Machine Learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch

Jennifer Meyer¹ , Thorben Jansen², Johanna Fleckenstein¹, Stefan Keller³ und Olaf Köller¹

¹Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel

²Institut für Pädagogisch-Psychologische Lehr- und Lernforschung (IPL), Kiel

³Pädagogische Hochschule der Fachhochschule Nordwestschweiz (FHNW), Basel

Zusammenfassung: Argumentatives Schreiben ist eine bedeutsame Kompetenz in der Fremdsprache Englisch. Entsprechende Schreibaufgaben sind Teil von Schulabschlussprüfungen in der Sekundarstufe II und von Zugangstests für Hochschulen (z.B. TOEFL®). Trotz ihrer Bedeutung wurden diese komplexen Schreibleistungen bisher im Kontext großer Schulleistungsuntersuchungen kaum empirisch untersucht. Ein Grund dafür ist die aufwendige Auswertung der Essays, für die eine große Anzahl speziell trainierter Kodiererinnen und Kodierer zur Beurteilung benötigt wird. Um den Aufwand der Auswertung zu reduzieren, können Machine Learning Verfahren eingesetzt werden, welche die Urteile der Kodiererinnen und Kodierer approximieren. Dabei werden linguistische Eigenschaften der Essays automatisiert erfasst, die dann genutzt werden, um mit Hilfe von statistischen Verfahren des maschinellen Lernens die menschlichen Urteile vorherzusagen. In der vorliegenden Arbeit soll dieses Vorgehen dargestellt und das Potenzial solcher automatisierten Prozeduren in Bezug auf die Vorhersagegenauigkeit untersucht werden. Dazu lagen Texte von $N = 2179$ Schülerinnen und Schülern der 11. Jahrgangsstufe in Deutschland und der Schweiz vor. Zur Kodierung der Texte wurde die open source-Software *The Common Text Analysis Platform* (CTAP) eingesetzt, die linguistische Textmerkmale automatisch kodiert. Auf Basis dieser Textmerkmale wurden die vorliegenden Urteile von trainierten Kodiererinnen und Kodierern des Educational Testing Service (ETS) vorhergesagt. Die Genauigkeit der Vorhersage erwies sich als zufriedenstellend ($r = .75$; Anteil genauer Übereinstimmung: 42 %) und konnte im Vergleich mit einer etablierten kommerziellen Software des ETS (e-rater®; $r = .81$; Anteil genauer Übereinstimmung: 42 %) bestehen. Es wurden vergleichbare Ergebnisse für die lineare Regression sowie Gradient Boosting als Analysestrategien zur Vorhersage der menschlichen Urteile gefunden. Möglichkeiten und Limitationen der automatisierten Textbeurteilung und deren Anwendung in Forschung und Praxis werden diskutiert.

Schlüsselwörter: Diagnostik von Schreibleistungen, Automatisierte Textkodierung, Schreibkompetenzen, Englisch als Fremdsprache, Machine Learning

Machine Learning in the educational context: Evidence of prediction accuracy considering essays in English as a foreign language

Abstract: Essay writing is an important skill in both first and foreign language learning. Argumentative writing in particular is an important aspect of final examinations in upper secondary school as well as university entrance exams (e.g., TOEFL®). Despite of their importance, argumentative writing competencies have rarely been investigated empirically in the context of large-scale educational assessments. One reason is that the process of rating the essays, which is needed to obtain valid scores, is both time-consuming and expensive, requiring a large amount of well-trained human raters. To reduce the cost of scoring essays, computerized automated scoring techniques can be applied to approximate the score given by the expert human raters. For that purpose, linguistic text features can be analyzed with computer-based algorithms and then combined using machine learning techniques (i.e., gradient boosting or regression analyses) to predict human scores. The present research illustrates this approach, highlighting the potential of automated scoring methods by applying regression analyses and gradient boosting to an existing data set of students' essays. We analyzed a sample of $N = 2179$ essays written by students in upper secondary schools in Germany and Switzerland (grade 11). We used the open-source software CTAP to code 173 linguistic features automatically. These linguistic features were used to predict the scores given by expert human raters trained by the Educational Testing Service (ETS). Results showed the accuracy of the prediction to be satisfactory ($r = .75$; percentage of exact agreement: 42 %) and comparable to the scores computed by a commercial software by ETS (e-rater®; $r = .81$; percentage of exact agreement: 42 %). Our study shows similar results for linear regression analysis and gradient boosting as two different strategies to predict essays scores. Opportunities and challenges of automated essay scoring in the context of foreign language assessment and its application in the school context are discussed.

Keywords: writing assessment, automated essay scoring, writing competencies, English as a foreign language, machine learning

Schreibkompetenzen in der Fremdsprache Englisch sind insofern besonders relevant für die Bildungsforschung, als dass sie die berufliche und gesellschaftliche Teilhabe in einer globalisierten Welt ermöglichen. Sie sind für die erfolgreiche Aufnahme eines Studiums unverzichtbar, bilden eine zentrale Voraussetzung für das wissenschaftliche Arbeiten und sind daher ein wichtiger Bestandteil des Abiturs in Deutschland (KMK, 2012) bzw. der Matura in der Schweiz (EDK, 2017). Weit verbreitet ist hier das argumentative Essay: eine Textsorte, die von Schülerinnen und Schülern in der Oberstufe beherrscht werden sollte (Keller, 2013). Ein wichtiges Ziel des Unterrichts in der Sekundarstufe II und die Aufgabe von Lehrkräften ist es somit, die Schülerinnen und Schülern für das erfolgreiche Schreiben in diesem Genre sorgfältig vorzubereiten. Das erfordert neben der Vermittlung der entsprechenden Kompetenzen des argumentativen Schreibens im Unterricht auch die korrekte Beurteilung der Essays, um darauf aufbauend individuelles Feedback für die Lernenden zur Verfügung zu stellen. Genaue Urteile von Lehrkräften gelten deshalb als integraler Teil eines guten Unterrichts (Elliott, Lee & Tollefson, 2001), als zentral für die Sicherung der Qualität von schulischer Bildung (Terhart, 2002), und als Grundlage für Fairness innerhalb des Bildungssystems (Kronig, 2007).

Diese anspruchsvolle Aufgabe der objektiven, reliablen und validen Textbeurteilung als Hauptgütekriterien erfordert kognitive und zeitliche Ressourcen der Lehrkräfte. Der hohe Ressourcenaufwand in der Beurteilung erschwert auch die wissenschaftliche Untersuchung komplexer Schreibleistungen. In wissenschaftlichen Untersuchungen besteht der aktuelle „Goldstandard“ zur Beurteilung von Essays darin, Texte durch mindestens zwei unabhängige, speziell trainierte Expertinnen und Experten bewerten zu lassen (siehe Rupp et al., 2019). Eine Möglichkeit zur Reduzierung des Aufwands in der Beurteilung für Lehrkräfte und in wissenschaftlichen Untersuchungen liegt im Einsatz von Machine Learning Methoden, um damit Algorithmen zur automatischen Kodierung von Texten zu gewinnen (im Überblick siehe Shermis & Burstein, 2013; Yan, Rupp & Foltz, 2020). Durch Machine Learning kann die Qualität von Texten auf Basis bereits existierender Urteile vorhergesagt werden (Chen & Meurers, 2016). Dazu werden die Urteile trainierter menschlicher Kodiererinnen und Kodierer approximiert, indem eine große Anzahl linguistischer Textmerkmale durch automatisierte computerbasierte Auswertungen bestimmt wird, um dann unter Nutzung statistischer Methoden die Urteile der Expertinnen und Experten für neue Texte vorherzusagen. Die methodischen Fortschritte in den letzten Jahren ermöglichen es bereits heute, dieses Vorgehen in

der Praxis anzuwenden, etwa um Lehrkräfte bei der Beurteilung komplexer Schreibleistungen im Schulalltag zu entlasten und es Forscherinnen und Forschern zu ermöglichen, in Schulleistungsstudien komplexe Schreibleistungen sowohl in Bezug auf die Hauptgütekriterien (Objektivität, Reliabilität und Validität) als auch ökonomisch zu beurteilen. Somit ist die Frage der automatisierten Beurteilung komplexer Schreibleistungen nicht nur für die Forschung, sondern auch für die Praxis relevant.

Ergebnisse zu den Algorithmen der Software e-rater® des Educational Testing Service (ETS) zeigten bereits, dass es möglich ist, Essays auch im Kontext der Fremdsprache Englisch auf diese Art und Weise in Bezug auf die Hauptgütekriterien zu bewerten (Rupp et al., 2019). Es gibt jedoch bisher nur wenige empirische Ergebnisse, welche die Tauglichkeit frei verfügbarer Software zur automatischen Textbeurteilung im Kontext der Fremdsprache überzeugend belegen können. An diese Forschungslücke anschließend soll in dieser Arbeit die psychometrische Qualität solcher automatisch generierter Urteile untersucht und kritisch diskutiert werden. Dazu benutzen wir linguistische Texteigenschaften, die automatisch mit frei verfügbarer Software kodiert werden, um die Textqualität, operationalisiert über Expertenurteile ($N = 2$ pro Text¹) vorherzusagen. Wir verwenden mit multipler linearer Regression und Gradient Boosting zwei etablierte Methoden, die wir in Bezug auf ihre Vorhersagegenauigkeit der Textqualität vergleichen. Schließlich wird geprüft, ob es möglich ist, eine Vorhersagegenauigkeit zu erreichen, die eine automatisierte Rückmeldung der Textqualität erlaubt, um Schülerinnen und Schülern direkt Feedback zu ihrer Schreibleistung zu geben. Dazu werden die Ergebnisse der vorliegenden Studie mit Ergebnissen aus der Literatur bzw. den Ergebnissen des e-rater® des ETS verglichen, um Aussagen über die Zuverlässigkeit des Algorithmus treffen zu können. ETS ist für die Administration und Auswertung des Test of English as a Foreign Language (TOEFL®) zuständig und auf Basis dieser umfassenden Erfahrung führender Anbieter auf dem Gebiet der Textbeurteilung und maschinellen Kodierung englischsprachiger Texte. Die in dieser Untersuchung verwendeten „Benchmarks“ des e-rater® können somit als Optimalstandard für die Genauigkeit maschineller Urteile herangezogen werden.

Mit diesem Artikel demonstrieren wir, wie Machine Learning zur Textbewertung im Forschungskontext großer Schulleistungsstudien eingesetzt werden kann und illustrieren anhand eines Beispielkorpus, wie eine große Menge an Texten anhand frei verfügbarer Programme reliabel und valide automatisch kodiert werden kann. Ziel

¹ Bei einer Abweichung von mehr als einer Stufe zwischen den beiden Expertinnen/Experten wurde ein dritter Rater einbezogen (siehe Rupp et al., 2019).

ist es, die Relevanz der Anwendung künstlicher Intelligenz in diesem Kontext sowohl für das Bildungsmonitoring der Schreibleistungen als auch für die praktische Umsetzung an den Schulen im Unterricht und bei der Förderung von Schreibfähigkeiten hervorzuheben.

Die Beurteilung von Textqualität

Die korrekte Beurteilung komplexer Schreibleistungen ist eine kognitiv anspruchsvolle Aufgabe, die die Integration einer Vielzahl von Schlussfolgerungen, Entscheidungen und Kriterien erfordert. Welche Texteigenschaften einbezogen, welche Kriterien höher gewichtet und welche Standards bzw. Vergleichsmaßstäbe angelegt werden, kann stark zwischen Personen variieren (Eckes, 2008). Aufgrund dessen variieren auch Urteile der Textqualität stark zwischen Personen (siehe Birkel & Birkel, 2002; Huot, 1990, 1996; Jansen et al., 2019; McNamara et al., 2015; Meadows & Billington, 2010).

Die Operationalisierung von Textqualität durch Expertenurteile

Um diese Unterschiede zu verringern, werden häufig Urteile trainierter Expertinnen und Experten eingesetzt, die auf der Basis von Beurteilungskriterien (Rubrics) vergeben werden (Bridgeman, 2013; Leckie & Baird, 2011). Beim TOEFL® werden zum Beispiel sechsstufige Rubrics eingesetzt, um ein holistisches Gesamturteil über die Textqualität zu generieren. Dazu werden in den Stufenbeschreibungen des Rubrics unterschiedliche Aspekte der Textqualität thematisiert. Darunter fallen beispielsweise die Ausarbeitung des Textes in Hinblick auf das Thema und die Argumentation, die Organisation des Textes im Hinblick auf Kohärenz und Stringenz der Ideen, sowie sprachliche Merkmale wie syntaktischer Variabilität, angemessene Wortwahl oder Qualität der Idiomatik. Kleinere sprachliche Fehler werden dabei explizit zugelassen, auch auf den höheren Stufen. Eine ausführliche Darstellung des Rubrics für argumentative Texte des ETS ist im Internet frei abrufbar (www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf; Abruf 09.11.2020). Dieses Rubric für das argumentative Schreiben im internetbasierten TOEFL® wurde auch in dieser Studie verwendet. Die Kodiererinnen und Kodierer des ETS werden bei jeder neuen Anwendung des Rubrics auf einen Datensatz umfassend trainiert und die Beurteilungen täglich neu kalibriert, um die Beurteilungsgenauigkeit zu optimieren (vgl. Rupp et al., 2019). Insgesamt sind also die Beurteilung der Texte, das Training und die Qualitätssicherung zeitaufwändig und

kostspielig. Dafür erhält man in diesem Verfahren hochgradig robuste Textgüteeinschätzungen, deren Reliabilität und Validität bereits für den europäischen Kontext gezeigt werden konnte (vgl. Fleckenstein et al., 2020; Rupp et al., 2019).

Die computerisierte, automatische Kodierung der Textgüte ist ein Ansatz, um das Problem des hohen Aufwands und der Kosten anzugehen, ohne jedoch Hauptgütekriterien wie Objektivität, Reliabilität oder Validität geringer zu gewichten. Dadurch wird einerseits die Erfassung von Schreibfähigkeiten im Kontext großer Schulleistungstudien ermöglicht, andererseits können auch Lehrkräfte bei der Beurteilung komplexer Schreibleistungen von Schülerinnen und Schülern in der Praxis unterstützt werden. Aufgrund des stärker zunehmenden Interesses an der Forschung zu Schreibleistungen in der Fremdsprache, sowie der ständig differenzierteren Möglichkeiten zur Textkodierung durch Open-Source Software, lässt sich ein Anstieg der Forschungsarbeiten im Bereich der automatischen Textbeurteilung verzeichnen (McNamara et al., 2015), wobei ein großer Anteil der Arbeiten weiterhin auf den amerikanischen Kontext sowie auf die Bewertung von Texten in der Erstsprache bezogen ist (z.B. Dikli, 2006; Graesser & McNamara, 2012; Shermis & Burstein, 2013; Xi, 2010; siehe auch Weigle, 2013).

Machine Learning bei der Textbeurteilung

Die Anwendung von Machine Learning auf die automatisierte Kodierung von Texten zur Vorhersage menschlicher Urteile stellt sich folgendermaßen dar. Alle Texte eines Korpus müssen über ein Gesamturteil verfügen, dessen Qualität dem „Goldstandard“ entsprechen sollte (umfassendes Training der Kodiererinnen und Kodierer sowie Mehrfachkodierung durch mindestens zwei Kodiererinnen bzw. Kodierer; siehe Rupp et al., 2019). Im nächsten Schritt werden die Texte in eine Software, das automatische Kodierungstool, eingelesen und es werden computerlinguistische Kodierungen von Textmerkmalen vorgenommen (s. unten). Das Korpus wird weiterhin in Training/Model Building- und Test-Sets aufgeteilt. Das Trainingsset wird dazu verwendet, einen Algorithmus zu finden, der das Gesamturteil auf Basis der Textmerkmale optimal vorhersagen kann. Hierfür werden häufig lineare multiple Regressionsansätze verwendet (z.B. Almond, Deane, Quinlan, Wagner & Sydorenko, 2012; Deane, 2014; Zhang & Deane, 2015), aber auch andere statistische Verfahren, wie beispielsweise Boosting (Sinharay et al., 2019). Im nächsten Schritt wird der so gewonnene Algorithmus benutzt, um die Urteile der übrigen Texte im Testset-Korpus vorherzusagen. Diese durch den Algorithmus vergebenen Urteile werden dann mit den Urteilen der

Expertinnen und Experten verglichen, um die Beurteilungsgenauigkeit zu bewerten. Die Schätzung kann dann als erfolgreich angesehen werden, wenn diese Urteile sich hinreichend ähnlich sind (Bridgeman, 2013).

Über verschiedene Studien hinweg finden sich exakte Übereinstimmungen von 30 bis 60 Prozent sowie Korrelationen von menschlichen und computerbasierten Urteilen zwischen $r = .60$ und $r = .85$ (Attali & Burstein, 2006; Rudner et al., 2006; Rupp et al., 2019; Shermis et al., 2010; Warschauer & Ware, 2006; für einen Überblick siehe Bridgeman, 2013). Übereinstimmungen, die bis auf mindestens eine Stufe genau sind (*adjacent agreement*), erreichen in den verschiedenen Studien Werte von 85 bis 100 % (Attali & Burstein, 2006; Rudner et al., 2006; Shermis et al., 2010; Warschauer & Ware, 2006; Rupp et al., 2019). Diese Zahlen zeigen, dass der Einsatz automatisierter Textbeurteilungssysteme bereits auf einem hohen Niveau erfolgt. Solche Systeme lassen sich insbesondere im Forschungskontext von großen Schulleistungsstudien sinnvoll einsetzen, in denen eine schnelle und ökonomische Auswertung des Textmaterials von besonderer Bedeutung ist.

Viele bisherige Studien verwenden lineare Regressionsmodelle als statistische Methode, um Urteile zur Textqualität mithilfe linguistischer Textmerkmale vorherzusagen (z.B. Almond et al., 2012; Deane, 2014; Zhang & Deane, 2015). Allerdings können auch komplexere und rechenintensivere Vorhersagemethoden, wie zum Beispiel Boosting, dafür eingesetzt werden, die zu besseren Vorhersagen führen können. Sinharay et al. (2019) fanden Evidenz für die Überlegenheit des Boosting bei der Vorhersage von Urteilen zur Textqualität. Aufbauend auf diese Studie vergleichen wir die Vorhersagegenauigkeit beider Methoden unter Verwendung der durch die Software CTAP vergebenen Textmerkmale.

Software zur automatisierten Textbeurteilung

e-rater®: Eine Software für automatisierte Textbeurteilung im amerikanischen Kontext

Im amerikanischen Kontext wurde die automatisierte Beurteilung komplexer Schreibleistungen bereits unter Verwendung verschiedener, teilweise kommerzieller Tools umfassend untersucht (z.B. Coh-Metrix, Writing Assessment Tool WAT, Linguistic Inquiry and Word Count LIWC; siehe McNamara et al., 2015; für umfangreiche Übersichtsarbeiten zur automatisierten Beurteilung siehe Shermis & Burstein, 2013 und Yan, Rupp & Foltz, 2020). Es existieren jedoch kaum empirische Untersuchungen im europäischen Kontext, und nur ein Teil der bisherigen Forschung bezieht sich auf Texte im Bereich der Fremdsprache Englisch. Es kann jedoch nicht grundsätzlich von einer Übertragbarkeit

der Ergebnisse im Bereich Englisch als Erstsprache auf die Fremdsprache ausgegangen werden, da diesen Schreibleistungen (zumindest teilweise) unterschiedliche psychologische Prozesse zu Grunde liegen und auch die entstehenden Texte sich in ihren linguistischen Merkmalen stark unterscheiden, insbesondere, da die Sprachkompetenzen mit der Kommunikation der Schreibziele konfundiert sein können. Beispielsweise fällt es in der Fremdsprache oft schwerer, beim Schreiben die geeigneten Vokabeln und syntaktischen Strukturen abzurufen, so dass für den Ausdruck von Ideen, Argumentationen oder Inhalten weniger kognitive Ressourcen zur Verfügung stehen (siehe Chenoweth & Hayes, 2001; Weigle, 2013; Weigle, 2005). Das Projekt Measuring English Writing Skills (MEWS; Keller et al., 2020; Köller et al., 2019) untersucht erstmals im europäischen Kontext an einer großen Stichprobe die Schreibleistungen von Lernenden im 11. Schuljahr in der Fremdsprache Englisch und setzt dabei systematisch die automatisierte Textbeurteilung ein. Dabei wurde sowohl „sachliches“ wie „argumentatives“ Schreiben geprüft (Fleckenstein et al., 2020). Die Beurteilung der in MEWS geschriebenen fast 10 000 Schülertexte wurde von geschulten Expertinnen und Experten des ETS vorgenommen (insgesamt $N = 56$; siehe Rupp et al., 2019, für die detaillierte Beschreibung der Rating-Prozeduren und Training der Expertinnen und Experten). Weiterhin wurde der für den TOEFL® entwickelte Scoring-Algorithmus e-rater® auf die Zielpopulation der Studie (Schülerinnen und Schüler an allgemeinbildenden Gymnasien in Deutschland und der deutschsprachigen Schweiz) adaptiert. Im Rahmen der MEWS Studie wurde auch die automatische Kodierung der Textqualität mit jener der trainierten Kodiererinnen und Kodierer verglichen sowie die Generalisierbarkeit der Vorgehensweisen auf unterschiedliche Populationen und Testkontexte überprüft. Es zeigte sich, dass eine Anpassung der ETS-Algorithmen notwendig war und die bereits verfügbaren Routinen des e-raters® nicht direkt auf die Zielpopulation anwendbar waren. Weiterhin machten die Ergebnisse die Promptspezifität der automatisierten Auswertung deutlich (vgl. auch Guo, Crossley & McNamara, 2013). Das bedeutet, dass ein Algorithmus zur Beurteilung einer bestimmten Textsorte bzw. Aufgabenstellung nicht auf andere Texte übertragen werden kann. Aus diesem Grund beschränken wir uns für die aktuelle Arbeit auf jenen Teil des MEWS-Textkorpus, welcher argumentative Texte enthält.

CTAP: Open Source Software zur automatisierten Kodierung lexikalischer und syntaktischer Komplexitätsmaße in Texten: Es existieren bereits unterschiedliche Tools zur automatischen Kodierung von Texten. Teilweise handelt es sich, wie bereits beschrieben, um kommerzielle Software (e-rater®, ETS, 2014; Burstein et al., 2003). In den letzten Jahren wurde aber auch vermehrt frei verfügbare Software entwickelt (z.B. Cohmetrix; siehe Graesser,

McNamara, Louwerse & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy & Cai, 2014). Wir fokussieren hier auf die Common Text Analysis Plattform (CTAP, Chen & Meurers, 2016) als ein umfangreiches Tool zur Kodierung linguistischer Textmerkmale, welches über eine hohe Benutzerfreundlichkeit verfügt und das Einlesen und Auswerten einer größeren Anzahl von Texten ermöglicht. Damit eignet sich CTAP auch in einer kostenfreien Version zum systematischen Einsatz in großen Schulleistungsuntersuchungen. Interessierte Nutzerinnen und Nutzer können sich online registrieren und das Tool selbstständig nutzen. Alle verfügbaren Textmerkmale werden detailliert beschrieben und können durch die Nutzerinnen und Nutzer selbstständig ausgewählt werden, um einen eigenen Textkorpus anhand der Merkmale zu kodieren. Im Wesentlichen handelt es sich um Komplexitätsmaße in den Bereichen lexikalische Variation und Angemessenheit sowie syntaktische Komplexität. Im Bereich des Wortschatzes gehört dazu u. a. die „Type Token Ratio“ (Verhältnis von im Text vorkommenden Wortformen und zu Grunde liegenden abstrakten sprachlichen Einheiten) oder korpusbasierte Schätzungen der Häufigkeit bestimmter Wörter im Gesamtwortschatz einer Sprache (z. B. SUBTLEX Index). Auf syntaktischer Ebene gehört dazu u. a. die Anzahl von parataktischen und hypotaktischen Satzkonstruktionen (z. B. Relativsätze) oder die Häufigkeit und Komplexität der verwendeten Modalverben (vgl. dazu auch „Kodierung von Textmerkmalen“, unten). Ein Überblick über die verwendeten Textmerkmale inklusive Beschreibung findet sich im elektronischen Anhang 1.

Fragestellungen

In der hier vorliegenden Arbeit prüfen wir die Beurteilungsgenauigkeit, die unter Verwendung von CTAP erreicht werden kann. Dazu nutzen wir Textbeurteilungen, die bereits aus der MEWS-Studie vorliegen. Die dort gewonnenen Ergebnisse der Beurteilung durch menschliche Kodiererinnen und Kodierer und e-rater® können im Hinblick auf die umfangreiche Erfahrung in der TOEFL®-Auswertung als Goldstandard in diesem Bereich angesehen werden. In Rupp et al. (2019) wurde eine exakte Übereinstimmung der Urteile der Kodiererinnen und Kodierer (auf einer sechsstufigen Skala) mit den automatisch generierten Urteilen von 40 Prozent erreicht. McNamara et al. (2013) finden eine exakte Übereinstimmung von 44 Prozent.

Ein – wie oben beschrieben – gewonnener Vorhersagealgorithmus ermöglicht es, Texte und Antworten von Schülerinnen und Schülern zu dieser spezifischen Auf-

gabenstellung automatisch innerhalb weniger Sekunden ohne die Beteiligung menschlicher Kodiererinnen und Kodierer zu bewerten. Um diese praktische Anwendbarkeit in der Schulpraxis und Bildungsforschung produktiv zu nutzen, ist es wichtig, Reliabilität und Validität solcher automatisierten Urteile umfassend zu untersuchen. Dazu wurde in der aktuellen Studie die automatische Kodierung mit der CTAP-Software für das Textkorpus zu einer Aufgabe des MEWS-Projekts an den Urteilen der menschlichen Kodiererinnen und Kodierer des ETS validiert. Die Studie fokussierte sich auf drei übergeordnete Fragestellungen:

1. Wie präzise können Expertenurteile zur Güte von geschriebenen englischsprachigen Texten von Schülerinnen und Schülern durch linguistische Textmerkmale, die in der open-source Software CTAP kodiert werden, vorhergesagt werden?

Zur Beantwortung dieser Fragestellung werden die Vorhersagemäße den entsprechenden Maßen der ETS Software e-rater® (Rupp et al., 2019) gegenübergestellt. Die Ergebnisse aus der MEWS Studie wurden hier als Benchmark angelegt. Ziel war es, eine Präzision in der Vorhersage zu erreichen, die den e-rater®-Algorithmen annähernd entspricht.

2. Kann das Verfahren des Gradient Boosting die Vorhersagegenauigkeit der Textqualität gegenüber dem traditionellen Verfahren der linearen Regression verbessern?

Dazu analysieren wir, inwiefern die Vorhersagegenauigkeit der Textqualität durch unterschiedliche statistische Verfahren verbessert werden kann. Die traditionelle Methode der linearen Regression wurde mit Gradient Boosting verglichen, um Teile der Forschungsergebnisse von Sinharay et al. (2019) im europäischen Kontext zu replizieren. Es wurden analog zu Sinharay et al. (2019) leichte Vorteile des Boosting Verfahrens bei der Vorhersagegenauigkeit erwartet.

3. Wie hoch sind die Zusammenhänge der mit Hilfe von CTAP vorhergesagten Urteile mit Außenkriterien?

Zur Überprüfung der Validität untersuchen wir die Zusammenhänge der durch die Kodierungen von CTAP vorhergesagten Urteile mit dem von e-rater® vorhergesagten automatisiertem Urteil und der Englischnote der Schülerinnen und Schüler.

Methode

Textkorpus

Die analysierten Texte der vorliegenden Arbeit stammen aus dem Projekt *Measuring English Writing Skills at Secondary Level (MEWS)*; siehe Keller et al., 2020, Köller et al., 2019), das als Kooperationsprojekt zwischen dem Leib-

niz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) und der Pädagogischen Hochschule der Fachhochschule Nordwestschweiz (PH FHNW) durchgeführt wurde. MEWS wurde von der Deutschen Forschungsgemeinschaft (DFG) und vom Schweizer Nationalfonds (SNF) gefördert und untersucht in einem mehrbenenanalytischen Messwiederholungsdesign Englischleistungen in der Qualifikationsphase der gymnasialen Oberstufe (11. Jahrgangsstufe im achtjährigen Gymnasium) in deutschen und Schweizer Oberstufen. In diesem Projekt wurden von $N = 2420$ Schülerinnen und Schülern (58.1% weiblich, Alter $M = 17.7$ Jahre) Texte zum Prompt „Television Advertising“, einer Aufgabe zum argumentativen Schreiben aus dem TOEFL iBT®, verfasst. Bei dieser Aufgabe werden Schülerinnen und Schüler mit folgender Aussage konfrontiert: „Television advertising directed toward young children (aged two to five) should not be allowed.“ Anschließend sollen sie ihre Zustimmung oder Ablehnung zu dieser Aussage in einem Essay ausführen und mit geeigneten Argumenten begründen. Die Beschreibung der sechs Stufen zur Qualitätsbeurteilung der Texte (Rubrics) des ETS ist im Internet frei abrufbar (www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf; Abruf 16.09.2020). Auf der Basis des Rubrics wurden die Texte durch jeweils zwei trainierte Kodiererinnen und Kodierer des ETS auf einer holistischen Beurteilungsskala von 0 – 5 eingeschätzt, welche in vorherigen TOEFL-Durchführungen zu den 30 % der „genauesten“ Kodiererinnen und Kodierer gehört hatten. Um eine hohe Güte bei der Bewertung zu erreichen, wurden die Kodiererinnen und Kodierer vor jedem Bewertungsdurchgang kalibriert. Dabei mussten sie Texte, deren Urteile aus früheren Studien bekannt waren, einschätzen. Erreichten sie hier eine unzureichende Übereinstimmung, so wurden sie von den Kodierungen ausgeschlossen (zum genauen Vorgehen Rupp et al., 2019). Um die Reliabilität der beurteilten Textqualität zu erhöhen, wurde der Mittelwert beider Kodiererinnen / Kodierer als Maß für Textqualität verwendet und durch die von CTAP kodierte Textmerkmale vorhersagt. Bei einer Abweichung von mehr als einer Stufe zwischen beiden Kodiererinnen / Kodierern wurde ein „Adjudicator-Score“ vergeben, das heißt, der Text wurde von einer weiteren Expertin bzw. einem weiteren Experten beurteilt. Wenn dies der Fall war, wurde der Text von unserer Analyse ausgeschlossen ($n = 42$). Weiterhin wurden die Texte ausgeschlossen, die ein Urteil von 0 hatten (z.B. in einer anderen Sprache verfasst, anderes Thema; $n = 7$). Es wurden auch Texte von Schülerinnen und Schülern ausgeschlossen, für die kein e-rater® Wert vorlag ($n = 43$) oder keine Englischnote der betreffenden Schülerinnen und Schüler im Datensatz vorhanden war ($n = 172$), sowie Texte, für die mindestens ein CTAP Score nicht berechnet werden konnte ($n = 23$). Dadurch resultierte eine

finale Stichprobengröße von $N = 2179$ Texten, die in die Textkodierungssoftware CTAP (Chen & Meurers, 2016) eingelesen wurden.

Kodierung von Textmerkmalen

Da wir an einer möglichst genauen Vorhersage der holistischen Beurteilungen durch Kodiererinnen und Kodierer interessiert waren, wurden alle verfügbaren Textmerkmale von CTAP angefordert. Es wurden insgesamt 173 lexikalische und syntaktische Textmerkmale ausgegeben. Das lexikalische Textmerkmal *NumberOfTokenswithMoreThan2Syllable* kodiert beispielsweise die Anzahl der Token mit mehr als 2 Silben. Das *LexicalSophisticationFeat reSUBTLEXWordFrequencyAWTtype* berechnet die lexikalische Komplexität durch einen Vergleich mit einer Worthäufigkeitsliste auf Basis der SUBTLEX Datenbank (z.B. siehe Brysbaert & New, 2009). Das Feature *SD-TokenLengthinLetters* berechnet die Standardabweichung der Tokenlänge in Anzahl der Buchstaben. Das syntaktische Textmerkmal *Number of Syntactic Constituents: Verb Phrase* zählt die Anzahl der Verbalphrasen im Text. Alle verwendeten Indizes werden im elektronischen Supplement (ESM) 1 aufgeführt und beschrieben.

Statistische Analysen

Zur Vorhersage des holistischen Gesamturteils wurde Gradient Boosting mit einer schrittweisen linearen Regression verglichen. Alle Analysen wurden mit frei verfügbaren R-Paketen durchgeführt (Paket gbm, Ridgeway, 2007). Das Vorgehen wurde analog zu Sinharay et al. (2019) gewählt: Es wurden zufällig zwei Drittel der Texte gezogen, um das Vorhersagemodell zu gewinnen („Modell-Set“), das letzte Drittel diente als „Test-Set“ zum Validieren der gefundenen Vorhersage. Sowohl die lineare Regression als auch das Boosting erfolgte für 100 Replikationen, um eine robustere Vorhersage zu erreichen. Als Indizes zum Vergleich der Vorhersagegüte des Modells wurden Root-Mean-Square-Deviations (RMSD), Korrelationen, Prozent exakter Übereinstimmung (exact agreement) und angrenzender Übereinstimmung (adjacent agreement) berechnet. Als Grundlage für die Berechnung dieser Indizes zur Vorhersagegenauigkeit diente der Test-Set Datensatz. Als zusätzliche Indikatoren für die Validität der automatisierten Urteile wurde jeweils die Korrelation mit der Vorhersage des e-raters® und der Englischnote berechnet. Aufgrund der unterschiedlichen Notensysteme in Deutschland und der Schweiz wurden zunächst beide Noten z-standardisiert und dann als eine gemeinsame Variable behandelt. Um die Unterschiede zwischen Korrelationen auf Signifi-

kanz zu testen, wurde das R-Paket *cocor* (Diedenhofen & Musch, 2015) eingesetzt. Für die Berechnung der Deskriptiva und Korrelationen wurde der Gesamtdatensatz zugrunde gelegt, in dem die Vorhersageformel, die für das Modell-Set gefunden wurde, auf die Gesamtstichprobe angewandt wurde.

Gradient Boosting (GB) ist eine Methode zur Klassifikation und Regression von Daten in großen Datensätzen und gehört damit zu den Data Mining Verfahren. Dabei wird ein Vorhersagemodell durch die schrittweise Integration von Entscheidungsbäumen verbessert. Ein Entscheidungsbaum ist ein Verfahren zur Vorhersage von Werten mithilfe einer aufeinander aufbauenden Abfolge von dichotomen Entscheidungen, die durch den Vergleich eines Prädiktors mit einem kritischen Wert getroffen werden. Zu Beginn des GB in dieser Studie besteht das Vorhersagemodell aus dem mittleren Expertenurteil über alle Texte für dieses Merkmal. Daraufhin wird für alle Texte der Abstand zwischen der Vorhersage mit diesem Modell und dem Expertenurteil für jeden Text berechnet. Diese Abstände für jeden Text, die Residuen, werden in dem nächsten Schritt mithilfe eines Entscheidungsbaumes mit den von CTAP analysierten Texteigenschaften vorhergesagt. Dieser Entscheidungsbaum wird daraufhin mit einem Faktor multipliziert, der sogenannten *Learning Rate*, und mit der anfänglichen Vorhersage, dem mittleren Expertenurteil, addiert, wodurch ein neues Vorhersagemodell entsteht. Im nächsten Schritt werden die Residuen für jeden Text des neuen Vorhersagemodells berechnet. Ist die Summe der Residuen kleiner als im vorherigen Schritt, hat sich die Vorhersage verbessert und es werden im nächsten Schritt die Residuen erneut durch einen Entscheidungsbaum vorhergesagt. Der neue Entscheidungsbaum wird ebenfalls multipliziert mit der *Learning Rate* und zu der Summe aus dem mittleren Expertenurteil und dem ersten

Entscheidungsbaum addiert. Dieses Verfahren wird bis zu einer vordefinierten Anzahl von Schritten fortgeführt oder bis ein neuer Entscheidungsbaum die Summe der Residuen nicht mehr reduziert, also der Baum die Vorhersage nicht mehr verbessert. Eine wichtige Entscheidung besteht also in der Festlegung dieser Einstellungen für Anzahl der Entscheidungsbäume, Interaction Depth und Learning Rate (*Tuning Parameters*; siehe Sinharay et al., 2019). In Orientierung an Sinharay et al. (2019) benutzten wir folgende Einstellungen für unsere Analysen: die Anzahl der Entscheidungsbäume pro Durchgang lag bei $n = 500$, die Komplexität der Entscheidungsbäume bzw. die Anzahl der Entscheidung (Interaction Depth) wurde auf 3 festgesetzt. Shrinkage beschreibt die Learning Rate und wurde auf 0.01 festgesetzt. Eine detaillierte Beschreibung des Boosting Ansatzes findet sich bei Friedman (2002) und Sinharay et al. (2019) oder für eine allgemeinere Einführung in das Machine Learning siehe auch James et al. (2013).

Ergebnisse

Korrelationen

Deskriptive Statistiken zu den Urteilen der menschlichen Kodierenden und Kodierer, Scores des e-rater® sowie Resultate der automatischen Kodierungen durch Regression und Boosting auf der Basis von CTAP sind in Tabelle 1 dargestellt. Weiterhin werden die Korrelationen der verschiedenen automatisch generierten Urteile mit den Englischnoten als weiteres Validitätskriterium angegeben (Fragestellung 3). Dabei fällt auf, dass die Korrelationen mit Noten für den e-rater® ($r = .45$) leicht höher ausfallen

Tabelle 1. Deskriptive Statistiken für die Textgüteurteile und Korrelationen zwischen den Maßen

	Deskriptive		Korrelationen (p -Werte unter der Diagonalen)			
	<i>M</i>	<i>SD</i>	e-rater®	CTAP – Regression	CTAP – Boosting	Englisch Note
Menschliche Urteile	3.17	0.73				
E Rater Score – beste Vorhersage	3.16	0.66	e-rater®	.72	.73	.45
CTAP – Regression	3.16	0.63	CTAP-Regression	<.001	.80	.38
CTAP – Boosting	3.16	0.57	CTAP- Boosting	<.001	<.001	.37
Note D/CH	9.19/ 4.70	2.54/ 0.54	Englisch Note	<.001	<.001	<.001

Anmerkungen: Deskriptive und Korrelationen beruhen auf den gerundeten Ergebnissen der Vorhersage des Gesamtdatensatzes; Die Vorhersageformel wurde nach der Prüfung der Vorhersagegenauigkeit (siehe Tabelle 2) auf Basis des Testsets auf den Gesamtdatensatz angewandt, um die Korrelationen zu berechnen; Die Noten in Deutschland sind auf einer Skala von 0 – 15, die Noten in der Schweiz auf einer Skala von 0 bis 6. In beiden Ländern stehen höhere Werte für bessere Leistungen; Um die Korrelationen zu berechnen, wurden beide Variablen erst z-standardisiert und danach zusammengeführt; D: Deutschland; CH: Schweiz.

als für die Ergebnisse mit CTAP ($r = .37$ für das Boosting bzw. $r = .38$ für die Regression). Der Unterschied ist statistisch signifikant (Vergleich der Korrelation des e-rater® mit Noten und der Korrelation der Vorhersagen mit CTAP auf Basis der multiplen linearen Regression mit Noten: $z = 4.52$, $p < .001$; für den Vergleich mit der Vorhersage auf Basis von CTAP mit Boosting: $z = 5.56$, $p < .001$).

Die Korrelationen des e-rater® mit dem Urteil der menschlichen Kodiererinnen und Kodierer sind höher als die Korrelationen der CTAP Vorhersagen mit dem Urteil der menschlichen Kodiererinnen und Kodierer (für die Regression $z = -3.55$, $p < .001$; für Boosting $z = -4.22$, $p < .001$). Zwischen den Korrelationen der menschlichen Urteile mit den Vorhersagen durch CTAP für Boosting und Regression bestehen keine signifikanten Unterschiede ($z = -0.86$, $p = .3911$).

Die Korrelation der gerundeten Werte des e-rater® mit den Vorhersagen durch CTAP liegt bei $r = .72$ bzw. $r = .73$ und ist damit jeweils signifikant geringer als die Übereinstimmung der Vorhersagen durch die verschiedenen CTAP-Werte (Boosting und Regression, $r = .80$; $z = -6.83$, $p < .001$ bzw. $z = -8.09$, $p < .001$). Die Unterschiede in den Korrelationen der CTAP Ergebnisse (Boosting vs. Regression) mit der Vorhersage des e-rater® sind nicht signifikant ($z = 1.38$, $p = .169$). Insgesamt deuten diese Ergebnisse darauf hin, dass die Qualität der Vorhersagen auf Basis von CTAP mit linearer Regression und Boosting vergleichbar ist, und dass diese signifikant unter der Qualität der Vorhersagen des e-rater® liegt. Trotz der statistischen Signifikanz der Unterschiede ist das Ausmaß als eher klein zu beschreiben. Auch die mit CTAP erstellen Analysen dürfen damit als valide Vorhersagen des menschlichen holistischen Scores betrachtet werden.

Vorhersagegenauigkeit

Die Ergebnisse der Vorhersagegenauigkeit für Boosting und lineare Regression sind in Tabelle 2 dargestellt. In Bezug auf Fragestellung 1 zeigt sich, dass die automatisierte

Kodierung durch CTAP mit der Übereinstimmung der Vorhersage des e-raters® und der Urteile der menschlichen Kodiererinnen und Kodierer aus der MEWS Studie vergleichbar ist. In MEWS lag die exakte Übereinstimmung über das Gesamtsample für den Prompt „TV Advertising“ bei 41.8%, und die Korrelation bei $r = .81$ (Rupp et al., 2019). Für das *adjacent agreement* fanden Rupp et al. (2019) eine Genauigkeit von 99.3%. Hier finden wir unter Verwendung des Boostings ähnliche Ergebnisse mit einer exakten Übereinstimmung von 41.1% und einer Korrelation von $r = .75$ (siehe Tabelle 2) sowie einem adjacent agreement von 99%. Für den Vergleich von Boosting mit linearer Regression in Fragestellung 2 zeigen sich analog zu den Ergebnissen von Sinharay et al. (2019) leichte Vorteile für das Boosting in allen Indizes der Vorhersagegenauigkeit (siehe Tabelle 2). Eine Ausnahme ist hier die Genauigkeit des adjacent agreement, die für beide Verfahren bei 99% liegt.

Diskussion

In dieser Arbeit wurde die freiverfügbare Software CTAP (Chen & Meurers, 2016) eingesetzt, um eine große Anzahl komplexer Texte in Englisch als Fremdsprache automatisch hinsichtlich lexikalischer und syntaktischer Merkmale zu kodieren und mit den so gewonnenen Textmerkmalen die Urteile trainierter Kodiererinnen und Kodierer vorherzusagen. Die Genauigkeit der Vorhersage war vergleichbar mit dem Ergebnis der kommerziellen Software e-rater® des ETS. Dies ist insofern erstaunlich, als e-rater® über verschiedene Features verfügt, welche spezifisch für die Verwendung mit den argumentativen TOEFL-Prompts programmiert und evaluiert wurden (Rupp et al., 2019). In der vorliegenden Studie konnte ein Vorhersagemodell mit nur leicht geringerer Validität mit einem unspezifischeren Open-Source-Analysetool erzeugt werden.

Der Vergleich unterschiedlicher Analyseverfahren zeigte, dass die Vorhersagegenauigkeit durch Gradient Boosting im Vergleich zur schrittweisen linearen Regression

Tabelle 2. Ergebnisse Boosting vs. Lineare Regression für die Vorhersagegenauigkeit der CTAP-Parameter für die Textqualität

	Regression		Boosting	
	M über 100 Replikationen	SD über 100 Replikationen	M über 100 Replikationen	SD über 100 Replikationen
RMSD (root mean square deviation)	0.51	0.09	0.48	0.01
Korrelation	0.72	0.06	0.75	0.02
Anteil exakte Übereinstimmung	0.41	0.02	0.42	0.02
Anteil adjacent agreement	0.99	0.03	0.99	0.00

Anmerkungen: adjacent agreement = Abweichung des Urteils um maximal eine Stufe; Die Vorhersagegenauigkeit bezieht sich auf das Testsample (1 / 3 des Gesamtdatensatzes).

etwas höher liegt, und repliziert somit Teile der Forschungsergebnisse von Sinharay et al. (2019). Diese Unterschiede schlagen sich jedoch nicht in den Korrelationen der gerundeten Werte mit Validitätsindikatoren (Englischnote, Urteile der menschlichen Kodiererinnen und Kodierer; Vorhersage des e-rater®) nieder. Hier zeigte sich, dass die Ergebnisse des e-rater® etwas höhere Korrelationen mit den Englischnoten aufwiesen als die Vorhersagen durch CTAP, und zwar sowohl auf Basis der linearen Regression als auch des Boosting.

Die leichten Vorteile des Boosting im Vergleich zur linearen Regression beim Vergleich der Indizes zur Vorhersagegenauigkeit bedeuten nicht, dass Boosting unter allen Bedingungen und bei beliebigen Textstimuli zu einer erhöhten Vorhersagegenauigkeit führt (vgl. Sinharay et al., 2019; insbesondere für die Rolle der Tuning Parameter). Weiterhin ist der Vorteil in der Vorhersagegenauigkeit im Vergleich zur schrittweisen linearen Regression als gering einzuschätzen, so dass Aufwand und Nutzen für den jeweiligen Kontext abgewogen werden sollten. Insbesondere wenn es um die Generierung von Vorhersagealgorithmen für den Einsatz in der Praxis geht, sollten verschiedene Methoden gegeneinander getestet werden, um die für den Einzelfall bestmögliche Vorhersage zu bestimmen. Dieser Prozess erfordert jedoch ein hohes Maß an Ressourcen und Expertise (vgl. Rupp et al., 2019). Unsere Studie zeigt, dass die Vorhersage auch für herkömmliche Verfahren wie der schrittweisen linearen Regression nah an den Optimalstandard eines spezialisierten Forschungsinstituts wie dem ETS heranzureichen scheint. Dennoch sprechen die Ergebnisse für die Zusammenhänge mit Englischnote als Außenkriterium für eine leichte Überlegenheit der Software e-rater®.

Limitationen

Die in dieser Studie verwendeten Vorhersagealgorithmen wurden spezifisch für diesen Textkorpus angepasst und sind nicht direkt auf ähnliche Textkorpora übertragbar. Es ist im Moment noch schwierig, generische Modelle zu finden, die sich auf unterschiedliche Aufgaben und Schülerpopulationen anpassen lassen, ohne signifikant an Vorhersagegüte zu verlieren (siehe Rupp et al., 2019). Ein wichtiges Ziel der weiteren Forschung ist es deshalb, generische Modelle zur Vorhersage von Textqualität zu ermöglichen, um die Einsatzmöglichkeiten der automatisierten Beurteilung auszubauen. Eine weitere Möglichkeit zur Entwicklung generischer Modelle könnte in der Verwendung von Prozessdaten in Form von Keystroke Log Daten (siehe Sinharay et al., 2019) liegen, die das Tipp- und Korrekturverhalten der Schreibenden abbilden, und nicht nur das fertige Textprodukt. Obwohl die Vorhersagegenauigkeit für diese Prozessdaten in der Studie von Sinharay et

al. (2019) im Vergleich zur Analyse der Texteigenschaften etwas geringer war, ist es denkbar, dass diese Prozessmerkmale zu Urteilen führen könnten, die weniger durch die spezifische Aufgabe beeinflusst und deshalb stärker zu verallgemeinern wären. Diese Hypothese muss jedoch zunächst empirisch überprüft werden. Eine Alternative ist die Generierung weiterer Textkorpora, die wie in dieser Studie für das Machine Learning eingesetzt werden können und somit eine Verbesserung des Algorithmus ermöglichen können. Wichtig ist hierbei die Qualität der Expertenurteile. Die automatisierte Textbeurteilung kann jeweils nur so gut sein wie der zugrundeliegende Expertenscore (siehe Rupp et al., 2019). Zukünftige Forschung sollte somit einen Fokus auf die qualitativ hochwertige und nachvollziehbare Beurteilung der Textkorpora legen, um damit reliable und ökonomische Auswertungen von Texten in Forschung und Praxis für eine Vielzahl unterschiedlicher Textgenres und Aufgaben zu ermöglichen.

Weiterhin spielt die Operationalisierung der Textqualität eine wichtige Rolle. In dieser Studie wurde Textqualität durch jene Merkmale der lexikalischen und syntaktischen Komplexität operationalisiert, welche CTAP standardisiert zu erfassen vermag. Dahinter steht die linguistisch plausible Annahme, dass kompetentere Schreiberinnen und Schreiber in der Fremdsprache auch komplexere Texte hervorbringen (Weigle, 2013). Allerdings unterschied sich das Konstrukt von Textqualität von jenem vieler Lehrkräfte, welche typischerweise auch sprachliche Fehler (z.B. bei Rechtschreibung und Grammatik) in ihre Textbeurteilungen einbeziehen (Vögelin et al., 2018). Analog zu den Rubrics der durch die Expertinnen und Experten verwendeten TOEFL® Skala werden kleine sprachliche Fehler nicht berücksichtigt und auch auf den höchsten Stufen toleriert. Da es sich hier allerdings spezifisch um Schreiben in der Fremdsprache handelt, ist dies eine fachdidaktisch durchaus plausible Operationalisierung von Textqualität. Nicht zielsprachenkonforme Verwendungen, die über die genannten singulären Fehler in Grammatik und Lexis hinausgehen, beeinflussen die Textqualität negativ, indem die Wörter oder Satzkonstruktionen für die Software nicht mehr erkennbar machen und somit die lexikalische und linguistische Beurteilung des Textes negativ beeinflussen. Dennoch sollte in zukünftigen Studien die Rolle solcher sprachlichen Fehler explizit untersucht werden, um die Frage zu beantworten, ob eine Analyse von Rechtschreib- und Grammatikfehlern durch automatisierte Programme die Vorhersagekraft der Modelle möglicherweise verbessern könnte.

Implikationen

Die vorliegende Studie hat Implikationen für Forschung und Praxis. Zum einen zeigt sie ein Vorgehen, mit dem jede

Forschungsgruppe mit Hilfe von Open Source Software Schülertexte automatisiert auswerten lassen kann, wenn geeignete Urteile von Expertinnen und Experten beziehungsweise gut geschulten Kodiererinnen und Kodierern vorliegen. Diese Art der ökonomischen Auswertung kann in Zukunft erlauben, mehr Zeit und Energie in qualitativ hochwertige menschliche Urteile zu investieren und dann automatisierte Auswertung zu benutzen, um damit eine vielfache Anzahl an Texten (eventuell auch späteren Zeitpunkten) zu kodieren und somit die Gesamtinvestition gering zu halten. Die Studie kann somit als Beispiel für ökonomische Textkodierung in großen Schulleistungstudien angesehen werden. Allerdings beurteilen alle momentan verfügbaren Methoden von Machine Scoring Texte im Wesentlichen auf der Basis linguistischer Indizes, während Aspekte wie inhaltliche Adäquatheit oder Qualität der Argumentation kaum ins Urteil einfließen. Solche Methoden sollen deshalb immer nur in Kombination mit mindestens einem menschlichen Urteil angewendet werden, zumindest wenn daraus sinnvolle Rückmeldungen für Schülerinnen und Schüler resultieren sollen.

Des Weiteren ist in Bezug auf die praktischen Implikationen anzumerken, dass das Ziel der automatischen Kodierung von Schülertexten insbesondere in der raschen automatisierten Rückmeldung für Schülerleistungen besteht. Das bedeutet, dass es in Zukunft bei hoher Beurteilungsgenauigkeit möglich sein wird, Schülerinnen und Schülern in Echtzeit und mit geringem Aufwand Feedback zur Qualität des gerade geschriebenen Textes zu geben. Dabei ist zu berücksichtigen, dass für die Praxis einer solchen automatisierten Rückmeldung gerade ihre Fehleranfälligkeit das entscheidende Kriterium darstellt: das Feedback kann im Einzelfall Schülerinnen und Schüler verwirren, wenn es aufgrund einer nicht hinreichenden Beurteilungsgenauigkeit fehlerhaft ist. Damit kann es zum einen die Akzeptanz des Feedbacks verringern, zum anderen auch die Motivation und den Leistungsfortschritt der Schülerinnen und Schüler einschränken (siehe aber Bai & Hu, 2017). Gleichzeitig stellt sich die Frage möglicher Alternativen. Im Bereich der Urteils-genauigkeit von Lehrkräften ist bereits bekannt, dass auch Lehrkräfte Urteilsfehler und Beurteilungsheuristiken unterliegen, insbesondere bei der Beurteilung komplexer schriftlicher Leistungen (Jansen et al., 2019). Der Einsatz automatisierter Rückmeldungen zu schriftlichen Schülerleistungen in der Schulpraxis sollte somit nicht davon abhängen, ob eine perfekte Beurteilungsgenauigkeit durch diese Systeme erreicht werden kann. Vielmehr sollte in Betracht gezogen werden, ob diese Systeme die Urteile von Lehrkräften sinnvoll unterstützen oder komplementieren können, oder aber mindestens Urteile mit einer vergleichbaren Genauigkeit ermöglichen. Wenn man die Kriterien für die Nützlichkeit und Qualität solcher Systeme sinnvoll an-

setzt, und weitere Textkorpora zu verschiedenen Aufgaben und Textgenres mit verlässlichen menschlichen Urteilen zur Generierung eines Beurteilungsalgorithmus zur Verfügung gestellt werden können, kann der Einsatz künstlicher Intelligenz im Sinne einer automatisierten Textbewertung in der Praxis geprüft werden. Darauf aufbauend sollte in zukünftiger Forschung verstärkt untersucht werden, inwiefern solches Feedback die Schreibleistungen von Schülerinnen und Schülern verbessert (z. B. Li et al., 2014). Hierzu gibt es bereits Forschung, die sich aber hauptsächlich mit Effekten der Rückmeldung von Oberflächenmerkmalen des Schreibprodukts beschäftigt (für einen aktuellen Überblick siehe Strobl et al., 2019). Möglicherweise sind hier zusätzlich spezifischere Urteile zu Facetten der Textqualität, beispielsweise zum Aufbau und der Struktur des Textes, notwendig, um das Feedback für die Schülerinnen und Schüler nutzbar zu machen. Beispielsweise schlagen Hattie und Timperley (2007) vor, dass erfolgreiches Feedback drei Fragen beantworten muss, unter anderem, was der nächste Schritt zur Verbesserung und Erreichung des Ziels sein muss. Dazu können spezifischere Rückmeldungen sinnvoll sein. Um diese Art von Rückmeldungen zu automatisieren, wäre ein Vorgehen analog zu dem in dieser Studie beschriebenen sinnvoll. Dafür wäre es aus schreibdidaktischer Sicht zentral, neben einer „holistischen“, d. h. ganzheitlichen Rückmeldung den Lernenden auch Informationen zu unterschiedlichen Qualitätsaspekten ihrer Text rückmelden zu können. Dazu ist es notwendig, dass auch für einzelne Textmerkmale wie Vokabular, Syntax, aber auch Argumentation, Organisation usw. automatisierte Vorhersagemodelle generiert werden können.

Weiterhin muss zukünftig auch für den deutschsprachigen Kontext untersucht werden, inwiefern automatisiertes Feedback zum einen von den Schülerinnen und Schülern, aber auch von den Lehrkräften akzeptiert wird. Für den amerikanischen Raum und die Erstsprache Englisch gibt es bereits einige Befunde zum Einsatz automatisierten Feedbacks und den Effekten auf die Leistungen der Schülerinnen und Schüler (z. B. Strobl et al., 2019; Wilson & Czik, 2016). Es zeigen sich positive Effekte auf das Feedback der Lehrkräfte, insbesondere auf die Zeitersparnis, sowie auf die Motivation der Schülerinnen und Schüler. Weitere Forschung ist notwendig, um diese Ergebnisse auf den fremdsprachlichen und europäischen Kontext zu übertragen. Unterschiede aufgrund anderer schulischer Traditionen in Bezug auf standardisierte Tests und die automatisierte Auswertung sind zu erwarten, die überprüft werden müssen. Es ist vorstellbar, dass Schülerinnen und Schüler das Feedback der automatisierten Rückmeldung weniger ernst nehmen, oder aber gerade aufgrund der Automatisierung für genauer und weniger fehleranfällig halten im Vergleich zu Lehrerfeedback. Für Lehrkräfte ist

anzunehmen, dass das automatisierte Feedback in der didaktischen Ausbildung bisher noch weitgehend unbekannt ist und deshalb zunächst kritisch gesehen wird. Hier sollte der Fokus zukünftiger Forschungen auf der Frage liegen, wie und unter welchen Umständen Lehrkräfte die automatisierte Textauswertung für sich und ihren Unterricht möglichst effektiv nutzen und einsetzen können. Weiterhin sollten Effekte automatisierten Feedbacks auf die Schreibmotivation untersucht werden.

Zusammenfassend können wir mit dieser Studie zeigen, dass automatische Textkodierung für fremdsprachige schriftliche Englischkompetenzen unter Verwendung von Open Source Software zur Textkodierung mit einer hinreichenden Vorhersagegenauigkeit einsetzbar ist. Wir zeigen weiterhin, dass mit Open Source Softwarelösungen wie CTAP zur Kodierung linguistischer Merkmale eine ähnliche Vorhersagegenauigkeit erzielt werden kann, wie mit den etablierten und kommerziellen Programmen des ETS. Und schließlich konnten wir darstellen, dass diese Möglichkeiten der automatisierten Textbeurteilung sich auch für den europäischen Kontext im Fremdsprachenunterricht eignen, womit sich spannende Perspektiven für die ökonomische Textbeurteilung in großen Schulleistungstudien wie auch die Entwicklung neuer Feedbacksysteme im Unterricht eröffnen.

Elektronisches Supplement (ESM)

Das elektronische Supplement ist mit der Online-Version dieses Artikels verfügbar unter <https://doi.org/10.1024/1010-0652/a000296>

ESM 1. Liste der in CTAP verwendeten linguistischen Textmerkmale und Beschreibung auf Englisch, sowie deutsche Übersetzung (Table E1)

Literatur

- Almond, R., Deane, P., Quinlan, T., Wagner, M. & Sydorenko, T. (2012). A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report Series*, 2012 (2), i-61.
- Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4 (3).
- Bai, L. & Hu, G. (2017). In the face of fallible AWE feedback: how do students respond? *Educational Psychology*, 37 (1), 67–81. <https://doi.org/10.1080/01443410.2016.1223275>
- Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49 (3), 219–224.
- Bridgeman, B. (2013). Human Ratings and Automated Essay Evaluation. In Shermis, M. D. & Burstein, J. (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 221–232). New York: Routledge.
- Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Burstein, J., Chodorow, M. & Leacock, C. (2003). CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *IAAI* (S.3–10).
- Chen, X. & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (S.113–119).
- Chenoweth, N. A. & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18 (1), 80–98. <https://doi.org/10.1177/0741088301018001004>
- Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*, 2014 (1), 1–23. <https://doi.org/10.1002/ets2.12002>
- Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One*, 10 (3), e0121945. <https://doi.org/10.1371/journal.pone.0121945>
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5 (1).
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25 (2), 155–185. <https://doi.org/10.1177/0265532207086780>
- EDK (2017). *Empfehlungen zum Fremdsprachenunterricht (Landessprachen und Englisch) in der obligatorischen Schule*. Verfügbar unter https://edudoc.ch/record/128697/files/empfehlungen_sprachenunterricht_d.pdf
- Elliott, J., Lee, S. W. & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills-Modified. *School Psychology Review*, 30 (1), 33–49.
- Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. & Köller, O. (2020). Linking TOEFL iBT® Writing Rubrics to CEFR Levels: Cut Scores and Validity Evidence from a Standard Setting Study. *Assessing Writing*, 41. <https://doi.org/10.1016/j.asw.2019.100420>
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38 (4), 367–378.
- Graesser, A. C. & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Hrsg.), *APA handbooks in psychology®. APA handbook of research methods in psychology, (Foundations, planning, measures, and psychometrics, Bd. 1, S.307–325)*. American Psychological Association. <https://doi.org/10.1037/13619-017>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36 (2), 193–202. <https://doi.org/10.3758/BF03195564>
- Guo, L., Crossley, S. A. & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18 (3), 218–238. <https://doi.org/10.1016/j.asw.2013.05.002>
- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77 (1), 81–112. <https://doi.org/10.3102/003465430298487>
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational research*, 60 (2), 237–263. <https://doi.org/10.3102/00346543060002237>
- Huot, B. (1996). Toward a new theory of writing assessment. *College composition and communication*, 47 (4), 549–566.

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Bd. 112, S.3 – 7). New York: Springer.
- Jansen, T., Vögelin, C., Machts, N., Keller, S. D. & Möller, J. (2019). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen. *Psychologie in Erziehung und Unterricht*, 66 (4), 303 – 315.
- Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe. Theorie, Prozessgestaltung, Empirie*. Tübingen: Narr.
- Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O. & Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing*. <https://doi.org/10.1016/j.jslw.2019.100700>
- KMK (2012). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz von 18. 10. 2012*. Berlin: KMK. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Fortgef-FS-Abi.pdf
- Köller, O., Fleckenstein, J., Meyer, J., Paeske, A. L., Krüger, M., Rupp, A. A. & Keller, S. (2019). Schreibkompetenzen im Fach Englisch in der gymnasialen Oberstufe. *Zeitschrift für Erziehungswissenschaft*, 22 (6), 1281 – 1312. <https://doi.org/10.1007/s11618-019-00910-3>
- Kronig, W. (2007). *Die systematische Zufälligkeit des Bildungserfolgs. Theoretische Erklärungen und empirische Untersuchungen zur Lernentwicklung und zur Leistungsbewertung in unterschiedlichen Schulklassen*. Bern: Haupt.
- Leckie, G. & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48 (4), 399 – 418. <https://doi.org/10.1111/j.1745-3984.2011.00152.x>
- Li, Z., Link, S., Ma, H., Yang, H. & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, 44, 66 – 78. <https://doi.org/10.1016/j.system.2014.02.007>
- McNamara, D. S., Crossley, S. A. & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45 (2), 499 – 515. <https://doi.org/10.3758/s13428-012-0258-1>
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K. & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35 – 59. <https://doi.org/10.1016/j.asw.2014.09.002>
- McNamara, D. S. & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In *Applied natural language processing: Identification, investigation and resolution* (S.188 – 205). IGI Global. <https://doi.org/10.4018/978-1-60960-741-8.ch011>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M. & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511894664>
- Meadows, M. & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: AQA Centre for Education Research and Policy.
- Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, 1 (1), 2007.
- Rudner, L. M., Garcia, V. & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4 (4).
- Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S. & Köller, O. (2019). Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. *ETS Research Report Series*, 4 (3), 1 – 23. <https://doi.org/10.1002/ets2.12249>
- Shermis, M. D. & Burstein, J. (Hrsg.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge. <https://doi.org/10.1558/wap.v6i2.437>
- Shermis, M. D., Burstein, J., Higgins, D. & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education*, 4 (1), 20 – 26.
- Sinharay, S., Zhang, M. & Deane, P. (2019). Prediction of Essay Scores From Writing Process and Product Features Using Data Mining Methods. *Applied Measurement in Education*, 32 (2), 116 – 137. <https://doi.org/10.1080/08957347.2019.1577245>
- Strobl, C., Ailhard, E., Benetos, K., Devitt, A., Kruse, O., Proske, A. & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, 131, 33 – 48. <https://doi.org/10.1016/j.compedu.2018.12.005>
- Terhart, E. (2002). *Standards für die Lehrerbildung. Eine Expertise für die Kultusministerkonferenz*. Münster: Waxmann Verlag.
- Vögelin, C., Jansen, T., Keller, S. & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal*. <https://doi.org/10.1080/09571736.2018.1522662>
- Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research*, 10 (2), 157 – 180. <https://doi.org/10.1191/1362168806lr190oa>
- Weigle, S. C. (2005). Second language writing expertise. In: K. Johnson (Hrsg.), *Expertise in language learning and teaching* (S.128 – 149). Hampshire, UK: Palgrave Macmillan.
- Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18 (1), 85 – 99. <https://doi.org/10.1016/j.asw.2012.10.006>
- Wilson, J. & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94 – 109. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27 (3), 291 – 300. <https://doi.org/10.1177/0265532210364643>
- Yan, D., Rupp, A. A. & Foltz, P. W. (Hrsg.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press. <https://doi.org/10.1201/9781351264808>
- Zhang, M. & Deane, P. (2015). Process features in writing: Internal structure and incremental value over product features. *ETS Research Report Series*, 2015 (2), 1 – 12. <https://doi.org/10.1002/ets2.12075>

Historie

Manuskript eingereicht: 24.04.2020

Manuskript nach Revision angenommen: 19.11.2020

Onlineveröffentlichung: 10.12.2020

Förderung

Die Daten der vorliegenden Studie stammen aus dem Projekt „Measuring English Writing at Secondary Level (MEWS)“, das von der Deutschen Forschungsgemeinschaft (DFG; GZ:K01513/12-1) und vom Schweizerischen National Fonds (SNF: Fördernr. 100019L162675) gefördert wird.

ORCID

Jennifer Meyer

<https://orcid.org/0000-0002-5714-3198>

Dr. Jennifer Meyer

Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN)
Olshausenstr 62
24118 Kiel
Deutschland
jmeyer@leibniz-ipn.de