Freitextaufgaben in Online-Tests: Bewertung und Bewertungsunterstützung

Cornelia Kiefer und Ulrike Pado
Hochschule für Technik Stuttgart
Schellingstr. 24, 70174 Stuttgart

{cornelia.kiefer|ulrike.pado}@hft-stuttgart.de
Telefon: 0711 - 8926 2811
Fax: 0711 - 8926 2553

Zusammenfassung

Der Einsatz von eLearning-Szenarien bietet viele innovative Möglichkeiten für die Wissensvermittlung. Spezielle eLearning-Tools dienen dazu, Lernressourcen, interaktive Elemente sowie Interaktions- und Kommunikationsmöglichkeiten bereitzustellen und zu kombinieren. So wird selbstgesteuertes, asynchrones Lernen möglich, methodisch erschließen sich neue Wege und hohe Aufwände für große Lerngruppen können sinken. In diesem Zusammenhang stellt sich die Frage, welchen Nutzen die computergestützte Umsetzung von Lernstandsüberprüfungen (Tests und Klausuren) für Dozenten und Lernende haben kann.

Stark assoziiert mit Tests im eLearning-Bereich sind Multiple-Choice-Aufgaben. Als automatisch korrigierbare Fragen können sie im eLearning-Umfeld schnell und objektiv bewertet werden und liefern auch bei großen Teilnehmerzahlen schnell Feedback an Lernende und Dozenten. Gleichzeitig zweifeln viele Dozenten daran, dass diese Frageform die geforderten Kenntnisse und Fähigkeiten wirklich wiederspiegeln und befürchten ungerechtfertigte Erfolge durch Raten. Freitextfragen umgehen diese Probleme und bieten den Mehrwert einer klareren Einsicht in die Denkweise des Prüflings, doch ist ihre Korrektur zeitaufwändig und oft subjektiv. Wir geben Hinweise für die Praxis, die die Bewertung von Freitextaufgaben verbessern und beschleunigen helfen, und illustrieren unsere Überlegungen an einem realen Datensatz von Freitextfragen und Antworten, der im Verlauf einer Einführungsveranstaltung in die Programmierung für Informatiker und Wirtschaftsinformatiker gewonnen wurde.

Abschließend stellen wir unsere noch andauernde Arbeit an einem System zur halbautomatischen Bewerterunterstützung vor, das vom computerbasierten Umfeld im eLearning-Bereich profitiert und sowohl den Zeitaufwand für die manuelle Bewertung als auch die Replizierbarkeit der Bewertungen weiter optimieren soll.

Schlüsselwörter eAssessment, Freitextfragen, objektive Bewertung, Bewerterunterstützung

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

1 Prüfen im eLearning-Kontext

eLearning, das Lehren und Lernen mit Hilfe des Computers, erfreut sich zunehmender Beliebtheit: Asynchrones, nicht ortsgebundenes Lernen wird möglich. Auch große Lernergruppen können unabhängig vom lokalen Raumangebot arbeiten und betreut werden. Besonders attraktiv ist aus Sicht der Lehrenden auch die Möglichkeit, Tests und Prüfungen ganz oder teilweise automatisch korrigieren zu lassen.

Wo zentral installierte Lern-Management-Systeme (LMS, wie z.B. Moodle[1], ILIAS[2] oder OLAT[3]) existieren, erlauben sie es auch technisch wenig versierten Dozenten, vielseitige eLearning-Angebote bereitzustellen. Diese Systeme helfen, Informationen zu einem Lehrangebot strukturiert online abzulegen, so dass sie jederzeit und von überall her erreichbar sind. Sie ermöglichen auch die Kommunikation mit und unter den Lernenden und erleichtern die Durchführung von automatischen Tests. So können die Möglichkeiten des eLearnings auch ergänzend zu einer traditionellen Lehrveranstaltung genutzt werden. Das eLearning erschließt aber ebenfalls methodisch neue Wege und ermöglicht neue Arbeitsweisen, bei denen die klassische frontale Wissensvermittlung zugunsten interaktiver, selbst-gesteuerter Erarbeitung von Lerninhalten in den Hintergrund tritt.

Aber natürlich ist eLearning kein Selbstzweck - das Hauptziel bleibt der möglichst effiziente Wissenserwerb. Dieser Artikel betrachtet einen wichtigen Teilaspekt dessen, nämlich das Überprüfen des Lernerfolgs, und fragt, ob und wie Prüfungen in einem eLearning-Umfeld qualitativ besser durchgeführt werden können.

Zunächst werden daher Prüfungen aus didaktischer Sicht betrachtet und die Auswirkungen der Überführung verschiedener Aufgabentypen in ein computergestütztes Umfeld werden analysiert. Wir präsentieren Hinweise zum geeigneten Vorgehen und analysieren ihre Wirksamkeit anhand einer realen Datensammlung.

Es fällt auf, dass die reine Computerunterstützung allein noch keine drastischen Auswirkungen auf die Prüfungsqualität hat. Es ergeben sich aber durch die Computerunterstützung neue Möglichkeiten, die Nachteile bestimmter Aufgabentypen abzuschwächen, so dass ihre Vorteile effektiver genutzt werden können. Wir stellen abschließend unsere noch andauernde Arbeit zu diesem Thema vor.

2 Lernstandsüberprüfungen

Nach der Vermittlung von Faktenwissen und der Anwendung dieses Wissens in Übungen ist die Lernstands-überprüfung der klassische Abschluss einer Lehrveranstaltung. Prüfungen haben unterschiedliche Ziele (vgl. Schmees 2011): Bei der summativen Prüfung am Kursende wird über das Bestehen des Kurses bzw. die Note entschieden. Sinn der Prüfung ist hier, festzustellen, dass die Prüflinge sich den Stoff der Veranstaltung in hinreichendem Maß angeeignet haben.

Ein anderes Ziel hat das formative Prüfen: Hier soll das Prüfungsergebnis den Lernenden frühzeitig Rückmeldung über ihren Leistungsstand geben. Üblicherweise werden solche Prüfungen mehrmals während der Lehrveranstaltung gestellt. Regelmäßige Lernstandsüberprüfungen bieten einem Anreiz zum kontinuierlichen Arbeiten durch konkrete Zwischenziele und motivieren durch regelmäßiges Feedback. So können die Lernenden eventuelle Wissenslücken frühzeitig identifizieren. Im Kontext mediengestützter Lehre erlauben (eventuell automatisch korrigierte) formative Prüfungen den Dozenten, auch in großen Veranstaltungen oder in Veranstaltungen, die vorwiegend im Selbststudium durchgeführt werden, regelmäßiges Feedback zu geben. Darüber hinaus erhalten die Lehrenden wichtige Hinweise auf Defizite bei den Lernenden, so dass diese im Verlauf der Veranstaltung aufgearbeitet werden können.

Grundsätzlich können beide Arten der Prüfung vollständig am Computer durchgeführt werden - allerdings ist bei summativen Prüfungen der technische Aufwand bei der Durchführung der Prüfung hoch, da eine manipulations- und ausfallsichere Prüfungsumgebung gewährleistet werden muss. Bei rein formativen Prüfungen ist ein vorübergehender Systemausfall dagegen zu verschmerzen, und wer manipuliert, betrügt sich in erster Linie selbst.

Um zu analysieren, welche Auswirkung die computergestützte Durchführung von Prüfungen hat, betrachten wir drei Qualitätsmerkmale für Prüfungen, die aus didaktischer Sicht entscheidend sind: Objektivität, Reliabilität und Validität (s. z.B. Hartig u. Jude 2007). Objektiv ist eine Prüfung, wenn ihr Ergebnis nur von den Fähigkeiten des Prüflings und nicht vom Durchführungs- oder Korrekturkontext abhängt. So ist Objektivität in mündlichen

1 www.moodle.de

2 www.ilias.de

3 www.olat.org

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

Prüfungen eine große Herausforderung, da die Fragen und die Reaktionen des Prüfers auf die Antworten von Prüfung zu Prüfung variieren. Auch die Aufteilung der Korrektur einer Prüfungsfrage auf mehrere Bewerter kann die Objektivität beeinträchtigen.

Reliabilität bezieht sich auf die Messgenauigkeit einer Prüfung: Bilden die Ergebnisse die Fähigkeiten des Prüflings korrekt ab? Dies zeigt sich z.B. in der Replizierbarkeit von Testergebnissen - derselbe Prüfling mit denselben Fähigkeiten sollte in derselben oder einer sehr ähnlichen Prüfung immer (annähernd) dasselbe Ergebnis erreichen. Unter der Annahme einer objektiven, validen Prüfung hängt dieses Kriterium also von den Inhalten und der Formulierung der Prüfungsaufgaben ab.

Validität schließlich heißt, dass die Prüfung tatsächlich die Kenntnisse und Fähigkeiten bewertet, die sie bewerten soll. Dieses Kriterium ist ebenfalls zentral an den Inhalt der (objektiven, reliablen) Prüfung gebunden. Allerdings ist auch der Durchführungskontext relevant, falls er besondere Kenntnisse und Fähigkeiten erfordert (so sollte sichergestellt werden, dass eine Online-Prüfung nicht implizit die Tippgeschwindigkeit des Prüflings oder die Vertrautheit mit der Prüfungsumgebung mitbewertet).

Im eLearning-Kontext ist nun interessant, ob und wie die computerbasierte Durchführung von Prüfungen ihre Qualität in Bezug auf Objektivität, Reliabilität und Validität verbessern kann.

2.1 Auswirkungen der computerbasierten Durchführung

Die Durchführung von Lernstandsüberprüfungen am Computer hat auf verschiedene Typen von Aufgaben unterschiedliche Auswirkungen. Abb. 1 (nach Hartig u. Jude 2007) spannt ein Spektrum an Aufgabentypen auf, die sich darin unterscheiden, wie stark die erwartete Antwort vorgegeben ist (geschlossene Aufgabetypen geben die erwartete Antwort dabei sehr genau vor). Wir diskutieren die Fragetypen an den Extremen des Spektrums: Auf der einen Seite das Antwort-Wahl-Verfahren (Multiple Choice) mit vorgegebenen Antworten, das klassischerweise mit computerbasierten Prüfungen assoziiert ist, und auf der anderen Seite Freitextfragen, die mit selbst ausformulierten Texten zu beantworten sind.

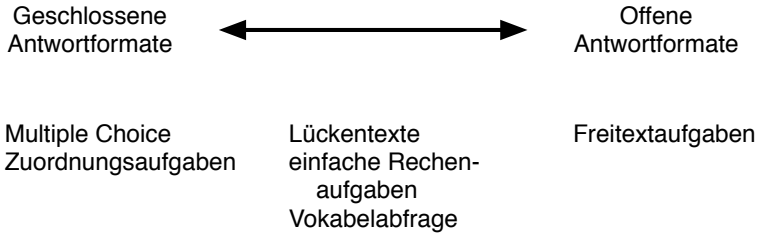

Abb. 1: Spektrum von Aufgabentypen nach Hartig u. Jude (2007)

Antwort-Wahl-Verfahren (Multiple Choice) Da bei Fragen im Antwort-Wahl-Verfahren die erwartete richtige Antwort genau vorgegeben ist (es ist nur eine Auswahl aus angebotenen Antwortoptionen möglich, dabei ist klar definiert, welche davon als richtig gelten), können solche Fragen sehr einfach automatisch korrigiert werden.

Allerdings wird die einfache Korrigierbarkeit nach der Prüfung durch eine hohe Komplexität beim Erstellen der Fragen erkauft (Ehlers u. a. 2013): Die Antwortoptionen müssen so formuliert sein, dass sie alle gleich plausibel sind, denn Antworten, die sich in Länge, Konkretheit oder Thema stark von den anderen unterscheiden, fallen auf und können auch ohne fundiertes Wissen leichter als korrekt oder falsch erkannt werden (*test-wiseness*). Auch ist das Auswählen einer vorgegebenen Antwort nicht dasselbe wie die zu prüfende Fähigkeit, das Wissen in einer realen Situation anzuwenden (Hartig u. Jude 2007). Hier kann also die Validität der Prüfung leiden. Aus didaktischer Perspektive ist ebenfalls relevant, dass die Überlegungen des Prüflings bei der Antwort-Wahl nicht nachzuvollziehen sind. Gerade für formatives Prüfen ist dieser Aufgabentyp daher für die Lehrenden nicht besonders informativ.

Die Durchführung am Computer verbessert bei Fragen im Antwort-Wahl-Verfahren hauptsächlich die Objektivität (durch die nachvollziehbare, fehlerfreie Auswertung). Reliabilität und Validität bleiben im Vergleich zu einer Durchführung auf Papier gleich (Jude u. Wirth 2007), da sie sich auf den Frageinhalt beziehen.

Freitextaufgaben Offene Aufgaben haben ein inverses Profil: Sie sind unkompliziert zu stellen, dafür müssen sie zeitintensiv manuell korrigiert werden. Im Gegensatz zu geschlossenen Fragen können auch Analysefähigkeit und kreative Prozesse einfach geprüft werden. Darüber hinaus ist Raten weitgehend ausgeschlossen und die selbst formulierten Antworten liefern gute Einblicke in Verständnis und Missverständnis des Stoffs, selbst wenn

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

sie relativ kurz bleiben. In unserer Datensammlung (s. Abschnitt 3.1) zeigt sich dies gut: Auf die Frage *Was unterscheidet in JAVA Klassenvariablen von Instanzvariablen?* antworten 14 von 52 Teilnehmern korrekt, dass Klassenvariablen einmal pro Klasse und unabhängig von Objektinstanzen existieren, wohingegen Objektvariablen zu einer bestimmten Objektinstanz gehören. Weitere sechs Teilnehmer antworten gar nicht, und von den verbleibenden 32 falschen Antworten lassen 15 erkennen, dass die Studierenden mit der Sichtbarkeit der Variablen argumentieren. Dieses weit verbreitete Missverständnis kann in der Lehrveranstaltung aufgegriffen und ausgeräumt werden.

Ein Vorteil der computerbasierten Durchführung von Freitextaufgaben ist das einheitliche Schriftbild. Das Entziffern von hastig hingeworfenen Texten und das Interpretieren unklarer Streichungen entfällt. Dadurch wird der Zeitaufwand bei der Korrektur reduziert. Schulz u. Apostolopoulos (2011) nennen 33% Zeitersparnis durch das reine Überführen einer Prüfung in den computergestützten Modus. Objektivität, Reliabilität und Validität bleiben beim computerbasierten Stellen von offenen Fragen zunächst unbeeinflusst. Bezüglich der Validität ist aber zu bedenken, dass die computerbasierte Durchführung von Freitextaufgaben implizit auch die Fähigkeiten im Umgang mit elektronischer Texterstellung misst (Hartig u. Jude 2007).

Wir betrachten nun genauer, wie sich die Objektivität der Bewertung von Freitextfragen verbessern lässt. Außerdem stellen wir fest, wie hoch der Zeitaufwand für die Korrektur unter realistischen Bedingungen ist. Zunächst präsentieren wir methodische Überlegungen, die unabhängig von der computerbasierten Durchführung wirksam sind. Dann beschreiben wir in Abschnitt 4 ein System, das halbautomatisch die Korrektur von Freitextaufgaben unterstützen soll.

3 Manuelle Bewertung von Freitextaufgaben

Im Folgenden geben wir methodische Hinweise zum praktischen Umgang mit Freitextfragen und analysieren die erreichbare Korrekturgeschwindigkeit und Objektivität. Die methodischen Überlegungen sind unabhängig vom Durchführungskontext. Für die Geschwindigkeitsschätzungen verwenden wir am Computer erstellte Antworten.

3.1 Datensammlung und Methodik

Wir verwenden einen Datensatz, der in einer einsemestrigen Vorlesung *Einführung in die Programmierung* erhoben wurde. Die Vorlesung wendet sich an Informatik- und Wirtschaftsinformatik-Erstsemester.

Während des Semesters konnten die Studierenden wöchentlich freiwillige formative Kurztests ablegen. Bonuspunkte für die Tests sollten zum kontinuierlichen Arbeiten motivieren. Außerdem gab es drei über das Semester verteilte summative Teilprüfungen. Aus diesen Lernstandsüberprüfungen wurden alle Freitextaufgaben (neun Freitextfragen) mit den zugehörigen Freitextantworten erhoben. Tabelle 1 zeigt die Antwortmenge pro Frage. Diese schwankt bei den wöchentlichen Tests mit dem Kursbesuch; an den Prüfungen nahmen jeweils mehr Studierende teil als an den wöchentlichen Tests. Exam 2 und 3 stammen aus summativen Teilprüfungen, bewertet wird in 0,5-Punkte-Schritten bis zur Maximalpunktzahl. Die Fragen überprüfen das Verständnis für grundlegende Konzepte der objektorientierten Programmierung, wie sie sich in Java zeigen. Die Antworten sind ein bis zwei Sätze lang.

Frage-ID	Antworten	Maximalpunkte
w4	60	1
w5	32	1
w6	52	1
w7	51	1
w9	52	2
w10	53	1
w11	37	2
exam 2	83	1
exam 3	71	2

Tabelle 1: Anzahl an Freitextantworten pro und maximal erreichbare Punktzahl

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

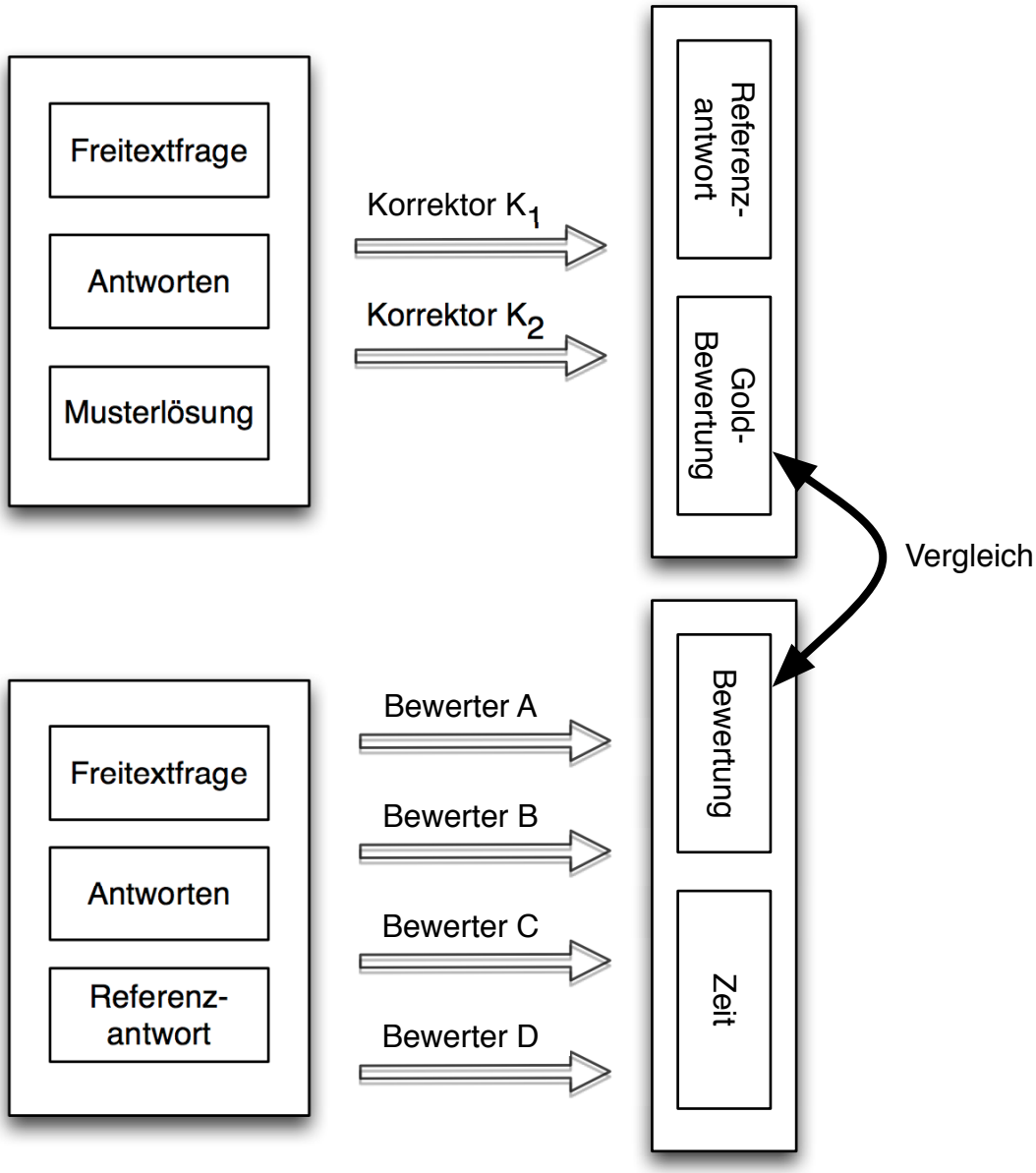

Abb. 2: Erstellung der Gold-Bewertung und Referenzantworten (oben) und Ermittlung von Bewertungszeit und Bewerter-Übereinstimmung (unten)

Zu jeder Frage gibt es eine Referenzantwort (s. Abschnitt 3.2), die die erwarteten Antwortelemente und die dafür jeweils zu vergebenden Teilpunkte auflistet. Wie in Abb. 2 (oben) gezeigt, wurde die Referenzantwort und die Punktebewertung für jede Frage von zwei Korrektoren erarbeitet. Sie bewerteten die Antworten zunächst unabhängig voneinander anhand einer rudimentären Musterlösung. Im Fall von Abweichungen wurden die unterschiedlichen Stimmen diskutiert und eine endgültige Punktzahl (Gold-Bewertung) festgelegt. In dieser Adjudikationsphase wurden von den beiden Bewertern ebenfalls gemeinsam die Referenzantworten festgelegt. Dieses in der normalen Bewertungspraxis ungewöhnlich aufwändige Vorgehen sichert die Qualität der Bewertungen.

Für die Analysen in diesem Artikel wurden die Freitextantworten nochmals von vier weiteren erfahrenen Bewertern anhand der Referenzantwort bewertet (s. Abb. 2 unten). Hierbei wurden die benötigte Zeit und die Bewertung erhoben. Die Bewertungen können dann mit der Gold-Bewertung verglichen werden.

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

3.2 Objektivere Bewertung

Eine objektivere, besser replizierbare Punktevergabe für Freitextantworten lässt sich unabhängig von der Durchführung am Computer oder auf Papier durch ein sauberes methodisches Vorgehen bei der Korrektur erreichen: Wenn im Vorhinein anhand einer Referenzantwort feststeht, welche Elemente eine korrekte Antwort haben soll und welche Teilpunkte es für die einzelnen Elemente gibt, lässt sich auch mit mehreren Bewertern konsistenter arbeiten. Darüber hinaus können die Referenzantworten auch beim formativen Testen hilfreich sein: Wenn sie veröffentlicht werden, dienen sie den Prüflingen als detailliertes Feedback, ohne dass dem Bewerter zusätzlicher Aufwand entsteht.

Turner u. Shellard (2004) empfehlen, beim Definieren von Bewertungsvorlagen mit der idealen Antwort zu beginnen und zu überlegen, welche inhaltlichen und formalen Elemente sie zu einer korrekten Antwort machen. Dann werden die Beschreibungen von schwächer zu bewertenden Antworten erstellt. Tabelle 2 demonstriert die Schritte beim Erstellen der Referenzantworten für eine Frage aus unserem Datensatz. Neben der Frage zeigen wir eine stichwortartige Musterantwort und dann die ausgearbeitete Referenzantwort mit Punkteangaben. Anhand dieser Anleitung können die Freitextantworten konsistent bewertet werden; Spielraum gibt es nur noch bei der Frage, ob die Teilaspekte korrekt genannt wurden.

Stichwortartige Musterlösung	Klassenvariable: existiert ohne Instanz der Klasse, bzw. einmal je Klasse. Instanzvariable: existiert, so lange die Instanz existiert, bzw. einmal pro Objekt.
Richtlinie zur Punktevergabe	Klassenvariable existiert einmal pro Klasse, ohne Instanz: 0,5 P. Instanzvariable existiert einmal pro Objekt: 0,5 P.
Ergänzung nach Erstkorrektur	Klassenvariablen werden mit static deklariert: 0,5 P. Instanzvariablen werden ohne static deklariert: 0,5 P.

Tabelle 2: Schritte bei der Entwicklung von Bewertungsrichtlinien für Antworten auf die Frage “Was unterscheidet in JAVA Klassenvariablen von Instanzvariablen? “

Während der Bewertung müssen die Referenzantworten manchmal angepasst und erweitert werden, um unvorhergesehene Antworten abzudecken. In Tabelle 2 ist dies in der letzten Zeile dokumentiert: Auch der Hinweis auf die Deklaration mit bzw. ohne `static`, der sich in einigen Freitextantworten fand, soll als korrekt gelten. Mit der Erweiterung der Referenzantwort ist die Bewerterentscheidung dokumentiert und für weitere Verwendungen der Frage hinterlegt.

Bei komplexen Fragestellungen oder für die Verwendung als Feedback beim summativen Prüfen können sogenannte *anchors* aus den vorliegenden Antworten ausgewählt werden; dies sind Beispielantworten auf jeder Leistungsstufe, die zur Orientierung des Bewerters bzw. der Prüflinge dienen. Wir verwenden hier wegen der Kürze der erwarteten Antworten keine anchors.

Der Nutzen der Referenzantworten zeigt sich beim Vergleich der Erstbewertung durch die Korrektoren K_1 und K_2 mit den Ergebnissen unserer vier Bewerter *A* bis *D*. Wir benutzen die Bewertungen des Korrektors K_2, da Korrektor K_1 die Fragen selbst gestellt und somit auch ohne Referenzantwort eine genaue Auffassung von der erwarteten Antwort hat. Wir betrachten, für wie viel Prozent der Fragen die Gold-Punktzahl mit dem Punktevorschlag des jeweiligen Bewerters genau übereinstimmt. Die Übereinstimmung von K_2 mit der endgültigen Gold-Bewertung liegt bei 65%. In 35% der Fälle hat Bewerter K_2 die Bewertung anhand der stichwortartigen Musterlösung also anders aufgefasst, als es die finale Referenzantwort vorschreibt. Die Bewerter *A* bis *D* verwendeten die finale Referenzantwort und zeigen Übereinstimmungen mit Gold von 70-75% (Median 74,5%). Die detaillierte Referenzantwort erleichtert also die Standardisierung der Bewertungen und verbessert damit die Objektivität der Prüfung. Der Unterschied zwischen der Übereinstimmung von K_2 mit Gold und den Übereinstimmungen von *C* und *D* mit Gold ist statistisch signifikant ($p<0.02$).

Die Übereinstimmung unserer Bewerter *A* bis *D* mit der Gold-Punktzahl ist auch im Vergleich mit der Literatur hoch; Mohler u. a. (2011) geben an, dass bei der Bewertung von kurzen Freitextantworten ohne Referenzantwort oder Bewertungsanleitung zwei Bewerter nur in 58% der Fälle dieselbe Punktzahl vergeben haben. Allerdings verwendete diese Studie für alle Fragen fünf Bewertungsstufen, während unsere Bewerter in den meisten Fällen nur drei Stufen (0, 0,5 und 1 Punkt) zur Verfügung hatten.

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

3.3 Korrekturzeit

Ein wichtiger Grund für den Verzicht auf Freitextfragen ist der Zeitaufwand für die manuelle Korrektur. Wir möchten einschätzen, wie hoch dieser Aufwand ist und haben daher den Zeitbedarf bei der Bewertung der neun Fragen durch unsere vier Bewerter *A* bis *D* erhoben. Je nach Bewerter werden pro Antwort durchschnittlich acht bis zwölf Sekunden für die Bewertung benötigt (Median: zehn Sekunden). Auffällig ist die Schwankung um bis zu 50% zwischen den einzelnen Personen, die in unserer praktischen Erfahrung typisch ist.

Bei diesen Korrekturdauern ergeben sich für unsere Daten sieben bis zehn Minuten Korrekturzeit pro Frage für einen wöchentlichen Kurztest und zehn bis fünfzehn Minuten pro Frage für die summativen Teilprüfungen. Diese Zeiten stellen eine untere Schranke bei idealen Bedingungen dar: Erfahrene Bewerter mit tiefer Fachkenntnis arbeiten ohne signifikante Zeitaufwände für die Punkteverwaltung oder die Interaktion mit dem verwendeten LMS. Außerdem liegen ausgearbeitete Referenzantworten mit Teilpunktangaben vor.

Noch auffälliger ist aber die Schwankung der durchschnittlichen Antwort-Korrekturdauer je nach Frage: Im Durchschnitt über die Bewerter variiert diese zwischen sechs und 15 Sekunden pro Antwort. Einige Fragen benötigen also fast drei Mal so viel Korrekturzeit wie andere..

Eine Analyse der durchschnittlichen Korrekturzeiten für die einzelnen Fragen gibt einige Hinweise darauf, wie die Formulierung von Freitextfragen die Korrekturdauer beeinflusst. Lange Korrekturzeiten fanden wir besonders bei Fragen mit mehreren zu nennenden Teilantworten, zum Beispiel bei der Frage *Beschreiben Sie die Klassenhierarchie unterhalb von Throwable. An welcher Stelle befinden sich Ihre selbstgeschriebenen Ausnahmen (meistens)?*. Diese Frage besteht schon in der Formulierung aus zwei Teilfragen. Bei der Korrektur sind viele Einzelaspekte zu überprüfen, was die Bewerter offensichtlich Zeit kostete. Ähnlich steht es bei der Frage *Sie möchten ein protected-Feld und mehrere unimplementierte Methoden vererben. Benutzen Sie eine abstrakte Klasse oder eine Schnittstelle? Warum?*, in der ebenfalls neben der rein faktischen Antwort eine mehrteilige Begründung gefordert wird.

Freitextfragen variieren also beträchtlich hinsichtlich der zu erwartenden Korrekturdauer, und so kann es sinnvoll sein, mehrteilige Fragen aufzuteilen bzw. ihnen durch eine entsprechend hohe Punktzahl einen Anteil an der Gesamtprüfung einzuräumen, der die längere Antwort- und Bewertungsdauer rechtfertigt.

4 Schnelle und objektive Bewertung: Computerunterstützung

Wir entwickeln ein automatisiertes System zur Bewerterunterstützung, um zusätzlich zur oben beschriebenen Beschleunigung und Verbesserung der Bewertung durch methodische Überlegungen vom computerbasierten Umfeld im Bereich eLearning zu profitieren.

Wir streben bewusst keine vollautomatische Bewertung von Essays und Freitextantworten an. Der Stand der Technik (Ziai u. a. 2012 gibt eine Übersicht) macht momentan noch für jede Frage einen sehr großen Anfangsaufwand nötig, bevor die Freitextantworten mit ausreichend hoher Genauigkeit klassifiziert werden können. Dies ist für die meisten Nutzungsszenarien, vor allem für den flexiblen Einsatz mit wechselnden Fragen, unrealistisch. Im Gegensatz dazu ist unser System nicht an ein bestimmtes Sachgebiet oder gar spezifische Fragen gebunden, sondern kann zur Bewerterunterstützung bei Freitextfragen aus unterschiedlichen Bereichen verwendet werden.

In der Literatur schlagen Wolska, Horbach u. Palmer (2014) vor, zur Unterstützung bei der manuellen Korrektur von Freitextantworten die Prüflingsantworten automatisch nach Ähnlichkeit in Klassen einzuteilen (zu clustern) und dem Bewerter für jede Antwortklasse nur noch den repräsentativsten Vertreter (den Zentroiden) zu zeigen. Die dafür vergebenen Punkte werden auch an die anderen Klassenmitglieder vergeben. Die Bewertung nur einer Teilmenge der Antworten sorgt so für die beschleunigte Korrektur. Uns ist allerdings vor dem Hintergrund summativer Prüfungen wichtig, dass der Bewerter tatsächlich alle Antworten liest. Schließlich liegt unser Augenmerk neben der Optimierung der Korrekturzeit auch auf der verbesserten Objektivität der Bewertungen.

Wir wollen Korrekturdauer und Objektivität durch eine geschickte Präsentation der Antworten verbessern. Zum einen sind die Antworten nach Güte sortierbar. Dies soll helfen, ähnliche Antworten auch ähnlich zu bewerten und im Zweifelsfall schnell Vergleiche ziehen zu können. Zum anderen erleichtert eine Markierung von zentralen Konzepten das Erfassen der Antwort.

4.1 Sortieren der Studentenantworten nach der Ähnlichkeit zur Referenzantwort

Der Algorithmus zur Sortierung nach Antwortgüte stützt sich auf die Annahme, dass besonders gute Antworten sehr ähnlich zur Referenzantwort sind. Schlechtere Antworten werden entsprechend immer unähnlicher zur Referenzantwort sein.

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

Wir nutzen grundlegende Ideen zur Ermittlung der Ähnlichkeit zwischen zwei Texten aus dem Feld der Computerlinguistik. Die Verfahren werden zum Beispiel im Bereich der maschinellen Übersetzung genutzt (vgl. Papieni u. a. (2002). Hier wird zur Bewertung der Qualität eines maschinellen Übersetzungssystems unter Anderem die Ähnlichkeit der automatisch erstellten Übersetzung zu einer (oder mehreren) vom Übersetzer und Muttersprachler formulierten Übersetzungen betrachtet. Eine verwandte Aufgabenstellung ist das Erkennen von Plagiaten (s. Potthast u. a. 2010 für eine Übersicht).

Wir verwenden für den Vergleich von Referenzantwort und Prüflingsantwort einen Algorithmus, der alle Teilzeichenketten identifiziert, die in beiden zu vergleichenden Texten übereinstimmen (Wise, 1996). Um in gewissem Rahmen abweichende Formulierungen zu erlauben, führen wir jedes Wort auf seine Grundform zurück und löschen alle Wörter, die bereits in der Frage auftauchen und somit für den Studenten bei der Beantwortung der Frage bekannt sind. Außerdem eliminieren wir Stoppwörter[4], das sind vornehmlich Funktionswörter wie Hilfsverben oder Artikel, die aufgrund ihrer Häufigkeit die Ähnlichkeitsbestimmung anhand von Substring-Vergleichen verfälschen würden.

4.2 Hervorhebung wichtiger Konzepte in den Studentenantworten

Die Hervorhebung von wichtigen Konzepten, die in einer Freitextantwort genannt werden, kann dem Bewerter helfen, die Vollständigkeit der Antwort schnell zu erfassen. Ein wichtiges Konzept ist in unserem Ansatz zunächst ein Begriff, der sowohl in der Referenzantwort als auch in der zu bewertenden Freitextantwort zu finden ist. Um diese Begriffe zu bestimmen, werden die beiden Texte auf der Ebene der Wortstämme verglichen. Ähnlich wie bei der Bestimmung des Ähnlichkeitsmaßes werden vor dem Abgleich alle Stoppwörter sowie alle Wörter, die in der Frage auftauchen, gelöscht.

4.3 Integration in ein LMS

Onlinetests werden häufig in LMS gestellt und korrigiert. Im LMS Moodle kann der Prüfer für die manuelle Korrektur von Freitextantworten auf eine benutzerfreundliche Präsentation der relevanten Daten zugreifen. Unsere Optimierungen in der Darstellung der zu korrigierenden Antworten wird nahtlos in die bisherige Ansicht in Moodle integriert. Der in Moodle hierfür vorgesehene Weg erfolgt mit Hilfe eines Plugins zur Integration von neuen Berichtsformaten. Der Bewerter verwendet also eine Abwandlung der herkömmlichen Moodle-Bewertungsansicht für Freitextaufgaben, bei der die Änderungen bezüglich Sortierung und Hervorhebungen realisiert sind.

Um das System dennoch möglichst unabhängig vom jeweils verwendeten LMS zu machen, wurden die neuen Funktionalitäten als Webservice zur Verfügung gestellt. Das Moodle-Plugin liest dann die zu einem Test gehörenden Daten aus der Moodle-Datenbank aus und schickt die relevanten Informationen an den Webservice. Der Webservice führt die linguistischen Analysen der Freitextantworten durch und gibt die Ergebnisse an das Moodle-Plugin zurück. Die Analyseergebnisse werden daraufhin im Moodle-Plugin für die Darstellung der Korrekturansicht verwendet. Die Integration des Webservice in andere LMS ist analog möglich.

Sowohl der Webservice als auch das Moodle-Plugin sind ab März 2015 auf der Moodle-Webseite im Bereich Plugins öffentlich verfügbar.

5 Fazit: Freitextaufgaben werden attraktiver

Wir haben betrachtet, welchen Nutzen Dozenten und Lernende aus der computerbasierten Durchführung von Lernstandsüberprüfungen ziehen können. Im Bereich der geschlossenen Fragen, deren Antworten klar definiert sind, kann automatisch korrigiert werden, wodurch sich der Zeitaufwand für die Dozenten auch bei großen Lerngruppen drastisch reduziert. Darüber hinaus erhöht sich möglicherweise die Objektivität, da Aufgabenpräsentation und Korrektur garantiert für alle Prüflinge gleich ist.

Andererseits eignen sich geschlossene Fragen nicht für die Prüfung jeder Fähigkeit. Daher sind offene Fragen mit frei zu formulierenden Antworten meist unverzichtbar; zusätzlich liefern sie didaktisch hochinteressante Einblicke in das Verständnis der Prüflinge. Achillesferse dieses Fragetyps ist neben dem hohen Zeitaufwand für die Korrektur die Objektivität der Bewertungen. Hier haben wir demonstriert, wie durch methodisch geschicktes Vorgehen beim Stellen der Aufgaben und bei der Vorbereitung der Bewertung sowohl der Zeitaufwand für das Korrigieren als auch die Objektivität verbessert werden kann. Diese Vorteile lassen sich auch bei papierbasierten Prüfungen nutzen, da sie nicht vom Computerkontext abhängen. Dabei ist klar, dass alle Überlegungen zur Optimierung der Korrekturzeit durch die Art der Aufgabenstellung gegenüber den

4 Eine Stoppwortliste für das Deutsche findet sich unter http://snowball.tartarus.org/algorithms/german/stop.txt

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.

inhaltlichen und didaktischen Überlegungen zweitrangig bleiben müssen. Zudem ist auch die Erstellung und Pflege von Referenzantworten, die die Objektivität der Bewertung verbessern, mit Aufwand verbunden. Sie ist sicherlich in summativen Prüfungen sinnvoll, insbesondere wenn die Frage mehrmals genutzt werden soll, der Bewerter nicht identisch mit dem Aufgabensteller ist oder mehrere Bewerter zusammen arbeiten.

Desweiteren haben wir unseren Ansatz zu einer automatisch gestützten Bewertung von Freitextfragen präsentiert, die aktuell entwickelt wird und Korrekturdauer und Objektivität weiter verbessern soll. Durch die Integration in das LMS Moodle soll unser Werkzeug einem breiten Nutzerkreis leicht zugänglich werden.

Literatur

Ehlers JP, Guetl C, Höntzsch S, Usener CA, Gruttmann A (2013) Prüfen mit Computer und Internet. In: Schön S, Ebner M (Hrsg.) L3T Lehrbuch für Lernen und Lehren mit Technologien

Hartig J, Jude N (2007) Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In: Hartig J, Klieme E (Hrsg.) Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik Bd. 20, Bundesministerium für Bildung und Forschung, Bonn, Berlin, S 17-32

Jude N, Wirth J (2007) Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen. In: Hartig J, Klieme E (Hrsg.) Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik Bd. 20, Bundesministerium für Bildung und Forschung, Bonn, Berlin, S 49-56

Mohler M, Bunescu R, Mihalcea R (2011) Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, S 752–762

Papieni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, S 311-318

Potthast M, Barrón-Cedeno A, Eiselt A, Stein B, Rosso P (2010) Overview of the 2nd Internation Competition on Plagiarism Detection. In: Braschler M, Harmann D (Hrsg.) Notebook Papers of CLEF 2010 LABs and Workshops.

Schmees M (2011) eAssessment an Hochschulen. Hamburger E-Learning Magazin Nr. 7, S 31-33

Schulz A, Apostolopoulos N (2011) Potenziale computergestützter Prüfungen. Hamburger E-Learning Magazin Nr. 7, S 37-39

Turner J, Shellard E (2004) Developing and using instructional rubrics. ERS Focus On, Educational Research Service

Wise MJ (1996) YAP3: Improved Detection of Similarities in Computer Program and Other Texts. In: SIGCSE Bulletin, ACM Special Interest Group on Computer Science Education, S 130-134

Wolska M, Horbach A, Palmer A (2014) Computer-assisted Scoring of Short Responses: The Efficiency of a Clustering-based Approach in a Real-life Task. In: Przepiórkowski A, Ogrodniczuk M (Hrsg.), Advances in Natural Language Processing, Lecture Notes in Computer Science Bd. 8686, Springer, S 298-210

Ziai R, Ott N, Meurers D (2012) Short Answer Assessment: Establishing Links between Research Strands. In: Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications, S 190-200

This article appeared in HMD Praxis der Wirtschaftsinformatik (http://link.springer.com/journal/40702), January 2015.The final publication is available at Springer via http://dx.doi.org/10.1365/s40702-014-0104-2.