Bewertungskriterien schriftlicher Lernerproduktionen B2 und C1 und ihre Validität

Dissertation, eingereicht von Anna Chita
bei der Philosophisch-Historischen Fakultät
der Universität Augsburg
zur Erlangung der Würde einer Dr. phil.

Dezember 2008

Bewertungskriterien schriftlicher Lernerproduktionen B2 und C1 und ihre Validität

Dissertation,
eingereicht von Anna Chita (geboren in Korbach, Hessen)
bei der Philosophisch-Historischen Fakultät
der Universität Augsburg
zur Erlangung der Würde einer Dr. phil.

Betreuung der Arbeit und Erstgutachter: Prof. Dr. Hans Jürgen Heringer
Zweitgutachter: Prof. Dr. Evelyn Frey
Tag der mündlichen Prüfung: 04. Juni 2009

Inhaltsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abs. = Absatz

APA = American Psychological Association

Bd. = Band

bzw. = beziehungsweise

ca. = circa

d.h. = das heißt

DAF = Deutsch als Fremdsprache

DESI = Deutsch-Englisch-Schülerleistungen International

DIALANG = Beurteilungssystem für Sprachlernende

e.V. = eingetragener Verein

FACETS = Multifacetten-Rasch-Modell

GDS = Großes Deutsches Sprachdiplom

GER = Gemeinsamer Europäischer Referenzrahmen

ggf. = gegebenenfalls

GI = Goethe-Institut

GmbH = Gesellschaft mit beschränkter Haftung

HV = Hörverstehen

i. d. R. = in der Regel

IELTS = International English Language Testing System

IRT = Item Response Theorie

Kap. = Kapitel

KDS = Kleines Deutsches Sprachdiplom

KLP = Lernlehrplan

KMK = Kultusministerkonferenz

KPG = Das griechische Staatszertifikat für Sprachen

KTT = Klassische Testtheorie

L1 = Erstsprache

L2 = Zweit- bzw. Fremdsprache

lat. = lateinisch

LV = Leseverstehen

MA = Mündlicher Ausdruck

NRW = Nordrhein-Westfalen

s. = siehe

SA = Schriftlicher Ausdruck

SI = Schriftliche Interaktion

SP = Schriftliche Produktion

StADaF = Ständige Arbeitsgruppe Deutsch als Fremdsprache

TDN = TestDaF-Niveau

Telc = The European Language Certificates

TestDaF = Test Deutsch als Fremdsprache

TOEFL = Test of English as a Foreign Language

u.Ä. = und Ähnliches

u.a. = unter anderem/und andere

vgl. = Vergleich

vs. = versus

z.B. = zum Beispiel

z.T. = zum Teil

Danksagung

Für das Gelingen dieser Arbeit trug eine Vielzahl von Personen bei. Zunächst möchte ich mich ganz herzlich bei Herrn Prof. Dr. Heringer bedanken, der sich durch sein Vertrauen in mich bereit erklärte mein Doktorvater zu sein, obwohl er wusste, dass es mir zu dieser Zeit unmöglich war nach Deutschland zu ziehen. Das Flugzeug war für mich das ständige Fortbewegungsmittel, um mich zwischen Griechenland und Deutschland und zwischen Beruf und Dissertation bewegen zu können.

Weiterhin möchte ich Frau Prof. Dr. Evelyn Frey danken, die sich einverstanden erklärte die Zweitkorrektur dieser Arbeit zu übernehmen. Sie ermöglichte mir im Februar 2008 im Goethe-Institut in München meine Dissertationsthematik vorzustellen und mit den Experten der Prüfungsabteilung darüber zu sprechen und zu diskutieren. Frau Freys Initiative ehrte mich besonders und erwies sich für mich als sehr fruchtbar.

Ein großer Dank gilt ebenso meiner lieben Freundin und Kollegin Frau Dr. Marieluise Ernst-Vidalis, die mir in dieser Zeit immer mit Rat und Tat zur Seite stand, sich nie gegen die endlosen fachlichen Diskussionen bis in die Nacht hinein beschwerte und zudem noch für mein leibliches und seelisches Wohl sorgte.

Ein immer offenes Ohr und viel Geduld und Verständnis hatte auch Andreas Bülow während der endlosen Telefonate. Er erlebte zum einen meine Impulsivität, was ihn oft zum Schmunzeln brachte und zum anderen war er stets bemüht mir auf jegliche meiner Fragen Hilfestellungen und Antworten zu geben, indem er mir oft andere Perspektiven aufzeigte.

Bedanken möchte ich mich ebenfalls bei meinem langjährigen Freund Oliver, bei dem ich in Augsburg während meiner Aufenthalte immer ein Zuhause haben konnte. Auch meinem Freund Günter sei gedankt, der in meinem Auftrag das Organisatorische und Logistische dieser Dissertation übernahm.

Abschließend möchte ich mich aus ganzem Herzen bei meiner Familie und all denjenigen bedanken, die mir in dieser Zeit sehr viel Verständnis entgegenbrachten, Geduld mit mir hatten und mich entbehren mussten.

Und zu guter letzt danke ich meinem Körper, der in dieser Zeit mit wenig Schlaf, Überstunden, und unregelmäßigen Essenszeiten ausgekommen ist und nicht nach Erholung bat - etwas, das ihm jetzt aber zu Recht zusteht.

Ioannina, Dezember 2008

Vorwort

Aus Gründen der sprachlichen Vereinfachung werde ich in dieser Arbeit Ausdrücke wie „standardisierte Prüfungen“ und „Test“, „Bewerter“ und „Rater“, „Lerner“ und „Prüfungsteilnehmer“, um einige zu nennen, synonym verwenden. Die kursiven Auszeichnungen verwende ich hauptsächlich für die Definition der Deskriptoren der einzelnen Bewertungskriterien. Aus stilistischen Gründen benutze ich nur die generischen Formen, z.B. beinhaltet der Begriff „Lerner“ sowohl Lernerinnen als auch Lerner - die Emanzipation der Frauen soll dabei nicht in Frage gestellt werden.

1 Einleitung

Meine Auseinandersetzung mit dem Thema der vorliegenden Dissertation „Bewertungskriterien schriftlicher Lernerproduktionen B2/C1 und ihre Validität“ hatte bereits in meiner Praxis als DaF-Lehrerin begonnen. Hierbei kommt man in Kontakt mit Sprachanfängern und hat dabei die Aufgabe die deutsche Sprache zu vermitteln. Mit einer Vielfalt von verschiedenen Methoden und Strategien wird im Unterricht der Versuch unternommen, den Deutschlernern die von ihnen „ausgewählte“ Zielsprache so nah wie möglich zu bringen.

Bezüglich des erworbenen Sprachstandes durch den Unterricht muss zwischen einer Bewertung innerhalb des Klassenzimmers und der im Rahmen standardisierter Prüfungen[1] unterschieden werden. Im Klassenzimmer geht es dem Lehrer in erster Linie darum, ein Feedback für seinen geleisteten Unterricht zu erhalten, die entsprechenden Fortschritte der Gruppe und einzelne Leistungen in Bezug dazu zu erkennen, damit er im Curriculum oder mit den von ihm gesetzten Lehrzielen fortfahren kann. Dies kann als Normorientierung verstanden werden. In standardisierten Prüfungen wird in der Regel ein Kriterium aufgestellt, das für die Sprachzertifizierung erfüllt sein muss – in diesem Fall spricht man von Kriteriumsorientierung. Der Unterricht wird somit testorientiert.

In fast allen Ländern werden Sprachzertifizierungsprüfungen indirekt von den jeweiligen Bildungssystemen abverlangt. Man muss nachweisen, was man kann, denn ohne Zertifizierung wird keinerlei Kompetenz zugesprochen. Anliegen und Ziel des Lehrers ist in jedem Falle, dass seine Schüler die nötigen Sprachkompetenzen erwerben. Mit der Existenz des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GER) wird nunmehr ein Raster zur Verfügung gestellt, mithilfe dessen das Können der Lerner auf eine von sechs Niveaustufen (A1-C2) zugeordnet werden soll.

Auf den Schablonen der sechs Niveaustufen versuchen die verschiedenen Testanbieter ihre Sprachzertifizierungsprüfungen aufzubauen, um die entsprechende Kompetenz zu zertifizieren. Das Goethe-Institut deklariert im Sinne des GER neue Prüfungen, die bestimmten Niveau-Stufen zugewiesen werden. Dazu gehören die in dieser Arbeit behandelten neuen Zertifikate des Goethe-Instituts B2 und C1, die zum ersten Mal weltweit im Herbst 2007 zur Anwendung kamen. Da es sich um neu erstellte Prüfungen handelt, war zu Beginn dieser Dissertation der Zugang relativ schwierig, denn selbst das Goethe-Institut befand sich noch in der Testentwicklung und Testerprobung. So wurden die Konturen dieser neuen Zertifikate erst im Laufe der Zeit deutlich.

Weiterhin wird die Prüfung des TestDaF-Instituts zum Gegenstand dieser Arbeit, da anhand einer einzigen Prüfung verschiedene Niveauzuweisungen erfolgen können, die schließlich für die „Hochschultauglichkeit“ eines nicht-muttersprachlichen potentiellen Bewerbers im sprachlichen Bereich für eine deutsche Universität oder Hochschule entscheidend sind.

[1] In dieser Arbeit liegt das Hauptaugenmerk auf den kriteriumsorientierten Prüfungen der Testanbieter Goethe-Institut und TestDaF-Institut.

Die zentrale Thematik dieser Dissertation ist die schriftliche Lernerproduktion auf den Niveaustufen B2 und C1 und die Validität der hierzu angewandten Bewertungskriterien der jeweiligen Testanbieter. Die Idee, diese Thematik anzugehen, rührt, wie bereits angeführt, aus meiner eigenen Praxis als Lehrerin, betrifft im Generellen aber auch die Problematik, dass Lerner in Sprachprüfungen nicht die Ergebnisse erzielen, die man ihnen zugetraut oder zugesprochen hätte. Dann stellt sich nämlich die Frage: Was kann der Lerner wirklich? Im Extremfall mag es heißen: „0 Punkte im schriftlichen Ausdruck“. Aber: Wie kommen diese 0 Punkte zustande? Entspricht dies der tatsächlichen Leistung eines Prüfungskandidaten oder Lerners?

- Was wird eigentlich getestet?
- Nach welchen Kriterien wurde bewertet?
- Sind die Bewertungskriterien als valide anzusehen?
- Wie kommt es zu x Punkten?
- Wie und von wem wurde die Lernerproduktion bewertet?
- Welche anderen Faktoren könnten zusätzlich zu diesem Resultat geführt haben?
- Ist dieses das Abbild der tatsächlichen Leistung eines Prüfungskandidaten bzw. Lerners?

In dieser Arbeit steht neben diesen Fragen die Problematik der Bewertung des schriftlichen Ausdrucks im Mittelpunkt, um wichtige Aussagen und mögliche Verbesserungsvorschläge in diesem Bereich zu machen. Die zurzeit in der Praxis anwendbaren Bewertungskriterien für den schriftlichen Ausdruck sollen auf das Maß ihrer Validität hin untersucht werden. Untersucht werden sollen die Bewertungsraster des Goethe-Instituts und des TestDaF-Instituts. Bei den Schwellenniveaus B2 und C1, die von einer bestimmten Sprachkomplexität charakterisiert werden, wie im Laufe dieser Arbeit zu sehen sein wird, kann die Subjektivität menschlicher Rater durch die Definition der Bewertungskriterien zunehmend an Freiraum gewinnen. Um dem entgegenzuwirken, müssen die Bewertungsraster, die aus unterschiedlich festgesetzten Kriterien bestehen, so engmaschig wie möglich definiert und entwickelt sein. Ziel dieser Arbeit ist es, die potentiellen Schwachstellen der verschiedenen Bewertungskriterien aufzudecken. Des Weiteren sollen daraus Verbesserungsvorschläge gemacht werden. Die Möglichkeit anderer Perspektiven oder auch anderer Ansätze soll diese Arbeit abrunden. Dies geschieht unter der Annahme, dass höchstmögliche Validität erst dann erreicht werden kann, wenn sich die Bewertung der schriftlichen Sprachkompetenz, unabhängig von externen Faktoren, als stabil und objektiv erweist.

2 Referenzrahmen und Testanbieter

Dieses Kapitel wird den theoretischen Hintergrund für die vorliegende Arbeit bereit stellen und sich mit sprachpolitischen Fragen und testtheoretischen Standardisierungen verschiedener Institutionen befassen. Als erstes wird der Europäische Rat und der daraus resultierende Gemeinsame Europäische Referenzrahmen für Sprachen (GER) vorgestellt, welcher das europäische Referenzniveau für Sprachkenntnisse und Sprachstandstests darstellt. Auf dessen Basis sind die hier aufgestellten Normen und Rahmenbedingungen das Instrument der in diesem Kapitel präsentierten Testanbieter, um durch Qualitätsmanagement standardisierte Tests hinsichtlich internationaler Ansprüche zu erstellen und schließlich Sprachkenntnisse unter Zuhilfenahme und Reflexion bestimmter Bewertungskriterien zu zertifizieren.

Im weiteren Verlauf werde ich mich mit der „American Psychological Association" befassen, die sich seit mehreren Jahrzehnten mit dem Teilbereich der Teststandardisierung auseinandersetzt. Die Standards der APA werden ebenso als Bezugssystem bei der kritischen Betrachtung der Bewertungskriterien schriftlicher Textproduktionen fungieren. Darauf folgend wird die sich seit Anfang der 90er Jahre etablierende Vereinigung von Sprachprüfungsanbietern innerhalb Europas, die so genannte ALTE vorgestellt, die unter anderem die Homogenität und Vergleichbarkeit der Sprachtests ihrer Mitglieder anhand definierter Standards erreichen will.

Abgerundet wird dieses Kapitel mit den Testanbietern Goethe-Institut e.V. und TestDaF-Institut, deren Bewertungskriterien schriftlicher Lernerproduktionen für die Referenzniveaus B2/C1 im Rahmen dieser Arbeit untersucht, dokumentiert und ggf. Kritisiert wurden.

2.1 Der Europarat und der Gemeinsame Europäische Referenzrahmen

Seit seiner Gründung definierte der Europäische Rat zur Förderung von Sprachen eine neue Dimension beim Lehren und Lernen von Sprachen. Der von ihm herausgegebene *Gemeinsame Europäische Referenzrahmen für Sprachen: Lernen, lehren und beurteilen* (GER)[2] wurde 2001 als Grundsatzdokument des Europarats[3] auch in deutscher Sprache veröffentlicht[4]. Die Ziele hierbei sind facettenreich. Unter anderem versteht sich der GER als „richtungweisend, umfassend, kohärent und transparent" (GER 2001: 19)[5] und als Basis für die Entwicklung von Lehrplänen, Unterrichtsplanungen und Materialerstellungen (GER 2001: 14), zugleich als Bezugssystem für Lehrer, Lerner, Curriculumentwickler, Bildungsträger und Testentwickler. Im Kontext dieser Arbeit dient der GER dem Zweck

2 Deutsche Übersetzung der endgültigen englischen Fassung des Common European Framework of Reference, Straßburg: Europarat, 2000 (Übersetzung von Prof. Jürgen Quetz, Frankfurt, Umsetzung der Eurodidaktik des Gemeinsamen Europäischen Referenzrahmens) gfl-journal, No. 3/2002

3 Der Europarat ist auch nicht zu verwechseln mit dem Europäischen Rat und dem Rat der Europäischen Union (Ministerrat), Quelle: www.wikipedia.de

4 http://www.goethe.de/z/50/commoneoro/

5 Die Literaturangabe für die deutsche Fassung Europarat-Rat für kulturelle Zusammenarbeit: Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen, Straßburg, Langenscheidt 2001 wird im Fließtext mit GER ausgezeichnet werden

der Sprachprüfungsplanung bezüglich des Prüfungsinhalts und der Beurteilungskriterien, die eher die positiven Facetten einer Lernersprache als die Defizite beleuchten sollen (GER 2001:18). Als „umfassend" sieht sich der GER, weil er den Versuch unternimmt, ein großes Spektrum der Sprachkompetenzen und deren Verwendungen so detailliert wie möglich zu definieren. Transparenz versucht der GER insofern zu sichern, indem sämtliche Definitionen wie auch andere Informationen für jedermann klar formuliert sein sollen und so als normative, gemeinsame Bezugsgröße dienen können (Perlmann-Balme 2006:5-13). Kohärent ist der Referenzrahmen erst dann, wenn seine Beschreibungen und Ausführungen „frei von inneren Widersprüchen" sind (GER 2001:19). Die Kombination dieser drei Merkmale impliziert kein „einziges" oder „einheitliches" System, folglich gehört die Definition und die Eingrenzung individueller Ziele und Methoden nicht zu den Aufgabenbereichen des GER. Dieser Referenzrahmen kann als deskriptiver Ansatz betrachtet werden, der lediglich das zur Verfügung stellt, was benötigt wird, um sich über gesetzte Ziele, Inhalte und die erforderlichen Methoden und hinsichtlich ihrer Realisierung Gedanken zu machen (Glaboniat/Müller 2006: 15-21). Die definierten Ziele reichen dabei in der Bandbreite von einzelnen sprachlichen Teilkompetenzen bis hin zur „kompletten" Sprachbeherrschung.[6]

2.1.1 Die Kompetenzbeschreibungen im Gemeinsamen Europäischen Referenzrahmen

Der GER beschreibt anhand von sechs definierten Niveaustufen (A1-C2) fremdsprachliche Kompetenzen und Fertigkeiten, deren Erwerb den FS-Lerner im Sinne eines handlungsorientierten Ansatzes (GER 2001:21) als soziale Person handlungsfähig machen sollen. Dabei ist das Definieren gemeinsamer Referenzniveaus der Sprachkompetenz aus verschiedenen Gründen nützlich. Einerseits könnten anhand der Erwartungen der einzelnen Niveaus Lernziele konkretisiert werden. Andererseits könnte der Vergleich zwischen verschiedenen Systemen oder auch Lernzielen durch gemeinsame Sprachkompetenzbeschreibungen erleichtert werden (GER 2001:28).

Im Mittelpunkt stehen verschiedene Arten von Kompetenzen, die einen Lerner befähigen sollen, kommunikativ aktiv zu sein. Dieser Tatsache Rechnung tragend benötigt ein Lerner für gewisse Kontexte und die Umstände oftmals lediglich Teilkompetenzen. Aus diesem Grund zeigt sich in der Taxonomie der Kompetenzen im GER eine Aufgliederung.[7] Die Aktivierung der verschiedenen Ausprägungen von Sprachkompetenz, die rezeptiv, produktiv, interaktionistisch oder sprachmittelnd vonstatten gehen kann, erfolgt mithilfe allgemeiner Kompetenzen, kommunikativer Kompetenzen und verschiedenen Strategien (GER 2001:21ff.). Folgende Tabelle soll die benötigten Elemente und Komponenten zur Sprachverwendung im Sinne eines handlungsorientierten Ansatzes verdeutlichen:

6 http://www.goethe.de/z/commeuro/i1.htm

7 http://www.goethe.de/z/commeuro/i1.htm

Allgemeine Kompetenzen	Kommunikative Sprach-kompetenzen	Kommunikative Sprachaktivitäten	Lebensbereiche / Domäne	Kommunikative Aufgaben, Strategien, Texte
deklaratives Wissen, prozedurales Wissen, persönlichkeits-bezogene Kompetenzen, Lernfähigkeit	Linguistische Kompetenz, soziolinguistische Kompetenz, pragmatische Kompetenz	Rezeption, Produktion, Interaktion, Sprachmittlung	öffentlicher Bereich, privater Bereich, beruflicher Bereich, Bildungswesen	z.B. Kommunikations- und Lernstrategien

Tabelle1: Komponenten eines handlungsorientierten Ansatzes

Diese im GER als horizontal definierte Dimension (GER 2001:25 ff.) gibt an, über welche Kompetenzen Lerner verfügen müssen, um kommunikativ handlungsfähig zu sein. Die jeweiligen Stufen des GER werden aufsteigend spezieller und erfordern zunehmend automatisierte Sprachbenutzung. An dieser Stelle sei vermerkt, dass die allgemeinen Kompetenzen und die Strategien keiner Stufe zuzurechnen sind, sondern für alle beschriebenen Niveaustufen als vorausgesetzt gelten. Das meiste benötigte Wissen bzw. Weltwissen ist eine latente Voraussetzung. Zu hinterfragen hierbei ist, in welcher Weise und in welchem Umfang dieses implizit vorausgesetzt wird, in dem Sinne, dass es bereits aus der L1 resultiert. Das bedeutet, dass bei dem Erlernen einer neuen Sprache das meiste benötigte Wissen (z.B. bestimmte Themen) bereits aus der Muttersprache hervorgehen kann. Laut GER verfügen Erwachsene über ein ausdifferenziertes Modell, „das mit dem Vokabular und der Grammatik der Muttersprache eng verbunden ist“, wobei diese eng aufeinander bezogen sind. Die Kommunikation eines Menschen wird im Laufe der Entwicklung durch die „Übereinstimmung zwischen den von den Beteiligten internalisierten Weltmodellen und der Sprache“ abhängig gemacht (GER 2001:103). Für das Erlernen einer Fremd- bzw. Zweitsprache geht man davon aus, dass Lerner über ein hinreichendes Weltwissen verfügen. Dabei umfasst Weltwissen zum Beispiel Sachwissen über das Land der gesprochenen Sprache oder die Einteilung von Dingen in Klassen und Funktionen. Im handlungsorientierten Ansatz des GER wird der Lerner als sozial Handelnder betrachtet, der als Teil der Gesellschaft kommunikative Aufgaben zu bewältigen hat, die jedoch nicht immer sprachlicher Natur sind (GER 2001:21). Der GER versucht diesbezüglich aus methodischer Sicht zu beantworten, wann „die zu prüfende Sprache ein Weltwissen involviert, das über den Reifegrad der Lernenden hinausgeht“ (GER 2001:145). Diesbezüglich stellt sich die Frage, ob man überhaupt von einem einheitlichen oder gemeinsamen Konzept ausgehen kann. Besonders wichtig ist diese Frage für den Bereich des schriftlichen Ausdrucks, wenn es darum geht, welche Textsorten von Testanbietern erwartet werden und ob diese unter Berücksichtigung aller

Faktoren schließlich gerechtfertigt sind. Dies betrifft auch die Definition der Bewertungskriterien hinsichtlich der als erforderlich definierten Kompetenzen.

Man kann sich das vom GER definierte Modell als ein funktionales Koordinatensystem vorstellen, in dem die verschiedenen Kompetenzen und Strategien aus obiger Tabelle die x-Achse und die im Folgenden skizzierten Niveaus A1-C2 die y-Achse darstellen. Je nach Ausprägung werden verschiedene Werte in dieses System eingetragen, Werte der Sprachbeherrschung, die sich funktional ab dem Wert Null aufwärts bewegen können. Man kann der angeführten Tabelle des Weiteren die Kongruenz der einzelnen Komponenten entnehmen. Kompetenzen treten mittels kommunikativer Sprachaktivitäten hervor. Die primären Prozesse Rezeption und Produktion, die Interaktion und die Sprachmittlung finden in den vier vom GER skizzierten Domänen ihre Anwendung. Da gemäß des GER die entscheidenden Lebensbereiche eines Individuums abgedeckt sind, wird er in dieser Hinsicht seinem Anspruch, „umfassend“ zu sein, gerecht. Die erarbeiteten und angeführten Kompetenzbeschreibungen sollen in diesem Sinne „kohärent und transparent“ (GER 2001:29) sein. Bedarfsanalysen und individuelle Systeme sollen anhand der kalibrierten Kompetenzbeschreibungen des GER Lernziele, Inhalte und Leistungsevaluation adäquat und im jeweils definierten Rahmen entsprechend formulieren. Lerner benötigen außerdem Strategien, um die Aktivierung der angeführten Kompetenzen erfolgreich zu bewältigen. Laut des GER werden Strategien von Lernern insofern benötigt und eingesetzt, um „die eigenen Ressourcen zu mobilisieren und ausgewogen zu nutzen, Fertigkeiten und Prozesse zu aktivieren, um die Anforderungen der Kommunikation in einem Kontext zu erfüllen und die jeweilige Aufgabe erfolgreich und möglichst ökonomisch der eigenen Absicht entsprechend zu erledigen“ (GER 2001:62). An anderer Stelle werde ich mich mit der Thematik der kommunikativen Strategien und insbesondere der Vermeidungsstrategien auseinander setzen, obgleich der GER Kommunikations- und Kompensationsstrategien nicht im Sinne eines Defizitmodells betrachtet. Durch das Einsetzen von Strategien versuchen Lerner ihr sprachliches Potential so zu nutzen, dass sie der Arbeitsanforderung gerecht werden. Der GER definiert Strategien als Gelenkstellen zwischen den Ressourcen bzw. der Kompetenzen der Lerner und dem, wie sie kommunikativ damit umgehen (GER 2001:38). Dieses wird im Referenzrahmen als Produktionsstrategie definiert, welche „das Mobilisieren von Ressourcen und das Ausbalancieren verschiedener Kompetenzen, d.h. das Ausnutzen der Stärken und Herunterspielen der Schwächen, um das vorhandene Potential im Sinne der jeweiligen Aufgabe einzusetzen, involviert“ (GER 2001:68).

2.1.2 Die Niveaustufen des Gemeinsamen Europäischen Referenzrahmens

Nach Krumm (2006) legen die Niveaustufenbeschreibungen des GER fest, was Lerner oder Prüfungsteilnehmer auf bestimmten Stufen können sollen, sodass das Resultat einer Prüfung zum Beispiel nicht inhaltslos bleibt (vgl. Krumm 2006). Die Oberbegriffe werden entsprechend „Elementare Sprachverwendung“ für Niveaubereich A, „Selbständige Sprachverwendung“ für Niveaubereich B und „Kompetente Sprachverwendung“ für Niveaubereich C benannt. Diese Niveaustufen können als Messeinheit der vertikalen Dimension (GER 2001:25ff.) unseres Koordinatensystems, der y-Achse also, verstanden werden. Dennoch weist der GER darauf hin, dass „Lernfortschritt nicht einfach das Vorankommen auf einer vertikalen Skala ist. Es gibt keinen zwingenden logischen Grund

dafür, dass Lernende sämtliche niedrigeren Stufen einer Teilskala durchlaufen müssen (…) Man sollte sich schließlich davor hüten, Niveaus und Sprachkompetenzskalen als eine lineare Messskala – wie z.B. einen Zollstock – zu interpretieren“ (GER 2001:15ff.).

Die schon 1975 definierte Stufe „Threshold“ (B1) von J. A. van Ek (vgl. van Ek 1976) „war die erste konsequente Durchführung eines lernerzentrierten, funktionalen und nationalen Ansatzes zur Lehrzielbestimmung für den Fremdspracherwerb“ (Baldegger/Müller/Schneider 1981:5) und wird nun durch weitere darunter und darüber liegenden Stufen vervollständigt, wobei die Abstände zwischen ihnen variieren und die Niveaustufenskala eher im Eistütenformat darzustellen wäre, da sprachliche Aktivitäten, Fertigkeiten und sprachliche Mittel aufsteigend komplexer werden (GER 2001:29):

A1	A2	*B1*	B2	C1	C2
Breakthrough	Waystage	*Threshold*	Vantage	Effective Operational Proficiency	Mastery

Tabelle 2: Niveaustufenbeschreibung

Für die vorliegende Arbeit sind die Stufen B2 und C1 von Interesse. Für diese auf einer vertikalen Dimension des GER angesiedelten Niveaustufen (GER 2001:27) sind so genannte Kann-Beschreibungen oder auch Can-Do-statements[8] entwickelt worden, die einen positiven und handlungsorientierten Charakter aufweisen. Sie stellen nicht die Sprachdefizite, sondern das fremdsprachliche Können auf entsprechender Lernerstufe zunächst für alle Sprachen in den Mittelpunkt (vgl. Glaboniat/Müller 2006). Dadurch soll konkretisiert werden, welche Anforderungen jedes Niveau für sich hat.[9] Dennoch verweist der GER darauf, dass es bis zu einem bestimmten Grad willkürlich ist, die Niveaustufen an Sprachkompetenz festzumachen und zu definieren (GER 2001: 28). In diesem Zusammenhang erweist sich der Aufsatz von Frey (2004) „Die Kompetenzbeschreibungen des Europäischen Referenzrahmens: Beobachtungen zur Trennschärfeproblematik“[10] als eine sehr hilfreiche Quelle. Durch die Ergebnisse eines Workshops hinsichtlich der Niveaubeschreibungen des GER und deren Training kommt die Autorin zum Resultat, dass die Zuordnung von Kompetenzen zu bestimmten Niveaustufen nicht eindeutig ist. Daher sei es wichtig, die *Exaktheit in der Beschreibung der Kompetenzniveaus* zu fordern, indem man zum einen mehr Konsistenz in der begrifflichen Anwendung der Niveaus und zum anderen mehr Konsistenz bei der Auswahl der Signalebenen voraussetzt.

Wie Skalen mit ihren Deskriptoren aufgestellt werden steht letztlich immer mit der jeweiligen Zweckerfüllung in Verbindung. Für die Beschreibung der Kompetenzniveaus ist es wichtig zu definieren, wie Kompetenz im jeweiligen Kontext zu verstehen ist. Die

8 Die Begriffe Kann-Beschreibungen und Can-Dos/Can-Do-statements werden in dieser Arbeit abwechselnd synonym verwendet werden

9 Die definierten Can-Dos haben konkreten Inhalt. Es stellt sich aber die Frage, ob dieser Inhalt die Anforderungen der Niveaus akkurat wiedergibt

10 www.hueber.de/sixcms/media.php/36/referenzrahm-frey.pdf., Zugriff am 20.08.2008

Niveaus sind anhand von kalibrierten und empfohlenen, jedoch nicht verpflichtenden Deskriptoren definiert, die mittels „intuitiver, quantitativer und qualitativer“ (GER 2001: 33). Messmethoden teilweise empirisch skaliert und schließlich durch Lehrende validiert worden sind (vgl. Schneider 2001). Für die Formulierung von Deskriptoren wurden im GER die Erfahrungen und die Datenbanken vieler verschiedener Institutionen herangezogen (GER 2001:217ff.). Dazu zählt auch die Zusammenarbeit mit der Organisation von Sprachprüfungsanbietern in Europa (ALTE), die im Weiteren vorgestellt wird.

Im so genannten Can-Do-Projekt der ALTE entwickelte und validierte man eine große Bandbreite an Deskriptoren für die Domänen von Erwachsenen, die die des GER ergänzen (GER 2001:33). Die Tatsache, dass sämtliche Projekte ihre Deskriptorendatenbanken hinsichtlich des Lebens von Erwachsenen erstellen, mag für diese Arbeit ein Indiz dafür sein, dass das Alter und die Reife potentieller Prüflinge für Sprachprüfungen verschiedener Testanbieter keine Berücksichtigung zu finden scheint.[11] Wichtig hierbei ist die Frage, wie der GER zunächst den Begriff Domäne definiert. Domänen werden hier sehr abstrakt kategorisiert und auf die wesentlichsten Kategorien beschränkt, in denen ein Lerner sozial agieren soll (GER 2001:22). Da sich die Domänen auf der x-Achse des Koordinatensystems befinden, muss die Frage beantwortet werden, worauf der thematische Input bei Sprachtests beruht. In den Komponenten der Sprachverwendung scheint das Alter nur indirekt innerhalb der Domänen, die Bereiche des Erwachsenenlebens definieren, begründet zu sein.

Deskriptoren sollten klar, transparent, kriteriumsbezogen und autonom sein. Da sich der Referenzrahmen als umfassend und transparent definiert, können gleiche Deskriptoren auch in anderen Systemen eingebettet werden, um diese dann auf die Aspekte kommunikative Aktivität, Strategie und kommunikative Sprachkompetenz zu beziehen (GER 2001:38ff.). In Globalskalen oder in nach Domänen kategorisierten Skalen für die einzelnen Niveaus werden Vorschläge gemacht, wie Beschreibungen aussehen könnten und welchem Zweck sie jeweils dienen sollten. Das als einheitlich verstandene System soll den Vergleich von Lernzielen, Niveaustufen, Materialien, Tests und Lernerfolgen in verschiedenen Systemen, Kontexten und Situationen erleichtern. Das Beschreibungssystem und die Referenzniveaus des GER basieren darauf, dass jeder Benutzer dadurch sein eigenes System formuliert. Aus diesem Grund sollte eine Skala von Referenzniveaus im Idealfall sowohl Kriterien, die ihre Beschreibung einschließt, als auch Fragen bezüglich der Messverfahren berücksichtigen (GER 2001:32ff.). Eine gemeinsame Referenzskala muss im Sinne des GER kontextfrei sein. Dies bedeutet, dass verschiedene kontextuelle Gegebenheiten darin Platz finden, so dass man zu einer Generalisierung kommt. Trotz dieses Umstandes müssen die Deskriptoren dennoch kontextrelevant sein, sie müssen funktional alle möglichen und potentiellen Kontexte abdecken können. Der GER stellt weiterhin den Anspruch, dass die Beschreibung einer Referenzskala auf Sprachkompetenztheorien basieren sollte. Die verschiedenen Theorien hinsichtlich der Sprachkompetenz, die letztlich aus Ansätzen des Zweitspracherwerbs hervorgehen, müssten meines Erachtens aber hinreichend selektiert werden, um eine Basis zu schaffen, auf der die gemeinsamen Referenzniveaus beschrieben werden können. Inwieweit dies realisiert werden kann, stelle ich insofern in Frage, weil innerhalb der Zweitspracherwerbsforschung keine universelle Positionierung feststellbar ist.

11 Aus analytischer Sicht ist dies insofern bemerkenswert, da das Alter das zweidimensionale Koordinatensystem sprengt

Betrachtet man nun die Referenzniveaus aus der Perspektive des Threshold Levels (B1), so wird ersichtlich, dass sich die Grenze nicht zwischen der selbständigen und der kompetenten Sprachverwendung befindet (Niveau B und C), denn „das Niveau B2 weicht ganz erheblich von den bisherigen Inhalten ab und ist als eine neue Schwelle zu betrachten“ (GER 2001:44). Das Vantage Level B2 definiert den Lerner, „der langsam aber sicher das mittlere Lernplateau durchschritten hat und merkt, dass er jetzt an einen Punkt angekommen ist, von dem aus die Dinge in einem anderen Licht erscheinen und sich neue Perspektiven eröffnen“ (GER 2001:44). Gerade in seiner starken Form grenzt sich das Niveau B2 vom Threshold Level (B1) ab, denn hier werden nunmehr grundlegende Aspekte bereit gestellt, wie zum Beispiel Argumentationsfähigkeit und stärkeres Sprachbewusstsein. Das Niveau C1 kennzeichnet schließlich die Intensivierung und Festigung der Diskursfertigkeiten und der sprachlichen Mittel des vorangegangenen Niveaus, wobei das Gewicht aber insgesamt auf den Aspekt der Flüssigkeit gelegt wird. Es wird ersichtlich, dass die Abstände zwischen den Niveaus nicht identisch sind. Eine Diskrepanz zwischen B2 und C1 ist geringer als die zwischen B1 und B2, obwohl B2 den erforderlichen Standard von C1 noch nicht erreicht.

Zur Formulierung einheitlicher Referenzskalen müssen Aspekte von Messverfahren berücksichtigt werden. Der GER stellt den Anspruch objektiver Skalen, um subjektiven Konventionen, so weit es geht, Stand halten zu können. Ich werde diesen Umstand im Verlauf der Arbeit immer wieder aufgreifen. Die jeweils definierten Kompetenzskalen sollen schließlich objektive Aussagen bezüglich der Leistungen geben können und dabei so konzipiert sein, dass durch ihr Ausmaß menschliche Rater oder Korrektoren *konsistente Unterscheidungen* treffen können. Die Forderung nach der Entwicklung einer gemeinsamen Referenzskala scheint berechtigt, obwohl dieser Idealzustand natürlich schwer zu erzielen ist. Anders ausgedrückt muss der Frage nachgegangen werden, ob diese bezüglich des gegebenen und definierten Kontextes das Kriterium der Validität erfüllen. Der GER hat diesen Validierungsprozess für die Entwicklung von Skalen, die aus Deskriptoren bestehen, mithilfe einer Kombination verschiedener Forschungsmethoden durchlaufen (GER 2001:33). Ziel der Validierung, die als ein *permanenter und theoretisch unendlicher Prozess* anzusehen ist, soll vor allem die explizite Beschreibung von Kompetenzen durch die Deskriptoren sein.

Die sechs aufgestellten Niveaus des GER, die sich außerdem durch subjektive Kategorisierungen aufspalten lassen können, definieren in Form einer Globalskala zunächst die grundlegenden Erfordernisse der Can-Dos.[12] Diese globale Auflistung von Kompetenzen soll im Sinne des GER definieren, *wie gut* jemand in einer Fremdsprache sprachliche Handlungen ausführen bzw. bewältigen kann. Es werden demnach Aussagen über die qualitativen Aspekte der Sprachkompetenz gemacht (GER 2001:35):

12 Skalen übersetzt von Prof. Günther Schneider (Universität Fribourg, Vertreter der Schweizerischen Konferenz der kantonalen Erziehungsdirektion/EDK). Ausführlichere Informationen enthält man unter www.goethe.de/referenzrahmen

	Kompetente Sprachverwendung
C2	**Kann praktisch alles, was er/sie liest oder hört, mühelos verstehen. Kann Informationen aus verschiedenen schriftlichen und mündlichen Quellen zusammenfassen und dabei Begründungen und Erklärungen in einer zusammenhängenden Darstellung wiedergeben. Kann sich spontan, sehr flüssig und genau ausdrücken und auch bei komplexeren Sachverhalten feinere Bedeutungsnuancen deutlich machen.**
C1	**Kann ein breites Spektrum anspruchsvoller, längerer Texte verstehen und auch implizite Bedeutungen erfassen. Kann sich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Kann die Sprache im gesellschaftlichen und beruflichen Leben oder in Ausbildung und Studium wirksam und flexibel gebrauchen. Kann sich klar, strukturiert und ausführlich zu komplexen Sachverhalten äußern und dabei verschiedene Mittel zur Textverknüpfung angemessen verwenden.**
	Selbständige Sprachverwendung
B2	**Kann die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen verstehen; versteht im eigenen Spezialgebiet auch Fachdiskussionen. Kann sich so spontan und fließend verständigen, dass ein normales Gespräch mit Muttersprachlern ohne größere Anstrengung auf beiden Seiten gut möglich ist. Kann sich zu einem breiten Themenspektrum klar und detailliert ausdrücken, einen Standpunkt zu einer aktuellen Frage erläutern und die Vor- und Nachteile verschiedener Möglichkeiten angeben**
B1	**Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Kann sich einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete äußern. Kann über Erfahrungen und Ereignisse berichten, Träume, Hoffnungen und Ziele beschreiben und zu Plänen und Ansichten kurze Begründungen oder Erklärungen geben.**
	Elementare Sprachverwendung
A2	**Kann Sätze und häufig gebrauchte Ausdrücke verstehen, die mit Bereichen von ganz unmittelbarer Bedeutung zusammenhängen (z. B. Informationen zur Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). Kann sich in einfachen, routinemäßigen Situationen verständigen, in denen es um einen einfachen und direkten Austausch von Informationen über vertraute und geläufige Dinge geht. Kann mit einfachen Mitteln die eigene Herkunft und Ausbildung, die direkte Umgebung und Dinge im Zusammenhang mit unmittelbaren Bedürfnissen beschreiben.**
A1	**Kann vertraute, alltägliche Ausdrücke und ganz einfache Sätze verstehen und verwenden, die auf die Befriedigung konkreter Bedürfnisse zielen. Kann sich und andere vorstellen und anderen Leuten Fragen zu ihrer Person stellen – z. B. wo sie wohnen, was für Leute sie kennen oder was für Dinge sie haben – und kann auf Fragen dieser Art Antwort geben. Kann sich auf einfache Art verständigen, wenn die Gesprächspartnerinnen oder Gesprächspartner langsam und deutlich sprechen und bereit sind zu helfen.**

Tabelle 3: Globale Kann-Beschreibung

Die globale Fragestellung, *wie gut* man etwas in der Fremdsprache beherrscht, ist in der hier im Mittelpunkt stehenden Bewertung schriftlicher Lernerproduktionen auf den Niveaus B2 und C1 problematisch. Im Folgenden soll die Globalskala dieser beiden Niveauabstufungen betrachtet und Stellung bezüglich der globalen Deskriptorendefinition genommen werden. Ich werde im Verlauf dieser Arbeit die so genannten Stolpersteine anhand der konkreten Bewertungskriterien des Goethe-Instituts und des TestDaF-Instituts erörtern und des Weiteren versuchen, alternative Vorschläge zu machen, wie diese zu umgehen wären. Diese holistische Aufbereitung der erwarteten Kompetenzen im Fremdsprachenbereich kann als Orientierungshilfe betrachtet werden. Insgesamt schließt der GER nicht aus, dass die Formulierungen der existierenden Skalen verfeinerungswürdig sind (GER 2001:34). Zu dem Oberbegriff *selbständige Sprachverwendung* zählt das Referenzniveau B2, welches global folgendermaßen definiert wird und Aufschluss über die Sprachkompetenz geben soll:

„Kann die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen verstehen; versteht im eigenen Spezialgebiet auch Fachdiskussionen. Kann sich so spontan und fließend verständigen, dass ein normales Gespräch mit Muttersprachlern ohne größere Anstrengung auf beiden Seiten gut möglich ist. Kann sich zu einem breiten Themenspektrum klar und detailliert ausdrücken, einen Standpunkt zu einer aktuellen Frage erläutern und die Vor- und Nachteile verschiedener Möglichkeiten angeben".

Im Folgenden werde ich erläutern, wieso meines Erachtens mittels dieser holistisch geprägten Definition der Niveaustufe B2 keine allgemeingültige Aussage bezüglich der Sprachkompetenz gemacht werden kann. Nach meiner Auffassung bedürfte bereits der erste Deskriptor auf dieser Niveaustufenbeschreibung einer engeren Begriffsbestimmung. Dass gerade in einer standardisierten Prüfungssituation[13], die in der vorliegenden Arbeit den Mittelpunkt ausmacht, konkrete bzw. abstrakte Themen vorgegeben werden können, ist zwar eine Tatsache, klärt aber meines Erachtens in dieser holistischen Beschreibung nicht den Umstand, um welche Themengebiete es sich handeln kann. Ich habe bereits angeführt, dass die Klassifizierung von Dingen, d.h. konkrete oder abstrakte Themen, Teil des Weltwissens sind. Zunächst bilden diese Themen die Oberfläche. Es stellt sich in diesem Zusammenhang aber die Frage, wie definierte Themengebiete bearbeiten werden, wenn sie nicht einmal das Weltwissen der L1 ausmachen. Neben dem kulturellen Umfeld sind aber ebenso das Alter und der persönliche Hintergrund hinsichtlich der Domänen insgesamt wichtig und zu berücksichtigen. Ich bin davon überzeugt, dass das *eigene Spezialgebiet* eines 16jährigen mit dem eines 30jährigen nicht vergleichbar gemacht werden kann. Dabei spielen unter anderem die verschiedenen Wissensbestände einer Person eine Rolle.[14] Zentral ist demnach, was ein Test zu messen vermag. Wenn es sich lediglich um die Erhebung von Sprachkenntnissen handelt, dann findet das Kriterium der Fairness seine Berechtigung insofern, dass der Faktor Alter beispielsweise eine untergeordnete Rolle spielen würde. Der GER definiert in seinem handlungsorientierten Ansatz spezifische Umgebungen und Handlungsfelder, in denen man die Vielzahl von Kompetenzen (vgl. oben) strategisch und planvoll einsetzt (GER 2001:21). Demnach ist die Handlungsfähigkeit in einem bestimmten Kontext bzw. Domäne als eine vorausgesetzte Kompetenz zu betrachten. Insofern wäre nachzuweisen, ob derartige Kompetenzen, die weit über die sprachliche Kompetenz hinaus gehen, bei allen Prüfungsteilnehmern als vorausgesetzt betrachtet werden können, so dass der erstellte

13 High-stakes (empirische Kalibrierung) vs. low-stakes Tests

Test seine Berechtigung hinsichtlich der Testgütekriterien findet. Es geht also in erster Linie darum, dass Sprachkompetenz im Sinne Bachmans/Palmers (1996) durch Interaktivität gekennzeichnet wird. Die verschiedenen Ebenen, auf denen Lerner interagieren sind sprachlich, strategisch, affektiv und auf Weltwissen bezogen. Innerhalb der Testsituation sollte aber darauf geachtet werden, die sprachliche Ebene hingegen der anderen drei Ebenen soweit zu maximieren, um dem was gemessen wird bzw. der Validität gerecht zu werden.

Das *eigene Spezialgebiet* ist eine kontroverse Definition. Während ein Spezialgebiet als ein Fachgebiet betrachtet werden kann, kann das eigene Spezialgebiet aber nicht erfassbar sein. Krekeler (2005) beschäftigt sich mit der Thematik des Fachbezugs in Sprachtests für den Hochschulzugang, allerdings konzentriert er sich auf das Leseverstehen.[15] Sprachtests ohne Fachbezug prüfen lediglich das Kriterium der Sprachkompetenz und werden im Sinne der Testtheorie der Testökonomie und der Testfairness gerecht (vgl. Kapitel 4.3.2). Bei Sprachtests mit Fachbezug werden Fachkenntnisse als Teil des Testkonstrukt angesehen bzw. als Kompetenz vorausgesetzt. Krekeler stellt verschiedene Positionen gegenüber. Während einerseits die Meinung vorherrscht, dass sprachliche Leistungen je nach Kontext und Aufgabentypus variieren können (vgl. Douglas 2000), existiert andererseits die Position, dass man Studienbewerber nicht mit Kommunikationssituationen aus dem Hochschulalltag konfrontieren sollte. Während Krekeler die Auswirkungen auf die Testvorbereitung und die Authentizität für die wichtigsten Vorteile von Sprachtests mit Fachbezug hält, stellt er kontrastiv dazu die Frage auf, ob sich Hochschulsprache durch Sprachtests mit Fachbezug überhaupt ausreichend differenzieren lassen kann. Diesbezüglich ist auch der Einfluss der Vorkenntnisse eines Prüfungskandidaten von Bedeutung und inwieweit sich daraus sprachliche Niveaus bzw. Schwellen ableiten lassen können. In diesem Sinne ist die zu messende sprachliche Kompetenz meines Erachtens als eine *Kann-Beschreibung des Spracheinsatzes* zu betrachten. Ob dieser Spracheinsatz ein gemeinsames bzw. universelles Hintergrundwissen voraussetzt, bleibt zu klären. Die Definition des B2-Niveaus auf dieser Skala kann Benutzer des GER dazu verleiten, dass sie unter der Berücksichtigung, welche Kompetenzen als vorausgesetzt gelten, Sprachtests mit Fachbezug erstellen (kann an Fachdiskussionen teilnehmen) könnten. Ziel hierbei wäre die Kompetenz der Sprachverwendung in bestimmten Kontexten zu eruieren. In diesem Sinne kann die Leistung in Abhängigkeit des Ausmaßes dieser erforderten Kompetenz variieren. Dass man auf diesem Niveau die Kompetenz einräumt, Argumentationen zu führen, ist meines Erachtens eine vom *breiten Themenspektrum* unabhängige Komponente. Dennoch bleibt der *klare und detaillierte Ausdruck* eine unzureichend definierte Kompetenz. Worauf bezieht sich der *klare und detaillierte Ausdruck*?

Sprache verwendet man auf der ersten Stufe des C-Niveaus kompetent, wenn gemäß des GER folgendes globale Definitionskonstrukt erfüllt ist:

„Kann ein breites Spektrum anspruchsvoller, längerer Texte verstehen und auch implizite Bedeutungen erfassen. Kann sich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen. Kann die Sprache im gesellschaftlichen und beruflichen Leben oder in Ausbildung und Studium wirksam und flexibel gebrauchen. Kann sich klar, strukturiert und ausführlich zu komplexen

15 Krekeler, C. (2005): Grammatik und Fachbezug in Sprachtests für den Hochschulzugang. Dissertationsschrift. Universität Duisburg Essen http://dueplico.uni-duisburg-essen.de/servlets/DocumentServlet?id_12458

Sachverhalten äußern und dabei verschiedene Mittel zur Textverknüpfung angemessen verwenden".

Zunächst gilt es innerhalb dieser globalen Kann-Beschreibung zu klären, welchen Anspruch man an den Lerner hat, wenn von *anspruchsvollen, längeren Texten* die Rede ist. Texte bzw. Textsorten können als mehr oder weniger anspruchsvoll empfunden werden. Dabei spielen unter anderem die verschiedenen Wissensbestände einer Person eine Rolle. Weiterhin besagt die spontane und flüssige Sprachbeherrschung längst nichts darüber, ob der Anspruch diesem Niveau, d. h. Sprache kompetent zu verwenden, gerecht wird. Ebenso bezieht sich der Sprachgebrauch dieser Niveaubeschreibung auf das gesellschaftliche und berufliche Leben. Auch hier kann man den Einwand erheben, dass Themeninhalte bezüglich des Studierens beispielsweise nicht unbedingt die Interessen eines 16jährigen definieren. Abschließend bezieht sich dieser holistische Deskriptor auf die Kompetenz, verschiedene sprachliche Mittel hinsichtlich der Äußerungsabsicht komplexer Sachverhalte angemessen zu verwenden. Man geht bei diesem globalen Deskriptor demnach der Frage nach, wie gut man sich zu *komplexen Sachverhalten äußern kann.* Die Antwort darauf lautet an dieser Stelle *klar, strukturiert und ausführlich.*

Abgesehen von der qualitativen Beschreibung der Sprachkompetenz, definiert der GER auch die Kehrseite, die Quantität. Hierbei geht es darum, *was* jemand in der Fremdsprache kann bzw. welche sprachlichen Handlungen im Mittelpunkt stehen (Glaboniat/Müller 2006:16). Auf diese so genannten „detaillierten Kann-Beschreibungen" wird im 5. Kapitel dieser Arbeit zurückgegriffen werden, wenn es darum geht, Kritik an verschiedenen Bewertungskriterien zu üben.

Nach Alderson (1991) gibt es verschiedene Arten von Skalen, deren Ziel und Hintergrund jeweils ein anderer ist. Benutzerorientierte Skalen sind in der Regel holistisch, um dem Lerner die Selbsteinschätzung zu erleichtern. Durch derartige einfache Skalen, die meist positiv formuliert sind, kann der Lerner ermitteln, was er bereits fähig ist, mittels Sprache zu tun (GER 2001:46). Beurteilungsorientierte Skalen konzentrieren sich auf den Aspekt der Qualität der erwarteten oder der zu messenden Leistung. Diese Skalen sind für den Bewertungs- und Beurteilungsprozess gedacht und können sowohl holistisch, analytisch oder gar als Kombination beider Eigenschaften auftreten, wobei auch auf höheren Niveaus meistens negativ formuliert wird. Holistische Skalen für Beurteilende sind mit nur einem Deskriptor ausgestattet, hingegen beziehen sich analytische Skalen auf verschiedene Kriterien der Sprachkompetenz oder auch –leistung. Skalen mit vielen Kategorien und Deskriptoren eignen sich allerdings weniger, denn dadurch wird die Beurteilungsobjektivität eingeschränkt. Rater empfinden es als eine Überforderung, wenn sie mehr als drei bis fünf Deskriptoren für die Bewertung zu Rate ziehen müssen (GER 2001:47). Der Schwerpunkt beurteilungsorientierter Skalen liegt schließlich auf der adäquaten Sprachverwendung eines Lerners. Der GER gibt in dieser Hinsicht den Ratern den Ratschlag „zu bedenken, inwieweit sich ihr Interesse auf eine verbesserte Konsistenz von Beurteilungen bezieht, indem gut definierte Kriterien für die verschiedenen Fertigkeitsniveaus angeboten werden" (GER 2001:49). Schließlich benutzen Testautoren aufgabenorientierte Skalen, um Tests zu entwickeln bzw. zu erstellen. Hier wird der Frage nachgegangen, wie ein Lerner mit Sprache umgehen kann. Auch für den Idealzustand der Testerstellung und ihrer Bewertung übernimmt der GER nicht die Verantwortung. Anstatt vorgefertigter Schablonen definiert er in diesem Zusammenhang testtheoretische Fragestellungen und Ansätze, die Anreiz dafür sein sollen, je nach Notwendigkeit methodisch adäquat vorzugehen (Glaboniat/Müller

2006:16f.)). Der GER weist darauf hin, dass man bei den verschiedenen Skalen jedoch einen Unterschied zwischen den Sprachkompetenzniveaus und der Bewertung der erzielten Leistung immer hinsichtlich des gesetzten Ziels machen muss. Gemeinsam ist allen Projekten jedenfalls, dass die Standards und die Relation zwischen Punktwerten und zugewiesener Kompetenzstufe sich so weit annähern, dass sie vergleichbar gemacht werden können (GER 2001:49).

Der GER bietet Skalen an, die jede Facette des kommunikativen Handelns berücksichtigen sollen. Bachman (1990:325f.) definiert in diesem Zusammenhang die „real-life"-Skala, wonach das abgebildet wird, was ein Lerner auf einer bestimmten Niveaustufe im realen Leben kommunikativ tun kann. Dabei beinhaltet „die Bewältigung einer kommunikativen Aufgabe die strategische Aktivierung spezieller Kompetenzen, um innerhalb eines bestimmten Lebensbereichs (...) zielgerichtet Handlungen mit einem klar definierten Ziel (...)auszuführen" (GER 2001:153). Im Vorfeld sind aber in dieser Hinsicht Fragen zu beantworten, die sich damit befassen, was ein Lerner erwerben muss, das heißt jeder Benutzer des GER muss die Lernziele für seinen eigenen Bedarf definieren. Dafür muss zunächst eine Bedarfsanalyse der Lerner oder gar der Gesellschaft gemacht werden. Der nächste Schritt besteht darin, zu ermitteln, anhand welcher Aufgaben, Prozesse und Aktivitäten diese Bedürfnisse befriedigt werden können, in welchen Domänen sich all das widerspiegelt und welche Kompetenzen und Strategien ein Lerner letztlich dafür benötigt. Vor allem ist die Strategieentwicklung ein wichtiger Punkt, dem der GER einen großen Platz einräumt (Glaboniat/Müller 2006:16). Schließlich skizziert der GER dadurch ein methodisch-didaktisches Konzept, das den Lerner, seine Bedürfnisse und den Fortschritt seiner Sprachkompetenz in den Mittelpunkt stellt. Die Kriterien für die Beschreibung der Kompetenzen unterliegen objektiven Kriterien, was schließlich dazu führen soll, dass eine Grundlage für die Angleichung verschiedener Lehr- und Lernzielen durch die Sprachen Europas und deren Zertifizierung definiert wird, indem ein gemeinsames Bezugssystem als Außenkriterium etabliert wird (Perlmann-Balme 2006:6).

Derartiges galt lange als überfällig für das Lehren, Lernen und Beurteilen von Sprachen, auch wenn trotzdem diverse Kritikpunkte hinsichtlich des GER definiert worden sind. Dennoch sind durch den GER gerade im Bereich Sprachprüfungen Veränderungen eingetreten, sodass sich Konsequenzen für Testinstitutionen, in unserem Fall im deutschsprachigen Raum, ergeben haben. Das äußert sich darin, dass existierende Prüfungen überarbeitet oder revidiert werden mussten, um den Vorgaben des GER und auch der ALTE und ihren Standards gerecht zu werden (Perlmann-Balme 2006:7).

In der vorliegenden Arbeit werde ich mich mit der Sprachaktivität schriftlicher Produktion von Lernern im Hinblick darauf befassen, welche Kompetenzen und kommunikativen Zwecke je nach Niveau und Testanbieter vom Lerner erwartet werden und wie diese zu entwickeln und fördern sind. Weiterhin ist die Fragestellung interessant, in welchem Kontext vom Lerner bzw. vom Prüfling verlangt wird, die schriftliche Produktion zu aktivieren und einzusetzen. Außerdem bleibt zu beantworten, ob der GER die geeigneten Grundlagen dafür schafft und wie diese zunächst von den Testanbietern und schließlich vom Endabnehmer Lerner realisiert werden. Zentraler Punkt ist zum einen die Validität der Bewertungskriterien schriftlicher Lernerproduktionen, um die konstante und konsistente Testvalidität insgesamt zu gewährleisten. Zum anderen ist die Vorgehensweise hinsichtlich der Korrektur und im weiteren der Bewertung durch Rater, die als das letzte Glied in der Testkette gelten, mittels Ratingverfahren und anhand verschiedener Kriterienkataloge entscheidend, um

schließlich der Verantwortung für die Beibehaltung der Testvalidität gerecht zu werden. Die im handlungsorientierten Ansatz des GER angeführten Kompetenzen und Kann-Beschreibungen werden in Kapitel 5 in der Diskussion der bestimmten Bewertungskriterien der verschiedenen Testanbieter berücksichtigt.

Des Weiteren werden die für die schriftliche Produktion nötigen Skalen des GER in Verhältnis zu den von den jeweiligen Testanbietern verwendeten Kann-Beschreibungen bzw. Deskriptoren gesetzt. Als mittelndes Instrument soll die speziell für die deutsche Sprache erarbeitete Fassung *Profile* (Glaboniat et al. 2002) samt ihren definierten auf dem GER beruhenden und umgesetzten Kann-Beschreibungen, soweit wie möglich, dienen. In dieser für die deutsche Sprache entstandenen Fassung sind zudem die empfohlenen sprachlichen Mittel, die Grammatik, verschiedene Textsorten und Lerner- und Kommunikationsstrategien aufgeführt. Wie der GER auch, so versteht sich auch *Profile Deutsch* nicht als verpflichtend. Lernziele und sprachliche Mittel werden hier zu einem Werkzeugkasten zusammengestellt und sollen vielmehr Richtlinien sein, die flexibel eingesetzt und zudem erweitert werden können (Perlmann-Balme 2006:10). Anregungen können hierdurch zum Beispiel für Curriculumentwicklung, Lehrkonzepten und Testentwicklung gegeben werden. Das sprachenpolitische Konzept des Europarats steht unter anderem auch dafür ein, dass ihre Mitgliedsstaaten mit ausreichenden Kommunikationskompetenzen ausgestattet werden, um untereinander interagieren zu können. Dadurch könnte es zu einer wachsenden Mobilität im Berufsbereich zum Beispiel innerhalb Europas kommen. Die geförderte Mehrsprachigkeit und Plurikulturalität soll ebenso kulturelle Barrieren und Vorurteile zwischen den verschiedenen europäischen Staaten abbauen (GER 2001:16). Der Europarat gibt den Mitgliedsstaaten Hilfestellung zur Anwendung neuer sprachlichen Programme und motiviert des Weiteren zur Innovation beim Sprachenlehren und - lernen, welche zudem weiterentwickelt werden soll, um die Kooperation zwischen Bildungsträgern verschiedener europäischer Länder zu fördern. Ein weiteres Ziel des GER besteht darin, die Vergleichbarkeit von validen Sprachqualifikationen oder Sprachzertifizierungen zu erlangen, d.h. eine Grundlage zu schaffen, Sprachzertifizierungen europaweit anzugleichen und anzuerkennen.

Der GER soll ebenfalls als Hilfsmittel für die Erstellung von standardisierten Sprachprüfungen und den dazu benötigten Bewertungskriterien fungieren, wobei deren Vergleichbarkeit letztlich nur dadurch erreicht werden kann, wenn Standards bei der Testentwicklung, -durchführung und –bewertung eingehalten werden (Perlmann-Balme 2006:6f.).

2.2 American Psychological Association

Die American Psychological Association (APA) gründete schon in den späten 40er Jahren ein Komitee für ethische Standards in der Psychologie und entwickelte daraus die ersten ethischen Prinzipien.[16] 1954 begann sie dann damit, sich mit dem Bereich der Entwicklung und Anwendung von Tests zu befassen und entsprechende Guides samt Standards in diesem Bereich zu veröffentlichen. Die erstmals 1966 und dann 1985 herausgegebenen „Standards for Educational and Psychological Testing“ und das dazugehörige Manual wurden zwischen 1991 und 1996 überarbeitet. Diese Standards wurden entwickelt und überdacht, um anderen Kriterien konsistent gegenüber zu sein. Die Absicht der *Standards* besteht darin, den ethischen Testgebrauch zu fördern und einen Maßstab bereitzustellen, mit dessen Hilfe die Testqualität evaluiert werden kann. Die APA ist eine exzellente Quelle für verschiedene testtheoretische Konzepte[17] und Qualifikationen, die sie für die Kompetenzen und den verantwortungsvollen Testgebrauch für wichtig hält.[18] Sie betont als optimale Voraussetzung hinsichtlich des Testgebrauchs Faktoren wie Wissen, Kenntnis, Fähigkeit, Schulung und Erfahrung. Für die APA bedeutet die Qualifikation eines Testbenutzers weniger eine Zertifizierung, als vielmehr das Aufzeigen von Kompetenz. Die seit August 2000 neu definierten und erprobten Richtlinien des APA-Vorstandes beschreiben zweierlei Kompetenzen. Zum einen bilden allgemeine Kompetenzen die Basis für den üblichen Testgebrauch. Der optimale und spezielle Testgebrauch bedarf aber speziellerer Kompetenzen. Anders ausgedrückt, sehen die Richtlinien der APA für die verschiedensten Kontexte und deren Gebrauch unterschiedliche Kompetenzen vor.

Die APA–Standards[19] werden im Kernkapitel der vorliegenden Arbeit der Analyse und Kritik der Bewertungskriterien schriftlichen Ausdrucks insofern hilfreich sein, indem die definierten Richtlinien als Basis verstanden werden, um das Gütekriterium der Validität zu unterstützen.

16 American Psychological Association. 1950. Ethical standards the distribution of psychological tests and diagnostic aids. American Psychologist 5, S. 620-626

17 Report of the Task Force on Test User Qualifications 2-88. Practice on Science Directorates APA. Approvedby the APA Council of Representatives. August, 2000.

18 DeMers, S.Y., Turner, S.M. (Cochairs), Andberg, M. Foote, W. Hough, L. Ivnik, R. Meier, S. Moreland, K. & Rey-Casserly, C.M. (2000). Report of the Task Force on Test User Qualifications. Washington, D.C.: Practice and Science Directorates, American Psychological Association - aus dem Original von mir übersetzt.

19 Die Literaturangabe oder der Verweis auf die „Standards der American Psychological Association“ wird in der Arbeit mit APA bzw. APA-Standard gekennzeichnet sein. Ersteres bezieht sich auf allgemeine Referenz während letzteres auf definierte Standards verweisen soll.

2.3 Association of Language Testers in Europe (ALTE)

Die Vereinigung Association of Language Testers in Europe (ALTE) setzt sich aus verschiedenen Testanbietern im Bereich der Fremdsprachen in Europa zusammen. Initialisiert wurde dieses Konzept 1989 anfangs durch die Universitäten von Cambridge und von Salamanca. Mittlerweile zählt ALTE 31 Mitglieder. Darunter fallen aus dem deutschsprachigen Raum neben der Telc GmbH, das Goethe-Institut und das TestDaF-Institut.

1991 begann ein langfristiges ALTE Rahmen-Projekt, in dem alle Mitglieder ihre Prüfungen nach dem gleichen Prinzip und Format definierten. Ziel der ALTE ist es, vergleichbare Sprachprüfungen und Zertifizierungen in Europa herzustellen. Um diese Homogenität unter den unterschiedlichsten Sprachprüfungen innerhalb Europas zu erreichen, hat ALTE einen Referenzrahmen der Niveaus entwickelt[20], welcher durch eine große Palette von Kann-Beschreibungen definiert wird. Der Rahmenplan der ALTE besteht auf den ersten Blick, wie der GER auch, aus sechs Niveaustufen hinsichtlich der Sprachbeherrschung. Dennoch sind es lediglich fünf, weil die erste Niveaustufe nicht nummeriert ist, sondern eben nur ein „Breakthrough" ist. An folgender globalen Tabelle sei der Vergleich beider Referenzrahmen deutlich gemacht:[21]

C 2 – ALTE 5	Fähigkeit, mit akademisch oder kognitiv anspruchsvollem Material umzugehen und Sprache mit gutem Erfolg auf einem Leistungsniveau zu benutzen, das in mancher Hinsicht fortgeschrittener sein mag als das eines durchschnittlichen Muttersprachlers.
C 1 – ALTE 4	Fähigkeit zu kommunizieren, mit Betonung darauf, wie gut etwas erledigt wurde im Hinblick auf Angemessenheit und Feingefühl und die Fähigkeit, mit nicht vertrauten Themen umzugehen.
B 2 – ALTE 3	Fähigkeit, die meisten Ziele zu erreichen und sich über eine Vielzahl von Themen auszudrücken.
B 1 – ALTE 2	Fähigkeit, sich auf begrenzte Weise in vertrauten Situationen auszudrücken und auf allgemeine Weise nicht-routinemäßige Informationen zu bewältigen.
A 2 – ALTE 1	Fähigkeit, mit einfachen, unkomplizierten Informationen umzugehen und der Beginn der Fähigkeit, sich in vertrauten Kontexten auszudrücken.
A 1 – ALTE Breaktrough	Elementare Fähigkeit, auf einfache Weise zu kommunizieren und Informationen auszutauschen.

Tabelle 4: Vergleich der Niveaustufen des GER und ALTE

20 Es werden seit 1998 Vergleiche und Korrelationsberechnungen der Kann-Beschreibungen und Skalen zwischen dem GER und ALTE angestellt

21 Durch statistische Verfahren wurde die Vergleichbarkeit der Stufen ALTE und GER erwiesen (www.alte.org)

Außerdem wurden 1994[22], ähnlich wie bei der APA[23], allgemeine Standards, der so genannte ALTE Code of Practice[24], für den gesamten Testprozess von Sprachprüfungen definiert. ALTE-Mitglieder bekennen und verpflichten sich dabei, diese bei der Testerstellung und – durchführung einzuhalten. Es wird von den Mitgliedern erwartet, dass ihre Prüfungserstellung und Qualitätskontrolle dokumentiert wird.

2.4 Das Goethe-Institut

Das Goethe-Institut e.V. wurde am 9. August 1951 in München gegründet und ist heute die bekannteste weltweit tätige Organisation zur Vermittlung deutscher Sprache und auswärtiger Kulturpolitik.[25] Das Institut, dem der Beiname *Institut zur Pflege der deutschen Sprache im Ausland und zur Förderung der internationalen kulturellen Zusammenarbeit*[26] beigefügt wird, ist also nicht nur Vermittler der deutschen Sprache, sondern leistet im Auftrag der Bundesrepublik Deutschland als gemeinnütziger Verein[27] auch einen großen Beitrag im Bereich der Kulturarbeit, der sich zum Beispiel in Ausstellungen zu Themen deutscher Geschichte oder der Organisation von Konzerten niederschlägt.[28]

Während der Zeit des so genannten Wirtschaftswunders wurden weitere Institute eröffnet. Schon 1951 wurden in Athen/Griechenland und in anderen Metropolen von im Ausland lebenden Deutschen erste Deutschkurse angeboten (Apelt o.J.:4). 1953 gab es in Deutschland das erste Fortbildungsangebot für ausländische Deutschlehrer. Der Schwerpunkt lag hier in erster Linie auf dem Erstellen von Lehrbüchern, Lehrplänen und Unterrichtsmethoden (Apelt o.J.:4). Zu erwähnen wäre an dieser Stelle, dass die ersten Lehrveranstaltungen für ausländische Studenten in Deutschland bereits im Sommersemester 1898 an der Friedrich-Wilhelms-Universität von Berlin eingeführt und angeboten wurden. Landeskundliche Inhalte wurden in diesem Zusammenhang in Deutschland also erstmals im Sommersemester 1903 Teil dieser Unterrichtsveranstaltungen.[29] Das Goethe-Institut erweitert seinen Tätigkeitsbereich von der Sprachförderung zur Programmarbeit (Vortragsreihen usw.) und zur Entwicklung neuer Lernmethoden oder Sprachtestentwicklungen, etwas was die Fusion mit Inter Nationes am 21. September 2000 zusätzlich bestärkt. Dieses Institut wurde 1952 gegründet, um die Präsenz Deutschlands im Ausland durch Informationsverbreitung deutlich zu machen.[30] Seitdem arbeitet das Goethe-Institut mit verschiedenen

22 www.alte.org

23 APA ist umfassender und bezieht sich nicht speziell auf Sprachstandstests

24 www.alte.org

25 Eckard Michels, Goethe-Institut, in: Historisches Lexikon Bayerns, URL:<http://www.historisches-lexikon bayerns.de/artikel/artikel_44721>

26 Satzung und Rahmenvertrag. Rechtliche Grundlagen des eingetragenen Vereins. Herausgegeben vom Goethe- Institut, München o.J., S. 6

27 http://de.wikipedia.org/wiki/Goethe-Institut

28 vgl. Kulturprogramme der Pädagogischen Verbindungsarbeit 1991 - 1997. Dokumentation. Herausgegeben vom Goethe-Institut, München 1997. Außerdem gibt das Goethe- Institut die Zeitschrift Goethe- Institut aktuell heraus, in der vierteljährlich die Kulturprogramme vorgestellt werden.

29 Günther, Roswitha: Das Deutsche Institut für Ausländer an der Universität Berlin in der Zeit von 1922 bis 1945. Ein Beitrag zur Erforschung des Lehrgebiets Deutsch als Fremdsprache. In: Beiträge zur Geschichte der Humboldt-Universität zu Berlin, Nr. 19. Berlin 1988

30 http://de.wikipedia.org/wiki/Goethe-Institut

universitären oder privaten Bildungsträgern im In- und im Ausland zusammen, indem unter anderem Projekte und Curricula erstellt werden.[31]

Seit seiner Gründung war das Goethe- Institut unabhängig von der Bundespolitik, obwohl es durch das Bundesaußenministerium finanziert wurde (vgl. Apelt o.J.:4). Durch diese Liquidität war dem Goethe-Institut der Zugang zu anderen Ländern möglich. Nicht unbeachtet darf die Tatsache gelassen werden, dass das Goethe- Institut in den letzten Jahren finanzielle Kürzungen, Umstrukturierungen und Schließungen von Zweigstellen hinnehmen musste.

Der zentrale Gedanke und das Motto des Goethe-Institutes besteht in der Vermittlung der deutschen Sprache und Kultur in Form einer Bereicherung der jeweiligen Kultur vor Ort. Alle Veranstaltungen der Institute haben den Hintergrund und das Konzept, dass sie sowohl für Freunde der deutschen Sprache und Kultur als auch für deutschneutrale Erstentdecker attraktiv sein können. Heute umfasst das weltweite Netz des Goethe-Instituts über 140 Kulturinstitute in 77 Ländern. Die insgesamt 15 Institute in Deutschland runden diese Arbeit ab und ermöglichen Auslandsaufenthalte der Sprachlerner in Deutschland.

Die deutsche Sprache im Ausland wird vom Goethe-Institut durch viele verschiedene Maßnahmen und Produkte betrieben, wie z.B. :

- Pädagogische Verbindungsarbeit zur Unterstützung des Deutschunterrichts in privaten und öffentlichen Institutionen im Ausland
- Entwicklung und Bereitstellung von Lehr- und Lernmaterialien
- Fortbildungsangebote für Lehrkräfte
- Sprachpolitische Aktivitäten
- Sprachkurse unterschiedlicher Zielgruppen

Anhand dieser Beiträge fördert das Goethe-Institut die deutsche Sprache und leistet somit auch einen Beitrag zur Förderung der Mehrsprachigkeit im Rahmen der europäischen Sprachpolitik. Dabei orientiert sich das Goethe-Institut an den Grundsätzen, die die Institutionen der Ständigen Arbeitsgruppe Deutsch als Fremdsprache (StADaF) beschlossen haben. Ein Beispiel hierfür wäre das Ziel „Vermittlung eines aktuellen Deutschlandbildes und Anregungen zu interkultureller Auseinandersetzung“. Die Arbeit der einzelnen Goethe-Institute im In- und Ausland beschränkt sich aber keinesfalls nur auf Sprachvermittlung. Die Kulturarbeit der Goethe-Institute im Ausland soll die Akzeptanz der deutschen Sprache fördern sowie dabei helfen, Vorurteile gegenüber Deutschland abzubauen.

Die meines Erachtens zentralste Aufgabe und Funktion aller Goethe-Institute ist die Bereitstellung von Prüfungsangeboten. Die ersten entstandenen Prüfungen (1963) waren die Oberstufenprüfungen KDS und GDS. In den letzten 50 Jahren sind aber immer wieder neue Prüfungen initialisiert, aber auch überarbeitet worden. Alle Prüfungen werden in der Zentrale in München erstellt.[32] Die Prüfungen decken die verschiedenen Niveaustufen des GER ab, die jeweils als aufbauend zueinander betrachtet werden können. Es muss im

31 Berthold Franke (Hrsg.): Jahrbuch 1998/1999 des Goethe-Instituts, S. 26

32 Writing Tasks: Pilot Samples. In: Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR: learning, teaching, assessment. 1995

Vorfeld betont werden, dass das Goethe-Institut die Qualitätsstandards der ALTE anwendet, wenn es um Testentwicklung, Durchführung, Bewertung und ihrer Analyse geht.[33] Aktuell bietet das Goethe Institut folgende Prüfungen an:[34]

Stufe des GER	Prüfungen des Goethe-Instituts
A1	Goethe-Zertifikat Start Deutsch 1, Goethe-Zertifikat Fit in Deutsch 1
A2	Goethe-Zertifikat Start Deutsch 2, Goethe-Zertifikat Fit in Deutsch 2
B1	Goethe-Zertifikat Deutsch, Goethe-Zertifikat Deutsch für Jugendliche
B2	Goethe-Zertifikat Deutsch für den Beruf, Goethe-Zertifikat B2
C1	Goethe-Zertifikat Prüfung Wirtschaftsdeutsch, Goethe-Zertifikat C1
C2	Goethe-Zertifikat Zentrale Oberstufenprüfung, Goethe-Zertifikat Kleines Deutsches Sprachdiplom
C2+	Goethe-Zertifikat Großes Deutsches Sprachdiplom

Tabelle 5: Prüfungen des Goethe-Instituts

Im Herbst 2007 wurden auf den Niveaus B2 und C1 weltweit die neu erstellten Prüfungen eingeführt. Genau diese zwei neuen Prüfungen des Goethe-Instituts werden neben der TestDaf-Prüfung zum Hauptgegenstand dieser Arbeit. Dabei werden die offenen Aufgabenformate und die erwarteten Textsorten des schriftlichen Ausdrucks und die dafür bereitgestellten Bewertungskriterien im Vordergrund stehen

33 Goethe-Zertifikat C1. Handbuch. Prüfungsziele. Testbeschreibung. 050707. S. 4f.

34 www.goethe. de

2.5 Das TestDaF-Institut

Das TestDaF-Institut ist eine gemeinnützige wissenschaftliche Institution, die von der Gesellschaft für Akademische Testentwicklung[35] mit der finanziellen Hilfe des Auswärtigen Amtes und des Bundesministeriums für Bildung und Forschung[36] ins Leben gerufen wurde. Der Hintergrund für die Gründung einer derartigen Instanz lag in der Diskussion, einen internationalen Hochschulzugangssprachtest für deutsche Universitäten zu entwickeln, der mit dem IELTS (International English Language Testing System) und dem TOEFL (Test of English as a Foreign Language)[37] verglichen werden könnte. 2001 wurde dann schließlich der standardisierte TestDaF-Test herausgebracht, bei dem es um den Sprachnachweis ausländischerer Studienanwärter bzw. Studienbewerber für die Hochschulzugangsberechtigung an deutschen Universitäten geht. Dabei wird der Sprachstand eines Kandidaten bezüglich des „akademischen Kontextes im oberen Leistungsspektrum“[38] überprüft. Der nach den vier Fertigkeiten in Subtests „getrennte“ TestDaF orientiert sich an einer mit drei verschiedenen Niveaustufen (TDN 3, TDN4, TDN5)[39] erarbeiteten Skala. Trotzdem orientiert sich der TestDaF sowohl an den Skalen des GER als auch an die der ALTE. Ein Charakteristikum dabei ist der hochschulbezogene Kontext bei den Stufenbeschreibungen.[40] An folgender Tabelle sei die Zuordnung angeführt:

GER	A1.1	A1.2	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1.1	C1.2	C2.1	C2.2
TestDaF							TDN 3		TDN 4		TDN 5	
ALTE	ALTE Breakthrough		ALTE 1		ALTE 2		ALTE 3		ALTE 4		ALTE 5	
TestDaF							TDN 3		TDN 4		TDN 5	

Tabelle 6: TestDaF im Vergleich zum GER und zu ALTE

35 Hierzu zählen: Hochschulrektorenkonferenz, DAAD e.V. Bonn, Goethe-Institut e.V. München, Fernuniversität/Gesamthochschule Hagen, Ruhr-Universität Bochum, Universität Leipzig, Fachverband Deutsch als Fremdsprache e.V.. Aus: www.testdaf.de

36 Grotjahn, R./Kleppin, K.: TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In: Germanistisches Jahrbuch der GUS „Das Wort“ 2000/2001, S. 267

37 IELTS und TOEFL sind die wichtigsten Sprachtests für englischsprachige Universitäten und Einrichtungen

38 Arras, U./Grotjahn, R.: TestDaf: Aktuelle Einwicklungen. Eine erweiterte Fassung eines Vortrages auf der 22. Arbeitstagung in Chemnitz, 28.02.2002.

39 TDN steht für TestDaF-Niveau (siehe auch Abkürzungsverzeichnis)

40 Grotjahn, R./Kleppin, K.: TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In: Germanistisches Jahrbuch der GUS „Das Wort“ 2000/2001, S. 26

Durch die verschiedenen Ergebnisse der Kompetenzen in den entsprechenden Subtests sollen sich Hochschulen einen „Eindruck“ vom Profil des Studienbewerbers verschaffen. Demzufolge kann je nach Profil jede Fakultät einer Universität ihrer Hochschulordnung entsprechend differenzieren und bedingt zulassen.[41] Erreicht ein Kandidat beispielsweise in allen vier Subtests die TDN-Stufe 4, so hat er die sprachliche Zugangsvoraussetzung für deutsche Hochschulen, wobei manche bereits Studienbewerber aufnehmen, die mindestens das TDN-Niveau 3 erreicht haben.

Da sich der TestDaf-Test als ein standardisierter Test versteht, der die Einhaltung der Gütekriterien der Testerstellung garantiert, wird er am TestDaF-Institut sowohl erstellt als auch bewertet.[42] Er wird mittlerweile in über 80 Ländern in mehr als 300 lizenzierten Testzentren abgenommen. Der TestDaF-Test wird in dieser Arbeit hinsichtlich des Subtests „schriftlicher Ausdruck“ untersucht werden. Bewertungskriterien und deren Realisierung werden im Rahmen der Validität beleuchtet werden.

41 Informationsmaterial des TestDaF-Instituts: Empfehlungen für Kurse und Materialien zur Vorbereitung auf die Prüfung TestDaF. 04/2005

42 www.testdaf.de

3 Modelle des Spracherwerbs

Zentrales Thema dieses Kapitels ist die fremdsprachliche Lernerproduktion im schriftlichen Ausdruck. Zunächst sollen Fragen hinsichtlich des Lernens einer Fremdsprache aufgestellt werden: Wie geht dieser Prozess vonstatten? Was geht im Lerner vor? Die wichtigsten Grundlagen und Hypothesen, die die Forschung des Zweit- und Fremdspracherwerbs geprägt haben, sollen die Grundlage zur Beantwortung dieser Fragen sein, indem sie kritisch gegenübergestellt, betrachtet und auf Relevanz hinsichtlich der Thematik dieser Dissertation untersucht werden. Weiterhin wird spezifisch auf die schriftliche Lernerproduktion und ihre Problematik im Fremdsprachenbereich eingegangen werden. Des Weiteren werden sich daraus, unter anderem Aspekte der sprachlichen oder auch kulturellen Identität und der Lernermotivation abzeichnen. Der zweite Teil dieses Kapitels wird sich dann mit der Definition des Kompetenzbegriffs beschäftigen, um zunächst die erforderlichen Sprachkompetenzen zu skizzieren und dann speziell auf die produktive Kompetenz Schreiben einzugehen. Kontrastiv dazu wird Bezug auf ein primärsprachliches Kernlehrcurriculum genommen, um darauf aufbauend die schriftliche Lernerproduktion, insbesondere in Testsituationen, samt ihren Strategien, Methoden und internen sowie externen Einflüssen zu beleuchten.

3.1 Der Spracherwerb

Spracherwerb ist der Prozess des Erlernens einer Sprache, was als das Spezifikum des Menschen gilt. Dieser ist prinzipiell in der Lage, jede Sprache zu erwerben, da er über einen angeborenen Sprachmechanismus verfügt. Die universale Sprachfähigkeit entfaltet sich im Laufe des Heranwachsens einer Person. Die Art und Weise wie Menschen Sprachen verarbeiteten und erlernten, vollzog sich über Millionen von Jahren bis vor einigen Jahrzehnten unsystematisch (Klein 1984:31). Die sich aus der Psycholinguistik entwickelte Spracherwerbsforschung versucht in den letzten Jahrzehnten eine Erklärung der Spracherwerbsprozesse zu geben, indem sie biologisch gegebene Sprachlernfähigkeiten, Charakteristika einer Person, Inputbedingungen oder vorausgesetztes Wissen für den Spracherwerb berücksichtigt.

3.2 Erstsprach- und Zweit- bzw. Fremdspracherwerb

3.2.1 Gesteuerter vs. ungesteuerter Fremdspracherwerb

In erster Linie wird zwischen den Lernkontexten (vgl. Bausch/Kasper 1979) Erstsprach- und Zweit- bzw. Fremdspracherwerb differenziert. Erstspracherwerb bezieht sich auf das Erlernen der Muttersprache, Zweit- bzw. Fremdspracherwerb differenziert zwischen gesteuertem und ungesteuertem Erlernen bzw. Erwerben einer weiteren Sprache. Das Erlernen einer zweiten Sprache ist differenzierter und daher muss schon die Definition dieses Rahmens per se entsprechend explizit gemacht werden. In der Literatur gibt es verschiedene Bezeichnungen dafür. Die Erforschung dieses Bereichs wird sowohl unter sprachlichen als auch unter außersprachlichen Gegebenheiten betrachtet. Während ersteres sich, global betrachtet, auf die Rolle und die Beziehung zweier Sprachen bezieht, deckt letzteres Faktoren wie Persönlichkeit oder Motivation ab (Egger 1995:81).

Der ungesteuerte Fremdspracherwerb ist mit dem nicht systematischen Lernen einer Fremdsprache gleichzusetzen (Klein 1984:28). Da es sich nicht um einen unterrichtsbegleiteten bzw. institutionellen Prozess handelt, wird dieses im Folgenden als Zweitspracherwerb bezeichnet werden. Ich lasse an dieser Stelle unberücksichtigt, dass auch im ungesteuerten Spracherwerb durch Interaktion gesteuert werden kann (Rösler 1995:150). Dieser natürliche Zweitspracherwerb erfolgt in der alltäglichen Kommunikation, zu der auch das Fernsehen oder das Zeitungslesen eines Migranten beispielsweise, gezählt werden können (Merten 1997:66ff.).

Beim gesteuerten Fremdspracherwerb, der für diese Arbeit interessant ist, wird eine zweite Sprache in der Regel außerhalb ihres normalen Verwendungsbereichs, das heißt nicht im Land, in dem sie Verkehrssprache ist, erlernt.[43] Im Gegensatz zum Zweitspracherwerb, der in erster Linie der Alltagsbewältigung dient und Fehler eine untergeordnete Rolle spielen (vgl. Rösler 1984), ist der Fremdspracherwerb[44] durch Strategieentwicklung zur Fehlervermeidung gekennzeichnet, wobei Fehler jedoch auch Anzeichen dafür sein können, dass die Sprache samt ihren Regeln erworben wird (Kielhöfer 1995:36).

Der handlungsorientierte Ansatz des GER lässt darauf schließen, dass es in Zukunft durch die wachsende Mobilität innerhalb Europas zu verschiedenen Erwerbsformen und entsprechenden Definitionen kommen wird. Hinsichtlich der Thematik der vorliegenden Arbeit könnte demnach die Homogenität von Lernern und ihren Produktionen zukünftig durch verschiedene Mischformen in Frage gestellt werden, da die unter Kapitel 3.2 vorgestellten Hypothesen zum Zweit- bzw. Fremdspracherwerb keine Entsprechung im definierten Sinne hätten.

43 Angemerkt sei an dieser Stelle, dass sich die wenigsten Zweitspracherwerbshypothesen aufgrund ihrer sehr theoretischen Grundlage auf den Fremdspracherwerb außerhalb des Zielsprachenlandes beziehen

44 Synonym dazu werden in dieser Arbeit als Variation des Begriffs Fremdspracherwerb auch Zielsprache, fremdsprachlich, u.ä. verwendet werden

3.2.2 Die Motivation als Reizelement beim Sprachenlernen

Ein entscheidendes Element bei dem Erlernen einer Sprache ist die Motivation. Dittmar (1995) spricht in diesem Zusammenhang von Lernenergie, wobei ihre Intensität und ihr Ausmaß das Spracherwerbsresultat bestimmt (Dittmar 1995:109 ff.). Die Gründe, aus denen man sich mit einer Sprache auseinandersetzt, können ganz unterschiedlicher Natur sein. Einerseits kann sowohl die Akkulturationsbereitschaft[45] als auch ein externer, sozialer Zwang Anlass sein, sich eine Sprache anzueignen. In diesem Fall ist die zu erlernende Sprache das einzige Kommunikationsmedium des Umfeldes. Dies impliziert aber nicht nur die Verständigung, sondern auch die Aneignung von Werten, Normen und Interpretationen dieser für die Gesellschaft typischen Sprache. In diesem Sinne durchläuft man gezwungenermaßen oft eine zweite Sozialisation im Bereich der Persönlichkeitsentwicklung, der Handlungsfähigkeit und konkreten Interpretationsschemata (Steinmüller 1995:161ff.). Auf der anderen Seite gibt es den Fall des mehr oder weniger freien Willens, eine weitere Sprache zu erwerben. Setzt man sich unter diesen Bedingungen mit einer neuen Sprache auseinander, so unterliegt dieser Prozess konkreten Prämissen, Funktionen und Zwecken (Beispiel Student in Griechenland). Ein Motiv bzw. eine Zielsetzung für die Spracherlernung könnte in diesem Fall der Sprachnachweis zur beruflichen Qualifikation oder die Attraktivität der Weiterbildung an einer Universität im deutschsprachigen Raum sein. In diesem Zusammenhang soll man gemäß des GER darüber nachdenken, „in welcher Beziehung kommunikative Aktivitäten und Lernaktivitäten zu den Antrieben, Motivationen und Interessen der Lernenden stehen“ (GER 2001:57).

Der Aspekt der Persönlichkeitsbildung ist eng mit dem Erlernen einer Fremdsprache gekoppelt. Neben den funktionalen und instrumentellen Zwecken gibt uns eine Sprache Aufschluss über Land und Leute. In ihr spiegeln sich Werte und Normen der entsprechenden Gesellschaft wider.

Im Fall des Fremdspracherwerbs durchliefe man nicht wie im ersten Beispiel eine weitere Sozialisation, sondern würde seinen Horizont erweitern und bisher unbekannte Normen und Wertvorstellungen modifizieren. Erst durch Sprache können Menschen miteinander in Interaktion treten und ihre Umwelt gestalten. Phipps und Gonzales sprechen in diesem Zusammenhang von *Languaging* (2004:2). Sprachenlernen erweitert den Bewusstseinshorizont der Lernenden in Richtung Toleranz, Verständnis des Anderen und fördert somit den Abbau existierender Vorurteile und damit in Zusammenhang stehender Lernblockaden (Phipps/Gonzalez 2004:168):

> „To be intercultural is to be beyond the captivities of culture“

45 zum Begriff der Akkulturation vgl. Herskovits, M.J.: Acculturation. New York, 1938

3.3 Hypothesen zum Zweitspracherwerb[46]

In der Zweitspracherwerbsforschung gibt es mehrere Erklärungsansätze, die sich in ihren Voraussetzungen klar voneinander abheben. Psycholinguistisch sollen Spracherwerbsabläufe und -sequenzen beschrieben und erklärt werden, indem die biologisch existierende Sprachlernfähigkeit und andere Faktoren den Rahmen bilden. Im Vordergrund steht entweder die zu erwerbende Sprache oder der Lerner und seine soziokulturelle Situation. Die wichtigsten Hypothesen werden im Folgenden kurz vorgestellt, ohne Anspruch auf Vollständigkeit. Ich beschränke mich hierbei auf relevante Teile der Hypothesen, die unmittelbar mit der Thematik dieser Arbeit in Beziehung zu stehen scheinen.

3.3.1 Die Kontrastivhypothese

Von Charles C. Fries (1947) initiiert und von Robert L. Lado (1957) fortgeführt, gilt diese Hypothese behavioristischen Ansatzes als die erste entwickelte (Merten 1997:73). Ihre ursprüngliche starke Version lautete sinngemäß:

> „Die Grundsprache des Lerners beeinflusst seinen Erwerb einer Zweitsprache in der Weise, dass in Grund- und Zweitsprache identische Elemente und Regeln leicht und fehlerfrei zu erlernen sind, unterschiedliche Elemente und Regeln dagegen Lernschwierigkeiten bereiten und zu Fehlern führen“ (Bausch/Kasper 1979:5).

Es wird ersichtlich, dass es hier nicht um ein lernerzentriertes Modell geht, das heißt im Mittelpunkt steht nicht der Spracherwerber, sondern Basis- und Zielsprache stehen im Vordergrund und machen demnach den Untersuchungsgegenstand aus. *Kontrastiv* bedeutet in diesem Zusammenhang das Gegenüberstellen zweier Sprachsysteme, um Gemeinsamkeiten und Unterschiede zu ermitteln, die dann durch Systematisierung den Lernerfolg eines Lerners im Fremdsprachenunterricht gewährleisten sollen. Juhász (1970:9), ein wichtiger Vertreter der Kontrastivhypothese, spricht in diesem Sinne von ,,Interferenzen" oder ,,negativem Transfer" bei fehlerhaften „Übertragungen“ und von ,,positivem Transfer", wenn die Umsetzung fehlerfrei erfolgt. Decken sich Strukturen der Erst- und der Zielsprache, so ist eine zweite Sprache gemäß der Kontrastivhypothese leicht erlernbar.

Dem „Input-Outputverhalten“ (Bausch/Kasper 1979:4) der Kontrastivhypothese kann nur im syntaktischen Bereich eine Bedeutung beigemessen werden, denn inhaltliche oder gar thematische Gegenüberstellungen zweier Sprachen sind nicht realisierbar. Die kontrastive Spracherwerbstheorie kann sich für die Didaktik sicherlich teilweise als nützlich erweisen. Der Anspruch jedoch aus dieser prognostischen Version Strukturidentitäten und Strukturdivergenzen zu ermitteln und daraus bestimmte Lernprozesse zu erwarten, erweist sich als nicht valide (Wienold 1973:93ff.). Dieser behavioristische Ansatz sagt beispielsweise nichts über zur Verfügung gestelltes Lernmaterial aus, mit dessen Hilfe man eine Sprache erwerben kann. Es geht hier einzig und allein um die sprachlichen Charakterzüge und Eigenschaften. Da es in diesem Ansatz prinzipiell um das „Umlernen“ der Muttersprache geht, kann er für unsere Arbeit nicht in Anspruch genommen werden, um schriftliche Lernerproduktionen über den Kontrast von

46 Diese Hypothesen machen keine Unterscheidung zwischen Zweit- und Fremdsprache. Es geht hier um eine weitere Sprache, die neben der Muttersprache erworben wird

erster und zweiter Sprache zu definieren (Kielhöfer 1995:35ff.). Auch Edmonson und House (1993:210) führen einige Schwachstellen der kontrastiven Analyse an. Selbst ein Kontrastmangel zwischen zwei Sprachsystemen könne zu interlingualen Fehlern führen. Fehler könnten aber auch intralingualer Natur sein, d. h. sowohl in der Zweit- als auch in der Fremdsprache selbst präsent sein. Es treten schließlich auch jene Fehler auf, die gänzlich unabhängig von beiden Sprachsystemen sind und „als kreative Lernerschöpfungen gesehen werden könnten, die Rückschlüsse auf Lernprobleme, Lernprozesse, Verarbeitung von Erklärungen und Entstehung eigener Lernerregeln" (Kielhöfer 1995:35ff.) zulassen. Ein letzter Kritikpunkt besteht darin, dass die Kontrastivitätshypothese zur Generalisierung aller Lerner gleicher Erst- bzw. Muttersprachen neigt. Es zeigt sich aber, dass diese Lerner nicht alle identische Fehler verzeichnen und somit keine allgemeingültigen Regeln aufgestellt werden können. Die abgeschwächte Form der Kontrastivhypothese fungiert mittlerweile nur zur „Erklärung von Lernschwierigkeiten und -fehlern" (Apeltauer 1987:32) und bietet damit eine Grundlage für das Begreifen der „Fehlergenese" (Corder 1973:293), bedenkt aber nicht den Fall, dass sich Strukturunterschiede zwischen den Sprachen nicht bedingt in der Lernschwierigkeit oder in Fehlern äußern müssen, sondern vielmehr in der Vermeidungsstrategie.

Bezogen auf die vorliegende Arbeit impliziert dieses, dass die wirkliche Kompetenz[47] eines Testkandidaten dadurch nicht ermittelt werden könnte. Unsicherheit darüber, ob eine Struktur korrekt angewandt wird, kann demnach dazu führen, dass man sich als Prüfling weniger komplexer syntaktischer Zusammensetzungen bedient. Derartige „verdeckte" oder auch „latente" Fehler, die durch Vermeidungs- oder Übergeneralisierungsstrategien nicht auf den ersten Blick zum Vorschein kommen, sollen in Kapitel 5 anhand der Kompetenzbeschreibungen des GER diskutiert werden, indem noch zusätzlich ausführlich Bezug auf die Bewertungskriterien und der tatsächlichen und existierenden Sprachkompetenz einer Person genommen werden soll (Rieck 1980:44).

3.3.2 Die Monitor-Theorie

Der Begründer Stephen D. Krashen sieht in dieser Theorie die Möglichkeit die Relation zwischen ungesteuertem und gesteuertem Spracherwerb zu definieren. Krashen untersuchte hauptsächlich erwachsene Lerner, die seiner Meinung nach über zwei verschiedene Systeme verfügen, um eine Sprache zu erlernen: den unbewussten Spracherwerb und das bewusste Sprachlernen. In den Mittelpunkt stellt Krashen das Erwerbsphänomen, d. h. die unbewussten Prozesse des Lerners. Demnach bedarf es keiner Regelaufstellung. Der bewusste Spracherwerb benötigt den so genannten „Monitor", mit dessen Hilfe der Lerner sein Wissen abspeichert und bei Bedarf abrufen kann. In Krashens radikaler und viel diskutierter Theorie gibt es verschiedene Lerntypen, die den Monitor je nach Notwendigkeit unterschiedlich nutzen (vgl. Krashen 1985).[48]

47 Der Begriff der Kompetenz wird in Kapitel 3.6 definiert

48 http://www.sdkrashen.com/SL_Acquisition_and_Learning/index.html

In „The Natural Approach. Language Acquisition in the Classroom" (vgl. Krashen/Terell 1983) werden fünf Thesen skizziert, um Krashens Ansatz zu festigen:

Spracherwerb und Sprachlernen: Es wird zwischen dem unbewussten Spracherwerb und dem Sprachenlernen unterschieden. Unbewusster Spracherwerb findet in realer Kommunikation statt, während Sprachenlernen mit bewusstem Einprägen von Sprachstrukturen in Verbindung gebracht wird. Da nach Krashen der unbewusste Spracherwerb weitaus effektiver als das Sprachenlernen ist, sollte der Fremdsprachenunterricht darin bestehen, Sprachaktivitäten zu stimulieren, statt Regelwissen einzuüben (Apeltauer 1987:9).

Natürliche Ordnung: Die Reihenfolge, in der grammatische Strukturen erlernt werden, ist vorhersehbar. Der Fremdsprachenunterricht berücksichtigt allerdings nicht diese zeitliche und natürliche Erwerbsreihenfolge, sondern richtet sich nach den zu behandelnden sprachlichen Phänomenen eines Curriculums oder Lehrwerks (Dulay/Burt/Krashen 1982:17).

Der Monitor: Zweitsprache wird über das unbewusste Erwerbssystem initiiert. Der Monitor als Kontrollinstanz und bewusstes System schaltet sich nur dann ein, wenn zum Beispiel eine notwendige korrekte syntaktische Form von Bedeutung ist.

Das Input: Sprachen werden erlernt, indem man verständliches Input bekommt. Je nach Niveau ist das Input demnach mehr oder weniger erforderlich. Einfache Sprachstrukturen stellen zum Beispiel das ideale Input für jemanden dar, der sich noch in der Grundstufe seiner Zielsprache befindet.

Der affektive Filter: Affektiv wird im Sinne dieser Theorie zum Beispiel die Lernermotivation genannt. Ideale Sprachlernsituationen hemmen Angst oder Scheu vor der Zielsprache.

Wissenschaftlich lässt sich Krashens Theorie nicht nachweisen, denn es bleibt unklar, welcher der beiden angeführten Prozesse gerade abläuft. Obwohl sie unüberprüfbar ist, kann sie dennoch hilfreich für den Fremdsprachenunterricht sein, da affektive Faktoren, wie die Lernermotivationen, eine Hauptrolle spielen (Kohn 1990:18ff.). Die Krashen-Theorie basiert auf der Annahme, dass der Zweit- bzw. Fremdspracherwerb bewusst vom Lerner beeinflussbar ist. In der Literatur ist diese Annahme aber bislang nicht validiert worden. Im Zusammenhang der vorliegenden Arbeit ist diese definierte Theorie in erster Linie irrelevant, denn die verschiedenen von Krashen benannten „User" könnten nicht den gleichen Sprachniveaus zugeteilt werden. Der „Monitor" ist ein latentes Kontrollelement, wodurch das Bewusstsein des Spracherwerbs meines Erachtens in Frage gestellt werden kann, denn wie Butzkamm (1989:97) definiert „ist Bewusstsein kein einheitlicher Zustand, der entweder da ist oder nicht".

Trotzdem ist diese Theorie auf einer anderen Art und Weise für die vorliegende Arbeit nützlich. Die bereits angeführten Thesen Krashens könnten den Hintergrund für die hier im Mittelpunkt stehende Textproduktion und die dafür nötigen Kompetenzen bereit stellen. Im entsprechenden Abschnitt werde ich Krashens Monitor-Theorie und seine fünf Thesen mit den Textproduktionsmodellen koppeln.

3.3.3 Identitätshypothese

Noam Chomskys Theorie, jeder Mensch habe einen angeborenen Spracherwerbsmechanismus, ist der Wechsel von den behaviorististischen zu den kognitiven Spracherwerbstheorien. Demnach spielt es keinerlei Rolle, ob ein Lerner Sprachkenntnisse hat, denn sowohl bei der Erst- als auch bei der Zielsprache handelt es sich um die gleichen grammatikalischen universalen Strukturen, die der Mensch als genetische Information mitbringt. Folglich ist der Ablauf jedes Spracherwerbs nach dem gleichen Muster aufgebaut (Kupfer-Schreiner 1994:40). Corder (1967) spricht in diesem Zusammenhang von einem „eingebauten Lehrplan", welcher den Lerner dabei unterstützt, Hypothesen über das Regelwerk der Zielsprache zu bilden. Die Regeln und Elemente aus syntaktischer und morphologischer Perspektive werden also durch einen aktivierten angeborenen mentalen Prozess sowohl in der Erst- als auch in der Zielsprache gleichermaßen erworben. Die Erstsprache hat laut dieser Theorie keinerlei Einfluss auf das Erwerben einer weiteren Sprache. Obwohl verschiedene empirische Studien ähnliche Entwicklungssequenzen im syntaktischen Regelerwerb aufgewiesen haben[49], lässt sich diese Theorie trotzdem nicht validieren, da der völlige Ausschluss der Erstsprache nicht bewiesen werden kann.[50] Der Erwerbsprozess verschiedener Strukturen, z. B. der Plural oder die Negation, erfolgt zu einem Teil aus einer Variation beider Sprachformen (Klein 1984:36ff.). An dieser Stelle kann der Aspekt der kognitiven und sozialen Entwicklung angesprochen werden. Jemand, der eine Zweitsprache erlernt, kennt gemäß seiner kognitiven Entwicklung (Piaget) in aller Regel die semantischen Konzepte von Wörtern bereits von seinem Erstspracherwerb. In diesem Sinne werden diese, in verkürzter Form ausgedrückt, auf die Zielsprache übertragen. Ginge man in der Tat von universalen Strukturen aus, so wäre die Diskussion bezüglich der zu erwerbenden und notwendigen Kompetenzen im Fremdsprachenbereich nicht gegeben. In der in dieser Arbeit im Mittelpunkt stehenden schriftlich produktiven Lernerkompetenz gäbe es demnach keinen Unterschied zur Primärsprache. Auf diesem Argument basierend kann diese Theorie nicht vertreten werden.

3.3.4 Interlanguage-Hypothese

Ein differenzierteres Modell des Zweitspracherwerbsprozesses ist die ‚Interlanguage-Hypothese. Man geht davon aus, dass der Lerner beim Erwerb einer zweiten Sprache ein spezifisches Sprachsystem, die so genannte Interlanguage entwickelt. Dieses Sprachkonstrukt setzt sich aus Merkmalen und Strukturen zusammen, die sowohl Rekonstruktionen beider Sprachen (Erst- und Zielsprache) aber auch eigene unabhängige Assoziationen eines Lerners beinhalten. Diese Hypothese macht als erste den Versuch, auch sozialpsychologische Faktoren einzubeziehen (Bausch/Kasper 1979:15ff.).

Larry Selinker (1972), der den Begriff dieser Hypothese entscheidend geprägt hat, betont die Regelhaftigkeit des Zweitspracherwerbs und kann systematisch erscheinende Fehler erklären, indem er die Charakterisierung der ,,Interlanguages" bzw. spezifischer Lernersprachen durch fünf verschiedene psycholinguistische Prozesse bestimmt sieht (Bausch/Kasper 1979:23ff.):

49 Verschiedene empirische Studien mit dem Ziel verschiedene Lernergruppen gegenüberzustellen versuchten die Identitätshypothese zu validieren (vgl. Dulay/Burt 1974)

50 Im Weiteren wird dieses durch die Resultate empirischer Studien mittels introspektiver Verfahren deutlich werden, z. B. bei Krings (1986)

- Der Lerner übernimmt Regeln, Muster und Gewohnheiten aus der Erstsprache und überträgt sie auf die Zweitsprache. (language transfer)
- Der Lerner wendet bestimmte falsche Strukturen an, die durch ungeeignete Lernmaterialien oder anderen Regelbildungen entstehen. (transfer of training)
- Der Lerner entwickelt eigenständig Regeln und Strategien, die er überprüft, bestätigt oder gar revidiert. (strategies of second language learning)
- Der Lerner versucht seine unzulänglichen und nicht ausreichenden Sprachkenntnisse in einer bestimmten Kommunikationssituation durch Strategien zu kompensieren. Dabei geht es nicht um die perfekte Beherrschung grammatikalischer Strukturen, sondern um die ausschließlich verständliche Kommunikationsfähigkeit. (strategies of second language communication)
- Der Lerner wendet erworbene Regeln auch auf Ebenen an, für die sie nach zielsprachlichen Normen nicht gültig sind. (overgeneralisation of target language material)

Entscheidend ist hier der Begriff der ,,Fossilierung". Ist der Lerner nach Kohn (1990:13) der Ansicht, er beherrsche die Sprache ausreichend, da er im Stande ist gut zu kommunizieren, vernachlässigt er es, Fehler zu korrigieren und fällt dadurch oftmals in ein früheres Stadium seiner „Interlanguage" (so genanntes „back-sliding"). Nach Klein (1984:40) beruhen „Interlanguages" bzw. Lernersprachen auf einer zweifachen Systematik. Jede einzelne Lernersprache besitzt eine innere Systematik, auch wenn sie von vielen instabilen Komponenten geprägt sein kann.

Die Interlanguage-Theorie räumt dem Lerner ein, in bestimmten Phasen seines Spracherwerbs Fehler zu machen und Zwischensprachen zu benutzen. Ziel bleibt hier dennoch die Perfektionierung der Zielsprache. Der gesamte Spracherwerb ist im Prinzip nichts Anderes als eine Reihe von Übergängen von einer Lernersprache zur nächsten. Diese Abfolge definiert demnach die Systematik, der eine Lernersprache zugrunde liegen kann. Lernersprachen sollten nicht als fehlerbehaftete Formen der Zielsprache angesehen werden. Sie sind ein eigenes Ausdruckssystem. Die Variabilität, Dynamik und Durchlässigkeit für Regeln und Strategien ermöglichen die stufenweise Annäherung an die Zielsprache (Apeltauer 1987:34).

Fazit

All diese globalen Erklärungsansätze stimmen darin überein, dass es sich beim Zweit- bzw. Fremdspracherwerb um einen komplexen Vorgang handelt. Der Lerner stellt eigene Hypothesen über die zu erlernende Sprache auf, die er dann bestätigen oder revidieren muss (Merten 1997:78ff.). Klein (1984:49) verweist darauf, dass Menschen einen Sprachverarbeiter besitzen, sodass Sprachproduktion und Sprachverstehen an das jeweils zu verarbeitende Material angepasst werden kann. Trotzdem bleiben bei diesen Hypothesen die sprachliche und soziale Realität eines Lerners völlig unberücksichtigt. Im Mittelpunkt stehen die Sprache und deren Erwerbsprozess. Die Zweitspracherwerbsforschung sollte sich demnach mit allen für das Lernen wichtigen Faktoren (linguistische, soziale, entwicklungspsychologische Aspekte) befassen. Jede Theorie des Zweitspracherwerbs versucht, die bekannte Tatsache zu erklären, warum der Erwerb einer zweiten oder auch dritten Sprache in den meisten Fällen weit vor dem Niveau stehen bleibt, dass Kinder beim Erwerb ihrer Muttersprache erreichen. Zur Beantwortung dieser Frage ist es zuallererst notwendig, viele Erwerbsprozesse aufzuzeichnen, zu analysieren, um die wirksamen Faktoren im Erwerbsprozess zu finden und Vorschläge für die Zweitsprachvermittlung formulieren zu können. Auch wenn in der zurückliegenden Zeit wichtige Erkenntnisse zum Zweitspracherwerb gewonnen wurden, kann die oben gestellte Frage noch nicht zufrieden stellend beantwortet werden. Betrachtet man den Zweitspracherwerb unter systematischen Gesichtspunkten, kann man die kaum überschaubare Vielzahl an Faktoren auf einige wenige eingrenzen. Im Kern geht es um folgende Größen, die im Zweitspracherwerb eine wichtige Rolle spielen: zunächst das Verhältnis schon erworbener Sprachen zu der zu lernenden, sodann die Zweitsprache als Lernobjekt mit ihren Strukturen und Regeln, weiterhin die biologischen Grundlagen und die psychischen Mechanismen und Strategien im Erwerbsprozess sowie schließlich das sprachliche Handeln der Lernenden in der Kommunikation mit Sprechern der Zielsprache. Es dürfte weiterhin anerkannt sein, dass diese vier Grundgrößen den Zweitspracherwerb beeinflussen. Allerdings gibt es wenig Übereinstimmung in der Bestimmung ihres Gewichts und der daraus ableitbaren Konsequenzen für die L2-Vermittlung. Der Zweitspracherwerbsprozess kann nicht bei allen Lernern gleichermaßen ablaufen, denn verantwortlich und eine tragende Funktion hierbei haben die divergierenden kognitiven Entwicklungen, sozialpsychologische und affektive Faktoren, die entweder positiv oder negativ für den Erwerb der Zielsprache sein können (Merten 1997:90ff.).

Zweitspracherwerb ist ein nie endender Prozess, der in einzelnen Schritten abläuft. Er impliziert nicht nur das Erlernen eines fremden Regelsystems. Es ist das anfangs langsame Herantasten und Hineinfühlen in eine fremde Sprache, Kultur und Gesellschaft. Wer sich in einer Sprache äußern und in einer anderen Sprache verstehen will, muss zweimal den Bezugsrahmen wechseln: den kulturellen und den sprachlichen (Steinmüller 1995:161ff.).

Nach Hüllen (1983:164) wird der Lerner als Ganzes beim Spracherwerb und durch den Sprachkontakt beeinflusst:

„(...)Der ganze Mensch lernt- mit allen Befindlichkeiten Bedingungen seines Körpers, seiner Gefühle, seiner intellektuellen Zurüstung, seiner sozialen Situation, mit allen darin aufgeschichteten Lebenserfahrungen und daraus abgeleiteten Lebenserwartungen“.

Die Zweitspracherwerbsforschung und die Sprachlehr- und Sprachlernforschung sollten sich mehr aufeinander abstimmen. Erstere sollte sich mehr mit der Praxis hinsichtlich des gesteuerten Fremdspracherwerbs beschäftigen, um dann letzterer die Basis zu liefern, sich mehr im Bereich des Sprachlernens im In- und Ausland samt diversester gesteuerter bzw. ungesteuerter sozial bedingter Mischformen zu bewegen. Testerstellungen und Konzeptionen hinsichtlich von Sprachkompetenz sollten die verschiedenen Dimensionen der Zweitspracherwerbsforschung berücksichtigen, um einen geeigneten Rahmen zu bilden. Der GER spricht im Sinne Hüllens (1983) lediglich von einem „handlungsorientierten Ansatz“ und sieht den Sprachenlerner als „sozialen Akteur“ innerhalb des europaweiten Kontinuums. Die Frage des Anspruches in Sprachtests, die auf dem GER beruhen, beantwortet meines Erachtens nicht die Frage der Sprachnorm. Zwar werden Kann-Beschreibungen definiert und teilweise empirisch skaliert, aber dennoch bleibt es zu klären, ob Spracherwerbsprobleme oder auch –regeln berücksichtigt worden sind, um bestimmte Kompetenzen vorauszusetzen und zu definieren. Es muss zunächst und erstrangig der Frage nachgegangen werden, was jede Kompetenz impliziert und worin sie bestehen soll.

Die im Vorfeld vorgestellten wichtigsten Hypothesen zum Zweitspracherwerb können in dieser Sache nicht alle Anwendung finden. Die grundlegendste Theorie bezüglich des Fremdspracherwerbs im Sinne dieser Arbeit scheint zunächst die Interlanguage-Hypothese zu sein, da die wichtigste Norm bei dieser hierbei das zielsprachige Input ist (Blommaert/Lutjeharms 2003:126). Unsicherheiten (oder grammatische Fehler, die der Korrektheit der Zielsprachennorm nicht genügen) auf einem Niveau der Interlanguage können Aufschluss darüber geben, in welchem Maße der Lerner bereits mit der zielsprachlichen Norm vertraut ist oder nicht (Kohn 1990:31, 50). Ähnlichkeiten zur Funktion von Interlanguages sind meines Erachtens auch in den Kann-Beschreibungen des GER zu finden, wobei die Abstufungen zwischen den Niveaubeschreibungen als Stadien bzw. Interlanguages betrachtet werden könnten. Inwieweit sich diese Assoziation letztlich durch die Benutzung von Kann-Beschreibungen für die Bewertung schriftlichen Ausdrucks deckt, wird im 5. Kapitel noch ausführlich zu sehen sein. Die Monitor-Theorie kann mittels seiner definierten Facetten hilfreich für das Verständnis von Schreibprozessmodellen sein, wie der nächste Abschnitt verdeutlichen wird.

3.4 Der schriftliche Ausdruck in der Fremdsprache

Frühere Fremdsprachenmethodenkonzepte rechneten die Fertigkeit des Schreibens lange Zeit zur vierten und damit neben Leseverstehen, Mündlichem Ausdruck und Hörverstehen zur letzten (Teil)Kompetenz (Krings 1989:377ff.). Im Laufe der Zeit stieg aber das Interesse an der Schreibforschung, und damit rückte die Fertigkeit Schreiben sowohl in der Mutter- als auch in der zu erlernenden Fremdsprache in den Mittelpunkt.

Seit den 70er Jahren verfolgt die sich etablierte und weiter etablierende Schreibdidaktik hauptsächlich das Produkt des Schreibens, d. h. sie untersucht die Qualitäten von Lernerproduktionen. Die Schreibforschung ist im Gegensatz zur Schreibdidaktik bemüht, durch empirische Untersuchungen und Schreibprozessmodelle den mentalen Prozessen während der Textproduktion und weiteren Einflussfaktoren auf den Grund zu gehen (Molitor-Lübbert 1989:278ff.). Anfang der 80er Jahre machte sich ein Interesse hinsichtlich der Thematik Sprach- und Textproduktion sowohl in der Mutter- als auch in der Fremdsprachenlehrforschung bemerkbar (Antos/Krings 1989:3ff.). Crystal (1987:180) merkt an dieser Stelle bezüglich der Auseinandersetzung schriftlichen Ausdrucks an:

„(...) The analogous study of written language is less advanced, but has just a promising future."

Ich werde mich in diesem Teil der Arbeit der fremdsprachlichen Textproduktion widmen. Ziel der verstärkt in den Mittelpunkt rückenden „fremdschreiblichen" Kompetenz ist es, wie auch in der Muttersprache, etwas auszudrücken oder mitzuteilen. Etwas selbst produktiv schriftlich zu verfassen, kann als die komplexeste und die am schwierigsten zu erlernende sprachliche Teilkompetenz in der Zielsprache bezeichnet werden. Ich definiere den schriftlichen Ausdruck in der Zielsprache ganz bewusst als sprachliche Teilkompetenz, denn meiner Ansicht nach erfordert diese weitere Kompetenzen, die aber eigentlich latent zu sein scheinen. Die diversen Kompetenzen sind nach Rost (2004) latente Variablen, die vorhanden sind und die Einfluss auf das beobachtbare Ergebnis haben. Ein bestimmtes Testverhalten wird auf eine oder mehrere latente Variablen zurückgeführt. Items bzw. Aufgaben werden als manifeste Variablen betrachtet, wobei die beobachteten Zusammenhänge unter ihnen auf den Einfluss der latenten Variablen zurückzuführen sind (Rost 1996:30). Die Aufgabenformate standardisierter Prüfungen können beispielsweise rezeptivlastig sein, das heißt der Arbeitsauftrag muss rezipiert werden können, damit mit der schriftlichen Produktion begonnen werden kann. Die Problematik einer schriftlichen Lernerproduktion ergibt sich außerdem nicht unbedingt aus der Fähigkeit einer logischen Textgliederung, sondern aus der Erwartung, einen normgerechten Text in fremder Sprache zu verfassen (Börner 1989:351ff.). Normgerecht einen Text zu produzieren, heißt gemäß eines allgemein anerkannten Standards zu verfahren. Im vorliegenden Fall gilt es den Standard für den schriftlichen Ausdruck auf den Niveaus B2/C1 zu bestimmen.

Es muss aber zunächst definiert werden, was ein Text ist. Es gibt in der Textlinguistik verschiedene Ansätze, die den Begriff Text definieren, abgrenzen und klassifizieren. Während der sprachsystematische Ansatz die syntaktische Beziehung zwischen Sätzen ausdrückt (Brinker 2001:14ff.), ist ein Text im Sinne des kommunikationsorientierten Ansatzes funktions- und themenabhängig (Schmidt 1973:150ff.). Diese Ansätze konzentrieren sich jeweils auf unterschiedliche Merkmale, schließen sich dennoch nicht aus. Ein Text wird in der Textlinguistik zunächst als „das Produkt aus der Verbindung

mehrerer Sätze zu einem Ganzen begriffen" (Linke/Nussbaumer/Portmann-Tselikas 2004:215). Um zu diesem Schluss zu kommen, wird das Augenmerk insofern auf die systematischen Bezüge zwischen den Sätzen gerichtet, dass unterschiedliche Formen grammatischer Verknüpfungen und Bezug schaffende sprachliche bzw. kohäsionsstiftende Mittel in den Mittelpunkt rücken. In der fortlaufenden Entwicklung und Umorientierung der Textlinguistik ist man aber noch einen Schritt weiter gegangen, als dass man Text nicht mehr als eine systematisch verbundene Satzmenge betrachtet, „sondern als eine eigenständige Größe, die ihren eigenen Organisationsprinzipien verpflichtet ist und von der ausgehend der Satz als Textbaustein betrachtet werden kann" (Linke/Nussbaumer/Portmann-Tselikas 2004:224). Ein Text ist in diesem Sinne die oberste Organisationsform innerhalb einer kommunikativen Situation, sofern eine kommunikative Funktion zugesprochen werden kann. Im Sinne dieser Definition werden Kriterien zur Texthaftigkeit benötigt, „die weniger an der linearen Verknüpfung von Element zu Element orientiert sind als vielmehr am Textganzen als einer komplex strukturierten und sowohl thematisch als auch konzeptuell zusammenhängenden sprachlichen Einheit" (Linke/Nussbaumer/Portmann-Tselikas 2004:224). Diese ganzheitliche Betrachtung eines Textes erfasst zudem die kommunikative Funktion, was auch im Sinne des handlungsorientierten Ansatzes des GER ist. Die Textdefinition kann also funktional folgendermaßen erweitert werden (Linke/Nussbaumer/Portmann-Tselikas 2004:245):

„Ein Text ist eine komplex strukturierte, thematisch wie konzeptuell zusammenhängende sprachliche Einheit, mit der ein Sprecher eine sprachliche Handlung mit erkennbarem kommunikativen Sinn vollzieht".

Diese Erweiterung des Textbegriffs führt zu der Gegenüberstellung des Begriffspaars Kohäsion vs. Kohärenz bzw. analog dazu zu dem linguistischen Modell der Oberflächenstruktur vs. der Texttiefenstruktur. Die Oberflächenstruktur eines Textes wird durch die sprachliche Realisierung von Informationseinheiten definiert, die durch kohäsionsstiftende Mittel miteinander verknüpft sind. Darüber hinaus betrachtet man was unter der Oberfläche eines Textes liegt, d. h. die Texttiefenstruktur, welche Informationen der Textoberfläche komplex miteinander verbindet. Um im Sinne vorliegender Dissertation schriftliche Lernerproduktionen verstehen und im Weiteren bewerten zu können, muss definiert werden, ob die Aneinanderreihung von Sätzen auf einen zusammenhängenden bzw. kohärenten Text schließen lässt. Um diese Texttiefenstruktur erschließen zu können, müssen die bereits erwähnte Textoberfläche bzw. die lineare Abfolge der Textbausteine, die Textverknüpfung und das Einbeziehen und Aktivieren von allgemeinem außersprachlichen Wissen betrachtet werden. Nach Linke/Nussbaumer/Portmann-Tselikas (2004:251) manifestieren sich textinterne Kriterien entsprechend an der Textoberfläche (Wortschatz, syntaktische Muster) oder an der Texttiefenstruktur (Thema, Textstruktur).

Die Texttiefenstruktur bzw. der Begriff der Kohärenz spielt in den Bewertungskriterien schriftlicher Lernerproduktionen (vgl. Kap. 5) eine entscheidende Rolle. Der GER spricht in diesem Zusammenhang von Diskurskompetenz, d. h. „von der Fähigkeit der Sprachverwendenden/ Lernenden, eine Satzsequenz so zu arrangieren, dass kohärente sprachliche Textpassagen entstehen" (GER 2001:123). Für die Textlinguistik spielen für den Prozess und die Gewährleistung der Kohärenz die bereits unter Kapitel 2.2.1 im Rahmen des GER definierten außersprachlichen Wissensbestände eine große Rolle. Zentral ist in diesem Zusammenhang aus textlinguistischer Sicht die Frage, wie konkret

im schriftlichen Ausdruck Elemente der Textoberfläche mit sprachlichem und außersprachlichem Wissen angereichert werden, so dass man von Textkohärenz sprechen kann (Linke/Nussbaumer/Portmann-Tselikas 2004:229). Eine weitere Voraussetzung, dass eine Reihe von Sätzen als kohärent empfunden wird, äußert sich im Thema eines Textes. Dabei geht es um die Handlung eines Textes oder anders ausgedrückt um den Kerngedanken. Da außersprachliche Wissensbestände wie Welt- oder Handlungswissen natürlich auch hier zum Tragen kommen, ist das Thema zunächst keine sprachliche Größe, denn man bezieht sich in erster Linie auf einen bekannten oder unbekannten Sachbereich (Linke/Nussbaumer/Portmann-Tselikas 2004:237). Linke/Nussbaumer/Portmann-Tselikas (2004:246, 250) betonen in diesem Zusammenhang die Wichtigkeit textexterner Faktoren wie zum Beispiel Textfunktion, Trägermedium, Textadressat, Situations- und Kommunikationszusammenhang, Beziehung der Kommunikationspartner, geteiltes Weltwissen der Kommunikationspartner und Handlungswissen. Sobald ein bestimmter kommunikativer Handlungswert einer sprachlichen Äußerungen zugeordnet werden kann, kann man dies als kohärent empfinden. Anders ausgedrückt kann nicht verhindert werden, „dass eine zufällige, nicht als Text intendierte Satzfolge kohärent und damit als Text verstanden wird“ (Linke/Nussbaumer/Portmann-Tselikas 2004:247). Ein weiteres Kriterium für Texthaftigkeit bzw. der Definition eines Textes besteht auch aus dem Umstand, dass eine Reihe von Sätzen, eine bestimmte Textsorte erkennen lassen (Linke/Nussbaumer/Portmann-Tselikas 2004:254). Interessant ist außerdem die funktionale Satzperspektive, wenn es um die Bestimmung eines Themas geht. Die so genannte satzlinguistische Thema-Rhema-Struktur definiert zum einen, *worüber* etwas ausgesagt und zum anderen *was* ausgesagt wird (Linke/Nussbaumer/Portmann-Tselikas 2004:238). Dieser textlinguistische Ansatz stellt mit anderen Worten die Struktur und die Handlung eines Textes dar.

Eine meines Erachtens integrative Definition im Sinne der Textlinguistik ist die von Brinker (2001:17), wobei es nicht um die grammatische Aneinanderreihung von Satzverknüpfungen, sondern vielmehr um die Komplexität einer sprachlichen Handlung geht:

„Der Terminus „Text“ bezeichnet eine begrenzte Folge von sprachlichen Zeichen, die in sich kohärent ist und die als Ganzes eine erkennbare kommunikative Funktion signalisiert".

Andere Autoren versuchen den Textbegriff anhand von Merkmalsausprägungen zu definieren. Helbig (1986:166) benennt in diesem Sinne fünf Textualitätskriterien, um eine Definition bereit zu stellen:

a) Text als Komplex von Sätzen (Komplexitätskriterium)
b) Text als kohärente Folge von Sätzen (Kohärenzkriterium)
c) Text als thematische Einheit (thematisches Kriterium)
d) Text als relativ abgeschlossene Einheit (Abgeschlossenheitskriterium)
e) Text als Einheit mit erkennbarer kommunikativer Funktion (kommunikatives Kriterium)

Der GER benutzt den Textbegriff „zur Bezeichnung aller sprachlichen Produkte, die Sprachlernende empfangen, produzieren oder austauschen“ (GER 2001:95). Dabei werden in Anlehnung daran alle sprachlichen Aktivitäten und Prozesse hinsichtlich ihrer kommunikativen Intention und auf den Text bezogen analysiert und taxonomiert (GER 2001:95). Im handlungsorientierten Ansatz des GER heißt Text „jeder Diskurs (mündlich oder schriftlich), der sich auf einen bestimmten Lebensbereich bezieht. Texte werden während der Ausführung einer Aufgabe Anlass für Sprachaktivitäten, indem sie diese unterstützen oder sogar als Prozess oder als Produkt Ziel der Aktivitäten sind“ (GER 2001:21). Während Texte von Medien getragen werden, definiert die „Art und die Struktur ihres Inhalts die verschiedenen Textsorten“ (GER 2001:96). Die im Zusammenhang dieser Arbeit und im Mittelpunkt stehenden Textsorten dieser Definition, sind geschriebene Texte. Für das B2-Zertifikat des Goethe-Instituts beispielsweise würde die Textsorte im schriftlichen Ausdruck durch einen Leserbrief repräsentiert. Um diese Textsorte bearbeiten zu können, müssen jedoch bestimmte Regeln und Muster erlernt werden. Wie Bearbeitungsanweisungen und Regeln auf die zu bearbeitende Textsorte Leserbrief in der Ziel- bzw. Fremdsprache umgesetzt und realisiert werden, hängt letztendlich vom Wissen und der Strategieanwendungen der Lerner bzw. der Prüfungsteilnehmer ab. Nach Hayes/Flower (1980:11) besteht der Schreibprozess folglich aus drei konkreten und stabilen Komponenten:

Aufgabenumfeld
Langzeitgedächtnis des Schülers
Schreibprozess (Planen, Formulieren, Überarbeiten)

Tabelle 7: Komponenten des Schreibprozesses

Ich werde im Folgenden dieses Modell auf die schriftliche Lernerproduktion im Fremdsprachenbereich beziehen und stellenweise weiter ausführen, indem ich es anhand der konkreten Prüfungen von Testanbietern im Bereich B2/C1 skizziere.

Das Aufgabenumfeld impliziert im Sinne Hayes/Flower (1980) eine auszuführende Handlung und auf meine Thematik bezogen eine zu bewältigende Aufgabe im schriftlichen Ausdruck. Je nach Testanbieter und das zu prüfende Niveau (gemäß des GER A1-C2) soll ein Aufgabeninput schriftlich bearbeitet werden (vgl. Kap. 5). Es soll also eine bestimmte Textsorte produziert werden, die den Handlungsrahmen bildet. Das stellt zunächst eine Tatsache dar, wobei meines Erachtens das Schreiben nicht nur als Produkt sondern auch als Prozess betrachtet werden sollte (in welchem Rahmen). Ob dies realisiert werden kann, hängt vom zweiten Faktor in dieser Schreibprozesskette ab, nämlich vom Langzeitgedächtnis einer Person, die kommunikativ bzw. schreibproduktiv tätig werden soll. Das Langzeitgedächtnis bewahrt Informationen auf, die unser Weltwissen bilden, für den Abruf zu einem späteren Zeitpunkt (Zimbardo 1992:270). Das zur Aufgabenbewältigung erforderliche Material bzw. Wissen und die entsprechende Strategie sind mehr oder weniger im Speicher Langzeitgedächtnis vorhanden. Die

Bearbeitung eines Inputs, worauf eine schriftliche Reaktion folgen soll, erfordert den Zugriff auf das im Langzeitgedächtnis mehr oder weniger verfügbare Weltwissen, Textwissen, Sprachwissen und Adressatenwissen. Was das Textwissen beim fremdsprachlichen Schreiben anbelangt, so muss man zunächst in Erfahrung bringen, ob es sich mit dem in der Erstsprache erworbenen deckt und folglich kompatibel sein kann. Es kann nicht grundsätzlich vorausgesetzt werden, dass Textsorten und ihre Merkmale universell und kulturunabhängig sind. Diesbezüglich hält es Glück (1988:32) für möglich, dass „Textsorten und ihre Ausprägung kulturspezifisch sein können". Verfügt ein Lerner nicht über das nötige Textwissen, welches Schemata und Textstrukturen beinhaltet, so kann er nicht mit dem Prozess der schriftlichen Produktion beginnen. Während das Weltwissen über das Langzeitgedächtnis aktiviert werden kann, kann das Textwissen im Falle der fehlenden textuellen Kompetenz nicht entsprechend angewendet werden. Raimes (1987) benennt in diesem Zusammenhang neben fehlenden Fremdsprachenkenntnissen auch das fehlende Strategiewissen und das fehlende Wissen über die zielsprachlichen Textkonventionen. Im kommunikativen Kompetenzmodell von Bachman/Palmer (1996:67ff.) besteht Sprachfähigkeit aus sprachlichem Wissen und strategischer Kompetenz. Während sprachliches Wissen aus strukturellem und pragmatischem Wissen besteht, soll die strategische Kompetenz aufzeigen, ob Sprachwissen in der Kommunikation angewandt werden kann (vgl. Faerch/Kasper 1983). Textwissen gehört hierbei zum strukturellen Wissen. Die Frage, die sich im Rahmen dieser Dissertation insgesamt und speziell hinsichtlich der schriftlichen Lernerproduktion stellt, ist, inwiefern das Wissen und dessen Anwendung bei zielsprachlichen Textkonventionen Aussagen über die Sprachkompetenz eines Lerners bzw. Prüflings erlaubt und folglich machen kann und ob dies schließlich normgerecht ist.

Der dritte und in sich untergliederte Aspekt bezieht sich auf den Schreibprozess an sich. Um diesen abzuschließen, müssen drei Phasen durchlaufen werden. Dabei wird Rücksicht auf die bereits erwähnten Komponenten genommen. Der Planungsvorgang setzt Schreibziele und -inhalte fest. In dieser Planungsphase setzt sich der Schreiber mit der Frage auseinander, *was, wie* und *womit* er schreiben soll, wobei sich diese normativen Aspekte in einer Wechselbeziehung befinden. Der GER definiert diesen dritten Bereich in der Kette der kommunikativen Aktivität Schreiben als das Anwenden metakognitiver Prinzipien (GER 2001:73). Es finden sich im Referenzrahmen empirisch nicht kalibrierte Beispielskalen hinsichtlich der sprachlichen Aktivität und Strategie für die schriftliche Produktion allgemein und für das Bericht- und Aufsatzschreiben.

Die für die Niveaus B2/C1 interessanten Deskriptorendefinitionen, die aus anderen Skalen zusammen gesetzt wurden, will ich an dieser Stelle kurz anführen (GER 2001:67ff.):

	Schriftliche Produktion allgemein	Berichte und Aufsätze schreiben
C1	Kann klare, gut strukturierte Texte zu komplexen Themen verfassen und dabei die entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden.	Kann klare, gut strukturierte Ausführungen zu komplexen Themen schreiben und dabei die entscheidenden Punkte hervorheben. Kann Standpunkte ausführlich darstellen und durch Unterpunkte, geeignete Beispiele oder Begründungen stützen.
B2	Kann klare, detaillierte Texte zu verschiedenen Themen aus seinem/ihrem Interessengebiet verfassen und dabei Informationen und Argumente aus verschiedenen Quellen zusammenführen und gegeneinander abwägen.	Kann einen Aufsatz oder Bericht schreiben, in dem etwas systematisch erörtert wird, wobei entscheidende Punkte angemessen hervorgehoben und stützende Details angeführt werden. Kann verschiedene Ideen oder Problemlösungen gegeneinander abwägen Kann in einem Aufsatz oder Bericht etwas erörtern, dabei Gründe für oder gegen einen bestimmten Standpunkt angeben und die Vor- und Nachteile verschiedener Optionen erläutern. Kann Informationen und Argumente aus verschiedenen Quellen zusammenführen.

Tabelle 8: Kann-Beschreibungen für den schriftlichen Ausdruck auf den Niveaus B2 und C1

Auch für den Schreibprozess nach Hayes/Flower (1980) bzw. der Produktionsstrategien Planen, Formulieren, Überarbeiten existieren im GER Beispielsskalen, wobei für die Zwecke dieser Arbeit manche irrelevanten Deskriptoren ausgelassen bzw. auf die kommunikative Aktivität hin konkretisiert wurden (GER 2001:70):

	Planen	Kompensieren/*Formulieren*	Kontrolle und Reparaturen/*Überarbeiten*
C1	Wie B2	Wie B2+	Kann bei Ausdrucksschwierigkeiten neu ansetzen und umformulieren, ohne die Äußerung ganz abreißen zu lassen.

B2	**Kann planen, was und wie er/sie etwas *ausdrücken* will, und dabei die Wirkung auf *des Adressaten* berücksichtigen**	**Kann etwas paraphrasieren und umschreiben, um Wortschatz- und Grammatiklücken zu überbrücken**	**Kann Fehler normalerweise selbst korrigieren, wenn sie/ihm bewusst werden. Kann eigene Fehler korrigieren, wenn sie zu Missverständnissen *führen können*. Kann sich seine Hauptfehler merken und sich bewusst in Bezug auf diese Fehler kontrollieren.**

Tabelle 9: Produktionsstrategien für den schriftlichen Ausdruck für die Niveaus B2 und C1

Um einen Leserbrief des Niveaus B2 bezüglich eines vorgegebenen Inputs schreiben zu lassen, muss man zunächst den erwünschten Inhalt und die Thematik definieren und eingrenzen: *Was* und *wie* soll der Lerner schreiben? Das gegebene Aufgabenfeld setzt der Lerner bzw. Prüfling insofern um, indem er sich auf der nächsten Ebene bewegend, Gedanken über die erforderlichen sprachlichen Mittel macht und durch die entsprechende schriftliche Produktion die Frage des „womit soll der Lerner schreiben?“ beantwortet. [51] **Sind die Phasen der Planung und des Formulierens abgeschlossen, dann wird der Schreibprozess mit der Überarbeitungsphase abgerundet. Hierbei soll nun überdacht werden, ob die Schreibziele und die zu erfüllenden Bedingungen erreicht wurden.**

Nach Feilke (1993a:17) ist das Erreichen der *textuellen Handlungskompetenz* ein langwieriger Aneignungsprozess. Ziel des Schreibens und seiner fortlaufenden Entwicklung ist, dass „der geschriebene Text in diesem Sinn als Produkt eines problemorientierten und problemlösenden Schreibprozesses, in dem die SchreiberInnen

- **ihre subjektive Involviertheit**
- **die sachliche Komplexität des Themas**
- **die formale Homogenität ihres Textes und**
- **die antizipierten Erwartungen eines Adressaten**

„unter einen Hut“ bringen müssen (...)“ (Feilke 1993a:23) [52]

Textuelle Handlungskompetenz ist wie im kommunikativen Kompetenzmodell von Bachman/Palmer (1996) auch nach Feilke (1993a) ziel-, situations- und adressatenspezifisch und involviert sowohl Wissen als auch Strategieanwendungen. Das Schreibprozessmodell von Hayes/Flower (1980) habe ich für die Zwecke vorliegender Dissertation speziell in Form einer Pyramide den konkreten Prüfungen angepasst. Das heißt, dass der Schreibprozess für die Bearbeitung der Aufgabe des schriftlichen Ausdrucks im B2-Zertifikat des Goethe-Instituts (Leserbrief) in diesem Sinne folgendermaßen von Statten geht:

51 Nach Hayes/Flower (1980) heißt dieser Prozess „Translating“

52 Hervorhebungen im Original

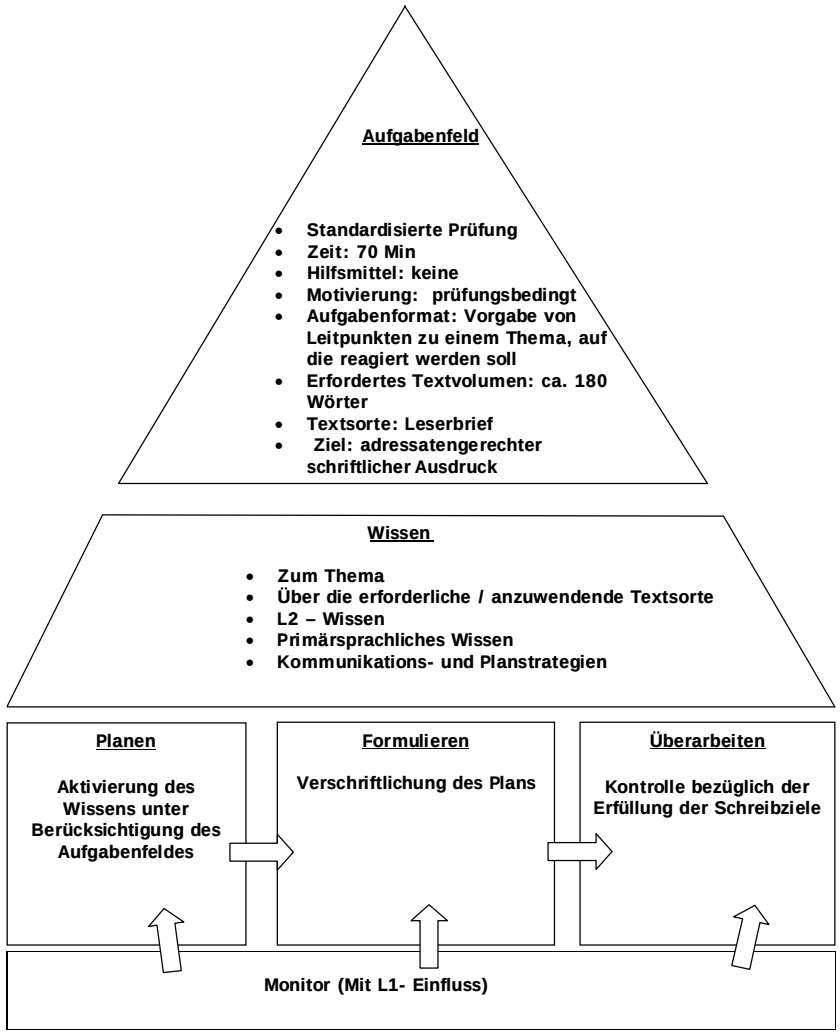

Tabelle 10: Das Schreibprozessmodell von Hayes/Flower (1980) am Beispiel des schriftlichen Ausdrucks des Niveaus B2 in Form einer Pyramide

In dem von mir erweiterten Schreibprozessmodell von Hayes/Flower (1980) hinsichtlich der einzelnen Prüfungen, die in dieser Arbeit behandelt werden, sollten alle Komponenten einer standardisierten Prüfung hinsichtlich des Schreibprozesses eingeflochten werden. Dabei habe ich beim Aufgabenfeld sämtliche Daten und Gegebenheiten, die ein Prüfling bei der jeweiligen Prüfung zu beachten hat, eingebracht. Dem Prüfling muss also in diesem ersten Schritt bewusst sein, um was es in diesem Prüfungsteil geht und welchen konkreten Gegebenheiten er ausgesetzt ist. Dabei spielen sowohl Zeit, Textsortenbewusstsein und der Zweck der Prüfung insgesamt eine Rolle. In der nächsten Kategorie sind dann explizit die erforderlichen Wissensbestände angeführt, die zur Bewältigung der gestellten Aufgabe benötigt werden, um schließlich den Schreibprozess zu planen, zu formulieren und zu überarbeiten. Der Einfluss des Monitors wird im GER als strategische Komponente betrachtet, die „dafür zuständig ist, die mentalen Aktivitäten und Kompetenzen (...) fortlaufend zu aktualisieren" (GER 2001:95). Es stellt sich demnach erneut die Frage, in welchem Maß welche Kompetenzen erwartet und vorausgesetzt werden, um die kommunikative Aufgabe des schriftlichen Ausdrucks zu bewältigen. Daher betont der GER die zentrale Rolle der inhaltlichen Ebene, während Lerner ihre kommunikativen Intentionen realisieren (GER 2001:154). Wichtig ist diesbezüglich also, aufgrund welcher Faktoren die Textproduktion vom Lerner ausgeführt werden kann (vgl. erweitertes Schreibprozessmodell in Form einer Pyramide).

Zu einer der ersten durchgeführten empirischen Untersuchungen hinsichtlich der Textproduktion im Fremdsprachenbereich gehören die von Krings (1986). Dabei wurde auf das introspektive Verfahren des „Lauten Denkens" zurückgegriffen, welches es erlaubt, Sprachverwendungs- und Sprachproduktionsprozesse insbesondere schriftlicher Art nachzuzeichnen und zu verstehen.[53] Hierbei sollten die Testpersonen jegliche Gedanken während der Bearbeitung einer Vorlage ungebundener bzw. freier Schreibaufgabe in der Fremdsprache verbalisieren, die dann anhand der angefertigten Transkriptionen analysiert wurden, um der Frage näher zu kommen, was in den Köpfen der Lerner einer Fremdsprache im Allgemeinen und insbesondere während einer schriftlichen Textproduktion vor sich geht (Krings 1989:394ff). Die Versuchspersonen, die an der ersten Pilot-Untersuchung teilnahmen, waren Französischstudenten im Hauptstudium der Universität Bochum, die auch Erfahrungen mit dem Land der Zielsprache hatten. Ihre Aufgabe bestand darin, auf einen französischen Anzeigetext in Form einer Bewerbung um die Anstellung als Au-pair zu schriftlich zu reagieren. Die zweite Pilotuntersuchung bestand darin, Bildergeschichten schriftlich nachzuerzählen. Ausgewählt wurden hierzu Sprachstudenten verschiedener Muttersprachen (fünf deutsche, zwei spanische, ein französischer und ein italienischer Muttersprachler), die sich im Hauptstudium der Universität Bochum befanden.

Im Zusammenhang mit der vorliegenden Dissertation ist es wichtig, das Profil und das sprachliche Niveau der Versuchspersonen zu klären, damit die Ergebnisse dieser empirischen Untersuchungen auf die Referenzniveaus B2 und C1 übertragen werden können. Da der GER zum Zeitpunkt dieser Untersuchung noch nicht definiert war, kann ich an dieser Stelle lediglich davon ausgehen, dass sich das Sprachniveau der Versuchspersonen mit den Schwellenniveaus B2 und C1 decken kann. Ein Indiz dafür ist der Aufgabentyp Bewerbungsschreiben, der in der ersten Pilotstudie angewandt wurde, welcher als Textsorte gerade in diesen Bereichen des GER präsent ist.

53 Anmerkung: Als Hilfsmittel wurden ein- und zweisprachige Wörterbücher und eine einzige Grammatik zugelassen.

Teile aus Krings erstelltem Fragenkatalog für das Ergründen der fremdsprachlichen Textproduktionsprozesse werden an dieser Stelle herangezogen und teilweise für das Ziel dieser Arbeit, die Bewertungskriterien von Lerner(text)produktionen nach ihrer Validität zu beurteilen, erweitert (Krings 1989:380). Ich habe hierbei die Ursachen und die Probleme bei Textproduktionsproblemen in interne, d. h. den Prüfling betreffend, und externe, d.h. von außen hervorgerufene Faktoren, eingeteilt:

Interne Ursachen und Probleme	Externe Ursachen und Probleme
Wie geht ein Lerner bei der Bearbeitung bzw. Produktion der ihm gestellten schriftlichen Aufgabe vor?	Welche sprachlichen oder nichtsprachlichen Probleme können auftreten?
Gibt es Lernerstrategien? Wenn ja welche?	Welche Textsorten werden vorausgesetzt?
Welche Rolle spielen primärsprachliche Kenntnisse bezüglich der Teildisziplin Schreiben?	Welche Gemeinsamkeiten und Unterschiede gibt es zwischen freier fremdsprachlicher Textproduktion und verschiedenen gebundenen Typologien des Schreibens?
Wie interagieren sprachliche und nichtsprachliche Wissensbestände im Prozess der Textproduktion?	Wie beeinflussen unterschiedliche Aufgabentypologien den schriftlichen Ausdruck in der Fremdsprache?
Welche intra-individuellen und inter-individuellen Unterschiede kennzeichnen in der Textproduktion den Lerner (Alter, Kompetenzgrad der Fremdsprache, vertraute Aufgabentypologie, Anwendung von Strategien, etc.)?	Gibt es eine Beziehung zwischen Textproduktionsprozess und dem daraus resultierenden Produkt?
	Wie sollte unter Berücksichtigung aller Faktoren ein Fremdsprachenunterricht aufgebaut sein?

Tabelle 11: Interne und externe Textproduktionsprobleme

Man könnte einen derartigen Fragenkatalog je nach Intention und Zielsetzung komprimieren oder auch erweitern. Für den Rahmen dieser Arbeit sind an dieser Stelle die grundlegendsten Überlegungen und Fragestellungen hinsichtlich des Prozesses und schließlich des Testens und Bewertens schriftlicher Lernerproduktionen ausgewählt worden, um diese im zentralen Kapitel 5 zur Diskussion zu stellen. Dabei finden sich Parallelen zu dem erweiterten Schreibprozessmodell wieder, die Krings unter anderem anhand seiner empirischen Forschungsarbeit teilweise durch seine Ergebnisse

beantworten und belegen konnte. Die Frage nach der Existenz sprachlicher oder nichtsprachlicher Probleme ließ sich durch die hauptsächliche Häufung in der Zielsprache beantworten, wo bestimmte semantische Einheiten im Gegensatz zur Erstsprache nicht aktiviert werden konnten. Der Mangel an spontanen fremdsprachlichen Konzeptualisierungen wirkt sich aufgrund der „muttersprachlichen Blockierung“ (Krings 1989:415ff) mit 44% auf den Bereich der Textproduktion aus. Somit kann davon ausgegangen werden, dass die Koppelung muttersprachlicher und fremdsprachlicher Prozesse zusammenhängt (vgl. Portmann 1991). Im Fall der schriftlichen Textproduktion bedeutet dies, dass sich der Prozess der Textplanung sowohl auf der muttersprachlichen als auch auf der zielsprachlichen Ebene vollzieht (vgl. Börner 1987). Im Sinne des bereits angeführten Schreibprozessmodells von Hayes/Flower (1980) bezieht sich diese Feststellung auf die Planung des Schreibprozesses. Ebenfalls wurde in dieser empirischen Studie nachgewiesen, dass spontane Assoziationen in der Muttersprache statt finden (vgl. Jones/Tetro 1987). Da man den äquivalenten Ausdruck oder Begriff in der Zielsprache nicht kennt, bedient man sich eines Wörterbuchs.[54] Die mehrfachen Vermerke darin lassen aber nicht auf die korrekte Verwendungsform im bestimmten Kontext schließen (es könnte sich z. B. um eine Redewendung handeln, die nicht einfach transferiert werden kann). Weitere Probleme (40%), die sich in der Fremdsprache äußern, decken die Palette der Orthografie, des Plurals, des Genussystems, des Tempus, des Modus und der Syntax ab, um die wichtigsten zu nennen. Bei der Beantwortung der sprachproblematischen Frage stellt sich zudem deutlich heraus, dass lexikosemantische Probleme so prägnant sind, dass man darauf schließen kann, dass das so genannte Bedeutungslernen keinerlei Platz im Fremdspracherwerb zu haben scheint (vgl. Levenston 1979).

Durch die verschiedenen Ausprägungen der Kompetenz im schriftlichen Ausdruck konnten weiterführend sowohl die Lernerstrategien zur Bearbeitung der gestellten Aufgabe erkannt, als auch die Aufgabenschwierigkeit durch die Problemkonzentration indiziert werden. Als wichtigste Lernerstrategie stellte sich in vielerlei Hinsicht die Muttersprache als Steuerungselement bzw. als Einsatzstrategie heraus. Angenommen, ein Lerner sucht z. B. nach einer entsprechend des Kontextes notwendigen Versprachlichung. Benützte er in diesem Fall ein Wörterbuch, so wäre es nach Krings (1989:420ff.) nicht möglich, dass die gefundenen Sprachverweise realisiert werden könnten, da ihre Funktion in einem gegebenen Sprachkonstrukt unbekannt wären (Vermeidungsstrategie). Man würde demnach strategisch auf die Muttersprache zurückgreifen, in der der äquivalente Ausdruckskomplex präsent ist. Ist die Übertragung dennoch nicht realisierbar, wird die muttersprachliche Version dermaßen umstrukturiert, so dass die Struktur leichter in die Fremdsprache zu transferieren ist. Vermeidungsstrategien wie jene im angeführten Beispiel gibt es nach Varadi (1983:65ff.) in einer schwächeren und stärkeren Form. Das so genannte *message adjustement* ist die Abschwächung des *message abandonment*. Ersteres veranlasst den Lerner dazu, seine Ausdrucksintention zu verändern und sie schließlich seiner Ausdruckskompetenz anzupassen, statt die Sprachmittel auf seine Ausdrucksintention abzustimmen. Die starke Version einer Vermeidungsstrategie impliziert die völlige Aufgabe der Äußerungsintention des Lerners, da die Ressourcen nicht ausreichen, um die fremdsprachliche Verbalisierung zu realisieren.

54 Berücksichtigt man die Ergebnisse dieser empirischen Studie, in der der Wörterbuchgebrauch gestattet war, dann stellt sich die ganz große Frage, wie Lerner in der Prüfungssituation ohne Hilfsmittel Texte produzieren. Im Sinne Hayes/Flower (1980) ist in diesem Fall das Langzeitgedächtnis das einzige Hilfsmittel, um einen schriftlichen Text zu produzieren.

In Krings empirischer Forschung kristallisierte sich weiterhin die Frage heraus, inwiefern die Muttersprache den fremdsprachlichen Textproduktionsprozess „mitsteuert“ bzw. lenkt. 40,6 % der Planrealisierungen, d. h. wie eine schriftliche Aufgabenstellung angegangen wird, vollzogen sich in der Muttersprache. Muttersprachliche Versprachlichungen sind demnach als automatisiert zu betrachten und finden sich wiederholt auf den Gebieten des Wortschatzes oder der Syntax im Prozess der fremdsprachlichen Textproduktion wieder. In diesem Fall wird das in der Muttersprache „erstellte“ sprachliche Teilkonstrukt in die Fremdsprache transferiert und mit weiteren fremdsprachlichen Elementen angereichert. An dieser Stelle sollte die Problematik des von Selinker (1972) definierten Interferenzbegriffes angesprochen werden, wobei zunächst in der Muttersprache konstruiert wird, dann in die Zielsprache übertragen und dieses gegebenenfalls noch mit zielsprachlichen Elementen angereichert wird.

In der Auswertung dieses empirischen Projekts beobachtet man eine mehr oder weniger starke Koppelung von Textproduktionsplanungsprozessen sowohl in der Mutter- als auch in der Zielsprache, jedoch sind die muttersprachlichen Einflüsse in diesen ausgewerteten Daten in fast jeder fremdsprachlichen Planung zu verzeichnen. Probleme, wie z. B. Verbalisierungen, die in der fremdsprachlichen Textproduktion zu finden sind, können im gleichen Prozess der Muttersprache nicht erkannt werden. Demnach ist bereits ein Unterschied zwischen Erstsprache und Zielsprache im Hinblick auf den Textproduktionsprozess definiert. Außerdem läuft der Textproduktionsprozess in der Muttersprache in einer viel höheren Geschwindigkeit als jener in der Fremdsprache ab.
Das führt zur Annahme, dass im schriftlichen Ausdruck der Muttersprache kaum eine Alternativplanung zu beobachten ist. Cummins (1994) führte eine Studie durch, indem sie Vergleiche im Schreibverhalten erfahrener und nicht erfahrener Schreiber in der Fremdsprache anstellte. Dabei kam sie zu der Schlussfolgerung, dass Schreibfähigkeit, die so genannte *writing expertise*, und das Niveau der Zielsprache unterschiedliche Wissensaspekte ausmachen.

3.5 Feststellungen und Beobachtungen für die Praxis

Schriftliche Übungen, Tests oder Prüfungen gehören zum Alltag des Fremdspracherwerbs. Welche Bedingungen, Ziele, Normen oder auch Formen des Schreibens in der Zielsprache sind hierbei gegeben? Wie erstellt ein Lerner einen Text in der jeweils erwarteten Form? In dieser Arbeit geht es erstrangig um kriteriumsorientierte Sprachprüfungen von verschiedenen Testanbietern im Bereich Deutsch als Fremdsprache. Bedenkt man, dass in der Regel das Erlernen der 1. Fremdsprache im Durchschnitt im Alter von 10 Jahren beginnt (Börner 1989:351), dann muss man der Frage nachgehen, inwiefern die kognitive Entwicklung und Reife des Lerners dem Anspruch der fremdsprachlichen Textproduktion insbesondere von Sprachzertifizierungsprüfungen genügen soll. Folglich müssen diejenigen Textformen behandelt oder auch erwartet werden, die der entsprechenden kognitiven Entwicklung analog sind. Standardisierte Sprachprüfungen neigen oft dazu, die von der Prüfungsordnung vorgesehene Altersbegrenzung herunterzusetzen. Das war bis zum Herbst 2008 z. B. bei Prüfungen des Goethe-Instituts der Fall, die das Mindestalter für die Niveaustufen B1-C1 laut offizieller Prüfungsordnung auf 16 Jahre festsetzten. Dem entgegen war das Mindestalter für diese Prüfungen in Griechenland, um ein prägnantes Beispiel zu nennen, mit einem Sonderstatus versehen. Hier bekamen bereits Jugendliche zwischen 13 und 15 Jahren Zulassungen für die ursprünglich für Erwachsene konzipierten Prüfungen.

Wie etwas in der Fremdsprache zu Papier gebracht werden soll, ist viel komplexer als in der Muttersprache. Die Kognition und das Weltbild eines Lerners mag sehr breit gefächert sein, aber die mangelnde sprachliche Ausdrucksfähigkeit hemmt den reibungslosen Prozess des Schreibens (Börner 1989:353). Als erstes stellt sich die Frage nach der Notwendigkeit schriftlicher Produktion in der Fremdsprache. Wie bereits erläutert, gehört die schriftsprachliche Kompetenz mittlerweile zu den Fertigkeiten, durch die Sprachkenntnisse überprüft und einem Prädikat zugeordnet werden. Ausgangslage ist demnach die Existenz dieser Teildisziplin in Sprachprüfungen. Welche Textsorten muss ein Lerner dafür beherrschen lernen? Welchen Normen ist zu folgen? In welcher Form wird der schriftliche Ausdruck in den verschiedenen Sprachprüfungen abgeprüft? Handelt es sich bei der Aufgabentypologie um einen Brief oder gar um einen Aufsatz (Börner 1989:356)? Was prüft der schriftliche Ausdruck schließlich und worauf basieren die Anforderungen? All diese Fragen sind im Rahmen des handlungsorientierten Ansatzes des GER zu beantworten. Er versteht sich nur als Bezugsrahmen und ist demnach nicht verpflichtend. Basierend auf den Kann-Beschreibungen und dem kommunikativen Ansatz des GER jedoch entwickeln verschiedene Testanbieter Zertifikats- bzw. Sprachnachweisprüfungen, in denen die schriftliche Produktion einen Teil ausmacht.

Schriftliche Texte in der Fremdsprache zu produzieren bedeutet wie in der Erstsprache auch, Informationen aus dem Langzeitgedächtnis zu aktivieren. Am Anfang steht der Kontext, in dem der Lerner einen sprachlichen Input erhält, den er bearbeiten soll. Dafür versucht er sprachliches Wissen, Weltwissen und strategisches Wissen anzuwenden. Diese Wissensarten sind miteinander verbunden und Defizite der einen Art können durch Kompetenzen in anderen Gebieten teilweise ausgeglichen werden. Die generelle Anwendung setzt das Vorhandensein voraus. Dieser Prozess kann für die ganze Verarbeitungsabfolge belastend sein. Lerner können mehr oder weniger auf das in der Erstsprache erlernte Wissen hinsichtlich Planstrategien, Textsorten, Kohärenz, Kohäsion und Stilmittel, um einige zu nennen, zurückgreifen. Gewährleistet sein muss jedoch, dass dieses Wissen zum einen kompatibel mit der Zielsprache ist und zum anderen, dass es

adäquat benutzt werden kann. Hinzu kommt der Faktor, dass jeglicher fremdsprachliche Wortschatz abrufbar sein muss, um diesen mit Hilfe grammatischer und syntaktischer Abstraktionen in ein Textgeflecht umzuwandeln (Börner 1989:359). Es muss an dieser Stelle aber die Überlegung angestellt werden, was passiert, sobald der Lerner weder Sach- bzw. Strategiewissen aus dem Langzeitgedächtnis noch Elemente aus dem Fremdsprachenregister bzw. seinem Lexikon abrufen kann (Antos 1989:22). Von Interesse ist auch die Variante, bei der das sprachliche Wissen stark defizitär ist und mithilfe von Welt- und Strategiewissen ausgeglichen wird. Dann stellt sich natürlich die Frage danach, was gemessen wird und in welchem Umfang das Welt-, Hintergrund und Fachwissen schließlich die Schreibkompetenz beeinflussen. Auf die Lesekompetenz bezogen behauptet Clapham (1996, 2000) im Rahmen ihrer doppelten Schwellenhypothese, dass das Welt-, Hintergrund und Fachwissen bei einem mittleren Sprachniveau nachhaltig das Testergebnis beeinflussen (vgl. Clapham 1996, Clapham 2000).

Bachman/Palmer (1996:60ff.) differenzieren in ihrem interaktiven Kommunikationsmodell zwischen Merkmalen innerhalb und außerhalb einer Testsituation. In einer Sprachprüfung bekommt der Prüfling einen sprachlichen Input. Demnach könnte bereits im Unterricht Übungsmaterial mit sprachlichen Inputs und Vorgaben bereit gestellt werden (vgl. Nation (2001)/Löschmann (1992)). Aus methodisch-didaktischer Sicht könnten Lerner dadurch ihr Sprachwissen aufbauen. Es gilt hier aber die Testsituation von der Unterrichtssituation abzugrenzen. Die Reaktion auf einen sprachlichen Input mag mithilfe der momentanen Interlanguage eines Lerners oder des Übernehmens einzelner referierter Sachverhalte des Inputs realisiert werden. Zu klären bleibt aber, wie dann latente Fehler, die auf Vermeidungsstrategien zurückzuführen sind, in einer Testsituation bewertet werden (Kohn 1990:15). Anders ausgedrückt stellt sich die Frage, ob Bewertungskriterien hinsichtlich der konkreten Aufgabenstellung im schriftlichen Ausdruck der verschiedensten Testanbieter Rücksicht auf die schon erwähnten Probleme und die Schwierigkeiten nehmen. In diesem Zusammenhang führt Cummins (1994:175) eine meines Erachtens treffende Aussage an:

„Differences in performance appear to arise - while writing in a second language – for the knowledge, procedures, or strategies people use to produce their writing (...)".

Die angeführte Problematik des Schreibens in der Fremdsprache ist sicherlich nicht zu ignorieren. Der Inhalt einer gestellten und konkreten Aufgabentypologie für die Textproduktion kann mehr oder weniger adäquat für die Zielgruppe sein.

Im Folgenden werde ich den Begriff der Kompetenz bezüglich des Schreibens anführen. Außerdem sollen die verschiedenen und für uns relevanten Textsorten definiert werden, die jeweils aufsteigend zu beherrschen sind. In diesem Zusammenhang werde ich die erforderlichen Kompetenzen für den schriftlichen Ausdruck der verschiedenen Niveaus für Sprachprüfungen einem primärsprachlichen Curriculum für das Fach Deutsch gegenüber stellen. Dieses kontrastive Aufzeigen soll den natürlichen und kognitiven Aufbau im Bereich der Textsorten und der Entwicklung in der Schreibarbeit muttersprachlicher Schüler demonstrieren. Außerdem sollen weitere Komponenten aufgezeigt werden, die Teile des Schreibprozesses sind.

3.6 Der Kompetenzbegriff

Das Erlernen einer Sprache und die Kommunikationsfähigkeit insgesamt erfordern gewisse Kompetenzen. Der GER definiert in seinem kommunikativen Ansatz die notwendigen Voraussetzungen und Kompetenzen für Sprachanwendung bzw. kommunikativer Kompetenz. Im Fremdsprachenbereich wird in der Regel von den vier klassischen Kompetenzen Leseverstehen, Hörverstehen, schriftlicher und mündlicher Ausdruck ausgegangen. Bolton (1982:55) unterscheidet an dieser Stelle zwischen informationsentnehmender und informationsverarbeitender Kommunikationsfähigkeit. Die hier im Mittelpunkt stehende Kompetenz ist der schriftliche Ausdruck. Um schriftlich produktiv zu werden, muss zunächst die Aufgabenstellung rezipiert bzw. der kommunikative Rahmen verdeutlicht werden (Bolton 1982:71). Das heißt, wie bereits im Pyramidenmodell erläutert, dass verschiedene Kompetenzen aktiviert werden müssen, die der Kategorie Wissen unterliegen. Es stellt sich folglich die Frage, worauf die Schreibkompetenz letztlich beruht und wodurch sie definiert wird. Aus diesem Grund soll eine adäquate Definition und Eingrenzung der Kompetenz schriftlicher Lernerproduktionen erarbeitet werden. Die vorausgesetzten Kompetenzen bei Sprachprüfungen des Goethe-Instituts beispielsweise sollen dokumentiert und zudem diskutiert werden, indem kontrastiv dazu ein kompetenzorientiertes Curriculum bzw. eine Unterrichtsvorgabe des primärsprachlichen Unterrichts herangezogen wird.

Der Begriff *Kompetenz* (lat. *competere* - zu etwas fähig sein) definiert aus psychologischer Sicht in erster Linie die Fähigkeit eines Menschen, bestimmte Aufgaben bzw. Anforderungen selbstständig auszuführen. Auf die Sprache bezogen, impliziert Kompetenz die „Fähigkeit eines Sprechers, in seiner Muttersprache eine unbegrenzte Zahl von grammatischen Sätzen zu erzeugen und zu verstehen sowie grammatische von ungrammatischen Sätzen unterscheiden können; (unbewusstes) Wissen eines Sprechers um die Regeln des Systems einer Sprache“ (Herbst 1991:18)..

Weinert (2001:29) definiert den Kompetenzbegriff als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen[55] und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“. Auf dieser Definition basiert das Deutsch-Modul im Kernlehrplan Nordrhein-Westfalens (KLP NRW)[56], der hier wie bereits erwähnt, als Vergleichsgrundlage dienen soll, um die für bestimmte Textsorten erforderlichen Kompetenzen für die Sprachprüfungen des Goethe-Instituts und des TestDaf-Instituts zu diskutieren. Diese Kompetenzdefinition ist nichts Anderes, als die funktionale Verbindung von Wissen, Verstehen, Können und Wollen. Nach Klieme (2004:11) „entwickeln sich Kompetenzen durch systematischen Aufbau, intelligente Vernetzung und variierende situative Einbettung von Wissen“. Wie aus dem bereits angeführten und erweiterten Schreibprozessmodell von Hayes/Flower (1980) deutlich wird, ist der Faktor Wissen die Basis für jegliche Form der Kompetenz bzw. der Performanz. Diese wird in unserem Zusammenhang als die Sprachverwendung definiert, die aber zudem durch spezifische Prozesse und Strategien beschrieben ist, um die Regeln der Kompetenz entsprechend zu verwenden (Kohn 1990:73):

55 Als Volition wird in der Psychologie der Prozess der Willensbildung bezeichnet. www.wikipedia.de (Zugriff am 02.01.2008)

56 http://db.learnline.de/angebote/deutschunterrichtsentwicklung/module/teil-2.pdf , Zugriff am 01. November 2007

„Tatsächlicher Gebrauch von Sprache in Äußerungen in konkreten Situationen, wozu auch Erscheinungen wie abgebrochene Konstruktionen, Zögerungsphänomene oder Regelverstöße zu rechnen sind“ (Herbst 1991:18),

Diese Definition nähert sich schon eher den Problematiken, denen Lerner ausgesetzt sind. Spricht man von Regelverstößen, so impliziert das Performanzfehler. Anders ausgedrückt ist das die nicht korrekte Umsetzung der Kompetenz hinsichtlich der Zielsprachennorm. Kompetenzfehler werden im GER als eine „Erscheinung von Lernersprachen“ betrachtet. Diese Interlanguages lassen demnach Zweifel aufkommen, ob L2-Lerner im Sinne Chomskys (1965:3) schließlich als „ideal speaker/listener in a completely homogeneous speech-community“ angesehen werden können.[57]

Um den Schreibprozess in einer Fremdsprache zu aktivieren, wird das grundlegende Element Wissen nach Edmonson & House (1993:267 ff.) in deklaratives und prozedurales Wissen eingeteilt. Dabei ist deklaratives Wissen das statische, sprachliche Wissen auf den verschiedenen sprachlichen Beschreibungsebenen, welches folgendermaßen differenziert werden kann:

- *implizites vs. explizites Wissen*

  Implizites bzw. explizites Wissen gibt an, inwieweit etwas unbewusst oder bewusst erworben wird.[58] Sprachwissen wird in diesem Zusammenhang durch das deklarative Wissen realisiert. Dabei ist das sprachliche Wissen eines Lerners nicht die Kompetenz eines Muttersprachlers.

- *analysiertes vs. unanalysiertes Wissen*

  Die gänzliche oder partielle Einprägung sprachlicher Elemente findet sich in dem Begriffspaar analysiertes und nicht-analysiertes Wissen wieder.

- *integriertes vs. nicht-integriertes Wissen*

  Inwiefern erworbenes Wissen bei verschiedenen Aufgabenbewältgungen verfügbar ist, definiert die Subkategorie integriertes/ nicht-intergriertes Wissen.

- *automatisiertes vs. nicht-automatisiertes Wissen*

  Automatisiertes bzw. nicht-automatisiertes Wissen gibt zum Beispiel den Anstrengungsgrad bei der Sprachanwendung an.

Diese Differenzierung des deklarativen Wissens zeigt Paralellen zu dem Schreibprozessmodell auf, indem auf all die Faktoren eingegangen wird, die für die schriftliche Lernerproduktion entscheidend sind (vgl. Pyramiden). Der GER fasst unter der Kategorie des deklarativen Wissens, welches zu den allgemeinen Kompetenzen zählt, das Weltwissen, das soziokulturelle Wissen und das interkulturelle Bewusstsein zusammen. Dabei wird den Benutzern des GER geraten, zu definieren, welche Unterkategorien des deklarativen Wissens von den Lernern auf den jeweiligen Niveaubeschreibungen erwartet werden und welche man von ihnen einfordern kann (GER 2001:103ff.).

Wie die Sprachkompetenz unter Berücksichtigung dieser Faktoren überhaupt realisiert werden kann, wird über das prozedurale Wissen beantwortet. An dieser Stelle ist von

57 In diesem Sinne ist die Kompetenz eines Lerners nicht mit der eines Muttersprachlers vergleichbar, was dazu führen sollte, dass die Bewertungskriterien (vgl. Kapitel 5) darauf achten sollten, welche Kompetenzen sie eigentlich messen wollen.

58 Krashen unterscheidet zwischen unbewusstem Spracherwerb und bewusstem Sprachlernen

dynamischem Wissen die Rede, durch welches das deklarative Wissen seine Anwendung findet. Prozedurales Wissen impliziert verschiedene Sprachrezeptions-, Sprachproduktions- und Interaktionsverfahren, wie z. B. Kommunikationsstrategien (Edmonson/House 1993:270). Der GER (2001:105ff.) definiert prozedurales Wissen als die Kombination aus praktischen und interkulturellen Fertigkeiten. Vervollständigt sieht der GER die allgemeinen Kompetenzen durch zwei weitere Faktoren: die persönlichkeitsbezogene Kompetenz und die Lernfähigkeit (s. Kapitel 2.2.1). Weinert (2001) definiert verschiedene Faktoren, die den individuellen Ausbildungsgrad der Kompetenz zu bestimmen scheinen:

- Motivation
- Erfahrung
- Handeln
- Können
- Verstehen
- Wissen
- Fähigkeit

Das Ziel standardisierter Sprachzertifizierungsprüfungen ist es, Sprachkompetenz zu messen. Das wird anhand der vier klassischen Fertigkeiten getan, für welche Aufgaben konstruiert werden:

Hörversten	Leseverstehen	rezeptiv
Sprechen	Schreiben	produktiv

Tabelle12: Die klassischen Kompetenzen

Was das Hör- und das Leseverstehen angeht, so handelt es sich um rezeptive Sprachleistungen, denn hier bedeutet Verstehen eine besonders aktive Tätigkeit. Mündlicher und Schriftlicher Ausdruck liegen den produktiven Sprachleistungen zugrunde. Wie im Vorfeld schon angedeutet, stellt sich meines Erachtens die Frage auf, ob diese klassischen Kompetenzen tatsächlich separat und unabhängig voneinander existieren können. Die im Mittelpunkt dieser Dissertation stehende Teilkompetenz ist die schriftliche Lernerproduktion. Der KLP definiert Schreibkompetenz als „eine zielgerichtete Fähigkeit, Texte herzustellen, indem das Schreiben fortlaufend und bewusst durch die folgenden Elemente gesteuert wird:[59]

59 http://db.learnline.de/angebote/deutschunterrichtsentwicklung/module/teil-2.pdf , S. 41 Zugriff am 1. November 2007, S. 48

- thematische und kommunikative Ziele
- gesammelte und geordnete Sachverhalte
- das metakognitive Wissen um die Prozessschritte des Schreibers
- Auswahl angemessener Schreibstrategien
- Kenntnis geeigneter Textordnungsmuster
- Beherrschung spezieller Prozeduren (Planung, Gliederung)

Der Kernlehrplan NRW ist für die Kompetenzerwartungen am Ende der Jahrgangsstufen 6, 8 und 10 kompetenzorientiert. Ziel der definierten Kompetenzen ist es, dass sie den Kern des erworbenen Wissens und Könnens bilden. Die im KLP definieren Kompetenzen für den gymnasialen primärsprachlichen Unterricht der Klassen 5-10 basieren auf den in der Grundschule erworbenen Kompetenzen und sollen eine Progression über die Jahrgangsstufen aufzeigen.[60] „Diese für den Deutschunterricht in Nordrhein-Westfalen verbindlichen Fachkompetenzen“ beruhen auf den schulformübergreifenden Bildungsstandards der Kultusministerkonferenz (KMK), wodurch „die Vergleichbarkeit der fachlichen Anforderungen in allen Schulformen der Sekundarstufe I gesichert werden soll“[61].

Die erwarteten Kompetenzen werden in Zwei-Jahres-Intervallen zum Ziel gesetzt. Dabei beschränkt sich dieses Modell auf wesentliche Kenntnisse und Fähigkeiten. Für die vorliegende Arbeit sind Aufgaben und Ziele des Deutschunterrichts, speziell des Schreibens von Textsorten von Interesse. Laut des Kernlehrplanes, soll „am Ende der Sekundarstufe I Sprache, sowohl schriftlich als auch mündlich, bewusst und differenziert verwendet werden können. Die Schüler sollen sach-, situations- und adressatengerecht sprechen und schreiben und die Wirkung der Sprache einschätzen können. Sie sollen über unterschiedliche Schreibformen verfügen, deren Funktion kennen und mit ihrer Hilfe ihre Argumentations- und Analysefähigkeiten entwickeln. Es ist aber ebenso wichtig, Schreibformen kennen zu lernen, die die kreativen Anlagen entwickeln.[62] Interessant hierbei ist die Realisierung all dieser Zielsetzungen für das Fach Deutsch konkret und für die fächerübergreifende Funktion der Sprache.[63] Die Idee diesen kompetenzorientierten Rahmen anzuführen, besteht darin, herauszufinden, weshalb die Testanbieter von Sprachprüfungen im DaF-Bereich zum einen bestimmte Kompetenzen und zum anderen konkrete Textsortenformate für den schriftlichen Ausdruck abverlangen. Das könnte in der Annahme begründet liegen, dass auf ein derartiges primärsprachliches Curriculum Bezug genommen wird, wo die Kompetenzen entsprechend der kognitiven Entwicklung und des Reifeprozesses der Schüler aufeinander aufbauen.

60 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=4&doc=d-gy, Zugriff am 1. November 2007

61 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=3&doc=d-gy, Zugriff am 4. November 2007

62 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=2&doc=d-gy, Zugriff am 4. November 2007

63 Das ist insofern wichtig, als dass in den Sprachprüfungen zum Beispiel fachspezifische Themen herangezogen werden, um schreibproduktiv zu werden.

Die Kompetenz des Schreibens wird im Kernlehrplan als ein Prozess betrachtet. Die für die vorliegende Arbeit wichtigsten Punkte des Schreibprozesses dieses Rahmens sind[64]:

- gemäß den Aufgaben und der Zeitvorgabe einen Schreibplan erstellen, sich für die angemessene Textsorte entscheiden und Texte ziel-, adressaten- und situationsbezogen, ggf. materialorientiert konzipieren und gestalten
- Aufbau, Inhalt und Formulierungen eigener Texte hinsichtlich der Aufgabenstellung überprüfen (Schreibsituation, Schreibanlass)
- Strategien zur Überprüfung der sprachlichen Richtigkeit und Rechtschreibung anwenden

Die Kompetenzerwartung Schreiben des Kernlehrplans ist unterteilt in die Kategorien Schreiben als Prozess, Texte schreiben und produktionsorientiertes Schreiben. Die für vorliegende Arbeit relevanten Inhalte jeder Kategorie sollen anhand einer Tabelle aufgezeigt werden:

	Ende der Jahrgangsstufe 6	Ende der Jahrgangsstufe 8	Ende der Jahrgangsstufe 10
Schreiben als Prozess	Schreibzielsetzung und Anwendung elementarer Methoden der Textplanung (z. B. Notizen, Stichwörter), Textformulierung und Textüberarbeitung.	Die Schülerinnen und Schüler gestalten Schreibprozesse zunehmend selbstständig. (z. B. den Text nach den Normen der Sprachrichtigkeit überarbeiten, stilistische Varianten erproben und Formulierungsent-scheidungen begründen)	Die Schülerinnen und Schüler beherrschen Verfahren prozesshaften Schreibens (z. B. Texte ziel-, adressaten- und situationsbezogen, ggf. materialorientiert konzipieren; strukturiert, verständlich, sprachlich variabel und stilistisch stimmig zur Aussage schreiben; sprachliche Mittel einsetzen; Aufbau, Inhalt und Formulierungen hinsichtlich der Aufgabenstellung überprüfen; Texte inhaltlich und sprachlich überarbeiten; Strategien der Überprüfung der sprachlichen Richtigkeit und Rechtschreibung anwenden; über die notwendige fachspezifische Begrifflichkeit verfügen; in gut lesbarer handschriftlicher Form und in einem der Situation entsprechenden Tempo schreiben;

64 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=3&doc=d-gy, Zugriff am 4. November 2007

Texte schreiben	Sie erkennen und bewerten Formen appellativen Schreibens in Vorlagen und verfassen einfache appellative Texte. (z. B. für die eigene Auffassung in einem Leserbrief in der Schülerzeitung werben) Sie entwickeln und beantworten Fragen zu Texten und belegen ihre Aussagen Sie formulieren Aussagen zu diskontinuierlichen Texten. (z. B. einfache Tabellen, Grafiken	Sie informieren, indem sie in einem funktionalen Zusammenhang berichten (über ein Ereignis, einen Missstand in Form einer Reportage) Sie erklären Sachverhalte und Vorgänge in ihren Zusammenhängen differenziert. (z.B. die Bedeutung und Aufgabe von Organisationen, Maßnahmen und Veranstaltungen, das Verhalten von Figuren) Sie gestalten appellative Texte und verwenden dabei verschiedene Präsentationstechniken. (z. B. zu Umweltfragen, schulischen Konflikten einen kritischen Kommentar, einen Aufruf verfassen) Sie formulieren Aussagen zu diskontinuierlichen Texten und werten die Texte in einem funktionalen Zusammenhang an Fragen orientiert aus. (z. B. Diagramme, Übersichten, Grafiken) Sie kennen und verwenden einfache standardisierte Textformen. (z. B. Anträge, Anfragen, Anzeigen)	Sie verfassen unter Beachtung unterschiedlicher Formen schriftlicher Erörterung argumentative Texte. (Thesen entwickeln, Argumente sammeln, nach Gewichtigkeit ordnen; Argumente durch Beispiele veranschaulichen, Schlussfolgerungen ziehen; für eine eigene Auffassung mithilfe wertender Akzentuierung argumentieren, Argumente überlegt anordnen; Gegenargumente zurückweisen (z. B. in einem Leserbrief Sie nutzen Formen appellativen Schreibens bewusst und situationsangemessen. (z. B. für Vorlagen bzw. in Anlehnung an Vorlagen werbende Texte verfassen - Lektüre eines Buches, eine Theaterveranstaltung; den appellativen Charakter eines Textes beschreiben, bewerten) Sie verfassen formalisierte kontinuierliche/diskontinuierliche Texte und setzen diskontinuierliche Texte funktional ein. (z. B. Protokoll, sachlicher Brief, Annonce, Cluster, Mindmap, Grafiken, Schaubilder, Statistiken in Referaten). Sie kennen, verwenden und verfassen Texte in standardisierten Formaten. (z. B. Bewerbungsschreiben, Geschäftsbrief, auch unter Nutzung diskontinuierlicher Texte: Diagramme, Übersichten u. Ä.)

Produktions-orientiertes Schreiben	Sie verfassen Texte nach Textmustern		

Tabelle 13: Die Kompetenzerwartung Schreiben nach dem Kernlehrplan NRW[65]

Diese Tabelle veranschaulicht die Kompetenzerwartungen des Kernlehrplanes NRW hinsichtlich des Schreibens als Prozess und als Text in den einzelnen Abstufungen. Das prozessorientierte Schreiben von Schülern im primärsprachlichen Unterricht geht von der Methodik elementarer Texplanung (Ende der Jahrgangsstufe 6) über die selbständige Gestaltung eines Schreibprozesses (Ende der Jahrgangsstufe 8) über zur Beherrschung des Verfahrens prozesshaften Schreibens (Ende der Jahrgangsstufe 10). Anhand der Kompetenzen des Schreibprozesses zeigt sich folglich die Kompetenz, Texte zu schreiben. Hierbei geht es um das Verfassen verschiedener Textsorten, die aufbauend beherrscht werden sollen. Einen einfachen appelativen Text bzw. einen Leserbrief oder Aussagen zu diskontinuierlichen Texten zu verfassen, soll bereits am Ende der Jahrgangstufe 6 geleistet werden können. Das Produzieren eines Leserbriefes begegnet uns in den hier diskutierten Prüfungen erstmalig beim B2-Zertifikat des Goethe-Institus. Aussagen zu diskontinuierlichen Texten schriftlich zu bearbeiten, ist der Inhalt schriftlicher Aufgabenstellungen auf dem Niveau C1 des Goethe-Instituts und im TestDaF zu finden. Die Textsorte, die Muttersprachler im Sinne des Kernlehrplans bereits am Ende der 6. Klasse beherrschen können müssen, erscheint erstmalig bei dem B2-Zertifikat des Goethe-Instituts, wobei ab Herbst 2008 laut Prüfungsordnung das 16. Lebensjahr nicht mehr vorausgesetzt sein wird.[66] Dennoch bleibt die Frage völlig unbeantwortet, welcher Schwierigkeitsgrad und welche Inhalte diesbezüglich gegeben werden und ob diesbezüglich die Paralellen bereits zu der 6. Jahrgangsstufe gezogen werden können oder ob diese später anzusiedeln sind (etwa am Ende der 8. Jahrgangsstufe), wo differenzierteres und funktionales Schreiben mehr im Mittelpunkt steht. Am Ende des Moduls der nächsten zwei Jahrgangsstufen sollen bereits behandelte Textformen der vorherigen Etappe differenzierter bearbeitet werden können. Zum Beispiel sollen nunmehr nicht nur Aussagen zu diskontinuierlichen Texten formuliert, sondern zusätzlich noch in einem funktionalen Zusammenhang ausgewertet werden. Diese differenziertere Kompetenzerwartung und Bearbeitung einer schriftlichen Aufgabe wird also am Ende der 8. Jahrgangsstufe angesetzt. Es ist jedoch der Umstand nicht geklärt, ob die Prüfungskandidaten das Verfassen derartiger Textformen auch in ihrem primärsprachlichen Unterricht erworben haben, um derartige Aufgabenformate sinngemäß zu bearbeiten. Der Kernlehrplan rundet sich am Ende der 10. Jahrgangsstufe ab. Während appelative Texte in Form eines Leserbriefes nun auch argumentativ betrachtet werden sollen, erwartet man von den Schülern dieser Klassenstufe, dass sie diskontinuierliche Texte nun auch funktional einsetzen. Der Komplexitätsgrad der

65 Die Kompetenzerwartungen sind hier auf die Relevanz dieser Arbeit beschränkt und zusammen gefasst. Der Kernlehrplan NRW für den primärsprachlichen Unterricht umfasst natürlich viel mehr Kompetenzerwartungen als die hier angeführten.

66 Das Goethe-Institut Athen, als Beispiel angeführt, hatte bislang eine Sonderregelung, die für das Niveau B2 ein Mindestalter von 15 Jahren ansetzte. Dieser Ausnahmezustand war meines Wissens aber in keinerlei Prüfungsordnung oder Bekanntmachung dokumentiert und gerechtfertigt. Die Altersbegrenzungen werden aber ab Herbst 2008 nun völlig aufgehoben werden.

fachlichen Anforderungen ist sowohl im Unterricht als auch bei der Leistungsbewertung altersgemäß und mit Bezug auf die Anforderungen der Schulformen zu konkretisieren.[67]

Der auf dem Schreibprozessmodell von Hayes/Flower beruhende Kernlehrplan definiert neben den für diese Arbeit interessanten und zu erwartenden Kompetenzen für den schriftlichen Ausdruck den Anspruch an die Sprachreflexion. Hierbei geht es um die Sprache als Kommunikationsmedium und um ihre Funktion. Die zu erwerbenden sprachlichen Mittel und Kompetenzen auf den jeweiligen Jahrgangsstufen, die funktional in unserem Zusammenhang zu sein scheinen, sollen in einer Tabelle zusammen gefasst werden:

	Ende der Jahrgangsstufe 6	Ende der Jahrgangsstufe 8	Ende der Jahrgangsstufe 10
Sprache als Mittel der Verständigung	Erkennen der Abhängigigkeit der Verständigung von der Situation. Erschließen von Äußerungsabsichten einer sprachlichen Form	Erkennen verschiedener Sprachebenen und Sprachfunktionen in mündlichen und schriftlichen Texten; Vergleichen und Unterscheiden von Ausdrucksweisen und Wirkungsabsichten sprachlicher Äußerungen, worüber in eigenen Texten begründet entschieden wird	Kenntnis verbaler/nonverbaler Kommunikations-strategien und gezieltes Einsetzen. Kenntnis verschiedener und grundlegender Textfunktionen
Sprachliche Formen und Strukturen in ihrer Funktion	Wortarten erkennen und unterscheiden und Funktion bestimmen; Kennen und Anwenden von Flexionsformen; Beschreibung von Satzstrukturen; Untersuchen von Wortbildung; Verfügen und Anwenden von operationalen Verfahren	Sicherer und funktionaler Gebrauch von Wortarten; Verbflexionsformen, deren funktionaler Wert erkannt wird; Bezeichnen und Bilden komplexer Satzgefüge; Sichere Erschließung und korrekte Anwendung von Wortbedeutungen; Anwendung operationaler Verfahren der Satz- und Textstruktur	Kenntnisse in Bezug auf Funktion, Bedeutung und Funktion von Wörtern; Beherrschen der Verbflexionsformen und deren Funktionen und Anwendung beim Schreiben eigener Texte; Differenzieren und Erweitern der syntaktischen Kenntnisse zum Schreiben eigener Texte; Beherrschen sprachlicher Verfahren
Sprachvarianten und Sprachwandel	Unterscheidung zwischen mündlichem und schriftlichem Sprachgebrauch; Erkennen und	Unterscheidung von Sprachvarianten; Erkennen von Zusammenhängen zwischen Sprachen und Nutzen zum Erlernen fremder Sprachen	Reflexion von Sprachvarianten; Reflexion der eigenen Sprache und ihre Bedeutung für das Erlernen von

67 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=4&doc=d-gy, Zugriff am 15. November 2007

	Nutzen verschiedener stilistischer Ebenen; Sprache im Kontrast als Mittel zur Erle rnung einer Fremdsprache		Fremdsprachen
Richtig Schreiben - Laut-/Buchstabenebene	Übervertieftes Wissen der Laut-Buchstaben-Zuordnung	Weitgehende sichere Anwendung des Wissens lautbezogener Regelungen, auch in schwierigen Fällen	Beherrschen von lautbezogenen Regelungen
Richtig Schreiben - Wortebene	Beherrschen von wortbezogenen Regelungen und deren Ausnahmen	Verfügen über weitere wortbezogene Regelungen	Sichere Verwendung wortbezogener Regelungen
Richtig Schreiben – Satzebene	Kennen und Beachten satzbezogener Regelungen	Kennen und Beachten satzbezogener Regelungen	Beherrschen weiterer satzbezogener Regelungen
Richtig Schreiben - Lösungsstrategien	Korrektur und Fehlervermeidung durch richtiges Abschreiben, Sprech- und Schreibproben, Fehleranalyse und Wörterbücher	Kontrolle der Schreibung mithilfe eines Wörterbuchs, der Benutzung von Textverarbeitungsprogammen und der Fehleranalyse nach individuellen Fehlerschwerpunkten	Korrektur und Fehlervermeidung mittels eines Wörterbuchs, Computerprogrammen und selbständiger Fehleranalyse

Tabelle 14: Die Kompetenzerwartung Sprachreflexion nach dem Kernlehrplan NRW[68]

Das Kriterirum der Sprachreflexion verweist im Sinne dieser Arbeit auf verschiedene Aspekte. Zum einen wird ersichtlich, welche Anforderungen im primärsprachlichen Deutschunterricht gestellt werden. Die verschiedenen Untergliederungen des Kriteriums Sprachreflexion definieren die unterschiedlichen Ebenen von Sprache. Sprache soll als Kommunikationsmedium aufsteigend insofern erschlossen werden, dass Sprachnormen, Sprachintentionen und Textfunktionen ihre Berechtigung finden. Die nächste Etappe will die sprachlichen Formen und Strukturen funktional näher bringen, mit dem Ziel, diese sicher und entsprechend der Kommunikationssituation anzuwenden. Eine weitere Unterkategorie der Sprachreflexion besteht darin, Sprachstile und -varianten zu unterscheiden. „Richtig Schreiben" bezieht sich hier auf die orthografischen Aspekte bzw. das Bild einer Sprache.

68 http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=4&doc=d-gy, Zugriff am 15. November 2007

An dieser Stelle besteht meines Erachtens eine Paralelle zu den von Testanbietern aufgestellten Bewertungskriterien hinsichtlich der schriftlichen Sprachproduktion, die im Kernkapitel dieser Dissertation angeführt wird (Kapitel 5). Während die für die sprachkommunikative Kompetenz sowohl die „Kenntnis verbaler bzw. nonverbaler Kommunikationsstrategien und das gezielte Einsetzen" als auch „die Kenntnis von Textfunktionen" erst am Ende der Jahrgangsstufe 10 für den primärsprachlichen Unterricht angesetzt werden, findet sich dieses in den jeweiligen Sprachprüfungen der Niveaus B2/C1 in den oberen Deskriptoren mit den maximal zu erreichenden Punkten eines Bewertungskatalogs wieder. Das führt aber meines Erachtens zu einer paradoxen Lage, denn wie können im fremdsprachlichen Bereich jene Kompetenzen abverlangt werden, die altersmäßig zeitgleich oder später im primärsprachlichen Unterricht als Lernziel in den Mittelpunkt rücken?[69]

Weiterhin ist das Wissen über Bedeutungen oder grammatische Strukturen eine unabdingliche Komponente der Sprachkompetenz. Diese Klassifikation ist nach Wienold (1973:78ff.) nicht dagegen abgesichert, dass anhand ihrer in *mechanischer* Weise bestimmte Kompetenzen geübt werden, ohne ein an Kommunikation orientiertes Lernziel, wie z. B. freie Interaktion mit Muttersprachlern, durchzusetzen. Es stellt sich an dieser Stelle demnach die Frage, ob diese Form der Sprachkompetenz tatsächlich ausreicht, dass man in einer Sprachprüfung Werte erreicht, die aussagen, dass man eine Sprache gemäß der festgesetzten Fähigkeitstabellen[70] auch tatsächlich *kann*. In diesem Zusammenhang ist für Wilkinson (1971:115ff.) das Beherrschungsausmaß einer Kompetenz von Bedeutung, welches in Korrelation zu dem anderer Kompetenzen steht. Weitere einfließende Komponenten wie die soziolinguistische Kompetenz spielen im Rahmen der Spracherlernung oder auch Sprachbeherrschung eine entscheidende Rolle. Auch wenn ein Lerner beispielsweise den Plural im Deutschen einwandfrei beherrscht, bleibt dennoch ungeklärt, ob er weiß, wie man sich zum Beispiel in bestimmten Situationen kommunikativ verhält bzw. in dem Kulturkreis der Zielsprache adressatengerecht ausdrückt. Da Normen und Verhaltensmuster nicht in Grammatiken zu finden sind, bedarf es landeskundlicher Informationen, um die soziolinguistische Kompetenz zu maximieren (vgl. Nodari 2002). Ein weiterer entscheidender Punkt, der angesprochen werden sollte, ist die sprachlogische Kompetenz. Dieser Begriff ist geprägt von Kohärenz und Komplexitität der Zielsprache. Komplexes Textverständnis, Verfassen kohärenter Texte oder Briefe im schriftlichen Ausdruck, oder im mündlichen Ausdruck zu einem Thema Argumente finden und diese zu vertreten, definieren diese Art der Kompetenz. Man könnte die sprachlogische Kompetenz als eine Basiskompetenz betrachten, da erworbene Fertigkeiten der Muttersprache oft in die Zielsprache transferiert werden. Ein Beispiel hierfür wäre das Verfassen eines Leserbriefes für das Niveau B2 (vgl. Pyramide in Kap. 3.4). Voraussetzung hierbei ist, dass man wissen sollte, dass ein Leserbrief sachbezogen zu sein hat. Ob diese Textform in der Muttersprache der Lerner denselben Normen folgt, kann nicht als gegeben betrachtet werden. Trotzdem wird in der erstellten Prüfungen (z.B. des Goethe-Instituts) verlangt, eine derartige Aufgabe zu erfüllen. Konträr dazu könnte man jetzt aber die Überlegung anstellen, dass die Kritik nicht nur an den Testanbietern zu üben ist. Die Curricula bzw. Rahmenlehrpläne der jeweiligen Länder sollten derart erstellt sein, dass der Unterricht auf das Können diverser Kompetenzen und schließlich die Bewältigung dieser in der Testsituation abzielt.

69 In der bis Herbst 2008 geltenden Prüfungsordnung des Goethe-Instituts ist das Mindestalter für das Ablegen einer Prüfung B2/C1 auf 16 Jahre gesetzt. In der neuen PO ist diese Regelung wie bereits angeführt allerdings aufgehoben.

70 Hiermit sind die Can-Dos des GER gemeint

Weitere Grundbedingungen hierfür sind die nötige sprachliche Kompetenz, das Alter und die damit verbundene Reife und schließlich soziolinguistische Normen, wie z.B. Briefaufbau oder Anredeformeln, die es im Rahmen der gestellten Textsorte konkret einzuhalten gilt. Spricht man des Weiteren von strategischer Kompetenz, so fasst man darunter die Fähigkeit, Verständigungsprobleme oder Erwerbsprobleme mit Hilfe von Strategien zu lösen (Nodari 2002:12). Der Bereich der klassisch definierten Sprachkompetenz beinhaltet noch weitere Unterbereiche, die verschiedene Perspektiven und Komponenten einer Sprache aufzeigen. Aus diesem Grund können die sprachlichen vier Fertigkeiten im hier definierten Sinne meines Erachtens nicht ausreichen, um Sprache zu „begreifen" und zu produzieren. Je nach Zielsetzung sollten die verschiedensten Kompetenzen der Priorität entsprechend definiert werden.

Auf dem Stand heutiger Sprachtests, die gemäß der vom GER definierten Niveaus Anwendung finden, werden die vier Fertigkeiten Leseverstehen, Hörverstehen, Schreiben als Testinhalt verstanden. Dabei beachtet der GER (2001:103), „dass Sprachverwendende und Sprachlernende eine Reihe von Kompetenzen einsetzen, um die in kommunikativen Aufgaben und Aktivitäten auszuführen". In Kapitel 2.1.1 und 2.1.2 sind die nötigen Elemente und Komponenten zur Sprachverwendung im Sinne eines handlungsorientierten Ansatzes bereits angeführt worden. Darüber hinaus wurden vom GER verschiedene Skalen entworfen, die zur Sprachverwendung benötigten Elemente beschreiben sollen. Somit ist vom GER ein Stufenkonstrukt entwickelt worden, welches eine bestimmte „Sprachnorm" vorgibt und die „Korrektheit" von Lernerproduktionen oder gar Lernersprachen[71] bestimmen soll. Diese nach Kohn (1990:68ff.) definierte außenweltliche Perspektive definiert die objektiven Fehler eines Testkandidaten, wobei die Korrektheit die Feststellung eines externen Beobachters ist. Die innenweltliche Ansicht ist demnach eine vollkommen subjektive Wahrnehmung oder das eigene Korrektheitswissen des Lerners bzw. Prüflings.

71 vgl. dazu die Interlanguagehypothese

4 Von der Testtheorie bis zur Testbewertung

Im vierten Kapitel der vorliegenden Arbeit soll der Frage nachgegangen werden, was Tests sind, was ihnen zugrunde liegt und sie charakterisiert, welchem geschichtlichen Hintergrund sie unterliegen und welche Rolle die Disziplin der Testtheorien hinsichtlich Tests bzw. Sprachtests und ihres Qualitätsstandards einnimmt. Dazu sollen zunächst wichtige Begriffe der verschiedenen Testtheorien definiert und kontrastiv gegenüber gestellt werden. Verschiedene Kriterien für die Entwicklung, Auswahl und Bewertung eines Tests sind Kernpunkt der Testtheorien. Die wichtigsten Ansätze und Schlüsselwörter werden die Basis und das nötige Verständnis bereitstellen, um Testerstellung, Testdurchführung und Testauswertung im engeren Sinne zu begreifen. Zentral ist die Frage, ob Tests die Zieleigenschaften oder auch -merkmale messen, das heißt inwiefern von Testvalidität ausgegangen werden kann und wodurch diese letztlich gekennzeichnet ist. Auf diesem Hintergrund sollen die Standards der APA den Bezugsrahmen bilden, um Prüfverfahren samt ihren Gütekriterien zunächst generell und dann speziell zu diskutieren. Außerdem sollen sich die Standards der APA für die Testentwicklung als notwendig erweisen. Abschließend sollen im zweiten Teil dieses Kapitels die dafür erforderlichen und erwarteten Kompetenzen, die das zu messende Merkmal in einer Testsituation ausmachen, in den Mittelpunkt gestellt werden.

4.1 Was ist ein Test[72]?

Die Begriffsdefinition von Test hängt von dem jeweiligen Zweck ab. Oft werden die Begriffe „Test" und „Prüfung" weitgehend synonym verwendet.[73] Wenn Unterschiede vorgenommen werden, dann bezeichnet „Test" weniger formalisierte Testverfahren, welche vor Ort konzipiert und durchgeführt werden, „Prüfungen" sind hingegen eher formalisiert und standardisiert (vgl. Perlmann-Balme, 2001).[74] Ein Test kann zum Beispiel aus methodischer Sicht als ein Verfahren gekennzeichnet werden, das Aufschluss über einzelne Personen, Objekte und Situationen gibt. Kennzeichnend hierbei ist, dass die Durchführung von Tests und Prüfungen im Regelfall eine konkrete Absicht oder Zielsetzung verfolgt (Wottawa 1980:11).

Eine klassische und relativ zeitlos ausgewählte testtheoretische Definition ist die von Lienert (1961:7), wobei deutlich wird, dass der Begriff „Test" als Fachbegriff mit vielfältigen Bedeutungen verwendet wird:

72 Ich werde für die Zwecke dieser Arbeit von Prüfungen sprechen, wenn es um das Goethe-Institut und das TestDaF-Institut geht, da es sich bei beiden Testanbietern um standardisierte Tests handelt. Dennoch wird in diesem Kapitel der Begriff „Test" benutzt, wenn es generell um testtheoretische Fragen geht (vgl. Fußnote 1).

73 Dem Zweck dieser Arbeit entsprechend müsste geklärt werden, ob es einen Unterschied zwischen einer Sprachprüfung und einem Sprachtest gibt. Nach dem Multilingual glossary of language testing terms versteht man unter einem Test die „Prozedur zur Feststellung der fremdsprachlichen Leistungsfähigkeit" (1998:127). Außerdem werden zwei weitere Bedeutungsvarianten angegeben. Zum einen der Test als Bezeichnung für den Teil einer Prüfung und zum anderen der Test als informelles Prüfverfahren. Der Bedeutungsumfang von „Prüfung" ist enger, er deckt sich mit der ersten Variante von „Test", d.h. der Prozedur zur Feststellung der Leistungsfähigkeit oder des Kenntnisstandes von Personen durch mündliche und/oder schriftliche Aufgaben. Das Erreichen einer Qualifikation (z.B. durch ein Zertifikat bestätigt) oder der Zugang zu einem Studium kann von dem Ergebnis abhängen (Multilingual glossary of language testingt terms, 1998:119).

74 Auf diese Unterscheidung trifft man auch beim fachlichen Sprachgebrauch. Dort unterscheidet man jedoch noch weitere Bedeutungen von „Test"

„Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung[75].“

Routineverfahren beruhen somit auf Tests, die bereits hinreichend in der Praxis angewendet und definiert worden sind. Wissenschaftlich ist ein Test dann, wenn er nach bestimmten Regeln konstruiert bzw. entwickelt wurde und dadurch gewährleistet ist, dass gezielte Aussagen über Persönlichkeitsmerkmale, Fähigkeiten oder Fertigkeiten (z. B. Sprachkompetenz) zunächst operationalisiert und schließlich formuliert werden können, um der Testtheorie zu entsprechen (Rost 1996:17). Durch die wesentlichen Bestimmungsstücke dieser Definition soll die Testdurchführung aller Testteilnehmer durch das Prüfverfahren, das als hoch standardisiert gilt, vergleichbar gemacht werden (Sommer 1971:192). Als Einwände für diese Theorie könnte man erheben, dass nicht immer nur quantitative, sondern auch qualitative Merkmalsausprägungen Ziel des Testens sein können. Während quantitative Merkmale, die es zu messen gilt, Personen anhand der entsprechenden Ausprägung des zu messenden Merkmals differenzieren, unterscheiden qualitative Merkmale hingegen ganze Gruppen voneinander (vgl. Norm- und Kriteriumsorientierung, Kap. 4.2).

Die APA hat Standards definiert und veröffentlicht, um in erster Linie die Basis bzw. den Referenzrahmen bereitzustellen und im Weiteren die Qualität in der Testpraxis zu bewerten. Grundlegendes Ziel dabei ist die Gewährleistung der Kriterien für den gesamten Testprozess. Dabei sind die für bestimmte Bedingungen relevanten Standards vor der Testanwendung definiert. Die APA führt des Weiteren mehrere Indikatoren an, die am Testprozess beteiligt sind. Die im Zusammenhang dieser Arbeit interessanten von der APA angeführten Indikatoren sind: Testentwickler, Personen, die Tests vermarkten, Testbewerter und schließlich diejenigen, die entweder freiwillig oder notwendigerweise Tests ablegen. Die Standards sind für die unterschiedlichen Indikatoren definiert, aber zugleich beschreibt die APA die Schwierigkeit, ihnen die definierten Standards zuzuordnen. Dennoch fokussiert sie in erster Linie auf die Verantwortung bestimmter Indikatoren:

„(…) the test development process, which focus primarily on the responsibility of test developers (…), specific uses of applications, which focus primarily on responsibilities of test users (…) and the rights and responsibilities of test takers” (APA 2004:2).

Die Standards basieren auf der Prämisse, dass alle Indikatoren des Testprozesses die jeweils maximale Leistung erbringen sollten, damit von effektivem Testen die Rede sein kann:

“(…) that all participants in the testing process possess the knowledge, skills, and abilities relevant to their role in the testing process, as well as awareness of personal and contextual factors that may influence the testing process” (APA 2004:2).

Im Zusammenhang dieser Arbeit sollte demnach das adäquate Wissen psychometrischer Prinzipien, wie das der Validität, von Testentwicklern und Ratern gegeben sein. Neben anderen Gütekriterien der Testtheorie soll insbesondere die Relevanz der Validität in Kap. 4.4.1 ausführlich angeführt und kritisiert werden.

75 Diese Definition Lienerts lehnt sich an: Warren, H.C.: Dictionary of Psychology. Boston 1934.

Die verwendeten Prüfverfahren sollten wissenschaftlich begründet sein, um systematisch, kontrollierbar und wiederholbar zu sein (Grubitzsch 1999:30). Die APA sieht den Test als einen evaluierenden Prozess an, der eine Verhaltensstichprobe abgibt und mithilfe dessen man unter standardisierten und routinierten Prozessbedingungen bewertet (APA 2004:3). Ein Test sollte Aufschluss über das Resultat und damit verbundenen Faktoren der zu testenden Merkmalsausprägung des Testteilnehmers geben. Zuletzt sollte ein Test als Indikator für empirisch abgrenzbare Eigenschaften, Verhaltensdispositionen, Fähigkeiten oder Kenntnisse verwendet werden (Lienert/Raatz 1994:1). Die APA unterscheidet den Begriff des Testens anhand verschiedener Dimensionen. Dabei standardisiert ein Test den Prozess, bei dem Testteilnehmer auf einen Input reagieren und dieses des Weiteren bewertet wird. Dafür benötigt man laut APA folgende Grundbedingungen, angefangen von der Materialaufbereitung bis hin zur Standardisierung eines Tests (APA 2004:3):

„(…) the mode in which test materials are presented, the degree to which stimulus materials are standardized, the type of response format and the degree to which test materials are designed to reflect or stimulate a particular context”.

Grotjahns Definition (2000:304) fasst den „Test“ als jegliches Prüfverfahren auf, das Individuen unter kontrollierten Bedingungen zu bestimmten Handlungs- und Verhaltensweisen veranlasst, die Rückschlüsse auf existierende Persönlichkeitsmerkmale auf dem Stand bzgl. bestimmter Maßstäbe (Lehrziele) erlauben soll. Nach Schneewind (1969:211) sollte die Testtheorie, die sich mit dem Testen und Messen beschäftigt, gesellschaftliche Bedingungen, formale Testerstellungsmethoden und -prinzipien und schließlich den Testinhalt insgesamt berücksichtigen, um die Entwicklung standardisierter Tests zu gewährleisten. Diese werden zum Beispiel dadurch definiert, dass durch eine Standardisierungsstichprobe die Zielpopulation bestimmt wird (Schelten 1980:73). Somit entstehen Normen, an denen die Leistungen oder Merkmalsausprägungen einzelner Testteilnehmer gemessen werden können.

Was impliziert aber der Begriff des Messens? Die APA definiert in ihren Standards das standardisierte Messen als Synonym zum Begriff Test. Standardisiertes Messen ist der Referenzrahmen der entwickelten Standards und setzt sich folgendermaßen zusammen (APA 2004:3):

„(…) measures of ability, aptitude, achievement, attitudes, interests, personality, cognitive functioning and mental health (…)”.

In diesem Zusammenhang sind Testpersonen die empirisch vorfindbaren Messobjekte, die innerhalb dieser empirischen Beziehungen existieren. Diese müssen jedoch anhand einer Messtheorie bzw. einer Testtheorie begründet werden. Dabei soll unter anderem der Frage nachgegangen werden, wie zum Beispiel die Reaktion von Testteilnehmern auf eine zu bewältigende Aufgabe mit dem zu messenden Kriterium zusammenhängt. Hinsichtlich dessen definiert Rost (2004:21), dass sich die Testtheorie mit dem Zusammenhang von Testverhalten und dem zu erfassenden psychischen Merkmal beschäftigt.

4.2 Sprachtests: Intentionen und Ziele

Die Entscheidung, in welchen sozialen Ort hinein ein Individuum mit einem bestimmten Testresultat klassifiziert wird, wird nicht eigens vom Test gefällt, denn nach Tenopyr (1981:1121) „sind es nicht die Tests, die gegenüber verschiedenen Gruppen voreingenommen sind, sondern die Tester". Die Entscheidung der Tester, oder in unserem Sinne der Rater, wird höchstens vom Test vorbereitet. Die Existenz der Testsituation ist also eine Folge der Absicht, über die Zuordnung zu sozialen Orten zu entscheiden. Cronbach (1980:103) fordert die Rechtfertigung des ganzen Selektionssystems, nicht nur die des Tests. In diesem Sinne ist es ganz irrelevant, welche Formen von Tests benutzt werden und nach welchen Testtheorien diese konzipiert sind.

Es wurde bereits definiert, was einen Test im generelleren Sinne ausmacht. Hinsichtlich der vorliegenden Arbeit muss nun der Frage nachgegangen werden, was ein Sprach(stands)test messen oder gar erfahren möchte. Jung (2001:221) definiert, dass man Sprachtests durchführt, um einen Einblick in die Sprachkompetenz einer Person bzw. eines Lerners zu bekommen. Grotjahn (2000:304f.) hat den Testbegriff „wissenschaftliches Routineverfahren" der am Anfang dieses Kapitel initiierter Definition Lienerts im Sinne formeller und informeller Sprachprüfungen erweitert:

„Unter Test soll jegliches Prüfverfahren gefasst werden, das Individuen unter kontrollierten Bedingungen zu bestimmten Handlungs- und Verhaltensweisen veranlasst, die Rückschlüsse ermöglichen sollen auf zugrunde liegende Persönlichkeitsmerkmale wie Sprachfähigkeit oder Wissenstrukturen, auf spezifische Fertigkeiten wie das Schreiben von fremdsprachigen Zusammenfassungen und/oder auf den Stand in Bezug auf einen bestimmten Maßstab, wie z.B. Lehrziele oder Leistung in einer Vergleichsgruppe."

Diese Definition kann für verschiedene Arten von Sprachtests und ihrem jeweiligen Verwendungszweck stehen und Anwendung finden. Es gibt Sprachtests, die innerhalb einer Gruppe vonstatten gehen und wobei diese den Bezugsrahmen bzw. die kollektive Norm bilden. Die Leistung einer Person wird also am Leistungsniveau der entsprechenden Gruppe gemessen. Diese normorientierte Testart gibt lediglich Auskunft über das Ranking innerhalb der Gruppe und sagt prinzipiell nichts darüber aus, wie man hinsichtlich des zu messenden Kriteriums abgeschnitten hat. Der normorientierte Test kann je nach Gruppe variieren, anders ausgedrückt bekäme man für die gleiche Leistung in zwei verschiedenen Gruppen unterschiedliche Bewertungen, da die Gruppe selbst den Maßstab festsetzt. Derartige Tests sind im Prinzip aussagelos, machen aber den Alltag in der Schulpraxis aus (vgl. Glaboniat/Müller 2006). Lehrer stellen derartige informelle Tests zusammen, um einerseits den Lernerfolg und die Effektivität des Unterrichts innerhalb einer bestimmten Periode definieren zu können und um andererseits daraus resultierend die curricularen Lehrinhalte weiterhin zu planen. Diese informelle Testerstellung und die von Lehrern angesetzten Bewertungskriterien sind im Sinne dieser Arbeit nicht von Bedeutung und werden nicht weiter ausgeführt werden. Zentrales Anliegen ist die Auseinandersetzung mit der Thematik formeller Tests. Diese gelten zunächst als standardisiert und haben im Gegensatz zu der oben erwähnten Testkategorie ein gesetztes Kriterium als Maßstab. Man spricht hierbei von einer kriterienbezogenen Norm. Dabei wird keinerlei Rücksicht auf die interpersonellen Leistungen genommen. Der Sprachstand wird daran gemessen, ob und inwieweit der Kandidat das zu messende Kriterium erfüllt, um anhand dessen seine Sprachkompetenz zu „erkennen" (Vollmer 2003:273). Laut des GER (2001:30) ist die Transparenz und die Vergleichbarkeit von

Leistungen nur dann zu erreichen, wenn das zu messende Kriterium schon im Vorfeld festgesetzt ist. Anhand von kriteriumsorientierten Tests wird direkt ein Ziel definiert. Das Ziel ist in unserem Zusammenhang ein Kriterium, eine ganz bestimmte Kompetenz, nämlich die Schreibkompetenz in der Fremdsprache, zu messen. Kriteriums- oder auch Kompetenzorientierung ist solange kein Synonym für Prüfungsorientierung, wie der „wahre Wert" bzw. die „wahre Kompetenz" eines Prüfungsteilnehmers nicht stabil als objektiv, valide und reliabel ermittelt wird, wenn also nicht von einem akkuraten Rating ausgegangen werden kann. Insgesamt muss sich der Testentwickler einer standardisierten Sprachstandsprüfung im Vorfeld über seine Grundlage und die Testabsicht bewusst sein. Er muss ein Modell der Sprachfähigkeit definieren, worauf er schließlich den Test aufbauen kann. Eine Basis für die Definition der zu messenden Kompetenzen stellt das kommunikative Kompetenzmodell des GER dar (vgl. Kap. 2.2.1, Tabelle der Kompetenzen). Das Sprachfähigkeitsmodell, das zunächst von Bachman (1990) definiert und von Bachman/Palmer (1996) leicht modifiziert wurde, ist bereits exemplarisch angeführt worden. In dieser Arbeit wird Bezug auf das vom GER definierte Modell genommen, welches nach Grotjahn (2001:84) „relativ detailliert die der Verwendung und dem Erlernen von Sprachen zugrunde liegenden allgemeinen und sprachbezogenen Kompetenzen beschreibt". Auf derartigen kommunikativen Modellen basierend, entwickeln Testanbieter wie das Goethe-Institut und das Test DaF-Institut demnach ihre Prüfungen.[76] Speziell im Vorfeld erarbeitete Prüfungscurricula sollen folglich nicht mehr das Wissen an sich überprüfen, sondern Aufschluss darüber geben, welche Kompetenzen und sprachlichen Handlungen von einem Prüfungskandidaten erwartet werden. Dabei werden für das Messen von Sprache bzw. das jeweilige Interlanguagestadium der Lerner die klassischen vier Kompetenzen überprüft, um das Kriterium der Reliabilität zu erfüllen (Wiedenmeyer 2006:56). Erwähnt sei an dieser Stelle auch das DIALANG-Projekt, das ein virtueller Sprachtest ist, der auf den Kompetenzstufen des GER basierend Lernern die Möglichkeit bietet, ihre Sprachkompetenz in den klassischen vier Fertigkeiten (LV, HV, SA, MA) zu überprüfen.[77] Wie kann Sprachkompetenz schließlich ermittelt bzw. gemessen werden? Unter 4.3.1 wird angeführt, welche Itemarten ausgewählt werden müssen, um nach der Bedarfsanalyse und dem zugrunde liegenden kommunikativen Modell einen guten Test zu erstellen, der allen testtheoretischen Gütekriterien sowohl hinsichtlich des Aufbaus als auch der Bewertung gerecht wird. Diese Koppelung ist besonders wichtig. Geschlossene Items implizieren beispielsweise eine prädestinierte Bewertung, d.h. es gibt einen konkreten Lösungsvorschlag, auf den die Rater keinerlei Einfluss haben (können). Geht es aber um offene Aufgabenformate, wie das beim schriftlichen Ausdruck der Fall ist, so ist das zu messende Kriterium auch ein Auftrag an die Rater selber. Es gibt dabei keinerlei Lösungsvorschläge oder Ansätze, die das Ratingverfahren vollständig objektivieren könnten. Wie bei informellen Prüfungen auch, bleibt den Ratern bei diesen Aufgabenformaten standardisierter Prüfungen, die etwas „Produktives" erfordern, ein gewisser Bewertungsspielraum. Standardisierte Tests sollen aber international vergleichbar gemacht werden, um allgemein gültige Aussagen über den Sprachstand bzw. die Struktur von Sprachkompetenz einer Person zu machen (vgl. Perlmann-Balme 2006/Glaboniat/Müller 2006).

76 Das Goethe-Institut basiert seine Prüfungen auf dem Modell der Kommunikationsfähigkeit von Bachman/Palmer, vgl. Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 14

77 http://www.goethe.de/Z/50/commeuro/c.htm Stand 20. 06. 2007

Weiterhin wird bei Sprachtests zwischen Performanz- und Kompetenztests unterschieden. Wie in Kapitel 3 bereits angeführt, ist die Performanz die Realisation der Kompetenz. Demnach prüfen Performanztests das Sprachoutput unter bestimmten Bedingungen und in bestimmten Situationen, während Kompetenztests eher als Sprachwissenstests bezeichnet werden könnten. Grotjahn (2000:322) beschreibt die Modellierung zielsprachlicher Aufgaben und Verwendungssituationen durch Items und die Reaktion darauf als Performanztest. Hingegen liegt den Testaufgaben bei einem Kompetenztest ein konkretes Modell der beim Sprachgebrauch beteiligten Kompetenzen zugrunde. McNamara (1996:43ff.) unterscheidet analytisch zwischen zwei verschiedenen Formen von Performanztests. Während die abgeschwächte Form die sprachliche Performanz in den Mittelpunkt stellt, definiert die starke Version eines Performanztests die Problemlösung und nicht die gezeigte Sprachleistung. Die starke Version eines Performanztests kann meiner Ansicht nach aber die Raterobjektivität in großen Maße bezüglich der Bewältigung der Aufgabenstellung einschränken.

4.3 Verschiedene Ansätze der Testtheorie

Es haben sich verschiedene testtheoretische Modelle herausgebildet, die eine Verbesserung der Testentwicklung anstreben. Während die klassische Testtheorie annimmt, dass sich der beobachtete Wert aus dem wahren Wert und dem Fehlerwert zusammensetzt, geht die probabilistische oder auch stochastische Testtheorie davon aus, dass die Wahrscheinlichkeit der Reaktion einer Testperson funktional aus ihrer Fähigkeit und der Aufgabenschwierigkeit entsteht. Diese zwei und ein weiterer alternativer testtheoretischer Ansatz sollen im Folgenden vorgestellt und der APA gegenüber gestellt werden. Dabei wird auf die im Rahmen dieser Arbeit wichtigsten Aussagen eingegangen, wobei aber die Standards der APA den Bezugsrahmen bilden.

Über 95% aller Tests auf dem Markt werden anhand der Skalierungsmethode der Itemanalyse der Klassischen Testtheorie (KTT) entwickelt. Dabei macht die Klassische Testtheorie weniger Aussagen darüber, ob eine Beziehung zwischen dem Verhalten der Testperson und der latenten Merkmale besteht. Sie fasst eher Annahmen über erzielte Messwerte zusammen. Das Resultat einer Messung im Sinne der KTT wird von Messfehlern überlagert, die unter anderem mit der Itemauswahl zu tun haben (vgl. Kap. 4.3.1). Die Weiterentwicklung der Klassischen Testtheorie findet sich im Begriff der Modernen Testtheorie wieder. Mittels statistischer Prüfungen sollen Modellgültigkeiten in Erfahrung gebracht werden. Während das deterministische Testmodell der KTT dem Alles-oder-Nichts-Prinzip unterliegt, d.h. der Testkandidat löst eine Aufgabe bzw. ein Item oder nicht, basieren probabilistische Testmodelle, wie z.B. die Item Response Theorie auf einem Wahrscheinlichkeitsprinzip. In diesen testtheoretischen Modellen geht man davon aus, dass die Wahrscheinlichkeit einer bestimmten Reaktion auf die festgesetzten, einzelnen Items von einer testunabhängigen latenten Dimension abhängig ist. Beobachtetes Verhalten definiert hier nur einen Indikator für ein latentes Merkmal, auf dessen Ausprägungsgrad geschlossen werden soll. Dieses Modell basiert folglich auf der Annahme eines latenten Kontinuums, auf dem jeder eine bestimmte Ausprägung aufweist (vgl. Müller 1999). Somit wird die Wahrscheinlichkeit einer manifesten Reaktion in Abhängigkeit von der Ausprägung des latenten Merkmals einer Person beschrieben (Embretson/Reise 2000:46f.). Rost/Spada (1982:60) definieren diesbezüglich, dass dem

beobachteten Verhalten eine latente Fähigkeit zugrunde liegt, die das Testverhalten steuert.

Man konnte bereits sehen, welchen Annahmen die KTT und die IRT zugrunde liegen. Erstere nimmt an, dass sich der beobachtete Wert aus einem wahren Wert und einem Fehlerwert zusammensetzt. Letztere geht davon aus, dass die Wahrscheinlichkeit einer Reaktion der Testperson als eine Funktion aus der Fähigkeit der Testperson und der Schwierigkeit der zu bearbeitenden Aufgabe dargestellt werden kann. Das in unserem Sinne wichtige kriteriumsorientierte Messen richtet sich nach der Frage, ob ein Testteilnehmer ein bestimmtes Kriterium erfüllt hat oder nicht. In einem kriteriumsorientierten Test werden Testleistungen eines Testteilnehmers mit inhaltlich genau definierten Zielen verglichen (vgl. auch Klauer 1987). Mithilfe statistischer Verfahren und testtheoretischen Ansätzen wird dann überprüft, was und wie gut das Kriterium getestet, gemessen und bewertet wird.

Interessant ist das Beispiel lehrzielorientierte Tests. Normorientierung definiert die APA (2004:50) als „norms (that) assist in the classification or description of examinees". Kriterienorientiert sind der APA entsprechend die Tests, die sich nicht auf die Fähigkeiten und Leistungen anderer Testteilnehmer, sondern auf die individuelle Leistung eines jeden Testteilnehmers hinsichtlich des zu messenden Kriteriums beziehen. Das aus dem Griechischen stammende „κριτήριο" wird als das Maß zur Bewertungs- bzw. Meinungsbildung definiert. Demnach kann man für die Zwecke dieser Arbeit weitere spezielle Bedeutungen ergänzen. Ein Kriterium kann ein zu erreichendes Lehrziel, ein Leistungskontinuum oder ein Leistungsstandard sein, der gewissen Normen unterliegt. Nach Vollmer (2003:273) müssen kriteriumsorientierte Tests idealnormiert sein, denn das Ziel dieser Leistungsmessung ist die Erfassung der Leistung im Rahmen eines definierten Aufgabenbereichs und von außen rekurriertem Kriterium (im hiesigen Sinne die Kompetenz im schriftlichen Ausdruck).

4.3.1 Das Itemuniversum

Will man ein bestimmtes Merkmal mithilfe eines Tests messen, so sollte man zunächst eruieren, ob ein derartiger Test bereits existiert. Sollte dies nicht der Fall sein, kann mit der Planung einer eigenen Testkonstruktion begonnen werden. Laut APA (2004:37) wird die Testentwicklung vom Testzweck und dem zu messenden Konstrukt gelenkt:

"The process of developing educational and psychological tests commonly begins with a statement of the purpose(s) of the test and the construct or content domain to be measured".

Dabei berücksichtigen die Standards sowohl Inhalt, Format, Testkontext und potentielle Konsequenzen der Testanwendung, als auch konkrete Bedingungen der Testanleitung und der Bewertungskriterien. Die Definition eines Testzwecks führt laut APA erst dann zum Produkt Test, wenn folgende vier Phasen berücksichtigt werden (APA 2004:37):

- **“delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured”**
- **“development and evaluation of the test specifications”**
- **“development, field testing, evaluation, and selection of the items and scoring guides and procedures”**
- **“assembly and evaluation of the test of operational use”**

Die erste Phase besteht demnach darin, das Testkonstrukt als erstes deutlich zu beschreiben und zu definieren. Im zweiten Schritt muss das Konstrukt auf einen Test hin spezifiziert werden, damit es gemessen werden kann. In diesem definierten Geltungsbereich oder auch Referenzrahmen muss im Sinne der Thematik dieser Arbeit der Frage nachgegangen werden, was alles im schriftlichen Ausdruck des Niveaus B2 bzw. C1 gefordert wird. Der nächste Schritt besteht nach der Definition der KTT darin, Items zu entwerfen, mit denen das Konstrukt überprüft werden kann. Diesen Schritt nennt die APA Testspezifizierung (APA-Standard 3.2:43):

„The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgements can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent”.

Darin ist aber nicht nur die Definition der Items, sondern sind zudem Faktoren wie Norm- oder Kriteriumsorientierung enthalten, welche wie folgt definiert werden (APA-Standard 3.4:43):

„The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented”.

Das ist insofern wichtig, denn normorientierte Werte würden an Populationswerten gemessen und eingestuft (qualitativ), während kriteriumsorientierte Werte an dem zu messenden Kriterium festgemacht werden und die individuelle Kompetenz widerspiegeln (quantitativ). Anders ausgedrückt, sollte die Produktion jedes einzelnen Testteilnehmers mithilfe der Aufgabenstellung im schriftlichen Ausdruck B2/C1 der Kompetenz entsprechend abgebildet werden.

Items werden für Testzwecke definiert. Ein Item ist die kleinste Einheit eines Tests, die das interessierende Kriterium inhaltlich repräsentiert (z.B. Aufgabe für den schriftlichen Ausdruck). Eine Menge von Items macht demnach den Test aus. Items können in den verschiedensten Formen auftreten, d.h. ihre Intention bzw. ihr Zweck kann variieren. Items definieren die Art und Weise, in der die Bearbeitung einer Testaufgabe erfolgt.

Die Testentwickler sollen die Items so zusammenstellen, dass sie den Anforderungen der Testspezifizierung im Sinne der APA genügen. Die ausgewählten Items sollten sowohl gut durchführbar, objektiv auswertbar als auch so beschaffen sein, dass die Subjektivität des Bewerters so stark wie möglich begrenzt und auch die Testökonomie berücksichtigt wird (Lienert 1961:22). Die für die vorliegende Arbeit zentralen Aufgabenformate sind offen bzw. frei, wobei keine Alternativen vorgegeben sind. Derartige Itemgestaltungen erfassen eine große Breite des zu messenden Konstrukts. Die Antwort bzw. Reaktion auf dieses Aufgabenformat kann mithilfe des Aufgabeninputs (auch Legende genannt) realisiert werden. Gegen die Benutzung bzw. der teilweisen Übernahme der Legende ist meines Erachtens aber nichts einzuwenden, wenn dies situations- und kontextspezifisch passiert.

Es erweist sich als schwierig, produzierte schriftliche Aufgaben zu standardisieren. Derartige Itemformen bilden den Rahmen für die Kritik ihrer Konstruktion und der Ihnen zur Verfügung stehenden Bewertungskriterien. Das offene Itemkonstrukt des schriftlichen Ausdrucks wird in Kapitel 5 ausführlich und anhand von konkreten Konzepten und existierenden Kriterienkatalogen von Testanbietern für den DaF-Bereich analysiert und beurteilt, sodass die so genannten und in Kap. 4.4 beschriebenen Gütekriterien wie Objektivität, Reliabilität und Validität ihre Gültigkeit beibehalten können.

Gute Items können durch inhaltliche Kriterien definiert werden. Sie sollten die wesentlichen Aspekte des zu messenden Kriteriums erfüllen. Abschließend ist ein gutes Item von seiner speziellen, konkreten und wirklichkeitsnahen Gestaltung abhängig (Lienert 1961:38). Den probabilistischen Testtheorien entsprechend existiert für jedes Item ein so genannter kritischer Wert, ab dem ein Item als gelöst gilt. Die Itemcharakteristik wird in diesem Sinne nicht deterministisch definiert (Wottawa 1980:46ff.). Sie ist hier die Funktion, die jedem Wert auf dem latenten Kontinuum eine Lösungswahrscheinlichkeit zuordnet. Probabilistisch existieren bestimmte Grundannahmen für die Itemcharakteristik. Zum einen lässt sich eine Person hinsichtlich ihrer Fähigkeit, ein bestimmtes Item zu bewältigen, durch den so genannten Personenparameter auf einer eindimensionalen Skala charakterisieren. Unabhängig davon, ob der latente Fähigkeitswert qualitativ oder quantitativ ist, muss er anhand beobachtbarer Variablen geschätzt werden, um zum Beispiel das Kontinuum zu bestimmen.

Tests müssen den Kriterien der Objektivität, der Reliabilität und der Validität gerecht werden (vgl. Lienert 1961/Lienert/Raatz 1998). Das kann aber nur dann erreicht werden, wenn auch seine Items diese Qualitäten besitzen. Objektiv ist ein Item, wenn es unter den Ratern Übereinstimmung darin gibt, ob eine Lösung, in unserem Sinne eine schriftliche Lernerproduktion, richtig oder falsch ist. Von einem reliablen Item spricht man dann, wenn es auch bei Wiederholung auf die gleiche Art und Weise vom Testteilnehmer gelöst wird. Natürlich muss man an dieser Stelle bedenken, dass sich ein zu messendes Kriterium, konkret die schriftliche Kompetenz in der Fremdsprache, nach einer bestimmten Periode sowohl verbessert als auch verschlechtert haben kann. Das Sprachniveau kann entweder fossilieren, zurück gehen oder sich verbessern.[78] Die Validität eines Items ist im Sinne der Trennschärfen zu verstehen. Dabei wird Trennschärfe als ein zentrales Gütekriterium für Items betrachtet, das zwischen guten und schlechten Testteilnehmern bzw. ihren Merkmalsausprägungen trennt (Lienert 1961:36, Rost 2004:369). Im Unterpunkt Itemanalyse wird noch näher auf diese und andere-Itemkennwerte eingegangen.

78 Inwieweit die Reliabilität einer gestellten Aufgabe auch für das offene Aufgabenformat geltend gemacht werden kann, wird sich im 5. Kapitel der vorliegenden Arbeit herausstellen.

4.3.1.1 Itemrevision

Nachdem für die Itemkonstruktion verschiedene Quellen zur Hilfe herangezogen wurden, müssen die Items nun überdacht werden. Bei der Itemrevision soll unter anderem auf sprachliche Verständlichkeit und klaren Formulierungen geachtet werden. Die Items werden laut APA nach ihrer Qualität, Klarheit und dem Mangel an Mehrdeutigkeit überarbeitet und neu formuliert. Der Testentwickler ist schließlich verantwortlich, wenn es um die Gewährleistung der Testspezifizierung geht (APA 2004:39). Der Standard 3.6 des Kapitels *Test Development and Revision* definiert (APA-Standard 3.6:44):

„The type of items, the response formats, scoring procedures, and the test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented".

Gute Verständlichkeit und Eindeutigkeit der Items richtet sich immer nach der Zielpopulation, was eine notwendige Voraussetzung ist (Lienert 1961:60ff.). In Bezug auf die hier vorliegende Arbeit sind die Regeln der Itemerstellung und –revision und folglich der Gewährleistung der Inhaltsvalidität (s. 4.4.1.1) insofern von Bedeutung, als die daraus entstehende Aufgabentypologie und ihre Formulierung entscheidend für das Verständnis der Aufgabe und folglich ihrer Bearbeitung sind (siehe Kapitel 5). Aus diesem Grunde sollte folgender Standard nicht unberücksichtigt gelassen werden (APA-Standard 3.7:39):

„The procedures used to develop, review and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented".

4.3.1.2 Itemanalyse

Die letzte Phase der Testkonstruktion ist schließlich die Itemanalyse, welche zur Erhöhung der Reliabilität und Validität beiträgt, indem nicht adäquate Items ausgeschlossen werden (Lienert 1961:67ff.). Die APA betont, dass für diese Phase der Testkonstruktion das angewandte testtheoretische Modell explizit angegeben werden sollte (APA-Standard 3.9:44):

„When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity fort he procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented".

In dieser Phase der Testerstellung gilt es zu klären, ob die Items der Testabsicht entsprechen. Es wird der Frage nachgegangen, welche Items entweder von allen oder von niemandem gelöst wurden, denn genau derartige Items lassen keinen Rückschluss auf die dem Lösungsverhalten unterstellte Kompetenz der Testperson zu, da es scheint, dass alle Testteilnehmer die Aufgabe gleichermaßen behandelt haben (Grubitzsch 1999:133). Verschiedene Itemkennwerte überprüfen Items nach verschiedenen Kriterien. Die Schwierigkeit eines Items differenziert in der KTT lediglich zwischen hohen und niedrigen Merkmalsausprägungen. Dabei ist der Schwierigkeitsindex die Wahrscheinlichkeit, mit der ein bestimmtes Item innerhalb der Eichstichprobe gelöst wird. Wird ein Item von vielen gelöst, so ist sein Schwierigkeitsgrad eher leicht. Demgegenüber gilt ein Item, das von wenigen gelöst wird, als eines schwierigen Niveaus. Was die Itemschwierigkeit in der probabilistischen Testtheorie betrifft, so spricht man von Aufgabenmerkmal, wenn jedes Item auf einer latenten Dimension einen bestimmten latenten Wert hat.

Dass das Kriterium der Trennschärfe von erheblicher Bedeutung ist, betonte bereits Ziehen (1897), der Prinzipien und Methoden der Intelligenzprüfung definierte (Lienert 1961:8). Trennschärfe ist ein Indikator dafür, wie gut ein einzelnes Item das gesamte Testergebnis repräsentiert. Ein trennscharfes Item soll leistungsfähigere von leistungsschwächeren Testteilnehmern trennen können. Hierbei gibt der Trennschärfekoeffizient Auskunft darüber, ob ein einzelnes Item oder gar der gesamte Test in der Lage ist, die Spreu vom Korn zu trennen. Die Trennschärfe gibt also an, inwieweit die Menge der Lösungen über alle Items identisch bleibt und wird demnach als die Korrelation zwischen Item- und Testscore[79] definiert. Testpersonen, die einen hohen Wert in einem Item erzielt haben, sollten auch in den anderen Testitems hohe Werte erbringen. Dadurch würde die Homogenität der Items gewährleistet werden. Balancierte Items, d. h. mit mittlerer Schwierigkeit, besitzen die größte Trennschärfe. Items, deren Schwierigkeitsgrade klein oder groß sind, sind weniger trennscharf, denn es ist offensichtlich, welche Verteilungen entstehen würden. Bei der Frage wie stark alle Items das gleiche Merkmal messen, bezieht man sich auf die Homogenität (Grubitzsch 1999:136). Dabei wird unterschieden zwischen der Homogenität pro Item und der Homogenität des Tests. Itemhomogenität heißt nach Rost (2004:100), dass alle Items dieselbe latente Variable ansprechen. Homogen sind die Tests, deren Items zwar keine identischen jedoch ähnliche Merkmalsfacetten repräsentieren. Heterogene Tests hingegen besagen, dass Items unterschiedliche Merkmalsfacetten erfassen.

Die Folge der Itemanalyse ist die Itemselektion, wo man verschiedenartig vorgehen kann. Selektiert man inhaltlich, so berücksichtigt man die Anschaulichkeit, den theoretischen Hintergrund, den Aufgabentyp und andere inhaltsgleiche Items. Erfolgt die Auswahl statistisch, dann eliminiert man Items z.B. aufgrund ihrer niedrigen Trennschärfe.

Die hier angeführten Itemkennwerte charakterisieren einen Test von seinen Items her. Ausschlaggebend für einen Test sind zudem verschiedene Gütekriterien, die im Folgenden angeführt werden. Gütekriterien geben Aufschluss darüber, ob ein konstruierter Test die nötigen Qualitäten aufweist. Man unterscheidet klassisch zwischen Haupt- und Nebengütekriterien. Die Begriffe seien im Folgenden aus verschiedenen Perspektiven beleuchtet und entsprechend definiert.

79 Score wird im Sinne eines Werts gebraucht.

4.4 Testtheorien und Gütekriterien

Es wurden bereits die Rahmenbedingungen für Testkonstruktion, Testdurchführung und Testbewertung beschrieben und definiert. Testtheoretische Ansätze und der Referenzrahmen der APA versuchen die Entwicklung und den Gebrauch von Tests zu evaluieren. Worin bestehen aber die praktischen Probleme, die zur Beschäftigung mit Testtheorien führen? Um diese Frage beantworten zu können, sei zunächst ein Beispiel aus der Schulpraxis angeführt:

Ein DaF-Lehrer möchte eine Klassenarbeit im schriftlichen Ausdruck in Form eines persönlichen Briefes zur Überprüfung des Sprachstandes schreiben lassen. Bei der Planung dieses Tests tauchen plötzlich diverse technische und praktische Grundprobleme auf, da ihm leider das Grundwissen über Formen der Testerstellung und Leistungsmessung fehlt (vgl. Bolton/Perlmann-Balme 2006). Der Lehrer stellt sich, in der Funktion des Testentwicklers und Testanwenders, sowohl Fragen bezüglich der Durchführungsobjektivität des Tests und des Aufgabenformats als auch der Bewertung im nachhinein. Zunächst einmal sollte sich der Testentwickler darüber klar sein, dass ein Testaufbau nicht nur Aufgabenerstellung impliziert. Um von einer guten Testkonstruktion auszugehen, müssen verschiedene, auch retrospektive oder formale, Aspekte bedacht werden. Ob auf den Unterricht, das Lernziel, den Testaufbau, den Lerneffekt oder das Verhalten der Aufsichtsperson bezogen, alle diese Aspekte müssen, um einen objektiven, validen, reliablen und neben anderen Kriterien fairen Test zu erhalten, bedacht und entsprechend eingehalten werden. Würden die folgenden Fragestellungen bei einer Testkonstruktion nicht aufkommen oder berücksichtigt werden, dann wäre im Sinne des GER zum einen das Kriterium der Transparenz und Vergleichbarkeit zu anderen Tests (vgl. Glaboniat/Müller 2006) und zum anderen die von der APA definierte Testspezifizierung nicht gegeben:

- Welche Kompetenzen will man im schriftlichen Ausdruck messen bzw. prüfen, um eine valide Aussage über den Sprachstand[80] der Schüler zu bekommen? (vgl. Konstruktvalidität)
- Sind die Prüfungsteile ausreichend und entsprechend formuliert, um die zu messende Kompetenz abzudecken? (vgl. Konstruktvalidität)
- Können anhand der erzielten Testwerte Aussagen über den Grad der zu messenden Kompetenz gemacht werden? (vgl. Konstruktvalidität)
- Sind die Aufgabenformate und deren Schwierigkeit der Kompetenzen und Strategien adäquat? (vgl. Konstruktvalidität)
- Soll jede Aufgabe einen Punktwert bekommen? Wie gewichtet man die Punkteverteilung? (Inhaltsvalidität)
- Ist der Inhalt der gestellten Aufgabe für die Repräsentation des gesuchten Kriteriums ideal ausgewählt? (vgl. Inhaltsvalidität)

80 Zu beachten ist, dass mithilfe eines Tests zum schriftlichen Ausdruck lediglich Aussagen zur schriftlichen Kompetenz gemacht werden können, nicht aber zur mündlichen Kompetenz, da ein derartiger Test dafür gar nicht sensitiv ist. Dieser Fehler wird häufig gemacht, z.B. bei Einstufungstests, die reine Grammatiktests sind und von denen man dann auf das Niveau der Lerner schließen will.

- Eignen sich alle Aufgaben zur guten Messung des Sprachstandes im schriftlichen Ausdruck? Wird zusätzlich noch etwas Anderes gemessen? (vgl. Konstruktvalidität)
- Was würde man messen, wenn man die Aufgaben anders stellen würde? Könnte man zwei parallele Tests erstellen, die austauschbar wären? (vgl. Reliabilität)
- Ist jeder Testteilnehmer über den Verlauf und die Bewertung des Tests informiert? (vgl. Durchführungsobjektivität)
- Ist dieser Test für alle Testteilnehmer gleichermaßen fair? (vgl. Fairness)
- Ist die angesetzte Zeit für diesen Test ausreichend? (vgl. Praktikabilität)

Nachdem der Begriff des Tests bereits definiert und erweitert wurde, kann man an dieser Stelle nunmehr, nach der Betrachtung der Problematik bei der Testerstellung, der Frage nachgehen, was denn einen guten Test, eine gute Testkonstruktion und folglich eine gute Bewertung auszeichnet. Was wird vorausgesetzt, damit von einem guten Test ausgegangen werden kann? Der Deutschlehrer in unserem Beispiel hat vermutlich schon eine Vorstellung darüber, was er bei seinen Schülern testen will, ist sich aber über die korrekte Testerstellung, die Qualität der Messmethode bzw. des Tests und der Einhaltung bzw. Existenz der testtheoretischen Gütekriterien weder im Klaren noch bewusst.

Ein guter Test hat als erstes den Anspruch objektiv, zuverlässig und gültig zu sein. Diese drei und weitere Kriterien sollten sich aus der Sicht der KTT spätestens nach der Itemselektion ergeben, sodass das entwickelte Testkonstrukt entweder beibehalten oder revidiert werden kann. Die als Hauptgütekriterien eines Tests geltenden Schlüsselkonzepte Objektivität, Reliabilität und Validität bedingen sich logisch. Zunächst werden die Hauptgütekriterien auf der APA und den Standards basierend definiert werden. Das Kriterium der Validität stellt für diese Arbeit das wichtigste Kriterium dar, denn die Thematik verlangt die Beantwortung der Frage, ob die Bewertungskriterien schriftlicher Lernerproduktionen valide sind. Es folgen anschließend die Kriterien Reliabilität und Objektivität. Weitere Kriterien, über die es in der Literatur Unstimmigkeiten bezüglich ihrer Rangordnung und Wichtigkeit gibt (vgl. Tschirner 2001), sollen im Sinne der Thematik dieser Arbeit aufgezeigt werden.

4.4.1 Validität

Laut APA ist Validität „the most fundamental consideration in developing and evaluating tests" (APA 2004:9). Validität soll angeben, wie zuverlässig ein Test das misst, was er vorgibt zu messen. Auf die vorliegende Arbeit bezogen heißt dies, dass auch tatsächlich das zu testende Kriterium schriftlicher Ausdruck gemessen werden soll. Der GER definiert ein Beurteilungs- bzw. Messverfahren dann als valide, wenn nachgewiesen werden kann, dass die Information des im jeweiligen Kontext gemessenen Kriteriums eine genaue Abbildung der Kompetenz eines Prüflings ist (GER 2001:172). Testziele müssen demnach klar definiert sein und das in Abhängigkeit zur spezifischen Testverwendung (Grotjahn 2000:312). Die APA definiert den Validierungsprozess folgendermaßen (APA 2004:9):

„The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations".

Dabei sollen sich die bereit gestellten Bewertungsprinzipien auf das zu messende Kriterium beziehen. Der Anwendungsbereich der Bewertungskriterien muss in diesem Sinne beschrieben und explizit auf das zu messende Kriterium bezogen und eingegrenzt werden. In diesem Zusammenhang betont die APA explizit die Wichtigkeit eines Referenzrahmens, der anzeigen soll, wie Testwerte zu verstehen sind (APA 2004:9):

„The detailed description provides a conceptual framework for the test, delineating the knowledge, the skills, abilities processes, or characteristics to be assessed. The framework indicates how the representation of the construct is to be distinguished from other constructs and how it should relate to other variables".

Während im Sinne der APA die Validierung die beabsichtigten Bewertungen bezüglich ihres Gebrauchs unterstützen soll, soll der konzeptuelle Referenzrahmen anzeigen, wie die Testwerte, hier die schriftlichen Lernerproduktionen, zu bewerten sind. Dabei kann es natürlich zu Revisionen kommen, wenn zum Beispiel wichtige Aspekte des schriftlichen Ausdrucks auf Niveau B2/C1 nicht einbezogen wurden und das Kriterium dadurch unterrepräsentiert ist. Entsprechendes gilt für ein überrepräsentiertes Konstrukt. Diesbezüglich wird Konstruktirrelevanz von der APA folgendermaßen definiert (APA 2004:10):

„(...) refers to the degree to which test scores are affected by processes that are extraneous to its intended construct".

Die Bewertung einer schriftlichen Lernerproduktion kann folglich durch Komponenten beeinflusst werden, die nichts mit dem zu messenden Kriterium gemein haben. Aus diesem Grund sollte der Validierungsprozess sorgfältig durchgeführt werden, um möglichen Verzerrungen und Fehlbewertungen auszuweichen. Das 5. Kapitel wird aufzeigen, dass die Revision der Bewertungskriterien für den schriftlichen Ausdruck B2/C1 im Sinne der APA Teil des Validierungsprozesses ist. Zu untersuchen ist in diesem Zusammenhang jedoch, inwiefern die revidierten Bewertungskriterien dem Validitätsbeweis genügen, welcher als „the joint responsibility of test developer and test user" (APA 2004:11) zu betrachten ist.

Die Validität eines Tests ist im Sinne dieser Arbeit eng mit den Lehrzieldefinitionen gekoppelt, die wiederum mit den Modellen des Fremdspracherwerbs und den Strukturen von Sprachkompetenz interagieren (Vollmer 2003:274).

In der Tradition der Testtheorie ist Validität ein Oberbegriff, der verschiedene Auslegungen und Formen zulässt, welche nach dem Verwendungszweck oder dem methodischen Vorgehen unterschieden werden. Die verschiedenen Validitätskonzepte, die Aufschluss über die Validität verschiedener Formen von Schlussfolgerungen geben, werden sehr inflationär verwendet (APA 2004:11). Die Überarbeitung der Standards aus dem Jahre 1985, die uns aktuell vorliegt, bezieht sich weniger auf das Differenzieren verschiedener Validitätsarten, als um Arten des Validitätsbeweises (APA 2004:11):

„To emphasize this distinction, the treatment (...) does not follow traditional nomenclature".

Die traditionellen Validitätsarten, die einer Nomenklatur folgen, sollen im folgenden den verschiedenen Definitionsansätzen des Validitätsbeweises der APA gegenübergestellt werden.

4.4.1.1 Arten des Validitätsbeweises

Die APA definiert Validität im Gegensatz zur klassischen Testtheorie nicht nach der Art sondern nach den verschiedenen Definitionsansätzen. Im Folgenden sollen die wichtigsten Validitätsbegriffe im Sinne dieser Arbeit aus verschiedenen Blickwinkeln betrachtet werden.

Die nach der klassischen Testtheorie definierte Inhaltsvalidität soll angeben, ob der Inhalt der ausgewählten Items das zu messende Kriterium grundsätzlich und erschöpfend erfasst. Der APA entsprechend wird das Inhaltsspektrum eines Tests folgendermaßen aufgefasst (APA 2004:11):

„Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring".

Testinhalt wird von der APA also viel weiter gefasst, als es der klassische testtheoretische Ansatz tut. Dennoch sollten sich Testinhalt und Testzweck im Rahmen des schriftlichen Ausdrucks auf den Niveaus B2/C1 decken. Dabei hängt die Angemessenheit des Testinhalts laut APA mit den Schlussfolgerungen der Testwerte zusammen. Der Testinhalt kann das Prüfungsziel schriftliche Kompetenz mehr oder weniger berücksichtigen. Aus diesem Grund sollte die Bewertung der Testleistung eines Prüflings im Sinne der APA sowohl den berücksichtigten als auch den nicht berücksichtigten Inhalt (*the content neglected and the content addressed)* mit einbeziehen (APA 2004:12). Als wichtig erweist sich in diesem Zusammenhang für die APA ebenso der Umfang, in dem die adäquate bzw. nicht adäquate Konstruktrepräsentation Vor- und Nachteile bei Testteilnehmern auslösen kann (APA 2004:12):

„(...) construct underrepresentation or construct-irrelevant components may give an unfair advantage or disadvantage to one or more subgroups of examinees".

Nach der KTT gibt die so genannte Konstruktvalidität den Grad der Präzision an, mit der ein Kriterium gemessen wird. Die APA nennt diesen Validitätsbeweis *Evidence based on consequences of testing.* Es werden Hypothesen hinsichtlich des zu messenden Kriteriums aufgestellt, die anhand der Testwerte Bestätigung finden sollen. Wenn sich die Konstruktvalidität nicht bestätigt, d.h. sie nicht gewährleistet ist, dann können Schlüsse gezogen werden, wie: das Konstrukt bzw. das Kriterium ist nicht existent bzw. hat keinerlei empirische Bedeutung. In diesem Fall würde der Test alles Andere als das zu messende Konstrukt messen und wäre demnach konstruktirrelevant. Das Kriterium wäre unterrepräsentiert, d.h. wichtige Dimensionen würden gänzlich fehlen (vgl. Messick 1989). Es ist folglich absolut notwendig, das zu messende Kriterium und den Testinhalt sorgfältig zu überdenken und zu überarbeiteten. Die Inhaltsanalyse muss, auch auf der klassischen Testtheorie basierend, erkennen lassen, welches Merkmal erfasst werden soll.

In vorliegender Arbeit geht es um den schriftlichen Ausdruck auf den Niveaubeschreibungen B2 und C1. Der Inhalt dieses Subtests variiert dabei sowohl innerhalb der Niveaus, als auch unter den Testanbietern, wie man im 5. Kapitel sehen wird. Es wird ein Input vorgegeben, auf den man unter Berücksichtigung gegebener Informationen schriftlich reagieren soll. Folglich wird ein konkreter Ausschnitt der Kompetenz hinsichtlich des Schreibens verlangt. Der Inhalt der Aufgabe bezieht sich auf eine bestimmte Thematik, die man bearbeiten soll. Der geforderte Wortschatz ergibt sich demnach aus dem Thema oder der Aufgabenstellung. Die Kann-Beschreibungen des GER

wollen Aufschluss darüber geben, was man auf welcher Stufe können muss bzw. soll (vgl. Kapitel 2.2.1). Diesbezüglich definiert die APA im Kapitel *Validity* folgenden Standard APA-Standard 1.6:18):

„When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified".

Laut dieses Standards sollte die Beziehung zwischen Aufgabenformat und der schriftlichen Kompetenz, die ermittelt werden soll, verdeutlicht werden. Dieterich (1973:99ff) spricht an dieser Stelle von repräsentativer Validität und drängt zur Forderung, dass ein gegebener Test das zu erfassende Kriterium im gesamten Bedeutungsumfang oder in seiner Repräsentativität wiedergibt. Es geht folglich darum, ob die gestellten Aufgaben auch tatsächlich das abverlangen, was man als Zielsetzung versteht. In Kapitel 5 wird erörtert, inwiefern die gesetzten Bewertungskriterien tatsächlich das zu messende Kriterium bzw. eine erwartete Kompetenz in diesem Sinne abbilden können. Der Standard 3.14 hält es für eine notwendige Voraussetzung die Bewertungskriterien gerade für den schriftlichen Ausdruck explizit zu machen (APA-Standard 3.14:46):

„The criteria used for scoring test takers' performance on extended-response items should be documented. This documentation is especially important for performance assessments, such as scorable portfolios and essays, where the criteria for scoring may not be obvious to the user".

Die APA hält des Weiteren die Beobachtung der Performanzstrategien bzw. der Antwortprozesse von Testteilnehmern für eine gute Beweisquelle hinsichtlich der Kriteriumsdefinition. Während einerseits diese Art des Validitätsbeweises dazu beitragen kann, die Bewertungsunterschiede zwischen Testteilnehmern zu hinterfragen, hängen die Bewertungen andererseits jedoch von den Ratern ab. Die zentrale Frage und Aufgabe des Validitätsbeweises ist in diesem Fall, den Bereich einzugrenzen, in dem die Rater in ihrer Bewertung konsistent sind. Es stellt sich dementsprechend die Frage, inwieweit und wie Rater die zur Verfügung stehenden Bewertungskriterien anwenden. Dabei muss jedoch ebenso sicher gestellt werden, dass die Bewertungskriterien nicht durch andere äußere Faktoren beeinflusst werden. Der folgende Standard definiert sehr deutlich (APA-Standard 1.7:19):

„When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgements or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth".

Die erstrangige und zentrale Frage dieser Arbeit ist, wie gut die Bewertungskriterien definiert sind, so dass Bewertungen als valide betrachtet werden können. Sowohl die Bewertungskriterien als auch die Rater, die diese umsetzen, müssen dem Validitätsbeweis gerecht werden. Durch den Validitätsbeweis der internen Struktur eines Tests soll der APA entsprechend Aufschluss darüber gegeben werden, ob die Wechselbeziehung zwischen Testitems und anderen Testkomponenten mit dem zu messenden Kriterium (hier: schriftlicher Ausdruck) und den bereit gestellten Bewertungskriterien übereinstimmen. Es kann durchaus sein, dass der von der APA definierte konzeptuelle Referenzrahmen verschiedene Komponenten testet, diese aber trotzdem das Kriterium der Homogenität erfüllen. In unserem Fall werden schriftliche Lernerproduktionen anhand von verschiedenen Kriterien, wie z.B. Inhalt oder Ausdrucksfähigkeit fest gemacht. Die Summe dieser und anderer Komponenten impliziert folglich die Kompetenz im schriftlichen Ausdruck. An dieser Stelle könnte man die Unterscheidung zwischen holistischen und analytischen Bewertungsprozessen anführen (APA 2004:38):

„Both of the procedures require explicit performance criteria that reflect the test framework.(...) Under the analytical scoring procedure, each critical dimension of the performance criteria is judged independently, and separate scores are obtained for each of these dimensions in addition to an overall score. Under the holistic scoring procedure, the same performance criteria may implicitly be considered, but only one overall score is provided".

Der analytische Ansatz zeigt Stärken und Schwächen eines Testteilnehmers auf, während der holistische Ansatz auf eine allgemeine Bewertung ausgerichtet ist. Unabhängig von diesen beiden unterschiedlichen Bewertungsansätzen sieht die APA die Item- und Bewertungsentwicklung jedoch als einen integrierten Prozess an (APA 2004:39). Im 5. Kapitel wird zu sehen sein, ob sich der Validitätsbeweis der internen Struktur behaupten kann und welche Bewertungsansätze die einzelnen Testanbieter bevorzugen. Dabei soll der Standard 3.22 des Kapitels *Test Development and Revision* nicht außer Acht gelassen werden (APA-Standard 3.22:47):

„Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally".

Auch in diesem Zusammenhang versucht die APA den Validitätsbeweis zu erbringen und unterscheidet zwischen *convergent* und *discriminant evidence.* Der konvergente Beweis wird insofern geliefert, wenn die Beziehung zwischen Testwerten und verschiedenen Messungen ähnliche Konstrukte festsetzt. Gegensätzliches ist bei dem Diskriminanzverfahren zu erwarten. Als hilfreich können sich gemäß der APA verschiedene Messverfahren entpuppen, wenn es darum geht, Testwerte zu erstellen. Eine weitere und häufig benutzte Methode, um Validität zu ermitteln, zeichnet sich dadurch aus, dass man sich zur Überprüfung dieser eines Außenkriteriums bedient. Die in der KTT definierte Kriteriumsvalidität, von vielen Autoren auch empirische Validität genannt,[81] vergleicht also, ob sich die Messungen des vorausgesetzten latenten

81 Diese Validierungsmethode der KTT wird anhand von Testwerten ermittelt. Daraus folgt oftmals der Begriff „empirische Validität"

Kriteriums mit denen des manifesten Außenkriteriums decken. Eine derartige Validierungsmethode bedingt die Validität sowohl des latenten als auch des manifesten Kriteriums. Die APA gibt bezüglich dieses Validitätsbeweises folgende Definition (APA 2004:13):

„Evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations".

Validität kann des Weiteren auch unter Beweis gestellt werden, wenn man der fundamentalen Frage nachgeht, wie genau Testwerte die Kriteriumsleistung voraussagen. In der KTT bedient man sich in dieser Hinsicht der Begriffe der prognostischen Validität und der Übereinstimmungsvalidität[82]. Die APA sieht diese Differenzierung als eine Methode an, um das Verhältnis zwischen Test und Kriterium zu definieren (APA 2004:14):

„A predictive study indicates how accurately test data can predict criterion scores that are obtained at a later time. A concurrent study obtains predictor and criterion information about the same time".

Während bei der prognostischen Validität das manifeste Kriterium zu einem späteren Zeitpunkt überprüft wird, können latentes und manifestes Kriterium mittels der Übereinstimmungsvalidität gleichzeitig ermittelt werden. Was die Beweisführung der Validität insgesamt angeht, so wäre es erstrebenswert, dass man sie generalisieren könnte. Der Anspruch, Validitätsbeweise universell zu machen, kann meines Erachtens auf die heutige Testerstellung und insbesondere die Testbewertung im DaF-Bereich bezogen, insofern nicht realisiert werden, als dass es Faktoren gibt, die inkonsistent sind.

4.4.2 Objektivität, Reliabilität und Nebengütekriterien

Die Hauptgütekriterien Objektivität und Reliabilität, die in der Testtheorie neben der Validität existieren, sollen im Weiteren synoptisch angeführt werden, da sie für das Ziel dieser Arbeit, die Validität der Bewertungskriterien zu untersuchen, eine eher untergeordnete Rolle spielen. Dennoch besteht zwischen den drei Gütekriterien eine logische Beziehung (Rost 1996:33). Zur Einschätzung der empirischen Validität ist es wichtig zu wissen, dass die Objektivität einen Einfluss auf die Reliabilität hat und dass die Reliabilität wiederum eine Obergrenze für die empirische Validität darstellt. Dies bedeutet, dass ein wenig objektiver und wenig reliabler Test nicht gleichzeitig valide sein kann. Dieser wichtige Sachverhalt wird von Praktikern häufig übersehen. Umgekehrt bedeutet eine hohe Objektivität und Reliabilität keineswegs, dass der entsprechende Test auch valide ist, d.h. das erfasst, was er erfassen soll:

- Die Objektivität und Reliabilität sind notwendige, jedoch nicht hinreichende Voraussetzungen für eine zufrieden stellende Validität
- Ein Test kann kriterienbezogen nicht valider als reliabel sein
- Sowohl die Paralleltest- als auch die Retestreliabilität können nicht höher sein als die innere Konsistenz und die Auswertungs- und Durchführungsobjektivität

82 Die englischen Bezeichnungen dafür sind predictive validity und concurrent validity

- Im Fall einer kriterienbezogenen Validität kann deshalb häufig auf eine Überprüfung der Objektivität und Reliabilität verzichtet werden
- Ein Test mit ausreichender Validität und einer geringen Reliabilität hat ausgezeichnete Verbesserungschancen, da sich die Reliabilität und damit zugleich die kriterienbezogene Validität zumeist testtechnisch leicht erhöhen lässt (z.B. durch Aussondern und Hinzufügen von Aufgaben)
- Ein Test mit geringer Validität und hoher Reliabilität eignet sich zwar zur Differenzierung zwischen Individuen, jedoch nur sehr bedingt zur Vorhersage des jeweiligen Kriteriums (Test und Kriterium messen nur sehr bedingt das Gleiche). Die kriterienbezogene Validität eines solchen Tests kann nur über eine inhaltliche Überarbeitung verbessert werden
- Um eine zufrieden stellende kriterienbezogene Validität zu erreichen, muss nicht nur der Test, sondern auch das Kriterium hinreichend objektiv und reliabel sein

Des Weiteren werden auch einige der mir im Zusammenhang dieser Thematik am wichtigsten erscheinenden Nebengütekriterien vorgestellt werden, um einen globalen Überblick der Testtheorie und was es zu berücksichtigen gilt, zu gewährleisten.

Das Kriterium der Objektivität geht der Frage nach, wie unabhängig das Testresultat von der Testsituation und dem Testbewerter bzw. Rater ist. Anders ausgedrückt sollte der Test unter anderen Umständen jedoch mit denselben Testteilnehmern zum gleichen Resultat führen. Verschiedene Rater würden in diesem Fall bei exakt denselben Personen das gleiche Ergebnis erlangen. Die Unabhängigkeit der Ergebnisse vom Anwender soll durch eine weitgehende Standardisierung von Durchführung, Auswertung und Interpretation erreicht werden (Kranz 2001:4). Testentwickler versuchen genau festzulegen, auf welche Weise und unter welchen Bedingungen die einzelnen Aufgaben gestellt werden, wie die Reaktionen darauf zu bewerten sind und welche Aussagen aufgrund der vorliegenden Resultate über das zu messende Kriterium zu treffen sind. Ein Test wird also konzipiert, um Aussagen über den Testteilnehmer und nicht über den Rater zu machen. Objektivität kann auch als Standardisierung des Testablaufs und seinen Phasen definiert werden. Ingenkamp (1985:34) bemerkt dazu: „Wenn wir bei einem Messergebnis nicht mehr unterscheiden können, wie weit es Merkmale des Gemessenen oder des Messenden kennzeichnet, wenn wir annehmen müssen, dass ein anderer Beobachter zu einem ganz anderen Ergebnis gekommen wäre, dann können wir aus diesem Messergebnis keine Aussagen und Folgerungen ableiten, die von über den Zufall hinausgehender Bedeutung sind."

Im Folgenden sollen die einzelnen Phasen des testdiagnostischen Prozesses aufgezeigt werden, welche unter Einhaltung der Vorschriften das Kriterium der Objektivität erhöhen (Lienert/Raatz 1998:8). Es wird häufig zwischen Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität unterschieden (Ingenkamp 1985:34ff, Lienert/Raatz 1998:8). Die so genannte Durchführungsobjektivität betrifft die Bedingungskonstanz in der Testdurchführung. Diese wird dann gewährleistet, wenn der Testteilnehmer nach den vorgegebenen Testanweisungen relativ autonom die Aufgaben bearbeiten kann. Das standardisierte Testmaterial und die einheitliche Anweisung müssen in diesem Sinne gleichermaßen gut und schnell von den Testteilnehmern erfasst werden (Schelten 1997:125). Eine standardisierte Testdurchführung könnte durch unvorhersehbare Fragen der Testteilnehmer bezüglich der Aufgabenbearbeitung

erschwert werden. Derartige Fragen wären zwar zu beantworten, jedoch so, dass keine über die standardisierten Instruktionen hinausgehenden Hilfestellungen in den Antworten enthalten sind. Die in Fragestellung der Durchführungsobjektivität könnte durch standardisierte Testanweisungen vermieden werden. Die oberste Regel für die Gewährleistung der Durchführungsobjektivität ist, die gegebenen Anweisungen genau einzuhalten.

Um von Auswertungsobjektivität zu sprechen, sollte das Auswertungsergebnis eines Tests unabhängig davon sein, welche Person den Test auswertet. Es muss genau angegeben werden, was und wie zu bewerten ist. Der Grad der Auswertungsobjektivität hängt außerdem von den Itemformen ab. Während bei geschlossenen Aufgabenformaten die Auswertungsobjektivität leicht eingehalten werden kann, rufen offene Aufgabentypen (hier: schriftlicher Ausdruck) einen Mangel an Auswertungsobjektivität hervor. Die Auswertungsobjektivität sollte in derartigen Fällen an vor der Bewertung zu operationalisierenden Kriterien gebunden sein, um sowohl technische Eigenschaften als auch die subjektive Testbewertung vor Fehlerbehaftung zu schützen. Die so genannte Signierobjektivität (Rost 1996:39) bezieht sich in unserem Zusammenhang auf die Objektivität bei der Kodierung schriftlicher Lernerproduktionen. Da der subjektive Ermessensspielraum eines Raters dabei groß ist, müssen bereits im Vorfeld Bewertungskriterien aufgestellt werden, die bestimmte Richtlinien vorgeben (Birkel 1976:43), um dadurch sowohl die Interpretationsobjektivität als auch die Validität zu gewährleisten (Bolton 1982:113). Die Interpretationsobjektivität[83] ist zwar schwieriger zu erreichen, jedoch entscheidend für die Testvalidität. Um Interpretationsobjektivität handelt es sich, wenn verschiedene Rater aufgrund desselben Testwerts zu den gleichen Testresultaten gelangen, ohne den Einfluss individueller Interpretationen. Lienert (1961:14) definiert sie als „vollkommen und zugleich trivial, wenn es sich um normierte Leistungstests handelt". Die völlige Interpretationsobjektivität ist gerade im schriftlichen Ausdruck schwer erzielbar. Erhöhen lässt sich die Interpretationsobjektivität zum Beispiel dadurch, dass zu dem vorgelegten Test Normen angegeben sind. Somit würde die Beurteilung aufgrund des gelegten Maßstabes nicht mehr der Subjektivität des Raters ausgesetzt sein.[84]

Ein weiteres Hauptgütekriterium der Testtheorie ist die Reliabilität. Diese wird als die Messpräzision sowohl stabiler als auch instabiler Merkmale definiert, wobei das passende Messinstrument ermittelt werden muss. Reliabilität ist anders ausgedrückt der Grad der Genauigkeit, mit dem der Test ein bestimmtes Merkmal misst, abgesehen davon, ob der Test auch das misst, was er vorgibt zu messen (Lienert/Raatz 1998:9). Um die Reliabilität eines Tests abzuschätzen, werden korrelationsstatistische Methoden angewendet,[85] da sich die Varianz des wahren Wertes bzw. der tatsächlichen Kompetenz empirisch nicht ermitteln lässt.[86] Dabei soll festgestellt werden, in welchem Ausmaß Testergebnisse miteinander unter den verschiedensten Umständen übereinstimmen können.

83 Anmerkung: Diese Form der Objektivität könnte unseres Erachtens auch als eine Form der Validität betrachtet werden

84 In Kapitel 4.8.1/4.8.2 setzt sich mit dem Rating und dem Raterverhalten auseinander

85 Die statistischen Messverfahren Testwiederholungsreliabilität, Paralleltestreliabilität, Testhalbierungsreliabilität und interne Konsistenz werden im Rahmen vorliegender Dissertation nicht näher ausgeführt werden, da sie den Rahmen sprengen würden

86 Das Gütekriterium Reliabilität basiert auf den Axiomen der Klassischen Testtheorie. Die Genauigkeit einer Messung wird folglich spezifiziert als der Anteil der Abweichung (Varianz) des wahren Wertes an der Gesamtabweichung (Gesamtvarianz) der Messung

Die Reliabilität eines Tests soll so hoch wie möglich sein. Sie wird zudem von der Objektivität in der Form beeinflusst, dass sie nicht höher sein kann als die Objektivität. Reliabilität kann außerdem aufgrund äußerer Faktoren wie Konzentrationsschwierigkeiten oder Ermüdungserscheinungen beeinträchtigt werden. Verschiedene Möglichkeiten zur Reliabilitätsschätzung sollten bereits während der Testentwicklung berücksichtigt werden (vgl. Raatz 2001). Dennoch reicht es nicht aus, nur die Messgenauigkeit eines Tests unter Beweis zu stellen. Dieser muss sich auch als valide erweisen. Da Reliabilität immer auch eine Voraussetzung der Validität ist, kann aus der Reliabilität einer Messung auch ein Maß für deren Validität abgeleitet werden.

Es gibt in der Literatur Uneinigkeit darüber wie die Rangfolge bestimmter Kriterien zu sein hat. Das im Zusammenhang mit dieser Arbeit wichtigste Kriterium ist das der Validität. Ich habe die anderen zwei Kardinalkriterien bereits definiert. Um das testtheoretische Konstrukt abzurunden, sollen die Nebengütekriterien Fairness, Normierung, Ökonomie und Nützlichkeit angeführt werden.

Zwischen dem Kriterium der Fairness und dem Begriff der Validität besteht meines Erachtens ein enger Zusammenhang. Da Tests interpersonell sind, kann das Kriterium der Fairness sehr leicht beeinflusst werden. Ein Beispiel dafür wäre die unerlaubte Interaktion zwischen Testleitern und Testteilnehmern. Die APA behandelt in diesem Zusammenhang in Kapitel 7 *„Fairness in testing and test use"* die Thematik, dass die faire Behandlung aller Testteilnehmer nicht nur die faire Betrachtung des Tests als Ganzes, sondern auch des Zwecks und seiner Bewertung impliziert. Ein bereits angeführtes Beispiel waren die Sprachtests ohne Fachbezug, wobei der Einfluss der Vorkenntnisse eine geringe Rolle spielt und der Testfairness gerecht werden. Testteilnehmer sollten demnach vergleichbare oder gleiche Möglichkeiten haben, um ihre Kompetenz testen zu lassen. Dass Testwerte verschiedener oder gleicher Testteilnehmer in verschiedenen Tests vergleichbar gemacht werden sollen, beschäftigt das etwas pragmatischere Kriterium der Normierung (Rost 1996:41). Die so genannte Standardisierung[87] erfolgt mittels statistischer Verfahren, in denen der individuelle Testwert in Relation zu den Leistungen der Zielpopulation beurteilt werden kann (Grotjahn 2000:317). Von Bedeutung ist Standardisierung besonders bei formellen Tests, wie die in Kapitel 5 untersuchten des Goethe-Instituts und des TestDaF-Instituts.

Das Nebengütekriterium Ökonomie ist wichtig in Bezug auf die leichtere, unkompliziertere Erstellung und Handhabung eines Tests (Lienert/Raatz 1998:12). Ein ökonomischer Test sollte zum Beispiel mit so wenig wie möglichen Items eine Merkmalsausprägung erschließen lassen. Eine Rolle hierbei spielen Testlänge und dementsprechende Bearbeitungszeit, benötigtes Material, Art der Testdurchführung und die zeitsparende Bewertung.

Nach dem Nützlichkeitsprinzip „usefulness" von Bachman/Palmer (1996:17ff.) ist Ökonomie die Relation zwischen zur Verfügung stehenden Ressourcen (z.B. ein Rater) und den benötigten Ressourcen (Grotjahn 2001:107ff.).[88] Unter dem Begriff der Praktikabilität verstehen sie einen vernünftigen Zusammenhang zwischen Aufwand und Ergebnissen einer Prüfung. In diesem Zusammenhang muss die Prüfung selbst zum einen eine vernünftige Länge haben und zum anderen sollte der Aufwand, der für die Prüfung

87 Standardisierung definiere ich nicht als einen statistischen Begriff, sondern als die adäquate Durchführung, Auswertung und Interpretation eines Tests, wobei die testtheoretischen Gütekriterien ihre Geltung haben müssen

88 Nützlichkeit kann nach Bachman/Palmer durch Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Rückwirkung und Praktikabilität definiert werden.

und zugleich für die Bewertung zu betreiben ist, nicht unzumutbar hoch sein. Bei der Frage, wann ein Test als nützlich gilt, gibt es auseinander gehende Meinungen bezüglich der Definition. Es muss zunächst ein Bedarf an einen Test existieren. Ist diese Voraussetzung erfüllt, dann kann die Nützlichkeit entsprechend hoch oder niedrig sein. Bachman/Palmer (1996:17) betonen in ihrem Modell, dass die wichtigste Überlegung bei der Testentwicklung der beabsichtigte Nutzen ist und daraus resultierend das Kriterium der Nützlichkeit als das wichtigste Testgütekriterium zu betrachten ist. Von hoher Nützlichkeit ist bei Lienert/Raatz (1998:13) dann die Rede, wenn es keinen äquivalenten Test gibt, der das zu messende Kriterium messen kann. Genau an dieser Stelle ist aber der Einwand und die berechtigte Frage einzubringen, ob denn die verschiedenen von Testanbietern formulierten Tests im Rahmen von Sprachstandsprüfungen für das Niveau B2 und C1 als äquivalente Tests betrachtet werden können. Der GER soll die Basis bereit stellen, dass Niveaus und Kompetenzen vergleichbar gemacht werden sollen. Werden aber Eigenschaften oder Merkmale von zahlreichen Tests evaluiert, dann hat der sich in der Entwicklung befindliche Test keine Nützlichkeit. Bachmann/Palmer (1996:18) definieren in ihrem Modell das Gütekriterium der Nützlichkeit aus der Summe weiterer sechs Gütekriterien bestehend:

Nützlichkeit =
Reliabilität+Konstruktvalidität+Authentizität+Interaktivität+Effekt+Praktikabilität

Diese kollektive Definition ist in dem Sinne zu verstehen, dass bei ihrer Einhaltung folgende Prinzipien gelten würden. Hier gilt es die Gesamtnützlichkeit eines Tests und nicht die Komponenten der Nützlichkeit wie z.B. die Reliabilität zu maximieren. Es ist des Weiteren zu betonen, dass das Kriterium der Nützlichkeit und das angemessene Gleichgewicht unter den ihr zustehenden Kriterien nur in bestimmten Testsituationen definiert werden können. Authentizität definiert in diesem Modell die Charakteristika eines Items als genuine, authentische Situationsaufgaben (Grotjahn 2000:318). Authentizität der Items kann zum Beispiel durch aktuelle reale Berichte gewährleistet werden. Dieses Gütekriterium könnte durch kommunikativ relevante Situationen im Rahmen eines handlungsorientierten Ansatzes ebenso Aufschluss über Lernerkompetenz und pragmatisch angemessenem sprachlichen Verhalten außerhalb der Testsituation geben (Apeltauer 1987:129). Die Interaktivität, die nach Bachman/Palmer ebenso Teil des Oberbegriffs Nützlichkeit ist, kann als das Ausmaß und die Art der Wechselwirkung zwischen den Testaufgaben und den im Hinblick auf das zu messende Konstrukt relevanten kognitiven Merkmalen der Kandidaten betrachtet werden, wobei als Merkmale die sprachliche Kompetenz, thematisches Wissen und affektive Schemata zu definieren sind (Bachman/Palmer 1996:25ff.). Die Wirkung eines Test auf die Makro- und die Mikroebene soll das Unterkriterium Effekt beschreiben. Makroebene schließt die Gesellschaft und das Erziehungssystem ein, während die Mikroebene durch Individuen definiert wird. Dieses Kriterium gilt als erfüllt, sobald aus Testergebnissen Konsequenzen bezüglich Prüfungsverlauf und Leistungsevaluation gezogen werden (Grotjahn 2001:108). Praktikabilität eines Tests heißt, er soll durchführbar sein. Nach Bachman/Palmer (2006) ist der ökonomische Faktor eine Eigenschaft dieses Unternebenkriteriums. Das Nebengütekriterium Nützlichkeit, das von Bachman/Palmer (2006) aus 6 Funktionen besteht, spielt bei der Testbewertung eine entscheidende Rolle. Authentizität und Interaktivität sind wichtige Faktoren für die Testpersonen. Das spiegelt

sich im Unterricht und im weiteren Sinne in der Gesellschaft wider, was schließlich eine Gewährleistung der Testvalidität bedeutet (Wiedenmeyer 2006:30).

Fazit

Haupt- bzw. Nebengütekriterien tragen alle zur Testentwicklung, Testdurchführung und Testauswertung bei. Im folgenden 5. Kapitel soll ausgeführt werden, inwieweit die Bewertungskriterien beim schriftlichen Ausdruck und deren Umsetzung diesem Prinzip unterliegen können. Notwendigerweise wird der Frage nachgegangen werden müssen, ob entwickelte und existierende Tests tatsächlich durch die zu messenden Fähigkeiten oder Merkmale beurteilt werden oder ob den gesellschaftlichen Anforderungen, in unserem Fall der Sprachzertifizierungsprüfungen gemäß des Modells des GER, entsprochen werden muss. Aus diesem Grund erscheint es zunächst sinnvoll, nach der Art des zu erfassenden Merkmals zu unterscheiden, d.h. die Tests zu klassifizieren. Man muss an dieser Stelle die Frage nach dem Hintergrund stellen, auf dem Tests erstellt worden sind, ob sie die nötigen testtheoretische Bedingungen erfüllen und was letztlich ihre Prüfungsintention ist.

4.5 Rater und Ratingverfahren

Während im Folgenden verschiedene Thesen bezüglich der Rater und der Ratingverfahren aufgestellt werden, soll versucht werden zu verdeutlichen, welche Faktoren die Subjektivität der Rater minimieren, sodass Raterurteil und Testergebnis gekoppelt bleiben könnten. Die Reflexion der Bewertungskriterien ist insofern wichtig, als man von einer Gewährleistung der Gütekriterien sprechen kann, um den „tatsächlichen“ Sprachstand eines Prüfungskandidaten innerhalb einer standardisierten Prüfung eruieren zu können.

4.5.1 Ratingverfahren

Nachdem ein Test inhaltlich mit den analogen Items ausgestattet worden ist und den Testgütekriterien und den jeweiligen Standards zur Testdurchführung zu genügen scheint, muss nun das entsprechende Werkzeug für die Testbewertung bearbeiteter Items im Sinne der kommunikativen Kompetenz beispielhaft zur Verfügung gestellt werden. Ratingverfahren bieten die Möglichkeit, aus Beobachtungen einen Messwert zu ermitteln. Einem Ratingverfahren unterliegen zunächst verschiedene Elemente, damit es im Sinne der Gütekriterien existieren kann. Als erstes steht das zu messende Kriterium im Mittelpunkt. Allein dieses Merkmal soll entscheidend für das richtige Ratingverfahren bzw. für seine Reliabilität sein. Wichtiger entpuppt sich aber die Genauigkeit der Entscheidungen, die getroffen werden, denn es sind nicht nur die Messungen die valide sein müssen, sondern genauso die daraus abgeleiteten Schlussfolgerungen. Anders ausgedrückt: Wie genau ist eine Entscheidung x auf einer festgesetzten Skala y? Das hängt zum einen von der Validität des betreffenden Kriteriums im betreffenden Situationskontext und von der Validität der gesetzten Bewertungskriterien und zum anderen von der Reliabilität und Auswertungsobjektivität der Rater ab (GER 2001:172). Eckes (2004) thematisiert in diesem Zusammenhang in seinem Aufsatz, dass das

Problem bei der Messung eines Kriteriums anhand von Ratingskalen oftmals durch die unzureichende Interraterreliabilität gekennzeichnet ist. Die APA definiert in dieser Sache (APA-Standard 3.22:2004):

„Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions of using rating scales or of deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally."

Prozeduren müssen also für die Bewertung und deren Kriterien durch den Testentwickler detailliert und präzise präsentiert werden, um die Genauigkeit des Ratings zu maximieren und ebenso müssen Instruktionen zur Benutzung der Ratingskalen klar gemacht werden. In dem Standard 3.23 geht es darum, dass „der Prozess der Auswahl, des Trainings und der Qualifizierung der Rater durch den Testentwickler dokumentiert werden sollten und dass das Trainingsmaterial (...) und der Prozess des Ratertrainings aus einem Ausmaß der Zustimmung zwischen den Ratern resultieren sollte, sodass bewertet werden kann, wie der Testentwickler es ursprünglich vorgesehen hat. Die Bewertungsreliabilität und das Motivationspotential sollten evaluiert und durch verantwortliche Leitungspersonen der Trainingseinheit dokumentiert werden". Die Originalversion lautet (APA-Standard 3.23:2004):

„The process for selecting, training, and qualifying scores should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on a score scale, and the procedures for training scorers should result in degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session".

Die APA besteht demnach darauf, dass Testentwickler sowohl Prozesse, Materialien, konkrete Richtlinien und Ratertrainingsmaßnahmen bereitstellen um Tests zu bewerten, als auch auf deren Kontrolle bezüglich Korrektheit des ganzen Ratingprozesses.

In wechselseitiger Beziehung stehen das zu messende Kriterium, die dafür formulierte Skala und schließlich die Rater, die man als Realisatoren des Ganzen betrachten könnte. Damit ein Rater eine zu messende schriftliche Lernerproduktion bewerten kann, müssen ihm Instruktionen gegeben werden. Diese Instruktionen müssen operationalisiert werden. Anders ausgedrückt, müssen so genannte Deskriptoren einerseits angeben, was und andererseits wie das schriftliche Konstrukt zu messen ist (Langer/Schulz 1974:17). Die Frage im Zusammenhang vorliegender Dissertationsthematik lautet demnach: Wie operationalisiert man geschriebene Sprache auf den Niveaus B2/C1? Wie ist diese Kompetenz zu definieren, um die Objektivität zu sichern und sich den „wahren Werten" von Sprachlernern zu nähern? Die produktive Schreibkompetenz könnte sicher eine ganze Liste von Merkmalen mit sich führen, die sie charakterisieren soll. Welche Aspekte sind aber die wichtigsten, um beispielsweise Urteilsfehler auszuschließen, die die Interraterreliabilität senken würden? Eckes (2004:489) stellt in diesem Kontext die Fragestellung auf, inwieweit „ein detailliert ausgearbeiteter Kriterienkatalog nebst intensiven Schulungen von Beurteilern im konsistenten Gebrauch der Kriterien bei der Beurteilung von schriftlichen Sprachleistungen zu einer zufrieden stellenden Urteilsgenauigkeit verhelfen würde".

Wie im 5. Kapitel zu sehen sein wird, müssen die Kriterienkataloge samt ihren Deskriptoren für das offene Aufgabenformat des schriftlichen Ausdrucks sehr eng gestrickt sein, um die Subjektivität der Rater auf das Mindeste zu reduzieren. Theoretisch betrachtet würde das keine Probleme bereiten, die Deskriptoren so weit zu fächern bzw. zu „objektivieren", dass der Willkür der Rater kein Raum gelassen würde. Praktisch würde das aber zweierlei Nachteile mit sich bringen: Einerseits bestünde die Gefahr, dass man den Bezug zum theoretischen Ausgangsbegriff bzw. Kriterium aus den Augen verliert und gerade dadurch eine Leistung oder ein zu messendes Kriterium dem Rater „verfällt". Andererseits vermerkt auch der GER, dass mehr als 4-5 Kategorien zur kognitiven Überlastung menschlicher Rater führen können und dass psychologisch betrachtet 7 Kategorien die Obergrenze bilden (GER 2001:187). Zudem kommen noch die Urteilsfehler hinzu, die als Mangel an Interraterreliabilität definiert werden könnten (vgl. Hoyt 2000). Damit das Ergebnis einer Messung (hier: die des schriftlichen Ausdrucks) einerseits nicht nur von der Beschaffenheit und der subjektiven Interpretation der benutzten Skala und andererseits von der subjektiv geprägten Raterwahrnehmung des zu messenden Kriteriums abhängt, muss eine objektive Umgebung geschaffen werden, die keinerlei Ausweichmöglichkeiten zulässt und die Objektivität der Bewertung garantiert (Eckes 2004:488f.).

Es gibt verschiedene Skalentypen, die je nach Perspektive für die Beurteilung herangezogen werden. Ich habe unter Kapitel 2.1.2 bereits die Skalendifferenzierung von Alderson (1991) vorgestellt. Die hier wichtigste und von ihm definierte „beurteilerorientierte" Skala sollte für Konsistenz im Ratingprozess stehen (vgl. North 1993). Eines der Ziele dieser Arbeit ist es herauszufinden, inwieweit dies realisierbar ist. Für die Bewertungsskalen des schriftlichen Ausdrucks unterscheiden wir aber zunächst zwischen holistischer und analytischer Beurteilungsskala, ohne auf ihre statistische Gegebenheit und Grundlage einzugehen. Einerseits geht man der Frage nach „was beurteilt werden soll". Dabei zielt die holistische Beurteilung darauf ab, ein globales Urteil zu erhalten. Es wird auf die intuitive Kompetenz der Rater vertraut, um verschiedene Aspekte zu gewichten und folglich eine Gesamtsituation wahrzunehmen. Der Beurteiler vergibt die Punktwerte danach, wie die Gesamtwirkung der schriftlichen Arbeit bewertet wird. Bei diesem Bewertungsverfahren werden die Punktwerte nicht auf verschiedene Kriterien für den schriftlichen Ausdruck verteilt, aber natürlich werden auch hier bestimmte Kriterien berücksichtigt. Von Bedeutung ist hierbei, wie genau seine Bewertung ausfällt, also „bestanden vs. nicht bestanden", „Zuordnung der Niveaustufe des GER", „Punktwert" usw.. Es ist demnach dem Rater überlassen, welche Interpretation er dem zu messenden Kriterium zuweist und schließlich seinem Urteil zugrunde legt (Langer/Schulz 1974:21). Der analytische Ansatz hingegen hat den Anspruch der Bewertung verschiedener Aspekte des zu messenden Kriteriums. Eine einzelne Aufgabe, z. B. einen Leserbrief zu verfassen, wird also anhand mehrerer Kategorien beurteilt. Es wird zunächst das zu messende Kriterium definiert und dann versucht anhand von Deskriptoren die zugehörigen Facetten in ihren Abstufungen zu beschreiben (GER 2001:185). Dabei scheint ein wesentlicher Anhaltspunkt die hierarchische Anordnung bzw. Gewichtung der einzelnen Facetten und ihren Ausprägungsgraden zu sein – vom Grundlegendsten zum Spezielleren (Langer/Schulz 1974:160). An dieser Stelle sei aber nochmals darauf hingewiesen, dass menschliche Rater nur mit einer begrenzten Anzahl von Deskriptoren umgehen können (Langer/Schulz 1974:29).

4.5.2 Deskriptorenuniversum

Deskriptoren sollten so objektiv wie möglich, zunächst global, dann fein und nach Komplexität aufgestellt werden, um eine Merkmalsausprägung zu definieren (North 1993:32). Dabei sollten die zur Beschreibung der Niveauzuweisung der einzelnen Kriterien benutzten Begriffe klar, unmissverständlich und präzise sein. Alderson (1991:82) fragt zurecht kritisch: „ Is ‚some' more than ‚a few' but fewer than ‚several' or ‚considerable' or ‚many' – and how many is ‚many'?"[89]. Deskriptoren, die vage Definitionen beinhalten, können, wie wir noch im 5. Kapitel bei den Bewertungskatalogen der einzelnen Testanbieter bemerken werden, verschiedene Interpretationen bei Ratern hervorrufen (vgl. Trim 1978). Sie sollten daher weder zu allgemein noch zu speziell in ihrer Formulierung sein (Langer/Schulz 1974:52). Weir (2005:2ff) hält die Definition der auf dem GER beruhenden Deskriptoren für nicht konsistent und nicht transparent, um Tests auf diesem Hintergrund zu entwickeln. Deskriptoren sollten konkrete Aufgaben beschreiben. Es wurde bereits erwähnt, dass die Deskriptoren, die die Schreibfertigkeit beschreiben, vom GER nicht empirisch kalibriert, sondern lediglich „durch eine Kombination von Elementen aus anderen Skalen erstellt wurden" (GER 2001:67).

Für Clark (1985:348) sind Deskriptoren nichts Anderes als die Beschreibung erwarteter Werte des zu messenden Kriteriums, die auf einem hypothetischen Konstrukt eines Kontinuums platziert sind und keine Garantie dafür sein können, dass ein angesetztes Kriterium auf der gewählten Skala anhand dieser Deskriptoren akkurat und valide ermittelt werden kann. Deskriptoren dürfen entscheidende Informationen, die unabdinglich für die Bewertung sind, nicht auslassen. Die Gestaltung der Deskriptoren soll die Ratersubjektivität soweit es geht minimieren (vgl. Alderson 1991). Des Weiteren dürfen sich keine Überschneidungen zwischen den Beschreibungen der einzelnen Unterkriterien ergeben, so dass nicht „normgerechte" Produktionen falsch bzw. doppelt zugeordnet würden (die so genannte Doppelsanktionierung) (Apeltauer 1987:186). Demnach muss jedes einzelne Unterkriterium genauestens definiert sein, so dass nicht normgerechte schriftliche Äußerungen von allen Ratern ausschließlich der entsprechenden Kategorie zugeordnet werden. North (1996) stellt fest, dass die Formulierung von Deskriptoren normorientiert ist, denn es wird immer Bezug auf andere Deskriptoren bzw. Stufen genommen. Demnach sind Deskriptoren keineswegs als selbständig zu betrachten.[90] Im zentralen Kapitel sollen die Bewertungskriterien der einzelnen Niveaustufen unterschiedlicher Testanbieter analysiert werden, indem die Definitionen der einzelnen Kategorien und folglich ihrer Deskriptoren anhand ihrer Eindeutigkeit und folglich ihrer Validität untersucht und kritisiert werden sollen.

89 Alderson, J.C. (1991a): Bands and Scores. In: Alderson, J.C:/North, B. (eds.): Language Testing in the 1990s. Modern English Publications/British Council. London. Macmillan. S. 71-86

90 North, B. (1996): Language Proficiency Descriptors. Presentation at the Language Testing Research Colloquium in Tampere, Finland in 1996.

4.5.3 Der menschliche Rater

Rater gelten als das Exekutivorgan von Ratingverfahren bzw. von definierten Bewertungskriterien. Setzen wir diese als valide voraus, wie das die einzelnen Testanbieter wohl bei ihren standardisierten Prüfungen tun, dann ist die Frage berechtigt, welche Voraussetzungen Rater auf dieser Basis erfüllen müssen, um die Validität beizubehalten. Man sollte sich sehr gründlich mit der Frage auseinandersetzen, wie die entsprechende Selektion von Ratern vorgenommen werden sollte und ob nur dieses Faktum den Zweck der internen Validität erfüllt. Deshalb erscheint es zunächst sinnvoll, eine Art Idealprofil eines menschlichen Raters zu erstellen. Man kann nur Vermutungen darüber anstellen, welchen Hintergrund die von den Testanbietern ausgewählten Rater haben. Möglicherweise handelt es sich um Lehrkräfte aus dem Fremdsprachenbereich. Ist dem so, kann dies sowohl positive als auch negative Auswirkungen auf die Bewertung einer sprachlichen Leistung haben. Versetzt man sich in die Lage einer Lehrperson, kann man davon ausgehen, dass gerade der grammatische, syntaktische oder auch der morphologische Bereich zum Augenmerk wird. Auf der anderen Seite hat man als Lehrkraft durch jahrelange Erfahrung das Privileg, einen besseren Einblick in textlinguistische Facetten, Fehler und Verständlichkeit eines geschriebenen Lernertextes zu haben. Nichtsdestotrotz sollte als erstes konkret definiert werden, welchen Anspruch man an die Rater hat. Auf die Thematik dieser Arbeit bezogen bedeutet dies, in welche Rolle Rater hinsichtlich der Sprache und ihrer Bewertung schlüpfen müssten. Rater sollten über mindestens zweierlei Kompetenzen verfügen. Zum einen sollten sie die eigene L1 aus der Sicht der L2 sehen können (Fremdsprachenkompetenz) und zum anderen sollten sie Test- und Bewertungskompetenz aufweisen.

Hinsichtlich der Wahrnehmung und Denkweise der Rater ist von Vaughan (1991:116) eine empirische Forschung betrieben worden, die zum Resultat kam, dass Rater keine „tabula rasa" seien. Daraus folgt, dass man eine gänzliche Minimierung der Subjektivität nicht erreichen kann, denn Rater verfügen ebenso über Weltwissen, Hintergrundinformationen, Erwartungen, Werte, Sensibilität und weiteren Faktoren (vgl. Wolfe/Feltovich 1994). In diesem Sinne dürften Rater nichts Anderes können, als das angesetzte Niveau der zu prüfenden Sprache - schlicht: sie müssten doch eine tabula rasa sein! Es sei ein Beispiel angeführt, um dieses Paradoxon zu verdeutlichen: Angenommen ein Rater, der für die Bewertung des schriftlichen Ausdrucks in Athen lokal eingesetzt wird, beherrscht die Muttersprache der Prüfungskandidaten, demnach Griechisch. Wäre sein Verständnis gegenüber fehlerhaften Äußerungen (z.B. Interferenzen) nicht bereits dadurch beeinflusst, dass er Griechisch kann? Dieser Umstand könnte ihn zu einer milderen und damit subjektiveren Einschätzung der schriftlichen Lernerproduktion verleiten. Der Grund könnte darin bestehen, dass er keine Beeinträchtigung beim Rezipieren erführe und dadurch Fehler latent blieben. Er würde also auf der Grundlage der Kontrastivhypothese unbewusst den Einfluss der Muttersprache rezipieren, der aber sein Verständnis und den Lesefluss nicht stören würde. Welche Vorkehrungen könnten in diesem Sinne getroffen werden, sodass „Eindrücke" von den Ratern objektiv aufgefasst werden? Nach Wolfe & Feltovich (1994) haben Eindrücke eines Raters weder mit dem *realen* Text noch mit den Eindrücken anderer Rater etwas gemein.[91] Sicherlich versucht man das Problem der Subjektivität von Ratern und folglich ihren Urteilen anhand verschiedener Methoden wie zum Beispiel

91 Wolfe, E.W./Feltovich, B. (1994): Learning to rate essays: a study of scorer cognition. Report presented at the annual meeting of the American Educational Research Association in New Orleans, LA, 4.-8. April 1994

Prüferschulungen oder Workshops in den Griff zu bekommen. Sowohl das Goethe-Institut[92] als auch das TestDaF-Institut[93] scheinen der professionellen Schulung von Ratern sowohl mit Hilfe der Materialbereitstellung als auch des Fortbildungsangebots einen großen Platz einzuräumen, um sowohl die Intra- als auch die Interraterreliablität, d.h. die Konsistenz zwischen verschiedenen Ratern und bei einzelnen Rater selber, zu maximieren (vgl. Lumley/McNamara 1993). Die Raterinkonsistenz könnte mit normorientierten Tests verglichen werden, da dort die Leistung einer Person im Hinblick auf das dortige Kontinuum definiert wird und in Wahrheit keine Aussage über den tatsächlichen *wahren Wert* machen kann. Lunz/Stahl (1990) stellen in diesem Zusammenhang fest, dass Rater ihre eigenen Standards haben, die sie schwer ablegen können. Was Ratertraining und ihre Effektivität bezüglich der Operationalisierung von Raterverhalten oder Raterstrenge betrifft, beschreibt Eckes (2008:155):

„Research on rater effects in language performance assessments has provided ample evidence for a considerable degree of variability among raters. Building on this research, I advance the hypothesis that experienced raters fall into types or classes that are clearly distinguishable from one another with respect to the importance they attach to scoring criteria. To examine the rater type hypothesis, I asked 64 raters actively involved in scoring examinee writing performance on a large-scale assessment instrument to indicate on a four-point scale how much importance they would attach to each of nine routinely used criteria. The criteria covered various performance aspects, such as fluency, completeness and grammatical correctness. In a preliminary step, many- facet Rasch analysis revealed that raters differed significantly in their views on the importance of the various criteria. A two-model clustering technique yielded a joint classification of raters and criteria, with six rater types emerging from the analysis. Each of these types was characterized by a distinct scoring profile, indicating that raters were far from dividing their attention evenly among the set of criteria. Moreover, rater background variables were shown to partially account for the scoring profile differences. The findings have implications for assessing the quality of large-scale rater-mediated language testing, rater monitoring, and rater training".

In Raterschulungen sollten also Klarheit hinsichtlich der Kriterien und ihrer Anwendung im Rahmen der definierten Skala geschaffen werden (vgl. Lunz/Stahl 1990). Es geht demnach um die Stabilisierung der Intraraterreliabilität statt um die Beseitigung von verschiedenen Strengeprofilen zwischen Ratern (vgl. Eckes 2008). Insgesamt sollten die angeführten Probleme berücksichtigt werden, um die Frage „Messen Skala und Rater wirklich das, was sie messen sollen?" beantworten zu können. Es handelt sich hier um ein zweidimensionales Problem. Auf der einen Seite existiert die Skala, die dem Rater als Instrument dienen soll, und auf der anderen Seite geht es um den Rater, den man als das letzte Glied in der Kette betrachten könnte, das für die Gewährleistung der Testvalidität verantwortlich ist. Ein Rater kann aber in seiner Bewertung eines Textes nicht besser sein, als die zugrunde liegenden Bewertungskriterien (Skala). In diesem Fall würde er mit einem Maßstab bewerten, der nicht vorgegeben ist. In diesem Sinne sagt die Reliabilität der Ratingskala nämlich nichts über ihre Validität aus. Die Qualität der Bewertungskriterien ist somit völlig irrelevant, denn Rater werden individuell dadurch gekennzeichnet, dass sie einen bestimmten Bewertungsstil aufweisen (vgl. Huot 1993).

92 Information aus einem Gespräch am 14.12.2006 mit Frau Dr. Michaela Perlmann-Balme, Testentwicklung Goethe-Institut München

93 www.testdaf.de Zugriff am 01.07.2007, unter der Rubrik „Seminare und Workshops"

Der Bewertungsstil kann allgemein betrachtet milder oder strenger bzw. objektiver oder subjektiver sein. Zudem können verschiedene Ratingfehler registriert werden, die die Inter- und Intraraterreliabilität minimieren. North (1993) führt in seinem Aufsatz „The development of descriptors on scales of language proficiency" drei klassische Ratingfehler an: den Halo-Effekt, die Zentraltendenz und die Strengevariation. Dabei wird der Halo-Effekt entweder als der Transfer von Urteilen einer holistschen Bewertung auf spezielle Kategorien oder zwischen zwei Kategorien oder als der Einfluss einer Textproduktion A auf Textproduktion B usw. verstanden. Arras/Grotjahn (2002:69) sprechen in diesem Zusammenhang von „seriellem Effekt". Unter Zentraltendenz versteht man das Raterverhalten dahingehend, dass mehrheitlich Werte des mittleren Bereichs für Leistungen vergeben werden. Das könnte in der unbewussten Annahme der Rater begründet liegen, Bewertungsfehler würden durch die Streuung im mittleren Bereich vermieden werden. Strengevariation zeichnet sich durch die konsistente Tendenz eines Raters aus, höher oder niedriger im Vergleich zu anderen Ratern zu bewerten (Wilson/Case 1997:4ff). In diesem Zusammenhang verwendet das TestDaF-Institut beispielsweise das computerbasierte Programm FACETS, das auf einem statistischem Verfahren beruht, das so genannte Multifacetten-Raschmodell von Linacre (1989), um unter anderem den Strengekoeffizienten der Rater zu ermitteln. Daneben wird mit diesem Programm versucht, weitere Aspekte, wie zum Beispiel die Fähigkeit einer Person, die Schwierigkeit des Items bzw. des Kriteriums oder die Eigenschaften des Raters, zu ermitteln (vgl. Lumley/McNamara 1993). Was die Fähigkeit der Person betrifft, so sollten leistungsstärkere Personen höhere Bewertungen und leistungsschwächere niedrigere Bewertungen erhalten. Schließlich werden die Eigenschaften des Raters hinsichtlich seiner „strengen" bzw. „milden" Bewertung qualitativ ermittelt, denn dieser Faktor entpuppt sich als einflussreich (Eckes 2004:492f). Es sei ein Beispiel angeführt: Nimmt man den Fall, dass zwei Rater gleichen „Strengeprofils" schriftliche Lernerproduktionen korrigieren und folglich bewerten. Sind beide als „mild" in ihrer Bewertung einzustufen, dann würden die Textproduktionen bessere Resultate und Bewertungen erzielen, als wenn das Raterduett ein strenges Profil hätte. Eckes (2004:488f) thematisiert in seiner Arbeit „Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen" sowohl die Inter- als auch die Intraraterreliabilität, die er anhand von Fallbeispielen belegt und stellt diesbezüglich die Frage nach den „Leistungen" von Prüfungskandidaten, bedingt durch die jeweiligen Ratingverfahren. Dabei unterscheidet er zwischen traditionellen und dem auf dem Multifacetten-Raschmodell basierenden Ratingverfahren. Unter traditionell versteht Eckes die klassische „Drittkorrektur" und das „arithmetische Mitteilungsverfahren". Diese werden als traditionell betrachtet, da die „ermittelten" Daten von Ratern unmittelbar eingeschätzt werden. Im Gegensatz dazu bietet die Alternative der Multifacetten-Korrektur unter Berücksichtigung verschiedener Facetten Aussagen darüber, ob es sich um „faire" Bewertungen handelt oder nicht.

Weiterhin kann der Frage nachgegangen werden, inwieweit denn eine exakt und objektiv erarbeitete Bewertungsskala die Urteilsgenauigkeit von Ratern gewährleisten könnte (vgl. Eckes 2004/Eckes 2008). Es gibt noch weitere Faktoren, die das nicht adäquate Bewerten unterstützen. Müdigkeit oder Erschöpfung können sich ebenso subjektiv auf das nicht adäquate Bewerten der erzielten Leistung eines Prüfungskandidaten auswirken wie die falsch verstandene Absicht der gesetzten Bewertungskriterien oder die Handschrift eines Prüflings (vgl. Wilson/Case 1997). Standardisierte Prüfungen und deren Resultate sagen laut Kritiker von

Leistungsstandards wenig darüber aus, ob die erzielten Werte unabdinglich einem in der Realsituation gleichen Wert zugewiesen würden (vgl. Perlamann-Balme 2006). Hauptziel sind jedoch in jedem Fall standardisierte Sprachtests, auf den testtheoretischen Gütekriterien basierend, zu gestalten. Testanbieter für Sprachstandstests, wie das Goethe-Institut oder das TestDaF-Institut, versuchen durch Qualitätssicherung den mittlerweile länderübergreifenden Ansprüchen des Europarats gerecht zu werden. Pragmatisch betrachtet muss man jedoch an dieser Stelle die Frage aufstellen, was die durch Bewertungsverfahren erzielten Werte bedeuten. Um valide Aussagen über die Sprachkompetenz eines Testteilnehmers machen zu können, muss der Einfluss leistungsirrelevanter und konstruktirrelevanter Faktoren so weit wie möglich minimiert werden.

5 Umsetzung definierter Bewertungskriterien[94]

Nachdem die theoretischen Grundlagen in den vorangegangenen Kapiteln beschrieben worden sind, werden in diesem Kapitel die zentralen Fragestellungen dieser Dissertation diskutiert und analysiert. Die für den schriftlichen Ausdruck erstellten Prüfungsformate einzelner Testanbieter und Lernerreaktionen für die Niveaustufen B2 und C1 werden zunächst separat aufgezeigt. Dabei soll verdeutlicht werden, worauf diese genau abzielen und ob Prüfungskandidaten den Aufgabenstellungen entsprechend reagieren. Weiterhin werden die jeweiligen und der Niveaustufe entsprechenden Bewertungskriterien angeführt, beschrieben und dokumentiert. Abschließend werden die Lösungsvorschläge schriftlichen Ausdrucks von Prüfungsteilnehmern zunächst der Aufgabenstellung gegenüber gestellt und darauf bezogen betrachtet, um schließlich die originalen Bewertungen in Beziehung zu den Bewertungskriterien und schließlich zu den Lernerproduktionen zu setzen. Hilfestellungen sollen die Standards der APA sein, um die eventuell aufkommenden Mängel und unkonkreten Definitionen der Bewertungskriterien zu begründen. Die kritische Betrachtungsweise soll in Verbindung mit alternativen Lösungsvorschlägen dazu beitragen, die Kriterienkataloge und die Vorgehensweisen bei Bewertungen schriftlicher Lernerproduktionen neu zu überdenken.

5.1 Das B2-Zertifikat des Goethe-Instituts

Das B2-Zertifikat des Goethe-Instituts ist eine neu erstellte Prüfung, die weltweit erstmals im Herbst 2007 zur Anwendung gekommen ist.[95] Inhalt dieser Prüfung sind die klassischen Kompetenzen Leseverstehen, Hörverstehen, schriftlicher und mündlicher Ausdruck. In einer 190-minütigen schriftlichen Gruppenführung sind die Prüfungsteile Leseverstehen, Hörverstehen und schriftlicher Ausdruck von den Testteilnehmern zu bearbeiten. Der mündliche Ausdruck erfolgt in einer 15-minütigen Paarprüfung bzw. einer 10-minütigen Einzelprüfung.[96] Insgesamt sind maximal 100 Punkte zu erzielen, wobei die Bestehensgrenze bei 60% liegt. Im Falle des Bestehens dieser Prüfung haben die Prüfungsteilnehmer, laut des Goethe-Instituts nachgewiesen, „dass sie die überregionale deutsche Standardsprache für ihre persönlichen Belange im privaten, gesellschaftlichen, akademischen und beruflichen Leben einsetzen können“. Die Kann-Beschreibungen für die Stufe B2 des GER, welche die selbständige Sprachverwendung charakterisiert, sind folgende:[97]

[94] Im Vorfeld sollte erwähnt werden, dass während der Zeit, in der diese Dissertation verfasst wurde, die Bewertungskriterien für die neuen Prüfungen des B2- und C1-Zertifikats des Goethe-Institus verschiedene Fassungen durchlaufen haben. Die Grundlage, die Validität der Bewertungskriterien zu untersuchen, zu diskutieren und gegebenenfalls Vorschläge zu ihrer Verbesserung zu machen, bilden schließlich die aktuellen Bewertungskriterien des Goethe-Instituts und die des TestDaF für die Niveaus B2/C1

95 www.goethe.de

96 Goethe-Zertifikat B2 Modellsatz 100707, S. 1

97 Goethe-Zertifikat B2 Modellsatz 100707, S. 1

- **Verstehen von komplexer gesprochener Standardsprache am Telefon und in Radiosendungen und dabei zu konkreten und abstrakten Themen die Hauptinhalte und für sich relevante Informationen entnehmen**
- **Eine Bandbreite von verschiedenen Texten verstehen**
- **Klar strukturierter Ausdruck in Briefen über komplexe Sachverhalte und Korrigieren fehlerhafter Briefe**
- **Klar strukturierte mündliche Darstellungen zu allgemeinen Themen sowie zu Themen aus dem eigenen Interessengebiet**
- **Aktive Beteiligung an informellen Diskussionen innerhalb vertrauter Kontexte und dabei Stellung nehmen und die eigenen Standpunkte darlegen und vertreten**

Im folgenden soll das Goethe-Zertifikat B2 tabellarisch hinsichtlich der zu prüfenden Kompetenzen, der jeweiligen Prüfungsziele, der Textsorten, der Aufgabentypen und der Punkteverteilung angeführt werden:[98]

	Aufgabe	Prüfungsziel	Textsorte	Aufgabentyp	Punkte
Leseverstehen,	1	Selektive Informationsentnahme	Kürzere Artikel, Anzeigen u.a.	Zuordnung	5
	2	Entnahme von Hauptaussagen und Einzelheiten	Artikel, Sachtext u.a.	Multiple-Choice (dreigliedrig)	5
	3	Erkennen von Meinungen oder Standpunkten	Stellungnahme, Kommentar u.a.	Alternativ-antwort	5
	4	Syntaktisch oder semantisch korrekte Textergänzung	Bericht u.a.	Lückentext (mit offenen Lücken)	10

Tabelle15: Kompetenz Leseverstehen im B2 Zertifikat des Goethe-Instituts

98 Goethe-Zertifikat B2 Modellsatz 100707, S. 2

	Aufgabe	Prüfungsziel	Textsorte	Aufgabentyp	Punkte
Hörverstehen	1	Selektive Informationsentnahme	Gespräch oder Nachricht auf Anrufbeantworter	Raster mit Lücken	10
	2	Entnahme von Hauptaussagen und Einzelheiten	Radiosendung (z.T. monologisch)	Multiple-Choice (dreigliedrig)	15

Tabelle16: Kompetenz Hörverstehen im B2 Zertifikat des Goethe-Instituts

	Aufgabe	Prüfungsziel	Textsorte	Aufgabentyp	Punkte
Mündlicher Ausdruck	1	Produktion: Monologisches Sprechen zu einem Thema	Statement	Text und drei Leitpunkte	12, 5
	2	Interaktion: Diskussion der Vor- und Nachteile eines Vorschlags und Aushandeln einer Entscheidung	Gespräch	Drei Fotos und drei Leitpunkte	12, 5

Tabelle17: Kompetenz mündlicher Ausdruck im B2 Zertifikat des Goethe-Instituts

	Aufgabe	Prüfungsziel	Textsorte	Aufgabentyp	Punkte
Schriftlicher Ausdruck, 80	1	Berichten, informieren, vergleichen, Ratschläge geben, Meinungen äußern	Leserbrief	Freies Schreiben nach Vorgabe von 4 Leitpunkten	15
	2	Erkennen und korrigieren von morphologischen, syntaktischen und semantischen Fehlern	Formeller Brief	Korrektur lesen	10

Tabelle18: Kompetenz schriftlicher Ausdruck im B2 Zertifikat des Goethe-Instituts

Der für diese Arbeit als wichtigster und zentral geltender Teil einer Prüfung ist die schriftliche Lernerproduktion. Der schriftliche Ausdruck beim B2-Zertifikat des Goethe-Instituts besteht aus zwei Teilen. In der ersten Aufgabentypologie wird von den Prüflingen ein Leserbrief gefordert. Es werden den Prüflingen zwei Themen zur Auswahl bereit gestellt, wobei eines davon zu bearbeiten ist. Beim zweiten Teil der Fertigkeit schriftlicher Ausdruck handelt es sich um die Korrektur eines „fehlerhaften" formellen Briefes. Mittelpunkt der Betrachtung ist der schriftliche Ausdruck, da die Bewertung offener Aufgabenformate hier von Interesse ist.

5.1.1 Aufgabenstellung für den schriftlichen Ausdruck im B2-Zertifikat des Goethe-Instituts

Die Aufgabenstellung des ersten Teils der textproduktiven Kompetenz in einer B2-Prüfung des Goethe-Instituts besteht darin, einen Leserbrief zu einem Thema und nach Vorgabe von vier Inhaltspunkten zu schreiben. In einer Beispielskala des GER für die schriftliche Produktion B2 allgemein wird die Kann-Beschreibung folgendermaßen definiert: „Kann (...) klare, detaillierte Texte zu verschiedenen Themen aus seinem/ihrem Interessengebiet verfassen und dabei Informationen und Argumente aus verschiedenen Quellen zusammenführen und gegeneinander abwägen" (GER 2001:67). Um dieses Können zu realisieren, wird eine so genannte Legende als Input gegeben, wobei es sich um eine Zeitungs- oder Internetmeldung handeln kann. Ein Original-Modellsatz aus den Trainingsmaterialien für Prüfende des Goethe-Instituts[99] soll einen Einblick geben, wie eine Aufgabenstellung zur schriftlichen Lernerproduktion aussehen kann[100]:

Aufgabe 1B **Dauer: 65 Minuten**

Im Internet lesen Sie folgende Meldung:

Große Mehrheit der Deutschen für strengere Kindererziehung

Für Kinder brechen schlechte Zeiten an: 62 Prozent der Deutschen finden, dass die lieben Kleinen wieder strenger erzogen werden sollten. Nur 31 Prozent sind einer Umfrage unter 1.000 Befragten zufolge mit den derzeitigen Erziehungsmethoden zufrieden, wie der Fernsehsender RTL am Samstag mitteilte. Mit 95 Prozent sprachen sich die meisten Befragten dafür aus, dass Kinder Pflichten wie Aufräumen und Einkaufen erfüllen sollten. 87 Prozent finden, dass Kinder regelmäßig über ihre Schularbeiten berichten sollten und 56 Prozent waren dafür, Kindern das Kaugummikauen in der Schule zu untersagen. Dagegen fanden nur 14 Prozent den Vorschlag gut, Kinder in eine Schuluniform zu stecken.

Schreiben Sie als Reaktion auf diese Meldung an die Online-Redaktion.

99 www.goethe.de/intern, Goethe-Zertifikat B2 Prüfertraining 090707

100 Das ist nicht das Originallayout. Die Arbeitsanweisung wird im Original anders präsentiert, sodass die Reihenfolge nicht als vorgegeben erscheint.

Sagen Sie,

- mit welchen der erwähnten Erziehungsmaßnahmen Sie persönlich (nicht) einverstanden sind.

- ob Kinder früher strenger erzogen wurden.

- welche Vorschläge für die Kindererziehung Sie machen möchten.

- wer für die Erziehung der Kinder zuständig ist.

Hinweise:

Vergessen Sie bitte nicht Anrede und Gruß.

Die Adresse der Internetredaktion brauchen Sie nicht anzugeben.

Bei der Beurteilung wird u.a. darauf geachtet,

– ob Sie alle vier angegebenen Inhaltspunkte berücksichtigt haben,

– wie korrekt Sie schreiben,

– wie gut Sätze und Abschnitte sprachlich miteinander verknüpft sind.

Schreiben Sie etwa 180 Wörter.

Goethe-Zertifikat B2 **Prüfertraining 090707 Seite 6**

Anhand des angeführten Modellsatzes des Goethe-Instituts wird ersichtlich, was einem Prüfungskandidaten genau vorgelegt wird. Oben rechts auf dem Kandidatenblatt wird die Zeitdauer 65 Minuten für die Bearbeitung dieser Aufgabe vorgeschlagen. Insgesamt dauert der SA 80 Minuten, wobei das Goethe-Institut 15 Minuten für den 2. Teil des schriftlichen Ausdrucks vorschlägt. Der Prüfungskandidat ist aber in seiner eigenen Zeiteinteilung völlig frei und kann die ihm zur Verfügung stehende Zeit individuell organisieren. Die situative Einbettung des Arbeitsauftrags „Im Internet lesen Sie folgende Meldung" verweist den Prüfling eindeutig darauf, dass er den Text zunächst lesen und folglich rezipieren muss. Es stellt sich dennoch die Frage, ob auf die formale Richtigkeit hingewiesen wird oder ob dies zu den erforderlichen Kompetenzen gehört, die vorausgesetzt werden. In diesem Sinne geben die Standards der APA Hilfestellung, indem darauf verwiesen wird, dass die Anweisungen zur Bearbeitung der Aufgaben so detailliert sein sollten, dass Testteilnehmer so darauf reagieren können, wie der Testentwickler dieses vorsieht (APA 2004:39). Die Aufgabenstellung muss demnach insofern rezipiert werden, dass die darin enthaltenen Informationen entnommen werden, um die geforderten vier Inhaltspunkte entsprechend zu bearbeiten. Der Testkandidat müsste anhand des Schlüsselsatzes „Reaktion auf diese Meldung an die Online-Redaktion" den konkreten Arbeitsanweisungen Folge leisten können. Es stellt sich hier zunächst und erstrangig die Frage nach dem Verständnis der zugrunde liegenden Aufgabe, um diese überhaupt bewältigen zu können. Das Gerüst der Aufgabenstellung und –anleitung muss so gut beschaffen sein, um von einer Gewährleistung der fairen Bedingungen hinsichtlich

des Testprozesses sprechen zu können. Der Arbeitsauftrag muss folglich für alle gleichermaßen verständlich sein, um Performanz unter Beweis zu stellen.

In der neu erstellten Prüfung B2 des Goethe-Instituts wird am Seitenende der Aufgabenstellung unter *Hinweise*[101] erläutert, was bei der Aufgabenbewältigung zu beachten ist. Es wird darunter vermerkt, dass Rater in derartigen Aufgabenformaten darauf achten werden, ob alle gestellten Inhaltspunkte bearbeitet wurden und wie *korrekt* und kohärent das schriftliche Konstrukt letztlich ist. Es muss verdeutlicht werden, ob es sich bei dem korrekten Schreiben um den richtigen Ausdruck oder um die korrekte syntaktische, grammatische, orthografische und morphologische Sprachverwendung handelt. Wenn das gerade für die Rater nicht explizit und ersichtlich ist, dann kann dies zu falschen Bewertungen bzw. Bewertungsgewichtungen führen. Zu erwähnen wäre schließlich, dass kein eindeutiger Hinweis bezüglich des Kriteriums Ausdrucksfähigkeit zu finden ist. Es wird lediglich vermerkt, dass bei der Bewertung darauf geachtet wird, wie korrekt man schreibt. Wie korrektes Schreiben zu verstehen ist, soll in der folgenden Diskussion aufgegriffen und erläutert werden.

Vordergründig muss aber zunächst das zu messende festgesetzte Kriterium oder auch Konstrukt und die darauf aufbauenden Bewertungsanleitungen nach APA ganz klar beschrieben und definiert werden (APA-Standard 1.2:2004):

„The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described".

Ob optimale Bewertungskriterien für das Testkonstrukt zu finden sind, ist eine zentrale Frage innerhalb der Diskussion in diesem Kapitel. Aus diesem Grund werden im Folgenden zunächst die als analytische Skala aufgestellten Bewertungskriterien angeführt, dann zunächst separat diskutiert, um schließlich die mithilfe dieser Kriterien bewerteten Lernerproduktionen kritisch zu betrachten. Ziel dabei ist es, Verbesserungsvorschläge hinsichtlich potenzieller Schwachstellen und unklarer Deskriptorendefinitionen anzuführen.

101 Siehe Modelsatz des GI

5.1.2 Bewertungskriterien für den schriftlichen Ausdruck für das B2-Zertifikat des Goethe-Instituts

Im Folgenden soll die Bewertungsskala für das B2-Zertifikat angeführt, beschrieben und analysiert werden:

KRITERIUM I Inhaltliche Vollständigkeit	3 Punkte	2,5 Punkte	2 Punkte	1 Punkt	0 Punkte
Inhaltspunkte schlüssig und angemessen dargestellt	Alle Inhaltspunkte	Drei Inhaltspunkte	Zwei Inhaltspunkte	Inhaltspunkte sind nur ansatzweise behandelt, an mehreren Stellen unklar	Thema verfehlt
KRITERIUM II Textaufbau+Kohärenz	4 Punkte	3 Punkte	2 Punkte	1 Punkte	0 Punkte
• Gliederung des Textes • Konnektoren, Kohärenz	Liest sich sehr flüssig	Liest sich noch flüssig	Stellenweise guter Aufbau, an einigen Stellen sprunghaft	Aneinanderreihung von Sätzen ohne erkennbare Gliederung	durchgängig unlogischer Text
KRITERIUM III Ausdrucksfähigkeit	4 Punkte	3 Punkte	2 Punkte	1 Punkt	0 Punkte
• Wortschatzspektrum • Wortschatz-beherrschung	Sehr gut und angemessen	Gut und angemessen	Stellenweise gut und angemessen	In ganzen Passagen nicht angemessen	In großen Teilen völlig unverständlich
KRITERIUM IV Korrektheit	4 Punkte	3 Punkte	2 Punkte	1 Punkt	0 Punkte
• Morphologie • Syntax • Orthografie, Interpunktion	kaum feststellbare Fehler	Einige deutliche Fehler, die das Verständnis aber nicht beeinträchtigen	Einige Fehler, die den Leseprozess stellenweise behindern	Unzählige Fehler, die das Verständnis erheblich stören	Unzählige Fehler, die das Verständnis unmöglich machen

Tabelle 19: Bewertungskatalog für das B2- Zertifikat des Goethe-Instituts

Diese analytische Bewertungsskala besteht aus vier Kriterien, die aber nicht die gleiche Gewichtung haben. Während das Kriterium der inhaltlichen Vollständigkeit 20% der Gesamtbewertung ausmacht, decken die Kriterien Textaufbau/Kohärenz, Ausdrucksfähigkeit und Korrektheit jeweils 26,666667 % der maximal zu erreichenden 15 Punkte. Außerdem ist die Verteilung zwischen den einzelnen Punkten unter den vier Kriterien nicht einheitlich. Interessant ist der Umstand, dass wenn eins der vier Kriterien mit 0 Punkten bewertet wird, kein Ausgleich durch die anderen Kriterien zu erzielen ist. Mit anderen Worten wird dann die gesamte Lernerproduktion mit 0 Punkten bewertet,[102] obwohl die Gewichtung der einzelnen Kriterien nicht einheitlich ist. Das erscheint mir als ein sehr wichtiger Punkt, der hinsichtlich der tatsächlichen Sprachkompetenz eine zentrale Rolle innerhalb der Diskussion darstellt.

Anhand dieses Rasters von Bewertungskriterien sollen Rater die schriftlichen Arbeiten unabhängig und separat voneinander korrigieren.[103] Im Folgenden sollen die einzelnen Kriterien der analytischen Bewertungsskala separat dokumentiert und untersucht werden. Es ist erstrebenswert, Verbesserungsvorschläge für die Definition der Deskriptoren zu liefern, um dem Gütekriterium der Validität so nah wie möglich zu kommen.

5.1.2.1 Kriterium: Inhaltliche Vollständigkeit

KRITERIUM I Inhaltliche Vollständigkeit	3 Punkte	2,5 Punkte	2 Punkte	1 Punkt	0 Punkte
Inhaltspunkte schlüssig und angemessen dargestellt	Alle Inhaltspunkte	Drei Inhaltspunkte	Zwei Inhaltspunkte	Inhaltspunkte sind nur ansatzweise behandelt, an mehreren Stellen unklar	Thema verfehlt

Tabelle 20: Inhaltliche Vollständigkeit im B2 Zertifikat

Die inhaltliche Vollständigkeit bezieht sich auf die *korrekte* Bearbeitung der Inhaltspunkte, die dem Prüfling vorgegeben sind. Bereits der oberste Definitionsdeskriptor ist derart formuliert, dass es zu einer subjektiven Raterbewertung führen kann. Es stellt sich die Frage was die Begriffe *schlüssig* oder *angemessen* implizieren und wie ein Korrektor diese *stringente Anweisung* zu verstehen hat. *Schlüssig* wird auf der CD-ROM Langenscheidts wie folgt definiert:[104]

„logisch und überzeugend = folgerichtig -> eine Argumentation, ein Beweis“

102 Goethe-Zertifikat B2 Modellsatz 100707, S. 31

103 www.goethe.de/intern, Goethe-Zertifikat B2: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 8

104 Langenscheidt: e-Großwörterbuch Deutsch als Fremdsprache. 2003 Langenscheidt KG Berlin und München. (CD-ROM)

Der Begriff oder die Definition *angemessen*, wodurch im gleichen Zuge die Darstellung des Kriteriums der inhaltlichen Vollständigkeit geprägt wird, bedeutet laut Langenscheidt „den Gegebenheiten, Umständen entsprechend = adäquat“.[105] Werden also die erforderlichen Inhaltspunkte schlüssig und angemessen dargestellt, dann müssen sie in diesem Sinne logisch, überzeugend und den Gegebenheiten entsprechend bearbeitet worden sein. Es geht hier also um eine allgemein-kognitive Komponente des Schreibens, wie ein Lerner in seinem Kopf Themen, Dinge u.Ä. strukturieren kann. Die inhaltliche Vollständigkeit impliziert in der hier definierten Form, dass die Inhaltspunkte zum einen logisch, überzeugend und folgerichtig und zum anderen adäquat bearbeitet werden. Meines Erachtens müsste es sich jedoch bei diesem Kriterium, das vielleicht durch die Klassifizierung einer allgemeinen Komponente die niedrigste Gewichtung hat, lediglich um die thematische Bearbeitung der Inhaltspunkte handeln. So könnte der Oberdeskriptor dann auch benannt werden. Denn nur unter diesem Nenner wäre z.B. der letzte Deskriptor Thema verfehlt (0 Punkte) gerechtfertigt.

Es wird also in dieser Bewertungsskala nicht ersichtlich, was das Kriterium der inhaltlichen Vollständigkeit genau zu messen vorgibt. Das Definitionsproblem, das sich aus der nicht eindeutigen Formulierung bezüglich der angesetzten und zu bearbeitenden Punkte ergibt, stellt demnach die Validität in Frage. Die Definition „*alle Inhaltspunkte schlüssig und angemessen dargestellt*“ fungiert hier als Oberbegriff für die einzelnen Abstufungen und Punktebewertungen. Darüber hinaus werden die ersten drei Deskriptoren (3 Punkte, 2,5 Punkte und 2 Punkte) sehr knapp formuliert, offensichtlich jedoch immer auf diesen Oberdeskriptor bezogen. Weiterhin sind die Intervalle zwischen den Punkteabstufungen nicht gleichmäßig, das heißt dass zwischen den ersten drei Deskriptoren im 0,5-Takt gestaffelt wird, während die Übergänge vom dritten zum vierten und zum letzten Deskriptor im 1-Punktetakt abfallen. Interessant ist bei diesen Intervallen auch der Umstand, dass, wenn zwei der vier Inhaltspunkte bearbeitet werden, also 50% des Arbeitsauftrags, es nur zu einem Punkt Abzug kommt. Dies bedeutet mathematisch ausgedrückt, dass obwohl die Lernerproduktion nur zur Hälfte den Arbeitsauftrag erfüllt hat, es nur zu einer Bewertungsminderung von 33, 33 % kommt.

Wenn alle Inhaltspunkte *schlüssig* und *angemessen* dargestellt werden, bekommt man gemäß dieses Kriteriums die maximale Punktzahl (3 Punkte). Während man zwei von drei Punkten dafür bekommt, dass man zwei von vier Inhaltspunkten auf schlüssige und angemessene Art und Weise bearbeitet hat, lautet die nächste und vorletzte Deskriptorendefinition *„Inhaltspunkte sind nur ansatzweise behandelt, an mehreren Stellen unklar“.* Zunächst müsste man an dieser Stelle definieren, was der Begriff *ansatzweise* ausdrückt. Man könnte den Begriff *ansatzweise* durch das Synonym *in geringem Maße*[106] ersetzen, dennoch müsste geklärt werden, woran sich das Maß der inhaltlichen Vollständigkeit insgesamt festmachen bzw. bestimmen lässt. Unter *Hinweise* wird der Prüfling nur darüber informiert, dass er darauf achten soll, ob *„alle vier angegebenen Inhaltspunkte berücksichtigt worden sind“*. Es wird außer der erwarteten Textlänge (ca. 180 Wörter) nichts bezüglich der Textdichte impliziert und vorgegeben. Außerdem wird nicht explizit gemacht, worauf sich die Unklarheit der Inhaltspunktebearbeitung bezieht. Auch dieser Teil des Deskriptors scheint sich eher auf

105 Langenscheidt: e-Großwörterbuch Deutsch als Fremdsprache. 2003 Langenscheidt KG Berlin und München. (CD-ROM)

106 Langenscheidt: e-Großwörterbuch Deutsch als Fremdsprache. 2003 Langenscheidt KG Berlin und München. (CD-ROM)

den subjektiven Eindruck eines Raters verlassen zu wollen. Interessant und von Bedeutung ist ebenso die Tatsache, dass die Möglichkeit ausgelassen bzw. nicht gegeben worden ist, lediglich einen Leitpunkt *schlüssig und angemessen* bearbeitet zu haben. Von der Definition des vierten Deskriptors *„Inhaltspunkte sind nur ansatzweise behandelt, an mehreren Stellen unklar“* ist der Übergang zum letzten und kritischsten Deskriptor *„Thema verfehlt“* (0 Punkte) daher sehr abrupt. Selbst wenn ein Prüfungskandidat die Leitpunkte nicht im Sinne der Aufgabe bearbeitet hat, dennoch Anrede, Einleitung, Hauptteil, Schluss und Gruß textsortengerecht erfüllt hat, sollte dies nicht einem leeren Blatt bzw. einer Themaverfehlung gleichgesetzt werden. Die Thematik der Textsorte findet aber nicht in diesem Kriterium sondern im zweiten so genannten *Textaufbau und Kohärenz* seine Anwendung. Diesbezüglich tragen Anrede, Einleitung, Schluss und Grußformel nicht dazu bei, ob das Thema erfüllt wurde oder nicht. In diesem Kriterium geht es lediglich um die Bearbeitung der erforderten Inhaltspunkte, die vom Prüfling ein bestimmtes thematisches Output erwarten. Fraglich bleibt lediglich, ob die erwartete Textlänge (180 Wörter) sich nicht auch durch die textsortenspezifischen Teile zusammensetzt. Ich will diesen Umstand an dieser Stelle jedoch völlig unberücksichtigt lassen und die Einteilung und die Inhalte der Kriterien als gegeben akzeptieren.

Die Problematik der Themaverfehlung ist weiterhin ein zweischneidiges Thema. Einerseits bezieht sich dies lediglich auf die Bearbeitung der Inhaltspunkte und nach diesen Deskriptorendefinitionen ist die Bearbeitung lediglich eines Inhaltspunktes eine Themaverfehlung, denn ansonsten ist diese Option nirgendwo aufgeführt. Interessant ist an dieser Stelle zu hinterfragen, warum es nur um die Annahme geht, dass es Schreiber gibt, die mindestens zwei Inhaltspunkte bearbeiten können, was nicht ausschließt, dass „Chaos“ produziert wird. Der Fall, dass es Schreiber gibt, die sich lediglich über einen Inhaltspunkt schriftlich äußern, bleibt völlig unberücksichtigt. Zum anderen werden das unterschiedliche Profil und die kognitive Reife der Prüfungskandidaten nicht berücksichtigt und dadurch wird die Prädikatenvergabe *Thema verfehlt* bereits subjektiv. Zur Verdeutlichung möchte ich an dieser Stelle ein Beispiel anführen:

Ein 15jähriger Prüfling soll einen Leserbrief auf ein Input schreiben, dessen Thematik aber nichts mit seinen Interessen und seinem Allgemeinwissen zu tun hat. „Kann über eine Vielzahl von Themen, die ihn/sie interessieren, klare und detaillierte Berichte schreiben“ lautet eine der Kann-Beschreibungen für das Referenzniveau B2. Wenn sich das GI nach dem GER richtet, dann sollte dieser Deskriptor Berücksichtigung bezüglich der Forderung finden. Bis zum Herbst 2008 galt noch die Prüfungsordnung des Goethe-Instituts, wobei *sich die Prüfung an Erwachsene und Jugendliche, die mindestens 16 Jahre alt sind, richtet.*[107] Wie bereits erwähnt fallen ab Herbst 2008 weltweit die Alterbegrenzungen weg. Dennoch bin ich der Meinung, dass Testanbieter sich in jedem Fall über die Alterunterschiede der Kandidaten, ihre Interessengebiete und schließlich über die Qualität und den Inhalt des entsprechenden Outputs bewusst sein müssen. Das Goethe-Institut ist dennoch bemüht, laufend neue Prüfungssätze zu entwickeln und zu veröffentlichen, um möglichst bald einen ausreichenden Fundus zu haben, aus dem dann das durchführende Institut die geeigneten Themen für die Prüfungsklientel auswählt.[108] Das ist in jedem Fall eine notwendige Voraussetzung, um die inhaltliche Bearbeitung und ihre faire Bewertung gemäß dem zu prüfenden Niveau zu gewährleisten.

107 www.goethe.de/athen>Prüfungen>Goethe-Zertifikat B2 bzw. C1. Zugriff am 13.06.2007

108 Aus einem Gespräch, das ich am 25. Februar 2008 mit Mitarbeitern des Goethe-Instituts München geführt habe

Die nächste Frage, die sich stellt ist, wer und unter welchen Bedingungen Aussagen über die inhaltliche Angemessenheit einer Lernerproduktion macht. Eine Entscheidung darüber zu treffen, ob jemand die gestellten Aufgaben inhaltlich angemessen oder auch adäquat bearbeitet hat, scheint zunächst eine subjektive Wahrnehmung der Rater zu sein. Die APA definiert in diesem Zusammenhang den Standard 3.20. Sobald die Realisierung der Anleitungsbedingungen zwischen den Testteilnehmern variieren kann müssen die erlaubten Variationen in den Anleitungsbedingungen identifiziert und dokumentiert sein. Meines Wissens sind in den Prüfungsunterlagen des Goethe-Instituts für das B2-Zertifikat keine möglichen Variationen definiert oder dokumentiert. Lediglich am Seitenende des Arbeitsauftrags wird unter Hinweise auf in der Bewertung beachtende Punkte, die es einzuhalten gilt, verwiesen. Dass Rater weltweit ständig und immer wieder trainiert werden, ist ein nicht zu ignorierender Fakt. Aus internen Quellen ist mir durchaus bekannt, dass das Goethe-Institut das Prüfertraining und was zudem noch dazu gehört sehr ernst nimmt und alles Erdenkliche unternimmt, um dem gerecht zu werden. Natürlich kann man davon ausgehen, dass dem Goethe-Institut die hier zur Diskussion und Kritik gestellten Schwachstellen der Bewertungskriterien durchaus geläufig sind und dass gerade solche Probleme, wie „was ist angemessen“ und „ab wann ist ein Thema verfehlt“ immer wieder an unzähligen praktischen Beispielen trainiert werden.

Die Gefahr, dass bereit gestelltes Inputmaterial für die schriftliche Produktion übernommen wird, ist nicht auszuschalten. Natürlich scheint dieses aufgrund des Aufgabenformats sehr schwierig zu sein. Aus diesem Grund muss bewusst gemacht werden, dass nicht immer von einer absoluten und autonomen Sprachkompetenz die Rede sein kann. Trotzdem ist die Gefahr gegeben, dass das schließlich resultierende schriftliche Konstrukt als der individuelle Ausschnitt der Schreibkompetenz gedeutet, interpretiert und schließlich bewertet werden wird. Dabei werden unter anderem die allgemeinen Kompetenzen Teil dieses schriftlichen Produktionsabschnittes, obwohl die Thematik, die für die verschiedenen Prüfungskandidaten sowohl aus einem geläufigen oder unbekannten Sachbereich stammt, keine sprachliche Größe ist. Dennoch wirkt sich diese außersprachliche Komponente im Sinne der Handlungsorientierung jedoch ohne Berücksichtigung der Interessengebiete und der Wissensbestände auf die Bewertung einer schriftlichen Lernerproduktion aus.

Es ist meines Erachtens aber nichts dagegen einzuwenden, wenn das bereit gestellte Inputmaterial von Prüfungskandidaten sprachlich modifiziert werden kann und nicht lediglich „übernommen“ wird. Das prozedurale Lernen umfasst Strategiewissen und dessen Anwendung. Wenn ein Lerner aufzeigen kann, dass er in seiner schriftlichen Produktion das Inputmaterial adäquat und entsprechend einbindet, so ist aus meiner Sicht nichts dagegen einzuwenden. Ganz im Gegenteil kommen hier Lernerstrategien zum Vorschein, die Teil des Lernens ausmachen.

5.1.2.2 Kriterium: Textaufbau und Kohärenz

KRITERIUM II Textaufbau+Kohärenz	4 Punkte	3 Punkte	2 Punkte	1 Punkte	0 Punkte
• Gliederung des Textes • Konnektoren, Kohärenz	Liest sich sehr flüssig	Liest sich noch flüssig	Stellenweise guter Aufbau, an einigen Stellen sprunghaft	Aneinanderreihung von Sätzen ohne erkennbare Gliederung	durchgängig unlogischer Text

Tabelle 21: Textaufbau und Kohärenz im B2 Zertifikat

Das Kriterium *Textaufbau und Kohärenz* hat eine Gewichtung von 26,66 %. Die maximal zu erreichende Punktzahl beträgt 4 Punkte und fällt im 1-Punkte-Takt innerhalb der Deskriptoren ab. Dieses Kriterium definiert sich in der endgültigen Fassung der Bewertungsskala für das B2-Zertifikat des Goethe-Instituts nunmehr eigenständig anhand zwei zusammenhängender Unterpunkte und es kann davon ausgegangen werden, dass dieses neu hinzugekommene autonome Kriterium neben den Kriterien Ausdrucksfähigkeit und Korrektheit ebenbürtig da steht:

- Gliederung des Textes
- Konnektoren, Kohärenz

Der GER hat für dieses Kriterium auf der Niveaustufe B2 zwei Kann-Beschreibungen, die es charakterisieren sollen (GER 2001:125):

- Kohärenz und Kohäsion

  „Kann verschiedene Verknüpfungswörter sinnvoll verwenden, um inhaltliche Beziehungen deutlich zu machen".

  „Kann eine begrenzte Anzahl von Verknüpfungsmitteln verwenden, um seine/ihre Äußerungen zu einem klaren, zusammenhängenden Text zu verbinden; längere Beiträge sind möglicherweise etwas sprunghaft".

Die Kann-Beschreibung der GER-Skala für dieses Kriteriums ist derart übergreifend, dass sie die ersten drei vom Goethe-Institut verwendeten Deskriptoren für die Bewertung dieses Kriteriums einschließt, das heißt den Bereich vier (4) bis einschließlich zwei (2) Punkten. Konkreter versucht Profile die globale Kann-Beschreibung der schriftlichen Produktion dieses Kriteriums für die deutsche Sprache speziell auf zwei Ebenen folgendermaßen zu definieren (Glaboniat et al. 2005:156, 165): Einerseits „Kann er/ sie in seinen/ihren schriftlichen Texten eine Reihe von Konnektoren und anderen Mitteln zur Textverknüpfung anwenden, um seine/ihre Ausführungen zu einem klaren, zusammenhängenden Text zu verbinden, wobei thematische Übergänge aber noch

sprunghaft bleiben können", während andererseits „er/sie bei relativ guter Grammatik eine Reihe von Konnektoren und anderen Mitteln der Textverknüpfung anwenden kann, um seine/ihre Ausführungen zu einem klaren, zusammenhängenden Text zu verbinden, wobei thematische Übergänge dabei auch noch sprunghaft bleiben können".

Hingewiesen sei an dieser Stelle auf die Verweise in der zweiten globalen Kann-Beschreibung, wobei zum einen auf schriftliche Texte und zum anderen auf die Beherrschung der Grammatik verwiesen wird. Es wird zwar auf die Textgliederung verwiesen, abgesehen jedoch von den syntaktischen Eigenschaften der Konnektoren kann dies in ihrer textlinguistischen Funktion zu finden sein. Definiert ist demnach weder im Oberbegriff noch in den einzelnen Deskriptoren, dass Textaufbau auch die Einhaltung von Textsorten einschließt. Im von mir als Pyramide abgewandelten Schreibprozessmodell von Hayes/Flower (1980) beinhaltet die Komponente Wissen auch das Textsortenwissen (vgl. Kap. 3.4). Des Weiteren sollte in den Deskriptoren explizit gemacht werden, wie Konnektoren zu bewerten sind, folglich deutlich voneinander abgegrenzt werden, so dass es nicht zu doppelten Bewertungen bzw. Doppelsanktionierungen hinsichtlich dessen kommt. Es ist meines Wissens in den mir zur Verfügung gestellten Materialien nichts darüber dokumentiert, wie dieser Doppelbewertung aus dem Weg gegangen werden kann, d.h. der Anspruch der Kohärenz könnte sich durch kohäsive Mittel problemlos im Unterpunkt Syntax der Kategorie Korrektheit wieder finden. Die Begriffe Kohärenz und Kohäsion sind unter textlinguistischen Gesichtspunkten bereits erläutert worden. Wenn einer bestimmten Satzfolge Kohärenz nicht immer zugesprochen werden kann, dann sollte zum Beispiel die so genannte Konzessivität bereits an der Textoberfläche in Form von Kohäsionsmitteln markiert werden.

Eine Lernerproduktion erlangt die volle Punktzahl 4, wenn sich die Textproduktion *sehr flüssig liest*. Es gilt zu klären, ob Lesefluss als universell definiert werden kann und welche Konnektoren und Redemittel ihn mehr oder weniger beeinflussen. An dieser Stelle stellt sich die Frage, ob Lesefluss nur an Syntax und an Morphologie festgemacht werden kann. Was ist mit dem Lesefluss im Beispielssatz „Farblose grüne Vorstellungen schlafen ruhig, weil der Schnee die Milch weggeschmolzen hat", der das syntaktische und morphologische Kriterium vollkommen deckt, aber kein Sinngehalt registriert werden kann? Es ist zudem nicht offensichtlich und ersichtlich, wann Textaufbau etwas über das geprüfte Niveau aussagt oder anders ausgedrückt an welchen kohärenten Mitteln man den Kompetenzbereich B2 festmachen kann. Ich behaupte an dieser Stelle, dass eine gut gegliederte und aufgebaute schriftliche Textproduktion des B1-Niveaus auf diesem Bewertungsraster Platz finden würde, gerade wenn geeignete kohärente Mittel für den Textaufbau benützt würden, die den Lesefluss, abgesehen von einem nicht zu enkodierenden Schriftbild, nicht im Geringsten behinderten. Sicherlich darf an dieser Stelle nicht unberücksichtigt gelassen werden, dass das jeweilige Aufgabenformat eine entscheidende Rolle spielt und dass das Goethe-Institut diesbezüglich die Aufgaben bzw. die Tasks dem Anspruch der jeweiligen Niveaus entsprechend formuliert. Ziel ist es sicherlich, dass die Aufgaben das elizitieren, was intendiert und bewertet werden will.

Interessant ist in diesem Zusammenhang ebenso die Frage, ob die Verwendung verschiedener sprachlicher Verknüpfungen als Unterscheidungsmerkmal zwischen den Niveaus B1 und B2 fungieren können. Während sprachliche Verknüpfungen (z.B. lexikalische oder strukturelle Beziehungen) einem Text Struktur verleihen und folglich die Textoberflächenstruktur ausmachen, sind außersprachliche Faktoren dafür

verantwortlich, dass die Texttiefenstruktur bzw. die Sinnzusammenhänge erkannt werden und folglich von Kohärenz gesprochen werden kann (vgl. Linke/Nussbaumer/Portmann-Tselikas 2004).

In der von Profile angeführten Kann-Beschreibung bezüglich des Textaufbaus und der Kohärenz für Niveau B2 wird von der Möglichkeit gesprochen, dass auf diesem Niveau die thematischen Übergänge sprunghaft bleiben können. Diese Einschränkung wird im Bewertungsraster dieses Kriteriums erst im dritten Deskriptor, der 50% der maximal zu erreichenden Punktzahl (2 Punkte) vergibt, definiert. Es können in diesem Fall dennoch passende kohärente Mittel eingesetzt worden sein, obwohl Rater selber eine Sprunghaftigkeit registrieren würden. Auch an dieser Stelle stellt sich erneut die Frage, ob alle Rater diese Sprunghaftigkeit registrieren würden. Die Gliederung und der Zusammenhang eines Textes haben jedoch nichts mit dem Verständnis zu tun, sofern kohärente Mittel textlinguistisch fungieren.

Einen Punkt bekommt eine Lernerproduktion, *wenn Sätze ohne erkennbare Gliederung aneinandergereiht sind*. Auch hier stellt sich die Frage nach dem *wer* etwas erkennt oder auch nicht. Mit Null Punkten wird ein *durchgängig unlogischer Text* gewertet. In Kap. 3.4 habe ich bereits Bezug auf die textlinguistische Definition Text genommen. Im vierten Deskriptor, der 1 Punkt vergibt, ist im Gegensatz zum letzten Deskriptor, der die Vergabe von Null Punkten beschreibt, nicht die Rede von Text. Bei der Definition des vorletzten Deskriptors, dass *Sätze ohne erkennbare Gliederung aneinandergereiht sind,* stellt sich die Frage, ob von einem Text gesprochen werden kann. Um die so genannte Texttiefenstruktur festzustellen, müssen die lineare Abfolge der Textbausteine, die Textverknüpfung und das Einbeziehen und Aktivieren von allgemeinem außersprachlichen Wissen betrachtet werden. Dennoch ist meines Erachtens die Definition, die für diesen Deskriptor benutzt wird, für Rater bzw. Bewerter schriftlicher Lernerproduktionen irreführend. Paradox erscheint, dass die 0-Punkte-Marke den Begriff *Text* einführt, auch wenn er als *durchgängig unlogisch* charakterisiert wird. Entweder wird ein Text im Sinne der Textlinguistik oder anderer integrativer Definitionen produziert oder es handelt sich lediglich um eine Aneinanderreihung von Sätzen, die zwar eine Einheit bilden, der Texttiefenstruktur bzw. der Kohärenz aber nicht gerecht werden kann.

Man kann in offenen Aufgabenformaten natürlich nicht den Anspruch erheben, dass Denkweisen zu Schablonen werden. Die Arbeitsvorgabe ist nicht durchnummeriert, sodass man zu einem bestimmten Produktionsmuster verpflichtet wäre. Wie und nach welchen Kriterien ein Prüfungskandidat die zur Verfügung stehenden Elemente für seine eigene Textproduktion gewichtet, ist aber eine persönliche Beurteilung. Dennoch ist es durchaus richtig, dass ein Textaufbau einer Lernerproduktion nach Logik und nach Zusammenhang untersucht und schließlich bewertet wird.

5.1.2.3 Kriterium: Ausdrucksfähigkeit

KRITERIUM III Ausdrucksfähigkeit	4 Punkte	3 Punkte	2 Punkte	1 Punkt	0 Punkte
Wortschatz-spektrum Wortschatz-beherrschung	Sehr gut und angemessen	Gut und angemessen	Stellenweise gut und angemessen	In ganzen Passagen nicht angemessen	In großen Teilen völlig unverständlich

Tabelle 22: Ausdrucksfähigkeit im B2 Zertifikat

Das Kriterium zur Bewertung des schriftlichen Ausdrucks auf B2-Niveau hat zunächst eine Reduzierung seiner Gewichtung erfahren. Von einst in der Erprobungsfassung 40% bewegt sich die Ausdrucksfähigkeit nun auf der gleichen Gewichtungsschiene (26,666667%) wie das zuvor behandelte Kriterium Textaufbau und Kohärenz und das im Folgenden noch aufgeführte Korrektheitskriterium. Bei diesem Kriterium soll das Wortschatzspektrum und die Wortschatzbeherrschung untersucht werden. Zunächst gilt zwischen den Begriffen Wortschatzspektrum und Wortschatzbeherrschung zu differenzieren. Das Wortschatzspektrum scheint sich auf die Variation und die Vielfalt des Wortschatzes in Form von kontextbezogenen Registern zu beziehen, welcher dann in der Wortschatzbeherrschung Anwendung findet. Es handelt sich also um den Umfang (breath) und die Tiefe (depth) des benutzten Wortschatzes. Ganz explizit sollte aber auch hier sicher gestellt werden, dass das jeweils vorliegende Wortschatzspektrum bzw. die Wortschatzbeherrschung einer Lernerproduktion gemessen wird. Dies beruht auf dem Umstand, dass ein Test lediglich ein kleiner Ausschnitt der fremdsprachlichen Kompetenz ist. Im GER lauten die Kann-Beschreibung für das allgemeine Spektrum sprachlicher Mittel folgendermaßen (GER 2001:110):

„Kann sich klar ausdrücken, ohne dabei den Eindruck zu erwecken, sich in dem, was er/sie sagen möchte, einschränken zu müssen".

„Verfügt über ein hinreichend breites Spektrum sprachlicher Mittel, um klare Beschreibungen, Standpunkte auszudrücken und etwas zu erörtern; sucht dabei nicht auffällig nach Worten und verwendet einige komplexe Satzstrukturen".

Der GER definiert aber auch detaillierte Skalen der lexikalischen Kompetenz, welche als die Fähigkeit verstanden werden kann, dass lexikalische und grammatische Elemente verwendet werden. Das Wortschatzspektrum lexikalischer Elemente (z.B. idiomatische Wendungen, Funktionsverbgefüge) und grammatischer Elemente (z.B. Artikel, Präpositionen, Modalpartikel) und deren Beherrschung werden in den vorhandenen Skalen des GER folgendermaßen beschrieben (GER 2001:112ff.):

- Wortschatzspektrum

„Verfügt über einen großen Wortschatz in seinem Sachgebiet und in den meisten allgemeinen Themenbereichen. Kann Formulierungen variieren, um häufige Wiederholungen zu vermeiden, Lücken im Wortschatz können dennoch zu Zögern und Umschreibungen führen".

- **Wortschatzbeherrschung**

„Die Genauigkeit in der Verwendung des Wortschatzes ist im Allgemeinen groß, obgleich einige Verwechslungen und falsche Wortwahl vorkommen, ohne jedoch die Kommunikation zu behindern".

Es wird hierbei die Unterscheidung zwischen deklarativem und prozeduralem Wissen sichtbar. Es ist durchaus möglich, dass man eine große Wortschatzbreite aufweist und dieses Wissen auch konkret und richtig, entsprechend und situationsspezifisch einsetzen kann. Im Sinne dieser Skalen scheinen auch die von Goethe-Institut verfügbaren Deskriptoren der ersten drei Abstufungen (4-2 Punkte) definiert zu sein. Abgesehen davon, dass sich diese Kann-Beschreibung unter anderem auf *allgemeine Interessen* bezieht, stellt sich die Frage, wie man das Wortschatzspektrum zu messen vermag. Laut Referenzrahmen soll man sich auf die vier Domänen privater Bereich, öffentlicher Bereich, beruflicher Bereich und bildender Bereich beschränken. Der GER betont hinsichtlich der Benutzung dieser zur Verfügung stehenden Skalen, dass man bedenken und angeben sollte (GER 2001:113),

- **welche lexikalischen Elemente die Lernenden erkennen und/oder verwenden müssen, auf welche lexikalischen Elemente sie vorbereitet werden sollen und welche Anforderungen in dieser Hinsicht an sie gestellt werden;**
- **wie lexikalische Elemente ausgewählt und angeordnet werden.**

Was die Ausdrucksfähigkeit auf Niveau B2 insgesamt ausmacht, geht aus den deskriptiven Abstufungen nicht hervor. Man kann allerdings davon ausgehen, dass Rater in ihren Schulungen mit dem Anspruch des Wortschatzes des jeweiligen Niveaus (in diesem Fall B2), den der GER vorschreibt, vertraut gemacht werden und diesen anhand von Lernerproduktionen soweit wie möglich trainieren und gegebenenfalls untereinander diskutieren. Eine Problematik, die sich an dieser Stelle entpuppt, ist, wenn der zu Papier gebrachte Wortschatz der Aufgabenvorgabe entlehnt ist. Völlig legitim ist dies meines Erachtens, wenn die Arbeitsvorlage situationsadäquat eingebunden ist und somit als eigenständiges Produkt bewertet werden kann.

Mit dem ersten und höchstbewerteten Deskriptor werden 4 Punkte erzielt, wenn Wortschatzspektrum und Wortschatzbeherrschung *sehr gut und angemessen* sind. Der GER betont bezüglich der Entwicklung von Deskriptoren, dass konkrete Aufgaben bzw. konkrete Fertigkeitsgrade bei der Ausführung von Aufgaben beschrieben werden sollen. Aus diesem Grund sollen die definierten Deskriptoren keine Vagheiten enthalten wie es in den Deskriptoren dieses Kriteriums der Fall ist (GER 2001:201). Anders ausgedrückt ist es erstrebenswert, keine quantitativen sondern qualitativen Bezeichnungen zu verwenden. Was bedeutet *gut* und *angemessen* in ihrer ganzen Bandbreite bezogen auf das Niveau B2 und wie wird dieses schließlich von den Ratern interpretiert? Es wird deutlich, dass im hiesigen Fall nicht explizit gemacht werden kann, worauf sich die Angemessenheit des Kriteriums Ausdruck bezieht. Der dritte Deskriptor wird um das Adverb *stellenweise* erweitert, das die Adjektive *gut* und *angemessen* modifiziert. Anders formuliert bekommt man als Testkandidat 2 Punkte, sobald die Ausdrucksfähigkeit nur an einigen Stellen gut und angemessen ist. Während man einen Punkt für den Ausdruck, der *in ganzen Passagen nicht angemessen ist,* bekommt, ist die Ausdrucksfähigkeit der letzten Etappe *in großen Teilen völlig unverständlich* mit 0 Punkten gekennzeichnet. Auch wodurch *ganze Passagen* und *große Teile* definiert sind, geht aus dem Deskriptor nicht

hervor. In diesem Sinne müssten Rater auch die mathematische Kompetenz mitbringen, die Stellen und Passagen zählen und die Prozentsätze dafür berechnen zu können (immer auf die Textlänge einer Lernerproduktion bezogen). Zudem kann der Eindruck über die bestehende oder nicht bestehende und noch angemessene Ausdrucksfähigkeit und Verständlichkeit einer Lernerproduktion ganz subjektiv sein. Unterscheidungen sollten auch gemäß des GER nicht davon abhängen, dass man Graduierungen wie *einige* oder *ein paar* auf der nächst höheren Stufe durch *viele* oder *die meisten* ersetzt (GER 2001:201). Auf diese Deskriptoren bezogen, wird aufsteigend folgendermaßen graduiert:

Stellenweise gut und angemessen - gut und angemessen - sehr gut und angemessen

Diese quantitativen Bezeichnungen, die in analytischen Skalen oft verwendet werden, verleiten dazu, dass man sein Augenmerk auf die Schwächen in einer Lernerproduktion richtet.

Aus der Aufgabenstellung für das Niveau B2 ergibt sich sicherlich ein bestimmter Kontext und folglich ein bestimmtes Register. Die Frage bezüglich der Angemessenheit des Wortschatzes ergibt sich demnach in erster Linie aus der Aufgabe selbst. Das Goethe-Institut gibt in den Abstufungen des Kriteriums der Ausdrucksfähigkeit keinen Aufschluss darüber, was die B2-Ebene ausmacht. Meines Erachtens ergibt sich das Register aus der Aufgabenstellung erst daraus, was in einem Brief bzw. Leserbrief behandelt werden soll. Natürlich können an dieser Stelle die internen Bewertungsrichtlinien und die analogen Trainingseinheiten für die Rater, die wie bereits betont sicherlich diesem Kriterium eine besondere Gewichtung schenken, dazu nicht dokumentiert werden, dennoch müsste das Goethe-Institut selbst laut APA hinsichtlich der Bewertungskriterien eine komplette Durchführungsdokumentation zur Verfügung stellen. In den Duchführungsbestimmungen steht über die Bewertung des schriftlichen Ausdrucks lediglich, dass sie „nach den Bewertungskriterien aus den Prüferblättern (S.8) erfolgt und dass als Hilfe für die Bewertung dieser Aufgabe Kandidatenbeispiele in den Trainingsmaterialien für Prüfende zur Verfügung stehen"[109]. Das bezieht sich natürlich auf den ganzen Kriterienkatalog. Des Weiteren scheint es *irreal*, gar unmöglich, dass Rater bzw. Korrektoren den erforderten, produktiv anzuwendenden Input für das Niveau B2 kennen. Wäre dem so, müsste man der Frage nachgehen, welcher Freiraum für synonyme Wortverwendungen einer niedrigeren Referenzskala gegeben wäre. Testet man die Ausdrucksfähigkeit des Niveaus B2, dann wird das gesamte Wortschatzspektrum A1 bis einschließlich B2 eingeschlossen. Wie ist die Bewertung demnach zu handhaben, wenn der Prüfling das Wortschatzspektrum aller vier Niveaustufen beherrschen soll?

***Profile* hat den Niveaustufen entsprechend Wortschatzlisten für Rezeption und Produktion erstellt. Inwieweit diese Ausarbeitung einem korpuslinguistischem Fundament zugrunde liegt bleibt fraglich. Meines Erachtens handelt es sich um ein willkürliches Konstrukt, auf das man das Wortschatzspektrum und die Wortschatzbeherrschung nicht beziehen darf.**

Bereits ab Niveau B2 scheinen die Niveaukategorisierungen von Ausdruck und grammatikalischen Strukturen eine Schwierigkeit zu bereiten. *Profile* selbst deutet auf die Schwierigkeit der Beschreibung der Referenzniveaus hin, denn „je höher das Niveau, desto weniger lassen sich niveauspezifische sprachliche Mittel definieren", obwohl die Komplexität der sprachlichen Handlungsabläufe je nach Niveau ansteigt" (Glaboniat et al. 2005:46). Positionen, die Wortschatz zu kategorisieren vermögen, wie *Profile* dies tut,

109 Goethe-Zertifikat B2: Prüfungsordnung. Stand: 100707. S. 6

sind meines Erachtens daher nicht hieb- und stichfest. Mittels Tests kann der Wortschatzumfang von Lernern bestimmt werden, das heißt wie weit die mentale Wortschatzverknüpfung fortgeschritten ist. Auf niedrigen Niveaus ist der Wortschatzerwerb noch intentional. Mit aufsteigender Sprachkompetenz nimmt die Wortschatzerweiterung auch dadurch zu, dass anhand der unterschiedlichen Thematiken ein Wechsel vom Konkreten zum Abstrakten statt findet.[110] Dennoch teile ich keineswegs die Position von *Profile*, dass man Wortschatz in Kategorien und in Niveaus fassen kann.

Dennoch will ich an dieser Stelle auf diese Ausarbeitung von *Profile* basierend folgende Frage zur Diskussion stellen: Was wäre, wenn sich die produzierten Wörter auf dem Niveau B1 befänden? Interessant zu eruieren ist, ob ein Kandidat, der sich in diesem Streuungsbereich bewegt, dabei aber trotzdem *sehr gut* formuliert, nicht doch 4 Punkte bekommen würde, obwohl der Wortschatz laut Deskriptor *angemessen* sein muss.

Skepsis ist meines Erachtens bei der Bewertung von Lernerproduktionen angebracht, wenn sich diese auf den Wortschatzinput und die Klassifizierung von *Profile* bezieht. Diese erarbeiteten und endlosen Wortschatzlisten können keineswegs eine Hilfestellung sein, um eine schriftliche Leistung (hier: B2) zu bewerten. Dies begründe ich damit, dass zum einen die Basis dieser Erarbeitung willkürlich und subjektiv ist und zum anderen ein Rater diese Wortschatzlisten auswendig kennen müsste, um Lernerproduktionen den Niveaustufen entsprechend zuzuordnen. Die Befürworter von *Profile* könnten an dieser Stelle das Argument anführen, dass Rater den erforderten Input mithilfe der *Profile* CD-ROM oder Wortschatzlisten ermitteln könnten. Selbst wenn Profile einem soliden Fundament unterläge, wäre dieser Lösungsweg testtheoretisch sicherlich nicht im Sinne der Gütekriterien Praktikabilität und Ökonomie. Zum einen würde dies sehr viel Zeit in Anspruch nehmen, folglich wäre das für die Testanbieter finanziell nicht tragbar und zum anderen würde die Bewertung dadurch nicht erleichtert. Ein weiteres Argument hinsichtlich der Schwierigkeit, Ausdrucksvermögen gemäß der Wortschatzlisten nach Niveaus zu ordnen, scheint der Überforderungsschwellenwert der Rater ab 5 Deskriptoren aufwärts zu liegen (vgl. Kap. 4.5.2). Die Kontroverse, die hier ganz deutlich wird ist: Wie kann ein Rater mit einem großen „definierten" Wortschatzinput vertraut sein, während er schon durch die Anzahl von 5 Deskriptoren aufwärts überfordert zu sein scheint?

Bei diesem Kriterium handelt es sich meines Erachtens um eine schwer zu definierende Kompetenz und das ganz unabhängig von dem Niveau, das abgeprüft wird. Natürlich lässt sich auf elementarer Ebene dieses Kriterium viel einfacher und leichter definieren, denn man weiß, man hat auf Niveau A1 beispielsweise einen sehr konkreten und elementaren Input und demnach entsprechenden Output zu erwarten. Aufsteigend wird es komplexer, denn Sprache wird reicher an Struktur, an Wortschatz und an Verständnis (GER 2001:28f.). Folglich bezieht man alles bisher Erlernte in einen Sprachlernprozess mit ein. Wenn man also die B2-Prüfung ablegen und in schriftlicher Produktion seine Sprachkompetenz unter Beweis stellen möchte, so wird das Resultat eine Verschmelzung aus elementaren und übergreifenden Sprachverwendungen sein. Die Aufgabenstellung muss sich auf einer sprachlichen Ebene bewegen, die allen Prüfungsteilnehmern gegenüber fair ist. Dabei ist es dennoch möglich, dass der Wortschatz sowohl bekannt als auch unbekannt ist. Die bewusste oder unbewusste Verwendung des initiierten Wortschatzes ist aber für den Leser nicht unbedingt ersichtlich. Es stellt sich weiterhin die Frage, ob im Rahmen der Bewertungskriterien dieser Umstand Berücksichtigung findet, dass Wörter aus der Legende übernommen werden, um textproduktiv zu werden.

110 Anhand der Kernlehrpläne für L1 habe ich dies bereits ausführlich dargestellt

Es wurde bereits erwähnt, dass Wortverwendungen und Sprachhandlungen verschiedener Niveaus aufgelistet würden, die von *Profile* nach Niveaus kategorisiert worden sind und an dieser Stelle von mir als synonyme Wendungen gegenüber gestellt werden, um die Problematik synonymer Wörter aufzuzeigen, wobei diese von *Profile* unterschiedlichen Niveaus zugeordnet werden. Die linke Spalte der Tabelle definiert den Wortschatz, den man laut Profile auf Niveaus B2 produktiv verwenden können sollte. Konträr dazu befinden sich in der rechten Spalte gleicher Tabelle Synonyme, die sich rezeptiv und produktiv auf anderen Niveaustufen befinden:

B2-produktiv	Andere Niveaueinteilungen
Türklinke (+B1 rezeptiv)	Türgriff -> B1 produktiv/rezeptiv
Geländer (+ B2 rezeptiv)	Treppenhaus -> B1 produktiv/rezeptiv
Gardine (+B2 rezeptiv)	Vorhang -> B1 produktiv/rezeptiv
mager (+ B1 rezeptiv)	dünn -> A2 produktiv/ A1 rezeptiv
Tiergarten (+ B1 rezeptiv)	Zoo -> B1 produktiv/ A2 rezeptiv
falls (+ B1 rezeptiv)	wenn.....dann....-> B1 produktiv/ A2 rezeptiv
beinahe (+B1 rezeptiv)	fast -> A2 produktiv/rezeptiv
dankbar sein (+ B1 rezeptiv)	jdm. Danken -> B1 produktiv/A2 rezeptiv
gebürtig (+ B2 rezeptiv)	geboren sein -> A1 produktiv/rezeptiv
sich langweilen (+ B1 rezeptiv)	jdm. Langweilig sein -> B1 produktiv/A2 rezeptiv
sich ärgern über/darüber...(+B1 rezeptiv)	sich ärgern, weil/dass....-> B1 produktiv/A2 rezeptiv
jdn. (nicht) leiden können (+B1 rezeptiv)	jdn. (nicht) mögen -> A2 produktiv/A1 rezeptiv
scheinbar/anscheinend (+B1 rezeptiv)	Es scheint, dass.....-> B1 produktiv/A2 rezeptiv
eventuell (+B1 rezeptiv)	vielleicht -> A2 produktiv/A1 rezeptiv
Denken Sie an/daran.....! (+B1 rezeptiv)	Vergessen Sie nicht....! -> B1 produktiv/A2 rezeptiv
tatsächlich (+B1 rezeptiv)	wirklich -> A2 produktiv/rezeptiv
keinesfalls (+ B1 rezeptiv)	Auf keinen Fall -> B1 produktiv/A2 rezeptiv
selbstverständlich (+ B1 rezeptiv)	natürlich -> B1 produktiv/ A1 rezeptiv
Es fällt mir nicht ein...(+B1 rezeptiv)	Ich habe vergessen, ob....-> B1 produktiv/A2 rezeptiv
etw. für gut/schlecht halten (+B1 rezeptiv)	etw. gut/schlecht finden -> B1 produktiv/A2 rezeptiv

Der Ansicht sein/meiner Ansicht nach… (+B1 rezeptiv)	Der Meinung sein/meiner Meinung nach -> B1 produktiv/A2 rezeptiv
ausgezeichnet! (+B2 rezeptiv)	phantastisch! -> B1 produktiv/rezeptiv
absichtlich (auch als Verneinung)-> (+A2 rezeptiv)	Mit Absicht (auch als Verneinung) -> B1 produktiv/ A2 rezeptiv
weshalb (+ A2 rezeptiv)	wieso -> A1 produktiv/rezeptiv
Ja, mag sein. (+ A2 rezeptiv)	Ja, kann sein. -> A2 produktiv/A1 rezeptiv
vor allem (+B1 rezeptiv)	besonders -> B1 produktiv/A2 rezeptiv
Das ist furchtbar! (+B1 rezeptiv)	Das ist schrecklich! -> B1 produktiv/A2 rezeptiv
Das macht mir Angst (+ B1 rezeptiv)	Ich habe Angst, weil….->B1 produktiv/A2 rezeptiv
Würden Sie mal…..tun? (+ B1 rezeptiv)	Würden Sie bitte…..tun? -> B1 produktiv/A2 rezeptiv

Tabelle 23: Gegenüberstellung synonymer Ausdrücke nach *Profile*

Profile **versteht sich wie der GER auch als ein offenes, transparentes und kohärentes System von Niveaubeschreibungen, welches sich nicht auf „endgültige Fassungen" beschränkt (Glaboniat et al. 2005:53). Die erarbeiteten Listen der verschiedenen Bereiche wurden aus Lernzielkatalogen verschiedener Arbeitsgruppen zusammengestellt (Glaboniat et al. 2005:43). Wie schon mehrfach angedeutet wurde, ist dennoch nicht ersichtlich, auf welchen linguistischen bzw. korpuslinguistischen Grundlagen die Arbeit von *Profile* begründet liegt. Dennoch habe ich mittels der CD-ROM von *Profile* diese Synonymliste erstellt, um im Rahmen des GER die Ausdrucksfähigkeit zu diskutieren. Interessant ist, wie die Bewertung einer schriftlichen Lernerproduktion, die mehrheitlich aus einem Wortschatzfundus unterhalb des Niveaus B2 besteht (siehe rechte Spalte obiger Tabelle), ausfallen würde. Wie Korrektoren die „nicht angemessene" Wortschatzverwendung anhand des analytischen Bewertungsrasters dokumentieren, einordnen und schließlich bewerten würden, bleibt schließlich zu klären. Man sieht an dieser Stelle, wie vorsichtig diese Thematik anzugehen ist, um die Validität nicht nur der Bewertungskriterien, sondern auch des gesamten Testes, zu gewährleisten.**

Das Vokabular des Arbeitsauftrages für das B2-Zertifikat sollte keinerlei Schlüsselwörter verwenden, die rezeptiv über das Niveau B1 reichen, damit dem Anspruch der APA, dass niemand einen unfairen Vorteil hat, wenn angemessen konstruierte und angewendete Tests die sozialen Ziele der Fairness und die Gleichheit der Gelegenheiten fördern, Rechnung getragen werden kann (vgl. APA 2004). In diesem Sinne kann bei Testkonzepten, die das Wissen oder die Fähigkeit eines Testteilnehmers festsetzen möchten durch Standardisierung gewährleistet werden, dass alle Testteilnehmer die gleichen Möglichkeiten haben, um ihre Kompetenz zu demonstrieren. Sicherlich ist das Goethe-Institut in seiner Testerstellung bemüht, dass sich die Items bzw. die Aufgaben für die B2-Prüfung auf einem sprachlichen Niveau bewegen, das sich im A2/B1-Bereich bewegt. Dadurch kann gewährleistet werden, dass die

Aufgabenstellung von jedem B2-Prüfungskandidaten rezipiert und schließlich bearbeitet werden kann. Es stellt sich aber dennoch die Frage, nach welchen Kriterien das Wortschatzspektrum ausgewählt wird und welchem Fundament dies zugrunde liegt.

5.1.2.4 Kriterium: Korrektheit

KRITERIUM IV Korrektheit	4 Punkte	3 Punkte	2 Punkte	1 Punkt	0 Punkte
Morphologie Syntax Orthografie, Interpunktion	kaum feststellbare Fehler	Einige deutliche Fehler, die das Verständnis aber nicht beeinträchtigen	Einige Fehler, die den Leseprozess stellenweise behindern	Unzählige Fehler, die das Verständnis erheblich stören	Unzählige Fehler, die das Verständnis unmöglich machen

Tabelle 24: Korrektheit im B2 Zertifikat

Das vierte und letzte Kriterium der Bewertungsskala für das B2-Zertifikat des Goethe-Instituts beinhaltet in seiner überarbeiteten und endgültigen Version die Unterbereiche Morphologie, Syntax, Orthografie und Interpunktion und macht 26,66 % der Gesamtbewertung aus. Da bezüglich der genauen Gewichtung der Unterbereiche nichts dokumentiert wird, wird davon ausgegangen, dass die drei Unterkriterien gleichwertig sind. Der GER kann bezüglich dieses Kriteriums die grammatische Korrektheit und die Beherrschung der Orthografie anhand von Kann-Beschreibungen in Skalen fassen (GER 2001:114 ff):

- **Grammatische Korrektheit**

„Gute Beherrschung der Grammatik; gelegentliche Ausrutscher oder nicht-systematische Fehler und kleinere Mängel im Satzbau können vorkommen, sind aber selten und können oft rückbildend korrigiert werden".

„Gute Beherrschung der Grammatik; macht keine Fehler, die zu Missverständnissen führen".

In dieser Skala ist von guter Beherrschung der Grammatik die Rede, es wird jedoch nicht explizit eingegrenzt, aus welchen grammatikalischen Komponenten und grammatischen Phänomenen der deutschen Sprache das Referenzniveau B2 ausgemacht wird. Außerdem schließt diese Kann-Beschreibung des GER hier mehr als nur den vom Goethe-Institut formulierten obersten Deskriptor für dieses Kriterium ein. Solange *Fehler nicht zu Missverständnissen führen,* können laut der Definition der hier benutzten Deskriptoren im schlechtesten Fall bis zu zwei (2) Punkte von maximal vier (4) zu erreichenden erzielt werden.

In Profile lauten die globalen Kann-Beschreibungen für schriftliche Interaktion (SI) und schriftliche Produktion (SP) des B2-Niveaus für die Unterpunkte des Kriteriums Korrektheit folgendermaßen (Glaboniat et al. 2005:156, 165):

- „Kann in Texten seine/ihre Kenntnisse in der deutschen Sprache bei relativ guter Beherrschung der Grammatik so anwenden, dass kaum Fehler entstehen bzw. kann viele Fehler selbst korrigieren". (SP)
- „Kann seine/ihre schriftlichen Texte weitgehend grammatikalisch korrekt verfassen, wobei gelegentlich nicht systematische Fehler und syntaktische Mängel vorkommen". (SI)
- „Kann Orthografie und Interpunktion weitgehend regelkonform anwenden". (SP)
- „Kann Orthografie und Interpunktion so korrekt anwenden, dass aus eventuellen Fehlern keine Missverständnisse entstehen". (SI)

Als erstes muss angenommen werden, dass die Kann-Beschreibungen generell - aber auch für das vorliegende Kriterium speziell - sicherlich nur die maximale Punktbewertung definieren. Sie stehen prinzipiell für das geforderte Interlanguagestadium auf einem bestimmten Niveau. Wie schon angedeutet, umfasst das Korrektheitskriterium viele Bereiche und daher sollte explizit gemacht werden, wie die Prioritäten innerhalb dieser Kategorie gesetzt sind und was das für die Bewertung zu bedeuten haben könnte. Obwohl meines Erachtens diese Unterbereiche nicht zusammengefasst werden dürften, wie in Kapitel 6 ausführlich erläutert werden wird, gilt es anhand der Gegebenheiten herauszufinden, was für die Korrektheit der Schriftsprache am repräsentativsten und am gewichtigsten ist. Anders ausgedrückt ist zu hinterfragen, ob alle Unterpunkte gleichwertig sind oder ob es eine Rangfolge gibt. Es wird aus keiner Quelle des Goethe-Instituts deutlich, wodurch Korrektheit definiert wird. Hier stellt sich folglich die Frage nach der Definition der Korrektheit. Nach Langenscheidt sollen „bestimmte (gesellschaftliche) Normen genau eingehalten werden".[111] Demnach bekommt man bei dem ersten Deskriptor die maximale Punktzahl 4, wenn es *kaum feststellbare Fehler* gibt. Natürlich sind Fehler in der Regel feststellbar, es sei denn, es handelt sich um latente Fehler, die zum Beispiel durch Vermeidungsstrategien hervorgerufen werden und zustande kommen. Gerade im Bereich der Morphologie, der Syntax, der Orthografie und der Interpunktion sind Fehler dennoch am offensichtlichsten. Wird hier mit *kaum feststellbar* angedeutet, dass es sich um nicht gravierende Fehler handelt? Dazu müsste man erstmal definieren, was ein gravierender, schwerer Fehler in diesem Bereich ist.

Erstmals wird in diesem Bewertungskatalog der Begriff Fehler eingeführt. In Sprachstandsprüfungen verschiedener Anbieter scheinen Fehler an der zielsprachlichen Norm orientiert zu sein. Dabei muss man sich meines Erachtens aber darüber im Klaren sein, dass die eindeutige und allgemeingültige Zuweisung von richtig oder falsch nicht möglich zu sein scheint. Während B2 impliziert, dass man bezogen auf ein L1-Niveau Fehler machen darf, macht das Goethe-Institut für die maximale Punktzahl von 4 Punkten allerdings fest, dass es kaum Fehler gibt, die feststellbar sind. Da es sich bei den Niveauzuschreibungen für mich um Interlanguages handelt, können die verschiedensten Fehlerquellen und –ursachen innerhalb dieser nicht absolut eruiert werden. Es wird in

111 Langenscheidt: e-Großwörterbuch Deutsch als Fremdsprache. 2003 Langenscheidt KG Berlin und München. (CD-ROM)

diesem Deskriptor zudem nicht explizit, welche Niveaus und Can-Dos darunter fallen. Zur Verdeutlichung sei folgendes Beispiel angeführt: Angenommen man benützte Kausaladverbien der Niveaustufe B1, konditionale Subjunktionen und Temporaladverbien der Elementarstufe A2, und all das *korrekt* - wie viele Punkte würde man für eine derartig tadellose Leistung bekommen? Die Definition des ersten und höchstbewertenden Deskriptors ist sehr allgemein und kann selbst korrekte Formen niedrigerer Niveaus beinhalten. Die Frage, die sich hier stellt, ist, ob B2 alle darunter liegenden Stufen einschließt und wo die Fehler, falls sie gemacht werden, vorkommen. Diese dürften sich in diesem Sinne nur im Bereich B2 befinden, denn alles darunter liegende müsste beherrscht werden. Ich bezweifle dennoch, dass das im absoluten Sinne der Fall ist. Es ist auf der einen Seite natürlich nicht anzuzweifeln, dass Sprache ein Ineinanderfließen ist, das im Laufe des Lernprozesses verstärkt und ausgebaut wird. Auf der anderen Seite aber führt der Zwang der Kategorisierung und der Einteilung in Niveaustufen und Kann-Beschreibungen jedoch dazu, dass man den Anspruch und den Konsens daran festzumachen hat. Es stellt sich die Frage danach, ob ein Prüfungsteilnehmer oder Lerner der selbständigen Sprachverwendung, die das B-Niveau ausmacht, gerecht wird. Dabei muss festgelegt werden, welche Fehler, Formen und Ausdrucksweisen ihn eher als *elementaren Sprachverwender* der Niveaustufe A definieren und durch welche Fähigkeitskombinationen er gute bzw. bessere Bewertungen, die in den B-Bereich fallen, erzielt.

Krings (1988) belegte bereits empirisch, dass die schriftliche Produktion in der Zielsprache eingeschränkte Automatisierung aufwirft (vgl. Kap. 3.4). Nach Bart (1999:89) wird die Fehlerbewertung dennoch an akzeptablen bzw. nicht akzeptablen zielsprachlichen Konstrukten gemessen. In diesem Sinne beinhaltet der nächste Deskriptor insgesamt 3 Punkte, wenn in der schriftlichen Produktion „einige deutliche Fehler, die das Verständnis nicht beeinträchtigen" auftreten. Auch an dieser Stelle ist das mathematische Verständnis der Rater erforderlich, um den Maßstab für *einige* einstimmig zu setzen. Es muss zunächst die Frage beantwortet werden, um wessen Verständnis es sich hier eigentlich handelt, das nicht beeinträchtigt wird. Im Sinne des Kapitels 4.5.3 muss erneut deklariert werden, dass menschliche Rater in erster Linie individuell und folglich unterschiedlich sind. Natürlich betrifft dies, wie bereits herausgestellt, die Frage der Bewertungsübereinstimmung und Bewertungsreliabilität, was nicht mit den Bewertungskriterien an sich gleichzusetzen ist, obwohl diese Einfluss auf die beiden erst genannten nehmen.

Die nächste Abstufung des Kriteriums scheint verglichen zu dem gerade dokumentierten Deskriptor etwas schwächer in ihrer Definition zu sein: *einige Fehler, die den Leseprozess stellenweise behindern*. An dieser Stelle kristallisiert sich meines Erachtens eine Lücke zwischen dem zweiten und dritten Deskriptor heraus. *Verständnis* wird hier durch *Leseprozess* ersetzt. Die Behinderung des Leseprozesses könnte zweifellos auch durch ein schlechtes Schriftbild verursacht werden, obwohl die sprachliche Qualität im Sinne des Korrektheitsanspruchs der Anforderung entspricht. Zudem muss man auch hier die Frage aufstellen, ob und unter welchen Umständen der *Leseprozess* bei allen Ratern gleichermaßen behindert bzw. gestört wird. Während der 3 Punkte erzielende Deskriptor keine Verständnisbeeinträchtigung durch *einige deutliche Fehler* erfährt, wird beim nächsten der Leseprozess bereits durch die Fehlerfrequenz lediglich *einiger Fehler* behindert. Zu definieren gilt hier, was unter *Leseprozess* verstanden wird. Das Leseverstehen kann als ein Prozess betrachtet werden, der zum Produkt Leseverständnis führt (vgl. Grotjahn 2000a). Ist demnach der Prozess des

Leseverstehens, also der Lesefluss stockend, dann wird das Verständnis beeinträchtigt. Dennoch muss dies keine allgemeingültige Aussage für alle Rezipienten und das evtl. gestörte Leseverständnis sein. Außerdem gilt es zu klären, wie Fehler hier nach Kategorien zu gewichten wären. Es wird weder etwas über den Fehlertypus noch über den zugehörigen Bereich ausgesagt.[112] Der vierte Deskriptor vergibt einen Punkt, wenn *unzählige Fehler das Verständnis erheblich stören*. Es muss ausdrücklich gemacht werden, welche Arten von Fehlern eine erhebliche Verständnisstörung hervorrufen können und zudem bei wem (bezogen auf die Rater) dies schließlich eintritt. Es müsste sich demnach um Fehler handeln, die bei keinem Korrektor eine Struktur erkennen lassen würden. In der Definition dieses Deskriptors wird das nicht explizit gemacht. Dadurch stellt sich berechtigterweise die Frage, ab welchem Moment eine Lernerproduktion derart „chaotisch" ist, dass das Verständnis erheblich gestört wird. Auch beim letzten Deskriptor wird nicht die Fehlerart sondern die Quantität der Fehler zum Mittelpunkt, wodurch *das Verständnis unmöglich gemacht wird* (0 Punkte). Man muss erneut darauf hinweisen, dass die Fehlertypologie in der Definition des Deskriptors nicht festgelegt zu sein scheint. Handelte es sich z.B. lediglich um schwere orthografische Fehler, die syntaktische Struktur wäre aber nicht zu bemängeln, dann wäre interessant, ob dies dennoch zur Einstufung auf den letzten Deskriptor führen würde. Es wird also aus der definierten Kategorie Korrektheit nicht ersichtlich, welche Unterkriterien das Verständnis und zu welchem Grad beeinflussen können.

Generell ist bei diesem Bewertungskriterium und seinen Abstufungen für den schriftlichen Ausdruck auf B2-Niveau nicht eindeutig, welche Arten von Strukturen und zu welchem Grad überhaupt untersucht werden. Auch wenn ein Testkandidat ein fehlerfreies aber syntaktisch schlichtes Konstrukt produziert bleibt es ungeklärt, ob er die maximale Punktzahl 4 erzielt, da der oberste Deskriptor derartige Fälle bereits im Vorfeld nicht explizit ausschließt. Das Kriterium der Fairness käme jedoch damit ins Spiel, wenn ein weiterer Testkandidat mit den Formulierungen und dem Gebrauch komplexerer und eloquenterer Elemente, aufgrund des Risikos der falschen Anwendung zum Beispiel den Leseprozess des Rezipienten behinderte.[113] Dieser Testkandidat müsste den Kriterien entsprechend und verglichen zu der schlichteren Lernerproduktion im besten Fall 3 Punkte bekommen. Eine mangelhafte Fairness würde sich durch nicht zu vergleichende komplexe und weniger komplexe strukturelle Strickmuster geschriebener Texte äußern und demnach wäre die Validität dieses Kriteriums oder der Bewertung selber nicht gegeben.

112 Im 6. Kapitel soll auf die Kombination verschiedener Bereiche in diesem Kriterium eingegangen werden.

113 Das Gegenüberstellen verschiedener Lernerreaktionen soll hier nur zur Veranschaulichung für die Anwendung der Deskriptoren dienen. Natürlich stehen Sprachstandsprüfungen des Goethe-Instituts im Zeichen der Kriteriums- und nicht der Normorientierung

5.1.3 Diskussion von Lernerreaktionen auf die Aufgabenstellung und deren Originalbewertungen

Nachdem der Kriterienkatalog diskutiert worden ist, sollen im Folgenden zwei Originalbewertungen des Goethe-Instituts angeführt werden, indem und unter Berücksichtigung der bereits angeführten Schwachstellen und Kritikpunkte Stellung von mir genommen werden soll.

Lösungsvorschlag A **202 Wörter**

Große Mehrheit der Deutschen für strengere Kindererziehung

Sehr geehrte Damen und Herren,

Mit großem Interesse habe ich die Meldung über Erziehung der Kinder gelesen. Ich freue mich sehr, daß dieses Thema heute so aktuell ist. Meiner Meinung nach, ist es nicht nur interessant, sondern auch ein großes Problem für die Eltern heute. Deswegen möchte ich etwas dazu schreiben.

Die wichtigste Rolle bei der Erziehung der Kinder spielen die Eltern. Manche Probleme in der Zukunft, zum Beispiel die schlechte Benehmung in der Schule oder ein schlechter Umgang mit den anderen Kindern kommen aus der Kindheit.

Die Atmosphäre in der Familie spielt eine große Rolle. Ich finde, daß es gut ist, daß die Kinder früher strenger erzogen wurden. Sie bekamen von der Familie viel mehr, als heute und waren auch in der Zukunft selbstständiger.

Ich bin einverstanden, daß die Kinder beim Einkaufen den Eltern helfen sollten. Und die Eltern sollten unbedingt auf das Studium der Kinder achten. Ich glaube auch, daß sehr gute Idee ist, Kinder in eine Schuluniform zu stecken und selbstverständlich das Kaugummikauen in der Schule untersagen. Die Eltern heute beschäftigen sich sehr mit ihrer Arbeit, aber sie sollten mehr Zeit für die Kinder haben.

Mit freundlichen Grüßen,

Maria K.

Diese Leistung erzielte ein Ergebnis von 12,5 von maximal zu erreichenden 15 Punkten. Im Trainingsmaterial für Prüfende für das B2-Zertifikat lautet der Kommentar für diese Lernerproduktion:[114] *(...) eine gute Leistung. (...) schreibt einen klar gegliederten Brief. (...) beherrscht Lexik und Grammatik gut, auch wenn (...) beim Aufbau noch an den Vorgaben der Aufgabe „entlanggehangelt" [wird].* Wie sich die 12,5 Punkte bei dieser Lernerproduktion zusammen setzen und wie dies kommentiert ist, soll zunächst angeführt und im Folgenden diskutieren werden:

114 Goethe-Zertifikat B2: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707. S. 11

Kriterium	Kommentar	Bewertung
Inhaltliche Vollständigkeit	Die Textlänge ist ausreichend. Alle vier Inhaltspunkte sind behandelt, jedoch zwei nur sehr knapp, deswegen gibt es einen halben Punkt Abzug	2,5 Punkte
Textaufbau und Kohärenz	Der Text besitzt eine gute Einleitung und ist durchgängig flüssig lesbar. Kaum verknüpfte, einzelne Absätze, deshalb Entscheidung für drei Punkte.	3 Punkte
Ausdrucksfähigkeit	Es bestehen noch Unsicherheiten im Sprachgefühl. Sie übernimmt Ausdrucksweisen aus dem Text, ohne sie an den Stil ihres Textes anzupassen: *in eine Schuluniform zu stecken, untersagen*. Außerdem bestehen Unregelmäßigkeiten, z.B. *Benehmung, Studium* ist falsch eingesetzt. Deswegen nach Diskussion einen Punkt Abzug.	3 Punkte
Korrektheit	Wenige Fehler, die beim Lesen kaum auffallen. Nach Diskussion Entscheidung für vier Punkte, da die Gesamtfehlerzahl gering ist.	4 Punkte

Tabelle 25: Originalbewertung einer B2 - Produktion

In der vorangegangenen Diskussion der Bewertungskriterien des B2-Zertifikats sind viele Fragen gestellt worden, die sich hier nun beantwortet lassen können.

Was das Kriterium der inhaltlichen Vollständigkeit und dessen Bewertung im vorliegenden Fall anbelangt, so erscheint der Kommentar mit dem zuständigen Deskriptor nicht kompatibel zu sein. Diese Lernerproduktion hat zwar *alle vier Inhaltspunkte behandelt, jedoch zwei nur sehr knapp*. Dennoch erfolgt an dieser Stelle lediglich ein Abzug von 0,5 Punkten. Dieser Zuordnung d.h. 2,5 Punkten nach müssten statt zwei drei Inhaltspunkte schlüssig und angemessen dargestellt worden sein. Man sieht bereits an dieser Stelle, dass Rater genügend Freiraum innerhalb der definierten Deskriptoren haben, diese subjektiv zu besetzen. Was die Textlänge und ihre Dichte anbelangt, habe ich bereits unter 5.1.2.1 Stellung dazu genommen. Dass die hier als ausreichend befundene Textlänge jedoch zu ¼ aus der Einleitung besteht, die wiederum in diesem Kriterium keine Anwendung bezüglich des Textsortenwissens in der Bewertung findet, scheint paradox. Dies findet im zweiten Kriterium jedoch raterintern Berücksichtigung, ohne dass in den Deskriptoren jedoch davon die Rede ist, ob der Text eine gute oder weniger gute Einleitung besitzt.

Obwohl diese Lernerproduktion als *durchgängig flüssig lesbar* befunden wurde, finden sich laut der Originalbewertung *kaum verknüpfte, einzelne Absätze*. Dennoch wird diese Feststellung entgegen der Definition des Deskriptors *liest sich noch flüssig* mit der hier angesetzten Punktevergabe (3 Punkte) honoriert. Der Kommentar für die Bewertung des Ausdrucksvermögens scheint 3 Punkte aufgrund der Formulierung nicht zu bestätigen. Das Existieren von *Unsicherheiten im Sprachgefühl* und die *Übernahme von*

Ausdrucksweisen aus der Legende werden dem zweiten Deskriptor des Kriteriums zugerechnet, welches von *guter und angemessener* Ausdrucksfähigkeit ausgeht. Die im Kommentar eher als negative Elemente definierten Eigenschaften werden dennoch einer positiven Kann-Beschreibung zugeordnet. Was das Korrektheitskriterium anbelangt, deckt sich der Bewertungskommentar mit der Definition des Deskriptors, dem die Korrektheit vorliegender Lernerproduktion zugeordnet wurde.[115]

Es muss betont werden, dass es hier in erster Linie nicht um das Lernerprodukt selbst geht, sondern erstrangig darum, wie die definierten Bewertungskriterien samt ihren Deskriptoren stabil für das Ermitteln der Schreibkompetenz eines Lerners eingesetzt, benutzt und genutzt werden können. Das genau ist der Schlüssel, um die Validität der Bewertungskriterien zu ermitteln. Diese Lernerproduktion ist zweifelsohne eine sehr schöne Leistung. Die Frage, die sich an dieser Stelle jedoch stellt, ist, ob diese unter Berücksichtigung vorliegender Bewertungsskala immer als die gleiche sehr schöne Leistung bewertet würde.

Lösungsvorschlag B **241 Wörter**

Sehr geehrte Damen und Herren,

heute habe ich im Internet euere folgende Meldung gelesen und ich möchte euch sagen was ich davon halte. Ich möchte das beurteilen als eine reife Frau, welsche wird auch in der Zukunft Kinder haben.

Der grösste Einfluss auf die Kinder haben natürlich die Eltern und seine Erziehung, aber auch der Bekanntenkreis und die Lehrer, welsche geben die Kinder fast jeden Tag die Unterrichten. Sie geben den Kindern eine Persönlichkeitform.

Die Kinder früher waren strenger gezogen als ehute zu zeit. Das ist kein Wunder, das war ganz anderes Zeit, das Leben war nicht so weit technologisch, die Leute haben auch anders gedacht, sie konnten nicht alles haben was für uns ganz normal ist. Die Kinder haben auch nicht so viel Spielzeug wie jetzt und Unterhaltung.

Jede Generation bringt was neues, neue Erfahrungen, ist auch dadurch intelligenter, hat mehr Toleranz und Verständnis für andere Menschen. Ich finde, dass die Kinder zusammen mit den Eltern aufräumen und einkaufen sollen, dass/damit sie später sich richtig im Leben finden könnten. Die Schularbeiten sollen auch regelmäßig gemacht werden und was das Kaugummikauen betrıfft, ich finde korrekt nur wenn die Kinder Pause (in Unterricht) haben, sie dürfen dann Kaugummi kauen.

Ich kann allen Eltern nur vorschlagen, dass sie viel mit den Kindern sich unterhalten sollen, über Probleme reden und dem Kind das Verständnisgefühl geben. Sie müssen wissen, dass, sie ein Freund auch zu Hause habe.

Mit freundlichen Grüßen

Anna D.

115 Obwohl die Kriteriumskonstellation meines Erachtens nicht angemessen ist, so gilt dennoch festzuhalten, dass die Bereiche Morphologie, Syntax, Orthografie und Interpunktion bestimmten und unabänderlichen Regeln unterliegen. Dies allein müsste die Validität in diesem Bereich gewährleisten.

Im Prüfertraining B2 wird diese Leistung als ausreichend eingestuft und mit 9 von 15 Punkten bewertet. Es wird im Kommentar sogar darauf verwiesen, dass die Textsorte Leserbrief erkennbar eingehalten wurde, obwohl dies nirgends in der Bewertungsskala berücksichtigt zu sein scheint, wie bereits in der Diskussion der einzelnen Kriterien und ihren Deskriptoren angemerkt worden ist. Das kommentierte Bewertungsraster für diese Lernerproduktion setzt sich folgendermaßen zusammen:[116]

Kriterium	Kommentar	Bewertung
Inhaltliche Vollständigkeit	Textlänge ist mit 241 Wörtern mehr als ausreichend. Alle vier Inhaltspunkte sind angemessen dargestellt.	3 Punkte
Textaufbau und Kohärenz	Der Text besitzt eine adäquate Einleitung und ist als Leserbrief gestaltet. Er lässt sich flüssig lesen, jedoch beeinträchtigen die sprachlichen Fehler die Lesbarkeit. Es gibt einen Punkt Abzug, da die einzelnen Abschnitte nicht gut verknüpft sind	3 Punkte
Ausdrucksfähigkeit	Der Wortschatz ist stellenweise angemessen und gut. Vereinzelt werden falsche oder unpassende Ausdrücke verwendet (reife Frau; Persönlichkeitform; das Leben war nicht so weit technologisch). Das Verständnis ist durch diese Fehler zwar nicht beeinträchtigt, dennoch Entscheidung für zwei Punkte.	2 Punkte
Korrektheit	Häufige Fehler in Morphologie und Syntax, die vereinzelt das Verständnis behindern (welsche geben die Kinder fast jeden Tag die Unterrichten). Das Verständnis bleibt noch erhalten, aber auf Grund der Fehlerzahl Entscheidung für einen Punkt.	1 Punkt

Tabelle 26: Originalbewertung einer B2 - Produktion

Was die Bewertung für die inhaltliche Länge anbelangt, ist bezüglich der Anwendung der Deskriptoren in diesem Fall nichts auszusetzen. Tatsächlich wurden in dieser Lernerproduktion die gestellten Inhaltspunkte entsprechend bearbeitet und abgedeckt. Im Kommentar wird zudem dokumentiert, dass die Textlänge mehr als ausreichend ist. In der Diskussion dieses Kriteriums habe ich bereits darauf hingewiesen, dass nichts bezüglich der Dichte eines Textes vermerkt wird. Offensichtlich ist dies dennoch ein latentes Unterkriterium in dieser Kategorie. In den Deskriptoren wird lediglich Bezug auf die schlüssige und angemessene Darstellung der Inhaltspunkte genommen. Was das zweite Kriterium anbelangt, so findet man auch im Kommentar dieser Lernerproduktion den Vermerk bezüglich der Textsorte, was aber im Sinne der APA, dass alles dokumentiert sein muss, in den Deskriptoren nicht zu finden ist. Die Bewertung liegt hier

116 Goethe-Zertifikat B2: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707. S. 21

im zweiten Deskriptor, der drei Punkte dafür vergibt, dass sich der Text *noch flüssig liest*. Das wird im Kommentar damit begründet, dass *sprachliche Fehler die Lesbarkeit beeinträchtigen*. Der Begriff Fehler tritt aber, wie bereits erwähnt wurde, lediglich im Kriterium Korrektheit auf. Demnach dürften die sprachlichen Fehler an dieser Stelle keine Berücksichtigung finden. Hier soll es ausschließlich um die Textgliederung und die durch Satz verknüpfende Elemente geschaffene Kohärenz gehen. Der Punktabzug wird dennoch auf der Basis sprachlicher Fehler kommentiert und gerechtfertigt. Trotzdem ist es nicht offensichtlich, worauf es bei der so genannten Lesbarkeit bzw. dem in den Deskriptoren definierten Lesefluss ankommt, denn diese Beeinträchtigung wird als durch Fehler verursacht kommentiert, welche aber in den Deskriptoren dieses Kriteriums weder definiert noch erwähnt wird. Im Kriterium der Ausdrucksfähigkeit werden nur zwei Punkte vergeben, obwohl die vereinzelt auftretende falschen oder unpassenden Ausdrücke laut Kommentar das Verständnis nicht beeinträchtigen. Der *stellenweise angemessene und gute Ausdruck* und die *vereinzelte Verwendung falscher Ausdrücke* finden sich im dritten Deskriptor des Ausdruckskriteriums *stellenweise gut und angemessen* wieder, obwohl dort nichts hinsichtlich des Verständnisses definiert ist. Dennoch besteht in diesem Fall eine Kompatibilität zwischen der Bewertungsskala und ihrer Anwendung im Ernstfall.

Obwohl sich Fehler ganz objektiv betrachtet bestimmen lassen können, muss dies anhand der Anwendung des Kriteriums Korrektheit revidiert werden. Vorliegende Lernerproduktion erzielt lediglich einen Punkt. Kommentiert wird, dass häufige Fehler (...) das Verständnis vereinzelt behindern (...), das aber noch erhalten bleibt. Demzufolge existiert kein eindeutiger Deskriptor. Während der 2-Punkte-Deskriptor definiert, dass *einige Fehler, die den Leseprozess stellenweise behindern,* sind es beim nächsten Deskriptor (1 Punkt) *unzählige Fehler, die das Verständnis erheblich stören*. Begründet wird die Vergabe eines Punktes mit häufigen Fehlern. Es sind aber dennoch nicht unzählige, wie dieser Deskriptor definiert. Unbeachtet ist außerdem, wie Wiederholungsfehler gewertet werden (z.B. welsche), um die tatsächliche Fehlerfrequenz zu ermitteln. Nirgendwo ist dies dokumentiert. Auch an dieser Stelle kann erneut festgestellt werden, dass die Definitionen von Bewertungskriterien und Deskriptoren nicht vollkommen objektiviert werden können, um alle Fälle darauf zu beziehen. Wäre ein Rater strenger in seiner Bewertung, dann bekäme diese Lernerproduktion möglicherweise 0 Punkte, da die unzähligen Fehler, sein Textverständnis unmöglich machten. Somit würde die gesamte Aufgabe des schriftlichen Ausdrucks mit Null Punkten bewertet werden. Zu hinterfragen ist genau an dieser Stelle, wie sinnvoll eine derartige Festlegung im Rahmen einer analytischen Bewertung schließlich ist.

Schließlich soll eine als nicht ausreichend befundene Leistung angeführt und diskutiert werden.

Lösungsvorschlag C 158 Wörter

Sehr geehrte Damen und Herren,

heute habe ich Ihre eine Meldung im Internet gelesen. Thema:“Große Mehrheit für strengere Kindererziehung“ interessiert mich sehr, weil ich selbst fünf Kinder habe. Auf meine Meinung um Kinder müssen nicht die Eltern kümmern, aber die Schule auch.

Das bedeutet, dass Kindern im Grunde frei sollgen sein. Im Moment bin ich mit den derzeitigen Erziehungsmethoden zufrieden, genauso wie 31 Prozent der Deutschen. 95 Prozent sprechen sich dafür aus, dass Kinder Pflichten wie Aufräumen und Einkaufen erfüllen sollten. Ich finde, dass es nicht besonders toll sind. Das kommt später, wann sie schon eine Lebenserfahrung bekommen. Ich bin auch gegen regelmäßige Schularbeite für Kindern, weil sie im disen Jahren andere Aufgabe haben sollen. Die Kinder in eine Schuluniform zu stecken? Was könnte schlimmste sein? Für eine bessere Kindererziehung möchte ich gern weitere Vorschläge machen. Kindern sollen selbst seine Lebensstiell wahlen. Die Eltern und die Schule mussen nur Kindern helfen.

Mit freundlichen Grüßen

Lidiya S.

Der globale Kommentar bezieht sich zunächst auf die unzureichende Leistung, die unter das B2-Niveau eingestuft wird. Obwohl der Brief klar gegliedert zu sein scheint, ist der Text mit 158 Wörtern zu kurz (verlangt werden in der Aufgabenstellung mindestens 180 Wörter). Zudem werden die einzelnen Inhaltspunkte nicht angemessen behandelt. Zur Verdeutlichung werde ich den Bewertungskommentar für diese Lernerproduktion anführen:

Kriterium	Kommentar	Bewertung
Inhaltliche Vollständigkeit	Der Text ist zu kurz. Die vier Inhaltspunkte sind zwar knapp behandelt, jedoch wiederholt die Teilnehmerin vieles aus der Aufgabenstellung wortwörtlich. Entscheidung für einen Punkt, da keiner der Inhaltspunkte schlüssig und angemessen dargestellt ist.	1 Punkt
Textaufbau und Kohärenz	Der Text liest sich trotz der Fehler noch flüssig. Nach Diskussion Entscheidung für zwei Punkte, da Lexik und Grammatik bei den anderen Kriterien bewertet werden.	2 Punkte

Ausdrucksfähigkeit	Die Teilnehmerin zeigt bei der Wortschatzbeherrschung kaum Eigenleistung, sondern übernimmt viel aus der Aufgabenstellung wie *mit den derzeitigen Erziehungsmethoden zufrieden; 95 Prozent sprechen*...Selbst Ausdrücke wie meiner Meinung nach sind fehlerhaft. Entscheidung für einen Punkt, da das Wortschatzspektrum im ganzen Text nicht angemessen ist.	1 Punkt
Korrektheit	Es treten gehäufte Fehler auf (auf meine Meinung; dass es toll sind; im disen Jahre), der Gesamtzusammenhang bleibt undeutlich, das Verständnis ist stellenweise gestört.	1 Punkt

Tabelle 27: Originalbewertung einer B2 - Produktion

In der Diskussion der Bewertungsskala für das B2-Zertifikat ist darauf aufmerksam gemacht worden, dass in den Kriterien nichts über die erforderliche Mindesttextlänge vermerkt ist. Lediglich auf dem Aufgabenblatt wird unter Hinweise angemerkt, dass eine Textproduktion von ca.180 Wörtern verlangt wird. Punktabzüge für das Nicht-Erreichen dieses Textvolumens werden meines Wissens offiziell nirgends dokumentiert. Die Frage der Textlänge findet dennoch *intern* im Kriterium der inhaltlichen Vollständigkeit Beachtung. In der hiesigen Punktevergabe wird berücksichtigt, dass die Aufgabenstellung als Hilfsmittel gebraucht wurde, um die Inhaltspunkte zu bearbeiten. Verwiesen wurde bereits darauf, dass diese Möglichkeit nicht unbeachtet gelassen werden darf. In den Deskriptoren des Kriteriums der inhaltlichen Vollständigkeit kann sich jedoch kein Indiz dafür finden, dass dieser Umstand Berücksichtigung findet. Für die inhaltliche Vollständigkeit bekam diese Lernerproduktion einen Punkt, wobei *Inhaltspunkte nur ansatzweise behandelt sind, an mehreren Stellen unklar*. Dieser Deskriptor besagt aber weder etwas über Textlänge, die nicht eingehalten worden ist, noch über Vermeidungsstrategieanwendungen anhand der Aufgabenstellung.

Wenn im Korrektorenkommentar notiert ist *der Text liest sich trotz der Fehler noch flüssig*, dann müsste sich dies mit der Deskriptorendefinition *liest sich noch flüssig* decken. Aus der Punktevergabe wird aber deutlich, dass trotz allem nicht dieser Deskriptor für das Bewerten des Textaufbaus und der Kohärenz herangezogen wurde, sondern der darunter liegende, der *von einem stellenweise guten Aufbau, der an einigen Stellen sprunghaft ist* ausgeht. Meines Erachtens wurde an dieser Stelle ganz subjektiv bewertet, ohne den Deskriptorendefinitionen Folge zu leisten.[117] Interessant ist, aus welchem Grund bei diesem Kriterium vermerkt wird, dass lexikalische und grammatikalische Fehler keine Berücksichtigung finden, zumal diese Bereiche anderen Kriterien unterliegen.

117 In Kapitel 6 wird bezüglich des Raterverhaltens und der Bewertungskriterien ausführlich eingegangen werden.

Diese Lernerproduktion müsste für das Kriterium *Textaufbau und Kohärenz* gemäß des Bewertungskatalogs insgesamt drei Punkte erzielen. Was die Ausdrucksfähigkeit betrifft, so wird im Kommentar erneut auf den Zugriff vorgefertigter Strukturen aus der Aufgabenstellung hingewiesen. Es besteht an dieser Stelle also die Gefahr der Doppelsanktionierung, denn bereits im Kriterium der inhaltlichen Vollständigkeit wird die Vergabe eines Punktes auch dadurch begründet. Auch beim letzten Kriterium scheinen sich Zuschreibungen anderer Kriterien zu finden. Der *Gesamtzusammenhang bleibt undeutlich* gehört in das Kriterium *Textaufbau und Kohärenz*. Der Anspruch an das Korrektheitskriterium soll lediglich darin bestehen, Morphologie, Syntax, Orthografie und Interpunktion auf Fehler hin zu untersuchen. Was gehäufte Fehler sind, müsste wie im Vorfeld diskutiert worden ist, zunächst einmal quantitativ bestimmt werden. Ob die laut Kommentar *gehäuften Fehler* mit der Definition im angesetzten Deskriptor, wo von *unzähligen Fehlern* die Rede ist, gleichzusetzen sind, müsste beantwortet werden.

Fazit

Die Gewichtung der einzelnen Kriterien der Bewertungsskala für das B2-Zertifikat wurde bereits insofern erläutert, dass die vier Kriterien unterschiedlich stark gewichtet sind. Während die inhaltliche Vollständigkeit insgesamt 20% ausmacht, decken die Kriterien Textaufbau und Kohärenz, Ausdrucksfähigkeit und Morphologie mit jeweils 26,66 % die weitere Punktepalette ab. Was die einzelnen Kriterien anbelangt, so müssen abschließend nochmals einige Punkte angesprochen werden.

Das Kriterium der inhaltlichen Vollständigkeit spricht in seinem letzten Deskriptor von einer Themaverfehlung. Eine Themaverfehlung im schriftlichen Ausdruck müsste demnach zu Null Punkten führen. Wie bereits angeführt, entnimmt man diesem Kriterium auch nicht den Fall der nicht angemessenen Textlänge von 180 Wörtern und die entsprechenden Konsequenzen daraus. Ob bei Nichteinhaltung der Wörtergrenze letztlich Punkte abgezogen werden, kann dem Bewertungsraster nicht entnommen werden. Vermerkt ist darüber hinaus weder etwas in den Prüferblättern noch in der Prüfungsordnung. Das zweite Kriterium Textaufbau und Kohärenz wurde in seiner Struktur bereits ausführlich diskutiert und bezüglich seiner subjektiv geprägten Deskriptoren kritisiert. Was die letzten beiden Kriterien Ausdrucksfähigkeit und Korrektheit anbelangt, so muss sichergestellt werden, dass Fehler entsprechend zugeordnet werden und das nur einmal. Auch zu diesem Punkt kann man in den verschiedensten Dokumenten des Goethe-Instituts keine Rückmeldung darüber bekommen, wie Rater damit umzugehen haben. Ob etwas als syntaktisch inkorrekt, als ausdrucksmäßig inkorrekt oder als inkohärent von Ratern verstanden wird, ist nicht eindeutig. Hätte man beispielsweise ein präpositionales Verb, das im falschen Kasus oder mit der falschen Präposition verbunden worden ist, stellt sich die Frage welchem Kriterium diese nicht normgerechte Verwendung zugeschrieben würde: dem Kriterium Ausdrucksfähigkeit, dem Kriterium Korrektheit oder wohl unbewusst beiden? Diesen Fall gilt es für die Rater insofern auszuschließen, dass die so genannte Doppelsanktionierung nicht zum Tragen kommt und ein schriftliches Konstrukt dadurch schlechter bewertet würde als nötig. Das Goethe-Institut vermerkt diese Problematik jedoch explizit in seinem Skript für das Prüfertraining des C1-Niveaus, „dass es bei auftretenden Fehlern wichtig ist, diese dem richtigen Kriterium zuzuordnen, um einen Doppelpunktabzug für ein und denselben Fehler zu vermeiden“[118].

118 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. 090707. S. 9

In der Kompetenz schriftlicher Ausdruck auf B2-Niveau kann man maximale 15 Punkte erreichen, die auf die vier vorgestellten Kriterien verteilt sind. Dabei wird die schriftliche Lernerproduktion von zwei Ratern unabhängig voneinander korrigiert. In der Prüfungsordnung des Goethe-Instituts für das B2-Zertifikat wird vermerkt, dass sich die Rater im Falle abweichender Ergebnisse auf ein Ergebnis einigen müssen. Kann kein gemeinsamer Nenner unter den Ratern gefunden werden, so entscheidet der Prüfungsverantwortliche eventuell mit einer Drittkorrektur.[119] Das führt zu der Annahme, dass man in diesem Sinne versucht, die Interraterreliabilität durch einen weiteren Rater zu sichern.

Der schriftliche Ausdruck setzt sich aus den Punkten der ersten und zweiten Aufgabe zusammen, welche bei 25 Punkten angesetzt sind. Der zweite Teil des schriftlichen Ausdrucks ist eine geschlossene Aufgabenstellung mit 10 zu erreichenden Punkten, die mit einem Lösungsschlüssel korrigiert wird. Bei der Korrektur der Tests wird laut Prüferblättern das „Gesamtergebnis des Prüfungsteils *Schriftlicher Ausdruck* auf das Formblatt *Gesamtergebnis* übertragen. Halbe Punkte werden nicht aufgerundet (...)“.[120] Nach § 15 werden die Prüfungsleistungen in Form von Punkten und Noten dokumentiert[121] und erfordern zum Bestehen der gesamten Prüfung die Summe aus mindestens 45 Punkte der schriftlichen und 15 Punkte der mündlichen Prüfung, folglich 60 von maximal 100 zu erreichenden Punkten. Was unter § 16 der Prüfungsordnung bezüglich der Zertifizierung und dem Prädikat als Notenskala angeführt wird, soll an dieser Stelle angeführt werden:

100 – 90 Punkte = sehr gut
89,5 – 80 Punkte = gut
79,5 - 70 Punkte = befriedigend
69,5 –60 Punkte = ausreichend
unter 60 Punkte = nicht bestanden

Tabelle 28 : Noten- und Prädikatenskala aus der Prüfungsordnung[122]

Die Prüfungsordnung ist laut des Vermerks darauf am 05.07.07 definiert worden. Die Noten- und Prädikatenskala beinhaltet die Punktestreuung aller vier Fertigkeiten (LV, HV, SA, MA) der neuen B2-Prüfung des Goethe-Instituts. Die Bestehensgrenze liegt bei 60% (60 von 100 Punkten). Es wird nichts darüber ausgesagt, ob jede Kompetenz für sich eine bestimmte Punktzahl erfordert oder nicht. Folglich werden die Leistungen in jeder Teilkompetenz aufsummiert und müssen insgesamt mindestens 60 Punkte erreichen.[123] Bei Tests wird im Allgemeinen eine Mindestpunktzahl angegeben, um den Status *bestanden* zu erwerben. Es stellt sich aber die Frage, wie bei Sprachzertifizierungstests, die ja aus verschiedenen Kompetenzbereichen bestehen, die erforderte Leistung oder auch der Erwartungswert an die Testkandidaten insgesamt fair bewertet werden kann. Die Festlegung der Bestehensgrenze von 60% ist zunächst einmal interessant und ein Thema für sich. Denn würde ein Test kalibriert, d.h. würde Aufgabe A durch andere Aufgaben verschiedener Schwierigkeitsgrade ersetzt, bestünde dennoch die Bestehensgrenze stabil bei 60%.

119 Prüfungsordnung Goethe-Institut. B2-Zertifikat. S. 11

120 Prüfungsordnung Goethe-Institut. B2-Zertifikat S. 11

121 http://www.goethe.de/lrn/prf/pro/b2_pruefungsordnung.pdf

122 http://www.goethe.de/lrn/prf/pro/b2_pruefungsordnung.pdf

123 http://www.goethe.de/lrn/prf/pro/b2_pruefungsordnung.pdf

5.2 Das C1-Zertifikat des Goethe-Instituts

Dieses Niveau ist die erste Stufe der kompetenten Sprachverwendung, welche mit dem C2-Niveau abgerundet wird. Kennzeichnend hierfür ist, dass ein breites Spektrum sprachlicher Mittel vorhanden ist, sodass generell betrachtet eine flüssige und spontane Kommunikation ermöglicht wird. Im Gegensatz zum darunter liegenden B2-Niveau liegt das Gewicht hier nun mehr auf dem Aspekt größerer Flüssigkeit und Komplexität (GER 2001:44). Das C1-Niveau wird in den Kann-Beschreibungen der Globalskala des GER im Sinne der Thematik dieser Dissertation wie folgt definiert:

- **„Kann sich spontan und fließend ausdrücken, ohne öfter deutlich erkennbar nach Worten suchen zu müssen".**
- **„Kann die Sprache im gesellschaftlichen und beruflichen Leben oder in Ausbildung und Studium wirksam und flexibel gebrauchen".**
- **„Kann sich klar, strukturiert und ausführlich zu komplexen Sachverhalten äußern und dabei verschiedene Mittel zur Textverknüpfung angemessen verwenden".**

Diesbezüglich wird das Prüfungsziel der Lerner „Bestehen der Prüfung" vom Goethe-Institut dadurch definiert, „dass (...) die überregionale deutsche Standardsprache geläufig ist,(...) dass sie die deutsche Sprache sicher verwenden und ihre persönlichen Belange im privaten, gesellschaftlichen, akademischen und beruflichen Leben adäquat ausdrücken können"[124]. Diese neue C1-Prüfung des Goethe-Instituts besteht im produktiven schriftlichen Teil aus zwei Aufgabenteilen. Der offene Aufgabentypus, der in dieser Arbeit im Vordergrund steht, behandelt auf diesem Niveau die Beschreibung einer Grafik, wobei zwei alternative Aufgaben zur Auswahl bereit stehen. Die zweite Aufgabe im schriftlichen Ausdruck des C1-Zertifikats ist eine C-Test ähnliche Aufgabe. Hier soll der Testkandidat sein Sprachvermögen anhand der zu füllenden Lücken verdeutlichen. Ich werde mich im Sinne dieser Dissertation dem ersten Teil schriftlicher Lernerproduktionen widmen, bei der der Testkandidat in der Lage sein sollte, „sich über komplexe Sachverhalte schriftlich klar und strukturiert auszudrücken und ein dem Leser angemessenes Register zu wählen"[125]. Der erste Teil des schriftlichen Ausdrucks wird in der Prüfungszielbeschreibung des Goethe-Instituts folgendermaßen dargestellt:[126]

Aufgabe	Prüfungsziel	Textsorte/Textstruktur	Aufgabentyp	Punkte
1	Produktion: Informationen referieren, etwas berichten/ vergleichen, Meinung äußern	Schriftliche Äußerung zu einem Thema	Freies Schreiben nach Vorgabe von 5 Leitpunkten	20

Tabelle29: Prüfungszielbeschreibung des schriftlichen Ausdrucks im C1–Zertifikats des Goethe-Instituts

124 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 8

125 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 8

126 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 26

Um eine Aufgabe schriftlichen Ausdrucks für einen Test zu erstellen, muss als erstes die Basis dafür geschaffen werden. Die vom GER definierten und übernommenen Kann-Beschreibungen sollen demnach als Referenzrahmen fungieren, um in diesem Sinne dem Prüfungsziel entsprechend *die Fähigkeit zum Verfassen eines schriftlichen Textes zu einem vorgegebenen Thema zu überprüfen.*[127] Dabei basiert die Testerstellung des Goethe-Instituts nicht auf dem handlungsorientierten kommunikativen Ansatz des GER, sondern auf dem Modell der Kommunikationsfähigkeit von Bachman und Palmer (1996).[128] Für die Itemerstellung jedoch hat sich das Goethe-Institut auf die vom GER nicht empirisch kalibrierten Beispielsskalen für sprachliche Aktivität und Strategien gestützt (GER 2001:67ff):[129]

- **<u>Schriftliche Produktion allgemein</u>**

  „Kann klare, gut strukturierte Texte zu komplexen Themen verfassen und dabei die entscheidenden Punkte hervorheben, Standpunkte ausführlich darstellen und durch Unterpunkte oder geeignete Beispiele oder Begründungen stützen und den Text durch einen angemessenen Schluss abrunden".

- **<u>Berichte und Aufsätze schreiben</u>**

  „Kann klare, gut strukturierte Ausführungen zu komplexen Themen schreiben und dabei zentrale Punkte hervorheben".

  „Kann Standpunkte ausführlich darstellen und durch Unterpunkte, geeignete Beispiele oder Begründungen stützen".

- **<u>Schriftliche Interaktion allgemein</u>**

  „Kann sich klar und präzise ausdrücken und sich flexibel und effektiv auf die Adressaten beziehen".

In der im folgenden vorgestellten Aufgabenstellung für den schriftlichen Ausdruck soll der Prüfungskandidat zudem nachweisen, ob er in der Lage ist, sich innerhalb eines breiten thematischen Spektrums (z. B. persönliche Daten und Verhältnisse, Wohnen/Umwelt, tägliches Leben/Arbeit, Freizeit/ Unterhaltung, Reise, Beziehung zu anderen Menschen/Kultur/Tradition, Gesundheit und Hygiene, Erziehung/ Ausbildung/ Lernen, Konsum/Handel, Ernährung, Dienstleistungen, Orte, Sprache/Kommunikation, Klima[130]) „ausführlich, kohärent sowie partner- und situationsadäquat schriftlich zu äußern"[131].

127 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 27

128 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 14

129 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 11

130 Basierend auf die im GER vorgestellten Kategorien im Rahmen der Klassifikation des Threshold Levels 1990 übernimmt das Goethe-Institut für die Überprüfung sprachlichen Handelns mögliche Themengebiete als Prüfungsinhalt. Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 19

131 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 27

5.2.1 Aufgabenstellung für den schriftlichen Ausdruck im C1-Zertifikat des Goethe-Instituts[132]

Dauer: 65 Minuten

Schreiben Sie eine Stellungnahme zu folgenden Punkten:

- **Freizeitverhalten der älteren Generation**
- **Unterschiede zwischen Jungen und Mädchen**
- **Vergleich der Ergebnisse mit dem Heimatland**
- **Persönliche Freizeitaktivitäten**
- **Ergebnisse der Grafik**

Hinweise:

Bei der Beurteilung wird u. a. darauf geachtet,

– ob Sie alle angegebenen Inhaltspunkte berücksichtigt haben,

– wie korrekt Sie schreiben,

– wie gut Sätze und Abschnitte sprachlich miteinander verknüpft sind.

Schreiben Sie etwa 200 Wörter.

Goethe-Zertifikat C1 Prüfertraining 090707

132 Goethe-Zertifikat C1: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707. S. 6

In diesem Teil der C1-Prüfung steht das kommunikative Prüfungsziel im Mittelpunkt, das dadurch gekennzeichnet ist, Informationen zu referieren, etwas zu berichten/ vergleichen, zu informieren und eine persönliche Meinung zu äußern.[133] Der Prüfungskandidat hat 65 Minuten Zeit, wie oben links auf dem Aufgabenblatt vermerkt ist.[134] Der Input für die Bearbeitung der folgenden fünf Inhaltspunkte in der originalen Beispielaufgabe besteht in der angeführten Statistik in Form eines Balkendiagramms. Demnach wird das Verständnis eines diskontinuierlichen Inputs vorausgesetzt. Diese Prüfung richtet sich der Prüfungsordnung nach an Erwachsene und Jugendliche, wobei das Mindestalter bis zum Herbst 2008 auf 16 Jahre angesetzt war.[135]

Das zu bearbeitende Thema für die Prüfungskandidaten lautet *Freizeit der Jugend.* Dennoch haben nicht alle zu bearbeitenden Leitpunkte direkt mit der vorgelegten Statistik zu tun. Der Prüfungskandidat soll zudem auch Stellung zum Freizeitverhalten der älteren Generation beziehen. Das bedeutet prinzipiell, dass sich diese Aufgabe nicht ausschließlich mit dem Freizeitverhalten der Jugend beschäftigt, sondern dass das Themen- und Wortschatzspektrum durch eine kontrastive Gegenüberstellung erweitert wird. Der Bezug auf die Situation im eigenen Heimatland, stellt eine weitere kontrastive Facette dar. Nicht ganz eindeutig ist der Unterschied der Inhaltspunkte *Unterschiede zwischen Jungen und Mädchen* und *Ergebnisse der Grafik*. Ich führe dies an, denn es könnte sehr leicht dazu führen, dass beides schriftlich zu einem Punkt zusammen gefasst wird. Das könnte zur Folge haben, dass es zu Minderungen in der Punktzahl bei der inhaltlichen Bewertung käme, wenn ein Inhaltspunkt als nicht bearbeitet erkannt würde. Schließlich soll auch die eigene Person bezüglich ihrer Freizeitvorlieben und –aktivitäten im schriftlichen Konstrukt eingebracht werden.

Diese schreibproduktive Aufgabe des C1-Zertifikats zielt offensichtlich auf die Textsorte Aufsatz ab, obwohl dies nicht explizit dokumentiert ist. Der Arbeitsauftrag lautet lediglich: *Schreiben Sie eine Stellungnahme zu folgenden Punkten*. Unter Hinweise gilt all das, was unter 5.1.1 bezüglich der Aufgabenstellung des B2-Zertifikats erörtert und diskutiert wurde. Hier soll sich die Lernerproduktion jedoch auf 200 Wörter erstrecken, auch deshalb, weil sich im Gegensatz zum B2-Zertifikat die zu bearbeitenden Inhaltspunkte um einen erhöht haben. Zudem ist es natürlich viel anspruchsvoller, wenn man bedenkt, dass die aufgeführte Legende mathematischen Charakter hat und vom Prüfungsteilnehmer schriftlich erarbeitet werden muss. Demnach sind 200 Wörter eine nicht zu unterschätzende Anforderung, denn die Möglichkeit der Verwendung ganzer Einheiten aus der Legende ist durch die vorliegende Aufgabenform hier nicht mehr gegeben.[136] Gemäß des bereits angeführten Kernlehrplans in NRW sollen diskontinuierliche Texte Ende der Jahrgangsstufe 10 funktional eingesetzt werden können. Das heißt, dass sich das Mindestalter für die C1-Prüfung (16 Jahre), das bis zum Herbst 2008 galt, sich mit dem Alter der 10. Jahrgangsstufe deckt, wobei es sich hier allerdings um den primärsprachlichen Unterricht handelt. Diese Gegenüberstellung wurde bereits unter Kap. 3.6 zur Diskussion gestellt.

133 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 26ff.

134 Das Goethe-Institut macht lediglich einen Vorschlag hinsichtlich der Zeiteinteilung. Auch hier gilt wie beim B2-Zertifikat, dass der Prüfungskandidat die Zeit für den schriftlichen Ausdruck, der aus zwei Aufgaben besteht, individuell gestaltet.

135 Goethe-Zertifikat C1. Prüfungsordnung, Durchführungsbestimmungen Stand: 050707. § 1, S.1. Diese Alterbegrenzung entfällt ab Herbst 2008.

136 Somit wird Vermeidungsstrategien und Entlehnungen aus der Legende wenig Platz eingeräumt

Die Bewertung des schriftlichen Ausdrucks erfolgt nach den durch Deskriptoren definierten Bewertungskriterien. Dabei wird in der Prüfungsordnung darauf hingewiesen, dass Kandidatenbeispiele in den Trainingsmaterialien für Prüfende als Hilfestellung dienen sollen.[137] Zunächst soll aber die Bewertungsskala unabhängig davon separat erörtert und dokumentiert werden.

5.2.2 Bewertungskriterien für den schriftlichen Ausdruck im C1-Zertifikat des Goethe-Instituts

KRITERIUM I Inhaltliche Vollständigkeit	4 Punkte	3 Punkte	2 Punkte	1-0,5 Punkte	0 Punkte
• Inhaltspunkte schlüssig und angemessen dargestellt	Alle Inhaltspunkte	vier Inhaltspunkte	drei Inhaltspunkte	Ein bis zwei Inhaltspunkte bzw. alle Inhaltspunkte nur ansatzweise	Thema verfehlt
KRITERIUM II Textaufbau+Kohärenz	5 Punkte	4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Gliederung des Textes • Konnektoren, Kohärenz	Liest sich sehr flüssig	Liest sich noch flüssig	Liest sich stellenweise sprunghaft, und einige fehlerhafte Konnektoren	Aneinanderreihung von Sätzen fast ohne logische Verknüpfung	Über weite Strecken unlogischer Text
KRITERIUM III Ausdrucksfähigkeit	5 Punkte	4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Wortschatzspektrum • Wortschatzbeherrschung	Sehr gut und angemessen	Gut und angemessen	Stellenweise gut und angemessen	Begrenzte Ausdrucksfähigkeit, Kommunikation stellenweise gestört	Text in großen Teilen völlig unverständlich
KRITERIUM IV Korrektheit	6 Punkte	5-4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Morphologie • Syntax • Orthografie, Interpunktion	Nur sehr kleine Fehler	Einige Fehler, die das Verständnis aber nicht beeinträchtigen	Einige Fehler, die den Leseprozess stellenweise behindern	Häufige Fehler, die den Leseprozess stark behindern	Text wegen großer Fehlerzahl unverständlich

Tabelle 30: Bewertungskatalog für das C1–Zertifikat des Goethe-Instituts

137 Goethe-Zertifikat C1. Prüfungsordnung, Durchführungsbestimmungen. Stand: 050707. S. 6

Auch auf dieser Niveaustufe beinhaltet das analytische Bewertungsraster vier Bewertungskriterien:[138]

- Inhaltlichen Vollständigkeit
- Textaufbau und Kohärenz
- Ausdrucksfähigkeit
- Korrektheit

Jedes der vier Kriterien ist unterschiedlich gewichtet. Das als stärkstes Kriterium geltende ist das der Korrektheit, das sechs (6) von maximal zwanzig (20) zu erreichenden Punkten in Anspruch nimmt. Prozentual bedeutet dies eine Gewichtung von 30%. Gleichwertig in ihrer Gewichtung folgen die Kriterien Ausdrucksfähigkeit und Textaufbau/Kohärenz mit jeweils 25%. Das Kriterium der inhaltlichen Vollständigkeit macht den kleinsten prozentualen Teil aus (20%). Nicht jedes Kriterium ist also für die Bewertung *kompetenter Sprachverwendung* auf dem Niveau C1 gleichermaßen bedeutend. Dennoch besteht paradoxerweise wie in der Prüfungsordnung des B2-Zertifikats auch hier die Regel, dass wenn ein Testteilnehmer die 0-Punkte-Marke erreicht, die Punktzahl für diese Aufgabe innerhalb des Tests Null beträgt. Dies scheint mir insofern unfair, als die Kriterien zum einen nicht die gleiche Gewichtung innerhalb des Bewertungskatalogs und der Bewertung insgesamt haben und zum anderen bei einer „Themaverfehlung“ die Sprachkompetenz nicht existent erscheint. Ich werde im Folgenden die Kriterien separat dokumentieren, um ggf. Änderungsvorschläge zu initiieren.

138 Im Anschluss der C1-Diskussion werden die zwei Zertifikate B2 und C1 gegenübergestellt werden, da es viele Parallelen in den Kriterien und den Deskriptoren gibt. Das soll insofern diskutiert werden, dass man den Unterschied im Anspruch dieser Niveaus hinterfragt.

5.2.2.1 Kriterium: Inhaltliche Vollständigkeit

KRITERIUM I Inhaltliche Vollständigkeit	4 Punkte	3 Punkte	2 Punkte	1- 0,5 Punkte	0 Punkte
• Inhaltspunkte schlüssig und angemessen dargestellt	Alle Inhaltspunkte	vier Inhaltspunkte	drei Inhaltspunkte	Ein bis zwei Inhaltspunkte bzw. alle Inhaltspunkte nur ansatzweise	Thema verfehlt

Tabelle 31: Inhaltliche Vollständigkeit im C1–Zertifikat des Goethe-Instituts

Dieses Kriterium hat bis auf wenige Unterschiede den gleichen Aufbau hinsichtlich der Deskriptorendefinitionen wie das im bereits dokumentierten und diskutierten B2-Zertifikat. Auch hier wird die *schlüssige und angemessene Darstellung der Inhaltspunkte* der Mittelpunkt der Bewertung dieses Kriteriums. Diese nicht stringente und subjektive Definition wurde bereits in der Diskussion des B2-Zertifikats ausführlich erläutert und gilt auch für hiesigen Fall. Die Deskriptoren fallen bei diesem Kriterium zunächst jeweils um einen (1) Punkt ab (4-3-2) und beim vorletzten, dem vierten Deskriptor, tritt eine Kombination der Punktevergabe 1 und 0,5 auf. Diese basiert auf der Definition *ein bis zwei Inhaltspunkte bzw. alle Inhaltspunkte nur ansatzweise*. Folglich bedeutet dies, dass man einen (1) oder 0,5 Punkte dafür bekommt, wenn *ein bis zwei Inhaltspunkte schlüssig und angemessen* oder *alle Inhaltspunkte nur ansatzweise dargestellt sind*. Sehr vage ist meines Erachtens zunächst das Wort *bis* im ersten Teil dieser Definition. Was heißt denn *ein bis zwei*? Man sagt beispielsweise: Ich gehe ein bis zwei Mal in der Woche in die Universitätsbibliothek. Gehe ich immer nur ein Mal oder manchmal auch zwei Mal pro Woche in die Bibliothek? Es wird deutlich, dass ein gewisser Spielraum darin gegeben ist, wie oft man es letztlich schafft, in die Bibliothek zu kommen. So betrachtet könnte das Wort *bis* auch *oder* bedeuten. Demnach geht aus dieser Definition nicht eindeutig hervor, ob die *schlüssige und angemessene Darstellung* eines Inhaltspunktes einen (1) oder 0,5 Punkte bekommt - gleiches gilt auch für die *schlüssige und angemessene Darstellung zweier Inhaltspunkte*. Außerdem stellt sich an dieser Stelle die Frage, ob von einer Gleichwertigkeit gesprochen werden kann, wenn *alle Inhaltspunkte nur ansatzweise dargestellt sind* gegenüber der *schlüssigen und angemessenen Darstellung eines oder zweier Inhaltspunkte*. Aufgefasst kann diese Definition aber auch anders: Wenn man *ein bis zwei Inhaltspunkte schlüssig und angemessen darstellt*, erlangt man einen (1) Punkt, während für die ansatzweise behandelten fünf Inhaltspunkte lediglich 0,5 Punkte vergeben werden. Explizit und definitiv kann aber hier nicht gesagt werden, ob dem tatsächlich so ist und ob die Bearbeitung eines oder zweier Inhaltspunkte gleichermaßen gewichtet wird. Möglich wäre, dass darin auch eine normorientierte Bewertung versteckt ist. Es geht aus dieser Deskriptorendefinition ganz deutlich hervor, dass der Ratersubjektivität nicht nur im letzten Deskriptor *Thema verfehlt*, sondern bereits an dieser Stelle sehr viel Platz eingeräumt wird. Weiterhin geht aus den Deskriptoren dieses

Kriteriums nicht hervor, ob es eine Auswirkung auf die Bewertung hat, wenn sich die Lernerproduktion unter den geforderten 200 Wörtern bewegt. In den originalen Bewertungen des Goethe-Instituts für das C1-Niveau wird zu sehen sein, dass dieses Detail dennoch in der Bewertung Berücksichtigung findet, auch wenn in keinem Deskriptor davon die Rede ist.

5.2.2.2 Kriterium: Textaufbau + Kohärenz

KRITERIUM II Textaufbau +Kohärenz	5 Punkte	4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Gliederung des Textes • Konnektoren, Kohärenz	Liest sich sehr flüssig	Liest sich noch flüssig	Liest sich stellenweise sprunghaft, und einige fehlerhafte Konnektoren	Aneinanderreihung von Sätzen fast ohne logische Verknüpfung	Über weite Strecken unlogischer Text

Tabelle 32: Textaufbau und Kohärenz im C1–Zertifikat des Goethe-Instituts

Dieses Kriterium macht 25% der Gesamtbewertung in diesem Teil des schriftlichen Ausdrucks auf dem C1-Niveau aus. Dabei geht es um die Textgliederung und die Anwendung von Konnektoren, um Textkohärenz zu erhalten. Für die Formulierung dieses Kriteriums hat sich das Goethe-Institut an den definierten Kann-Beschreibungen des GER orientiert (GER 2001:125):

„Kann klar, sehr fließend und gut strukturiert sprechen und zeigt, dass er/sie die Mittel der Gliederung sowie der inhaltlichen und sprachlichen Verknüpfung beherrscht".

Diese Kann-Beschreibung des GER scheint ein konkretes Interlanguagestadium für dieses Kriterium zu definieren, welches sich im obersten Deskriptor diese Kriteriums wieder findet. Zudem decken sich die ersten zwei Deskriptorendefinitionen, für die jeweils fünf (5) und vier (4) Punkte vergeben werden, vollständig mit denen aus dem Bewertungskatalog des B2-Zertifikats. Demnach erübrigt es sich, die Schwachstellen erneut ausführlich anzuführen und kritisch zu dokumentieren. Der Lesefluss und die mögliche Implikation in seinen Steigerungsformen ist auch hier durch den subjektiven Eindruck der Rater gekennzeichnet. Auf Unterschiede bzw. Schnittstellen der zwei Niveaustufen B2 und C1 wird später noch Stellung genommen werden.

Der dritte Deskriptor vergibt drei von fünf Punkten, wenn sich die Textproduktion *stellenweise sprunghaft liest, und einige fehlerhafte Konnektoren* verzeichnet werden. Die stellenweise Sprunghaftigkeit scheint eng gekoppelt mit der korrekten Verwendung der Konnektoren zu sein. Die fehlerhafte Verwendung von Konnektoren kann aber syntaktischer und weniger textlinguistischer Natur sein. In diesem Fall sollte ihre Bewertung dann innerhalb des Kriteriums Korrektheit berücksichtigt werden. Der vierte Deskriptor vergibt entweder zwei (2) oder einen (1) Punkt, wenn Rater eine

Aneinanderreihung von Sätzen fast ohne logische Verknüpfung registrieren. Dabei kann aber nicht von Text die Rede sein, wie bereits in Kapitel 3.4 diskutiert wurde. Auch bei diesem Deskriptor ist nicht ersichtlich, wann zwei (2) und wann ein (1) Punkt dafür vergeben werden. Das Schlüsselwort scheint in dieser Definition das Wort *fast* zu sein. *Fast ohne logische Verknüpfung* heißt dann wohl, dass es an der harten Annahme angrenzt, durch fehlende logische Satzverknüpfungen auf einen unlogischen Text zu schließen, dass der Text unlogisch wird oder dass es lediglich um eine Satzaneinanderreihung geht, die jedoch dem C1-Niveau nicht gerecht wird. In Kapitel 5.1.2.2 habe ich den Kohärenz- und Textbegriff bereits ausführlich angeführt. Was den letzten Deskriptor anbelangt, definiert er *über weite Strecken unlogische Texte*. Bereits die Definitionskonstellation *unlogischer Text* lässt aus textlinguistischer Sicht zu wünschen übrig, denn entweder wird ein Text produziert oder es handelt sich nur um eine Aneinanderreihung von Sätzen, ohne im Zeichen der Kohärenz zu stehen.

5.2.2.3 Kriterium: Ausdrucksfähigkeit

KRITERIUM III Ausdrucksfähigkeit	5 Punkte	4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Wortschatzspektrum • Wortschatzbeherrschung	Sehr gut und angemessen	Gut und angemessen	Stellenweise gut und angemessen	Begrenzte Ausdrucksfähigkeit, Kommunikation stellenweise gestört	Text in großen Teilen völlig unverständlich

Tabelle 33: Ausdruck im C1–Zertifikat des Goethe-Instituts

Das dritte Kriterium, das die Bewertungsskala für die C1-Prüfung des Goethe-Instituts ausmacht, ist die Ausdrucksfähigkeit. Das Spektrum sprachlicher Mittel allgemein wird im GER für das C1-Niveau folgendermaßen definiert (GER 2001:110):

„Kann aus seinen/ihren umfangreichen Sprachkenntnissen Formulierungen auswählen, mit deren Hilfe er/sie sich klar ausdrücken kann, ohne sich in dem, was er/sie sagen möchte, einschränken zu müssen".

Detaillierter ausgedrückt geht es hier um das Wortschatzspektrum und die Wortschatzbeherrschung. Im GER lauten die Kann-Beschreibungen für diese beiden Aspekte jeweils (GER 2001:112ff):

- **Wortschatzspektrum**

  „Beherrscht einen großen Wortschatz und kann bei Wortschatzlücken problemlos Umschreibungen gebrauchen; offensichtliches Suchen nach Worten oder der Rückgriff auf Vermeidungsstrategien sind selten. Gute Beherrschung idiomatischer Ausdrücke und umgangssprachlicher Wendungen".

- **Wortschatzbeherrschung**

  „Gelegentliche kleinere Schnitzer, aber keine größeren Fehler im Wortgebrauch".

Die ersten drei Deskriptoren decken sich zum einen mit der vom GER definierten Kann-Beschreibung, zudem aber auch mit den ersten drei des B2-Zertifikats. Erneut stellt sich die Frage, inwiefern sich Wortschatzspektrum und Wortschatzbeherrschung innerhalb der Schwellenniveaus B2 und C1 unterscheiden. Darauf wird am Ende der Diskussion beider Niveaus eingegangen.

Der vierte der fünf Deskriptoren vergibt für die Definition *begrenzte Ausdrucksfähigkeit, Kommunikation stellenweise gestört* entweder zwei (2) oder einen (1) Punkt. Die Vergabe dieser Punkte kann sich meines Erachtens hier nur an dem strengeren oder milderen Urteil des Raters orientieren. Das hängt im Einzelnen davon ab, wie begrenzt den jeweiligen Ratern die Ausdrucksfähigkeit erscheint und an wie vielen Stellen sie die Kommunikation als gestört empfinden, damit sie zwei oder einen Punkt/e vergeben. Der Übergang zum Deskriptor, der mit Null (0) Punkten bewertet, ist schließlich sehr konsequent. An dieser Stelle stoßen Rater auf einen *Text*, der ihnen *in großen Teilen völlig unverständlich* erscheint. Es gilt zu klären, wodurch ein Text unverständlich wird, denn wie im nächsten Kriterium Korrektheit zu sehen sein wird, vergibt auch hier der letzte Deskriptor Null (0) Punkte, wenn der *Text wegen großer Fehlerzahl unverständlich ist*. Unklar ist bislang, worin der Unterschied der Unverständlichkeit eines Textes zwischen den zwei Kriterien begründet liegt. Es scheint, als ginge es lediglich um die Quantität der Ausdrucksfehler und der Fehler insgesamt. Im Folgenden soll aber das Kriterium Korrektheit ganzheitlich aufgezeigt und diskutiert werden.

5.2.2.4 Kriterium: Korrektheit

KRITERIUM IV Korrektheit	6 Punkte	5-4 Punkte	3 Punkte	2-1 Punkte	0 Punkte
• Morphologie • Syntax • Orthografie, Interpunktion	Nur sehr kleine Fehler	Einige Fehler, die das Verständnis aber nicht beeinträchtigen	Einige Fehler, die den Leseprozess stellenweise behindern	Häufige Fehler, die den Leseprozess stark behindern	Text wegen großer Fehlerzahl unverständlich

Tabelle 34: Korrektheit im C1–Zertifikat des Goethe-Instituts

Für das C1-Zertifikat des Goethe-Instituts wird dieses Kriterium mit einer Gewichtung von insgesamt 30% (maximal 6 von 20 zu erreichenden Punkten) repräsentiert. Es kann also davon ausgegangen werden, dass auf dieser Stufe nun Morphologie, Syntax, Orthografie und Interpunktion die entscheidende Rolle hinsichtlich der schriftlichen Textproduktion ausmachen. Dennoch kann diese Annahme trotz der Gewichtung dieses Kriteriums revidiert werden, denn die 0-Punkte-Marke jedes Kriteriums dieser Skala führt zu keiner Möglichkeit eines Ausgleichs. In diesem Kriterium ist der Punktabfall inhomogen. Der erste Deskriptor ist mit sechs (6), der zweite mit fünf bis vier (5-4), der dritte mit drei (3), der vierte mit zwei bis einen (2-1) und der letzte schließlich mit Null (0) Punkten besetzt. Auch für dieses Kriterium definiert der GER Kann-Beschreibungen (GER 2001:114,118):

- **Grammatische Korrektheit**

  „Kann beständig ein hohes Maß an grammatischer Korrektheit beibehalten. Fehler sind selten und fallen kaum auf".

- **Beherrschung der Orthografie**

  „Die Gestaltung, die Gliederung in Absätze und die Zeichensetzung sind konsistent und hilfreich".

  „Die Rechtschreibung ist, abgesehen von gelegentlichen Verschreiben, richtig".

Die maximal zu erreichende Punktzahl von sechs Punkten erlangt man, wenn es sich um *nur sehr kleine Fehler handelt*. Es muss zunächst erwähnt werden, dass die Untergliederungen in diesem Kriterium keine Aussage darüber machen, ob sie als gleichwertig zu betrachten sind. Unbeachtet der Tatsache, dass ich diese Kombination für ungerechtfertigt halte, ist es nicht ersichtlich, welche der Unterkriterien Morphologie, Syntax, Orthografie und Interpunktion eine größere oder geringere Rolle bei dem Korrektheitsanspruch spielen.[139] Man könnte die Vermutung anstellen, dass sie mit jeweils 33,33 % als gleichwertig zu betrachten sind. Das erscheint aber für die Bewertung und im Sinne der Testentwicklung als weniger praktikabel, was durchaus nachvollziehbar ist und auf einen schriftlichen Text bezogen vom Auge des Raters eine derartige prozentuale Aufteilung des Kriteriums nicht abverlangt werden kann. Trotzdem bleibt es unklar, was man unter sehr kleinen Fehlern zu verstehen hat. Ist eine falsche Adjektivendung, ein falsch gesetztes Komma oder ein fehlendes Dehnungs-h ein sehr kleiner Fehler? Interessant wäre die Antwort auf die Frage, wie die einzelnen Rater diese Definition interpretieren, denn es ist durchaus denkbar, dass es zu verschiedenen Interpretationen eines Fehlers und insgesamt dieses Deskriptors kommen kann. Im nächsten Deskriptor gibt es für die Definition *einige Fehler, die das Verständnis aber nicht beeinträchtigen* zwei alternative Punkte zu vergeben (5-4). In welchem Fall dem Korrektheitsanspruchs einer C1-Leistung fünf oder vier Punkte genügen, geht aus diesem Deskriptor nicht hervor. Wie auch im B2-Zertifikat erzielt eine schriftliche Leistung auf C1-Niveau ganze drei (3) Punkte, wenn es sich um *einige Fehler, die den Leseprozess stellenweise behindern* handelt. Diese Deskriptorendefinition ist bereits in der B2-Diskussion erörtert und hinterfragt worden (s. Kap. 5.1.2.4). Der vorletzte Deskriptor vergibt zwei bis einen Punkt (2-1), wenn *häufige Fehler, die den Leseprozess stark behindern*, auftreten. Zunächst müsste definiert werden, was unter häufigen Fehlern zu verstehen ist. Sind es immer wiederkehrende gleiche Fehler (Wiederholungsfehler) oder ist damit die Fehlerfrequenz gemeint? Was Wiederholungsfehler anbelangt, so ist aus den Deskriptoren nicht ersichtlich, ob sie doppelt bewertet werden oder nicht. Geht es andererseits um die Frequenz der Fehler, dann müsste lediglich geklärt werden, welche Fehlerarten den Leseprozess derart stark behindern. Außerdem muss explizit definiert werden, wie stark die *Leseprozessbehinderung* sein muss, dass man im besten Fall zwei (2) und im schlimmsten Fall einen (1) Punkt vergibt, obwohl die Definition dieses Deskriptors für beide Alternativpunkte gleichbedeutend ist. Geht man zur nächsten und letzten Deskriptorendefinition *Text wegen großer Fehlerzahl unverständlich* über, dann stellt sich die Frage, ob denn jegliche Fehlerarten notwendigerweise zur Unverständlichkeit einer Lernerproduktion führen müssen. Es muss hinterfragt werden,

139 In Kapitel 6 werde ich genauer Bezug dazu nehmen, warum die Kombination dieser Unterkriterien nicht angemessen scheint.

ob viele Anomalien auch automatisch die Unverständlichkeit für alle Rater implizieren, so dass man Null (0) Punkte dafür bekommt. Diesbezüglich erscheint mir dieser Deskriptor eher als eine sehr subjektiv wahrzunehmende Untergruppierung des Kriteriums Korrektheit, wobei die beträchtliche Fehlerzahl das Textverständnis möglich oder unmöglich macht.

5.2.3 Diskussion von Lernerreaktionen auf die Aufgabenstellung und deren Originalbewertungen

Im Weiteren sollen nun originale C1-Lernerproduktionen, die vom Goethe-Institut anhand der vorgestellten Bewertungskriterien korrigiert und bewertet worden sind, angeführt werden. Es soll untersucht werden, ob die aufgestellten Bewertungskriterien anhand der definierten Deskriptoren für dieses Niveau der Validität gerecht werden und ob die Bewertung durch die Rater insgesamt gerechtfertigt erscheint.

Lösungsvorschlag A **193 Wörter**

Laut der Statistik mögen die meisten Jugendlichen zwischen 12 und 25 Jahren sich mit Leuten treffen und fernsehen.

Diese Freizeitbeschäftigung erfreuen sich großer Beliebtheit sowohl bei Mädchen, als auch bei den Jungen. Auf dem dritten Platz sind Bücher, wobei sie mehr von Mädchen gelesen werden (32%).

Es lassen sich auch andere Unterschiede zwischen Jungen und Mädchen erkennen. Die überwiegende Mehrheit von Mädchen shoppt gern (27%) und unternimmt etwas mit der Familie (21%), während Jungen sich mehr fürs Internet, Computer interessieren (34 und 33). Sport ist bei den beiden Geschlechter sehr beliebt, obwohl die Anzahl der spotinteressierten Jungen ein bisschen überwiegt.

Meiner Meinung nach ist diese Grafik typisch für die jungen Leute überall auf der Welt, denn es sind hier die verbreitesten und beliebtesten Freizeitaktivitäten von Jugendlichen dargestellt, die unabhängig von der Nationalität sind.

Heutzutage bleibt den jungen Menschen ziemlich wenig Zeit für ihre Lieblingsbeschäftigungen übrig. Es gibt nämlich so viele Herausforderungen, die man meistern muss und im Vergleich zu der älteren Generation stehen die Jugendlichen mehr unter dem Zeitdruck. Ich persönlich verbringe meine Freizeit so, wie die anderen jungen Leute überall auf der Welt. Am liebsten treffe ich mich mit Freunden oder lese.

Mit einer erzielten Bewertung von 19 von maximal zu erreichenden 20 Punkten gilt diese Lernerproduktion als eine sehr gute Leistung. Kommentiert wird, dass *die Teilnehmende einen flüssig lesbaren, in sich klar strukturierten Text verfasst und trotz der Kürze des Textes eine überzeugende Leistung ihrer Sprachbeherrschung auf C1-Niveau gezeigt hat.*[140] Der ausführliche und auf die einzelnen Bewertungskriterien bezogene Kommentar und wie sich die erreichten 19 Punkte zusammen setzen sei im Folgenden aufgezeigt und diskutiert:[141]

Kriterium	Kommentar	Bewertung
Inhaltliche Vollständigkeit	Die Textlänge ist gerade noch ausreichend. Nur ein Inhaltspunkt (*Freizeitverhalten der älteren Generation*) wird zu knapp, alle anderen werden ausreichend behandelt.	3 Punkte
Textaufbau und Kohärenz	Der Text liest sich flüssig und ist klar strukturiert. Eine eindeutige Einleitung fehlt zwar, aber dies fällt kaum auf. Deswegen kein Punktabzug	5 Punkte
Ausdrucksfähigkeit	Der Wortschatz des Textes ist dem Niveau entsprechend gewählt (*erfreuen sich großer Beliebtheit; die Anzahl überwiegt; Herausforderungen, die man meistern muss*) und es gibt keine falsch verwendeten Ausdrücke.	5 Punkte
Korrektheit	Es treten nur sehr vereinzelt Fehler auf (*bei beiden Geschlechter*). Nach Diskussion Entscheidung für sechs Punkte.	6 Punkte

Tabelle 35: Originalbewertung einer C1- Produktion

Ganz gezielt wurde diese Lernerproduktion als erste ausgewählt, um aufzuzeigen, dass die Validität der Bewertungskriterien prinzipiell nur durch tatsächlich objektiv zu betrachtende einwandfreie Leistungen beibehalten werden kann. Es gibt die definierten Bewertungskriterien und die Rater, die diese anwenden. Die Bewertung eines Textes kann nicht besser sein, als die Bewertungskriterien bestimmen. Die definierten Bewertungskriterien beziehen sich auf das zugrunde liegende Konstrukt, was hier der schriftliche Ausdruck ist. Die Validität der Bewertungskriterien ergibt sich damit daraus, wie gut diese das Konstrukt widerspiegeln. Die Umsetzung dieser Bewertungskriterien soll aufzeigen, ob Rater diesem Validitätsniveau mittels ihrer Bewertung entsprechen können. Das Paradoxon, das sich aber an dieser Stelle zeigt ist, dass es offensichtlich einerseits Texte gibt, die besser zu den Kriterien passen und andererseits Texte, die nicht sehr kompatibel mit den Bewertungskriterien zu sein scheinen. Die Bewertungskriterien sollten aber derart stabil sein, dass mit ihrer Hilfe aufgezeigt werden kann, wie gut ein Text den Kriterien entspricht und wie die Kompetenz im schriftlichen Ausdruck als Teil der gesamten Sprachkompetenz schließlich eingeschätzt werden kann. Eine Bewertungsskala kann also nicht nur dann valide sein, wenn die Leistung vorbehaltlos

140 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 11

141 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 11

und einwandfrei ist und durch das nicht in Anspruch nehmen eines Korrekturstiftes von jedermann als eine fehlerlose Leistung betrachtet werden würde. Es handelt sich um verschiedene Abstufungen von Interlanguages einzelner Lerner eines bestimmten Niveaus, auf die innerhalb des Bewertungskatalogs insofern Rücksicht zu nehmen ist, dass auch diese von allen gleichermaßen wahrgenommen und entsprechend eingestuft werden.

Lösungsvorschlag B **285 Wörter**

Freizeit der Jugend

Wie jeder die Freizeit verbringt, ist eine ganz persönliche Sache. Heutzutage gibt es besonders viele Möglichkeiten, sich zu erholen und dem eignen Interesse entsprechend Zeit zu verbringen.

Was mir auf dem ersten Blick in der Statistik auffällt, ist der Wunsch einer Mehrheit der Jugend möglichst viel Zeit miteinander zu verbringen; sogar mehr so bei den Mädchen als bei Jungen. Dass das Fernsehen die nächst beliebteste Tätigkeit der Jugend ist, ist kaum zu erstaunen, wie auch die Tatsache, dass Jungen viel länger fernsehen als Mädchen.

Was unter den markanten Unterschieden zwischen Jungen und Mädchen ins Auge sticht, ist die Beschäftigung mit dem Computerspiel und Internet surfen. Beide sind Bereiche der Jungen; während nur 8 % der Mädchen am Computer spielen, interessieren sich 33 % der Jungen daran! Auch beim Internetsurfen ist die Zahl der Jungen das Doppelte als der Mädchen.

Ganz erwartet war die Sache mit dem Einkaufen, das typisch weiblich ist. In den Geschäften sieht man doch fünfmal mehr Mädchen als Jungen! Das Lesen als Freizeitbeschäftigung interessiert Mädchen viel mehr als Jungen: in Zahlen sind sie beziehungsweise 32 % und 18 %. Diese Grafik könnte man als fast typisch für die Jugend überall auf der Welt nennen. Die junge Menschen genießen selbstverständlich viel mehr Freizeit im Vergleich zu den Älteren. Mit Alter kommen mehrere Verantwortungen: der Beruf, der Haushalt, die Familie, die Kinder usw. Die ältere Generation besonders in Indien hat viel weniger Freizeit. Damals gab es keine technische Geräte als Hilfe wie zum Beispiel Waschmaschine, Geschirrspüler, Elektroherd, u.ä. Man wohnte früher in Großfamilien und das machte viel Arbeit und kaum Freizeit. In meiner Freizeit schreibe ich gern Briefe – die sind oft E-mails, außerdem höre ich Musik, lese Bücher oder unterhalte mich mit Freunden.

Die als zweite hier angeführte Lernerproduktion einer C1-Prüfung wird vom Goethe-Institut als eine gute Leistung kommentiert:[142] *Der Text ist gut gegliedert und lesbar, nur vereinzelt fehlen Verknüpfungen. Jedoch gibt es beim Ausdruck noch einige unpassende Wortverwendungen, die den Gesamteindruck etwas trüben*.

142 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 15

Inhaltliche Vollständigkeit	**Der Text ist fast um ein Drittel länger als gefordert. Alle Inhaltspunkte werden behandelt, bis auf einen alle ausführlich. Aufgrund des ausgewogenen Textes wird für die kurze Aussage zum dritten Inhaltspunkte (Vergleich Heimatland) kein Punkt abgezogen.**	**4 Punkte**
Textaufbau und Kohärenz	**Der Text ist überwiegend flüssig lesbar, an einigen Stellen fehlen Verknüpfungen zwischen den einzelnen Abschnitten.**	**4 Punkte**
Ausdrucksfähigkeit	**Die Wortschatzkenntnisse werden differenziert eingesetzt. An einigen Stellen fehlen jedoch adäquate Ausdrücke und es werden „nahe" Ausdrücke verwendet (*kaum zu erstaunen; ganz erwartet; Verantwortungen statt „Verpflichtungen"*). Das Verständnis bleibt gesichert, jedoch wird ein Punkt angezogen.**	**4 Punkte**
Korrektheit	**Es gibt vereinzelt Fehler (*interessieren daran*), die beim Lesen jedoch kaum auffallen und den Leseprozess nicht behindern.**	**4 Punkte**

Tabelle 36: Originalbewertung einer C1- Produktion

Erstrangig geht es bei dieser kommentierten Bewertung darum, ob die angesetzten Bewertungskriterien des Goethe-Instituts für das C1-Niveau auch adäquat benutzt werden, sodass sich der Kommentar und die reale Bewertung mit dem Skalenkonstrukt tatsächlich deckt.

Das Kriterium der inhaltlichen Vollständigkeit scheint sich unabhängig von der Definition der einzelnen Deskriptoren nicht nur auf die zu bearbeitenden fünf Inhaltspunkte zu beziehen, sondern auch auf die Mindestanforderung der Textlänge (hier: ca. 200 Wörter). Dies geht lediglich aus der kommentierten Bewertung hervor, die sich auf die Textlänge bezieht. Kein einziger Deskriptor dieses Kriteriums erwähnt neben der Inhaltspunktedarstellung diesen Umstand. Inwiefern aus den Begriffen *schlüssig, angemessen* oder sogar *ansatzweise* auf die Textlänge geschlossen werden kann, ist sehr fraglich. Laut Bewertungskommentar werden die maximal zu erreichenden vier (4) Punkte vergeben, da *der Text ausgewogen ist*. Die Tatsache, dass lediglich vier (4) von fünf (5) Inhaltspunkten ausführlich behandelt worden sind, müsste laut der definierten Deskriptoren in der Bewertungsskala zu drei (3) Punkten führen. Dies impliziert, dass die Definition des konkret zu benutzenden Deskriptors keinerlei Einfluss auf die Bewertung hat. Statt dessen wurde spontan und wohl ratersubjektiv eine interne Ausnahmeklausel für vorliegenden Fall geschaffen. Dies ist aber nirgendwo im Definitionsraster des Kriteriums inhaltliche Vollständigkeit verzeichnet. Ganz strikt und objektiv betrachtet, müssten im Sinne der Definition und der Deskriptorenkategorisierung dieses Kriteriums drei (3) Punkte vergeben werden, denn wie bereits von den Ratern selbst kommentiert wurde, sind nur vier (4) von fünf (5) Inhaltspunkten ausführlich behandelt worden. Das

deckt sich absolut mit dem zweiten Deskriptor, der für die *schlüssige und angemessene Darstellung von vier Inhaltspunkten* steht.

Bei der Betrachtung des zweiten Kriteriums wird der Text als *überwiegend flüssig lesbar* kommentiert und bekommt dafür die Punktzahl vier (4), die dem zweiten Deskriptor zu verdanken ist. Meines Erachtens decken sich die Definitionsabschnitte *überwiegend flüssig* und *liest sich noch flüssig*. Es folgt aber der Zusatz, dass *an einigen Stellen Verknüpfungen zwischen den einzelnen Abschnitten fehlen*. Wenn dies von geschulten Ratern so verzeichnet wird, stellt sich die Frage, wo dies auf der definierten Skala dieses Kriteriums am besten Platz finden würde. Es ist nicht eindeutig, ob dieser feine Makel bereits in diesem Deskriptor oder in der Teildefinition *einige fehlerhafte Konnektoren* zu suchen ist. Außerdem gilt zu klären, ob fehlende Verknüpfungen gleichzeitig auch als fehlerhaft gelten.

Betrachtet man des Weiteren die Bewertung der Ausdrucksfähigkeit, so entdeckt man innerhalb der kommentierten Bewertung einen Widerspruch. Obwohl den Ratern die *Wortschatzkenntnisse differenziert eingesetzt zu sein scheinen, fehlen an einigen Stellen adäquate Ausdrücke*. Für diese Gegensätzlichkeit werden vier (4) Punkte vergeben, die aber unter den Deskriptor *gut und angemessen* fallen. Die Definition *gut und angemessen* enthält aber keinerlei negative Einschränkung oder jegliche Form von inkompetenter Verwendung. Diese Raterbeobachtung fände besser Platz im Deskriptor *stellenweise gut und angemessen*, was jedoch lediglich mit drei (3) Punkten honoriert würde.

Diese Leistung erzielt bei dem am stärksten gewichteten Bewertungskriterium Korrektheit vier (4) Punkte. Folglich werden die *vereinzelten Fehler, die beim Lesen kaum auffallen und den Leseprozess nicht behindern* dem zweiten Deskriptor zugeschrieben. Interessant ist an dieser Stelle jedoch, dass dieser Deskriptor entweder fünf (5) oder vier (4) Punkte für *einige Fehler, die das Verständnis aber nicht beeinträchtigen*, vergibt. Der angewandte Maßstab oder das Kriterium gehen aus der Punktevergabe von fünf (5) oder vier (4) Punkten nicht hervor, da in beiden Fällen die gleiche Definition gilt. Man könnte zum Beispiel an dieser Stelle konkret hinterfragen, wieso die vorliegende Lernerproduktion nicht fünf (5) Punkte erzielt. Es scheint, dass gerade bei Deskriptoren, die aus zwei Punkten bestehen, dem Rater ein sehr großer Freiraum gegeben wird, diese ganz individuell und funktional einzusetzen. Man sieht, dass sobald Mängel innerhalb der angesetzten Kriterien zu verzeichnen sind, die Schwierigkeit beginnt, diese den definierten Deskriptoren genau zuzuordnen. Die Begründungen, die teilweise für die Vergabe von Punkten gegeben werden, sind ratersubjektive Ausnahmedefinitionen, die in den Deskriptoren nicht einmal ansatzweise angeführt sind.

Lösungsvorschlag C **274 Wörter**

Ich beschäftige mich mit dem Thema „Freizeit und Jugend“. Dafür habe ich vor mir die Statistiken Angaben. Die vorliegende Statistik verdeutlicht, dass die meisten Jugendliche ihrer Freizeit sich mit den Leuten zu treffen verbringen. Im Vergleich zu den Jungen verbringen die Mädchen ihrer Freizeit damit. In diesem Zusammenhang fällt mir auf, dass ihre Interessen an etwas mit der Familie zu unternehmen steht an der siebten Stelle. Es gibt einen großen Unterschied zwischen den Interessen der Mädchen und der Jungen. Während 33 Jungen von den 100 Befragten sich für den Computer interessieren, interessieren sich nur 8 Mädchen dafür. Merkwürdig ist, dass nur 5 Jungen gerne einkaufen, wogegen 27 Mädchen gern zum Einkaufen gehen. Aber meiner Meinung nach ist diese Grafik nicht für die jungen leute überall auf der Welt typisch. Mit den Interessen der Jugend kann man sich nicht verallgemeinern. Dafür gibt es verschiedene Gründe nähmlich wie viel Zeit man für sich selbst hat, woran hat man Interesse oder wie viel Geld man für solche aktivität ausgeben willst oder kannst usw. dafür spielt der Alter auch eine große Rolle. Ich bin der Meinung, dass die ältere Generation mehr Zeit zur Verfügung hat um sich zu erholen oder einen Hobby zu treiben, besonders wenn man pensioniert oder in der Rente ist. Im solchen Zeitraum hat man viel Zeit für sich selbst. Auf der anderen Seite, wenn ein Junge gar nicht arbeiten will, hat er auch viel Zeit zur Verfügung.

Zum Schluss möchte ich Ihnen von meinem eigenen Interesse erzählen. Obwohl ich kein Hobby habe, beschäftige ich ich mit der Vorbereitung der Arbeitsblätter für die Unterrichtsstunden. Das gefällt mir und deshalb nenne ich diese Tätigkeit als mein Hobby.

Diese Leistung bekommt 13 von 20 Punkten und gilt als auf *niedrigem C1-Niveau* begründet, obwohl *der Text in sich gegliedert und gut verknüpft ist*. Kommentiert wird aber, *dass der Gesamteindruck durch die hohe Fehleranzahl beeinträchtigt wird.*[143] Zu vermerken wäre an dieser Stelle, wieso von Gesamteindruck die Rede ist, wenn es nicht um eine holistische Bewertungsskala geht, die dem subjektiven Eindruck der Rater genügend Entfaltungsmöglichkeiten lässt. Konkreter werden die einzelnen Bewertungskriterien wie folgt kommentiert und aufgezeigt:[144]

Inhaltliche Vollständigkeit	**Die Textlänge ist mehr als ausreichend. Alle Inhaltspunkte werden angemessen behandelt**	**4 Punkte**
Textaufbau und Kohärenz	**Der Text ist strukturiert und flüssig lesbar, es gibt sowohl eine Einleitung als auch einen Schluss. Im gesamten Text sind die Sätze und Abschnitte miteinander verknüpft. Die Lesbarkeit wird durch die hohe Fehlerzahl beeinträchtigt und bei diesem Kriterium bewertet.**	**4 Punkte**

143 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 23

144 Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707, S. 23

Ausdrucksfähigkeit	**Der Wortschatz ist überwiegend angemessen, es gibt nur vereinzelt Fehler (*kann man sich nicht verallgemeinern*) bzw. unklare Ausdrücke (*wenn ein Junge gar nicht arbeiten will*).**	**3 Punkte**
Korrektheit	**Es gibt zahlreiche Fehler im Bereich der Syntax und der Morphologie, auch beim Genus (*dass die meisten Jugendlichen ihrer Freizeit sich mit den Leuten treffen verbringen; man für solche aktivitäten ausgeben willst; keinen Hobby*). Nach Diskussion Entscheidung für zwei Punkte.**	**2 Punkte**

Tabelle 37: Originalbewertung einer C1- Produktion

Die inhaltliche Vollständigkeit dieser Textproduktion wird mit vier (4) Punkten bewertet und dem ist nichts auszusetzen, denn es wurde der entsprechende Deskriptor dafür angewandt, ohne dass irgendwelche Einschränkungen von Seiten der Rater zu verzeichnen sind.

Das Kriterium, das sich auf den Textaufbau und seine Kohärenz bezieht, wird dem Deskriptor *liest sich noch flüssig*, der vier (4) Punkte vergibt, zugeordnet. Im Kommentar dieser Bewertung werden sowohl Textstruktur und Lesefluss als auch Textsortenmerkmale (z.B. Einleitung, Schluss) und Verknüpfungen innerhalb des Textes positiv gekennzeichnet. Dennoch wird auf der Grundlage, dass die Lesbarkeit durch die hohe Fehlerzahl beeinträchtigt wird, ein Punkt abgezogen. Es wurde bereits in vorangegangenen Kapiteln diskutiert, was die Lesbarkeit ausmacht. Dieses Kriterium hat meines Erachtens keinerlei Funktion, um Fehler zugerechnet zu bekommen. Zudem wird im Kommentar doch explizit darauf hingewiesen, dass *der Text strukturiert und flüssig lesbar ist*. Konträr dazu steht am Ende des Kommentars, dass *die Lesbarkeit durch die hohe Fehlerzahl beeinträchtigt wird*. Ob für die Rater diese Textproduktion schließlich *flüssig lesbar* ist oder ob dadurch *ihr Lesefluss gestört wird*, kann diesen Bewertungskommentaren nicht entnommen werden, da sie sich widersprechen.

Drei (3) von fünf (5) Punkten erzielt diese Lernerproduktion für das Kriterium der Ausdrucksfähigkeit. Die Erklärung des Kommentars, *dass der Wortschatz überwiegend angemessen ist* und es nur *vereinzelt zu Fehlern oder unklaren Ausdrücken kommt* wird dem Deskriptor *stellenweise gut und angemessen* zugeordnet. Der Zusatz der vereinzelten Fehlerstreuung wird in dieser Deskriptorendefinition jedoch nicht explizit aufgeführt. Diese latente Information mag im Begriff *stellenweise* inbegriffen sein. Aus den fünf Deskriptoren, die das Kriterium der Ausdrucksfähigkeit definieren, ist wohl der ausgewählte Deskriptor der passende, jedoch lediglich hinsichtlich der Betrachtung und Wahrnehmung der Rater.

Das Kriterium Korrektheit soll auch hier, wie der Begriff selbst besagt, die korrekte Anwendung der Syntax, der Morphologie, der Orthografie und der Interpunktion untersuchen. Von maximal sechs (6) zu erreichenden Punkten erzielt diese Leistung für dieses Kriterium zwei (2) Punkte. Das wird dermaßen gerechtfertigt, dass *es zahlreiche Fehler im Bereich der Syntax und der Morphologie gibt*. In der Bewertung des Ausdrucks wurde die hohe Fehlerfrequenz aber bereits miteinbezogen, da die Lesbarkeit des Textes

dadurch beeinträchtigt wird. Explizit wird im Kommentar vermerkt, dass dieser Umstand in diesem Kriterium Anwendung für die Punktevergabe findet. Daher dürfte die Fehlerfrequenz nicht erneut im Kriterium Korrektheit bewertet werden, denn dann würde der Fall der Doppelsanktionierung eintreten. Kommentierte man nicht, dass die Fehler im Ausdruckskriterium verbucht werden, dann würden dennoch vier (4) Punkte vergeben. Anders beim Korrektheitskriterium. Die Fehlerfrequenz, die hier Anwendung findet, wird dem vierten Deskriptor (2-1 Punkte) zugeordnet. Dieser definiert häufige Fehler, *die den Leseprozess stark behindern*. Kommentiert wird unter dem Kriterium Ausdruck, dass die Lesbarkeit beeinträchtigt wird. Es wird aber nichts über die Stärke dieser Beeinträchtigung besagt. Daher könnte diese Leistung für den Korrektheitsanspruch auch drei (3) Punkte erzielen. Unklar bleibt des Weiteren, wann laut des vierten Deskriptors häufige Fehler einer Lernerleistung zwei (2) bzw. einen (1) Punkt erlangen.

Fazit

Es wurde der gesamte Kriterienkatalog für die freie schriftliche Produktion des C1-Zertifikats präsentiert und kritisch betrachtet. Außerdem habe ich anhand der Originalbewertungen von Lernerproduktionen versucht aufzuzeigen, wie diese Bewertungskriterien eingesetzt werden, so dass schließlich auch von Leistungsvalidität ausgegangen werden kann.

Bei den Originalkommentaren und Bewertungen der ausgewählten Lernerreaktionen wurde darauf hingewiesen, dass die Bewertungskriterien nicht immer dieselbe Anwendung finden. Außerdem gehen aus den definierten Deskriptoren viele von Ratern dokumentierte Merkmale nicht hervor, trotzdem finden sie bei der Bewertung Beachtung. Ein Beispiel ist die Nicht-Einhaltung der erforderlichen Textlänge. In keinem der Deskriptoren wird dieser Punkt angesprochen. Rater ziehen dennoch Punkte ab, wenn dieses Erfordernis nicht erfüllt scheint[145].

In der Testbeschreibung und in den Prüfungszielen für das C1-Zertifikat wird auf Seite 15 vermerkt, dass das Textwissen im Kriterium Textaufbau und Kohärenz berücksichtigt und bewertet wird.[146] Einsicht in diese Tatsache, auf die bereits ausführlich Bezug genommen wurde, konnte lediglich durch die originalen Bewertungskommentare der vom Goethe-Institut geschulten Rater erzielt werden. Das Kriterium selber gibt keinen Anhaltspunkt dafür, dass es darin Anwendung zu finden hat. Des Weiteren wurde die Problematik angesprochen, dass es sehr schwierig ist, Fehler den Kriterien Ausdruck oder Korrektheit zuzuordnen. Im Handbuch für das C1-Zertifikat wird angeführt, dass das grammatische Wissen mit dem Kriterium Ausdruck und Korrektheit bewertet wird.[147] Schwarz auf Weiß wird diese Problematik also zusätzlich verstärkt. Eindeutige Regelungen diesbezüglich scheinen aber nicht definiert zu sein. Aber allein durch den Umstand, dass das Kriterium Korrektheit die Unterpunkte Morphologie und Syntax beinhaltet, dürfte dieser Toleranzbereich nicht gegeben sein. Die Gefahr der Doppelsanktionierung ist prinzipiell voraussehbar und, wie aufgezeigt, oft auch präsent.

145 Nach internen Informationen wird dies ab dem Prüfungssatz 3 ausdrücklich erwähnt. Im vorliegenden Modellsatz ist jedoch noch nicht die Rede davon, wie Rater das Einhalten oder Nicht-Einhalten der Textlänge handhaben sollen.

146 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 15

147 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 15

In der Diskussion einzelner Lernerbewertungen ging es bezüglich der Wortschatz- und Grammatikkenntnisse auch darum, welche Kenntnisse von Strukturen und Wortschatzlisten wie zu bewerten und welchen Deskriptoren zuzurechnen sind. In dem vom Goethe-Institut formulierten Handbuch zu Prüfungszielen und zur Testbeschreibung steht paradoxerweise folgendes:[148]

„Wortschatz- und Grammatikinventare zum Goethe-Zertifikat C1 gibt es (...) nicht".

Dies wird damit begründet, dass sich durch die Benutzung authentischer Texte keine verbindliche Wortschatzeingrenzung vornehmen lassen kann. Dass Wortschatz nicht eingegrenzt werden kann, widerspricht aber den einzelnen Deskriptorendefinitionen des Kriteriums Ausdrucksfähigkeit. In der Diskussion wurde bereits ausführlich Bezug darauf genommen. Auf der Basis des kommunikativen Modells von Bachman/Palmer (1996) wird das funktionale und soziolinguistische Wissen dem Kriterium der Ausdrucksfähigkeit zugerechnet. Es soll bei diesem Kriterium demnach berücksichtigt werden, wie flexibel und angemessen die Sprache je nach kommunikativer und kontextueller Situation, sowie nach Ziel und Adressaten Anwendung findet.[149] Die Definition dieses Kriteriums wurde bereits ausführlich erörtert und besteht lediglich aus den Unterpunkten Wortschatzspektrum und Wortschatzbeherrschung. Die entsprechende Sprachkomplexität ist meines Erachtens in diesem Kriterium nirgendwo definiert.[150]

Abschließend lässt sich feststellen, dass, je höher das zu prüfende Niveau ist, Sprache auch komplexer wird. Dies allein führt schon dazu, dass der Subjektivität der Rater durch die teilweise defizitären Bewertungskriterien mehr Freiraum gegeben wird. Es kann schon aus dem Grund nicht von Objektivität der Sprachwahrnehmung ausgegangen werden, denn sowohl Profile als auch das Goethe-Institut können Wortschatz und Grammatik bei ansteigenden Niveaus nicht eingrenzen und definieren. Im nächsten Kapitel soll auf diese Problematik kontrastiv eingegangen werden.

5.3 Kontrastiver Ausblick und Neuansatz der Kriterien für das B2- und C1-Zertifikat des Goethe-Instituts

Im Herbst 2007 hatten das B2- und C1-Zertifikat des Goethe-Instituts ihre weltweite Premiere. Zum ersten Mal haben sich Lernende aus aller Welt diesen neuen Prüfungen unterzogen. Die Kriterienkataloge sind Werkzeuge für Rater, um den schriftlichen Ausdruck dieser zwei Niveaus zu bewerten. Die APA definiert in ihrem Unterkapitel „Supporting documentation for tests" diesbezüglich (APA-Standard 6.13:70).:

„When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or caution".

148 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 20

149 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 15

150 Die Thematik der Sprachkomplexität werde ich abschließend in Kapitel 6 behandeln, um der Problematik der Validität von Bewertungskriterien entgegenzuwirken, um dadurch möglicherweise Verbesserungsvorschläge und weiterführende Optionen zu erreichen

Prüfer werden ganz sicher über Veränderungen bezüglich des Testprozesses informiert und hinsichtlich der Bewertungskriterien und den gesetzten Anforderungen intensivst geschult. In der Prüfungsordnung der offiziellen und schließlich endgültigen Fassung des B2-Zertifikats wird mit dem § 18 darüber informiert, dass „die Mitglieder durch Trainingsmaterialien und –seminare auf ihre Aufgabe vorbereitet werden“.[151] Es stellt sich natürlich allgemein und speziell hinsichtlich von internen Veränderungen trotzdem die Frage, inwieweit Rater innerhalb der einzelnen Bewertungskriterien eine Interraterreliabilität bzw. eine Homogenität erreichen werden.[152] Es wird in § 18 zwar betont, dass Rater auf ihre Aufgabe vorbereitet werden, dennoch habe ich bereits in Kapitel 4.5.3 die Thematik der Rater angeführt und entsprechend diskutiert. Die APA definiert in diesem Zusammenhang im Kapitel *Test administration, scoring and reporting* (APA-Standard 5.9:64f):

„ When test scoring involves human judgement, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented”.

Es wurde bei der Definition der einzelnen Bewertungskriterien hinsichtlich der Niveaus B2 und C1 aufgezeigt, dass nicht nur die Bewertungskriterien identisch sind, sondern dass auch einige der Deskriptoren deckungsgleich sind. Die Frage, die sich in diesem Zusammenhang stellt, ist demnach, worin die Unterschiede beider Niveaus zu suchen sind. Während das B2-Niveau die oberste Stufe der selbständigen Sprachkompetenz darstellt, grenzt es dennoch an die unterste Ebene der kompetenten Sprachverwendung, die durch das C1-Niveau repräsentiert wird. Insofern wäre es interessant, wenn explizit definiert werden könnte, was die Unterschiede genau dieser zwei Schwellenniveaus sind. Das Goethe-Institut beschreibt in seinem Handbuch für Prüfungsziele und Testbeschreibung, dass der Unterschied sprachlichen Könnens dieser zwei Niveaus darin liegt, dass Sprache aufsteigend (also C1) eine größere Bandbreite syntaktischer Strukturen zur Verfügung hat und dass sprachliche Mittel flexibler eingesetzt werden können.[153] Dennoch liegt die Antwort in der Aufgabenstellung, die letztlich darauf abzielt, dass diese z.B. bei C1 durch Abstraktionen sprachlich bewerkstelligt werden kann. Diese so definierte Unterscheidung kann aber insofern keine Gültigkeit haben, denn es wurde bereits explizit hervorgehoben, dass bei aufsteigenden Niveaus Kategorisierungen im grammatischen und vokabularen Bereich schwierig sind. Interessant wäre in diesem Zusammenhang ebenfalls, wie und ob sich die Bewertung schriftlicher Lernerproduktionen ändern würde, wenn man die Bewertungskataloge für B2 und C1 entsprechend auswechselte. Da die Unterschiede oberflächlich minimal zu sein scheinen, könnte bis auf ein paar Feinheiten eine Einheitsskala für beide Niveaus geltend gemacht werden. Zwar wurde in der Diskussion angeführt, dass sich die Kriteriengewichtungen innerhalb dieser zwei Niveaus leicht unterscheiden, aber der Umstand, dass im Falle des Erreichens der 0-Punkte-Marke eines einzelnen Kriteriums in beiden Fällen zum Nichtausgleich der Null Punkte führt, lässt die Gewichtungen für den Fall, dass Null Punkte erreicht werden, offensichtlich wertlos erscheinen. Hätten die gewichtigsten Kriterien tatsächlich die berechnete Dominanz, dann dürfte das Erreichen der 0-Punkte-Marke schwächerer Kriterien nicht unausgleichbar sein. Außerdem verleiten die Definitionen der letzten Deskriptoren (Null Punkte) der einzelnen und gleichen Kriterien

151 Goethe-Zertifikat C1. Prüfungsordnung, Durchführungsbestimmungen. Stand 050707. § 18, S. 3

152 Siehe dazu Kapitel 6

153 Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707. S. 20

der zwei Niveaus meines Erachtens sehr leicht zu den von North (1993) definierten Ratingfehlern[154] (vgl. Kap. 4.5.2), wodurch die Subjektivität der Rater begünstigt werden kann. Das Goethe-Institut selber betont, dass es nirgendwo die Nicht-Existenz der Ratersubjektivität anführt.[155]

In internen Prüferunterlagen des Goethe-Instituts[156] für die Niveaus B2 und C1 wurden jeweils Fokuspunkte für die einzelnen Kriterien definiert, die die Grundsätze der Bewertung betreffen sollen. In einer Tabelle sollen diese kommentierten Anhaltspunkte für die Niveaus B2 und C1 zusammenfassend kontrastiv gegenüber gestellt werden:[157]

	B2	C1
Inhaltliche Vollständigkeit	• Berücksichtigung der Textlänge (ist der Brief zu kurz, d.h. weniger als 150 Wörter, führt dies zu Punktabzug, ist er zu lang, so bleibt dies unberücksichtigt) • Entscheidend: angemessene und ausführliche Darstellung der Inhaltspunkte (ausführlich bedeutet jeoch nicht, dass zu jedem Inhaltspunkt ein voller Textabschnitt zu schreiben ist)	• Berücksichtigung der Textlänge (ist der Brief zu kurz, d.h. weniger als 175 Wörter, führt dies zu Punktabzug, ist er zu lang, so bleibt dies unberücksichtigt) • Entscheidend: angemessene und ausführliche Darstellung der Inhaltspunkte (ausführlich bedeutet jeoch nicht, dass zu jedem Inhaltspunkt ein voller Textabschnitt zu schreiben ist)
Textaufbau und Kohärenz	• Textsorte: Leserbrief • Text muss Briefform aufweisen (Anrede, Gruß, Bezug zur Legende) • Für die Kohärenz spielen insbesondere Satzanfänge eine entscheidende Rolle (Bezug zu Vorhergehendem, Abwechslungsreichtum)	• Klare und gute Struktur der Texte (Hervorhebungen, Beispiele) • Für die Kohärenz spielen insbesondere Satzanfänge eine Rolle sowie ein passender Schluss, der den Text abrundet.

154 Halo-Effekt, Zentraltendenz, Strengevariation

155 Dies wurde mir am 23. Februar 2008 in einem Gespräch mit Mitarbeitern des Goethe-Instituts München, die im Bereich der Prüfungserstellung tätig sind, vermittelt.

156 Diese wurden mir vom Goethe-Institut München zur Verfügung gestellt mit der Erlaubnis diese benutzen zu dürfen.

157 Ich habe zur Hervorhebung die Unterschiede zwischen den Niveaus fett und unterstrichen markiert. Die restlichen Anhaltspunkte sind für beide Niveaus identisch.

Ausdrucks-fähigkeit	• <u>Differenziertes und nuanciertes Ausdrucksvermögen</u> • Bewertung verschiedener Aspekte: a) falsche Verwendung von Ausdrücken je nach Quantität und Qualität bewertet b) der verwendete Wortschatz liegt nicht oder nur stellenweise auf dem Niveau c) Stilbrüche im Text (z.B. nicht angemessene Einbettung der Vorgabe)	• <u>Breites Spektrum von Redemitteln</u> • Bewertung verschiedener Aspekte: a) falsche Verwendung von Ausdrücken je nach Quantität und Qualität bewertet b) der verwendete Wortschatz liegt nicht oder nur stellenweise auf dem Niveau c) Stilbrüche im Text (z.B. nicht angemessene Einbettung der Vorgabe)
Korrektheit	• Fehler dem richtigen Kriterium zuordnen • Überkorrektur und Wiederholungsfehler im Text sind zu vermeiden • Fehlerklassifizierung spielt eine untergeordnete Rolle • Wichtig und im Vordergrund: (behinderte) Verständlichkeit	• <u>Derart gute Sprachbeherrschung, dass Fehler vereinzelt vorkommen</u> • Fehler dem richtigen Kriterium zuordnen • Überkorrektur und Wiederholungsfehler im Text sind zu vermeiden • Fehlerklassifizierung spielt eine untergeordnete Rolle • Wichtig und im Vordergrund: (behinderte) Verständlichkeit

Tabelle 38: Kontrastive Gegenüberstellung interner Bewertungsrichtlinien für die Niveaus B2 und C1 des Goethe-Instituts

Diese Fokuspunkte, die Rater innerhalb ihrer Prüferschulung vorgelegt bekommen, sind keineswegs Inhalte der definierten Deskriptoren der einzelnen Bewertungskriterien und Niveaus. Das könnte unter anderem dadurch erklärt werden, dass z.B. die Textsorte gelegentlich im Bewertungskommentar angeführt wird. Weiterhin bestätigt sich meine These, dass sich die Kriterienkataloge oder auch der Fokus auf die Niveaus B2 und C1 minimal unterscheiden. Dennoch sollen sich die Textproduktionen hinsichtlich ihrer Komplexität (z.B. Kohäsion etc.)[158], die aus unterschiedlichen Schwierigkeitsgraden der Aufgabenstellung (Vergleich vs. Abstraktion) resultieren, unterscheiden.

158 Der Begriff der Komplexität wird in Kapitel 6 angeführt.

In den Originalbewertungen sind manche Kommentare hinsichtlich ihrer in den Deskriptoren zu suchenden Relevanz diskutiert worden. Wie es scheint haben diese ihre Legitimität nur in den Prüferunterlagen, denn der jedem zugängliche Kriterienkatalog gibt keine Auskunft über latentes Bewertungsvorgehen. Bereits im ersten Kriterium *Inhaltliche Vollständigkeit* sieht man, dass die Nicht-Einhaltung der minimalen Textlänge (B2:150 vs. C1: 175) zu Punktminderungen führt. Diskutiert wurde dies bereits, denn die Deskriptoren sind nicht entsprechend formuliert und danach ausgerichtet. Außerdem wurde in der vorangegangenen Diskussion der Punkt der Textsorte angesprochen, der in keinem der Kriterien und seinen deskriptiven Abstufungen berücksichtigt scheint. Die Aufgabenstellung des schriftlichen Ausdrucks für das B2-Zertifikat besteht darin, einen Leserbrief zu verfassen. Ich stellte die Frage, in welchem Kriterium diese Schreibproduktion Anwendung findet. Ich konnte lediglich mögliche Erklärungen finden, weshalb und wo diese Textkompetenz zu berücksichtigen wäre. Aus den internen Prüferunterlagen geht eindeutig hervor, dass im Kriterium Textaufbau und Kohärenz darauf geachtet wird, dass diese Textsorte formgerecht umgesetzt wird. Dies ist natürlich von erheblicher Bedeutung. Dennoch geht es in dieser Arbeit in erster Linie um die Bewertungskriterien und die definierten Deskriptoren, die in dieser Hinsicht aber nichts darüber aussagen. Es bleibt daher zu klären, ob Rater lediglich durch die Prüferunterlagen diesen Umstand berücksichtigen, gerade wenn ausschließlich das analytische Bewertungsraster für die Bewertung des schriftlichen Ausdrucks als Instrument dienen soll. Wenn ein Deskriptor grundlegende Bedingungen nicht explizit macht, wirkt sich dies negativ auf die Bewertungsreliabilität und –validität aus. Für das C1-Niveau sollen die Rater lediglich darauf achten, dass die Struktur klar und gut ist (was diese Definition wiederum impliziert, soll hier nicht weiter ausgeführt werden, da Ähnliches bereits in diesem Kapitel fortlaufend diskutiert wurde) und dass der Text abgerundet wird. Der Begriff der Kohärenz soll sowohl bei B2 als auch bei C1 exakt die gleiche Beachtung bekommen. Es wird hier kein Niveauunterschied verzeichnet. Wert gelegt soll auf Verknüpfungen, gerade bei Satzanfängen, die auf Vorheriges Bezug nehmen. Die Definition und die Schwachstellen der einzelnen Deskriptoren beider Niveaus (B2/C1) wurden bereits ausführlich diskutiert. Für das Kriterium der Ausdrucksfähigkeit gilt insgesamt, dass aus den Materialien, die die Rater während ihrer Schulung bekommen, hervorgeht, dass sowohl für das Niveau B2 als auch für das Niveau C1 die gleichen Aspekte bewertet werden sollen. Definiert wird lediglich der Anspruch des Ausdrucksvermögens etwas unterschiedlich. An dieser Stelle muss aber erneut der Frage nachgegangen werden, wo die Grenze des Wortschatzinventars überhaupt zu ziehen ist und ob die Unterscheidung letztlich so offensichtlich sein kann, dass man von *differenziert/nuanciert* oder von einem *breiten Spektrum* ausgeht. In erster Linie gilt es, sich die unterschiedlichen Aufgabenstellungen für derartige Ansprüche je nach Niveau zu betrachten. Außerdem geht es hierbei um die Qualität und Quantität des Wortschatzes eines Lerners. Dabei soll der Wortschatz bei aufsteigendem Sprachniveau vom Konkreten zum Abstrakten übergehen. Während also die Aufgabenstellung für das B2-Niveau eine Reaktion auf eine Internet- oder Zeitungsmeldung erwartet, behandelt die Aufgabenstellung für das C1-Zertifikat übergreifende Themen, die unter anderem auch voraussetzen, dass man einen diskontinuierlichen Input (z.B. eine Grafik oder ein Balkendiagramm) soweit versteht, dass man darauf schriftsprachlich reagiert.

Interessant sind auch die zu beachtenden Punkte und die Vorgaben für das Kriterium Korrektheit. Der Zusatz für das C1-Niveau *derart gute Sprachbeherrschung, dass Fehler vereinzelt vorkommen,* kann lediglich die ersten zwei Deskriptoren bzw. die oberen Stufen dieses Kriteriums vollkommen decken. Rater werden weiterhin darüber aufgeklärt, dass die Fehlerzuweisung explizit und nicht doppelt erfolgen darf. Inwiefern die Unterkriterien dieses Kriteriums bei beiden Niveaus gewichtet sind, wird durch Einsicht in diese Schulungsmaterialien beantwortet: *Fehlerklassifizierung spielt eine untergeordnete Rolle*. Das wichtigste Kriterium dieses Kriteriums (!) heißt *Verständlichkeit*. Inwiefern aber Verständlichkeit immer dem Korrektheitsanspruch genügen muss, sei unbeantwortet gelassen. Wie gut etwas verstanden wird, ist eine ganz subjektive Angelegenheit. Die Rater lesen den Text, der von den Prüfungskandidaten verfasst wurde. In diesem Zusammenhang sieht Urquhart (1987:389) die Variation im Produkt des Lesens in zwei Dimensionen wirken:

- Durch die Interpretation besteht zum einen die Möglichkeit, dass es sich um Leser aus verschiedenen Kulturen handelt und zum anderen kann ein und derselbe Leser zu unterschiedlichen Zeiten mit unterschiedlichem Wissen einen Text unterschiedlich interpretieren.
- Was das Verständnis betrifft, so ergibt sich die Variation im Leseprodukt aus den unterschiedlichen Lesezielen und den damit verbundenen Lesestilen. Das Problem der Verständlichkeit könnte damit zusammenhängen, dass der Rater den Text nicht mehr liest, sondern nur noch interpretieren kann, weil aufgrund der Fehler sein Verständnis beeinträchtigt wird.

Also sind die Persönlichkeit und das Profil des Raters maßgebilch dafür, ob der Leseprozess oder das Verständnis beeinträchtigt werden. Obwohl dieses Kriterium als das objektivste gelten müsste, denn Syntax, Morphologie und Orthografie unterliegen Regeln, erweist sich auch dieses als der Subjektivität eines Raters offen gegenüber. Das Problem, das sich aus der nicht adäquaten Definition der Bewertungskriterien und ihrer Deskriptoren ergibt, verletzt die Standards des wichtigsten Gütekriteriums der Testtheorie, der Validität. Es ist eine Verzerrung, wenn das Kriterium nicht adäquat gemessen wird, was zu Punktabzug und zur Minderung des Testergebnisses führen kann. Validität sollte als die fundamentalste Erwägung in der Testentwicklung und Testevaluation angesehen werden. Die APA zieht die mehrfachen Bewertungsausführungen in Betracht, doch zu diesem Zweck muss jede beabsichtigte Interpretation valide sein. Der Testinhalt bezieht sich nach APA sowohl auf das Thema und das Aufgabenformat als auch auf die Richtlinien der Prozesse hinsichtlich des Bewertens. Grundlegend für die Testentwicklung ist es folglich zu ermitteln, ob die passenden Kriterien angewandt werden, um von einem relevanten Validitätsbeweis sprechen zu können und inwieweit dieser für verschiedene Testsituationen generalisiert werden kann. Die APA definiert hinsichtlich dieses Problems (APA-Standard 1.3:18):

„If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations".

5.4 Der TestDaF

Der „Test Deutsch als Fremdsprache" bzw. der TestDaF wurde erstmals im Jahre 2001 eingeführt. Bei der Überprüfung der vier klassischen Fertigkeiten soll die entsprechende sprachliche Kompetenzstufe des Testteilnehmers für den Hochschulzugang definiert werden. Es stellt sich in diesem Zusammenhang die Frage, inwieweit die geforderten Leistungen mit den tatsächlichen Anforderungen im Studium korrelieren und somit das Kriterium der Authentizität decken, wobei es echte Authentizität in einem Test nie zu geben scheint (vgl. Arras/Grotjahn (2002). Wie in Kapitel 2.5. bereits angeführt, lehnen sich die TestDaF-Niveaustufen (im Weiteren TDN) an die Kann-Beschreibungen der ALTE und des GER. Das Leistungsspektrum des TestDaF bewegt sich gemäß des GER im Bereich B 2.1 – C 1.2 und auf der ALTE-Skala auf der Stufe 3. Ein erheblicher Unterschied besteht zudem noch im zu überprüfenden sprachlichen Kontext. Während der GER handlungsorientiert die allgemeine Sprachfähigkeit erfasst, bezieht der TestDaF seine „Messung" auf die Sprachverwendung im akademischen Kontext.

Da es sich um eine Hochschulzugangssprachprüfung handelt, ist der TestDaF im Sinne von Qualitätskriterien wie die der APA sehr um die Konstruktion, Analyse und Evaluation bemüht. Darauf basierend sollen sich deutsche Hochschulen auf den TestDaF verlassen können, um potentielle nicht-muttersprachliche Studienanfänger anhand ihrer „ermittelten" Stärken und Schwächen in der deutschen Sprache entsprechend der jeweiligen Zulassungsordnung einzustufen, um sie dann entweder zu immatrikulieren oder aber nicht. In der Prüfungsordnung des TestDaF steht diesbezüglich unter §1 Absatz 2:[159]

„Wenn alle Teilprüfungen mindestens mit der TestDaF-Niveaustufe (TDN 4) abgelegt worden sind, gilt dies gemäß §4 Abs. 5 RO-DT als Nachweis der sprachlichen Studierfähigkeit für die uneingeschränkte Zulassung oder Einschreibung zu allen Studiengängen und Studienabschlüssen.(...) Gemäß § 1, Abs. 3, 4 und 5 in Verbindung mit § 4, Abs. 7 RO-DT können auf Beschluss der jeweiligen Hochschule für bestimmte Studienzwecke auch geringere sprachliche Eingangsvoraussetzungen festgelegt werden."

Die Entwicklung eines TestDaF durchläuft verschiedene testtheoretische Phasen, bevor dieser schließlich zur Überprüfung der verschiedenen Kompetenzen bereit gestellt wird. Interessant ist die Tatsache, dass in der Phase der Vorerprobung die Kontrollgruppe, im Gegensatz zu den potentiellen Kandidaten, aus Muttersprachlern besteht. Dadurch soll beispielsweise die Trennschärfe der Items sichergestellt werden (vgl. Kapitel 4.3.1.2). Es ist nicht ersichtlich, ob Muttersprachler die graue Masse der Durchschnittsbürger oder ebenso potentielle Studienbewerber sind, mit denen die ausländischen Prüfungskandidaten „konkurrieren". Von Bedeutung ist auch die Tatsache, anhand welcher testtheoretischen Methode der TestDaF samt seinen „Facetten" überprüft und bewertet wird. Mit dem Bewusstsein, dass Ratingverfahren in der Regel mit „Urteilsfehlern" oder „rater bias" (siehe Kapitel 4.5.1) behaftet sind, und die Interraterreliabilität meist nicht gewährleistet werden kann, bedient sich der TestDaF der Hilfe eines probabilistischen Modells, dem so genannten Multifacetten-Raschmodell (vgl. Kapitel 4.5.3). Dabei werden verschiedene Faktoren definiert, die die Leistungsbeurteilung bestimmen (Eckes 2003:57):

159 http://www.testdaf.de/teilnehmer/pdf/pruefungsordnung.pdf

- Fähigkeit der Testperson
- Schwierigkeit des Kriteriums/Schwierigkeit des Items
- Strenge der Rater

Diese drei „Facetten“ bedingen sich gegenseitig, wenn es um den Ratingprozess geht. Die Interpretation der Fähigkeit einer Testperson sollte in ihrer Einstufung adäquat sein. Was die Schwierigkeit der Kriterien oder auch der Items angeht, so implizieren diese, dass die „Streuung“ der Einstufungen daran festgemacht werden kann. Der dritte Punkt betrifft die Strenge eines Raters. Anders ausgedrückt gibt es „mildere“ und „strengere“ Rater, was dazu führt, dass man nicht von einer fairen Bewertung sprechen kann. Anhand der Multifacetten-Analyse wird für den TestDaF im Sinne des „fairen Durchschnitts“ zum einen der Strengekoeffizient eines jeden Raters ermittelt, zum anderen auch die Schwierigkeit von Kriterien und Items (Eckes 2004:501).

Was schließlich das Testformat des TestDaF angeht, so handelt es sich, wie bereits angeführt, um eine Testform, die die vier klassischen Kompetenzen (LV, HV, SA, MA) in vier verschiedenen Subtests überprüft. Laut § 2 der Prüfungsordnung „werden auf dem TestDaF-Zeugnis die Prüfungsergebnisse nach Fertigkeiten getrennt ausgewiesen, um den Hochschulen ein differenzierteres Leistungsprofil des Studienbewerbers zu vermitteln“. Die erreichten Leistungen in den jeweiligen Subtests werden den TDN-Stufen 3, 4 und 5 zugeordnet. Das Niveau „unter TDN 3“ informiert lediglich darüber, dass das Eingangsniveau dieses Tests nicht erreicht wurde (Arras/Grotjahn 2002:65).

5.4.1 Der schriftliche Ausdruck im TestDaF

Im Mittelpunkt dieser Arbeit steht der schriftliche Ausdruck und die dafür erstellten Kriterienkataloge, um „schreibproduktive Kompetenz“ zu definieren. Schriftlicher Ausdruck kann dabei den drei TestDaF-Niveaustufen (TDN) zugeordnet werden, welche generisch folgendermaßen beschrieben werden:[160]

- TDN 3: „Kann sich in studienbezogenen Alltagssituationen (u.a. Bericht für Stipendiengeber) weitgehend verständlich und zusammenhängend schriftlich äußern; kann sich im fächerübergreifenden wissenschaftlichen Kontext (u.a. Protokolle, Thesenpapiere) vereinfacht äußern, sprachliche und strukturelle Mängel können das Textverständnis beeinträchtigen.“
- TDN 4: „Kann sich in studienbezogenen Alltagssituationen (u.a. Bericht für Stipendiengeber) sowie im fächerübergreifenden wissenschaftlichen Kontext (u.a. Protokolle, Thesenpapiere) weitgehend zusammenhängend und strukturiert sowie weitgehend angemessen äußern; sprachliche Mängel beeinträchtigen das Textverständnis nicht.“
- TDN 5: „Kann sich in studienbezogenen Alltagssituationen (u.a. Bericht für Stipendiengeber) sowie im fächerübergreifenden wissenschaftlichen Kontext (u.a. Protokolle, Thesenpapiere) zusammenhängend und strukturiert sowie sprachlich angemessen und differenziert äußern.“

160 http://www.testdaf.de/teilnehmer/tn-info_nivea.php

Beim schriftlichen Ausdruck des TestDaf-Tests geht es lediglich um eine Aufgabe hoher Komplexität, die unterschiedliche Schreibhandlungen umfasst und die es anhand einer Skala mit Hilfe definierter Einzelkriterien und Deskriptoren zu bewerten gilt (Eckes 2004:486). Die schreibproduktive Leistung bzw. das Bewertungsresultat kann nicht durch die Leistungen in den anderen Subtests kompensiert werden (Arras/Grotjahn 2002:65f). Um hinter die Kulissen des TestDaF-Tests und seiner Bewertung zu schauen, habe ich am 3./4.3.2007 an dem Workshop „Beurteilung schriftlicher und mündlicher Leistungen im TestDaF“ teilgenommen, den das TestDaF-Institut in Fulda organisiert hat. Neben den einführenden Teilen wurde der schriftliche Ausdruck und dessen Bewertung zum Schwerpunkt dieses Workshops. Durch die schriftliche Lernerproduktion des TestDaF soll Aufschluss darüber gegeben werden, ob der Testkandidat „hochschultauglich“ ist. Die Prüfungsordnung definiert das Ziel dieser Teilprüfung folgendermaßen: “Der Kandidat soll zeigen, dass er Schreibhandlungen, die im hochschulbezogenen Kontext relevant sind, angemessen ausführen kann.“ Dabei werden anhand der vorgelegten Aufgabe zur schriftlichen Produktion unter anderem vom Kandidaten folgende Teilkompetenzen, Strategien und Techniken erwartet:[161]

- Logischer Aufbau und Strukturierung schriftlicher Äußerungen
- Beschreibung statistischer Daten in Grafiken oder Tabellen
- Unterscheidung zwischen Beschreibung und Interpretation
- Argumentationen entwickeln können
- Unterscheidung zwischen sachlicher Information und persönlicher Stellungnahme
- Stellung zu kontroversen Fragen zu nehmen
- verschiedene Standpunkte begründet darzustellen
- Vor- und Nachteile gegeneinander abzuwägen
- Verwendung von kohäsionsstiftend sprachlichen Mitteln
- Kompensationsstrategien für nicht bekannte Redemittel

Der TestDaF versteht sich als eine kriteriumsorientierte Prüfung, deren Ziel der Vergleich einer individuellen Leistung mit der gewünschten Fähigkeit ist. Es werden verschiedene Bewertungskriterien aufgestellt, die dabei behilflich sein sollen. Im Folgenden wird eine Aufgabenstellung des schriftlichen Ausdrucks angeführt und diskutiert. Im Anschluss daran werden Lernerproduktionen und die vom TestDaF begründete Bewertung folgen, welche dokumentiert und auf ihre Gültigkeit überprüft werden. Der Kriterienkatalog soll das Ganze abrunden, indem die Einzelkriterien zunächst separat und dann als Ganzes betrachtet und analysiert werden sollen, um diese schließlich auf die Lernerproduktion zu beziehen und anzuwenden.

161 Workshop „Beurteilung schriftlicher und mündlicher Leistungen im TestDaF“ am 3./4. 3. 2007 in Fulda

5.4.1.1 Aufgabenstellung im schriftlichen Ausdruck

Der Subtest „schriftlicher Ausdruck“ möchte Schreibfertigkeiten, die im Hochschulkontext von Bedeutung sind, überprüfen. Anhand einer Grafik, die beschrieben werden soll, wird eine Stellungnahme zu konkreten Aspekten des Themas verlangt. Anreiz dafür können Zitate oder verschiedene Statements sein.[162] Es sei die Aufgabenstellung aus dem Modellsatz 02 des TestDaF-Instituts angeführt, um den Typus, seine Thematik und die Eindeutigkeit zur Bearbeitung zu dokumentieren:

Schreiben Sie einen Text zum folgenden Thema: 60 Minuten

Wahl des Hochschulorts

Wo soll ich studieren? Diese Frage stellt sich, wenn man sich entschlossen hat, ein Hochschulstudium aufzunehmen. Ist eine große Hochschule in einer Großstadt oder aber eine Hochschule mit weniger Studierenden in einer kleineren Stadt empfehlenswert? Für die Entscheidung ist nicht nur die Attraktivität der Universitätsstadt ausschlaggebend. Auch das Betreuungsverhältnis an der Hochschule ist ein wichtiges Entscheidungskriterium, d. h. die Frage: Wie viele Studierende werden von einer Professorin oder einem Professor betreut?

Bevölkerung sowie Studierende und ProfessorInnen an ausgewählten Hochschulstandorten in Deutschland

	Anzahl Einwohner-Innen	Anzahl Student-Innen	Anzahl Professor-Innen	Betreuungs-verhältnis
Greifs-wald	54.000	6.970	160	ca. 44:1
Freiburg	205.000	17.520	320	ca. 54:1
Leipzig	493.000	24.820	405	ca. 61:1
Hamburg	1.715.400	36.140	630	ca. 57:1
Köln	963.000	60.300	410	ca. 147:1

Beschreiben und vergleichen Sie, wie sich die unterschiedlichen Universitäten

- hinsichtlich der Anzahl an Studierenden und ProfessorInnen sowie
- hinsichtlich des Betreuungsverhältnisses unterscheiden

Beziehen Sie dabei auch die Größe des Hochschulorts ein.

162 TestDaF: Bewertungsanleitung zum Modellsatz 02. Bezogen aus dem Workshop „ Beurteilung schriftlicher und mündlicher Leistungen im TestDaF“ am 3./4.3.2007 in Fulda

Was die Wahl des Hochschulorts betrifft, so gibt es unterschiedliche Meinungen:

Das Studium an einer Hochschule in einer Großstadt ist sehr viel interessanter, denn man hat dort ein reiches kulturelles Angebot und bessere Chancen, einen Nebenjob oder einen Praktikumsplatz zu finden.

Das Studium an einer kleinen Hochschule fernab der Großstädte ist effektiver, weil man sich besser auf das Studium konzentrieren kann und einen engeren Kontakt zu den Lehrkräften und Mitstudierenden pflegt.

- Geben Sie beide Aussagen mit eigenen Worten wieder.
- Nehmen Sie Stellung zu beiden Aussagen und begründen Sie Ihre Meinung.
- Gehen Sie auf die Situation in Ihrem Heimatland ein.

Das Aufgabenformat für den schriftlichen Ausdruck bei TestDaF besteht aus einer Texterstellungsaufgabe und fordert eine deskriptive und argumentative Schreibhandlung.[163] Zum einen wird ein einführender Text mit hoher Informationsdichte, im vorliegenden Fall „Wahl des Hochschulorts“, zur Verfügung gestellt, um einen Einblick in die Thematik anhand von bereit gestellten Hintergrundinformationen zu liefern. Dieser als Stimulus fungierender Einführungstext sollte auf TDN 3 platziert sein, damit alle Prüfungskandidaten im Sinne der APA die gleichen Bedingungen, folglich Fairness, haben. Darüber hinaus kommt es erst durch die Bereitstellung der Tabelle zu der eigentlichen Aufgabenstellung. Verlangt wird an dieser Stelle von den Prüfungskandidaten zunächst, die Tabelle unter bestimmten Bedingungen „zu beschreiben und zu vergleichen“. Die weiter unten angeführten Statements sollen in den Text mit eingebunden werden, indem sie mit eigenen Worten wiedergegeben, dann abgewogen und diskutiert werden sollen. Die eigene Meinung wird insofern verlangt, als man sich für das eine oder andere Statement argumentativ entscheiden muss. Auch spielt die kulturkontrastive Facette eine Rolle, indem die definierte Situation auf das jeweilige Heimatland zu beziehen ist. Ein wichtiger Punkt ist, inwieweit bei einer derartigen Aufgabenstellung für die Prüfungsteilnehmer eindeutig ist, was genau sie schriftlich produzieren sollen. Diese Frage stellt sich auf, denn das Aufgabenformat ist in verschiedene Subkategorien unterteilt, was sehr leicht dazu führen kann, dass man zum Beispiel durch den „Zeitdruck“ bedingt den Überblick verlieren kann, obwohl die bereit gestellte Information sehr detailliert ist. Interessant ist an dieser Stelle ebenso zu erwähnen, dass keine bestimmte Textsorte gefordert wird, wie das zum Beispiel beim Goethe-Institut der Fall ist.

5.4.1.2 Lernerproduktion

Im Folgenden soll eine Lernerproduktion auf die Aufgabenstellung des Modellsatzes 02 des TestDaF-Instituts aufgezeigt und anhand der zu bearbeitenden Aufgabe diskutiert werden. Daraufhin wird die dokumentierte Bewertung der Lernerproduktion des TestDaF-Instituts angeführt werden. Die Bewertungsskala und ihre Deskriptoren sollen darauf aufbauend schließlich hinsichtlich ihrer Eindeutigkeit untersucht werden, um sie des Weiteren zu analysieren und zu diskutieren.

163 Zur Bewältigung dieses Subtests stehen 60 Minuten zur Verfügung, wobei keinerlei Hilfsmittel zugelassen sind.

Lernerproduktion 1 auf die Aufgabe des schriftlichen Ausdrucks in Modellsatz 02:[164]

Die Tabelle zeigt, wie die Anzahle der Bevölkerung sowie Studierende und ProfessorInnen in verschiedenen Städten bzw. die Betreuungsverhältnis in verschiedenen Universitäten aussehen. Ganz deutlich ist es, dass es in einer Großstadt zwar viel mehr Studierende und ProfessorInnen gibt, aber die Betreuungsverhältnis ist auch relativ groß. Z.B. in Köln mit 963.000 EinwohnerInnen gibt es 60.300 Studierende und 410 ProfessorInnen. Daraus folgt die Betreuungsverhältnis 147:1. Im Gegenteil dazu ist die Betreuungsverhältnis in einer Kleinstadt viel kleiner, obwohl es dort wenigere Studierende und ProfessorInnen. Z.B. in Stadt Greifswald mit nur 54.000 EinwohnerInnen gibt es 6.970 Studierende und 160 ProfessorInnen. Die Betreuungsverhältnis ist aber 44:1.

Das Studium an einer Großhochschule in einer Großstadt ist sehr interessant. Man fühlt sich nie langweilig. Denn man kann immer viele Kontakte mit Leute aus verschiedene Kulturen lernen, sonder auch lernt, wie man mit einem anderen Mensch besser umgehen soll. Und es ist allen bekannt, dass die Kommunikation eine große Rolle spielt. Außerdem gibt es hier in einer Großstadt zahlreiche Firmen und Geschäften. Deshalb existieren über kein Problem, einen Nebenjob sowie Praktikumsplatz zu finden. Die beide Sache sind unheimlich wichtig. Mit einem Nebenjob kann ein Studierender seine finanzielle Lastung erleichtern. Bei Praktikum kann er wertvolle Erfahrungen sammeln.

Das Studium in einer Kleinstadt ist im Vergleich zu einer Großstadt viel effektiver Denn es herrscht in einer Kleinstadt immer nur Ruhigkeit. Studierende können sich für Ihr Studium richtig Mühe geben. So kann man sein Studium früher abschließen und mit seiner Karriere besser anfangen. Weil es nicht viele Leute in einer Kleinstadt gibt, kann man die Beziehungen zwischen ProfessorInnen und KomilitoneInnen gut pflegen. Und man kann schnell die Hilfe beim Studium bekommen und voneinander besser lernen. Die Situation in meinem Heimatland sieht ganz anders aus. Fast alle Studierende möchte nur in einer Großstadt gehen und studieren. Weil es nur in Großstadt die Universitäten mit hohe Ruf gibt, und die Ruf spielt eine große Rolle, deshalb möchte fast jeder an Uni mit hohe Ruf studieren.

Als aller erstes fällt auf, dass in der Aufgabenstellung nicht explizit gemacht wird, was für eine Textlänge erfordert wird. Obwohl die facettenreiche Aufgabenstellung wissenschaftliches und hochschulbezogenes Schreiben voraussetzt, definiert das TestDaF-Institut nicht das „Ausmaß" der zu überprüfenden Schreibfertigkeit in diesem Subtest. Im besuchten Workshop wurde bezüglich dieser Fragestellung ebenfalls nichts Konkretes geäußert. Wichtig sei vielmehr, ob der Lernertext den Anforderungen auf textueller Ebene genüge.[165] Demnach kann man in diesem Sinne davon ausgehen, dass der schriftliche Ausdruck nicht anhand seiner Länge und Quantität „gemessen" wird. Ich werde im Weiteren die definierten und gesetzten „Anforderungen" für den Subtest schriftlicher Ausdruck dokumentieren.

164 TestDaF, Bewertungsanleitung zum Modellsatz 02, Text 3, 10/2005. S. 16.

165 Informationen aus dem Workshop „Beurteilung schriftlicher und mündlicher Leistungen im TestDaF" am 3./4. 3. 2007 in Fulda

Global betrachtet scheint diese Textproduktion anfangs solide zu sein. Setzt man sich aber nun näher mit der Legende, dem tabellarischen Input und den angeführten Statements auseinander, so wird ersichtlich, dass der Kandidat sprachproduktiv die vorgelegten Informationen nicht strukturiert und exakt wiedergegeben hat. Ebenso scheinen die Argumentationsdarlegungen nicht ausreichend gegenüber gestellt und begründet worden zu sein.

Das TestDaF-Institut hat nicht den Anspruch eines grammatisch und orthografisch perfekten Textes. Es geht hier vielmehr um

- **die gute Textstruktur und ihre Verständlichkeit**
- **die vollständige, sachliche und folgerichtige Bearbeitung der Aufgabenpunkte**
- **das Aufzeigen, dass der Prüfungskandidat in der Lage ist auf Hochschulniveau eine schriftliche Arbeit anzufertigen.**[166]

Wie eine derartige Produktion vom TestDaF bewertet wird, soll im Weiteren demonstriert werden. Im Vorfeld sollen zunächst die zur Bewertung definierten und benötigten Kriterien angeführt und im Sinne der Validität diskutiert werden.

5.4.2 Bewertungskriterien für den schriftlichen Ausdruck im TestDaF

Für die Bewertung des schriftlichen Ausdrucks im TestDaF werden gleich gewichtete Bewertungskriterien herangezogen, die je nach TDN- Stufe (TDN 3, TDN 4 oder TDN 5) bewertet werden und jeweils aus drei Unterkriterien bestehen. Anders ausgedrückt, können die einzelnen Kriterien verschiedenen Stufen zugeordnet werden. Diese differenzierte Bewertung dient dem TestDaF dazu, Prüfungskandidaten den jeweiligen Niveaustufen zuzuordnen um daraus eine facettenreichere Sprachstandsdiagnose zu ermitteln. Die Ansprüche und Can-Dos dieser drei Stufen decken den Bereich B 2.1. – C 1.2. ab (siehe Tabelle in Kapitel 2.5.). Es gibt für die Bewertung des Subtests schriftlicher Ausdruck zunächst drei Hauptkriterien, die wie folgt definiert sind:[167]

- **Gesamteindruck**

  „Dieses Kriterium erfasst den Text als Ganzes in seiner Wirkung auf den Rezipienten bzw. auf die Rezipientin. Die Bewertung erfolgt danach, inwieweit MuttersprachlerInnen den Text flüssig lesen und dem Gedankengang folgen können".

- **Behandlung der Aufgabe**

  „Dieses Kriterium erfasst die Ausführlichkeit und Komplexität, mit der die Aufgabenstellung behandelt wird. Die Bewertung erfolgt danach, inwieweit die geforderten Schreibhandlungen elaboriert sind und auf die Punkte der Aufgabenstellung eingegangen wird".

166 http://www.testdaf.de/teilnehmer/pdf/modellsatz02/tipps02_sa.pdf

167 TestDaF-Institut: Bewertungsanleitung zum Modellsatz 02. 10/2005. S. 8

- **Sprachliche Realisierung**

„Dieses Kriterium erfasst die sprachlichen Mittel des Textes. Dazu gehören das Maß an Kohäsion und Ausdrucksfähigkeit sowie die Richtigkeit der verwendeten Sprachmittel. Die Bewertung erfolgt nach Breite, Korrektheit und Angemessenheit der eingesetzten sprachlichen Mittel“.

5.4.2.1 Das Kriterium Gesamteindruck

Das Kriterium des Gesamteindrucks scheint mit insgesamt 33,33 % holistisch und vom individuellen Eindruck eines Raters geprägt zu sein. Einwände gegen die Kriteriumsdefinition sind hinsichtlich der Wirkung auf muttersprachliche Rezipienten aus zweierlei Gründen zu erheben. Zum einen ist die Wirkung eines Textes auf eine Person nicht zu ergründen, denn wie bereits in Kapitel 4.5.3 mehrfach erwähnt wurde, ist die Wahrnehmung bzw. das Auffassungsvermögen eines jeden Raters individuell und wenn man so will, letztendlich subjektiv. Wie flüssig ein Text gelesen werden kann und ob man dem Gedankengang folgen kann, ist meiner Meinung nach ein nicht zu objektivierender Umstand. Zum anderen stellt sich die Frage was die muttersprachliche Norm auszumachen scheint. Abgesehen von all dem bekommen speziell geschulte Rater bei diesem Kriterium vom TestDaF-Institut folgende Anleitungen:[168]

- **Wie liest sich der Text? Ist er gut lesbar oder muss man manche Textstellen zweimal lesen? (*Unterkriterium: Lesefluss*)**
- **Ist der Gedankengang in Ordnung oder gibt es Widersprüche und Gedankensprünge, so dass man manchmal nicht weiß, was eigentlich ausgedrückt werden soll? (*Unterkriterium: Gedankengang*)**
- **Wie ist der Text aufgebaut? Gibt es eine Einleitung? Gibt es Überlegungen zwischen den Abschnitten, also z.B. zwischen der Grafikbeschreibung und dem argumentativen Teil? Gibt es eine Schlussfolgerung, ein Fazit? (*Unterkriterium: Textaufbau*)**

Obwohl dieses Kriterium lediglich den Gesamteindruck eruieren soll, so wird dieser trotzdem an den einzelnen Defiziten festgemacht. Das Unterkriterium „Lesefluss“ ist eine potentielle Doppelsanktionierungsquelle, denn dieser findet sich erneut bei der Kohäsion der sprachlichen Realisierung wieder. Es ist auffällig, wie die im Folgenden aufgezeigten Bewertungskriterien in ihrer Form eher analytisch sind, und zum Teil Sprachfacetten behandeln, die bereits im Kriterium „Gesamteindruck“ eine Rolle zu spielen scheinen.

168 http://www.testdaf.de/teilnehmer/pdf/modellsatz02/tipps02_sa.pdf

5.4.2.2 Das Kriterium: Behandlung der Aufgabe

Das zweite Kriterium wird „Behandlung der Aufgabe“ genannt und scheint ebenfalls mit insgesamt 33, 33 % auf die inhaltliche Realisierung der Aufgabe fokussiert zu sein. Hier geht es um die Frage, wie die gestellte Aufgabe inhaltlich bearbeitet wurde. In diesem Sinne sollten Rater auf folgende Richtlinien Acht geben:[169]

- **Sind alle Punkte der Aufgabenstellung ausreichend behandelt oder fehlt etwas? (*Unterkriterium: Punkte der Aufgabenstellung*)**
- **Wie ist die Grafik beschrieben? Sind alle wichtigen Informationen folgerichtig zusammengefasst? Oder sind die Informationen der Grafik ungenau und eventuell fehlerhaft wiedergegeben? Werden Entwicklungen aufgezeigt? Kann man die Beschreibung der Grafik verstehen? (*Unterkriterium: Beschreibung*)**
- **Wie ist der argumentative Teil bearbeitet? Sind die vorgegebenen Meinungen mit eigenen Worten wiedergegeben, oder wurden sie abgeschrieben? Sind die Argumente für oder gegen ein Problem immer begründet? Oder ist einfach nur die persönliche Meinung ohne Begründung geschrieben worden? Sind Vor- und Nachteile einer Frage begründet vorgebracht, oder nur aufgezählt. Ist der Text sachlich? Oder wurden nur ganz persönliche Ansichten vorgebracht? Ist die Situation im Heimatland verständlich beschrieben und in die Argumentation eingebaut? (*Unterkriterium: Argumentation*)**

Was das Unterkriterium der „Punkte der Aufgabenstellung“ anbelangt, so wurde bereits erwähnt, dass über das Ausmaß der schriftlichen Textproduktion nichts vermerkt scheint. Zu hinterfragen ist, wie man als Rater dennoch von „ausreichender Behandlung der Aufgabe“ ausgehen kann oder auch nicht, und worauf man diese Definition bezieht. Beim Unterkriterium „Beschreibung“ geht es um die korrekte Auffassung der Grafik. Der Fall, dass ein Prüfungskandidat die Grafik durch seine L1 zwar verstanden hat, jedoch aufgrund fehlender Sprachkompetenz Fehler an die Oberfläche gelangen lässt, sollte jedoch vorsichtig angegangen werden. Derartige Fälle führen leicht zur Doppelsanktionierung, so dass die Gefahr der falschen Bewertung einer Lernerproduktionen besteht. Interessant ist auch die Frage, was „gemessen“ wird, wenn der Schreiber selbst aus seinem muttersprachlichen Verständnis heraus die Grafik nicht verstanden hat und dementsprechend sprachlich nichts Adäquates produzieren kann. Zu betonen ist an dieser Stelle, dass das Verständnis einer Grafik eine rein kognitive und keine sprachliche Leistung ist. Wenn also die Grafik vom Prüfungskandidaten nicht verstanden wird, so kann der sprachliche Output nicht die erforderte Leistung erbringen. Leider wird hier Kognition und Verständnis einer sprachunabhängigen Komponente (in unserem Fall die Grafik) mit der Sprachproduktion gleichgesetzt. Wichtig ist es deshalb die Kriterien untereinander sehr strikt zu trennen, so dass die Einstufung von Leistungen bzw. Fehlleistungen der richtigen Kategorie zugeordnet werden können. Im Unterkriterium „Argumentation“ wird auf die Möglichkeit verwiesen, dass ein Prüfungskandidat den Wortschatzinput der Legende und der Grafik für seine Produktion übernehmen kann, diese jedoch in aller Regel nicht ausreichend ist. Es wird aber nichts darüber ausgesagt, was die Konsequenz daraus ist und wie Derartiges schließlich zu bewerten ist.

169 http://www.testdaf.de/teilnehmer/pdf/modellsatz02/tipps02_sa.pdf

5.4.2.3 Das Kriterium: sprachliche Realisierung

Mit den letzten 33, 33 % der Gesamtbewertung des schriftlichen Ausdrucks soll das Kriterium der sprachlichen Realisierung Aufschluss darüber geben, welche sprachlichen Mittel eine Textproduktion eines TestDaF aufweist. Rater sollen dabei in diesem Sinne allgemein auf folgende Fragestellungen Rücksicht nehmen:[170]

- **Sind die Sätze im Text miteinander verbunden, d. h. ist der Text kohärent? Wird stets nur „und" verwendet oder werden auch andere Konjunktionen benutzt? Variieren die Konjunktionen sinnvoll? (*Unterkriterium: Kohäsion*)**
- **Werden immer die gleichen einfachen Sätze (z. B. Hauptsätze) geschrieben oder auch Nebensätze verwendet? Werden immer die gleichen Nebensätze geschrieben oder variieren die Konstruktionen? (*Unterkriterium: syntaktische Strukturen*)**
- **Wie breit und genau ist der Wortschatz? Werden z. B. immer die gleichen Verben benutzt oder variiert der Wortschatz? Werden die treffenden Ausdrücke benutzt? (*Unterkriterium: Wortschatz*)**
- **Wie viele sprachliche Fehler gibt es in dem Text? Treten oft Fehler auf, oder nur manchmal? Kann man den Text trotz einiger Fehler noch verstehen? Oder kann man ihn wegen der Fehler nicht immer verstehen? (*Unterkriterium: Korrektheit*)**

Bei dem Unterkriterium Kohäsion der sprachlichen Realisierung geht es um kohäsive Mittel. Dabei wird aber zunächst nicht deutlich, welche Verknüpfungselemente jede einzelne TDN definieren. Es zeigt sich wiederum, dass die Annahme, dass diese und jene sprachlichen Mittel eine bestimmte Stufe ausmachen würden, nicht gelten kann. Diese Problematik lässt sich eher dadurch klären, wenn man darauf achtet, welche sprachlichen Mittel und wie diese im Gesamten eingesetzt werden. Die Referenzebene dafür ist damit stets der Text. Es stellt sich die Frage hinsichtlich der Benutzung kohäsiver Mittel und der Einstufung dessen auf den TDN-Stufen 3, 4 oder 5. Gleiches gilt für die syntaktischen Strukturen. Es gilt ebenso zu klären, wie viel Komplexität Sätze aufweisen müssen, dass man entsprechende Stufenzuweisung erlangt. Was den Wortschatz anbelangt, so ist ein sehr großes Gebiet angesprochen, dass meines Erachtens vollkommen subjektiv im Empfinden der einzelnen Rater ist. Es wurde bereits angesprochen, dass Wortschatz nicht kategorisiert werden kann, auch wenn aufsteigend eine Wortschatzvariation erwartet wird.

170 http://www.testdaf.de/teilnehmer/pdf/modellsatz02/tipps02_sa.pdf

5.4.2.4 Der Kriterienkatalog

Um einen besseren Eindruck darüber zu bekommen, wird im Folgenden der ausführliche Bewertungskatalog für den schriftlichen Ausdruck des TestDaF-Instituts vorgestellt:[171]

	TDN 5	TDN 4	TDN 3	Unter TDN 3
Gesamt-eindruck	1. Der Text liest sich durchgängig flüssig 2. Der Gedanken-gang kann problemlos nachvollzogen werden 3. Der Text ist klar strukturiert	1. An einzelnen Stellen gerät der Lesefluss ins Stocken. 2. Der Gedan-kengang kann nachvollzogen werden, wenn auch vereinzelt die Rezeption verzögert wird. 3. Der Text ist insgesamt noch strukturiert.	1. An manchen Stellen ist wiederholtes Lesen erforderlich. 2. Der Gedan-kengang kann von einem kooperativen Leser nachvollzogen werden. 3. Der Text weist Brüche auf.	1. Der Text liest sich insgesamt nicht flüssig. 2. Der Gedan-kengang kann nur mühsam oder bruchstückhaft nachvollzogen werden. 3. Der Text ist nicht klar strukturiert.
Behandlung der Aufgabe	*Der Text wird der Aufgabenstellung inhaltlich gerecht:* 1. Alle in der Aufgabenstellung genannten Punkte werden in ausreichendem Umfang behandelt 2. Die Infor-mationen der Grafik(en) werden zusammen-gefasst; sie werden klar und folgerichtig dargestellt 3. Im argumen-tativen Teil wird	*Der Text wird der Aufgabenstellung inhaltlich weitgehend gerecht:* 1. Alle in der Aufgabenstellun g genannten Punkte werden behandelt, manche jedoch zu knapp. 2. Die Informationen der Grafik(en) werden klar und folgerichtig wiedergegeben. 3. Im argumen-tativen Teil wird sachlich begründet, z.T. nur knapp, und	*Der Text wird der Aufgabenstellung inhaltlich noch gerecht:* 1. Fast alle in der Aufgabenstellung genannten Punkte werden behandelt. 2. Die Informationen der Grafik(en) werden überwiegend aufzählend wiedergegeben. 3. Im argumen-tativen Teil werden Standpunkte/Über legungen deutlich und ggf. Durch persönliche	*Der Text wird der Aufgabenstellung inhaltlich nicht gerecht:* 1. Nur einige in der Aufgabenstellung genannte Punkte werden behandelt. 2. Die Beschreibung der Grafik(en) ist nicht verständlich. 3. Im argumen-tativen Teil werden Standpunkte/ Überlegungen nicht oder nur in Ansätzen verdeutlicht.

171 TestDaF-Institut: Bewertungsanleitung zum Modellsatz 02. 10/2005. S. 11

	sachlich und ausführlich genug begründet und ggf. Werden Beispiele als Belege angeführt.	ggf. Werden Beispiele als Belege angeführt.	Wertungen verstärkt.	
Sprachliche Realisierung	*Die sprachliche Realisierung ist der Aufgabenstellung angemessen:* 1.Der Text hat - ein breites Spektrum an kohäsions-stiftenden Mitteln - ein breites Spektrum an syntaktischen Strukturen 2.Der Wortschatz ist weitgehend differenziert und präzise. 3.Der Text enthält vereinzelt morpho-syntaktische, lexikalische und orthografische Fehler.	*Die sprachliche Realisierung ist der Aufgabenstellung weitgehend angemessen:* 1. Der Text hat ein begrenztes Spektrum an kohäsions-stiftenden Mitteln ein begrenztes Spektrum an syntaktischen Strukturen 2. Der Wortschatz ist breit, teilweise jedoch nicht präzise 3. Der text enthält gelegentlich (nicht-systematische) morphosyntaktische, lexikalische und orthografische Fehler, die das Verstehen jedoch nicht beeinträchtigen	*Die sprachliche Realisierung ist der Aufgabenstellung nicht immer angemessen:* 1. Der Text hat - einfache Verknüpfungs-elemente - einige Variationen bei den syntaktischen Strukturen 2. Der Wortschatz ist ausreichend 3. Der text enthält morpho-syntaktische, lexikalische und orthografische Fehler, die das Verstehen deutlich erschweren.	*Die sprachliche Realisierung ist der Aufgabenstellung nicht angemessen:* 1. Der Text hat - kaum Verknüpfungs-elemente - nur wenige Variationen bei den syntaktischen Strukturen 2. Der Wortschatz ist eingeschränkt. 3. Der Text enthält morpho-syntaktische, lexikalische und orthografische Fehler, die das Verstehen deutlich erschweren.

Tabelle 39: Kriterienkatalog für den TestDaF

Ich habe bereits die einzelnen Kriterien separat vorgestellt. Im Weiteren sollen die Kriterien bzw. Einzelkriterien anhand ihrer Deskriptoren diskutiert werden. Dabei werde ich horizontal vorgehen, indem die Definitionen der Einzelkriterien je nach TDN-Stufe aufgezeigt werden. Das Stufenniveau deckt den Bereich TDN 5 bis einschließlich „unter TDN 3“. Letzteres bringt die nicht ausreichende Sprachkenntnis für den Hochschulzugang mit sich und kann daher meines Erachtens als das Extrem der maximal zu erreichenden TDN 5-Stufe angesehen werden.

5.4.2.4.1 Das Kriterium Gesamteindruck

Als erstes sollen die Unterkriterien bzw. Einzelkriterien des holistischen „Gesamteindrucks“ betrachtet werden. Das erste Einzelkriterium nennt sich „Lesefluss“ und es stellt sich die Frage, was darunter zu verstehen ist und ob dieses Kriterium „universal“ fungieren kann. Man könnte „Lesefluss“ als den ungehinderten Verlauf während des Lesens oder des Rezipierens definieren. Dieser kann aber meines Erachtens nach nicht objektiviert bzw. standardisiert werden. Auf Stufe TDN 5 sollte sich „der Text durchgängig flüssig lesen“. Anders ausgedrückt, hängt die Einstufung dieses Kriteriums auf dieser Skala mit der Persönlichkeit des Raters zusammen. Ist „sein“ Lesefluss „durchgängig flüssig“, dann bekommt der Prüfungskandidat das Niveau TDN 5 zugewiesen. Bei einem anderen Rater könnte sich dieser Umstand ganz anders ausdrücken und derselbe Kandidat bekäme eine andere Stufenzuweisung, z.B. TDN 4, da „an einzelnen Stellen der Lesefluss ins Stocken gerät“. Es gilt zu definieren, durch welche Faktoren Lesefluss ins Stocken geraten kann. Eine unleserliche Handschrift kann jedem Rater Probleme bereiten. Lesefluss kann auch durch „Übermüdung“ oder mangelnde Konzentration eines Raters in einer Art und Weise beeinträchtigt werden, was aber eine gravierende Verletzung der Reliabilität wäre. Der vorletzten Stufe wird eine Lernerproduktion dann zugeordnet, wenn „an manchen Stellen wiederholtes Lesen erforderlich ist“. Dennoch ist nicht klar, weshalb wiederholtes Lesen und von welchem Rater erforderlich zu sein hat und ob dies schließlich zum erwünschten Verständnis führt. An dieser Stelle könnte man die Frage stellen, ob beim Stocken nicht automatisch bestimmte Textpassagen erneut gelesen werden, auch wenn der Störfaktor nur ein einziges Wort ist. Schließlich erreicht das sprachliche Niveau eines Prüfungskandidaten, das am Eindruck der Rater festgemacht wird, wenn „sich der Text insgesamt nicht flüssig liest“ lediglich das Prädikat „unter TDN 3“.

Zu dem holistischen Kriterium des Gesamteindrucks gehört auch das Einzelkriterium „Gedankengang“. Wenn Rater das Urteil abgeben, dass „der Gedankengang problemlos nachvollzogen werden kann“, dann impliziert dieser Umstand TDN-Stufe 5. Auch hier stellt sich erneut die Frage, ob alle Rater dasselbe Urteil über eine Lernerproduktion hinsichtlich des Gedankenganges abgeben würden.[172] Einer Stufe darunter, d.h. TDN 4, wird das schriftliche Lernerkonstrukt zugeordnet, wenn „der Gedankengang zwar nachvollzogen werden kann, auch wenn vereinzelt die Rezeption verzögert wird“. Bei

172 Eckes (2008) stellt in seinem Aufsatz „Rater types in wring performance: a classification approach to rater variability“ fest, dass das Verhalten der Rater ganz individuell und unabhängig von gesetzten Normen und Bewertungskriterien vonstatten geht. In Kapitel 6 werde ich ausführlicher darauf eingehen.

dem einen Rater mag dieses zutreffen, beim nächsten allerdings nicht. Interessant ist der Deskriptor dieses Einzelkriteriums für die Stufe TDN 3: „Der Gedankengang kann von einem kooperativen Leser nachvollzogen werden". Prinzipiell bedeutet dies zunächst, dass es kooperative und nicht kooperative „Leser" gibt, die den schriftlichen Ausdruck bewerten sollen. Es mag durchaus sein, dass die Definition „kooperativ" etwas Anderes impliziert, was ich aber nur bezogen auf den Rater deuten kann. Bereits dieser Umstand scheint die Profile der einzelnen Rater zu differenzieren. Dennoch ist es wichtig zu definieren, was einen „kooperativen Leser" ausmacht. Wenn man davon ausginge, er „denkt mit", dann stellt sich sofort die nächste Frage auf, ob dies nur auf Stufe TDN 3 zwingend erforderlich ist. Konträr dazu muss aber auch berücksichtigt werden, ob man dem Gedankengang eines Prüfungskandidaten nicht folgen könnte, wenn man kein „kooperativer Leser" ist. Die letzte Niveauzuweisung definiert das Einzelkriterium Gedankengang insofern, als dieser „nur mühsam oder bruchstückhaft nachvollzogen werden kann". Obwohl nicht explizit ist, ob der Gedankengang für alle Rater „mühsam" ist oder nicht, wird an dieser Stelle mit „unter TDN 3" bewertet.

Das dritte und letzte Einzelkriterium des Oberkriteriums „Gesamteindruck" nennt sich Textaufbau, auch wenn der TestDaF keine bestimmte Textsorte voraussetzt. Es wird sich noch bestätigen, dass der Textaufbau erneut im Unterkriterium Kohäsion des Bereichs der sprachlichen Realisierung, Anwendung und Berücksichtigung findet. Es soll zunächst betrachtet werden, wie sich „Textaufbau" holistisch auf den Ratingprozess auswirkt. Der TDN 5 werden diejenigen Prüfungskandidaten zugeordnet, deren „Text klar strukturiert ist". Es ist natürlich auch an dieser Stelle nicht ersichtlich, für wen ein Text „klar strukturiert" ist und folglich so definiert wird. Eine Stufe darunter ist „der Text insgesamt noch strukturiert". Auch hier ist es meiner Meinung nach eine Ermessensfrage, was man als „insgesamt noch strukturiert" definiert. Jegliche Struktur ab diesem Moment scheint laut Kriterienkatalog und Deskriptoren nicht auszureichen, um die sprachlichen Voraussetzungen für ein Hochschulstudiumsbeginn zu erfüllen. Die TDN-Stufe 3 definiert den Textaufbau nämlich damit, dass „der Text Brüche aufweist". Es wird aber nichts darüber ausgesagt, wie Brüche sich äußern und welcher Art sie sind. Die eigentlich nicht-existente Stufe „unter TDN 3" besagt das absolute Gegenteil von TDN 5: „Der Text ist nicht klar strukturiert". Auch hier stellt sich die Frage, ob es eine bestimmte einzuhaltende Form gibt, selbst wenn das Format in diesem Subtest überhaupt nicht definiert werden kann.

Das Kriterium „Gesamtausdruck" macht samt seinen drei bereits angeführten Einzelkriterien 1/3 der Gesamtbewertung des schriftlichen Ausdrucks für den TestDaF aus. Es handelt sich um ein holistisch geprägtes Kriterium ist, d. h. die Rater haben bei diesem Kriterium trotz der definierten Deskriptoren ihren eigenen Ermessensspielraum. Es kann folglich nicht von einer objektiven Bewertung ausgegangen werden, da Rater durch dieses Kriterium ihren persönlichen Gesamteindruck preisgeben, indem sie sich an die jeweiligen Deskriptoren der einzelnen Abstufungen halten. Diese Deskriptoren sind aber nicht objektiv, denn ob ein Text flüssig ist, man dem Gedankengang des Testteilnehmers folgen kann oder der Text eine Struktur aufweist, ist eine vollkommen subjektive Einschätzung. Der TestDaF erlaubt Ratern demnach anhand dieses holistischen Kriteriums zu einem Drittel der Gesamtbewertung, willkürlich oder auch nicht, über die schriftliche Leistung eines Testteilnehmers zu „bestimmen" oder gar zu „entscheiden".

5.4.2.4.2 Das Kriterium „Behandlung der Aufgabe"

Das zweite analytische Kriterium auf der Bewertungsskala schriftlichen Ausdrucks des TestDaF besteht ebenfalls aus drei Einzelkriterien. Insgesamt sollen die Einzelkriterien aus verschiedenen Perspektiven der Frage nachgehen, ob der Text einen Bezug zur Aufgabenstellung aufweist. In jeder TDN-Stufe unterliegen diese drei Einzelkriterien zusammengefasst einem anderen Motto. Darauf werde ich abschließend noch eingehen. Das erste Einzelkriterium behandelt die Thematik „Punkte der Aufgabenstellung". Die in der Aufgabenstellung gegebenen Aufgaben sollen innerhalb von maximal 60 Minuten bearbeitet werden. Diese Realisierung wird ihren Abstufungen entsprechend den Niveaus TDN 5 bis „unter TDN 3" zugeordnet. Hat ein Prüfungskandidat „alle in der Aufgabenstellung genannten Punkte in ausreichendem Maße behandelt", so impliziert dieses TDN 5. Zu definieren wäre an dieser Stelle der Begriff „ausreichend", wenn selbst das TestDaF-Institut keinen Aufschluss über die zu erbringende quantitative Leistung bzw. die Wortanzahl geben kann. Es gibt Kandidaten, die drücken sich knapper aus als andere, aber dafür „kompakter" und „effizienter". Dennoch stellt sich die Frage, ob Derartiges ausreicht, um das Prädikat TDN 5 zu bekommen. Wahrscheinlich würde eine derartige Leistung der TDN-Stufe 4 zugerechnet, denn zwar wären „alle in der Aufgabe genannten Punkte behandelt, manche jedoch zu knapp". Interessant ist die Definition dieses Einzelkriteriums für TDN 3: „Fast alle in der Aufgabenstellung genannten Punkte werden behandelt". Hier wird nichts von der Quantität der zu bearbeitenden Punkte gesagt, lediglich, dass „fast alle Punkte" bearbeitet wurden, wie viele es sind, geht aber hieraus nicht hervor. Abgerundet wird dieses Einzelkriterium auf der Bewertungsskala damit, dass „nur einige in der Aufgabenstellung genannten Punkte behandelt werden" (unter TDN 3). Auch hier ist weder die Rede davon, wie viele Punkte behandelt werden oder nicht, noch von der „ausführlichen Aufgabenbearbeitung". Es ist doch im Rahmen des Möglichen, dass trotzdem Teile der Aufgabenstellung laut der Definition des TestDaF „ausführlich" behandelt werden.

Das zweite Einzelkriterium nennt sich „Beschreibung". Die Deskriptoren auf TDN 5 und TDN 4 hierfür lassen kaum Unterschiede erkennen. Während auf TDN 5 „die Informationen der Grafik(en) zusammengefasst und klar und folgerichtig dargestellt werden", definiert der Deskriptor TDN 4 „lediglich", dass „die Informationen der Grafik(en) klar und folgerichtig wiedergegeben werden". Man fällt demnach auf Niveau TDN 4 ab, sobald man die Informationen der Grafik(en) lediglich wiedergibt, aber nicht zusammenfasst. In dieser Hinsicht muss man sich auf die Aufgabenstellung besinnen, um zu eruieren, ob dieses überhaupt abverlangt wird. Bezüglich der Grafikbeschreibung wird in der Aufgabenstellung unseres Modellbeispiels folgendes verlangt:

Beschreiben und vergleichen Sie, wie sich die unterschiedlichen Universitäten

- *hinsichtlich der Anzahl an Studierenden und ProfessorInnen sowie*
- *hinsichtlich des Betreuungsverhältnisses unterscheiden*

Beziehen Sie dabei auch die Größe des Hochschulorts ein.

Der Arbeitsauftrag besagt nichts über eine Zusammenfassung der Grafikdaten bzw. -informationen. Der Prüfungskandidat soll lediglich unterschiedliche Universitäten hinsichtlich verschiedener Punkte „beschreiben“ und „vergleichen“. Es wird aus der Aufgabenstellung ebenso wenig deutlich, ob „alle“ Universitäten untereinander „beschrieben und verglichen“ werden sollen. Der Deskriptor des nicht ausreichenden Niveaus TDN 3 definiert in diesem Zusammenhang, dass „die Informationen der Grafik(en) überwiegend aufzählend wiedergegeben werden“. Was diese Stufe von einer schriftlichen Lernerproduktion abverlangt, ist nicht explizit und bleibt zu klären. Man muss davon ausgehen, dass „aufzählend wiedergegeben“ nicht „beschreiben“ oder „vergleichen“ bedeutet. Erneut stellt sich auch an dieser Stelle die Frage, wie eine derartige Grafik in ihrer Beschreibung anzugehen ist bzw. was vom Testteilnehmer hinsichtlich dessen erwartet wird. Der Deskriptor der letzten Stufe findet „die Beschreibung der Grafik(en) nicht verständlich“. Wenn etwas nicht verständlich ist, dann hat dieses zum einen mit Wortschatz, sprachlichen Mitteln und Korrektheit und zum anderen damit zu tun, dass die Grafik falsch interpretiert bzw. erst gar nicht verstanden wurde. Dennoch bleibt zu klären, wie im Einzelkriterium „Beschreibung“ dennoch von Nicht-Verständlichkeit ausgegangen werden kann. Wenn es darum geht, dass etwas nicht verstanden wird, dann könnte das genauso als holistisch betrachtet werden. In dem Fall würde sich dieses Einzelkriterium in dieser Abstufung mit dem holistischen Unterkriterium Gedankengang decken. Das dritte Unterkriterium bei „Behandlung der Aufgabe“ nennt sich „Argumentation“. Der argumentative Teil ist hier Inhalt der jeweiligen Deskriptoren. Dabei wird in der Aufgabenstellung des vorgestellten Modellsatzes folgendes von den Testteilnehmern abverlangt:

Was die Wahl des Hochschulorts betrifft, so gibt es unterschiedliche Meinungen:

Das Studium an einer Hochschule in einer Großstadt ist sehr viel interessanter, denn man hat dort ein reiches kulturelles Angebot und bessere Chancen, einen Nebenjob oder einen Praktikumsplatz zu finden.

Das Studium an einer kleinen Hochschule fernab der Großstädte ist effektiver, weil man sich besser auf das Studium konzentrieren kann und einen engeren Kontakt zu den Lehrkräften und Mitstudierenden pflegt.

- ***Geben Sie beide Aussagen mit eigenen Worten wieder.***
- ***Nehmen Sie Stellung zu beiden Aussagen und begründen Sie Ihre Meinung.***
- ***Gehen Sie auf die Situation in Ihrem Heimatland ein.***

Die verschiedenen Realisierungen dieser Aufgabenstellung haben die Deskriptoren des Argumentationskriteriums zum Gegenstand. Das Prädikat TDN 5 bekommt ein Testkandidat dann, wenn „im argumentativen Teil sachlich und ausführlich genug begründet wird und ggf. Beispiele als Belege angeführt werden“. Auch hier entpuppt sich die Definition der Begründung als vage, wenn sie „ausführlich genug“ ist und wer schließlich darüber entscheiden darf, ob „ausführlich genug“ argumentiert worden ist oder nicht. Einen entscheidenden Einschnitt gibt es zum TDN 4-Deskriptor. Während die TDN-Stufe 5 vom Testteilnehmer erwartet, dass die Argumentation „sachlich und ausführlich genug“ begründet wird, bedarf es auf der Basis dieser Prüfung auf TDN 4, der sprachlichen Zulassungsvoraussetzung für deutsche Hochschulen, lediglich der „sachlichen und z.T. knappen“ Begründung im argumentativen Teil. Die Definitionsdiskrepanz zwischen TDN 5 und TDN 4 ist meines Erachtens unabhängig von der Argumentationsweise groß. Interessant ist die Definition des TDN 3-Deskriptors: „Im argumentativen Teil werden Standpunkte/Überlegungen deutlich und ggf. durch persönliche Wertungen verstärkt“. Es ist anzunehmen, dass man dieser Stufe zugeordnet wird, weil man nicht sachlich, sondern subjektiv wertet. Dieser stark positiv definierte Deskriptor wird jedoch lediglich der Stufe TDN 3 zugerechnet. Unter Niveau TDN 3 bewegt sich ein Prüfungsteilnehmer dann, wenn „im argumentativen Teil Standpunkte/Überlegungen nicht oder nur in Ansätzen verdeutlicht werden“. Auch wenn man eine Argumentation in Ansätzen durchführt, wird dies der Nicht-Behandlung argumentativer Vorgehensweise gleichgesetzt.

Interessant ist bei dem Kriterium „Behandlung der Aufgabe“, dass bei jeder TDN-Stufe eine Vorgabe für alle drei Unterkriterien gemacht wird. Anders ausgedrückt, definiert TDN 5 für die Einzelkriterien „Punkte der Aufgabenstellung“, „Beschreibung“ und „Argumentation“ die Überschrift: „Der Text wird der Aufgabenstellung inhaltlich gerecht“. Für TDN 4 lautet das Äquivalent: „Der Text wird der Aufgabenstellung inhaltlich weitgehend gerecht“. TDN 3 liegt unter der Basis und die Einzelkriterien müssen folgendes erfüllen: „Der Text wird der Aufgabenstellung inhaltlich noch gerecht“. Bei dem als letztes definierten Niveau auf der Bewertungsskala „unter TDN 3“ „wird der Text der Aufgabenstellung inhaltlich nicht gerecht“. Es stellt sich an dieser Stelle die Frage, inwieweit denn die Leistung in diesen drei Einzelkriterien gleich sein muss. Es ist durchaus denkbar und zu erwarten, dass ein Prüfungskandidat beim Unterkriterium „Beschreibung“ hervorragend abschneidet (TDN 5), im argumentativen Teil aber den Erwartungen nicht gerecht wird. Folglich können verschiedene Variationsmöglichkeiten innerhalb der einzelnen Einzelkriterien existieren, was ihre erforderte und schließlich erbrachte Leistung anbelangt.

5.4.2.4.3 Das Kriterium: sprachliche Realisierung

Das letzte gleichwertige und analytisch zu bewertende Kriterium des Bewertungskatalogs des TestDaF ist ebenfalls durch drei Einzelkriterien definiert. Das erste Einzelkriterium heißt „sprachliche Mittel“ und beinhaltet die Thematik der Kohäsion und der syntaktischen Strukturen. Mit TDN 5 wird eine Lernerproduktion dann bewertet, wenn der Text „ein breites Spektrum an kohäsionsstiftenden Mitteln und syntaktischen Strukturen“ aufweist. Es eröffnet sich aber direkt die Frage, was „breit“ bedeutet und wie breit „breit“ sein kann. Auf TDN 4 benötigt eine Lernerproduktion lediglich „ein begrenztes Spektrum an kohäsionsstiftenden Mitteln und syntaktischen Strukturen“. Es bleibt zu klären, worauf

sich die Begrenztheit des Spektrums bezieht, wenn zum Beispiel kohäsionsstiftende Mittel und syntaktische Strukturen verwendet werden, die lediglich der Stufe TDN 4 zuzuweisen sind. Obwohl eine Sprachstandsprüfung immer ein kleiner Ausschnitt aus der Kompetenz eines Testteilnehmers ist, ist es interessant zu eruieren, ob eine Lernerproduktion letztlich etwas darüber aussagen wird. Dennoch kann man ein gezeigtes „begrenztes Spektrum“ sprachlicher Mittel nicht als falsch definieren. Es wird in den ersten zwei Deskriptoren natürlich nichts über Fehler hinsichtlich dessen erläutert. Fehler sind aber in diesem Fall latent, wenn man „bestimmte“ sprachliche Mittel, die der Testanbieter oder der Rater selbst erwartet, nicht verwendet. Bei Deskriptor für TDN 3 hat „der Text einfache Verknüpfungselemente und einige Variationen bei den syntaktischen Strukturen“, was nicht ausreichend ist und unter der insgesamt benötigten Basis von TDN 4 liegt. Man kann an dieser Stelle lediglich die Vermutung anstellen, dass einfache Verknüpfungselemente zum Beispiel eher die so genannten „adusos“[173] sind, die Hauptsätze miteinander verbinden. Was die Variation in der syntaktischen Struktur betrifft, so wird nicht definiert auf welchem textlinguistischen Niveau sich dieser Vorgang bewegt. Es kann auch von komplexeren syntaktischen Strukturen mit wenig Variation die Rede sein.[174] Unter die TDN-Stufe 3 gelangt man, wenn „kaum Verknüpfungselemente und nur wenige Variationen bei den syntaktischen Strukturen“ auftreten. Der Testkandidat benutzt diesem Deskriptor zufolge demnach einfache Hauptsätze, denen sich beispielsweise ein Relativsatz anhängt. Es zeichnet sich bei diesem Kriterium für den TestDaF bereits ab, dass es im Sinne der Komplexität der Sprache und ihrer Produktion auf Unterschiede im syntaktischen Sprachgebrauch bezieht. Im 6. Kapitel werde ich die Thematik der Komplexität beschreiben. Es wird deutlich werden, wie schwierig es ist, dass menschliche Rater derartig Komplexes bewerten können.

Das zweite Einzelkriterium in dieser Kategorie ist der Wortschatz. Schon der Übergang von Deskriptor TDN 5 zu Deskriptor TDN 4 ist meines Erachtens sehr gegensätzlich definiert. Während ein Testkandidat aus der Sicht des Wortschatzes TDN 5 erreicht, wenn dieser „weitgehend differenziert und präzise ist“, wird das Wortschatzspektrum auf TDN 4 zwar als „breit aber als teilweise nicht präzise“ definiert. Es stellt sich die Frage, ob der verwendete Wortschatz auf seine kontextuelle Anwendung hin untersucht wird. Man müsste hier unter anderem nach „Register“ bewerten, wenn es um „akademische bzw. hochschulbezogene Kontexte“ geht. Natürlich kann man aber auch beim TestDaF nicht erwarten, dass die Lernerproduktion aus rein akademischem Vokabular besteht, da der TestDaF lediglich Bezug zu allgemeinen Kommunikationssituationen aus dem Hochschulleben herstellt, die aber ohne Fachbezug sind.[175] Selbst dieses scheint aber ein schwieriges Vorhaben zu sein. Bereits unter 5.1.2.3. wurde diskutiert, welche Probleme die Bewertung des Wortschatzes mit sich bringt. Auch in diesem Zusammenhang scheint die Bewertung dieses Einzelkriteriums in der Willkür der Rater zu liegen. Im Gegensatz zur „breiten aber teilweise nicht präzisen“ Wortschatzverwendung auf TDN 4 ist diese auf TDN 3 „ausreichend“. Man muss direkt fragen: ausreichend wofür? Da diese Stufe erst einmal nicht ausreichend ist, um die sprachliche Hochschulzugangsberechtigung zu erlangen, steht dieser Deskriptor konträr dazu und muss konkretisiert werden. Mit der Note „ausreichend“ wird der Grenzbereich „Basis“ definiert. Wenn der TDN 3-Deskriptor

173 Die Konnektoren aber, denn, und, sondern, oder

174 Der Begriff der Komplexität wird im 6. Kapitel ausführlich angeführt werden.

175 Vgl. Krekeler, C. (2005): Grammatik und Fachbezug in Sprachtests für den Hochschulzugang. Dissertationsschrift. Universität Duisburg Essen. http://dueplico.uni-duisburg-essen.de/servlets/DocumentServlet?id_12458

den aktivierten Wortschatz für ausreichend befindet, dann wird er dem Anspruch gerecht. Der letzte Deskriptor für „unter TDN 3“ benennt den Wortschatz als „eingeschränkt“. Auch an dieser Stelle können Überlegungen angestellt werden, worauf Bezug genommen wird. Eine mögliche Interpretation wäre, dass der Wortschatz auf den akademischen Kontext bzw. die „Hochschultauglichkeit“ bezogen eingeschränkt ist. Er wird demnach dem Anspruch nicht gerecht. Wenn diese Definition korrekt und akzeptabel wäre, dann könnte der ausreichende Wortschatz auf TDN 3 ebenso auf den Hochschulkontext bezogen werden. In diesem Zusammenhang kann von einer Fehldefinition bzw. Kompetenzverschiebung auf der Skala gesprochen werden, denn mit diesem Deskriptor der Skala TDN 3 wird nichts erreicht.

Das letzte Einzelkriterium im Bereich der sprachlichen Realisierung nennt sich Korrektheit und wird ebenso analytisch bewertet. Der höchsten Stufe gehört man diesbezüglich an, wenn die Textproduktion eines Prüfungskandidaten „vereinzelt morphosyntaktische, lexikalische und orthografische Fehler“ enthält. Welcher Art die Fehler sind, scheint an dieser Stelle nicht definiert zu sein.[176] Möglicherweise schließt dieser Deskriptor eventuell auch „schwere“ Fehler ein, die vereinzelt auftreten. Daraus kann es aber auch an anderen Stellen zu Sanktionen kommen, was aber keineswegs eintreten darf.

Die Stufe TDN 4 erreicht man, wenn der Text „gelegentlich (nicht-systematische) morphosyntaktische, lexikalische und orthografische Fehler enthält, die das Verstehen jedoch nicht beeinträchtigen“. Erst an dieser Stelle im Bewertungsraster dieses Unterkriteriums wird der Zusatz „Verstehensbeeinträchtigung“ geliefert. Nach meinem Verständnis gibt es keinen großen Unterschied zwischen Deskriptor TDN 5 und TDN 4, denn:

a) Man kann nicht messen was wann „vereinzelt“ oder „gelegentlich“ ist

b) Der Zusatz „Verstehen wird jedoch nicht beeinträchtigt“ liefert in diesem Sinne keinerlei neue Information. Auch auf dieser Stufe wird das Verständnis „nicht“ beeinträchtigt.

Im Deskriptor TDN 3 geht es bereits um die Beeinträchtigung des Verstehens. Es wird aber nicht ersichtlich, ob es schließlich um die Rezeption eines jeden Raters geht. Über die Quantität der Fehler wird nichts besagt, es geht lediglich um den Umstand, dass das Verstehen beeinträchtigt wird. Auch auf welche Art und Weise das Verstehen beeinträchtigt wird, kann diesem Deskriptor nicht entnommen werden. Die letzte Niveauzuweisung spricht in diesem Zusammenhang von Fehlern, „die das Verstehen deutlich erschweren“. Auch hier gilt zu klären, um wessen Verstehensprobleme es geht und welche Fehler dazu führen. Es ist nach meinem Verständnis ersichtlich und logisch begründbar, dass nicht „die gleichen“ Fehler bei „allen“ Ratern die „gleichen“ Verstehensprobleme mit sich führen. Demnach besteht auch an dieser Stelle erneut eine verkappte „holistische“ Bewertung, denn diese Umstände sind im Auge des Betrachters zu suchen und zu finden

176 Auf dem von mir besuchten Workshop des TestDaF-Instituts am 3./4. 2007 in Fulda wurde unter anderem angesprochen, dass zu Korrektheitsfehlern unter anderem falsche Artikel oder falsche Endungen zählen.

Im Folgenden soll die Bewertung der bereits vorgestellten Lernerproduktion angeführt werden. Es soll diskutiert werden wie Rater des TestDaF-Instituts diese Textproduktion im Rahmen der standardisierten Prüfung eingestuft haben.

5.4.3 Bewertung einer schriftlichen Textproduktion

Die Aufgabenstellung des Modellsatzes 02 wurde bereits unter Kapitel 5.4.1.1 angeführt und dokumentiert. Die Lernerreaktion darauf ist im Folgenden im Kapitel 5.4.1.2. dargestellt worden (Text 3). An dieser Stelle soll die Bewertung dieser Textproduktion betrachtet werden. Es ist nicht ganz klar, ob die anzuführende Bewertung die Summe aus zwei Raterurteilen oder nur das Ergebnis eines einzelnen Raters ist. An keiner Quelle konnte das explizit festgemacht werden. Dieser Umstand soll aber an dieser Stelle unberücksichtigt gelassen werden. Hauptaugenmerk soll sein, wie und ob selbst ein einzelner Rater die „Vorschriften" des TestDaF-Instituts einhält bzw. verzerrt. Ich werde die Kriterien samt ihren Einzelkriterien jeweils separat anführen und bezüglich ihrer Bewertung erörtern.

Gesamteindruck

Einzelkriterien	TDN	Begründung / Beispiele
1. Lesefluss	4	Missverständlicher Gebrauch sprachlicher Mittel, z.B. 3 ff. Zwar…aber…auch (?), 8 obwohl (=weil?), 11 aber (?), 16 über (=überhaupt?) Fehlerhafter Satzbau sowie andere Fehler beeinträchtigen den Lesefluss gelegentlich
2. Gedankengang	3	Kooperation notwendig, um dem Gedankengang folgen zu können, da die Abschnitte inhaltlich zusammenhangslos bleiben. Irritation aufgrund fehlender Markierung der Fremdmeinung (12ff & 20 ff). Dies führt dazu, dass bei der Lektüre der Eindruck entsteht, es handele sich in beiden Fällen um seine/ihre- freilich widersprüchliche- Meinung. 26 sieht ganz anders aus- wieso?; 26 f und 27 Großstadt/Ruf (Wiederholung)
3. Textaufbau	3	Einleitung und Überleitungen fehlen, allerdings kann der erste Satz als Einführung in die zu beschreibende Tabelle betrachtet werden. Text weist Brüche auf; keine Verbindung der einzelnen Abschnitte (10, 20). Überleitung zur Situation im Heimatland ist zwar sprachlich markiert und hat somit kohäsive Funktion (26 sieht ganz anders aus), inhaltlich/logisch jedoch unklar: Anders als wo? Kein Bezug zu dem zuvor Gesagten (-> Gedankengang).

Tabelle 40: Kriterium Gesamteindruck im TestDaF

Ich habe bereits meine Ansicht darüber ausgedrückt, dass das holistische Bewertungskriterium „Gesamteindruck" den Ratern einen großen Ermessensspielraum gewährt. Je nach Auffassungsvermögen und Profil urteilt jeder Rater individuell. Dabei besteht aber hinsichtlich des Bewertungskatalogs die Gefahr, dass teilweise doppelt bewertet wird, denn in diesem holistischen Kriterium sind Elemente der analytischen Kriterien bzw. Unterkriterien inbegriffen. Interessant ist auch die Information, die ich am teilgenommenen Workshop des TestDaF-Instituts bekommen habe:[177] „Rater dürften während der Bewertung nichts auf den Textproduktionen vermerken bzw. korrigieren, um den Halo-Effekt so gut es geht auszuschließen. Aus diesem Grund werde ihnen eine Tabelle gereicht, in der sie für jedes Einzelkriterium ihre Bewertungen anhand von Begründungen samt Beispielen notieren können".

Dem Einzelkriterium des Leseflusses wird für vorliegende Lernerproduktion das Prädikat TDN 4 gegeben. Der Rater vermerkt, welche sprachlichen Mittel und syntaktische Strukturen „seinen" Lesefluss behindern. Dieser holistische Eindruck findet sich aber erneut in dem Kriterium der sprachlichen Realisierung, im Unterbereich Kohäsion und syntaktische Strukturen wieder. Das zweite Einzelkriterium „Gedankengang" wird hier auf der Stufe TDN 3 definiert. Der Rater vergibt aus seiner Sicht demnach ein Prädikat, das unter dem erforderlichen Niveau für die Universitätszulassung liegt. Der Vermerk „Kooperation notwendig, um dem Gedankengang folgen zu können" lässt die Frage aufkommen, mit wem der Rater zu kooperieren vermag. Unverständlich ist ebenso die Irritation des Raters bezüglich der Meinungsmarkierung. Auch das wird im analytischen Unterkriterium „Argumentation" wiederholt vermerkt und folglich entscheidend bei der Bewertung sein. Die Vermutung, dass der Gesamteindruck schließlich an einzelnen „Fehlern" festgemacht wird, die dann analytisch erneut angeführt werden, bestätigt sich erneut. Es stellt sich die Frage, was unter Gesamteindruck zu verstehen ist, wenn es als holistisches Kriterium gilt. Das dritte Unterkriterium „Textaufbau" wird ebenfalls holistisch bewertet. Dabei ist nicht offensichtlich, ob es um die Textkohärenz geht, die bereits an dieser Stelle eruiert werden soll, obwohl das Kriterium der sprachlichen Realisierung diese sicherlich unter „sprachliche Mittel" mit einbezieht.

Die Bewertung hinsichtlich des Textaufbaus bei vorliegender Textproduktion wird lediglich mit der Definition der holistischen TDN-Stufe 3 „der Text weist Brüche auf" bewertet. Diese Feststellung kann sich an dieser Stelle jedoch nicht auf ein Textformat beziehen, da das TestDaf-Institut darüber explizit nichts besagt und fordert. Kommentiert wird diesbezüglich dass „Einleitungen und Überleitungen fehlen" oder dass „einzelne Abschnitte nicht verbunden sind". Im Unterkriterium „sprachliche Mittel" des Kriteriums „sprachliche Realisierung" wird der Bereich der Kohäsion aber eigenständig behandelt.

177 Informationen aus dem Workshop „Beurteilung schriftlicher und mündlicher Leistungen im TestDaF" am 3./4. 3. 2007 in Fulda

Behandlung der Aufgabe

Einzelkriterien	TDN	Begründung / Beispiele *(Die inhaltliche Umsetzung wirkt „abgearbeitet", kein diskursiver Text*
1. Punkte der Aufgabenstellung	4	Fremdmeinungen paraphrasiert, jedoch nicht als solche markiert; die Stellungnahme fehlt bzw. bleibt implizit; insgesamt zu kurz behandelt; die Aufgabe wird „abgearbeitet"
2. Beschreibung	3	Thema der Grafik wird genannt; Städte werden als Beispiele angegeben, jedoch lückenhaft; Hamburg fehlt z.B. als Gegenbeispiel Die Daten der Grafik werden aufzählend und nicht vollständig genannt, falsche Kohäsionsmittel verfälschen die Aussage (s. auch Lesefluss)
3. Argumentation	3	Meinungen gut paraphrasiert, zu beiden Positionen werden Argumente angeführt. Insgesamt jedoch irritiert, dass keine Markierung der Fremdmeinungen erfolgt. Dadurch bleibt schließlich unklar, welche Haltung er/sie vertritt, eigene Meinung bleibt unklar (s. Gedankengang). Daher findet kein Abwägen statt, die Schreibhandlung „argumentieren" ist nur teilweise umgesetzt. Situation im Heimatland ist nicht in Argumentation eingebunden.

Tabelle 41: Kriterium Behandlung der Aufgabe im TestDaF

Das nächste Kriterium, nach dem bewertet werden soll, ist die „Behandlung der Aufgabe". Hier soll „die Ausführlichkeit und Komplexität, mit der die Aufgabenstellung behandelt wird" erfasst werden. Bewertet wird danach, „inwieweit die geforderten Schreibhandlungen elaboriert sind und auf die Punkte der Aufgabenstellung eingegangen wird"[178]. Insgesamt wird für vorliegende Lernerproduktion „die inhaltliche Umsetzung als abgearbeitet empfunden, so dass kein diskursiver Text zustande kommt". Wie sich dies konkret auf die drei Unterbereiche dieses Kriteriums äußert werde ich im Folgenden dokumentieren. Zunächst wird nicht allen Unterkriterien dasselbe Prädikat zugeordnet. Für die „Punkte der Aufgabenstellung" wird die TDN-Stufe 4 vergeben, die durch den Deskriptor „alle in der Aufgabenstellung genannten Punkte werden behandelt, manche jedoch zu knapp" vertreten wird. Der Rater kommentiert, dass der Prüfungskandidat Fremdmeinungen paraphrasiert, dass seine eigene Stellungnahme fehlt bzw. implizit bleibt und dass im Großen und Ganzen die Aufgabe abgearbeitet wirkt. Die Gegenüberstellung der beiden Stellungnahmen, die in dieser Lernerproduktion eingebunden ist, indem Bezug auf die Wichtigkeit eines „Nebenjobs" oder einer „Praktikumsstelle" genommen wird, ist der Bewertung zufolge nicht explizit genug. Das nächste Unterkriterium „Beschreibung" untersucht die Behandlung der Grafik und ihrer Informationen. Diese Lernerproduktion wird hinsichtlich dieses Unterkriteriums auf TDN 3 eingestuft. Das wird damit begründet, dass das Thema der Grafik und Städte zwar

178 TestDaF, Bewertungsanleitung zum Modellsatz 02. 10/2005. S. 8

genannt werden, jedoch lückenhaft und zudem aufzählend sind. Wichtig ist bei diesem Kommentar der Hinweis, dass durch „falsche Kohäsionsmittel" die Aussage verfälscht wird und zudem noch auf das Unterkriterium des Leseflusses hingewiesen wird. Weshalb an dieser Stelle auf Kohäsion eingegangen wird, ist nicht angemerkt, obwohl diese im Kriterium sprachliche Realisierung berücksichtigt wird. Die Gefahr der Doppelsanktionierung bestätigt sich in hiesigem Fall, denn der Gebrauch der Kohäsionsmittel wird in drei verschiedenen Kriterien erwähnt, die unabhängig voneinander sind. Irreführend ist für mich zudem der Kommentar für das Unterkriterium Argumentation des Deskriptors, der die TDN-Stufe 3 definiert. Der Deskriptor, dem die argumentative Leistung zugeordnet wurde, definiert: *Im argumentativen Teil werden Standpunkte/Überlegungen deutlich und ggf. durch persönliche Wertungen verstärkt*. In der Diskussion der einzelnen Deskriptoren wurde bereits hinterfragt und darauf eingegangen, wieso gerade dieser Deskriptor der nicht ausreichenden Stufe TDN 3 zugeteilt wird. Im vorliegenden Fall gibt es sowohl einen Widerspruch zwischen Kommentar und Deskriptor, als auch innerhalb des Kommentars selber. Während der Bewerter „Meinungen als gut paraphrasiert und zu beiden Positionen Argumente angeführt" sieht, ist er insgesamt jedoch irritiert, was die eigene oder die fremde Meinung innerhalb der Lernerproduktion anbelangt. Die „Irritation" hat der Bewerter bereits im holistischen Unterkriterium Gedankengang (Oberkriterium: Gesamteindruck) in die Bewertung einbezogen. Der Widerspruch, dass gut paraphrasierte Meinungen zu Irritationen führen, bleibt unklar.

Sprachliche Realisierung

Einzelkriterien	TDN	Begründung / Beispiele
1. Sprachliche Mittel	4	
• Kohäsion	4	Verknüpfungselemente sind vorhanden, werden jedoch oft falsch angewendet (s. Lesefluss): 7, 13, 14, 23, 28
• Syntaktische Strukturen	3	Einerseits oft Hauptsätze, andererseits begrenzte Variationsbreite (dass, weil, denn, nicht nur…sondern auch)
2.Wortschatz	4	Breit, aber häufig nicht ganz treffend verwendet, jedoch i. d. R. verständlich (s. Korrektheit: Lexikfehler): 14 Kontakt aufnehmen, 15 verschiedene Kulturen, 17 Kommunikation, 26 Karriere Nicht immer angemessen: 10 unheimlich, 25 richtig
3. Korrektheit	3	Wiederholt morphosyntaktische Fehler, die das Verstehen z.T. beeinträchtigen: 14f kennt…lernen, 18 existiert über kein Problem, 21 finanzielle Lastung erleichtern Lexik: 24 Ruhigkeit (=Übergeneralisierung), 7 im Gegenteil dazu (statt Gegensatz)

Tabelle 42: Kriterium sprachliche Realisierung im TestDaF

Es soll zunächst auf die verschiedenen sprachlichen Mittel eingegangen werden. Was die Kohäsion betrifft, so wird im Kommentar auf die Existenz von Verknüpfungselementen hingewiesen, diese „jedoch oft falsch angewendet". Dennoch wird entgegen des Kommentars das Prädikat TDN 4 dafür vergeben. Der Deskriptor dieser Stufe für das Kriterium der Kohäsion besagt nichts über die richtige oder falsche Anwendung. Es wird, wie im Vorfeld bereits diskutiert, lediglich auf die Begrenztheit kohäsionsstiftender Mittel Bezug genommen. Der Kommentar und seine Stufenzuordnung ist insofern nicht nachvollziehbar. Prinzipiell existieren Verknüpfungselemente, jedoch werden diese, dem Kommentar entsprechend, oft falsch angewendet. Die Frage ist an dieser Stelle, ob entgegen der Kompatibilität des Deskriptors und des Kommentars, eine derartige Feststellung als „angemessen" betrachtet werden kann. Im Bereich der syntaktischen Strukturen wird das Prädikat TDN 3 vergeben, da es zum einen zwar Hauptsätze gibt, diese aber begrenzt variiert werden. Der Deskriptor dieser Stufe lautet: „Der Text hat einige Variationen bei den syntaktischen Strukturen". Hingegen könnte anhand dieses Kommentars auch der Deskriptor der TDN 4 geltend gemacht werden, wo „der Text ein begrenztes Spektrum an syntaktischen Strukturen" aufweist. Insgesamt werden Kohäsion und syntaktische Strukturen aufsummiert und für das Kriterium sprachlicher Mittel erlangt diese Lernerproduktion die TDN-Stufe 4.

Das Kriterium des Wortschatzes wird für die vorliegende schriftliche Lernerproduktion als „breit, aber häufig nicht ganz treffend verwendet, jedoch in der Regel verständlich" dokumentiert und mit TDN 4 honoriert. Der Deskriptor dieser Stufe definiert einen breiten Wortschatz, der teilweise jedoch nicht präzise ist. Der Zusatz im Kommentar „häufig nicht ganz treffend, jedoch in der Regel verständlich" ist erneut ein Widerspruch, der aber für die Stufenzuordnung nicht maßgeblich zu sein scheint, denn es wird speziell auf Lexikfehler verwiesen. Hier wird nämlich Prädikat TDN 3 vergeben, denn aufgrund wiederholter morphosyntaktischer und lexikalischer Fehler wird das Verstehen zum Teil beeinträchtigt.[179] Die Definition des Deskriptors bezieht sich jedoch auf die Beeinträchtigung des Verstehens, ohne zu verdeutlichen in welchem Maße und bei wem dies eintritt. Die Definitionen der einzelnen Deskriptoren und die Schwierigkeit der Zuordnungen wurden im Vorfeld bereits ausführlich diskutiert.

Fazit

Ich habe in der vorangegangenen Diskussion des TestDaF versucht, die Schwachstellen und die Widersprüchlichkeit innerhalb der Kriterien und dem Auffassungsvermögen von Korrektoren anzuführen. Was nun die vorgestellte Lernerproduktion und die Bewertung anhand der neun Kriterien betrifft, so kann mit bloßem Auge das Gesamtergebnis der einzelnen Stufenzuordnungen je nach Kriterium erfolgen. Das TestDaF-Institut kalkuliert die Einzelergebnisse der neun Bewertungskriterien, die unabhängig voneinander sind, auf der Basis testmethodischer Berechnungen, um die Stufe der Kompetenz des schriftlichen Ausdrucks anzuzeigen.[180] Benutzt wird hierbei die Multifacetten-Raschanalyse, der ein probabilistisch testtheoretisches Modell zugrunde liegt. Dabei will dieses testmethodische Konstrukt eine „möglichst objektive und präzise Information über die Elemente der

179 Im 6. Kapitel sollen die Problematik und Schwierigkeit von Kategorisierungen von Lexik und Morphologie im Sinne ihrer ganzen Komplexität aufgezeigt werden

180 TestDaF, Bewertungsanleitung zum Modellsatz 02. 10/2005. S. 8

betrachteten Facetten gewinnen"[181]. Dazu zählen die Leistungsfähigkeit der beurteilten Personen, die Strenge der Beurteiler und die Schwierigkeit der Aufgaben bzw. der Kriterien. Eckes (2006) stellt das Multifacetten-Korrekturverfahren den traditionellen Messverfahren (Drittkorrekturverfahren und arithmetisches Mittelungsverfahren) gegenüber. Während die traditionellen Messverfahren die Rohdaten, d. h. die Bewertungen der Rater, unmittelbar für die Stufenzuweisung verwenden, werden im Multifacetten-Korrekturverfahren die Bewertungen der einzelnen TDN-Stufen nach den einzelnen Kriterien kalibriert. Somit sollen Aussagen über die Strenge bzw. Milde der einzelnen Bewerter gemacht werden, um faire Durchschnitte berechnen zu können.[182] Ich werde das genaue Vorgehen des TestDaF-Instituts nicht weiter ausführen, da sich mein Fokus in erster Linie in den gesetzten Bewertungskriterien, ihrer Definition und schließlich ihrer Validität liegt. Die Probleme, die sich hieraus ergeben, sind angeführt worden und auch die Facette der Rater wurde in Kapitel 4.5.3 ausführlich dokumentiert.

181 http://www.testdaf.de/html/publikationen/pdffiles/Eckes_FaDaF_Essen.pdf, S. 12, Zugriff am 25.10.2006

182 http://www.testdaf.de/html/publikationen/pdffiles/Eckes_FaDaF_Essen.pdf, S. 25, Zugriff am 25.10.2006

6 Resümee und Ausblick

In diesem abschließenden und auf einen weiterführenden Gedanken weisenden Kapitel soll das bisher Erarbeitete zunächst zusammengefasst werden. Ausblickend soll hier nun der Fokus nicht mehr auf die testtheoretische Betrachtungsweise gelegt werden. Es wurde im Laufe dieser Arbeit an vielen Stellen deutlich, dass anhand der testtheoretischen Modelle die Schwierigkeit, Komplexitäten einer Sprache zu eruieren, nicht einfach zu bewältigen ist. Deshalb soll weiterführend eine linguistische Betrachtungsweise aufgezeigt werden, die exemplarisch die Problematik der Sprachkomplexität aufzeigt. Anhand des Kriteriums der Korrektheit soll dies verdeutlicht werden. Außerdem wird das alternative Bewertungssystem des griechischen Staatszertifikats vorgestellt und kurz umrissen. Zusammenfassend werden die verschiedenen und wichtigsten Facetten der in der Praxis üblichen Bewertung schriftlicher Lernerproduktionen im Sinne der Validität nochmals aufgezeigt.

Zur Aufgabe dieser Arbeit gehört die Untersuchung der Validität der Bewertungskriterien für den schriftlichen Ausdruck von Lernerproduktionen auf den Niveaus B2 und C1 der GER. Messick (1983:13) definiert: „Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment".

Für die Thematik „Bewertungskriterien schriftlicher Lernerproduktionen B2/C1 und ihre Validität" sind verschiedene Bereiche bearbeitet worden. Zu Anfang wurde der Gemeinsame Europäische Referenzrahmen für Sprachen vorgestellt, der unter anderem als Basis für die Erstellung von Sprachprüfungen gilt. Dabei wurden seine Kompetenz- und Niveauzuschreibungen für die Zwecke dieser Arbeit erläutert. Im Anschluss daran habe ich die American Psychological Association (APA) angeführt, auf die ich mich in der gesamten Arbeit als Referenzrahmen hinsichtlich testtheoretischer und testpraktischer Fragen und Gegebenheiten berufe. Zusätzlich wird die Association of Language Testers in Europe (ALTE) knapp umrissen. Die Arbeit und Zwecke der Testanbieter, deren Bewertungskriterien für den schriftlichen Ausdruck in dieser Arbeit untersucht werden, stehen in den Kapiteln 2.4 und 2.5 im Mittelpunkt. Das dritte Kapitel beschäftigt sich mit Modellen des Spracherwerbs, wobei von der generellen Theorie zu den spezifischeren für den Fremd- und Zweitspracherwerb übergegangen wird. Außerdem bildet der schriftliche Ausdruck in der Fremdsprache das Unterkapitel 3.4, dem die Definition des Kompetenzbegriffs folgt (Kap. 3.6). Eine wichtige Grundlage für die Thematik dieser Dissertation sind Testtheorie und ihre Gütekriterien. Das 4. Kapitel bildet zudem noch den Schwerpunkt der Ratingverfahren und der menschlichen Rater. Das eigentliche und zentrale Kapitel ist das 5. Kapitel. Hier werden die einzelnen Prüfungen für das B2- und C1-Niveau vorgestellt. Zentraler Punkt ist der schriftliche Ausdruck. Dabei werden sowohl die Aufgabenstellungen als auch die zugrunde liegenden Bewertungskriterien vorgestellt, erörtert und wo nötig im Sinne der Validität kritisiert.

In diesem Kapitel soll die Arbeit abgerundet werden, indem selektive Schwachstellen nun am Beispiel fundierter Ansätze der Komplexität im Mittelpunkt stehen. Hauptziel ist es Verbesserungsvorschläge zu machen, um der Validität so nah wie möglich zu kommen. Wie gemessene Leistungen interpretiert werden, ist die Kernaussage der Validität. Ich habe in dieser Arbeit die Bewertungskriterien des Goethe-Instituts und des TestDaF-Instituts für den schriftlichen Ausdruck zunächst vorgestellt, dokumentiert, kritisiert und schließlich anhand originaler Interpretationen diskutiert. Zu betonen ist, dass diese Arbeit nicht empirisch fundiert ist, sodass die Resultate lediglich einer kleinen authentischen Stichprobe unterliegen. Es handelt sich vielmehr um eine kritische Auseinandersetzung, wobei die einzigen Daten, die mir zur Verfügung standen, lediglich die jeweiligen Bewertungsraster der Testanbieter waren. Insofern habe ich versucht aus der simplen Betrachtung der Bewertungsraster die Schwachstellen herauszuarbeiten. Zusätzliche aus internen Quellen erhaltene Informationen habe ich zwar mit eingeflochten, jedoch bin ich der Ansicht, dass derartige latente Informationen, die nicht in den Rastern beinhaltet sind, unberücksichtigt gelassen werden müssten. Für die Bewertung schriftlicher Lernerproduktionen muss ein System geschaffen werden, dass keiner weiteren Anweisung und zusätzlicher interner Richtlinien bedarf, um der Validität möglichst nah zu kommen. Ziel dieser Arbeit ist die höchstmögliche Validität zu eruieren und ggf. Vorschläge zu machen, auf welche Art und Weise diese erreicht werden kann.

Die Validität eines Tests kann natürlich bereits im Vorfeld beeinträchtigt werden, wenn das zu messende Konstrukt anhand der Testentwicklung und der Aufgaben nicht entsprechend repräsentiert und abgebildet wird. Im Sinne dieser Arbeit bedeutet dies, dass eine fremdsprachliche Leistung in Abhängigkeit vom Kontext und vom Aufgabentyp variieren kann. Daher gilt es genau festzulegen, wodurch das zu messende Konstrukt und worauf bezogen repräsentiert wird. Am Anfang dieser Dissertation wurde diesbezüglich der GER vorgestellt, der seit 2001 für Sprachzertifizierungen in Europa als Referenzrahmen fungiert. Demnach müssen sich Testanbieter im Bereich von Sprachprüfungen daran orientieren und den definierten Kann-Beschreibungen (Can-Dos) der jeweiligen Niveaus Rechnung tragen, wenn es darum geht, Tests bzw. Prüfungen zu erstellen. Dabei sollen die Aufgaben auf das zu prüfende Niveau abzielen. Im Sinne der Validität müssen die gesetzten und definierten Bewertungskriterien in erster Linie im Zusammenhang mit der Aufgabenstellung bzw. dem angestrebten Niveau stehen. Weiterhin gilt zu überdenken, ob die definierten und bereits vorgestellten in der heutigen Praxis üblichen Bewertungskriterien diesen Ansprüchen gerecht werden. Was zu fehlen scheint, ist ein linguistischer Ansatz. Dazu zählt die Komplexität einer Sprache, die sich in verschiedenen Bereichen wie z.B. im Wortschatz, in den syntaktischen Strukturen und in der Morphologie äußert. Den Begriff der Komplexität gebraucht der TestDaF nach Eckes hinsichtlich seiner Anforderung der schriftlich zu erarbeitenden Aufgabe, die verschiedene Schreibhandlungen abverlangt, dabei bleibt aber unbeantwortet, wie diese zu verstehen ist.[183] Nach Edmonds wird Komplexität auf abstrakt-theoretischer Ebene definiert „als die Eigenschaft eines Modells, das es schwierig macht, das gesamte Verhalten in einer gegebenen Sprache zu formulieren, auch wenn die gesamte angemessene Information über Teilkomponenten und ihrer Beziehung zueinander gegeben ist"[184]. Anhand dieses allgemeingültig theoretischen Ansatzes wird ersichtlich, dass Sprache derart komplex ist, dass sie sehr schwer zu erfassen ist. Somit wird sie in

183 Vgl. S. 156 ff in dieser Arbeit

184 In Zusammenhang dieser Arbeit bezieht sich der abstrakte Sprachbegriff von Edmonds auf die zu prüfende Fremdsprache, in hiesigem Fall Deutsch

Teilkomponenten aufgebrochen. Das Goethe-Institut und das TestDaF-Institut unterteilen Sprache in diesem Sinne folgendermaßen:[185]

Goethe-Institut	TestDaF-Institut
• Inhaltliche Vollständigkeit	• Gesamteindruck a) Lesefluss b) Gedankengang c) Textaufbau
• Textaufbau und Kohärenz	• Behandlung der Aufgabe a) Punkte der Aufgabenstellung b) Beschreibung c) Argumentation
• Ausdrucksfähigkeit	• Sprachliche Realisierung a) Kohäsion b) Syntaktische Strukturen c) Korrektheit
• Korrektheit	

Tabelle 43: Überblick der Kriterien beim Goethe-Institut und TestDaF- Institut

Unabhängig davon, ob es sich um analytische oder holistische Bewertungssysteme handelt, bilden diese unterschiedlich definierten Teilkomponenten für die Testanbieter das Abbild des schriftlichen Ausdrucks ab. Dabei muss aber deren Interaktion und Zusammenhang gewährleistet werden. Man stelle sich das Abbild als ein Puzzle vor, das erst durch das Verbinden der zugehörigen Teile zum Vorschein kommt. Somit sollte die Gesamtheit die Summe der sie ausmachenden Teile sein. Im Sinne Edmonds gilt es die Komplexität auf ein Sprachmodell anzuwenden. Dabei wird sie von Nicht-Wissen unterschieden.[186] Folglich gilt es, ein Bewertungssystem zu definieren, das das Ziel der Prüfung zum Inhalt hat. Im Mittelpunkt der vorliegenden Arbeit ist das zu messende Kriterium die Kompetenz im schriftlichen Ausdruck. Tschirner (2001:121ff.) definiert in diesem Zusammenhang die verschiedenen Komponenten als „Basis für die Bewertungssysteme“, wobei diese so bestimmt sein müssen, dass sie „am besten unterschiedliche Niveaustufen unterscheiden können“. Zur Veranschaulichung dieser Definition soll das Kriterium „Korrektheit“ des Goethe-Instituts des B2-Zertifikats, das sich aus den Komponenten Morphologie, Syntax, Interpunktion und Orthografie

185 Es wurde bereits ausführlich zu diesen analytischen bzw. holistischen Bewertungsrastern Bezug genommen. An dieser Stelle werden die gesetzten Kriterien nochmals angeführt, um den Begriff der Komplexität hinsichtlich von Sprache zu definieren und auszuführen.

186 Im Zusammenhang dieser Arbeit könnte das als Can-NOT-Do definiert werden.

zusammen setzt, angeführt werden. Die Zusammensetzung dieses Kriteriums stellt bereits die Konstruktvalidität in Frage. Syntax und Morphologie sollten keineswegs mit Interpunktion und Orthografie vermengt werden.[187] Diesbezüglich stellt sich die Frage, was man mittels derartiger Bewertungskriterien zu messen vermag, wenn z.B. nicht von syntaktischer oder morphologischer Komplexität die Rede ist, sondern lediglich von Deskriptoren, die anhand von mehr oder weniger feststellbaren Fehlern entsprechende Punktzuordnungen erlauben. Fehler werden beim Goethe-Institut an der Verständnisstörung und der Verständnisbeeinträchtigung festgemacht.[188]

KRITERIUM IV Korrektheit	4 Punkte	3 Punkte	2 Punkte	1 Punkt	0 Punkte
*Morphologie *Syntax *Orthografie, Interpunktion	kaum feststellbare Fehler	Einige deutliche Fehler, die das Verständnis aber nicht beeinträchtigen	Einige Fehler, die den Leseprozess stellenweise behindern	Unzählige Fehler, die das Verständnis erheblich stören	Unzählige Fehler, die das Verständnis unmöglich machen

Tabelle 44: Kriterium Korrektheit des B2 Zertifikats des Goethe-Instituts

Wenn eine zu messende bzw. zu bewertende Lernerproduktion richtige aber einfache syntaktische Strukturen nach dem Prinzip Subjekt-Prädikat-Objekt anwendet, bekommt sie gemäß der gesetzten Deskriptoren dieses Kriteriums die maximale Punktzahl. Anders ausgedrückt kommt es bei einer derartigen Lernerproduktion für das Niveau B2 zu einem Punktabzug, obwohl der höher zu erwarteten Komplexität nicht gerecht wird. Es stellt sich die Frage, wie dem entgegen eine komplexere aber fehlerbehaftete Produktion bewertet würde, die diesem Niveau eher entspräche. Nach Chomskys Generativer Grammatik wird syntaktische Komplexität bei Crystal (1991:151) als der Bezug innerhalb eines Satzes und als der Bezug von Satz zu Satz definiert:

„(…) a generative grammar is a set o FORMAL RULES which PROJECTS a finite set of sentences upon the potentially infinite set of sentences that constitute the language as a whole, and it does this in an EXPLICIT manner, ASSIGNING to each a set of STRUCTURAL DESCRIPTIONS (…)”.

Das DESI-Projekt bezieht bei der Betrachtung grammatischer Strukturen, deren Umfang und das Maß an Korrektheit, mit dem sie eingesetzt werden, ein. Es werden bei diesem Bewertungssystem, das eine semi-kreative schriftliche Produktion bewerten soll, neben anderen, verschiedene Satzmuster, wie Hypotaxe und Parataxe, und Flexionsphänomene berücksichtigt.[189] Auch das TestDaF-Institut zeigt bei der

187 Auf Seite 119 ff vorliegender Arbeit wurde diese Problematik erläutert. Auch wenn Profile die Bereiche Grammatik, Orthografie und Interpunktion unter dem Oberbegriff „Korrektheit“ zusammen fasst, wird nichts über die Vereinbarung dieser Elemente erwähnt.

188 Goethe-Zertifikat B2: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707. S. 7

189 Aus dem Word-Dokument des DESI-Kodierhandbuchs von 2004, S. 9 (Heringer, Personalkommunikation)

sprachlichen Realisierung im Unterkriterium syntaktische Strukturen Ansätze, wenn Rater den Fragen nachgehen sollen, ob „immer die gleichen einfachen Sätze (z. B. Hauptsätze) geschrieben oder auch Nebensätze verwendet werden bzw. immer die gleichen Nebensätze geschrieben oder die Konstruktionen variieren“[190].

Rater bzw. Bewerter können im Sinne der Testtheorie nicht besser sein als das ihnen zur Verfügung gestellte Bewertungsraster, das ihnen als Messinstrument dienen soll. Wenn Rater unbeachtet der definierten Kriterien bewerten, dann wird automatisch die Kriteriumsvalidität verletzt, unabhängig davon, ob sie besser oder auch objektiver als das zugrunde liegende Bewertungsraster in ihrer Bewertung sind. Das definierte Bewertungssystem eines jeden Testanbieters stellt in der Kriteriumsorientierung die Basis für die Konstruktvalidität dar. Diesbezüglich wird das anhand der Bewertungskriterien definierte Konstrukt als gegeben betrachtet. Insgesamt gilt es ein Bewertungssystem zu definieren, das das Ziel der Prüfung zum Inhalt hat. Dabei ist Komplexität nach Edmonds die globale Charakteristik eines Modells, die relativ zur angewandeten Sprache, zur Identifikation der Komponenten und dem allgemeinen Verhalten (Verb, Valenz, morphologische Komplexität etc.) ist. Es gibt verschiedene Komplexitätsarten, die auf verschiedenen Schwierigkeitsgraden basieren und somit den Abstand zwischen Wissen der einzelnen Elemente und dem Gesamtwissen definieren. Sobald Schwierigkeiten vorhanden sind, so ist sicher auch Komplexität gegeben. Wenn man die abstrakte Ebene der Komplexität von Edmonds[191] auf die Bewertungskriterien bezieht, indem man den pragmatischen Ansatz der Komplexität von Cutler (1983) berücksichtigt, dann muss zunächst eruiert werden, ob in den einzelnen zugrunde liegenden Bewertungskriterien und deren Deskriptoren von Komplexität ausgegangen werden kann. Dennoch muss man ganz nüchtern ins Auge fassen, dass Komplexität von Menschen kaum erfassbar ist, lediglich von einem Automaten.[192] Unbeachtet dieser Tatsache könnten die Deskriptoren daraufhin betrachtet werden, ob sich Edmonds abstrakte Komplexitätsdefinition in den Bewertungskriterien in irgendeiner aufgespalteten Form deskriptiv äußert. Cutler setzt sich in ihrem Aufsatz „Lexical Complexity and Sentence Processing“ mit der semantischen und morphologischen Komplexität auseinander. Die Komplexität lexikalischer Repräsentationen kann durch verschiedene Dimensionen zum Vorschein kommen: die semantische, die syntaktische und die morphologische. Cutler (1983:43) betont in diesem Zusammenhang: “(…) The existence of complex representations of all three types has been specifically claimed:

- Both (all) interpretations of an ambiguous word are always activated
- Idioms are stored and accessed as lexical items
- Morphological decomposition is involved in the storage and retrieval of lexical items”

190 TestDaF-Institut, Bewertungsanleitung zum Modellsatz 02. 10/2005, S. 8

191 Edmonds Dissertationsschrift unter: http://66.102.1.104/scholar?hl=de&lr=&safe=off&q=cache:QZZeMCkzfyIJ:demo.cs.brandeis.edu/~pablo/papers/edmon99-56.pdf+syntactic+complexity+pdf+-informa

192 Vgl. Edmonds (S.86) In diesem Zusammenhang weist Edmonds mittels Chomskys Hierarchie darauf hin, dass ein Computerprogramm fähig ist, selbst höchste Komplexität zu durchschauen: „(…) the speed and capacity of components is growing exponentially (…)“

Die Komplexität mehrdeutiger Wörter oder auch idiomatischer Redewendungen wird bereits dadurch erzielt, dass sie verschiedene semantische lexikalische Einträge verzeichnen.[193] Während die morphologische Komplexität von Wörtern durch eine einzige semantische Repräsentation gekennzeichnet wird, können verschiedene Morpheme oder Teile eines Wortes verschiedene semantische Repräsentationen haben. Wenn idiomatische Wendungen als einzelne Einheiten dargestellt werden, dann wird die syntaktische Komplexität durch deren lexikalische Repräsentation begründet. Cutler betont, dass lexikalische Komplexität keinen Effekt auf die Schwierigkeit lexikalischen Zugangs ausübt. Was die lexikalische Mehrdeutigkeit als Beispiel syntaktischer Komplexität anbelangt, so besteht sie zunächst aus der systematischen und der unsystematischen Mehrdeutigkeit Die systematische Mehrdeutigkeit bezieht sich auf Wörter, die zwar verwandt sind aber verschiedenen Klassenzuordnungen unterliegen. Unsystematische Mehrdeutigkeit definiert gleiche Wörter, die aber unabhängige Bedeutungen haben (Beispiel: die Bank). Weiterhin kann lexikalische Mehrdeutigkeit auch durch Wörter definiert werden, die fast gleiche Bedeutungen haben, aber auf verschiedene Sachen Referenz nehmen (Beispiel: der See, die See). Zudem kommt Cutler in ihrem Aufsatz zum Schluss, dass lexikalische Mehrdeutigkeit nicht mit der ansteigenden Schwierigkeit des lexikalischen Zugangs gekoppelt sein kann. Es sei schwieriger eine Reihe von Wörtern als einen akzeptablen Satz zu bewerten, wenn dieser mehrdeutige als einfache Wörter enthalte. Im Zusammenhang der in dieser Arbeit diskutierten Bewertung von schriftlichen Lernerproduktionen weist *Profile* auf die Schwierigkeit der Beschreibung der Referenzniveaus hin, denn „je höher das Niveau, desto weniger lassen sich niveauspezifische sprachliche Mittel definieren“ obwohl die Komplexität der sprachlichen Handlungsabläufe je nach Niveau ansteigt.[194] Nach Cutler geht es außerdem darum, dass man idiomatische Wendungen oder Funktionsverbgefüge nicht mit der Bedeutung der Verkettung der einzelnen Wörter gleichsetzt (z.B. das Funktionsverbgefüge *einen Antrag stellen* statt *beantragen*). Cutler hat idiomatische Wendungen als Kontrollinstanz fungierende Wendungen untersucht (z.B. Hals- und Beinbruch vs. Hals- und Armbruch). Weiterhin stellt Cutler (1983:44) kontrastiv gegenüber, wodurch „Komplexität und Einfachheit“[195] gekennzeichnet: „A negative definition is that lexical complexity occurs wherever lexical entries are not simple; lexical simplicity is the case when a phonetic representation of a word evokes a single lexical entry which contains only a single word class representation and a single semantic representation”.

Zur semantischen Komplexität führen nach Cutler auch negative Elemente, die je nach Vorkommen einen Satz schwer verständlich machen, auch wenn die syntaktische Struktur fehlerlos ist. Ich möchte ein Beispiel anführen, um diese Problematik zu verdeutlichen:

Einige Pfälzer Winzer könnten die Tatsache ihres Zweifelns zu verneinen nicht schaffen, dass in manchen Weinsorten die Existenz von Schwefel fehle.

193 Im 5. Kapitel dieser Arbeit habe ich die von Profile erstellten Synonyme in Wortlisten und dem zugesprochenen Niveau gegenüber gestellt und hinterfragt. Die Antwort auf das willkürliche Konstrukt von Profile, Wörter verschiedenen Niveaus zuzuweisen ergibt sich aus Cutlers Aufsatz, der die Schwierigkeit und die Komplexität synonymer, mehrdeutiger oder auch gleichbedeutender Wörter aufzeigt.

194 vgl. S. 116 ff , Diskussion der Bewertungskriterien des B2-Zertifikats des Goethe-Instituts

195 Im Original spricht Cutler von complexity vs. simplicity. Oben genannte Begriffe wurden von mir entsprechend ins Deutsche übersetzt.

Während in diesem Beispiel die syntaktische Korrektheit nicht verletzt wird, führen vermehrte negierende Wörter bzw. Negationen zu einer komplexen Semantik bzw. zu einer Verständnisbeeinträchtigung. Der Begriff der Verständnisbeeinträchtigung wird in den Deskriptoren des Kriteriums Korrektheit, das das Goethe-Institut für die Bewertung des schriftlichen Ausdrucks zur Hand nimmt, an Fehlern festgemacht. Es sind aber nicht immer Fehler, die zur Verständnisbeeinträchtigung führen. Das Beispiel zeigt ganz deutlich, dass auch eine fehlerlose syntaktische Struktur dennoch aus der Kombination verschiedener morphologischer und auch semantischer Elemente so komplex erscheinen kann, dass die Aussageabsicht nicht sofort erkennbar ist. Frey/Heringer (2007:334) definieren Schwerverständlichkeit als ein Anzeichen höheren sprachlichen Niveaus. Es wird bereits an dieser Stelle deutlich, wie sich Komplexität äußert. Demnach stellt sich die Frage, wie und ob sie wahrgenommen wird, um schließlich eine Bewertung abzugeben.

Komplexität ist nach Cutler auch in der Morphologie einer Sprache gegeben, unter anderem in Wörtern mit Präfixen und Suffixen. Morphologisch komplexe Wörter beinhalten in ihrer lexikalischen Repräsentation die Details ihrer morphologischen Struktur. Nehmen wir als Beispiel das Wort *unternachten* statt *übernachten*. Es wird an dieser Stelle ein falsches Präfix bei der Wortbildung verwendet. Cutler (1983:57) erwähnt in diesem Zusammenhang die Feststellung Fays: "(...) substitution errors often occur in which a prefixed word is replaced by another word with the same stem, but different prefix or a non-occurring combination of prefix with the target stem". Ähnlich ist es auch bei der Verwendung eines falschen Suffixes, der auch zu semantischen Fehlern führen kann (z.B. wunderlich vs. wunderbar). Im Sinne Cutlers scheint das DESI-Projekt in seinem Kodierhandbuch die Lexik einer Lernerproduktion zu betrachten. Hauptaugenmerk sind nicht nur lexikalische Elemente sondern auch der Morphologie angehörende Teilbereiche wie zum Beispiel Wortanschlüsse, Wortvalenzen, Kollokationen und der Gebrauch idiomatischer Mittel.[196] Auch das TestDaF-Institut zeigt sich im Unterkriterium Wortschatz bemüht, indem es den Fragen nachgeht, „ob immer die gleichen Verben benutzt werden oder der Wortschatz variiert und ob die treffenden Ausdrücke benutzt werden"[197].

Die morphologische Komplexität kann nach Cutler (1983:63) auch durch die produktive Morphologie bedingt werden: „(...) speakers make errors of word formation, they also regularly create their own neologisms, that is, use their internalized knowledge of morphological structure". Im Sinne der Zweitspracherwerbstheorien, die im dritten Kapitel dieser Arbeit angeführt worden sind, können derartige Lernerstrategien nicht unberücksichtigt gelassen werden. Internalisiertes morphologisches Wissen wird angewandt und es kommt zu Übergeneralisierung[198] des Erlernten. Ein Lerner kann im Deutschen zum Beispiel die Regel, dass alle Verben auf –ieren bei der Partizipbildung ein –t bekommen, auf das Verb „verlieren" anwenden. Die produktive Morphologie „er hat verliert" ist also ganz und gar nicht abwegig, denn der Lerner stützt dieses auf eine Regel, wenngleich diese hier keine Anwendung finden darf. Ein weiteres Beispiel ist das Genus. Während in einschlägigen Grammatiken von der Regel die Rede ist, dass Substantive auf –ur feminin sind, ist es sehr wahrscheinlich, dass ein Lerner dem Wort „Abitur" den Artikel „die" davor setzt.

196 Aus dem Word-Dokument des DESI-Kodierhandbuchs von 2004, S. 8 (Heringer, Personalkommunikation)

197 TestDaF-Institut, Bewertungsanleitung zum Modellsatz 02. 10/2005, S. 8

198 siehe dazu Kapitel 3, Interlanguagehypothese, S. 46 ff

Das Gebiet der Komplexität und insbesondere der sprachlichen Komplexität ist ein sehr weites und schwierig zu beschreibendes Feld. Crystal (1991:68) definiert in diesem Sinne: „(...) it has not yet proved feasible to establish independent measures of complexity defined in purely linguistic terms, largely because of controversy over the nature of the linguistic measures used (...)". Dennoch könnten die abstrakte Komplexitätstheorie von Edmonds und Cutlers pragmatischer Ansatz anhand von syntaktischen Mitteln ein erster Schritt sein, um die Bewertung des schriftlichen Ausdrucks aus einer anderen Perspektive zu betrachten. Zunächst sind Kriterien absolut notwendige Elemente, um Sprachkompetenz zu messen. Allerdings hängt alles von der Beschaffenheit und Grundlage der einzelnen Kriterien ab, nach denen Lernerproduktionen bewertet werden sollen. Diesbezüglich haben sich Heringer/Frey in ihrem Forschungsprojekt „Automatische Bewertung schriftlicher Lernerproduktionen" ausführlich damit auseinander gesetzt. Frey/Heringer (2007:331) erzeugten „mittels textueller Parameter einen Score, der möglichst hoch mit der Bewertung durch menschliche Rater korreliert". Der Validitätsbeweis wurde in diesem Sinne einer linguistischen Diskussion ausgesetzt. Ich möchte die Herangehensweise von Heringer/Frey am Beispiel der lexikalischen Kompetenz kurz umreißen. Während die allgemeine Kann-Beschreibung des GER diese lediglich als die Verwendung lexikalischer und grammatischer Elemente definiert, benennen Frey/Heringer (2007:336ff) in ihrer automatischen Bewertung schriftlicher Lernerproduktionen acht Parameter, die korpuslinguistischen Fundus haben: Wortschatzkomplexität, lexikalische Komplexität, morphologische Tiefe, lexikalische Tiefe, lexikalische Elaboriertheit, lexikalische Varianz, lexikalische Breite und lexikalische Ladung. Dieser korpuslinguistische Ansatz behandelt Sprachkomplexität, wie sie wünschenswert wäre. Erstrebenswert ist in der Praxis üblichen Bewertung von Lernerproduktionen folglich ein Ausschnitt statt eines Abbildes der Komplexität für das zu messende Konstrukt, wobei nach Crystal (1991:68) Komplexität wie folgt definiert ist:

"A central theme is the nature of the interaction between levels of difficulty in cognitive and linguistic STRUCTURES (...)"

Es stellt sich allerdings die Frage, ob die so definierte Komplexität in den Kriterien und in ihren deskriptiven Abstufungen Anwendung finden kann, so dass sie für Rater sichtbar, handhabbar und anwendbar ist. Es kommt nach Eckes (2008) jedoch bereits zuvor zum ersten Konflikt mit dem Validitätsbeweis. In seinem Aufsatz „Rater types in writing performance assessments: a classification approach to rater variability" setzt er sich mit den Bewertungsstilen und den Gewichtungen der einzelnen Kriterien von Ratern auseinander. In seiner empirischen Arbeit, die sich auf die Bewertung des schriftlichen Ausdrucks im TestDaF begrenzt, kommt er zu dem Resultat, dass es verschiedene Ratertypen gibt, die unabhängig von den gesetzten Bewertungskriterien fungieren. Das TestDaF-Institut wendet diesbezüglich das so genannte Multifacettenmodell an, dass die Strenge bzw. Milde eines Raters ermittelt. Er betont aber weiterhin, dass selbst intensivste Schulungen die Ratervariabilität nicht zu dem Maß minimieren können, wie es wünschenswert wäre: „...raters typically remained far from functioning interchangeably even after extensive training sessions...." (Eckes 2008:156). Weiterhin führt Eckes verschiedene empirische Studien in diesem Bereich an, die auf eine Art und Weise alle darauf hinauslaufen, dass Rater trotz gleicher und intensiver Schulung auf verschiedene Aspekte bzw. Kriterien in einer schriftlichen Lernerproduktion fokussieren. Mit seiner „rater type hypothesis" betont auch er, dass Rater, die auf eine bestimmte Bewertungsskala hin trainiert werden, in ihrem Bewertungsverhalten sehr stark variieren

(Eckes 2008:161). In seiner empirischen Untersuchung kommt er zudem zu dem Ergebnis, dass Rater nach Kriteriengewichtung und folglich der Bewertung klassifiziert werden können. Feststellend bemerkt Eckes (2008:178), dass Rater an sich schon verschieden bewerten. Kommen zudem aber die Bewertungskriterien hinzu, was unabdinglich ist, so entsteht eine Kombination eines Rater-Kriteriums-Klassifikationssystems („a joint of rater x criterion classification system"). Meines Erachtens kann diese Feststellung einer Matrix in der mathematischen Vektorenmultiplikation gleichgesetzt werden, wobei aus den verschiedenen Kombinationsmöglichkeiten, die aufgrund der unterschiedlichen Ratertypen und den jeweiligen individuellen Kriteriengewichtungen, beliebige „Produkte" möglich sind.

Je freier interpretierbar also die gesetzten Kriterien bzw. ihre Realisierung mittels der Deskriptoren sind, desto instabiler entpuppt sich die Bewertung und das bedeutet nichts Anderes als einen zweiten Bruch im Validitätsbeweis. Ich möchte aber dennoch nochmals explizit machen, dass für mich das Grundprinzip lauten muss: Rater können nicht besser sein, als die vorgegebene Bewertungsskala bzw. ihre Deskriptoren.

Das offene Aufgabenformat „Schriftlicher Ausdruck" ist hinsichtlich der Bewertung ein Problem, das wie bereits erläutert wurde, aus vielen Facetten besteht. Oberstes Ziel ist es nach Frey (2004:9) zunächst, prägnante Kompetenzbeschreibungen zu definieren, um dem Validitätsbeweis der Bewertungskriterien gerecht zu werden[199]. Dafür sei es notwendig die Niveaubeschreibungen im Hinblick auf die genannten Kriterien der Trennschärfe durch eine neue Auflage des GER anzustreben. Das griechische Staatszertifikat für Fremdsprachen (KPG) basiert auf einem holistisch-analytischen Bewertungsmodell, das sich auf die Kann-Beschreibungen des GER bezieht:

A. Bewältigung der kommunikativen Aufgabe	B. Textaufbau und lexikalische Kompetenz	Grammatische und orthografische Korrektheit	Item
Umfassend bewältigt	befriedigend	befriedigend	15
		Nicht befriedigend	14
	Nicht befriedigend	befriedigend	13
		Nicht befriedigend	12
Mit Mängeln, aber befriedigend bewältigt	befriedigend	befriedigend	10
		Nicht befriedigend	9
	Nicht befriedigend	befriedigend	8
		Nicht befriedigend	7
Nicht befriedigend bewältigt	befriedigend	befriedigend	5
		Nicht befriedigend	4
	Nicht befriedigend	befriedigend	3
		Nicht befriedigend	2
Entspricht nicht der Aufgabenstellung, keine Antwort			1

Tabelle 45: Bewertungskatalog des griechischen Staatszertifikats für Sprache

199 http://www.hueber.de/sixcms/media.php/36/referenzrahm-frey.pdf

Das Bewertungsraster, das für das griechische Staatszertifikat für die Bewertung schriftlichen Ausdrucks entwickelt wurde,[200] ist eine Kombination analytischen und holistischen Bewertungssystems. Auffällig ist zunächst, dass es bei diesem Bewertungsraster keine quantitativen Bezeichnungen gibt, wie das zum Beispiel bei Deskriptoren analytischer Modelle der Fall ist (vgl. Goethe-Institut). Es wird lediglich eine qualitative Unterscheidung zwischen „befriedigend" und „nicht befriedigend" gemacht. Auch hier scheint Sprachkomplexität nicht begründet zu sein. Die qualitativen Bezeichnungen für die jeweiligen drei Kriterien, die als analytisch betrachtet werden können, muss der Rater in Beziehung zu den Kann-Beschreibungen des GER setzen. Das heißt nichts Anderes, als dass der Rater die von GER definierten Kann-Beschreibungen jedes Niveaus sehr gut verinnerlichen muss. Dieses Raster ist unter Berücksichtigung der drei Bewertungskriterien und denn jeweiligen Kann-Beschreibungen des GER relativ gut zu handhaben. Das Manko hierbei ist, wenn Rater die Kann-Beschreibungen nicht berücksichtigen und intuitiv bewerten. Die Gefahr ist gegeben, denn dieses Raster ist einheitlich für alle Niveaustufen. Wenn ein Rater eine Lernerproduktion bewerten soll und die Beschreibungen des vorliegenden Niveaus nicht „respektiert", kann es zu Fehlinterpretationen kommen und das Kriterium der Reliabilität wird verletzt. Unabdingbare Voraussetzung ist demnach die drei analytisch gesetzten Bewertungskriterien *Bewältigung der kommunikativen Aufgabe, Textaufbau und lexikalische Kompetenz und grammatische und orthografische Korrektheit* immer in Bezug auf das zu prüfende Niveau und die jeweilige Aufgabenstellung zu betrachten. Wird dies gewährleistet, so ist dem Rater nicht viel Freiraum in seiner Subjektivität erlaubt, da er lediglich zwischen den qualitativen Abstufungen befriedigend vs. nicht befriedigend entscheiden muss. Der Rater muss sich dementsprechend zwischen dem Bewertungsraster und der entsprechenden Kompetenzbeschreibung des Niveaus X für die drei definierten Kriterien *Bewältigung der kommunikativen Aufgabe, Textaufbau und lexikalische Kompetenz und grammatische und orthografische Korrektheit* orientieren.

An dieser Stelle kommt erneut der Faktor *menschlicher Rater* zum Tragen, den die APA mit dem Standard 1.2 berücksichtigt (APA-Standard 1.2:17):

„The test developer should set forth clearly how test scores are intended to be interpreted and used. (...) and the construct that the test is intended to assess should be clearly described".

Die Definition dieses Standards fordern Testanbieter zur Angabe auf, nach welchen Kriterien Sie ihre Rater auswählen und welche Qualifikationen und Erfahrungen diese mitbringen müssen. In diesem Zusammenhang steht ebenso (APA-Standard 3.23:47):

„The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person (s) responsible for conducting the training session".

200 Basierend auf dem Raster Tsopanoglous (2000), das er in seinem Buch „Methodologie wissenschaftlicher Forschung und ihre Anwendungen in der Bewertung von Sprachkompetenz" vorstellt (Titel wurde von mir aus dem Griechischen übersetzt), entwickelte die wissenschaftliche Arbeitsgruppe für die Staatszertifikatsprüfungen DaF dieses Bewertungsraster für die im Moment existierenden Niveauprüfungen einheitlich

Die Subjektivität eines Raters kann „durch die genaue Festlegung eines Bewertungssystems und die fachgerechte Raterschulung“[201] teilweise begrenzt werden. Dabei ist das oftmals in den Bewertungskriterien erwähnte Verständnis bzw. die Verständlichkeit von Lernerproduktionen keineswegs objektiv. Prüfungsanbieter wie das Goethe-Institut und das TestDaF-Institut begründen ihre nach verschiedenen Systemen definierten Kriterien nicht auf linguistischen Konzepten, so dass man die Komplexität in den verschiedenen Bereichen erfassen könnte. Derartig komplexe Modelle würden menschliche Rater insofern überfordern. Es liegt nicht in ihrer Natur, Sprachkomplexitäten zu definieren, nach Niveau zu sortieren und entsprechend numerisch zuzuordnen (Punktevergabe). Zudem bleibt zu klären, worauf man menschliche Rater eigentlich schult, was sie daraufhin beurteilen und was sie schließlich in der Lage sind zu beurteilen. Möglicherweise ist das eine Frage, die empirisch angegangen werden müsste. In diesem Sinne kann abschließend festgestellt werden, dass Derartiges lediglich maschinell vonstatten gehen kann, um die Gütekriterien der Testtheorie und die hier im Mittelpunkt stehende Validität nicht zu verletzen. Messick (1983:13) sieht den Schlüssel der Testvalidität u.a. in der Interpretation und im funktionalen Wert der Bewertung hinsichtlich der sozialen Konsequenz ihres Gebrauchs. In dieser Arbeit ist versucht worden möglichst viele Facetten der Problematik der Bewertung und ihrer Validität im schriftlichen Ausdruck kritisch zu beleuchten. Dieses Untersuchungsfeld ist allerdings sehr weit. Diesbezüglich gibt es im Zusammenspiel vieler Faktoren grundlegende Desiderate, die noch in weiteren Forschungen angegangen werden müssten, um konkretere Aussagen treffen zu können. Ob linguistische Ansätze ausreichend sind, um die Validität gewährleisten zu können, gilt es in einer anderen wissenschaftlichen Arbeit unbedingt empirisch zu fundieren und nachzuweisen.

201 Materialien zur Prüferschulung für die staatlichen Sprachzertifikate des griechischen Bildungsministeriums, Athen, 25.05.2008

7 Literaturverzeichnis

Alderson, J.C.(1991): Bands and scores. In: Alderson, J.C./North, B. (eds.): Language testing in the 1990s. London: British Council/Macmillian, Developments in ELT: 71-86

American Psychological Association. 1950. Ethical standards the distribution of psychological tests and diagnostic aids. American Psychologist 5

Antos, G./Krings, H.P.(1989): Einleitung. In: Textproduktion. Ein interdisziplinärer Forschungsüberblick. (Hg): Antos, G./Krings, H.P.: Max Niemeyer Verlag. Tübingen 1989, 1-4

Apelt, Hans Peter (o.J.): Am Anfang stand der Sprachunterricht: Streifzüge durch die Geschichte des Goethe- Instituts, München.

Apeltauer, E. (1987): Gesteuerter Zweitspracherwerb: Voraussetzungen und Konsequenzen für den Unterricht. (Hg.) Apeltauer, E.. Hueber. München 1987, 35-50

Arras, U., Grotjahn, R. (2002): TestDaf: Aktuelle Einwicklungen. Eine erweiterte Fassung eines Vortrages auf der 22. Arbeitstagung in Chemnitz, 28.02.2002.

Bachman, L. F. /Palmer, A.S. (1996): Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press

Bachman, L.F. (1990): Fundamental considerations in language testing. Oxford. OUP.

Baldegger, M./ Müller, M./ Schneider, G. (1981): Kontaktschwelle Deutsch als Fremdsprache. (=Europarat-Rat für europäische Zusammenarbeit). Langenscheidt, Berlin/München.

Bart, K.-M. (1999): Annäherung an die Fremdsprache und Interferenzwirkung der Muttersprache: Interimsprachenanalyse und Strategien der Genuszuweisung bei fortgeschrittenen spanischsprachigen Deutschlernenden. Inaugural-Dissertation. Philosophische Fakultät der Albert-Ludwig-Universität, Freiburg.

Bausch, K.-R., Kasper, G. (1979): Der Zweitsprachenerwerb. Gutachten erstellt im Auftrag des Gesprächskreises „Frankreichkunde“ bei der Robert-Bosch-Stiftung Stuttgart. Mai 1979.

Bausch, K.R./Kasper, G. (1979): Der Zweitspracherwerb, Möglichkeiten und Grenzen der „großen Hypothesen“. In: Linguistische Berichte 64, 3-35

Berthold, F. (Hrsg.): Jahrbuch 1998/1999 des Goethe- Instituts

Birkel, P. (1976) Glossar wichtiger testtheoretischer Begriffe. In: Kultusministerium Rheinland-Pfalz. Schulversuche und Bildungsforschung. Berichte und Materialien. Beltz Verlag. Weinheim, 27-48

Blommaert, M.-R./Lutjeharms, M. (2003): Lernersprache aus der Sicht der Lernenden: Fehler und Norm in der Mutter- und Fremdsprache. In: (Hg.) Pürschel, H./Tinnefeld, T.: Moderner Fremdsprachenerwerb zwischen Interkulturalität und Multimedia. Reflexionen und Anregungen aus Wissenschaft und Praxis. AKS-Verlag: Bochum, 126-137.

Bolton, S. (1982): Die Gütebestimmung kommunikativer Tests. Inauguraldissertation zur Erlangung des Grades eines Doktors der Philosophie im Fachbereich Neuere Philologien der Johann Wolfgang Goethe-Universität zu Frankfurt am Main.

Bolton, S./Perlmann-Balme, M. (2006): Schulische Abschlussprüfungen konzipieren - wie macht man das? Ein Werkstattgespräch zum Thema. In: : Zeitschrift für die Praxis des Deutschunterrichts. Fremdsprache Deutsch. Heft 34-2006. Goethe-Institut. Klett, 58-60

Börner, W. (1987): Schreiben im Fremdsprachenunterricht. Überlegungen zu einem Modell. In: Lörscher, W./Schulze, R. (Eds.). Perspectives on Language in Performance. Tübingen. Gunter Narr, 1336-1349 (Bd. 2)
Börner, W. (1989): Didaktik schriftlicher Lernerproduktion in der Fremdsprache. In: Textproduktion. Ein interdisziplinärer Forschungsüberblick. (Hg): Antos, G./Krings, H.P.. Max Niemeyer Verlag. Tübingen, 348-375
Brinker, K. (2001): Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. 5. Auflage. Berlin: Erich Schmidt.
Butzkamm, W. (1989): Psycholinguistik des Fremdsprachenunterrichts. Natürliche Künstlichkeit. Von der Muttersprache zur Fremdsprache. Tübingen
Canadian Modern Language Review, 42/2
Chomsky, N. (1965): Aspects of the theory of syntax. Cambridge, Mass.: MIT Press.
Clapham, C. (1996): The developments of IELTS: a study of the effects of backround knowledge on reading comprehension (Studies in Language Testing 4). Cambridge. Cambridge University Press.
Clapham, C. (2000): Assessment for academic purposes: where next? System 28 (4)
Clark, J. (1985): Curriculum renewal in second language learning: An overwiew.
Corder, P.S. (1967): The significance of learner's errors. In: (Hg.) Richards, J.C.: Error analysis, perspectives on second language acquisition. London., 19-27
Corder, P.S. (1973): Introducing applied linguistics. Harmondswort.
Cronbach, L.J. (1980): Validity on parole: How can we go straight? In: Schrader, W. (Hg.): New directions for testing and measurement. San Francisco, 99-108
Crystal, D. (1987): The Cambridge Encyclopedia of Language. University Press. Cambridge
Crystal, D. (1991): A dictionary of linguistics and phonetics. 3rd Edition. Blackwell Publishers
Cummins, A. (1994): Writing expertise and second-language proficiency. In: Cumming, A. H. (ed.): Bilingual Performance in Reading and Writing. Ann Arbor MI: Benjamins [The Best of Language Learning], 173-221
Cutler, A. (1983): Lexical Complexity and Sentence Processing. In: Flores d´Arcais, G. B. / Jarvella, R. J. (Ed.): The Process of Language Understanding. John Wiley & Sons Ltd., 43-79
DeMers, S.Y., Turner, S.M. (Cochairs), Andberg, M. Foote, W. Hough, L. Ivnik, R. Meier, S. Moreland, K. & Rey-Casserly, C.M. (2000). Report of the Task Force on Test User Qualifications. Washington, D.C.: Practice and Science Directorates, American Psychological Association
Dieterich, R. (1973): Psychodiagnostik. München. Reinhardt
Dittmar, N. (1995): Was lernt der Lerner und warum? Was DaF-Lehrer schon immer über den Zweitsprachenerwerb wissen sollten. In: Deutsch als Zweit- und Fremdsprache. Methoden und Perspektiven einer akademischen Disziplin. (Hg): Dittmar, N. u.a.. Peter Lang Verlag. Frankfurt a. M., Bd. 52, 107-137
Dulay, H./Burt, M. (1974): Goofing: a indicator of childrens' s second language learning strategies. In Language Learning 22.
Dulay, H.:/Burt, M. /Krashen, S. (1982): Language Two. Oxfort University Press, New York..
Eckes, T. (2003): Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. In: Fremdprachen und Hochschule, 2003, Heft 69, 43-68

Eckes, T. (2004): Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In: Wolf, A./Ostermann, T./Choloste, C. (Hg.): Integration durch Sprache (Materialien Deutsch als Fremdsprache, Bd. 73). Regensburg. Fachverband Deutsch als Fremdsprache, 485-518
Eckes, T. (2008): Rater types in writing performance assessments: a classification approach to rater variability. SAGE Publications, 155-185
Edmonson, W./House, J. (1993): Einführung in die Sprachlehrforschung. Tübingen/Basel: Francke
Egger, K. (1995): Muttersprachenerwerb und Zweitsprachenerwerb. Gemeinsamkeiten und Unterschiede. In: Zweitsprachenlernen in einem mehrsprachigen Gebiet. Grundlagen und Perspektiven für ein neues Curriculum. (Hg): Augusto Carli u.a. Provincia Autonoma di Bolzano, 77-83
Embretson, S. E. /Reise, S. P. (2000): Item Response Theory for Psychologists. Lawrence Erlbaum. 1. edition.
Europarat-Rat für kulturelle Zusammenarbeit: Gemeinsamer europäischer Referenzrahmen für Sprachen (2001): lernen, lehren, beurteilen, Straßbourg, Langenscheidt
Faerch, C./Kasper G. (1983). Plans and strategies in foreign language communication. In: Faerch/Kasper (Hrsg.): Strategies in Interlanguage Communication. London/New York: Longman, 20-60
Feilke, H. (1993a): Schreibentwicklungsforschung. Ein kurzer Überblick unter besonderer Berücksichtigung der Entwicklung prozeßorientierter Schreibfähigkeiten. In: Diskussion Deutsch 24/1993, Heft 129
Frey, E./Heringer, H.J. (2007): Automatische Bewertung schriftlicher Lernerproduktionen. In: Linguistische Berichte 211, 331-345
gfl-journal, No. 3/2002
Glaboniat et al (2005): Profile deutsch. Gemeinsamer europäischer Referenzrahmen. Langenscheidt KG. Berlin. München
Glaboniat, M./Müller, M.(2006): Note „sehr gut!" - aber in Bezug worauf? In: Zeitschrift für die Praxis des Deutschunterrichts. Fremdsprache Deutsch. Heft 34-2006. Goethe-Institut. Klett. 14-21
Glück, H. (1988): Schreiben in der Fremdsprache. Eine Einführung. In: Lieber, M./Posset, J. (Hrsg.): Texte schreiben im Germanistikstudium. München: Iudicium. 25-43.
Goethe-Zertifikat B2. Modellsatz. 100707
Goethe-Zertifikat B2. Prüfungsordnung. 050707
Goethe-Zertifikat B2: Trainingsmaterial für Prüfende. Schriftlich-Mündlich. Prüfertraining 090707.
Goethe-Zertifikat C1. Prüfungsordnung. Durchführungsbestimmungen. 050707.
Goethe-Zertifikat C1. Trainingsmaterial für Prüfende. Schriftlich-Mündlich. 090707.
Goethe-Zertifikat C1: Prüfungsziele. Testbeschreibung. Handbuch. 050707.
Grotjahn, R. (2000): Testtheorie: Grundzüge und Anwendungen in der Praxis. In: (Hg.) Wolff, A./Tänzer, H.: Sprache-Kultur-Politik. Beiträge der 27. Jahrestagung Deutsch als Fremdsprache vom 3.-5. Juni 1999 an der Universität Regensburg. Universität Regensburg: Fachverband Deutsch als Fremdsprache (=Materialien Deutsch als Fremdsprache, Bd. 53, 304-341
Grotjahn, R. (2000a): Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Frundlagen und Konsequenzen für die Entwicklung des TestDaF. In: Bolton. S. (Hg.): TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar. Köln: VUB Gilde, 7-55

Grotjahn, R. (2001): Leistungsmessung und Leistungsbeurteilung. Band A. Patras. EAP (Fernuniversität)
Grotjahn, R./Kleppin, K. (2000/2001): TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In: Germanistisches Jahrbuch der GUS „Das Wort“ 2000/2001
Grubitzsch, S. (1999): Testtheorie-Testpraxis: psychologische Tests und Prüfverfahren im kritischen Überblick. 2. Auflage der vollständig überarbeiteten und erweiterten Neuausgabe. Eschborn bei Frankfurt a. M.: Klotz
Günther, R. (1988): Das Deutsche Institut für Ausländer an der Universität Berlin in der Zeit von 1922 bis 1945. Ein Beitrag zur Erforschung des Lehrgebiets Deutsch als Fremdsprache. In: Beiträge zur Geschichte der Humboldt-Universität zu Berlin, Nr. 19. Berlin, 71-75
Hayes, J./Flower, L. (1980): Identifying the Organization of Writing Processes. In: Gregg, L.W./Steinberg, E.R. (Hg.): Cognitive processes in writing. Hillsdale, N.J., 3-30
Helbig, G. (1986): Entwicklung der Sprachwissenschaft seit 1970. Leipzig: VEB Bibliographisches Institut.
Herbst, T. (1991): Terminologie der Sprachbeschreibung: ein Lernwörterbuch für das Anglistikstudium. (Hg.) Herbst, T., Stoll, R., Westermayr, R.. 1. Auflage, 1. Dr. – Ismaning: Hueber (Forum Sprache)
Herskovits, M.J. (1938): Acculturation. New York
Hoyt, W.T. (2000): Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methods, 5.
Hüllen, W. (1983): Über das allmähliche Verfertigen von Sprachregeln. In: Der fremdsprachliche Unterricht 8/83, KILIAN, V./NEUNER, G./SCHMITT, W. (Hg.), Deutsch als Zweitsprache in der Erwachsenenbildung. München
Huot, B.A. (1993): The influence of holistic procedures on reading and rating student essays. In: Williamson, M.M. & Huot, B.A. (eds.): Validity holistic scoring for writing assessment. Cresskill, NJ: Hampton Press, 207-236
Informationsmaterial des TestDaF-Institus: Empfehlungen für Kurse und Materialien zur Vorbereitung auf die Prüfung TestDaF. 04/2005
Ingenkamp, K. (1985): Lehrbuch der pädagogischen Diagnostik. Weinheim: Beltz
J.A. van Ek (1976): The threshold level for modern language learning in schools. The Council of Europe. Longman. Strasbourg
Jones, S./Tetro, J. (1987): Composing in a foreign language. In: Matsuhashi, A. (Eds.): Writing in Real Time: Modeling the Production Processes. Norwood NJ: Ablex, 34-57
Juhász, J. (1970): Probleme der Interferenz. Max Hueber Verlag. Ismaning.
Jung, L. (2001): 99 Stichwörter zum Unterricht Deutsch als Fremdsprache. Ismaning: Max Hueber Verlag:
Kielhöfer, B. (1995): Die Rolle der Kontrastivität beim Fremdspracherwerb. In: Deutsch als Zweit- und Fremdsprache. Methoden und Perspektiven einer akademischen Disziplin. (Hg): Dittmar, N. et al. Peter Lang Verlag. Frankfurt a. M., Bd 52, 35-49
Klauer, K.J. (1987): Kriteriumsorientierte Tests: Lehrbuch der Theorie und Praxis lehrzielorientierten Messen. Göttingen: Hogrefe.
Klein, W. (1984): Zweitsprachenerwerb. Königstein
Klieme, E.(2004): Was sind Kompetenzen und wie lassen sie sich messen? In: Pädagogik 6, 2004. Landesinstitut für Schule, 10-13
Kohn, K. (1990): Dimensionen lernersprachlicher Performanz: theoretische und empirische Untersuchungen zum Zweitsprachenerwerb. (Hg.) Kohn, K.. Tübingen: Narr

Kranz, H.T. (2001): Einführung in die klassische Testtheorie. 5. Aufl.. Eschborn bei Frankfurt a. M.: Klotz.
Krashen, S. (1985): The Input Hypothesis: Issues and Implications. London. Longman
Krashen, S./ Terell, T.D. (1983): The natural approach. Language acquisition in the classroom. Oxford.
Krekeler, C. (2005): Grammatik und Fachbezug in Sprachtests für den Hochschulzugang. Dissertationsschrift. Universität Duisburg Essen
Krings, H.P. (1989): Schreiben in der Fremdsprache-Prozessanalysen zum „vierten skill“. In: Textproduktion. Ein interdisziplinärer Forschungsüberblick. (Hg): Antos, G./Krings, H.P.. Max Niemeyer Verlag. Tübingen, 377-436
Krumm, H. – J. (2006): Müssen jetzt alle dasselbe können? Vor- und Nachteile der Globalisierungsprozesse im Sprachunterricht. In: Zeitschrift für die Praxis des Deutschunterrichts. Fremdsprache Deutsch. Heft 34-2006. Goethe-Institut. Klett, 30-33
Kupfer-Schreiner, C.(1994): Sprachdidaktik und Sprachentwicklung im Rahmen interkultureller Erziehung- Das Nürnberger Modell. Ein Beitrag gegen Rassismus und Ausländerfeindlichkeit. Deutscher Studien Verlag. Weinheim.
Langenscheidt e-Großwörterbuch Deutsch als Fremdsprache. 2003 Langenscheidt KG Berlin und München (CD-ROM)
Langer, H./Schulz v. Thun, F. (1974): Messung komplexer Merkmale in Psychologie und Pädagogik: Ratingverfahren. In: Beihefte der Zeitschrift Psychologie in Erziehung und Unterricht. Heft 68. Ernst Reinhardt Verlag. München-Basel, 13-60
Levenston, E.A. (1979): Second Language Acquisition: Issues and Problems. In: Interlanguage Studies Bulletin 4. 1979, 147-160
Lienert, G. A./Raatz, U. (1998): Testaufbau und Testanalyse. 6. Auflage. München Weinheim: Beltz-Psychologie-Verlags-Union.
Lienert, G.A. (1961): Testaufbau und Testanalyse. Verlag Julius Beltz. Weinheim.
Linke, A./Nussbaumer, M./Portmann-Tselikas, P.R. (2004): Studienbuch Linguistik, 5. erw. Auflage, Tübingen: Niemeyer, 215-255
Löschmann, M. (1992): Effiziente Wortschatzarbeit. Alte und neue Wege – integrativ, kommunikativ, interkulturell, kreativ (Deutsch als Fremdsprache in der Diskussion). Frankfurt/Main, New York: Peter Lang
Lumley, T./McNamara, T.F. (1993): Rater characteristics and Rater bias: Implications for training. Language Testing Research Colloquium (15th Cambridge, England, United Kingdom, August 1993)
Lunz, M.E./Stahl, J. (1990): Judge consistency and severity across grafing periods. Evaluation and health professions 13,4.
McNamara, T. F. (1996): Measuring second language performance. London: Longman
Merten, S. (1997): Wie man Sprache(n) lernt. Eine Einführung in die Grundlagen der Erst- und Zweiterwerbsforschung mit Beispielen für das Unterrichtsfach Deutsch. Peter Lang. Frankfurt a. M., 65-117
Messick, S. (1989): Validity. In: Linn, R.L. (Ed.): Educational measurement (3rd edition). New York: American Council of Education, 13-103
Molitor-Lübbert, S. (1999): Schreiben und Kognition. In: Textproduktion. Ein interdisziplinärer Forschungsüberblick. (Hg): Antos, G./Krings, H.P..Max Niemeyer Verlag. Tübingen, 278-296
Müller, H. (1999): Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen. Bern. Huber.
Nation, Paul (2001): Learning vocabulary in another language (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Nodari, C. (2002): Was heißt eigentlich Sprachkompetenz? In: Barriere Sprachkompetenz. Dokumentation zur Impulstagung vom 2. Nov. 2001 im Volkshaus Zürich, SIBP Schriftenreihe 18
North, B. (1993): The development of descriptors on scales of language proficiency. NFLC occasional papers. John Hopkins University, Washington DC. National Foreign Language Center.
North, B. (1996): Language Proficiency Descriptors. Presentation at the Language Testing Research Colloquium in Tampere, Finland in 1996.
Perlmann-Balme, M. (2006): "Das alles kann ich schon!" Kompetenzen testen, prüfen, zertifizieren. In: Zeitschrift für die Praxis des Deutschunterrichts. Fremdsprache Deutsch. Heft 34-2006. Goethe-Institut. Klett, 5-13
Phipps &Gonzalez, (2004). Modern languages: Learning and teaching in an intercultural field. London, California, New Delhi: Sage
Portmann, P.R. (1991): Schreiben und Lernen: Grundlagen der fremdsprachlichen Schreibdidaktik. Tübingen: Niemeyer
Raatz, U. (2001): Leistungsmessung und Leistungsbeurteilung, Bd. D. Patras: EAP (Fernuniversität)
Raimes, A. (1987): Language proficiency, writing ability, and composition strategies: A study of ESL college student writers. In: Language Learning: A Journal of Applied Linguistics 3(37), 439-468
Report of the Task Force on Test User Qualifications 2-88. Practice on Science Directorates APA. Approvedby the APA Council of Representatives. August, 2000.
Rieck, B.-O.J. (1980): Fehler beim ungesteuerten Zweitspracherwerb ausländischer Arbeitnehmer. In: (Hg.) Cherubim, D.: Fehlerlinguistik. Tübingen: Niemeyer, 43-60
Rösler, D. (1984): Lernerbezug und Lehrmaterial Deutsch als Fremdsprache. Heidelberg 1984. Rösler, D. (1995): Deutsch als Fremd- und Zweitsprache: Gemeinsamkeiten und Unterschiede. In: Deutsch als Zweit- und Fremdsprache. Methoden und Perspektiven einer akademischen Disziplin. (Hg): Dittmar, N. Et al.. Peter Lang Verlag. Frankfurt a. M., Bd.52, 149-159
Rost, J. (2004): Lehrbuch. Testtheorie-Testkonstruktion. 2. vollständig überarbeitete und erweiterte Auflage. Verlag Hans Huber. Bern
Rost, J. /Spada, H. (1982): Probabilistische Testtheorie. In: Klauer, K.J. (Hg.): Handbuch der pädagogischen Diagnostik (Vol. 1). Düsseldorf: Schwann, 59-97
Satzung und Rahmenvertrag. Rechtliche Grundlagen des eingetragenen Vereins. Herausgegeben vom Goethe- Institut, München o.J.
Schelten, A. (1980): Grundlagen der Testbeurteilung. Quelle und Meyer. Heidelberg
Schelten, A. (1997): Testbeurteilung und Testerstellung. Stuttgart: Franz Steiner
Schmidt, Siegfried F. (1973): Texttheorie. Probleme einer Linguistik der sprachlichen Kommunikation. München: Fink.
Schneewind, K.A. (1969): Methodisches Denken in der Psychologie. Bern: Huber
Sommer, J. (1971): Diagnostische Psychologie. In: Rogge, K.E. (Hg.): Steckbrief der Psychologie. Heidelberg: Quelle & Meyer, 170-195
Steinmüller, U. (1995): Korreferat zum Beitrag Dietmar Rösler: "Deutsch als Fremdsprache und Deutsch als Zweitsprache: Unterschiede und Gemeinsamkeiten" – vor allem jedoch Unterschiede. In: Deutsch als Zweit- und Fremdsprache. Methoden und Perspektiven einer akademischen Disziplin. (Hg.): Dittmar, Norbert u.a.. Peter Lang Verlag. Frankfurt a. M., Bd. 52., 161-164
Tenopyr, M.L. (1981): The realities of employment testing. American Psychologist, 36
TestDaF, Bewertungsanleitung zum Modellsatz 02. 10/2005

Thomas Eckes, T. (2008): Rater types in writing performance assessments: A classification approach to rater variability. Language Testing 2008/25,
Trim, J.I.M. (1978): Some possible lines of development of an overall structure for a european unit/credit scheme for foreign language learning by adults. Strasbourg: Council of Europe
Tschirner, E. (2001): Leistungsmessung und Leistungsbeurteilung. Band B. Patras: EAP (Fernuniversität)
Tütken, G. (1984): Selbstständiges zusammenhängendes Schreiben für Fortgeschrittene. In: InfoDaF, Nr. 1. Jahrgang 1984/85. DAAD. Mai 1985, 57-68
Urquhart, A.H. (1987): Comprehensions and interpretations. Reading in a Foreign Language 3
Varadi, T. (1983a): Strategies of Target Language Learner Communication: Message-adjustment. In: Faerch, C./ Kasper, G. (Hg.): Strategies in Interlanguage Communication. London/New York: Longman, 75-99
Vaughan, C. (1991): Holistic assessment: What goes on in the rater`s mind? In: Hamp-Lyons, L. (ed.): Assesing second language writing in academic context. Norwood, NJ: Ablex, 111-125
Vollmer, H.J. (2003): Leistungsmessung: Überblick. In: (Hg.) Krumm/Bausch/Christ: Handbuch Fremdsprachenunterricht. Tübingen und Basel: A. Francke Verlag, 273-277
Weinert, F.E. (2001): Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: (Hg.) Weinert, F.E.: Leistungsmessung in Schulen. Weinheim und Basel. Beltz Verlag, 17-31
Weir, C.J. (2005): Limitations of the Common European Framework for developing comparable examinations and tests. Centre for Research in Testing, Evaluation and Curriculum. Language Testing 22(3)University Roehampton
Wiedenmeyer, D. (2006): DaF-Testen. Testentwicklung und Testbeurteilung. DaF extra Verlag. Athen
Wienold, G. (1973): Die Erlernbarkeit der Sprachen. Eine einführende Darstellung des Zweitsprachenerwerbs. Kösel-Verlag GmbH & Co., München.
Wilkinson, A. (1971): The Foundations of Language: Talking and Reading in Young Children. London
Wilson, M./Case, H. (1997): An examination in Rater severity over time: a study in Rater drift. Berkeley Evaluation and Assessment Research (BEAR) Center. University of California, Berkeley. October 1997
Wolfe, E.W./Feltovich, B. (1994): Learning to rate essays: a study of scorer cognition. Report presented at the annual meeting of the American Educational Research Association in New Orleans, LA, 4.-8. April 1994
Wottawa, H. (1980): Grundriss der Testtheorie. Grundfragen der Psychologie. München: Juventa Verlag.
Writing Tasks: Pilot Samples. In: Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR: learning, teaching, assessment. 1995
Zimbardo, Philip G. (1992): Psychologie. 5., neu übersetzte und bearbeitete Auflage. Bearbeitet und herausgegeben von Hoppe-Graff, S./Keller, B. Springer-Lehrbuch. Berlin.

Internetadressen und Links

http://db.learnline.de/angebote/deutschunterrichtsentwicklung/module/teil-2.pdf
http://db.learnline.de/angebote/deutschunterrichtsentwicklung/module/teil-2.pdf,
Aus: Weinert, F.E. (Hg.): Leistungsmessung in Schulen. Weinheim und Basel. Beltz Verlag. 2001
http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=2&doc=d-gy
http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=3&doc=d-gy
http://db.learnline.de/angebote/kernlehrplaene/text.jsp?kap=4&doc=d-gy
http://de.wikipedia.org/wiki/Goethe-Institut
http://www.hueber.de/sixcms/media.php/36/referenzrahm-frey.pdf
http://www.goethe.de/Z/50/commeuro/c.htm.
http://www.goethe.de/z/50/commoneoro/
http://www.goethe.de/z/commeuro/i1.htmhttp://www.testdaf.de/html/publikationen/pdf files/Eckes_FaDaF_Essen.pdf
http://dueplico.uni-duisburg-essen.de/servlets/DocumentServlet?id_12458
http://www.sdkrashen.com/SL_Acquisition_and_Learning/index.html
http://www.testdaf.de/teilnehmer/pdf/pruefungsordnung.pdf
http://www.testdaf.de/teilnehmer/tn-info_nivea.php
http://www.historisches-lexikon_bayerns.de/artikel/artikel_44721
www.alte.org
www.goethe.de
www.goethe.de/athen>Prüfungen>Goethe-Zertifikat
www.goethe.de/intern
www.testdaf.de
www.goethe.de/referenzrahmen/Quetz 2002
www.goethe.de/z/50/commoneuro/deindex.htm

Vita

Anna Chita

- **06. Dezember 1971 in Korbach/Hessen geboren**
- **1990 Abitur**
- **1998 Magister Artium in den Fächern „Deutsche Philologie: DaF/DaZ“ (Hauptfach) „Psychologie“ und „Schulpädagogik“**
- **1998 bis heute DaF-Lehrerin in verschiedenen Bereichen (Integrationskurse, DaZ, Deutsch für Griechen, Prüfungsvorbereitung aller Niveaustufen)**
- **seit 2003 Prüferin und Korrektorin beim staatlichen Staatszertifikats für Fremdsprachen des griechischen Kultusministeriums. Wissenschaftliche Mitarbeiterin des staatlichen Staatszertifikats für Fremdsprachen des griechischen Kultusministeriums für die Fortbildung von Prüfern und Korrektoren**
- **2009 Promotion zum Thema: Bewertungskriterien schriftlicher Lernerproduktionen B2/C1 und deren Validität»**