesen (ab ca. 3. Klasse). Beim Erstlesen spielt die Leseflüssigkeit als „das Rekodieren (Lautieren) und das Dekodieren (Sinnentnahme des Gelesenen bzw. Lautierten)" eine Rolle (Hasselhorn et al. 2012: 63). Zunehmende Zügigkeit beim lauten Lesen zeigt die Automatisiertheit dieser Prozesse an (vgl. LaBerge / Samuels 1974). Hinzu kommt schrittweise eine adäquate wortübergreifende Gliederung (Herstellung von lokaler Kohärenz, vgl. Rosebrock / Nix 2014: 17). Demgegenüber beschäftigt sich der vorliegende Beitrag mit dem weiterführenden Lesen. Der Übergang kann z. B. dadurch gekennzeichnet sein, dass eine zügige Lesegeschwindigkeit erreicht ist und Verlesungen so rar geworden sind, dass sie das Leseverstehen nicht merklich beeinträchtigen.

Zu prosodischen Aspekten von Leseflüssigkeit gibt es eine Vielzahl von Definitionen. "Reading prosody" ist den Autorinnen der oben aufgeführten Definition zufolge „appropriate expression or intonation coupled with phrasing that allows for the maintenance of meaning" (Kuhn et al. 2010: 233). Über die Richtung des Zusammenhanges zwischen Prosodie und Leseverstehen besteht allerdings bis dato keine Einigkeit („Henne-oder-Ei-Dilemma", Nix 2011: 101). Dass hier aber hochrelevante, wenn auch weiter zu erforschende Zusammenhänge bestehen, ist Konsens:

> In der fluency-Forschung [werden] solche prosodischen Eigenschaften isoliert, die auch für das ausdrucksvolle, sinnkonstituierende und verstehende Lesen von Texten eine nachweisbar wichtige Rolle spielen. Dabei haben sich die Pausengestaltung, die Modulation der Tonhöhe sowie die Lautstärke […] bisher als die aussagekräftigsten Variablen erwiesen (Nix 2011: 96).

Aus funktional-linguistischer Perspektive stellt die Aussagekraft der von Nix aufgeführten Variablen allerdings ein Problem dar, denn die Rekonstruktion von Zusammenhängen zwischen prosodischen Merkmalen und deren Funktionen ist alles andere als trivial. Hierzu verweist Maas (1999: 86) auf eine oft konstatierte „komplexe Interaktion von prosodischen Indikatoren in der Äußerung, die es kaum möglich macht, die abstrakt postulierten Strukturen empirisch zu verifizieren." Dass die

Funktionen von Prosodie dennoch eine wichtige Rolle spielen, wird dann besonders deutlich, wenn man das Hörverstehen einbezieht. Hier kann die prosodische Dimension als differentia specifica gegenüber dem Textverstehen beim leisen Lesen angesehen werden:

> Die Art und Weise, wie Äußerungen mit der Stimme gestaltet werden, gibt zusätzliche Informationen darüber, wie die verbalen Propositionen zu verstehen und auszuwerten sind. Der Einbezug der prosodischen Dimension ist oft entscheidend, um zu einem angemessenen Hörverstehen zu gelangen (Zingg Stamm et al. 2016: 129).

In diesem Sinne gehen wir davon aus, dass die Untersuchung prosodischen Lesens besonders im Hinblick darauf aufschlussreich ist, dass ein Text für einen echten Zuhörer vorgelesen wird.

> Wer vorliest und sich damit richtig auf das Erklären des Textes für seine Hörer einläßt, der versteht auch, was er liest (Ockel 2011: 67).

Wobei ein solches Vorleseszenario in der vorliegenden Untersuchung nur simuliert werden kann. Eine verstehende Prosodie und eine Prosodie der Inszenierung gehen beim Vorlesen Hand in Hand. Prosodisches Lesen in diesem Sinne darf damit als Schnittstelle der didaktisch relevanten Aspekte Leseflüssigkeit, Leseverstehen und 'Lesen für andere' gelten.[1] Mit der Vorstellung eines sehr spezifischen diagnostischen Ansatzes kann der vorliegende Beitrag allerdings nur erste Schritte der Einbeziehung der angesprochenen Vielfalt aufzeigen. Das Hauptaugenmerk liegt dabei nicht nur auf der Erforschung des „Leseverhaltens" von SchülerInnen, sondern in besonderem Maße auch auf der Erforschung des „Hörverhaltens" von RaterInnen.

Im Mittelpunkt des vorliegenden Beitrags steht die experimentelle Evaluation einer innovativen Ratingprozedur, die sich im Vergleich zu vorliegenden Ansätzen (z. B. Pinnell et al. 1995) stärker an den Methoden orientiert, die in der Grundlagenforschung der perzeptiven Phonetik Tradition haben (vgl. Sappok / Arnold 2012a, 2012b; Überblick in Rietveld / Chen 2006). Dabei werden 144 kurze Audioaufnahmen einbezogen, von denen jeder Rater nur eine Auswahl nach nur einem Kriterium zu beurteilen hat, dafür aber in drei Durchgängen in jeweils neu randomisierter Reihenfolge. Diese Schwerpunktsetzung erlaubt es, neben der *Inter*-Raterreliablilität auch die *Intra*-Raterreliabilität zu untersuchen, und stellt neben der Kürze der Ratingstimuli (ca. 2 bis 9 Sekunden) und der Einfachheit der Hörinstruktion einen wichtigen methodischen Unterschied zu den bisher in didaktischen Kontexten verwendeten Ansätzen dar. Weiterhin haben wir einfache akustische Messungen an den

1 In sprachdidaktischen Curricula findet sich diese Schnittstelle verteilt auf zwei Kompetenzbereiche: In dem Bereich „Sprechen und Zuhören / vor anderen sprechen / Texte sinngebend und gestaltend vorlesen" wird eine Vorlesepräsentation dahingehend betrachtet, wie sie vom Vorleser (stimmlich) gestaltet und vom Zuhörer empfunden wird. Im Kompetenzbereich „Lesen – mit Texten und Medien umgehen / über grundlegende Lesefertigkeiten verfügen: flüssig, sinnbezogen […] lesen" wird das Vorlesen zuvorderst als Teil der allgemeinen Lesekompetenz als basale Kulturtechnik verstanden, über die jeder Mensch verfügen soll (vgl. KMK 2003: 10, 13).

Sprachsignalen durchgeführt und Ratings und Messungen im Hinblick auf die akustische Rekonstruierbarkeit von Sprechstilmitteln verglichen.

Die Ratingstimuli stammen aus 12 Audiotexten, gelesen durch die $N_K = 12$ leistungsstärksten von 48 Kindern aus vier dritten Klassen. Mit einer demgegenüber hohen Raterstichprobe von $N_R = 96$ wird ein Schwerpunkt gesetzt, der durch folgende Forschungsfragen bestimmt ist:

1. Inwiefern ist die Strategie der entwickelten Ratingprozedur aus methodologischer Perspektive positiv zu evaluieren?
2. Inwiefern sind die ermittelten Ratingscores und akustischen Messungen kompetenzdiagnostisch aussagekräftig und für die didaktische Perspektive hilfreich?

Der Beitrag gliedert sich in folgende Abschnitte:

Zunächst klären wir in einer Bestandsaufnahme die bisherigen Herangehensweisen an die Erforschung von prosodischem Lesen und die dabei aufgetretenen methodischen Probleme. Weiterhin konkretisieren wir unsere Ziele und Fragen, bevor wir unser eigenes Setting vorstellen und Schritt für Schritt evaluieren und die Ergebnisse interpretieren. Der Beitrag endet mit einer Zusammenfassung und einem Ausblick für die Lautlesedidaktik, der Ausgangspunkte für weitere Forschung und Diskussion aufzeigt.

1. Forschungsstand

1.1 Was verstehen wir unter prosodischen Aspekten und welches prosodische Phänomen untersuchen wir?

Die wichtigste Unterscheidung bei der Wahrnehmung von Sprachschall besteht aus phonetischer Perspektive in der Identifizierung von insgesamt vier Domänen:

> There [are] only four perceptual domains available to the human auditory system for differentiating the elements of speech. These [are] the domains of perceptual quality, duration, pitch and loudness (Laver 1994: 431, Herv. durch uns).

Hier lässt sich QUALITY ansehen als der linguistische Wortlaut, d. h. das, *was* gesagt wird, und Prosodie als Sammelbegriff für die prosodischen Basismerkmale DURATION, PITCH und LOUDNESS, d. h. dafür, *wie* etwas gesagt wird. Auf diese Weise zwischen einer linguistischen Domäne einerseits und drei prosodischen Domänen andererseits zu unterscheiden, ist nicht unproblematisch. Im Bereich der Prosodie kann vielmehr weiter differenziert werden zwischen „sinngemäßer" und damit eng mit linguistischen Merkmalen des Gesagten vernüpfter Prosodie und „sinngestaltender" Prosodie: Letztere ist relativ autonom gegenüber linguistischen Merkmalen, indem sie bedeutend offener ist für die Handhabung durch das Individuum. Derartige Phänomene erfahren zwar steigende Aufmerksamkeit in der einschlägigen Forschung (vgl. Fuchs et al. 2015), die zugrunde liegende Unterscheidung ist aber naiven HörerInnen und damit auch RaterInnen nur sehr schwer bewusst zu machen.

In der vorliegenden Arbeit haben wir uns deshalb auf eine sehr spezifische prosodische „Leseherausforderung“ beschränkt, um die Trennschärfe zu QUALITY zu maximieren. Untersucht wird die prosodische Markierung des Wechsels der sprechenden Figur bei einem von *einem* Sprecher vorgelesenen Dialog. Was wir damit erheben, bezeichnen wir heuristisch als „prosodische Diskursgliederungskompetenz“. Dieses Konstrukt wird als eine besonders aussagekräftige Facette von prosodischem Lesen angesehen, bei dem großer Spielraum für den Einsatz aller o. g. Basismerkmale von Prosodie besteht. Im vorliegenden Kontext sind bzgl. DURATION Pausendauern relevant und Aspekte von PITCH und LOUDNESS werden als akustische Korrelate von Stimme-Verstellen untersucht (Details in Kapitel 4.3). Genauer zu untersuchen bleibt, in welchem Maße das untersuchte Merkmal „prosodische Diskursgliederungskompetenz“ als ein indexikalisches Maß für den Bereich prosodischer Lesekompetenz und damit als ein relevanter Aspekt von weiterführender Leseflüssigkeit angesehen werden könnte (siehe Kapitel 5).

1.2 Bestehende Ansätze zum Assessment prosodischen Lesens und methodische Konsequenzen

Ein populäres und von deutschen AutorInnen (z. B. Rosebrock et al. 2016) vielfach adaptiertes Vorgehen ist die Untersuchung von Ratings nach der sog. „Pinnell-Skala“ (Pinnell et al. 1995). Bei diesem Verfahren werden Lese-Performances in ihrer Gesamtheit von geschulten Ratern hinsichtlich Leseflüssigkeit und dabei auch hinsichtlich prosodischer Aspekte auf einer vierstufigen Skala beurteilt. Auch wenn sich das Verfahren in vielen Kontexten bewährt hat, besteht Ergänzungsbedarf (vgl. Scheerer-Neumann 2015: 90). Kuhn et al. (2010) kritisieren an dieser und ähnlichen Skalen (z. B. Rasinski et al. 2009) mangelnde Präzision bzw. Reliabilität bei hohem Aufwand. So halten Kuhn et al. signalphonetische Messungen an Audiodaten für die soweit überlegene Herangehensweise und sehen Entwicklungsbedarf bei Ratingskalen vor allem in der differenzierteren Formulierung von Bewertungskategorien:

> Whether rating scales will ever have the precision necessary for them to add meaningfully to our measurement of reading fluency beyond text reading speed and accuracy […] is a concern, but it is an avenue that needs to be pursued. Still, we believe that these more complex scales are the general direction in which rating scales of prosody need to go (Kuhn et al. 2010: 236, Herv. durch uns).

Die Rater sollen also einen globalen Höreindruck zu einer Vorlese-Performance noch ausdifferenzierter beurteilen. In Orientierung an den Methoden der Grundlagenforschung zu psychoakustisch relevanten prosodischen Phänomenen erprobt die hier vorgestellte Untersuchung eine Herangehensweise, die dem Vorschlag von Kuhn et al. diametral entgegengesetzt ist, indem sie eine maximal einfache Skala, einen sehr simplen Hörauftrag und kurze Hörstimuli mit Fokus auf nur ein prosodisches Phänomen benutzt. Die Ergänzung von Ratingdaten durch Messungen im Audiosignal halten wir dabei ebenfalls für unverzichtbar.

2. Ziele und Fragen

Unser Ziel ist die Entwicklung eines validen, reliablen Verfahrens zur Erhebung von prosodischem Lesen. Es stellt die Voraussetzung dar für eine anschließende Modellierung der Entwicklung von prosodischer Kompetenz sowie für eine Entwicklung didaktischer Maßnahmen zur Förderung derselben.

Die eingangs formulierten Fragen lassen sich wie folgt konkretisieren:

1.) Zur Evaluation des entwickelten Verfahrens:

a) Gelingt es, bei der Ratingprozedur genuin prosodische Aspekte von anderen Aspekten von Leseflüssigkeit zu separieren?
→ Stichwort: Konstruktvalidität

b) Wie viele Rater sind nötig, um aussagekräftige Ergebnisse zu erzielen?
→ Stichwort: Ökonomie

2.) Zur Anwendung des Verfahrens:

a) Wie einheitlich nehmen verschiedene Rater das prosodische Lesen eines bestimmten Kindes wahr?
→ Stichwort: Reliabilität

b) Wie einheitlich sind die prosodischen Mittel, die verschiedene Kinder benutzen, um eine positive Bewertung zu bekommen?
→ Stichwort: Standardisierbarkeit von Vorlesekompetenz

Separiert betrachtet werden kann die Bearbeitung der Frage 2 b). Sie betrifft die Didaktik des lauten Lesens. Kuhn et al. (2010) stellen hierzu fest:

> If the development of expressiveness is important, what about it is important and what is it important for? If it is essential to reading, we may, indeed, wish to prioritize prosody in our instruction. If it is inessential or emerges without instruction, then we might decide not to (Kuhn et al. 2010: 234).

Durch eine Fundierung und Differenzierung der Auslegung von ‘gutem’ Vorlesen erwarten wir entsprechend Anknüpfungspunkte für didaktisches Handeln.

3. Untersuchung von „Prosodischer Diskursgliederungskompetenz“: Das Verfahren

Zum Stimulustext

Die Kinder haben in Einzelsitzungen dem Versuchsleiter einen vorwiegend dialogischen Text vorgelesen. Dabei war im Stimulus der Wechsel der sprechenden Figur mit nichtsprachlichen Mitteln (Konterfeis, Abb. 1) gekennzeichnet.

Abb. 1: **Stimulustext** (Ausschnitt) zur Elizitation von „prosodischer Diskursgliederung“ sowie Muster für die anschließende Extraktion von 12 überlappenden Doppelsequenzen (DS) mit Wechsel der sprechenden Figur im Zentrum.

Für die Betrachtung prosodischer Aspekte von Leseflüssigkeit ist das ein methodischer Kniff: Welche Figur gerade spricht, muss rein stimmlich gekennzeichnet werden. Der Einsatz prosodischer Mittel ist hier kein „ästhetisches Sahnehäubchen“ – er ist zwingend notwendig, um das Vorgelesene verständlich zu machen. Und gleichzeitig bietet sich reichlich Spielraum für stimmlich-stilistische Vielfalt. Auf der anderen Seite ergibt sich damit ein eigentümliches Leseszenario, zu dem geklärt werden muss, inwiefern es als repräsentativ für das Lesen an sich gelten kann. Aus theoretischer Sicht handelt es sich um eine Form von Redewiedergabe (z. B. Günthner 2002, Forschungsüberblick in Butterworth 2015: 57 ff.) bzw. „reported speech“ (z. B. Klewitz / Couper-Kuhlen 1999), wenn auch nicht in einem spontanen Konversations-, sondern in einem Lesekontext. Um die Eigentümlichkeit dieses Lesekontexts genauer einzuordnen, eignet sich die von Dirscherl / Pafel (2016) vorgeschlagene Vierfelder-Taxonomie der Rede- und Gedankendarstellung. Sie basiert auf den Merkmalen „zitierend“ und „referierend“ einerseits und „explizit“ und „implizit“ andererseits. Die dialogischen Sequenzen des Stimulustextes können hier eindeutig als zitierend eingeordnet werden. Hinsichtlich der Markierung der sprechenden Figuren aber muss im vorliegenden Fall weiter zwischen Lesendem und Zuhörendem unterschieden werden: Für den Lesenden ist die Markierung explizit (Konterfeis), und für den Hörenden ist sie explizit zu machen. Die spezifischen Herausforderungen, die sich damit für den Lesenden ergeben, stellen eine Innovation des vorliegenden Beitrags dar. Um sie zu meistern, ist es hilfreich, über mehr als

einen „prosodic/paralinguistic *habitus*“ (Couper-Kuhlen 1998: 6, Herv. i. Orig.) zu verfügen, was als ausgesprochen hierarchiehoher Aspekt von prosodischem Lesen gelten kann.

Zur Stichprobe Kinder (N_K = 12)

Im ersten Erhebungsschritt wurden 48 Drittklässler (24 Mädchen und 24 Jungen) aufgenommen. Einbezogen wurden dafür vier Klassenverbände aus drei Grundschulen. Aus dem Korpus wurden in einem zweiten Schritt 12 Aufnahmen ausgewählt, die das oberste Leistungsquartil repräsentieren. Für die Auswahl wurden die 48 Aufnahmen nach den NAEP-Kriterien *accuracy, rate* und *fluency* analysiert (NAEP 2005 bzw. Pinnell et al. 1995). Zudem spielte das Kriterium eine Rolle, mit 6 weiblichen und 6 männlichen Kindern weiterzuarbeiten. Somit wurde eine, was die Evaluation betrifft, konservative Bedingung eingeführt: Wenn es gelingt, sogar innerhalb von Level 4 der Pinnell-Skala fein zu differenzieren, sollte dies in breiteren Anwendungskontexten erst recht möglich sein.

In den Sitzungen sollten die Kinder den zuvor unbekannten Text (Abb. 1, insgesamt 205 Wörter) zweimal vorlesen. Für die weitere Arbeit wurde die jeweils zweite Aufnahme („secunda vista“) verwendet. So sollte es den Kindern erleichtert werden, sich auf ein Lesen für andere zu konzentrieren, und den Ratern sollte es durch möglichst stockungsfreie Vorträge erleichtert werden, sich auf genuin prosodische Merkmale zu konzentrieren.

Zur Stichprobe Rater (N_R = 96)

An den Ratingsitzungen nahmen insgesamt 96 Personen teil. Dabei wurden zwei Bedingungen eingerichtet:

Zum einen wurden Einzelsitzungen mit n_{LEHR} = 24 Lehrenden durchgeführt (Hochschullehrende aus den Bereichen Sprachdidaktik/-wissenschaft, Literaturdidaktik/-wissenschaft, Sprechwissenschaft sowie Schullehrkräfte). Diese Rater fungierten als „Experten“, die zwar nicht extra für das Raten geschult wurden, aber mit dem Gegenstand durch ihre berufliche Expertise vertraut sind. Daneben wurden Sammelsitzungen mit n_{STUD} = 72 Lehramtsstudierenden in universitären Computerräumen durchgeführt. Sie fungierten als „naive“ Rater.

Zur Aufbereitung der Audiodaten für die Ratingsitzung

Zu dem Lesestimulus wurde ein Muster entwickelt, wonach die dialogischen Passagen in 12 überlappende Doppelsequenzen eingeteilt wurden („DS“ in Abb. 1). Im Zentrum jeder Doppelsequenz stand eine Position, an der die sprechende Figur wechselt, flankiert von den Äußerungen der wiedergegebenen Gesprächspartner. Die Audioaufnahmen wurden entsprechend geschnitten, und die 144 Token, im Folgenden Sounds genannt, wurden dargeboten mit der Instruktion: „Wie deutlich ist

ein Sprecherwechsel zu hören?“. Die Antwortmöglichkeit bestand in einer vierstufigen Skala [2] mit den Bezeichnungen

"--" = sehr undeutlich = 1

"-" = undeutlich = 2

"+" = deutlich = 3

"++" = sehr deutlich = 4

Eine Ratingsitzung bestand darin, dass ein Rater eine Auswahl von 12 Sounds in randomisierter Reihenfolge in drei Durchgängen einzuschätzen hatte. Die Auswahl wurde dahingehend kontrolliert, dass sie inhaltlich die 12 verschiedenen Doppelsequenzen der Geschichte in randomisierter Reihenfolge umfasste. Weiterhin wurde sie dahingehend kontrolliert, dass jede Doppelsequenz von einem anderen der 12 Kinder gelesen war. Jeder Rater hörte also jedes Kind und jeden Inhalt mit nur einem Sound. Aus der Datenmenge ergeben sich 12 Pakete mit Auswahlen von Sounds (Schattierungen in Abb. 2). Waren die 12 Sounds abgehört, wurden sie in neu randomisierter Reihenfolge ein zweites und dann ein drittes Mal dargeboten. In einer Sitzung wurden also 36 Bewertungen abgegeben.

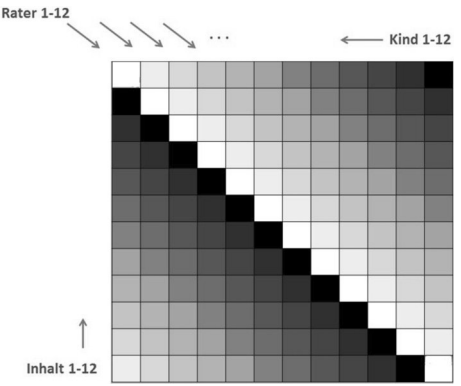


Abb. 2: **Dekontextualisierung der Audiosamples in einem Durchgang:** Jeder Rater ist mit jedem Kind und mit jedem Inhalt nur einmal konfrontiert – er hört ein Teilpaket einer „Schattierung“, z. B. Rater 1 = „weiß“ (oben links).

Für die Umsetzung wurde mittels der Anwendung Experiment MFC von Praat (Boersma / Weenink 2016) ein vollautomatischer Sitzungsablauf für einen Rechnerarbeitsplatz entwickelt, der mit Kopfhörer und Maus ausgestattet war. Für jedes der 36 Ratings konnte der betreffende Sound bei Bedarf bis zu dreimal wiederholt abgehört werden. Eine Sitzung dauerte inklusive einer mündlichen Einführung ca. 10 Minuten.

2 Die Daten sind damit ordinalskaliert, werden aber im Folgenden als intervallskaliert behandelt. Hierzu verweisen wir auf die Diskussion in Rietveld / Chen (2006: 303 f.), die dieses Vorgehen für Kontexte wie den vorliegenden als zulässig erachten.

Mit diesem Setting wird eine maximale Dekontextualisierung des Vorgelesenen erlangt. Natürlich kann ein Rater trotzdem aus dem, was gesagt wird, Wahrscheinlichkeiten hinsichtlich Position und Deutlichkeit des Wechsels der sprechenden Figur ableiten. Unser Weg, diesen Faktor zu kontrollieren, bestand darin, im Text das Ausmaß der linguistischen Abgrenzung zwischen linkem und rechtem Kontext möglichst konstant maximal hoch zu halten (z. B. Frage-Antwort, Frage-Gegenfrage, s. Abb. 1).

Zu den akustischen Messungen

Die 144 Samples wurden parallel zur Ratingprozedur signalphonetisch vermessen, und zwar mit dem Minimalanspruch, möglichst einfach zu erhebende akustische Korrelate von in Frage kommenden Sprechstilmitteln zu erfassen, nämlich Pausendauer sowie die absoluten Differenzen zwischen linker und rechter Sequenz bzgl. PITCH und LOUDNESS. Dahinter steht einmal die Annahme, dass die Pausendauer bei den hier untersuchten lesestarken SuS in einem angemessenen Rahmen kontrolliert gesetzt wird und in diesem Rahmen je länger, desto hörerfreundlicher ist. Die absoluten Tonhöhen- und Lautheitsdifferenzen zwischen linkem und rechtem Kontext wiederum werden als akustische Korrelate von „Stimme-Verstellen" angesehen, und zwar auch wieder je ausgeprägter, desto hörerfreundlicher (Abb. 3 stellt ein eindrückliches Beispiel vor). Für jeden der 144 Sounds wurden so mithilfe der phonetischen Analysesoftware Praat (Boersma / Weenink 2016) folgende Parameter gemessen:

– PAUSE [s] = Pausendauer zwischen „links" und „rechts"
– PITCH.DIFF [Hz] = |(Pitchmittelwert „links") – (Pitchmittelwert „rechts")|
– INT.DIFF [dB] = |(Amplitudenmittelwert „links") – (Amplitudenmittelwert „rechts")|

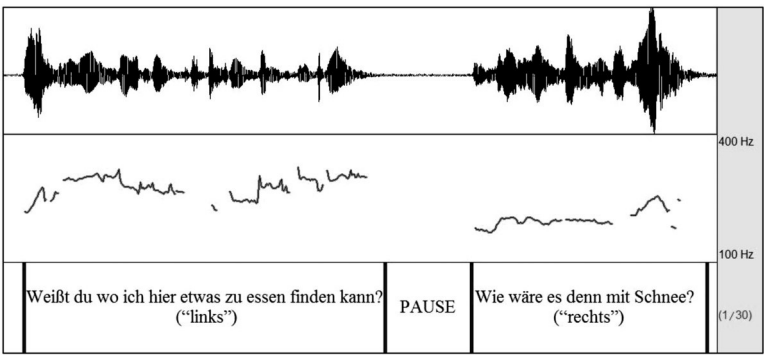


Abb. 3: **Beispiel für eine Doppelsequenz, gesprochen von einem Jungen (Praat-Screenshot)**. Im mittleren Bereich der Abbildung sieht man die automatisch extrahierte Grundfrequenzkontur. Der Mittelwert „links" beträgt 276 Hz, der Mittelwert „rechts" beträgt 203 Hz, woraus sich ein PITCH.DIFF von 73 Hz ergibt. Diesem deutlichen Messwert entspricht auch der Höreindruck. Der Junge verstellt für beide Figuren deutlich seine Stimme.

4. Analysen

Um den Rahmen für die Analysen deutlich zu machen, soll zum Einstieg eine organisatorische Übersicht über die Datenlage gegeben werden (siehe Abb. 4).

Insgesamt sind bezüglich der Ratingprozedur vier Datensätze (X, Y, Z und S) zu unterscheiden. Sie repräsentieren verschiedene Stadien der Zusammenfassung der erhobenen Daten durch Mittelwertbildung. Am Anfang stehen die Rohdaten (X) mit 3.456 Ausprägungen, und am Ende stehen 12 Scores (S) für die untersuchten Kinder. Dazu kommen Datensätze aus den Ergebnissen der signalphonetischen Analysen (unten in Abb. 4).

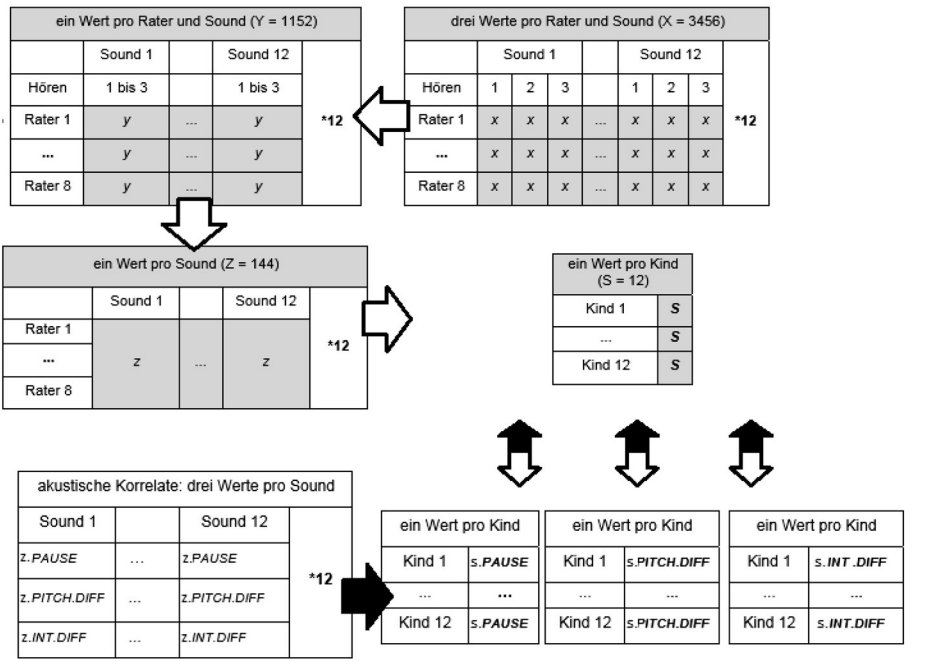


Abb. 4: **Überblick zur Datenlage.** Helle Pfeile: schrittweise Zusammenfassung der Rohdaten (X) zu einem Mittelwert (S) pro Kind. Dunkle Pfeile: Zusammenfassung der verschiedenen akustischen Korrelate zu einem Mittelwert pro Kind. Jeder Schritt erlaubt Analysen zur Robustheit der zusammengefassten Daten: bei Datensatz X: Intra-Rater-Reliabilität[3]; bei Datensatz Y: Inter-Rater-Reliabilität[4] und bei Datensatz Z: Bezug der Ratings zu verschiedenen unabhängigen Variablen[5]. Die S-Daten schließlich erlauben die Untersuchung kindspezifischer Erfolgsbedingungen anhand der akustischen Korrelate (Doppelpfeile; s. Abb. 10)[6].

3 Bezug zu Frage 2 a) und 1 b).
4 Bezug zu Frage 2 a) und 1 b).
5 Bezug zu Frage 1 a).
6 Bezug zu Frage 1 a) und 2 b).

4.1 Konstruktvalidität

Ad. 1 a) Gelingt es, bei der Ratingprozedur sinngestaltend-prosodische Aspekte von anderen Aspekten von Leseflüssigkeit zu separieren?

Um in einem ersten Schritt ein klares Bild dazu zu liefern, welche Rolle es spielt, *wer* spricht, und welche Rolle, *was* gesprochen wird, analysieren wir die Z-Daten. Sie repräsentieren die zentrale Messung der Ratingprozedur: Die Varianz durch Rater ist neutralisiert, es liegt ein Ratingwert pro Sound vor.[7]

Abb. 5 zeigt links die Mittelwerte der Kinder und rechts die Mittelwerte der Doppelsequenzen, beide Male als Ranking dargestellt.

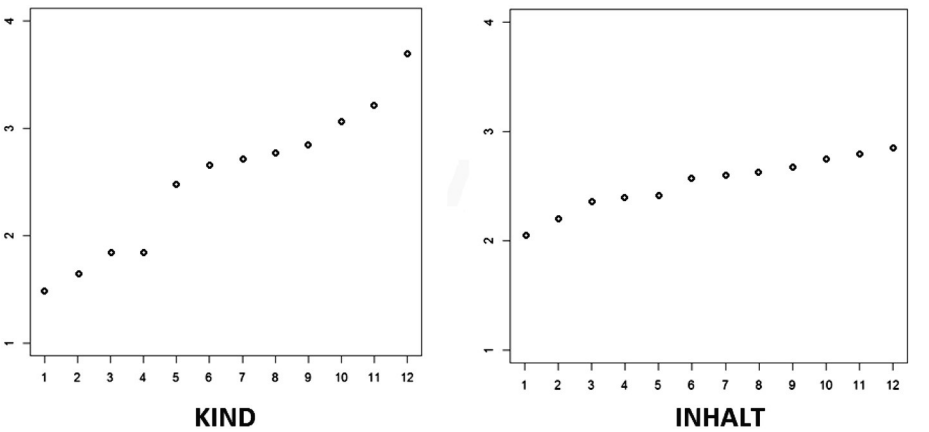


Abb. 5: **„Wie?" vs. „Was?":** Die Gegenüberstellung von kindspezifischen Rating-Mittelwerten (S-Daten) und inhaltsspezifischen Rating-Mittelwerten zeigt eine hohe Differenzierung bei den Kindern, während der Wortlaut fast bedeutungslos erscheint.

So wird auf den ersten Blick ersichtlich, dass der Einfluss der individuellen *Stimmen*, also aller prosodischer Aspekte zusammen, bedeutend größer ist als der Einfluss der einzelnen *Wortlaute*, also aller linguistischer Aspekte zusammen. Um diese Gegenüberstellung inferenzstatistisch auszuwerten, nutzen wir lineare Modelle (R Core Team 2014, vgl. Chambers 1992).

Zunächst soll so geklärt werden, inwiefern die in Abb. 5 repräsentierten Mittelwerte unterschiedliche Populationen repräsentieren (12 Z-Werte je Punkt). Die Analyse Z ~ KIND ergibt adj. $R^2 = .66$ ***. Die Analyse Z ~ INHALT ergibt adj. $R^2 = .01$ (n. s.). Der erste Wert ist als extrem hoch einzuordnen. Dazu kann man sich den konservativen Umstand vor Augen halten, dass jeder Rater mit einem bestimmten Kind einzig anhand nur eines einzigen, wenige Sekunden langen Audiosamples konfrontiert war.

7 Die zehn unkooperativsten Rater wurden ausgeschlossen, sodass alle teilnehmenden Rater einen ICC_{OCS}-Wert von > .3 haben (Erläuterungen dazu in Kap. 4.2).

Auf die Frage *1 a)* „*Gelingt es, bei der Ratingprozedur sinngestaltend-prosodische Aspekte von anderen Aspekten von Leseflüssigkeit zu separieren?*“ können wir daher antworten:

Hier ist die Prozedur erfolgreich. Der individuelle Mittelwert eines Kindes stellt eine *trennscharfe*, in einem insgesamt engen Leistungssegment (oberstes Quartil) *feindifferenzierende* Messung eines *validen* Personenmerkmals („prosodische Diskursgliederungskompetenz“) dar.[8] Der Faktor ‘INHALT’ ist demgegenüber vollkommen unbedeutend.

4.2 Ökonomie und Reliablilität

Ad. 1 b) Wie viele Rater sind nötig, um aussagekräftige Ergebnisse zu erzielen?

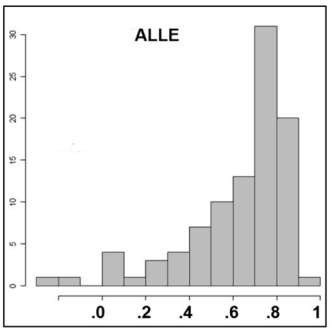
Ad. 2 a) Wie einheitlich nehmen verschiedene Rater das prosodische Lesen eines bestimmten Kindes wahr?

Intra-Rater-Reliabilität: Analyse von Datensatz X

Der hierfür herangezogene Intra-Klassen-Korrelations-Koeffizient nimmt nomalerweise eine Ausprägung zwischen 0 und 1 an und bezeichnet zusammenfassend keine bis vollständige Übereinstimmung in den raterspezifischen x1-, x2-, x3-Reihen. Meist werden ICC-Maße zur Angabe von *Inter*-Rater-Reliabilität verwendet (s. u.), sie können aber auch auf *Intra*-Rater-Reliabilität angewendet werden. Rechnerisch besonders hart „bestraft“ werden Abweichungen bei dem Verfahren, das Gwet zur Bestimmung der Intra-Rater-Reliabilität für ein vergleichbares Setting vorschlägt (vgl. Gwet 2008).[9] Die Verteilung der 96 resultierenden ICC_{ocs}-Ausprägungen ist durch einen Median von .73 gekennzeichnet (siehe Abb. 6, linke Spalte). Die Gruppe LEHR steht mit einem Median von .76 tendenziell besser da als die drei STUD-Gruppen, wobei die Mediane auch bei diesen > .7 ausfallen. Die Daten können als hochgradig reliabel gelten (vgl. Gwet 2008, der eine Schwelle von .5 als akzeptabel bezeichnet).

8 Anspruchsvollere Methoden zur Bestimmung der Konstruktvalidität (konfirmatorische Faktorenanalyse) werden unseres Erachtens erst bei größeren Stichproben sinnvoll.

9 In der von uns verwendeten Softwareumgebung „R“ (R Core Team 2014) wird im Rahmen des „IRR-package“ (Gamer et al. 2014) die betreffende Funktion durch das Setzen von ICC_{ocs} : model = "oneway", type = "consistency", unit = "single" aktiviert.



Rater	**ALLE N = 96**	LEHR n = 24	STUD1 n = 24	STUD2 n = 24	STUD3 n = 24
Median	**.73**	.76	.72	.71	.70

Abb. 6: **Intra-Rater-Reliabilität: Verteilung der ICC_{ocs}-Koeffizienten der einzelnen Rater.** Der links bis in den negativen Bereich auslaufende Teil des Histogramms zeigt, dass einige Rater als unkooperativ eingestuft werden müssen. Der Gruppenvergleich weist auf eine geringfügige Überlegenheit der Gruppe LEHR hin.

Die Intra-Rater-Reliablilität wurde unseres Wissens bei Ratingprozeduren zur Leseflüssigkeit noch nicht untersucht, sodass keine Vergleiche angestellt werden können. Neben dem positiven Ergebnis lässt sich festhalten, dass die Methode gut geeignet ist, unkooperative Rater zu identifizieren und systematisch auszuschließen. Sehr auffällig am Histogramm ist der linke Ausläufer, der sich bis in den negativen Bereich erstreckt. Die sich abzeichnende Schwelle deckt sich mit in ähnlichen Kontexten ermittelten Schwellen, wonach um 10 % bei studentischen Ratern unkooperativ zu sein pflegen und ausgeschlossen werden dürfen (Sappok / Arnold 2012 a, 2012 b).

Inter-Rater-Reliabilität: Analyse von Datensatz Y

Zu jedem Audiosample liegen nach der Zusammenfassung der x-Werte insgesamt 8 y-Werte vor. Mit dem hieraus generierten Mittelwert z wird eine einheitliche Höreigenschaft bzgl. eines Audiosamples vorausgesetzt und geschätzt (siehe Kap. 4.1). Die Robustheit dieser Schätzung wird mit der Inter-Rater-Reliabilität der Y-Daten ermittelt. Hierzu werden die Ratings derjenigen Rater in Gruppen zusammengefasst, die dasselbe Kontingent von 12 Samples abgehört hatten (Sortierung nach „Schattierungen“, s. Abb. 2).[11]

11 Bei dem nun herangezogenen ICC-Koeffizienten (vgl. Gamer 2014) handelt es sich um den Typ ICC_{tca} : model = "twoway", type = "consistency", unit = "average"
Im Unterschied zum ICC_{ocs} für die Intra-Rater-Reliabilität wird bei ICC_{tca} berücksichtigt, dass die Rater austauschbar sein und dass ihre Ratings zu einem Mittelwert zusammengefasst werden sollen (vgl. Hallgren 2012). Die Koeffizienten fallen bei ICC_{tca} weniger streng aus als bei dem oben herangezogenen Kennwert ICC_{ocs}. Aus rechnerisch-methodischen Gründen wurden hier die als unkooperativ identifizierten Rater noch mit einbezogen.

Die resultierenden 12 schattierungsspezifischen ICC_{tca}-Werte sind bis auf eine Ausnahme > .9 und der Median liegt bei .95 (siehe Abb. 7, linke Spalte). Dieses Ergebnis kann als exzellent eingestuft werden.

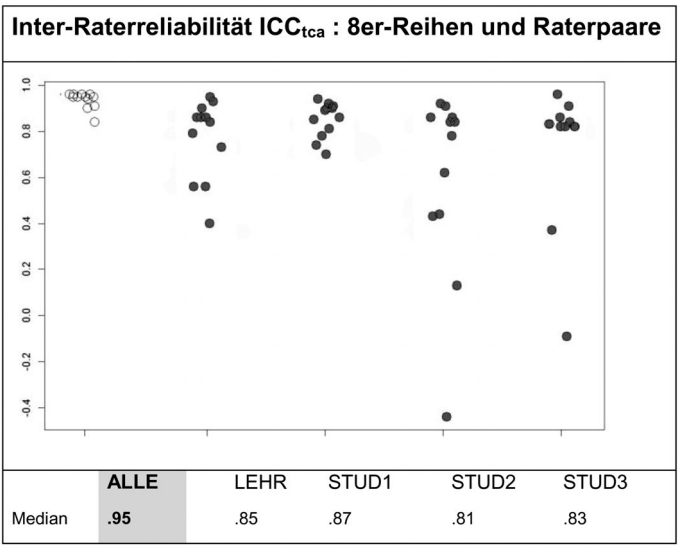


Abb. 7: **Inter-Raterreliabilität:** ICC_{tca}-Koeffizienten bei 8 Ratern pro Sound (leere Kringel) und ausdifferenziert bei 2 Ratern pro Sound (gefüllte Kringel). Bei 8 Ratern ist die Reliabilität durchweg sehr hoch, aber auch bei Raterpaaren bemerkenswert hoch ausgeprägt, wobei 10 Ausreißer (ICC_{tca} < .7) zu verzeichnen sind.

Auf die Frage *2 a) Wie einheitlich nehmen verschiedene Rater das prosodische Lesen eines bestimmten Kindes wahr?* können wir daher antworten:

Unterschiedliche Rater nehmen ein und denselben Sound in höchstem Maße einheitlich wahr.[12]

Das herangezogene Analyseformat kann nun auch dazu verwendet werden, Aufschluss im Hinblick auf eine Ökonomisierung des Verfahrens zu liefern und den Einfluss des Faktors LEHR vs. STUD weiter zu beleuchten (vgl. Abb. 7, gefüllte Punkte rechts). Hierzu werden nicht mehr, wie oben, acht y-Werte pro Sound

12 In der NAEP-Studie von 2002 wird die Inter-Raterreliabilität bei Raterpaaren mit einem ICC-Koeffizienten von .82 angegeben (vgl. NAEP 2005: 50). Ein Vergleich zu den im vorliegenden Kontext ermittelten Werten kann nur unter Vorbehalt erfolgen, da sich aus der gegebenen Beschreibung das genaue Zustandekommen des angegebenen Wertes nicht rekonstruieren lässt. Am ehesten lässt sich der angegebene Wert auf die in Abb. 7 angegebenen Mediane zu Raterpaaren beziehen, die tendenziell leicht höher ausfallen. Wichtig ist dabei die Berücksichtigung der Tatsache, dass sich der NAEP-Wert auf eine repräsentative Stichprobe über das gesamte Leistungsspektrum bezieht, während im vorliegenden Kontext eine Feindifferenzierung bei SuS erfolgt, die im Vorfeld Level 4 der NAEP-Skala zugeordnet wurden. Vor diesem Hintergrund ist das hier vorgestellte Verfahren als reliabler anzusehen.

geprüft, sondern nur zwei (Raterpaare). Es zeigt sich auch bei den ICC_{tca}-Werten der Raterpaare ein überraschend gutes Bild. Der Median ist in allen Gruppen > .8. Die Ausreißer nach unten deuten wieder auf eine Marge von ca. 10% unkooperativen Ratern hin.

Auf die Frage *1 b) Wie viele und welche Rater sind nötig, um aussagekräftige Ergebnisse zu erzielen?* können wir daher antworten:

Da die Rater prosodisches Lesen weitgehend unabhängig von Expertise und Erhebungsbedingungen einheitlich wahrnehmen, könnte in Zukunft eine bedeutend ökonomischere Variante des Verfahrens benutzt werden. Hierzu werden in Kapitel 5 generelle Empfehlungen für die Untersuchung auch anderer prosodischer Aspekte von Leseflüssigkeit mit Methoden aus der perzeptiven Phonetik ausgesprochen.

4.3 Standardisierbarkeit von Vorlesen; prosodische Register

Ad. 2 b) Wie einheitlich oder heterogen sind die prosodischen Mittel, die verschiedene Kinder benutzen, um eine positive Bewertung zu bekommen?

Für die Bearbeitung dieser Frage werden nun die Ergebnisse der Ratingprozedur und die akustischen Messungen der Vorlese-Performances kombiniert (s. Formeln im Kontext von Abb. 3 sowie Überblick in Abbildung 4). So können wir zunächst feststellen, in welchem Verhältnis die Intensität beim Einsatz von Pausendauer, Tonhöhendifferenz und Lautheitsdifferenz zu den Raterurteilen der Deutlichkeit des Wechsels der sprechenden Figur im Allgemeinen steht, und anschließend, welche dieser prosodischen Stilmittel die Kinder im Einzelnen in welchem Maße einsetzen.

In einem ersten Schritt wird anhand der Z-Daten untersucht, inwiefern die im Einzelnen wahrgenommene Diskursgliederung sich über die gemessenen akustischen Korrelate rekonstruieren, d. h. wie viel Varianz sich in den Ratings aller Sounds mittels der akustischen Parameter aufklären lässt. Aufschlussreich ist hierzu das lineare Modell

z.RATING ~ z.PAUSE + z.PITCH.DIFF + z.INT.DIFF

Ein adj. R^2 von .22 zeigt, dass 22% der soundspezifischen Varianz über die akustischen Korrelate aufgeklärt werden ($p < 0{,}001$).[13]

Noch einmal bedeutend klarer wird das Bild mit der Einbeziehung des Faktors KIND durch Mittelwertbildung bei den Ratings und bei den akustischen Daten (S-Daten, s. Abbildung 5). Hierzu werden zunächst die einzelnen Korrelate und dann ein kombiniertes Modell in den Blick genommen (Abb. 8).

13 Zur Bedeutung dieses Befundes ist noch einmal auf Abb. 6 zu verweisen. Dort wurde die überragende Bedeutung des Faktors KIND und damit „Stimme“ belegt. Der hier analysierte Zusammenhang zwischen Akustik und Höreindruck ignoriert diesen Faktor, indem „Stilmittel“ losgelöst von „Stimme“ betrachtet wird.

	adj. R^2
S ~ S.PAUSE	.36 *
S ~ S.PITCH.DIFF	.60 **
S ~ S.INT.DIFF	.50 **
S ~ S.PAUSE + S.PITCH.DIFF + S.INT.DIFF	.78 **

Abb. 8: Kinderspezifische Kennwerte bei den Ratings (S-Daten) und den prosodischen Mitteln. Die auf S-Niveau reduzierten Modelle übertreffen die Erwartungen in ihrer Prädiktionskraft und zeigen Wege auf, Ratings durch Messungen zu ersetzen.

Vor diesem Hintergrund kann als endgültiges Ergebnis die weiterführende Hypothese festgehalten werden, dass sich die über die Ratingprozedur ermittelte prosodische Diskursgliederungskompetenz eines Kindes über folgende Gleichung prädizieren lässt:

S.RATING = 0,63 + 1,92*S.PAUSE + 0,01*S.PITCH.DIFF + 0,33*S.INT.DIFF

Einschränkend muss darauf hingewiesen werden, dass die inferenzstatistische Auswertung bei einer so kleinen Stichprobe von Kindern nur bedingt aussagekräftig ist. Deshalb verzichten wir auf eine Aussage zum Stellenwert der einzelnen prosodischen Mittel in Relation zueinander (z. B. anhand der unterschiedlichen adj. R^2-Werte), sondern beziehen uns nur auf die Kombination der S-Daten mit S.PAUSE + S.PITCH.DIFF + S.INT.DIFF. Hier stellen wir fest: Die Varianzaufklärung ist mit 78 % als ausgesprochen hoch einzustufen, d. h. die gemessenen prosodischen Mittel spielen eine sehr ausgeprägte Rolle im Raterurteil.

Auf die Frage *2 b) Wie einheitlich oder heterogen sind die prosodischen Mittel, die verschiedene Kinder benutzen, um eine positive Bewertung zu bekommen?* können wir daher vorläufig antworten:

Die Kinder, die beim Vorlesen von den Ratern eine gute Bewertung bekommen, setzen prosodische Mittel intensiv und sehr heterogen ein. Aus dem kombinierten Modell lässt sich zusammenfassend die Hypothese ableiten: Je intensiver der Gebrauch von prosodischen Mitteln, desto besser der Ratingscore.

Im Folgenden ergänzen wir die Bearbeitung der Frage, indem wir die Rolle von Heterogenität in der Wahl der prosodischen Mittel deskriptiv visualisieren (vgl. Abb. 9).

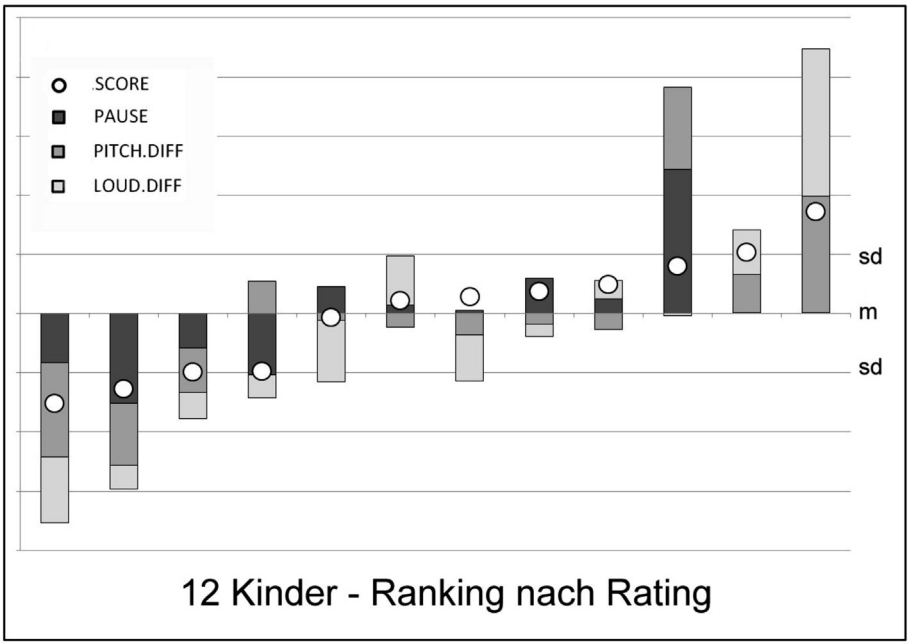


Abb. 9: Benchmarking von prosodischen „Registern" in Abhängigkeit vom Score. Dargestellt ist das Ranking der Kinder nach der Ratingprozedur und als gestapelte Balken die Intensität im Einsatz verschiedener prosodischer Mittel, signalphonetisch gemessen. Alle Parameter wurden z-normalisiert (m = 0, sd = 1).

Die Balken in Abb. 9 stellen in ihrer Zusammensetzung individuelle prosodische „Register" dar. Auslenkung nach unten steht für vergleichsweise defizitären Umgang, Auslenkung nach oben steht für bevorzugten Umgang. Im linken Bereich zeigt sich ein Mangel an intensivem Einsatz von Stilmitteln, der dann sukzessive abgebaut erscheint. Ab einem durchschnittlichen S-Score (0 in Abb. 9; vgl. Kind 5) kann von einem heterogenen Bild gesprochen werden. Im Mittelfeld (Kind 5 bis 9) zeichnet sich ein Flaschenhals ab, in dem Intensität und Scores durchschnittlich ausgeprägt sind, aber unterschiedliche Schwerpunkte erkennen lassen. Im Spitzenfeld (Kind 10 bis 12) ist Heterogenität stark ausgeprägt. Diese Verhältnisse ermöglichen eine Konkretisierung der o. g. Hypothese: Je besser der Score, desto heterogener die Stilmittel, sowohl was ihre *Intensität* als auch was ihre *Zusammensetzung* angeht.

Diese Hypothese sagt aufgrund der sehr ausschnitthaften Messungen nichts darüber aus, welche Stilmittel insgesamt beim prosodischen Lesen in welchem Maße relevant sind. Hierzu ist weitere Forschungsarbeit nötig. Die Frage nach unmittelbaren didaktischen Konsequenzen aus den vorgestellten Befunden kann hier nur anvisiert werden. Dazu und zu weiteren Forschungsperspektiven werden im folgenden Ausblick Überlegungen vorgestellt.

5. Zusammenfassung, Ausblick und didaktische Konsequenzen

Der vorliegende Beitrag hat sich mit der Konzeption eines Erhebungsinstruments zu prosodischen Aspekten von Leseflüssigkeit einem Problem gestellt, das nach aktuellem Forschungsstand als offenes Desiderat gilt (vgl. Scheerer-Neumann 2015: 90). Betrachtet wurde eine Facette von Kompetenz auf dem Gebiet des prosodischen Lesens, die „prosodische Diskursgliederungskompetenz". Offen bleibt mit der vorgestellten Untersuchung, in welchem Maße dieses Merkmal einen Index für eine generellere prosodische Kompetenz und damit einen Aspekt von Leseflüssigkeit darstellt. In unserer laufenden Forschungsarbeit wird eine solche über Dominanzpaarvergleiche (vgl. Bortz et al. 2008) ermittelt und mit differenzierteren akustischen Maßen zu prosodischer Diskursgliederungskompetenz korreliert. Endgültige Ergebnisse stehen noch aus.

Nach der Evaluation der eingeschlagenen, von Methoden der perzeptiven Phonetik geprägten Strategie lässt sich aber eine Reihe verallgemeinernder Empfehlungen aussprechen. Bei der Erhebung von Ratings zu prosodischen Aspekten von fortgeschrittener Leseflüssigkeit verspricht die Einhaltung folgender Kriterien unseren Ergebnissen zufolge die Aussicht auf maximierte Konstruktvalidität, Inter-Raterreliabilität und damit Objektivität bei minimiertem ökonomischen Aufwand:

– Die Rateranzahl sollte zwei- bis dreimal so hoch sein wie die Anzahl der Kinder.
– Die Ratingsitzungen können bei vollautomatisch-computergesteuertem Ablauf in Gruppen mit Ratern ohne Expertise und Training und einer Dauer von insgesamt 10 bis 15 Minuten durchgeführt werden.
– Der Lesestimulus sollte nur eine Sorte von prosodisch zu lösendem Problem 10- bis 15-mal stellen – im Hinblick auf:
– Die Hörinstruktion sollte sich auf eine bestimmte, einfach zu benennende prosodische Herausforderung beschränken, die Hörstimuli sollten kurz sein (max. 10 bis 15 Sekunden) und in zwei bis drei Durchgängen je Rater dargeboten werden – zum Einhören und zur Kontrolle von Intra-Raterreliablilität, Letzteres im Hinblick auf:
– Die Ratings der 10 unreliabelsten Prozent der Rater sollten aus der Analyse ausgeschlossen werden.

Diese Empfehlungen beziehen sich eng auf das spezifische Setting der vorliegenden Untersuchung. Auf allgemeiner Ebene besteht der Hauptunterschied zu gängigen Herangehensweisen in der besonderen Berücksichtigung des Faktors Textstimulus und der gezielten Beschränkung und Konzentration auf einzelne prosodische Herausforderungen. In einer Anschlussuntersuchung wurde mittlerweile ein Längsschnitt-Audiokorpus von ca. 1.000 Aufnahmen erhoben, bei dem eine große Bandbreite entsprechend präparierter Textstimuli verwendet wurde. Das LAUDIO-Korpus („Longitudinal Audio") zeichnet sich außerdem dadurch aus, dass die Lautlese-Entwicklung der beteiligten SuS (N = 46) über drei Jahre von der Grund-

schule bis maximal Jahrgangstufe 7 begleitet wurde. Auch die ersten Analysen dieses Materials weisen in die im Folgenden ausgeführte Richtung.

Neben den methodisch relevanten Befunden konnten wir mit der vorliegenden Untersuchung feststellen, dass Kinder bereits am Ende der dritten Klasse in der Lage sind, prosodische Mittel intensiv und dabei sehr heterogen einzusetzen, wobei in der Tendenz das Vorlesen umso positiver beurteilt wird, je intensiver und heterogener die prosodischen Mittel auftreten. Dieses Ergebnis möchten wir abschließend dem wichtigen didaktischen Anspruch zuführen, Konsequenzen für den (Laut-) Leseunterricht zu entwickeln, wohl wissend, dass die Datenbasis von zwölf Lesekindern keine großen Sprünge zulässt.

Kuhn et al. (2010) fragen in diesem Zusammenhang nach prosodierelevanten Instruktionen bei der Leseförderung (vgl. Zitat in Kap. 2). Bezieht man diese Frage auf unsere 12 Lesekinder, so ist zunächst einmal festzustellen, dass diejenigen, die in Abb. 10 rechts stehen (Kind 7 bis 12) keiner Instruktion zum Lautlesen mehr bedürfen. Sie sind auch nach vergleichendem holistisch-introspektiven Hören ihrer Gesamtaufnahmen guten Gewissens als Repräsentanten von Maximalstandards für die Primarstufe anzusehen; man hört diesen Kindern sehr gerne zu. Die aufgezeigte Heterogenität in ihren prosodischen Registern sollte in diesem Zusammenhang unbedingt als zu erhaltendes Merkmal angesehen werden. Womöglich sind hier Lob und Ermunterung zum häufigen Vorlesen schon Instruktion genug.

Vorleseaufgaben in Schulbüchern sind oft von einigen wenigen und wenig phantasievollen Begriffen geprägt: „Suche dir einen der zwei Texte aus und trage ihn *betont* vor“ (Brettschneider u. a. 2013: 154).

Derartige Arbeitsaufträge suggerieren eine – allseits bekannte – richtige Prosodie beim Vorlesen. Dieselbe Auffassung findet man auch in der Forschungsliteratur, wenn es heißt:

> Die prosodisch korrekte Wiedergabe des Gelesenen wird […] als höchste Anforderungsstufe des flüssigen Lesens angesehen, die die Beherrschung der engeren Teilfertigkeiten auf der Wortebene voraussetzt (Nix 2011: 95 mit Bezug auf Holle 2009; Herv. durch uns).

Und die in der Einleitung zitierten Autoren:

> As children learn to read with good prosody, they come to display an intonational pitch contour increasingly similar to the one used by adults when they read (Kuhn et al. 2010: 234).

Auch hier geht man von einer prosodischen Norm, einem ‘(adult) target model’, aus und nimmt darüber hinaus eine bestimmte Erwerbsreihenfolge an, bei der so etwas wie ‘Prosodielosigkeit’ den Anfang bildet und erwachsenes Vorlesen das Ende. Vorlesedidaktisch ist ein solcher Ansatz vor dem Hintergrund der hier dargelegten Überlegungen und Forschungsergebnisse eher unbefriedigend. Gleichzeitig ist es nicht einfach, geeignetere Instruktionen zum Vorlesen zu formulieren, bedenkt man die heterogenen Möglichkeiten der Umsetzung. „Gutes“ Vorlesen, so viel kann viel-

leicht festgehalten werden, besteht auch darin, seine Zuhörer durch Unerwartetes zum „guten" Zuhören zu animieren.

Insofern plädieren wir bzgl. der fortgeschrittenen Phasen des Leseerwerbs für eine Didaktik, in der das Zuhören beim Vorlesen durch starke Lesekinder einen festen Platz hat (verstärkt auch mithilfe entsprechender Audioaufnahmen). Auf diese Weise wird bei den Leselernern ein Bewusstsein dafür ermöglicht, dass Vorlesen immer bedeutet, *für jemanden* zu lesen. Eine explizite Reflexion prosodischer Stilmittel ist dann auf der Folie dieses 'Zuhörer-Bewusstseins' sinnvoll. Voraussetzung dafür sind unterrichtliche Vorlese-Settings, bei denen nicht nur der Lehrkraft zur Übung oder zu diagnostischen Zwecken, sondern 'echtem' Publikum vorgelesen wird.

Literatur

Beck, Rufus (2009): Gutes Vorlesen ist eben, wenn sich keiner langweilt. Online-Manuskript. Abrufbar unter: https://www.vorlesewettbewerb.de/files/vwb_tipps_beck_1.pdf (zuletzt abgefragt: 11.04.2018).

Boersma, Paul/Weenink, David (2016): Praat: doing phonetics by computer [Computer program]. Version 6.0.13. URL: http://www.praat.org/ (zuletzt abgefragt: 13.06.2017).

Bortz, Jürgen/Lienert, Gustav A./Boehnke, Klaus (2008): Verteilungsfreie Methoden in der Biostatistik. Heidelberg: Springer.

Brettschneider, Stephanie/Clasing, Silke/Diederichs, Saskia/Petersen, Katja (2013): Zebra/Lesebuch 3. Schuljahr. Stuttgart: Klett.

Butterworth, Judith (2015): Redewiedergabeverfahren in der Interaktion. Heidelberg: Winter.

Chambers, John M. (1992): Linear models. In: Chambers, John M./Hastie, Trevor J. (Hg.): Statistical Models in S. Cole. Belmont: Wadsworth & Brooks. S. 95–143.

Couper-Kuhlen, Elizabeth (1998): Coherent Voicing. On Prosody in Conversational Reported Speech. In: Interaction and Linguistic Structure 1. URL: https://kops.uni-konstanz.de/bitstream/handle/123456789/3722/453_1.pdf?sequence=3&isAllowed=y (zuletzt abgefragt 30.03.2018).

Dirscherl, Fabian/Pafel, Jürgen (2016): Die vier Arten der Rede- und Gedankendarstellung: Zwischen Zitieren und Referieren. In: Linguistische Berichte 241. S. 3–47.

Fuchs, Susanne/Pape, Daniel/Petrone, Caterina/Perrier, Pascal (Hg.) (2015): Individual Differences in Speech Production and Perception. Frankfurt a. M.: Peter Lang.

Gamer, Matthias/Lemon, Jim/Fellows, Ian/Singh, Puspendra (2014): irr: Various coefficients of interrater reliability and agreement [software]. Version 0.84 URL: http://CRAN.R-project.org/package=irr (zuletzt abgefragt: 13.06.2017).

Günthner, Susanne (2002): Stimmenvielfalt im Diskurs. Formen der Stilisierung und Ästhetisierung in der Redewiedergabe. In: Gesprächsforschung 3. S. 59–80.

Gwet, Kilem L. (2008): Intrarater Reliability. In: Wiley Encyclopedia of Clinical Trials. Hoboken, NJ. S. 473–485.

Hallgren, Kevin A. (2012): Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. In: Tutor Quant Methods Psychol. 2012; 8(1). S. 23–34.

Hasselhorn, Marcus/Roth, Hans-Joachim/Roßbach, Hans-Günther/Stanat, Petra/Schneider, Wolfgang/Baumert, Jürgen/Becker-Mrotzek, Michael/Kammermeyer, Gisela/Rauschenbach, Thomas/Rothweiler, Monika (2012): Expertise „Bildung durch Sprache und Schrift (BISS)" –

Bund-Länder-Initiative zur Sprachförderung, Sprachdiagnostik und Leseförderung. URL: http://www.biss-sprachbildung.de/pdf/BiSS-Expertise.pdf (zuletzt abgefragt: 31.03.2018).

Holle, Karl (2006): Flüssiges und phrasiertes Lesen (fluency). Lesethereoretische Gundlagen und unterrichtspraktische Hinweise. Unveröffentlichtes Vortragsmanuskript. Lüneburg.

Holle, Karl (2009): Psychologische Lesemodelle und ihre lesedidaktischen Implikationen. In: Garbe, Christine/Holle, Karl/Jesch, Tatjana (Hg.): Texte lesen. Textverstehen, Lesedidaktik, Lesesozialisation. Paderborn: Schöningh. S. 103–165.

Klewitz, Gabriele/Couper-Kuhlen, Elizabeth (1999): QUOTE – UNQUOTE? The role of prosody in the contextualization of reported speech sequences. In: Interaction and Linguistic Structure 12. S. 1–34.

Koriat, Asher/Greenberg, Seth N./Kreiner, Hamutal (2002): The extraction of structure during reading: Evidence from reading prosody. In: Memory & Cognition, 30. S. 270–280.

Kuhn, Melanie/Schwanenflugel, Paula/Meisinger, Elizabeth (2010): Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency. In: Reading Research Quarterly, 45. S. 230–251.

KMK (2003): Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss. Beschluss vom 04.12.2003. Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

LaBerge, David/Samuels, S. Jay (1974): Toward a theory of automatic information processing in reading. In: Cognitive Psychology, 6. S. 293–323.

Lauer-Schmaltz, Marie/Rosebrock, Cornelia/Gold, Andreas (2014): Lautlesetandems in der Grundschule – Bedingungen und Grenzen ihrer Wirksamkeit. In: Didaktik Deutsch 37. S. 44–61.

Laver, John (1994): Principles of Phonetics. Cambridge: CUP.

Maas, Utz (1999): Phonologie. Opladen/Wiesbaden: Westdeutscher Verlag.

NAEP (2005) – i. e.: Daane, Mary C./Campbell, Jay R./Grigg, Wendy S./Goodman, Madeline J./Oranje, Andreas (2005): Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington.

Nix, Daniel (2011): Förderung der Leseflüssigkeit. Theoretische Fundierung und empirische Überprüfung eines kooperativen Lautlese-Verfahrens im Deutschunterricht. Weinheim: Juventa.

Ockel, Eberhard (2011): Leselehre. In: Papst-Weinschenk, Marita (Hg.): Grundlagen der Sprechwissenschaft und Sprecherziehung. München: UTB. S. 81–90.

Pinnell, Gay S./Pikulski, John J./Wixson, Karen K./Campbell, Jay R./Gough, Phillip B./Beatty, Alexandra S. (1995): Listening to Children Read Aloud: Oral Reading Fluency, National Center for Educational Statistics, Washington.

R Core Team (2014): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 3.4.0. URL http://www.R-project.org/ (zuletzt abgefragt: 13.06.2017).

Rasinski, Timothy/Rikli, Andrew/Johnston, Susan (2009): Reading Fluency: More Than Automaticity? More Than a Concern for the Primary Grades? In: Literary Research and Instruction, 48. S. 350–361.

Rietveld, Toni/Chen, Aoju (2006): How to Obtain and Process Perceptual Judgments of Intonational Meaning. In: Sudhoff, Stefan/Lenertová, Denisa/Meyer, Roland/Pappert, Sandra/Augurzky, Petra/Mleinek, Ina/Richter, Nicole/Schließer, Johannes (Hg.): Methods in Empirical Prosody Research. Berlin u. a.: de Gruyter. S. 283–320.

Rosebrock, Cornelia/Nix, Daniel (2014): Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung. Baltmannsweiler: Schneider Verlag Hohengehren.

Rosebrock, Cornelia / Nix, Daniel / Rieckmann, Carola / Gold, Andreas (2016): Leseflüssigkeit fördern. Lautlese-Verfahren für die Primar- und Sekundarstufe. Seelze: Klett / Kallmeyer.

Sappok, Christopher / Arnold, Denis (2012a): On the Normalization of Syllable Prominence Ratings. In: Proceedings of Speech Prosody, 6th International Conference of the International Speech Communication Association, Shanghai, China, May 2012.

Sappok, Christopher / Arnold, Denis (2012b): More on the Normalization of Syllable Prominence Ratings. In: Proceedings of Interspeech, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 2012.

Scheerer-Neumann, Gerheid (2015): Lese-Rechtschreib-Schwäche und Legasthenie. Grundlagen, Diagnostik und Förderung. Stuttgart: Kohlhammer.

Schwanenflugel, Paula J. / Hamilton, Anne Marie / Kuhn, Melanie R. / Wisenbaker, Joseph M. / Stahl, Steven A. (2004): Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. In: Journal of Educational Psychology, 96. S. 119 – 129.

Zingg Stamm, Claudia / Behrens, Ulrike / Käser-Leisibach, Ursula / Krelle, Michael / Weirich, Sebastian (2016): Neue Aufgabenformate für die Messung von Zuhörkompetenzen. In: Keller, Stefan / Reintjes, Christian (Hg.): Aufgaben als Schlüssel zur Kompetenz: Didaktische Herausforderungen, wissenschaftliche Zugänge und empirische Befunde. Münster: Waxmann. S. 129 – 140.

Anschrift der Verfasser:

Dr. Christopher Sappok, Universität zu Köln, Institut für deutsche Sprache und Literatur II, Gronewaldstr. 2, 50931 Köln

csappok@uni-koeln.de

Prof. Dr. Johanna Fay, Europa-Universität Flensburg, Institut für Sprache, Literatur und Medien, Germanistisches Seminar, Auf dem Campus 1, 24919 Flensburg

Johanna.Fay@uni-flensburg.de