QUALITÄTSSICHERUNG BEIM TESTEN FREMDSPRACHLICHER LESEVERSTEHENSKOMPETENZ

Ulrike Arras
DAAD Venezuela/Universidad Central de Venezuela
arras.ulrike@gmail.com

1. Problemaufriss

Das Lesen gilt als hochkomplexe und hochindividualisierte Handlung. Zudem können wir das Lesen nicht direkt beobachten, etwa wie beim Schreiben, bei dem wir ein Produkt, einen geschriebenen Text, in Händen halten und in der Folge beurteilen können (Performanztest). „The process is normally silent, internal, private" (Alderson, 2000: 4). Dies hat Auswirkungen auf die Frage, wie wir Lesekompetenzen messen bzw. beurteilen können. Wir sind bei der Messung fremdsprachlicher Lesekompetenz auf Indizien angewiesen (Kompetenztest). Das bedeutet: Da wir von außen kaum untersuchen können, welche kognitiven Prozesse beim Lesen individuell tatsächlich angestoßen werden, worauf der Leser oder die Leserin ihr Augenmerk richtet, welche Lesestrategien sie tatsächlich anwendet und welches Vorwissen sie einbringt etc.,[1] müssen wir uns bei der Messung der Leseverstehenskompetenz (dies gilt zu einem guten Teil auch in Bezug auf die Messung der Hörverstehenskompetenz) auf indirekte Merkmale als Indizien für zugrunde liegende Kompetenzen verlassen. Wir müssen also beispielsweise anhand einer Antwort beurteilen können, ob die mit dem zugrunde liegenden Lesetext und der dazugehörigen Fragestellung elizitierte Handlung dem entspricht, was wir zu messen versuchen. Diese Beobachtung hat gewichtige Folgen für das Format eines Lesetests, also darauf, wie wir bestimmte zu messende kognitive Operationen elizitieren. Der Input (die Aufgabenstellung, also der Lesetext, die dazugehörigen Fragen sowie die Arbeitsanweisung) muss also so gestaltet sein, dass er das elizitiert, was wir tatsächlich messen wollen. Mit anderen Worten: Aufgabenstellung und die dadurch elizitierten Handlungen müssen valide sein. Hinsichtlich der Validität skizziert Weir (2005) die folgenden Fragen, die bei der Entwicklung eines Tests, also auch beim Design von Lesetests, geklärt werden müssen (s. auch Grotjahn/Tesch (im Druck)):

1 Hinsichtlich der Testteilnehmenden ist zu klären, inwiefern der Test die physischen, physiologischen und psychologischen Merkmale der Lernenden berücksichtigt.

1 Allerdings liegen gerade auch für das fremdsprachliche Leseverstehen Modelle vor, die uns dabei helfen, diesen hochkomplexen Verarbeitungsprozess zu verstehen und für den Fremdsprachenunterricht sowie für die Testentwicklung nutzbar zu machen. Einen aktuellen Überblick erstellen Grotjahn und Tesch (im Druck). Exemplarisch auch Ehlers (1989) sowie Alderson (2000). Ein jüngst vorgelegtes Modell liefern Weir/Khalifa (2009).

2 Hinsichtlich der Kontextvalidität (*context validity*) ist danach zu fragen, wie fair die Merkmale der Aufgaben sowie die Testdurchführung gegenüber den Testteilnehmenden sind.
3 Die *scoring validity* (Validität der Ergebnisermittlung) bestimmt, wie zuverlässig die durch den Test erzielten Ergebnisse sind.
4 Die *consequential validity* –im Deutschen existiert hierfür der Begriff Testwirkungsvalidität– geht der Frage nach, welche Wirkung ein Test „auf alle an ihm Beteiligten und von ihm Betroffenen" hat (Grotjahn/Tesch (im Druck)).
5 Ein weiteres Kriterium für das Validitätsmaß ist die kriterienorientierte Validität, also die Frage danach, ob externe Belege für die Qualität des Tests existieren.
6 Von besonderer Relevanz für die Testentwicklung ist die kognitive Validität (*cognitive validity*), denn sie fokussiert die Frage, ob die kognitiven Prozesse, die mit der Aufgabenstellung elizitiert werden und die zur Bearbeitung bzw. Lösung der Aufgabe erforderlich sind, der Sprachverwendungssituation entsprechen, die wir mit dem Test abbilden wollen.

Prinzipiell gilt: Wenn wir testen, dann entwickeln wir Hypothesen. Wir entwickeln ein dem Test zugrunde liegendes Konstrukt: Bestimmte Items oder Fragen zu einem Lesetext setzen bestimmte zu messende kognitive und textverarbeitende Operationen in Gang. Beim Testen müssen wir also reduzieren: Ein Test bildet genau genommen nicht reale Sprachhandlungen ab, sondern spiegelt sie wider. Anhand der Ergebnisse können wir deshalb nur Vermutungen anstellen über die tatsächliche Kompetenz bzw. darüber, wie ein Prüfling in einer echten Situation tatsächlich reagieren wird. Deshalb ist es erforderlich, die Merkmale realer Sprachverwendungssituationen möglichst genau im Test und in der durch den Test zu elizitierenden Sprachhandlungen wiederzugeben. Mit anderen Worten: Der Test bzw. die durch ihn elizitierten Sprachhandlungen sollen möglichst authentisch und valide sein. Dabei geht es nicht darum, beispielsweise in einem Leseverstehenstest möglichst unveränderte Lesetexte bereitzustellen, sondern in erster Linie darum, die zu elizitierenden Sprachhandlungen, also in unserem Beispiel die zu messenden Kompetenzen in Form von Lesestrategien und kognitiven Prozessen, zu erfassen. Dies kann auch mit für den Test eingerichteten Lesetexten geschehen. In der Regel müssen wir gerade auch die Texte für den Test bearbeiten, um Authentizität herbeizuführen. So ist ein zentrales Problem beim Messen der Lesekompetenz, dass die Prüflinge gegenüber dem Lesetext und seinem Thema sehr wahrscheinlich keine intrinsische Motivation haben: Sie lesen den Text nicht aus Interesse, sondern weil sie sich in einer Prüfung befinden, in der das Lesen von ihnen verlangt wird. Außerdem ist ihnen zuvor das Thema des zu lesenden Textes nicht bekannt, sie erhalten diese Information erst zum Zeitpunkt des Lesens selbst. Daher können auch im Vorfeld, wie in realen Sprachverwendungssituationen üblich, gar keine Erwartungen und Hypothesen zum Inhalt des Textes gebildet werden. Dies

ist ein entscheidendes Dilemma beim Testen der Lesekompetenz: Einerseits müssen wir, um valide zu testen, reale Sprachverwendungssituationen abbilden, andererseits müssen wir der Testsicherheit wegen gerade Authentizität reduzieren.[2]

2. Testkonstruktion

Wer einen Test konzipiert, sollte zunächst die *test specifications* festlegen, denn sie definieren –auch für andere einsehbar– Testziel und Zielgruppe, Textsorte, Itemtyp etc. Diese Informationen sind sozusagen der genetische Code eines Tests. Anhand dieser Informationen können auch in Zukunft für dieselbe Zielgruppe und dieselbe Testfunktion Prüfungsaufgaben entwickelt werden. Außerdem erlaubt die Festschreibung der Testspezifikationen ein gewisses Maß an Transparenz, sowohl für die Institution (Kollegium) als auch für die Lernenden bzw. Prüflinge.

Ein erster Schritt bei der Festlegung der Testspezifikationen ist die Entscheidung, welches Format der Test haben soll. Wichtig hierbei ist die Unterscheidung zwischen integriertem (*integrated*) und isoliertem (*discrete-point* oder *analytic*) Testen. Discrete-Point-Tests „zielen auf die Messung der Kenntnis spezifischer isolierter sprachlicher Phänomene" (Grotjahn, 2003: 37), d.h. das zu überprüfende Phänomen wird weitgehend eingegrenzt, etwa auf eine Teilfertigkeit oder die Beherrschung bestimmter sprachlicher Strukturen. Es handelt sich meistens um geschlossene Aufgabentypen wie Multiple-Choice- oder Zuordnungsaufgaben, deren Auswertung als ökonomisch und objektiv gilt. Bei integrativen Tests hingegen geht es darum, „to gain a much more general idea of how well students read" (Alderson, 2000: 207). Sie entsprechen eher „dem tatsächlichen Sprachgebrauch in realen Kommunikationssituationen" (Bolton, 1996: 103) und sollten daher gerade im Kontext Unterricht nicht fehlen. Integrative Tests versuchen, verschiedene für das Leseverstehen relevante Handlungen zu aktivieren. Zum Teil werden sehr komplexe Aufgaben wie etwa die schriftliche Zusammenfassung eines Lesetextes in der Zielsprache eingesetzt. Hierbei ist zu berücksichtigen, dass dieser Aufgabentyp nicht allein Lesekompetenz misst, sondern auch die schriftliche Ausdrucksfähigkeit sowie strategisches Wissen, etwa Aufbau der schriftlichen Zusammenfassung, Entscheidungen über Wichtigkeit der im Lesetext vorhandenen Informationen etc.) erfordert und ein aufwändiges Auswertungssystem verlangt.[3] Die Entscheidung für isoliertes oder integriertes Testen und damit die Konzipierung der Leseverstehensaufgaben basiert auf folgenden Einzelaspekten:

- den zu messenden Lese- und Verstehensstrategien,
- der inhaltlichen bzw. thematischen und sprachlichen Gestaltung des Textes,
- der Textsorte,

2 Lewkowicz (2000) erörtert den Stellenwert der Authentizität im Rahmen von Leistungsmessung auf der Grundlage empirischer Daten.

3 Diskussion hierzu s. Alderson (2000: 207).

- dem Itemtyp, mithilfe dessen das Verstehen beurteilt werden soll, und
- dem Beurteilungsverfahren.

Die genannten Faktoren bestimmen letztendlich den Schwierigkeitsgrad der Prüfung.

3. Schwierigkeitsfaktoren

Was nun aber ist eigentlich ein schwieriger Leseverstehenstest? Wie kann man die Schwierigkeit taxieren und ggfs. justieren? Dies ist insofern von besonderer Bedeutung, als dass wir beim Testen darauf achten müssen, dass die Tests annähernd gleich schwierig zu sein haben. Es ist nicht akzeptabel, dass wir im akademischen Jahr 2008/2009 Leseverstehenstests einsetzen, die einfacher oder schwieriger zu lösen sind und andere Kompetenzen messen als der Test im akademischen Jahr 2009/2010, denn das wäre unfair und nicht valide. Wenn wir einen Test auf dem Niveau A2 entwickeln, dann muss dieser Test immer A2 messen. Es darf nicht sein, dass er in einem Jahr A2, im nächsten B1 abdeckt. Also brauchen wir klare Kategorien und Testspezifikationen, die uns bei der Schwierigkeitsbestimmung, -justierung und -konstanthaltung helfen.

Prinzipiell ist Textverstehen in hohem Maße abhängig von der Person der Leserin oder des Lesers, also schlussendlich von etlichen „subjektiven Faktoren wie Interesse, Lesemotivation, Lesestil, aber auch vom thematischen, kulturellen und Weltwissen, von der Vertrautheit mit der Textsorte usw." (Arras, 2006: 84). Die Schwierigkeit hängt letztendlich von verschiedenen Faktoren ab, es handelt sich um eine komplexe „Wechselbeziehung von Merkmalen des jeweiligen Lesetextes, von Eigenschaften der zugehörigen Items, von Spezifika der Instruktion sowie von personenspezifischen Merkmalen" (Grotjahn, 2000: 23f.). Zudem zeigen Erkenntnisse aus der Lesbarkeitsforschung und aus Untersuchungen zur Bestimmung von Aufgabenschwierigkeiten[4], dass eine Einschätzung der Schwierigkeit von Leseverstehensaufgaben nur sehr eingeschränkt möglich ist. Was die Festlegung der Schwierigkeitsdeterminanten (und damit die Möglichkeit, die Schwierigkeit unterschiedlicher Leseverstehensaufgaben zu kontrollieren und konstant zu halten) anbelangt, so lässt sich das Problem auf zwei grundlegende Fragen reduzieren: Was ist ein schwieriger Lesetext, und was sind schwierige Fragen zur Überprüfung des Leseverstehens? Nützliche Zusammenstellungen von Schwierigkeitsdeterminanten für die Konstruktion von Leseverstehensaufgaben sind in Grotjahn (2000: 47) zu finden. Folgende Kategorien sind hierbei zu unterscheiden:

- Faktoren, die die Schwierigkeit des Textes bestimmen,
- Faktoren, die die Itemschwierigkeit determinieren und schließlich
- Faktoren, die die Text-Item-Relation bestimmen.

4 Überblick s. Grotjahn (2000).

Erst die Kombination aus den genannten Faktoren ermöglicht eine genauere Bestimmung der Schwierigkeit eines Lesetests. Wir können also nicht einen Aspekt herausgreifen und feststellen, dass er eine bestimmte Schwierigkeit verursacht. Erst die Analyse eines Aspekts im Zusammenspiel mit den anderen Faktoren erlaubt eine genauere Bestimmung. So kann die Schwierigkeit eines langen, komplexen Textes, der eine hohe *type-token-ratio*[5] aufweist und auch inhaltlich anspruchsvoll ist, im Kontext einer Leseverstehensprüfung dadurch relativiert werden, dass die dazugehörigen Items beispielsweise rasch fokussierbare und konkrete Informationen abfragen, so dass lediglich ein Wiedererkennen von Information, nicht aber das Verstehen des Gesamtzusammenhangs oder das Erfassen impliziter Bedeutung erforderlich ist. Das Item und die Text-Item-Relation elizitieren in diesem Fall also eine kognitive Operation, die einen geringen Schwierigkeitsgrad aufweist und auf dem *Gemeinsamen europäischen Referenzrahmen* (GeR) bereits im A-Bereich beschrieben ist.[6] Umgekehrt kann die Schwierigkeit eines Lesetests hoch sein, auch wenn der Lesetext selbst sprachlich und inhaltlich einfach ist, die dazugehörigen Items jedoch komplexe Fragen darstellen, also beispielsweise auf das Verstehen impliziter Informationen abzielen. Die Schwierigkeit einer Aufgabe ist jedoch nicht allein von den Text- und Itemfaktoren abhängig, sondern auch von der Beurteilung bzw. vom Beurteilungsmaßstab. Das bedeutet: Eine schwierige Aufgabe wird leichter, wenn die Beurteilung milde ist. Oder umgekehrt: Auch eine hinsichtlich der genannten Faktoren leichte Aufgabe kann sich als schwierig erweisen, wenn die Maßstäbe bei der Beurteilung streng sind. Das Beurteilungsverfahren ist prinzipiell von der Wahl des Itemtyps abhängig. So erfordern offene und in geringerem Ausmaß halboffene Itemtypen die Auswertung mit Hilfe von Beurteilungskriterien oder Musterlösungen. Die BeurteilerInnen müssen dabei Aussagen über die Leseverstehenskompetenz anhand einer meist schriftlichen Leistung treffen: Sie müssen den Text rezipieren, ggf. rekonstruieren, interpretieren, mit den Vorgaben abgleichen und schließlich ein Urteil fällen bzw. die erhobene Leistung einer bestimmten Leistungsstufe zuordnen. Bei geschlossenen Itemtypen hingegen werden die korrekt gelösten Items erfasst, der ermittelte Punktscore verweist sodann auf die erreichte (und zuvor festgelegte) Kompetenzstufe. Hierbei ist erforderlich, eine bestimmte zu erreichende Punktzahl einem bestimmten Leistungsniveau zuzuordnen, wozu statistische Auswertungsverfahren heranzuziehen sind.

Ein zentraler Schwierigkeitsfaktor beim Lesen und somit von Lesetests ist der in einem Lesetext verwendete Wortschatz. Um die Schwierigkeit sowie die

5 Die Type-Token-Ratio bezeichnet das Verhältnis zwischen der Summe verschiedener Wörter in einem Text zur Summe aller Wörter in einem Text. Je größer dieser Wert, desto komplexer ist der Text. Mit anderen Worten: Je mehr unterschiedliche Wörter in einem Text, desto schwieriger ist er zu rezipieren.

6 So nennt die Skala „Leseverstehen allgemein" des GeR (2001: 74f.) auf A1 die Sprachhandlungen „verstehen, indem er/sie bekannte Namen, Wörter und einfachste Wendungen heraussucht".

Angemessenheit und Geeignetheit eines Lesetextes als Grundlage für einen Leseverstehenstest beurteilen zu können, müssen wir also u.a. den darin verwendeten Wortschatz analysieren. Handelt es sich um frequente Wörter, handelt es sich um Wörter, die bereits erlernt wurden auf der anvisierten Kompetenzstufe, oder handelt es sich um Wörter, die kontextuell erschlossen werden müssen, um die Textaussage zu verstehen? Selbstredend werden in diesen Fällen jeweils andere Kompetenzen gemessen.

Ein Problem besteht nun darin, dass wir uns bei der Konzipierung von Lehrwerken und Prüfungen heute an veralteten Wortschatzlisten orientieren. Tatsächlich basieren moderne Prüfungen sowie Profile Deutsch, an dem sich moderne Lehrwerke orientieren, auf Wortschatzlisten des Deutschen, die als veraltet gelten müssen: Nach Tschirner (2006) „beruht ein Großteil der Lexikauswahl aktueller Grund- und Aufbauwortschätze wie auch von Lehrwerken auf einer Zählung, die mehr als 100 Jahre zurückliegt". Tschirner schlägt in der Konsequenz eine Neuorientierung am Herder/BYU-Korpus vor (Tschirner/Jones, 2005) bzw. am daraus entstandenen Häufigkeitswörterbuch (Jones/Tschirner, 2006).

4. Zusammenfassung: Was ist zu tun?

Prinzipiell gilt: Wir können alles testen, so lange wir *begründen,* was wir testen, in welcher Weise, aus welchen Gründen und mit welchem Ziel und dabei zentrale Testgütekriterien einhalten. Dieses Anliegen macht es erforderlich, ein gewisses Maß an Standardisierung einzuhalten, auch bei informellen Tests, im schulischen ebenso wie im Hochschulkontext. Dieses gewisse Maß an Standardisierung ist möglich, indem wir die Testspezifikationen definieren, also die Merkmale der Prüfung festlegen, damit möglichst wenig Varianz entsteht, also eine Prüfung (beispielsweise Leseverstehenskompetenz DaF am Ende des 1. Studienjahres oder am Ende des 1. Semesters) stets die gleichen Anforderungen aufweist, unter den gleichen Bedingungen abgelegt wird etc. So kann man verhindern, dass die Schwierigkeit einer Prüfung von den Lehrkräften, PrüferInnen oder vom Zeitpunkt abhängig ist, denn das sind für die Leistungsmessung irrelevante Merkmale. Ferner müssen wir die Beurteilungsmaßstäbe festlegen, ebenfalls um zu verhindern, dass über Gebühr starke Varianz zwischen verschiedenen Prüfungsterminen eintritt. Schließlich ist es notwendig, die Prüfung, ihre Anforderungen und ihre Beurteilungsmaßstäbe am *GeR* (Europarat/Rat für kulturelle Zusammenarbeit, 2001) zu verorten. All die skizzierten Aspekte können nicht individuell bestimmt werden. Vielmehr ist eine intensive Arbeit innerhalb der *peer group* (etwa dem Kollegium) erforderlich. Im Dialog-Konsens-Verfahren sollten Prüfungen und Beurteilungsmaßstäbe diskutiert und in der Gruppe den Niveaus des *GeR* zugeordnet werden. Dies sollte zum einen *intra*departamental, also beispielsweise innerhalb einer Deutschabteilung, geleistet werden, möglichst anhand konkreter Prüfungen. In einem zweiten

Schritt sollten im selben Dialog-Konsens-Verfahren aber auch *inter*departamental, also auf der Ebene der Fakultät, des Sprachenzentrums, d.h. der übergreifenden Institution, Kompetenzniveaus und Standards definiert werden. So kann am Ende des Evaluierungs- und Koordinierungsprozesses ein abteilungsübergreifendes Konzept entwickelt werden, das unabhängig von der gelehrten Einzelsprache und den individuellen Lehrkräften Konsens bzw. einen gewissen Standardisierungsgrad aufweist und den Qualitätsanforderungen genügt. Hilfestellung bei dieser, im Übrigen zeitaufwändigen, Arbeit bieten zum einen natürlich der GeR und seine Skalen, zum anderen ist eine Auseinandersetzung mit den Methoden und Arbeitsschritten des Europarat-Projekts zur Verlinkung von Sprachprüfungen am GeR empfehlenswert. Praktische Anleitung zur Kategorisierung und Schwierigkeitsbestimmung von Testaufgaben bietet dabei das so genannte *Manual*. Mit Hilfe der darin enthaltenen Raster können Sprachprüfungen den Niveaustufen des GeR zugeordnet können.[7] Hilfestellung bieten zudem die Handreichungen, die die *Association of Language Testers in Europe (ALTE)* bereitstellt (www.alte.org).

Die skizzierten Maßnahmen ermöglichen es uns, in größerem Maße zentrale Testgütekriterien wie Reliabilität, Objektivität und vor allem auch Validität einzuhalten, denn die Schwierigkeitsniveaus der Prüfungen sowie die Bedingungen der Testdurchführung bleiben weitgehend konstant und wir kontrollieren besser, ob wir das testen, was wir vorgeben zu messen. All dies führt dazu, dass unsere Leistungsmessungen, unsere Beurteilungen und Zertifikate aussagekräftiger werden und vor allem, dass wir unseren Studierenden ein höheres Maß an Transparenz und Fairness bieten. Dies sollte den Aufwand lohnen.

Literatur

ALDERSON, J. C. (2000): *Assessing Reading*. Cambridge: Cambridge University Press.

ARRAS, U. (2006): „Testen und Beurteilen des Leseverstehens in der Fremdsprache". In: *Babylonia* 3-4/2006, 81-86.

BACHMAN, L. F.; PALMER, A. (1996): *Language Testing in Practice*. Oxford: Oxford University Press.

BOLTON, S. (1996): *Probleme der Leistungsmessung: Lernfortschrittstests in der Grundstufe*. Berlin: Langenscheidt.

COHEN, A. D. (2000): "Exploring strategies in test-taking: fine tuning verbal reports from respondents". In: EKBATANI, G.; PIERSON, H. (eds.): *Learner-Directed Assessment in ESL*. Mahwah, N. J., London: Lawrence Erlbaum, 127-150.

DAVIDSON, F.; LYNCH, B. K. (2002): *Testcraft. A Teacher's guide to Writing and Using Language Test Specifications*. New Haven, London: Yale University Press.

7 Das *Manual* wurde im Kontext des vom Europarat 2003 in Auftrag gegebenen Projekts *Relating Language Examinations to the Common European Framework of Reference for Languages - CEFR* entwickelt (nähere Informationen unter http://www.coe.int/).

DOYÉ, P. (1988): *Typologie der Testaufgaben für den Unterricht Deutsch als Fremdsprache*. Berlin: Langenscheidt.

EHLERS, S. (1998): *Lesetheorie und fremdsprachliche Lesepraxis aus der Perspektive des Deutschen als Fremdsprache*. Tübingen: Narr.

EUROPARAT/RAT FÜR KULTURELLE ZUSAMMENARBEIT (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin *et al*.: Langenscheidt.

GROTJAHN, R. (2000): „Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF". In: BOLTON, S. (ed.): *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Köln: Gilde, 7 - 55.

GROTJAHN, R. (2003): *Leistungsmessung und Leistungsbewertung. Fernstudienbrief für den Weiterbildungs-Masterstudiengang „Deutschlandstudien. Schwerpunkt: Deutsche Sprache und ihre Vermittlung*. FernUniversität - Gesamthochschule in Hagen.

GROTJAHN, R.; TESCH, B. (im Druck): „Messung der fremdsprachlichen Leseverstehenskompetenz im Fach Französisch". In: POSCH, R.; TESCH, B.; KÖLLER, O. (eds.), *Standardbasierte Testentwicklung und Leistungsmessung: Französisch in der Sekundarstufe I*. Münster: Waxmann.

HUGHES, A. (2003): *Testing for Language Teachers*. Second edition. Cambridge: Cambridge University Press.

JONES, R.; TSCHIRNER, E. (2006): *A frequency dictionary of German: Core vocabulary for learners*. London: Routledge.

LEWKOWICZ, J. A. (2000): "Authenticity in language testing: Some outstanding questions". In: *Language Testing* 17 - 1, 43 - 64.

TSCHIRNER, E. (2006): „Häufigkeitsverteilungen im Deutschen und ihr Einfluss auf den Erwerb des Deutschen als Fremdsprache". In: CORIMA, E.; MARELLO, C.; ONESTI, C. (eds.): *Atti del XII Congresso Internazionale di Lessicografia*. Alessandria: Edizioni dell' Orso, 1277 - 1288.

TSCHIRNER, E.; JONES, R. (2005): *The Herder-BYU electronic corpus of contemporary German*. Leipzig: Herder-Institut.

WEIR, C. (2005): *Language Testing and Validation: An Evidence-based Approach*. Houndgrave, Hampshire: Palgrave Macmillan.

WEIR, C.; KHALIFA, H. (2009): *Examining Reading* (=Studies in Language Testing 29). Cambridge: Cambridge University Press.

Internetseiten

www.alte.org

www.coe.int