Studien zum Physik- und Chemielernen

M. Hopf und M. Ropohl [Hrsg.]

378

Tom Bleckmann

Formatives Assessment auf Basis von maschinellem Lernen

Eine Studie über automatisiertes Feedback zu Concept Maps aus dem Bereich Mechanik

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf und Mathias Ropohl

Diese Reihe im Logos Verlag Berlin lädt Forscherinnen und Forscher ein, ihre neuen wissenschaftlichen Studien zum Physik- und Chemielernen im Kontext einer Vielzahl von bereits erschienenen Arbeiten zu quantitativen und qualitativen empirischen Untersuchungen sowie evaluativ begleiteten Konzeptionsentwicklungen zu veröffentlichen. Die in den bisherigen Studien erfassten Themen und Inhalte spiegeln das breite Spektrum der Einflussfaktoren wider, die in den Lehr- und Lernprozessen in Schule und Hochschule wirksam sind.
Die Herausgeber hoffen, mit der Förderung von Publikationen, die sich mit dem Physik- und Chemielernen befassen, einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Verbesserung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Martin Hopf und Mathias Ropohl

Studien zum Physik- und Chemielernen

Band 378

Tom Bleckmann

Formatives Assessment auf Basis von maschinellem Lernen

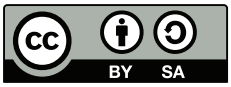
Eine Studie über automatisiertes Feedback zu Concept Maps aus dem Bereich Mechanik

Logos Verlag Berlin

Studien zum Physik- und Chemielernen
Martin Hopf und Mathias Ropohl [Hrsg.]

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.



Diese Publikation wurde von NiedersachsenOPEN, dem zentralen niedersächsischen Open-Access-Publikationsfonds, unterstützt.
We acknowledge the financial support of NiedersachsenOPEN, the Open Access Publishing Fund of Lower Saxony.

Logos Verlag Berlin GmbH 2024

ISBN 978-3-8325-5842-0
ISSN 1614-8967
DOI 10.30819/5842

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10
D-12681 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.de

Formatives Assessment auf Basis von maschinellem Lernen

Eine Studie über automatisiertes Feedback zu Concept Maps aus dem Bereich Mechanik

Der Fakultät für Mathematik und Physik
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
Dr. rer. nat.

vorgelegte Dissertation von

M.Ed. Tom Bleckmann

2024

M.Ed. Tom Bleckmann
Formatives Assessment auf Basis von maschinellem Lernen
Eine Studie über automatisiertes Feedback zu Concept Maps aus dem Bereich Mechanik

Referent: Prof. Dr. Gunnar Friege
Institut für Didaktik der Mathematik und Physik, Leibniz Universität Hannover
1. Korreferentin: Jun.-Prof. Dr. Peter Wulff
Institut für Naturwissenschaften, Geographie und Technik, Pädagogische Hochschule Heidelberg
2. Korreferentin: Prof. Dr. Eirini Ntoutsi
Institut für Datensicherheit, Universität der Bundeswehr München

Leibniz Universität Hannover
Institut für Didaktik der Mathematik und Physik
Arbeitsgruppe Physikdidaktik
Welfengarten 1A
30167 Hannover

Hinweise: Diese Arbeit wurde teilweise vom Niedersächsischen Ministerium für Wissenschaft und Bildung im Rahmen des Graduiertenkollegs „LernMINT: Datengestützter Unterricht in den MINT-Fächern“ (Projekt Nr. 51410078) unterstützt. Die Durchführung der Studien wurde von der Niedersächsischen Landesschulbehörde genehmigt.

Danksagung

An erster Stelle möchte ich mich bei meinem Doktorvater Prof. Dr. Gunnar Friege herzlich bedanken. Ihre Unterstützung und das stets sehr gute Betreuungsverhältnis haben maßgeblich zum Gelingen dieser Dissertation beigetragen. Ich bin besonders dankbar für die vielen schönen Momente auf Konferenzen, bei denen ich die Gelegenheit hatte, meine Arbeit vorzustellen. Sie waren eine entscheidende Unterstützung, und nicht nur die vielen Gespräche über Fußball werden mir immer in guter Erinnerung bleiben.

Ein herzlicher Dank gilt auch Jun.-Prof. Dr. Peter Wulff und Prof. Dr. Eirini Ntoutsi für die Begutachtung meiner Arbeit sowie die anregenden Diskussionen darüber. Prof. Dr. Siegfried Raasch danke ich für die Übernahme des Prüfungsvorsitzes und die angenehme Disputation.

Mein Dank richtet sich ebenfalls an alle ehemaligen und aktuellen Kolleg:innen der Arbeitsgruppe Physikdidaktik der Leibniz Universität Hannover. Die vielen wertvollen fachdidaktischen Diskussionen und der Austausch über mein Projekt haben mir meine Zeit sehr angenehm gemacht. Mein Dank gilt auch den studentischen Mitarbeiter:innen, vor allem Hannes und Antonia, die mich in vielen Dingen stark entlastet haben.

Dirk danke ich für die unzähligen schönen und lustigen Momente, auch nach Feierabend, sei es beim Bier trinken oder Bier brauen, sowie für die inspirierenden Gespräche, die mir noch lange positiv in Erinnerung bleiben werden.

Ein besonderer Dank geht an meinen Bürokollegen André. Du hast mir in schwierigen Phasen nicht nur mit Deinem Optimismus, sondern auch mit deinem einzigartigen Humor geholfen. Die gemeinsamen Kaffeepausen waren wahre Oasen im Arbeitsalltag, die mir geholfen haben, körperlich und seelisch durchzuhalten. Deine Unterstützung, auch außerhalb der Arbeit, war unbezahlbar, sei es bei spontanen oder geplanten Feierabendbieren. Ich werde unsere Gespräche über das Leben, die Arbeit und alles dazwischen immer in guter Erinnerung behalten und

freue mich auf viele weitere lustige Momente mit dir.

Ich möchte mich auch bei den vielen Mitgliedern des Graduiertenkollegs Lern-MINT bedanken. Die gemeinsamen Diskussionen haben meine Arbeit bereichert und ich konnte viel Neues lernen. Mein besonderer Dank gilt Wolfgang, der immer ein offenes Ohr für mich hatte und mich mit den richtigen Hinweisen unterstützt hat.

Ein großer Dank geht an alle Schüler:innen und Physiklehrer:innen, die sich bereit erklärt haben, an meiner empirischen Untersuchung teilzunehmen. Ohne Eure Hilfe und Kooperation wäre die vorliegende Arbeit nicht möglich gewesen.

Danke auch an die Kolleg:innen des ALTA Institute der Cambridge University für den tollen Forschungsaufenthalt in England, der es mir ermöglichte, meine Methoden aus einem neuen Blickwinkel zu betrachten und meine Arbeit zu verbessern.

Von ganzem Herzen danke ich meiner Familie, insbesondere meinen Eltern sowie meinen Freunden, die mich in meinem Vorhaben bestärkt und unterstützt haben. Ohne euch wäre das alles nicht möglich gewesen.

Abschließend möchte ich mich bei meiner Freundin Ann-Christin bedanken. Du hast mich durch deine Unterstützung, dein Verständnis und deinen Zuspruch gerade in schwierigen Phasen immer wieder motiviert, mein Ziel konsequent zu verfolgen. Du warst und bist meine größte Stütze und ohne dich hätte ich diese Reise nicht geschafft. Ich bin dankbar, dich an meiner Seite zu haben.

Zusammenfassung

Formatives Assessment hat sich in vielen Studien als eine lernförderliche Maßnahme herausgestellt. Im Gegensatz zum summativen Assessment werden beim formativen Assessment diagnostische Informationen während des Lernprozesses erhoben. Auf der Grundlage dieser Informationen soll der weitere Lehr-Lern-Prozess adaptiert werden. Im besten Fall wird das Lernen auf einer individualisierten Ebene optimiert. Dazu sind eine Vielzahl von Daten nötig, die zudem aufbereitet und ausgewertet werden müssen. Im regulären Physikunterricht mit großen Klassen ist dies oft zu zeitintensiv, sodass formative Assessments nur eingeschränkt möglich sind. Dies zeigt sich beispielsweise beim Einsatz von Concept Maps. Durch Concept Maps können Zusammenhänge zwischen verschiedenen Begriffen der Physik dargestellt werden, wodurch wertvolle Informationen u. a. über Wissenslücken und Missverständnisse gewonnen werden. Der Auswertungsprozess kann jedoch sehr zeitaufwendig sein, insbesondere wenn die Concept Maps umfangreich sind und eine große Anzahl von Begriffen enthalten.
In der vorliegenden Arbeit erfolgt daher eine automatische Auswertung von Concept Maps durch Techniken des maschinellen Lernens, um Lehrkräfte und Lernenden im Lehr-Lernprozess zu unterstützen. In Studie 1 wurde eine Concept Map zum Thema Mechanik entwickelt und im Physikunterricht eingesetzt. Sie wurden durch zwei menschliche Rater bewertet und mehrere Machine-Learning-Modelle wurden mit den Daten trainiert und getestet. Die Ergebnisse zeigen eine sehr gute Übereinstimmung zwischen der menschlichen Bewertung und den verschiedenen Machine-Learning-Modellen. Basierend auf diesen Ergebnissen und dem besten geeigneten Machine-Learning-Modell wurde in Studie 2 ein Feedback-Tool entwickelt. Ziel war es, Einblicke in die Integration der automatischen Auswertung in den alltäglichen Physikunterricht zu gewinnen. Zu diesem Zweck wurde das Feedback-Tool an zwei Zeitpunkten in verschiedenen Klassen eingesetzt, um Lehrkräften und Lernenden automatische Rückmeldungen bereitzustellen. Die Auswertung von Fragebögen und Interviews ergab, dass die Vorteile der automatischen Auswertung von den Lehrkräften erkannt wird, sie jedoch das vollständige Potenzial nicht ausschöpfen. Die Ergebnisse zeigten, dass Lehrkräfte vor allem

Rückmeldungen auf Klassenebene und nicht auf individueller Lernenden-Ebene nutzten.
Insgesamt konnte in der vorliegenden Arbeit aufgezeigt werden, dass Systeme, die auf maschinellem Lernen basieren, im Bildungsbereich eine Vielzahl von Chancen bieten. Die Kombination von menschlicher Lehrkraft und computergestützter Hilfe ermöglicht eine individuelle Unterstützung von Lernenden, die in der Regel sonst nicht möglich wäre. Allerdings konnten auch Anregungen für künftige Forschungen erarbeitet werden, die eine stärkere Zusammenarbeit zwischen Informatik, Fachdidaktik und Lehrkräften erfordern.

Abstract

Formative assessment has been shown in many studies to be a method that promotes learning. In contrast to summative assessment, formative assessment is designed to collect diagnostic information during the learning process. Based on this information, the further teaching-learning process should be adapted. In the best case, learning is optimized on an individual level. This requires a large amount of data to be processed and evaluated. In regular physics classes with large numbers of students, this is often too time-consuming, so that formative assessments are only possible to a limited extent. An example of this is the use of concept maps. Concept maps can be used to illustrate relationships between different concepts in physics, providing valuable information about gaps in knowledge and misunderstandings that can be used as part of formative assessment. However, the assessment process can be time-consuming, especially when concept maps are complex and contain numerous concepts.

In this thesis, concept maps are automatically evaluated using machine learning techniques to support teachers and learners in the teaching-learning process. In Study 1, a concept map on the topic of mechanics was developed and used in physics classes. It was evaluated by two human raters and several machine learning models were trained and tested on the data. The results show a very good agreement between the human evaluation and the machine learning models. Based on these results and the best machine learning model, a feedback tool was developed in Study 2. The goal was to gain insight into the integration of automatic evaluation into everyday physics teaching. To this end, the feedback tool was used twice in different classes to provide automatic feedback to teachers and learners. The evaluation of questionnaires and interviews showed that the advantages of automatic evaluation are recognized by teachers, but that they have not exploited its full potential. The results indicated that teachers mainly used feedback at the class level rather than at the individual learner level.

Overall, this thesis suggests that machine learning-based systems offer a wide range of opportunities in education. The combination of human teachers and computer-based assistance allows for individualized support of learners that would otherwise not be possible. However, suggestions for further research could also

be developed, which require a stronger cooperation between computer science, science education and teachers.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

> „AI [Artificial Intelligence] has the potential to revolutionize the education sector by enhancing learning experiences, supporting teachers and offering more personalized learning opportunities for students. We must equip teachers with the knowledge and strategies they will need to use this new technology to improve and streamline everyday processes as well as classroom implementation“ (Bojorquez & Vega, 2023).

Obiges Zitat hebt die Chancen von künstlicher Intelligenz (KI) für den Bildungsbereich hervor: KI hat das Potenzial, den Bildungsbereich zu revolutionieren, indem Lernerfahrungen verbessert, Lehrkräfte unterstützt und individualisierte Lernmöglichkeiten für Lernende geschaffen werden können. Ebenso wird darauf hingewiesen, dass es wichtig ist, Lehrkräfte mit dem nötigen Wissen und den Strategien auszustatten, um diese neue Technologie effektiv einzusetzen. Dies umfasst sowohl die Optimierung der Lehr-Lern-Prozesse als auch die Integration von KI in den alltäglichen Unterricht.

Trotz des Potenzials muss der Einsatz von KI im Unterricht auch immer kritisch reflektiert werden. So thematisieren Studien, die KI-Anwendungen im Bildungsbereich untersucht haben, Probleme wie eine unreflektierte Akzeptanz von KI-basierten Ergebnissen oder die Benachteiligung bestimmter Lernenden-Gruppen durch KI-basierte Systeme (z. B. Krupp et al., 2023; Yao et al., 2020). Zusätzlich sind KI-basierte Systeme oftmals so komplex, dass die Kontrollierbarkeit und Erklärbarkeit der Ergebnisse eine große Herausforderung darstellen (Steinert et al., 2023).

Für den Einsatz im Unterricht muss daher sichergestellt werden, dass KI-basierte Systeme Informationen von Lernenden und Lehrenden zuverlässig, fachlich korrekt und ethisch unbedenklich auswerten können. Dies erfordert eine domänenübergreifende Zusammenarbeit von Expert:innen aus Fachdidaktik und Informatik, um die Chancen und Risiken von KI für die Optimierung des Lehr-Lern-Prozesses zu erforschen.

Im Bildungsbereich gibt es eine Vielzahl von Anwendungen für KI wie eine automatische Rechtschreibkontrolle oder die automatische Übersetzung von Texten. Der Vorteil von KI-basierten Systemen ist, dass sie große Datenmengen in sehr kurzer Zeit analysieren. Dadurch können Einblicke in den Lernprozess der Lernenden gewonnen und zum Beispiel im Rahmen eines formativen Assessments genutzt werden.
Formatives Assessment gilt seit langer Zeit als eine pädagogische Maßnahme mit einem hohen lernförderlichen Effekt (z. B. Hattie, 2009). Dabei besteht formatives Assessment aus einer Abfolge von drei Handlungsschritten (Souvignier & Hasselhorn, 2018):

1. Erfassung des individuellen Lernstandes
2. Rückmeldung für Lehrkräfte und Lernende
3. Optimierung des weiteren Lehr-Lern-Prozesses, aufgrund der gesammelten diagnostischen Informationen

Aus dieser Abfolge wird deutlich, dass die diagnostischen Informationen, die im ersten Schritt erhoben werden, die Grundlagen für die weiteren Handlungen im formativen Assessment schaffen (Schütze et al., 2018). Im Schulalltag werden diese Informationen oft durch Fragen im Unterrichtsgespräch erhoben, wobei eine Vielzahl von unterschiedlichen Methoden dazu existiert (Wiliam, 2010).
Eine dieser Methoden stellen Concept Maps dar. Concept Maps sind grafische Hilfsmittel zur Darstellung und Organisation von Inhalten und Wissen (Novak & Cañas, 2008; Ryssel, 2018). Lernende müssen bei der Erstellung einer Concept Map die Zusammenhänge zwischen verschiedenen Begriffen organisieren und visualisieren und so ihr Wissen über die gelernten Inhalte reflektieren. Daher können wichtige Informationen über Wissenslücken und Missverständnisse gesammelt werden, die im Rahmen eines formativen Assessments zur Optimierung des Lehr-Lern-Prozesses genutzt werden können (Novak & Cañas, 2006).
Aber trotz der lernförderlichen Effekte werden formative Assessments im Unterricht kaum genutzt (Black & Wiliam, 1998). Die Ursachen hierfür sind vielfältig. In der Praxis werden häufig der zu hohe Zeitaufwand für die menschliche Auswertung der gesammelten Informationen sowie die große Anzahl an Lernenden in einer Klasse, die eine individuelle Rückmeldung erschwert, als Gründe genannt (Bennett, 2011; Black & Wiliam, 1998; Hunt & Pellegrino, 2002).

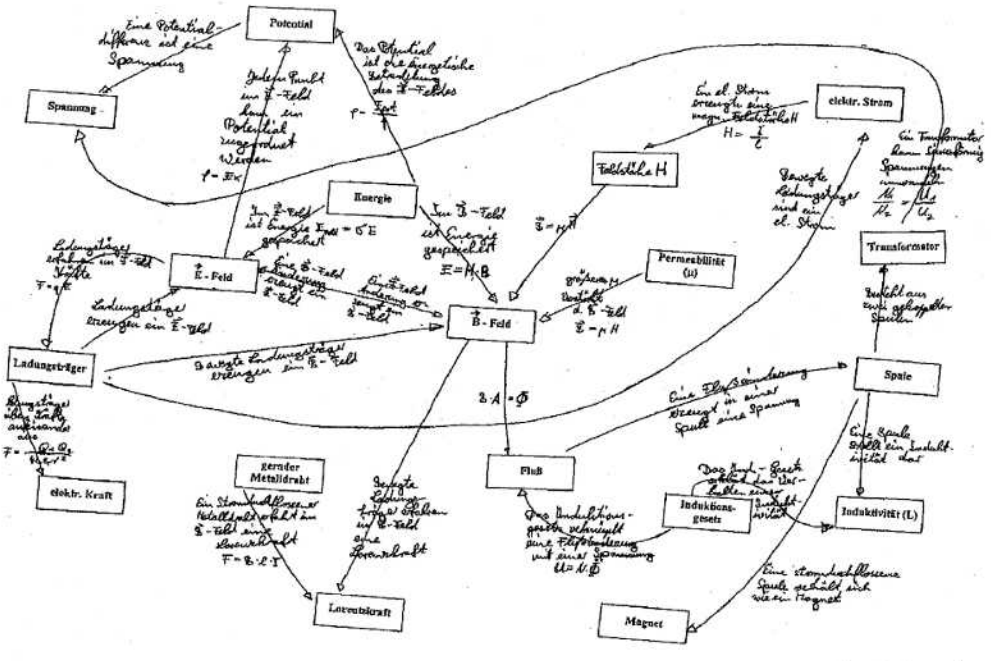


Abbildung 1.1: Concept Map zum Thema Elektrizitätslehre (Friege, 2001)

Diese Problematik wird auch bei der Betrachtung der Concept Map in Abbildung 1.1 deutlich. Zur Auswertung der dargestellten Concept Map kann z. B. die Anzahl der verwendeten Begriffe genutzt werden. Für eine detaillierte Auswertung können jedoch auch die Zusammenhänge zwischen den einzelnen Begriffen betrachtet und analysiert werden. Dies führt zu einem detaillierten Einblick in den aktuellen Wissensstand der Lernenden, was eine fundierte Optimierung des Lehr-Lern-Prozesses ermöglicht (Trumpower & Sarwar, 2010).

Allerdings kostet dieser Auswertungsansatz auch viel Zeit, da jeder Zusammenhang zwischen den Begriffen betrachtet werden muss und die Concept Maps innerhalb der Klasse bezüglich Anzahl und Qualität schwanken können (Hartmeyer et al., 2018). Die Verwendung von Concept Maps für ein qualitativ hochwertiges formatives Assessment ist daher im normalen Schulalltag nur schwer umsetzbar, da ein gesamter Klassensatz von Concept Maps kaum zeiteffizient ausgewertet werden kann (Ley, 2015).

In der vorliegenden Arbeit wird daher ein KI-basiertes System entwickelt, welches den zweiten Schritt im formativen Assessment *Rückmeldung für Lehrkräfte und Lernende* durch eine automatische Auswertung von Concept Maps übernehmen wird.

Das Ziel ist es, Erkenntnisse über eine KI-basierte Auswertung von Concept

Maps, die für ein formatives Assessment eingesetzt werden, zu erheben. Durch die automatische Auswertung kann der hohe Zeitaufwand, der sonst bei einer menschlichen Auswertung der Concept Maps entsteht, reduziert werden. Zudem kann die automatische Auswertung auch für eine große Anzahl an Lernenden ohne zusätzlichen Aufwand durchgeführt werden, was eine deutliche Unterstützung im Lehr-Lern-Prozess darstellen kann. Es wird überprüft, wie hoch die Übereinstimmung zwischen menschlicher und automatischer Auswertung ist, um Aussagen über die Leistungsfähigkeit der KI-basierten Auswertung treffen zu können. Zudem wird analysiert, wie zuverlässig die automatische Auswertung ist und ob die Auswertung unabhängig von Eigenschaften der Lernenden ist. So kann überprüft werden, ob eine gewisse Kontrollierbarkeit und Erklärbarkeit vorliegt, was für den Einsatz in Schulen wichtige Aspekte sind.
Die automatische Auswertung der Concept Map wird aber nicht nur auf einer theoretischen Ebene betrachtet. Ein weiteres Ziel der Arbeit ist die Untersuchung der KI-basierten Concept-Map-Auswertung im regulären Physikunterricht. Damit Aussagen zur Integration der automatischen Auswertung im alltäglichen Physikunterricht gemacht werden können, wird daher eine Studie an verschiedenen Gymnasien durchgeführt. In dieser Studie wird die automatische Auswertung an zwei unterschiedlichen Zeitpunkten des Physikunterrichts den Lehrkräften und Lernenden ein Feedback zur Verfügung stellen, welches auf der automatischen Auswertung von Concept Maps basiert. Anschließend werden über Interviews und Fragebögen Informationen zum Umgang mit der KI-basierten Auswertung erhoben und analysiert. Damit vielfältige Einblicke gesammelt werden können, wird der dritte Schritt des formativen Assessments *Optimierung des weiteren Lehr-Lern-Prozesses, aufgrund der gesammelten diagnostischen Informationen* nicht festgelegt.
Durch die vorliegende Arbeit können vielfältige Ergebnisse erzielt werden. Es werden nicht nur neue Erkenntnisse bezüglich der Leistung KI-basierter Auswertungen, die für ein formatives Assessment genutzt werden, erzielt, sondern wichtige Einblicke in die Integration von KI in den alltäglichen Physikunterricht ermittelt. Dies ermöglicht, neue Perspektiven einzunehmen und Strategien für einen effektiven Umgang mit KI-basierten Technologien zu ermitteln. Es können so wichtige Grundbausteine für einen zukünftigen Physikunterricht gelegt werden, denn KI wird in den nächsten Jahren das Lehren und Lernen prägen, oder wie es Seldon (2018) formulierte: „*AI the biggest thing to happen in education for 500 years*“.

Aufbau der vorliegenden Arbeit

Im Folgenden wird ein Überblick über die Inhalte der einzelnen Kapitel gegeben, welcher der Orientierung dienen soll:

In *Kapitel 2* wird zunächst das Thema formatives Assessment behandelt. Dabei wird versucht, den Begriff des formativen vom summativen Assessment abzugrenzen. Es wird außerdem auf die verschiedenen Schlüsselstrategien von Wiliam und Thompson (2008) eingegangen, wobei ein besonderes Augenmerk auf die lernförderliche Rückmeldung gelegt wird. Anschließend werden Befunde der Lehr-Lern-Forschung betrachtet und eingeordnet.

In *Kapitel 3* werden Concept Maps eingeführt, da sie, wie oben beschrieben, zur Erhebung des Lernstandes eingesetzt werden. Das Kapitel wird verschiedenen Concept-Map-Formate und Auswertungsmöglichkeiten diskutieren und einordnen.

In *Kapitel 4* wird das maschinelle Lernen (Machine Learning) beleuchtet, was ein Teilgebiet der künstlichen Intelligenz ist. In diesem Kapitel wird der Fokus vor allem auf das überwachte maschinelle Lernen gelegt, da die automatische Auswertung der Concept Maps durch ein Modell aus diesem Bereich des maschinellen Lernens erzeugt wird. Um das entwickelte Machine-Learning-Modell analysieren und vergleichen zu können, werden in Kapitel 4 aktuelle Forschungen aus dem Bildungsbereich untersucht.
In *Kapitel 5* werden die genaue Zielsetzung und das Erkenntnisinteresse der vorliegenden Arbeit beschrieben. Um die oben genannten Ziele der Arbeit zu erreichen, wird der empirische Teil in zwei Teilstudien unterteilt.

In *Kapitel 6* wird der Fokus auf der Entwicklung der Concept Map und die automatische Auswertung durch das Machine-Learning-Modell liegen. Zunächst wird daher die Konzipierung der Concept Map, die zur Erhebung des Lernstandes eingesetzt wird, beschrieben. Es wird außerdem die Entwicklung des Bewertungsschemas, welches die Grundlage für die lernförderliche Rückmeldung ist, dargelegt. Anschließend wird die erste Erhebungsphase beschrieben, bei der die entwickelte Concept Map in verschiedenen Klassen im Physikunterricht eingesetzt und von menschlichen Bewertern ausgewertet wird. Danach wird auf die Entwicklung des Machine-Learning-Modells eingegangen. Nach der Entwicklung werden

die Ergebnisse bezüglich der Übereinstimmung zwischen der menschlichen und automatischen Auswertung dargestellt und für den Anwendungsfall im regulären Physikunterricht kritisch diskutiert.

In *Kapitel 7* liegt der Fokus auf dem Einsatz des entwickelten Machine-Learning-Modells als Feedback-Tool. Daher wird zunächst auf der Grundlage der automatischen Auswertung ein lernförderliches Feedback konzipiert, welches in einer zweiten Erhebung in verschiedenen Klassen im Physikunterricht eingesetzt wird. Daher wird anschließend die Stichprobe und die Methodik dieser Erhebung erläutert. Abschließend werden auch in Kapitel 7 die Ergebnisse aus den Interviews und Fragebögen dargestellt und diskutiert.

In *Kapitel 8* werden die Ergebnisse aus den beiden Teilstudien zusammenfassend diskutiert.

In *Kapitel 9* endet die Arbeit mit einem kurzen Fazit und einem Ausblick.

2 Formatives Assessment

Wenn man im schulischen Bereich von Assessment spricht, meint man in der Regel die Erfassung des Lernstandes der Lernenden. Ein Assessment kann dabei verschiedene Zwecke und Ziele verfolgen, wie die abschließende Leistungsbeurteilung am Ende oder die Erhebung des Vorwissens der Lernenden am Beginn einer Unterrichtseinheit (Schütze et al., 2018). Eine grundlegende Unterscheidung wird oftmals zwischen dem summativen und formativen Assessment gemacht (Maier, 2010). Da aber vor allem der Begriff des formativen Assessments nicht klar definiert ist und oftmals unscharf benutzt wird (Bennett, 2011; Schmidt, 2020), soll in diesem Kapitel zunächst eine Begriffsbestimmung erfolgen. Anschließend werden die fünf Schlüsselmerkmale des formativen Assessments von Wiliam und Thompson (2008) vorgestellt. Da Feedback ein sehr zentrales Element des formativen Assessments ist, wird es in einem eigenen Abschnitt diskutiert. Abschließend werden in diesem Kapitel Ergebnisse zur Lernwirksamkeit und zu Herausforderungen bei der Implementierung von formativen Assessments vorgestellt.

2.1 Begriffsbestimmung

Die Unterscheidung zwischen dem summativen und formativen Assessment wurde erstmals von Scriven (1967) vorgenommen (vgl. auch Wiliam & Thompson, 2008). Bei dieser Differenzierung geht es vor allem um den Zweck der Evaluation und weniger um die Methode, Aufgabenstellung oder den Erhebungsrhythmus (Maier, 2010).
Das summative Assessment versteht sich als traditionelle Form des Leistungstests und wird als abschließende Leistungsbewertung genutzt (Scriven, 1967; N. Wolf, 2014). Es wird häufig am Ende einer Unterrichtseinheit eingesetzt und dient somit zur Überprüfung der zu erreichenden Lernziele. Auf Grundlage dieser Informationen können dann Noten vergeben oder Selektionsentscheidungen getroffen werden (Black & Wiliam, 2009; Buholzer et al., 2020). Eine typische Methode sind daher Klassenarbeiten (Pellegrino et al., 2001; N. Wolf, 2014). Aus den daraus resultie-

renden Noten und Entscheidungen werden in der Regel keine weiteren Schritte über den zukünftigen Lernprozess der Lernenden gezogen (N. Wolf, 2014). Es hat trotz der weitreichenden Konsequenzen für die Lernenden nur eine indirekte Auswirkung auf das eigentliche Lernen (Sadler, 1989). Daher wird das summative Assessment auch oft als *assessment of learning* bezeichnet (Maier, 2010).
Im Gegensatz dazu steht das formative Assessment, dessen Ziel es ist, mithilfe erhobener diagnostischer Informationen den weiteren Lehr-Lernprozess zu verbessern (Schütze et al., 2018). Nach Black und Wiliam (2009, S.7) lässt sich formatives Assessment wie folgt beschreiben: „*Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited*". Auch der Zeitpunkt des Einsatzes unterscheidet sich zum summativen Assessment, da formatives Assessment meistens unterrichtsbegleitend durchgeführt wird (Schütze et al., 2018). Beim formativen Assessment geht es daher nicht um Leistungsmessungen, die ein einfaches oder dichotomes Feedback bereitstellen, da es um den Zweck der Rückmeldung geht (Maier, 2010). Denn ein Assessment ist dann formativ, wenn die Informationen genutzt werden, um gezielte Fördermaßnahmen für Lernende zu entwickeln oder die nachfolgenden Unterrichtsstunden zu verbessern (N. Wolf, 2014). Aus diesem Grund können auch sogenannte *monitoring assessments* oder *diagnostic assessments*, die lediglich Informationen liefern, was z. B. richtig oder falsch beantwortet wurde, nicht mit einem formativen Assessment gleichgesetzt werden (Wiliam & Thompson, 2008). Allerdings können diese beiden Assessment-Methoden oder auch Klassenarbeiten trotzdem einen formativen Charakter haben, solange die diagnostischen Informationen zur Verbesserung des Lehr-Lernprozesses genutzt werden (Black & Wiliam, 2009). Formatives Assessment wird deshalb auch *assessemt for learning* genannt (Black, 2015).
Zusammenfassend kann festgehalten werden, dass summatives und formatives Assessment unterschiedliche Ziele verfolgen. Sie müssen sich daher nicht gegenseitig ausschließen. Vielmehr bilden sie aufgrund der unterschiedlich gewonnenen Informationen ein ausgeglichenes Assessmentsystem (Schmidt, 2020). Um den komplexen Prozess des formativen Assessments noch näher zu beschreiben, wird im nachfolgenden Abschnitt auf die fünf Schlüsselmerkmale des formativen Assessments von Wiliam und Thompson (2008) eingegangen.

2.2 Die fünf Schlüsselmerkmale

Formatives Assessment kann in unterschiedlichsten Varianten im Unterricht eingesetzt werden. Um dafür einen theoretischen Rahmen zu schaffen, haben Wiliam und Thompson (2008) fünf Schlüsselmerkmale herausgearbeitet (Black & Wiliam, 2009). Die fünf Schlüsselmerkmale des formativen Assessments basieren auf den drei Leitfragen beim Lernen und Lehren von Ramaprasad (1983) (siehe auch Schmidt, 2020):

(i) Wo stehen die Lernenden? (Lernstand)
(ii) Wohin sollen die Lernenden gehen? (Lernziel)
(iii) Welche Schritte müssen getan werden, um das Lernziel zu erreichen?

Die Beantwortung der drei Fragen gehört zu den klassischen Aufgaben von Lehrkräften. Wiliam und Thompson (2008) sehen dies aber auch als Aufgabenbereich der Lernenden und der Mitschülerinnen und Mitschüler (Peers) an. Durch die Kombination der drei Akteure mit den drei Leitfragen wurde von Wiliam und Thompson (2008) das Modell der fünf Schlüsselstrategien des formativen Assessments entwickelt, das in Abbildung 2.1 dargestellt ist.

Die fünf Schlüsselstrategien werden im Folgenden näher beschrieben. Da Feedback ein zentraler Bestandteil des formativen Assessments und auch für diese Arbeit relevant ist, wird die dritte Strategie, die lernförderliche Rückmeldung, in einem eigenen Abschnitt adressiert (siehe Abschnitt 2.3).

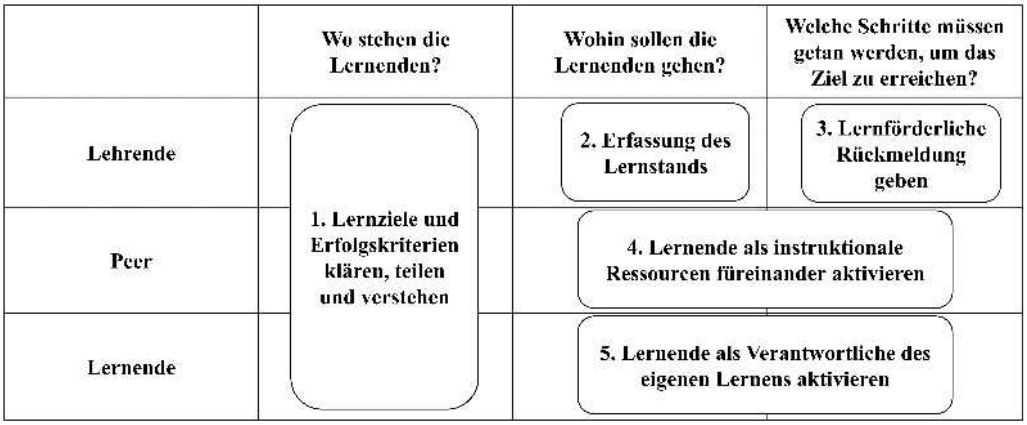


Abbildung 2.1: Die fünf Schlüsselmerkmale des formativen Assessments (Wiliam & Thompson, 2008) (Übersetzung durch Autor)

Klärung der Lernziele und Erfolgskriterien

Der erste Aspekt der fünf Merkmale des formativen Assessments bezieht sich auf die Lernziele und die damit verbundenen Erfolgskriterien. Diese dienen als Start in eine neue Lerneinheit und als Ausgangspunkt für ein formatives Assessment (Schmidt, 2020). Da in der Regel ein festgelegter Lehrplan oder verpflichtendes Kerncurriculum vorliegt, ist es meistens die Aufgabe von Lehrkräften, die Lernziele auszuarbeiten, wohingegen nach Wiliam (2010) dies nicht explizit festgelegt ist (vgl. Abbildung 2.1). So kann z. B. bei einem Projektunterricht eine gemeinsame Vorstellung entwickelt werden, welche Lernziele erreicht werden sollen und wie überprüft werden kann, ob die gesetzten Ziele erreicht wurden (Buholzer et al., 2020).

Unabhängig davon muss allerdings darauf geachtet werden, dass die Lernziele verständlich formuliert sind, sodass sie für alle Lernenden nachvollziehbar und transparent sind (Wiliam, 2010). Auf diese Weise kann ein Rahmen geschaffen werden, der für alle Beteiligten identisch ist und als Grundlage für eine formative Beurteilung dienen kann (Schmidt, 2020). Da Erfolgskriterien nicht für jeden Lerner gleich sind, sollten diese individuell und spezifisch entwickelt werden, damit jeder Lernende sie bestmöglich nutzen kann (Schütze et al., 2018). Eine hohe Qualität dieser Strategie zeichnet sich dadurch aus, dass die Kriterien kognitiv aktivierend sind und mit anwendungsbezogenen Beispielen verknüpft sind. Zudem sollen die Lernziele auf dem Vorwissen der Lernenden aufbauen (Buholzer et al., 2020).

Ein wesentlicher Aspekt der fünf Merkmale ist, dass die Art der Lernziele unabhängig von der formativen Beurteilung verstanden werden muss. Das bedeutet, dass z. B. nicht immer eine messbare Leistungssteigerung im Fokus stehen muss, sondern ebenso soziale Ziele betrachtet werden können, solange diese klar kommuniziert werden (Schmidt, 2020; Wiliam, 2010). Die Entwicklung von Einstellungen und sozialen Zielen lassen sich auch nicht mittels eines Leistungstests erfassen. „*In other words, a commitment to formative assessment does not entail any particular view of what the learning intentions should be, nor does it entail a commitment to any particular view of what happens when learning takes place*“ (Wiliam, 2010, S. 33).

Erfassung des Lernstandes

Die zweite Schlüsselstrategie *Lernstand durch Diskussionen, Fragen und Aufgaben erfassen* bezieht sich auf unterschiedliche Methodiken, welche während eines formativen Assessments verwendet werden können. Die dabei gewonnenen diagnostischen Informationen schaffen die Grundlagen für alle weiteren Handlungen im formativen Assessmentprozess (Schütze et al., 2018). Nach Wiliam (2010) werden diese Informationen oft durch Fragen erhoben, wobei aber auch betont wird, dass eine Vielzahl von Maßnahmen dafür genutzt werden kann. Typische Methoden sind Lerntagebücher, Portfolios oder sogenannte One-Minute-Paper, bei denen die Lernenden innerhalb einer Minute Kurzfragen beantworten sollen. Außerdem können computergestützte Systeme wie Student-Response-Systems oder intelligente tutorielle Systeme eingesetzt werden (siehe Schütze et al., 2018). Die unterschiedlichen Methoden lassen sich auch hinsichtlich des Planungsumfangs klassifizieren. Shavelson et al. (2008, S. 300) unterscheiden dabei zwischen *on-the-fly*, *planned-for-interaction* und *formal and embedded in curriculum* (siehe Abbildung 2.2).

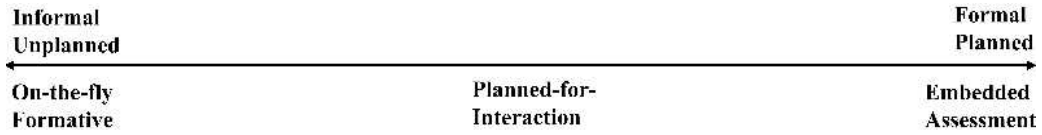


Abbildung 2.2: Kontinuum der formativen Assessment-Praktiken (Shavelson et al., 2008)

Bei *on-the-fly*-Assessments gewinnen die Lehrkräfte eher spontan, interaktiv und manchmal unerwartet Informationen, z. B. im Rahmen von Gruppendiskussionen oder im Unterrichtsgespräch (Schütze et al., 2018; Shavelson et al., 2008). Auf der gegenüberliegenden Seite dieser Einteilung (siehe Abbildung 2.2) liegen die *formal-and-embedded-in-curriculum*-Assessments, die eher stark formalisiert sind und im Vorfeld im Lehrplan platziert wurden, um zielgerichtete Assessmentmomente zu entwickeln (Shavelson et al., 2008).
Die unterschiedlichen formativen Assessment-Methoden können in einer Vielzahl von verschiedenen fachspezifischen Bereichen angewendet werden. Dabei ist jedoch zu beachten, dass nicht alle ermittelten Informationen und eingesetzten Verfahren gleichermaßen brauchbar sind (Wiliam, 2010). So lassen sich insbesondere motorische Kompetenzen eher durch Beobachtungen statt durch papierbasierte Methoden analysieren (Schütze et al., 2018). Nach Buholzer et al. (2020) weisen Realisierungen dieser Schlüsselstrategie dann eine hohe Qualität auf, wenn den

Lernenden nicht nur ausreichend Zeit zur Verfügung gestellt wird, sondern sie auch derart gefordert werden, dass sie eigene Ideen und Lösungen entwickeln müssen. Außerdem sollen die diagnostischen Informationen hinsichtlich der (gemeinsam) aufgestellten Lernziele und Erfolgskriterien betrachtet werden (Buholzer et al., 2020).
Dabei können formative Assessments nicht nur auf Individualebene erfolgen, sondern auch auf einer Klassenebene durchgeführt werden (Schütze et al., 2018). Es können demnach Informationen über den Leistungsstand der Klasse erhoben und z. B. für die Planung einer nachfolgenden Stunde genutzt werden (Black & Wiliam, 2009). Im Gegensatz dazu wird bei der Individualebene der Lernstand einzelner Lernender erfasst und formativ genutzt. Die Differenzierung zwischen den beiden Ebenen bedeutet aber nicht, dass ein formatives Assessment nicht beide Ebenen einbeziehen kann. Es ist möglich, mit einem formativen Assessment diagnostische Informationen auf beiden Ebenen zu gewinnen. Diese können dann für individuelle Rückmeldungen sowie die Optimierung nachfolgender Stunden genutzt werden (Schütze et al., 2018).

Lernförderliche Rückmeldung

Lernförderliche Rückmeldung oder auch Feedback[1] ist wohl das prominenteste Element des formativen Assessments und wird in vielen Forschungsarbeiten als die zentrale Schlüsselstrategie angesehen (u. a. Black & Wiliam, 2009; Bürgermeister et al., 2014; Heritage, 2007). Hattie und Timperley (2007) beschreiben Feedback als „*information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding*“ (Hattie & Timperley, 2007, S. 81). Eine andere Definition findet man bei Ramaprasad (1983), bei der der formative Charakter klarer wird: „*Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way*“ (Ramaprasad, 1983, S. 4). Daraus lässt sich ableiten, dass Feedback die Lücke zwischen dem bestehenden und dem angezielten Lernstand schließen soll (Schmidt, 2020). Das kann durch die Lehrkräfte erfolgen, in dem sie die erhobenen Informationen nutzen, um die nachfolgenden Unterrichtsstunden an den Lernstand der Lernenden anpassen, z. B. durch unterschiedliche Aufgabenblätter oder auf das Vorwissen abgestimmte

[1] In vielen Arbeiten werden Rückmeldung und Feedback als Synonyme verwendet, so auch in dieser Arbeit. Eine Diskussion über die Verwendungen der beiden Begriffe findet sich z. B. in Müller und Ditton (2014).

Unterrichtseinstiege. Eine individuelle Rückmeldung kann auch direkt für die Lernenden erfolgen, sodass diese ihren eigenen Lernprozess selbstständig optimieren können (Schütze et al., 2018). Daraus ergibt sich, dass eine formative Beurteilung eher prospektiv und nicht retrospektiv sein sollte (Wiliam, 2010). Dies bedeutet nicht, dass die Rückmeldung im Rahmen des formativen Assessments nicht die gesetzten Lernziele (feed up) oder den aktuellen Lernstand der Lernenden (feed back) beschreiben soll, sondern dass auf Basis der diagnostischen Informationen die nächsten Schritte im Lehr-Lernprozess festgelegt werden sollen (feed forward) (Buholzer et al., 2020; Hattie & Timperley, 2007).
Wie genau die diagnostischen Informationen ausgewertet und in einer lernförderlichen Rückmeldung dargestellt werden, ist nicht festgelegt. Es existiert eine Vielzahl von Feedbackstrategien, die förderlich für das Lernen sein können (Hattie & Timperley, 2007). Allerdings wirken viele verschiedene Aspekte wie die Feedbackquelle, Präsentationsform oder Feedbackebene auf die Qualität der lernförderlichen Rückmeldung ein (Schmidt, 2020). Um einen theoretischen Rahmen für das automatisch generierte Feedback, das in der zweiten Studie dieser Arbeit eingesetzt werden soll, zu schaffen, wird im Abschnitt 2.3 genauer auf die einzelnen Facetten von Feedback eingegangen.

Lernende als instruktionale Ressourcen füreinander aktivieren

Bei der vierten Schlüsselstrategie steht der Begriff des Füreinander im Vordergrund, da hier die Lernenden einander bewerten und rückmelden sollen. Da sowohl in der Literatur (Schmidt, 2020) als auch in dieser Arbeit diese Strategie des formativen Assessments eher weniger Bedeutung hat, soll das Peer-Assessment nur kurz beschrieben werden.
Betrachtet man Abbildung 2.1, lässt sich feststellen, dass Lernstandserhebung und Feedback kombiniert werden (Wiliam, 2010). Bei einem Peer-Assessment nehmen die Lernenden die Rolle der Lehrkraft ein und müssen so einerseits die Lernziele und Erfolgskriterien verstanden haben und andererseits in der Lage sein, die Lernstände zu erheben und darauf aufbauend Rückmeldung geben zu können. Sie werden als instruktionale Ressourcen füreinander aktiviert (Schmidt, 2020; Wiliam, 2010). Im Optimalfall können so Erkenntnisse über die eigenen Lernprozesse gewonnen werden, welche anschließend für den weiteren Lehr-Lernprozess nutzbar sind (Buholzer et al., 2020).

Lernende als Verantwortliche des eigenen Lernens aktivieren

Die fünfte und letzte Schlüsselstrategie nach Wiliam und Thompson (2008) stellt die Lernenden selbst und das eigenständige Lernen in den Vordergrund. Es zielt darauf ab, den Lernenden zu helfen, ihr Verständnis und Wissen eigenständig zu beschreiben und mit den im Vorfeld aufgestellten Leistungskriterien zu vergleichen. Resultierend daraus sollen die Lernenden selbstständig feststellen, wo potenzielle Wissenslücken sind und wie sie im Lernprozess fortschreiten wollen (Buholzer et al., 2020). Die Strategie, Lernende als Eigentümer ihres eigenen Lernens zu aktivieren, basiert unter anderem auf den theoretischen Annahmen der Motivation (Ryan & Deci, 2000), des Interesses (Hidi & Harackiewicz, 2000) und vor allem auf denen des selbstregulierten Lernens (Black & Wiliam, 2009).
Das Konzept des selbstregulierten Lernens wurde in einer Vielzahl von Forschungsarbeiten untersucht und ist auch für das formative Assessment ein relevanter Aspekt (Wiliam, 2010). Da selbstreguliertes Lernen eine komplexe Form des Lernens sein kann, die nicht nur innerhalb des Klassenzimmers stattfindet, muss die Lehrkraft sicherstellen, dass die Lernenden unterschiedliche effektive Lernstrategien und Self-Assessments kennen und nutzen können. Deshalb sollte die Lehrkraft sowohl didaktisch als auch methodisch Unterstützung zur Verfügung stehen (Andrade, 2010; Schmidt, 2020). Aber auch andere Einflussfaktoren wie die persönlichen Eigenschaften der Lernenden oder die sozialen Umstände innerhalb der Klasse können das selbstregulierte Lernen beeinflussen (Andrade, 2010).
Angesichts dessen hat nach Buholzer et al. (2020) das Self-Assessment dann eine hohe Qualität, wenn es von der Lehrkraft anregend gestaltet wurde und zu einer tieferen Selbstreflexion anregt (Wylie & Lyon, 2016). Damit es den formativen Charakter nicht verliert, ist zudem darauf zu achten, dass die Erkenntnisse primär für das weitere Lernen und eine mögliche Optimierung der nächsten Lernschritte von den Lernenden genutzt werden (Andrade, 2010).

2.3 Feedback

Wie im vorhergehenden Abschnitt beschrieben, ist Feedback eine der Schlüsselstrategien in einem formativen Assessment. Feedback kann in unterschiedlichen Formen auftreten. Im Unterricht erhalten Lernende oft Rückmeldung bezüglich ihrer Leistung. Dies kann durch Lehrkräfte, andere Lernende oder aber durch Self-Assessments erfolgen. Es ist praktisch nicht möglich, kein Feedback zu geben, da

	Bezeichnung	Beispiel
Einfaches Feedback	Knowledge of result (KR)	Richtig/falsch
	Knowledge of performance (KP)	80 % der Aufgabe korrekt gelöst
	Knowledge of correct result (KCR)	Angabe der richtigen Lösung
Elaboriertes Feedback	Knowledge on task constrains (KTC)	Hinweise auf die Art der Aufgabe
	Knowledge about concepts (KC)	Erklärung von Fachbegriffen
	Knowledge about mistakes (KM)	Ort/Anzahl der Fehler
	Knowledge on how to proceed (KH)	Hinweise auf Lösungsstrategien
	Knowledge on meta-cognition (KMC)	Metakognitive Leitfragen

Tabelle 2.1: Einfaches vs. elaboriertes Feedback (Narcis, 2006)

neben direkten verbalen oder nonverbalen Äußerungen auch Schweigen gewisse Informationen überliefern kann (Ditton & Müller, 2014). Die Art und Weise, wie Feedback gegeben wird, wird jedoch häufig nicht reflektiert (N. Wolf, 2014).
Feedback ist ein mehrdimensionales Konstrukt, das auf komplexen Verhaltensketten verschiedener Personen beruht (Ditton & Müller, 2014). Deshalb gibt es abhängig von dem wissenschaftlichen Ansatz und Fachgebiet unterschiedliche Klassifizierungen von Feedback (Müller & Ditton, 2014). Unterscheidungsmerkmale von Feedback können die Ausführlichkeit, der Zeitpunkt, die Informationsqualität, die Komplexität oder die soziale Dimension sein (N. Wolf, 2014, S. 20).
Zudem existieren äußere Einflüsse wie die Schulform oder das Unterrichtsfach, welche die Wirkung von Feedback zusätzlich beeinflussen können (N. Wolf, 2014). So lässt sich erklären, dass trotz diverser Feedbackstudien uneinheitliche Ergebnisse bezüglich der Effektgrößen von Feedback vorliegen (Narciss, 2014; N. Wolf, 2014).

2.3.1 Einfaches und elaboriertes Feedback

Eine Möglichkeit, Feedback einzuordnen, ist auf einer inhaltlichen Ebene. Man kann zwischen einfachem und elaboriertem Feedback unterscheiden (Narciss, 2006).
Die einfachste Variante des Feedbacks bezieht sich nur auf die Information, ob die Lösung der Aufgabe richtig oder falsch ist (knowledge of results). Darauf aufbauend kann auch die Anzahl der korrekt gelösten Antworten rückgemeldet werden (knowledge of performance). Weiteres einfaches Feedback wäre die Präsentation der richtigen Lösung der bearbeiteten Aufgabe (knowledge of correct result).
Eine Differenzierung findet sich ebenfalls beim elaborierten Feedback, allerdings ist diese Unterteilung nicht so eindeutig (Ryssel, 2012). Nach Narciss (2006)

zählen zusätzliche Informationen zu inhaltlichen Aspekten wie Hinweise zu der Art der Aufgabe (knowledge on task constrains) zum elaborierten Feedback. Weitere inhaltliche Hinweise, wie Erklärungen zu Fachbegriffen (knowledge about concepts) oder wo genau der Fehler gemacht wurde (knowledge about mistakes), zählen ebenfalls zum elaborierten Feedback (siehe Tabelle 2.1). Auch Informationen über konkrete Lösungsstrategien (knowledge on how to proceed) oder über die Regulierung des eigenen Lernprozesses (Knowledge on meta-cognition) gehen über das einfache Feedback hinaus (Müller & Ditton, 2014; Ryssel, 2012). Es lässt sich also erkennen, dass das elaborierte Feedback Informationen enthält, die nicht unmittelbar auf die richtige Lösung schließen lassen (Ryssel, 2018).
Die Trennung des Feedbacks auf inhaltlicher Ebene ist in der Forschung ein viel untersuchter Bereich (Hattie & Wollenschläger, 2014; Ryssel, 2012). So hat unter anderem Moreno (2004) die Lernwirksamkeit von zwei verschiedenen Feedbacks mit Studierenden eines Psychologiekurses untersucht. Folgt man der Klassifizierung von Narciss (2006) hat in der Studie von Moreno (2004) Gruppe 1 ein einfaches Feedback erhalten, bei dem die Studierenden erst die richtige Lösung mitgeteilt bekamen (KR) und anschließend wurde die korrekte Lösung präsentiert (KCR) (vgl. Tabelle 2.1). Gruppe 2 erhielt ein elaboriertes Feedback aus mehreren Schritten: Zunächst gab es eine Information, ob die Antwort korrekt war (KR). Darauf aufbauend wurde erklärt, warum die Lösung korrekt oder nicht korrekt war (KM-KC). Abschließend wurde die richtige Lösung präsentiert (KCR), sofern die gegebene Antwort falsch war. Angesichts der Gruppeneinteilung kann man sagen, dass elaboriertes Feedback auch Teile des einfachen Feedbacks enthalten kann. Moreno (2004) konnte letztlich nachweisen, dass das elaborierte Feedback KR-KM-KC-KCR zu einer höheren Lernwirksamkeit und einer Reduzierung der kognitiven Überlastung führt als das eingesetzte einfachere Feedback (KR-KCR). Dieses Ergebnis lässt sich auch in anderen Studien finden. Chase und Houmanfar (2009) setzten in ihrer Studie ebenfalls unterschiedliche Feedbacks ein. Untersuchungsgegenstand war der Einfluss von elaboriertem und einfachem Feedback auf die Lernleistung von Psychologiestudierenden. Beim einfachen Feedback wurden Informationen über den Anteil der richtig gelösten Aufgaben (KP) und die richtige Lösung (KCR) präsentiert. Beim elaborierten Feedback wurden Informationen zu den Fragen bereitgestellt (KM) und konzeptionelle Hinweise gegeben (KC/KTC). Chase und Houmanfar (2009) kommen zu dem Ergebnis, dass das elaborierte Feedback, insbesondere bei schwierigen Aufgaben, zu einer signifikanten Verbesserung der Leistung der Studierenden führt.

McKendree (1990) führte eine Studie mit Lernenden einer nordamerikanischen High-School durch, um zu untersuchen, ob einfaches oder elaboriertes Feedback lernwirksamer ist. Dabei bekam eine Kontrollgruppe ein einfaches Feedback (KR) und drei Versuchsgruppen ein elaboriertes Feedback (KM, KH und KM+KH). McKendree (1990) konnte zeigen, dass alle drei elaborierten Feedback-Gruppen eine signifikante niedrige Fehleranzahl im Nachtest hatten als die Gruppe mit dem einfachen Feedback (Ryssel, 2018).
Die Betrachtung von umfangreichen Meta-Analysen wie von Hattie (2009), Bangert-Drowns et al. (1991) oder Kluger und DeNisi (1996) zeigen jedoch auch andere Ergebnisse. Nach Hattie (2009) hat Feedback zwar einen der größten Einflüsse auf das schulische Lernen, jedoch weisen die Effektstärken eine erhebliche Variabilität auf (Hattie, 2009, S. 174). Auch Bangert-Drowns et al. (1991) konnten eine große Streuung der Effektstärken aufzeigen, wobei sogar ein Drittel der Studien eine negative Effektstärke des Feedbacks aufwies (Ryssel, 2018). Betrachtet man nur die diskutierten Unterschiede zwischen elaborierten und einfachen Feedbacks zeigt die Meta-Analyse von Bangert-Drowns et al. (1991), dass z. B. *knowledge of results* (KR) zu keinem Effekt, *knowlegde of correct result* (KCR) und elaboriertes Feedback zu einem mittleren Effekt führen. Möchte man elaboriertes Feedback einsetzen, empfiehlt sich eine sequenzielle Präsentation, bei der erst in einem zweiten Schritt die richtige Lösung präsentiert wird. So soll eine tiefergehende Verarbeitung des Feedbacks erreicht werden (Narciss & Huth, 2006; Ryssel, 2018). Darüber hinaus sollte zumindest bei komplexen Aufgaben ein effektives Feedback zeitversetzt erfolgen, um eine längere Beschäftigung mit den Inhalten zu ermöglichen (Clariana et al., 2000; Ryssel, 2018). Auch das Vorwissen hat einen Einfluss auf die Effektivität des Feedbacks. So kann elaboriertes Feedback für Lernende mit geringem Vorwissen vorteilhaft sein, während Lernende mit hohem Vorwissen kein detailliertes Feedback benötigen (Ryssel, 2018). Allerdings konnte kein genereller Vorteil vom elaborierten Feedback gegenüber dem einfachen Feedback KCR nachgewiesen werden, was erneut zeigt, dass Feedback von diversen Einflussfaktoren wie den Lerngruppen oder dem konkreten Anwendungsfall beeinflusst wird (Bangert-Drowns et al., 1991; Ryssel, 2018).

2.3.2 Ein Modell für wirksames Feedback im Unterricht

Wie im letzten Abschnitt skizziert, gibt es zahlreiche, teilweise inkonsistente Ergebnisse zur Effektivität von Feedback. Deshalb entwickeln unterschiedliche

Forschungsgruppen theoretische Modelle, um diese Faktoren detaillierter beschreiben und so die zentralen Merkmale genauer identifizieren zu können (Narciss, 2014).
Kulhavy und Stock (1989) entwickelten ein Feedbackmodell, bei dem als zentraler Faktor die individuelle Komponente der Antwortsicherheit im Mittelpunkt steht. Ein weiteres Modell findet sich bei Bangert-Drowns et al. (1991). Dieses fünfstufige Feedbackmodell fokussiert sich ebenfalls auf einen individuellen Faktor, nämlich die Feedbackverarbeitung. Der theoretische Ansatz von Kluger und DeNisi (1996) fokussiert die Leistung und den Einfluss des Feedbacks auf die Aufmerksamkeitsregulation. Bei dem multidimensionalen interaktiven Feedbackmodell von Narciss (2006) werden neben individuellen auch situative Faktoren identifiziert, die für die Wirkung des Feedbacks relevant sind.
Um das Feedback in dieser Arbeit genauer beschreiben und einteilen zu können, soll das Modell von Hattie und Timperley (2007) genauer betrachtet werden. Das Modell wurde speziell im Kontext des formativen Assessments entwickelt und auf einer breiten empirischen Basis aufgebaut. Zudem sollen nach Hattie und Timperley (2007) bei einem effektiven Feedback sowohl die Lernenden als auch die Lehrenden einbezogen werden: „*When teachers seek, or at least are open to, feedback from students as to what students know, what they understand, where they make errors, when they have misconceptions, when they are not engaged—then teaching and learning can be synchronized and powerful. Feedback to teachers helps make learning visible*“ (Hattie, 2009, S. 173). Aus diesen Gründen wurde das Modell von Hattie und Timperley (2007) für die vorliegende Arbeit ausgewählt (siehe Abbildung 2.3).
Aus Abbildung 2.3 geht hervor, dass effektives Feedback drei zentrale Fragen beantworten muss: Was sind die Ziele? Welcher Fortschritt wurde gemacht? Was muss gemacht werden, um das Ziel besser zu erreichen? Vergleicht man diesen Grundgedanken mit den drei Leitfragen des formativen Assessments (vgl. Abschnitt 2.2), fällt auf, dass der Ausgangspunkt für beide Ansätze fast identisch ist. Es lässt sich so herausarbeiten, dass Feedback ein zentraler Aspekt des formativen Assessments ist. Allerdings würde die Reduzierung des formativen Assessments nur auf Feedback zu kurz greifen, da immer die Verwendung und die Funktion der Rückmeldung mitbetrachtet werden müssen (Maier, 2010).
Nach Hattie und Timperley (2007) kann man von einer idealen Lernumgebung sprechen, wenn sowohl die Lernenden als auch die Lehrenden versuchen, die drei zentralen Fragen aus Abbildung 2.3 zu beantworten. Welche Auswirkungen

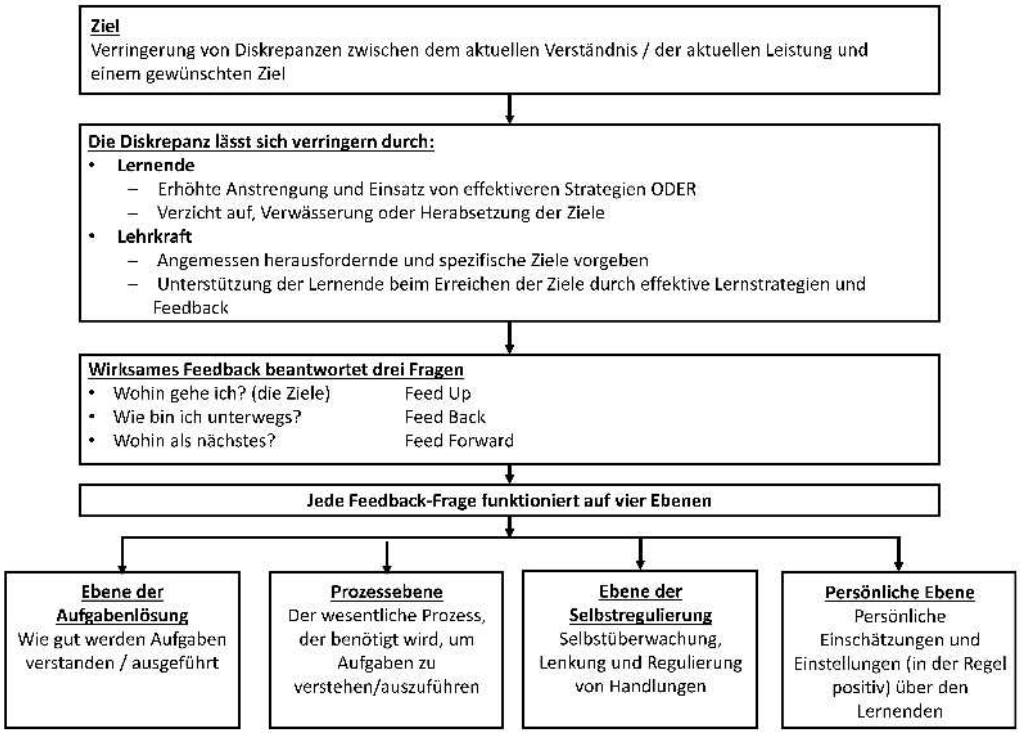


Abbildung 2.3: Das Feedbackmodell nach Hattie und Timperley (2007) (Übersetzung durch Autor)

diese Antworten auf die Verringerung der Differenzen zwischen dem aktuellen Lernstand und dem festgelegten Ziel haben, hängt von der Ebene ab, auf die das Feedback abzielt (Hattie, 2009). Die zu betrachtenden Ebenen sind die Ebene der Aufgabenlösung, die Ebene des Bearbeitungsprozesses, die Ebene der Selbstregulierung und die persönliche Ebene (vgl. Abbildung 2.3).

Feedback zur Aufgabenlösung enthält Informationen, ob eine Aufgabe richtig oder falsch gelöst wurde (z. B. knowledge of results, vgl. Tabelle 2.1). Diese Art des Feedbacks wird im Schulalltag häufig eingesetzt und wird auch korrigierendes Feedback genannt (Hattie & Timperley, 2007). Es weist eine sehr hohe Wirksamkeit auf, besonders bei einfacheren Aufgabentypen und wenn die Rückmeldung schnell erfolgt (N. Wolf, 2014). Außerdem ist Feedback auf dieser Ebene effektiv, wenn Lernende eine fehlerhafte Interpretation aufweisen und nicht ein Mangel an Informationen vorliegt (Hattie & Timperley, 2007). Falls ein Mangel an Informationen vorliegt, sind weitere Hinweise notwendig, um ein effektives Feedback zu gewährleisten. Allerdings kann diese Ebene des Feedbacks eine Grundlage für weiteres Feedback sein (Hattie & Timperley, 2007).

Feedback in der Prozessebene bezieht sich auf die zugrunde liegenden Lösungs-

methoden und -prozesse der einzelnen Aufgaben (Hattie & Timperley, 2007). Die Lernenden sollen mithilfe des Feedbacks ihre erstellten Lösungsweisen kritisch hinterfragen und reflektiert auf Fehlersuche gehen. Dadurch können Fehler beseitigt und neue Lösungsansätze entwickelt werden (N. Wolf, 2014). Durch prozessbezogenes Feedback kann, im Vergleich zum Feedback auf Aufgabenebene, ein tieferes Verständnis der Lerninhalte und ein nachhaltigeres Lernen erreicht werden, da kognitive Prozesse aktiviert werden (Hattie & Timperley, 2007; N. Wolf, 2014). Es ist besonders wirksam, wenn es verzögert eingesetzt wird (N. Wolf, 2014). Allerdings sollte man bedenken, dass beide Feedbackebenen einander beeinflussen können. Feedback zur Aufgabenlösung kann zu einem besseren Vertrauen in die Aufgabe und in sich selbst führen, was zu einer erhöhten Motivation für eine aktive Fehlersuche und Nutzung neuer Lösungsstrategien führt (Hattie & Timperley, 2007).

Selbstregulierung beinhaltet eine Vielzahl von unterschiedlichen Aspekten wie die Fähigkeit, sich selbst einzuschätzen, der Umgang mit Feedback, das Vertrauen in die Feedbackinformationen oder die Bereitschaft, Hilfe anzunehmen (Hattie & Timperley, 2007). Feedback in dieser Ebene beinhaltet demnach keine inhaltlichen Informationen zur Aufgabe und bezieht sich ebenfalls nicht auf den Lösungsprozess der Lernenden. Da der Effekt, der durch ein Feedback zur Selbstregulation ausgelöst wird, erst über einen größeren Zeitraum messbar wird, (N. Wolf, 2014) wird diese Ebene nicht weiter betrachtet werden.

Die von Hattie und Timperley (2007) eingeführte letzte Ebene Feedback zur Selbsteinschätzung verspricht den geringsten Effekt aller vier Ebenen, da die Informationen kaum inhaltlichen Wert haben (Hattie & Timperley, 2007). Eine typische Rückmeldung in dieser Ebene ist ein Lob an die Lernenden wie „gut gemacht“. Diese Information beantwortet nicht nur keine der drei Fragen (vgl. Abbildung 2.3), sondern enthält kaum inhaltsbezogene Informationen und führt ebenfalls nicht zu neuen Lernstrategien oder erhöhter Selbstregulation. Allerdings kommt diese Art der Rückmeldung vielfach im Unterrichtsalltag vor, weswegen sie im Modell von Hattie und Timperley (2007) aufgeführt wird (Hattie & Timperley, 2007). Lob oder Hinweise, die zu einem Vergleich mit anderen Lernenden führen, lösen zudem eher eine Auseinandersetzung mit sich selbst aus und nicht mit der Lösung der zu bearbeitenden Inhalte (N. Wolf, 2014).

Aus der Betrachtung des Modells von Hattie und Timperley (2007) lassen sich auch Konsequenzen für das formative Assessment ableiten. Feedback in einem formativen Assessment muss Lehrenden oder Lernenden diagnostische Informa-

tionen liefern, die eine der drei zentralen Fragen *Was sind die Ziele?*, *Welcher Fortschritt wurde gemacht?* oder *Was muss gemacht werden, um das Ziel besser zu erreichen?* in einer der vier Ebenen beantworten (vgl. Abbildung 2.3). Das bedeutet, dass Möglichkeiten während des Unterrichts geschaffen oder gezielt Aufgaben konzipiert werden müssen, die Hinweise über die Aufgabe, die Lösungsprozesse oder die Selbstregulation der Lernenden bereitstellen (Hattie & Timperley, 2007). Daraus lässt sich schließen, dass von Feedback in einem formativen Assessment beide Seiten profitieren können. Die Lernenden erhalten wichtige Informationen über ihre Kenntnisse und potenzielle Wissenslücken sowie Hinweise zu ihren Lösungsstrategien. Die Lehrenden können ihren Unterricht reflektieren und so weitere Schritte im Lehr-Lernprozess effektiv an die Bedürfnisse ihrer Lernenden anpassen.

2.3.3 Automatisches Feedback

Die bisherige Betrachtung hat gezeigt, dass es viele unterschiedliche Kriterien gibt, wie Feedback gestaltet werden kann. Die Sichtung des Forschungsstandes legt dar, dass aufgrund der vielfältigen Faktoren, die auf die Effektivität von Feedback einwirken können, keine konsistenten Forschungsergebnisse vorliegen. So konnte beispielsweise gezeigt werden, dass der Einsatz von elaboriertem Feedback keinen grundsätzlichen Vorteil gegenüber einfachem Feedback bietet. In dem empirischen Teil dieser Arbeit wird ein Feedback eingesetzt (siehe Kapitel 7). Die Grundlage dieses Feedbacks wird durch ein Machine-Learning-Modell erstellt, weswegen man von einem automatisch generierten Feedback sprechen kann. Um dieses Feedback zu konzipieren und auf einer theoretischen Grundlage aufzubauen, werden in diesem Abschnitt ausgewählte Faktoren und Forschungsarbeiten betrachtet, die speziell für ein automatisch generiertes Feedback relevant sind.
Für das formative Assessment ist eine transparente und klar verständliche Rückmeldung ein notwendiges Kriterium: „*Auch Sadler (1998) weist in diesem Zusammenhang darauf hin, dass es nicht nur auf die Art der Rückmeldung, sondern auch die Qualität ankommt, woraus sich für zukünftige Studien die Notwendigkeit ergibt, die einzelnen Methoden formativer Leistungsbeurteilung stärker qualitätsbezogen, möglicherweise qualitativ [. . .] zu erfassen*“ (Bürgermeister et al., 2014, S. 51). Demnach sind Qualitätsmerkmale wie Verständlichkeit der Informationen oder die Angemessenheit des Feedbacks, wichtig für eine positive Unterstützung des Lehr-Lernprozesses (N. Wolf, 2014). So finden sich in der Literatur Studien über

Leistungssteigerungen der Lernenden, aber auch über Unzufriedenheit mit dem Feedback (Cavalcanti et al., 2021). Nach Burke (2009) kann dies auf einen zu hohen Komplexitätsgrad, die Länge des Feedbacks und eine negative Formulierung des Feedbacks zurückgeführt werden. Dies verdeutlicht noch einmal die Komplexität und Mehrdimensionalität eines effektiven Feedbacks. An dieser Stelle muss auch darauf hingewiesen werden, dass viele Lernende nie gelernt haben, mit Feedback umzugehen. Weaver (2006) konnte nachweisen, dass mehr als die Hälfte der teilnehmenden Studierenden nie eine Einführung in das Verständnis und die Nutzung von Feedback erhalten hatten.
Ein vielfach genutztes automatisches Feedback ist der direkte Vergleich der Antworten der Lernenden mit der gewünschten Antwort der Lehrenden (Cavalcanti et al., 2021). Außerdem werden Visualisierungen in Form von Dashboards und KI-basierten Methoden wie Machine Learning eingesetzt (z. B. Botelho et al., 2023; Lee et al., 2021). Dabei ist zu beachten, dass Feedback nicht nur aus minimalen Informationen besteht, sondern, wie von Hattie und Timperley (2007) gefordert, auch neue Lernstrategien oder weiterführende Hinweise liefert. Dazu könnten Daten genutzt werden, die über die Lernprozesse der Lernenden Auskunft geben und die Lehrkräfte systematisch über den Fortschritt informieren. Auf Grundlage dieser Informationen können dann Lehrende Rückmeldung geben und weitere Unterrichtsschritte anpassen oder die Feedbackysteme verfassen direkt Rückmeldungen an die Lernenden (Blikstein et al., 2014; Pardo et al., 2019). Der Nachteil solcher KI-basierten Methoden ist allerdings, dass im Vorfeld eine große Datenmenge gebraucht wird, um die Modelle für ihre Aufgabe zu trainieren (Zhai et al., 2020).
Die meisten automatischen Feedbacksysteme werden eingesetzt, um entweder Rückmeldung über die Leistung der Lernenden bei einer bestimmten Aufgabe zu geben oder um Unterstützung zur Selbstregulation bereitzustellen (Cavalcanti et al., 2021). Ein Vorteil solcher Feedbacksysteme ist die Möglichkeit, die Anzahl der erneuten Lösungsversuche zu modifizieren. Eine Methode besteht darin, direkt nach der ersten falschen Antwort Feedback zu geben (single-try-feedback). Eine andere Möglichkeit besteht darin, den Lernenden eine vorher festgelegte Anzahl an Lösungsversuchen einzuräumen (multiple-try-feedback). Bei der dritten Option erhalten die Lernenden so lange Rückmeldung, bis sie die richtige Lösung gefunden haben (answers-until-correct) (Clariana, 1993; Narciss, 2008; Ryssel, 2018). Es konnten keine generellen Vorteile für eine der drei Möglichkeiten nachgewiesen werden (Ryssel, 2018). Allerdings zeigte sich, dass bei komplexeren Aufgaben

das multiple-try-feedback am effektivsten ist (Clariana & Koul, 2005). Ryssel (2018) geht davon aus, dass Lernende mit geringem Vorwissen mehr Fehler bei der Bearbeitung der Aufgabe machen und sie deshalb schneller aufgeben und eher eine direkte Präsentation der korrekten Lösung bevorzugen. Dagegen weisen Lernende mit einem hohen Vorwissen eher die kognitiven Fähigkeiten auf, die richtige Lösung selbstständig zu erarbeiten. Außerdem verlieren sie nicht die Motivation, wenn die richtige Lösung nicht direkt präsentiert wird (Clariana et al., 2000; Ryssel, 2018). Deshalb können Lernende mit einem hohen Vorwissen von einem Feedbacksystem profitieren, welches mehrere Lösungsversuche anbietet (Clariana, 1993).
Weiterhin hat sich gezeigt, dass beim computergestützten Feedback die Präsentation nicht unimodal, sondern besser bimodal, z. B. auditiv und visuell, durchgeführt werden sollte (Mayer & Moreno, 1998). Durch die Kombination mehrerer Präsentationsformen kann die kognitive Belastung der Lernenden reduziert werden (Mayer & Moreno, 2003; Ryssel, 2018). Hinzu kommt, dass eine Rückmeldung, die von einem Computer generiert wird, weniger angstinduziert ist als Rückmeldungen, die von Menschen übermittelt werden (Lenhard et al., 2012). Ein weiterer Punkt ist, dass verbales Feedback im Vergleich zu computergeneriertem Feedback die Aufmerksamkeit auf den Feedbackgebenden lenkt und nicht auf die zu übermittelnden Informationen (Kluger & DeNisi, 1996; Ryssel, 2018).
Man muss jedoch berücksichtigen, dass die Qualität automatischer Rückmeldung kritisch hinterfragt werden muss. Eine Lehrkraft wird in den meisten Fällen in der Lage sein, qualitativ hochwertigeres Feedback als ein automatisches Feedbackmodell zu erzeugen (Herding et al., 2010). Eine erfahrene Lehrkraft besitzt ein viel breiteres und individuell anpassbares Set an Aktionen als das meist vordefinierte automatische Feedback (Herding et al., 2010). Allerdings nehmen solche Systeme viel Arbeit, in Form von Auswertungen und Analysen, ab. Bei einer großen Lerngruppe ist es schwierig, jedem Lernenden ein hochwertiges individuelles Feedback zu geben. Speziell bei offenen Aufgaben kann eine solche Rückmeldung komplex und zeitaufwendig sein. Daher fordern Cavalcanti et al. (2021) mehr Forschungsarbeiten, die sich systematisch mit der Untersuchung und Analyse der Qualität von automatischem Feedback befassen.
Ein Großteil der von Cavalcanti et al. (2021) analysierten Studien befasste sich mit Feedback als Lernhilfe für Lernende. Lediglich drei Studien hatten das Ziel, die Lehrerenden mit einem automatischen Feedback zu unterstützen (Cavalcanti et al., 2021). Eine der Forschungsarbeiten ist der Artikel von Martin et al. (2009).

Die Forschenden schlugen das System MAGADI vor, das Daten aus einer E-Learningplattform sammelt und auswertet. Die Ergebnisse werden den Lehrkräften in einem Dashboard angezeigt und mögliche Verbesserungen vorgeschlagen (Martin et al., 2009). Diese Visualisierung von Informationen zur Unterstützung der Lehrenden kann ein wirksames Instrument sein, das nicht nur die Arbeitsbelastung der Lehrenden verringert, sondern auch zu einem besseren Verständnis des Verhaltens der Lernenden beiträgt (Cavalcanti et al., 2021; Lim et al., 2021). Cavalcanti et al. (2021) betonen aber auch, dass Lehr-Lern-Dashboards sorgfältig konzipiert werden müssen. Die Autoren weisen daher auf die Notwendigkeit hin, mehr Forschung zu betreiben, die sich stärker auf die Lehrenden konzentriert und untersucht, wie mit dem automatisierten Feedback umgegangen wird, wie es die Unterrichtspraxis verändert und wie die Qualität des Feedbacks ist (Cavalcanti et al., 2019, 2021).

2.4 Befunde der Lehr-Lern-Forschung

Im letzten Abschnitt werden Ergebnisse zur Lernwirksamkeit des formativen Assessments beleuchtet. Außerdem sollen Herausforderungen bei der Implementierung in den Schulalltag dargestellt werden.

2.4.1 Lernwirksamkeit

Das formative Assessment ist eine in vielen Studien mit unterschiedlichen Schwerpunkten untersuchte Methode. Wie in den vorherigen Abschnitten beschrieben, wird formatives Assessment oft zur Förderung des Lernens eingesetzt, weswegen viele Studien Effekte auf die Leistung der Lernenden untersucht haben (Schmidt, 2020).
Eine in diesem Kontext viel zitierte Studie ist die Arbeit von Black und Wiliam (1998). Die Autoren identifizierten formatives Assessment als eine der wirksamsten Methoden zur Optimierung schulischen Lernens, da sie eine hohe Effektstärke nachweisen konnten. Es konnte auch in weiteren Studien gezeigt werden, dass nicht nur leistungsschwächere Lernende von einem formativen Assessment profitieren können, sondern die Lernleistung von allen Lernenden gesteigert werden konnte (Maier, 2010).
In neuen Arbeiten werden jedoch die Ergebnisse von Black und Wiliam (1998) kritisch betrachtet. Dunn und Mulvenon (2009) und Bennett (2011) stellten methodische Schwächen in den von Black und Wiliam (1998) zitierten Studien fest.

Zudem kritisierten Dunn und Mulvenon (2009) die teilweise nicht konsistente Verwendung des Begriffs des formativen Assessments. Es wird weiterhin festgestellt, dass Black und Wiliam (1998) zu allgemeine Schlussfolgerungen auf Grundlage der zitierten Studien getroffen haben, da die Arbeiten sich nur auf eine bestimmte Gruppe von Lernenden oder unterschiedliche Aspekte des formativen Assessments beziehen (Schütze et al., 2018).
Jedoch konnten auch neue Forschungsarbeiten die positive Einschätzung des formativen Assessments bestätigen und einen kleinen bis mittleren positiven Effekt auf die Leistung der Lernenden nachweisen (z. B. Dunn & Mulvenon, 2009; Kingston & Nash, 2011, 2015). Studien, die sich mit einer Moderatoranalyse befasst haben, zeigen, dass die Effektstärke vom Schulfach beeinflusst werden kann. So wurde in naturwissenschaftlichen Fächern eine kleinere Effektgröße als im Englischunterricht nachgewiesen, was auf eine höhere Komplexität der Naturwissenschaften zurückzuführen ist (Kingston & Nash, 2011, S. 33; Schmidt, 2020). Die Klassenstufe scheint allerdings keinen Einfluss auf den Effekt zu haben, wohingegen der Einsatz von computerbasierten formativen Assessments Einfluss haben kann (Schmidt, 2020). So konnten in den USA Leistungseffekte von computerbasierten formativen Assessments nachgewiesen werden (z. B. Nunnery et al., 2006; Yeh, 2006). Lehrerende berichteten, dass durch die Verwendung solcher Systeme eine gezielte Optimierung des Unterrichts möglich war (Maier, 2010; Yeh, 2006). Auch die Literaturübersicht von McLaughlin und Yan (2017), die sich auf die Wirksamkeit von onlinebasierten formativen Assessments wie elektronischen Student-Response-Systemen oder Multiple-Choice-Tests konzentrierte, zeigte positive Effekte auf die Leistung der Lernenden.
Nach Wiliam und Leahy (2007) kann der größte Effekt auf die Leistung der Lernenden durch zeitnahe Rückmeldungen erreicht werden. Von sogenannten *Short-cycle*-Rückmeldungen kann gesprochen werden, wenn die Rückmeldung unmittelbar nach der Lernhandlung folgt. *Medium-cycle*-Rückmeldungen, die ebenfalls einen hohen Einfluss auf die Leistung haben, liegen zwischen den Schritten nur wenige Tage, z. B. bei zwei aufeinander folgenden Unterrichtsstunden (Maier, 2010).
Es konnten aber nicht nur Effekte auf die Lernleistung der Lernenden nachgewiesen werden. Das formative Assessment kann auch positiven Einfluss auf die Selbstregulation und die Motivation der Lernenden haben (siehe Schmidt, 2020). Zusammenfassend kann also festgehalten werden, dass formatives Assessment positive Einflüsse auf die Lernenden haben kann. Allerdings zeigen die Studien auch,

dass die konkrete Umsetzung und externe Faktoren wie das Schulfach Einfluss auf die Effekte haben können. Deshalb sind nach Maier (2010, S. 301) „*die empirischen Belege für die Effektivität formativer Leistungsmessung überzeugend*“, jedoch fordern Souvignier und Hasselhorn (2018) mehr Studien, in denen konkrete Umsetzungen von formativen Assessments evaluiert werden sollen.

2.4.2 Herausforderungen im Schulalltag

Trotz der überwiegend positiven Berichte über formatives Assessment wird es in vielen Klassenräumen nicht umgesetzt (Black & Wiliam, 1998). Dies lässt sich auf verschiedene Faktoren zurückführen.

Für ein formatives Assessment benötigen Lehrkräfte eine Vielzahl von Kompetenzen. Lehrkräfte müssen nicht nur den Inhalt kennen und verstehen, sondern sie müssen auch die verschiedenen Lernwege der Lernenden berücksichtigen können (Hunt & Pellegrino, 2002). Zudem müssen Lehrkräfte Möglichkeiten schaffen, diagnostische Informationen zu erheben. Diese müssen dann in einer effektiven und möglichst schnellen Weise ausgewertet werden, um aufbauend daraus Handlungsoptionen gewinnen zu können (Schildkamp et al., 2020). Lehrkräfte benötigen also neben fachlichen und fachdidaktischen Kompetenzen spezifische Assessment-Kompetenzen (diagnostische Informationen auswerten, interpretieren und nutzen), um ein formatives Assessment erfolgreich im Unterricht einsetzen zu können (Schütze et al., 2018).

Diese vielfältigen Kompetenzen beziehen sich nicht nur auf eine Art des formativen Assessments. Wie die letzten Abschnitte gezeigt haben, sind formative Assessments in vielfältiger Weise einsetzbar, weswegen Lehrkräfte für eine ganze Reihe verschiedener Methoden diese Abläufe beherrschen müssen. Van der Kleij und Eggen (2013) konnten zeigen, dass einige Lehrkräfte nicht in der Lage waren, Informationen, die durch eine Beobachtung der Lernenden erhoben worden, gewinnbringend zu interpretieren. Durch unzureichende Interpretation der gewonnenen Informationen kann es zu einer ineffektiven Nutzung des formativen Assessments kommen (Schildkamp et al., 2020).

Falls Lehrkräfte die diagnostischen Informationen verarbeiten und auswerten können, ist eine individuelle Rückmeldung aufgrund der großen Anzahl an Lernenden oftmals nicht möglich (Hunt & Pellegrino, 2002). Zwar kann eine Lehrkraft eine hohe Bewertungskompetenz haben, jedoch ist ein formatives Assessment wenig hilfreich, wenn die konkrete Umsetzung zeitlich nicht realisierbar ist. So

weisen Hunt und Pellegrino (2002) darauf hin, dass eine Lehrkraft, die in einem Schulalltag mehrere Klassen mit jeweils 30 Lernenden unterrichtet und zusätzliche administrative Aufgaben erledigen muss, nicht in der Lage ist, ein effektives formatives Assessment durchzuführen. Daher wird der Zeitaufwand als eine der größten Herausforderungen bei der Implementierung von formativen Assessments im Schulalltag angesehen (Bennett, 2011; Black & Wiliam, 1998). Deshalb wird die Nutzung von computergestützten Verfahren gefordert, die Informationen über die Lernenden sammeln und auswerten können. Auf diese Weise können Lehrkräfte Hinweise und Informationen erhalten, die sie für weitere Schritte nutzen können (Hunt & Pellegrino, 2002, S. 75).
Zeuch et al. (2017) zeigten außerdem, dass die Nutzung von erhobenen diagnostischen Informationen für die Planung der weiteren Unterrichtsstunden nicht zu den etablierten Routinen von Lehrkräften gehören. Dies erfordere Veränderungen im Lehr-Lern-Prozess und das Erlangen neuer, vorher unbekannter Prozesse und Methoden (Schütze et al., 2018). Daher benötigen Lehrkräfte oftmals ein Training oder gezielte Unterstützung, um das Potenzial von formativen Assessments auszuschöpfen (siehe Stecker, 2017; Schütze et al., 2018). Die Implementierung von formativen Assessments im Schulalltag kann also ein langwieriger Prozess sein, der in kleinen Schritten vorgenommen werden sollte (Angelo & Cross, 1993; Harlen, 2008; Schütze et al., 2018).

3 Concept Maps

Concept Maps wurden Anfang der 1970er-Jahre von J.D. Novak als grafisches Hilfsmittel zur Darstellung und Organisation von Inhalten und Wissen entwickelt (Novak & Cañas, 2008; Ryssel, 2018). Im deutschsprachigen Raum werden sie auch oftmals als Begriffskarte, Begriffslandkarte oder Begriffsnetz bezeichnet. Concept Maps bestehen in der Regel aus mehreren Begriffen (Concepts), die durch gerichtete und beschriftete Pfeile miteinander verbunden und in Beziehung gesetzt werden. Die beschrifteten Pfeile werden Relationen genannt und die Kombination aus Begriff – Relation – Begriff wird als Proposition bezeichnet (Ryssel, 2018; Stracke, 2004). Die beschrifteten Pfeile verknüpfen zwei Begriffe miteinander, um so den Sinnzusammenhang zwischen ihnen darzustellen (Ley, 2015).
In einer Concept Map stellen die Begriffe, welche meistens in Kästchen oder Kreisen dargestellt werden, Objekte, Beispiele oder Ereignisse dar. Relationen hingegen können mit Adjektiven, Verben, Formeln oder ganzen Sätzen formuliert werden, welche eine sinnvolle Aussage zwischen den Begriffen herstellen sollen (Novak & Cañas, 2008).
Durch diese Eigenschaften lassen sich Concept Maps auch von Mindmaps abgrenzen, da in einer Mindmap die Beziehungen zwischen den Begriffen unbestimmt bleiben und diese vor allem fürs Brainstorming eingesetzt werden (Ryssel, 2018). Concept Maps hingegen haben ein vielfältigeres Einsatzgebiet, so auch als Assessment-Methode. Ruiz-Primo und Shavelson (1996) haben zur Beschreibung von Concept Maps als Assessment-Methode die drei Kategorien Aufgaben-, Antwort- und Bewertungsformat eingeführt.
Unter dem Aspekt Antwortformat verstehen die Autoren die Wahl des Mediums, welches zur Bearbeitung der Concept Map genutzt wird. Typischerweise wird entweder Stift und Papier oder eine geeignete Concept-Map-Software genutzt (Ley, 2015). Da in der vorliegenden Arbeit die Concept Maps digital und automatisch ausgewertet werden sollen, wird auf eine Concept-Map-Software zurückgegriffen, die im Abschnitt 6.1 genauer beschrieben wird.
In diesem Kapitel werden daher nur die beiden anderen Aspekte, das Aufgaben- und das Bewertungsformat, thematisiert. Da die Concept Map in dieser Arbeit

als eine formative Assessment-Methode eingesetzt wird, wird der Einsatz von Concept Maps zu diesem Zwecke ebenfalls diskutiert.

3.1 Aufgabenformat

Nach Novak und Gowin (1984) besitzen Concept Maps primär eine hierarchische Struktur. Bei dieser Struktur steht ein allgemeiner, übergeordneter Begriff an der Spitze, welcher durch weitere Begriffe spezifiziert wird. Im Laufe der Jahre hat sich allerdings gezeigt, dass ein hierarchischer Aufbau nicht zwangsläufig notwendig oder sogar gar nicht erst möglich ist (Ryssel, 2018; Yin et al., 2005). So postulierte Kinchin et al. (2000) neben der hierarchischen *chain*-Struktur noch zwei weitere Strukturen *spoke* und *net* (siehe Abbildung 3.1).

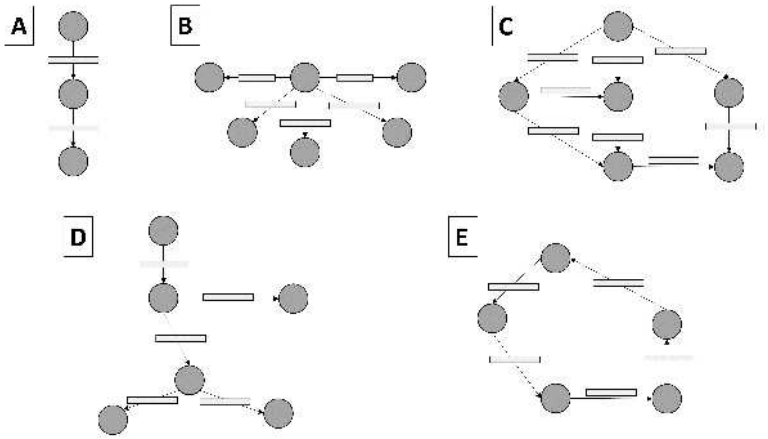


Abbildung 3.1: Concept-Map-Strukturen nach Kinchin et al. (2000) und Yin et al. (2005): A) chain, B) spoke, C) net, D) tree, E) circle

Die *chain*-Struktur entspricht der klassischen Concept Map nach Novak und Gowin (1984), bei der die Propositionen in einer linearen Abfolge aufgebaut sind und bei der jeder Begriff nur mit dem unmittelbaren darüber oder darunterliegenden Begriff verknüpft wird. Die radiale *spoke*-Struktur besitzt im Gegensatz dazu einen zentralen Begriff, der mit allen weiteren Begriffen verknüpft wird. Die *net*-Strutkur weist den höchsten Vernetzungsgrad auf, weswegen nach Kinchin et al. (2000) dies als Anzeichen für ein effektives Lernen gesehen werden kann.
Yin et al. (2005) ergänzten die Charakterisierung durch die *circle*- und *tree*-Struktur, um die vielfältigen Strukturen von Concept Maps noch besser erfassen zu können (siehe Abbildung 3.1). Die *tree*-Struktur ähnelt dabei der klassischen

chain-Struktur, mit dem Zusatz, dass Abzweigungen oder Äste hinzugefügt wurden. Bei der *circle*-Struktur werden die Propositionen aneinandergereiht, wobei deren Enden miteinander verbunden sind (Yin et al., 2005). Concept Maps können aus mehreren dieser fünf Grundstrukturen aufgebaut sein.
Eine andere Möglichkeit, Concept Maps zu charakterisieren, ist über den Grad der Vorstrukturierung (Ruiz-Primo, 2004). Bei sehr strikten Vorgaben erhalten die Lernenden gewisse Komponenten wie die Begriffe oder die Relationen bereits im Vorfeld. Das bedeutet, dass die Lernenden mit einer (teilweise) vorstrukturierten Concept Map arbeiten müssen (Ruiz-Primo, 2004; Ryssel, 2018). Die Vorgaben können immer weiter geöffnet werden, sodass gar keine Vorstrukturierung mehr durchgeführt wird. Die Lernenden müssen demnach eigenständig entscheiden, welche und wie viele Begriffe sie in ihrer Concept Map benutzen wollen und wie diese in Beziehung gesetzt werden können (Ruiz-Primo, 2004).
Cañas, Novak und Vanhear (2012) erweitern diese Charakterisierung bezüglich der strukturellen Vorgaben um eine weitere Dimension, die inhaltliche Vorgabe. Folglich unterscheiden die Autoren zwischen der inhaltlichen Freiheit, bei der es um die Freiheit bezüglich des Themas oder der Wahl der Begriffe geht und der strukturellen Freiheit, bei der es um grafische Strukturen geht (siehe Abbildung 3.2).
Die maximale inhaltliche und strukturelle Freiheit wäre demnach gegeben, wenn man eine Concept Map zu einem beliebigen Thema ohne Einschränkungen bzw. Vorgaben zu Begriffen und Relationen erstellt (obere rechte Ecke). Im Gegensatz dazu wäre das reine Auswendiglernen einer fertigen Concept Map ein völliger Mangel beider Freiheitsdimensionen (untere linke Ecke).
Zwischen diesen Extremen liegt eine Vielfalt von unterschiedlichen Concept-Map-Formaten. Wie der Überblicksartikel von Strautmane (2012, S. 2) zeigt, existieren über 700 verschiedene Möglichkeiten, eine Concept Map zu konstruieren. An dieser Stelle muss deshalb angemerkt werden, dass die Abbildung 3.2 keinen Anspruch auf eine vollständige Darstellung aller Variationen von Concept Maps hat. Die Abbildung dient eher einer exemplarischen Übersichtsgrafik über in der Bildungsforschung häufig eingesetzte Concept-Map-Aufgabenformate (Cañas, Novak & Reiska, 2012). Zudem stammt die Einordnung der in Abbildung 3.2 gezeigten Variationen eher aus einer qualitativen und keiner quantitativen Analyse. Die Positionen der einzelnen Formate entsprechen daher relativen anstatt absoluten Werten (Cañas, Novak & Reiska, 2012). Im Folgenden sollen daher nicht einzelne Formate im Detail diskutiert werden, sondern vielmehr die Vor- und Nachteile, die

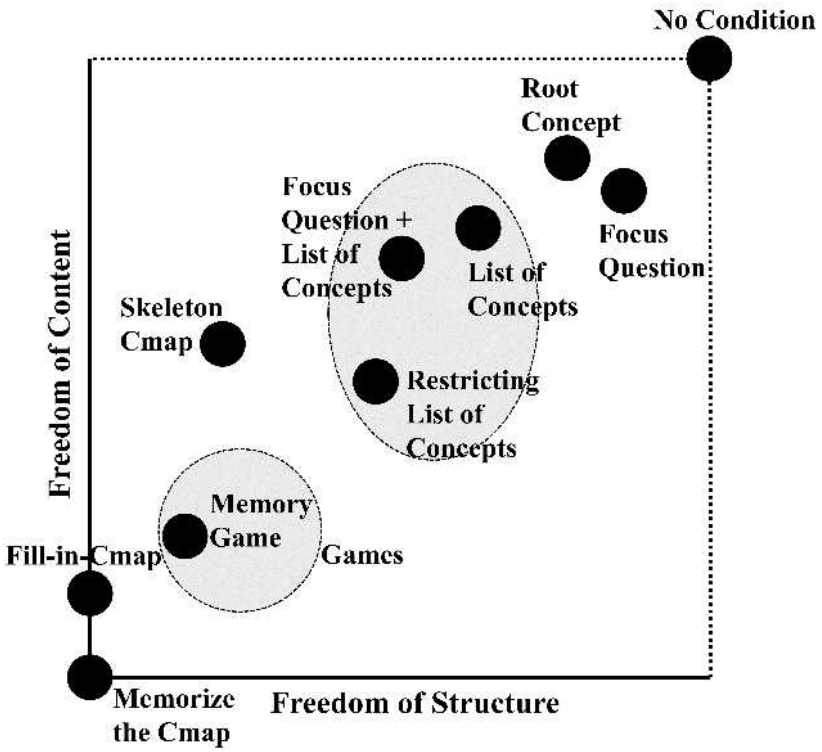


Abbildung 3.2: Charakterisierung von Concept Maps bezüglich der strukturellen und inhaltlichen Vorgaben (Cañas et al., 2023)

sich aus den unterschiedlichen inhaltlichen und strukturellen Vorgaben ergeben.

Offene vs. geschlossene Aufgabenformate

Nicht jede Concept Map kann dieselben Informationen über Wissensstrukturen von Lernenden liefern und nicht jede Concept Map ist gleich kognitiv herausfordernd (Ruiz-Primo, 2004; Ruiz-Primo & Shavelson, 1996; Yin et al., 2005). Offene Formate wie die eigenständige Konstruktion einer Concept Map (in Abbildung 3.2 z. B. *Root Concept* oder *Focus Question*) können tiefergehendes Lernen fördern, da die Lernenden eine aktive Rolle einnehmen müssen (Ryssel, 2018). Die Lernenden müssen sich komplett selbstständig mit den Inhalten beschäftigen und zentrale Propositionen identifizieren (Stracke, 2004). Zudem können die Concept Maps im Verlaufe des Unterrichts stetig weiterentwickelt werden, um die behandelten Inhalte wie Experimente, Ergebnisse oder Phänomene in einer übersichtlichen Art und Weise zu sichern (Novak & Cañas, 2008). Man muss dabei jedoch berücksichtigen, dass die Komplexität der Concept-Map-Erstellung durch den Grad der Offenheit beeinflusst wird. Je weniger Vorgaben gemacht

werden, desto höher ist der Schwierigkeitsgrad und desto schneller kann es zu einer kognitiven Überlastung kommen, speziell für schwächere Lernende. Außerdem wird die Vergleichbarkeit der Concept Maps erschwert (Chang et al., 2002; Ley, 2015; Ruiz-Primo et al., 2001; Ryssel, 2018). Zusätzlich dazu zeigen diverse Belege, dass Lernende, die keine Vorerfahrung mit Concept Maps haben, Schwierigkeiten mit der Konstruktion haben und es so zu einer weiteren kognitiven Last kommen kann (Becker, 2022; Cañas et al., 2003; Novak & Cañas, 2008). Speziell bei den offenen Formaten müssen Lernende nicht nur Begriffe sammeln und ordnen, sinnvolle und korrekte Propositionen bilden, sondern auch metakognitive Strategien anwenden wie das Planen der Concept Map (Becker, 2022). Damit Lernende trotzdem mit Concept Maps arbeiten können, empfehlen verschiedene Autoren ein methodisches Concept-Map-Training, um die kognitive Belastung zu reduzieren (u. a. Cañas et al., 2003). Dabei ist ein Concept-Map-Training sowohl für schwache als auch für leistungsstarke Lernende sinnvoll (Mintzes et al., 2011). Für die genaue Umsetzung eines solchen Trainings gibt es verschiedene Ansätze, die jedoch meistens die relevanten Schritte wie wichtige Begriffe identifizieren und sinnvolle Propositionen finden enthalten (Becker, 2022).

Um die Komplexität einer Concept-Map-Aufgabe zu reduzieren, können bestimmte Aspekte der Concept Map vorgegeben sein (in Abbildung 3.2 z. B. *List of Concepts*). Dadurch wird zwar der Freiheitsgrad eingeschränkt, allerdings können z. B. Lehrkräfte so steuern, welche Inhalte relevant sind und in der Concept Map enthalten sein sollen. Dieses Format schränkt die Kreativität der Lernenden ein, jedoch kann man nicht unbedingt von einem einfachen Format sprechen, da „*the most challenging and difficult aspect of constructing a concept map is constructing the propositions[. . .]*“ (Novak & Cañas, 2008, S. 20). Lehrkräfte können Einblicke in das aktuelle Verständnis der Lernenden gewinnen, da sie trotz der Vorgaben sehen, welche Inhalte in der Concept Map nur wenig integriert wurden (Novak & Cañas, 2008). Geschlossene Formate eignen sich deshalb gut für die Aktivierung von Wissen oder für eine Vorwissensabfrage (Strautmane, 2012).

Ein weiteres geschlossenes Format sind *Fill-in-the-map*-Formate (Cañas, Novak & Reiska, 2012). Bei dieser Art von Concept Maps bereiten Lehrkräfte im Vorfeld eine Concept Map vor, in der bestimmte Aspekte freigelassen werden. Eine Möglichkeit wäre, eine vorstrukturierte Concept Map zu entwickeln, bei der bestimmte Begriffe ergänzt werden müssen. Eine andere Möglichkeit wäre das Weglassen bestimmter Relationen (Abbildung 3.3). Die Lernenden müssen sich bei diesem Aufgabenformat also nur auf bestimmte Propositionen fokussieren, weswegen

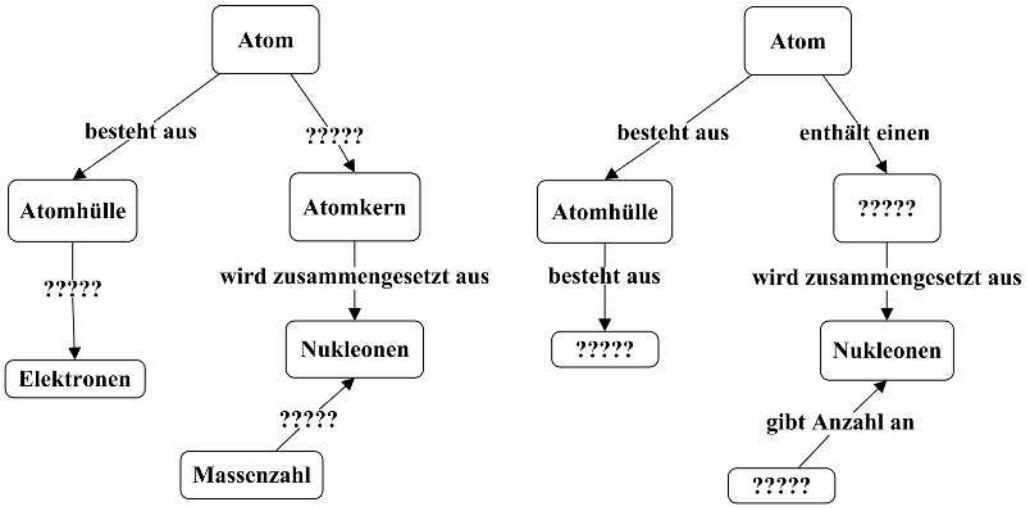


Abbildung 3.3: Beispiele für *Fill-in-the-map*-Format. Links: Festgelegte Begriffe, freie Relationen. Rechts: Festgelegte Relationen, freie Begriffe

sich *Fill-in-the-map*-Formate bei komplexen Inhalten oder wenig erfahrenen Lernenden anbieten (Novak & Cañas, 2008). So kann die Lehrkraft gezielt steuern, welche Inhalte und in welchen Umfang die Concept Map behandelt werden sollen (Stracke, 2004). Diese Art von Aufgabenformat kann allerdings auch so eingesetzt werden, dass die Lernenden weitere Begriffe oder Propositionen eigenständig ergänzen können (Ryssel, 2018).

3.2 Auswertungsformat

Concept Maps lassen sich mit verschiedenen charakteristischen Größen, z. B. graphentheoretische Maßen wie die Anzahl Begriffe oder die Anzahl von Zusammenhangskomponenten, beschreiben. Zur Auswertung einer Concept Map können diese Größen genutzt werden. Zudem ist der Vergleich individueller Concept Maps mit Expertenmaps in der Literatur zu finden; oder auch die Kombination aus beiden Auswertungsansätzen (Ruiz-Primo & Shavelson, 1996).
Die Betrachtung einzelner Größen zur Auswertung der Concept Map beruht auf der Arbeit von Novak und Gowin (1984), welche von vielen weiteren Forschungsarbeiten übernommen und immer weiter adaptiert wurde (siehe Anohina & Grundspenkis, 2009; Stracke, 2004). Die Kriterien von Novak und Gowin (1984) bieten sich primär bei hierarchischen Concept Maps an, da neben der Anzahl von Propositionen und der Anzahl von Querverbindungen auch die Hierarchieebenen in die Bewertung einfließen. Aus diesem Auswertungsansatz lassen sich

dann unterschiedliche Scores bilden. Betrachtet man verschiedene Concept Maps, lassen sich diese Kriterien bei sehr freien Ansätzen gut anwenden, bei denen die Lernenden eine Concept Map eigenständig entwickeln müssen. Durch diese quantitativen Kennzahlen kann man schnell einen Blick über die Strukturen der Concept Map gewinnen (Strautmane, 2012).
Kriterien über die Gesamtstruktur, die aus einer graphentheoretischen Betrachtung stammen, wie der Durchmesser der Concept Map, der die Anzahl der Propositionen zwischen den am weitesten entfernten Begriffen angibt, können für solch ein Concept-Map-Format ebenfalls angewendet werden (Ifenthaler, 2010). Ruiz-Primo und Shavelson (1996) weisen jedoch darauf hin, dass durch die wenigen Vorgaben ein Vergleich zwischen den einzelnen Concept Maps erschwert wird, was sich auch auf eine valide Auswertung auswirken kann. Zudem werden quantitative Kriterien für eine genaue Analyse der Concept Map eher kritisch gesehen und die Aussagekraft über den inhaltlichen Gehalt einer Concept Map wird mittlerweile hinterfragt (Ley, 2015; Ryssel, 2018).
Bei geschlossenen Formaten wie dem *Fill-in-the-map*-Format kann man zwar auch quantitative Kennzahlen berechnen, allerdings haben diese durch die festen Strukturen nur eine geringe Aussagekraft, da z. B. der Durchmesser der Concept Map vorgegeben ist. Es bieten sich daher qualitative Aspekte an wie eine Analyse der Qualität der Propositionen. Einige Arbeiten verwenden dazu dichotome Bewertungen der Propositionen, wohingegen andere eine vielschichtige qualitative Analyse durchführen (Strautmane, 2012). So analysierte z. B. Wadouh (2007) die Propositionen erstens auf fachliche Korrektheit, in dem sie die Propositionen in falsch, ungenau und richtig einordnete, und zweitens auf den Fachgehalt, der niedrig, mittel oder hoch sein konnte. So können die Zusammenhänge deutlich detaillierter betrachtet und falsche oder ungenaue Propositionen besser identifiziert werden, als das durch einen quantitativen Gesamtscore der Fall wäre (Ryssel, 2018). Jedoch muss bei der qualitativen Analyse auf den enormen Zeitaufwand hingewiesen werden (Stracke, 2004).
Zur Auswertung einer Concept Map kann man auch eine sogenannte Experten- oder Referenzmap nutzen. Bei der Verwendung einer Expertenmap wird angenommen, dass dadurch eine Art ideale Struktur vorliegt, die den Inhalt bestmöglich abbildet und die Lösungen der Lernenden sich dieser Idealstruktur mit fortschreitendem Lernprozess annähern (Stracke, 2004). Bei dem Vergleich der individuellen Concept Maps mit einer Expertenmap lassen sich sowohl quantitative als auch qualitative Bewertungen durchführen. Quantitative Kennzahlen werden meistens

über verschiedene Ähnlichkeitsmetriken bestimmt, die angeben, inwieweit die Lernenden-Concept-Map von der Expertenmap strukturell abweicht (Anohina & Grundspenkis, 2009; Strautmane, 2012). Einen eher qualitativen Ansatz schlägt Ruiz-Primo (2000) vor, bei dem das Verhältnis der richtigen Propositionen der Lernenden-Concept-Map im Verhältnis zu den insgesamt möglichen gültigen Propositionen in der Expertenmap betrachtet wird. So wird mehr auf einer qualitativen Ebene geschaut, anstatt lediglich strukturelle Aspekte zu betrachten. Bei dem Auswertungsansatz mittels Expertenmap stellt sich allerdings die Frage, inwiefern und von wem die Expertenmap konstruiert wurde. Da diese beim Vergleich als Idealstruktur angesehen wird, sollte sie möglichst nicht nur von einer Person, sondern von einer Gruppe von Experten erstellt werden. Außerdem sollten möglichst alle Propositionen identifiziert werden, die für den Inhaltsbereich der Concept Map als wichtig und für die Lernenden als relevant eingeschätzt werden (Ruiz-Primo, 2000). Es ist zudem fraglich, ob der Vergleich zwischen zwei Concept Maps ohne Weiteres durchgeführt werden kann, da in der Regel eine Proposition nicht nur auf eine Art und Weise ausgedrückt werden kann. So können Propositionen, die inhaltlich richtig sind, sich aber auf semantischer Ebene zu den Experten-Propositionen unterscheiden, als falsch bewertet werden (Hartmeyer et al., 2018). Dies ist auch der Grund, warum diese Art von Auswertung nur schwer zu automatisieren ist (Strautmane, 2012). Wenn man weitere automatische Concept-Map-Auswertungen anschaut, stellt man fest, dass diese in der Komplexität schwanken (Anohina & Grundspenkis, 2009). Ryoo und Linn (2016) entwickelten ein Web-based Inquiry Science Environment (WISE), bei dem Lernende eine Concept Map erstellen sollten, wie Pflanzen die Energie der Sonne aufnehmen und nutzen. Für die Auswertung dieser Concept Maps wurde im Vorfeld eine Reihe typischer Lernenden-Fehler gesammelt sowie Experten befragt. Das WISE-Tool beschränkte die Lernenden jedoch auf einen vorgefertigten Satz an Begriffen und Relationen, die die Lernenden für ihre Erstellung der Concept Maps nutzen konnten. Die automatische Auswertung ist demnach nur ein Abgleich mit den vordefinierten Propositionen und berücksichtigt keine eigenständigen Lösungen (Kroeze et al., 2021). Andere Arbeiten zur automatischen Auswertung von Concept Map wie das Artificial Intelligence-Based Student Learning Evaluation (AISLE) (G. P. Jain et al., 2014), der Reasonable Fallible Analyser (RFA) (Conlon, 2004) oder das Concept Map Assessment tool (COMPASS) (Gouli et al., 2004) nutzen eine Expertenmap zur Auswertung der angefertigten Concept Maps. Darauf aufbauend werden unterschiedliche Kennzahlen berechnet, die zur Auswertung

der Concept Maps genutzt werden. Neben den oben beschriebenen Problemen kritisieren Kroeze et al. (2021) ebenfalls diesen Ansatz: „*By reducing the structural and semantic quality of concept maps to a set of mathematical equations, these tools sacrifice some interpretability for increased portability. The use of a reference map also aims at portability, but assuming the reference map to be the absolute truth limits the ability of these tools to accurately assess students' knowledge.*“ (Kroeze et al., 2021, S. 1).
Insgesamt zeigt sich, dass viele (automatische) Ansätze zur Auswertung einer Concept Map eher einfache Bewertungsschema nutzen und im besten Fall die Validität der Begriffe und Propositionen durch den Vergleich mit einer Expertenmap bestimmt wird. Es kann schwierig sein, den Wissensstand einer Person zu beurteilen, wenn nur strukturelle oder quantitative Aspekte betrachtet und qualitative Aspekte außer Acht gelassen werden.

3.3 Concept Maps als formative Assessment-Methode

Concept Maps sind eine vielfältige Methode, welche in unterschiedlichen Varianten mit diversen Auswertungsansätzen eingesetzt werden können. Der Anwendungszweck in dieser Arbeit ist, eine Concept Map als Assessment-Methode einzusetzen.
Es gibt daneben noch eine Reihe von anderen Anwendungsgebieten wie die Unterrichtsplanung oder die Curriculumentwicklung (Stracke, 2004), welche im Folgenden aber nicht weiter thematisiert werden sollen.
In Kapitel 2 wurde beschrieben, dass bei formativen Assessments diagnostische Informationen über den Lernstand der Lernenden erfasst und auf Grundlage dessen weitere Schritte im Lehr-Lernprozess optimiert werden. Da Lernende beim Erstellen einer Concept Map die Zusammenhänge zwischen den Begriffen organisieren und visualisieren und so ihr Wissen über die gelernten Inhalte reflektieren müssen, können wichtige Informationen über Wissenslücken und Missverständnisse gesammelt werden (Novak & Cañas, 2006). So erhalten nicht nur Lernende wichtige Rückmeldungen, die sie für ihr weiteres Lernen nutzen können. Auch Lehrkräfte erhalten Einblick in den aktuellen Wissensstand der Lernenden, der z. B. für die Planung des weiteren Unterrichts genutzt werden kann (Anohina-Naumeca, 2015; Novak & Cañas, 2006). Angesichts dessen werden Concept Maps als eine gute formative Assessment-Methode angesehen (Buldu & Buldu, 2010; Hartmeyer

et al., 2018; Ruiz-Primo & Shavelson, 1996).
Durch Concept Maps können die Leistung der Lernenden ermittelt und so der Lernweg sichtbar gemacht werden (Hartmeyer et al., 2018). Zudem sind sie sehr flexibel einsetzbar, da sie auf dem Vorwissen der Lernenden aufbauen, die Lernmotivation steigern oder das reflektierende Denken fördern können (Buldu & Buldu, 2010). Denn je nach Concept Map stehen die Lernenden vor unterschiedlichen Herausforderungen wie dem Treffen einer geeigneten Auswahl von Begriffen oder sinnvollen Propositionen, die den aktuellen Inhalt oder Wissensstand am besten abbilden (Anohina-Naumeca, 2015).
Ein (formatives) Assessment mit Concept Maps lässt sich an unterschiedlichen Zeitpunkten im Unterricht platzieren. So kann zu Beginn des Unterrichts das Vorwissen der Lernenden erhoben und so eine Ausgangsbasis für die weitere Unterrichtsreihe geschaffen werden. Aber auch während der Unterrichtsreihe können über Concept Maps Veränderungen im Wissen der Lernenden sichtbar gemacht werden. Diese Informationen können anschließend für weitere Schritte im Lehr-Lernprozess genutzt werden (Ley, 2015; Stracke, 2004).
Hartmeyer et al. (2018) stellten in ihrem Review über Concept-Map-basiertes formatives Assessment fest, dass der Assessment-Prozess sich stark bezüglich der Formate unterscheidet. In den meisten Studien mussten die Lernenden eine Concept Map eigenständig erstellen und als Bewertungsgrundlage wurde eine Experten- bzw. Lehrkraft-Map genutzt. Yin et al. (2005) untersuchten in ihrer Studie zwei Concept-Map-Formate für die Verwendung als formative Assessment-Methode. Eine Gruppe sollte eine Concept Map mit vorgegebenen Relationen erstellen, wohingegen die andere Gruppe die Relationen eigenständig entwickeln sollten. Sie stellten fest, dass der etwas freiere Ansatz besser für ein formatives Assessment geeignet ist. Auch Vanides et al. (2005) stellten fest, dass eine *construct-a-map*-Methode eher für ein formatives Assessment genutzt werden soll als ein vorstrukturiertes *fill-in-a-map*-Format. Daraus lässt sich ableiten, dass Concept-Map-Formate, die weniger Einschränkung haben, besser für ein formatives Assessment geeignet sind, da sie den Lernenden mehr Freiraum bieten und so einen besseren Einblick in den Wissensstand geben können (Hartmeyer et al., 2018).
Allerdings sollte ein formativer Assessment-Prozess in den Unterricht einfach zu implementieren und zu nutzen sein, denn nach dem National Research Council (2000) sehen viele Lehrkräfte das formative Assessment als eine zusätzliche Arbeitsbelastung an (siehe auch Trumpower & Sarwar, 2010). Für ein formati-

ves Assessment mit Concept Maps könnte man deshalb einfach zu bestimmende Kennzahlen, die z. B. aus einer graphentheortischen Auswertung stammen, benutzen. Jedoch bietet diese Auswertung eher einen oberflächlichen Blick in den Wissensstand und liefert nicht genügend Informationen für ein formatives Assessment, weswegen dieser Ansatz eher für ein summatives Assessment geeignet ist (Trumpower & Sarwar, 2010).
Angesichts dessen weisen viele Studien auf Herausforderungen bei der Nutzung von Concept Maps als formatives Assessment hin (Buldu & Buldu, 2010; Hartmeyer et al., 2018). Betrachtet man nicht nur die oberflächlichen quantitativen Merkmale, sondern analysiert die einzelnen Lernenden-Concept-Maps genauer, kann dies zu einer erheblichen Mehrbelastung für die Lehrkräfte führen, was wiederum ein schnelles und effektives Feedback verhindern kann (Hartmeyer et al., 2018; Hwang et al., 2011). Eine zeitnahe, qualitativ hochwertige Auswertung von Concept Maps kann deshalb schwierig sein und so den Kriterien eines formativen Assessments widersprechen (Hartmeyer et al., 2018). Speziell offene Concept-Map-Formate werden als zeitintensiv in der Auswertung angesehen, vor allem wenn man wenig Erfahrung mit Concept Maps hat (Buldu & Buldu, 2010).

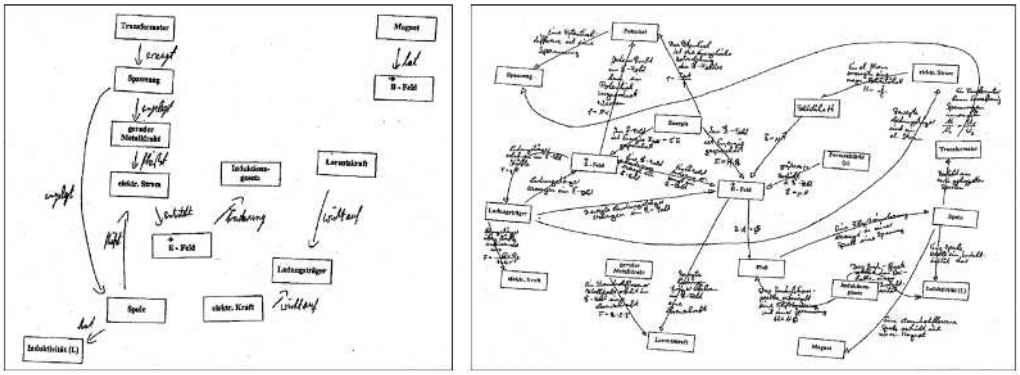


Abbildung 3.4: Zwei Concept Maps mit unterschiedlicher inhaltlicher Qualität aus Friege (2001)

Aus Abbildung 3.4 wird diese Problematik deutlich. Die Abbildung zeigt zwei qualitativ unterschiedliche Concept Maps aus der Arbeit von Friege (2001) und verdeutlicht die Vielfältigkeit von Concept Maps. Aus den Concept Maps wird nicht deutlich, an welcher Stelle eine qualitative Auswertung begonnen werden soll, da es bei einer fertigen Concept Map keinen Startpunkt gibt. Durch diese Eigenart wird eine schnelle Auswertung erschwert (Ley, 2015). Zudem zeigen die beiden Concept Maps auch, dass die Propositionen auf unterschiedliche Art und

Weise sowie inhaltlichen Niveaus gebildet werden können, was eine einfache und zeitökonomische Auswertung erschwert.
Zusammenfassend lässt sich sagen, dass Concept Maps das Potenzial für eine formative Assessment-Methode besitzen, da sie unter anderem Missverständnisse aufdecken und das Zusammenhangwissen darstellen können. Dagegen ist kritisch einzuwenden, dass die diagnostischen Informationen, die aus der Auswertung entstehen, einfach und zeitökonomisch gewonnen werden müssen. Trumpower und Sarwar (2010) warnen deshalb davor, dass ein benutzerunfreundliches formatives Assessment mit Concept Maps gar nicht erst eingesetzt wird und fordern daher, den gesamten Auswertungsprozess der Concept Map so weit wie möglich zu automatisieren.

4 Maschinelles Lernen

> „Künstliche Intelligenz wird unser Leben auf unglaubliche Weise verändern. Alle Bereiche unseres Zusammenlebens sind davon betroffen“ (Kitzmann, 2022, S. 1).

Unter künstlicher Intelligenz versteht man Computer, Maschinen oder Technologien, die so aufgebaut sind, dass sie wie Menschen lernen, denken und Handlungen ausführen können (Ertel, 2021). KI wird in einer Vielzahl von Branchen eingesetzt, da Vorhersagen und Prognosen aus großen Datenmengen erstellt werden können, Sprach- und Kommunikationsprozesse effizient gestaltet werden können oder Muster und Trends gefunden werden können (Ertel, 2021; Kitzmann, 2022). Dabei wird KI als ein Oberbegriff verstanden. Ein Teilgebiet von KI ist maschinelles Lernen oder auch Machine Learning (ML) genannt, bei dem Algorithmen eingesetzt werden, die von Daten lernen können. Klassische Aufgaben von ML ist die Klassifikation oder Regression (Pfannstiel, 2022; Richter, 2019). Ein Teilgebiet des maschinellen Lernens ist wiederum das sogenannte Deep Learning. Beim Deep Learning werden (komplexe) neuronale Netze mit künstlich erzeugten Neuronen eingesetzt. Die Struktur solcher neuronalen Netze ähnelt dem Aufbau des menschlichen Gehirns (Goodfellow et al., 2016).
Die Verfahren des maschinellen Lernens werden in drei Teilgebiete eingeordnet: überwachtes Lernen (supervised Learning), unüberwachtes Lernen (unsupervised Learning) und bestärkendes Lernen (reinforcement Learning) (Richter, 2019). Beim überwachten Lernen soll ein Modell die Beziehung zwischen einer Eingabe X und einer Ausgabe Y lernen. So soll ein Modell trainiert werden, das in der Lage ist, eine neue unbekannte Eingabe selbstständig einem Label zuzuweisen (Hirschle, 2022). Da diese Labels im Vorfeld feststehen müssen und meist von Menschen erzeugt werden, nennt sich dieses Verfahren überwachtes Lernen (Mahesh, 2019). Überwachtes Lernen wird bei einer Reihe von Problemen genutzt wie bei der automatischen Klassifizierung von E-Mails (Eingabe X) in Spam oder nicht Spam (Ausgabe Y) oder bei der Vorhersage des Aktienpreises (X Aktieninformationen, Y Preis in sechs Monaten) (Richter, 2019).

Im Gegensatz dazu ist beim unüberwachten Lernen nur die Eingabe X bekannt. Das Ziel bei solchen Ansätzen ist es, Muster in den Daten zu finden (Richter, 2019). Ein typisches Verfahren des unüberwachten Lernens sind Clusteranalysen (Plaue, 2021).
Bestärkendes Lernen wird als drittes Teilgebiet des maschinellen Lernens angesehen. In diesem Fall lernt das Modell selbstständig eine Strategie und optimiert diese durch Interaktion mit der Umwelt (Richter, 2019). Typische Beispiele aus diesem Bereich des bestärkenden Lernens sind Modelle, die eine gute Spielstrategie beim Schach finden oder den schnellsten Weg durch ein Labyrinth ermitteln können (Richter, 2019).
In dieser Arbeit soll ein Machine-Learning-Modell entwickelt werden, das automatisch Antworten von Lernenden in eine vorher definierte Feedback-Kategorie zuweist. Dieser Ansatz entspricht einem Klassifikationsproblem (Eingabe: Lernenden-Antworten, Ausgabe: Feedback-Kategorie), das zu den überwachten maschinellen Lernverfahren zählt. Da die Lernenden-Antworten in natürlicher Sprache (Textform) vorliegen, spricht man auch von natürlicher Sprachverarbeitung oder auch Natural Language Processing (NLP) genannt. NLP ist ein spezifisches Anwendungsgebiet des maschinellen Lernens, das sich auf die Verarbeitung und Analyse von natürlicher Sprache konzentriert (Hirschle, 2022). Angesichts dessen wird das folgende Kapitel sich speziell mit diesem Teil des überwachten maschinellen Lernens auseinandersetzen.

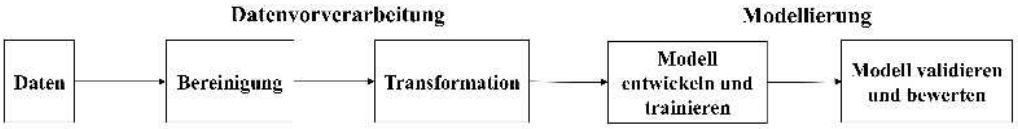


Abbildung 4.1: Arbeitsablauf eines Machine-Learning-Modells (Thevapalan, 2021)

Das folgende Kapitel wird nach einem klassischen Arbeitsablauf eines Machine-Learning-Modells aufgebaut (siehe Abbildung 4.1). Während des Trainings lernt das Modell die Beziehungen zwischen den Eingangsdaten und den entsprechenden Ausgaben. Der zugrunde liegende Algorithmus legt den Rahmen fest, wie das Modell trainiert wird und wie es die Beziehungen zwischen den Daten lernt. Das trainierte Machine-Learning-Modell repräsentiert die gelernten Beziehungen zwischen den Eingangsdaten und den entsprechenden Ausgaben und verfügt über eine große Anzahl an Parametern, die während des Trainings angepasst werden. Das Modell kann dann verwendet werden, um Vorhersagen für neue Daten zu treffen (siehe Abbildung 4.1). Daher wird im folgenden Kapitel zunächst

auf die Datenvorverarbeitung eingegangen und anschließend die Modellierung betrachtet. Abschließend werden in diesem Kapitel aktuelle Arbeiten aus dem Bildungsbereich zu diesem Thema diskutiert und eingeordnet.

4.1 Datenvorverarbeitung

Bevor ein Machine-Learning-Modell entwickelt werden kann, müssen die Daten vorverarbeitet werden, denn die Qualität der Daten hat direkten Einfluss auf die Vorhersagekraft des Modells (Sghir et al., 2022). Nach Brownlee (2020) ist dies ein wichtiger Schritt in jedem Machine-Learning-Projekt, der später über Erfolg oder Misserfolg entscheiden kann.
Die Qualität der Daten kann über unterschiedliche Eigenschaften wie Vollständigkeit, Eindeutigkeit, Richtigkeit oder Aktualität überprüft werden (Han et al., 2012; Plaue, 2021). Um diese Kriterien zu erfüllen, muss der verwendete Datensatz überprüft und gegebenenfalls bearbeitet werden. Die Datenvorverarbeitung versucht also, aus den Rohdaten einen qualitativ hochwertigen Datensatz zu erzeugen (Plaue, 2021). Dazu können Datenfehler wie fehlende Werte ergänzt oder gelöscht oder bestimmte Daten wie Ausreißer oder doppelte Einträge identifiziert werden (Han et al., 2012). Je nach Anwendungsfall und Datenstrukturen können so unterschiedliche Methoden zur Steigerung der Datenqualität angewendet werden.

Aus dem NLP-Bereich sind klassische Schritte bei der Datenvorverarbeitung die Entfernung von Sonderzeichen wie Ausrufe- oder Fragezeichen oder das Entfernen von sogenannten Stoppwörtern (Chai, 2023). Stoppwörter sind Wörter wie bestimmte oder unbestimmte Artikel oder Konjunktionen, die einen Text nicht von einem anderen Text unterscheiden (Ferilli et al., 2014). Durch das Entfernen von Stoppwörtern wird die Gesamtanzahl an Wörtern reduziert, jedoch die Zahl an eindeutigen Wörtern beibehalten, was zu besserer Leistung für maschinelles Lernen führen kann (Armano et al., 2015; Wallach, 2006; Zaman et al., 2011). Trotzdem muss abgewogen werden, welche und wie viele Wörter aus dem Korpus entfernt werden, da auch Stoppwörter einen bedeutungsvollen Inhalt und je nach Anwendung eine entscheidende Bedeutung haben können: „*It is inappropriate to say that stopwords are meaningless, just because their content does not differentiate text documents in the corpus*“ (Chai, 2023, S. 523).
Ein weiterer häufig genutzter Ansatz zur Datenvorverarbeitung ist die Normalisierung von Wörtern auf ihre Grundform (Stemming oder Lemmatisierung) (Manning

et al., 2012). So werden die Wörter *genießt* und *genossen* auf dieselbe Grundform *genießen* zurückgeführt, was den Wortschatz verkleinert (Chai, 2023). Durch diesen Normalisierungsschritt konnten ebenfalls Verbesserungen im maschinellen Lernen erzielt werden (Rajput & Khare, 2015; Torres-Moreno, 2014). Aber auch bei diesem Schritt muss eine sorgfältige Abwägung durchgeführt werden, da durch Normalisierungen die inhaltliche Bedeutung von einem Wort verloren gehen kann (Bao et al., 2014; Chai, 2023).
Unabhängig von der Wahl und dem Umfang der Datenvorverarbeitung müssen die Daten transformiert werden, denn ein Machine-Learning-Modell kann nur Daten verarbeiten, die in einer numerischen Form vorliegen. Deshalb müssen z. B. Text- oder Bilddaten in eine numerische Repräsentation transformiert werden, bevor das Machine-Learning-Modell diese interpretieren kann (Sghir et al., 2022) (siehe Abbildung 4.1).
Wörter haben eine eigenständige Bedeutung, doch stehen sie auch in semantischer und grammatikalischer Relationen zu anderen Wörtern (Hirschle, 2022). Daher gibt es verschiedene Verfahren zur Transformation von Texten, die unterschiedliche Aspekte betrachten. Diese Verfahren erstellen Abbildungen von Wörtern in numerische Vektoren aus reellen Zahlen, weswegen man von Textvektorisierung oder Embeddings spricht. Die dadurch entstehenden Vektoren können dann für das Klassifizieren oder Clustern von Texten genutzt werden (Raghav, 2019).
In der Literatur gibt es eine Vielzahl von Ansätzen, die zur Vektorisierung von Texten genutzt werden. Um die verschiedenen Verfahren einzuordnen, sollen ausgewählte und häufig genutzte Verfahren genauer vorgestellt werden. Durch die Beschreibung dieser Verfahren können die unterschiedlichen Möglichkeiten der Textvektorisierung gut dargestellt werden.

4.1.1 Traditionelle Verfahren

Bag-of-Words

Ein Beispiel für ein traditionelles Verfahren ist das *Bag-of-Word*-Verfahren (BoW). Bei dem BoW-Verfahren werden Dokumente[2] als eine Liste von Wörtern, unabhängig von der genauen Reihenfolge oder Grammatik, betrachtet. Zunächst werden daher alle Wörter, die in den gesamten Dokumenten vorkommen, notiert und eine

[2] Im Kontext der Textvektorisierung bezieht sich der Begriff „Dokument“ in der Regel auf eine Sammlung von Textdaten, auf die das Verfahren angewendet wird. Bei einem Dokument kann es sich um einen Satz, einen Absatz, einen Artikel, eine Website oder ein Buch handeln.

lange Lister der Wörter, die als Vokabular bezeichnet wird, gebildet (Hirschle, 2022). Betrachtet man die drei Sätze „*Die Sonne scheint am blauen Himmel*“, „*Es regnet draußen*“ und „*Die Sonne scheint draußen*“ sieht das Vokabular wie folgt aus: [am, blauen, die, draußen, es, Himmel, regnet, scheint, Sonne].

	am	blauen	die	draußen	es	Himmel	regnet	scheint	Sonne
Satz 1	1	1	1	0	0	1	0	1	1
Satz 2	0	0	0	1	1	0	1	0	0
Satz 3	0	0	1	1	0	0	0	1	1

Tabelle 4.1: Vektorisierung durch das *Bag-of-Word*-Verfahren für die drei Sätze: „*Die Sonne scheint am blauen Himmel*“, „*Es regnet draußen*“ und „*Die Sonne scheint draußen*“

Aus dieser Wortliste kann nun die Dokumenten-Wort-Matrix gebildet werden, bei der jedes Wort eine Spalte und jeder Satz eine Zeile bildet. Wenn ein Wort in einem Satz auftritt, wird dies mit einer 1, ansonsten mit einer 0 codiert (siehe Tabelle 4.1). Das BoW-Verfahren transformiert demnach jeden Satz in einen Vektor mit fester Länge, die das Vorhandensein eines bestimmten Wortes beschreibt (Birunda & Devi, 2021).

Term frequency–inverse document frequency

Ein anderer Ansatz, der auf dem BoW-Verfahren aufbaut, ist das *Term-frequency –inverse–document–frequency*-Verfahren (tf-idf), das zu den am häufigsten verwendeten Verfahren zählt (Aizawa, 2003). Anders als beim BoW-Modell wird beim *tf-idf*-Verfahren nicht mehr die reine Worthäufigkeit betrachtet, sondern die Wörter werden in Relation zu ihrer Häufigkeit im Gesamtkorpus gewichtet (Hirschle, 2022). Dazu werden zwei Hauptkomponenten berechnet: Die Termhäufigkeit $tf(ij)$ gibt an, wie häufig ein Wort i im Dokument j vorkommt. Um die Dokumentenlänge zu berücksichtigen, wird der Faktor der Termfrequenz oftmals normalisiert und durch die Gesamtzahl der Wörter in dem Dokument j geteilt (Aizawa, 2003; Trstenjak et al., 2014):

$$tf(i,j) = \frac{\text{Anzahl Wort i im Dokument j}}{\text{Gesamtanzahl an Wörtern (w) im Dokument j}} = \frac{f_j(i)}{\max_{w \in j} f_j(w)}$$

Durch die inverse Dokumentenhäufigkeit $idf(i,D)$ wird die Spezifität des Worts i im Gesamtkorpus D bestimmt. Die inverse Dokumentenhäufigkeit misst die Bedeu-

	am	blauen	die	draußen	es	Himmel	regnet	scheint	Sonne
Satz 1	0,18	0,07	0,18	0	0	0,18	0	0,07	0,07
Satz 2	0	0	0	0,14	0,33	0	0,33	0	0
Satz 3	0	0	0,10	0,10	0	0	0	0,10	0,10

Tabelle 4.2: Vektorisierung durch das *tf-idf*-Verfahren für die drei Sätze „*Die Sonne scheint am blauen Himmel*“, „*Es regnet draußen*“ und „*Die Sonne scheint draußen*“

tung des Wortes i anhand der Verwendung in der Gesamtheit aller Dokumente D und spiegelt so die Seltenheit oder Spezifität eines Wortes wider (Hirschle, 2022). So werden vielfach auftretende Wörter einer geringen Bedeutung zugewiesen als Wörter, die nur vereinzelt in den Dokumenten vorkommen. So sollen relevante von irrelevanten Dokumenten unterschieden werden (Hirschle, 2022). Zur Berechnung wird der Logarithmus von dem Verhältnis der Gesamtzahl der Dokumente D zur Anzahl der Dokumente, die das Wort i enthalten, bestimmt:

$$\begin{aligned} idf(i,D) &= \log\left(\frac{\text{Gesamtanzahl der Dokumente}}{\text{Anzahl der Dokumente, in dem das Wort i auftritt}}\right) \\ &= \log\left(\frac{D}{d \in D : i \in D}\right) \end{aligned}$$

Aus den beiden Komponenten kann dann das *tf-idf*-Verfahren berechnet werden:

$$\text{tf-idf}(i,j,D) = tf(i,j) \cdot idf(i,D)$$

Überträgt man das *tf-idf*-Verfahren auf die drei Beispielsätze, erhält man eine neue Dokumenten-Wort-Matrix, die nicht mehr aus Einsen oder Nullen besteht, sondern aus Gewichten (siehe Tabelle 4.2). So werden nicht nur die einfachen Häufigkeiten wie beim *BoW*-Verfahren betrachtet, sondern auch die Seltenheit der Wörter in der gesamten Dokumentensammlung bestimmt. Dies macht das *tf-idf*-Verfahren zu einer verbesserten Möglichkeit, eine numerische Repräsentation von Texten zu erzeugen.

Die traditionellen Verfahren *BoW* und *tf-idf* bieten eine einfache Möglichkeit, Texte in eine numerische Repräsentation in Form von Vektoren und Matrizen umzuwandeln. Zudem kann die numerische Repräsentation nach der Berechnung immer noch nachvollzogen werden. Abhängig von der Größe des gesamten Datensatzes können jedoch große und spärlich besetzte Matrizen entstehen. Für jedes Wort muss ein eigenes Gewicht berechnet werden, was das Trainieren eines Mo-

dells verlangsamen kann (Hirschle, 2022). Hier zeigt sich, dass das Vorverarbeiten der Daten helfen kann, die Qualität und die Performance des Modells zu verbessern. Auch Vorverarbeitungsschritte wie das Löschen aller Stoppwörter oder die Umwandlung in Kleinbuchstaben, können helfen, das Vokabular zu verkleinern (Hirschle, 2022). Zudem enthalten die einzelnen Vektoren keinerlei semantische Informationen oder Informationen über den Kontext des Textes, da nur reine statistische Metriken betrachtet werden. Auch die Reihenfolge der Wörter wird nicht berücksichtigt, die je nach Kontext unterschiedlich bedeutsam ist, was zu einem Informationsverlust führen kann (Aizawa, 2003).

4.1.2 Statische Embeddings

Bei statischen Embeddings werden die Bedeutung von einzelnen Wörtern und deren Beziehungen zu anderen Wörtern innerhalb eines Satzes berücksichtigt. Dazu können sogenannte flache neuronale Netze genutzt werden (Abubakar & Umar, 2022; Hirschle, 2022). Flache neuronale Netze sind neuronale Netze mit einem relativen einfachen Aufbau. Sie bestehen nur aus drei Schichten: Eingabe-, Zwischen- und Ausgabeschicht (Werner, 2021). Diese flachen neuronalen Netze können mit einem großen Datensatz trainiert werden. Während des Trainingsprozesses wird jedes Wort in einen mehrdimensionalen Vektorraum projiziert. Die daraus entstehenden Wortvektoren oder auch Embeddings genannt berücksichtigen zwar den Kontext eines Wortes während des Trainingsprozesses des neuronalen Netzes, jedoch sind die Vektoren nach dem Training statisch. Das bedeutet, dass die Wortvektoren unabhängig vom späteren Anwendungskontext identisch bleiben (Birunda & Devi, 2021). Im Vergleich zum *BoW*-Verfahren werden die Wörter

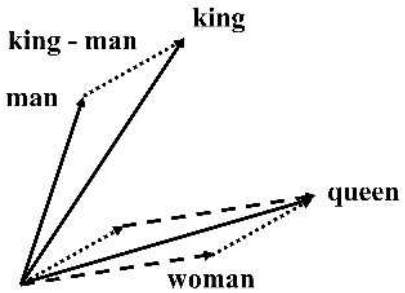


Abbildung 4.2: Darstellung algebraischer Operationen von Wortvektoren (Liang et al., 2022, S. 3)

durch eine bestimmte Anzahl von Parametern repräsentiert, die man als semantische Dimension verstehen kann (Hirschle, 2022). Das hat den Vorteil, dass man

nicht nur die semantische Information der Wörter abbilden kann, sondern über die vektorisierte Darstellung auch algebraische Operationen durchführen kann. Dadurch lassen sich semantische Beziehungen zwischen Wörtern darstellen und für spätere Aufgaben verwenden (Hirschle, 2022).
So konnten Mikolov, Yih und Zweig (2013) zeigen, dass die Rechnung der Wortvektoren $\overrightarrow{König} - \overrightarrow{Mann} + \overrightarrow{Frau}$ den Wortvektor $\overrightarrow{Königin}$ ergibt (siehe Abbildung 4.2).
Auch Analogien wie „Frankreich ist zu Paris, wie Deutschland zu X?“ konnten erfolgreich gelöst werden (Mikolov, Chen et al., 2013). Dies resultiert aus der Tatsache, dass ähnliche Wörter im Vektorraum näher zusammenliegen (Mikolov, Chen et al., 2013). Dabei kann die Ähnlichkeit zweier Vektoren $\vec{a}$ und $\vec{b}$ über den aufgespannten Winkel oder die sogenannte Kosinus-Ähnlichkeit bestimmt werden:

$$\text{Kosinus-Ähnlichkeit} = \cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|a\| \cdot \|b\|} = \frac{\vec{a} \cdot \vec{b}}{\sqrt{a^2 \cdot b^2}}$$

Eines der bekanntesten statischen Verfahren ist *Word2vec* von Google (Mikolov, Chen et al., 2013). Bei *Word2vec* gibt es zwei primäre Modelle, *continuous bag-of-words* (CBOW) und *Skip-Gram*. Dabei bestehen beide Modelle aus neuronalen Netzen, die jeweils drei Schichten haben (siehe Abbildung 4.3).

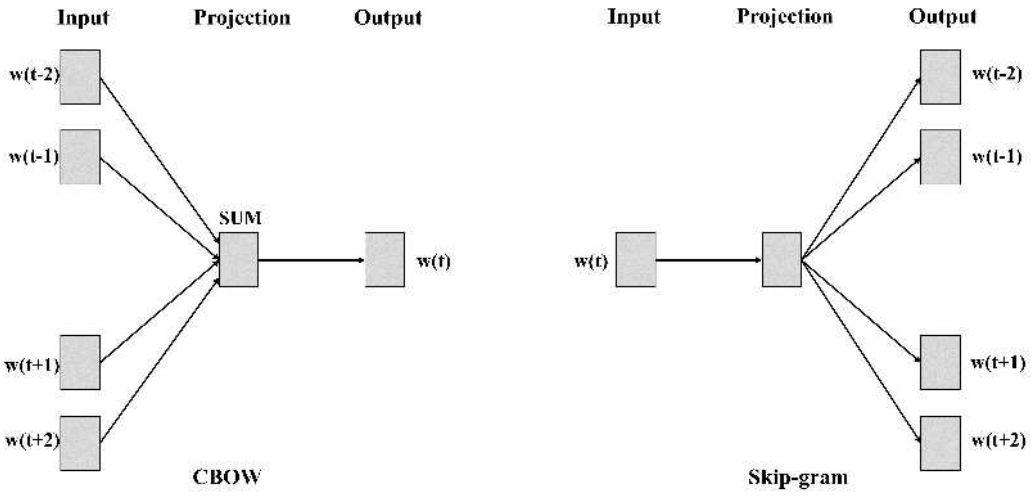


Abbildung 4.3: Architektur der CBOW und Skip-gram Modell (Mikolov, Chen et al., 2013, S. 5)

Beim CBOW-Modell wird das Zielwort anhand der danebenliegenden Wörter vorhergesagt (siehe Abbildung 4.4). Aus den sogenannten Kontextwörtern wird ein gemittelter Wortvektor berechnet, der zur Vorhersage des Zielwortes genutzt wird

(Data Basecamp, 2023b; Mikolov, Chen et al., 2013). Dabei spielt die Reihenfolge der Kontextwörter jedoch keine Rolle und wird bei der Berechnung vernachlässigt (Lilleberg et al., 2015).

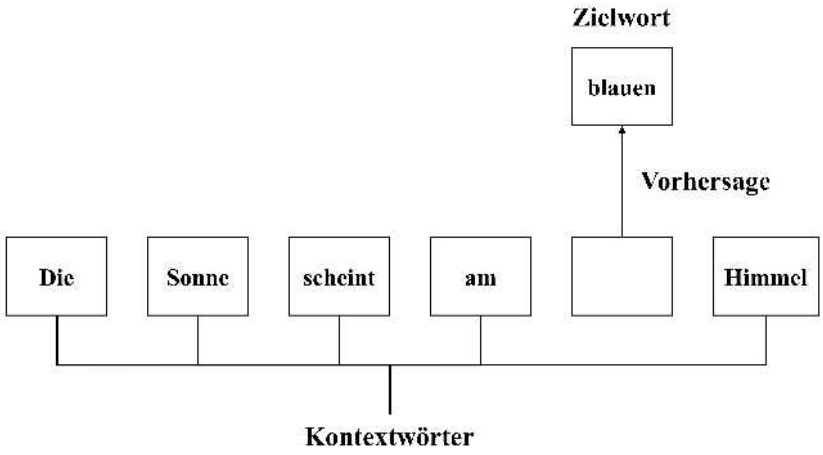


Abbildung 4.4: Beispiel eines CBOW-Modells

Im Gegensatz dazu werden beim *Skip-gram*-Modell die Kontextwörter auf Grundlage des zentralen Wortes vorhergesagt (siehe Abbildung 4.3). Es ist sozusagen eine Umkehrung des *CBOW*-Ansatzes (Mikolov, Chen et al., 2013).
Neben der *Word2vec*-Modellen existieren noch weitere Modelle wie das *Fasttext*-Modell von Facebook (Bojanowski et al., 2016) oder das *Glove*-Modell von der Stanford Universität (Pennington et al., 2014), die jedoch einen ähnlichen Aufbau haben und deshalb nicht weiter beschrieben werden.
Um solche Modelle zu trainieren, wird ein großer Textkorpus benötigt. Dabei wird angenommen, dass sich die Bedeutung der Wörter aus dem Kontext der Sätze innerhalb des Korpus erschließen lässt. Da der Rechenaufwand für das Training sehr groß sein kann, kann man auf vortrainierte Modelle zurückgreifen, die bereits auf einem großen Korpus, z. B. Wikipedia, trainiert wurden (Hirschle, 2022).
Da beim *Word2vec*-Ansatz die Relevanz der einzelnen Wörter nicht betrachtet wird, zeigten auch kombinierte Ansätze aus *tf-idf* und *Word2vec* gute Ergebnisse bei der Klassifikation von Texten (Lilleberg et al., 2015).

4.1.3 Kontextabhängige Embeddings

Die Beschreibung der statischen Embeddings hat gezeigt, dass jedes Wort durch einen bestimmten Vektor dargestellt werden kann. Dies kann jedoch bei Wörtern mit mehreren Bedeutungen schnell zu Problemen führen, da polyseme Wörter

dieselbe numerische Darstellung haben (Ethayarajh, 2019). Angesichts dieser Problematik wurden neue Verfahren entwickelt, bei denen kontextabhängige Wortvektoren erstellt werden. Ein bekanntes Beispiel für dieses Verfahren ist das von Google entwickelte BERT-Modell (BERT steht für Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018).
Kontextabhängige Embeddings werden mittels transformatorbasierte Sprachmodelle erstellt, die den Kontext sowohl während des Trainingsprozesses als auch bei der Anwendung berücksichtigen können (Ethayarajh, 2019). Transformatorbasierte Sprachmodelle bauen auf der Encoder-Decoder-Architektur auf (Vaswani et al., 2017). Dabei handelt es sich um zwei verschiedene neuronale Netze, die gemeinsam oder unabhängig voneinander trainiert und genutzt werden können (Hirschle, 2022). Das BERT-Modell ist ein Beispiel für ein Encoder-Modell und die GPT-Modellreihe (GPT steht für Generative Pretrained Transformer) von OpenAI (z. B. Brown et al., 2020) ein Beispiel für ein Decoder-Modell (Prince, 2023).
Encoder und Decoder haben dabei verschiedene Aufgaben. Encoder werden genutzt, um kontextualisierte Embeddings zu erstellen, die nicht nur die Bedeutung des einzelnen Wortes, sondern auch den Kontext berücksichtigen (Ethayarajh, 2019). Daher wird im Gegensatz zu den statischen Embeddings das Wort *Maus* im Satz „Die Maus läuft schnell über die Wiese“ anders eingebettet sein als im Satz „Der Wissenschaftler benutzt eine Maus, um seinen Computer zu steuern“. Die kontextualisierten Embeddings, die z. B. von einem BERT-Modell erzeugt werden, können dann für verschiedene Aufgaben wie Textklassifikation verwendet werden (Heidloff, 2023).
Der Decoder hat das Hauptziel, Text basierend auf einer gegebenen Eingabe oder einem gegebenen Kontext zu generieren. Der Prozess der Textgenerierung erfolgt schrittweise, wobei der Decoder bei jedem Schritt ein neues Wort basierend auf den bisherigen vorhergesagten Wörtern generiert. Das Ergebnis ist eine kontinuierliche Generierung von Text, die auf dem gegebenen Eingabekontext basiert (Heidloff, 2023; Prince, 2023). Diese autoregressiven Schritte werden so lange fortgeführt, bis das Ende des Inputs erreicht und der Output vollständig erstellt wurde (Platen, 2020).
Die Übersetzung zwischen zwei Sprachen ist ein Beispiel für ein Encoder-Decoder-Modell (siehe Abbildung 4.5). Der Encoder wird benutzt, um den zu übersetzenden englischen Satz „*I want to buy a car*“ durch eine Reihe von Transformationsblöcken in kontextualisierte Embeddings zu transformieren.

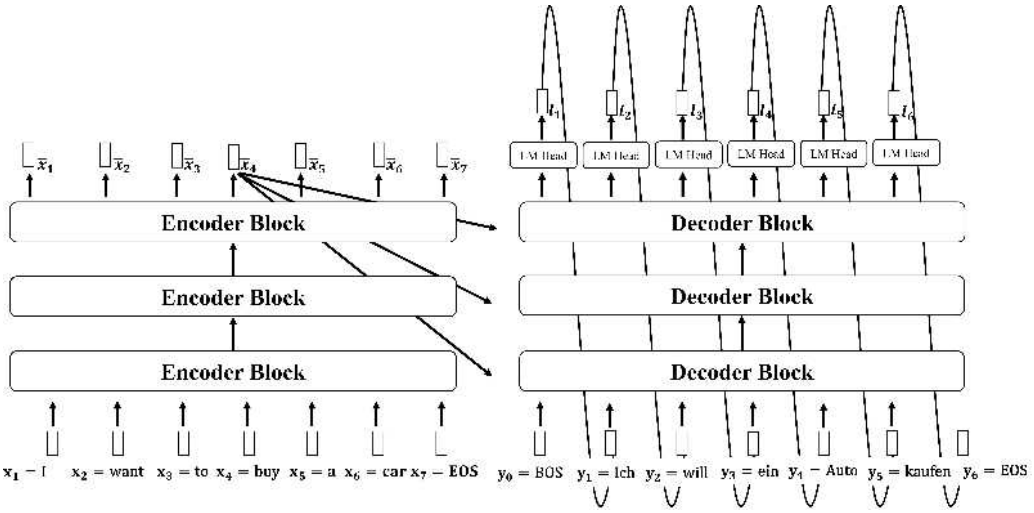


Abbildung 4.5: Vereinfachter Aufbau eines Transformers (Platen, 2020)

Der Decoder nutzt diese Embeddings, um den übersetzten deutschen Satz „*Ich will ein Auto kaufen*“ vorherzusagen. Die Decoderschichten beachten die Ausgabe des Encoders, weswegen jedes deutsche Ausgabewort von den vorherigen Ausgabewörtern und dem englischen Ausgangssatz abhängig ist (siehe Abbildung 4.5) (Prince, 2023, S. 227).
Dieser beschriebene Ablauf und der skizzierte Aufbau eines Encoder-Decoder-Modells in Abbildung 4.5 ist in der Realität deutlich umfangreicher, da die neuronalen Netze aus einer Vielzahl von unterschiedlichen Schichten und Parametern zusammengesetzt sind. So besteht z. B. das Modell GPT-3 von OpenAI aus 175 Milliarden Parametern, die trainiert werden können (Brown et al., 2020). Daher kann man vortrainierte Modelle, die man ohne weiteres Training nutzen kann, verwenden. Einige Modelle wie Variationen der BERT-Modelle sind als Open Source verfügbar und können z. B. auf der Plattform Hugging Face genutzt und heruntergeladen werden (T. Wolf et al., 2019). Bereits trainierte Modelle können trotzdem noch auf einen bestimmten Bereich oder spezifischen Anwendungsfall angepasst werden (Heidloff, 2023).

Der Austausch von statischen zu kontextabhängigen Verfahren hat zu erheblichen Verbesserungen bei einer Vielzahl von Anwendungen des maschinellen Lernens geführt (Ethayarajh, 2019). Die Modelle können gewissermaßen universal eingesetzt und müssen nicht für einen bestimmten Kontext trainiert werden (Hirschle, 2022). Allerdings benötigt man für das Trainieren solcher Modelle einen großen Satz von Daten, was die Reproduktion erschweren kann (Hirschle, 2022)

Außerdem muss darauf geachtet werden, dass sowohl bei den kontextabhängigen als auch bei statischen Verfahren die Trainingsdaten, mit denen die Modelle trainiert werden, ausgewogen und nicht verzerrt sind, da sonst z. B. Vorurteile gestärkt werden können (Uttamchandani & Quick, 2022). Die Trainingsdaten bestehen in der Regel aus Texten, die von Menschen erstellt wurden, weswegen es möglich sein kann, dass der Datensatz rassistische oder geschlechterspezifische Stereotypen enthält (Bender et al., 2021). Dies kann dazu führen, dass problematische Vorhersagen getroffen werden können (Caliskan et al., 2017). So zeigten Bolukbasi et al. (2016), dass Modelle, welche *Word2vec* nutzen, die Analogie „Vater ist zu Arzt, wie Mutter zu X“ mit dem Wort *Krankenschwester* lösten, was ein Beispiel für eine Verzerrung bezüglich der Geschlechter darstellt.
Zudem haben solche Modelle Probleme mit Wörtern, die nicht im Trainingsdatensatz enthalten sind, was zu einem Informationsverlust führen kann (Data Basecamp, 2023b). Auch durch die komplexe Architektur kann eine transparente Evaluation der Modelle herausfordernd sein (Hirschle, 2022). Außerdem benötigt man für das Trainieren solcher Modelle enorme Rechenleistungen, was Auswirkungen auf die Umwelt haben könnte: „*Training a single BERT base model (without hyperparameter tuning) on GPUs was estimated to require as much energy as a trans-American flight*“ (Bender et al., 2021, S. 612).

4.2 Trainieren eines Modells

Beim überwachten maschinellen Lernen muss das Modell die Beziehung zwischen einer Eingabe und der Zielvariable effektiv erfassen. Dafür wird ein Modell auf Grundlage von Daten trainiert. In der Regel werden die Daten in verschiedene Datensätze aufgeteilt, sodass ein Machine-Learning-Modell auf Trainingsdaten trainiert und auf einem Testdatensatz getestet werden kann. In diesem Abschnitt wird deshalb die Datenpartitionierung erläutert und es wird ein kurzer Überblick über Machine-Learning-Modelle zur Klassifikation gegeben.

4.2.1 Aufteilung der Daten

Ohne Aufteilung der Daten werden zum Trainieren eines Modells wie auch zum Testen des Modells dieselben Daten benutzt. Bei dieser Selbstvalidierung wird also keine Aufteilung durchgeführt, was den Prozess sehr praktikabel macht. Allerdings ist es schwierig festzustellen, wie das trainierte Modell sich bei neuen

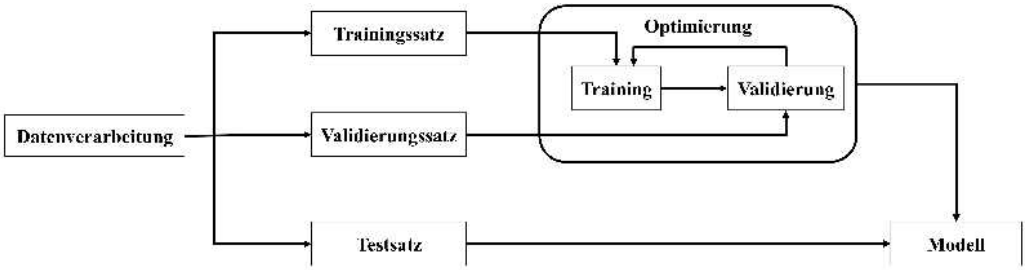


Abbildung 4.6: Aufteilung der Daten in drei distinkte Datensätze

ungesehenen Daten verhält (Zhai et al., 2020). Es besteht die Gefahr, dass das Modell die Zusammenhänge nicht lernt, sondern nur reproduziert und so unbrauchbar für die Praxis wird, da stets ungesehene Fälle auftreten werden. Das Ziel eines Trainingsprozesses ist es, ein Modell zu finden, was neue ungesehene Daten gut vorhersagen kann (Burkov, 2020).
Angesichts dessen teilt man die Daten in verschiedene Teildatensätze auf. Eine häufig gewählte Strategie ist die Aufteilung in drei distinkte Datensätze: Trainings-, Validierungs- und Testdatensatz (Burkov, 2020) (siehe Abbildung 4.6).
Der Trainingsdatensatz ist der größte der drei Datensätze und wird zum Trainieren des Modells genutzt. Durch diese Daten soll das Modell die Zusammenhänge in den Daten lernen und so in der Lage sein, neue unbekannte Daten vorherzusagen (Burkov, 2020). Die beiden anderen Datensätze werden verwendet, um statistische Kennzahlen zu berechnen, die die Performance des Modells widerspiegeln. Der Validierungsdatensatz wird benutzt, um die optimalen Parameter des Modells während des Trainingsprozesses zu finden. Dieser Schritt wird Hyperparameter-Optimierung genannt und in Abschnitt 4.3.2 aufgegriffen. Der Testdatensatz soll die neuen unbekannten Daten simulieren und so die Performance des Modells nach Abschluss des Trainings inkl. Optimierung messen (Burkov, 2020). Eine genaue Betrachtung der gängigen statistischen Kennzahlen, die zur Optimierung und zur Messung der Performance genutzt werden, wird in Abschnitt 4.3.1 dargestellt.
Durch die Aufteilung in verschiedene Datensätze wird die Aussagekraft der Leistung und die Generalisierbarkeit des Modells, im Gegensatz zur Selbstvalidierung, deutlich erhöht (Zhai et al., 2020). Um valide Aussagen über die Performance des Modells treffen zu können, müssen die Validierungs- und Testdaten ähnliche Merkmalsverteilungen aufweisen und möglichst identisch zu den erwarteten neuen Daten sein (Burkov, 2020). Für die genaue Verteilung der Daten in die drei Teildatensätze gibt es keine festen Vorgaben. Oftmals wird ein Ansatz gewählt, der 80 % der Daten in einen Trainingsdatensatz und jeweils 10 % der Daten in die

anderen beiden Sätzen aufteilt (Burkov, 2020).
Die gängigste Methode zur Aufteilung der Daten ist die Kreuzvalidierungsstrategie (Ertel, 2021; Zhai et al., 2020). Bei dieser Methode werden die Daten in mehrere gleich große Blöcke aufgeteilt und zum Trainieren und Testen des Modells genutzt. In Abbildung 4.7 ist eine Kreuzvalidierung mit fünf Blöcken dargestellt, also

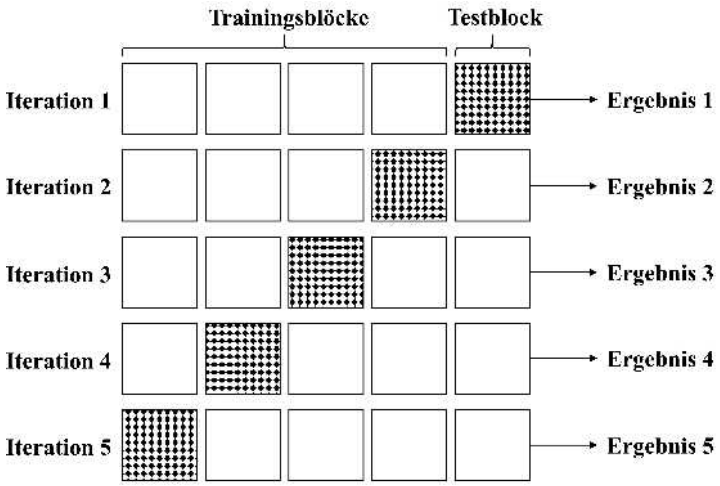


Abbildung 4.7: Beispiel einer Kreuzvalidierung mit 5 Blöcken

eine Verteilung von 80 % Trainings- und 20 % Testdaten. Das Modell wird dabei fünfmal trainiert und evaluiert, wobei immer vier Blöcke zum Trainieren und ein Block zum Testen genutzt werden. Daraus ergeben sich fünf einzelne Ergebnisse, die anschließend gemittelt werden und eine Aussage über die Performance des Modells liefern (Ertel, 2021). Der Vorteil einer Kreuzvalidierung ist, dass der gesamte Datensatz sowohl zum Trainieren als auch zum Testen benutzt wird (Pedregosa et al., 2011). Dieser Ansatz bietet sich primär bei kleineren Datensätzen an und kann sowohl bei der Aufteilung in Trainings- und Testdaten als auch zur Aufteilung in Trainings- und Validierungsdaten verwendet werden (Burkov, 2020). Für die Aufteilung der Daten in die verschiedenen Blöcke kann man verschiedene Vorgehensweisen wählen. Betrachtet man einen Datensatz, der Daten aus unterschiedlichen Klassen enthält, die das Machine-Learning-Modell später richtig vorhersagen soll, ist die einfachste Variante in Abbildung 4.7 dargestellt. Bei dieser *k-fold*-Variante werden die Daten abhängig von der Reihenfolge, aber unabhängig von der Klassenzugehörigkeit in möglichst gleich große *k*-Blöcke aufgeteilt (Pedregosa et al., 2011). Eine andere Möglichkeit ist der *Shuffle Split*, bei dem die Reihenfolge der Daten in die verschiedenen Blöcke zufällig verteilt wird. Es werden zwar weiterhin nicht die Klassen berücksichtigt, jedoch bietet

dieses Verfahren eine bessere Verteilung der Daten innerhalb der Trainings- und Testblöcke (Pedregosa et al., 2011). Wenn die eine Klasse im Datensatz sehr stark dominiert, die Daten also ungleich verteilt sind, bietet sich die *Stratified-Shuffle-Split*-Variante an. Bei diesem Ansatz werden die Daten zufällig in die Blöcke aufgeteilt und es wird darauf geachtet, dass in den Trainings- und Testblöcken die Verteilung der unterschiedlichen Klassen gleich verteilt ist.

Es existiert noch eine Vielzahl weiterer Möglichkeiten, die Daten in Trainings-, Validierungs- und Testdaten aufzuteilen. Die Wahl einer geeigneten Strategie hängt letztlich vom vorliegenden Datensatz und dem angestrebten Ziel ab. Zhai et al. (2020) fanden heraus, dass in der naturwissenschaftsdidaktischen Forschung die meisten Arbeiten eine Kreuzvalidierungsstrategie gewählt haben und die Modelle, im Vergleich zu anderen Strategien, die bessere Performance aufweisen. Daraus lässt sich schließen, dass durch eine Kreuzvalidierungsstrategie Machine-Learning-Modelle besser neue Daten verarbeiten können, was zu validieren Modellen führt (Zhai et al., 2020).

4.2.2 Modelltypen

Überwachtes maschinelles Lernen wird in der Regel für Klassifikations- oder Regressionsverfahren angewendet. Im folgenden Abschnitt werden nur Verfahren für die Klassifikation dargestellt, da in dieser Arbeit ein Modell zur Klassifizierung von Textdaten entwickelt wird.
Im Allgemeinen ist ein Trainingsdatensatz $(X,Y) = ((x_1,y_1),...,(x_N,y_N))$ gegeben, wobei $x_i \in \mathbb{R}$ die Merkmalsvektoren und $y_i \in \{0,1,...,K\}$ die zu klassifizierenden Klassen sind (Plaue, 2021). Bei einer binären Klassifikation mit den Klassen $Y = \{0,1\}$ wäre das Ziel des Machine-Learning-Modells, eine Entscheidungsregel $\hat{f} : \mathbb{R}^D \to \{0,1\}$ zu finden, mit der neue unbekannte Daten $x_* \in \mathbb{R}$ einer Klasse $\hat{y_*} = \hat{f}(x_*)$ zugeordnet werden (Plaue, 2021).
Für eine Klassifikation wurde bereits eine Vielzahl von Algorithmen entwickelt (Richter, 2019). Die Wahl eines passenden Algorithmus ist dabei nicht trivial, da je nach Anwendungsfall und Datensatz einige Algorithmen besser abschneiden als andere (Kuhn & Johnson, 2020). Daher ist es schwer, im Vorfeld einen geeigneten Algorithmus zu bestimmen, weswegen die Auswahl ein iterativer Prozess ist und mehrere Algorithmen getestet werden (Zhai et al., 2020). Man findet in der Literatur viele unterschiedliche Algorithmen. Die häufig verwendete Python-Bibliothek

für maschinelles Lernen, scikit-learn, bietet 13 verschiedene Algorithmenklassen an (Pedregosa et al., 2011). In diesem Abschnitt wird sich auf eine Auswahl von einigen Schlüsselmodellen fokussiert, die als Fundament für das Verständnis und die Anwendung von Klassifikationsmethoden dienen sollen.

Logistische Regression

Die logistische Regression ist ein Algorithmus zur Klassifizierung und kein Regressionsmodell. Der Algorithmus wird auch als *logit regression*, *maximum-entropy classification* (MaxEnt) oder als *log-linear classifier* bezeichnet (Pedregosa et al., 2011).
Der Algorithmus wird bei einem binären Klassifikationsproblem angewendet und liefert als Ausgabe nicht nur die Klasse, sondern auch eine Wahrscheinlichkeit für die Klassifizierung (Plaue, 2021). Betrachtet man die Beispieldaten aus Abbildung 4.8, fällt auf, dass man den Zusammenhang schlecht mit einer linearen Funktion modellieren kann. Es besteht weder ein stetiger Zusammenhang noch ist die lineare Funktion für die Zielvariable beschränkt (Hirschle, 2022).
Deswegen wählt man als Entscheidungsfunktion die nicht lineare Sigmoidfunktion, die auf dem Intervall $[0,1]$ beschränkt ist:

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

Die Klassifikation erfolgt dann mittels eines Grenzwertes, der, wegen der A-posteriori-Wahrscheinlichkeit meistens auf 0,5 gesetzt ist (Ertel, 2021):

$$\text{Klasse} = \begin{cases} 1 \text{ falls } \frac{1}{1+\exp(-(w_o+w_1x_1+w_2x_2+...+w_Nx_N))} \geq 0,5 \\ 0 \text{ sonst} \end{cases}$$

Die Gewichte w_i werden über eine Verlustfunktion und Optimierer bestimmt. Die Verlustfunktion misst die Abweichung zwischen den vorhergesagten und den tatsächlichen Werten. Der Optimierer versucht, die Gewichte des Modells iterativ anzupassen, um die Verlustfunktion zu minimieren. Ziel ist es, dass das Modell die Verteilung der Daten bestmöglich abbildet (Hirschle, 2022).

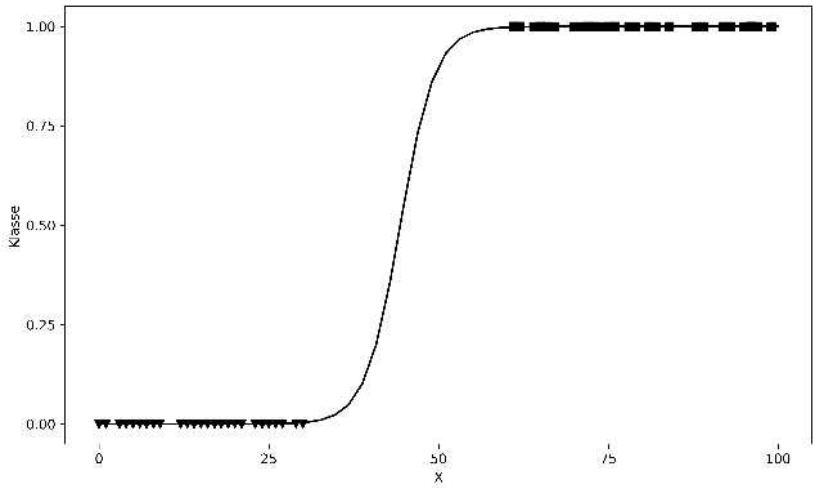


Abbildung 4.8: Binäre Klassifizierung durch eine logistische Regression

Support Vector Machine

Support Vector Machines (SVM) sind Verfahren zur linearen Separierung von Daten. SVMs gelten als ein sehr robuster Algorithmus, der hohe Leistung in verschiedenen Kontexten liefern kann (Bonaccorso, 2017; Zhai et al., 2020).
Das Ziel einer SVM ist es, die Daten mithilfe einer Hyperebene in zwei Klassen aufzuteilen. Der Algorithmus versucht dabei, den Abstand zwischen den Daten und der trennenden Hyperebene zu maximieren (Noble, 2006). Dabei werden die Datenpunkte, die der Hyperebene am nächsten liegen, Stützvektoren oder *support vectors* genannt (Ertel, 2021) (siehe Abbildung 4.9).
In einem zweidimensionalen Raum ist die trennende Hyperebene eine Linie. Die Hyperebene kann daher als lineare Entscheidungsfunktion definiert werden:

$$f(x) = x^T \beta + \beta_0 = 0$$

Dabei ist β der Normalenvektor und β_0 eine Konstante. Diese beiden Parameter der linearen Entscheidungsfunktion werden durch das Modell gelernt, um die optimale Trennung zwischen den Klassen zu finden.
Die Klassifizierung erfolgt basierend auf dem Vorzeichen von $f(x)$. Indem die Entscheidungsfunktion gleich null gesetzt wird, wird definiert, dass Daten, die auf der einen Seite der trennenden Linie liegen ($f(x) > 0$), einer Klasse zugeordnet werden, während Daten auf der anderen Seite ($f(x) < 0$) der anderen Klasse zugeordnet werden (siehe Abbildung 4.9).
In der linken Darstellung in Abbildung 4.9 findet man Beispieldaten und die tren-

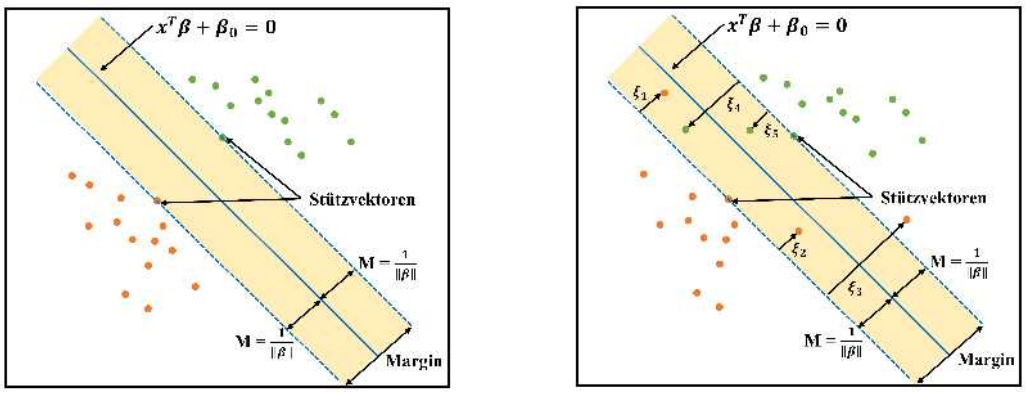


Abbildung 4.9: Darstellung einer Support Vectore Machine (SVM). Links: Linear trennende Hyperebene, die die Datenpunkte aufteilt und den entstehenden Abstandsraum (Margin). Rechts: Nicht vollständig linear trennbare Datenpunkten mit dem Toleranzparameter ζ (Hastie et al., 2009, S. 418)

nende Linie. Die gestrichelten Linien markieren die Stützvektoren, also die Daten, die der Linie am nächsten liegen und den dadurch entstehenden Abstandsraum oder auch *margin M* genannt. Die *margin* ist demnach der Abstand zwischen der Hyperebene und den Stützvektoren. Typischerweise wird die *margin* als $2M = \frac{2}{\|\beta\|}$ definiert, wobei die Formel sich aus einer geometrischen Überlegung ergibt (siehe Abbildung 4.9). Die Norm von β wird verwendet, da sie den Abstand von der Hyperebene zu den nächsten Punkten in Richtung des Normalenvektors repräsentiert. Das Ziel ist es nun, die Hyperebene zu finden, die den Abstandsraum zwischen den beiden Klassen maximiert. Denn ein großer Abstandsraum bedeutet, dass die trennende Linie weiter von den Trainingsdaten entfernt ist. Dadurch wird eine klare Trennung zwischen den Klassen erreicht, was eine bessere Generalisierungsfähigkeit und Robustheit des Modells gegenüber neuen Daten impliziert (Hastie et al., 2009).

Der in der linken Darstellung in Abbildung 4.9 gezeigte Fall geht von optimal trennbaren Daten aus. Bei realen Problemen können die Trainingsdaten der positiven Klassen aber neben Daten der negativen Klasse liegen (Richter, 2019). Um dennoch eine Klassifikation durchzuführen, führt man einen Toleranzparameter ζ_i ein, der auch *slack variable* genannt wird und den Grad der Falschklassifizierung auf einer Skala von 0 bis ∞ bestimmt (Hastie et al., 2009; Richter, 2019). Durch den Toleranzparameter kann berechnet werden, mit welchem Abstand ein Datenpunkt auf der „falschen“ Seite der trennenden Linie liegt (siehe Abbildung 4.9). Zudem kann durch die Hinzunahme einer weiteren Konstanten festgelegt werden, wie viele Datenpunkte auf der falschen Seite liegen dürfen (Richter, 2019). In der

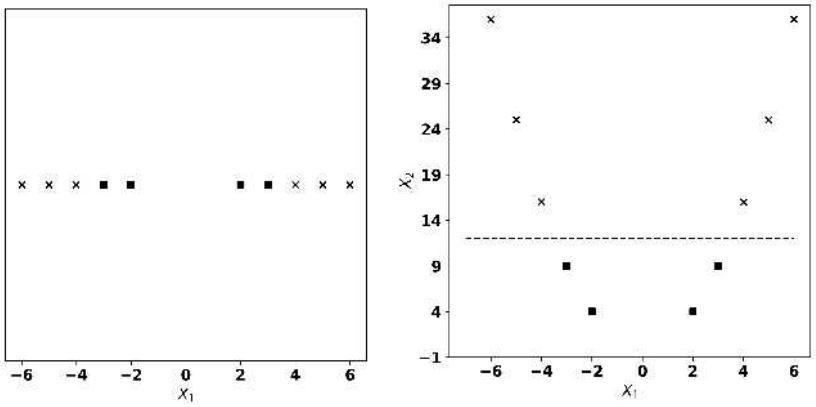


Abbildung 4.10: Kernel-Trick: Projektion von Daten in eine höhere Dimension. Hier mit der Funktion $\Phi(x) = x^2$

rechten Darstellung in Abbildung 4.9 ist ein solcher Fall dargestellt.
Falls die Daten sich nicht linear trennen lassen, kann eine SVM trotzdem angewandt werden (Noble, 2006). Durch den sogenannten Kernel-Trick können die ursprünglichen Vektoren in einen höherdimensionalen Raum transformiert und projiziert werden (siehe Abbildung 4.10). Durch diesen Schritt werden die Vektoren im höher dimensionalen Raum wieder linear separierbar und eine Hyperebene zur Trennung der Daten kann gefunden werden (Bonaccorso, 2017).
Daher gelten SVMs als ein sehr robuster Algorithmus, der in verschiedenen Kontexten eingesetzt werden kann (Bonaccorso, 2017; Zhai et al., 2020).

Decision Tree und Random Forest

Entscheidungsbäume oder auch Decision Trees genannt, sind eine nichtparametrische Methode zur Klassifizierung von Daten (Pedregosa et al., 2011). Sie gelten als ein einfaches und effizientes Verfahren. Das Ziel ist die Vorhersage der Zielvariable durch gelernte Entscheidungsregeln (Ertel, 2021; Pedregosa et al., 2011). Die Stärke von Decision Trees liegt in der Interpretierbarkeit, da der entstehende Entscheidungsbaum nachzuvollziehen und zu kontrollieren ist (Ertel, 2021).
Die (binäre) Klassifikation wird als ein sequenzieller Entscheidungsprozess durchgeführt. Das bedeutet, dass jedes Merkmal aus dem Datensatz bewertet und in Zweige aufgeteilt wird. Dies wird so lange wiederholt, bis die gesuchte Zielvariable erreicht ist (Bonaccorso, 2017). In Abbildung 4.11 sind die Merkmale *Entfernung*, *Wochenende* und *Sonne* visualisiert. Durch die Bewertung dieser Merkmale können Grenzen festlegt werden, welche dann Entscheidungszweige aufspalten, die letztlich zu der Zielvariable ja oder nein führen. Je nach Merk-

malsauswahl und den dazugehörigen Grenzwerten ändert sich die Struktur des Entscheidungsbaumes (Bonaccorso, 2017).

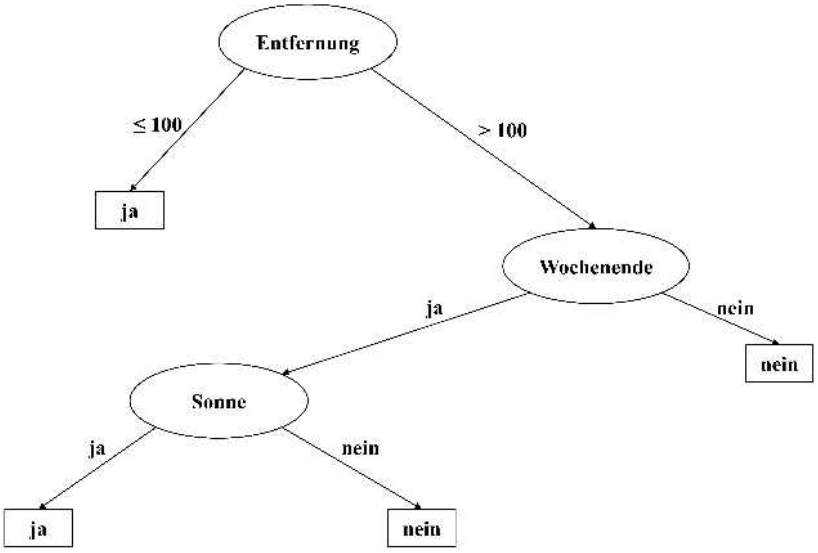


Abbildung 4.11: Entscheidungsbaum für das Klassifikationsproblem Skifahren (Ertel, 2021)

Betrachtet man die Trainingsdaten $X = \{x_1, x_2, ..., x_n\}$, wobei $x_i \in \mathbb{R}^m$ gilt. Das bedeutet, dass jeder Input aus m-Merkmalen besteht. Der Algorithmus eines Decision Trees arbeitet in rekursiven Schritten. Zunächst wird durch eine Auswahlfunktion ein Merkmal ausgesucht, das den Datensatz am besten aufteilt. Dafür kann man z. B. den Informationsgewinn der einzelnen Merkmale berechnen. Der Informationsgewinn wird bestimmt, indem man die Entropie zwischen dem ursprünglichen und dem aufgeteilten Datensatz bestimmt.
Die Entropie H misst die Unreinheit oder Unordnung einer beliebigen Menge von Merkmalen. Eine hohe Entropie ist gleichbedeutend mit einer höheren Unordnung der Daten und einer niedrigeren Vorhersagegenauigkeit des Entscheidungsbaums, da es schwieriger ist, klare Entscheidungen zu treffen (Ertel, 2021). Zur Berechnung der Entropie betrachtet man die Anzahl der Klassen N (z. B. Ja und Nein, siehe Abbildung 4.11) und die Wahrscheinlichkeiten p_i, dass ein Datenelement der Klasse i angehört (Ertel, 2021):

$$H(p) = H(p_1, ..., p_n) := -\sum_{i=1}^{N} \log_2 p_i$$

Daraus lässt sich der Informationsgehalt $I(D) := 1 - H(D)$ einer Datenmenge D

bestimmen, den man als Gegenwahrscheinlichkeit verstehen kann (Ertel, 2021). Aus dem Informationsgehalt lässt sich wiederum der Informationsgewinn definieren:

$$G(D,m) = \sum_{i=1}^{N} \frac{|D_i|}{|D|} I(D_i) - I(D)$$

Der Informationsgewinn $G(D,m)$ wird also durch die Differenz zwischen der Entropie des ursprünglichen Datensatzes und der gewichteten durchschnittlichen Entropie der Teil-Datensätze berechnet, die durch die Aufteilung des Datensatzes entstehen (Ertel, 2021). Ein Merkmal teilt den Datensatz umso besser, je höher sein Informationsgehalt ist. Nachdem ein Merkmal ausgesucht und der Datensatz aufgeteilt wurde, beginnt dieser Schritt wieder von vorn, bis keine Merkmale mehr existieren.

Der Random-Forest-Ansatz ist eine Erweiterung des Decision Trees, da bei diesem Verfahren mehrere Entscheidungsbäume trainiert werden (Breiman, 2001). Dabei wird jeder einzelne Entscheidungsbaum auf einem zufällig ausgewählten Teildatensatz trainiert. Dadurch entstehen mehrere disjunkte Datensätze mit einer zufälligen Teilmenge der Merkmale. Das resultiert in mehreren schwächeren Entscheidungsbäumen, die unterschiedliche Vorhersagen machen können (Bonaccorso, 2017). Das Klassifikationsergebnis kann dann z. B. über ein Mehrheitsvotum oder Mittelwertbildung bestimmt werden (Pedregosa et al., 2011). Durch die Bildung eines solchen Ensembles soll eine bessere Klassifikation erreicht werden, insbesondere bei hochdimensionalen Trainingsdaten (Richter, 2019).

K-Nearest-Neighbors

Der K-Nearest-Neighbor-Algorithmus (KNN) nutzt zur Klassifikation die Lage und Abstände der einzelnen Datenpunkte im Merkmalsraum (Plaue, 2021). Bei der KNN-Klassifikation kann als Abstandsmaß die euklidische Distanz genutzt werden (Ertel, 2021):

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Wählt man $k = 1$, wird zur Klassifikation nur der nächstgelegene Punkt betrachtet und dessen Klasse zugeordnet. Es wird also geschaut, welche Klasse der Datenpunkt mit der niedrigsten euklidischen Distanz hat und dessen Klasse wird für den neuen Datenpunkt übernommen. Bei einem höheren k wird eine Mehrheitsentscheidung unter den k nächsten Nachbarn getroffen. Dadurch entstehen

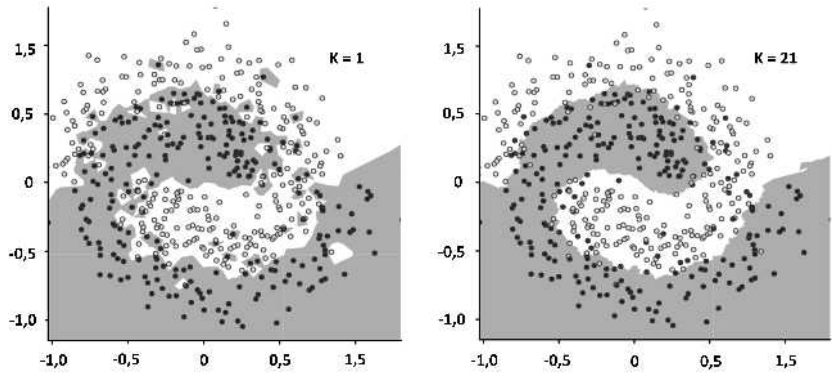


Abbildung 4.12: Darstellung eines KNN-Klassifikators mit $k = 1$ (links) und $k = 21$ (rechts) (Plaue, 2021, S. 226)

Entscheidungsgrenzen wie für den Fall einer binären Klassifizierung in Abbildung 4.12. Der Fall $k = 1$ ist in der linken Grafik dargestellt. Wenn ein neuer Punkt klassifiziert werden soll, wird er einer der beiden Klassen zugeordnet. Die hellen bzw. dunkelgrauen Flächen bilden dabei die Entscheidungsgrenze, die durch die euklidische Distanz entstanden sind. In der rechten Grafik ist der Fall mit $k = 21$ dargestellt. Durch die Betrachtung der 21 nächsten Nachbarn verändern sich die Entscheidungsgrenzen für die zu klassifizierenden Daten. Falls es zu keiner Mehrheit kommt, wird der Punkt einer zufälligen Klasse zugeordnet (Hastie et al., 2009). In der Praxis wird der genaue Wert für k mittels des Validierungssets getestet und so der KNN-Klassifikator optimiert. Obwohl der Algorithmus im Vergleich zu den anderen Verfahren recht simpel erscheint, wurde er in vielen Studien erfolgreich eingesetzt (Hastie et al., 2009).

Multilayer Perceptron (MLP)

Neuronale Netze spielen ebenfalls eine wichtige Rolle bei der Klassifikation, insbesondere bei der Verarbeitung von hochdimensionalen Daten wie bei der Bild- oder Musikerkennung (Richter, 2019). Ein recht einfaches, aber weitverbreitetes neuronales Netz ist das Multilayer Perceptron (MLP) (Goodfellow et al., 2016). Ein MLP besteht aus einer Eingabeschicht, mindestens einer verdeckten Schicht und einer Ausgabeschicht (siehe Abbildung 4.13). Jede Schicht besteht aus einer Anzahl von Neuronen. Die Eingabeschicht nimmt die Trainingsdaten auf und leitet sie zu der (ersten) verdeckten Schicht weiter. Hier werden alle Berechnungen des neuronalen Netzes durchgeführt. Die Ausgabeschicht generiert die Vorhersagen des MLPs (Goodfellow et al., 2016). Bei einem binären Klassifikationsproblem enthält die Ausgabeschicht nur ein einzelnes Neuron. Es enthält die aggregierten

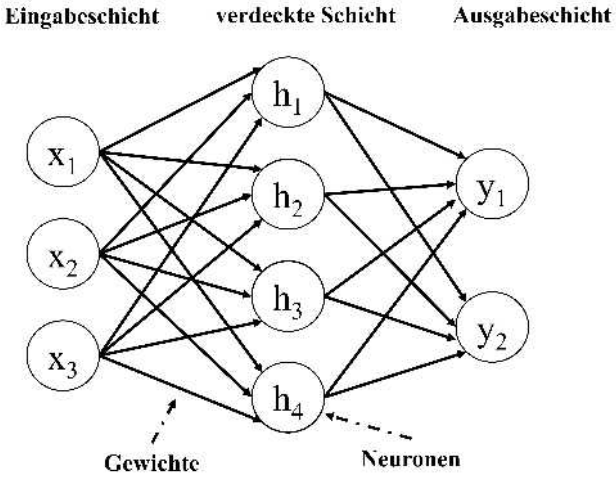


Abbildung 4.13: Darstellung eines Multilayer Perceptron (MLP) mit einer verdeckten Schicht (Prince, 2023, S. 36)

Informationen aus der verdeckten Schicht und erzeugt eine Entscheidung für eine der beiden Klassen. Das Ausgabeneuron gibt letztlich die Wahrscheinlichkeit an, dass die Eingabe zu einer der beiden Klassen gehört (Prince, 2023). Die einzelnen Neuronen in den jeweiligen Schichten sind miteinander verbunden. Jede Verbindung zwischen den Neuronen besitzt ein Gewicht, das während des Trainingsprozesses angepasst wird (siehe Abbildung 4.13) (Prince, 2023).

Die Betrachtung der beschriebenen Algorithmen des überwachten maschinellen Lernens hat gezeigt, dass es sehr unterschiedliche Ansätze zur Klassifikation gibt. Die Algorithmen weisen unterschiedliche Vorteile und Grenzen auf. Die logistische Regression oder KNN sind einfach zu verstehen und zu implementieren, wohingegen neuronale Netze wie das MLP aufgrund ihrer teilweise mehrschichtigen und komplexen Struktur als Blackbox-Modell aufgefasst werden, weswegen ihre Vorhersagen nur schwer nachzuvollziehen sind (Buhrmester et al., 2021). Dagegen besitzen neuronale Netze die Fähigkeit, komplexe nicht lineare Beziehungen zu modellieren, währenddessen eine logistische Regression nur lineare Entscheidungsgrenzen erstellen kann. Daher ist es im Vorfeld nicht leicht, einen passenden Algorithmus auszuwählen, weswegen man in der Regel verschiedene Algorithmen für die Entwicklung eines Machine-Learning-Modells nutzt (Kuhn & Johnson, 2020).
Neben den hier beschriebenen Verfahren existiert noch eine ganze Reihe anderer Ansätze wie Boosting-Ansätze oder der Naive-Bayes-Klassifikator (Pedregosa et al., 2011). Beim Boosting-Ansatz werden mehrere schwache Lerner zu einem

starken Lerner kombiniert. Das Ziel besteht darin, die Leistung des Modells kontinuierlich zu verbessern, indem die Fehler der vorherigen Modelle korrigiert werden. Dabei werden die Modelle nacheinander trainiert, wobei versucht wird, die Fehler des vorherigen Modells zu reduzieren (Data Basecamp, 2023a). Typische Boosting-Ansätze sind z. B. AdaBoost, XGBoost, oder Gradient Boost (Pedregosa et al., 2011). Eine detaillierte Beschreibung jedes verfügbaren Algorithmus oder aller neuronaler Netze kann diese Arbeit nicht leisten. Daher soll es bei der Darstellung der ausgewählten Algorithmen belassen werden.

4.3 Testen und Optimieren eines Modells

Um die Güte eines entwickelten Machine-Learning-Modells zu bestimmen, muss es evaluiert und getestet werden. Durch die Leistungsmessung können nicht nur die unterschiedlichen Modelle verglichen, sondern auch die einzelnen Algorithmen optimiert werden. Zunächst werden deshalb in Abschnitt 4.3.1 die gängigen Kriterien zur Messung eines Klassifikators vorgestellt und anschließend in Abschnitt 4.3.2 die Hyperparameter-Optimierung und der Verzerrungs-Aspekt thematisiert.

4.3.1 Gütekriterien eines Klassifikationsmodells

Die Wahl der passenden Evaluationskennzahlen ist ein entscheidender Schritt bei der Entwicklung eines Machine-Learning-Modells (Hossin & Sulaiman, 2015). Es wird deshalb eine Vielzahl von unterschiedlichen Kennzahlen eingesetzt, um die Qualität eines Modells zu bestimmen. Im Folgenden werden häufig verwendete Kennzahlen beschrieben und zur Einfachheit lediglich der Fall der binären Klassifikation mit den Klassen 0 (negativ) und 1 (positiv) genutzt (Burkov, 2020; Zhai et al., 2020). Eine umfangreichere Übersicht findet sich z. B. in Hossin und Sulaiman (2015) oder Ferri et al. (2009).

	Klasse 1	**Klasse 0**
Vorhersage Klasse 1	True Positive (TP)	False Positive (FP)
Vorhersage Klasse 0	False Negative (FN)	True Negative (TN)

Tabelle 4.3: Konfusionsmatrix einer binären Klassifikation mit den Klassen 1 und 0

Bei einem Klassifikationsansatz ist die Grundlage für die Messung der Performance die Mensch-Maschine-Übereinstimmung. Dabei werden die von Menschen

erstellten Vorhersagen oder Bewertungen als „Goldstandard“ angesehen und zum Vergleich mit dem Machine-Learning-Modell genutzt (Zhai et al., 2020). Viele Kennzahlen lassen sich daher über die Konfusionsmatrix (Confusion-Matrix) berechnen (siehe Tabelle 4.3).
Die Zeilen der Konfusionsmatrix stehen für die von dem Modell vorhergesagten Klassen. Die Spalten dagegen stehen für die tatsächlichen Klassen. Aus den richtig klassifizierten Fällen *TP* und *TN* lässt sich die Treffsicherheit oder Accuracy berechnen:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Die Accuracy ist eine einfache und eine oft genutzte Kennzahl, um die Qualität eines Modells zu messen, da sie den Anteil an richtig klassifizierten Fällen angibt (Ferri et al., 2009). Allerdings hat die Accuracy bei unausgeglichenen Datensätzen, die von einer Klasse stark dominiert werden, nur eine bedingte Aussagekraft. Wenn in einem Datensatz 90 % der Daten zu einer Klasse gehören, könnte ein einfaches Modell, das immer diese Klasse vorhersagt, eine Accuracy von 90 % erreichen.
In der Praxis benutzt man deshalb weitere Kennzahlen wie Recall und Precision (Hirschle, 2022):

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

Die Precision gibt das Verhältnis zwischen den richtig vorhergesagten positiven Fällen und der Zahl aller Fälle, die das Modell als positive Fälle vorhergesagt hat, an. Der Recall hingegen zeigt das Verhältnis der richtig vorhergesagten positiven Fälle und der Fälle, die eigentlich als positiv hätten vorhergesagt werden müssen (siehe Tabelle 4.3). Da Precision und Recall einander beeinflussen, wird je nach Aufgabe des Modells versucht, eine der beiden Kennzahlen zu maximieren. Möchte man z. B. eine Klassifikation von Spam-Nachrichten durchführen, wäre ein hohe Precision wünschenswert, da man so sicherstellt, dass weniger korrekte Nachrichten im Spam-Ordner landen (Burkov, 2020).
Eine weitere gängige Kennzahl, die sowohl die Precision, als auch den Recall berücksichtigt, ist der F1-Score (Hossin & Sulaiman, 2015):

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Der F1-Score ist das harmonische Mittel zwischen Recall und Precision und kann

Werte zwischen 0 und 1 annehmen. Bei einer Klassifizierung mit mehr als zwei Klassen werden oftmals Durchschnitts- oder gewichtete F1-Scores angegeben. Dazu berechnet man zunächst für jede Klasse einen F1-Score. Daraus kann dann ein Macro-F1-Score gebildet werden, der der durchschnittliche F1-Score ist. Beachtet man noch die Häufigkeitsverteilung der Daten der jeweiligen Klassen, kann ein gewichteter F1-Score berechnet werden, wobei die Anzahl der Daten für jede Klasse berücksichtigt wird.
Betrachtet man Forschungsarbeiten aus dem Bereich Science Education und Machine Learning (z. B. Zhai et al., 2020), findet man als Qualitätskennzahl für Machine-Learning-Modelle häufig Cohen's Kappa (Cohen, 1960):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Der Vorteil des Cohen's Kappas ist, dass nicht nur die Übereinstimmung zwischen Mensch und Modell gemessen werden kann, sondern dass auch die zufällige Klassifikation herausgerechnet wird. In der Definition des Kappas steht p_o für die beobachtete und p_e für die erwartete Übereinstimmung. Die beobachtete Übereinstimmung ist äquivalent zu der oben beschriebenen Accuracy. Die erwartete Übereinstimmung, die die hypothetische Wahrscheinlichkeit einer zufälligen Übereinstimmung angibt, ist definiert als:

$$p_e = p_{class1} + p_{class2} =$$

$$\frac{(TN+FP)\cdot(TN+FN)}{(TN+TP+FP+FN)^2} + \frac{(FN+FP)\cdot(FP+TP)}{(TN+TP+FP+FN)^2}$$

Das Cohen's Kappa kann maximal den Wert 1 annehmen. Ein Kappa von unter 0 bedeutet, dass keine Übereinstimmung vorliegt. Ein Kappa zwischen 0 und 0,2 deutet auf eine leichte Übereinstimmung, zwischen 0,21 und 0,40 auf eine ausreichende Übereinstimmung, zwischen 0,41 und 0,60 auf eine mittelmäßige Übereinstimmung, zwischen 0,61 und 0,80 auf eine beachtliche Übereinstimmung und zwischen 0,81 und 1 auf eine fast vollkommene Übereinstimmung hin (Landis & Koch, 1977).
Unabhängig von der Wahl der Kennzahl zur Messung der Leistung eines Machine-Learning-Modells müssen akzeptable Schwellenwerte definiert werden. Diese variieren je nach Kontext und sind schwer zu verallgemeinern (Burkov, 2020).

4.3.2 Hyperparameter und Verzerrung-Varianz-Dilemma

Um ein leistungsstarkes Modell zu entwickeln, müssen gewisse Parameter im Vorfeld des Trainingsprozesses gewählt werden. Das Klassifikationsergebnis kann je nach gewählten Hyperparametern variieren. Diese Hyperparameter sind gewissermaßen Stellschrauben, die das Modell nicht selbstständig lernt, sondern die ausgewählt werden müssen (Hirschle, 2022). Je nach Algorithmus können unterschiedlich viele Hyperparameter optimiert werden, z. B. die Anzahl der nächsten Nachbarn bei der Klassifikation mittels KNN. Um die besten Hyperparameter zu finden, wird das Validierungsset genutzt (vgl. Abbildung 4.6). Dazu wird eine Reihe von unterschiedlichen Hyperparametern ausgewählt, die zum Trainieren des Modells genutzt werden. Für die Auswahl können unter anderem zufällige Hyperparameter genutzt oder eine systematische Rastersuche durchgeführt werden (Plaue, 2021). Das Modell, das durch die Hyperparametersuche die beste Leistung auf dem Validierungsset erzielt hat, wird ausgewählt und anschließend mithilfe des Testdatensatzes final bewertet (Plaue, 2021).
Bei der Optimierung des Modells ist das Verzerrung-Varianz-Dilemma zu beachten (Richter, 2019). Die Varianz erfasst, wie stark sich das Machine-Learning-Modell verändert, wenn es mit einem anderen Trainingsdatensatz trainiert wird (Fortmann-Roe, 2012). Es gibt sozusagen die Schwankung der Klassifikation bei wiederholter Anwendung an (Plaue, 2021). Die Verzerrung oder auch Bias genannt, bezieht sich auf inhärente Fehler, die das Modell auch bei einer großen Anzahl an Trainingsdaten macht. Das Modell ist also für eine bestimmte Art von Lösung oder Daten voreingenommen und wiederholt den Fehler immer wieder (Fortmann-Roe, 2012). Modelle können sowohl eine hohe Varianz als auch eine hohe Verzerrung haben (siehe Abbildung 4.14).
Das Dilemma besteht darin, dass Modelle, welche eine komplexe Struktur haben und optimal an die Trainingsdaten angepasst wurden, anfällig für kleine Änderungen sind und deshalb eine hohe Varianz aufweisen. Die Modelle sind zu gut an die Trainingsdaten angepasst, weswegen man auch von Overfitting spricht (Plaue, 2021). Modelle, die nicht komplex genug sind, weisen jedoch oft eine hohe Verzerrung auf. In diesem Fall spricht man von Underfitting (Fortmann-Roe, 2012).
Daraus lässt sich schließen, dass durch die Reduzierung der Verzerrung die Varianz steigt und vice versa (Plaue, 2021). Bei der Optimierung der Hyperparameter muss der Punkt gefunden werden, an dem das Modell sowohl eine geringe Verzerrung

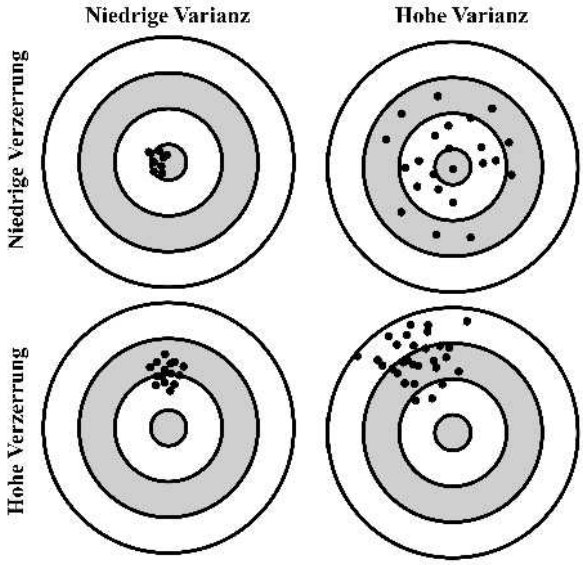


Abbildung 4.14: Darstellung von Verzerrung und Varianz (Fortmann-Roe, 2012)

als auch eine niedrige Varianz aufweist. Dadurch entsteht ein Modell, das die Trainingsdaten gut abbildet und auch auf neue Daten generalisierbar ist.
Der Aspekt der Verzerrung kann bei der Erstellung von Machine-Learning-Modellen an verschiedenen Punkten beachtet werden (Krist & Kubsch, 2023). Schon bevor das eigentliche Modell entwickelt wird, wird in der Regel die Aufgabe definiert. Diese Aufgabe kann im Bildungskontext z. B. eine Fragestellung sein, die Lernende bearbeiten sollen. Ein Machine-Learning-Model könnte dann auf Grundlage der Antworten für eine automatische Bewertung trainiert werden. Die Konzeption der Fragestellung kann bereits zu einer Verzerrung führen, da die Standpunkte und Absichten der entwickelnden Menschen in die Fragestellung einfließen können (Harper & Kayumova, 2023; Krist & Kubsch, 2023). Es kann aber auch durch die eigentlichen Daten zu Verzerrungen kommen, wenn z. B. bestimmte Gruppen unterrepräsentiert sind (Baker, 2019). Aber auch durch das eigentliche Modelllernen kann es zu einer Verzerrung führen, die beachtet werden muss (Krist & Kubsch, 2023).

4.4 Maschinelles Lernen in der Bildungsforschung

Künstliche Intelligenz ist auch im Bereich der Bildungsforschung ein aktuelles und aufstrebendes Forschungsgebiet (Sghir et al., 2022; Shin & Shim, 2020).

Obwohl es diesen Forschungsbereich schon seit ungefähr 30 Jahren gibt, wurde erst in den vergangenen Jahren intensiv geforscht, wie man KI-basierte Lösungen zur Unterstützung der Lernenden und Lehrenden sowie als Werkzeug für die Forschenden einsetzt (Zawacki-Richter et al., 2019). Durch die immer stärker werdende Verbreitung von digitalen Lernumgebungen kann eine Vielzahl von Daten erhoben werden, die z. B. für ein Feedback genutzt werden kann (Sghir et al., 2022). In der vorliegenden Arbeit wird ein Machine-Learning-Modell entwickelt, das für ein automatisiertes formatives Assessment verwendet wird. Daher werden in diesem Abschnitt speziell Forschungsarbeiten aus diesem Bereich betrachtet und diskutiert.

Liu et al. (2016) entwickelten das Machine-Learning-Modell *c-rater-ML*, das acht verschiedene Aufgaben zur Erklärung naturwissenschaftlicher Problemstellungen automatisch bewerten kann. Die Aufgaben stammten dabei aus den Klassenstufen 6 bis 8 und behandelten Themen wie Energieumwandlung, Energieübertragung, Plattenbewegung oder Photosynthese. Für die automatische Auswertung nutzten die Autoren ein überwachtes maschinelles Lernverfahren und konnten eine Übereinstimmung von einem Cohen's Kappa zwischen 0,62 und 0,90 zwischen dem Modell und den menschlichen Bewertungen erreichen. Die Autoren führten die Schwankungen auf die unterschiedlichen Stichprobengrößen und die möglichen nicht berücksichtigten lexikalischen Antwortvariationen zurück (Liu et al., 2016). Insgesamt bewerteten Liu et al. (2016) die Kappas als zufriedene Übereinstimmung und stellten das Potenzial ihres Modells für den Einsatz im Unterricht heraus.

Zu einem ähnlichen Ergebnis kommen auch Kayhan Moharreri et al. (2014). Die Forschungsgruppe entwickelte das formative Assessmentsystem *EvoGrader*, das für eine automatische Bewertung von schriftlichen Erklärungen im Biologie-Unterricht entwickelt wurde. Das System analysiert Antworten von Lernenden automatisch und liefert so den Lehrkräften detaillierte Informationen mittels verschiedener Diagramme und Tabellen. Für die Übereinstimmung zwischen Mensch und Modell konnten die Autoren ein durchschnittliches Cohen's Kappa von 0,84 angeben (Kayhan Moharreri et al., 2014). Dabei betonten Kayhan Moharreri et al. (2014), dass die Auswertungsdauer 99 % weniger Zeit in Anspruch nimmt und kostenlos ist.

Nakamura et al. (2016) untersuchten ein Machine-Learning-Modell zur automatischen Klassifizierung von Kurzantworten zu verschiedenen physikalischen Themen, welche aus einer interaktiven Lernumgebung stammen. Im Gegensatz

zu den anderen beiden Studien konnten die Autoren nur bei vier von neun automatisierten Bewertungen eine Überstimmung von 70 % angeben. Die schlechten Bewertungen führten Nakamura et al. (2016) auf die teilweise kleinen Datensätze von um die 100 Antworten zurück. Aus Sicht von Nakamura et al. (2016) stellt die Übereinstimmung von 70 % eine Grundlage für die Nutzung solcher Machine-Learning-Modelle für den Unterricht dar.
Steinert et al. (2023) entwickelten eine Plattform namens LEAP (Learning with AI about Physics), die ChatGPT nutzt, um Lernenden automatisch Feedback zu geben. Das Ziel der Studie war es, selbstreguliertes Lernen zu verbessern, indem die Plattform automatisch formatives Feedback bereitstellt. Dazu konnten Lehrende eigenständig Fragen in der Plattform implementieren und eine richtige Lösung bereitstellen. Auf der Plattform mussten die Lernenden eine Antwort abgeben, die dann an ChatGPT gesendet wurde, um ein formatives Feedback zu erhalten. Die Ergebnisse von Steinert et al. (2023) deuten darauf hin, dass die Verwendung von großen Sprachmodellen wie ChatGPT ein vielversprechender Weg für ein personalisiertes und effektives Lernen sei.
Die positiven Ergebnisse konnten Zhai et al. (2020) in ihrem Review bestätigen. Die untersuchten Studien konnten Übereinstimmungen von Kappa-Werten von 0,08 bis 0,99 zeigen, wobei der Durchschnittswert bei einem Kappa von 0,72 lag. Die besten Ergebnisse konnten Machine-Learning-Modelle erzielen, die eine Cross-Validation-Strategie genutzt hatten (Zhai et al., 2020). Die Autoren sind überzeugt, dass der Einsatz von Machine-Learning-Modellen den Bildungsbereich voranbringen kann und ein Potenzial für zeitnahes und effektives automatisches Feedback bietet. Durch genaue Rückschlüsse aus komplexen Daten können Lehrkräfte entlastet werden (Zhai et al., 2020, 2021). Zhai et al. (2020) weisen jedoch darauf hin, dass viele Studien sich nur auf die Validität der Machine-Learning-Modelle fokussieren und fordern deshalb mehr Arbeiten, die auch die Anwendung solcher Modelle im Unterricht erforschen.
Es existiert aber auch eine Vielzahl von Studien, die den Einsatz von Machine-Learning-Modellen im Unterricht kritisch sehen (Li et al., 2023). Nach Cheuk (2021) ist eines der Hauptprobleme bei dem Einsatz von KI im Bildungsbereich, dass die Trainingsdatensätze nicht ausreichend beschrieben und so mögliche Verzerrungen (Bias) nicht erkannt werden. So können Lernende, die einen Migrationshintergrund besitzen, nur wenig im Trainingsdatensatz vertreten sein, was zu einer schlechteren Performance und einer Verzerrung führen kann (Yao et al., 2020). Zudem konnte eine Diskrepanz zwischen mehrsprachigen und reinen eng-

lischsprachigen Lernenden nachgewiesen werden (Li et al., 2023). Dies lässt den Schluss zu, dass viele sprachliche Praktiken von Lernenden nicht in ausreichender Menge im Trainingsdatensatz enthalten waren, was zu einer Verzerrung des Modells führte (Li et al., 2023). Trotz der hohen Übereinstimmung zwischen Mensch und Machine-Learning-Modell sind Li et al. (2023) der Meinung, dass der Einsatz solcher Systeme sogar negative Auswirkungen auf die Lernenden haben kann, da im Gegensatz zu ausgebildeten Lehrkräften Machine-Learning-Modelle Aspekte wie die individuelle Geschichte des Lernenden oder die Dynamik im Klassenzimmer nicht berücksichtigen können (Li et al., 2023).
Es gibt aber auch Herausforderungen, die die Lehrkräfte betreffen. Lehrkräfte müssen bei der Verwendung von Machine-Learning-Modellen nicht nur in der Lage sein, die Ergebnisse in einer lernförderlichen Weise zu nutzen, sondern sie müssen auch nachvollziehen können, wie das Modell zu den Entscheidungen gekommen ist (Niemi, 2021). Denn nur dann können Lehrkräfte KI-basierte Tools auch in ihren Unterricht integrieren und mögliche Rückfragen beantworten (Niemi, 2021). Außerdem müssen noch Fragen bezüglich des Datenschutzs und der Privatsphäre beantwortet werden. Niemi (2021) weist darauf hin, dass weiterer Forschungs- und Diskussionsbedarf besteht, um zu klären, wer welche Informationen für welchen Zweck nutzen darf und wer letztlich die Verantwortung für Entscheidungen im Zusammenhang mit KI-basierten Systemen trägt.

5 Zielsetzung und Erkenntnisinteresse der Untersuchung

Die vorliegende Arbeit untersucht den Einsatz eines Machine-Learning-Modells (ML-Modells) zur Auswertung einer Concept Map und dessen Anwendung als automatisches formatives Assessment. Für diese Untersuchung wird der empirische Teil der Arbeit in zwei Studien aufgeteilt (siehe Abbildung 5.1).

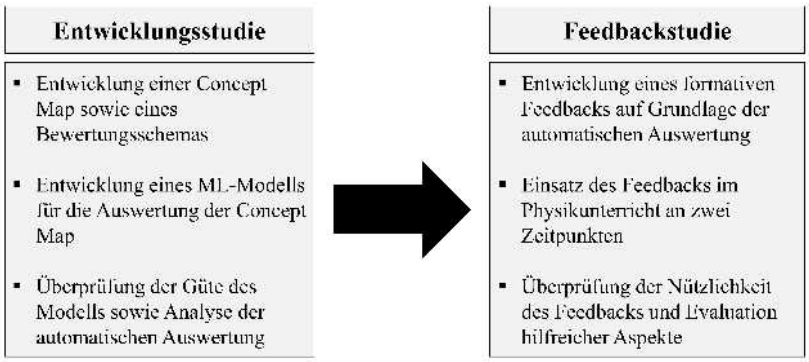


Abbildung 5.1: Aufbau des empirischen Teils der Arbeit

Zunächst werden in der Entwicklungsstudie eine Concept Map sowie ein geeignetes Bewertungsschema entwickelt, welches die spätere Grundlage für ein formatives Assessment sein wird. Das Ziel der Entwicklungsstudie besteht darin, ein Machine-Learning-Modell zu finden, welches die Concept Map automatisch auswerten kann. Damit dieses Modell in den Schulen überhaupt eingesetzt werden kann, wird in der Entwicklungsstudie die Qualität des Machine-Learning-Modells untersucht.
In der darauffolgenden Feedbackstudie wird das entwickelte Modell Teil eines formativen Assessments sein. Dazu wird die automatische Auswertung als Basis für ein Feedback genutzt, welches an zwei Zeitpunkten den Lehrkräften und Lernenden bereitgestellt wird. Der erste Erhebungszeitpunkt wird direkt am Anfang der Lerneinheit liegen, sodass die automatische Auswertung Informationen über das Vorwissen der Lernenden liefern kann. Die zweite Erhebungsphase wird kurz vor Klausur, also eher am Ende der Lerneinheit sein. So kann Feedback

zum aktuellen Wissen sowie der Veränderung zwischen den beiden Zeitpunkten erstellt werden. Den teilnehmenden Lehrkräften wird das Feedback ohne konkrete Handlungsempfehlung zur Verfügung gestellt. Es wird nur darauf hingewiesen, dass das Feedback im Rahmen eines formativen Assessments genutzt werden soll. Das Ziel der Feedbackstudie ist deshalb die Untersuchung der Praxistauglichkeit der automatisch generierten Auswertung sowie der Umgang mit den einzelnen Elementen des Feedbacks.
Im Folgenden werden die Untersuchungsschwerpunkte sowie die sich daraus ergebenden Forschungsfragen zu den beiden Studien erläutert. Zudem werden aus den aufgestellten Forschungsfragen explizite Hypothesen abgeleitet, die durch die erhobenen Daten überprüft werden.

Machine Learning in der Physikdidaktik

Der erste Schwerpunkt ist der Einsatz eines Machine-Learning-Modells im Rahmen einer physikdidaktischen Forschungsarbeit. Die Analyse der aktuellen Studienlage in Kapitel 4 hat verdeutlicht, dass diese Forschungsrichtung noch in den Anfängen steckt. Die vorliegende Arbeit zielt darauf ab, an diesem Punkt anzuknüpfen und neue Erkenntnisse sowie Potenziale für die zukünftige Gestaltung des Physikunterrichts zu erschließen.
Zur automatischen Auswertung der Concept Maps wird ein Machine-Learning-Modell, genauer gesagt ein Klassifikations-Modell, entwickelt (siehe Kapitel 4). Dieser Ansatz ermöglicht eine automatische Zuweisung von vorher festgelegten Kategorien. Um die Performance des Klassifikator-Modells zu evaluieren, wird im ersten Schritt die Übereinstimmung zwischen menschlicher und computergenerierter Auswertung untersucht. Das entwickelte Machine-Learning-Modell wird in der Feedbackstudie als ein formatives Assessment eingesetzt. Es wird demnach keine abschließende Leistungsbeurteilung durchgeführt. Außerdem werden die Ergebnisse des Machine-Learning-Modells nicht für weitreichende Entscheidungen wie die Versetzung in das nächste Schuljahr genutzt. Daher ist es nicht notwendig, eine perfekte Übereinstimmung des Modells mit der menschlichen Bewertung zu erreichen. Beim Blick in die aktuelle Forschungsliteratur wird ebenfalls deutlich, dass eine Übereinstimmung von 100 % kaum erreicht wird (vgl. Kapitel 4). Nakamura et al. (2016) postulierten deshalb eine Übereinstimmung von 70 % als Grundlage für die Nutzung eines Machine-Learning-Modells. Im Review von Zhai et al. (2020) wird eine durchschnittliche Übereinstimmung von $\kappa = 0,72$

festgestellt. Der Übereinstimmungswert zwischen $\kappa = 0,61$ und $\kappa = 0,80$ gilt nach Landis und Koch (1977) als eine beachtliche Übereinstimmung. Daher wird in dieser Arbeit der Wert von $\kappa = 0,70$ als Grenzwert für einen Einsatz in der Schule festgelegt. Durch diese Grenze wird sichergestellt, dass das Modell dem Standard anderer Arbeiten entspricht und ein Großteil der Concept Maps richtig ausgewertet wird. Zudem liegt dieses Kappa in der Mitte des beachtlichen Übereinstimmungsbereiches, was eine zuverlässige Anwendung in der Schule erwarten lässt.
Die Concept Maps sollen aber nicht auf einer graphentheoretischen Ebene ausgewertet werden. Um ein zielgerichtetes individuelles Feedback zu ermöglichen, bietet sich eine Auswertung der einzelnen Propositionen an, da hier die wertvollsten Informationen für ein formatives Assessment liegen (Trumpower & Sarwar, 2010). Wie in Kapitel 4 beschrieben, ist die Entwicklung eines Machine-Learning-Modells ein iterativer Prozess, bei dem verschiedene Modelle trainiert und getestet werden. Das Ziel ist es, ein Machine-Learning-Modell zu entwickeln, welches die geforderte Übereinstimmung erreichen kann. Daraus lässt sich die erste Forschungsfrage formulieren:

Forschungsfrage 1

Kann eine Übereinstimmung von mindestens $\kappa = 0,70$ bei einer Bewertung der Propositionen einer Concept Map zwischen der menschlichen Bewertung und der durch Techniken des maschinellen Lernens generierten Bewertung erreicht werden?

Wenn ein Machine-Learning-Modell gefunden wurde, das die Anforderungen an die Übereinstimmung erfüllt, wird dieses Modell anschließend in der Feedbackstudie an verschiedenen Schulen zu zwei Zeitpunkten eingesetzt (siehe Abbildung 5.1). Bevor das Modell jedoch eingesetzt werden kann, wird es noch weiter analysiert. Falls mehrere Modelle die Anforderung an die Übereinstimmung erfüllen, wird diese Analyse für eine nachvollziehbare Entscheidung für eins der Modelle genutzt. Die weiteren Forschungsfragen und Hypothesen beziehen sich also auf die Machine-Learning-Modelle, die eine Übereinstimmung von $\kappa = 0,70$ aufweisen konnten.
Die entwickelten Machine-Learning-Modelle werden mittels Trainingsdaten, die in der Entwicklungsstudie erhoben werden, trainiert und getestet. Die Trainingsdaten und damit die Grundlage des Machine-Learning-Modells werden in der

Entwicklungsstudie am Ende der Lerneinheit erhoben. Daher kann davon ausgegangen werden, dass die sprachlichen Praktiken und der inhaltliche Gehalt der Concept Maps aus der Feedbackstudie im Vergleich zum Trainingsdatensatz variieren können. Daraus lässt sich folgende Hypothese aufstellen:

Hypothese 1

Das entwickelte Machine-Learning-Modell wird nach dem Trainingsprozess eine niedrigere Übereinstimmung mit menschlichen Bewertungen aufweisen, wenn es neue Concept Maps auswertet, die das Vorwissen abfragen, im Vergleich zu Concept Maps, die am Ende der Lerneinheit erstellt wurden.

Gleichzeitig wird eine inhaltliche Überprüfung der Machine-Learning-Ergebnisse durchgeführt. Dazu werden primär die falsch klassifizierten Antworten genauer betrachtet und auf Fehlermuster untersucht. Diese Analyse soll eine umfassendere Aussage über die Qualität des Modells zulassen. Deshalb lautet die zweite Forschungsfrage:

Forschungsfrage 2

Können spezifische Strukturen in dem Antwortverhalten der Lernenden identifiziert werden, die einen Einfluss auf die Performance des Machine-Learning-Modells bei der automatischen Auswertung der Concept Maps haben?

Die Darstellung des Trainings eines Machine-Learning-Modells, das zur Auswertung von Textdaten genutzt wird, hat gezeigt, dass schriftliche Antworten immer in eine numerische Repräsentation umgewandelt werden müssen (siehe Kapitel 4). Dazu stehen diverse Methoden wie die transformatorbasierten Sprachmodelle zur Verfügung (Birunda & Devi, 2021). Diese Modelle sind auf einem großen Textkorpus trainiert, um die Semantik einzelner Wörter und ganzer Sätze zu erfassen und zu verarbeiten. Auch in dieser Arbeit wird ein solches Modell genutzt werden, da die Betrachtung in Kapitel 4 gezeigt hat, dass damit gute Ergebnisse bei der Klassifikation erwartet werden können. Außerdem steht eine Vielzahl von vortrainierten Modellen zur Verfügung, die bereits in einer Vielzahl von Arbeiten eingesetzt wurden und eine einfache Verwendung ermöglichen.
In der Physik sind zur Beschreibung von Zusammenhängen aber auch Formeln, Zahlen oder Einheiten relevant. Deshalb kann davon ausgegangen werden, dass

in den Concept Maps ebenfalls Propositionen mit Formeln oder Zahlen enthalten sind. Die vortrainierten Sprachmodelle verfügen jedoch über ein festes Vokabular, das während des Trainings auf natürlicher Sprache gelernt wurde. Da kein eigenes Sprachmodell genutzt wird, können auch keine neuen Vokabeln wie Formeln und Zahlen direkt in das Vokabular aufgenommen werden. Im Gegensatz zu natürlicher Sprache haben Formeln und Zahlen auch keine grammatikalische Bedeutung. Das lässt die Hypothese zu, dass die automatische Auswertung von Propositionen, die strukturierte Daten wie Formeln oder Zahlen enthalten, eine Herausforderung für das Machine-Learning-Modell darstellen können:

Hypothese 2

Die automatische Auswertung von Propositionen, die Formeln oder Zahlen enthalten, führt zu einer geringeren Performance des Machine-Learning-Modells im Vergleich zu Propositionen, die ausschließlich textuellen Inhalt aufweisen.

Als abschließende Betrachtung der Qualität des Machine-Learning-Modells wird auch auf den Aspekt der Verzerrung eingegangen. Durch die Betrachtung des Verzerrung-Varianz-Dilemmas (siehe Abschnitt 4.3.2) ist deutlich geworden, dass Machine-Learning-Modelle systematische Fehler machen können. Cheuk (2021) stellte fest, dass die Verzerrung von Machine-Learning-Modellen in vielen Arbeiten zu wenig Beachtung geschenkt wird und speziell bei Arbeiten im Bildungskontext ein wichtiger Punkt ist. Es ist daher wichtig sicherzustellen, dass die automatische Auswertung nicht durch das Geschlecht oder die schulischen Leistungen der Lernenden beeinflusst wird. Eine ungleichmäßige oder voreingenommene Auswertung aufgrund dieser Faktoren könnte zu einer Benachteiligung bestimmter Lernender führen. Weiterhin sollte die automatische Auswertung objektiv und zuverlässig sein, um valide und konsistente Bewertungen der Propositionen der Lernenden zu ermöglichen. Angesichts dessen wird nicht nur die Struktur des Trainingsdatensatzes analysiert, sondern auch die Klassifikationen des Modells mit Merkmalen der Lernenden verglichen. Daher lautet die nächste Forschungsfrage:

Forschungsfrage 3

Zeigt die automatische Auswertung der Concept Maps durch ein Machine-Learning-Modell eine Verzerrung (Bias) in Bezug auf das Geschlecht der Lernenden sowie schulische Leistungen in den Fächern Mathematik, Physik und Deutsch?

Formatives Assessment und Feedback

Der zweite Schwerpunkt des Erkenntnisinteresses dieser Arbeit ist das formative Assessment. Die Ausarbeitung in Kapitel 2 hat gezeigt, dass das formative Assessment eine der effektivsten Methoden zur Optimierung schulischen Lernens ist. Dabei wurde besonders auf die fünf Schlüsselstrategien von Wiliam und Thompson (2008) eingegangen. Dieser Teil der Arbeit wird sich auf die dritte Schlüsselstrategie, die lernförderliche Rückmeldung, fokussieren.
Denn Souvignier und Hasselhorn (2018) stellten fest, dass es weitere Forschungsarbeiten geben muss, „*in denen konkrete Maßnahmen zur Umsetzung des Konzepts formativen Assessments praktisch erprobt und wissenschaftlich evaluiert werden*“ (Souvignier & Hasselhorn, 2018, S. 695). Außerdem konnte ausgearbeitet werden, dass eine zeitnahe effektive Rückmeldung ein großes Problem für Lehrkräfte darstellen kann (siehe Kapitel 2).
Durch die automatische Auswertung sollen die Lehrkräfte kaum zusätzliche Zeit für die Auswertung der Concept Maps aufwenden müssen, sodass ein zeitnahes formatives Assessment möglich ist. Außerdem sind die Lehrkräfte durch die automatische Auswertung in der Lage, ihren Lernenden eine schnelle und einfache (individuelle) Rückmeldung geben zu können.
Um eine Grundlage für die Bewertung der automatischen Auswertung zu haben, werden zunächst die generelle Einstellung und Nutzungshäufigkeit der Lehrkräfte bezüglich formativer Assessments erhoben. Dies ermöglicht die Erhebung eines Ausgangspunkts und liefert erste Ergebnisse bezüglich der Vor- und Nachteile des formativen Assessments im Physikunterricht. Die dazugehörige Forschungsfrage lautet:

Forschungsfrage 4

In welchem Umfang nutzen die teilnehmenden Lehrkräfte formatives Assessment in ihrem Physikunterricht?

Aufgrund der dargestellten Problematiken des formativen Assessments in Abschnitt 2.4 wie des Zeitmangels oder der Schwierigkeit bei einer großen Klasse, jedem Lernenden ein individuelles Feedback zu ermöglichen, werden trotz der hohen Lernwirksamkeit selten formative Assessments eingesetzt (Black & Wiliam, 1998). Daher kann davon ausgegangen werden, dass auch die teilnehmenden Lehrkräfte dieser Studie wenig formative Assessment-Methoden anwenden. Das lässt die folgende Hypothese zu:

Hypothese 3

Die teilnehmenden Lehrkräfte berichten über ein recht eingeschränktes Spektrum an formativen Assessment-Methoden.

Automatische Feedbacksysteme können Lehrkräften viel Arbeit abnehmen. Allerdings hat die Ausarbeitung in Abschnitt 2.3.3 gezeigt, dass die Qualität dieser Systeme kritisch hinterfragt werden muss. Deshalb forderten Cavalcanti et al. (2021) mehr Studien, die sich mit der genauen Analyse von automatischem Feedback befassen sollen. Zudem formulierten Zhai et al. (2020) in ihrem Überblicksartikel die Forschungslücke, dass zu wenig Arbeiten die explizite Anwendung von Machine-Learning-Modellen im Unterricht fokussieren. Deshalb lautet die nächste Forschungsfrage:

Forschungsfrage 5

Inwiefern kann das automatische Feedback dazu beitragen, formatives Assessment erfolgreich in den Physikunterricht zu integrieren?

Das automatische Feedback wird aus mehreren unterschiedlichen Aspekten bestehen. Es werden Informationen sowohl auf Klassen- als auch auf einer Individualebene den Lehrkräften zur Verfügung gestellt. Es kann deshalb evaluiert werden, welche Elemente der automatischen Auswertung als besonders unterstützend empfunden wurden. Weiterhin kann analysiert werden, zu welchem Zwecke die Lehrkräfte die bereitgestellten Informationen genutzt haben. Da eine individuelle Rückmeldung für jeden Lernenden sehr zeitintensiv sein kann und die Lehrkräfte durch die automatische Auswertung einfach und schnell Informationen über jeden Lernenden erhalten, kann erwartet werden, dass die automatische Auswertung auch für individuelle Rückmeldungen genutzt wird. Durch die automatische Auswertung kann demnach die Problematik des zeitintensiven formativen Assessments

gelöst werden. Deshalb wird die folgende Hypothese formuliert:

Hypothese 4
Die Lehrkräfte nutzen die automatische Auswertung für eine gezielte individuelle Rückmeldung.

Die Lernenden können aber nicht nur ein individuelles Feedback über die Lehrenden erhalten. Das Machine-Learning-Modell erstellt nach dem zweiten Zeitpunkt in der Feedbackstudie auch für die Lernenden ein eigenes direktes Feedback. Da aufgrund der großen Anzahl der Lernenden in einer Klasse und des angesprochenen zeitlichen Aufwands die Lernenden kaum individuelle Feedbacks erhalten, kann davon ausgegangen werden, dass die Lernenden die automatische Auswertung und das individuelle Feedback positiv bewerten. Deswegen lautet die nächste Hypothese:

Hypothese 5
Die Lernenden werden die automatische Auswertung positiv bewerten, da sie dadurch ein individuelles Feedback erhalten.

Concept Maps

Kapitel 3 hat gezeigt, dass Concept Maps als eine gute formative Assessment-Methode angesehen werden, da sie reichhaltige Informationen über den Leistungsstand der Lernenden liefern (Buldu & Buldu, 2010; Hartmeyer et al., 2018; Ruiz-Primo & Shavelson, 1996). Allerdings gelten Concept Maps auch als sehr herausfordernd, sowohl für leistungsstarke als auch für leistungsschwache Lernende (siehe Kapitel 3). Daher wird die eingesetzte Concept Map untersucht und die daraus entstandenen Vor- und Nachteile analysiert. Die dazugehörige Forschungsfrage lautet:

Forschungsfrage 6
Was sind die wahrgenommenen Herausforderungen und potenziellen Nutzen, die Lernende bei der Erstellung der Propositionen in den Concept Maps erleben?

Concept Maps liefern nicht nur wichtige Informationen, die Lehrkräfte für ein formatives Assessment nutzen können, sondern sie geben den Lernenden auch

einen Überblick über ihr eigenes Wissen. Bei der Bearbeitung einer Concept Map müssen die relevanten Zusammenhänge erschlossen werden, was wiederum potenzielle Wissenslücken offenbaren kann. Daher wird ermittelt, ob auch mit der in dieser Studie gewählten Concept Map die Lernenden in der Lage waren, ihre eigenen Wissensstrukturen besser zu erkennen. Die Hypothese lautet:

Hypothese 6
Die Erstellung der Concept Maps fördert die Selbstreflexion der Lernenden über ihr eigenes Wissen.

Für die automatische Auswertung müssen die Concept Maps in einer digitalen Form vorliegen. Aus ökonomischen Gründen ist es deshalb sinnvoll, die Concept Maps direkt digital zu erheben. Dies hat den Vorteil, dass die Antworten nicht erst digitalisiert werden müssen und direkt weiter verarbeitet werden können. Außerdem bietet eine digitale Erhebung den Vorteil, dass Interaktionsmuster durch Log-Daten aufgezeichnet werden können. Durch die Log-Daten können Einblicke in die Bearbeitung der Lernenden gewonnen werden, die bei einer Erhebung mit Stift und Papier nur schwer messbar wären. Es können Rückschlüsse auf Probleme bei der Bearbeitung gezogen werden und wichtige Erkenntnisse über den Aufbau der genutzten Concept Map gewonnen werden. Allerdings hat Abschnitt 3.3 auch gezeigt, dass es keine feste Leserichtung bei einer Concept Map gibt, was unter anderem eine schnelle und einfache Auswertung erschwert (Ley, 2015). Es kann daher davon ausgegangen werden, dass auch die Bearbeitungsschritte der Lernenden variieren werden. Deshalb lässt sich folgende Hypothese formulieren:

Hypothese 7
In den anfänglichen Bearbeitungsschritten der Lernenden lassen sich keine Muster identifizieren.

Eine Vielzahl von Studien hat gezeigt, dass Concept Maps für unterschiedliche Assessmentzwecke eingesetzt werden können (Buldu & Buldu, 2010; Hartmeyer et al., 2018; Ruiz-Primo & Shavelson, 1996). In der Feedbackstudie werden Concept Maps zum Beginn und zum Ende einer Lerneinheit erhoben. Da an beiden Erhebungszeitpunkten dieselben Lernenden die Concept Maps bearbeiten werden, kann eine zeitliche Veränderung untersucht werden. Dies führt zu der nächsten Forschungsfrage:

Forschungsfrage 7

Wie entwickeln sich die Concept Maps der Lernenden zwischen dem ersten und zweiten Erhebungszeitpunkt?

Die Lernenden werden die Concept Map an beiden Zeitpunkten unter identischen Bedingungen bearbeiten. Wie bereits beschrieben, liegt der erste Erhebungszeitpunkt direkt zu Beginn und der zweite Erhebungszeitpunkt eher am Ende der Lerneinheit. Es kann somit angenommen werden, dass die Lernenden im Verlauf des Unterrichts einen Lernzuwachs erfahren und bei der Bearbeitung der zweiten Concept Map weniger Fehler machen. Dies lässt die folgende Hypothese zu:

Hypothese 8

Die Anzahl der falschen Propositionen wird beim zweiten Erhebungszeitpunkt signifikant geringer sein als beim ersten.

6 Entwicklungsstudie

Das Ziel der Entwicklungsstudie ist die Entwicklung eines Machine-Learning-Modells, das in der späteren Feedbackstudie eingesetzt wird. Um dieses Ziel zu erreichen, ist die Entwicklungsstudie in drei aufeinanderfolgende Phasen untergliedert (siehe Abbildung 6.1). In der ersten Phase in Abschnitt 6.1 wird zunächst eine geeignete Concept Map konzipiert sowie ein passendes Bewertungsschema entwickelt. In Phase zwei (Abschnitt 6.2) wird eine erste Schulerhebung durchgeführt. Dafür wird die entwickelte Concept Map in mehreren Klassen eingesetzt und so Daten für die Entwicklung des Machine-Learning-Modells gesammelt. Anschließend werden mit diesen Daten mehrere Machine-Learning-Modelle für die automatische Auswertung entwickelt und getestet. An dieser Stelle wird überprüft, ob eines dieser Modelle die geforderte Übereinstimmung von $\kappa = 0,70$ erreicht hat. So kann eine erste Auswahl an Modellen erfordern, die dann anschließend in Phase drei (Abschnitt 6.3) näher analysiert werden. Es wird dabei eine mögliche Verzerrung betrachtet und die automatische Auswertung auf Fehlermuster untersucht.
Die Entwicklungsstudie betrachtet demnach den ersten Schwerpunkt des Erkenntnisinteresses dieser Arbeit. In Abschnitt 6.4 werden die Ergebnisse zusammengefasst und die aufgestellten Forschungsfragen und Hypothesen aus Kapitel 5 beantwortet und diskutiert. Abschließend werden Schlussfolgerungen für die anschließende Feedbackstudie gezogen.

Abbildung 6.1: Ablauf der Entwicklungsstudie in drei Phasen

6.1 Phase 1: Inhaltliche Vorbereitung der Studie

In diesem Abschnitt werden einige wichtige Grundlagen für die vorliegende Arbeit festgelegt. Um ein Machine-Learning-Modell zu entwickeln, das eine Concept Map automatisch auswerten kann, muss zunächst eine geeignete Concept Map konzipiert werden. Wie in Kapitel 3 bereits beschrieben wurde, gibt es unterschiedliche Concept Maps, die von sehr offenen bis zu vollständig vorgegebenen Strukturen variieren können. Deshalb soll in der ersten Phase, speziell mit dem Blick auf die Machine-Learning-Modelle, eine Concept Map ausgewählt werden. Dazu werden zwei Concept-Map-Studien betrachtet und analysiert. Die daraus gewonnenen Ergebnisse werden als Grundlage für das Design der in der vorliegenden Arbeit verwendeten Concept Map dienen. Zudem wird ein Bewertungsschema festgelegt, welches die Grundlage der automatischen Auswertung und des späteren Feedbacks darstellt. Abschließend wird im Rahmen einer curricularen Analyse das inhaltliche Thema der Concept Map bestimmt und diskutiert.

6.1.1 Analyse der Studie „Physik und Physiologie“

Kubin (o. D.) führte analog zur Studie von Plomer (2011) eine Concept-Map-Studie zum Thema Physik und Physiologie (im weiteren Verlauf PhyPhy) durch. Plomer (2011) entwickelte Alternativexperimente für das physikalische Praktikum der Medizinstudierenden der Ludwig-Maximilians-Universität München. Dabei untersuchte er die Lernwirksamkeit der neu konzipierten Experimente mithilfe eines selbst entwickelten Wissenstests. Der Test bestand dabei aus einer Concept Map und Multiple-Choice-Fragen aus dem Bereich Elektrizitätslehre und Physiologie der Nervenzellen (Plomer, 2011, S. 2). Insgesamt nahmen ca. 300 Studierende teil, die in zwei Gruppen aufgeteilt wurden. Die Kontrollgruppe absolvierte die traditionellen Experimente, wohingegen die Treatmentgruppe neue adressatenspezifische Experimente absolvieren musste. Da ein Pre-Test aus organisatorischen Gründen nicht durchgeführt werden konnte, wurde der beschriebene Wissenstest nur als Post-Test eingesetzt (Plomer, 2011, S. 34).
Die Studie von Kubin (o. D.) hatte ähnliche Ziele und ein ähnliches Studiendesign und wurde an der Medizinischen Hochschule Hannover durchgeführt. Insgesamt nahmen 324 Studierende der Human- und Zahnmedizin (22,4 Jahre im Schnitt; 205 weiblich, 115 männlich, 4 keine Angabe) an der Untersuchung teil. Da für

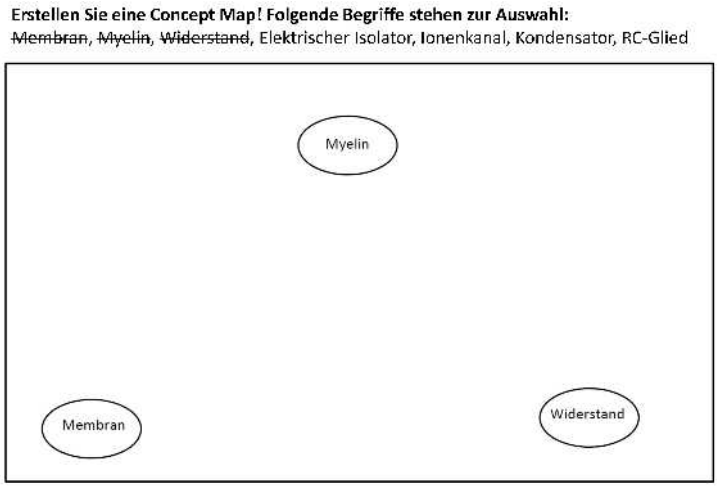


Abbildung 6.2: Verwendete Concept Map der PhyPhy-Studie zum Thema Elektrizitätslehre und Physiologie der Nervenzellen (Kubin, o. D.)

diese Arbeit nur die Concept Map von Relevanz ist, werden die anderen Tests der Studie im Folgenden nicht weiter betrachtet.

Als Einführung in die Thematik Concept Map bekamen die Studierenden zunächst eine Beispiel Concept Map zum Thema Tiere mit den Begriffen Fell, Maus, Katze, Elefant und Sendung mit der Maus. Die dazugehörigen Propositionen wurden mit Zahlen codiert und unter der Concept Map ausgeschrieben, z. B. Eine Maus hat ein Fell (Kubin, o. D. Plomer, 2011).

Die eigentliche Concept-Map-Aufgabe bestand aus einer fast leeren Concept Map (siehe Abbildung 6.2). Die Begriffe Membran, Myelin, Widerstand, elektrischer Isolator, Ionenkanal, Kondensator, RC-Glied wurden als Liste bereitgestellt, wobei die Begriffe Myelin, Membran und Widerstand als Konzepte in der Concept Map bereits vorgegeben waren (siehe Abbildung 6.2). Die dazugehörigen Propositionen sollten von den Studierenden separat auf der Rückseite des Blattes erstellt und innerhalb der Concept Map mit einer Nummer versehen werden. Beispiele für erstellte Propositionen der PhyPhy-Studie sind in Tabelle 6.1 dargestellt.

Kurzform	Proposition
KoWi	Ein Kondensator hat einen Widerstand
RCIo	Die Ionenkanäle sind in Reihe geschaltet und somit mit einem RC Glied vergleichbar
MeKo	Eine Membran wirkt wie ein Kondensator
KoeI	Im Kondensator ist ein Dielektrikum welches elektrisch isoliert
MeIo	Eine Membran besitzt Ionenkanäle

Tabelle 6.1: Auszüge von Propositionen aus dem Datensatz der PhyPhy-Studie (Kubin, o. D.)

Die erstellten Propositionen wurden mithilfe eines Bewertungsschemas analysiert, das aus sechs Kategorien bestand, wobei die inhaltliche Korrektheit der Propositionen im Vordergrund stand (Plomer, 2011, S. 44):

K1 Zwischen zwei Begriffen wurde weder eine Linie eingezeichnet noch findet sich eine Erklärung dazu.

K2 Zwei Begriffe wurden mit einer Linie verknüpft, die Erklärung dazu ist nicht richtig.

K3 Zwei Begriffe wurden mit einer Linie verknüpft und der Zusammenhang korrekt erklärt.

K4 Zwei Begriffe wurden mit einer Linie verknüpft, allerdings wurde keine Erklärung formuliert.

K5 Zwei Begriffe wurden mit einer Linie verknüpft, die Erklärung dazu ist nicht richtig, findet sich in dieser Formulierung jedoch häufiger.

K6 Zwischen zwei Begriffen wurde keine Linie eingezeichnet, allerdings geht aus den restlichen Antworten offensichtlich hervor, dass Wissen um diesen Zusammenhang vorhanden ist und dieser auch formuliert werden kann.

Auf Grundlage dieses Kategoriensystems wurden unterschiedliche Scores gebildet, um die Concept Maps zu analysieren. So nutzte Plomer (2011) unter anderem einen Score für die Concept Map $S_{Map} \in \{0, \ldots, 19\}$, der die Anzahl der korrekten Antworten abbildet oder einen Score, der die korrekten Propositionen aus dem Bereich Physik zählt $S_{Physik} \in \{0, \ldots, 5\}$.
Für die weitere Analyse der Ergebnisse aus der PhyPhy-Studie soll das Kategoriensystem vereinfacht werden, da bei einer Machine-Learning-Auswertung nur die Propositionen relevant sind, die von den Studierenden inhaltlich vollständig erstellt wurden. Deshalb können die Kategorien K1, K4 und K6 aus dem Datensatz gelöscht werden. Die Kategorien K2 und K5 lassen sich als eine neue Kategorie *falsch* und die Kategorie K3 als Kategorie *richtig* zusammenfassen.
Für den Datensatz aus der PhyPhy-Studie von Kubin (o. D.) bedeutet das, dass 2.660 unterschiedliche Propositionen in den beiden neu gebildeten Kategorien verbleiben. Betrachtet man die Verteilung der Kategorien *richtig* und *falsch*, fällt auf, dass der Datensatz ungleichmäßig verteilt ist. Von den 2.660 Antworten im Datensatz wurden 89 % als *richtig* bewertet und lediglich 11 % als *falsch*. Die

Studierenden haben mehrere Propositionen formuliert, die von den Bewertern fast ausschließlich als korrekte Antworten bewertet wurden. So findet man für die Propositionen Ionenkanal – Kondensator, Ionenkanal – RC-Glied und Ionenkanal – Widerstand weniger als 3 % der Antworten in der Kategorie *falsch* wieder. Hinzu kommt, dass die Häufigkeiten der gebildeten Propositionen deutlich variieren. So findet man Zusammenhänge wie Membran – Ionenkanal (MeIo) oder Myelin – elektrischer Isolator (MyeI), welche über 250-mal gebildet wurden, jedoch auch Propositionen wie elektrischer Isolator – RC-Glied (eIRC) oder Ionenkanal – RC-Glied (IoRC), die eine Häufigkeit von unter 50 Antworten aufweisen (siehe Abbildung 6.3).

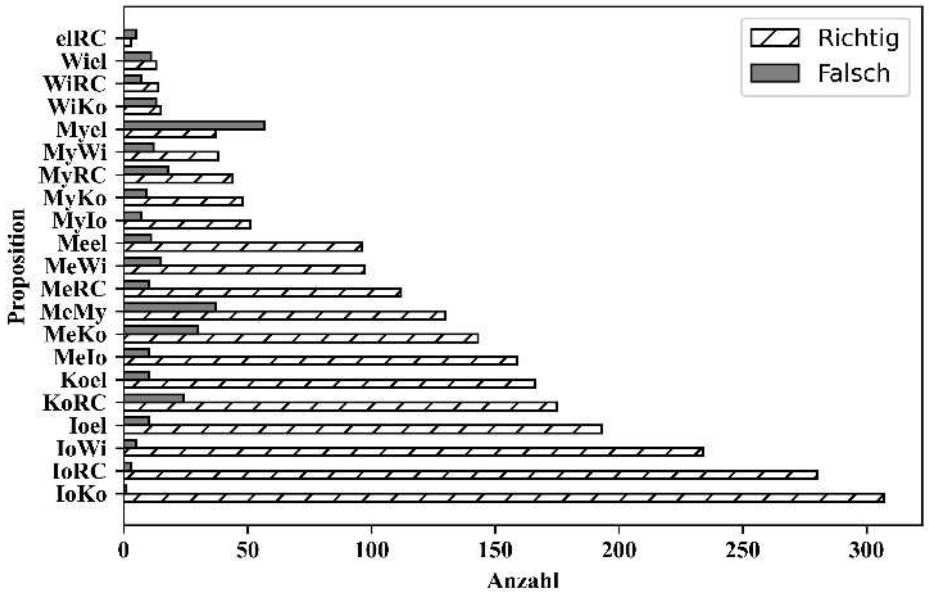


Abbildung 6.3: Antworthäufigkeiten der erstellten Propositionen aus der PhyPhy-Studie für die Kategorien richtig & falsch

6.1.2 Analyse der Studien von Friege

Der zweite Datensatz, der für die Entwicklung der Concept Map analysiert wird, stammt aus der Studie von Friege und Lind (2024). Der zentrale Kern dieser Forschungsarbeit liegt auf einer empirischen Untersuchung zum Thema Problemlösen auf Grundlage eines Experten-Novizenvergleichs (Friege, 2001, S. 2). Im Rahmen der empirischen Untersuchung und zur Beantwortung der Forschungsfragen

wurden unterschiedliche Testinstrumente entwickelt und angewandt. Darunter auch Concept Maps und der anschließende Vergleich zwischen Experten- und Novizen-Maps.
Im Folgenden werden nur die Ergebnisse einer Concept-Map-Auswertung zum Thema Himmelsmechanik berichtet. Es wurde eine Liste von 15 Begriffen (Masse, Planet, Gravitationsfeld, Energie, Sonne, Kraft, Beschleunigung, Bewegung, Drehimpuls, Geschwindigkeit, Umlaufzeit, Entfernung, Komet, Keplersche Gesetze und Gravitationspotenzial) zur Verfügung gestellt und die Concept Map musste eigenständig mit Papier und Stift erstellt werden. Zur Bewertung der einzelnen

Nr.	Kategorie	Beispiel
1	Ober-/Unterbegriffsrelation	„ist ein“, „gehört zu“
2	Charakteristisches Merkmal	„hat“, „besitzt“
3	Aktivitätsmerkmal	„macht“, „kreist um“, „bewegt sich“
4	Funktionsrelation – qualitativ	„ist abhängig von“, „bestimmt“
5	Funktionsrelation – quantitativ	„F=ma“, oder auch „F ist prop. zu“
6	Relationen ohne bestimmte Qualität	„hat zu tun mit“. Alle unbeschriftete Knoten und die nicht zu Kat. 1–5 passen.
7	Falsche Relation	Sinnlose Kantenbeschriftung

Tabelle 6.2: Bewertungsschema aus den Studien von Friege (z. B. 2001)

Concept Maps wurde ein System aus sieben verschiedenen Kategorien genutzt (siehe Tabelle 6.2). Dabei wurde das Bewertungsschema aus der Arbeit von Fischler und Peuckert (2000) adaptiert. In diesem Kategoriensystem (siehe Tabelle 6.2) werden die Propositionen auf einer qualitativen Ebene eingeordnet. Die Kategorien 1 bis 6 sind Differenzierungen von Propositionen, die alle zumindest physikalisch korrekt sind, aber unterschiedliche Qualitäten besitzen. So werden eher einfache Aussagen wie „... hat ein ...“ anders bewertet, als funktionale Zusammenhänge wie „... $s = v \cdot t$ ...“. Hieraus ergibt sich die Möglichkeit einer deutlich detaillierteren Auswertung der Concept Maps, die man für eine ausführlichere Rückmeldung nutzen kann als die reine Betrachtung von richtigen und falschen Propositionen. Die Kategorie 7 wird für physikalisch falsche und sinnlose Kantenbeschriftungen genutzt.
In der hier analysierten Studie haben 135 Lernende eine Concept Map erstellt. Der Datensatz besteht dabei aus 2.741 unterschiedlichen Propositionen. Von den 15 möglichen Begriffen wurden im Schnitt 14 Begriffe und 20 unterschiedliche Propo-

sitionen für die Concept Maps genutzt, wobei der Begriff „Kraft“ in jeder erstellen Concept Map zu finden war. Auch die Begriffe „Beschleunigung“, „Entfernung“, „Masse“ und „Planet“ wurden vielfach verwendet. Betrachtet man eine Concept

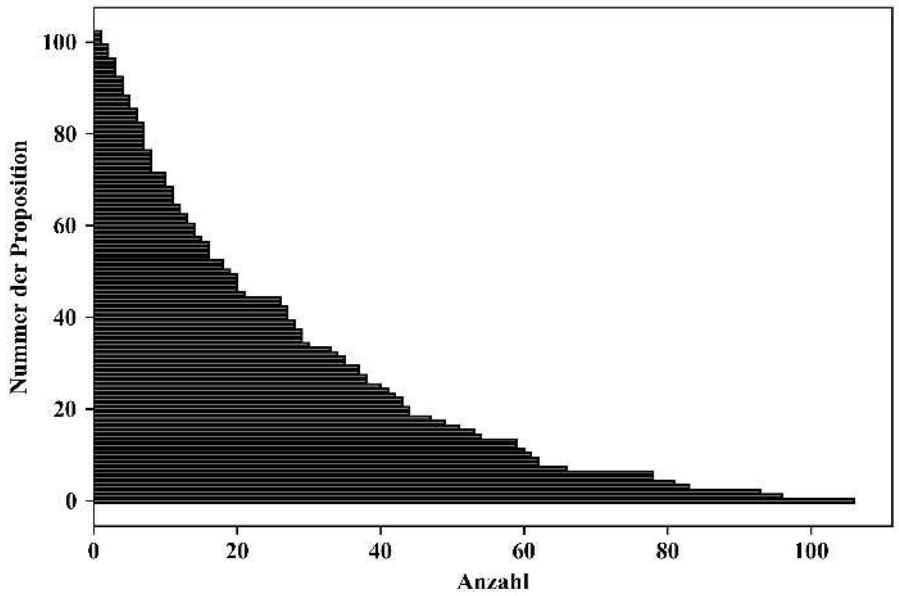


Abbildung 6.4: Häufigkeitsverteilung der Propositionen der Himmelsmechanik-Concept-Maps

Map mit n Begriffen, können $\frac{n(n-1)}{2}$ mögliche Propositionen erstellt werden, da eine Verknüpfung mit demselben Begriff ausgeschlossen ist und die Richtung der Verbindungspfeile keine Rolle spielt (siehe Kapitel 3). Für die Himmelsmechanik-Concept-Map bedeutet das, dass 105 verschiedene Propositionen hätten gebildet werden können. Analysiert man die erstellen Concept Maps genauer, stellt man fest, dass aus den 15 vorgegebenen Begriffen 103 unterschiedliche Propositionen gebildet wurden (siehe Abbildung 6.4). Im Mittel wurde jede Proposition 27-mal erstellt, wobei die Häufigkeiten stark schwanken. Einige Propositionen wie „Kraft-Beschleunigung“ oder „Masse-Kraft“ wurden von vielen Lernenden in ihrer Concept Map verwendet. Allerdings wurden 83 % der Propositionen von weniger als 50 Lernenden gebildet und 31 der 103 erstellten Propositionen wurden sogar nur von einer einstelligen Anzahl von Lernenden verwendet.
Die Verteilung der sieben Ratingkategorien in Abbildung 6.5 zeigt, dass die Kategorien 4 und 5 am häufigsten im Datensatz vorkommen. Mit 886 von den 2.741 Propositionen sind demnach knapp ein Drittel qualitative Funktionsrelationen, wie

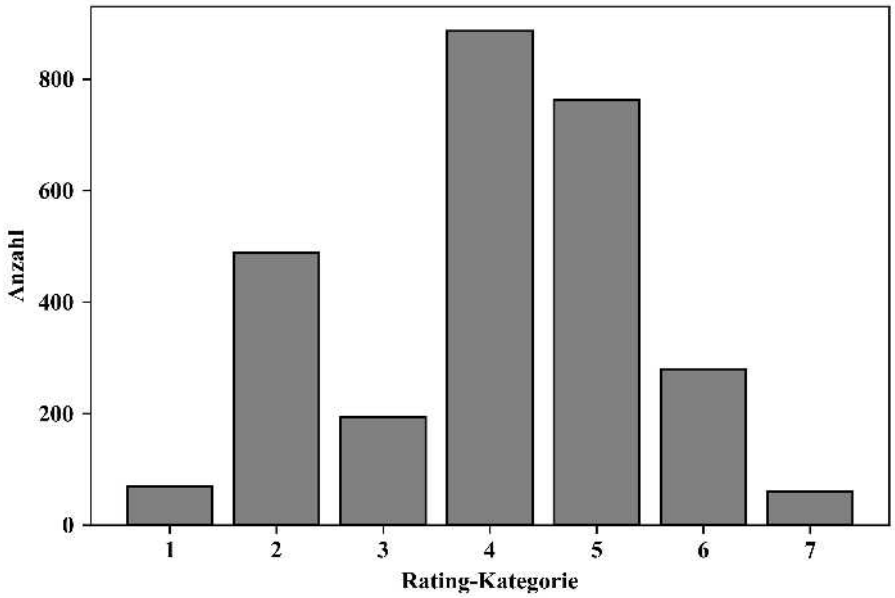


Abbildung 6.5: Verteilung der Ratingkategorien der Himmelsmechanik-Concept-Maps mit dem Bewertungsschema aus den Studien von Friege (z. B. 2001)

„Die Umlaufzeit hängt von der Geschwindigkeit ab.“ oder „Die Sonne erzeugt Gravitationsfeld.“. Auch die quantitativen Funktionsrelationen sind mit 763 Propositionen im Datensatz häufig vertreten. Typische Beispielantworten aus dieser Kategorie sind „Masse – $E = m \cdot c^2$ – Energie“ oder „Kraft – $F = m \cdot a$ – Beschleunigung“. Die Ergebnisse decken sich mit den häufig gebildeten Propositionen, da insbesondere bei „Masse-Kraft“ oder „Kraft-Beschleunigung“ einfache funktionale Beziehungen gebildet werden können. Weniger findet man Antworten aus den Kategorien *charakteristisches Merkmal* (488) oder *Relationen ohne bestimmte Qualität* (280) im Datensatz. Die Kategorie mit der niedrigsten Häufigkeit ist jedoch die Kategorie *falsche Relationen* mit lediglich 60 Antworten. Fasst man die Kategorien wieder in zwei Kategorien *richtig* und *falsch* zusammen, zeigt sich eine erneute starke Ungleichverteilung im Datensatz mit mehr richtig gebildeten Propositionen. Lediglich 2 % der Propositionen wurden von den Bewertern als falsche und demnach 98 % als richtige Propositionen eingestuft.

6.1.3 Schlussfolgerungen

Auf Grundlage der Analyse der PhyPhy-Studie und den Himmelsmechanik-Concept-Maps sowie weiterer Literatur wird das Design der in der Hauptstudie

verwendeten Concept Map festgelegt. Ebenso wird ein Bewertungssystem definiert, das später für die Auswertung der Concept Maps verwendet werden kann. Abschließend werden potenzielle inhaltliche Themen für die Concept-Map-Studie diskutiert.

Concept Map Format

Die PhyPhy-Studie hat ein recht offenes Concept-Map-Design gewählt, bei dem die Studierenden aus einer Liste von sieben Begriffen zum Thema Physik und Physiologie wählen konnten. Die Bewertung der erstellten Concept Maps konzentrierte sich hauptsächlich auf die inhaltliche Richtigkeit der Propositionen. Aus der Analyse wurde deutlich, dass die Studierenden vorwiegend inhaltlich korrekte Propositionen erstellten und bestimmte Propositionen priorisiert haben.
In den Studien von Friege (z. B. 2001) konnten die Lernenden ebenfalls aus einer Liste von Begriffen wählen, um so Propositionen für ihre Concept Map zu bilden. Der Unterschied zwischen den beiden Designs ist, dass in den Studien von Friege (z. B. 2001) prinzipiell die Möglichkeit bestand, weitere Begriffe hinzuzufügen und keine der Begriffe in der Concept Map bereits integriert waren (Friege, 2001, S. 114).
Beide Concept-Maps-Designs sind sowohl von dem Inhaltsaspekt als auch von dem Freiheitsaspekt im mittleren Bereich des Kontinuums von Concept Maps anzuordnen, wobei die Himmelsmechanik-Concept-Map einen etwas höheren Freiheitsgrad hat (siehe Abbildung 6.6). Den Vorteil solcher Designs sieht Plomer (2011) wie folgt: „*Durch eine Vorgabe von Begriffen wird nicht nur der Umfang der zu erstellenden Concept Map nach oben hin begrenzt, sondern es steht auch der Großteil der mentalen Leistung für das Formulieren der Propositionen zur Verfügung*“ (Plomer, 2011, S. 36).
Die gewählten Designs resultieren allerdings aus einem sehr heterogenen Datensatz, bei dem viele unterschiedliche Concept Maps erstellt wurden. Für eine Auswertung mittels Machine Learning sind die beiden Studien deshalb kaum geeignet. Aufgrund der Vielzahl an unterschiedlichen Propositionen, von denen einige nur selten vorkommen, gestaltet sich das Trainieren eines Machine-Learning-Modells als äußerst schwierig.
Bei der Betrachtung anderer Studien aus dem Bereich der Fachdidaktik erkennt man, dass die Datensätze, die zum Trainieren eines Machine-Learning-Modells verwendet wurden, zwischen 323 und 27.257 Antworten liegen (Zhai et al., 2020,

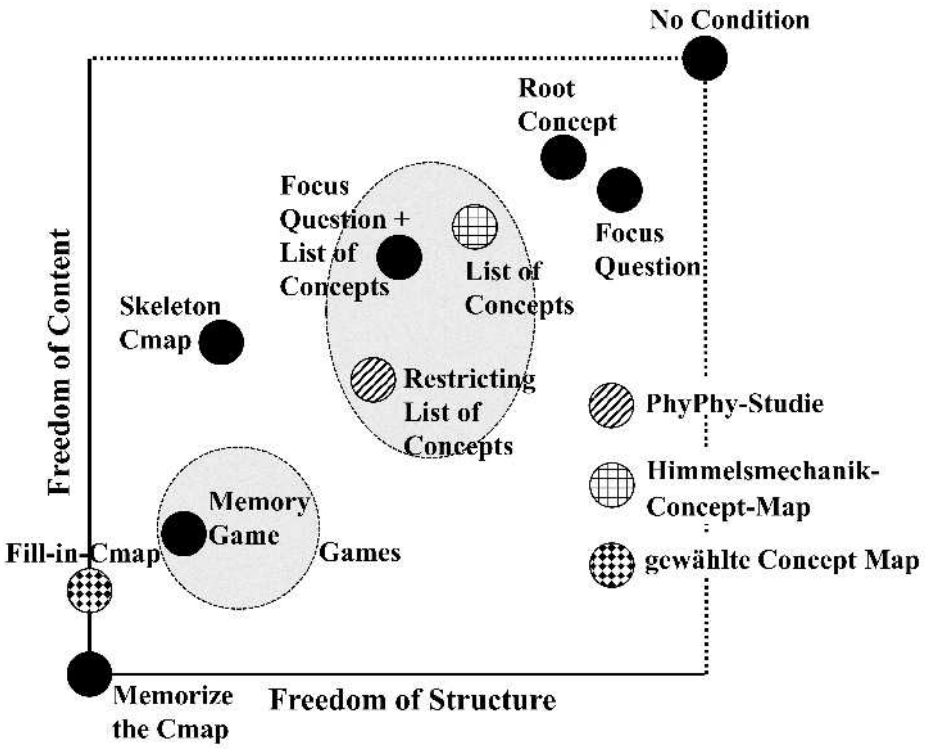


Abbildung 6.6: Einordnung der beiden Concept-Map-Studien und die in dieser Arbeit verwendeten Concept Map in das Schema von Cañas et al. (2023)

S. 9). Überträgt man das auf eine Concept-Map-Studie und wählt man das Format aus den Studien von Friege (z. B. 2001), benötigt man eine sehr hohe Anzahl an Concept Maps. Angenommen, man möchte von den 105 möglichen Propositionen jeweils 200 Antworten jeder Proposition im Datensatz haben, um eine automatische Auswertung auf Propositionsebene zu entwickeln, dann müssten 1.050 Concept Maps erhoben werden, wenn die Lernenden im Durchschnitt 20 Propositionen erstellen. Da bei dieser Abschätzung davon ausgegangen wird, dass jede Proposition gleich häufig erstellt wird, müssen sogar noch mehr Concept Maps erhoben werden. Da dies aus zeitlichen und finanziellen Gründen nur schwer umsetzbar ist, ist dieses Format für dieses Projekt nicht geeignet.
Insofern realistisch betrachtet in einer Erhebung lediglich einige hundert Concept Maps erhoben werden können, wurde sich für ein *Fill-in-CMap*-Format entschieden, das stärkere Vorgaben als die PhyPhy-Studie und die Studien von Friege (z. B. 2001) hat. Bei diesem Format müssen die Lernenden eine vorgefertigte Concept Map bearbeiten. Hierbei können entweder Begriffe oder Propositionen vorgegeben sein. Auch eine Kombination aus beiden Aspekten ist denkbar. So sind die Freiheitsgrade nach oben hin stark begrenzt, da es klare Vorgaben in der Concept Map

gibt. In dieser Studie sollen die Begriffe und deren Anordnung vorgegeben werden. Die Relation müssen die Lernenden allerdings selbstständig bestimmen und frei formulieren. So entstehen trotzdem individuelle Propositionen und Concept Maps. Das führt zu einer großen Anzahl strukturell ähnlicher Concept Maps, die sich durch das geschlossene Format einfacherer und valider (automatisch) auswerten lassen (Ruiz-Primo & Shavelson, 1997; Yin et al., 2005). Das Format reduziert auch die Komplexität und Schwierigkeit der Concept Map (Ley, 2015; Plomer, 2011). Durch die Vorgaben können sich die Lernenden direkt auf die Erstellung der Propositionen konzentrieren und verlieren keine Zeit durch die Komplexität des Themengebiets (Hartmeyer et al., 2018). Es ist auch schwer abzuschätzen, wie viele Lernende bereits mit Concept Maps gearbeitet haben und somit Erfahrung im Umgang mit solchen Mapping-Verfahren haben. Zudem kann ein vollkommen offenes Design ohne Concept-Map-Vorerfahrung kognitiv sehr anspruchsvoll sein (Bleckmann & Friege, 2023).
Ein Nachteil geschlossener Concept-Map-Formate ist, dass man von einem Verlust von didaktischen relevanten Informationen ausgehen muss. Formate wie das der Himmelsmechanik-Concept-Map können das Teilwissen und die Missverständnisse der Lernenden besser erfassen als geschlossene Formate, da sie den Lernenden mehr Freiraum bieten (Yin et al., 2005). Allerdings sollen die Lernenden bei der gewählten Concept Map die Propositionen selbstständig formulieren. Das eigenständige Entwickeln der Zusammenhänge liefert die meisten Erkenntnisse über mögliche Fehlvorstellung und das inhaltliche Verständnis (Ruiz-Primo et al., 1998; Yin et al., 2005). Somit können auch bei dem recht geschlossenen *Fill-in-CMap*-Format wichtige Informationen erhoben werden. Zudem können so die Qualitäten der einzelnen Aussagen analysiert und untersucht werden, was wiederum für ein formatives Assessment wichtige Einblicke liefert. Dieses Format eignet sich demnach gut für das Ziel dieses Projekts. Wenn es gelingt, eine automatische Auswertung für das *Fill-in-CMap*-Format zu entwickelt, kann in einem nächsten Schritt der Ansatz bei einer offeneren Concept-Map-Aufgabe getestet werden.
Daher wurde sich für eine Concept Map entschieden, welche durch die folgenden Punkte charakterisiert wird:

- Es werden Begriffe in einer Concept Map vorgegeben.
- Wie in der Himmelsmechanik-Concept-Map besteht die Möglichkeit, weitere Begriffe in die Concept Map zu ergänzen (sogenannte *Joker*).

- Die Anordnung der Begriffe sowie Verbindungslinien zwischen diesen werden ebenfalls vorgegeben, um dieselbe Struktur bei den Concept Maps gewährleisten zu können.
- Dadurch entstehen festgelegte Propositionen, die eine Teilmenge aller möglichen Propositionen sind.
- Die Relation müssen die Lernenden selbstständig bestimmen und frei formulieren, sodass trotzdem individuelle Concept Maps erstellt werden.
- Die Leserichtung der Propositionen (A $\rightarrow$ B oder B $\rightarrow$ A) ist den Lernenden überlassen.

Wie die Einordnung des gewählten Formats in Abbildung 6.6 gezeigt hat, kann man von einem recht geschlossenen Format sprechen. Durch die restriktiven Vorgaben für die Lernenden sind die einzelnen Begriffspaare zwar vorgegeben, jedoch werden die Relationen frei erstellt. Die daraus entstehenden Propositionen (Begriff - Relation - Begriff) werden immer noch sehr individuell und auf unterschiedlichen sprachlichen und inhaltlichen Niveaus sein, was eine einfache Auswertung per Hand erschwert. Dies kann für Lehrkräfte eine zusätzliche (zeitliche) Belastung darstellen, die eine direkte Rückmeldung an Lernende verhindern kann (Hartmeyer et al., 2018). Trotz der starken Vorgaben kann die Concept Map also nicht ohne Weiteres schnell ausgewertet und genutzt werden, was wiederum für das Ziel eines effektiven formativen Assessments hinderlich ist (Hartmeyer et al., 2018).

Thema der Concept Map

Nachdem das Format der Concept Map festgelegt wurde, muss nun der Inhalt der Concept Map bestimmt werden. Dafür stehen mehrere Themen der Schulphysik zur Verfügung. Die Studie soll in Niedersachsen stattfinden und mit gymnasialen Lernenden durchgeführt werden. Diese Lernenden beschäftigen sich in der Sekundarstufe I mit den Themenbereichen Energie, Magnetismus und Elektrizität, Mechanik, Thermodynamik, phänomenorientierte Optik, Atom- und Kernphysik (Niedersächsisches Kultusministerium, 2015). In der Sekundarstufe I wird Physik in Doppeljahrgängen unterrichtet. So wird z. B. der Themenbereich Mechanik im Jahrgang 7 oder 8 und der Themenbereich Thermodynamik im Jahrgang 9 oder 10 unterrichtet (Niedersächsisches Kultusministerium, 2015). Die Entscheidung, in welchem Jahrgang Physik unterrichtet wird, liegt bei den jeweiligen Schulen.

Die Sekundarstufe II besteht an niedersächsischen Gymnasien aus einer einjährigen Einführungsphase, die in der Regel noch im Klassenverband durchgeführt wird und einer zweijährigen Qualifikationsphase (Niedersächsisches Kultusministerium, 2022b). Inhaltlich beschäftigten sich alle 11. Klassen im ersten Schulhalbjahr zunächst mit dem Themengebiet der Mechanik, genauer gesagt mit der Dynamik. Im anschließenden zweiten Halbjahr des 11. Jahrgangs hat jede Klasse die Wahl eines Themas, mit dem sie sich gemeinsam mit der Lehrkraft auseinandersetzen möchte. In der zweijährigen Qualifikationsphase werden die Themen Elektrizität, Schwingungen und Wellen, Quantenobjekte und Atomhülle und Atomkern behandelt (Niedersächsisches Kultusministerium, 2022b).
Es ist essenziell für dieses Projekt, bei der Auswahl des Inhalts der Concept Map auf ein Thema zurückzugreifen, das in einer Vielzahl von Klassen unterrichtet wird, um eine hohe Stichprobe zu erzielen. Wie weiter oben beschrieben, sind Datensätze mit einer hohen Teilnehmendenzahl keine Seltenheit und notwendig für die Entwicklung eines Machine-Learning-Modells. Unter diesem Gesichtspunkt sind die Inhalte in der Qualifikationsphase nicht geeignet. Einerseits müsste man für die beiden Anforderungsniveaus zwei unterschiedliche Concept Maps erstellen und andererseits haben Physikkurse in der Oberstufe meist nur eine geringe Anzahl von Lernenden. Lediglich 5,45 % der Lernenden haben im Jahr 2022 ihr Abitur in Physik auf erhöhtem Anforderungsniveau gemacht (Niedersächsisches Kultusministerium, 2022a). Auch die Themen der Sekundarstufe I bieten sich nicht an, da die Schulen selbstständig entscheiden können, wann sie Physik in dem bestimmten Doppeljahrgang unterrichten. Das bedeutet, dass es Schulen geben wird, die Thermodynamik bereits im 9. Jahrgang unterrichten werden und andere Schulen wiederum erst im darauffolgenden 10. Jahrgang.
Folgt man dieser Logik, sticht das Thema Dynamik des 11. Jahrgangs heraus. Nach dem niedersächsischen Kerncurriculum für die gymnasiale Oberstufe soll jede 11. Klasse mit diesem Thema in das Halbjahr starten (Niedersächsisches Kultusministerium, 2022b, S. 27). Daher ist anzunehmen, dass dieses Thema mit einer hohen Stundenanzahl unterrichtet wird. Außerdem wird die 11. Klasse noch im Klassenverbund unterrichtet. Demnach ist die potenzielle Population für die Stichprobe sehr groß. Zudem sollten die Lernenden über ein gewisses Vorwissen zu diesem Thema verfügen, da gewisse Inhalte der Mechanik bereits in der Sekundarstufe I unterrichtet werden (siehe Niedersächsisches Kultusministerium, 2015).

Festlegung der Begriffe

Wirft man einen genaueren Blick in das Kerncurriculum für die gymnasiale Oberstufe, findet man zum Themengebiet Dynamik folgende zentralen Begriffe: freier Fall, waagerechter Wurf, Kraft, Newton'sche Axiome, gleichförmige Kreisbewegung, Umlaufdauer, Bahngeschwindigkeit, Zentripetalbeschleunigung und -kraft, kinetische Energie und Energieerhaltungssatz der Mechanik (Niedersächsisches Kultusministerium, 2022b, S. 28). Laut den prozessbezogenen Kompetenzen sollen die Lernenden in der Lage sein, u. a. die „*Ergebnisse auf ausgewählte gleichmäßig beschleunigte Bewegungen*" übertragen zu können, „*zwischen sprachlicher, grafischer und algebraischer Darstellung dieser Zusammenhänge*" übersetzen zu können oder den „*Ortsfaktor als Fallbeschleunigung*" deuten zu können. Damit die Concept Map nicht zu umfangreich wird, müssen die relevanten Inhalte und dazugehörigen Begriffe eingegrenzt werden. Da nicht alle Schulklassen im selben Tempo vorankommen werden, bieten sich Themen wie freier Fall und waagerechter Wurf an, da sie zu Beginn der Unterrichtsreihe behandelt werden (Niedersächsisches Kultusministerium, 2022b, S. 28). Dies hat den Vorteil, dass zumindest zu Beginn der Unterrichtseinheit alle teilnehmenden Klassen dieselben Inhalte behandeln sollten.
Um eine tiefere und bessere Übersicht über die genauen Inhalte zu erhalten, bietet sich ein Blick in die gängigen Schulbücher an. Die Schulbücher Dorn.Bader (2018) und Impulse Physik (2018) starten die Unterrichtseinheit Dynamik mit den Beschreibungen der beiden Bewegungsarten gleichförmige und gleichmäßig beschleunigte Bewegung. Dabei ist eine gleichförmige Bewegung durch eine konstante Geschwindigkeit und durch das Zeit-Ort-Gesetz $s(t) = v \cdot t$ definiert. Eine Bewegung mit einer konstanten Beschleunigung wird hingegen als gleichmäßig beschleunigte Bewegung bezeichnet. Das Zeit-Ort-Gesetz sowie das Zeit-Geschwindigkeit-Gesetz lauten: $s(t) = \frac{1}{2} \cdot a \cdot t^2$, $v(t) = a \cdot t$. Für die Beschreibung und Erklärung der beiden Bewegungsarten und der dazugehörigen Konzepte Geschwindigkeit und Beschleunigung benutzen beide Schulbücher Anwendungsbeispiele. Zur Veranschaulichung dieser Zusammenhänge und zur Untersuchung der Abhängigkeiten werden die drei Diagramme Zeit-Weg (t-s), Zeit-Geschwindigkeit (t-v) und Zeit-Beschleunigung (t-a) genutzt (Dorn.Bader, 2018).
Nachdem die beiden Bewegungsarten behandelt wurden, werden in den nachfolgenden Seiten unterschiedliche Fall- und Wurfbewegungen untersucht. Sowohl Dorn.Bader (2018) als auch Impulse Physik (2018) starten den Abschnitt mit

der Betrachtung des freien Falls. Dabei wird der freie Fall als reibungsfreie Bewegung definiert und die Fallbeschleunigung g als konstante Beschleunigung $a = g = 9,81 \frac{m}{s^2}$ für den freien Fall eingeführt. Der Fall mit der Luftreibung wird in beiden Schulbüchern ebenfalls kurz behandelt. Darauf aufbauend geht es um Wurfbewegungen, wobei der waagerechte Wurf ohne Reibung ausführlich untersucht wird. Von einem waagerechten Wurf spricht man, wenn der Körper horizontal, also waagerecht abgeworfen wird. Dabei setzt sich die Wurfbewegung aus einer gleichförmigen Bewegung in x-Richtung und einem freien Fall, also aus einer gleichmäßig beschleunigten Bewegung in y-Richtung zusammen.
Nach der Untersuchung der verschiedenen Bewegungsformen folgt die Analyse der Ursachen von Bewegungen. Es werden Kräfte und vor allem die Newton'schen Gesetze betrachtet. Impulse Physik (2018) führt den Kraftbegriff wie folgt ein: „Kräfte können Körper verformen und deren Geschwindigkeit und Bewegungsrichtung verändern. Ihre Wirkung hängt von ihrem Betrag, ihrer Richtung und ihrem Angriffspunkt am Körper ab. Kräfte lassen sich durch Vektoren beschreiben“ (Impulse Physik, 2018, S. 44). Ein besonderes Augenmerk legt Dorn.Bader (2018) auf die sogenannte Grundgleichung der Mechanik (zweites Newton'sches Gesetz): $\vec{F} = m \cdot \vec{a}$ mit der Einheit für die Kraft $1N = \frac{1kg \cdot m}{s^2}$. Anschließend geht es um das Trägheitsprinzip (erstes Newton'sches Gesetz) welches Dorn.Bader (2018) wie folgt definiert: „Jeder Körper behält Betrag und Richtung seiner Geschwindigkeit bei, solange auf ihn keine Kraft ausgeübt wird“ (Dorn.Bader, 2018, S. 44). Abschließend wird in diesem Kapitel noch das Wechselwirkungsprinzip (drittes Newton'sches Gesetz) eingeführt, welches besagt, dass Kräfte immer wechselseitig wirken. Nachdem das Thema der Kreisbewegungen behandelt wird, schließen beide Schulbücher mit einem Kapitel über Energie ab.
Wie voranstehend beschrieben, soll die Concept Map inhaltlich nach oben abgegrenzt werden. Zum einen soll so die Concept Map übersichtlich und verständlich gehalten werden und zum anderen soll die Concept Map für möglichst viele Klassen thematisch passend sein. Aus der Betrachtung des niedersächsischen Kerncurriculums für die gymnasiale Oberstufe (Niedersächsisches Kultusministerium, 2022b) und der Analyse der beiden Schulbücher Dorn.Bader (2018) und Impulse Physik (2018) werden die folgenden elf Begriffe für die Concept Map genutzt: Die ersten beiden Begriffe stellen die Grundlage der behandelten Themen dar, nämlich die beiden Bewegungsformen *gleichförmige Bewegung* und *gleichmäßig beschleunigte Bewegung*. Da für die Betrachtung der beiden Bewegungsarten in den Schulbüchern Graphen eine wichtige Rolle spielen und im Kerncurricu-

lum die grafische Betrachtung und der Transfer in andere Darstellungsformen explizit erwähnt werden, sind die Graphen *Zeit-Geschwindigkeit*, *Zeit-Weg* und *Zeit-Beschleunigung* drei weitere Begriffe, die in der Concept Map enthalten sind. Da die beiden Bewegungsarten mittels der Begriffe *Beschleunigung* und *Geschwindigkeit* definiert werden können, sind diese beiden Begriffe ebenfalls in der Concept Map enthalten. Damit auch die Newton'schen Axiome bzw. die Grundgleichung der Mechanik repräsentiert werden kann, ist der Begriff *Kraft* ebenso Teil der Concept Map. Abschließend soll die Concept Map noch zwei Beispiele für die verschiedenen Bewegungsarten enthalten. Es wurde sich aufgrund der ausführlichen Betrachtung in beiden Schulbüchern sowie der Erwähnung im Kerncurriculum für den *waagerechten Wurf* und *freien Fall* inklusive des *Massen*-Begriffs entschieden. Eine Übersicht der elf ausgewählten Begriffe findet sich in Tabelle 6.3.

#	**Begriff**
1	gleichförmige Bewegung
2	gleichmäßig beschleunigte Bewegung
3	Zeit-Geschwindigkeit-Graph
4	Zeit-Beschleunigung-Graph
5	Zeit-Weg-Graph
6	Beschleunigung
7	Geschwindigkeit
8	Kraft
9	Masse
10	waagerechter Wurf
11	freier Fall

Tabelle 6.3: Die elf ausgewählten Begriffe für die Concept Map

Durch diese Auswahl der Begriffe wurde versucht, das umfangreiche Thema der Dynamik in der Jahrgangsstufe 11 zu konzentrieren und sich dabei auf die zentralen Elemente der ersten Themen der Unterrichtseinheit zu fokussieren. Außerdem sollten den Lernenden Begriffe wie Geschwindigkeit, Zeit-Weg-Graph oder Zeit-Geschwindkeit-Graph zumindest für die gleichförmige Bewegung bereits aus den Jahrgangsstufen 7 und 8 bekannt sein (Niedersächsisches Kultusministerium, 2015). Dies hat den Vorteil, dass die Lernenden ein gewisses Vorwissen für die Bearbeitung der Concept Map mitbringen sollten.

Konzipierung der Propositionen

Wie in Abschnitt 6.1.3 beschrieben wurde, sollen die Lernenden die Propositionen in einer strukturell vorgegebenen Concept Map selbstständig entwickeln. Das bedeutet, dass die Verbindungen zwischen den elf Begriffen und deren Anordnungen innerhalb der Concept Map bereits im Vorfeld festgelegt werden müssen. Aus den elf Begriffen lassen 55 verschiedene Propositionen erstellen, wobei nicht alle Verbindungen inhaltlich Sinn ergeben (siehe Tabelle 6.4).

	Proposition
Physikalisch sinnvoll	Beschleunigung – F = m * a – Masse Masse – ist unabhängig beim – Freier Fall
Physikalisch weniger sinnvoll	Kraft - ??? - Zeit-Weg-Graph Masse - ??? - Gleichförmige Bewegung

Tabelle 6.4: Beispiele für physikalisch sinnvolle und weniger sinnvolle Propositionen

Zudem sollte die Concept Map auch übersichtlich und möglichst einfach gestaltet sein, damit alle teilnehmenden Lernenden, unabhängig von ihrer Vorerfahrung, die Concept Map bearbeiten können. Aus inhaltlicher Perspektive lassen sich mehrere Propositionen physikalisch sinnvoll bilden. Sowohl im Impulse Physik (2018) als auch im Dorn.Bader (2018) werden die Graphen Zeit-Weg/-Beschleunigung/-Geschwindigkeit genutzt, um die Bewegungsarten zu analysieren und zu charakterisieren. Daraus lassen sich sechs Propositionen bilden. Zur Charakterisierung von gleichförmiger und gleichmäßig beschleunigter Bewegung werden außerdem die Begriffe Geschwindigkeit und Beschleunigung genutzt. Daraus lassen sich vier weitere Propositionen bilden. Betrachtet man die Newton'schen Axiome, die in beiden betrachteten Schulbüchern ausführlich behandelt werden, lässt sich der Kraftbegriff mit den beiden Begriffen Geschwindigkeit und Beschleunigung verknüpfen. Dies sind alles kanonische Beispiele, aus denen man Propositionen bilden kann und die in allen Klassen mit großer Wahrscheinlichkeit behandelt werden sowie den Kern des Unterrichts darstellen. Damit die Concept Map übersichtlich bleibt, wurde sich letztlich für diese 19 verschiedenen Propositionen entschieden (siehe Tabelle 6.5).

Konkrete Darstellung der Concept Map

Nachdem die physikalisch sinnvollen Propositionen entwickelt wurden, muss im nächsten Schritt die konkrete Anordnung innerhalb der Concept Map erstellt wer-

#	Begriffspaar
1	gleichförmige Bewegung – Zeit-Beschleunigung-Graph
2	gleichförmige Bewegung – Zeit-Geschwindigkeit-Graph
3	gleichförmige Bewegung – Zeit-Weg-Graph
4	gleichmäßig beschleunigte Bewegung – Zeit-Beschleunigung-Graph
5	gleichmäßig beschleunigte Bewegung – Zeit-Geschwindigkeit-Graph
6	gleichmäßig beschleunigte Bewegung – Zeit-Weg-Graph
7	Beschleunigung – gleichförmige Bewegung
8	Beschleunigung – gleichmäßig beschleunigte Bewegung
9	Geschwindigkeit – gleichförmige Bewegung
10	Geschwindigkeit – gleichmäßig beschleunigte Bewegung
11	Beschleunigung – Geschwindigkeit
12	Kraft – Beschleunigung
13	Kraft – Geschwindigkeit
14	waagerechter Wurf – gleichförmige Bewegung
15	waagerechter Wurf – gleichmäßig beschleunigte Bewegung
16	freier Fall – gleichmäßig beschleunigte Bewegung
17	freier Fall – Beschleunigung
18	freier Fall – Geschwindigkeit
19	Masse – freier Fall

Tabelle 6.5: Die 19 ausgewählten Begriffspaare für die Concept Map

den. Dazu muss zunächst entschieden werden, ob die Concept Map mittels eines digitalen Tools oder per Stift und Papier erstellt werden soll. Da die Concept Maps später automatisch ausgewertet werden sollen und diese dazu digital vorliegen müssen, ist die Verwendung einer Concept-Map-Software naheliegend. Einerseits können die Concept Maps so direkt digital gespeichert und weiterverwendet werden und andererseits ist die Durchführung der Studie so ökonomischer und ökologischer.

Eine bekannte und in vielen Forschungsarbeiten genutzte Concept-Map-Software ist CmapTools, eine Client-Server-basierte Software, die am Institute for Human and Machine Cognition (IHMC) entwickelt wurde, um die Erstellung von Concept Maps zu unterstützen und die Zusammenarbeit und den Austausch von verschiedenen Benutzern zu ermöglichen (Cañas et al., 2004; Daley et al., 2007). Die Funktionen und die grafische Oberfläche von CmapTools wurden so konzipiert, dass sowohl Benutzer, die wenig technisches Wissen und Erfahrung mit Concept Maps haben, als auch erfahrende Benutzer die Software bedienen können (Cañas et al., 2004). Das Konzipieren einer Concept Map funktioniert in CmapTools über Mausklicks. So können Konzepte und Verbindungslinien einfach erstellt

und bearbeitet werden. Außerdem können Optionen für verschiedene Stile wie Schriftart, Farben, Linientypen eingeblendet werden. Die Software CmapTools ist für mehrere Geräte (u. a. iPads und Computer) verfügbar. Die Basissoftware ist aktuell für Bildungseinrichtungen kostenlos, wobei eine Beta-Testversion auch für andere Benutzer einschließlich kommerzieller Nutzung kostenlos angeboten wird[3]. Die Software kann unter https://cmap.ihmc.us/ herunterladen werden.
Überträgt man die Liste der Propositionen und die Vorgaben der Concept Map aus Abschnitt 6.1.3 in Cmaptools, erhält man die Concept Map, die in Abbildung 6.7 dargestellt ist. Im mittleren Block der Concept Map befinden sich die drei

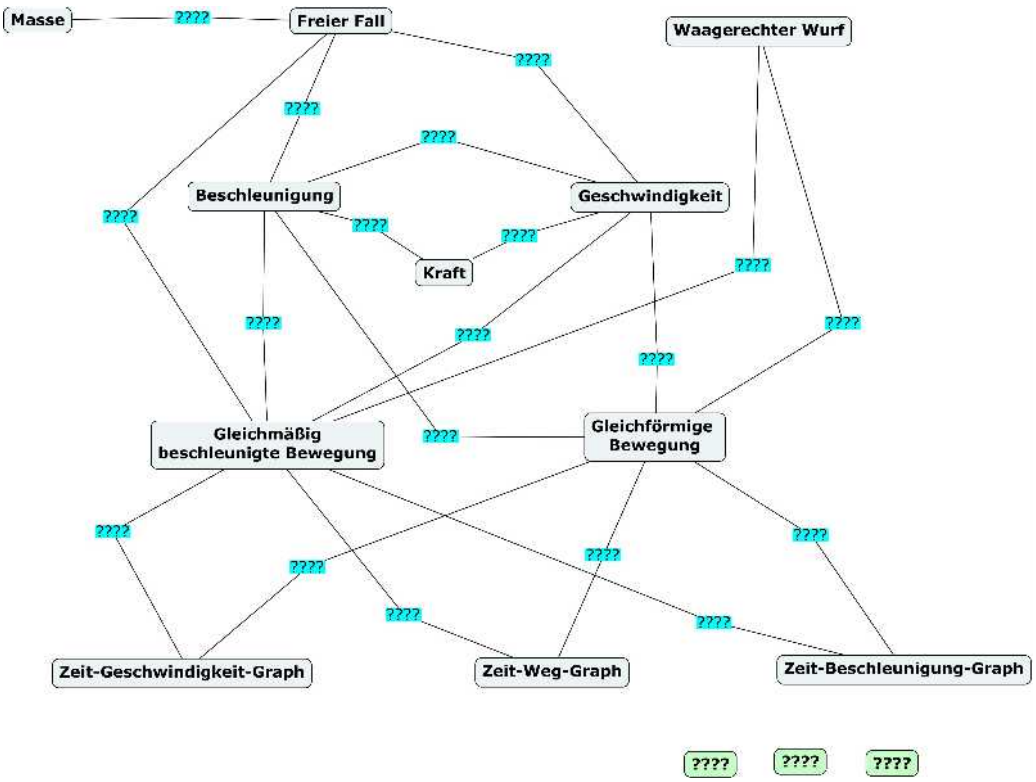


Abbildung 6.7: Konzipierte Concept Map mit den festgelegten 19 Propositionen in CmapTools

Begriffe Beschleunigung, Geschwindigkeit und Kraft sowie die beiden Hauptbewegungsarten gleichmäßig beschleunigte und gleichförmige Bewegung. Im unteren Teil der Concept Map sind die drei Graphen Zeit-Geschwindigkeit-Graph, Zeit-Beschleunigung-Graph und Zeit-Weg-Graph dargestellt. Ganz oben befinden sich die beiden Beispielbewegungen waagerechter Wurf und freier Fall und der Begriff Massen. Damit die Vorgabe, die Leserichtung ist frei wählbar, erfüllt ist,

[3] Stand 11.12.2023

wurden in der Concept Map keine Pfeile, sondern nur einfache Linien verwendet. Zudem wurden Joker-Begriffe in Form von drei andersfarbigen leeren Begriffen im unteren Bereich eingefügt. So können Lernende diese bei Bedarf ausfüllen und eigenständig neue Propositionen erstellen. Wie in Abschnitt 6.1.3 beschrieben, besteht die Hauptaufgabe der Lernenden darin, Relationen zwischen den bestehenden Begriffspaaren zu formulieren und daraus Propositionen herzustellen. Damit diese sich farblich von den Begriffen abgrenzen, wurden die leeren Felder blau hinterlegt und mit drei Fragezeichen als Platzhalter gefüllt (siehe Abbildung 6.7).

Entwicklung des Bewertungsschemas

Für die Bewertung einer Concept Map gibt es eine Vielzahl von Möglichkeiten (siehe Abschnitt 3.2). Die Wahl des Bewertungssystems ist entscheidend für die spätere Entwicklung der automatischen Machine-Learning-Anwendung, da das Modell mit den Concept Maps und den dazugehörigen Bewertungen trainiert wird. Sowohl bei der PhyPhy- als auch bei den Studien von Friege (z. B. 2001) wurde der Fokus der Bewertung auf die einzelnen Propositionen gelegt. Bei der PhyPhy-Studie wurde ursprünglich ein sechsstufiges Bewertungsschema genutzt, das die einzelnen Propositionen nach richtig und falsch differenziert. Für die Bewertung der Himmelsmechanik-Concept-Maps wurden sogar sieben verschiedene Ratingkategorien verwendet. Hier wurde der Fokus mehr auf eine qualitative Unterscheidung der Propositionen gelegt. So gab es nicht nur richtige oder falsche Aussagen, sondern z. B. auch Ober-/Unterbegriffsrelationen oder Funktionsrelationen.
Auch in dieser Arbeit sollen die einzelnen Propositionen und deren Qualität analysiert und für die spätere Rückmeldung genutzt werden. Bestimmte quantitative Merkmale der Concept Map können zwar trotzdem ermittelt werden, allerdings werden sie nicht für die Rückmeldung verwendet. Zum einen fallen bei solch einem geschlossenen Concept-Map-Format viele quantitative Kennzahlen weg, da bis auf die Joker-Begriffe die Anzahl der Begriffe und Propositionen festgelegt ist. Zum anderen wird es in der Literatur uneinheitlich gesehen, ob eine graphentheoretische Auswertung überhaupt Aufschluss über den Inhalt einer Concept Map geben kann (Ley, 2015, S. 24). Da die Concept Maps nicht für eine summative Bewertung genutzt werden, sondern die Grundlage für ein formatives Assessment darstellen, sind die Informationen, die aus den einzelnen Propositionen ermittelt werden können, relevant. Deshalb fällt auch in diesem Projekt die Wahl auf eine Auswertung der einzelnen Propositionen. Damit die Auswertung und das dar-

	Kategorie	Beschreibung	Schlüsselwörter
A	Falsch	physikalisch falsche Antworten	
B	einfacher Zusammenhang	Zuordnung von Beispielen oder Eigenschaften. Beschreibung von ungerichteten Zusammenhängen	ist ein, verändert sich, besitzt
C	gerichteter Zusammenhang	Beschreibung von oberflächlichen oder gerichteten Zusammenhängen	steigt/sinkt, x hängt von y ab
D	detaillierte und funktionale Zusammenhänge	Kann als Formel oder verbaler Zusammenhang formuliert werden	steigt / sinkt linear, je / desto

Tabelle 6.6: Das konzipierte Bewertungsschema mit Schlüsselwörtern für die vier Kategorien A, B, C und D

aus resultierende Feedback nicht nur aus der Anzahl von richtigen und falschen Propositionen besteht, sollen mehrere Kategorien entwickelt werden. Für diese Entwicklung wurde sich an dem Auswertungsschema der Studien von Friege (z. B. 2001) orientiert.

Die Verteilung der sieben Ratingkategorien der Himmelsmechanik-Concept-Map (siehe Abbildung 6.5) hat gezeigt, dass die Besetzung der Kategorien nicht gleich verteilt sind. Damit für die automatische Auswertung eine bessere Verteilung innerhalb der verschiedenen Kategorien entsteht, wurde das Bewertungsschema auf vier Kategorien zusammengefasst (siehe Tabelle 6.6). Die erste Kategorie des Bewertungsschemas ist selbsterklärend und wird für physikalisch falsche Propositionen genutzt. In die zweite Kategorie sollen alle Propositionen fallen, die als einfache Zusammenhänge gesehen werden können. Diese sind insbesondere Zuordnungen von Beispielen oder sehr einfache ungerichtete Zusammenhänge. Typische Antworten in dieser Kategorie könnten Propositionen in der Form „Concept-A ist ein Concept-B“ oder „Concept-A verändert sich bei Concept-B“ sein. Wenn die Propositionen gerichtet sind, also nicht nur „verändert“, sondern „steigt/sinkt“ als Verbindungswort genutzt wird, lässt die Proposition sich in der Kategorie C verorten. Diese Propositionen sind zwar weiterhin nicht detailliert, jedoch weisen sie mehr Informationen auf als die Propositionen in Kategorie B. Für sehr detaillierte Propositionen und funktionale Zusammenhänge steht die Kategorie D zur Verfügung. Anders als beim Bewertungsschema der Studien von Friege (z. B. 2001) wird in diesem Fall nicht zwischen qualitativen und quantitativen Funktionsrelationen unterschieden.

Proposition	Kat.	Beispielantwort
Geschwindigkeit – Freier Fall	A	steigt exponentiell
	B	ist vorhanden bei
	C	wird immer größer beim
	D	linear ansteigend
Gleichförmige Bewegung – Zeit-Weg-Graph	A	verläuft parallel zur x-Achse
	B	ist darstellbar als
	C	steigt an
	D	konstante Steigung
Geschwindigkeit – Gleichmäßig beschleunigte Bewegung	A	ist konstant
	B	verändert sich
	C	wird größer / kleiner
	D	steigt/sinkt linear

Tabelle 6.7: Auszug aus dem Codierleitfaden für drei Propositionen

Durch die Verwendung von nur vier unterschiedlichen Kategorien wird versucht, eine bessere und gleichmäßigere Verteilung der Propositionen zu erhalten. Das Bewertungsschema soll später die Basis für eine Rückmeldung sowohl für Lehrkräfte als auch für Lernende sein. Da die Rückmeldung Teil eines formativen Assessments ist und deshalb keine Noten vergeben werden oder eine andere Form einer abschließenden Bewertung durchgeführt wird, muss das Schema auch keine ordinale Skalierung haben. Wie die Rückmeldungen im Detail aussehen und explizit umgesetzt werden, wird in Kapitel 7 ausführlich beschrieben.
Damit die erstellten Propositionen auch gut von menschlichen Bewertern analysiert werden können, benötigt es einen Codierleitfaden. Dieser Leitfaden soll aus Beispielen für jede der 19 Propositionen und für jede der vier Bewertungskategorien gebildet werden. Zunächst werden die Beispiele und Codierrichtlinien deduktiv aus dem Bewertungsschema und aus den Arbeiten von Friege (2001) und Fischler und Peuckert (2000) entwickelt. Nachdem die ersten Concept Maps erhoben wurden, werden die Codierrichtlinien noch einmal angepasst und optimiert. Dazu sollen Codierbeispiele aus den Lernenden-Propositionen gewonnen werden, welche die Vielseitigkeit und den Sprachgebrauch der unterschiedlichen Lösungen widerspiegeln. So kann gleichzeitig überprüft werden, ob die gebildeten Kategorien eindeutig definiert und ob diese für das Auswertungsziel geeignet sind (Hammann & Jördens, 2014, S. 175). Nachdem das Schema gegebenenfalls

modifiziert wurde, können die Propositionen von zwei unabhängigen menschlichen Bewertern analysiert und eingeordnet werden, damit auf der Grundlage der Bewertung ein Machine-Learning-Modell gebildet werden kann. Ein Auszug des Bewertungsschemas inklusive Codierbeispiele findet sich in Tabelle 6.7 und der vollständige Leitfaden in Anhang A.

6.2 Phase 2: Entwicklung eines Machine-Learning-Modells

In der zweiten Phase wird die entwickelte Mechanik-Concept-Map in mehreren Schulen eingesetzt, um Daten für die Entwicklung eines Machine-Learning-Modells zu erheben. Der Abschnitt wird deshalb in zwei Teile unterteilt. Es wird zunächst die Erhebung in den Schulen beschrieben und auf den Datensatz eingegangen. Anschließend wird der Entwicklungsprozess des Machine-Learning-Modells dargestellt sowie die ersten Ergebnisse präsentiert.

6.2.1 Vorbereitung der Erhebung

Bevor die Mechanik Concept Map in den Schulen eingesetzt werden kann, muss eine Einführung in das Thema Concept Maps erarbeitet werden. Im Vorfeld konnte nicht abgeschätzt werden, wie viele Klassen bzw. Lernende bereits mit Concept Maps gearbeitet haben. Da das gewählte Concept-Map-Format ein recht geschlossenes Format ist, sollte die Bearbeitung deutlich einfacher sein als die Neuerstellung einer Concept Map auf einem leeren Blatt Papier, da bereits die Grundstruktur der Concept Map im Vorfeld feststand. Daher wird sich für eine kurze Einführung entschieden, die nicht länger als 15 Minuten dauern soll. So kann der Fokus der Erhebung und der Großteil der verfügbaren Zeit auf der Bearbeitung der Concept Map liegen.
Zunächst wurde in kurzen einleitenden Sätzen die grundlegende Idee hinter einer Concept Map erklärt. So sollte ein gemeinsames Verständnis geschaffen werden. Damit alle Lernenden die Grundstruktur einer Concept Map verstehen, wurde anschließend eine einfache Concept Map zum Thema Tiere an der Tafel gemeinsam erarbeitet (siehe linke Concept Map in Abbildung 6.8). Dabei wurde sich an der Studie von Plomer (2011) orientiert. Besonderen Wert wurde auf die Formulierung der Propositionen gesetzt, da dies die Hauptaufgabe der Lernenden ist. Es sollte

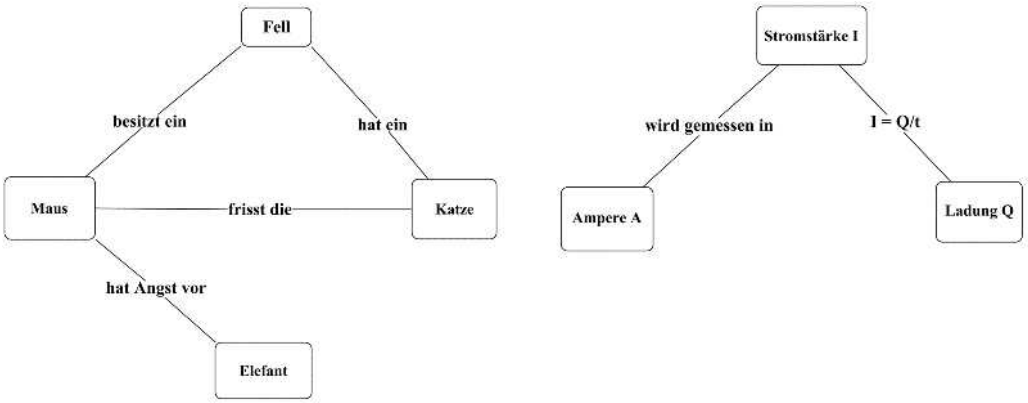


Abbildung 6.8: Einführungsbeispiele zum Thema Concept Map

verdeutlicht werden, dass Propositionen nicht aus langen Sätzen, sondern aus präzisen direkten Zusammenhängen, die im optimalen Fall eindeutig und richtig sind, bestehen. Anschließend wird ein zweites Beispiel gezeigt und an der Tafel erarbeitet (siehe rechte Concept Map in Abbildung 6.8). Die lediglich aus drei Begriffen bestehende Concept Map zum Thema E-Lehre hatte das Ziel, explizit zu zeigen, dass auch Formeln zur Erstellung von Propositionen genutzt werden können.

In der zu bearbeitenden Mechanik-Concept-Map ist keine Leserichtung vorgegeben. Deshalb wurde in beiden Beispielen darauf geachtet, ebenfalls keine Leserichtung vorzugeben und aufzuzeigen, dass man die Propositionen, wie in Physik üblich, in beide Richtungen formulieren kann. So kann unter anderem Katze-Maus sowohl als Proposition „Katze – frisst die – Maus“ als auch als „Maus – wird gefressen von – Katze“ formuliert werden. Damit die Einführung identisch ist und alle Lernenden die gleichen Informationen enthalten, wurde die Einführung in allen Klassen immer von der gleichen Person durchgeführt.

Nachdem in die Methode Concept Maps eingeführt wurde, bekamen die Lernenden die Instruktionen zum Ablauf der Erhebung. Dabei wurden vor allem die gewählte Concept Map und das Ziel der Studie thematisiert. Zudem wurden die Joker-Begriffe eingeführt, die zum Ergänzen der Concept Map genutzt werden konnten. Alle teilnehmenden Lernenden erhielten Laptops, auf denen bereits die Software CmapTools installiert und die Mechanik Concept Map erstellt war. Nach einer kurzen Einführung in CmapTools, bei der die wichtigsten Funktionalitäten erklärt wurden, bekamen die Lernenden 25 Minuten für die Bearbeitung der Concept Map, wobei darauf geachtet wurde, dass die Lernenden möglichst in Einzelarbeit arbeiten.

Abschließend wurde noch ein Fragebogen eingesetzt, der inhaltlich zweigeteilt ist: Zunächst wurden demografische Angaben wie Alter und Geschlecht erhoben. Anschließend wurde nach Noten in Mathematik, Physik und Deutsch gefragt und welcher Physikkurs in der Qualifikationsphase voraussichtlich besucht wird. Der zweite Teil bezog sich auf die bearbeitete Concept Map. Es sollte herausgefunden werden, welche Aspekte der Concept Map für die spätere Feedbackstudie verbessert werden können. Zudem sollten die Lernenden ihre Meinung äußern, ob Concept Maps öfter eingesetzt werden sollen und wie ihre persönliche Concept Map am besten ausgewertet werden soll. Durch den Fragebogen können erste Erkenntnisse zu Herausforderungen bei der Bearbeitung einer Concept Map erhoben werden. Für das Ausfüllen des Fragebogens wurde eine ausfüllbare PDF-Datei genutzt, die auf den Laptops zu finden war. Der vollständige Fragebogen befindet sich in Anhang B.

6.2.2 Soziodemografische Daten der Stichprobe

Die Lernenden-Concept-Maps bildeten die Grundlage für die Trainings- und Testdaten des Machine-Learning-Modells. Da eine möglichst homogene Verteilung der Propositionen erreicht werden soll, wurden als Erhebungszeitpunkt die letzten Wochen vor den Sommerferien gewählt. Es konnte so davon ausgegangen werden, dass die Lernenden die Inhalte der Concept Map bereits behandelt hatten und dementsprechend möglichst viele Propositionen eigenständig bilden konnten. Zudem sollte durch diesen Zeitpunkt eine möglichst große Teilnehmendenzahl erreicht werden, da der normale Lehrplan wenig gestört wird.
Die Erhebung fand im Sommer 2022 mit fünf unterschiedlichen Gymnasien in Niedersachsen statt. Insgesamt beteiligten sich 14 Klassen mit insgesamt 233 Lernenden an der Studie (siehe Tabelle 6.8). Von den 233 Lernenden bearbeiteten 230[4] eine Concept Map und 203 füllten den dazugehörigen Fragebogen aus. Bei der Klasse 14 gab es ein Datenübertragungsproblem, weswegen hier nur zwei Fragebögen zur Verfügung stehen. Die restliche Differenz lässt sich auf die Freiwilligkeit des Fragebogens am Anschluss der Concept Map zurückführen.
Alle Lernenden besuchten zum Erhebungszeitpunkt eine 11. Klasse. Demzufolge findet sich in der Stichprobe auch keine große Schwankung bezüglich des Alters. Das Durchschnittsalter der Lernenden lag bei knapp 17 Jahren. Aus Tabelle 6.8

[4] In drei Concept Maps wurden keine der 19 Propositionen bearbeitet, weswegen diese Concept Maps als nicht bearbeitet gewertet wurden.

Klasse	Schule	Lernende	Concept Maps	Fragebögen	Geschlecht (m/w/d)	∅Alter
1	A	20	20	19	9/9/1	16,78
2	A	9	9	9	8/1/0	16,56
3	A	19	19	19	12/7/0	16,4
4	B	19	19	17	5/12/0	17,4
5	B	20	19	20	8/12/0	16,95
6	C	16	15	16	7/9/0	17
7	C	14	13	14	4/9/1	16,86
8	D	14	14	12	7/5/0	16,75
9	D	22	22	22	10/9/2 $\pm$ 1 o.A.	17
10	D	24	24	22	8/14/0	16,95
11	D	20	20	15	9/5/0 + 1 o.A.	17,07
12	D	3	3	3	0/3/0	17
13	D	19	19	13	3/9/1	17
14	E	14	14	2	2/0/0	17
Σ	**5**	**233**	**230**	**203**	**92/104/5**	$\overline{Alter}$ **= 16,91**

Tabelle 6.8: Soziodemografische Daten der Stichprobe aus der Entwicklungsstudie

geht hervor, dass die Geschlechterverteilung fast gleich verteilt war. Bei der Erhebung identifizierten sich 51 % der Lernenden als weiblich, 45 % als männlich und 2 % als divers. Der restliche Anteil an Lernenden gab im Fragebogen kein Geschlecht an.
Die Noten in den Fächern Mathematik, Physik und Deutsch waren innerhalb der Stichprobe ähnlich verteilt (siehe Abbildung 6.9). Der Median für die Mathematik- und Physiknote betrug 10 Punkte und für die Deutschnote 9 Punkte, wobei 15 Punkte die beste und 0 Punkte die schlechteste Note ist. Betrachtet man den Interquartilsabstand der drei Boxplots in Abbildung 6.9, erkennt man eine identische Breite von 4 Punkten. Aus Abbildung 6.9 kann demnach festgestellt werden, dass die teilnehmenden Lernenden in allen drei Fächern eine ähnliche Notenverteilung aufwiesen und diese eher im oberen Bereich der Notenskala war.

6.2.3 Analyse der Concept Maps und menschliche Bewertung der Propositionen

Wie in Phase 1 beschrieben, war die Aufgabe der Lernenden, bis zu 19 Propositionen zu bearbeiten. Insgesamt lagen 230 bearbeitete Concept Maps mit 3.322 Propositionen vor. Das ergibt einen Durchschnitt von 14 Propositionen pro Concept Map.
Die Häufigkeiten der einzelnen Propositionen schwanken zwischen 153 und 218 (siehe Tabelle 6.9). Im Mittel wurden die Propositionen 174-mal erstellt. Die vier Propositionen, die den Begriff *freier Fall* enthalten, wurden dabei am häufig-

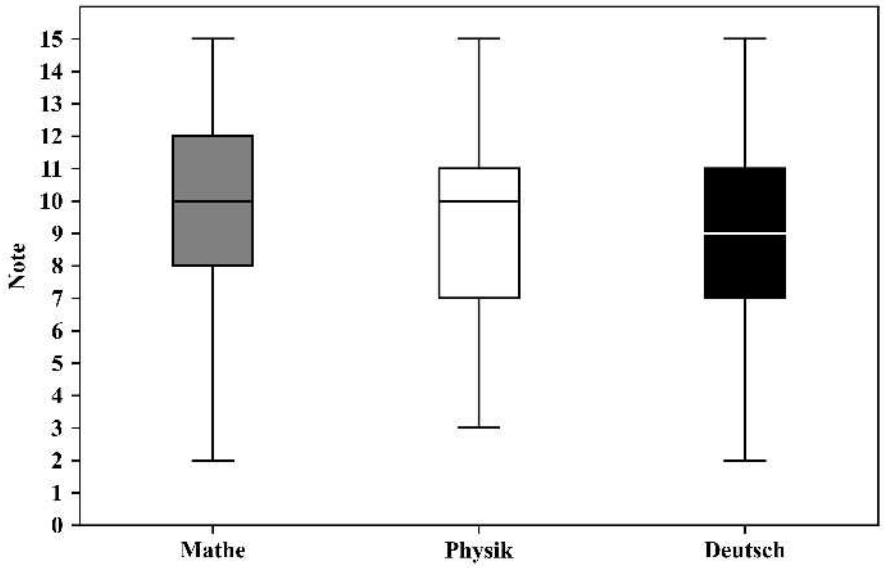


Abbildung 6.9: Notenverteilung der Lernenden aus der Entwicklungsstudie in den Fächern Mathematik, Physik und Deutsch

sten und die beiden Propositionen, die den Begriff *Zeit-Beschleunigung-Graph* enthalten, am wenigsten bearbeitet.

Die Möglichkeit, eigene Begriffe in die Concept Map einzufügen oder zusätzliche Propositionen zu erstellen, wurde nur von vier Lernende wahrgenommen. Es wurden die zusätzlichen Begriffe *Zeit*, *potenzielle Energie*, *Ortsfaktor*, *Gravitation* und *Verantwortung* benutzt. Aus diesen Begriffen haben die Lernenden unterschiedliche Propositionen gebildet, z. B. *Zeit - benötigt zur Berechnung von - Beschleunigung* oder *Potenzielle Energie erhöht die Beschleunigung beim - Freier Fall*. Zudem erstellten die Lernenden zusätzliche Propositionen aus bestehenden Begriffen wie *Masse - hat Einfluss auf die - Beschleunigung*. Da diese Joker-Propositionen unter 1 % der Gesamtdaten ausmachen, werden sie für die automatische Auswertung aus dem Datensatz gelöscht. Das bedeutet, dass der Datensatz sich auf 3.302 Propositionen reduziert.

Die verbleibenden Propositionen wurden mithilfe des Bewertungsschemas, welches in Phase 1 konzipiert wurde, bewertet. Dazu wurde der Codierleitfaden (siehe Anhang A) verwendet, welcher aus Beispielen für jede der vier Bewertungskategorien zu den 19 Propositionen besteht. Die Beispiele wurden zunächst deduktiv erarbeitet und mittels der Concept Maps aus Schule A (siehe Tabelle 6.8) ange-

passt, sodass die Codierbeispiele den sprachlichen Praktiken der Lernenden besser entsprechen. Zusätzlich wurde ein Handbuch für die Bewertung erstellt, welches alle relevanten Inhalte aufgreift und erklärt.
Wie in Abschnitt 6.1 beschrieben, sollen für die Bewertung nur die Propositionen und keine graphentheoretischen Maße der Concept Map betrachtet werden. Daher wurden sämtliche Concept Maps als Liste von Propositionen betrachtet, die von zwei menschlichen Bewertern ausgewertet werden können. Die Bewertung führte einerseits der Autor dieser Arbeit durch und andererseits ein Studierender der Nanotechnologie, der Erfahrung in Rating von physikalischen Datensätzen hat. Da der Inhalt der Concept Map der Physik aus der Sekundarstufe II entspricht, können beide menschlichen Bewerter als Experten angesehen werden.

Proposition	**H**	κ
Beschleunigung – Freier Fall	202	0,86
Kraft – Beschleunigung	174	0,78
Beschleunigung – Gleichmäßig beschleunigte Bewegung	173	0,86
Gleichförmige Bewegung – Zeit-Weg-Graph	167	0,48
Gleichmäßig beschleunigte Bewegung – Zeit-Geschwindigkeit-Graph	175	0,67
Gleichförmige Bewegung – Waagerechter Wurf	176	0,94
Gleichförmige Bewegung – Zeit-Geschwindigkeit-Graph	161	0,79
Gleichmäßig beschleunigte Bewegung – Zeit-Beschleunigung-Graph	153	0,87
Gleichmäßig beschleunigte Bewegung – Waagerechter Wurf	171	0,92
Beschleunigung – Gleichförmige Bewegung	155	0,87
Geschwindigkeit – Gleichmäßig beschleunigte Bewegung	161	0,88
Geschwindigkeit – Gleichförmige Bewegung	177	0,96
Gleichförmige Bewegung – Zeit-Beschleunigung-Graph	154	0,65
Masse – Freier Fall	218	0,87
Gleichmäßig beschleunigte Bewegung – Zeit-Weg-Graph	162	0,85
Geschwindigkeit – Freier Fall	205	0,94
Kraft – Geschwindigkeit	155	0,57
Beschleunigung – Geschwindigkeit	181	0,88
Gleichmäßig beschleunigte Bewegung – Freier Fall	182	0,83
Joker-Proposition	20	-
Summe	**3322**	-

Tabelle 6.9: Häufigkeiten (H) der 19 Propositionen und Cohen's Kappa für die Übereinstimmung zwischen den beiden menschlichen Ratern

Die Gesamtübereinstimmung zwischen den beiden menschlichen Bewertern lag bei 87 %, mit einem Cohen's Kappa von $\kappa = 0,83$. Nach Landis und Koch (1977) entspricht dieser Wert einer fast vollkommenen Übereinstimmung. Betrachtet man

die Übereinstimmung für die einzelnen Propositionen, findet man eine Schwankung zwischen $\kappa = 0,48$ und $\kappa = 0,96$. Lediglich die beiden Propositionen *gleichförmige Bewegung – Zeit-Weg-Graph* und *Kraft – Geschwindigkeit* weisen eine mittelmäßige Übereinstimmung auf. Fünf weitere Propositionen sind in dem Bereich der beachtlichen Übereinstimmung und demnach zwölf Propositionen im fast vollkommenen Übereinstimmungsbereich (siehe Tabelle 6.9).
Für die Entwicklung des Machine-Learning-Modells benötigt es jedoch nur eine menschliche Bewertung. Daher wurden die unklaren Propositionen noch einmal von beiden Bewertern gemeinsam gesichtet, sodass eine abschließende Bewertung festgelegt werden konnte. Wie in Kapitel 4 beschrieben, entspricht diese Bewertung nun dem Goldstandard, welcher für die Berechnung der Kennzahlen zur Überprüfung der Machine-Learning-Modelle genutzt wird. Insgesamt ergibt sich eine Verteilung der 3.302 Propositionen auf die vier Bewertungskategorien: Kategorie A - 32 % (1.054), Kategorie B - 19 % (615), Kategorie C - 26 % (867), Kategorie D - 23 % (766). Im Datensatz befindet sich demnach knapp ein Drittel Propositionen, die als falsch bewertet wurden. Kategorie B findet sich mit 19 % der Propositionen am wenigsten im Datensatz wider.
In Abbildung 6.10 findet man die Verteilung der Bewertungskategorien innerhalb der 19 Propositionen. Die Abbildung zeigt, dass keine Proposition eine Gleichverteilung der vier Bewertungskategorien aufweist. Es existieren sogar Propositionen, die eine sehr geringe Anzahl einer bestimmten Bewertungskategorie besitzen oder sogar lediglich aus drei Kategorien bestehen wie Proposition *Beschleunigung – Gleichförmige Bewegung*. Es lassen sich auch Propositionen finden, bei denen eine Bewertungskategorie deutlich dominiert. Bei der Proposition *gleichmäßig beschleunigte Bewegung – Freier Fall* findet man 70 % der Bewertungen in der Klasse B wieder, wohingegen bei *Geschwindigkeit – Gleichförmige Bewegung* Kategorie C mit 69 % dominiert (siehe Abbildung 6.10).

6.2.4 Entwicklungsschritte

Die erhobenen und bewerteten Propositionen dienten als Datengrundlage für die Entwicklung des Machine-Learning-Modells. Das Ziel war es, ein Modell zu entwickeln, das die Propositionen automatisch auswerten kann und dabei eine Übereinstimmung von mindestens $\kappa = 0,70$ mit den menschlichen Bewertungen aufweist (Forschungsfrage 1). Für die Entwicklung des Machine-Learning-Modells wurde der in Kapitel 4 beschriebene Ablauf genutzt.

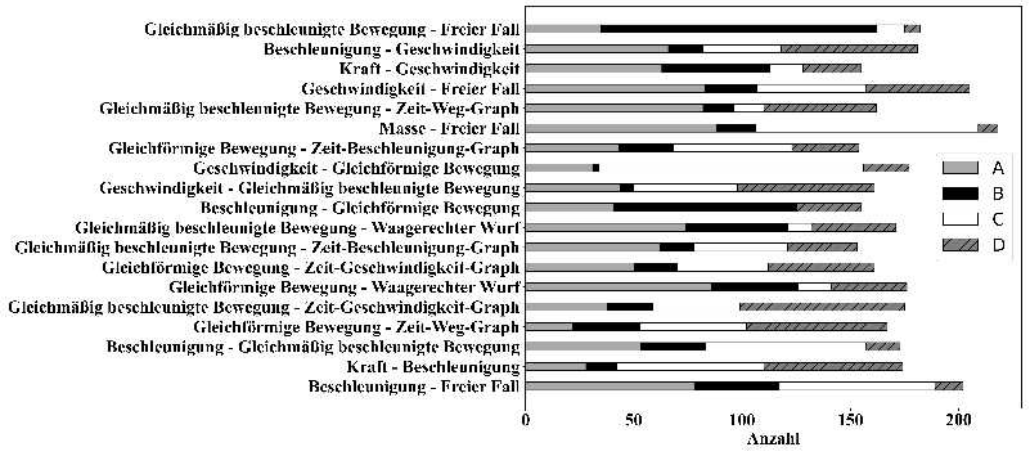


Abbildung 6.10: Häufigkeitsverteilung der 19 Propositionen bezüglich der vier Bewertungskategorien A, B, C und D

Vorverarbeitung der Proposition

Für die Vorverarbeitung können unterschiedliche Methoden angewendet werden, um die Qualität des Datensatzes zu erhöhen (siehe Abschnitt 4.1).
Die Beispiele aus Tabelle 6.10 zeigen, dass die Lernenden für ihre Propositionen nicht nur Wörter, sondern auch Zahlen, Formeln oder Einheiten genutzt haben. Von den 3.302 Propositionen enthielten 558 entweder eine Formel, eine Zahl, eine Einheit oder eine Kombination dessen. Dabei variierte die Schreibweise dieser strukturierten Daten teilweise erheblich, da z. B. als Divisionszeichen ein Doppelpunkt oder ein Schrägstrich verwendet wurde (siehe Beispiel 5 in Tabelle 6.10). Um die Struktur der Formeln nicht zu verändern oder sogar zu entfernen, wurden deshalb keine Satzzeichen, wie !,(,*;+,-,.,/,: oder Zahlen gelöscht. Es wurden auch keine Stoppwörter wie Präpositionen oder die Negation *nicht* entfernt, da die Propositionen im Datensatz teilweise sehr kurz sind und z. B. die Entfernung einer Verneinung den Inhalt der Proposition verändern kann (siehe Beispiel 4 in Tabelle 6.10). Zudem wird auf eine Normierung der Wörter verzichtet, da Kapitel 4 gezeigt hat, dass die Bedeutung der Wörter verloren gehen kann. Dadurch soll eine bessere Aussagekraft über die Performance des Machine-Learning-Modells erhalten werden.
Als einziger Vorverarbeitungsschritt wurde der Datensatz auf Duplikate untersucht. Für die Suche nach Duplikaten wurde ein selbst geschriebenes Python-Skript genutzt, das inhaltlich identische Propositionen herausfiltern kann. Um die Vielfältigkeit der Propositionen, die Formeln enthalten, beizubehalten, wurde bei der Suche nur auf identische Schreibweisen geachtet. Angesichts dessen werden inhalt-

Nr.	Proposition
1	Beschleunigung - ist $9,81 m/s^2$ - Freier Fall
2	Beschleunigung - V=a*t - Geschwindigkeit
3	Beschleunigung - verändert - Geschwindigkeit
4	Beschleunigung - ist nicht Teil einer - Gleichförmige Bewegung
5	Kraft a = F:m Beschleunigung

Tabelle 6.10: Fünf Beispiele aus dem Datensatz von Lernenden-Propositionen

lich identische Formeln wie $v = a \times t$, $v = a.t$ und $v = a \cdot t$ im Datensatz behalten, da sie eine andere Schreibweise aufweisen. Durch das Entfernen der Duplikate wird zwar in die Verteilung der Daten eingegriffen. Es sollte jedoch vermieden werden, dass das Machine-Learning-Modell mit denselben Propositionen trainiert und getestet wird. Durch die Entfernung von Duplikaten aus dem Datensatz kann das Modell besser generalisieren, da es nicht übermäßig auf bestimmte Beispiele trainiert wird. Zudem können zu viele Duplikate zu Overfitting führen, bei dem das Modell sich zu stark an die Trainingsdaten anpasst und die Leistung auf neuen Daten sinkt. Letztlich wurde die Entscheidung, Duplikate zu entfernen, getroffen, um eine validere Aussage über die Leistungsfähigkeit des Modells treffen zu können.

Durch die Entfernung von Duplikaten hat sich der Datensatz von 3.302 auf 2.315 Propositionen reduziert. Es wurden demnach 987 Propositionen gelöscht (siehe Tabelle 6.11). Die Häufigkeiten der einzelnen Propositionen schwanken nun zwischen 96 und 161, mit einer mittleren Häufigkeit von 122. Betrachtet man den neuen Datensatz, erkennt man, dass je nach Proposition zwischen 18 % und 47 % als Duplikate erkannt und demnach entfernt wurden. Am häufigsten konnte bei den Propositionen *Gleichmäßig beschleunigte Bewegung – Freier Fall*, *Geschwindigkeit – Gleichförmige Bewegung* und *Gleichförmige Bewegung – Waagerechter Wurf* mit über 40 % Duplikaten identifiziert werden. Wenn man einen genaueren Blick in den Datensatz und die Duplikate wirft, findet man bei *Gleichmäßig beschleunigte Bewegung – Freier Fall* 62-mal die Proposition *Gleichmäßig beschleunigte Bewegung ist ein Freier Fall* oder bei *Geschwindigkeit – Gleichförmige Bewegung* 24-mal die Proposition *Geschwindigkeit konstant Gleichförmige Bewegung*. Diese Beispiele zeigen, dass es charakteristische Propositionen gibt, die aus dem Datensatz gelöscht wurden.

Die Verteilung der vier Bewertungskategorien hat sich durch das Entfernen der Duplikate ebenfalls verändert. Von den 987 Propositionen waren 18 % mit Kate-

Nr.	Proposition	Duplikate	Neue Anzahl
1	Beschleunigung – Freier Fall	41	161
2	Kraft – Beschleunigung	52	122
3	Beschleunigung – Gleichmäßig beschleunigte Bewegung	47	126
4	Gleichförmige Bewegung – Zeit-Weg-Graph	56	111
5	Gleichmäßig beschleunigte Bewegung – Zeit-Geschwindigkeit-Graph	49	126
6	Gleichförmige Bewegung – Waagerechter Wurf	76	100
7	Gleichförmige Bewegung – Zeit-Geschwindigkeit-Graph	38	123
8	Gleichmäßig beschleunigte Bewegung – Zeit-Beschleunigung-Graph	33	120
9	Gleichmäßig beschleunigte Bewegung – Waagerechter Wurf	62	109
10	Beschleunigung – Gleichförmige Bewegung	47	108
11	Geschwindigkeit – Gleichmäßig beschleunigte Bewegung	38	123
12	Geschwindigkeit – Gleichförmige Bewegung	82	95
13	Gleichförmige Bewegung – Zeit-Beschleunigung-Graph	27	127
14	Masse – Freier Fall	74	144
15	Gleichmäßig beschleunigte Bewegung – Zeit-Weg-Graph	39	123
16	Geschwindigkeit – Freier Fall	56	149
17	Kraft – Geschwindigkeit	40	115
18	Beschleunigung – Geschwindigkeit	44	137
19	Gleichmäßig beschleunigte Bewegung – Freier Fall	86	96
	Summe	**987**	**2315**

Tabelle 6.11: Häufigkeitsverteilung der 19 Propositionen ohne Duplikate

gorie A, 25 % mit Kategorie B, 32 % mit Kategorie C, und 25 % mit Kategorie D bewertet. Insgesamt ergibt sich eine Verteilung der 2315 Propositionen auf die vier Bewertungskategorien: Kategorie A - 38 % (883), Kategorie B - 16 % (381), Kategorie C - 24 % (557), Kategorie D - 21 % (494). Dies spiegelt sich auch in der Verteilung innerhalb der einzelnen Propositionen wider (siehe Abbildung 6.11).

Transformation der Proposition

Damit das Machine-Learning-Modell die Propositionen verarbeiten kann, mussten diese zunächst in eine numerische Repräsentation umgewandelt werden. Wie die Ausarbeitung in Abschnitt 4.2 gezeigt hat, gibt es dazu unterschiedlich komplexe Ansätze. Die Beschreibung der verschiedenen Verfahren hat gezeigt, dass transformatorbasierte Sprachmodelle nicht nur in verschiedenen Kontexten angewandt werden können, sondern auch eine ausgezeichnete Performance ermöglichen. Daher wurde auch in dieser Arbeit solch ein Ansatz gewählt.
Für die Erzeugung einer numerischen Repräsentation der Propositionen mittels transformatorbasierter Sprachmodelle steht eine Vielzahl von Modellen zur Verfügung. Eine bekannte und viel genutzte Plattform für solche Transformer-Modelle ist Hugging Face (T. Wolf et al., 2019). Hugging Face ist eine Website mit einer Open-Source-Community, die sich auf KI und natürliche Sprachverarbeitung spe-

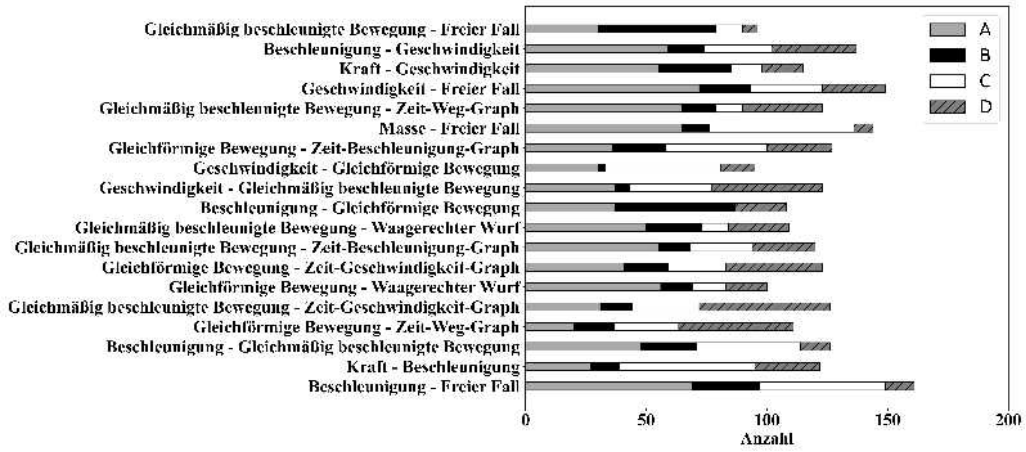


Abbildung 6.11: Häufigkeitsverteilung der 19 Propositionen ohne Duplikate bezüglich der vier Bewertungskategorien A, B, C und D

zialisiert hat. Über Hugging Face kann man auf eine umfangreiche Bibliothek von vorab trainierten Sprachmodellen zurückgreifen. Außerdem können Entwickler:innen Modelle und Datensätze auf Hugging Face teilen, diskutieren und für die Zusammenarbeit bereitstellen.

Für die Transformation der Propositionen wurde in dieser Arbeit ein SBERT-Modell genutzt. SBERT steht für Sentence-BERT, was eine Variante der BERT-Modelle darstellt (Reimers & Gurevych, 2019) (siehe auch Abschnitt 4.2). Die SBERT-Modelle verbessern traditionelle BERT-Modelle durch eine spezielle Architektur, was sich unter anderem in der Verarbeitungsgeschwindigkeit bemerkbar macht (Reimers & Gurevych, 2019). Im Prinzip nutzen die SBERT-Modelle ein normales BERT-Modell als Grundlage sowie eine zusätzliche Pooling-Schicht in der Architektur des neuronalen Netzes. Die Standardkonfiguration der Pooling-Schicht mittelt die durch das BERT-Modell entstehenden Wort-Vektoren. Durch diesen zusätzlichen Schritt entsteht nur ein Vektor pro Satz und nicht wie beim BERT-Modell mehrere Wortvektoren (Reimers & Gurevych, 2019). Da in dieser Arbeit die einzelnen Propositionen automatisch ausgewertet werden sollen, bietet sich solch ein Modell an. Es konnte so für jede Proposition ein Satzvektor (Sentence Embedding) erstellt werden, der dann als Input für das Modell dient.

Das in der vorliegenden Arbeit verwendete Modell *German Semantic STS V2* kann unter Hugging Face heruntergeladen werden[5]. Das Modell nutzt als Grundbau das *German-BERT-large*-Modell (Chan et al., 2020), das ebenfalls auf Hugging Face verfügbar ist. Das *German-BERT-large*-Modell wurde auf einer großen

[5]https://huggingface.co/aari1995/German_Semantic_STS_V2

Datenmenge deutscher Texte trainiert (163,4 GB) und stellte bei der Veröffentlichung 2020 das bisher beste deutsche Sprachmodell dar (Chan et al., 2020). Das *German-Semantic-STS-V2*-Modell nutzt für die Erstellung der Satzvektoren die Standardkonfiguration der Pooling-Schicht *mean-pooling*. Wie oben beschrieben, werden durch diesen Ansatz die generierten Wortvektoren gemittelt und dadurch ein Vektor für jede Proposition erstellt. Die entstehenden Vektoren haben eine Dimension von 1.024 und können für die Klassifikation der Propositionen genutzt werden.

Aufteilung der Propositionen

Für die Aufteilung der Daten in die verschiedenen Datensätze wurde eine Kreuzvalidierungsstrategie genutzt. Die Ergebnisse aus Kapitel 4 haben gezeigt, dass die Verwendung einer Kreuzvalidierungsstrategie robuste Ergebnisse bei der Entwicklung eines Klassifikations-Modells liefert. Zudem war der bereinigte Datensatz, im Vergleich zu anderen Studien aus diesem Bereich, nicht groß (Zhai et al., 2020). Durch die Anwendung der Kreuzvalidierung konnte das Modell auf dem gesamten Datensatz trainiert und getestet werden. Dadurch wurde eine verlässlichere Aussage über die Performance des Modells ermöglicht. Aus der Abbildung 6.11 geht hervor, dass im bereinigten Datensatz die Bewertungskategorien A bis D nicht gleich verteilt waren. Um zu ermöglichen, dass das Modell mit allen Bewertungskategorien trainiert und getestet wird, wurde eine *Stratified 10-Fold cross-validaton* zur Aufteilung in Trainings- und Testdaten verwendet (Train-Test-CV). Aus den 2.315 Propositionen wurden demnach zehnmal ein Trainingsdatensatz mit 90 % und ein Testdatensatz mit 10 % Propositionen erstellt (siehe Abbildung 6.12).
Zur Optimierung der Hyperparameter wurde ebenfalls eine Kreuzvalidierungsstrategie (Optimierung-CV) angewendet. Das bedeutet, dass der Trainingsdatensatz erneut zehnmal in einen Optimierungsdatensatz und einen Validierungsdatensatz aufteilt wird. Zur Aufteilung wurde dieselbe Strategie mit einer Aufteilung von 90 % zu 10 % gewählt (siehe Abbildung 6.12).

Trainings- und Optimierungsablauf

Aus Abbildung 6.12 geht der genaue Entwicklungsablauf des Machine-Learning-Modells hervor. Wie Kapitel 4 gezeigt hat, ist es schwer, im Vorfeld einen geeigneten Algorithmus zu bestimmen, weswegen die Entwicklung ein iterativer Prozess ist und mehrere Algorithmen getestet werden. Daher wird der skizzierte Ablauf

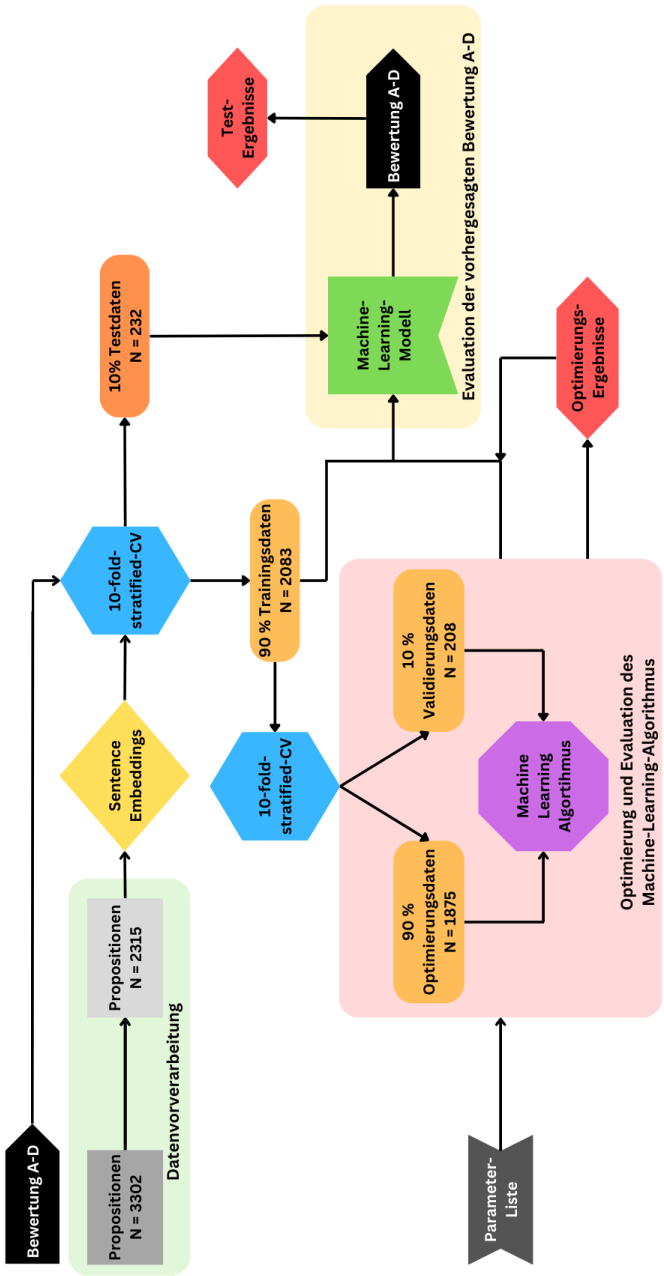


Abbildung 6.12: Entwicklungsprozess der Machine-Learning-Modelle

mit einer Auswahl von acht ausgewählten Algorithmen durchgeführt, um das beste Modell für die Klassifikation der Propositionen zu finden. Es werden folgende Algorithmen getestet: Support Vector Machine (SVM), Multi-Layer-Perceptron (MLP), K-Nearest-Neighbor (KNN), AdaBoost Classifier (ADA), Random Forest Tree (RFT), Decision Tree (DT), XGBoost Classifier (XGB) und Gradient Boosting Classifier (GBC). Die Wahl der Algorithmen war durch eine Analyse anderer Studien erfolgt (z. B. Krüger & Krell, 2020; Pedregosa et al., 2011; Wulff et al., 2020; Zhai et al., 2020). Für die Erstellung des Machine-Learning-Modells werden Python und die Bibliothek scikit-learn (Pedregosa et al., 2011) genutzt.
Nach der Aufteilung der Daten in den Trainings- und Testdatensatz wurden zunächst die Hyperparameter des gewählten Algorithmus optimiert. Dazu wurde im Vorfeld eine Liste von Parametern für jeden Algorithmus festgelegt (siehe Tabelle 6.12). Die Parameterkombinationen wurden dabei aus anderen Arbeiten übernommen oder nach Empfehlungen aus der Literatur gewählt (z. B. Banerjee, 2020; A. Jain, 2016; Shi et al., 2021; Vittorini et al., 2020).
Für die Parameteroptimierung wurde eine Rastersuche gewählt. Das hatte zur Folge, dass für jede Parameterkombination ein Modell erstellt wird. Das bedeutet z. B., für *K-Nearest-Neighbor* mussten 80 verschiedene Modelle trainiert und getestet werden. Dies ergibt sich aus den vier Möglichkeiten des ersten Parameters und den 20 Möglichkeiten des zweiten Parameters. Jedes dieser Modelle wurde durch die Optimierung-CV zehnmal auf den Optimierungsdaten trainiert und mit den Validierungsdaten getestet (siehe Tabelle 6.12).
Als Optimierungsparameter wurde Cohen's Kappa gewählt. Der F1-Score wäre auch eine Möglichkeit für einen Optimierungsparameter gewesen. Allerdings misst Cohen's Kappa nicht nur die Genauigkeit der Vorhersagen, sondern auch die Übereinstimmung über das hinaus, was durch Zufall zu erwarten wäre. Zudem ist es durch frühere Studien aus dem Forschungsbereich einfach zu interpretieren. Durch die Optimierung-CV entstanden für jede Parameterkombination zehn verschiedene Kappas, die nach dem zehnten Durchgang zu einem Optimierungs-Kappa gemittelt wurden. Die Parameter sowie das Optimierungs-Kappa werden als Optimierungs-Ergebnis ausgegeben und gespeichert (siehe Abbildung 6.12).
Nachdem alle Parameterkombinationen für das jeweilige Modell getestet wurden, wurde untersucht, welche Parameterkombination das beste Optimierungs-Kappa aufwies. Mit diesem Set an Parametern wurde ein neues *optimales* Modell auf den gesamten Trainingsdaten trainiert (siehe Abbildung 6.12). Dieses Modell wurde anschließend auf den zurückgehaltenen Testdaten getestet und final evaluiert. Da

Algorithmus	**Hyperparameter**	
	Parameter	Parameterbereich
Support Vector Machine (SVM)	C	$1^x, x \in [-5,5]$
	γ	$2^x, x \in [-15,-13,\ldots,2,4,]$, auto
	Kernel	linear, rbf, sigmoid, poly
Multi-Layer-Perceptron (MLP)	Activation	tanh, relu
	α	[0,0001, 0,001, 0,01, 0,1]
	Hidden Layer	[(100,0), (10,30,10), (20,0), (50,50,50), (50,100,50)]
	Learning Rate	constant, adaptive
	Solver	sgd, adam
K-Nearest-Neighbor (KNN)	Algorithm	auto, ball_tree, kd_tree, brute
	N neighbors	[1, 2, ..., 20]
AdaBoost Classifier (ADA)	Learning Rate	[0,001, 0,005, 0,01, ..., 1]
	N estimators	[1, 5, 10, ..., 300]
Random Forest Tree (RFT)	Max depth	[2, 4, 6, 8, 10, 15, 20, 25, 30], none
	Max features	sqrt, log2
	N estimators	[1, 5, 10, ..., 300]
Decision Tree (DT)	Max depth	[2, 4, 6, 8, 10, 15, 20, 25, 30], none
	Max features	sqrt, log2
	Splitter	best, random
XGBoost Classifier (XGB)	Booster	gbtree, dart, gblinear
	η	[0,001, 0,005, 0,01, ..., 1]
	Max depth	[2, 4, 6, 8, 10, 15, 20, 25, 30], none
Gradient Boosting Classifier (GBC)	Learning Rate	[0,001, 0,005, 0,01, ..., 1]
	Max depth	[2, 4, 6, 8, 10, 15, 20, 25, 30], none
	Max features	sqrt, log2
	N estimators	[1, 5, 10, ..., 300]

Tabelle 6.12: Hyperparameter der eingesetzten Algorithmen

das Modell diese Daten bisher nicht kannte, konnte dadurch eine verlässliche Aussage über die Performance getroffen werden.

Durch die Train-Test-CV wurde der beschriebene Prozess zehnmal wiederholt, sodass auf dem gesamten Datensatz sowohl optimiert als auch trainiert und getestet wurde. Als finales Ergebnis wurden die unterschiedlichen Kennzahlen (siehe nächster Abschnitt) sowie die optimalen Parameter ausgegebenen. Da die Struktur der Daten durch die Aufteilung in die zehn Blöcke variieren kann, können schwankende Ergebnisse und Parameter entstehen.

Als zusätzlicher Hyperparameter wurde eine Dimensionsreduktion mittels Hauptkomponentenanalyse (Principal Component Analysis, PCA) durchgeführt (Ringnér, 2008). Dieser Schritt hatte sich in einigen Arbeiten als hilfreich erwiesen und reduziert die hohe Dimension von 1.024 der Sentence Embeddings auf eine niedrigere Dimension. Durch die Reduktion können unter anderem Rauschen in den Daten unterdrückt und Redundanz reduziert werden (Ringnér, 2008). Dieser Schritt wurde

noch vor der Aufteilung der Daten in die verschiedenen Datensätze durchgeführt und zählt deshalb nicht zu den Hyperparametern der eigentlichen Algorithmen. Da die ursprüngliche Dimension der Sentence Embeddings 1.024 ist, wurden die Dimensionen 512, 256, 128, 64, 32, 16 und 8 getestet. Die Abstufung der Dimensionen folgte einem logarithmischen Muster, wobei jede Dimension halb so groß ist wie die vorherige.

Kennzahlen zur Bestimmung der Übereinstimmung

Zur Überprüfung wurden die in Kapitel 4 eingeführten Kennzahlen genutzt. Dazu zählen die Treffsicherheit (Accuracy), Cohen's Kappa, Precision, Recall und der F1-Score. Als Basis für die Berechnung diente die menschliche Bewertung, der als Goldstandard angesehen wird (siehe Abschnitt 6.2.3). Um das beste Modell zu finden, wurden die Durchschnittswerte der Kreuzvalidierung für die Treffsicherheit und Cohen's Kappa betrachtet. Zudem wurde ein gewichteter F1-Score angegeben, der die Anzahl der Proposition für jede Kategorie berücksichtigte. Zur Verbesserung der Übersichtlichkeit werden ausschließlich die Ergebnisse auf dem Testdatensatz und nicht die Optimierungswerte berichtet.
Um eine Grundlage für die Aussage über die Performance der Modelle zu erhalten, wurde zusätzlich noch ein sogenannter Dummy-Classifier genutzt. Ein Dummy-Classifier ist ein einfacher Klassifizierungs-Algorithmus, der als minimale Performancegrenze für die anderen Klassifizierungs-Algorithmen dient. Die Vorhersagen werden dabei auf simplen Regeln getroffen und nicht auf gelernten Zusammenhängen (Pedregosa et al., 2011). Der hier verwendete Dummy-Classifier erzeugt Vorhersagen, indem er die Verteilung der vier Bewertungskategorien in den Trainingsdaten nachbildet. Dies bedeutet, dass der Dummy-Classifier bei einem Anteil von 20 % einer Klasse im Trainingsdatensatz bei 20 % der Fälle der Testdaten diese Klasse zufällig rät.

6.2.5 Erste Ergebnisse

Phase 2 der Entwicklungsstudie sollte überprüfen, ob die ausgewählten Modelle die geforderte Übereinstimmung von $\kappa = 0,70$ erreichen konnten. In der anschließenden Phase 3 in Abschnitt 6.3 werden die Modelle, die diese Übereinstimmung erreichen konnten, genauer analysiert. Deshalb werden im folgenden Abschnitt nur Kennzahlen für die Gesamtleistung verglichen und z. B. keine genauere Untersuchung der Performance bezüglich der verschiedenen Bewertungskategorien durchgeführt.

	SVM	MLP	KNN	ADA	RFT	DT	XGB	GBC	Dummy
Acc.	**0,82** $\pm$ 0,03	0,80 $\pm$ 0,02	0,77 $\pm$ 0,02	0,55 $\pm$ 0,03	0,74 $\pm$ 0,02	0,60 $\pm$ 0,02	0,75 $\pm$ 0,03	0,76 $\pm$ 0,04	0,28 $\pm$ 0,02
κ	**0,75** $\pm$ 0,04	0,72 $\pm$ 0,02	0,69 $\pm$ 0,02	0,36 $\pm$ 0,04	0,64 $\pm$ 0,04	0,45 $\pm$ 0,04	0,65 $\pm$ 0,04	0,66 $\pm$ 0,04	0,02 $\pm$ 0,05
F1	**0,82**	0,80	0,77	0,54	0,74	0,60	0,75	0,76	0,28

Tabelle 6.13: Accuracy (Acc.), Cohen's Kappa und gewichteter F1-Score für die acht Machine-Learning-Modelle und Dummy-Classifier

Tabelle 6.13 zeigt die Leistung der acht verschiedenen Machine-Learning-Modelle auf dem reduzierten Datensatz. Die Support Vector Machine erzielte in allen drei ausgewählten Metriken die besten Werte. Eine durchschnittliche Treffsicherheit von 82 % und ein Cohen's Kappa von $\kappa = 0,75$ weisen auf eine gute Übereinstimmung zwischen dem Modell und den menschlichen Bewertungen hin. Der hohe gewichtete F1-Score von 0,82 zeigte zudem eine gute Balance zwischen Precision und Recall.

Eine ähnlich gute Leistung konnte das MLP erzielen. Durch ein Cohen's Kappa von $\kappa = 0,72$ und ein F1-Score von 0,80 kann man auch hier von einer guten Klassifizierungsfähigkeit sprechen. Die anderen Modelle KNN, RFT, XGB und GBC zeigten ebenfalls eine solide Leistung, die jedoch etwas niedriger im Vergleich zu SVM und MLP war. Der AdaBoost Classifier und der Decision Tree hatten die niedrigste Leistung bei der automatischen Auswertung der Propositionen erzielt. Sowohl die beiden Cohen's Kappa als auch die F1-Scores weisen auf eine im Vergleich zu den anderen Modellen schlechtere Genauigkeit bei der Klassifizierung der Propositionen hin.

Generell kann man in Tabelle 6.13 erkennen, dass die Standardabweichungen der jeweiligen Modelle sehr gering war. Daraus kann geschlossen werden, dass bei der Kreuzvalidierungsstrategie stabile Werte erzielt wurden. Zudem zeigen die Ergebnisse, dass alle entwickelten Modelle die Leistung des Dummy-Classifiers deutlich übertrafen. Das deutet darauf hin, dass die Modelle im Vergleich zum Dummy-Classifier durch den Trainingsprozess Zusammenhänge in den Daten gelernt haben. Es ist jedoch anzumerken, dass kein Modell die Mensch-Mensch-Übereinstimmung von $\kappa = 0,82$ erreichen konnte.

Insgesamt konnten mit der Support Vector Machine und dem Multi Layer Perceptron zwei Modelle entwickelt werden, die die geforderte Übereinstimmung von $\kappa = 0,70$ zwischen Mensch und Modell bei der Klassifikation der Propositionen erreichen konnten. Diese beiden Modelle werden nun im folgenden Abschnitt genauer analysiert, um letztlich eine begründbare Entscheidung für ein finales Machine-Learning-Modell zu treffen.

6.3 Phase 3: Analyse der automatischen Auswertung

Die Auswahl des richtigen Machine-Learning-Modells für die automatische Auswertung der Propositionen ist entscheidend für den Einsatz in der Schule. In der letzten Phase der Entwicklungsstudie wird daher die Leistung der beiden Modelle analysiert und verglichen.
Zunächst werden die automatischen Bewertungen beider Modelle in Bezug auf die vier Bewertungskategorien untersucht. Ziel soll es sein, einen besseren Einblick in die Bewertung der Proposition zu bekommen und zu schauen, ob es zwischen den Bewertungskategorien unterschiedliche Übereinstimmungswerte gibt. Hier soll eine erste Entscheidung für ein Modell getroffen werden. Anschließend werden die Propositionen, die das Modell falsch klassifiziert hat, betrachtet. Durch diese inhaltliche Überprüfung sollen spezifische Strukturen in dem Antwortverhalten der Lernenden identifiziert werden, die einen Einfluss auf die Leistung des Machine-Learning-Modells bei der automatischen Auswertung der Proposition hat. Abschließend werden noch die beiden Modelle mit Blick auf eine mögliche Verzerrung untersucht.
Diese Analysen werden helfen, die Stärken und Schwächen zu verstehen und fundierte Schlussfolgerungen darüber zu ziehen, ob ein Machine-Learning-Modell geeignet ist, um es in der Feedbackstudie als automatisches Feedback-Tool einzusetzen.

6.3.1 Ergebnisse der Machine-Learning-Modelle bezüglich der vier Bewertungskategorien

Für eine detaillierte Analyse der beiden Modelle reicht die Betrachtung von Cohen's Kappa und F1-Score für die gesamten Vorhersagen nicht aus, weil z. B. keine Aussagen über die automatische Bewertung einzelner Bewertungskategorien gemacht werden können. Durch die Confusion Matrix können die Vorhersagen der einzelnen Klassen betrachtet und so Precision, Recall und der F1-Score für die jeweiligen Bewertungskategorien bestimmt werden. Abbildung 6.13 zeigt die Confusion Matrix (CFM) der beiden Modelle SVM und MLP. Zeilenweise findet man die menschliche Bewertung und spaltenweise die Bewertung des jeweiligen Modells für die vier Bewertungskategorien. In beiden CFMs fällt auf, dass die Modelle den Großteil der Proposition den richtigen Kategorien zugeordnet hatten.

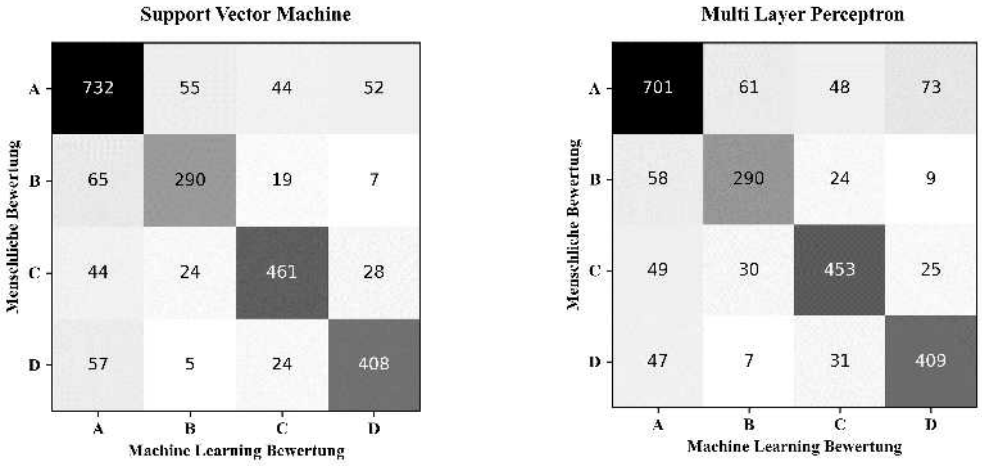


Abbildung 6.13: Confusion Matrix (CFM) der entwickelten Modelle Support Vector Machine (links) und Multi-Layer-Perceptron (rechts)

Um dennoch die Unterschiede herauszuarbeiten, können Precision, Recall und der F1-Score berechnet werden (siehe Tabelle 6.14).
Für die Bewertungskategorie A (falsche Propositionen) hatte die SVM einen leicht höheren Recall als Precision, wobei der Unterschied nur minimal ist. Beim MLP war die Differenz stärker ausgeprägt. Für die Bewertungskategorie A zeigte das Modell eine höhere Precision als Recall. Das bedeutet, dass der MLP dazu neigte, weniger falsche Propositionen in die anderen drei Bewertungskategorien zu klassifizieren, jedoch einige falsche Propositionen übersah. Es gab demnach einige falsche Propositionen, die der MLP nicht erkannt hatte.
Für die Bewertungskategorie B verhielt es sich genau andersherum. Die SVM hatte eine höhere Precision und der MLP einen höheren Recall, wobei der Unterschied marginal war.
Bei den Propositionen der Kategorie C zeigte die SVM wieder einen kleinen Unterschied, während der MLP einen identische Precision und Recall hatte.
Einen größeren Unterschied findet man hingegen bei der letzten Kategorie D. Die SVM zeigte erneut nur eine minimale Differenz, wohingegen der MLP einen höheren Recall hatte. Der höhere Recall weist darauf hin, dass der MLP wenige Propositionen der Kategorie D nicht erkannte. Das Modell neigte allerdings dazu, auch Proposition anderer Bewertungskategorien in Kategorie D zu bewerten. Insgesamt wiesen die F1-Scores trotz der (kleinen) Unterschiede zwischen Recall und Precision insgesamt für beide Modelle hohe Werte auf, wie in Tabelle 6.14 dargestellt. Auffällig ist jedoch, dass Kategorie B im Vergleich zu den anderen drei Bewertungskategorien abfiel. Sowohl bei der SVM als auch beim MLP hatte die

	SVM			MLP		
	Precision	Recall	F1	Precision	Recall	F1
A	0,82	0,83	0,82	0,82	0,79	0,81
B	0,78	0,76	0,77	0,75	0,76	0,75
C	0,84	0,83	0,83	0,81	0,81	0,81
D	0,82	0,83	0,83	0,79	0,83	0,81

Tabelle 6.14: Precision, Recall und F1-Scores der beiden Modelle SVM und MLP für die jeweiligen Bewertungskategorien A-D

Kategorie B den niedrigsten F1-Score. Das deutet darauf hin, dass die Kategorie B für beide Modelle schwieriger zu klassifizieren war. Die Propositionen, die mit Kategorie B bewertet wurden, hatten jedoch auch die geringste Häufigkeit im Datensatz.
Betrachtet man nur die falsch klassifizierten Propositionen, findet man zwischen den beiden Modellen SVM und MLP eine große Schnittmenge. Die SVM hatte insgesamt 424 Propositionen nicht in dieselbe Bewertungskategorie bewertet wie die menschlichen Bewerter. Beim MLP waren es 462 Propositionen. Vergleicht man diese Propositionen, findet man 323 Propositionen in beiden Fällen wieder. Die restlichen falsch klassifizierten Propositionen sind über die 19 Propositionen fast gleich verteilt.
Daher ist ein weiterer wichtiger Aspekt die Betrachtung der falsch klassifizierten Propositionen in der Bewertungskategorie A. Es wäre wünschenswert, wenn die Modelle möglichst wenig falsche Propositionen in eine der drei Bewertungskategorien B-D einordnen. Da das Bewertungsschema keine ordinale Skalierung aufweist, wird jede fehlerhafte Klassifizierung identisch behandelt. Aus diesen Gedanken heraus und aus den CFMs in Abbildung 6.13 kann die $False_A$-Rate für beide Modelle berechnet werden:

$$False_A^{SVM} = \frac{55+44+52}{732+55+44+52} = 0,17$$

$$False_A^{MLP} = \frac{61+48+73}{701+61+48+73} = 0,21$$

Aus diesen beiden Werten kann geschlossen werden, dass die SVM weniger falsche Propositionen in einer der drei Bewertungskategorien mit richtigen Propositionen klassifiziert hatte.

Obwohl der Großteil der Proposition von beiden Modellen in dieselben Bewer-

Nr.	Proposition	Mensch-Mensch	Goldstandard-SVM	Differenz
1	Beschleunigung – Freier Fall	0,86	0,69	0,17
2	Kraft – Beschleunigung	0,78	0,68	0,10
3	Beschleunigung – Gleichmäßig beschleunigte Bewegung	0,86	0,72	0,14
4	Gleichförmige Bewegung – Zeit-Weg-Graph	0,48	0,73	-0,25
5	Gleichmäßig beschleunigte Bewegung – Zeit-Geschwindigkeit-Graph	0,67	0,67	0,00
6	Gleichförmige Bewegung – Waagerechter Wurf	0,94	0,78	0,16
7	Gleichförmige Bewegung – Zeit-Geschwindigkeit-Graph	0,79	0,73	0,06
8	Gleichmäßig beschleunigte Bewegung – Zeit-Beschleunigung-Graph	0,87	0,81	0,06
9	Gleichmäßig beschleunigte Bewegung – Waagerechter Wurf	0,92	0,76	0,16
10	Beschleunigung – Gleichförmige Bewegung	0,87	0,71	0,16
11	Geschwindigkeit – Gleichmäßig beschleunigte Bewegung	0,88	0,78	0,10
12	Geschwindigkeit – Gleichförmige Bewegung	0,96	0,78	0,18
13	Gleichförmige Bewegung – Zeit-Beschleunigung-Graph	0,65	0,58	0,07
14	Masse – Freier Fall	0,87	0,85	0,02
15	Gleichmäßig beschleunigte Bewegung – Zeit-Weg-Graph	0,85	0,76	0,09
16	Geschwindigkeit – Freier Fall	0,94	0,75	0,19
17	Kraft – Geschwindigkeit	0,57	0,59	-0,02
18	Beschleunigung – Geschwindigkeit	0,88	0,85	0,03
19	Gleichmäßig beschleunigte Bewegung – Freier Fall	0,83	0,56	0,27

Tabelle 6.15: Cohen's Kappa der jeweiligen Propositionen für Mensch zu Mensch und Goldstandard (finale Bewertung der beiden menschlichen Bewerter) zu SVM-Übereinstimmung

tungskategorien wie von den menschlichen Bewertern klassifiziert wurden, zeigte die SVM eine insgesamt bessere Leistung. Dieses Modell hatte im Vergleich zum MLP für jede Bewertungskategorie eine bessere Leistung in Bezug auf das Cohen's Kappa, die Precision, den Recall und den F1-Score erzielt. Die Verbesserung in allen Metriken deutet darauf hin, dass die SVM besser in der Lage war, die Propositionen zu „verstehen", Muster zu erkennen und präzise Bewertungen durchzuführen. Besonders hervorzuheben ist die niedrigere $False_A$-Rate der SVM im Vergleich zum MLP. Es zeigt, dass die SVM weniger Propositionen, die eigentlich physikalisch falsch waren, in die Bewertungskategorie B-D zuordnen, was das Risiko von Fehlinterpretationen minimiert. Insgesamt zeigt die SVM also die bessere Leistung, weswegen im Folgenden dieses Modell für die weitere Analyse betrachtet wird.

6.3.2 Ergebnisse der SVM bezüglich der 19 Propositionen

Um weitere Stärken und Schwächen der SVM herauszufinden, werden in diesem Abschnitt die Übereinstimmungswerte für die 19 Propositionen analysiert. In Tabelle 6.15 sind die Cohen's Kappas für die Übereinstimmung zwischen Mensch und Mensch sowie zwischen Mensch und SVM für jede der 19 Propositionen aufgelistet.
Die Übereinstimmungswerte zeigen, dass bei 17 der 19 Propositionen die mensch-

lichen Bewerter eine höhere Übereinstimmung aufwiesen. Bei der Proposition *Gleichförmige Bewegung – Zeit-Weg-Graph* hatte jedoch die Mensch-SVM-Übereinstimmung einen deutlich höheren Wert. Hier wies das Machine-Learning-Modell eine beachtliche Übereinstimmung auf, wohingegen die menschlichen Bewerter nur eine moderate Übereinstimmung hatten. Bei der Proposition *Kraft – Geschwindigkeit* waren beide Übereinstimmungswerte im unteren Bereich, wobei die Mensch-SVM-Übereinstimmung ein wenig höher war. Insgesamt lag die Mensch-SVM-Übereinstimmung nur bei drei Propositionen in dem Bereich einer mittelmäßigen Übereinstimmung. Alle anderen Propositionen konnten eine beachtliche oder fast vollkommene Übereinstimmung nachweisen.
Die Häufigkeiten der drei problematischen Propositionen waren im Vergleich zu den anderen Propositionen nicht geringer. Deshalb wird ein genauerer Blick in den Datensatz geworfen, um Muster zu finden, die die schlechte Übereinstimmung nachvollziehbar machen.
Die mittlere Länge aller Proposition lag bei 61 ± 13 Zeichen. Die drei problematischen Propositionen hatten eine mittlere Länge von 68, 39 und 67. Das bedeutet, dass bei zwei der dreien Propositionen nicht die Länge Einfluss auf die Leistung brachte. Die Proposition *Kraft – Geschwindigkeit* mit einer mittleren Länge von 39 Zeichen war jedoch die Proposition mit der geringsten Anzahl an Zeichen. Jedoch hatte die Proposition *Kraft – Beschleunigung* nur eine leicht höhere mittlere Zeichenlänge, aber eine deutlich stärkere Übereinstimmung.
Bei der inhaltlichen Analyse der Propositionen kann erkannt werden, dass die drei problematischen Propositionen einen hohen Anteil an Formeln oder Zahlen aufwiesen. Deshalb wurden zur Überprüfung die Propositionen in zwei Gruppen aufgeteilt: Propositionen, die ausschließlich aus Text bestanden, wurden zu der Gruppe *Text* gezählt und Propositionen, die eine Formel, eine Zahl, eine Einheit oder eine Kombination davon enthielten, wurden zu der Gruppe *Formel* gezählt. Betrachtet man diese beiden Gruppen für den gesamten Datensatz, kann festgestellt werden, dass die Mehrheit der Propositionen mit 83 % zu der Gruppe *Text* und nur 17 % zur Gruppe *Formel* gehörten. Die drei problematischen Propositionen wiesen jedoch mit 27 %, 25 % und 19 % einen höheren Anteil an Formeln als der Durchschnitt (17 %) auf. Berechnet man für beide Gruppen Cohens's Kappa, zeigte sich eine deutliche höhere Übereinstimmung bei der *Text*-Gruppe:

$$\kappa_{Text} = 0.76 \quad \kappa_{Formel} = 0,61$$

Daraus kann geschlossen werden, dass die *Formel*-Propositionen einen Einfluss

auf die Leistung der SVM hatte.
Um die *Formel*-Gruppe noch genauer zu untersuchen ist in Abbildung 6.14, die Confusion Matrix für diese Gruppe dargestellt.

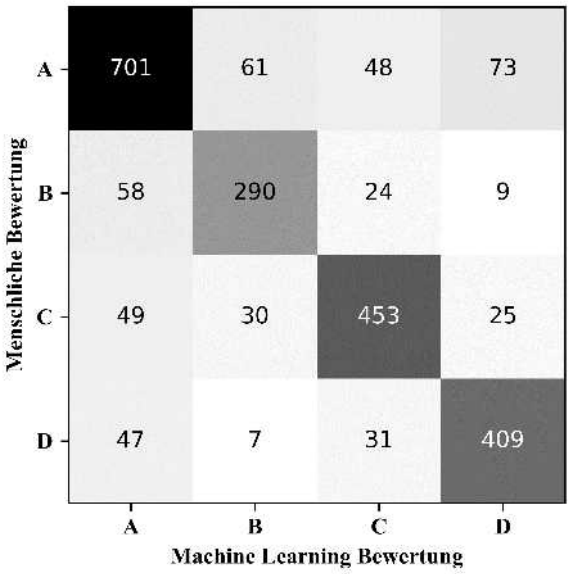


Abbildung 6.14: Confusion Matrix der SVM für die *Formel*-Gruppe

Aus Abbildung 6.14 ist ersichtlich, dass die meisten Propositionen der *Formel*-Gruppe die Bewertung A oder D hatten. Aufgrund des entwickelten Bewertungsschemas war diese Verteilung zu erwarten, da physikalisch richtige funktionale Zusammenhänge in Bewertungskategorie D bewertet werden sollten. Für die Bewertungskategorie D konnten auch die höchsten Werte für Precision, Recall und F1 erzielt werden (siehe Tabelle 6.16). Das deutet darauf hin, dass die SVM grundsätzlich eine gute Leistung bei der Klassifizierung von Propositionen der Klasse D erzielt hat. Die Kennzahlen für die Bewertungen B und C waren aufgrund der kleinen Datenmenge nur wenig aussagekräftig. Bei den falschen Propositionen der *Formel*-Gruppe hatte die SVM häufig die Kategorie D vorhergesagt. Die $False_A$-Rate war deshalb mit 0,43 im Vergleich zum Gesamtdatensatz deutlich höher. Für Propositionen der Kategorie D findet man einen zwar nicht ganz so ausgeprägten, aber ähnlichen Zusammenhang.
Angesichts dieser Ergebnisse liegt die Schlussfolgerung nahe, dass die SVM Probleme hatte, falsche Propositionen von Propositionen in der Bewertungskategorie D in der *Formel*-Gruppe zu unterscheiden.

	Precision	Recall	F1
A	0,74	0,70	0,72
B	0,57	0,59	0,58
C	0,69	0,71	0,70
D	0,80	0,82	0,81

Tabelle 6.16: Precision, Recall und F1-Score der SVM für die *Formel*-Gruppe

6.3.3 Ergebnisse der SVM bezüglich Lernendenmerkmale

Als letzter Analyseschritt werden die Bewertungen der SVM auf externe Attribute untersucht. Die Propositionen stammen aus der Erhebung mit 14 Klassen aus fünf verschiedenen Gymnasien (siehe Abschnitt 6.2.2). Um auszuschließen, dass die SVM bestimmte Klassen benachteiligt, ist in Tabelle 6.17 das Cohen's Kappa für jede Klasse dargestellt.
In Tabelle 6.17 erkennt man, dass Schwankungen bezüglich der Leistung der SVM zwischen den Klassen existieren. Der Mittelwert über alle Klassen liegt bei $\kappa = 0,74$. Die stärkste Abweichung vom Mittelwert haben demnach die Klassen 11 und 13. Unter dem gesetzten Grenzwert von $\kappa = 0,70$ liegen die drei Klassen 10, 11 und 12. Die Klassen 10 und 11 stammen von derselben Schule. Allerdings wurden die Klassen von unterschiedlichen Lehrkräften unterrichtet. Zudem waren die Klassen 6, 7, 8, und 9 ebenfalls von der Schule. Bei der Altersverteilung unterscheiden sich die drei Klassen nicht. Allerdings wurden in der Klasse 11 nur 30 Propositionen erstellt und der Anteil in der *Formel*-Gruppe ist mit 40 % am höchsten, was die schlechte Leistung erklären könnte (siehe Tabelle 6.8). Gegen diese Argumentation spricht die hohe Anzahl an Concept Maps in den anderen beiden Klassen. Auch der Anteil der *Formel*-Gruppe ist hier mit 24 % und 15 % deutlich geringer (siehe Tabelle 6.17). Angesichts dessen liegt eine Betrachtung weiterer Einflussfaktoren wie des Geschlechts und der schulischen Leistungen der Lernenden nahe.
Das Geschlecht ist innerhalb des Datensatzes mit 92 männlichen und 104 weiblichen Lernenden annähernd gleich verteilt. Auch die Übereinstimmung zwischen den menschlichen Bewertungen und der SVM ist für beide Geschlechter identisch:

$$\kappa_m = 0,74 \quad \kappa_w = 0,74$$

Dadurch kann gezeigt werden, dass das Geschlecht für die SVM keine Rolle bei der Bewertung der Propositionen spielt. Dies könnte bedeuten, dass menschliche

Klasse	Schule	Lehrkraft	Cohen's Kappa	Propositionen	Anteil *Formel*-Gruppe
1	A	L1	0,73	275	12 %
2	A	L1	0,73	200	3 %
3	A	L2	0,80	125	13 %
4	B	L3	0,80	213	18 %
5	B	L3	0,70	218	15 %
6	C	L4	0,81	159	18 %
7	C	L5	0,70	136	34 %
8	C	L6	0,70	123	27 %
9	C	L4	0,71	158	25 %
10	C	L4	0,69	182	24 %
11	C	L5	0,63	30	40 %
12	D	L7	0,64	204	15 %
13	D	L8	0,86	158	5 %
14	E	L9	0,81	165	18 %

Tabelle 6.17: Cohen's Kappa der SVM bezüglich der 14 Schulklassen

Bewertung ebenfalls geschlechtsneutral durchgeführt wurde oder andere Faktoren wie die schulische Leistung einen größeren Einfluss auf die Bewertungen der SVM hat.

Um den Einfluss der schulischen Leistung der Lernenden auf die Bewertung der SVM zu untersuchen, werden die Lernenden in zwei Gruppen aufgeteilt: Die Darstellung der Noten in Abbildung 6.9 hat gezeigt, dass der Median bei 10 Punkten liegt. Teilt man die Gruppen nach dem Median auf, erhält man ungefähr zwei gleich große Gruppen an Lernenden. Darauf lassen sich zwei Gruppen bilden: In der ersten Gruppe sind alle Lernenden, die eine Note von 10 Punkten oder besser angaben. Das sind also alles Lernende, die im jeweiligen Fach die Note *gut* oder *sehr gut* haben. In der zweiten Gruppe sind alle Lernenden, die eine Note von unter 10 Punkten angaben und dementsprechend eine schlechtere Note als *gut* haben.

Fach	Note $\geq$ 10	Note $<$ 10
Physik	0,74 (89)	0,74 (84)
Mathematik	0,74 (89)	0,74 (84)
Deutsch	0,74 (72)	0,74 (72)

Tabelle 6.18: Cohen's Kappa der SVM für die beiden Noten-Gruppen. In Klammern ist die jeweilige Anzahl der Lernenden angegeben.

Die Übereinstimmungswerte der SVM mit den menschlichen Bewertungen in Tabelle 6.18 zeigen, dass es keinen Unterschied zwischen den beiden Noten-

Gruppen gibt. Die Unterschiede sind nur in der vierten Nachkommastelle zu erkennen und deswegen nicht relevant. So kann gezeigt werden, dass auch die schulischen Leistungen keinen Einfluss auf die Bewertung der SVM haben.

6.3.4 Auswertung des Fragebogens

Da die Concept Map in der Feedbackstudie erneut eingesetzt werden soll, wurden neben den soziodemografischen Angaben im Fragebogen auch Verbesserungsvorschläge für die Concept Map erhoben. Insgesamt konnten so 152 Kommentare gesammelt werden. Der Großteil (63 Antworten) der Lernenden wünschte sich keine Verbesserung und war mit der Durchführung und Gestaltung der Concept Map zufrieden. Ein Teil der Lernenden (25 Antworten) kritisierte jedoch die fehlenden Pfeile in der Concept Map. Bei der Gestaltung in Abschnitt 6.1 wurden zur Verbindung zweier Begriffe Linien statt Pfeile gewählt, um den Lernenden einen gewissen Freiraum bei der Formulierung der Proposition zu lassen. Weitere 25 Lernende wünschen sich eine übersichtlichere Gestaltung der Concept Map. Vorschläge waren unter anderem, eine andere Schriftart zu wählen oder die Verwendung von unterschiedlichen Farben, um Zusammenhänge besser deutlich zu machen. Auch auf technischer Ebene konnten einige Verbesserungsvorschläge erhoben werden. Durch die Verwendung der Concept-Map-Software CmapTools waren die Lernenden in der Lage, die eingesetzte Concept Map auch strukturell zu bearbeiten. Das bedeutet, dass sowohl die Verbindungslinien als auch die Begriffe verschoben und verändert werden konnten. Diese Funktionen führten dazu, dass manche Lernende die Concept Map (unabsichtlich) verändert und deshalb Schwierigkeiten bei der Bearbeitung hatten. Der restliche Teil der Antworten bezog sich auf fachliche Schwierigkeiten und den Wunsch nach mehr Hilfestellung. So wurde z. B. angegeben, dass Auswahlmöglichkeiten für die Propositionen vorhanden sein sollten oder genauere (fachliche) Vorgaben gemacht werden sollten.

6.4 Beantwortung der Forschungsfragen und Diskussion der Ergebnisse

Die Entwicklungsstudie hatte zum Ziel, ein effektives Machine-Learning-Modell für die automatische Auswertung der Propositionen zu entwickeln und zu evaluieren. Dieses Modell soll in der nachfolgenden Feedbackstudie die Basis für ein

Feedback-Tool sein.
Bevor ein Machine-Learning-Modell trainiert und getestet werden konnte, wurden die Daten zunächst in drei Teilmengen aufgeteilt. Der Großteil (90 %) der Daten wurde zum Training und Optimieren der Modelle genutzt. Der zurückgehaltene Testdatensatz (10 %) wurde letztlich für die finale Evaluation genutzt. Durch die genutzte Kreuzvalidierung sollte erreicht werden, dass die Modelle auf dem gesamten Datensatz trainiert und getestet und so robuste Ergebnisse erzielt werden können. Für die Transformation der Textdaten in eine numerische Repräsentation wurden sogenannte Sentence Embeddings mit einem vortrainierten SBERT-Modell genutzt (siehe Kapitel 4 und Abschnitt 6.2).
Es wurden insgesamt acht Algorithmen genutzt. Durch eine Vielzahl von unterschiedlichen Hyperparametern konnten so verschiedene Modelle trainiert und getestet werden. In Kapitel 5 wurde ein Akzeptanzbereich für die Übereinstimmung zwischen menschlicher Bewertung und Bewertung, die durch ein Machine-Learning-Modell entstanden sind, festgelegt. Da die Anwendung der automatischen Auswertung im Rahmen eines formativen Assessments eingesetzt werden soll und nicht für Leistungstests, deren Ergebnisse mit Konsequenzen für die Lernenden verbunden sind, sollte ein Cohen's Kappa von mindestens $\kappa = 0,70$ erreicht werden. Es konnte gezeigt werden, dass mehrere Modelle eine gute Übereinstimmung mit den menschlichen Bewertungen hatten. Zwei Modelle, die Support Vector Machine und das Multi Layer Perceptron, erreichten mit ihrer Leistung die besten Übereinstimmungswerte der getesteten Modelle. Durch die verwendete Kreuzvalidierung konnte ebenfalls gezeigt werden, dass die Modelle konsistente Leistungen über verschiedene Teilmengen der Daten erreichen konnten. Die beiden Modelle waren daher in der Lage, die Übereinstimmungsgrenze von mindestens $\kappa = 0,70$ zu erreichen, weswegen die erste Forschungsfrage positiv beantwortet werden konnte: *Es kann mit der SVM und dem MLP eine Übereinstimmung von mindestens* $\kappa = 0,70$ *bei einer Bewertung der Propositionen einer Concept Map zwischen der menschlichen Bewertung und der durch Techniken des maschinellen Lernens generierten Bewertung erreicht werden.*
Durch eine genauere Analyse der beiden Machine-Learning-Modelle wurde gezeigt, dass die SVM leichte Vorteile bei der Klassifikation der einzelnen Bewertungskategorien aufweist. Die SVM hatte für jede Bewertungskategorie die bessere Precision und den besseren Recall und dementsprechend auch den höheren F1-Score. Zudem wurde die Fehlerrate bezüglich der Bewertungskategorie A, also dass eine Proposition eines Lernenden fälschlicherweise als korrekt vorhergesagt

wurde, bestimmt. Es wäre wünschenswert, wenn das Modell möglichst keine falschen Propositionen in eine der drei anderen Bewertungskategorien zuordnet. Dadurch könnten Lernende falsche Vorstellungen festigen, wenn das Machine-Learning-Modell die Lernenden in ihren Fehler bestärkt. Weiterhin könnte es zu Glaubwürdigkeitsproblemen für die Lehrkräfte führen, wenn das Machine-Learning-Modell den Lernenden zustimmt, obwohl ihre Propositionen eigentlich physikalisch falsch waren. Tritt der Fall ein, dass das Modell eine Proposition fälschlicherweise als falsche Proposition kennzeichnet, können die Lernenden immer noch selbstständig die Fehlklassifikation anmerken und die Bewertung des Machine-Learning-Modells anzweifeln. Im besten Fall entsteht so eine Diskussion über die Inhalte. Da die SVM auch die niedrige Fehlerrate hatte, bietet dieses Modell insgesamt die beste Leistung der getesteten Modelle und erfüllt am besten die Anforderungen an den spezifischen Anwendungsfall als Feedback-Tool. Allerdings konnte die SVM nicht die Übereinstimmungswerte der beiden menschlichen Bewerter von $\kappa = 0,82$ erreichen.

Durch die Verwendung des vortrainierten SBERT-Modells wurden die Propositionen in Sentence Embeddings transformiert. Die Verwendung von vortrainierten Sprachmodellen für die Klassifikation von Text ist eine weitverbreitete Praxis (siehe Kapitel 4). Dabei werden diese Sprachmodelle auf einem großen Datensatz trainiert, der Texte aus unterschiedlichen Bereichen enthält. Sentence Embeddings basieren dabei auf der Bedeutung von Wörtern und Sätzen. Physikalische Formeln und Zahlen haben jedoch keine Bedeutung im herkömmlichen Sinne. Außerdem kann davon ausgegangen werden, dass das genutzte SBERT-Modell auf Texten trainiert wurde, die kaum Formeln enthalten. Daher wurde die Hypothese formuliert, dass die Anwesenheit solcher Elemente negativ auf die Leistung des Modells wirken könnte. Bei der Analyse der Klassifikationen konnte festgestellt werden, dass die Propositionen, die eine Formel oder ähnliches enthalten, einen negativen Einfluss auf die Bewertung der Support Vector Machine hatten, während die Länge der Antworten keinen wesentlichen Einfluss zu haben scheint. Diese Ergebnisse können verwendet werdet, um die Forschungsfrage 2 zu beantworten: *Es können spezifische Strukturen in dem Antwortverhalten der Lernenden identifiziert werden, die einen Einfluss auf die Performance des Machine-Learning-Modells bei der automatischen Auswertung der Concept Maps haben*. Es kann zudem die Hypothese 2 bestätigt werden: *Die automatische Auswertung von Propositionen, die Formeln oder Zahlen enthalten, führt zu einer geringeren Performance*

des Machine-Learning-Modells im Vergleich zu Propositionen, die ausschließlich textuellen Inhalt aufweisen.
Der Anteil der Propositionen, die der *Formel*-Gruppe zugeordnet werden können, ist jedoch recht klein. Zudem ist der Unterschied in der Übereinstimmung zwischen der *Formel*- und *Text*-Gruppe zwar messbar, aber nicht so stark, dass es einen Einsatz in der Schule verhindern würde. Trotzdem könnten Lernende, die zum Beispiel durch eine hohe Expertise viele Zusammenhänge durch physikalische Formeln ausdrücken, von dem Modell benachteiligt werden. Diese Benachteiligung kann als eine Limitation des entwickelten Machine-Learning-Modells aufgezeigt werden. Bei der Entwicklung des Feedback-Tools muss daher auf diesen Aspekt geachtet werden. Bei der weiteren Betrachtung der Klassifikationen sind Zuordnungen der SVM aufgefallen, die keine nachvollziehbaren Gründe für die Entscheidung hatten. Für die Mehrzahl dieser Zuordnungen konnten keine Muster oder Gründe für die Entscheidung des Modells erkannt werden. Trotz der insgesamt zufriedenstellenden Übereinstimmung zeigt dies eine weitere Limitation des Modells.
Weitere Forschungsarbeiten könnten sich mit der Nutzung eines speziell auf physikdidaktische Kontexte abgestimmten Sprachmodells auseinandersetzen. In dieser Studie wurde ein SBERT-Modell genutzt, das zwar auf einer großen Bandbreite an deutschsprachigen Texten trainiert wurde, allerdings wurde der Hauptteil der Texte aus dem Internet genutzt. Sprachmodelle können Schwierigkeiten haben, mit Wörtern umzugehen, die nicht im Trainingskorpus enthalten waren. Diese Wörter könnten zu schlechteren Leistungen führen. Ein abgestimmtes Sprachmodell könnte die spezifischen sprachlichen Praktiken der Lernenden besser abbilden, was in einer besseren Leistung resultieren könnte. Ein weiterer Ansatz, den künftige Studien aufgreifen könnten, wäre ein zweistufiges Modell für die automatische Auswertung der Propositionen. In der ersten Stufe werden die zu klassifizierenden Propositionen mit einer vorher erstellten Datenbank verglichen. Diese Datenbank könnte aus den in dieser Studie gesammelten Propositionen und den dazugehörigen Bewertungen sein. Falls die Proposition bereits in der Datenbank enthalten ist, könnte die Bewertung aus der Datenbank übernommen werden. Dies hätte den Vorteil, dass Propositionen mit Formeln deutlich sicherer ausgewertet werden, da ein Großteil der Propositionen bereits in der Datenbank enthalten sind. Falls die Proposition nicht in der Datenbank ist, würde in der zweiten Stufe das Machine-Learning-Modell die automatische Auswertung übernehmen und eine Bewertung vorhersagen. Durch dieses Zwei-Stufen-Modell kann eine bessere

Übereinstimmung erwartet werden, was den Einsatz in Schulen, unabhängig von dem Zweck, verbessern würde.

Ein weiteres Problem, welches nicht nur in dieser Studie sichtbar wurde, ist die Notwendigkeit großer Datensätze. Wie die theoretische Ausarbeitung gezeigt hatte, gibt es Machine-Learning-Modelle, die mit einem enormen Datensatz von bis zu 27.000 Antworten trainiert wurden (Zhai et al., 2020). Oftmals ist es jedoch aus ökonomischen und ökologischen Gesichtspunkten schwierig, an eine solche große Datenmenge zu kommen. Daher wäre es für zukünftige Forschungsarbeiten hilfreich, gemeinsame Datensätze zu nutzen und die Erkenntnisse darüber zu teilen. Das würde nicht nur die Entwicklung von Machine-Learning-Modelle, die in Bildungskontext eingesetzt werden, deutlich vereinfachen, sondern die gesamte (didaktische) Evaluation transparenter und aussagekräftiger machen.

Die automatische Auswertung soll in der Feedbackstudie Hinweise für die Lehrkräfte und Lernenden bereitstellen, die für den weiteren Lehr-Lern-Prozess genutzt werden können. Daher musste ausgeschlossen werden, dass die SVM eine Verzerrung bezüglich bestimmter Attribute wie das Geschlecht der Lernenden aufweist. Die automatische Auswertung sollte nicht von dem Geschlecht oder den schulischen Leistungen der Lernenden beeinflusst werden, da jeder Lernende Zugang zu einem fairen Feedback haben sollte, unabhängig vom Geschlecht oder schulischen Leistungen. Wenn das Machine-Learning-Modell aufgrund dieser Faktoren voreingenommen wäre, könnte dies zu Ungleichheiten führen und bestimmte Lernendengruppen benachteiligen. Wenn Lernende und Lehrkräfte das Feedback-Tool als unfair wahrnehmen, könnte dies das Vertrauen in die automatische Bewertung beeinträchtigen. Ein ausgewogenes und gerechtes Feedback ist also entscheidend, um das Vertrauen teilnehmender Personen in das Tool zu stärken. Es wurde daher zunächst experimentell überprüft, ob die SVM unterschiedliche Übereinstimmungswerte auf Schulklassenebene aufweist. Dabei konnte eine Schwankung zwischen Klassen festgestellt werden. Jedoch konnte gezeigt werden, dass die SVM weder von dem Geschlecht noch von der schulischen Leistung der Lernenden beeinflusst wird. Es konnte eine gewisse Objektivität gezeigt und eine Verzerrung bezüglich dieser Attribute ausgeschlossen werden. Daher kann die dritte Forschungsfrage der vorliegenden Arbeit beantwortet werden: *Die automatische Auswertung der Concept Maps durch ein Machine-Learning-Modell zeigt keine Verzerrung (Bias) in Bezug auf das Geschlecht der Lernenden sowie*

schulische Leistungen in den Fächern Mathematik, Physik und Deutsch.
Durch die Analyse konnte nicht abschließend geklärt werden, warum diese Schwankungen auf Klassenebene aufgetreten sind. Mögliche Ursachen könnten die Verwendung von verschiedenen Lehrmethoden sein, was die Merkmale und Strukturen der Propositionen beeinflusst, die das Machine-Learning-Modell klassifiziert. Zudem könnten die Interaktionen zwischen Lernenden einer Klasse die Art und Weise beeinflussen, wie sie die Concept Maps bearbeiten. Um diese Fragen abschließend zu beantworten, können weitere Forschungsarbeiten vielfältigere Forschungsmethoden anwenden, um mehr Daten von Lernenden zu erheben. Zusätzlich sollten Informationen über die Lehrkräfte erhoben werden, um Effekte bezüglich der Lehrkraft auf die automatische Auswertung untersuchen zu können. Auf technischer Ebene kann überprüft werden, ob durch die Kreuzvalidierung nicht nur die Bewertungskategorien, sondern auch die Schulklassen gleich verteilt sind. So kann ausgeschlossen werden, dass das Modell bestimmte Schulklassen nur zum Trainieren oder Testen genutzt hat.
Es bedeutet aber nicht, dass die SVM generell keine Verzerrung aufweist. Es kann jedoch nicht ausgeschlossen werden, dass weitere Quellen von Verzerrung auftreten können. Bei einer weiteren Studie können deshalb auch Daten zum sozioökonomischen Status oder elterlicher Bildung erhoben werden, um weitere Verzerrungs-Quellen auszuschließen. Außerdem sind die Concept Maps in Schulen erhoben worden, die in einer ähnlichen geografischen Region sind. Die geografische Region, aus der die Lernenden stammen, könnte kulturelle, soziale und wirtschaftliche Unterschiede aufweisen, die sich bei der Bearbeitung der Concept Map auswirken könnten. Außerdem stammen die Trainingsdaten von Schulen, die als reines Gymnasium aufgebaut sind. Die Propositionen, die zum Trainieren und Testen des Modells verwendet wurden, könnten daher nicht die gesamte Bandbreite der möglichen Propositionen und Situationen abdecken, denen das Modell in der Praxis begegnen könnte. Ein mögliches Beispiel wäre der Einsatz an einer integrierten Gesamtschule, die ebenfalls eine 11. Klassenstufe hat, aber gegebenenfalls ein anderes Leistungsniveau aufweisen könnte. Die begrenzte Datenrepräsentativität kann die Fähigkeit des Modells einschränken, korrekte Vorhersagen für neue oder unerwartete Propositionen zu treffen, was ebenfalls eine Limitation des Machine-Learning-Modells ist. Daher müssen die Ergebnisse der SVM immer unter Berücksichtigung der verfügbaren Informationen verstanden werden.

Zusammengefasst konnte mit der SVM ein Machine-Learning-Modell entwickelt werden, das eine gute Übereinstimmung mit den menschlichen Bewertern aufweist. Diese Aussage ist jedoch nur für die verwendete Concept Map zum Thema Mechanik gültig. Eine Übertragbarkeit auf andere Concept Maps kann nicht ohne Weiteres erfolgen. Dazu müssen erst weitere Forschungen betrieben werden, die unterschiedliche Concept Maps zu verschiedenen Themen analysieren. Dennoch kann vermutet werden, dass auch in anderen Domänen ähnlich gute Ergebnisse erzielt werden können. Propositionen in Concept Maps folgen oftmals einem ähnlichen logischen Aufbau, bei dem ein Begriff mit einem anderen in Beziehung gesetzt wird, z. B. *Begriff A ist ein Begriff B*. Daher könnten die Ergebnisse auch für ähnlich strukturierte Concept Maps relevant sein.

Es wurden aber nicht nur die unterschiedlichen Machine-Learning-Modelle verglichen und das beste Modell genauer analysiert. Ebenfalls wurden die Fragebögen der Lernenden, die nach der Erhebung ausgefüllt wurden, betrachtet. Die Ergebnisse der Fragebögen zeigten, dass die Software CmapTools zu viele Funktionen bietet und die Lernenden mit dieser Vielzahl überfordert waren. Die Hauptaufgabe der Lernenden bestand in der Bearbeitung der 19 Propositionen. Zusätzliche Funktionen wie das Verschieben der Verbindungslinien oder die Änderung der Schriftart sind für diesen Zweck überflüssig und führten bei den Lernenden zu Verwirrungen. Für die Feedbackstudie kann daher überlegt werden, eine eigene Lernumgebung zu nutzen, die auf die eingesetzte Concept Map konzipiert wird. Außerdem zeigten die Ergebnisse, dass manche Lernende vorgegebene Pfeile in der Concept Map bevorzugen würden. Es wurde sich bewusst dagegen entschieden, um die Leserichtung der Propositionen offenzulassen und so den Lernenden einen gewissen Freiraum zu geben. Eine Möglichkeit, die Bearbeitung der Concept Map zu erleichtern, wäre also, die Leserichtung ebenfalls vorzugeben. Das bietet den Vorteil, dass Lernenden, die bisher nicht mit Concept Maps gearbeitet hatten, die Bearbeitung leichter fallen wird. Allerdings wird dadurch die Concept Map noch weniger Freiheiten aufweisen.
Weiterhin konnte festgestellt werden, dass Teile der Lernenden mit der Bearbeitung der Concept Map überfordert waren und sich mehr Hilfestellung gewünscht hätten. Verschiedene Auswahlmöglichkeiten für die Propositionen werden in der Feedbackstudie gleichwohl nicht umgesetzt werden. Künftige Arbeiten könnten verschiedene Differenzierungsmaßnahmen entwickeln, die unter anderem auch Hilfestellungen zu einzelnen Begriffen oder Auswahlmöglichkeiten enthalten können.

7 Feedbackstudie

Es existieren zahlreiche Studien, die die Performance von Machine-Learning-Modellen beschreiben, welche in einem bildungsnahen Kontext eingesetzt werden sollen. Zhai et al. (2020) kritisierten jedoch, dass es an Arbeiten fehlt, die diese Modelle auch tatsächlich in Klassenzimmern einsetzen und die pädagogische Wirkung untersuchen. Diese Kritik wird in der vorliegenden Arbeit aufgegriffen. Das entwickelte Machine-Learning-Modell und die daraus resultierenden automatischen Auswertungen der Propositionen werden als Feedback in Schulen eingesetzt und analysiert.

Der theoretische Hintergrund in Abschnitt 2 hat gezeigt, dass Feedback ein zentrales Element des formativen Assessments ist. Obwohl die Qualität und Wirksamkeit des Feedbacks von entscheidender Bedeutung sind, konzentriert sich die Feedbackstudie nicht auf eine umfassende Bewertung der Feedbackmechanismen. Vielmehr liegt der Fokus darauf, wie die teilnehmenden Lehrkräfte und Lernenden dieses Feedback in der Praxis nutzen. Es wird angenommen, dass das Feedback einen signifikanten Einfluss auf das Lernen hat, jedoch werden keine ausführlichen Forschungsergebnisse in Bezug auf die Lernförderlichkeit des Feedbacks präsentiert. Stattdessen werden Erfahrungen und Beobachtungen aus der praktischen Anwendung dieser Methode beleuchtet. Wie die Entwicklungsstudie ist die Feedbackstudie in drei Phasen gegliedert (siehe Abbildung 7.1).

In der ersten Phase wird das eigentliche Feedback entwickelt. Dabei dient die automatische Auswertung des Machine-Learning-Modells als Basis. Damit die

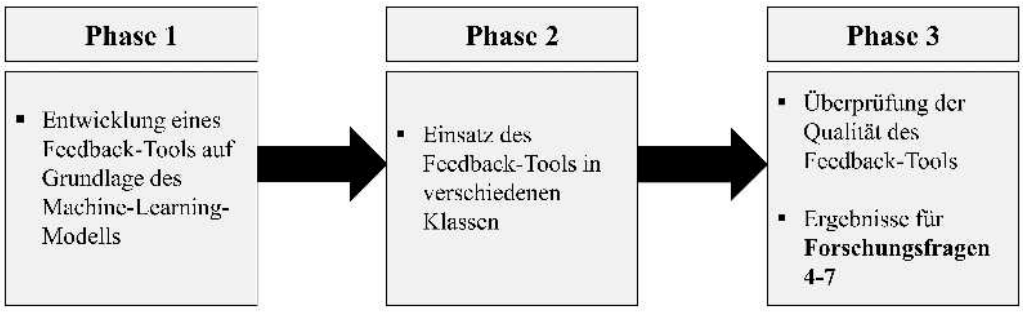


Abbildung 7.1: Ablauf der Feedbackstudie in drei Phasen

Lehrkräfte und die Lernenden nützliche Informationen aus der automatischen Auswertung erhalten können, wird auf Grundlage des theoretischen Hintergrunds (siehe Abschnitt 2.3) eine Umsetzung des Feedbacks erarbeitet.
Das daraus entstehende Feedback-Tool wird dann in der zweiten Phase in verschiedenen Klassen eingesetzt. Das Ziel dieser Phase ist es, Informationen über den Einsatz und die Nützlichkeit des Feedbacks zu erheben. Dazu werden Interviews durchgeführt und Fragebögen eingesetzt.
Abschließend werden in Phase 3 die gesammelten Daten analysiert und die restlichen Forschungsfragen beantwortet.
Bevor auf die drei Phasen näher eingegangen wird, werden die Implikationen der Schlussfolgerungen der Entwicklungsstudie beschrieben. Es hatte sich gezeigt, dass der technische Umfang der Software CMapTools teils negative Auswirkungen auf die Lernenden hatte. Angesichts dessen wird für die Feedbackstudie eine Eigenentwicklung genutzt, die speziell für die Bedürfnisse der Studie angepasst wird. Für die Feedbackstudie wird ein Webserver erstellt, auf dem eine eigene Lernplattform namens *Intelligent Physics Trainer* (IPT) integriert ist[6].

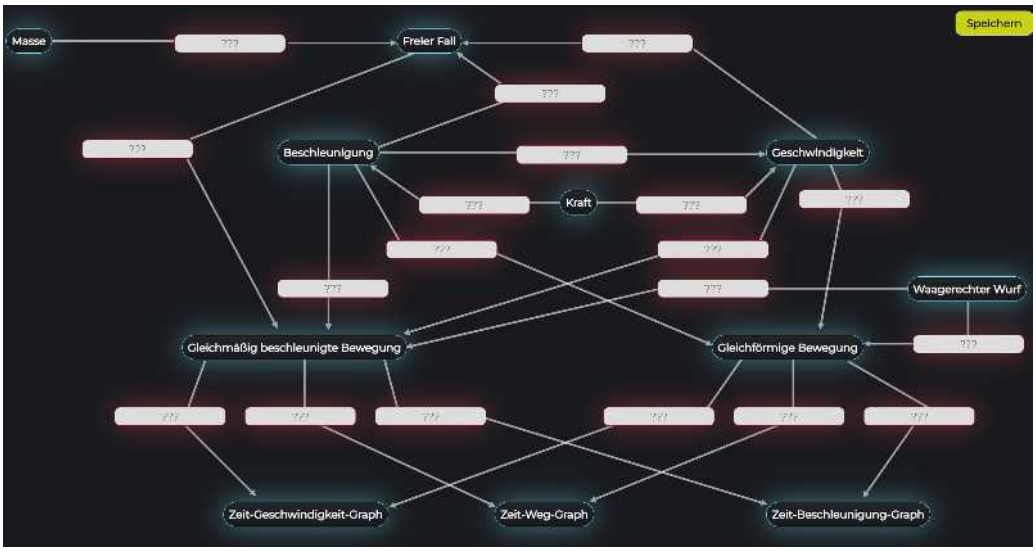


Abbildung 7.2: Neues Design der Concept Map auf der Lernplattform *Intelligent Physics Trainer* (IPT)

Die Verwendung von IPT bietet mehrere Vorteile: Die Lernenden können ihre eigenen Geräte wie iPads oder Laptops nutzen und sind nicht mehr auf die Leihgeräte angewiesen. Das bedeutet außerdem, dass durch den Webserver lediglich Anmeldedaten zur Verfügung gestellt werden müssen, sodass die Erhebung auch

[6] https://ipt.idmp.uni-hannover.de/

selbstständig von den Lehrkräften durchgeführt werden kann. Durch die erstellten Anmeldedaten können auch die bearbeiteten Concept Maps besser mit den Lernenden verknüpft werden, was ein individuelles Feedback ermöglichen kann. Die zu bearbeitende Concept Map wird außerdem nach den Verbesserungsvorschlägen der Lernenden angepasst. Es werden zur Verbindung der Begriffe statt Linien nun Pfeile genutzt. Zudem wird ein anderes Farbdesign gewählt und die Elemente lassen sich nicht mehr verschieben (siehe Abbildung 7.2). Durch diese Veränderungen sollte die Bearbeitung für die Lernenden vereinfacht werden. Zusätzlich bietet das eigene System die Möglichkeiten, auch Log-Daten aufzuzeichnen, die einen Einblick in die Bearbeitungsschritte der Lernenden liefern können. Diese Funktion war mit CMapTools in diesem Umfang nicht möglich.
Der genaue Ablauf der Feedbackstudie mit der Verwendung von IPT wird in Phase zwei genauer dargestellt. Im nächsten Abschnitt wird zunächst beschrieben, wie aus der automatischen Auswertung das eingesetzte Feedback entsteht.

7.1 Phase 1: Entwicklung des Feedbacks

Die Entwicklungsstudie hat gezeigt, dass die SVM die Proposition mit einer hinreichend guten Übereinstimmung automatisch auswerten kann. Diese automatische Auswertung soll den Lehrkräften und den Lernenden als Feedback zur Verfügung gestellt werden.
In Abschnitt 2.3 wurde dargelegt, dass Feedback ein mehrdimensionales Konstrukt ist und auf unterschiedliche Arten gebildet und an mehreren Zeitpunkten gegeben werden kann. In dieser Arbeit soll die automatische Auswertung der Concept Maps an zwei unterschiedlichen Zeitpunkten als Feedback eingesetzt werden. Das bedeutet, dass die Lernenden die Concept Map zweimal bearbeiten werden und das Machine-Learning-Modell jeweils die Concept Maps auswertet (siehe Abbildung 7.3).

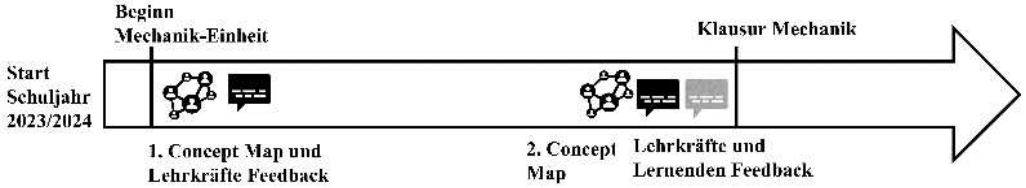


Abbildung 7.3: Erhebungszeitpunkte der Concept Map und Zeitpunkte des Feedbacks in der Feedbackstudie

Der erste Bearbeitungszeitpunkt wird zu Beginn der Unterrichtseinheit sein. Die resultierenden Informationen können als Hinweise über das Vorwissen der Lernenden verstanden werden. Da dieses Feedback vor allem für die Lehrkräfte wichtige Einblicke liefert, werden am ersten Zeitpunkt auch nur die teilnehmenden Lehrkräfte ein automatisches Feedback erhalten. Die Lehrkräfte können die Einblicke nicht nur für die weitere Unterrichtsplanung, sondern auch für individuelle Rückmeldungen nutzen. Da eine individuelle Rückmeldung durch die Lehrkräfte und nicht direkt durch das automatische Feedback-Tool erfolgt, können mögliche Fehlklassifikationen des Machine-Learning-Modells von den Lehrkräften abgefangen werden. Durch diese Eingrenzung auf die ausschließliche Rückmeldung für Lehrkräfte wird daher verhindert, dass die Lernenden ohne zusätzliche Kontrolle die automatische Auswertung erhalten. Durch Lehrkräfte in der Feedbackschleife kann z. B. die Fehlklassifikation aufgrund der *Formel*-Gruppe identifiziert werden. Zudem soll geschaut werden, zu welchem Zweck die Lehrkräfte das Feedback einsetzen und ob sie ggf. den weiteren Unterrichtsverlauf anpassen.
Am zweiten Zeitpunkt erhalten sowohl die Lehrkräfte als auch die Lernenden ein Feedback durch die automatische Auswertung der SVM. Durch den erneuten Einsatz des Machine-Learning-Modells können die Lehrkräfte erneut einen Einblick in den aktuellen Wissensstand der Lernenden bekommen. Da der zweite Zeitpunkt ein paar Wochen nach der ersten Bearbeitung liegt, kann zudem die Veränderung zwischen den beiden Concept Maps als Feedback genutzt werden. Im Optimalfall liegt der zweite Zeitpunkt in unmittelbarer Nähe zur Klausur, sodass die Lernenden die automatische Auswertung als Feedback für ihr eigenes Lernen nutzen können. Die Lehrkräfte erhalten zusätzlich Hinweise über Stärken und Schwächen, die sie vor der Klausur nochmals thematisieren können.
Damit die Stärken und Schwächen der automatischen Auswertung analysiert werden können, wird das Feedback aus verschiedenen Elementen bestehen. Es kann so geschaut werden, welche Aspekte die Lehrkräfte wofür genutzt haben. Um eine genauere Analyse durchzuführen, wird das Feedback in zwei Gruppen aufgeteilt. Die Aufteilung erfolgt nach der Differenzierung von Narciss (2006) in einfaches und elaboriertes Feedback (siehe Abschnitt 2.3). Die Studienlage hat gezeigt, dass es keinen generellen Vorteil für eine der beiden Arten gibt. Zudem kann durch diese Aufteilung untersucht werden, wie ausführlich die automatische Auswertung sein muss.
Das elaborierte Feedback basiert auf den vier Bewertungskategorien (siehe Abschnitt 6.1). Durch diese vier Kategorien können differenzierte Hinweise gegeben

werden. Für das einfache Feedback werden die vier Kategorien zu zwei Kategorien zusammengefasst. Die Bewertungskategorie A wird behalten und dient zur Aufdeckung falscher Propositionen. Die Kategorien B – D werden zu einer Kategorie *richtig* zusammengefasst. Es wird also keine stufenweise Differenzierung durchgeführt, sondern lediglich richtige Propositionen identifiziert. In beiden Fällen wird die entwickelte SVM genutzt, wobei beim einfachen Feedback die Kategorien B – D nach der Klassifizierung automatisch recodiert werden.
Die teilnehmenden Lehrkräfte werden zufällig zu einer der beiden Feedback-Gruppen zugewiesen. Damit die Lehrkräfte mit dem Feedback und den jeweiligen Bewertungskategorien arbeiten können, werden sie eine Erklärung zum Ablauf der Studie und zum eigentlichen Feedback erhalten. Die genaue Umsetzung des Feedbacks für beide Gruppen und für beide Zeitpunkte wird im Folgenden genauer erklärt und beschrieben.

7.1.1 1. Feedback – Vorwissen

In beiden Feedback-Gruppen erhalten die Lehrkräfte eine Excel-Tabelle mit unterschiedlichen Grafiken und Auswertungen. Für das erste Feedback füllen die Lernenden die Concept Map zum Beginn der Mechanik-Einheit aus, um Hinweise zum Vorwissen sammeln zu können. Zunächst wird in beiden Gruppen eine Übersicht auf Klassenebene bereitgestellt (siehe Abbildung 7.4). Die Übersicht nutzt die automatische Auswertung der Proposition und liefert den Lehrkräften einen schnellen Einblick, welche Propositionen richtig oder falsch ausgefüllt wurden. Um Unterschiede deutlich zu machen, wurde eine farbliche Abstufung gewählt. Da die Übersichtsgrafik möglichst einfach zu verstehen sein soll, wurde die Auswertung auf die einfache Auswertung beschränkt. Zudem wurden alle Propositionen, die von den Lernenden freigelassen wurden, als falsch gewertet. Daraus lassen sich erste Hinweise gewinnen, welche Zusammenhänge den Lernenden bekannt waren. Als zusätzliche Information erhalten die Lehrkräfte in beiden Feedback-Gruppen die Anzahl der teilnehmenden Lernenden und die durchschnittliche Anzahl an ausgefüllten Propositionen. Außerdem werden noch die Propositionen, die am wenigsten ausgefüllt und die am häufigsten falsch beantwortet wurden, aufgelistet. Diese Hinweise können den Lehrkräften helfen, die farblichen Abstufungen der Übersichtsgrafik besser einordnen zu können.
Die elaborierte Feedback-Gruppe erhält zusätzlich eine Analyse der 19 Propositionen. Die Analyse besteht aus einer Auflistung der 19 Propositionen und den

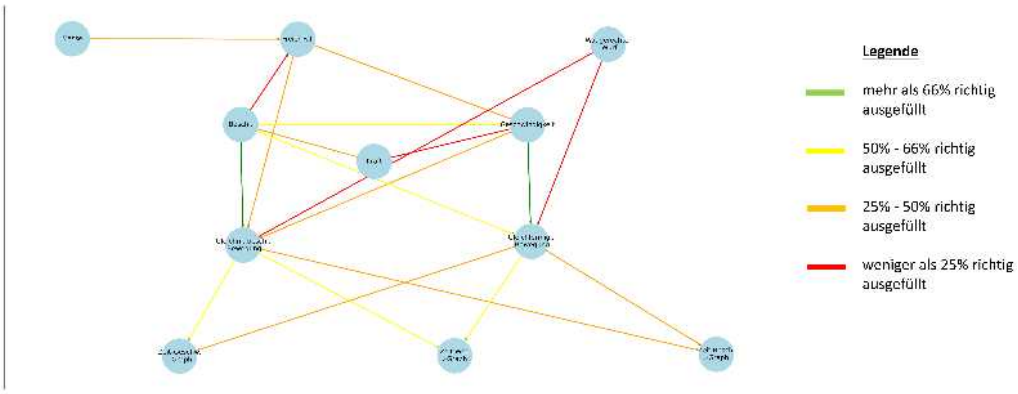


Abbildung 7.4: Übersicht der Concept Map auf Klassenebene (Auszug aus dem Lehrkräfte-Feedback)

Häufigkeiten der vier Bewertungskategorien. Um zwischen nicht bearbeiteten Propositionen und Kategorie A (falsche Propositionen) unterscheiden zu können, wird die Kategorie *leer* als fünfte Kategorie genutzt. Neben der tabellarischen Darstellung erhalten die Lehrkräfte ein Balkendiagramm zur Darstellung der Häufigkeiten (siehe Abbildung 7.5) sowie ein Spinnennetz-Diagramm zur Visualisierung der Verteilung der fünf Kategorien (siehe Anhang C).

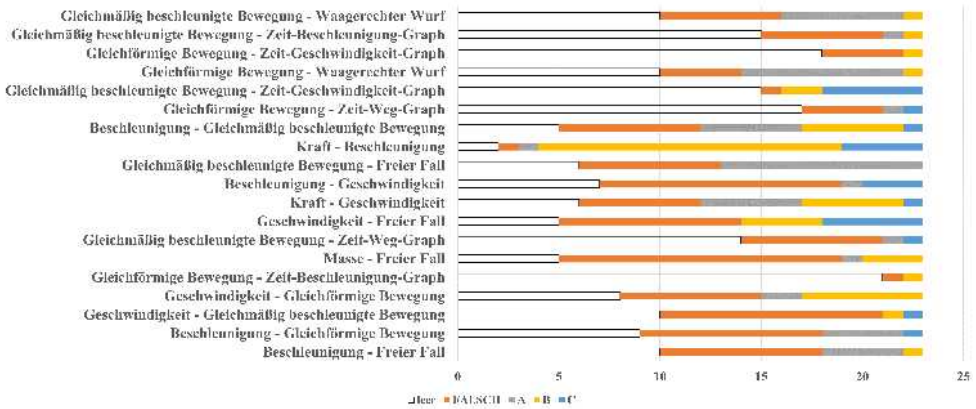


Abbildung 7.5: Analyse der 19 Propositionen bezüglich der vier Bewertungskategorien A, B, C und D sowie nicht bearbeitete Propositionen (Auszug aus dem Lehrkräfte-Feedback, elaborierte Feedback-Gruppe)

Als dritten Feedback-Aspekt werden beide Feedback-Gruppen eine Analyse auf Lernenden-Ebene bekommen. Die elaborierte Feedback-Gruppe erhält eine tabellarische Auflistung der teilnehmenden Lernenden und der Häufigkeiten der fünf Bewertungskategorien sowie eine Visualisierung dieser Daten in Form eines Balkendiagramms. Dadurch können die Lehrkräfte Einblicke in die Concept Maps

der Lernenden gewinnen und schauen, ob Auffälligkeiten bei einzelnen Lernenden existieren. Analog dazu erhalten die Lehrkräfte der einfachen Feedback-Gruppe eine tabellarische Übersicht der teilnehmenden Lernenden sowie die Anzahl an richtigen und falschen Propositionen. Zur Visualisierung wird kein Balkendiagramm wie in Abbildung 7.5, sondern ein Boxplot für die Anzahl an richtigen Propositionen genutzt (siehe Abbildung 7.6).
Zusätzlich wird die Häufigkeit der richtigen Proposition noch in einem Histogramm dargestellt, um eine einfache Übersicht zu ermöglichen.
Damit die Lehrkräfte auf besonders auffällige Propositionen hingewiesen werden, werden als vierter Schritt sogenannte „problematische Propositionen“ besonders markiert. Zu dieser Kategorie zählen Propositionen, die von mehr als 66 % der teilnehmenden Lernenden falsch oder gar nicht bearbeitet wurden. Diese Propositionen werden in einer eigenen Excel-Seite mit den dazugehörigen automatischen Bewertungen aufgelistet. Dies ermöglicht den Lehrkräften eine einfache und schnelle Übersicht über nicht bekannte oder falsche Zusammenhänge in den Concept Maps.

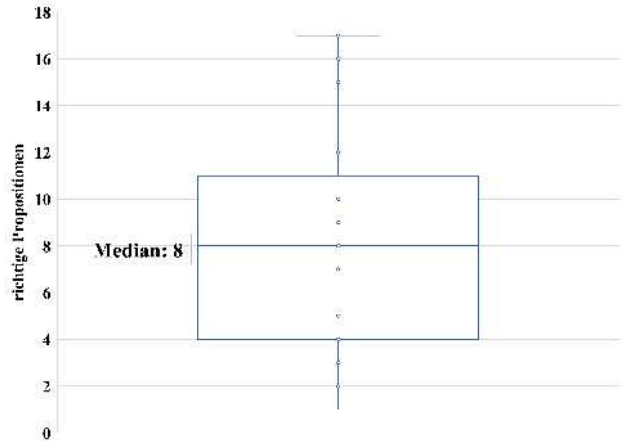


Abbildung 7.6: Boxplot bezüglich der richtigen Proposition auf Klassenebene (Auszug aus dem Lehrkräfte-Feedback, einfache Feedback-Gruppe)

Abschließend werden den Lehrkräften alle Propositionen ihrer Lernenden mit der dazugehörigen Bewertung zur Verfügung gestellt. So können die Lehrkräfte alle Propositionen eigenständig analysieren oder nach Propositionen bestimmter Lernenden suchen.

Elaboriertes Feedback	Einfaches Feedback
Übersichtsgraf (richtig/falsch)	Übersichtsgraf (richtig/falsch)
Information zur Klasse	Information zur Klasse
Analyse Propositionen (Bewertungsschema)	Analyse Propositionen (richtig/falsch)
Analyse Lernende (Bewertungsschema)	-
Problematische Propositionen (mehr als 66 % falsch oder gar nicht beantwortet)	Problematische Propositionen (mehr als 66 % falsch oder gar nicht beantwortet)
Alle Antworten (Bewertungsschema)	Alle Antworten (richtig / falsch)

Tabelle 7.1: Übersicht des Feedbacks für die Lehrkräfte

7.1.2 2. Feedback – vor der Klausur

Für das zweite Feedback werden die Lernenden dieselbe Concept Map erneut bearbeiten, ohne die erste Concept Map noch einmal zu sehen. Das Ziel des zweiten Feedbacks ist es, lernförderliche Rückmeldungen vor der Klausur zu generieren, sodass die Lernenden ihre Stärken und Schwächen reflektieren und die Lehrkräfte bestimmte Zusammenhänge wiederholen können. Daher werden dieses Mal sowohl die Lehrkräfte als auch die Lernenden ein Feedback durch die automatische Auswertung erhalten.
Das Feedback für die Lehrkräfte bleibt, bis auf eine kleine Erweiterung, identisch. Sie erhalten erneut eine Excel-Tabelle mit den in Abschnitt 7.1.1 beschriebenen Feedback-Aspekten. Bei der Analyse der Lernenden werden zusätzlich Lernende markiert, die mehr als 66 % der Proposition gar nicht oder falsch beantwortet hatten. Diese Hervorhebung erhalten beide Feedback-Gruppen. Dadurch sollen den Lehrkräften Hinweise auf schwächere Lernende gegeben werden. Das ermöglicht, nicht nur auf die Zusammenhänge der Concept Map einzugehen, sondern auch gezielter individuelleres Feedback geben zu können.
In Tabelle 7.1 sind die ausgearbeiteten Feedback-Elemente für beide Feedback-Gruppen noch einmal zusammengefasst aufgelistet.

Lernenden-Feedback

Das Feedback für die Lernenden wird wie die Bearbeitung der Concept Map auf der Lernplattform IPT zur Verfügung gestellt. Dazu erhalten die Lernenden eine Abbildung der Concept Map, bei der die Verbindungspfeile farblich markiert

sind. Die farbliche Markierung wird so gewählt, dass sowohl für Lernende mit Farbsehschwäche als auch für normale Sehende die Farben gut erkennbar sind. Zusätzlich werden die Lernenden ihre Propositionen in Tabellenform dargestellt bekommen. Da auch bei den Lernenden zwischen den beiden Feedback-Gruppen differenziert wird, ist die exakte Umsetzung des Feedbacks abhängig von der Feedback-Gruppe.
Beide Feedback-Gruppen erhalten Feedback zur ersten und zweiten Concept Map. So können Veränderungen zwischen den beiden Zeitpunkten dargestellt werden. Das Feedback zur ersten Concept Map wird für beide Gruppen identisch sein. Es wird aus einem einfachen Feedback bestehen, da die Concept Map das Vorwissen der Lernenden abbildet und davon ausgegangen werden kann, dass viele Propositionen nicht bearbeitet werden konnten. Die Darstellung der ersten Concept Map und die Präsentation des einfachen Feedbacks dient deshalb eher zur Visualisierung der Veränderung zum zweiten Zeitpunkt (siehe Abbildung 7.7). Beim Feedback für die zweite Concept Map erhalten die Lernenden der elaborierten Feedback-Gruppe ein abgestuftes Feedback auf Grundlage der vier Bewertungskategorien. In Abbildung 7.7 ist ein Ausschnitt des Lernenden-Feedbacks dargestellt. Um den Lernenden eine lernförderliche Rückmeldung zu ermöglichen, werden die vier Bewertungskategorien ausformuliert. So können die Lernenden individuelle Hinweise bezüglich ihrer Propositionen erhalten.
Für einen motivierenden Effekt werden zuerst die Propositionen präsentiert, die von der SVM in die Kategorie D eingeordnet wurden (grüne Linien in Abbildung 7.7). Diese Propositionen werden mit der Aussage „*Ausgezeichnete Arbeit! Diese Antworten sind bereits sehr detailliert oder enthalten einen richtigen funktionalen Zusammenhang!*“ dargestellt. Als Nächstes werden alle Propositionen der Kategorie C (magentafarbene Linien in Abbildung 7.7) sowie eine dazugehörige Proposition der Kategorie D präsentiert. Durch die Feedback-Aussage „*Diese Antworten sind richtig, jedoch noch zu oberflächlich. Versuche beim nächsten Mal genauer auf den Zusammenhang zwischen den beiden Begriffen einzugehen*“ und der beispielhaften Proposition aus Kategorie D können die Lernenden erkennen, warum ihre Proposition so bewertet wurde und wie eine bessere Proposition hätte aussehen können. Die orangen Linien markieren alle Propositionen der Kategorie B. Auch hier werden die Propositionen dieser Kategorie in Tabellenform dargestellt sowie eine dazugehörige Proposition der Kategorie D als Feedback bereitgestellt. Die dazugehörige Feedback-Aussage lautet: „*Diese Antworten sind richtig, sie könnten allerdings noch präziser ausgedrückt werden. Du könntest*

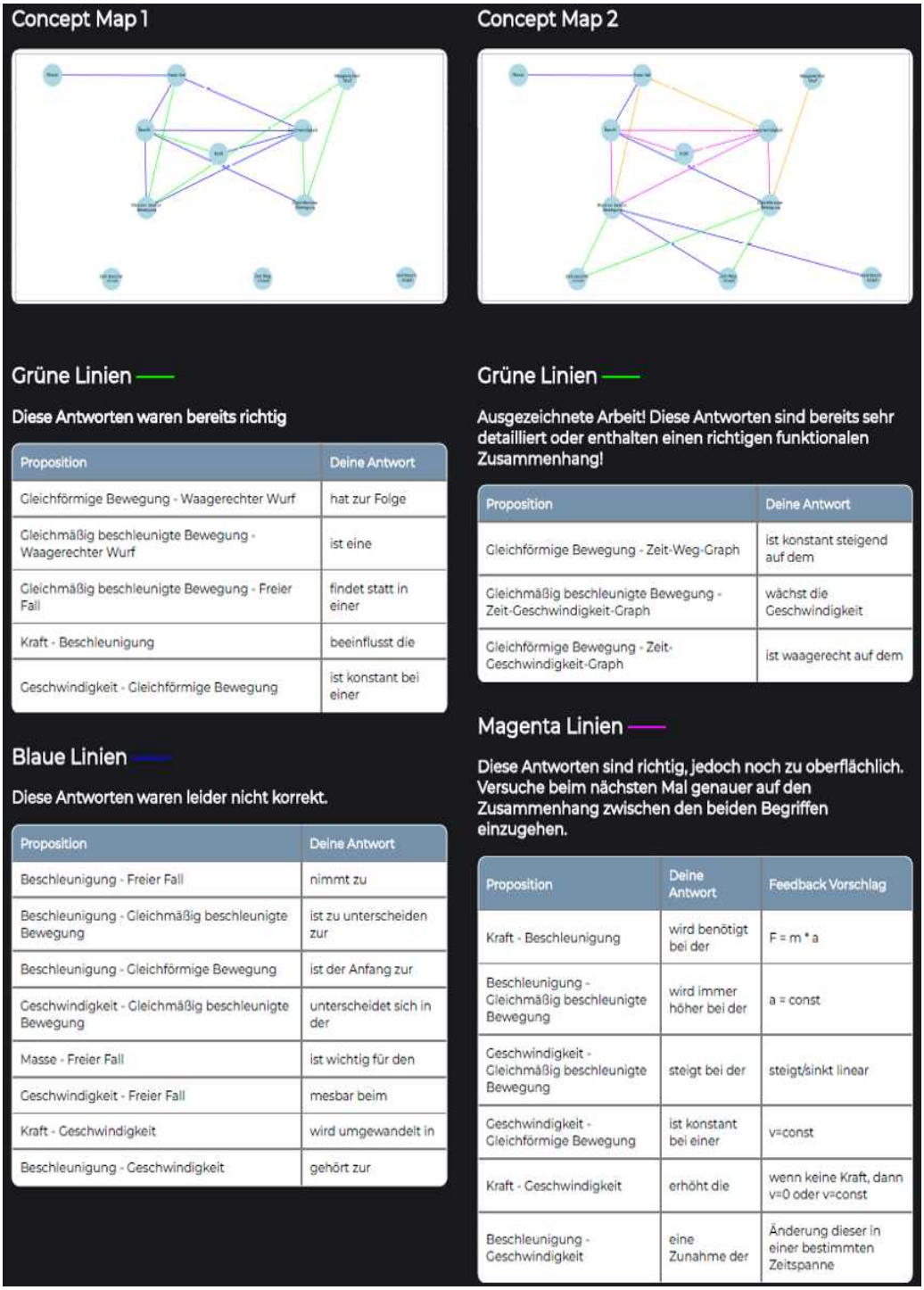


Abbildung 7.7: Auszug aus dem Lernenden-Feedback der elaborierten Feedback-Gruppe

mehr Details hinzufügen, z. B. wie genau sich etwas verhält, oder einen funktionalen Zusammenhang finden, um eine umfassendere Antwort zu liefern. Du bist auf dem richtigen Weg, weiter so!“. Durch die positive Formulierung der Feedback-Aussage sollen die Lernenden motiviert werden. Zudem verdeutlicht die Aussage und der Feedback-Vorschlag (Kategorie D Propositionen), welche Schritte die Lernenden in Zukunft machen können, um eine bessere Lösung zu finden. Als letztes Feedback werden die Propositionen aufgelistet, die die SVM als falsch (Kategorie A) bewertet hat (blaue Linien). Es wird zusätzlich der Hinweis gegeben, dass auf der nächsten Feedback-Seite eine Musterlösung präsentiert wird, die zum Vergleich der falschen Propositionen genutzt werden soll.
Die einfache Feedback-Gruppe erhält lediglich ein Feedback ohne weitere Informationen. Das bedeutet, dass für die erste und zweite Concept Map das Feedback identisch aussieht. Es werden beide Concept Maps mit einer binären farblichen Markierung als Übersicht gezeigt (linke Concept Map in Abbildung 7.7). Die richtigen Propositionen (grüne Linien) werden in Tabellenform mit der Aussage „Diese Antworten sind richtig“ präsentiert. Da es sich um ein einfaches Feedback handelt, wird auf weitere Hinweise verzichtet. Die falschen Propositionen werden analog zur elaborierten Feedback-Gruppe als blaue Linien mit dem Hinweis einer Musterlösung auf der nächsten Seite aufgelistet. Das vollständige Lernenden-Feedback ist für beide Gruppen in Anhang D dargestellt.

Die zweite Feedback-Seite ist für beide Gruppen fast identisch. Es wird eine Concept Map als Musterlösung sowie die dazugehörigen Propositionen in Tabellenform präsentiert (siehe Anhang D). Durch die Präsentation der richtigen Lösung erst im zweiten Schritt kann eine tiefergehende Verarbeitung des Feedbacks erreicht werden (siehe Abschnitt 2.3.1). Neben der Musterlösung wird die Übersichts-Grafik der zweiten Lernenden Concept Map dargestellt. Je nach Feedback-Gruppe ist die farbliche Markierung binär oder mehrfarbig. Zudem werden erneut alle Lernenden-Propositionen aufgelistet. Dieses Mal wird in beiden Gruppen nur zwischen richtig (Bewertungskategorie B-D) und falsch (Kategorie A) unterschieden.
Die zweite Feedback-Seite hat nicht nur das Ziel, eine Musterlösung für das weitere Lernen bereitzustellen. Es sollen durch den Vergleich mit der Musterlösung auch Fehler bei der automatischen Auswertung abgefangen werden. Da die SVM kein perfektes Klassifikationsmodell ist, werden fehlerhafte klassifizierte Propositionen auftreten. Die Lernenden können diese fehlerhaften Klassifikationen auf

der zweiten Feedback-Seite markieren (siehe Anhang D). Durch diesen Schritt kann nachvollzogen werden, bei welchen Propositionen die SVM Fehler macht. Es kann außerdem überprüft werden, ob und welche Fehler die Lernenden markieren.

Zusammenfassend bieten sowohl das einfache als auch das elaborierte automatisierte Feedback durch die SVM einen vielversprechenden Ansatz zur Unterstützung der Lernenden und Lehrkräfte. Das einfache Feedback zeichnet sich durch die direkte Kennzeichnung von richtigen und falschen Propositionen aus, wobei hier die Rückmeldung primär auf der Ebene der Aufgabenlösung erfolgt (siehe Kapitel 2). Das elaborierte Feedback liefert zusätzliche Informationen, weshalb hier nicht nur Feedback auf der Ebene der Aufgabenlösung, sondern auch auf der Prozessebene automatisch generiert wird. Die Entwicklung beider Feedback-Arten bietet somit eine differenzierte Herangehensweise, die es ermöglicht, das automatische Feedback und dessen Implementierung im Schulalltag untersuchen zu können.

7.2 Phase 2: Beschreibung der Stichprobe und Methodik

Die Feedbackstudie wurde analog zu der Entwicklungsstudie in mehreren 11. Gymnasialklassen durchgeführt. Bei dieser Studie wurde allerdings die Lernplattform IPT für die Erhebung der Concept Maps genutzt. Dazu bekamen die teilnehmenden Lehrkräfte eine Liste mit anonymen Benutzernamen und Passwörtern. Durch die Anonymisierung soll der Datenschutz gewährleistet werden, da eine nachträgliche Zuordnung nicht mehr möglich ist. Jedoch sollten die Lehrkräfte die Benutzernamen eigenständig an die Lernenden verteilen, sodass die Lehrkräfte wissen, welcher Lernende welchen anonymen Benutzernamen hat. So können die Lehrkräfte die automatische Auswertung auch für individuelles Feedback nutzen.
Durch die Verwendung der Lernplattform konnten die Lehrkräfte die Erhebung eigenständig durchführen. Um zu gewährleisten, dass die Erhebung in jeder Klasse gleich abläuft, wurde die Concept-Map-Einführung ebenfalls auf die Lernplattform übertragen. Außerdem wurden auf der Lernplattform der Zweck und die Inhalte der Studie sowie die genaue Aufgabe erklärt (siehe Abbildung 7.8).

Abbildung 7.8: Erklärung der Aufgabe für die Lernenden auf der Lernplattform *Intelligent Physics Trainer* (IPT)

1. Erhebungszeitpunkt

Wie in Abbildung 7.3 dargestellt, wurde die Feedbackstudie an zwei Zeitpunkten durchgeführt. Der erste Zeitpunkt war direkt zu Beginn der Unterrichtseinheit Mechanik. Die Lernenden mussten die Concept Map also nur mithilfe ihres Vorwissens beantworten.
Für die Bearbeitung der Concept Map bekamen die Lernenden 30 Minuten Zeit. Nach dieser Bearbeitungszeit sollten die Lernenden noch einen kurzen Fragebogen ausfüllen. Der Fragebogen war ähnlich aufgebaut wie der aus der Entwicklungsstudie. Zunächst sollten die Lernenden ihr Alter, Geschlecht und die Noten in den Fächern Mathematik, Physik und Deutsch angeben. Danach folgten mehrere Fragen zum allgemeinen Thema und zu der bearbeiteten Concept Map. Die Fragen sind teilweise aus der Arbeit von Stracke (2004) übernommen und adaptiert. Der eingesetzte Fragebogen bestand neben dem soziodemografischen Teil aus sieben Fragen mit einer 4-stufigen Likert-Skala und vier offenen Fragen. Der vollständige Fragebogen befindet sich im Anhang A.
Insgesamt nahmen 279 Lernende aus 16 Klassen an dem ersten Erhebungszeitpunkt teil (siehe Tabelle 7.2). Die 16 Klassen wurden von zwölf Lehrkräften unterrichtet. Die einfachen und elaborierten Feedback-Gruppen konnten so auf jeweils acht Klassen gleich verteilt werden. Manche Lehrkräfte unterrichteten meh-

Schule	Klasse	Lehrkraft	Feedback-Gruppe	Lernende
S1	S1_1	L1	elaboriert	16
	S1_2	L1	einfach	20
	S1_3	L2	elaboriert	20
S2	S2_1	L3	elaboriert	22
	S2_2	L3	einfach	16
	S2_3	L3	elaboriert	19
	S2_4	L4	einfach	14
	S2_5	L5	einfach	23
S3	S3_1	L6	einfach	14
S4	S4_1	L7	einfach	20
	S4_2	L7	elaboriert	22
S5	S5_1	L8	einfach	15
S6	S6_1	L9	elaboriert	6
S7	S7_1	L10	elaboriert	15
S8	S8_1	L11	elaboriert	15
S9	S9_1	L12	einfach	22
Summe	16	12	8 - 8	279

Tabelle 7.2: Stichprobe der ersten Concept Map in der Feedbackstudie

rere Klassen gleichzeitig, weswegen sie zu beiden Feedback-Gruppen zugeordnet wurden (z. B. Lehrkraft L1 in Tabelle 7.2). Durch diese doppelte Zuordnung kann später auch ein Vergleich beider Gruppen durchgeführt werden.
Der Altersdurchschnitt der Lernenden lag bei $16,07 \pm 0,73$ Jahren und die Verteilung der Geschlechter war wie folgt: weiblich 51 %, männlich 41 %, divers 2 % und keine Angabe 6 %. Die Mittelwerte der Noten nach der 10. Klasse lagen in Mathematik bei 2,5 und in Physik und Deutsch bei 2,6. Es konnte durch den Fragebogen außerdem festgestellt werden, dass 72 % der Lernenden das erste Mal mit einer Concept Map gearbeitet hatten.

2. Erhebungszeitpunkt

Der zweite Erhebungszeitpunkt sollte kurz vor Klausur liegen, damit das automatische Feedback noch für die Vorbereitung genutzt werden konnte. Die Feedbackstudie war so angelegt, dass dieselben Klassen die Concept Map erneut ausfüllen. Durch diverse Problematiken wie Unterrichtsausfälle oder Krankheit konnte die zweite Concept Map nicht mit jeder Klasse bearbeitet werden. Von den ursprünglichen 16 Klassen und 12 Lehrkräften haben nur noch elf Klassen

Schule	Klasse	Lehrkraft	Feedback-Gruppe	Lernende
S1	S1_1	L1	elaboriert	13
	S1_2	L1	einfach	15
S2	S2_5	L5	einfach	21
S3	S3_1	L6	einfach	11
S4	S4_1	L7	einfach	17
	S4_2	L7	elaboriert	8
S5	S5_1	L8	einfach	19
S6	S6_1	L9	elaboriert	5
S7	S7_1	L10	elaboriert	13
S8	S8_1	L11	elaboriert	7
S9	S9_1	L12	einfach	18
Summe	11	9	5 - 6	147

Tabelle 7.3: Stichprobe der zweiten Concept Map in der Feedbackstudie

die zweite Concept Map bearbeiten können (siehe Tabelle 7.3). Das bedeutet, dass nur noch neun Lehrkräfte ein automatisches Feedback erhalten haben, wobei fünf Lehrkräfte in der elaborierten Feedback-Gruppe und sechs in der einfachen Feedback-Gruppe waren.
Insgesamt haben 147 Lernende eine zweite Concept Map ausgefüllt. Von den 147 Lernenden haben 124 sowohl die erste als auch die zweite Concept Map bearbeitet. Nach der Bearbeitung der zweiten Concept Map sollten die Lernenden erneut einen kurzen Fragebogen ausfüllen. Dieser Fragebogen war ähnlich angelegt wie beim ersten Zeitpunkt und bestand aus fünf Fragen mit einer 4-stufigen Likert-Skala und einer offenen Frage. Inhaltlich bezogen sich die Fragen auf die Erstellung der zweiten Concept Map und auf einen Vergleich zwischen den beiden Erhebungszeitpunkten (siehe Anhang A).

Feedback-Phasen

Im vorherigen Abschnitt wurden die beiden Feedback-Gruppen beschrieben. Die Lehrkräfte erhielten unmittelbar nach der Bearbeitung der beiden Concept Maps das automatische Feedback und die Lernenden bekamen nach der zweiten Concept Map das automatische Feedback auf der Lernplattform IPT. Um die Forschungsfragen beantworten und das automatische Feedback analysieren zu können, wurden mit acht teilnehmenden Lehrkräften leitfadengestützte Interviews geführt (siehe Tabelle 7.4). Zwei der interviewten Lehrkräfte bekamen beide Feedback-Arten (L1 und L7), vier Lehrkräfte waren in der einfachen Feedback-Gruppe und die zwei

restlichen Lehrkräfte waren in der elaborierten Feedback-Gruppe. Die Fächerkombination war bei den meisten Lehrkräften Mathematik und Physik, wobei eine Lehrkraft neben Physik Englisch und eine weitere Lehrkraft Chemie unterrichten. Die leitfadengestützten Interviews sind im Rahmen einer Masterarbeit entstanden (Pohl, n. V.). Der Leitfaden bestand aus vier inhaltlichen Blöcken zu den Themen 1) allgemeine Einstellung zu KI im Bildungswesen, 2) formatives Assessment, 3) Nutzen des Feedbacks für die Lehrkraft und 4) mögliche Problemfelder der automatischen Auswertung. Zusätzlich gab es noch einen Block über weitere Einsatzmöglichkeiten von KI im Physikunterricht, der als Ergänzung angesehen werden kann.

Der erste Block zur allgemeinen Einstellung zu KI im Bildungswesen diente als Einstieg in das Interview. Es sollte abgefragt werden, welche Gedanken die Lehrkraft zur Nutzung von KI im Bildungswesen und ob sie bereits Erfahrung mit solchen Tools hat. Die Antworten können als Einordnung des weiteren Interviewverlaufs angesehen werden und helfen, eine Grundlage für die spätere Auswertung zu schaffen.

Die Fragen aus dem formativen Assessment-Block wurden aus dem theoretischen Kapitel 2 abgeleitet. Es sollten Erfahrungen mit dem Umgang von formativen Assessment-Methoden erfragt und auf aus der Literatur bekannte Probleme wie den Zeitmangel eingegangen werden. Die Antworten zu diesem Themengebiet liefern eine Einschätzung zum allgemeinen Umgang der Lehrkraft mit formativen Assessment-Methoden, die als Vergleich mit dem automatischen Feedback nützlich ist.

Der dritte und vierte Block zur eigentlichen Nutzung des automatischen Feedbacks und dessen Problemen basieren auf der Studie von Wulff et al. (2021). In dieser Studie wurde ein Akzeptanzfragebogen zur computerbasierten Rückmeldung entwickelt und eingesetzt, der aus den Bereichen Wirksamkeit, Nützlichkeit, Persönlichkeit und subjektive Richtigkeit bestand. Aus diesem Akzeptanzfragebogen wurden Fragen für das Interview adaptiert, sodass unter anderem erhoben werden konnte, welche Aspekte des automatischen Feedbacks besonders nützlich waren oder ob das automatische Feedback genug auf die Lerngruppe einging. Auch die möglichen Fehlklassifikationen der SVM konnten in diesen Blöcken thematisiert werden, sodass eine Einschätzung zur Praxistauglichkeit erhoben werden kann.

Im abschließenden Ergänzungsblock konnte die Lehrkraft noch potenzielle weitere Einsatzmöglichkeiten für KI-basiertes Feedback angeben. Die Aussagen können z. B. für weitere Forschungsarbeiten genutzt werden. Der gesamte eingesetzte

Schule	Lehrkraft	Feedback-Gruppe	Fächerkombination
S1	L1	elaboriert & elaboriert	Mathematik & Physik
S2	L5	einfach	Mathematik & Physik
S3	L6	einfach	Mathematik & Physik
S4	L7	elaboriert & elaboriert	Mathematik & Physik
S5	L8	einfach	Mathematik & Physik
S7	L10	elaboriert	Englisch & Physik
S8	L11	elaboriert	Chemie & Physik
S9	L12	einfach	Mathematik & Physik

Tabelle 7.4: Stichprobe der Interviews

Leitfaden findet sich in Anhang F.
Nachdem die Interviews geführt wurden, wurden die Audiodateien transkribiert. Dazu wurden die einfachen Transkriptionsregeln nach Dresing und Pehl (2011) gewählt, da für die inhaltliche Analyse die sprachlichen Besonderheiten der interviewten Lehrkräfte nicht von Bedeutung sind. Um die Interviews zu analysieren, wurde eine zusammenfassende Inhaltsanalyse nach Mayring (2022) durchgeführt. Die Auswertungseinheit sind in diesem Fall die acht Interviews. Daraus lässt sich ein einzelnes Interview als Kontexteinheit definieren und einzelne Aussagen der Lehrkräfte als Kodiereinheit auffassen (Mayring, 2022).
Da die Interviews und die zusammenfassende Inhaltsanalyse eher einem explorativen Ansatz entsprechen, wurde das Kategoriensystem induktiv aus den Interviews erstellt und mit dem Interviewleitfaden angepasst. Durch die induktive Kategorienbildung können die Erfahrungen und Beurteilungen der Lehrkräfte gut erfasst werden. Das fertige Kategoriensystem, das zur Auswertung genutzt wurde, findet sich in Anhang G.
Damit die Einschätzung der Lernenden zum automatischen Feedback auch analysiert werden konnte, wurde ein weiterer Fragebogen eingesetzt. Nachdem die Lernenden ihr automatisches Feedback auf der Lernplattform IPT erhalten haben, wurden sie automatisch zu einem Fragebogen weitergeleitet. Der Fragebogen bestand aus Fragen zur Wirksamkeit, Nützlichkeit, Persönlichkeit und Richtigkeit des automatischen Feedbacks. Die Fragen waren, wie die Blöcke drei und vier des Interviewleitfadens, aus dem Akzeptanzfragebogen zur computerbasierten Rückmeldung von Wulff et al. (2021) adaptiert. Insgesamt enthielt der Fragebogen zehn Fragen mit einer 4-stufigen Likert-Skala und drei offene Fragen (siehe Anhang A). Das automatische Feedback der Lernenden auf der Lernplattform IPT wurde von 47 Lernenden in Anspruch genommen. Dabei stammten 6 Lernende aus der

Proposition	1. Erhebungszeitpunkt		2. Erhebungszeitpunkt		Summe	
	Elaboriert	Einfach	Elaboriert	Einfach	Elaboriert	Einfach
Beschleunigung - Freier Fall	105	99	38	93	143	192
Beschleunigung - Gleichförmige Bewegung	88	78	34	79	122	157
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	79	81	35	75	114	156
Geschwindigkeit - Gleichförmige Bewegung	97	98	40	89	137	187
Gleichförmige Bewegung - Zeit-Beschleunigung-Graph	78	64	35	76	113	140
Masse - Freier Fall	127	120	42	93	169	213
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	89	73	38	84	127	157
Geschwindigkeit - Freier Fall	114	106	41	93	155	199
Kraft - Geschwindigkeit	98	105	32	81	130	186
Beschleunigung - Geschwindigkeit	114	108	36	93	150	201
Gleichmäßig beschleunigte Bewegung - Freier Fall	104	94	37	84	141	178
Kraft - Beschleunigung	102	108	34	86	136	194
Beschleunigung - Gleichmäßig beschleunigte Bewegung	94	85	36	83	130	168
Gleichförmige Bewegung - Zeit-Weg-Graph	87	71	37	81	124	152
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	96	78	37	84	133	162
Gleichförmige Bewegung - Waagerechter Wurf	86	85	33	73	119	158
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	82	66	36	79	118	145
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigung-Graph	88	70	37	76	125	146
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	70	71	29	66	99	137

Tabelle 7.5: Häufigkeitsverteilung der 19 Propositionen für beide Feedback-Gruppen und Erhebungszeitpunkte

elaborierten und 41 Lernende aus der einfachen Feedback-Gruppe. Den Feedback-Fragebogen füllten lediglich 26 Lernende aus, wobei 22 Lernende aus der einfachen und 4 aus der elaborierten Feedback-Gruppe stammten.

Bearbeitete Concept Maps

Um die Ergebnisse aus den Fragebögen und Interviews besser einordnen zu können, sind in Tabelle 7.5 die Häufigkeitsverteilung der 19 Propositionen für beide Feedback-Gruppen und Erhebungszeitpunkte dargestellt. Beim ersten Erhebungszeitpunkt lässt sich zwischen beiden Feedback-Gruppen eine kleine Differenz betrachten. Die Lernenden der elaborierten Feedback-Gruppe erstellen Propositionen mit einer durchschnittlichen Häufigkeit von 95. Die einfache Feedback-Gruppe hingegen mit einer durchschnittlichen Häufigkeit von 87. Beim zweiten Erhebungszeitpunkt ändert sich die Häufigkeitsverteilung zwischen den beiden Gruppen. Die durchschnittliche Häufigkeit der 19 Propositionen reduziert sich bei der elaborierten Gruppe auf 36, wobei die bei der einfachen Gruppe auf 83 steigt. Dies lässt sich durch die Unterschiede zwischen den Stichproben erklären, da beim zweiten Erhebungszeitpunkt nur noch 46 Lernende der elaborierten und 105 Lernende der einfachen Feedback-Gruppe an der Studie teilnahmen.

7.3 Phase 3: Ergebnisse

Die folgenden Ergebnisse werden auf Basis der Aussagen aus den Interviews und den Fragebögen präsentiert. Diese Aussagen wurden nicht verändert, weshalb etwaige Rechtschreibfehler oder umgangssprachliche Formulierungen übernommen wurden.

7.3.1 Formatives Assessment im Physikunterricht

Von den acht interviewten Lehrkräften konnten zwei (L6 und L10) die Definition des formativen Assessments erläutern. Beide Lehrkräfte stellten heraus, dass es sich um eine Ermittlung des aktuellen Wissensstandes der Lernenden handelt und die daraus gewonnenen Informationen für weitere Schritte im Lehr-Lern-Prozess genutzt werden sollen: „*Und ich versuche anhand dessen, was ich da herausfinde, Schülern konkrete Anleitungen zu geben, wo sie sozusagen weitermachen können [. . .]*“ (L10, Pos. 18). Die beiden Lehrkräfte L5 und L1 gaben an, dass sie den Begriff des formativen Assessments vorher nicht kannten. Allerdings erkannten Sie nach einer kurzen Erklärung das Konzept dahinter wieder. Ähnlich verhielt es sich mit den restlichen vier Lehrkräften, die ebenfalls keine exakte Beschreibung wiedergeben konnten, jedoch eine ungefähre Vorstellung von einem formativen Assessment hatten: „*Nein, aber ich kann mir was drunter vorstellen*“ (L8, Pos. 22).
Obwohl nicht allen Lehrkräften der Begriff des formativen Assessments auf theoretischer Ebene eindeutig bekannt war, gaben sechs Lehrkräfte an, Teile der fünf Schlüsselstrategien des formative Assessments (siehe Kapitel 2.2) im Unterricht zu nutzen. Die zweite Schlüsselstrategie „*Erfassung des Lernstands*“ konnte bei fünf Interviews wiedergefunden werden. Die Lehrkräfte L5, L12 und L6 gaben an, durch unbewertete schriftliche Lernzielkontrollen den Leistungsstand der Lernenden zu erheben, um möglichen Hürde zu identifizieren. Im Gegensatz dazu nutzten die Lehrkräfte L7 und L10 Unterrichtsgespräche oder Beobachtungen im Unterricht, um Einblicke in den aktuellen Wissensstand zu erhalten. Die Lehrkraft L10 argumentierte, dass es zu zeitaufwendig sei, von jedem Lernenden den Wissensstand zu erheben, weswegen sie eher auf Beobachtungen in Arbeitsphasen setzt. Die gesammelten Informationen nutzen drei Lehrkräfte (L5, L12, L11), um ihren weiteren Unterricht an die Bedürfnisse der Lernenden anzupassen. Die Lehrkraft L11 stellt sogar heraus, dass sie im Unterricht ständig das Geschehen

reflektiert und den Unterricht anpasst.

Interview	Aussage
L5, Pos. 31–32	Nur eben, dass das zur Verfügung stellen auch für Aufgaben für einzelne Schüler in allen Klassengruppen über [...] für alle Kinder ist natürlich vom rein zeitlichen Aufwand her schwierig.
L7, Pos. 24	Vorstellen kann ich mir das schon, aber das ist einfach nicht möglich arbeitstechnisch. Also, ich versuche immer mal wieder Feedback zu geben, aber nehmen wir mal, wenn es wirklich mal eine Unterrichtsstunde ist. Und wenn ich da jedem Feedback geben müsste bei einer Klasse von 30 Schülern, was mittlerweile gang und gäbe ist, das ist nicht möglich, also einzeln. Also das schafft man nicht.
L12, Pos. 24	Ja, auf jeden Fall. Ich wünschte, ich könnte das auch mehr und besser machen. Aber es ist für mich nicht, zumindest zum jetzigen Zeitpunkt, ich glaube auch, dass es grundsätzlich nicht möglich ist, dass man alle Schüler immer im Blick hat.
L6, Pos. 28–30	Nämlich, ich würde schon gerne. Aber wann?

Tabelle 7.6: Aussagen zum Zeitaufwand zu einer individuellen Rückmeldung für Lernende

Die Schlüsselstrategie „*lernförderliche Rückmeldung*“ lässt sich in zwei Interviews finden. Die Lehrkraft L11 nutzt eine direkte Rückmeldung, wobei nicht erkannt werden kann, ob die Rückmeldung tatsächlich lernförderlich im Sinne eines formativen Assessment ist und ob es sich um individuelle oder gruppenweise Rückmeldungen handelt. Die Lehrkraft weist jedoch darauf hin, dass eine individuelle Rückmeldung aufgrund der großen Klassengröße nicht machbar ist: „*Das schafft man eigentlich nicht wirklich gut im Unterricht.[...] eine persönliche Rückmeldung an jeden Schüler [ist] eigentlich kaum zu machen*“ (L11, Pos. 20). Eine ähnliche Aussage findet man auch bei Lehrkraft L10, die ebenfalls Rückmeldungen nutzt, um ihren Lernenden Hinweise für die Optimierung zu geben. Die Lehrkraft stellt heraus, dass solch ein Feedback für jeden einzelnen Lernenden zeitlich nicht realistisch ist: „*[...] dass man sich den sozusagen zur Seite nimmt und sagt hier, da und da könntest du noch mal nachgucken, das läuft schon gut, so*

geht es weiter. Aber das schafft man nicht mit allen, würde ich behaupten“ (L10, Pos. 20). Die Rückmeldungen werden daher eher auf einer Klassenebene und mit ausgewählten einzelnen Lernenden durchgeführt.
An dieser Stelle ist hinzuzufügen, dass auch die Lehrkräfte L5, L7, L12, L6 und L1 zwar gerne eine individuelle Rückmeldung geben möchten, jedoch den zusätzlichen Zeit- und Arbeitsaufwand als zu hoch einschätzen. Die Aussagen zu diesem Aspekt in Tabelle 7.6 zeigen deutlich, dass die Lehrkräfte es zeitlich nicht schaffen, jedem einzelnen Lernenden ein individuelles Feedback zu ermöglichen und sie diesen Aspekt bedauern. Dabei spricht Lehrkraft L6 außerdem an, dass viele Lernende nicht gelernt haben, mit einer Rückmeldung umzugehen und deshalb der positive Effekt eines formativen Assessments verpufft.
Eine Lehrkraft nutzte die gegenseitige Korrektur von Tests und die anschließende gemeinsame Besprechung der Lernziele als eine Art von Rückmeldung. Dadurch sprach die Lehrkraft die Schlüsselstrategie *Lernende als instruktionale Ressourcen für einander aktivieren* an. Die beiden Lehrkräfte L1 und L8 gaben an, kein formatives Assessment in ihrem Unterricht zu nutzen. Das von beiden genannte Argument war hier erneut der zu hohe Zeitaufwand „*Da braucht man natürlich auch Zeit [. . .]*“ (L8, Pos. 24).
Betrachtet man diese Aussagen, kann man zu dem Schluss kommen, dass die interviewten Lehrkräfte nicht genau wissen, was unter dem Begriff formatives Assessment verstanden werden kann. Sie setzen den Begriff oftmals mit einer normalen Rückmeldung gleich, wobei nicht explizit der Zweck dieser Rückmeldung genannt wird. Die Reduzierung des formativen Assessments auf eine Rückmeldung zeigt allerdings die nicht vollständige Sichtweise der Lehrkräfte zum formativen Assessment. In keinem Interview konnte daher eine vielfältige Menge an formativen Assessment-Methoden erkannt werden. Es kann angenommen werden, dass die Lehrkräfte entweder keine formativen Assessment-Methoden kennen oder sie nicht einsetzen. Denn das Hauptproblem des formativen Assessments ist nach den Interviews der enorme Zeitaufwand, speziell bei der individuellen Rückmeldung für die Lernenden. Aus dieser Problematik heraus wünschten sich drei Lehrkräfte (L5, L1, L8) mehr computergestützte Anwendungen, die sie bei einer individuellen Rückmeldung einsetzen können: „*Also wenn es jetzt beispielsweise durch KI-basierte Auswertungen erleichtert würde, auch wirklich individuelles Feedback zu generieren, dann kann ich mir das gut vorstellen, ja*“ (L5, Pos. 30).

7.3.2 Hilfreiche Elemente des automatischen Feedbacks

Den meisten Lehrkräften war die Concept Map als Methode für den Unterricht unbekannt. So gaben fünf Lehrkräfte an, noch nie mit einer Concept Map gearbeitet zu haben. Die restlichen drei Lehrkräfte nutzten Concept Maps zur Sicherung von Ergebnissen oder zur Zusammenfassung eines Themas. Es konnte jedoch erkannt werden, dass noch keiner der acht Lehrkräfte eine Concept Map für ein formatives Assessment eingesetzt hat.
Aus den Interviews ging jedoch hervor, dass die automatische Auswertung der Concept Maps generell als nützlich empfunden wurde. So haben fünf Lehrkräfte einen positiven Eindruck geschildert: „*Ja, ich würde sogar gerne mehr damit arbeiten tatsächlich. Und würde mir tatsächlich auch wünschen, dass das man das als Lehrkraft auch frei zugänglich hätte*“ (L12 einfaches Feedback, Pos. 32). Auffällig war, dass vor allem die Lehrkräfte der einfachen Feedback-Gruppe die Auswertung als zu detailliert beschrieben haben (siehe Tabelle 7.7)

Interview	**Aussage**
L5, Pos. 50	Ja, also. Okay, das ist ja schon mal interessant, dass das hier die kompakte[7] Version ist (lacht). Aber ich glaube, mehr Detailfülle hätte ich da wirklich nicht gebraucht
L12, Pos. 42	[. . .] also man muss ja auch immer das im Hinblick sehen, wie viel Zeit hat so eine Lehrkraft im Alltag, und fand es eigentlich ganz gut. Also bloß nicht zu viel, weil man liest es sich am Ende dann ja wieder nicht durch.
L8, Pos. 52	Nein. Ich fand das schon mal ausreichend, also wirklich ohne Schmus. Und ich fand das schon ausführlich

Tabelle 7.7: Aussagen zur Detailliertheit der automatischen Auswertung

Sie gaben an, dass es ihnen schwergefallen ist, jedes einzelne Diagramm und jede automatische Auswertung im Detail zu betrachten, da ihnen im Schulalltag dafür die Zeit fehlte. Dies spiegelte sich auch in der Bewertung der einzelnen Aspekte des automatischen Feedbacks wider. In den Interviews konnte erkannt werden, dass vor allem der Übersichtsgraph und die zusammenfassenden Aspekte

[7] Damit ist das einfache Feedback gemeint.

wie die *problematischen Propositionen* von den Lehrkräften genutzt wurden. Die beiden Lehrkräfte L6 und L1 betonten, dass die Tabelle mit allen Propositionen der Lernenden viel zu detailliert ist und stellten die Nützlichkeit der vorausgewählten Auswertung heraus: „*Dann gab es noch die Seite „alle Antworten" und die macht ihrem Namen alle Ehre so weit. Das scrollt man mal so durch, tatsächlich*" (L6, Pos. 44). Der Übersichtsgraph wurde vor allem aufgrund der farblichen Markierungen und der einfachen Art der Informationsgewinnung gelobt. Die *problematischen Propositionen* nutzten viele Lehrkräfte als Ergänzung, um auffällige Propositionen zu analysieren. Die Lehrkraft L10 forderte sogar noch mehr zusammenfassende Aspekte, die eine geeignete Vorauswahl der automatischen Auswertung liefert: „*[. . .] Würde es aber vielleicht noch mal cool finden, wenn drei oder vier Sätze zur Lerngruppe noch mal erscheinen, dass man vielleicht sagt okay, die Lerngruppe hat da und da Probleme oder ist da und da schon sehr gut. Ich meine, das kann man zwar auch irgendwie aus dieser Grafik erkennen, aber ich glaube, wenn es verbalisiert wäre, noch mal sozusagen als Statement auftaucht, wäre das glaube ich noch mal eine zusätzliche Info, die ich für hilfreich ansehen würde*" (L10 elaboriert, Pos. 60).

Insgesamt ist also zu erkennen, dass die Lehrkräfte die automatische Auswertung nicht auf einer individuellen Ebene, sondern eher auf einer Klassenebene angeschaut haben. Das wird nicht nur durch die Nutzung der Feedback-Aspekte deutlich, sondern es wurde ebenfalls von manchen Lehrkräften explizit betont. So beschreibt die Lehrkraft L10, die in der elaborierten Feedback-Gruppe war, den Kern des automatischen Feedbacks wie folgt: „*Und das fand ich sozusagen das Wesentliche an dem Feedback, dass ich einen schnellen Überblick darüber bekommen habe, wo die Klasse steht. Das individuelle Feedback habe ich mir zwar auch irgendwie angeschaut, aber habe es jetzt für mich nicht so stark benutzt*" (L10 elaboriert, Pos. 39).

7.3.3 Nutzung des automatischen Feedbacks

Trotz der positiven Einschätzung der acht interviewten Lehrkräfte wurde das automatische Feedback kaum im Sinne eines formativen Assessments genutzt. Drei Lehrkräfte (L6, L1 & L11) gaben an, die automatische Auswertung zwar angeschaut zu haben, allerdings keine weiteren Handlungsschritte daraus gebildet haben. Es wurde also weder der weitere Unterricht angepasst noch haben die Lehrkräfte das Feedback für eine (individuelle) Rückmeldung genutzt. Als Grund

nannten die Lehrkräfte, dass sie während des normalen Schulalltags keine zusätzliche Zeit hatten, sich intensiv mit der automatischen Auswertung zu beschäftigen und die Concept Map inhaltlich nicht gepasst hatten, weswegen die Auswertung keine richtige Relevanz hatte. Die Lehrkraft L10 nutzte die automatische Auswertung, um sich Gedanken über die Lerngruppe zu machen, jedoch nannte L10 keine expliziten Konsequenzen im Sinne eines formativen Assessments, die auf Grundlage der bereitgestellten Hinweise durchführt wurden.
Im Gegensatz dazu konnten bei den Lehrkräften L5, L7, L8 und L12 konkrete Anwendungsfälle erkannt werden. Die Lehrkraft L5 aus der einfachen Feedback-Gruppe hat die automatische Auswertung genutzt, um Wissenslücken auf Klassenebene zu finden. Mit diesen Informationen hatte L5 dann den weiteren Unterrichtsverlauf angepasst, um die Wissenslücken der Lernenden zu wiederholen und gemeinsam mit der Klasse zu schließen. Die Lehrkraft L12, die ebenfalls in der einfachen Feedback-Gruppe war, gab an, die zweite automatische Auswertung für die Vorbereitung der Lernenden auf die anstehende Klassenarbeit genutzt zu haben. Zudem nutzen L12 und auch L8 die Informationen, um die eigentliche Klassenarbeit besser an das Leistungsniveau der Lernenden anzupassen: „*[. . .] ich habe das zweite Feedback ja im Grunde genommen noch mal genutzt, um jetzt die auf die Klassenarbeit vorzubereiten und habe auch ein bisschen meine Klassenarbeit daran orientiert*“ (L12 einfaches Feedback, Pos. 30). Am deutlichsten beschrieb die Lehrkraft L7 ihre Konsequenzen. L7 unterrichtete zwei verschiedene Klassen, weswegen sie in beiden Feedback-Gruppen (einfach und elaboriert) war. Sie gab an, dass sie durch die automatische Bewertung Wissenslücken identifizieren konnte, die sie so nicht gesehen hätte: „*Also habe ich gesehen, dass da wirklich in der Lerngruppe, da gibt es einen Denkfehler, gibt es irgendwo ein Problem das zu verstehen. Genau, deswegen hat mir das wirklich sehr gut geholfen [. . .]. Und das war auf jeden Fall gut, dass ich das so gemacht habe. Und ohne die Hilfe des Feedbacks hätte ich das glaube ich nie nicht so in der Form registriert*“ (L7 elaboriertes und einfaches Feedback, Pos. 30,34). Diese Aussagen verdeutlichen gut, welches Potenzial in solch einer automatischen Auswertung stecken kann.
Durch diese zusätzlichen Informationen konnte sie zumindest mit einer Klasse eine auf die Lernenden abgestimmte Unterrichtsstunde durchführen. Die Lehrkraft L7 konnte dadurch sogar explizite Auswirkungen feststellen. Die Klasse, mit der L7 die angepasste Wiederholungsstunde vor der Klausur durchgeführt hatte, konnte im Vergleich mit der anderen Klasse bessere Ergebnisse erzielen: „*[. . .] die eine Gruppe, mit der ich es besprochen habe vorher, da haben das alle richtig gemacht*

und bei der anderen [. . .] haben sehr viele geschrieben, es ist ein exponentieller Zusammenhang, dann musste ich es bei allen falsch markieren“ (L7 elaboriertes und einfaches Feedback, Pos. 30,34). Die vier Lehrkräfte L5, L7, L8 und L12, die die automatische Auswertung für ihren weiteren Unterricht genutzt hatten sowie die Lehrkraft L10, die angab, zumindest sich Gedanken wegen der Auswertung zu machen, äußerten sogar, auch in Zukunft mit der automatischen Concept-Map-Auswertung zu arbeiten.
Aus den beschriebenen Anwendungsfällen der Lehrkräfte kann erkannt werden, dass eher das automatische Feedback zur zweiten Concept Map genutzt wurde. Fünf der interviewten Lehrkräfte beschrieben die zweite Rückmeldung auch als generell hilfreicher. Das Hauptargument der Lehrkräfte bezog sich auf die mangelnde Aussagekraft des automatischen Feedbacks zur ersten Concept Map, da die Lernenden durch das nicht vorhandene Vorwissen die Concept Map nur in Teilen oder falsch beantworteten: „*Auf jeden Fall beim zweiten Mal. Also, es war beim zweiten Mal viel hilfreicher, weil es natürlich auch viel aussagekräftiger war*“ (L5 einfaches Feedback, Pos. 46); „*Definitiv beim zweiten. Also zu Beginn, da war kaum was da. Vieles, was da also angekreuzt wurde, war wirklich jenseits von Gut und Böse.*“ (L7 elaboriertes und einfaches Feedback, Pos. 46). Lediglich die Lehrkraft L8 konnte in der ersten Concept Map einen positiven Effekt erkennen, da die Lernenden eine Steigerung des Lernzuwachses zwischen den beiden Concept Maps sehen können.

7.3.4 Subjektive Wahrnehmung des automatischen Feedbacks

Die Mehrheit der Lehrkräfte ist überzeugt, dass die automatische Auswertung mehr Fehler macht als eine menschliche Lehrkraft. Es wird angezweifelt, wie die automatische Auswertung mit Rechtschreibfehlern oder Fehlern bezüglich Groß- und Kleinschreibung umgeht. Zudem wird erwähnt, dass die Lernenden eine Schreibweise nutzen könnten, die die KI nicht kennt und es daher zu Fehlern kommen kann. Die Lehrkraft L10 ist auch der Auffassung, dass eine menschliche Lehrkraft die Antworten der Lernenden deutlich besser interpretieren und auch bei falschen Antworten richtige Ansätze erkennen und nutzen kann. L10 zweifelt an, ob eine KI diese Interpretationsfähigkeit besitzt. Die Lehrkraft L1 kritisiert auch die fehlende Transparenz der automatischen Auswertung. Es war für die Lehrkraft teilweise nicht ersichtlich, wie das Machine-Learning-Modell zu den

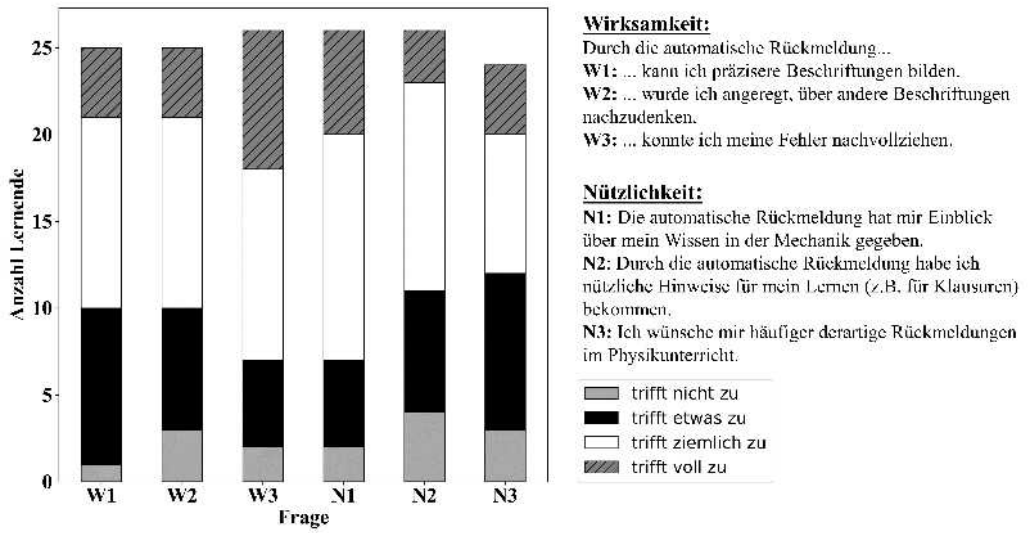


Abbildung 7.9: Ergebnisse des Fragebogens zur Wirksamkeit und Nützlichkeit des automatischen Feedbacks

Auswertungen gekommen ist. L1 stellt aber auch fest, dass durch die automatische Auswertung deutlich mehr Daten verarbeitet werden können, als das eine menschliche Lehrkraft schaffen könnte. Dieser Aspekt wird auch von L11, L12 und L8 aufgegriffen. Sie erkennen hier deutliche Vorteile gegenüber einer menschlichen Bewertung. Die Lehrkraft L8 ergänzte in diesem Zusammenhang, dass die automatische Auswertung auch nach einer großen Anzahl von Propositionen immer noch wie am Anfang bewerten würde und eine menschliche Lehrkraft müde wird und es so zu Fehlern in der Bewertung kommen kann. Eine differenzierte Meinung hatte die Lehrkraft L7. Sie ist überzeugt, dass eine KI zwar auf fachlicher Ebene deutlich besser sei, eine Lehrkraft jedoch didaktisch mehr Kompetenzen hat. Sie ist deswegen auch der Auffassung, dass es durch den Einsatz einer KI im Unterricht zu keiner Konkurrenzsituation kommt, sondern vielmehr Synergien entstehen können, wo beide Parts voneinander lernen können.

In diesem Abschnitt wird auch die subjektive Wahrnehmung der Lernenden betrachtet. Dafür wird sich primär auf den Fragebogen bezogen, der nach dem Feedback eingesetzt und von 26 Lernenden bearbeitet wurde (vgl. Abschnitt 7.2). Da für die Fragebogenabschnitte *Richtigkeit*, *Persönlichkeit* und *Abschluss* von den meisten Lernenden nicht bearbeitet wurde, bezieht sich die Auswertung nur auf die Abschnitte *Wirksamkeit* und *Nützlichkeit* (siehe Anhang A). Die Mehrheit der Lernenden ist überzeugt, dass sie durch die automatische Auswertung präzisere Beschriftungen für die Propositionen bilden können (Frage W1, siehe Abbildung 7.9). Zudem gaben 7 Lernende an, dass die Aussage „*Durch die automatische*

Rückmeldung wurde ich angeregt, über andere Beschriftungen nachzudenken“ etwas zutrifft, 11 Lernende gaben an, dass die Aussage ziemlich zutrifft, und 4 Lernende, dass die Aussage völlig zutrifft (Frage W2, siehe Abbildung 7.9). Die meisten der 26 Lernenden konnten durch die automatische Auswertung ihre Fehler in den Concept Maps nachvollziehen, was den positiven Eindruck der Lernenden verstärkt (Frage W3, siehe Abbildung 7.9). Die automatische Auswertung konnte ebenfalls dazu beitragen, Einblicke in den Wissensstand der Lernenden zu liefern (Frage N1, siehe Abbildung 7.9). Allerdings gaben vier Lernende an, dass die automatische Auswertung keine Hinweise für das weitere Lernen bereitgestellt hat, und sieben Lernende gaben an, dass diese Aussage nur etwas zutrifft (Frage N2, siehe Abbildung 7.9). Dieses Bild spiegelt sich auch bei der Frage N3 wider, bei der nach einer häufigeren Nutzung solcher Systeme gefragt wurde. Die Hälfte der Lernenden meinte, dass die Aussage „*Ich wünsche mir häufiger derartige Rückmeldungen im Physikunterricht*“ nicht oder nur etwas zutrifft (Frage N3, siehe Abbildung 7.9). Die beiden offenen Fragen in diesem Abschnitt wurden nur von einer geringen Anzahl Lernender ausgefüllt.

Bei fünf Aussagen konnte klar erkannt werden, dass die Lernenden die automatische Auswertung als nützlich empfanden (siehe Tabelle 7.8). Ein Lernender stellte heraus, dass im normalen Schulalltag keine Zeit für eine individuelle Rückmeldung bleibt und dies durch die automatische Auswertung nun möglich war. Zusätzlich wurde von einem Lernenden angemerkt, dass die automatische Auswertung geholfen hat, verschiedene Denkweisen zu erkunden und zu verstehen. Bei einer anderen Lernenden-Aussage konnte nicht genau erkannt werden, ob das Feedback als nützlich empfunden wurde. Es wurde angemerkt, dass die automatische Auswertung bei unterschiedlicher Schreibweise derselben Proposition fehlerhaft sei, jedoch die bereitgestellte Musterlösung nützlich für die Klausurvorbereitung war (siehe Tabelle 7.8). Bei den zwei weiteren Aussagen konnte festgestellt werden, dass die Lernenden das Feedback als nützlich empfanden, da Fehler bei der Rückmeldung aufgetreten sind und die automatische Auswertung keine neuen Erkenntnisse für den Lernenden geliefert hat (siehe Tabelle 7.8).

Um die Fehler der automatischen Auswertung zu kennzeichnen, konnten die Lernenden auf der zweiten Feedback-Seite fehlerhafte Zuordnungen markieren (vgl. Abschnitt 7.2). Insgesamt nutzen diese Funktion 35 Lernende, wobei fünf Lernende mindestens einen Fehler der automatischen Auswertung gefunden haben. Die Lernenden stammen alle aus der einfachen Feedback-Gruppe. Sieben Propositionen wurden von den Lernenden als fehlerhafte Auswertung gekennzeichnet.

Kategorie	Aussage
nützlich	Ja Ja, ich habe erst gedacht, dass sobald ein anderes Wort gewählt wird, es falsch ist. So ist es nicht. Es ist also wenig Aufwand für hilfreiche Rückmeldungen. Ja, weil man direkt wusste, wie es richtig gewesen wäre. Ja, sie war ein guter Weg, um selbst nach anderen Denkweisen zu sehen und nachvollziehen. Die automatische Rückmeldung finde ich nützlich, da im Unterricht oft keine Zeit ist, um individuelles Feedback zu erhalten und das hier möglich ist.
unklar	Da die richtige Aussage auf vielen verschiedenen Arten ausgedrückt werden können, werden sinngemäß richtige Antworten als schlicht falsch bewertet, was dem Schüler kein geeignetes Feedback vermittelt. Gut ist, dass eine fachlich korrekte Antwort gegeben wird, die auch für Klausuren nützlich ist.
nicht nützlich	Nein, da Fehler bei der Rückmeldung aufgetreten sind. Nicht wirklich, da sie mir nicht viel Neues gezeigt hat.

Tabelle 7.8: Aussagen der Lernenden zur offenen Frage „*Hast du die automatische Rückmeldung als nützlich empfunden?*“

Jede dieser sieben Propositionen wurde von der automatischen Auswertung in die Kategorie *falsch* zugeordnet. Das bedeutet, dass von den 35 Lernenden keine fehlerhafte automatische Auswertung einer anderen Bewertungskategorie markiert wurde. Dieses Ergebnis lässt sich auch in den Interviews identifizieren. Die Lehrkraft L1 konnte beobachten, dass die Lernenden vor allem die Proposition der Kategorie *falsch* betrachtet haben: „*Ja, aber da konnte ja irgendwie markieren, wenn irgendwas falsch war. Das haben die auch gemacht, wenn sie mal meinten, dass das eine falsche war. Sie haben es natürlich nur geguckt, glaube ich, hauptsächlich bei denen, die als falsch markiert wurden. Das heißt, ich glaube nicht, dass sie intensiv sich die richtig markierten angeguckt haben* [...]“ (L1, Pos. 56).

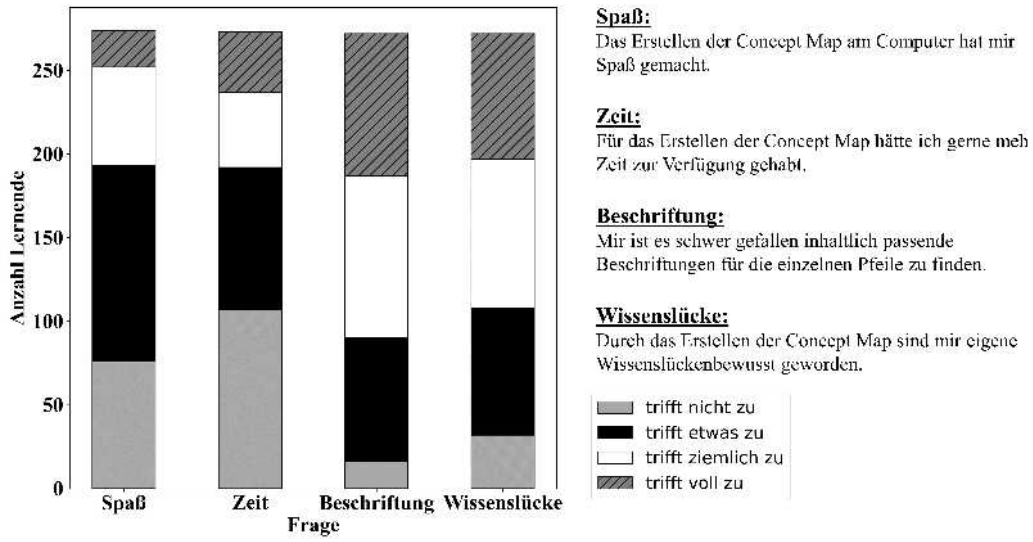


Abbildung 7.10: Ausgewählte Ergebnisse des Fragebogens nach der ersten Concept Map

7.3.5 Auswertung der Fragebögen zur eingesetzten Concept Map

Der Großteil der Lernenden aus der Feedback-Studie hat zum ersten Mal mit einer Concept Map gearbeitet. Dies ging aus den Antworten des ersten Fragebogens (siehe Anhang A) hervor, bei dem 72 % der Lernenden angaben, noch keine Erfahrung mit Concept Maps zu haben. Den Lernenden hat die erste Bearbeitung der Mechanik-Concept-Map nur teilweise Spaß gemacht: So gaben 28 % der Lernenden an, dass die Aussage *die Bearbeitung der Concept Map hat mir Spaß gemacht* nicht zutrifft, 43 % etwas zutrifft, 22 % ziemlich zutrifft und lediglich 8 % voll zutrifft (siehe Abbildung 7.10). Die Bearbeitungsdauer für die Concept Map schien für die Lernenden jedoch ausreichend zu sein. 70 % der Lernenden gaben an, dass die Aussage, noch mehr Bearbeitungszeit zu benötigen, teilweise oder gar nicht zutrifft (siehe Abbildung 7.10). Wenig überraschend ist es vielen Lernenden schwergefallen, passende Beschriftungen für die Propositionen in ihrer Concept Map zu finden. So gaben 61 % an, dass diese Aussage ziemlich oder sogar voll zutrifft (siehe Abbildung 7.10). Eine ähnliche Verteilung findet man auch bei der nächsten Aussage des Fragebogens. Hier sollten die Lernenden angeben, ob durch die Concept Map die eigenen Wissenslücken bewusst geworden sind (siehe Abbildung 7.10).
Zusätzlich zu den geschlossenen Items beantworteten die Lernenden in einer offenen Frage, ob Concept Maps öfter im Unterricht eingesetzt werden sollten. Um

die Antworten zu analysieren, wurden sie den Kategorien *öfter einsetzen* und *nicht öfter einsetzen* zugeordnet. Beim Blick in die Antworten zeigte sich ein deutliches Bild: Von den 232 abgegebenen Antworten konnten 191 Antworten der Kategorie *öfter einsetzen* und nur 41 der Kategorie *nicht öfter einsetzen* zugeteilt werden. Die Argumente für den häufigeren Einsatz waren vielfältig (siehe Tabelle 7.9). Die Lernenden stellten fest, dass durch die Concept Map die Zusammenhänge sichtbar werden und dies beim Lernen helfen kann. Die übersichtliche Darstellung der Inhalte könne nach den Lernenden auch zur Überprüfung des eigenen Wissens genutzt werden. Zudem wurde den Lernenden durch die Concept Map und die darin befindenden Propositionen ersichtlich, welche relevanten Inhalte und wichtigen Zusammenhänge für den Unterricht von Bedeutung sind. Aber nicht nur die Zusammenhänge des eigentlichen Themas wurden den Lernenden durch die Concept Map deutlich. Die Lernenden erkannten auch die Möglichkeit, verschiedene Inhalte miteinander zu verknüpfen. Die Argumente gegen einen Einsatz im Unterrichten ließen sich auf zwei Hauptargumente zurückführen, siehe auch Tabelle 7.9: Einerseits gaben Lernende an, dass die Concept Map zu anspruchsvoll war und nicht deutlich wurde, was genau gemacht werden sollte. Andererseits seien Concept Maps zu individuell, um sie im Unterricht einzusetzen. Sie bieten sich eher für das eigene Lernen und nicht für den Gebrauch im Klassenzimmer an. Zudem seien Concept Maps zu unübersichtlich, was den gemeinsamen Vergleich zusätzlich erschwert.

Öfter einsetzen ($N = 191$)	**Nicht öfter einsetzen** ($N = 41$)
Concept Maps können eine sehr überschaubare und verständliche Methode sein [...]. Im Unterricht wäre dies eine klare Methode zum besseren Verständnis des Themas.	Nein, es ist langweilig und auch zu anspruchsvoll, da Teile der Map teilweise vor mehreren Jahren im Unterricht Thema waren.
Ja, weil ich denke, dass man sich die wichtigen Fachbegriffe besser einprägen und das Thema besser nachvollziehen kann.	Nein, da es unklar ist, welche Antwort bei den leeren Feldern erwartet wird.
Ja, denn man kann sich dadurch selber testen, inwiefern man den Unterricht verstanden hat und mitgekommen ist.	Nein, da es sehr individuell und daher schwer zu vergleichen ist.

Es hilft 2 Begriffe besser miteinander zu verknüpfen und es daher hilft sich die Bedeutung der Begriffe zu merken.	Nein, weil Concept Maps bei umfangreichen Themen schnell, sehr unordentlich und unübersichtlich wirken.
Ja, da sie Zusammenhänge gut verdeutlichen können und so gut beim Lernen helfen.	Nein, Antworten können sehr unterschiedlich sein.
Ja, ich finde es ziemlich praktisch, um den Wissensstand der Schüler*innen zu prüfen und diesen anschließend zu fördern.	Eher nicht. Wenn ja, nur mit Beispielen oder vorgegebenen Wörtern, die man richtig ein-/ anordnen muss. So war es zu schwer.
Ich finde, es ist eine gute Art der Überprüfung, um zu ermessen, wie viel Wissen man aus den vorherigen Unterrichtsstunden mitgenommen hat.	Nein, weil man selbst auf die Antworten kommen muss und nicht weiß, ob es richtig ist. Der Lehrer erklärt uns nichts und wir lernen nichts Neues dazu.
Ich denke, dass der Einsatz von Concept Maps im Unterricht vorteilhaft sein könnte, vor allem vor Klassenarbeiten bzw. am Ende eines Themas, um sich die Beziehungen zwischen verschiedenen Formeln und den korrespondierenden Begriffen klarzumachen.	Nicht wirklich, da man für sich selber vieles erarbeiten muss, ohne dass einem ausführlich erklärt.
Ja, da es die eigenen Schwächen aufzeigt und man dann an diesen arbeiten kann. Am besten am Anfang des Schuljahres, vor der Arbeit und am Ende des Schuljahres, um den Lernfortschritt zu sehen.	Nein, für mich war es schwierig damit zu arbeiten, da nicht konkrete Fragen gestellt wurden und ich daher Schwierigkeiten damit hatte herauszufinden, was von einem verlangt wird.

Tabelle 7.9: Ausgewählte Antworten der Lernenden auf die offene Frage „*Bist du der Meinung, dass man Concept Maps im Unterricht öfter einsetzen sollte?*“

Nach der zweiten Concept Map der Feedbackstudie wurde ebenfalls ein Fragebogen eingesetzt (siehe Anhang A). Für die Darstellung der folgenden Ergebnisse werden nur die Lernenden betrachtet, die beide Fragebögen ausgefüllt haben

($N = 59$). Diese Teilstichprobe kann als repräsentativ für die gesamte Stichprobe angesehen werden, da die Verteilung des Geschlechts (weiblich 51 %, männlich 44 %, diverse 2 %, keine Angabe 3 %) und die Mittelwerte der Noten (Physik = 2,70, Mathematik = 2,61, Deutsch 2,74) vergleichbar sind (siehe Abschnitt 7.2).

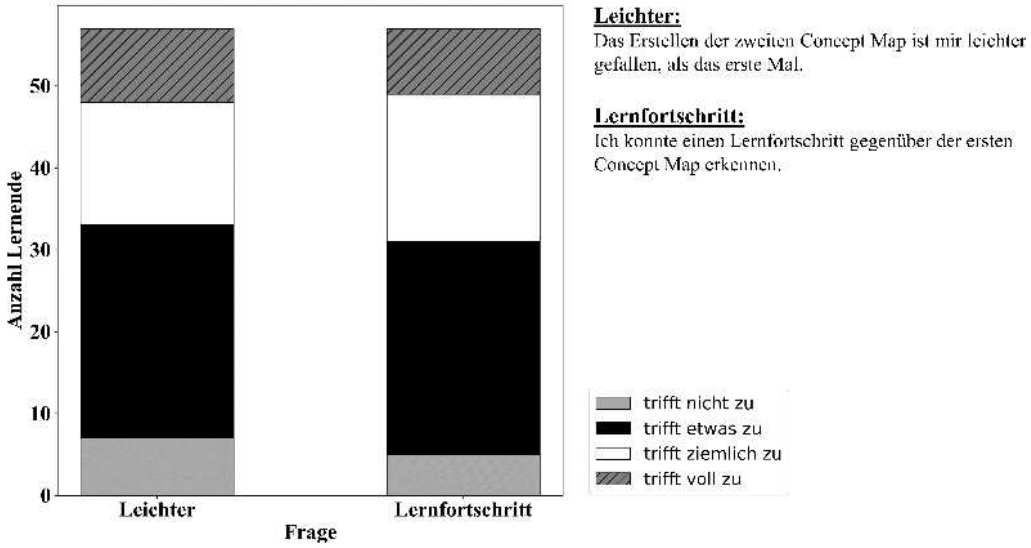


Abbildung 7.11: Ausgewählte Ergebnisse des Fragebogens nach der zweiten Concept Map

Im zweiten Fragebogen wurden die Lernenden gefragt, ob ihnen die Erstellung der zweiten Concept Map leichter gefallen ist als die der ersten und ob sie einen Lernfortschritt zwischen den beiden Zeitpunkten erkennen konnten. Abbildung 7.11 zeigt, dass über die Hälfte der Lernenden die Erstellung der zweiten Concept Map nicht oder nur etwas leichter fand. Zudem zeigt Abbildung 7.11, dass 46 % der Lernenden die Aussage bezüglich des Lernfortschritts für ziemlich oder sogar voll zutreffend halten. Allerdings waren genauso viele Lernende der Meinung, dass die Aussage nur etwas zutrifft. Die Ergebnisse zeigen also kein deutliches Bild.

Die Ergebnisse bezüglich des Spaßfaktors haben sich zwischen den beiden Erhebungszeitpunkten nicht wesentlich verändert (siehe Abbildung 7.12). Bezüglich der Schwierigkeiten beim Finden passender inhaltlicher Beschriftungen kann eine Verbesserung festgestellt werden. Es haben mehr Lernende der Aussage nur etwas zugestimmt, was darauf hindeutet, dass sie weniger Probleme hatten. Man kann jedoch erkennen, dass es immer noch einem Teil der Lernenden schwerfiel, passende Beschriftungen für ihre Propositionen zu finden (siehe Abbildung 7.12). Bei der Bearbeitungsdauer kann jedoch ein deutlicher Trend erkannt werden. Die

Ergebnisse zeigen, dass noch mehr Lernende mit der Bearbeitungszeit zufrieden waren (siehe Abbildung 7.12).

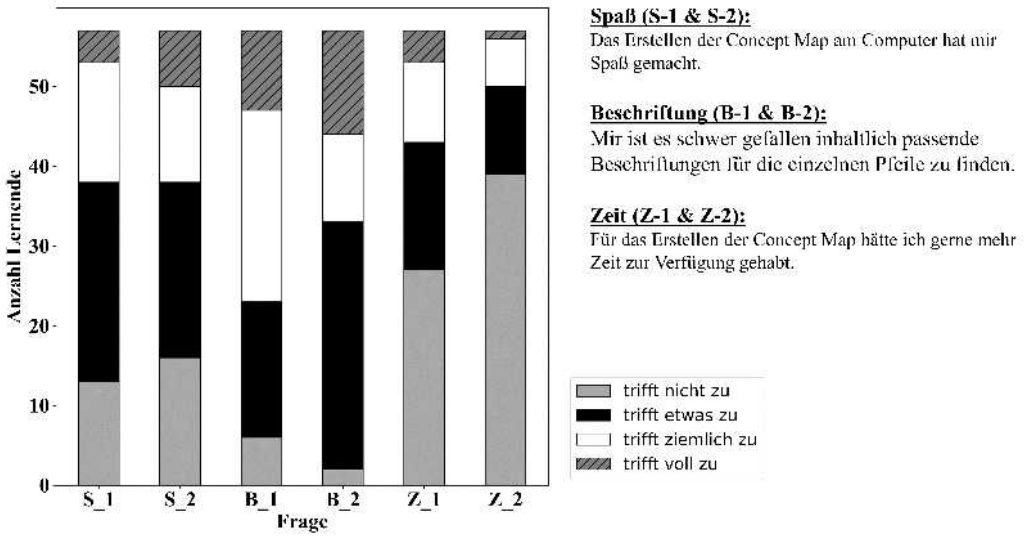


Abbildung 7.12: Vergleich der Ergebnisse beider Fragebögen

7.3.6 Analyse der Log-Daten

Durch die Verwendung der Lernplattform IPT konnten nicht nur die Concept Maps einfach und schnell automatisch ausgewertet werden, sondern auch diverse Log-Daten erhoben werden. Diese Möglichkeit wurde nur bei der zweiten Concept Map der Feedbackstudie genutzt, da die Implementierung dieser Erweiterung bei dem ersten Durchgang nicht fertiggestellt war. Mithilfe der Log-Daten kann nachvollzogen werden, in welcher Reihenfolge die Lernenden die Concept Map bearbeitet hatten. Da es sehr viele Wege gibt, die insgesamt 19 Propositionen zu bearbeiten, wird nur die Reihenfolge der ersten vier Propositionen betrachtet. Zusätzlich werden ab dem zweiten Bearbeitungsschritt nur die Wege analysiert, die von mindestens zwei Lernenden durchgeführt wurden, um eine übersichtliche Ergebnisdarstellung zu gewährleisten. In Abbildung 7.13 sind die Bearbeitungswege der Lernenden in einem Sankey-Diagramm dargestellt. Um die Grafik übersichtlicher zu gestalten, sind die bearbeiteten Propositionen abgekürzt und mit einer Nummer versehen. Die dazugehörige Proposition befindet sich unter dem Sankey-Diagramm. Zusätzlich werden noch die Häufigkeiten angegeben. Diese befinden sich hinter dem Doppelpunkt der Propositionsnummer. Die Zahl

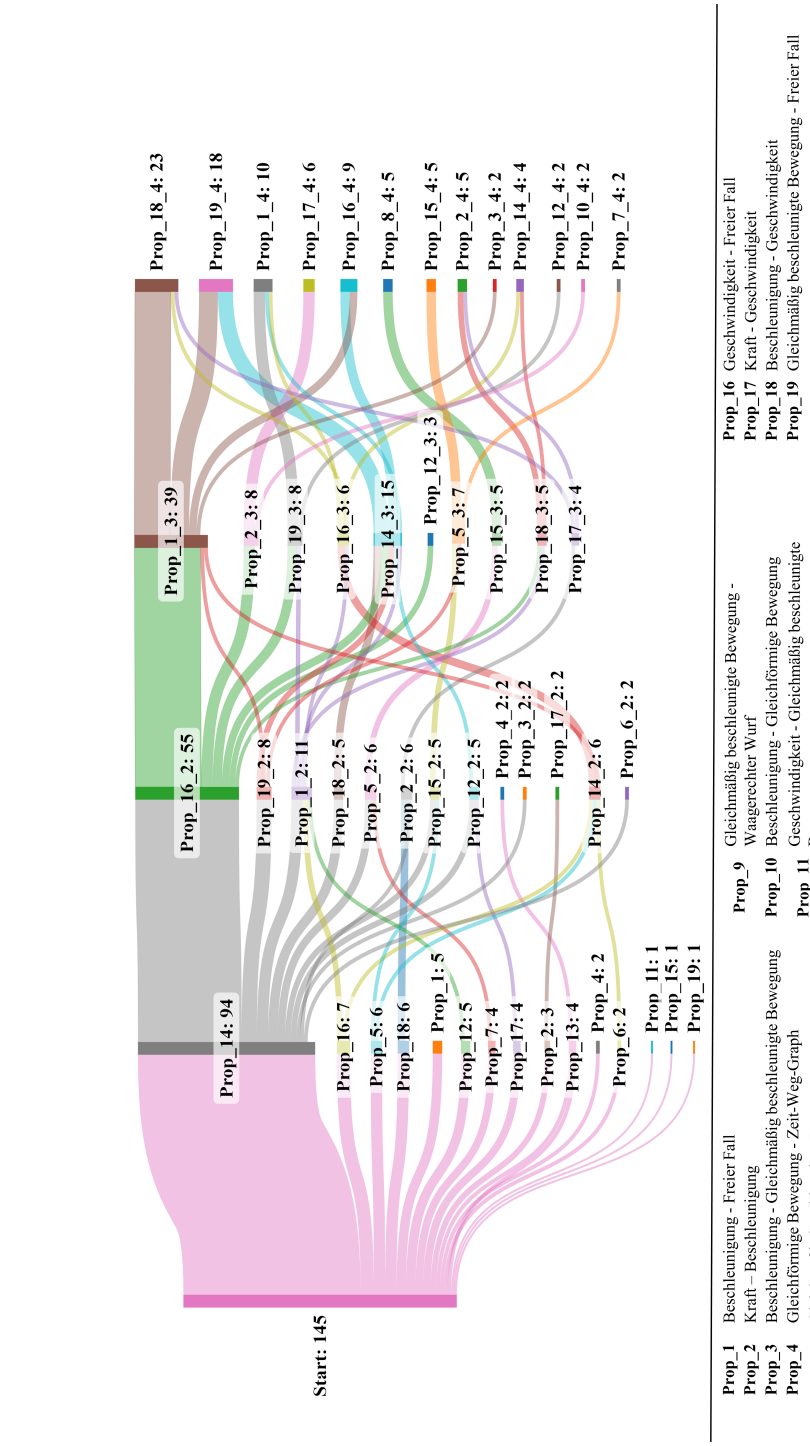


Abbildung 7.13: Bearbeitungswege der Lernenden für die ersten vier Propositionen der zweiten Concept Map

hinter dem Unterstrich zeigt den Bearbeitungsschritt auf und hat keine inhaltliche Bedeutung.
Die Log-Daten gaben Aufschluss über 145 Lernende. Durch die Abbildung 7.13 wird ersichtlich, dass ein Großteil der Lernenden zuerst die Proposition *Masse – Freier Fall* bearbeitet hat. Mit insgesamt 94 Lernenden haben demnach 65 % aller Lernenden diese Proposition zum Beginn ihrer Concept Map bearbeitet. Die restlichen Lernenden verteilen sich auf 14 weitere Propositionen, wovon *Geschwindigkeit – Freier Fall* mit sieben Lernenden noch die zweithäufigste Proposition darstellt. Demnach wurden die Propositionen *Gleichmäßig beschleunigte Bewegung – Zeit-Beschleunigung-Graph*, *Gleichmäßig beschleunigte Bewegung – Waagerechter Wurf*, *Beschleunigung – Gleichförmige Bewegung* und *Beschleunigung – Gleichmäßig beschleunigte Bewegung* von keinem Lernenden als erste Proposition bearbeitet. Beim zweiten Bearbeitungsschritt gibt es durch die Größe der Concept Map viele verschiedene Möglichkeiten. Es lässt sich doch erneut ein Weg identifizieren, der von einer Vielzahl von Lernenden genutzt wurde. Von den 94 Lernenden, die zunächst die Proposition *Masse – Freier Fall* bearbeitet hatten, hatten 55 im zweiten Bearbeitungsschritt sich mit der Proposition *Geschwindigkeit – Freier Fall* beschäftigt. Diese Proposition wurde ausschließlich von Lernenden erstellt, die vorher die Proposition *Masse – Freier Fall* bearbeitet hatten (siehe Abbildung 7.13). Acht andere Lernende bearbeiteten nach der Proposition *Masse – Freier Fall* die Proposition *Gleichmäßig beschleunigte Bewegung – Freier Fall*. Die Proposition *Beschleunigung – Freier Fall* wurde im zweiten Schritt von insgesamt elf Lernenden bearbeitet, wobei diese sich im ersten Schritt mit drei unterschiedlichen Propositionen beschäftigt hatten (siehe Abbildung 7.13). Im dritten Bearbeitungsschritt haben 39 Lernende die Proposition *Beschleunigung – Freier Fall* erstellt. Mit 35 Lernenden kam der Großteil der Lernenden von der Proposition *Geschwindigkeit - Freier Fall*. Die restlichen Lernenden bearbeiteten im dritten Schritt neun verschiedene Propositionen, wobei die Proposition *Masse – Freier Fall* von 15 Lernenden am zweithäufigsten erstellt wurde. Aus dieser Betrachtung kann auch geschlossen werden, dass ein kleiner Teil der Lernenden zu Beginn der Concept Map bestimmte Propositionen mehrmals bearbeitet haben muss. Im vierten und letzten betrachteten Schritt wurden insgesamt 13 verschiedene Propositionen erstellt. Dabei wurden die Propositionen *Beschleunigung – Geschwindigkeit* durch 23 und *Gleichmäßig beschleunigte Bewegung – Freier Fall* durch 18 Lernende am häufigsten erstellt.
In Abbildung 7.14 ist der häufigste Bearbeitungsweg der Lernenden grafisch dar-

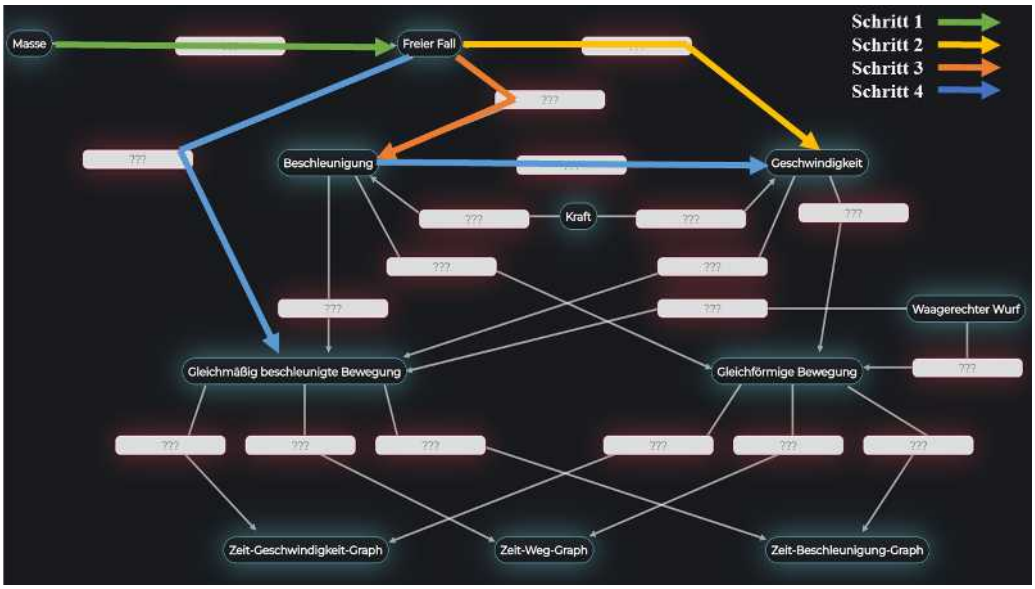


Abbildung 7.14: Der häufigste Bearbeitungsweg der zweiten Concept Map

gestellt. Für die ersten drei Schritte sind jeweils die am häufigsten bearbeiteten Propositionen markiert. Für den vierten Schritt wurden die beiden häufigsten Wege eingezeichnet. Man erkennt, dass die Lernenden die Concept Map von oben links aus angefangen haben zu bearbeiten und daher mit der Proposition *Masse – Freier Fall* starteten. Anschließend sind sie dieser Leserichtung gefolgt und haben die benachbarte Proposition *Geschwindigkeit – Freier Fall* bearbeitet. Im dritten Schritt ist weiterhin der Begriff des freien Falls der zentrale Begriff. Es wurde ausgehend davon die Proposition *Beschleunigung – Freier Fall* erstellt. Im Anschluss wurde von den meisten Lernenden entweder die benachbarte Proposition *Beschleunigung – Geschwindigkeit* oder erneut die vom freien Fall ausgehende Proposition *Gleichmäßig beschleunigte Bewegung – Freier Fall* bearbeitet.

7.3.7 Darstellung der Concept-Map-Entwicklung

Um die Concept Maps der Feedbackstudie analysieren zu können, wurden die Proposition mittels des Bewertungsschemas von einem menschlichen Bewerter bewertet. Dies ermöglicht die Überprüfung der Performance des Machine-Learning-Modells mit den für das Modell neuen Daten und es können Unterschiede sowie Gemeinsamkeiten zwischen den beiden Zeitpunkten festgestellt werden.
Beim ersten Zeitpunkt wurden insgesamt 3.458 Propositionen von den Lernenden erstellt. Im Schnitt enthielten die Concept Maps 13 Propositionen. Das bedeutet, dass sechs Propositionen durchschnittlich nicht bearbeitet wurden, was in 1.710

leere Proposition resultiert. Die dabei häufigste Proposition ist mit 247 *Masse – Freier Fall*, was durch die vorherige Analyse auch zu erwarten war. Mehr als die Hälfte (54 %) aller Propositionen wurde von den menschlichen Bewertern mit der Kategorie A, also als falsche Proposition markiert. Beim zweiten Zeitpunkt nahmen weniger Lernende teil, weswegen auch weniger Propositionen erstellt wurden. Im Datensatz befinden sich 2.255 ausgefüllte und nur noch 538 leere Propositionen. Das bedeutet, dass die Concept Maps im Schnitt 15 ausgefüllte und vier leere Propositionen enthielten, wobei auch hier die Proposition *Masse – Freier Fall* die meisterstellte Proposition ist.
Um die beiden Datensätze miteinander vergleichen und eine Veränderung feststellen zu können, werden erneut nur die Lernenden betrachtet, die an beiden Erhebungszeitpunkten an der Studie teilgenommen haben.

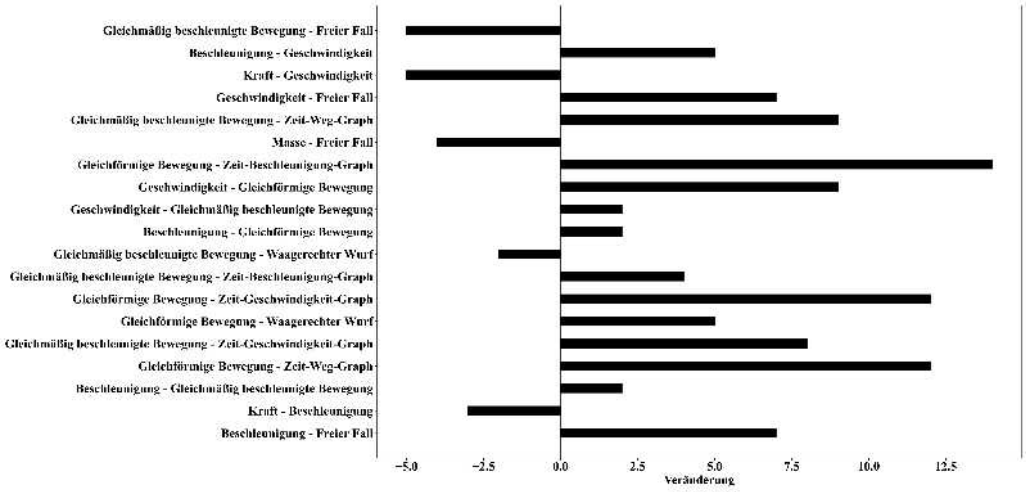


Abbildung 7.15: Veränderung der Häufigkeiten der 19 Propositionen zwischen den beiden Zeitpunkten der Feedbackstudie

In Abbildung 7.15 ist die Veränderung der Häufigkeiten der 19 Propositionen für die beiden Erhebungszeitpunkte der Feedbackstudie dargestellt. Die Abbildung zeigt, dass 79 Propositionen beim zweiten Zeitpunkt mehr erstellt wurden. Bei 14 von den 19 Proposition kann ein Anstieg verzeichnet werden. Lediglich bei *Gleichmäßig beschleunigte Bewegung – Freier Fall*, *Kraft – Geschwindigkeit*, *Masse – Freier Fall*, *Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf* und *Kraft – Beschleunigung* konnte ein Rückgang von bis zu fünf Propositionen festgestellt werden. Den stärksten Zuwachs findet man hingegen bei den Propositionen *Gleichförmige Bewegung – Zeit-Beschleunigung-Graph*, *Gleichförmige Bewegung – Zeit-Geschwindigkeit-Graph* und *Gleichförmige Bewegung – Zeit-Weg-Graph*

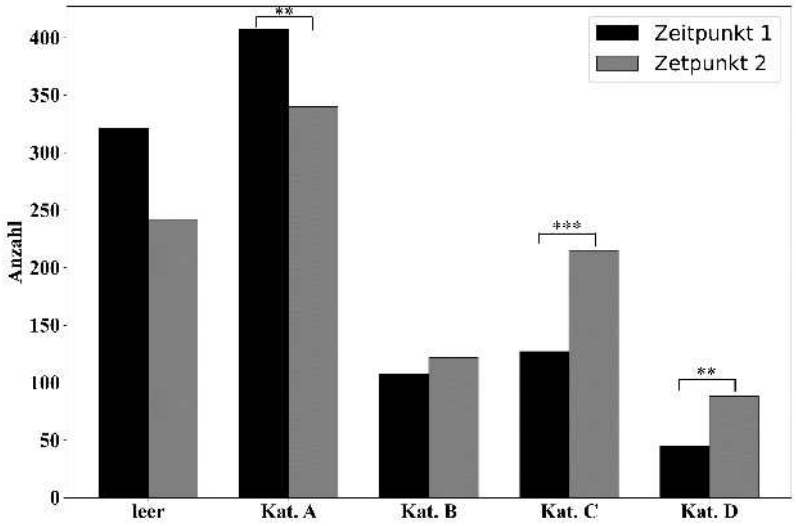


Abbildung 7.16: Häufigkeitsverteilung der vier Bewertungskategorien A, B, C und D sowie leere Propositionen für die beiden Zeitpunkte der Feedbackstudie und Darstellung der signifikanten Veränderungen

(siehe Abbildung 7.15). Auffällig ist, dass jede dieser vier Propositionen im unteren Drittel der Concept Map angeordnet ist und demnach einen weiten Abstand zur „Start-Proposition“ *Masse – Freier Fall* hat.
Betrachtet man nur die Häufigkeiten der vier Bewertungskategorien, erkennt man deutlichere Unterschiede als bei der Häufigkeitsverteilung der 19 Propositionen (siehe Abbildung 7.16). Beim ersten Zeitpunkt wurden die meisten Propositionen (407) mit der Kategorie A bewertet. Die Kategorien B und C waren mit 107 und 127 ähnlich verteilt, wohingegen die Kategorie D mit 45 den kleinsten Teil der Propositionen ausmachte. Zudem wurden insgesamt 321 Propositionen beim ersten Zeitpunkt von den Lernenden nicht bearbeitet (siehe Abbildung 7.16). Vergleicht man die Verteilung des ersten Zeitpunkts mit der des zweiten Zeitpunkts, sieht man einen Rückgang der falschen Propositionen (Kategorie A) auf 340. Die Anzahl der Propositionen, die mit der Kategorie B bewertet wurden, ist allerdings recht stabil geblieben. Im Gegensatz dazu sind die Propositionen der Kategorien C um 88 Propositionen gestiegen. Auch die Anzahl der Propositionen, die mit Kategorie D bewertet wurden, ist fast um das Doppelte gestiegen. Betrachtet man nur die nicht bearbeiteten Propositionen, erkennt man einen Rückgang auf 242 leere Propositionen beim zweiten Zeitpunkt. Aus Abbildung 7.16 geht demnach hervor, dass bei beiden Zeitpunkten die falschen Propositionen dominierten. Jedoch trat beim zweiten Zeitpunkt eine Verschiebung in Richtung der Bewertungskategorien C und D auf.

Um zu überprüfen, ob die Unterschiede auch signifikant sind, wurde ein Wilcoxon-Vorzeichen-Rangtest angewendet. Zur Berechnung des Wilcoxon-Tests wurden die einzelnen Bewertungskategorien separat betrachtet. Der Wilcoxon-Test wurde gewählt, da die Voraussetzungen für parametrische Tests nicht erfüllt und die Daten nicht normalverteilt waren sowie wegen der gepaarten Natur der Stichprobe. Der Test wurde mit der Python-Bibliothek SciPy durchgeführt.
Für die Berechnung des Wilcoxon-Tests wurde für jeden Lernenden die Anzahl der Propositionen für jede Bewertungskategorie der beiden Erhebungszeitpunkte bestimmt. Anschließend wurden die Differenzen berechnet, welche Aufschluss über die Veränderung zwischen den beiden Zeitpunkten geben können.
Die Anwendung des Wilcoxon-Tests zeigte einen signifikanten Unterschied zwischen den beiden Zeitpunkten für die Bewertungskategorien A, C und D (siehe Abbildung 7.16).
Die Ergebnisse deuteten darauf hin, dass die Anzahl der falschen Propositionen beim zweiten Zeitpunkt signifikant niedriger war als beim ersten Zeitpunkt ($z = -2,61$, $p < 0,01$). Die Effektstärke liegt bei $r = 0,37$ und entspricht nach Cohen (1988) einem mittleren Effekt. Für die Bewertungskategorie C konnte eine signifikante Steigerung nachgewiesen werden ($z = -3,90$, $p < 0,001$) (siehe Abbildung 7.16). Die Effektstärke von $r = 0,60$ entspricht einem starken Effekt (Cohen, 1988). Ebenso konnte für die Kategorie D eine signifikante Steigerung mit einem starken Effekt bestimmt werden ($z = -3,17$, $p < 0,01$, $r = 0,54$).
Um einen genaueren Blick in die Veränderungen der Propositionen zu erhalten, ist in Abbildung 7.17 ein Flussdiagramm der Bewertungskategorien zu finden. Hierzu wurde von jedem Lernenden für jede Proposition ermittelt, welche Bewertungskategorie die jeweilige Proposition beim ersten und zweiten Zeitpunkt hatte.
In Abbildung 7.17 ist zu erkennen, dass der größte Anteil der Propositionen, die beim ersten Zeitpunkt nicht bearbeitet wurden, auch beim zweiten Zeitpunkt nicht bearbeitet wurden (132). Ein weiterer großer Teil der leeren Propositionen wurde beim zweiten Zeitpunkt als falsche Proposition (Kategorie A) bewertet (91). Ein Großteil der Propositionen, der beim ersten Zeitpunkt falsch war, wurde auch beim zweiten Zeitpunkt als falsch bewertet (192). Es konnten aber auch Übergänge in die anderen Kategorien verzeichnet werden (siehe Abbildung 7.17).
Die Kategorie B erhält den größten Zuwachs von Propositionen, die beim ersten Zeitpunkt noch als falsch bewertet wurden (45), wobei auch hier viele Propositionen stabil geblieben sind (42) (siehe Abbildung 7.17).

Ebenso verhält es sich mit der Kategorie C, bei der der größte Zuwachs von den falschen Propositionen (86) kam. Zudem konnte neben den stabilen Propositionen (59) noch ein Übergang von leeren Propositionen in die Kategorie C verzeichnet werden (51).
Die Propositionen der Kategorie D erhalten fast in gleichen Teilen einen Zuwachs aus den leeren (23) und falschen (22) Propositionen, wobei auch hier die Hälfte der Propositionen stabil geblieben sind (23) (siehe Abbildung 7.17).

7.3.8 Machine-Learning-Auswertung der neuen Concept Maps

Die Bewertung der Leistung des Machine-Learning-Modells auf neuen Propositionen ist von entscheidender Bedeutung, da sie die tatsächliche Fähigkeit des Modells zur Verallgemeinerung seiner Lernmuster auf bisher ungesehene Beispiele widerspiegelt. Zudem können in den neuen Propositionen andere sprachliche Variationen oder Muster enthalten sein, die nicht im Trainingsdatensatz aus der Entwicklungsstudie vorhanden waren. Durch die Überprüfung mit neuen Daten wird die Robustheit des Modells gegenüber solchen Variationen und Störungen getestet.
In Abschnitt 6.2 und 6.3 wurden deshalb unterschiedliche Modelle trainiert und auf ungesehenen Testdaten getestet und evaluiert. Durch die Feedbackstudie und die daraus entstandenen neuen Propositionen kann die SVM erneut analysiert werden. Die Analyse liefert weitere Erkenntnisse über die Robustheit und Zuverlässigkeit des Modells bezüglicher Propositionen, die in einem realen Anwendungsfall entstanden sind.
Wie voranstehend beschrieben, wurden die neuen Propositionen von einem menschlichen Bewerter auf Grundlage der Daten aus der Entwicklungsstudie bewertet. Tabelle 7.10 zeigt die Performance der SVM mittels Accuracy, Cohen's Kappa und des gewichteten F1-Scores für die beiden Zeitpunkte der Feedbackstudie sowie den aggregierten Datensatz. Zum Vergleich sind in der Tabelle 7.10 noch einmal die Performance-Werte der SVM aus der Entwicklungsstudie dargestellt.
Für die Propositionen vom ersten Zeitpunkt konnte mit 0,83 ein leicht besserer F1-Score im Vergleich zur Entwicklungsstudie von 0,82 erreicht werden. Beim zweiten Zeitpunkt zeigt die SVM ein um 0,02 schlechteres Cohen's Kappa. Jedoch finden sich alle Werte in dem in Kapitel 5 gesetzten Akzeptanzbereich. Aggregiert man beide Datensätze und lässt die SVM diese automatisch auswerten, erhält

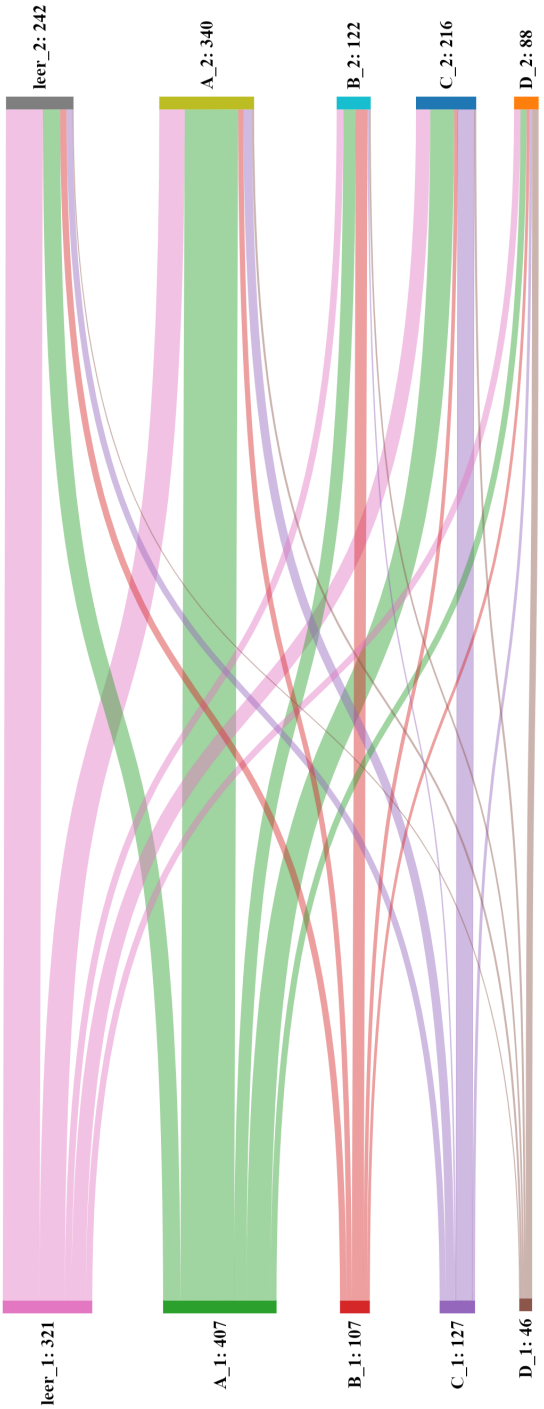


Abbildung 7.17: Veränderungen der Propositionen bezüglich der vier Bewertungskategorien A, B, C und D sowie leere Propositionen vom ersten zum zweiten Erhebungszeitpunkt

	Accuracy	Cohen's Kappa	F1-Score
Entwicklungsstudie	0,82	0,75	0,82
Zeitpunkt 1	0,83	0,74	0,83
Zeitpunkt 2	0,81	0,73	0,81
Zeitpunkt 1 + 2	0,82	0,74	0,82

Tabelle 7.10: Accuracy, Cohen's Kappa und gewichteter F1-Score der SVM bezüglich der neuen Propositionen beider Zeitpunkte

man eine identische Accuracy und einen identischen F1-Score im Vergleich zur Entwicklungsstudie. Das lässt insgesamt den Schluss zu, dass die SVM auch bei neuen Daten robust und zuverlässig Propositionen automatisch auswerten kann.

7.4 Beantwortung der Forschungsfragen und Diskussion der Ergebnisse

In der Entwicklungsstudie wurde die Grundlage für die Feedbackstudie gelegt, indem ein Machine-Learning-Modell zur automatischen Auswertung von Propositionen einer Concept Map zum Thema Mechanik entwickelt wurde. Durch die Analyse der unterschiedlichen Modelle und der detaillierten Untersuchung der SVM konnten bereits einige Forschungsfragen und Hypothesen beantwortet und diskutiert werden. Dieser Abschnitt dient dazu, die Ergebnisse der Feedbackstudie zusammenzufassen und die restlichen Forschungsfragen und Hypothesen zu beantworten und einzuordnen.

Machine-Learning-Modell

Im ersten Untersuchungsschwerpunkt Machine Learning in der Physikdidaktik wurde die Hypothese aufgestellt, dass das entwickelte Machine-Learning-Modell eine niedrigere Übereinstimmung mit menschlichen Bewertern aufweist, wenn es neue Concept Maps automatisch auswerten soll, die das Vorwissen abfragen, im Vergleich zu Concept Maps, die am Ende der Lerneinheit erstellt wurden. Die Hypothese wurde aufgestellt, da in der Entwicklungsstudie Concept Maps zum Trainieren und Testen der Modelle genutzt wurden, die nach der Unterrichtseinheit Mechanik erhoben wurden und sich bezüglich der sprachlichen Praktiken und des inhaltlichen Gehalts zur ersten Concept Map der Feedbackstudie unterscheiden könnten. Die Ergebnisse haben allerdings gezeigt, dass das entwickelte

Machine-Learning-Modell eine vergleichbare Leistung wie in der Entwicklungsstudie zeigen konnte. Das bedeutet, dass die SVM in beiden Zeitpunkten der Feedbackstudie die Propositionen der Lernenden mit einer gleichbleibenden Genauigkeit in dieselben Bewertungskategorien wie die menschliche Bewertung zuordnen kann. Daher scheint es, dass die Concept Maps aus der Feedbackstudie keinen negativen Einfluss auf die automatische Auswertung hatten und damit die Hypothese 1 *Das entwickelte Machine-Learning-Modell wird nach dem Trainingsprozess eine niedrigere Übereinstimmung mit menschlichen Bewertungen aufweisen, wenn es neue Concept Maps auswertet, die das Vorwissen abfragen, im Vergleich zu Concept Maps, die am Ende der Lerneinheit erstellt wurden* verworfen werden kann.
Die Ergebnisse haben sogar gezeigt, dass die Übereinstimmung beim ersten Zeitpunkt etwas höher war als beim zweiten Zeitpunkt. Das zeigt, dass die Hypothese möglicherweise sogar in die entgegengesetzte Richtung hätte formuliert werden können. Ein möglicher Grund könnte der niedrige Anteil der Kategorie-D-Propositionen im Datensatz des ersten Zeitpunkts sein. Die Analyse der SVM in der Entwicklungsstudie hatte gezeigt, dass die Performance der SVM durch Propositionen, die Formeln oder Zahlen enthalten, beeinflusst wird. Da das Bewertungsschema so konzipiert wurde, dass richtige funktionale Zusammenhänge in die Bewertungskategorie D eingeordnet werden sollten, kann der niedrige Anteil dieser Kategorie auch gleichbedeutend mit einer niedrigen Anzahl an Propositionen sein, die Formeln oder Zahlen enthalten. Dies würde auch die leicht schlechtere Übereinstimmung beim zweiten Zeitpunkt erklären, da hier der Anteil der Kategorie-D-Propositionen höher ist.
Insgesamt zeigen die Ergebnisse erneut, dass die SVM die Propositionen in dem im Vorfeld gesetzten Akzeptanzbereich automatisch auswerten kann und bestätigen so die in der Entwicklungsstudie gezeigten Ergebnisse. Dennoch muss weiter daran gearbeitet werden, einen ausgewogenen Trainingsdatensatz zu erhalten, damit das Machine-Learning-Modell noch robuster gegenüber Veränderungen in den Daten wird. Eine Vergrößerung des Trainingsdatensatzes ist dabei nur eine Möglichkeit. Ebenso denkbar ist die Erhebung von Concept Maps, die aus unterschiedlichen Regionen oder nicht ausschließlich von reinen Gymnasien stammen. Die Concept Maps aus der Feedbackstudie können ein erster Anfang sein und Teil eines neuen Trainingsdatensatzes werden, sodass ein neues noch besseres Modell trainiert werden kann.

Formatives Assessment

Der zweite Schwerpunkt des Erkenntnisinteresses dieser Arbeit war das formative Assessment. Um eine Grundlage für die Bewertung der automatischen Auswertung zu haben, sollte zunächst die generelle Einstellung und Nutzungshäufigkeit der Lehrkräfte bezüglich formativer Assessments im Physikunterricht untersucht werden. Vor diesem Hintergrund wurde die Hypothese aufgestellt, dass die teilnehmenden Lehrkräfte über ein eingeschränktes Spektrum an formativen Assessment-Methoden berichten, was sich auf die zeitintensive Nutzung zurückführen lässt.
Durch die Auswertung der Interviews der acht Lehrkräfte konnte gezeigt werden, dass den meisten Lehrkräften der Begriff des formativen Assessments nicht bekannt war. Dennoch konnten zumindest Teile der fünf Schlüsselstrategien des formativen Assessments im Unterricht wiedergefunden werden. Dabei wurde vor allem auf die Erfassung des Lernstands zurückgegriffen, z. B. in Form von unbewerteten schriftlichen Lernzielkontrollen oder Beobachtungen im Unterrichtsgeschehen. Die Mehrheit der beschriebenen Methoden können nach der Klassifizierung von Shavelson et al. (2008) als *On-the-fly*-Methoden bezeichnet werden.
Es konnte weiterhin erkannt werden, dass die Lehrkräfte formatives Assessment oftmals mit einer Rückmeldung gleichsetzen, wobei nicht explizit der Zweck dieser Rückmeldung durchdacht wird. Dies stellt noch heraus, dass Rückmeldungen das zentrale Element des formativen Assessments sind (siehe Kapitel 2).
Es stellte sich außerdem heraus, dass individuelle Rückmeldungen im Rahmen eines formativen Assessments aufgrund der großen Klassengröße nicht machbar und daher im normalen Schulalltag nicht realistisch sind. Die Aussagen aus den Interviews zeigten deutlich, dass die Lehrkräfte es zeitlich nicht schaffen, jedem einzelnen Lernenden eine individuelle Rückmeldung zu ermöglichen, obwohl sie den lernförderlichen Aspekt dahinter sehen. Aufgrund dieser Ergebnisse kann die Hypothese 3 *Die teilnehmenden Lehrkräfte berichten über ein recht eingeschränktes Spektrum an formativen Assessment-Methoden* zumindest in Teilen bestätigt werden.
Die Lehrkräfte nannten kaum explizite formative Assessment-Methoden aus ihrer eigenen Unterrichtserfahrung, weshalb die Forschungsfrage 4 wie folgt beantwortet werden kann: *Die teilnehmenden Lehrkräfte setzen formatives Assessment in geringem Umfang im Unterricht ein.*
Allerdings lässt sich nicht abschließend klären, ob dies an der geschilderten Zeitproblematik oder an der fehlenden Verknüpfung zwischen einzelnen Methoden

und dem Begriff des formativen Assessment lag. Daher wäre es für weitere Forschungsarbeiten wichtig, detailliert auf die einzelnen Aspekte des formativen Assessments einzugehen und mit den teilnehmenden Lehrkräften zu diskutieren. Zudem muss an dieser Stelle auch die kleine Stichprobe von acht Lehrkräften berücksichtigt werden, welche für weitere Studien erhöht werden sollte. So können validere Aussagen erhoben werden, die einen noch genaueren Einblick in die Unterrichtspraxis der Physiklehrkräfte bieten. Diese Ergebnisse könnten genutzt werden, um die angesprochenen Probleme zu verbessern oder eine Entwicklung von gezielten Weiterbildungen für Lehrkräfte zu konzipieren.

Durch die automatische Auswertung der einzelnen Propositionen sollte ein schnelle und einfache Bereitstellung von Informationen für Lehrkräfte und Lernende erfolgen. Diese Informationen könnten aufseiten der Lehrkräfte zum Beispiel für die Anpassung des weiteren Unterrichts oder individuelle Rückmeldungen für die Lernenden genutzt werden. Deshalb sollte untersucht werden, inwiefern die automatische Auswertung der Propositionen dazu beitragen kann, formative Assessments im Physikunterricht zu integrieren.
Aus der Literatur war bereits bekannt, dass gezielte individuelle Rückmeldungen für Lernende sehr zeitaufwendig sind, was durch die Interviews bestätigt wurde. Da die Lehrkräfte wenig zusätzliche Zeit für die Auswertung der Propositionen aufbringen mussten, wurde die Hypothese aufgestellt, dass die bereitgestellten Informationen für eine gezielte individuelle Rückmeldung genutzt werden. Durch die schnelle und vergleichsweise einfache Auswertung konnte erwartet werden, dass die angesprochene Zeitproblematik abgeschwächt wird und die Lehrkräfte neue Feedback-Möglichkeiten nutzen.
Aus den Interviews wurde ersichtlich, dass die Lehrkräfte die automatische Auswertung generell als nützlich für ihre Arbeit empfunden haben. Jedoch konnte festgestellt werden, dass die einzelnen Elemente des automatischen Feedbacks eher als zu detailliert beschrieben wurden. Dies galt sowohl für die Lehrkräfte, die das einfache Feedback erhalten haben als auch für die Lehrkräfte aus der elaborierten Feedback-Gruppe.
An dieser Stelle muss deshalb diskutiert werden, wie detailliert die automatische Auswertung sein muss, um optimal auf die Bedürfnisse der Lehrkräfte abgestimmt zu sein. Zwar bietet das elaborierte Feedback mehr und genauere Informationen über den Leistungsstand der Lernenden. Allerdings scheint die Fülle an Informationen die Lehrkräfte zu überfordern, sodass die Bereitstellung dieser präziseren

Informationen überflüssig zu sein scheint. Es wurden daher auch besonders die Auswertungen, die einen zusammenfassenden Charakter auf Klassenebene besitzen wie die *problematischen Propositionen* oder der Übersichtsgraph als hilfreich eingeschätzt. Feedback-Elemente, die auf die Leistung der einzelnen Lernenden eingegangen sind, wurden weniger bis gar nicht betrachtet.
Durch die Interviews konnten auch Einblicke in die Verwendung der automatischen Auswertung gewonnen werden. Die Studie wurde so geplant, dass die Lehrkräfte möglichst viel Freiraum haben sollten, um mittels eines explorativen Ansatzes das Nutzungsverhalten analysieren zu können. Die Ergebnisse zeigten, dass die meisten Lehrkräfte zwar die automatische Auswertung betrachteten, jedoch konnten kaum konkrete Handlungsschritte erkannt werden. Die wenigen Lehrkräfte, die angaben, das Feedback genutzt zu haben, haben es nicht für gezielte individuelle Rückmeldungen genutzt. Es wurde vielmehr genutzt, um Wissenslücken der gesamten Klasse zu thematisieren und zu schließen. Keine der acht interviewten Lehrkräfte hat die automatische Auswertung für ein gezieltes, individuelles Feedback genutzt. Daher kann die aufgestellte Hypothese 4 *Die Lehrkräfte nutzen die automatische Auswertung für eine gezielte individuelle Rückmeldung* verworfen werden.
Diese Ergebnisse verwundern, da nicht nur aus der Theorie bekannt war, dass individuelle Rückmeldungen sehr zeitintensiv sind, sondern die meisten Lehrkräfte dies auch im Interview schilderten. Die automatische Auswertung, die von den Lehrkräften als hilfreich erachtet wurde, ermöglichte es den Lehrkräften, auf ihre Klasse abgestimmte Informationen für den weiteren Lehr-Lern-Prozess auch auf einer individuellen Ebene zu nutzen. Es kann demnach geschlussfolgert werden, dass das Potenzial solcher automatischen Feedbacktools vorhanden ist und von manchen Lehrkräften auch gewünscht ist. Jedoch ist in dieser Studie die praktische Umsetzung in den Schulalltag, zumindest auf einer individuellen Ebene, kaum gelungen.
Die Ergebnisse der Untersuchung lassen den Schluss zu, dass die teilnehmenden Lehrkräfte formative Assessments nur in geringem Maße in ihrem Physikunterricht nutzen. Die automatische Auswertung der Concept Map hat zu keiner Veränderung dieses Verhaltens geführt. Daher kann die Forschungsfrage 5 wie folgt beantwortet werden: *Die automatische Auswertung der Concept Map konnte nur in geringem Maße dazu beigetragen, formative Assessments in den Physikunterricht zu integrieren*.
Ein möglicher Grund könnte die vorher festgelegte Struktur der Concept Map

gewesen sein. Durch diese geschlossene Variante hatten die Lehrkräfte keinen Spielraum, die Concept Map auf ihre Klasse anzupassen. So bemerkten einige Lehrkräfte, dass die eingesetzte Concept Map keine optimale Passung mit dem Leistungsstand der Lernenden hatte. Es wurde sich zwar bewusst für diese Concept Map entschieden, um eine verlässliche automatische Auswertung entwickeln zu können. Dennoch sollte in weiteren Forschungsarbeiten eine offenere Variante getestet werden, bei der die Lehrkräfte mehr Freiräume haben, die Concept Map auf die Bedürfnisse ihrer Lerngruppe anpassen zu können. Das Gleiche gilt auch für die einzelnen Feedback-Elemente. Es wurde sich bewusst für ein breites Spektrum an verschiedenen Auswertungen entschieden, um überprüfen zu können, welche Elemente für welchen Zweck genutzt werden. Bei zukünftigen Arbeiten könnten Lehrkräfte bereits bei der Entwicklung solcher automatischen Feedbacktools einbezogen werden. Dies ermöglicht eine bessere Passung an den Schulalltag und könnte zu einer häufigeren und gezielteren Nutzung der automatischen Auswertung der Concept Maps führen. Es muss dennoch erneut angemerkt werden, dass in der Feedbackstudie nur acht Lehrkräfte interviewt wurden und diese Ergebnisse keine allgemeinen Schlussfolgerungen zulassen und deswegen weiter offen geforscht werden sollte.

Der theoretische Hintergrund zum automatischen Feedback (siehe Kapitel 2) hat gezeigt, dass eine transparente und klar verständliche Rückmeldung ein notwendiges Kriterium für ein formatives Assessment ist. Außerdem sind Lehrkräfte nach Herding et al. (2010) in den meisten Fällen in der Lage, qualitativ hochwertigeres Feedback als ein automatisches Feedbackmodell zu erzeugen. Die Mehrheit der interviewten Lehrkräfte war überzeugt, dass die automatische Auswertung mehr Fehler macht als eine menschliche Lehrkraft, was die Annahme von Herding et al. (2010) bestätigt. Es wurde angemerkt, dass nicht klar war, wie die automatische Auswertung mit Rechtschreibfehlern oder Fehlern bezüglich Groß- und Kleinschreibung umgeht. Man kann also sehen, dass eine fehlende Transparenz der automatischen Auswertung aus Sicht der Lehrkräfte vorhanden ist. Es war für die Lehrkraft teilweise nicht ersichtlich, wie das Machine-Learning-Modell zu den Auswertungen gekommen ist. Eine Lehrkraft zweifelt zudem an, ob eine KI Interpretationsfähigkeit besitzt. Transparenz, Interpretationsfähigkeit und Nachvollziehbarkeit von KI-basierten Anwendungen sind insbesondere in sensiblen Anwendungsgebieten wie Bildung, wo Entscheidungen Auswirkungen auf Einzelpersonen haben können, ein wichtiger Aspekt. In dieser Studie wurde

die automatische Auswertung für ein formatives Assessment eingesetzt, dennoch sollte auch hier ein gewisses Vertrauen in die automatische Auswertung vorhanden sein. KI-Modelle sind oftmals sehr komplex und werden daher als „Blackbox“ angesehen. Es sollte demnach in zukünftigen Arbeiten versucht werden, die Lehrkräfte, die eine KI-basierte automatische Auswertung nutzen sollen, auch auf einer methodischen Ebene stärker einzubinden und zu informieren. Denn nur so kann ein Vertrauen in die für viele Lehrkräfte vielleicht noch unbekannten KI-basierten Systeme geschaffen und Vorurteile oder Ängste beseitigt werden. Des Weiteren wäre eine gezielte Ausbildung mit dem Umgang von KI-basierten Systemen in der Studienphase an Universitäten denkbar, um den Grundstein für eine zukünftige Arbeit mit diesen Systemen zu schaffen.
Zudem kann davon ausgegangen werden, dass an dieser Studie vor allem Lehrkräfte teilgenommen haben, die KI-basierten Systemen zumindest nicht komplett kritisch gegenüberstehen oder ein gewisses Interesse an solchen Anwendungen haben. Speziell für die restlichen Lehrkräfte könnte der Aspekt der Transparenz und des Vertrauens noch entscheidender sein, was den Ansatz nach einer stärkeren Einbeziehung nochmals stützt.

Zusätzlich zu den Lehrkräften erhielten die Lernenden ebenfalls ein automatisches Feedback, das aus der Auswertung beider Concept Maps bestand. Es wurde die Hypothese abgeleitet, dass die Lernenden das automatische Feedback positiv bewerten, da sie eine auf ihre Leistung abgestimmte Rückmeldung erhalten. Die Auswertung des Feedback-Fragebogens zeigte, dass die Lernenden die automatische Auswertung grundsätzlich als nützlich und wirksam einschätzten. Auch die Betrachtung der Antworten aus den offenen Fragen zeigte, dass die Mehrheit der Lernenden die automatische Auswertung als hilfreich betrachteten. Ein Lernender merkte sogar an, dass im normalen Unterricht für individuelle Rückmeldungen keine Zeit bliebe und dies durch die automatische Auswertung nun möglich sei.
Durch Unterrichtsausfälle, Krankheit oder Praktika sind während der Feedbackstudie allerdings viele Unterrichtsstunden ausgefallen oder nur mit einer geringen Anzahl an Lernenden durchgeführt worden. Folglich wurde das Lernenden-Feedback von den Lehrkräften teilweise als Hausaufgabe aufgegeben oder nur als optional betrachtet. Insofern muss an dieser Stelle festgehalten werden, dass aufgrund der geringen Stichprobe von 26 Lernenden keine validen Aussagen getroffen werden können. Auch die offene Frage wurde letztlich nur von acht Lernenden beantwortet, weswegen kein aussagekräftiges Ergebnis erzielt werden konnte. Die

aufgestellte Hypothese 5 *Die Lernenden werden die automatische Auswertung positiv bewerten, da sie dadurch ein individuelles Feedback erhalten* kann somit nicht abschließend beantwortet werden. Dennoch lässt sich aus der Analyse der Fragebögen ableiten, dass die automatische Auswertung für die Lernenden von Vorteil sein kann.
Bei einer nächsten Untersuchung müsste demnach mehr darauf geachtet werden, dass auch die Lernenden das Feedback zur Verfügung gestellt bekommen und letztlich ebenfalls nutzen. Erst dann können Ergebnisse erzielt werden, die eine aussagekräftige Schlussfolgerung über die Effekte der automatischen Rückmeldung für die Lernenden zulassen.

Concept Maps

Das dritte und letzte Erkenntnisinteresse lag auf der eingesetzten Concept Map. Durch Concept Maps können reichhaltige Informationen über den Leistungsstand der Lernenden ermittelt werden, die für eine lernförderliche Rückmeldung genutzt werden können (siehe Kapitel 3). Wie die Feedback-Studie gezeigt hat, sind Concept Maps allerdings immer noch eine recht unbekannte Methode, da fast drei Viertel der Lernenden vorher noch nie mit einer Concept Map gearbeitet haben. Aus dieser Perspektive ist es umso interessanter, die Herausforderungen und potenziellen Nutzen der Concept Map als einer formativen Assessment-Methode zu untersuchen.
Indem die Lernenden die Mechanik-Concept-Maps bearbeitet haben, wurden sie dazu angeregt, über ihr vorhandenes Wissen nachzudenken und es zu organisieren und zu strukturieren, da sie Beziehungen zwischen den verschiedenen Begriffen herstellen mussten. Sie mussten entscheiden, wie die Begriffe miteinander zusammenhängen und wie sie Zusammenhänge in Form von Propositionen erklären können. Dies erforderte eine kritische Auseinandersetzung mit dem eigenen Wissen und eine Bewertung der eigenen Verständnisse. Concept Maps ermöglichen also den Lernenden, ihr eigenes Wissen zu erkennen und zu reflektieren (siehe Kapitel 3). Daher wurde die Hypothese gebildet, dass durch die eingesetzte Mechanik-Concept-Map die Selbstreflexion der Lernenden über ihr eigenes Wissen gefördert wird.
Nach der Bearbeitung der ersten Concept Map gab mehr als die Hälfte der Lernenden an, dass es ihnen schwergefallen ist, passende Beschriftungen für ihre Propositionen in der Concept Map zu finden. Dennoch konnte durch den Fragebo-

gen festgestellt werden, dass den Lernenden ihre eigenen Wissenslücken bewusst geworden sind. Dieses Ergebnis lässt sich auch in den Aussagen der offenen Frage wiederfinden. Insgesamt waren 82 % der befragten Lernenden der Auffassung, dass Concept Maps öfter im Unterricht eingesetzt werden sollten. Ein Argument für den häufigeren Einsatz war, dass durch die Concept Map die Zusammenhänge des behandelten Themas sichtbar wurden und dies einen lernförderlichen Effekt habe. Darüber hinaus wurde von den Lernenden angemerkt, dass Concept Maps es ermöglichen, das eigene Wissen zu reflektieren und zu unterscheiden, welche Zusammenhänge für den Unterricht relevant sind. Es gab allerdings auch Lernende, die überfordert waren und die Bearbeitung der Concept Map als zu anspruchsvoll einstuften. Durch den Fragebogen nach der zweiten Concept Map konnte jedoch die Tendenz erkannt werden, dass beim Erstellen der Propositionen weniger Lernende Schwierigkeiten hatten.
Damit kann die Forschungsfrage 6 beantwortet werden: *Die Lernenden konnten mehrere potenzielle Vorteile wie das Bewusstwerden eigener Wissenslücken oder die Visualisierung von Zusammenhängen bei der Erstellung der Propositionen erkennen. Aber auch Herausforderungen wie ein zu hoher fachlicher Anspruch konnten identifiziert werden*.
Insgesamt lässt sich schlussfolgern, dass die Lernenden durch die Bearbeitung der beiden Mechanik-Concept-Maps ihr eigenes Wissen bewusst reflektieren und dessen Struktur erkennen konnten. Dies kann letztlich ihre Fähigkeit verbessern, das Gelernte zu verstehen, zu behalten und anzuwenden. Somit kann die Hypothese 6 *Die Erstellung der Concept Maps fördert die Selbstreflexion der Lernenden über ihr eigenes Wissen* bestätigt werden.

Trotzdem sollte auch der Einsatz der Concept Map kritisch reflektiert werden. Viele Lernende gaben an, dass es ihnen sowohl bei der ersten als auch bei der zweiten Concept Map schwerfiel, passende Beschriftungen für ihre Propositionen zu finden. Dieses Ergebnis ist nicht verwunderlich, da Concept Maps für schwache und starke Lernende als herausfordernd gelten (siehe Kapitel 3). Wenn Lernende Concept Maps zum ersten Mal bearbeiten, kann dies Auswirkungen auf den Bearbeitungsprozess und zu den Schwierigkeiten beigetragen haben. Die Lernenden mussten sich nicht nur inhaltlich mit der Concept Map auseinandersetzen, sondern auch kognitive Ressourcen für das Nachvollziehen der neuen Methode aufbringen. Es wurde sich zwar bewusst für eine sehr kurze Concept-Map-Einführung entschieden, da die Concept Map stark vorstrukturiert war, allerdings sollte in

späteren Untersuchungen mehr Zeit für den Einstieg genutzt werden. Eine bessere Grundlage ermöglicht es den Lernenden, sich mehr auf die Formulierung der Propositionen zu konzentrieren und weniger kognitive Ressourcen für die visuelle Interpretation der Concept Map aufzuwenden.
In der Feedbackstudie wurde der gesamte Ablauf digital auf der Lernplattform IPT durchgeführt. So auch die Einführung und die Erklärung der Studie. Durch die Interviews wurde ersichtlich, dass der Großtteil der Lehrenden bislang nicht mit einer Concept Map gearbeitet hatte. Das bedeutet, dass die Lernenden keine fundierten Antworten auf eventuelle Fragen erwarten konnten, da die Lehrkräfte sich selbst einarbeiten mussten. Daher sollte auch bei geschlossenen Concept Maps stärker auf ein Concept-Map-Training oder eine -Einführung geachtet werden, damit alle Lernenden und Lehrkräfte den Aufbau und Umgang mit einer Concept Map nachvollziehen können. Zudem könnte man auch das Bewertungsschema mit in die Einführung einbinden. Den Lernenden konnte im Vorfeld nicht klar sein, welche Grundlage zur Bewertung ihrer Proposition genutzt wurde. Daher könnte man für weitere Arbeiten, die sich mit einer automatischen Auswertung von Concept Maps beschäftigen, auch exemplarische Beispiele für die jeweiligen Bewertungskategorien mit in der Einführung berücksichtigen.

Eine Problematik bei der Verwendung von Concept Maps als einer formativen Assessment-Methode ist, dass die Auswertung anspruchsvoll und zeitaufwendig sein kann. Je nach Komplexitätsgrad und Umfang der Concept Map kann die Analyse der einzelnen Propositionen herausfordernd sein. Bei umfangreichen Concept Maps ist es nicht ersichtlich, an welcher Stelle der Auswertungsprozess begonnen werden soll, da es keine feste Leserichtung gibt. Daher wurde angenommen, dass dies auch für die Erstellung der eingesetzten Mechanik-Concept-Map gilt und die Hypothese formuliert, dass sich keine Muster in der anfänglichen Bearbeitung der Lernenden finden lassen.
Durch die Auswertung der Log-Daten der Lernenden konnte herausgefunden werden, dass es eine typische Herangehensweise an die eingesetzte Concept Map gibt. Fast zwei Drittel der Lernenden starteten mit der Proposition *Masse - Freier Fall* und damit in der linken oberen Ecke. Anschließend führte der Bearbeitungsweg ausgehend von dem Begriff des freien Falls weiter nach rechts und nach unten. Es lässt sich also eine typische Bearbeitungsrichtung erkennen, die einer Leserichtung eines Texts entsprechen könnte. Damit kann die Hypothese 7 *In den anfänglichen Bearbeitungsschritten der Lernenden lassen sich keine Muster*

identifizieren verworfen werden.
Diese Erkenntnisse können für die Planung einer weiteren Studie genutzt werden. Wenn im Vorfeld zu erwarten ist, an welcher Stelle der Concept Map die Lernenden ihre Bearbeitung starten, könnte man z. B. Begriffe wählen, die allen Lernenden aufgrund ihres Vorwissens bekannt sind. Dies könnte Einfluss auf die Motivation der Lernenden haben, wenn sie nicht direkt mit einer herausfordernden Proposition starten müssen. Es wäre auch denkbar, die Struktur der Concept Map auf die Leistung der Lernenden anzupassen, um eine individuellere Passung zu erreichen.

Eine weitere Untersuchung hat die Veränderung zwischen den beiden Concept Maps in der Feedbackstudie betrachtet. Die eingesetzte Concept Map hatte das Ziel, Informationen und Hinweise bereitzustellen, die die Lehrkräfte und Lernenden im Sinne eines formativen Assessments nutzen sollten. Die Studie und die Concept Map wurden daher nicht vor dem Hintergrund einer Überprüfung der Lernleistung oder zur Messung einer Kompetenzsteigerung konzipiert. Dennoch sollte herausgefunden werden, ob ein Unterschied in den Concept Maps zwischen den beiden Zeitpunkten erkennbar ist. Da die Concept Map Begriffe enthielt, die erst im Laufe der Lerneinheit behandelt wurden und der erste Erhebungszeitpunkt direkt am Beginn der Einheit war, konnte davon ausgegangen werden, dass zwischen den beiden Zeitpunkten eine signifikante Veränderung festgestellt werden konnte, wobei speziell auf falsche Propositionen geachtet werden sollte.
Die erste Analyse hat gezeigt, dass die meisten Propositionen in beiden Concept Maps mit einer ähnlichen Häufigkeit erstellt wurden. Der größte Unterschied zwischen beiden Zeitpunkten konnte bei der Proposition *Gleichförmige Bewegung – Zeit-Beschleunigung-Graph* beobachtet werden. Hier gab es eine Zunahme von 14 Propositionen. Es konnte zudem ein Zuwachs von drei Propositionen verzeichnet werden, die ebenfalls im unteren Drittel der Concept Map waren. Dies könnte sich auf die Erfahrung der Lernenden mit der Bearbeitung der Concept Map zurückführen lassen. Geht man davon aus, dass der gezeigte typische Bearbeitungsweg auch für die erste Concept Map gilt, kann der Zuwachs der unteren Proposition mit der erneuten Bearbeitung zusammenhängen.
Die Ergebnisse zeigen weiterhin, dass sich die Anzahl falscher Propositionen beim zweiten Erhebungszeitpunkt deutlich verkleinert. In der Folge konnte durch einen Wilcoxon-Test nachgewiesen werden, dass sich die Anzahl der falschen Propositionen signifikant zwischen der ersten und zweiten Concept Map unter-

scheidet. Damit kann die Hypothese 8 *Die Anzahl der falschen Propositionen wird beim zweiten Erhebungszeitpunkt signifikant geringer sein als beim ersten* bestätigt werden. Zudem wurden insgesamt weniger Proposition unbearbeitet gelassen. Es konnte auch ein signifikanter Anstieg der Bewertungskategorien C und D verzeichnet werden.
Dadurch kann die Forschungsfrage 7 beantwortet werden: *Die Concept Maps haben sich bezüglich der Bewertungskategorien A, C, und D signifikant verändert.* Obwohl die Lehrkräfte sich für die zweite Concept Map aussprachen, zeigte die Analyse, dass die Kombination aus beiden Concept Maps auch für eine Rückmeldung dienen kann. Die Kritik der Lehrkräfte war, dass die erste Concept Map kaum Aussagekraft hat, da das Vorwissen nicht ausreichte, um die Concept Map adäquat zu bearbeiten. Dennoch zeigten die Analysen der Lernenden-Aussagen, dass auch die erste Concept Map hilfreich für das Verständnis war. Es wurde außerdem erwähnt, dass Concept Maps an unterschiedlichen Stellen in der Lerneinheit hilfreich seien, um den eigenen Lernfortschritt zu überprüfen. Daraus kann abgeleitet werden, dass bei dem Einsatz der automatischen Auswertung der ersten Concept Map die Darstellung des Feedbacks angepasst werden muss. Es könnte bereits ausreichend sein, den Lehrkräften in einer kurzen Übersicht die wichtigsten Informationen bereitzustellen und erst bei einem späteren Zeitpunkt detaillierte Informationen präsentieren. Für das Lernenden-Feedback sollten beide Auswertungen beibehalten werden, da die eingesetzte Concept Map einen Lernfortschritt sichtbar machen kann. Eine andere Überlegung wäre, die erste Concept Map erst in der Mitte der Lerneinheit einzusetzen. Durch diese Verschiebung können die Lernenden mehr Inhalte lernen, was zu einer besseren Bearbeitung der Concept Maps führen kann. Ob dies ausreichen würde, die Kritik der Lehrkräfte abzuschwächen, kann an dieser Stelle nicht beantwortet werden.

8 Zusammenführende Diskussion

Durch die Entwicklung des Machine-Learning-Modells in der Entwicklungsstudie und dessen Einsatz in der Feedbackstudie konnten vielfältige Erkenntnisse gesammelt werden.
Nach den Schlüsselstrategien des formativen Assessments von Wiliam und Thompson (2008) wurde durch die Concept Map eine Möglichkeit der Erfassung des Lernstands geschaffen und durch die automatische Auswertung und des Feedback-Tools lernförderliche Rückmeldungen erstellt. Durch die feste Struktur der Concept Map kann ebenfalls gesagt werden, dass die Erfolgskriterien verständlich formuliert waren, da durch die vorgegebenen Propositionen ersichtlich wurde, welche Erfolgskriterien erreicht werden sollten. Allerdings muss kritisch diskutiert werden, dass diese Erfolgskriterien für alle Lernenden identisch waren. Es war kaum möglich zu differenzieren und für die Lernenden individuelle und spezifische Kriterien zu entwickeln, sodass jeder Lernende die Concept Map bestmöglich nutzen kann. In Zukunft könnte demnach an einer Concept Map gearbeitet werden, die eine Differenzierung zulässt.
In der Entwicklungsstudie wurde als ein Kriterium für den Einsatz des Machine-Learning-Modells die Fehlerrate ($False_A$-Rate) bezüglich der Bewertungskategorie A, also dass eine Proposition eines Lernenden fälschlicherweise als korrekt vorhergesagt wurde, bestimmt. Damit sollte sichergestellt werden, dass sich keine falschen Vorstellungen verfestigen und keine Glaubwürdigkeitsprobleme für die Lehrkräfte entstehen.
Die Ergebnisse der Feedbackstudie legen nahe, dass Lernende primär die von dem Machine-Learning-Modell als falsch bewerteten Propositionen betrachten. Eine Fehlklassifikation bei einer Proposition, die von dem Modell in die Kategorien B, C oder D bewertet wurde, wurde von keinem der Lernenden angemerkt. Dieses Ergebnis konnte auch durch eine Aussage in einem Interview gestützt werden.
Da die Lernenden die richtigen Propositionen nicht betrachteten, muss sichergestellt werden, dass das Modell hier wenig Fehler macht. Dies bestärkt noch einmal die Betrachtung der $False_A$-Rate bei der Entwicklung eines Machine-Learning-Modells. Daher sollten zukünftige Forschungsarbeiten darauf abzielen,

eine niedrige $False_A$-Rate zu erreichen.
An dieser Stelle soll die Problematik mit Propositionen der *Formel*-Gruppe erneut aufgegriffen werden. Eine Anpassung an das Feedback-Tool könnte erfolgen, indem das Tool darüber informiert, wie sicher die automatische Auswertung ist. Auf diese Weise könnte der Lehrkraft signalisiert werden, dass es sich um eine unsichere automatische Auswertung, z. B. durch das Auftreten einer Formel, handelt, die ihr Eingreifen erfordert. Dies könnte eine Möglichkeit sein, das Machine-Learning-Modell und das Feedback-Tool weiter zu verbessern und eine niedrige $False_A$-Rate zu erzielen.

Das genutzte Bewertungsschema (siehe Abschnitt 6.1) wurde entwickelt, um eine Grundlage für das automatische Feedback zu haben. Man sollte jedoch bedenken, dass jede Form der Kategorisierung inhärente Einschränkungen aufweist und nicht alle möglichen Nuancen vollständig abbilden kann. Dies kann zu einem Informationsverlust führen. Obwohl das Schema aus älteren Arbeiten adaptiert wurde, bleibt es unvermeidlich, dass bestimmte Propositionen möglicherweise nicht klar in eine der vorgegebenen Kategorien passen, da die Kategorien nicht vollständig trennscharf sind. Es ist wichtig, sich dieser Einschränkungen bewusst zu sein, da sie unmittelbar Auswirkung auf die Entwicklung des Machine-Learning-Modells hat. Je objektiver und trennschärfer das Bewertungsschema ist, desto besser können die Ergebnisse des Machine-Learning-Modells analysiert und interpretiert werden. Zukünftige Forschungen könnten darauf abzielen, das Schema weiter zu verfeinern oder alternative Methoden zu erforschen. Eine potenzielle Möglichkeit könnten Ansätze aus dem unüberwachten maschinellen Lernen wie Clustering sein. Durch Clustering-Verfahren könnten Muster in den Daten gefunden werden, die für eine Entwicklung eines Bewertungsschemas oder sogar für direkte Rückmeldungen genutzt werden können.
Die eingesetzte Concept Map und das Bewertungsschema wurden in der Entwicklungsstudie mit dem Ziel konzipiert, diagnostische Informationen bereitzustellen, die im weiteren Lehr-Lern-Prozess genutzt werden können. Dieser Ansatz wurde daher nicht vor dem Hintergrund einer Messung von Lernzuwachs konzipiert. Concept Maps wurden allerdings in vielen Arbeiten bereits erfolgreich als Diagnoseinstrument eingesetzt (z. B. Ley, 2015; Ruiz-Primo, 2004). Die Ergebnisse der Feedbackstudie zeigten, dass auch mit der konzipierten Concept Map und dem entwickelten Bewertungsschema eine Veränderung zwischen den Erhebungszeitpunkten nachgewiesen werden kann. Zwischen den beiden Erhebungszeitpunkten

wurde der Physikunterricht regulär fortgeführt. Man kann also davon ausgehen, dass die Lernenden innerhalb dieser Woche neue Inhalte gelernt haben und daher in der Lage waren die zweite Concept Map besser zu bearbeiten. Insbesondere die signifikante Abnahme der falschen Propositionen und der signifikante Anstieg der Bewertungskategorien C und D kann auf einen Lernzuwachs hindeuten.
Die Konstanz der Bewertungskategorie B kann auf das Bewertungsschema zurückgeführt werden. Schon in der Entwicklungsstudie ist aufgefallen, dass die vier Bewertungskategorien nicht gleich verteilt sind und die Kategorie B den kleinsten Teil im Datensatz ausmacht. Allerdings ist zu berücksichtigen, dass die Verschiebung der Bewertungskategorien auch durch Erinnerungseffekte bedingt sein kann, da die Lernenden dieselbe Concept Map an beiden Zeitpunkten bearbeiteten. Es könnte außerdem sein, dass die Veränderung auf eine gesteigerte Concept-Map-Kompetenz zurückzuführen ist. Der Großteil der Lernenden hatte im Vorfeld nicht mit einer Concept Map gearbeitet, was eine zusätzliche kognitive Herausforderung darstellen kann. Durch den erneuten Einsatz der Concept Map könnte die Bearbeitung weniger herausfordernd gewesen sein, was zu einer insgesamt besseren Bearbeitung geführt hat. Dies würde auch den Zuwachs der Propositionen im unteren Drittel der Concept Map erklären.
Insgesamt kann also nicht abschließend geklärt werden, ob die eingesetzte Concept Map und das Bewertungsschema als Diagnoseinstrument geeignet sind. Die Ergebnisse zeigen, dass das Potenzial vorhanden ist, obwohl genauer ausgearbeitet werden muss, ob die Veränderung durch ein größeres Zusammenhangwissen oder einer gesteigerten Bearbeitungskompetenz entstanden ist. Zukünftige Forschung, die auch eine größere Stichprobe verwenden sollte, könnte daher ein zusätzliches Testinstrument verwenden, das von einer Vergleichsgruppe bearbeitet werden muss. Die Ergebnisse und der Vergleich zwischen der Concept-Map-Gruppe und der Referenzgruppe können dann zur Validierung verwendet werden, um die Frage der Diagnosefähigkeit abschließend zu klären.

Der theoretische Hintergrund zum formativen Assessment hat gezeigt, dass formatives Assessment nicht zu den etablierten Routinen von Lehrkräften gehöre und sie gezielte Unterstützung benötigen, um das volle Potenzial von formativen Assessments ausschöpfen zu können (siehe Kapitel 2). Der Einsatz der automatischen Auswertung in der Feedbackstudie zeigte, dass das Potenzial des Feedback-Tools von den Lehrkräften kaum ausgeschöpft wurde und somit die Aussagen bestätigt werden können. Ein möglicher Grund dafür könnte die gewählte Concept Map

oder die exakte Umsetzung des Feedback-Tools sein. Es ist jedoch zu erkennen, dass weiterer Forschungsbedarf notwendig ist, um solche automatischen Feedback-Tools besser in den Schulalltag zu integrieren und damit auf die Bedürfnisse der Lehrkräfte abzustimmen. Diese Ergebnisse gelten möglicherweise nicht nur für die automatische Auswertung von Concept Maps, sondern für ein viel größeres Spektrum an KI-basierten Methoden. Die Entwicklungen auf diesem Gebiet werden immer schneller voranschreiten, weswegen es wichtig ist, sinnvolle und gezielte Methoden für den Einsatz im Bildungsbereich zu entwickeln und didaktisch zu erforschen. Eine gemeinsame Diskussion von Fachdidaktiken und Lehrkräften könnte die zielgerichtete Entwicklung von KI-basierten Systemen voranbringen und die Qualität und Fairness solcher Systeme erhöhen.
Die gewählte Concept Map war durch die Vorgaben recht geschlossen, was zum Verlust von didaktisch wichtigen Informationen führen kann. Es wurde sich bewusst für eine geschlossene Concept Map entschieden, um möglichst strukturell ähnliche Concept Maps zu gewinnen, die dann automatisch ausgewertet werden können. Die Ergebnisse aus der Entwicklungsstudie haben gezeigt, dass der gewählte Concept-Map-Ansatz für eine automatische Auswertung sehr gut geeignet war. Aus der Analyse der Interviews konnte allerdings die Notwendigkeit einer offeneren Concept Map erkannt werden. Die Lehrkräfte kritisierten die festen Strukturen und die nicht vorhandene Möglichkeit, die Concept Map an ihre Lerngruppe anpassen zu können. In einem nächsten Schritt könnte daher die automatische Auswertung einer offeneren Mechanik-Concept-Map getestet werden, bei der z. B. die in der vorliegenden Arbeit verwendeten Begriffe nur als Liste zur Verfügung gestellt werden. Dadurch müssten die Lernenden noch eigenständiger Propositionen bilden, was zu detaillierten Informationen über den aktuellen Wissensstand führen kann. Außerdem wurden die „Joker-Propositionen“ (siehe Abschnitt 6.1), die die Lernenden zusätzlich zu den festen Propositionen erstellt haben, nicht weiter betrachtet. Zukünftige Analysen können auch diese Propositionen aufgreifen und analysieren, um den Lehrkräften eine noch detaillierte Rückmeldung zu ermöglichen.
In diesem Zusammenhang kann auch der Einsatz von generativen Sprachmodellen wie ChatGPT diskutiert und geprüft werden. Diese Sprachmodelle könnten in der Lage sein, Concept Maps aus unterschiedlichen Themenfeldern und Strukturen automatisch zu bewerten und Rückmeldungen zu generieren. Die Verwendung solcher generativer KI muss jedoch kritisch diskutiert und reflektiert werden, da diese Systeme nicht für einen didaktischen Einsatz in der Schule konzipiert

wurden. Anhand dieser Ergebnisse kann überprüft werden, ob Feedback-Tools, die wie in der vorliegenden Arbeit gezielt mit didaktischem Hintergrund entwickelt wurden, notwendig sind oder ob auf große Sprachmodelle zurückgegriffen werden kann. Sprachmodelle sind zwar generalisierbarer, d. h. sie können für die Analyse verschiedener Concept Maps verwendet werden, aber sie sind oft nicht transparent, da z. B. Informationen über den Trainingsdatensatz fehlen (siehe Kapitel 4). Außerdem ist nicht direkt ersichtlich, was mit den Daten passiert, die in Systeme wie ChatGPT eingegeben werden (Sallam, 2023). Speziell bei Anwendungen im Bildungsbereich ist der Aspekt des Datenschutzes entscheiden, da oftmals sensible Informationen von Lernenden und Lehrkräften verarbeitet werden. Zudem ist ebenfalls nicht klar, wer letztlich die Verantwortung für die getätigten Aussagen oder Entscheidungen trägt (Niemi, 2021).
Daher gibt es immer mehr Open-Source-Modelle, die auf Plattformen wie Huggingface[8] zur Verfügung gestellt werden (T. Wolf et al., 2019). Diese Modelle bieten das Potenzial von generativen Sprachmodellen und weisen gleichzeitig eine höhere Transparenz auf, was die Nachvollziehbarkeit der Modelle deutlich vereinfachen kann. Daher können Open-Source-Modelle, speziell für den Bildungsbereich konzipiert werden, was ein interessanter Ansatz für weitere Forschung darstellt.

[8] https://huggingface.co/

9 Fazit und Ausblick

Im theoretischen Teil der vorliegenden Arbeit wurde zunächst das formative Assessment behandelt. Es konnte herausgestellt werden, dass das Hauptziel des formativen Assessments die Förderung und Unterstützung des Lehr-Lern-Prozesses darstellt. Es bietet Lehrkräften Einblicke in den Lernfortschritt der Lernenden und hilft ihnen, ihre Unterrichtsmethoden und -strategien zu verbessern, um die Lernergebnisse der Lernenden zu optimieren. Lernende können die bereitgestellten Informationen nutzen, um ihr eigenes Lernen zu reflektieren und zu adaptieren. Allerdings konnte ebenfalls gezeigt werden, dass formatives Assessment trotz der lernförderlichen Effekte kaum in Schulen eingesetzt wird. Als ein Hauptgrund konnte der enorme Zeitaufwand für die Lehrkräfte herausgearbeitet und Handlungsbedarf an weiterer Forschung erkannt werden.
Durch Concept Maps können wertvolle Informationen über den Lernfortschritt der Lernenden erhoben werden, weswegen sie für die Erhebung des Lernstandes ausgewählt wurden. Die Vielzahl von diagnostischen Informationen, die durch die Bearbeitung einer Concept Map entstehen, wurde in dieser Arbeit durch maschinelle Lernverfahren ausgewertet. Daher erfolgte im theoretischen Teil der Arbeit außerdem eine ausführliche Erörterung von Machine-Learning-Modellen und dessen Einsatz im Bildungskontext.
Ziel des empirischen Teils der vorliegenden Arbeit war die Entwicklung eines Machine-Learning-Modells zur automatischen Auswertung von Concept Maps und dessen Einsatz als Feedback-Tool in Schulen. Zunächst wurde in der Entwicklungsstudie eine Concept Map zum Thema Mechanik und ein vierstufiges Bewertungsschema ausgearbeitet, welches die Basis für eine lernförderliche Rückmeldung war. Das zentrale Ergebnis der ersten Teilstudie war, dass mit der SVM ein Machine-Learning-Modell entwickelt werden konnte, das die Anforderungen für den Einsatz als Feedback-Tool in Schulen erfüllte. Dieses Feedback-Tool wurde anschließend an zwei Zeitpunkten in verschiedenen Schulen eingesetzt, um Lehrkräften und Lernenden automatische Rückmeldungen zu den bearbeiteten Concept Maps bereit zustellen, die für den weiteren Lern-Lehr-Prozess genutzt werden konnten. Durch Fragebögen und Interviews konnte herausgefunden wer-

den, dass die Vorteile der automatischen Auswertung erkannt wurden, jedoch das vollständige Potenzial nicht ausgeschöpft werden konnte. Die Ergebnisse sprachen dafür, dass Lehrkräfte vor allem Rückmeldung auf Klassenebene und nicht auf individueller Lernenden-Ebene nutzten. Die Ergebnisse demonstrierten jedoch auch, dass Lehrkräfte nicht substituierbar sind. Aus den Interviews ging zudem hervor, dass formative Assessments nicht zu den alltäglichen Routinen der teilnehmenden Lehrkräfte gehören und sich dies durch die automatische Auswertung nicht wesentlich geändert hat. Insbesondere dieses Ergebnis eröffnet die weiterführende Frage, welche Weiterbildungsmaßnahmen für Lehrkräfte entwickelt werden können und sollten, um den Einsatz (automatischer) formativer Assessments im Klassenzimmer zu etablieren.

Im Rahmen der Arbeit konnten zahlreiche Ergebnisse auf theoretischer und empirischer Ebene zusammengetragen werden. Mit der Entwicklung und Evaluation des Machine-Learning-Modells wurde gezeigt, dass das Potenzial von KI-basierten Systemen für den Bildungsbereich vorhanden ist. Durch die Synergie zwischen menschlicher Lehrkraft und computergestützter Hilfe können in Zukunft viele unterschiedliche Lernende profitieren. An dieser Stelle muss jedoch nochmals auf den explorativen Charakter insbesondere der Feedbackstudie hingewiesen werden. Die vorliegende Arbeit kann daher keine allgemeingültigen Ergebnisse liefern. Vielmehr dienen die Ergebnisse als Anregung für weitere Forschung, um die Potenziale und Risiken KI-basierter Systeme im Bildungsbereich auszuloten. Künftige Forschung wird sich mit der Frage befassen müssen, wie KI-basierte Systeme in den Lehrplan und in das Schulsystem integriert werden können. Es ist von entscheidender Bedeutung, dass Bildungseinrichtungen Strategien entwickeln, die eine sinnvolle und effektive Einführung dieser Tools ermöglichen. Dies beinhaltet nicht nur die oben erwähnten Weiterbildungsmaßnahmen für Lehrkräfte, sondern auch die Entwicklung von Richtlinien für einen verantwortungsvollen Umgang. Es muss sichergestellt werden, dass die Privatsphäre der Lernenden und Lehrkräfte geschützt wird und die Daten verantwortungsvoll und transparent behandelt werden. Diesbezüglich ist eine offene Diskussion über ethische und rechtliche Fragen im Kontext des Einsatzes von KI im Bildungsbereich erforderlich.
Die präsentierten Ergebnisse werfen zudem weiterführende Fragen hinsichtlich der langfristigen Auswirkungen auf. Es konnte gezeigt werden, dass der kurzzeitige Einsatz zumindest teilweise zu einer Verbesserung der Lehr-Lern-Situation beitragen konnte. Es bleibt daher offen, inwiefern KI-basierte Systeme auch eine

langfristige Auswirkung haben. Diese Erkenntnis sollte zum Anlass genommen werden, um zu überprüfen, wie sich die regelmäßige Nutzung KI-basierter Systeme auf das Lernverhalten und die Leistung der Lernenden auswirkt und welche Anpassungen möglicherweise erforderlich sind, um ihre langfristige Wirksamkeit zu maximieren.
Insgesamt lässt sich aus den Ergebnissen der vorliegenden Arbeit ableiten, dass in Zukunft eine verstärkte Zusammenarbeit zwischen den Disziplinen Informatik, Fachdidaktik und Bildungsinstitutionen notwendig sein wird, um die Integration von KI-basierten Systemen in den Lehrplan zu optimieren und einen Mehrwert für alle Beteiligten zu schaffen.

Literaturverzeichnis

Abubakar, H. D., & Umar, M. (2022). Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, *4*(1&2), 27–33. https://doi.org/10.56471/slujst.v4i.266

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Andrade, H. (2010). Students as the Definitive Source of Formative Assessment: Academic Self-Assessment and the Self-Regulation of Learning. *NERA Conference Proceedings 2010*, (Paper 25). http://digitalcommons.uconn.edu/nera_2010/25

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers. (2nd ed.)* San Francisco, CA: Jossey-Bass.

Anohina, A., & Grundspenkis, J. (2009). Scoring concept maps. In B. Rachev & A. Smrikarov (Hrsg.), *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing - CompSysTech '09* (S. 1). ACM Press. https://doi.org/10.1145/1731740.1731824

Anohina-Naumeca, A. (2015). Justifying the usage of concept mapping as a tool for the formative assessment of the structural knowledge of engineering students. In *Knowledge Management & E-Learning* (S. 56–72, Bd. 7(1)). https://doi.org/10.34105/j.kmel.2015.07.005

Armano, G., Fanni, F., & Giuliani, A. (2015). Stopwords Identification by Means of Characteristic and Discriminant Analysis. *Proceedings of the International Conference on Agents and Artificial Intelligence*, 353–360. https://doi.org/10.5220/0005194303530360

Baker, R. S. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. https://doi.org/10.5281/ZENODO.3554745

Banerjee, P. (2020). *A Guide on XGBoost hyperparameters tuning*. Zugriff 19. Januar 2023 unter https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-like Events. *Review of Educational Research*, *61*(2), 213. https://doi.org/10.2307/1170535

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The Role of Pre-processing in Twitter Sentiment Analysis. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, D.-S. Huang, K.-H. Jo & L. Wang (Hrsg.), *Intelligent Computing Methodologies* (S. 615–624, Bd. 8589). Springer International Publishing. https://doi.org/10.1007/978-3-319-09339-0_62

Becker, L. B. (2022). *Die Kunst, Concept Mapping zu trainieren und einzusetzen: Ein Vergleich unterschiedlicher Ansätze und ihre Bedeutung für kognitive Prozesse, kognitive Belastung und Lernleistung* [Diss., Universität zu Köln]. Mathematisch-Naturwissenschaftlichen Fakultät.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Birunda, S. S., & Devi, R. K. (2021). A Review on Word Embedding Techniques for Text Classification. In J. S. Raj, A. M. Iliyasu, R. Bestak & Z. A. Baig (Hrsg.), *Innovative Data Communication Technologies and Application* (S. 267–281, Bd. 59). Springer Singapore; Imprint Springer. https://doi.org/10.1007/978-981-15-9651-3_23

Black, P. (2015). Formative assessment - an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 161–177. https://doi.org/10.1080/0969594x.2014.999643

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bleckmann, T., & Friege, G. (2023). Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method. *Knowledge Management & E-Learning: An International Journal*, 433–447. https://doi.org/10.34105/j.kmel.2023.15.025

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (2014). Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. *Journal of the Learning Sciences*, *23*(4), 561–599. https://doi.org/10.1080/10508406.2014.954750

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. https://doi.org/10.48550/arXiv.1607.04606

Bojorquez, H., & Vega, M. M. (2023). *The Importance of Artificial Intelligence in Education for All Students*. Zugriff 12. April 2024 unter https://www.idra.org/resource-center/the-importance-of-artificial-intelligence-in-education-for-all-students

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. https://doi.org/10.48550/arXiv.1607.06520

Bonaccorso, G. (2017). *Machine learning algorithms: A reference guide to popular algorithms for data science and machine learning*. Packt.

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12793

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery.

Buholzer, A., Baer, M., Zulliger, S., Torchetti, L., Ruelmann, M., Häfliger, A., & Lötscher, H. (2020). Formatives Assessment im alltäglichen Mathematikunterricht von Primarlehrpersonen: Häufigkeit, Dauer und Qualität. *Unterrichtswissenschaft*, *48*(4), 629–661. https://doi.org/10.1007/s42010-020-00083-7

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction*, *3*(4), 966–989. https://doi.org/10.3390/make3040048

Buldu, M., & Buldu, N. (2010). Concept mapping as a formative assessment in college classrooms: Measuring usefulness and student satisfaction. *Procedia - Social and Behavioral Sciences*, *2*(2), 2099–2104. https://doi.org/10.1016/j.sbspro.2010.03.288

Bürgermeister, A., Klieme, E., Rakoczy, K., Harks, B., & Blum, W. (2014). Formative Leistungsbeurteilung im Unterricht: Konzepte, Praxisberichte und ein neues Diagnoseinstrument für das Fach Mathematik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 41–60). Hogrefe.

Burke, D. (2009). Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education*, *34*(1), 41–50. https://doi.org/10.1080/02602930801895711

Burkov, A. (2020). *Machine learning engineering*. True Positive Inc.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cañas, A. J., Coffey, J., Carnot, M. J., Feltovich, P., Hoffmann, R. R., Feltovich, J., & Novak, J. D. (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support*. The Institute for Human and Machine Cognition. https://www.ihmc.us/users/acanas/Publications/ConceptMapLitReview/IHMC%20Lit

Cañas, A. J., Reiska, P., & Shvaikovsky, O. (2023). Improving learning and understanding through concept mapping. *Knowledge Management & E-Learning: An International Journal*, 369–380. https://doi.org/10.34105/j.kmel.2023.15.021

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., Eskridge, T. C., Arroyo, M., & Carvajal, R. (2004). CMAPTOOLS: A KNOWLEDGE MODELING AND SHARING ENVIRONMENT. In A. J. Cañas, J. D. Novak & F. M. González (Hrsg.), *Concept Maps: Theory, Methodology, Technology* (S. 13–20).

Cañas, A. J., Novak, J. D., & Reiska, P. (2012). FREEDOM VS. RESTRICTION OF CONTENT AND STRUCTURE DURING CONCEPT MAPPING - POSSIBILITIES AND LIMITATIONS FOR CONSTRUCTION AND ASSESSMENT. In A. J. Cañas, J. D. Novak & J. Vanhear (Hrsg.), *Concept maps* (S. 247–257). University of Malta.

Cañas, A. J., Novak, J. D., & Vanhear, J. (Hrsg.). (2012). *Concept maps: Theory, methodology, technology : proceedings of the Fifth International Conference on Concept Mapping, Volume 2*. University of Malta.

Cavalcanti, P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Cavalcanti, P., Ferreira Leite de Mello, R., Rolim, V., Andre, M., Freitas, F., & Gasevic, D. (2019). An Analysis of the use of Good Feedback Practices in Online Learning Courses. *2019 IEEE 19th International Conference on*

Advanced Learning Technologies (ICALT), 153–157. https://doi.org/10.1109/ICALT.2019.00061

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

Chan, B., Schweter, S., & Möller, T. (2020). German's Next Language Model. http://arxiv.org/pdf/2010.10906v4

Chang, K.-E., Sung, Y.-T., & Chen, I.-D. (2002). The Effect of Concept Mapping to Enhance Text Comprehension and Summarization. *The Journal of Experimental Education*, *71*(1), 5–23. https://doi.org/10.1080/00220970209602054

Chase, J. A., & Houmanfar, R. (2009). The Differential Effects of Elaborate Feedback and Basic Feedback on Student Performance in a Modified, Personalized System of Instruction Course. *Journal of Behavioral Education*, *18*(3), 245–265. https://doi.org/10.1007/s10864-009-9089-2

Cheuk, T. (2021). Can AI be racist? Color–evasiveness in the application of machine learning to science assessments. *Science Education*, *105*(5), 825–836. https://doi.org/10.1002/sce.21671

Clariana, R. B. (1993). A review of multiple-try feedback in traditional and computer-based instruction. In *Journal of Computer-Based Instruction* (S. 67–74).

Clariana, R. B., & Koul, R. (2005). Multiple-Try Feedback and Higher-Order Learning Outcomes. *International Journal of Instructional Media*, *32*(3). https://www.learntechlib.org/p/63335

Clariana, R. B., Wagner, D., & Roher Murphy, L. C. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, *48*(3), 5–22. https://doi.org/10.1007/BF02319855

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Erlbaum. http://www.loc.gov/catdir/enhancements/fy0731/88012110-d.html

Conlon, T. (2004). 'BUT IS OUR CONCEPT MAP ANY GOOD?': CLASSROOM EXPERIENCES WITH THE REASONABLE FALLIBLE ANALYSER. In A. J. Cañas, J. D. Novak & F. M. González (Hrsg.), *Concept Maps: Theory, Methodology, Technology* (S. 159–166).

Daley, B. J., Cañas, A. J., & Stark-Schweitzer, T. (2007). CmapTools: Integrating teaching, learning, and evaluation in online courses. *New Directions for Adult and Continuing Education*, *2007*(113), 37–47. https://doi.org/10.1002/ace.245

Data Basecamp. (2023a). *Was ist Gradient Boosting?* Zugriff 5. April 2024 unter https://databasecamp.de/ki/gradient-boosting

Data Basecamp. (2023b). *Was ist Word2Vec?* Zugriff 5. Februar 2024 unter https://databasecamp.de/ki/word2vec

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Ditton, H., & Müller, A. (2014). Einleitung. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen* (S. 7–10). Waxmann.

Dorn.Bader. (2018). *Physik Sek II Einführungsphase: Gymnasium Niedersachsen*. Westermann.

Dresing, T., & Pehl, T. (Hrsg.). (2011). *Praxisbuch Transkription: Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen* (2. Aufl.). Dr. Dresing und Pehl GmbH.

Dunn, K. E., & Mulvenon, S. W. (2009). A Critical Review of Research on Formative Assessments: The Limited Scientific Evidence of the Impact of Formative Assessments in Education. https://doi.org/10.7275/jg4h-rb87

Ertel, W. (2021). *Grundkurs Künstliche Intelligenz*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32075-1

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. https://doi.org/10.48550/arXiv.1909.00512

Ferilli, S., Esposito, F., & Grieco, D. (2014). Automatic Learning of Linguistic Resources for Stopword Removal and Stemming from Text. *Procedia*

Computer Science, *38*, 116–123. https://doi.org/10.1016/j.procs.2014.10.019

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fischler, H., & Peuckert, J. (Hrsg.). (2000). *Concept mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie*. Logos-Verl.

Fortmann-Roe, S. (2012). *Understanding the Bias-Variance Tradeoff*. Zugriff 16. Februar 2024 unter https://scott.fortmann-roe.com/docs/BiasVariance.html

Friege, G., & Lind, G. (2024). Materialien aus dem DFG-Projekt "Wissenszentriertes Problemlösen in Physik"(LI 829/1-3). Persönliche Mitteilung.

Friege, G. (2001). *Wissen und Problemlösen - eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*. Logos Verl.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Gouli, E., Gogoulou, A., Papanikolaou, K., & Grigoriadon, M. (2004). COMPASS: AN ADAPTIVE WEB-BASED CONCEPT MAP ASSESSMENT TOOL. In A. J. Cañas, J. D. Novak & F. M. González (Hrsg.), *Concept Maps: Theory, Methodology, Technology* (S. 295–302).

Hammann, M., & Jördens, J. (2014). Offene Aufgaben codieren. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 169–178). Springer Berlin Heidelberg.

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining*. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Harlen, W. (2008). Editior´s introduction. In W. Harlen (Hrsg.), *Student Assessment and Testing* (S. xix–xlvi).

Harper, A., & Kayumova, S. (2023). Invisible multilingual Black and Brown girls: Raciolinguistic narratives of identity in science education. *Journal of Research in Science Teaching*, *60*(5), 1092–1124. https://doi.org/10.1002/tea.21826

Hartmeyer, R., Stevenson, M. P., & Bentsen, P. (2018). A systematic review of concept mapping-based formative assessment processes in primary and secondary science education. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 598–619. https://doi.org/10.1080/0969594X.2017.1377685

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/b94608

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (First published 2009). Routledge Taylor & Francis Group.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hattie, J., & Wollenschläger, M. (2014). A conceptualization of feedback. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen* (S. 135–149). Waxmann.

Heidloff, N. (2023). *Foundation Models, Transformers, BERT and GPT*. Zugriff 20. April 2024 unter https://heidloff.net/article/foundation-models-transformers-bert-and-gpt/

Herding, D., Zimmermann, M., Bescherer, C., & Schroeder, U. (2010). Entwicklung eines Frameworks für semi-automatisches Feedback zur Unterstützung bei Lernprozessen. In *DeLFI 2010 - 8. Tagung der Fachgruppe E-Learning der Gesellschaft für Informatik e.V* (S. 145–156). Gesellschaft für Informatik e.V.

Heritage, M. (2007). Formative Assessment: What Do Teachers Need to Know and Do? *Phi Delta Kappan*, *89*(2), 140–145. https://doi.org/10.1177/003172170708900210

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the Academically Unmotivated: A Critical Issue for the 21st Century. *Review of Educational Research*, *70*(2), 151–179. https://doi.org/10.3102/00346543070002151

Hirschle, J. (2022). *Deep Natural Language Processing: Einstieg in Word Embedding, Sequence-to-Sequence-Modelle und Transformer mit Python*. Hanser. 10.3139/9783446473904

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining &*

Knowledge Management Process, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Hunt, E., & Pellegrino, J. (2002). Issues, Examples, and Challenges in Formative Assessment. *New Directions for Teaching and Learning*, *2002*, 73–85. https://doi.org/10.1002/tl.48

Hwang, G.-.-J., Shi, Y.-.-R., & Chu, H.-.-C. (2011). A concept map approach to developing collaborative Mindtools for context–aware ubiquitous learning. *British Journal of Educational Technology*, *42*(5), 778–789. https://doi.org/10.1111/j.1467-8535.2010.01102.x

Ifenthaler, D. (2010). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*, *58*(1), 81–97. https://doi.org/10.1007/s11423-008-9087-4

Impulse Physik. (2018). *Einführungsphase 11: Gymnasium Niedersachsen*. Ernst Klett Verlag GmbH.

Jain, A. (2016). *Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python*. Zugriff 19. Januar 2023 unter https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

Jain, G. P., Gurupur, V. P., Schroeder, J. L., & Faulkenberry, E. D. (2014). Artificial Intelligence-Based Student Learning Evaluation: A Concept Map-Based Approach for Analyzing a Student's Understanding of a Topic. *IEEE Transactions on Learning Technologies*, *7*(3), 267–279. https://doi.org/10.1109/TLT.2014.2330297

Kayhan Moharreri, Minsu Ha & Ross H Nehm. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, *7*(1). https://doi.org/10.1186/s12052-014-0015-2

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, *42*(1), 43–57. https://doi.org/10.1080/001318800363908

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kingston, N., & Nash, B. (2015). Erratum. *Educational Measurement: Issues and Practice*, *34*(2), 55. https://doi.org/10.1111/emip.12075

Kitzmann, A. (2022). *Künstliche Intelligenz*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-37700-7

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Krist, C., & Kubsch, M. (2023). Bias, bias everywhere: A response to Li et al. and Zhai and Nehm. *Journal of Research in Science Teaching*, *60*(10), 2395–2399. https://doi.org/10.1002/tea.21913

Kroeze, K., van den Berg, S., Veldkamp, B., & de Jong, T. (2021). Automated Assessment of and Feedback on Concept Maps during Inquiry Learning. *IEEE Transactions on Learning Technologies*, 1. https://doi.org/10.1109/TLT.2021.3103331

Krüger, D., & Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *26*(1), 157–172. https://doi.org/10.1007/s40573-020-00118-7

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2023). Unreflected Acceptance – Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education. http://arxiv.org/pdf/2309.03087v1

Kubin, D. (o. D.). *Studien zur Verbesserung der Physikausbildung in der Medizin. Unveröffentlichte Ergebnisse*. MHH und AG Physikdidaktik an der Leibniz Universität Hannover.

Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press LLC. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5838737

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. https://doi.org/10.1007/BF01320096

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09889-7

Lenhard, W., Baier, H., Endlich, D., Lenhard, A., Schneider, W., & Hoffmann, J. (2012). Computerunterstützte Leseverständnisförderung: Die Effekte automatisch generierter Rückmeldungen. *Zeitschrift für Pädagogische Psychologie*, *26*(2), 135–148. https://doi.org/10.1024/1010-0652/a000066

Ley, S. L. (2015). *Concept Maps als Diagnoseinstrument im Physikunterricht und deren Auswirkung auf die Diagnosegenauigkeit von Physiklehrkräften* [Diss.]. https://duepublico.uni-due.de/servlets/DocumentServlet?id=38141

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, *60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

Liang, W., Wang, L., She, J., & Liu, Y. (2022). Detecting Resource Release Bugs with Analogical Reasoning. *Scientific Programming*, *2022*, 1–9. https://doi.org/10.1155/2022/3518673

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 136–140.

Lim, L.-A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., & Dawson, S. (2021). What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, *72*, 101202. https://doi.org/10.1016/j.learninstruc.2019.04.003

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Mahesh, B. (2019). Machine Learning Algorithms - A Review. https://doi.org/10.21275/ART20203995

Maier, U. (2010). Formative Assessment - Ein erfolgversprechendes Konzept zur Reform von Unterricht und Leistungsmessung? *Zeitschrift für Erziehungswissenschaft*, *13*(2), 293–308. https://doi.org/10.1007/s11618-010-0124-9

Manning, C. D., Raghavan, P., & Schütze, H. (2012). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Martin, M., Alvarez, A., Ruiz, S., Fernandez-Castro, I., & Urretavizcaya, M. (2009). Helping Teachers to Track Student Evolution in a B-Learning Environment. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, 342–346. https://doi.org/10.1109/ICALT.2009.152

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*(2), 312–320. https://doi.org/10.1037/0022-0663.90.2.312

Mayer, R. E., & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (13., überarbeitete Auflage). Beltz. https://www.content-select.com/index.php?id=bib_view&ean=9783407258991

McKendree, J. (1990). Effective Feedback Content for Tutoring Complex Skills. *Human-Computer Interaction*, *5*(4), 381–413. https://doi.org/10.1207/s15327051hci0504_2

McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning*, *33*(6), 562–574. https://doi.org/10.1111/jcal.12200

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In L. Vanderwende, H. Daumé III & K. Kirchhoff (Hrsg.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (S. 746–751). Association for Computational Linguistics. https://aclanthology.org/N13-1090

Mintzes, J. J., Canas, A., Coffey, J., Gorman, J., Gurley, L., Hoffman, R., McGuire, S. Y., Miller, N., Moon, B., Trifone, J., & Wandersee, J. H. (2011). Comment on Retrieval practice produces more learning than elaborative studying with concept mapping. *Science (New York, N.Y.)*, *334*(6055), 453, author reply 453. https://doi.org/10.1126/science.1203698

Moreno, R. (2004). Decreasing Cognitive Load for Novice Students: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science*, *32*(1/2), 99–113. https://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Müller, A., & Ditton, H. (2014). Feedback: Begriff, Formen und Funktionen. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen* (S. 11–28). Waxmann.

Nakamura, C. M., Murphy, S. K., Christel, M. G., Stevens, S. M., & Zollman, D. A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Physical Review Physics Education Research*, *12*(1). https://doi.org/10.1103/PhysRevPhysEducRes.12.010122

Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Bd. 56). Waxmann.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector (Hrsg.), *Handbook of research on educational communications and technology* (S. 125–143). Lawrence Erlbaum Associates.

Narciss, S. (2014). Modelle zu den Bedingungen und Wirkungen von Feedback in Lehr-Lernsituationen. In H. Ditton (Hrsg.), *Feedback und Rückmeldungen* (S. 43–82). Waxmann.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, *16*(4), 310–322. https://doi.org/10.1016/j.learninstruc.2006.07.003

National Research Council. (2000). *How People Learn Brain, Mind, Experience, and School: Expanded Edition*. National Academies Press. https://doi.org/10.17226/9853

Niedersächsisches Kultusministerium. (2015). Kerncurriculum für das Gymnasium Schuljahrgänge 5-10: Naturwissenschaften - Sekundarstufe I. http://www.cuvo.nibis.de

Niedersächsisches Kultusministerium. (2022a). Auswertung Zentralabitur 2022 - Fächer (Niedersachsen). Zugriff 12. August 2023 unter https://www.nibis.de/uploads/mk-bolhoefer/2022/3_Abitur-Auswertung_2022_Faecher-Nds.pdf

Niedersächsisches Kultusministerium. (2022b). Kerncurriculum für das Gymnasium – gymnasiale Oberstufe: Physik. http://www.cuvo.nibis.de

Niemi, H. (2021). AI in learning. *Journal of Pacific Rim Psychology*, *15*. https://doi.org/10.1177/18344909211038105

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Novak, J. D., & Cañas, A. J. (2008). The Theory Underlying Concept Maps and How to Construct and Use Them. http://cmap.ihmc.us/Publications/

Novak, J. D., & Gowin, D. B. (1984). *Learning How to Learn*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173469

Novak, J. D., & Cañas, A. J. (2006). The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool. *Information Visualization*, *5*(3), 175–184. https://doi.org/10.1057/palgrave.ivs.9500126

Nunnery, J. A., Ross, S. M., & McDonald, A. (2006). A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6. *Journal of*

Education for Students Placed at Risk (JESPAR), *11*(1), 1–18. https://doi.org/10.1207/s15327671espr1101_1

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128–138. https://doi.org/10.1111/bjet.12592

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pellegrino, J. W., Chudosky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. National Academies Press. https://doi.org/10.17226/10019

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In Q. C. R. I. Alessandro Moschitti, G. Bo Pang & U. o. A. Walter Daelemans (Hrsg.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (S. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pfannstiel, M. A. (2022). *Künstliche Intelligenz im Gesundheitswesen*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-33597-7

Platen, P. v. (2020). *Transformer-based Encoder-Decoder Models*. Zugriff 12. Februar 2024 unter https://huggingface.co/blog/encoder-decoder

Plaue, M. (2021). *Data Science*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-63489-9

Plomer, M. (2011). *Physik physiologisch passend praktiziert: eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*. Logos Verlag Berlin.

Pohl, A. (n. V.). *KI und formatives Assessment – eine Analyse von Interviews mit Lehrkräften und Schüler:innenantworten* [Masterarbeit]. Leibniz Universität Hannover. IDMP.

Prince, S. J. D. (2023). *Understanding Deep Learning*. The MIT Press. http://udlbook.com

Raghav, P. (2019). *Understanding NLP Word Embeddings — Text Vectorization*. Zugriff 31. Januar 2024 unter https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223

Rajput, B. S., & Khare, N. (2015). A survey of Stemming Algorithms for Information Retrieval. In *International Organization of Scientific Research – Journal of Computer Engineering* (S. 76–80, Bd. 17 (3)).

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Richter, S. (2019). *Statistisches und maschinelles Lernen*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-59354-7

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Ruiz-Primo, M. A. (2000). *On the use of concept maps as an assessment tool in science: What we have learned so far*. https://www.redalyc.org/pdf/155/15502103.pdf

Ruiz-Primo, M. A. (2004). Examining concept maps as an assessment tool: Concept Maps: Theory, Methodology, Technology, Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain (September 14-17, 2004), Editorial Universidad Pública de Navarra.

Ruiz-Primo, M. A., Schulz, S. E., Li, M., & Shavelson, R. J. (1998). *Comparison of the Reliability and Validity of Scores from Two Concept-Mapping Techniques.* Educational Resources Information Center. https://files.eric.ed.gov/fulltext/ED422378.pdf

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*(6), 569–600. https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1<3.0.CO;2-M

Ruiz-Primo, M. A., & Shavelson, R. J. (1997). *Concept-Map Based Assessment: On Possible Sources of Sampling Variability*. Educational Resources Information Center. https://api.semanticscholar.org/CorpusID:55360020

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the Validity of Cognitive Interpretations of Scores From Alternative Concept-Mapping Techniques. *Educational Assessment*, *7*(2), 99–141. https://doi.org/10.1207/S15326977EA0702_2

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary educational psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryoo, K., & Linn, M. C. (2016). Designing automated guidance for concept diagrams in inquiry instruction. *Journal of Research in Science Teaching*, *53*(7), 1003–1035. https://doi.org/10.1002/tea.21321

Ryssel, J. (2012). Die Lernwirksamkeit von einfachem und elaboriertem Feedback in Verbindung mit dem Erstellen von Concept Maps im Planspielunterricht. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Berufs- und wirtschaftspädagogische Analysen: Aktuelle Forschungen zur beruflichen Bildung* (S. 89–101). Verlag Barbara Budrich.

Ryssel, J. (2018). *Feedback zu Concept Maps im betriebswirtschaftlichen Planspielunterricht – eine empirische Untersuchung* [Dissertation]. Technische Universität Dresden.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, *11*(6). https://doi.org/10.3390/healthcare11060887

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Schmidt, C. (2020). *Formatives Assessment in der Grundschule: Konzept, Einschätzungen der Lehrkräfte und Zusammenhänge* (1st ed. 2020). Springer Fach-

medien Wiesbaden; Imprint: Springer VS. https://doi.org/10.1007/978-3-658-26921-0

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort - Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*(4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Scriven, M. (1967). The methodology of evaluation. *Perspectives of curriculum evaluation, ed. R.W. Tyler, R.M. Gagne, and M. Scriven*, 39–83. Chicago, IL: Rand McNally.

Seldon, A. (2018). *Are Robots Really Set to Replace Teachers? Sir Anthony Seldon Discusses AI and Agency*. Zugriff 16. April 2024 unter https://www.teachwire.net/news/are-robots-really-set-to-replace-teachers-sir-anthony-seldon-discusses-ai-and-agency/

Sghir, N., Adadi, A., & Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: A decade systematic review (2012-2022). *Education and information technologies*, 1–35. https://doi.org/10.1007/s10639-022-11536-0

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. *Applied Measurement in Education*, *21*(4), 295–314. https://doi.org/10.1080/08957340802347647

Shi, J., Bian, J., Richter, J., Chen, K.-H., Rahnenführer, J., Xiong, H., & Chen, J.-J. (2021). MODES: model-based optimization on distributed embedded systems. *Machine Learning*, *110*(6), 1527–1547. https://doi.org/10.1007/s10994-021-06014-6

Shin, D., & Shim, J. (2020). A Systematic Review on Data Mining for Mathematics and Science Education. *International Journal of Science and Mathematics Education*. https://doi.org/10.1007/s10763-020-10085-7

Souvignier, E., & Hasselhorn, M. (2018). Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*, 693–696. https://doi.org/10.1007/s11618-018-0839-6

Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2023). Harnessing Large Language Models to Enhance Self-Regulated Learning via Formative Feedback. http://arxiv.org/pdf/2311.13984v2

Stracke, I. (2004). *Einsatz computerbasierter concept maps zur Wissensdiagnose in der Chemie: empirische Untersuchungen am Beispiel des chemischen Gleichgewichts* [Dissertation]. Kiel.

Strautmane, M. (2012). Concept map-based knowledge assessment tasks and their scoring criteria: An overview. In A. J. Cañas, J. D. Novak & J. Vanhear (Hrsg.), *Concept Maps: Theory, Methodology, Technology*. https://cmc.ihmc.us/cmc2012papers/cmc2012-p113.pdf

Thevapalan, A. (2021). *The Machine Learning Workflow Explained (and How You Can Practice It Now)*. Zugriff 29. Januar 2024 unter https://towardsdatascience.com/the-machine-learning-workflow-explained-557abf882079

Torres-Moreno, J.-M. (2014). *Automatic Text Summarization*. ISTE; Wiley. http://www.loc.gov/catdir/enhancements/fy1413/2014947781-d.html

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

Trumpower, D. L., & Sarwar, G. S. (2010). Formative Structural Assessment: Using Concept Maps as Assessment for Learning. In J. Sánchez, A. J. Cañas & J. D. Novak (Hrsg.), *Fourth International Conference on Concept Mapping* (S. 132–136).

Uttamchandani, S., & Quick, J. (2022). An Introduction to Fairness, Absence of Bias, and Equity in Learning Analytics. In C. Lang, G. Siemens, A. F. Wise, D. Gašević & A. Merceron (Hrsg.), *The Handbook of Learning Analytics* (S. 205–212). SoLAR. https://doi.org/0.18608/hla22.020

Vanides, J., Yin, Y., Tomita, M., & Ruiz-Primo, M. A. (2005). *Using concept maps in the science classroom*. Science Scope, 28, 27-31.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Vittorini, P., Menini, S., & Tonelli, S. (2020). An AI-Based System for Formative and Summative Assessment in Data Science Courses. *International Jour-*

nal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-020-00230-2

Wadouh, J. (2007). *Vernetzung und kumulatives Lernen im Biologieunterricht der Gymnasialklasse 9.* [Dissertation]. Universität Duisburg-Essen. Fakultät für Biologie. https://d-nb.info/993499937/34

Wallach, H. M. (2006). Topic modeling. In W. Cohen & A. Moore (Hrsg.), *Proceedings of the 23rd international conference on Machine learning - ICML '06* (S. 977–984). ACM Press. https://doi.org/10.1145/1143844.1143967

Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, *31*(3), 379–394. https://doi.org/10.1080/02602930500353061

Werner, M. (2021). Flache Neuronale Netze für die Klassifizierung. In M. Werner (Hrsg.), *Digitale Bildverarbeitung* (S. 349–382). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-22185-0_12

Wiliam, D. (2010). An Integrative Summary of the Research Literature and Implications for a New Theory of Formative Assessment. In H. Andrade & G. J. Cizek (Hrsg.), *Handbook of Formative Assessment*. Routledge.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment: In J. H. McMillan (Hrsg.), Formative classroom assessment: Theory into practice (S. 29–42). New York: Teachers College Columbia University.

Wiliam, D., & Thompson, M. (2008). Integrating Assessment with Learning: What Will It Take to Make It Work? In C. A. Dwyer (Hrsg.), *The Future of Assessment: Shaping teaching and learning* (S. 53–82). Routledge. https://doi.org/10.4324/9781315086545-3

Wolf, N. (2014). *Formative Leistungsmessung im naturwissenschaftlichen Unterricht* [Diss., Pädagogischen Hochschule Schwäbisch Gmünd]. https://phsg.bsz-bw.de/frontdoor/deliver/index/docId/14/file/Dissertation_FORMAL_Nicole_Wolf_Digital_Ohne_externe_Links.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. v., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. http://arxiv.org/pdf/1910.03771v5

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*, *30*(1), 1–15. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2021). Stärkung praxisorientierter Hochschullehre durch computerbasierte Rückmeldung zu Reflexionstexten. *die hochschullehre*, *7*. https://doi.org/10.3278/HSL2111W

Wylie, C., & Lyon, C. (2016). Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice (Revised) (Washington: FAST, SCASS & CCSSO., Hrsg.). Zugriff 26. Mai 2023 unter https://center.ncsu.edu/ncfalcon/pluginfile.php/2/course/section/57/FAROP_Revised_2016.pdf

Yao, L., Cahill, A., & McCaffrey, D. F. (2020). The Impact of Training Data Quality on Automated Content Scoring Performance. https://api.semanticscholar.org/CorpusID:210999398

Yeh, S. S. (2006). High-Stakes Testing: Can Rapid Assessment Reduce the Pressure? *Teachers College Record: The Voice of Scholarship in Education*, *108*(4), 621–661. https://doi.org/10.1111/j.1467-9620.2006.00663.x

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2), 166–184. https://doi.org/10.1002/tea.20049

Zaman, A. N. K., Matsakis, P., & Brown, C. (2011). Evaluation of stop word lists in text retrieval using Latent Semantic Indexing. *2011 Sixth International Conference on Digital Information Management*, 133–136. https://doi.org/10.1109/ICDIM.2011.6093315

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0171-0

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing Teachers' Competencies to Read and Interpret Graphs from Learning Progress Assessment: Results

from Tests and Interviews. *Learning Disabilities Research & Practice*, *32*(1), 61–70. https://doi.org/10.1111/ldrp.12126

Zhai, X., Krajcik, J., & Pellegrino, J. (2021). On the Validity of Machine Learning-based Next Generation Science Assessments: A Validity Inferential Network. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09879-9

Zhai, X., Yin, Y., Pellegrino, J., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Anhang

A - Concept Map Bewertung

	Kategorie	Beschreibung	Schlüsselwörter	Beispiele
A	Falsch	Physikalisch falsche Relation		Masse → *„hat Einfluss“* → Freier Fall
B	Oberbegriff, charakteristisches Merkmal, einfache Zuordnung und Zusammenhänge	Zuordnung von Beispielen oder Eigenschaften. Beschreibung von einfachen, ungerichteten Zusammenhängen	„ist ein“, „hat“, „gehört zu“, „besitzt“, „hat zu tun mit“, „benötigt xy zum berechnen“, „verändert sich“	Gleichmäßig beschleunigte Bewegung → *„ist ein“* → Freier Fall Waagerechter Wurf → *„besitzt eine“* → Gleichförmige Bewegung Geschwindigkeit → *„ändert sich“* → Freier Fall
C	Oberflächliche Abhängigkeit Einfache **gerichtete** Zusammenhänge	Oberflächliche physikalische Abhängigkeiten zw. den Begriffen Zusammenhänge sind gerichtet beschrieben.	„steigt/sinkt“ „nimmt zu/ab“ „x hängt von y ab“	Gleichmäßig beschleunigte Bewegung → „steigt“ → Zeit-Geschwindigkeit-Graph Geschwindigkeit → *„nimmt zu bei der“* → Gleichmäßig beschleunigte Bewegung
D	**Detaillierte** Zusammenhänge und Abhängigkeiten Klare Funktionsbeziehung	Kann als Formel oder verbaler Zusammenhang formuliert sein. Die Zusammenhänge sind detailliert beschrieben.	„je … desto…“, „steigt linear/exponentiell …“, „… proportional zu…“	Gleichmäßig beschleunigte Bewegung → *„v=a*t“* → Zeit-Geschwindigkeit-Graph Geschwindigkeit → „Linear ansteigender Graph“ → Gleichmäßig beschleunigte Bewegung

Props	Kat.	Beispielantwort
Beschleunigung - Freier Fall	D	ist konstant mit g = 9.81; a = g
	C	hat eine gleichmäßige; ändert sich nicht; von g abhängig, Ortsfaktor g = 9.81 **m/s^2**
	B	hat eine; mit Ortsfaktor g; 9.81
	A	lässt ansteigen, ist eine
Beschleunigung - Geschwindigkeit	D	Änderung dieser in einer bestimmten Zeitspanne
	C	wenn steigt/sinkt gibt es eine; steigert die
	B	verändert
	A	hat eine
Beschleunigung - Gleichförmige Bewegung	D	a = 0
	C	
	B	ist nicht vorhanden; null; Kostant
	A	hat eine
Beschleunigung - Gleichmäßig beschleunigte Bewegung	D	a = const; a = v/t
	C	ändert sich nicht; ist konstant/gleichmäßig
	B	hat eine; a ist nicht null
	A	ändert sich durch
Geschwindigkeit - Freier Fall	D	linear ansteigend
	C	wird immer größer beim
	B	ist vorhanden bei / verändert sich
	A	steigt exponentiell
Geschwindigkeit - Gleichförmige Bewegung	D	v = const; v = s/t
	C	ändert sich nicht; konstant
	B	hat eine
	A	verändert sich
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	D	steigt/sinkt linear ; gleichmäßig verändert; v = a * t
	C	wird größer / kleiner
	B	hat eine; verändert sich
	A	ist konstant

Gleichförmige Bewegung - Waagerechter Wurf	D	in X-Richtung
	C	Bewegung zur Seite
	B	besitzt eine
	A	ist keine
Gleichförmige Bewegung - Zeit-Beschleunigung-Graph	D	Graph liegt auf x-Achse
	C	gerader Verlauf; ist konstant
	B	lässt sich erstellen
	A	steigt konstant
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	D	verläuft parallel zur X-Achse; v = const
	C	ist konstant
	B	darstellbar; gerade
	A	v steigt exponnentiell
Gleichförmige Bewegung - Zeit-Weg-Graph	D	linear ansteigender Graph; gleiche/konstante Steigung
	C	steigt an
	B	ist darstellbar als
	A	Verläuft parallel zur X-Achse
Gleichmäßig beschleunigte Bewegung - Freier Fall	D	s = 1/2 *a*t^2
	C	v abhängig von t; mit a=g; Ortsfaktor 9.81 m/s^2
	B	ist ein Beispiel für
	A	ist keine
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	D	in Y-Richtung
	C	Bewegung nach unten
	B	besitzt eine
	A	ist keine
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigung-Graph	D	verläuft parallel zur X-Achse; a = const.
	C	ist konstant
	B	darstellbar; gerade
	A	steigt exponnentiell

Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	D	eine lineare Funktion; konstanten Steigung
	C	steigt an; steigt an; ansteigende Gerade, {nur} konstant/linear
	B	ist darstellbar als
	A	Verläuft parallel zur X-Achse
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	D	steigt quadratisch
	C	nimmt zu
	B	lässt sich erstellen; der Graph verläuft ungleichmäßig
	A	steigt exponentiell
Kraft - Beschleunigung	D	F = m * a; mal Masse ist
	C	beeinflusst; a hängt ab von F
	B	ein Faktor von; Masse
	A	ist eine kraft
Kraft - Geschwindigkeit	D	wenn keine Kraft, dann v = 0 oder v = const
	C	erhöht
	B	verändert
	A	ist eine
Masse - Freier Fall	D	egal wie schwer der Körper alle fallen gleich schnell
	C	hat keinen Einfluss auf
	B	fällt im
	A	beeinflusst

B - Fragebögen

Fragebogen der Entwicklungsstudie

Frage	Antwortmöglichkeit
1. Persönliche Angaben	
Alter	offenes Item
Geschlecht	männlich, weiblich, divers
2. Noten	
Welche Zensur hattest du in 11.1 in ...	
Mathematik, Deutsch, Physik	offenes Item
Welche Zensur hattest du in der 10. Klasse in ...	
Mathematik, Deutsch, Physik	offenes Item
3. Physikunterricht	
Welchen Physik-Kurs wirst du voraussichtlich in der Oberstufe besuchen?	Leistungskurs Physik, Grundkurs Physik, gar keinen Physikkurs
4. Concept Maps	
Was kann man deiner Meinung nach die Concept verbessern (z. B. technische Aspekte, Dauer, Vorgaben in der Concept Map etc.)?	offenes Item
Bist du der Meinung, dass man Concept Maps im Unterricht öfter einsetzen sollte? Bitte begründe deine Antwort (z. B. wenn ja: wie, wann, warum ...)!	offenes Item
Wie sollte deiner Meinung nach deine fertige Concept Map ausgewertet und im Unterricht weiter verwendet werden?	offenes Item
Möchtest du noch etwas ergänzen?	offenes Item

Fragebogen 1 der Feedbackstudie

Skala Feedbackstudie: trifft nicht zu, trifft etwas zu, trifft ziemlich zu, trifft voll zu

1. Persönliche Angaben	
Alter	offenes Item
Geschlecht	männlich, weiblich, divers
2. Noten	
Welche Zensur hattest du in der 10. Klasse in ...	
Mathematik, Deutsch, Physik	Notenskala 1-6

3. Concept Maps	
Ich habe zum ersten Mal mit einer Concept Map gearbeitet.	Skala Feedbackstudie
Das Erstellen der Concept Map am Computer hat mir Spaß gemacht.	Skala Feedbackstudie
Bist du der Meinung, dass man Concept Maps im Unterricht öfter einsetzen sollte? Bitte begründe deine Antwort (z. B. wenn ja: wie,wann, warum ...)!	offenes Item
Die Handhabung der Lernplattform und der Concept Map war mit intuitiv klar.	Skala Feedbackstudie
Es hat mir keine Probleme bereitet, mit der Lernplattform und der Concept Map zu arbeiten.	Skala Feedbackstudie
Für das Erstellen der Concept Map hätte ich gerne mehr Zeit zur Verfügung gehabt.	Skala Feedbackstudie
Was kann man deiner Meinung nach verbessern (z. B. technische Aspekte, Vorgaben in der Concept Map, etc.)?	offenes Item
Mir ist es schwergefallen, inhaltlich passende Beschriftungen für die einzelnen Pfeile zu finden.	Skala Feedbackstudie
Durch das Erstellen der Concept Map sind mir eigene Wissenslücken bewusst geworden.	Skala Feedbackstudie
Wie sollte deiner Meinung nach deine fertige Concept Map ausgewertet und im Unterricht weiter verwendet werden?	offenes Item
4. Abschluss	
Möchtest du noch etwas ergänzen?	offenes Item

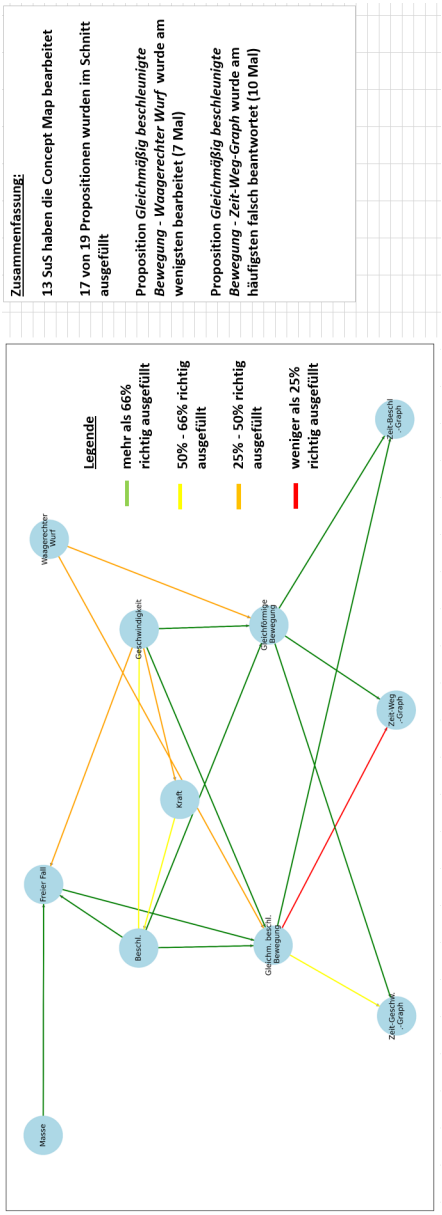
Fragebogen 2 der Feedbackstudie

1. Concept Maps	
Das Erstellen der zweiten Concept Map ist mir leichter gefallen als das erste Mal.	Skala Feedbackstudie
Ich konnte einen Lernfortschritt gegenüber der ersten Concept Map erkennen.	Skala Feedbackstudie
Mir ist es schwergefallen, inhaltlich passende Beschriftungen für die einzelnen Pfeile zu finden.	Skala Feedbackstudie
Das Erstellen der zweiten Concept Map hat mir Spaß gemacht.	Skala Feedbackstudie
Für das Erstellen der zweiten Concept Map hätte ich gerne mehr Zeit zur Verfügung gehabt.	Skala Feedbackstudie
2. Abschluss	
Möchtest du noch etwas ergänzen?	offenes Item

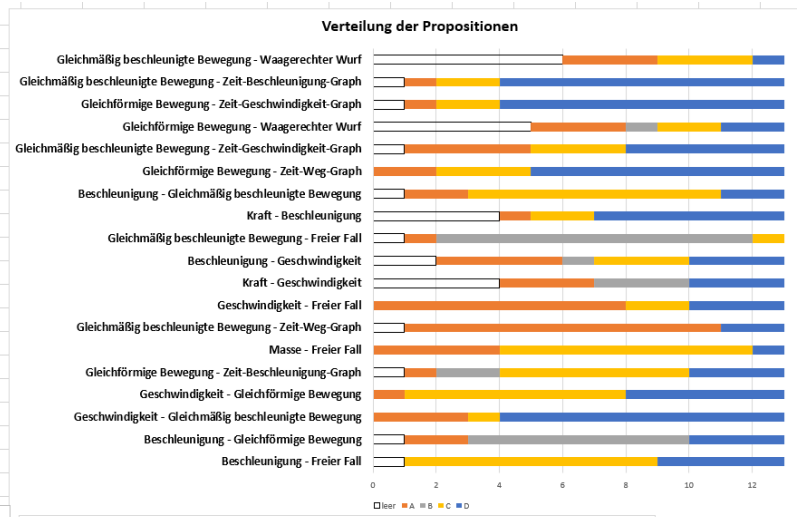
Fragebogen 3 der Feedbackstudie

1. Wirksamkeit	
Durch die automatische Rückmeldung ...	
... kann ich präzisere Beschriftungen bilden.	Skala Feedbackstudie
... wurde ich angeregt, über andere Beschriftungen nachzudenken.	Skala Feedbackstudie
... konnte ich meine Fehler nachvollziehen.	Skala Feedbackstudie
2. Nützlichkeit	
Die automatische Rückmeldung hat mir Einblick über mein Wissen in der Mechanik gegeben.	Skala Feedbackstudie
Durch die automatische Rückmeldung habe ich nützliche Hinweise für mein Lernen (z. B. für Klausuren) bekommen.	Skala Feedbackstudie
Ich wünsche mir häufiger derartige Rückmeldungen im Physikunterricht.	Skala Feedbackstudie
Hast du die automatische Rückmeldung als nützlich empfunden? Bitte begründe deine Antwort.	offenes Item
Welche Veränderungen oder Verbesserungen würdest du dir für die automatische Rückmeldug wünschen?	offenes Item
3. Persönlichkeit	
Ich finde, dass die automatische Rückmeldung ...	
... gut auf meine Stärken in den Concept Maps eingeht.	Skala Feedbackstudie
... gut auf meine Schwächen in den Concept Maps eingeht.	Skala Feedbackstudie
4. Richtigkeit	
Ich bin der Meinung, dass ...	
... die automatische Rückmeldung meine Beschriftungen richtig identifiziert hat.	Skala Feedbackstudie
... die automatische Rückmeldung nicht mehr Fehler macht als eine menschliche Rückmeldung.	Skala Feedbackstudie
18. Abschluss	
Möchtest du noch etwas ergänzen?	offenes Item

C - Automatisches Feedback für die Lehrenden

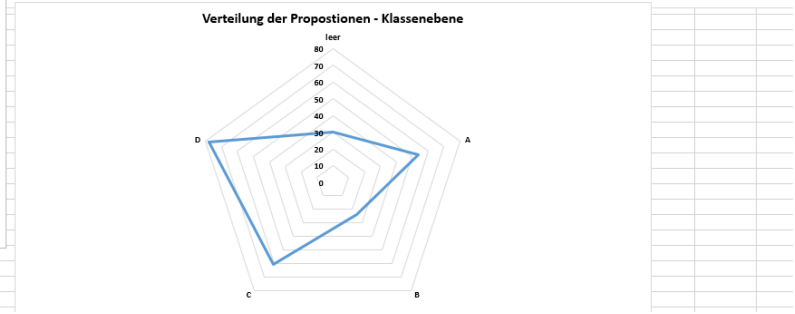


Props	leer	A	B	C	D
Beschleunigung - Freier Fall	1	0	0	8	4
Beschleunigung - Gleichförmige Beweg	1	2	7	0	3
Geschwindigkeit - Gleichmäßig beschle	0	3	0	1	9
Geschwindigkeit - Gleichförmige Bewe	0	1	0	7	5
Gleichförmige Bewegung - Zeit-Beschle	1	1	2	6	3
Masse - Freier Fall	0	4	0	8	1
Gleichmäßig beschleunigte Bewegung	1	10	0	0	2
Geschwindigkeit - Freier Fall	0	8	0	2	3
Kraft - Geschwindigkeit	4	3	3	0	3
Beschleunigung - Geschwindigkeit	2	4	1	3	3
Gleichmäßig beschleunigte Bewegung	1	1	10	1	0
Kraft - Beschleunigung	4	1	0	2	6
Beschleunigung - Gleichmäßig beschle	1	2	0	8	2
Gleichförmige Bewegung - Zeit-Weg-Gr	0	2	0	3	8
Gleichmäßig beschleunigte Bewegung	1	4	0	3	5
Gleichförmige Bewegung - Waagerecht	5	3	1	2	2
Gleichförmige Bewegung - Zeit-Geschw	1	1	0	2	9
Gleichmäßig beschleunigte Bewegung	1	1	0	2	9
Gleichmäßig beschleunigte Bewegung	6	3	0	3	1
SUMME	**30**	**54**	**24**	**61**	**78**



Hier sehen Sie eine Analyse der Propositionen bezogen auf die vier Ratingkategorien "falsch, A, B , C". Die beiden Grafiken sollen Ihnen Informationen über das Vorwissen liefern.

Zur Erinnerung:
A - einfache Zusammenhänge
B - gerichtete Zusammenhänge
C - detailierte Zusammenhänge & richtige Formeln

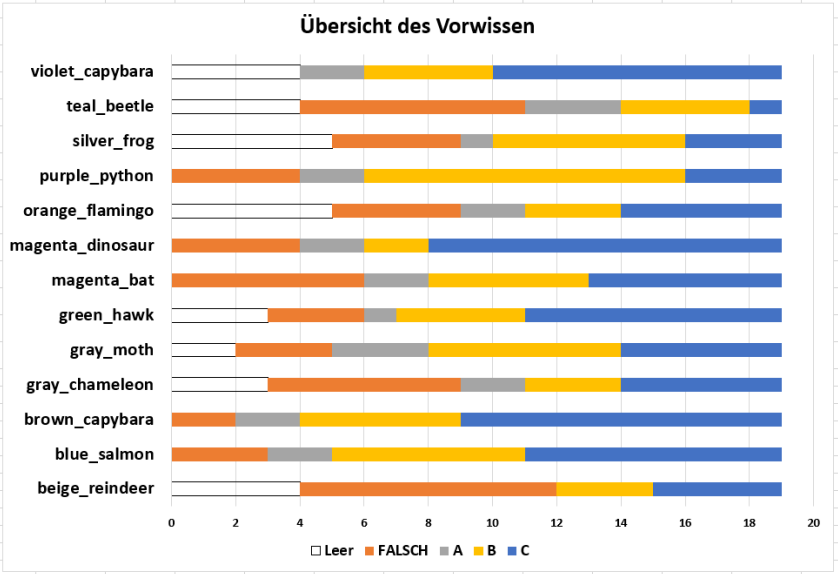


SuS_ID	Leer	FALSCH	A	B	C
beige_reindeer	4	8	0	3	4
blue_salmon	0	3	2	6	8
brown_capybara	0	2	2	5	10
gray_chameleon	3	6	2	3	5
gray_moth	2	3	3	6	5
green_hawk	3	3	1	4	8
magenta_bat	0	6	2	5	6
magenta_dinosaur	0	4	2	2	11
orange_flamingo	5	4	2	3	5
purple_python	0	4	2	10	3
silver_frog	5	4	1	6	3
teal_beetle	4	7	3	4	1
violet_capybara	4	0	2	4	9

Hier sehen Sie eine Analyse Ihrer SuS bezogen auf die vier Ratingkategorien "falsch, A, B , C". Die beiden Grafiken sollen Ihnen Informationen über das Vorwissen liefern.

Zur Erinnerung:
A - einfache Zusammenhänge
B - gerichtete Zusammenhänge
C - detailierte Zusammenhänge & richtige Formeln

Student at risk: 0 SuS



"Problematic Propositions" sind Propositionen, die mehr als 66 % falsch oder gar nicht beantwortet wurden.
Die Zahlenreihenfolge entsprechen der Reihenfolge der Kategorien: A = 1, B = 2, C = 3. Leere Antworten sind hier nicht aufgelistet.

Hier sehen Sie die dazu gehörigen SuS-Antworten. Eine Übersicht der Antworten für jede Proposition finden Sie unter "Alle Antworten"

Proposition	**leer o. falsch**	**richtig**		**Proposition**	**leer o. falsch**	**richtig**
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	**11**	**2**		*Gleichmäßig beschleunigte Bewegung -*	**9**	**4**
SuS-Antwort	**Bewertung**			**SuS-Antwort**	**Bewertung**	
1:2at^2+v0•t•s0	**0**			wird gleichmäßig beschleunigt	**0**	
Exponentiell	**0**			vertikaler Vektor	**2**	
quadratisch	**3**			In y-Richtung	**3**	
quadratische Funktion	**3**			Vertikal	**2**	
Exponentiell	**0**			vertikaler Richtung	**2**	
s(t)=1/2at2+v0t+s0	**0**			/	**0**	
s(t)=½•at²+v0t+s0	**0**			y(t)=-12*g*t2+h	**0**	
(1/2)•a•t^2+V0•t+S0	**0**					
s(t) = 1/2*a*t^2+v0*t+s	**0**					
s(t)=1/2*a*t^2+v0*t+s0	**0**					
Exponentiell, s(t)=1/2at^2+v0t+s0	**0**					
Geschwindigkeit und Zeit	**0**					

Hier sehen Sie alle Antworten Ihrer SuS zu den 19 Propostionen der Concept Map.
Die Antwort "empty" und die dazugehörige Bewertung "-1" bedeutet, dass keine Antwort eingetragen. Die Zahlenreihenfolge entsprechen der Reihenfolge der Kategorien: A = 1, B = 2, C = 3.

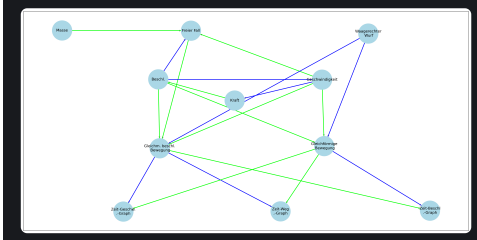
SuS_ID	Proposition	SuS-Antwort	Bewertung
magenta_bat	Beschleunigung - Freier Fall	beschleunigt konstant	2
green_hawk	Beschleunigung - Freier Fall	a=9,81m/s2	3
purple_python	Beschleunigung - Freier Fall	a=9,81m/s^2	3
violet_capybara	Beschleunigung - Freier Fall	ist konstant beim	2
brown_capybara	Beschleunigung - Freier Fall	9,81 m/s^2	2
silver_frog	Beschleunigung - Freier Fall	9,81m/s^2	2
blue_salmon	Beschleunigung - Freier Fall	a=g=9,81m/s2	3
gray_moth	Beschleunigung - Freier Fall	9,81m/s²	2
gray_chameleon	Beschleunigung - Freier Fall	Gleichmäßige Beschleunigung	2
orange_flamingo	Beschleunigung - Freier Fall	empty	-1
magenta_dinosaur	Beschleunigung - Freier Fall	a=g=9,81m/s^2	3
beige_reindeer	Beschleunigung - Freier Fall	Konstant	2
teal_beetle	Beschleunigung - Freier Fall	gleichbleibend	2
magenta_bat	Beschleunigung - Gleichförmige Bewegung	a=0	3
green_hawk	Beschleunigung - Gleichförmige Bewegung	empty	-1
purple_python	Beschleunigung - Gleichförmige Bewegung	Null	1
violet_capybara	Beschleunigung - Gleichförmige Bewegung	ist null	1
brown_capybara	Beschleunigung - Gleichförmige Bewegung	Konstant 0	1
silver_frog	Beschleunigung - Gleichförmige Bewegung	a=/0	3
blue_salmon	Beschleunigung - Gleichförmige Bewegung	keine Beschleunigung	1
gray_moth	Beschleunigung - Gleichförmige Bewegung	keine	1
gray_chameleon	Beschleunigung - Gleichförmige Bewegung	Keine Veränderung	1
orange_flamingo	Beschleunigung - Gleichförmige Bewegung	ist gleich 0	3
magenta_dinosaur	Beschleunigung - Gleichförmige Bewegung	a=v/t	0
beige_reindeer	Beschleunigung - Gleichförmige Bewegung	Formelzeichen a	0
teal_beetle	Beschleunigung - Gleichförmige Bewegung	0	1
magenta_bat	Geschwindigkeit - Gleichmäßig beschleunigte	Geschwindigkeit linear	3
green_hawk	Geschwindigkeit - Gleichmäßig beschleunigte	Verläuft linear ansteigend im 45* Grad Winkel	3
purple_python	Geschwindigkeit - Gleichmäßig beschleunigte	linear	2
violet_capybara	Geschwindigkeit - Gleichmäßig beschleunigte	nimmt gleichmäßig zu	3

D - Automatisches Feedback für die Lernenden

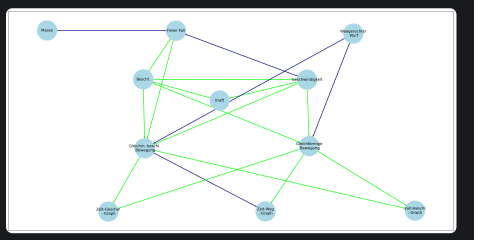
Einfache Feedbackgruppe

Feedback Seite 1

Concept Map 1



Concept Map 2



Grüne Linien

Diese Antworten waren bereits richtig.

Proposition	Deine Antwort
Kraft - Beschleunigung	Verändernde Kraft
Beschleunigung - Gleichmäßig beschleunigte Bewegung	Keine Veränderung
Gleichförmige Bewegung - Zeit-Weg-Graph	Linearer Graph
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	Glatte Linie
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigung-Graph	Immer verwendbar
Beschleunigung - Gleichförmige Bewegung	Keine Vorhanden
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	Gleichmäßig verändernd
Geschwindigkeit - Gleichförmige Bewegung	Gleich bleibend
Masse - Freier Fall	Gravitation
Geschwindigkeit - Freier Fall	9,81 m/s (ohne Widerstand)
Gleichmäßig beschleunigte Bewegung - Freier Fall	Ohne Luftwiderstand

Grüne Linien

Diese Antworten sind richtig.

Proposition	Deine Antwort
Beschleunigung - Freier Fall	9.81 m/s (Ohne Luftwiderstand)
Kraft - Beschleunigung	Desto mehr, desto höher
Beschleunigung - Gleichmäßig beschleunigte Bewegung	Wenn Konstant
Gleichförmige Bewegung - Zeit-Weg-Graph	v * t (Linear steigend)
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	a * t (Linear steigend)
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	V (Konstant)
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigungs-Graph	v/t (Konstant)
Beschleunigung - Gleichförmige Bewegung	Wenn nicht vorhanden
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	Wenn steigend mit der Zeit
Geschwindigkeit - Gleichförmige Bewegung	Wenn Konstant
Gleichförmige Bewegung - Zeit-Beschleunigungs-Graph	0 (Konstant)
Kraft - Geschwindigkeit	Desto mehr, desto höher
Beschleunigung - Geschwindigkeit	Erhöht mit der Zeit
Gleichmäßig beschleunigte Bewegung - Freier Fall	Ohne Luftwiderstand (9.81 m/s)

Blaue Linien

Diese Antworten waren leider nicht korrekt.

Proposition	Deine Antwort
Beschleunigung - Freier Fall	35 km/h die sekunde (ohne Widerstand)
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	Niedrige Beschleunigung
Gleichförmige Bewegung - Waagerechter Wurf	Wurf Mensch
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	Wurfkraft abhängig
Gleichförmige Bewegung - Zeit-Beschleunigung-Graph	Nicht notwendig
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	Hohe Beschleunigung
Kraft - Geschwindigkeit	Gleichbleibende Kraft
Beschleunigung - Geschwindigkeit	Je höher desto mehr

Blaue Linien

Diese Antworten sind leider nicht korrekt. Du findest auf der nächsten Seite eine Musterlösung, die du zum Vergleichen und zum Lernen nutzen kannst.

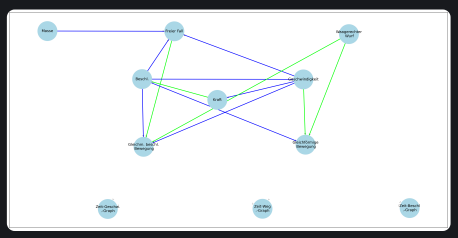
Proposition	Deine Antwort
Gleichförmige Bewegung - Waagerechter Wurf	Ohne Gravitation
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	Menschlicher Wurf auf der Erde
Masse - Freier Fall	In der Luft ohne Haltung
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	1/2a * t^2 (Exponentiell Steigend)
Geschwindigkeit - Freier Fall	Beschleunigt sich jede Sekunde

Weiter

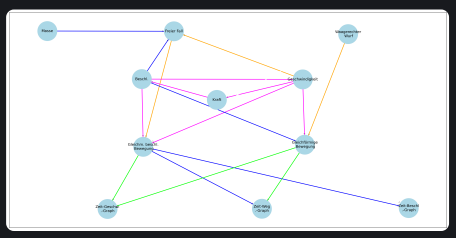
Elaborierte Feedbackgruppe

Feedback Seite 1

Concept Map 1



Concept Map 2



Grüne Linien

Diese Antworten waren bereits richtig

Proposition	Deine Antwort
Gleichförmige Bewegung - Waagerechter Wurf	hat zur Folge
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	ist eine
Gleichmäßig beschleunigte Bewegung - Freier Fall	findet statt in einer
Kraft - Beschleunigung	beeinflusst die
Geschwindigkeit - Gleichförmige Bewegung	ist konstant bei einer

Blaue Linien

Diese Antworten waren leider nicht korrekt.

Proposition	Deine Antwort
Beschleunigung - Freier Fall	nimmt zu
Beschleunigung - Gleichmäßig beschleunigte Bewegung	ist zu unterscheiden zur
Beschleunigung - Gleichförmige Bewegung	ist der Anfang zur
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	unterscheidet sich in der
Masse - Freier Fall	ist wichtig für den
Geschwindigkeit - Freier Fall	mesbar beim
Kraft - Geschwindigkeit	wird umgewandelt in
Beschleunigung - Geschwindigkeit	gehört zur

Grüne Linien

Ausgezeichnete Arbeit! Diese Antworten sind bereits sehr detailliert oder enthalten einen richtigen funktionalen Zusammenhang!

Proposition	Deine Antwort
Gleichförmige Bewegung - Zeit-Weg-Graph	ist konstant steigend auf dem
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	wächst die Geschwindigkeit
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	ist waagerecht auf dem

Magenta Linien

Diese Antworten sind richtig, jedoch noch zu oberflächlich. Versuche beim nächsten Mal genauer auf den Zusammenhang zwischen den beiden Begriffen einzugehen.

Proposition	Deine Antwort	Feedback Vorschlag
Kraft - Beschleunigung	wird benötigt bei der	F = m * a
Beschleunigung - Gleichmäßig beschleunigte Bewegung	wird immer höher bei der	a = const
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	steigt bei der	steigt/sinkt linear
Geschwindigkeit - Gleichförmige Bewegung	ist konstant bei einer	v=const
Kraft - Geschwindigkeit	erhöht die	wenn keine Kraft, dann v=0 oder v=const
Beschleunigung - Geschwindigkeit	eine Zunahme der	Änderung dieser in einer bestimmten Zeitspanne

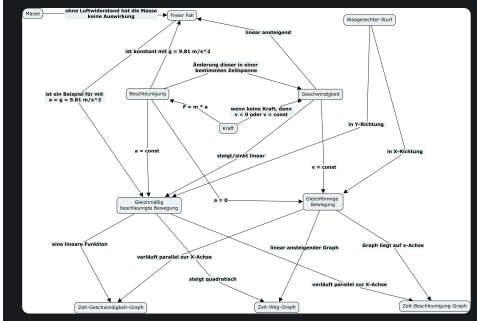
Orange Linien

Diese Antworten sind richtig, sie könnten allerdings noch präziser ausgedrückt werden. Du könntest mehr Details hinzufügen, z.B. wie genau sich etwas verhält oder einen

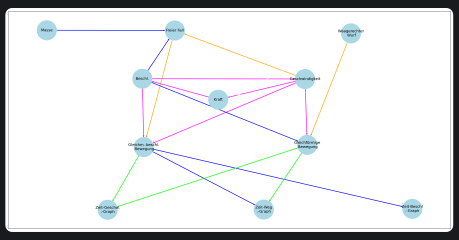
Zweite Feedback-Seite für beide Gruppen

Feedback Seite 2

Muster Concept Map



Deine Concept Map



Diese Antworten können Beispiele für eine Musterlösung sein.

Proposition	Vorschlag
Gleichförmige Bewegung - Waagerechter Wurf	in X-Richtung
Geschwindigkeit - Freier Fall	linear ansteigend
Gleichmäßig beschleunigte Bewegung - Freier Fall	ist ein Beispiel für ... mit a=g=9,81m/s^2
Kraft - Beschleunigung	F = m * a
Beschleunigung - Gleichmäßig beschleunigte Bewegung	a = const
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	steigt/sinkt linear
Geschwindigkeit - Gleichförmige Bewegung	v=const
Kraft - Geschwindigkeit	wenn keine Kraft, dann v=0 oder v=const
Beschleunigung - Geschwindigkeit	Änderung dieser in einer bestimmten Zeitspanne
Gleichförmige Bewegung - Zeit-Weg-Graph	linear ansteigender Graph
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	eine lineare Funktion
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	verläuft parallel zur X-Achse
Beschleunigung - Freier Fall	ist konstant mit g = 9.81 m/s^2
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigungs-Graph	verläuft parallel zur X-Achse
Beschleunigung - Gleichförmige Bewegung	a=0
Masse - Freier Fall	ohne Luftwiderstand hat die Masse keine Auswirkung
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	steigt quadratisch
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	in Y-Richtung
Gleichförmige Bewegung - Zeit-Beschleunigungs-Graph	Graph liegt auf x-Achse

Deine richtigen Antworten

Proposition	Deine Antwort	Fehler?
Gleichförmige Bewegung - Waagerechter Wurf	ist eine	☐
Geschwindigkeit - Freier Fall	hat eine zumessende	☐
Gleichmäßig beschleunigte Bewegung - Freier Fall	beginnt bei einer	☐
Kraft - Beschleunigung	wird benötigt bei der	☐
Beschleunigung - Gleichmäßig beschleunigte Bewegung	wird immer höher bei der	☐
Geschwindigkeit - Gleichmäßig beschleunigte Bewegung	steigt bei der	☐
Geschwindigkeit - Gleichförmige Bewegung	ist konstant bei einer	☐
Kraft - Geschwindigkeit	erhöht die	☐
Beschleunigung - Geschwindigkeit	eine Zunahme der	☐
Gleichförmige Bewegung - Zeit-Weg-Graph	ist konstant steigend auf dem	☐
Gleichmäßig beschleunigte Bewegung - Zeit-Geschwindigkeit-Graph	wächst die Geschwindigkeit	☐
Gleichförmige Bewegung - Zeit-Geschwindigkeit-Graph	ist waagerecht auf dem	☐

Deine falsche oder nicht ausgefüllten Antworten

Proposition	Deine Antwort	Fehler?
Beschleunigung - Freier Fall	wird beim längeren Fallweg im schneller	☐
Gleichmäßig beschleunigte Bewegung - Zeit-Beschleunigungs-Graph	ist eine gleichbleibende Steigung im	☐
Beschleunigung - Gleichförmige Bewegung	ist der Beginn der	☐
Masse - Freier Fall	verschnellert den	☐
Gleichmäßig beschleunigte Bewegung - Zeit-Weg-Graph	steigt die Länge des Weges immer weiter an	☐
Gleichmäßig beschleunigte Bewegung - Waagerechter Wurf	empty	☐
Gleichförmige Bewegung - Zeit-Beschleunigungs-Graph	empty	☐

E - Lernplattform *Intelligent Physics Trainer* (IPT)

Liebe Schülerinnen und Schüler,

mein Name ist Tom Bleckmann und ich bin Doktorand in der Physikdidaktik an der Leibniz Universität Hannover.

Für mein Promotionsprojekt benötige ich Eure Hilfe! Ich möchte gemeinsam mit Euch eine neue Feedback-Methode ausprobieren.

Dieses Feedback soll eine wertvolle Ergänzung zum Unterricht sein und Euren Lehrkräften helfen, Euch beim Lernen von Physik besser zu unterstützen. **Das bedeutet, dass ich Eure Leistung nicht benoten möchte, sondern Euch und Euren Lehrkräften nur hilfreiches Feedback zur Verfügung stellen möchte. Eure Teilnahme wird also keine negativen Auswirkungen für Euch haben!**

Das Feedback wird automatisch erstellt. Ich interessiere mich für die Qualität des Feedbacks und wie Ihr und Eure Lehrkräfte es nutzt. Ihr sollt zweimal eine Concept Map zum Thema Mechanik bearbeiten. Zunächst heute und später noch einmal kurz vor der Klausur.

Was genau Eure Aufgabe ist und wie überhaupt eine Concept Map aussieht, erfahrt Ihr auf den nächsten Seiten.

Vielen Dank für Eure Teilnahme und Unterstützung

Tom Bleckmann

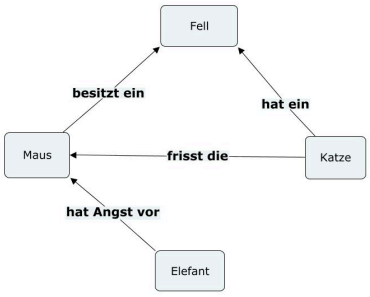
Weiter

In einer Concept Map werden Wissen und Inhalte strukturiert dargestellt. Dabei werden zusammengehörige Begriffe durch beschriftete Pfeile verbunden. So entsteht ein Netzwerk. Es ermöglicht einen einfachen Überblick über die Zusammenhänge eines Themengebiets.

Eine Concept Map besteht aus drei zentralen Elementen:

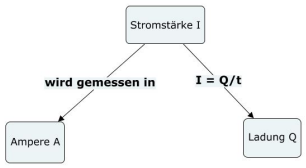
1. **Das Thema der Concept Map**
2. **Die zentralen Begriffe des Themas, die in Kästchen dargestellt werden**
3. **Die beschrifteten Pfeile, die eine sinnvolle Verknüpfung zwischen zwei Begriffen herstellen**

Ein Beispiel für eine Concept Map zum Thema **Tiere** könnte demnach so aussehen:



Wie Ihr in diesem Beispiel erkennen könnt, werden die einzelnen Begriffe (Fell, Maus, Katze, Elefant) mit beschrifteten Pfeilen („besitzt ein", „hat ein", etc.) verbunden, um so den Zusammenhang zwischen den Begriffen zu verdeutlichen. Dabei wird jeweils in Pfeilrichtung gelesen, z. B. „Maus – besitzt ein – Fell" und jeder Pfeil verbindet immer nur zwei Begriffe. Ihr seht ebenfalls, dass es in den meisten Fällen, mehr als ein passendes Verbindungswort gibt.

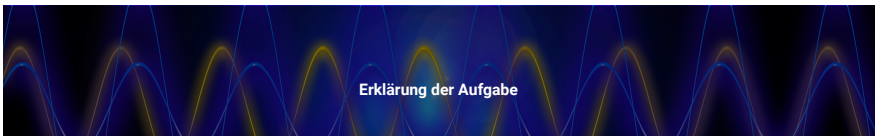
Auch für jedes Physikthema kann eine Concept Map erstellt werden, z. B. mit den Begriffen „Stromstärke I", „Ladung Q" und „Ampere A":



Wie Ihr erkennen könnt, bestehen die Verbindungen zwischen den Begriffen nicht aus langen Sätzen, sondern aus präzisen direkten Zusammenhängen, die im Optimalfall (physikalisch) richtig formuliert werden.

Man kann auch eine Formel als Verbindungswort nutzen, wie zwischen Stromstärke I und Ladung Q.

Weiter



Nun seid Ihr an der Reihe:

Eure Aufgabe wird es gleich sein, eine Concept Map zum **Thema Mechanik** zu bearbeiten.

Auf der nächsten Seite erscheint eine vorgefertigte Concept Map mit festgelegten Begriffen und Pfeilen, die sich auch nicht verschieben oder ändern lassen. **Eure Aufgabe ist es, die Verbindungswörter zwischen den Begriffen sinnvoll zu beschriften und in die passenden Lücken einzutragen**. Ihr sollt also die Zusammenhänge zwischen den einzelnen Mechanik-Begriffen herstellen. Wie Ihr in den beiden Beispielen gesehen habt, könnt Ihr kurze Sätze, einzelne Wörter oder Formeln nutzen.

Ihr habt für die Concept Map **30 Minuten Zeit.** Bitte bearbeitet die Concept Map in **Einzelarbeit**. Nur so kann ein individuelles Feedback erfolgen.

Nachdem Ihr fertig seid, könnt Ihr Eure Antworten speichern. Dazu findet Ihr einen Button im oberen rechten Bereich.

ACHTUNG: Ihr könnt Eure Antworten nur einmal speichern.

Falls Ihr eine Lücke nicht ausfüllen könnt, ist das natürlich kein Problem. Wie eingangs erwähnt, wird Eure Concept Map nicht benotet!

Anschließend gelangt Ihr noch zu einem kleinen Fragebogen, den Ihr bitte ebenfalls ausfüllt.

Wenn Du ein iPad benutzt und es noch nicht im Querformat hast, drehe es bitte jetzt, damit die Concept Map richtig angezeigt wird!

Viel Spaß 🙂

ZUR CONCEPT MAP MECHANIK

F - Interviewleitfaden aus der Masterarbeit Pohl (n. V.)

Einleitung

1. Für Teilnahmebereitschaft bedanken
2. Vorstellung (Antonia Pohl, 25 Jahre alt, Studierende M. Ed. Mathe und Physik für das Lehramt an Gymnasien, Interviews mit Lehrkräften für MA)
3. Erklärung des Vorgehens (ca. 15-20 Minuten Gespräch)
4. Erlaubnis für Audioaufnahme einholen, Erläuterung Datenschutz und Anonymisierung der Daten

Kurzfragebogen

1. An welcher Schulform unterrichten Sie?
2. Welche Fächer unterrichten Sie?
3. Seit wie vielen Jahren sind Sie Lehrer:in?

Befragung

Fragen	Mögliche Antworten	Bemerkungen/Steuerungsfragen
1. Allgemeine Einstellung zu KI im Bildungswesen Sie haben ein KI-generiertes Feedback zu den bearbeiteten Concept-Maps Ihrer SuS erhalten. Deswegen möchte ich zu Beginn erst einmal allgemein auf künstliche Intelligenz eingehen.		
1.1 Welche Gedanken haben Sie zur Nutzung von KI im Bildungswesen?	- Hilfreich: Möglichkeit, Unterricht individuell an SuS anzupassen	- Bei eher positiver Argumentation: Sehen Sie auch Hindernisse oder mögliche Probleme bei der Nutzung von KI?

	- Hilfreich: Entlastung der Lehrkraft, z. B. bei der Bewertung von schriftlichen Arbeiten oder bei der Planung von Unterricht - Mögliche Problematik: Datenschutz - Skepsis in Bezug auf Qualität der KI-Systeme	- Bei eher negativer Argumentation: Können Sie sich vorstellen, dass die Nutzung von KI-Systemen auch positive Aspekte haben könnte?
1.2 Nutzen Sie KI manchmal in Ihren Tätigkeiten als Lehrkraft?	nie/selten/häufig	Selten/häufig: - Was genau haben Sie genutzt? - Inwiefern? Wobei genau konnte KI entlasten/unterstützen? - Wie waren Ihre Erfahrungen diesbezüglich? Nie: - Auch nicht ChatGPT? (sonst Übersetzungstools, Grammatikkontrolle) - Aus was für Gründen noch nie genutzt?

2. Formatives Assessment Kommen wir nun noch einmal auf das Feedback zurück, was Sie erhalten haben. Feedback stellt einen wichtigen Aspekt von formativem Assessment dar.		
2.1 Kennen Sie den Begriff „Formatives Assessment“?	ja/nein	Wenn nein: Erklärung Formatives Assessment beschreibt die lernprozessbegleitende Erhebung von Schüler:innenleistungen zur Optimierung des Lehr-/Lernprozesses, nicht zur Leistungsmessung. Auf diesem Weg können mögliche Wissenslücken aufgedeckt und damit die Unterrichtsgestaltung angepasst werden. Das fortlaufende Feedback soll außerdem die SuS unterstützen, ihren Lernprozess zu verbessern.
2.2 Wenn ja: Wie würden Sie formatives Assessment definieren?	- Feedback während des Lernprozesses - Lernprozess optimieren - Abgrenzung von summativem Assessment - Verschiedene Umsetzungsmöglichkeiten - Rückmeldung für SuS	- Haben Sie Ideen, wie sich formatives Assessment im Unterricht umsetzen lässt? *-kurz halten-*
2.3 Wie oft nutzen Sie formatives Assessment in Ihrem Unterricht?	- häufig/selten/nie	- Beschreiben Sie, wie Sie formatives Assessment nutzen. - Individuell oder auf Gruppen-/Klassenebene?

2.4 Können Sie sich vorstellen, zu bestimmten Aufgaben regelmäßig individuelles Feedback für die SuS zu erstellen?	- Ja, da sehr hilfreich für Lernprozess - Nein, keine zeitliche Kapazität	- Wo sehen Sie Vorteile und Hürden?
3. Nutzen des Feedbacks für die Lehrkraft Sie haben automatisch generiertes Feedback zu den bearbeiteten Concept Maps der SuS erhalten. Darum soll es im Folgenden nun spezifisch gehen.		
3.1 Haben Sie schon einmal mit Concept Maps gearbeitet?	ja/nein	*Wenn ja*: - Wofür genau haben Sie Concept Maps genutzt? - Wie sind Ihre Erfahrungen diesbezüglich?
3.2 Haben Sie das erhaltene automatisch generierte Feedback genutzt?	-ja/nein	*Wenn ja:* - Können Sie beschreiben, wie genau Sie das Feedback zur Planung der Unterrichtseinheit genutzt haben? - Hat Sie das erhaltene Feedback dazu angeregt, Konsequenzen für die Planung der Unterrichtseinheit zu ziehen? - Konnten Sie durch das gegebene Feedback Zeit sparen (sowohl im Unterricht, als auch bei der Planung)? - Nutzen auf Klassenebene oder individueller Ebene *Wenn nein:* - Warum nicht?

3.3 Halten Sie das Feedback allgemein für nützlich?	ja/nein	*Wenn ja*: Können Sie sich vorstellen, öfters mit so einem Feedback zu arbeiten? *Wenn nein*: Wie sollte Feedback aussehen, dass für Sie hilfreich wäre?
Welche Aspekte halten Sie für besonders hilfreich?	- z. B. nur Übersicht etc.	
Gibt es auch Teile des Feedbacks, die sich nicht angeschaut haben oder bei der Planung der Unterrichtseinheit nicht berücksichtig haben?	- z. B. Problematic Propositions	Warum?
Sie haben zu zwei Zeitpunkten Feedback über die Bearbeitungen Ihrer SuS erhalten. An welchem Zeitpunkt war das Feedback für Sie hilfreicher/nützlicher?	- erster/zweiter Zeitpunkt	Warum?
3.4 Würden Sie sich wünschen, das im Feedback genauer auf die Lerngruppe eingegangen wird?	ja/nein	→ Feedback elaboriert/kompakt Um Begründen bitten *Falls Lehrkraft beide Feedback-Arten erhalten hat:* - Können Sie beschreiben, ob das elaborierte Feedback hilfreicher war oder ob das kompakte Feedback alle nötigen Informationen enthält?

3.5 Denken Sie es ist notwendig bzw. hilfreich, dass die SuS ebenfalls Feedback zu ihren Bearbeitungen erhalten?	ja/nein	Um Begründung bitten
4. Mögliche Problemfelder		
4.1 Glauben Sie, dass die automatische Auswertung mehr oder weniger Fehler als Sie selber macht?	- mehr/weniger	
4.2 Können Sie Probleme identifizieren, die bei KI erzeugtem Feedback auftreten könnten?	- Auftretende Fehler → Verwirrung der SuS - Ungenaue Auswertungen	
4.3 Haben Sie oder Schüler:innen Fehler im Schüler:innenfeedback bemerkt?	ja/nein	Wenn nein: Erläuterung, dass Fehler auftreten.
5. Ergänzungen		
5.1 Würden Sie sich wünschen, dass so ein KI-basiertes Feedback für weitere Aufgabenstellungen im PU möglich wäre?		z. B. für Versuchsprotokolle
5.2 Ich habe damit alle meine Fragen gestellt. Möchten Sie etwas ergänzen?		

Abschluss

- Für Teilnahme bedanken
- Fragen zum Interview?

G - Auszug aus dem Codesystem der Masterarbeit Pohl (n. V.)

Übersicht Codesystem

- **1. Künstliche Intelligenz**
  - 1.1. KI im Bildungswesen
    - 1.1.1. Positiv
      - 1.1.1.1. Individuell
      - 1.1.1.2. Entlastung
      - 1.1.1.3. Lernhürden erkennen
      - 1.1.1.4. Potenziale erkennen
      - 1.1.1.5. Differenzierung
      - 1.1.1.6. Variation
      - 1.1.1.7. Einsatz im Unterricht
    - 1.1.2. Negativ
      - 1.1.2.1. Fehlende Ausbildung
      - 1.1.2.2. Schüler:innenlösungen ersetzen
      - 1.1.2.3. Falschaussagen
      - 1.1.2.4. Unspezifisch
      - 1.1.2.5. Kein Hinterfragen
  - 1.2. Nutzung KI als Lehrkraft
    - 1.2.1. Genutzt
      - 1.2.1.1. Texte/Aufgabenstellungen
      - 1.2.1.2. Rechtschreibprüfung
      - 1.2.1.3. Differenzierung
      - 1.2.1.4. Medienkompetenz
      - 1.2.1.5. Inspiration
    - 1.2.2. Nicht genutzt
      - 1.2.2.1. Ungewohnt
      - 1.2.2.2. Keine Notwendigkeit
    - 1.2.3. Positive Erfahrung
    - 1.2.4. Negative Erfahrung
- **2. Formatives Assessment**
  - 2.1. Bekanntheit Formatives Assessment
    - 2.1.1. Bekannt
    - 2.1.2. Unbekannt
    - 2.1.3. Ungefähre Vorstellung
  - 2.2. Nutzung von formativem Assessment
    - 2.2.1. Keine Nutzung
    - 2.2.2. Unterricht anpassen
    - 2.2.3. Lernhürden identifizieren
    - 2.2.4. Lernstand erfassen
    - 2.2.5. Wiederholung
    - 2.2.6. Feedback
    - 2.2.7. Ebene
      - 2.2.7.1. Klassenebene
      - 2.2.7.2. Individualebene
      - 2.2.7.3. Klassen- und Individualebene
  - 2.3. Regelmäßiges, individuelles Feedback
    - 2.3.1. Abhängig vom Aufgabentypus
    - 2.3.2. Mündlich möglich
    - 2.3.3. Mit Unterstützung KI
    - 2.3.4. Hoher Zeitaufwand
    - 2.3.5. Keine Wertschätzung
- **3. Feedback**
  - 3.1. Concept Maps
    - 3.1.1. Nicht genutzt

- 3.1.2. Genutzt
  - 3.1.2.1. Zusammenfassung
  - 3.1.2.2. Positive Erfahrung

- 3.2. Nutzung des Feedbacks
  - 3.2.1. Nicht genutzt
    - 3.2.1.1. Inhaltliche Passung
    - 3.2.1.2. Keine Zeit
    - 3.2.1.3. Geringe Teilnahme SuS
  - 3.2.2. Genutzt
    - 3.2.2.1. Zusätzliche Stunde
    - 3.2.2.2. Planung Unterrichtseinheit
    - 3.2.2.3. Aspekte Wiederholen
    - 3.2.2.4. Wissensstand erfassen
    - 3.2.2.5. Klassenebene
    - 3.2.2.6. Vorbereitung Klassenarbeit
    - 3.2.2.7. Konzipierung Klassenarbeit

- 3.3. Einschätzung zum Feedback
  - 3.3.1. Allgemein nützlich
  - 3.3.2. Kritik
    - 3.3.2.1. Zu umfangreich
    - 3.3.2.2. Verfälschung durch Unklarheiten
    - 3.3.2.3. Flexibilität
    - 3.3.2.4. Zuverlässigkeit
    - 3.3.2.5. Elaboriert: Einordnung
  - 3.3.3. Regelmäßig
    - 3.3.3.1. Zustimmung
    - 3.3.3.2. In reduzierter Form
  - 3.3.4. Aspekte des Feedbacks
    - 3.3.4.1. Alle
    - 3.3.4.2. Übersicht
    - 3.3.4.3. Analyse-SuS
    - 3.3.4.4. Problematic Propositions
    - 3.3.4.5. Alle Antworten
  - 3.3.5. Zeitpunkt
    - 3.3.5.1. Erster hilfreicher
    - 3.3.5.2. Zweiter nicht aussagekräftig
    - 3.3.5.3. Beide
  - 3.3.6. Lerngruppe
    - 3.3.6.1. Ausreichend genau
    - 3.3.6.2. Detaillierter
  - 3.3.7. SuS-Feedback
    - 3.3.7.1. Sinnvoll
    - 3.3.7.2. Nicht angeschaut

- 3.4. Weitere Aufgabenstellungen
  - 3.4.1. Allgemeine Zustimmung
    - 3.4.1.1. Anderer Aufgabentyp
    - 3.4.1.2. Weitere Themen

4. Mögliche Problemfelder

- 4.1. Fehler KI
  - 4.1.1. Mehr Fehler
    - 4.1.1.1. KI überflüssig
    - 4.1.1.2. Didaktisch
    - 4.1.1.3. Interpretation
  - 4.1.2. Weniger Fehler
    - 4.1.2.1. Fachlich
    - 4.1.2.2. Objektiver

Lebenslauf

Wissenschaftlicher Werdegang

Promotion zum Dr. rer. nat., Fakultät für Mathematik & Physik, Leibniz Universität Hannover
01/2021 - 06/2024

Master of Education, Fächerkombination: Mathematik & Physik, Fakultät für Mathematik & Physik, Leibniz Universität Hannover
10/2018 - 09/2020

Bachelor of Science, Fächerkombination: Mathematik & Physik, Fakultät für Mathematik & Physik, Leibniz Universität Hannover
10/2013 – 10/2018

Bachelor of Science, Fakultät für Maschinenbau, Leibniz Universität Hannover
09/2012 – 09/2013

Allgemeine Hochschulreife, Goethe Gymnasium Hildesheim
08/2004 – 08/2012

Publikationsliste

Paper und Tagungsbandbeiträge

Bleckmann, T., & Friege, G. (2023). Automatische Auswertung von Concept Maps: Wie kann Machine Learning helfen? Tagungsband GDCP 2022

Bleckmann, T., & Friege, G. (2023). Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method. Knowledge Management & E-Learning, 15(3), 433–447. https://doi.org/10.34105/j.kmel.2023.15.025

Bleckmann, T. & Friege, G. (2024). Feedback durch Machine Learning – Automatische Rückmeldung zu Concept Maps. Tagungsband GDCP 2023

Bleckmann, T., Meyer, A., Oldag, J., Stamatakis, M., Markovnikova, A. (2024). LernMINT: Datengestützter Unterricht in den MINT-Fächern: Postersymposium. Tagungsband GDCP 2023

Poster und Vorträge

Bleckmann, T., Dieckhoff, L., Friege, G. (2021). LernMINT interdisziplinär, innovativ und zukunftsorientiert. GDCP Schwerpunkttagung Maschinelles Lernen und computerbasierte Textanalysen. Online. Posterpräsentation

Bleckmann, T., Gritz, W., Friege, G. (2022). Analysis of Concept Maps for the use in formative assessment: Can Machine Learning help? NARST ANNUAL INTERNATIONAL CONFERENCE. Vancouver, BC, Kanada. Vortrag

Bleckmann, T., Friege, G. (2022). Using Machine Learning to analyze Concept Maps for formative assessment. 9th International Conference on Concept Mapping.

Valetta, Malta. Vortrag

Bleckmann, T., Friege, G. (2022). Automatische Auswertung von Concept Maps: Wie kann Machine Learning helfen? GDCP Jahrestagung. Aachen. Vortrag

Bleckmann, T., Friege, G. (2023). Verwendung von Machine Learning zur Auswertung von Concept Maps in der Mechanik. Deutsche Physikalische Gesellschaft Jahrestagung. Hannover. Posterpräsentation

Bleckmann, T., Gritz, W., Friege, G. (2023). Using Machine Learning for a qualitive evaluation of Concept Maps: New opportunities for formative assessment? NARST ANNUAL INTERNATIONAL CONFERENCE. Chicago, IL, USA. Vortrag

Bleckmann, T., Friege, G. (2023). Feedback durch Machine Learning – Automatische Rückmeldung zu Concept Maps. GDCP Jahrestagung. Hamburg. Posterpräsentation

Bleckmann, T., Friege, G. (2024). Evaluation of Machine Learning generated Feedback for Concept Maps. NARST ANNUAL INTERNATIONAL CONFERENCE. Denver, CO, USA. Posterpräsentation

Bisher erschienene Bände der Reihe
Studien zum Physik- und Chemielernen

ISSN 1614-8967

Vollständige Übersicht auf unserer Website

https://www.logos-verlag.de/spcl

Aktuelle Bände

300 Amany Annaggar (2020): A Design Framework for Video Game-Based Gamification Elements to Assess Problem-solving Competence in Chemistry Education
ISBN 978-3-8325-5150-6 52.00 EUR

301 Alexander Engl (2020): CHEMIE PUR – Unterrichten in der Natur. Entwicklung und Evaluation eines kontextorientierten Unterrichtskonzepts im Bereich Outdoor Education zur Änderung der Einstellung zu „Chemie und Natur“
ISBN 978-3-8325-5174-2 59.00 EUR (open access)

302 Christin Marie Sajons (2020): Kognitive und motivationale Dynamik in Schülerlaboren. Kontextualisierung, Problemorientierung und Autonomieunterstützung der didaktischen Struktur analysieren und weiterentwickeln
ISBN 978-3-8325-5155-1 56.00 EUR (open access)

303 Philipp Bitzenbauer (2020): Quantenoptik an Schulen. Studie im Mixed-Methods Design zur Evaluation des Erlanger Unterrichtskonzepts zur Quantenoptik
ISBN 978-3-8325-5123-0 59.00 EUR (open access)

304 Malte Ubben (2020): Typisierung des Verständnisses mentaler Modelle mittels empirischer Datenerhebung am Beispiel der Quantenphysik
ISBN 978-3-8325-5181-0 43.50 EUR (open access)

305 Wiebke Hinrike Kuske-Janßen (2020): Sprachlicher Umgang mit Formeln von LehrerInnen im Physikunterricht am Beispiel des elektrischen Widerstandes in Klassenstufe 8
ISBN 978-3-8325-5183-4 47.50 EUR (open access)

306 Kai Bliesmer (2020): Physik der Küste für außerschulische Lernorte. Eine Didaktische Rekonstruktion
ISBN 978-3-8325-5190-2 58.00 EUR (open access)

307 Nikola Schild (2021): Eignung von domänenspezifischen Studieneingangsvariablen als Prädiktoren für Studienerfolg im Fach und Lehramt Physik
ISBN 978-3-8325-5226-8 42.00 EUR (open access)

308 Daniel Averbeck (2021): Zum Studienerfolg in der Studieneingangsphase des Chemiestudiums. Der Einfluss kognitiver und affektiv-motivationaler Variablen
ISBN 978-3-8325-5227-5 51.00 EUR

309 Martina Strübe (2021): Modelle und Experimente im Chemieunterricht. Eine Videostudie zum fachspezifischen Lehrerwissen und -handeln
ISBN 978-3-8325-5245-9 45.50 EUR

310 Wolfgang Becker (2021): Auswirkungen unterschiedlicher experimenteller Repräsentationen auf den Kenntnisstand bei Grundschulkindern
ISBN 978-3-8325-5255-8 50.00 EUR

311 Marvin Rost (2021): Modelle als Mittel der Erkenntnisgewinnung im Chemieunterricht der Sekundarstufe I. Entwicklung und quantitative Dimensionalitätsanalyse eines Testinstruments aus epistemologischer Perspektive
ISBN 978-3-8325-5256-5 44.00 EUR (open access)

312 Christina Kobl (2021): Förderung und Erfassung der Reflexionskompetenz im Fach Chemie
ISBN 978-3-8325-5259-6 41.00 EUR (open access)

313 Ann-Kathrin Beretz (2021): Diagnostische Prozesse von Studierenden des Lehramts. eine Videostudie in den Fächern Physik und Mathematik
ISBN 978-3-8325-5288-6 45.00 EUR (open access)

314 Judith Breuer (2021): Implementierung fachdidaktischer Innovationen durch das Angebot materialgestützter Unterrichtskonzeptionen. Fallanalysen zum Nutzungsverhalten von Lehrkräften am Beispiel des Münchener Lehrgangs zur Quantenmechanik
ISBN 978-3-8325-5293-0 50.50 EUR (open access)

315 Michaela Oettle (2021): Modellierung des Fachwissens von Lehrkräften in der Teilchenphysik. Eine Delphi-Studie
ISBN 978-3-8325-5305-0 57.50 EUR (open access)

316 Volker Brüggemann (2021): Entwicklung und Pilotierung eines adaptiven Multistage-Tests zur Kompetenzerfassung im Bereich naturwissenschaftlichen Denkens
ISBN 978-3-8325-5331-9 40.00 EUR (open access)

317 Stefan Müller (2021): Die Vorläufigkeit und soziokulturelle Eingebundenheit naturwissenschaftlicher Erkenntnisse. Kritische Reflexion, empirische Befunde und fachdidaktische Konsequenzen für die Chemielehrer*innenbildung
ISBN 978-3-8325-5343-2 63.00 EUR

318 Laurence Müller (2021): Alltagsentscheidungen für den Chemieunterricht erkennen und Entscheidungsprozesse explorativ begleiten
ISBN 978-3-8325-5379-1 59.00 EUR

319 Lars Ehlert (2021): Entwicklung und Evaluation einer Lehrkräftefortbildung zur Planung von selbstgesteuerten Experimenten
ISBN 978-3-8325-5393-7 41.50 EUR (open access)

320 Florian Seiler (2021): Entwicklung und Evaluation eines Seminarkonzepts zur Förderung der experimentellen Planungskompetenz von Lehramtsstudierenden im Fach Chemie
ISBN 978-3-8325-5397-5 47.50 EUR (open access)

321 Nadine Boele (2021): Entwicklung eines Messinstruments zur Erfassung der professionellen Unterrichtswahrnehmung von (angehenden) Chemielehrkräften hinsichtlich der Lernunterstützung
ISBN 978-3-8325-5402-6 46.50 EUR

322 Franziska Zimmermann (2022): Entwicklung und Evaluation digitalisierungsbezogener Kompetenzen von angehenden Chemielehrkräften
ISBN 978-3-8325-5410-1 49.50 EUR

323 Lars-Frederik Weiß (2021): Der Flipped Classroom in der Physik-Lehre. Empirische Untersuchungen in Schule und Hochschule
ISBN 978-3-8325-5418-7 51.00 EUR

324 Tilmann Steinmetz (2021): Kumulatives Lehren und Lernen im Lehramtsstudium Physik. Theorie und Evaluation eines Lehrkonzepts
ISBN 978-3-8325-5421-7 51.50 EUR

325 Kübra Nur Celik (2022): Entwicklung von chemischem Fachwissen in der Sekundarstufe I. Validierung einer Learning Progression für die Basiskonzepte „Struktur der Materie“, „Chemische Reaktion“ und „Energie“ im Kompetenzbereich „Umgang mit Fachwissen“
ISBN 978-3-8325-5431-6 55.00 EUR

326 Matthias Ungermann (2022): Förderung des Verständnisses von Nature of Science und der experimentellen Kompetenz im Schüler*innen-Labor Physik in Abgrenzung zum Regelunterricht
ISBN 978-3-8325-5442-2 55.50 EUR

327 Christoph Hoyer (2022): Multimedial unterstütztes Experimentieren im webbasierten Labor zur Messung, Visualisierung und Analyse des Feldes eines Permanentmagneten
ISBN 978-3-8325-5453-8 45.00 EUR

328 Tobias Schüttler (2022): Schülerlabore als interessefördernde authentische Lernorte für den naturwissenschaftlichen Unterricht nutzen
ISBN 978-3-8325-5454-5 50.50 EUR

329 Christopher Kurth (2022): Die Kompetenz von Studierenden, Schülerschwierigkeiten beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-5457-6 58.50 EUR

330 Dagmar Michna (2022): Inklusiver Anfangsunterricht Chemie. Entwicklung und Evaluation einer Unterrichtseinheit zur Einführung der chemischen Reaktion
ISBN 978-3-8325-5463-7 49.50 EUR

331 Marco Seiter (2022): Die Bedeutung der Elementarisierung für den Erfolg von Mechanikunterricht in der Sekundarstufe I
ISBN 978-3-8325-5471-2 66.00 EUR

332 Jörn Hägele (2022): Kompetenzaufbau zum experimentbezogenen Denken und Arbeiten. Videobasierte Analysen zu Aktivitäten und Vorstellungen von Schülerinnen und Schülern der gymnasialen Oberstufe bei der Bearbeitung von fachmethodischer Instruktion
ISBN 978-3-8325-5476-7 56.50 EUR (open access)

333 Erik Heine (2022): Wissenschaftliche Kontroversen im Physikunterricht. Explorationsstudie zum Umgang von Physiklehrkräften und Physiklehramtsstudierenden mit einer wissenschaftlichen Kontroverse am Beispiel der Masse in der Speziellen Relativitätstheorie
ISBN 978-3-8325-5478-1 48.50 EUR (open access)

334 Simon Goertz (2022): Module und Lernzirkel der Plattform FLexKom zur Förderung experimenteller Kompetenzen in der Schulpraxis. Verlauf und Ergebnisse einer Design-Based Research Studie
ISBN 978-3-8325-5494-1 66.50 EUR

335 Christina Toschka (2022): Lernen mit Modellexperimenten. Empirische Untersuchung der Wahrnehmung und des Denkens in Analogien beim Umgang mit Modellexperimenten
ISBN 978-3-8325-5495-8 50.00 EUR (open access)

336 Alina Behrendt (2022): Chemiebezogene Kompetenzen in der Übergangsphase zwischen dem Sachunterricht der Primarstufe und dem Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-5498-9 40.50 EUR (open access)

337 Manuel Daiber (2022): Entwicklung eines Lehrkonzepts für eine elementare Quantenmechanik. Formuliert mit In-Out Symbolen
ISBN 978-3-8325-5507-8 48.50 EUR

338 Felix Pawlak (2022): Das Gemeinsame Experimentieren (an-)leiten. Eine qualitative Studie zum chemiespezifischen Classroom-Management
ISBN 978-3-8325-5508-5 46.50 EUR

339 Liza Dopatka (2022): Konzeption und Evaluation eines kontextstrukturierten Unterrichtskonzeptes für den Anfangs-Elektrizitätslehreunterricht
ISBN 978-3-8325-5514-6 69.50 EUR

340 Arne Bewersdorff (2022): Untersuchung der Effektivität zweier Fortbildungsformate zum Experimentieren mit dem Fokus auf das Unterrichtshandeln
ISBN 978-3-8325-5522-1 39.00 EUR (open access)

341 Thomas Christoph Münster (2022): Wie diagnostizieren Studierende des Lehramtes physikbezogene Lernprozesse von Schüler*innen?. Eine Videostudie zur Mechanik
ISBN 978-3-8325-5534-4 44.50 EUR (open access)

342 Ines Komor (2022): Förderung des symbolisch-mathematischen Modellverständnisses in der Physikalischen Chemie
ISBN 978-3-8325-5546-7 46.50 EUR

343 Verena Petermann (2022): Überzeugungen von Lehrkräften zum Lehren und Lernen von Fachinhalten und Fachmethoden und deren Beziehung zu unterrichtsnahem Handeln
ISBN 978-3-8325-5545-0 47.00 EUR (open access)

344 Jana Heinze (2022): Einfluss der sprachlichen Konzeption auf die Einschätzung der Qualität instruktionaler Unterrichtserklärungen im Fach Physik
ISBN 978-3-8325-5553-5 42.00 EUR (open access)

345 Jannis Weber (2022): Mathematische Modellbildung und Videoanalyse zum Lernen der Newtonschen Dynamik im Vergleich
ISBN 978-3-8325-5566-5 68.00 EUR (open access)

346 Fabian Sterzing (2022): Zur Lernwirksamkeit von Erklärvideos in der Physik. Eine Untersuchung in Abhängigkeit von ihrer fachdidaktischen Qualität und ihrem Einbettungsformat
ISBN 978-3-8325-5576-4 52.00 EUR (open access)

347 Lars Greitemann (2022): Wirkung des Tablet-Einsatzes im Chemieunterricht der Sekundarstufe I unter besonderer Berücksichtigung von Wissensvermittlung und Wissenssicherung
ISBN 978-3-8325-5580-1 50.00 EUR

348 Fabian Poensgen (2022): Diagnose experimenteller Kompetenzen in der laborpraktischen Chemielehrer*innenbildung
ISBN 978-3-8325-5587-0 48.00 EUR

349 William Lindlahr (2023): Virtual-Reality-Experimente. Entwicklung und Evaluation eines Konzepts für den forschend-entwickelnden Physikunterricht mit digitalen Medien
ISBN 978-3-8325-5595-5 49.00 EUR

350 Bert Schlüter (2023): Teilnahmemotivation und situationales Interesse von Kindern und Eltern im experimentellen Lernsetting KEMIE
ISBN 978-3-8325-5598-6 43.00 EUR

351 Katharina Nave (2023): Charakterisierung situativer mentaler Modellkomponenten in der Chemie und die Bildung von Hypothesen. Eine qualitative Studie zur Operationalisierung mentaler Modell-komponenten für den Fachbereich Chemie
ISBN 978-3-8325-5599-3 43.00 EUR

352 Anna B. Bauer (2023): Experimentelle Kompetenz Physikstudierender. Entwicklung und erste Erprobung eines performanzorientierten Kompetenzstrukturmodells unter Nutzung qualitativer Methoden
ISBN 978-3-8325-5625-9 47.00 EUR (open access)

353 Jan Schröder (2023): Entwicklung eines Performanztests zur Messung der Fähigkeit zur Unterrichtsplanung bei Lehramtsstudierenden im Fach Physik
ISBN 978-3-8325-5655-6 46.50 EUR

354 Susanne Gerlach (2023): Aspekte einer Fachdidaktik Körperpflege. Ein Beitrag zur Standardentwicklung
ISBN 978-3-8325-5659-4 45.00 EUR

355 Livia Murer (2023): Diagnose experimenteller Kompetenzen beim praktisch-naturwissenschaftlichen Arbeiten. Vergleich verschiedener Methoden und kognitive Validierung eines Testverfahrens
ISBN 978-3-8325-5657-0 41.50 EUR (open access)

356 Andrea Maria Schmid (2023): Authentische Kontexte für MINT-Lernumgebungen. Eine zweiteilige Interventionsstudie in den Fachdidaktiken Physik und Technik
ISBN 978-3-8325-5605-1 57.00 EUR (open access)

357 Julia Ortmann (2023): Bedeutung und Förderung von Kompetenzen zum naturwissenschaftlichen Denken und Arbeiten in universitären Praktika
ISBN 978-3-8325-5670-9 37.00 EUR (open access)

358 Axel-Thilo Prokop (2023): Entwicklung eines Lehr-Lern-Labors zum Thema Radioaktivität. Eine didaktische Rekonstruktion
ISBN 978-3-8325-5671-6 49.50 EUR

359 Timo Hackemann (2023): Textverständlichkeit sprachlich variierter physikbezogener Sachtexte
ISBN 978-3-8325-5675-4 41.50 EUR (open access)

360 Dennis Dietz (2023): Vernetztes Lernen im fächerdifferenzierten und integrierten naturwissenschaftlichen Unterricht aufgezeigt am Basiskonzept Energie. Eine Studie zur Analyse der Wirksamkeit der Konzeption und Implementation eines schulinternen Curriculums für das Unterrichtsfach „Integrierte Naturwissenschaften 7/8“
ISBN 978-3-8325-5676-1 49.50 EUR

361 Ann-Katrin Krebs (2023): Vielfalt im Physikunterricht. Zur Wirkung von Lehrkräftefortbildungen unter Diversitätsaspekten
ISBN 978-3-8325-5672-3 65.50 EUR (open access)

362 Simon Kaulhausen (2023): Strukturelle Ursachen für Klausurmisserfolg in Allgemeiner Chemie an der Universität
ISBN 978-3-8325-5699-0 37.50 EUR (open access)

363 Julia Eckoldt (2023): Den (Sach-)Unterricht öffnen. Selbstkompetenzen und motivationale Orientierungen von Lehrkräften bei der Implementation einer Innovation untersucht am Beispiel des Freien Explorierens und Experimentierens
ISBN 978-3-8325-5663-1 48.50 EUR (open access)

364 Albert Teichrew (2023): Physikalische Modellbildung mit dynamischen Modellen
ISBN 978-3-8325-5710-2 58.50 EUR (open access)

365 Sascha Neff (2023): Transfer digitaler Innovationen in die Schulpraxis. Eine explorative Untersuchung zur Förderung der Implementation
ISBN 978-3-8325-5687-7 59.00 EUR (open access)

366 Rahel Schmid (2023): Verständnis von Nature of Science-Aspekten und Umgang mit Fehlern von Schüler*innen der Sekundarstufe I. Am Beispiel von digital-basierten Lernprozessen im informellen Lernsetting Smartfeld
ISBN 978-3-8325-5722-5 53.50 EUR (open access)

367 Dennis Kirstein (2023): Individuelle Bedingungs- und Risikofaktoren für erfolgreiche Lernprozesse mit kooperativen Experimentieraufgaben im Chemieunterricht. Eine Untersuchung zum Zusammenhang von Lernvoraussetzungen, Lerntätigkeiten, Schwierigkeiten und Lernerfolg beim Experimentieren in Kleingruppen der Sekundarstufe I
ISBN 978-3-8325-5729-4 50.50 EUR (open access)

368 Frauke Düwel (2024): Argumentationslinien in Lehr-Lernkontexten. Potenziale englischer Fachtexte zur Chromatografie und deren hochschuldidaktische Einbindung
ISBN 978-3-8325-5731-7 63.00 EUR (open access)

369 Fabien Güth (2023): Interessenbasierte Differenzierung mithilfe systematisch variierter Kontextaufgaben im Fach Chemie
ISBN 978-3-8325-5737-9 48.00 EUR (open access)

370 Oliver Grewe (2023): Förderung der professionellen Unterrichtswahrnehmung und Selbstwirksamkeitsüberzeugungen hinsichtlich sprachsensibler Maßnahmen im naturwissenschaftlichen Sachunterricht. Konzeption und Evaluation einer video- und praxisbasierten Lehrveranstaltung im Masterstudium
ISBN 978-3-8325-5738-6 44.50 EUR (open access)

371 Anna Nowak (2023): Untersuchung der Qualität von Selbstreflexionstexten zum Physikunterricht. Entwicklung des Reflexionsmodells REIZ
ISBN 978-3-8325-5739-3 59.00 EUR (open access)

372 Dominique Angela Holland (2023): Bildung für nachhaltige Entwicklung (BNE) kooperativ gestalten. Vergleich monodisziplinärer und interdisziplinärer Kooperation von Lehramtsstudierenden bei der Planung, Durchführung und Reflexion von Online-BNE-Unterricht
ISBN 978-3-8325-5760-7 47.00 EUR (open access)

373 Renan Marcello Vairo Nunes (2024): MINT-Personal an Schulen. Eine Untersuchung der Arbeitssituation und professionellen Kompetenzen von MINT-Lehrkräften verschiedener Ausbildungswege
ISBN 978-3-8325-5778-2 51.00 EUR (open access)

374 Mats Kieserling (2024): Digitalisierung im Chemieunterricht. Entwicklung und Evaluation einer experimentellen digitalen Lernumgebung mit universeller Zugänglichkeit
ISBN 978-3-8325-5786-7 45.50 EUR

375 Cem Aydin Salim (2024): Die Untersuchung adaptiver Lernsettings im Themenbereich „Schwimmen und Sinken“ im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-5787-4 49.00 EUR (open access)

376 Novid Ghassemi (2024): Evaluation eines Lehramtsmasterstudiengangs mit dem Profil Quereinstieg im Fach Physik. Erkenntnisse zu Eingangsbedingungen, professionellen Kompetenzen und Aspekten individueller Angebotsnutzung
ISBN 978-3-8325-5789-8 41.50 EUR (open access)

377 Martina Flurina Cavelti (2024): Entwicklung und Validierung eines Messinstruments zur Erfassung der Schülerkompetenzen im Bereich des wissenschaftlichen Skizzierens im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-5829-1 45.00 EUR (open access)

378 Tom Bleckmann (2024): Formatives Assessment auf Basis von maschinellem Lernen. Eine Studie über automatisiertes Feedback zu Concept Maps aus dem Bereich Mechanik
ISBN 978-3-8325-5842-0 46.50 EUR (open access)

379 Jana Marlies Rehberg (2024): Das physikspezifische Mindset zum Studienbeginn. Fragebogenentwicklung und Aufbau einer Online-Intervention
ISBN 978-3-8325-5850-5 59.50 EUR (open access)

380 Florian Trauten (2024): Entwicklung und Evaluation von automatisierten Feedbackschleifen in Online-Aufgaben im Fach Chemie
ISBN 978-3-8325-5859-8 47.00 EUR (open access)

Vollständige Übersicht unter: https://www.logos-verlag.de/spcl

Studien zum Physik- und Chemielernen

Herausgegeben von Martin Hopf und Mathias Ropohl

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung im deutschsprachigen Raum.
Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Univ.-Prof. Dr. Martin Hopf
Universität Wien,
Österreichisches Kompetenzzentrum
für Didaktik der Physik,
Porzellangasse 4, Stiege 2,
1090 Wien, Österreich,
Tel. +43-1-4277-60330,
e-mail: martin.hopf@univie.ac.at

Prof. Dr. Mathias Ropohl
Didaktik der Chemie,
Fakultät für Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen,
Tel. 0201-183 2704,
e-mail: mathias.ropohl@uni-due.de

Formatives Assessment hat sich als effektive Methode zur Unterstützung des Lehrens und Lernens erwiesen. Allerdings werden formative Assessments in der Praxis häufig durch die Größe der Lerngruppen sowie zeitlichen Aufwand erschwert. Vor diesem Hintergrund wurde mithilfe maschinellen Lernens eine automatische Auswertung von Concept Maps für ein formatives Assessment entwickelt und evaluiert.

In einer ersten Studie wurde eine vorstrukturierte Concept Map zum Thema Mechanik entwickelt und von 230 Lernenden der 11. Jahrgangsstufe bearbeitet. Die Propositionen der bearbeiteten Concept Maps ($N = 3322$) wurden anschließend von zwei menschlichen Experten anhand eines vierstufigen Bewertungsschemas bewertet. Basierend auf den Propositionen und den Bewertungen wurden mehrere Machine-Learning-Modelle zur automatischen Auswertung der Propositionen entwickelt, die eine hohe Übereinstimmung mit den menschlichen Bewertungen aufwiesen.

In der Folge wurde ein Feedback-Tool entwickelt, welches Lehrkräften und Lernenden automatische Rückmeldungen bereitstellt. Die Integration des Tools in den Physikunterricht wurde in einer zweiten Studie untersucht. Dabei zeigte sich, dass Lehrkräfte die Vorteile der automatischen Auswertung erkennen, das volle Potenzial jedoch noch nicht ausschöpfen.

Die Ergebnisse verdeutlichen das Potenzial von Machine-Learning-Systemen im Bildungsbereich und heben die Notwendigkeit einer engeren Zusammenarbeit zwischen Informatik, Fachdidaktik und Lehrkräften hervor.

Logos Verlag Berlin

ISBN 978-3-8325-5842-0