AkaTex Working Papers | Nr. 5

Beurteilen von Texten mittels Ratingverfahren im Projekt AkaTex – Methoden – Lena Decker, Ina Kaplan

zum BMBF-Forschungsprojekt

Akademische Textkompetenzen bei Studienanfängern und fortgeschrittenen Studierenden des Lehramtes unter besonderer Berücksichtigung ihrer Startvoraussetzungen

Projektleiterinnen:

Prof. Dr. Gesa Siebert-Ott
Universität Siegen
Philosophische Fakultät
Germanistisches Seminar
Adolf-Reichwein-Straße 2
D-57068 Siegen

PD Dr. Kirsten Schindler
Universität zu Köln
Philosophische Fakultät
Institut für Deutsche Sprache und Literatur II
Innere Kanalstraße 15 | Triforum
D-50823 Köln

Projekthomepage:

http://www.uni-siegen.de/phil/akatex/

KOORDINIERUNGSSTELLE

GEFÖRDERT VOM

AkaTex Working Papers | Nr. 5

Lena Decker, Ina Kaplan

Beurteilen mittels Ratingverfahren im Projekt AkaTex – Methoden

Kontakt:

Universität Siegen
Fakultät I
Germanistisches Seminar
Adolf-Reichwein-Str. 2
D-57068 Siegen

decker@germanistik.uni-siegen.de
kaplan@germanistik.uni-siegen.de

Bibliographische Angabe:

Decker, Lena/ Kaplan, Ina (2014): Beurteilen mittels Ratingverfahren im Projekt AkaTex – Methoden (AkaTex Working Papers, 5). 2., korrigierte Auflage. Siegen und Köln: Universität Siegen und Universität zu Köln.

Die erste Auflage erschien ebenfalls 2014

Beurteilung von Texten mittels Ratingverfahren im Projekt AkaTex - Methoden

Lena Decker, Ina Kaplan

1 Einleitende Bemerkungen

Die Beurteilung der Qualität von Texten mittels Rating- bzw. Beurteilungsverfahren ist sowohl in schulischen als auch in außerschulischen Kontexten weit verbreitet und zur gängigen Praxis geworden (vgl. auch Eckes 2004: 485). Beispiele für den systematischen Einsatz von Ratern stellen die nationalen und internationalen Schulleistungsstudien dar. So wurde in der *DESI-Studie* („Deutsch-Englisch-Schülerleistungen-International") die Qualität der Schülertexte – in diesem Falle Briefe – über ein Ratingverfahren erfasst (vgl. auch Neumann 2007: 86):

> „Die Erfassung der Qualität der Schülertexte erfolgte mehrperspektivisch und zwar einerseits durch die Kodierung von Merkmalen des Inhalts und der Form und andererseits durch die gestufte Beurteilung der sprachlich-textuellen Eigenschaften. Jeder Brief wurde von zwei geschulten Personen beurteilt" (DESI-Konsortium 2006: 7).

Auch die offenen Aufgabenformate des in *PISA 2000* eingesetzten Lesekompetenztests wurden von geschulten Ratern – in Deutschland handelte es sich dabei um Studierende des Lehramts höherer Semester – anhand von detaillierten Codieranweisungen beurteilt (vgl. Baumert/Stanat/Demmrich 2001: 42 und Artelt et al. 2001: 81).

Weitere Beispiele für den Einsatz von Ratern sind Fremdsprachentests wie der „Test Deutsch als Fremdsprache" (TestDaF). Hier werden die Texte des Prüfungsteils „Schriftlicher Ausdruck" von geschulten Beurteilerinnen und Beurteilern bewertet (vgl. www.testdaf.de).

Ihrer großen Beliebtheit zum Trotz sind Ratingverfahren zur Beurteilung der Qualität von Texten – und zur Leistungsmessung überhaupt – mit einer Reihe von Schwierigkeiten verbunden. Das Hauptproblem dieser Verfahren ist die in vielen Fällen unzureichende Übereinstimmung zwischen den Ratern, auch Interraterreliabilität genannt[1]. Als wesentliche Ursache einer geringen Interraterreliabilität ist nach Eckes (2004: 486) die unterschiedlich ausgeprägte Tendenz zur Strenge bzw. Milde anzusehen: Strenge Rater tendieren generell zu niedrigeren Bewertungen und damit zu einer Unterschätzung der Qualität der Texte, wohingegen milde Rater generell höhere Bewertungen vergeben und somit dazu neigen, die Qualität der Texte zu überschätzen (vgl. Eckes 2004: 494). Dies soll im Folgenden an einem Beispiel aus der weiter oben erwähnten TestDaF-Prüfung verdeutlicht werden:

In dem Prüfungsteil „Schriftlicher Ausdruck" wurden in einem Teilrating insgesamt einundzwanzig Texte von zwei Ratern (in diesem Fall Rater 13 und Rater 03) bewertet. Die Ergebnisse dieses Ratings sind in Abb. 1 dargestellt. Wie man erkennen kann, liegen insgesamt fünf Übereinstimmungen (grau unterlegt) und sechzehn Nichtübereinstimmungen vor. Das ergibt eine Übereinstimmungsrate von nur 24% (Kappa= 0.21[2]). Diese unzureichende Interraterreliabilität liegt darin begründet, dass Rater 03 die Texte systematisch höher eingestuft hat als Rater 13. So hat beispielsweide letzterer vier Texte nach TDN 3 – was der Niveaustufe B2.1 des europäischen Referenzrahmens entspricht – eingestuft, ersterer dieselben Texte aber nach TDN 5 (=Niveaustufe C1.2)[3] (vgl. Eckes 2004: 494).

---

[1] Für einen detaillierten Überblick über die unterschiedlichen Methoden zur Bestimmung der Interratereliabilität vgl. Wirtz/Casper (2002).

[2] Zur Verdeutlichung: Eine gute Übereinstimmung erfordert κ-Werte über 0.7 (vgl. Bortz/Döring 2002: 277).

[3] Für eine genaue Niveaustufenübersicht vgl. https://www.testdaf.de/teilnehmer/tn-info_nivea_stufen.php.

	Beurteiler 03				
Beurteiler 13	unter TDN 3	TDN 3	TDN 4	TDN 5	Zeilensumme
unter TDN 3	**1**	**2**	**2**		5
TDN 3			**6**	**4**	10
TDN 4			**2**	**2**	4
TDN 5				**2**	2
Spaltensumme	1	2	10	8	21

Abbildung 1: Bewertungen der Rater 13 und 03 in der TestDaF-Prüfung (Prüfungsteil „Schriftlicher Ausdruck“) nach Eckes (2004: 493)

Eine unzureichende Interraterreliabilität – wie sie gerade im Beispiel verdeutlicht wurde – ist sehr problematisch, da dies bedeutet, dass die Ergebnisse nicht vom Rater unabhängig sind. Aus diesem Grund wurde die Übereinstimmungsrate im Ratingverfahren des Projektes AkaTex regelmäßig überprüft, um ggf. sofort mit einer Nachschulung beginnen zu können. Im Folgenden soll genauer auf dieses Ratingverfahren mit seinen beiden Teilratings eingegangen werden.

2 Das Ratingverfahren des Projektes AkaTex

Das Ratingverfahren des Projektes AkaTex zur Beurteilung der Diskursreferate[4] bestand aus zwei Teilen: Begonnen wurde mit dem Raten der Kategorie „Wissenschaftliches Formulieren“ (Teilrating 1), anschließend wurde der „Fachliche Gehalt und die Argumentation“ der Diskursreferate beurteilt (Teilrating 2).

Für die Beurteilung der insgesamt 193 Texte standen zwei Rater, welche zuvor im Rahmen einer Raterschulung angeleitet wurden, zur Verfügung. Bei diesen handelt es sich um wissenschaftliche Mitarbeiter der Fakultät I (Germanistisches Seminar) der Universität Siegen, welche seit mehreren Semestern Lehrveranstaltungen im Bereich Deutsch als Zweitsprache leiten und somit auch erfahren im Beurteilen von studentischen Texten sind.

Grundlage für alle Textbeurteilungen bildeten zum einen Ratingbögen, welche die einzelnen Bewertungskriterien enthalten, und zum anderen Kodierhandbücher – auch Manuals genannt –, welche die Anweisungen für die Rater beinhalten. Im Folgenden soll detailliert auf die zwei Teilratings eingegangen werden. Im Fokus stehen dabei die Ratingbögen und die Kodierhandbücher.

2.1 Teilrating 1: Wissenschaftliches Formulieren (Schwerpunkt: diskursstrukturierende Prozeduren)

Im ersten Teilrating „Wissenschaftliches Formulieren“ (Schwerpunkt diskursstrukturierende Prozeduren[5]) sollten die Rater beurteilen, ob die Studierenden die Fähigkeiten besitzen, diskursstrukturierende Prozeduren kontextuell passend zu verwenden. Bevor genauer auf den Ratingbogen mit den einzelnen Kriterien und auf das Kodierhandbuch mit den Informationen zu den einzelnen Werten eingegangen wird, soll im Folgenden zunächst kurz beschrieben werden, wie die zwei Rater im Rahmen der Raterschulung[6] auf das Beurteilen der Diskursreferate im Hinblick auf das „Wissenschaftli-

[4] Für detaillierte Informationen zu dieser Textform vgl. Working Paper 4.
[5] Unter „diskursstrukturierenden Prozeduren“ verstehen wir diejenigen wissenschaftlichen Textprozeduren, die man in einem besonderen Maße benötigt, um an wissenschaftlichen Diskursen schreibend partizipieren zu können, nämlich intertextuelle Prozeduren wie „Nach X“ oder „X dagegen kritisiert“ und Positionierungsprozeduren wie „Meines Erachtens“.
[6] Diese Raterschulung wurde von mir und einer weiteren Mitarbeiterin des Projektes „AkaTex“, Ina Kaplan, durchgeführt.

che Formulieren“ vorbereitet wurden. Der Ablauf der Schulung orientierte sich an dem Vorschlag von Neumann (2007: 87).

Um die Rater in ihre Arbeit einzuführen, wurden diese zunächst mit dem Forschungsprojekt AkaTex vertraut gemacht, d.h. es wurden der theoretische Rahmen, die Zielsetzungen und die zentralen Forschungsfragen des Projekts erläutert. Anschließend ging es um die Textform „Diskursreferat“: Mithilfe einer zuvor erstellten Graphik wurden den Ratern die zentralen Anforderungen, welche diese Textform an die Studierenden stellt, dargelegt. Danach wurden den Ratern die wesentlichen Informationen zu dem zu ratenden Textkorpus übermittelt. Neben der Anzahl der Texte ging es hier vor allem die Besprechung der drei Aufgabenstellungen und der dazugehörigen Primärtexte. Kern der Schulung bildete die detaillierte Erörterung zum einen des Ratingbogens und zum anderen des Kodierhandbuchs. Zur Verdeutlichung wurde gemeinsam ein Diskursreferat, welches nicht zum zu ratenden Textkorpus gehört, im Hinblick auf das „Wissenschaftliche Formulieren“ beurteilt, d.h. es wurde für dieses Diskursreferat auf Basis des Kodierhandbuchs ein Ratingbogen ausgefüllt. Den Abschluss der Raterschulung bildete ein Proberating: Die zwei Rater bekamen jeweils zehn (5x Prätest, 5x Posttest) identische, nicht zum Korpus gehörende Diskursreferate, welche sie innerhalb einer Woche beurteilen sollten. Fragen, die während dieses Proberatings aufkamen, wurden entweder persönlich oder per Mail geklärt. Die Beurteilungen der Rater wurden dann mit „Mustercodierungen“ abgeglichen. So konnten Probleme ermittelt werden, die anschließend im Rahmen einer individuellen Nachschulung gelöst wurden.

Kommen wir nun zum Ratingbogen der Kategorie „Wissenschaftliches Formulieren“ mit dem Schwerpunkt auf den diskurstrukturierenden Prozeduren (vgl. Abb. 2). Dieser enthält zum einen Kriterien, welche sich auf die „Domänentypik“ der diskursstrukturierenden Prozeduren beziehen. Die Kriterien 1a und 2a betreffen dabei die intertextuellen Prozeduren, Kriterium 3a die Positionierungsprozeduren.

Ein Beispiel für die Verwendung von domänenuntypischen Positionierungsprozeduren stellt der folgende Ausschnitt aus einem Diskursreferat dar:

> „**Meiner Meinung nach** hat Stern mit dieser Aussage Recht, da es wirklich genügend Kinder gibt, die in ihrer Familie als einziges Mitglied die deutsche Sprache beherrschen. Somit wäre eine „Deutschpflicht“ für diese Kinder eventuell sehr wichtig und hilfreich bei der Entwicklung ihrer deutschen Sprachkenntnisse. Abschließend kann ich weder Prof. L. Hoffmann mit seiner These, dass Mehrsprachigkeit ein kostbares gut und jeder Förderung würdig sei, noch der „Deutschpflicht“ auf dem Schulhof“ eindeutig zustimmen. **Ich denke**, dass beide Ansichten Aspekte enthalten, die vertretbar sind [...].

Die Fähigkeit, diskursstrukturierende Prozeduren kontextuell passend einzusetzen, wird jedoch nicht nur daran festgemacht, ob diese Prozeduren domänentypisch sind, sondern auch daran, wie häufig sie gebraucht werden (vgl. auch Steinhoff 2007). Aus diesem Grund enthält der Ratingbogen zum anderen Kriterien, welche sich auf die Varianz der diskursstrukturierenden Prozeduren beziehen: Die Kriterien 1b und 2b betreffen dabei die Varianz der intertextuellen Prozeduren, Kriterium 3b die Varianz der Positionierungsprozeduren.

Ein Beispiel für eine fehlende Varianz der diskursstrukturierenden Prozeduren stellt der folgende Ausschnitt aus einem Diskursreferat dar:

> **Spinners Meinung (2004, S.5ff.) nach** hat die Kultusministerkonferenz (kurz: KMK) für die einzelnen Lernbereiche eine große Anzahl von Standards erstellt. **Ihm nach** bietet die Standardisierung den Vorteil der Überprüfung des Gelernten. Parallel dazu kritisiert er jedoch auch die daraus resultierenden Nachteile. **Spinners Ansicht nach** werden nicht nur die Bildung, sondern auch die Schülerinnen und Schüler standardisiert. Für die Lehrkräfte ergibt sich daraus die Konsequenz, dass die Möglichkeit verschiedener Interpretationen nicht zustande kommt. Daraus wiederum folgt eine Reduzierung der Schülerkompetenzen, da sie den Aufgabenstellungen gerecht gemacht

> werden. **Spinners Erklärungen nach** wird den Schülerinnen und Schülern die Entfaltung ihrer Individualität im Unterricht vorenthalten. Er (Spinner 2004, S.10) verweist darauf, dass sich die Schulbuchverlage der neuen bildungspolitischen Sichtweise angepasst, und somit auch neue Terminologien geschaffen haben. **Nach Spinner** ist das neue Leitbilder Bildungspolitik „[…] der planende, seine Verhaltensweise kontrollierende, metakognitiv sich steuernde, sich seiner Zielsetzungen bewusste und über einsetzbare Strategien verfügende Mensch" (Spinner 2004, S.10). […] **Spinners Meinung (2004, S.13) nach** werden den Lernsituationen ihre Komplexität entzogen, während die Subjektivität verfällt und sich das selbstständige Lernen daher zu einem angeleiteten Training entwickelt.

Kriterium 4 bildet ein „Globalurteil" des gesamten Diskursreferates hinsichtlich des „Wissenschaftlichen Formulierens": Hier geht es nicht nur um die diskursstrukturierenden Prozeduren, sondern vor allem darum, ob in dem Diskursreferat andere wissenschaftliche Textprozeduren – beispielsweise Gliederungsprozeduren wie „In diesem Text soll zunächst X thematisiert werden" – verwendet werden.

VERBUNDPARTNER: UNIVERSITÄT SIEGEN Universität zu Köln
KOORDINIERUNG: KoKoHs
GEFÖRDERT VOM: Bundesministerium für Bildung und Forschung

akatex

Rating der Kategorie *Wissenschaftliches Formulieren*

Code des Diskursreferats: N.J1.M1.1.A2.0.P2
Rater: A

1 a) Werden die wiedergegebenen Positionen und Argumente durch angemessene wissenschaftssprachliche Mittel gekennzeichnet?	1	2	3	4
1 b) Variieren die sprachlichen Mittel, die zur Kennzeichnung der wiedergegebenen Positionen und Argumente verwendet werden?	1	2	3	4
2 a) Werden die wiedergegeben Positionen und Argumente durch angemessene wissenschaftssprachliche Mittel vergleichend aufeinander bezogen?	1	2	3	4
2 b) Variieren die sprachlichen Mittel, die zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente verwendet werden?	1	2	3	4
3 a) Wird die eigene Position durch angemessene wissenschaftssprachliche Mittel gekennzeichnet?	1	2	3	4
3 b) Variieren die sprachlichen Mittel, die zur Kennzeichnung der eigenen Positionen verwendet werden?	1	2	3	4
4 Wird eine „Alltägliche Wissenschaftssprache“ im Sinne von Ehlich verwendet?	1	2	3	4

Kommentar:

1

Abbildung 2: Beispiel Ratingbogen "Wissenschaftliches Formulieren“

Ob Studierende die Fähigkeit besitzen, diskursstrukturierende Prozeduren kontextuell passend einzusetzen, wurde also an insgesamt sechs Kriterien überprüft. Zusätzlich wurde ein siebtes Kriterium als holistisches Urteil des gesamten Diskursreferates bezüglich des „Wissenschaftlichen Formulierens" mit in das Rating aufgenommen[7].

Die sieben Kriterien sollten von den Ratern anhand einer vierstufigen Ratingskala mit aufsteigenden Werten von eins bis vier bearbeitet werden. Die Wahl einer vierstufigen Ratingskala liegt zum einen darin begründet, dass den Ratern ein „Skalenmittelpunkt" vorenthalten werden sollte: Gibt man eine Mittelkategorie vor, so läuft man Gefahr, dass diese von den Ratern vermehrt als „Fluchtkategorie" genutzt wird. Man spricht bei diesem Verhalten auch von einer „Tendenz zur Mitte" (vgl. auch Porst 2008: 81, Raab-Steiner/Benesch 2012: 63). Zum anderen wurde sich für eine vierstufige Ratingskala entschieden, da es bei einer fünf- oder sechsstufigen Skala erheblich schwieriger ist, die einzelnen Werte hinreichend voneinander abzugrenzen.

Damit die Rater die Kriterien bearbeiten konnten, musste ihnen natürlich mitgeteilt werden, was sich hinter den einzelnen Werten „verbirgt". Aus diesem Grund erhielten sie – wie bereits erwähnt – ein Kodierhandbuch mit den Informationen zu den einzelnen Werten. Im Folgenden sollen diese Informationen mit Hilfe der folgenden Tabellen dargestellt werden.

I. Kriterium 1a

Kriterium 1a bezieht sich auf die Frage, ob die wiedergegebenen Positionen und Argumente durch angemessene wissenschaftssprachliche Mittel gekennzeichnet werden.	
Wert 1	Ein Diskursreferat erhält den Wert 1, wenn während des gesamten Textes nicht deutlich wird „wer spricht", d.h. die wiedergegebenen Positionen und Argumente **nicht bzw. kaum** durch sprachliche Mittel gekennzeichnet werden.
Wert 2	Im Unterschied zu Wert 1 werden bei Wert 2 die wiedergegebenen Positionen und Argumente im Diskursreferat zwar durch sprachliche Mittel gekennzeichnet, diese sind aber **überwiegend wissenschaftsuntypisch** (beispielsweise „X meint", „X findet") **und/ oder** weisen **überwiegend Formulierungsbrüche** auf (beispielsweise „X gibt Beispiele auf", „X verweist darauf hin").
Wert 3	Wert 3 liegt vor, wenn im Diskursreferat die wiedergegebenen Positionen und Argumente **überwiegend** durch **wissenschaftstypische Mittel** (beispielsweise „X stellt die These auf", „X weist darauf hin") gekennzeichnet werden. Wissenschaftsuntypische Mittel und/ oder Formulierungsbrüche kommen bei diesem Wert aber dennoch – wenn auch nur vereinzelt – vor.
Wert 4	Hier werden die wiedergegebenen Positionen und Argumente im Diskursreferat **durchgehend** durch **wissenschaftstypische Mittel** gekennzeichnet.

[7] Inwiefern diese Kriterien die interessierende Fähigkeit angemessen operationalisieren, wurde im Dialog mit Experten aus der Schreibforschung validiert (Expertenvalidierung). Im Ergebnis wurde die Operationalisierung als schlüssig und vollständig bewertet.

II. Kriterium 1 b

Kriterium 1b bezieht sich auf die Frage, ob die sprachlichen Mittel zur Kennzeichnung der wiedergegebenen Positionen und Argumente variieren.	
Wert 1	Ein Diskursreferat erhält den Wert 1, wenn die sprachlichen Mittel zur Kennzeichnung der wiedergegeben Positionen und Argumente **nicht bzw. kaum** variieren, also ein bestimmtes Muster rekurrent verwendet wird.
Wert 2	Bei Wert 2 variieren die sprachlichen Mittel zur Kennzeichnung der wiedergegebenen Positionen und Argumente **teilweise**, d.h. bestimmte sprachliche Mittel werden noch immer übermäßig gebraucht, daneben finden sich im Diskursreferat aber auch einige Mittel, welche nicht wiederholend verwendet werden.
Wert 3	Hier liegt eine **überwiegende** Variation der sprachlichen Mittel zur Kennzeichnung der wiedergegebenen Positionen und Argumente vor: Die wiederholende Verwendung bestimmter sprachlicher Mittel stellt bei diesem Wert eine Ausnahme dar, kommt aber dennoch vor
Wert 4	Diskursreferate, welche den Wert 4 erhalten, zeichnen sich durch eine **durchgehende** Variation der sprachlichen Mittel zur Kennzeichnung der wiedergegebenen Positionen und Argumente aus, d.h. es wird kein sprachliches Mittel übermäßig gebraucht.

III. Kriterium 2a

Kriterium 2a bezieht sich auf die Frage, ob die wiedergegebenen Positionen und Argumente durch angemessene wissenschaftssprachliche Mittel vergleichend aufeinander bezogen werden.	
Wert 1	Bei diesem niedrigsten Wert werden die wiedergegebenen Positionen und Argumente im Diskursreferat **nicht bzw. kaum** durch sprachliche Mittel vergleichend aufeinander bezogen. Die Schreibaufgabe „Diskursreferat“ ist hier also additiv-referierend gelöst worden.
Wert 2	Diskursreferate mit dem Wert 2 zeichnen sich dadurch aus, dass die wiedergegebenen Positionen und Argumente **überwiegend** durch **wissenschaftsuntypische Mittel** (beispielsweise „X bemängelt an Y, dass...“) **und/ oder überwiegend** durch sprachliche Mittel mit **Formulierungsbrüchen** (beispielsweise „X kritisiert auf Y, dass...“) vergleichend aufeinander bezogen werden.
Wert 3	Wert 3 wird dann vergeben, wenn im Diskursreferat die wiedergegebenen Positionen und Argumente **überwiegend** durch **wissenschaftstypische Mittel** (beispielsweise „Ähnlich wie X vertritt auch Y die These, dass...“) vergleichend aufeinander bezogen werden. Wissenschaftsuntypische Mittel oder sprachliche Mittel mit Formulierungsbrüchen werden bei diesem Wert demnach nur vereinzelt verwendet.
Wert 4	Ein Diskursreferat erhält den Wert 4, wenn die wie-

	dergegebenen Positionen und Argumente **durchgehend** durch **wissenschaftstypische Mittel** vergleichend aufeinander bezogen werden. Wissenschaftsuntypische Mittel oder sprachliche Mittel mit Formulierungsbrüchen kommen bei diesem Wert also nicht vor.

IV. Kriterium 2b

Kriterium 2b bezieht sich auf die Varianz der sprachlichen Mittel, die zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente verwendet werden. Beurteilbar ist dieses Kriterium erst ab drei verwendeter sprachlicher Mittel, da sonst das Ergebnis verzerrt werden würde. Ein Beispiel: In einem Diskursreferat findet sich nur ein einziges sprachliches Mittel, welches die Positionen und Argumente vergleichend aufeinander bezieht. Streng genommen müsste in diesem Fall der beste Wert – also Wert 4 – vergeben werden, da kein sprachliches Mittel rekurrent verwendet wird.	
Wert 1	Ein Diskursreferat erhält den Wert 1, wenn die sprachlichen Mittel zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente **nicht bzw. kaum** variieren.
Wert 2	Bei Wert 2 variieren die sprachlichen Mittel zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente **teilweise**, d.h. bestimmte sprachliche Mittel werden noch immer übermäßig gebraucht, daneben finden sich im Diskursreferat aber auch einige Mittel, welche nicht rekurrent verwendet werden.
Wert 3	Hier liegt eine **überwiegende** Variation der sprachlichen Mittel zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente vor, d.h. die wiederholende Verwendung bestimmter sprachlicher Mittel stellt bei diesem Wert eine Ausnahme dar, kommt aber dennoch vor.
Wert 4	Diskursreferate, welche den Wert 4 erhalten, zeichnen sich durch eine **durchgehende** Variation der sprachlichen Mittel zum vergleichenden Aufeinanderbeziehen der Positionen und Argumente aus, d.h. es wird kein sprachliches Mittel übermäßig gebraucht.

V. Kriterium 3a

Kriterium 3a bezieht sich auf die Frage, ob die eigene Position durch angemessene wissenschaftssprachliche Mittel gekennzeichnet wird.	
Wert 1	Dieser Wert wird dann vergeben, wenn während des gesamten Textes nicht deutlich wird „wer spricht“, d.h. die eigene Position **nicht bzw. kaum** durch sprachliche Mittel gekennzeichnet wird.
Wert 2	Im Unterschied zu Wert 1 wird bei Wert 2 die eigene Position im Diskursreferat zwar durch sprachliche Mittel gekennzeichnet, diese sind aber **überwiegend wissenschaftsuntypisch** (beispielsweise „Ich finde“, „Meiner Meinung nach“) **und/ oder** weisen **überwiegend Formulierungsbrüche** auf (beispielsweise „Meines Erachtens nach“)

Wert 3	Bei Diskursreferaten, welche den Wert 3 erhalten, wird die eigene Position **überwiegend** durch **wissenschaftstypische Mittel** (beispielsweise „Ich vertrete die Position, dass...“) gekennzeichnet. Wissenschaftsuntypische Mittel oder sprachliche Mittel mit Formulierungsbrüchen kommen bei diesem Wert nur selten vor.
Wert 4	Hier wird die eigene Position im Diskursreferat **durchgehend** durch **wissenschaftstypische Mittel** (beispielsweise „meines Erachtens“ bzw. „m.E.“) gekennzeichnet.

VI. Kriterium 3b

Kriterium 3b bezieht sich auf die Frage, ob die sprachlichen Mittel zur Kennzeichnung der eigenen Position variieren. Wie das Kriterium 2b ist auch dieses Kriterium erst ab drei verwendeter sprachlicher Mittel beurteilbar.	
Wert 1	Ein Diskursreferat erhält den Wert 1, wenn die sprachlichen Mittel zur Kennzeichnung der eigenen Position **nicht bzw. kaum** variieren, also ein bestimmtes Muster rekurrent verwendet wird.
Wert 2	Bei Wert 2 variieren die sprachlichen Mittel zur Kennzeichnung der eigenen Position **teilweise**, d.h. bestimmte sprachliche Mittel werden noch immer übermäßig gebraucht, daneben finden sich im Diskursreferat aber auch einige Mittel, welche nicht wiederholend verwendet werden.
Wert 3	Hier liegt eine **überwiegende** Variation der sprachlichen Mittel zur Kennzeichnung der eigenen Position vor: Die wiederholende Verwendung bestimmter sprachlicher Mittel stellt bei diesem Wert eine Ausnahme dar, kommt aber dennoch vor.
Wert 4	Diskursreferate, welche den Wert 4 erhalten, zeichnen sich durch eine **durchgehende** Variation der sprachlichen Mittel zur Kennzeichnung der eigenen Position aus, d.h. es wird kein sprachliches Mittel übermäßig gebraucht.

VII. Kriterium 4

Kriterium 4 stellt ein Globalurteil des gesamten Diskursreferates bezüglich des „Wissenschaftlichen Formulierens“ dar und bezieht sich auf die Frage, ob eine „Alltägliche Wissenschaftssprache“ im Sinne von Ehlich (1999) verwendet wird.	
Wert 1	Dieser Wert wird dann vergeben, wenn im Diskursreferat **keine bzw. kaum** eine „Alltägliche Wissenschaftssprache“ verwendet wird, sondern eine Alltags- bzw. Umgangssprache.
Wert 2	Hier wird **überwiegend** eine Alltags- bzw. Umgangssprache **und/ oder** eine „Alltägliche Wissenschaftssprache“, die Formulierungsbrüche aufweist (beispielsweise „Folgend werde ich die These von Hoffmann diskutieren“), verwendet.
Wert 3	Ein Diskursreferat erhält den Wert 3, wenn **überwiegend** eine „Alltägliche Wissenschaftssprache“ im

	Sinne von Ehlich verwendet wird. Alltags- bzw. Umgangssprache und eine „Alltägliche Wissenschaftssprache“, welche Formulierungsbrüche aufweist, kommen bei diesem Wert selten vor.
Wert 4	Diskursreferate, welche den Wert 4 erhalten, zeichnen sich dadurch aus, dass **durchgehend** eine „Alltägliche Wissenschaftssprache“ verwendet wird.

Neben dieser Beschreibung der einzelnen Werte erhielten die Rater zu jedem Wert Ankerbeispiele und zusätzlich eine – unter anderem in Anlehnung an Jakobs (1999: 94) – erstellte Liste mit domänentypischen, domänenuntypischen und falsch verwendeten diskursstrukturierenden Prozeduren. So wurde die Chance erhöht, dass die Rater bezüglich der „Domänentypik“ einer diskursstrukturierenden Prozedur zu einer übereinstimmenden Einschätzung gelangen.

Die Interratereliabilität wurde zum ersten Mal nach vierzig gerateten Texten pro Rater mit Hilfe von SPSS überprüft. Da die Werte zufriedenstellend bis gut waren, wurde das Rating ohne Zwischenschulung fortgesetzt.

Zusammenfassend lässt sich der Ablauf des Ratingprozesses für das Teilrating 1 tabellarisch folgendermaßen darstellen (vgl. auch Neumann 2007: 87):

Zeitpunkt	Arbeitsschritte
ca. zwei Wochen vor Beginn des Ratings - Raterschulung	- kurze Einführung in das Forschungsprojekt „AkaTex“ und in die Textform Diskursreferat - Informationen zum zu ratenden Textkorpus (Anzahl, Aufgabenstellungen, Primärtexte) - Besprechung der Bewertungskriterien und des Kodierhandbuchs - Gemeinsames Raten eines Beispieltextes - Klärung von Fragen - Proberating an Beispieltexten durch die Rater (10 Texte) - Abgleich mit Musterkodierung - Individuelle Nachschulung der Rater, Klärung von Fragen, die während des Proberatings aufgetreten sind
Beginn des Ratings	- ständiges Monitoring durch Klärung spezifischer Fragen der Rater - Rating, bis vierzig Texte pro Rater vorliegen
Erste Überprüfung der Interratereliabilität	- Überprüfung der Interratereliabilität mit Hilfe von SPSS
Fortsetzung des Ratings	- ständiges Monitoring durch Klärung spezifischer Fragen der Rater

Abbildung 3: Ablauf des Ratingprozesses Teilstudie 1

2.2 Teilrating 2: Fachlicher Gehalt und Argumentation

Im zweiten Teilrating sollten die Rater nicht den Sprachgebrauch, sondern den fachlichen Gehalt der Diskursreferate mit Hilfe von insgesamt sechs Kriterien beurteilen: Die Kriterien 1a und 1b beziehen sich auf die innere Struktur, die Kriterien 2a und 2b auf die Relevanz und die sachliche Richtigkeit der dargestellten Positionen und Argumente und Kriterium 3 betrifft die fachliche Begründetheit der eigenen Position. Kriterium 4 stellt – in Anlehnung an das Kriterium 4 der Kategorie „Wissenschaftliches Formulieren“ – ein Globalurteil des gesamten Diskursreferates bezüglich des fachlichen Gehalts und der Argumentation dar (vgl. Abb. 4).

VERBUNDPARTNER: UNIVERSITÄT SIEGEN Universität zu Köln

KOORDINIERUNG: KoKoHs

GEFÖRDERT VOM: Bundesministerium für Bildung und Forschung

akatex

Rating der Kategorie *Fachlicher Gehalt und Argumentation*

Code des Diskursreferats: N.J1.M1.1.A2.0.P2
Rater: A

1 a) Ist der Text inhaltlich sinnvoll strukturiert (im Hinblick auf die gestellte Aufgabe)?	1	2	3	4
1 b) Weist der Text eine diskursive Struktur auf?	1	2	3	4
2 a) Enthält der Text die für das Thema / die Aufgabenstellung relevanten Positionen und Argumente der Autoren?	1	2	3	4
2 b) Gibt der Text diese Positionen und Argumente sachlich richtig wieder?	1	2	3	4
3 Bezieht sich die eigene Position auf relevante Argumente der Autoren?	1	2	3	4
4 Gesamteindruck / Globalurteil	1	2	3	4

Kommentar:

1

Abbildung 4: Beispiel Ratingbogen "Fachlicher Gehalt und Argumentation"

Die Frage, ob diese Kriterien das interessierende Merkmal – also den fachlichen Gehalt und die Argumentation der Diskursreferate – angemessen operationalisieren, wurde im Dialog mit Experten validiert. Auch diese Operationalisierung wurde als schlüssig und vollständig bewertet.

Ebenso wie die sieben Kriterien zum „Wissenschaftlichen Formulieren" sollten auch die Kriterien zum fachlichen Gehalt und zur Argumentation von den Ratern anhand einer vierstufigen Ratingskala mit aufsteigenden Werten von eins bis vier bearbeitet werden. Grundlage war auch hier ein Kodierhandbuch.

Im Folgenden sollen die einzelnen Werte mit Hilfe der folgenden Tabellen erläutert werden.

I. Kriterium 1a

Kriterium 1a bezieht sich auf die Frage, ob der Text im Hinblick auf die gestellte Aufgabe inhaltlich sinnvoll strukturiert ist (Kohärenz). Eine inhaltlich sinnvolle Strukturierung liegt zum einen dann vor, wenn die Reihenfolge der Darstellung der Positionen und Argumente logisch ist, also die Chronologie des wissenschaftlichen Diskurses deutlich gemacht wird. So wäre es beispielsweise nicht logisch, bei dem Posttest mit der Position von Michael Becker-Mrotzek zu beginnen, da dessen Position zum wissenschaftlichen Diskurs nur vor dem Hintergrund der Positionen von Peter Sieber und vor allem Kaspar H. Spinner verständlich wird. Zum anderen zeigt sich eine inhaltlich sinnvolle Strukturierung an dem Vorhandensein strukturierender Elemente (auf der Makroebene *Einleitung und Schluss*, auf der Mikroebene *wissenschaftliche Textprozeduren* wie „Im Folgenden werde ich..." oder „Zusammenfassend lässt sich festhalten, dass...").	
Wert 1	Dieser Wert wird dann vergeben, wenn die Reihenfolge der Darstellung der Positionen und Argumente **nicht** logisch ist **und keine bzw. kaum** (1x) sinnvolle strukturierende Elemente verwendet werden.
Wert 2	Ein Diskursreferat enthält den Wert 2, wenn die Reihenfolge der Darstellung der Positionen und Argumente nicht logisch ist **und/ oder nur teilweise** sinnvolle strukturierende Elemente verwendet werden. **Teilweise** bedeutet, dass das Diskursreferat beispielsweise zwar eine Einleitung aufweist, die den Leser in das Thema einführt und ihm einen Einblick in die Struktur der Arbeit ermöglicht, aber keinen Schlussteil, der z.B. die wesentlichen Positionen zusammenfasst und einen Bezug zur Einleitung herstellt.
Wert 3	Hier ist die Reihenfolge der Darstellung der Positionen und Argumente logisch **und** es werden **überwiegend** sinnvolle strukturierende Elemente verwendet. **Überwiegend** bedeutet, dass das Diskursreferat beispielsweise eine Einleitung und einen Schlussteil aufweist, aber keine bzw. nur wenige strukturierende Elemente auf der Mikroebene.
Wert 4	Bei diesem höchsten Wert ist die Reihenfolge der Darstellung der Positionen und Argumente logisch **und** es werden **durchgehend** sinnvolle strukturierende Elemente verwendet. **Durchgehend** bedeutet, dass das Diskursreferat sowohl eine Einleitung als auch einen Schlussteil und sinnvolle strukturierende Elemente auf der Mikroebene aufweist.

II. Kriterium 1b

Kriterium 1b bezieht sich auf die Frage, ob das Diskursreferat eine diskursive Struktur aufweist. Diskursiv ist eine Struktur dann, wenn im Diskursreferat ein wissenschaftlicher Diskurs dargestellt wird, also zunächst das Thema bzw. der Gegenstand dieses Diskurses deutlich gemacht wird und anschließend die verschiedenen Positionen und Argumente zu diesem Diskurs vergleichend dargestellt werden.	
Wert 1	Hier wird das Thema des Diskurses **nicht** deutlich gemacht **und** die verschiedenen Positionen und Argumente zu diesem Diskus werden **nicht bzw. kaum** (1x) vergleichend dargestellt. Bei diesem Wert weist das Diskursreferat demnach keine diskursive Struktur auf, sondern es liegen unverknüpfte Zusammenfassungen der einzelnen Texte vor (additiv-referierend).
Wert 2	Dieser Wert wird dann vergeben, wenn das Thema des wissenschaftlichen Diskurses nicht deutlich gemacht wird **und/ oder** die verschiedenen Positionen und Argumente zu diesem Diskus nur **teilweise** vergleichend dargestellt werden. **Teilweise** bedeutet, dass nur wenige Bezüge (max. 3) zwischen den Positionen und Argumenten hergestellt werden.
Wert 3	Bei diesem Wert wird das Thema des wissenschaftlichen Diskurses deutlich gemacht **und** es werden die verschiedenen Positionen und Argumente zu diesem Diskus **überwiegend** vergleichend dargestellt. **Überwiegend** bedeutet, dass mehr Bezüge als bei Wert 2 hergestellt werden, zur angemessenen Darstellung eines wissenschaftlichen Diskurses aber noch einige wenige Bezüge fehlen.
Wert 4	Ein Diskursreferat erhält den Wert 4, wenn das Thema des wissenschaftlichen Diskurses deutlich gemacht wird **und** die verschiedenen Positionen und Argumente zu diesem Diskus **durchgehend** vergleichend dargestellt werden. Hier liegt also eine angemessene Darstellung eines wissenschaftlichen Diskurses vor.

III. Kriterium 2a

Kriterium 2a bezieht sich auf die Frage, ob der Text die für das Thema/ die Aufgabenstellung relevanten Positionen und Argumente der Autoren enthält. Diese wurden zuvor in einem „Erwartungshorizont“, welcher den Rater als Orientierung ausgehändigt wurde, zusammengestellt.	
Wert 1	Bei diesem Wert enthält das Diskursreferat **keine bzw. kaum** (1x) für das Thema/ die Aufgabenstellung relevante(n) Positionen und Argumente. Das bedeutet, es werden fast ausschließlich Positionen und Argumente dargestellt, welche sich nicht auf das Thema/ die Aufgabenstellung beziehen **oder** es werden **alle** Positionen und Argumente dargestellt, d.h. die relevanten Positionen und Argumente werden nicht von den weniger relevanten getrennt.
Wert 2	Hier enthält das Diskursreferat **teilweise** die für das Thema/ die Aufgabenstellung relevanten Positionen und Argumente. **Teilweise** bedeutet, dass es neben relevanten Positionen und Argumenten auch nicht

	relevante Positionen und Argumente enthält **oder keine bzw. kaum** (1x) nicht relevante(n) Positionen und Argumente, **aber zu wenig** relevante Positionen und Argumente.
Wert 3	Dieser Wert wird dann vergeben, wenn das Diskursreferat **überwiegend** die für das Thema/ die Aufgabenstellung relevanten Positionen und Argumente enthält. **Überwiegend** bedeutet, dass es **keine bzw. kaum** nicht relevante(n) Positionen und Argumente enthält **und nur wenige** relevante Positionen und Argumente nicht enthält.
Wert 4	Hier enthält das Diskursreferat **keine** nicht relevanten und **alle** relevanten Positionen und Argumente.

IV. Kriterium 2b

Kriterium 2b bezieht sich auf die Frage, ob das Diskursreferat die für die Aufgabenstellung/ das Thema relevanten Positionen und Argumente sachlich richtig wiedergibt. Die sachliche Richtigkeit der Darstellung konnten die Rater dem jeweiligen Erwartungshorizont und/ oder den entsprechenden Primärtexten entnehmen.	
Wert 1	Das Diskursreferat erhält diesen Wert, wenn es die Positionen und Argumente **nicht bzw. kaum** (1x) sachlich richtig wiedergibt. Das bedeutet, dass die Positionen und Argumente fast ausschließlich sachlich falsch wiedergeben werden **oder** die Positionen und Argumente den Autoren fast ausschließlich falsch zugeordnet werden.
Wert 2	Hier werden die Positionen und Argumente **teilweise** sachlich richtig wiedergegeben. **Teilweise** bedeutet, dass das Diskursreferat neben sachlich richtig wiedergegebenen **auch sachlich falsch** wiedergegebene Positionen und Argumente enthält **und/ oder** die Positionen und Argumente im Diskursreferat den Autoren **in einigen Fällen falsch** zugeordnet werden.
Wert 3	Bei Wert 3 werden die Positionen und Argumente **überwiegend** sachlich richtig wiedergegeben. **Überwiegend** bedeutet, dass im Diskursreferat **keine bzw. kaum** sachlich falsche(n) Positionen und Argumente vorkommen, sondern die Positionen und Argumente **weitgehend** sachlich richtig wiedergegeben und den Autoren richtig zugeordnet werden. Maximal zwei Abweichungen sind bei diesem Wert zulässig.
Wert 4	Wert 4 wird dann vergeben, wenn im Diskursreferat **alle** Positionen und Argumente sachlich richtig wiedergegeben und **alle** Positionen und Argumente den Autoren richtig zugeordnet werden.

V. Kriterium 3

Kriterium 3 bezieht sich auf die Frage, ob sich die eigene Position auf relevante Positionen und Argumente der Autoren bezieht und somit fachlich begründet ist. Hier geht es darum, ob der Verfasser sich am jeweiligen wissenschaftlichen Diskurs beteiligt, also an geeigneten Stellen im Diskursreferat oder am Ende die dargestellten Positionen und Argumente der Autoren aufgreift und sich beispielsweise der Position eines Autors anschließt bzw. sich von dieser distanziert.	
Wert 1	Wert 1 wird dann vergeben, wenn sich die eigene

	Position **nicht bzw. kaum** (1x) auf relevante Positionen und Argumente der Autoren bezieht. Das bedeutet, dass die eigene Position entweder nichts mit dem Thema des Diskurses zu tun hat **oder** zwar auf das Thema des Diskurses Bezug nimmt, nicht aber auf die relevanten Positionen und Argumente der Autoren (stattdessen: Verfasser berichtet von eigenen Erfahrungen und Erlebnissen).
Wert 2	Hier bezieht sich die eigene Position **teilweise** auf relevante Positionen und Argumente der Autoren. **Teilweise** bedeutet, dass neben Bezügen zu den Positionen und Argumenten der Autoren auch von eigenen Erfahrungen und Erlebnissen berichtet wird.
Wert 3	Ein Diskursreferat erhält den Wert 3, wenn sich die eigene Position **überwiegend** auf relevante Positionen und Argumente der Autoren bezieht. **Überwiegend** bedeutet, dass **nicht bzw. kaum** (1x) von eigenen Erfahrungen und Erlebnissen berichtet wird, sondern sich auf Positionen und Argumente der Autoren bezogen wird, diese Bezüge aber noch zu wenig sind.
Wert 4	Dieser höchste Wert wird dann vergeben, wenn sich die eigene Position **durchgehend** auf relevante Positionen und Argumente der Autoren bezieht, also fachlich begründet dargelegt wird.

VI. Kriterium 4

Kriterium 4 stellt – in Anlehnung an das Kriterium 4 der Kategorie „Wissenschaftliches Formulieren“ – ein Globalurteil des gesamten Diskursreferates bezüglich des „Fachlichen Gehalts und der Argumentation“ dar und orientiert sich an den Schulnoten.	
Wert 1	Wert 1 bedeutet, dass der fachliche Gehalt und die Argumentation des Diskursreferates **ungenügend bis mangelhaft** sind.
Wert 2	Wert 2 wird dann vergeben, wenn der fachliche Gehalt und die Argumentation des Diskursreferates **ausreichend bis befriedigend** sind.
Wert 3	Hier sind der fachliche Gehalt und die Argumentation des Diskursreferates **gut**.
Wert 4	Ein Diskursreferat erhält den höchsten Wert, wenn der fachliche Gehalt und die Argumentation des Textes **sehr gut** sind.

Der Ablauf des Ratingprozesses für das Teilrating 2 gleicht dem des Teilratings 1 (vgl. Abb. 3). Lediglich die Einführung in das Projekt „AkaTex“ und in die Textform „Diskursreferat“ sowie die Informationen zu dem zu ratenden Textkorpus wurden weggelassen.

Literatur

Artelt, Cordula/Stanat, Petra/Schneider, Wolfgang/Schiefele, Ulrike (2001): Lesekompetenz: Testkonzeption und Ergebnisse. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. S. 69-137.

Baumert, Jürgen/Stanat, Petra/Demmrich, Anke (2001): PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. S. 15-68.

Bortz, Jürgen/Döring, Nicola (2002): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 3., überarbeitete Aufl. Berlin et al.: Springer.

DESI-Konsortium (Hrsg.) (2006): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch Englisch Schülerleistungen International (DESI). Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung.

Eckes, Thomas (2004): Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In: Wolff, Armin/ Ostermann, Torsten/ Chloster, Christoph (Hrsg.): Integration durch Sprache. Regensburg: FaDaF (Materialien Deutsch als Fremdsprache, Bd. 73), S. 485-518.

Neumann, Astrid (2007): Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen. Münster et al.: Waxmann.

Porst, Rolf (2008): Fragebogen. Ein Arbeitsbuch. Wiesbaden: VS Verlag für Sozialwissenschaften.

Raab-Steiner, Elisabeth/Benesch, Michael (2012): Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung. 3., aktualisierte und überarbeitete Auflage. Wien: facultas wuv.

Steinhoff, Thorsten (2007): Wissenschaftliche Textkompetenz. Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten. Tübingen: Niemeyer.

Wirtz, Markus/Casper, Franz (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen et al.: Hogrefe.

Abbildungsverzeichnis