Le|se|räu|me
ZEITSCHRIFT FÜR LITERALITÄT
IN SCHULE UND FORSCHUNG

SARA REZAT / SEBASTIAN KILSBACH / RABIA KARABEY / NADINE MICHEL / MAJA STAHL / HENNING WACHSMUTH

# Didaktische Modellierung automatisierten adaptiven Feedbacks zu argumentativen Lerner*innentexten

**Abstract**

Der Beitrag beschäftigt sich mit der didaktischen Modellierung automatisierten adaptiven Feedbacks zu argumentativen Lerner*innentexten. An einem Korpus aus diesen Texten wird gezeigt, wie adaptives Feedback auf Grundlage einer Mehrebenenannotation, einem mehrdimensionalen Qualitätsrating und Varianzanalysen modelliert werden kann. Im Beitrag wird zunächst ein Überblick über computergestützte Unterstützungssysteme zum Argumentieren gegeben und anschließend die Annotation und das Rating des Korpus beleuchtet. Daran anknüpfend wird das Feedbackverständnis dargelegt, das den Ausgangspunkt für die Modellierung des Feedbacks bildet. Die Ergebnisse der Varianzanalysen zeigen exemplarisch, welche argumentativen Textstrukturen typisch für eine bestimmte Qualitätsstufe von Texten sind und indizieren damit, ab welcher Qualitätsstufe zu welchen Kategorien Feedback gegeben werden sollte.

## 1 | Einleitung

Im Januar 2024 hat die Ständige Wissenschaftliche Kommission der Kultusministerkonferenz ein Impulspapier zu Large Language Models und den Potenzialen dieser Modelle veröffentlicht (KMK 2024). Darin wird unter anderem die Entwicklung domänenspezifischer Feedback-Tools gefordert, die mit fachspezifischen und qualitativ hochwertigen Daten trainiert werden. Im DFG-Projekt *Computergestütztes Lernen argumentativen Schreibens in der digitalen Schulbildung*[1], das bereits 2020 beantragt und bewilligt wurde, wird dieser Forderung nachgekommen; derzeit wird der Prototyp eines Verfahrens der maschinellen Sprachverarbeitung (Natural-Language-Processing, NLP) entwickelt, mittels dessen Schüler*innen unter Berücksichtigung ihres individuellen Kenntnisstandes durch automatisiertes Feedback beim Ausbau argumentativer Kompetenzen unterstützt werden.

Im vorliegenden Beitrag wird gezeigt, wie durch eine Mehrebenen-Strukturannotation und ein mehrdimensionales Rating eines argumentativen Lernendentextekorpus ein adaptives und automatisiertes Feedback modelliert werden kann, das gezielt zur Förderung argumentativer Kompetenzen genutzt werden kann. Zunächst wird ein Überblick über den aktuellen Stand computerbasierter Unterstützungssysteme gegeben mit dem Ziel, das vorgestellte Projekt in

1 Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – 453073654.

den Forschungskontext einzuordnen. Dafür werden wir darlegen, wie das zugrundeliegende Textkorpus annotiert, mit Ratings versehen und analysiert wurde (Kapitel 2 und 3). Auf Grundlage der annotierten und um Ratings erweiterten Daten wurden mithilfe von Varianzanalysen Niveaustufen der Texte abgeleitet. Die Ergebnisse dieser Analysen sind Kern des Kapitels 4. Wie ein adaptives automatisiertes Feedback im Rahmen von NLP-Verfahren auf Basis der vorangegangenen Schritte modelliert werden kann, wird in Kapitel 5 gezeigt. Im abschließenden Fazit (Kapitel 6) wird skizziert, welche Herausforderungen mit der Modellierung automatisierten Feedbacks einhergehen.

## 2 | Computerbasierte Unterstützungssysteme zum Argumentieren

Die Entwicklung algorithmischer Methoden für die automatische Analyse persuasiver Essays oder Argumenttexte sowie die Entwicklung entsprechender Unterstützungssysteme startete bereits Anfang des Jahrtausends; einen Überblick dazu gibt der Beitrag von Benetos (2023, 83). Neben Software-Werkzeugen zur grafischen Darstellung von Argumentationen (*diagramming tools*) sind mittlerweile komplexere digitale Argumentationssysteme bzw. -plattformen entwickelt worden. Diese Systeme bieten zum einen grafische Darstellungen von Argumentationen und zum anderen Unterstützungsangebote, die mit dem Ziel verbunden sind, kognitive (z. B. Strategieinstruktion) und metakognitive Fähigkeiten beim Schreiben argumentativer Texte zu verbessern (z. B. *Rationale, Endoxa Learning, Kialo, C-SAW*; vgl. Benetos 2023).

An dieser Stelle soll exemplarisch für solche Systeme die C-SAW-Software (Computer-Supported Argumentative Writer) vorgestellt werden, die von Benetos und Bétrancourt (2020) entwickelt wurde und Schreibende bei der Generierung von Argumenten und der Strukturierung ihrer Argumente bzw. ihres Textes unterstützt. Die C-SAW-Software umfasst allerdings keine qualitative Bewertung des Textes. Das System bietet jeweils ein Textfeld für die Einleitung, für den Hauptteil und die Konklusion. Zu jedem dieser Textfelder werden strukturelle Hilfen angeboten, unter anderem werden Argumentbausteine und eine Visualisierung der eigenen Argumentstruktur zur Verfügung gestellt. Das System bietet auch Formulierungshilfen (Prompts) ebenso wie Feedback zur kognitiven und metakognitiven Auseinandersetzung mit dem zu verfassenden argumentativen Text. Beispielsweise wird in den Prompts die Rolle von Gegenargumenten expliziert oder die Lernenden haben die Aufgabe anzugeben, wie stark sie ihr gewähltes Argument einschätzen und auf welcher Grundlage bestimmte Argumente hergeleitet wurden (z. B. statistische Argumente etc.). Im Rahmen einer DBR-Studie wurde der Einsatz der C-SAW-Software mit dem Einsatz eines klassischen Texteditors verglichen. Dabei zeigte sich, dass die Argumentationen, die im Rahmen der C-SAW-Software verfasst wurden, stärker ausgebaut wurden als im Vergleichssetting.

Wambsganss et al. (2021) haben für das Argumentieren den sogenannten *ArgueTutor* entwickelt, ein Lernsystem (auch: *pedagogical conversational agent* [PCAs]), das dialogisch angelegt ist und durch adaptives Feedback Lernende beim Schreiben argumentativer Texte unterstützt. Grundlegend für die Entwicklung dieses Systems sind die Schritte, die in Abbildung 1 zu sehen sind.

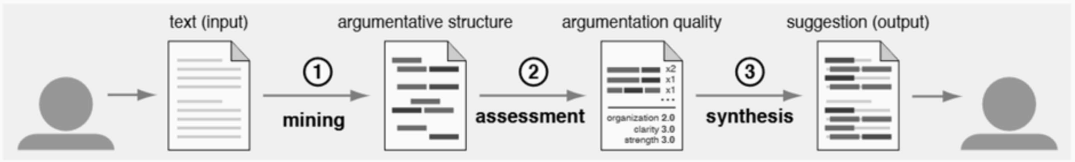


Abbildung 1: Schritte zum schreibunterstützenden Feedback-Output (Wachsmuth et al. 2016, 1681)

Im ersten Schritt wird ein Text in ein System eingegeben und es werden mit Hilfe einer Text-Mining-Software (d. h. einer Software, die mit Hilfe von NLP-Verfahren automatische

Textstrukturen erkennt und bestimmt) spezifische Strukturen des Textes (unter anderem Argumente, Gegenargumente, argumentative Textprozeduren) erschlossen. Für diesen automatisierten Prozess des Argument Minings ist es aber zuvor notwendig, ein geeignetes Textkorpus auszuwählen und die Daten so aufzubereiten, d. h. zu annotieren, dass sie mit computerlinguistischen Verfahren analysiert werden können. Im zweiten Schritt (Assessment) bewertet das System den Text qualitativ anhand von Kriterien, etwa zum Strukturaufbau, Inhalt oder der argumentativen Stichhaltigkeit. Auch dafür muss das Korpus entsprechend händisch aufbereitet sein, in der Regel durch ein Rating der Texte. Das Mining und das Assessment sind die Grundlage für die Synthese, d. h. die Generierung einer Einschätzung des jeweiligen Lernniveaus auf Grundlage der strukturellen und qualitativen Analyse des Textes (vgl. Wachsmuth et al. 2016).

Das automatische Erkennen und Generieren von Argumenten spielen hier also eine zentrale Rolle. Dadurch unterscheidet sich der *ArgueTutor* von den Lernsystemen, die Benetos (2023) in ihrem Beitrag darstellt. In einer Studie mit 55 Proband*innen wurde der *ArgueTutor* mit einem technologiebasierten Argumentationsfeedback verglichen. Hier zeigte sich, dass der Einsatz von *ArgueTutor* zu überzeugenderen Texten und einer besseren Textqualität führt. Abbildung 2 gibt einen Eindruck von dem System.

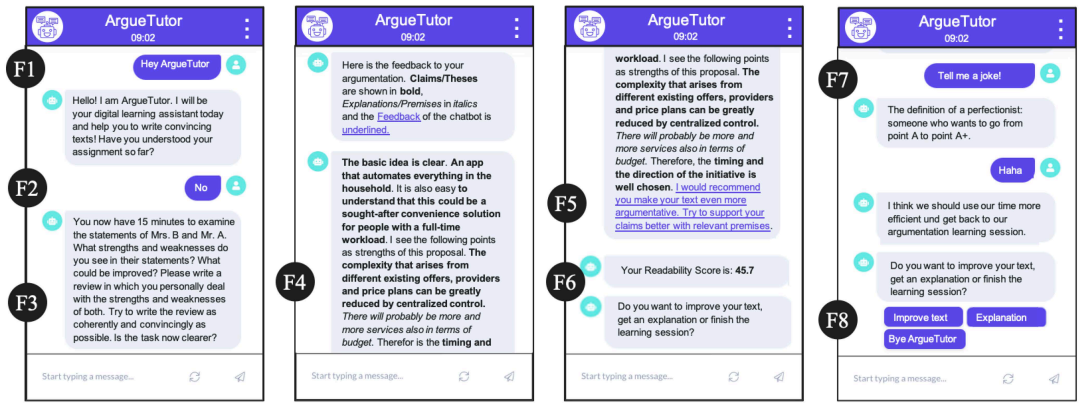


Abbildung 2: Screenshot aus dem Lernsystem *ArgueTutor* (entnommen aus Wambsganss et al. 2021, 1)

Der *ArgueTutor* ist auf das Schreiben englischsprachiger Essays ausgerichtet. Zielgruppe sind Studierende. Dem System liegt allerdings keine explizite sprach- und schreibdidaktische Ausrichtung zugrunde.

Das in diesem Beitrag vorgestellte DFG-Projekt bezieht sich auf aktuelle Entwicklungen im Bereich der NLP- und der KI-Forschung (vgl. etwa Lehnen 2023), aber auch auf didaktische Förderbedarfe im Bereich des argumentativen Schreibens. Im Vergleich zu den dargestellten Systemen und auch bisherigen Forschungen in diesem Bereich (vgl. Kilsbach et al. 2025) zeichnet sich das in diesem Beitrag vorgestellte DFG-Projekt dadurch aus, dass erstens mit einem deutschsprachigen argumentativen Lernendenkorpus gearbeitet wird, und dass zweitens ein adaptives und explizit schreibdidaktisch fundiertes Feedback modelliert wird. Annotation und Rating des Korpus werden nachfolgend detaillierter vorgestellt.

# 3 | Annotation und Rating des Textkorpus

## 3.1 | Informationen zum Korpus

Für die Annotation und das Rating wurde ein Korpus bestehend aus 1320 argumentativen Lerner*innentexten zu drei Schreibaufgaben zusammengestellt. Dafür wurde auf das bereits bestehende FD-LEX aus dem Scriptoria-Korpus zurückgegriffen, das im Rahmen der Studie von Becker-Mrotzek und Grabowski (2018) zu den Teilkomponenten der Schreibkompetenz von Schüler*innen der 5. und 9. Klasse 2013 und 2015 erstellt wurde. Insgesamt besteht es aus 2814 Texten. Für unsere Zwecke erfolgte die Textauswahl gleichmäßig verteilt nach Schreibaufgabe, Klassenstufe und Geschlecht (vgl. Kilsbach et al. 2025).

## 3.2 | Annotation des Korpus

Im Folgenden wird ein kurzer Überblick über die Annotationsebenen und -kategorien gegeben. Eine ausführliche Darlegung des Annotationsvorgehens bieten Stahl et al. (2024) und Kilsbach et al. (2025).

Es wurde eine Mehrebenenannotation durchgeführt, bei der zwischen Makro- und Mikroebene unterschieden wird. Diese beiden Ebenen wurden noch einmal in jeweils zwei weitere Ebenen untergliedert. Abbildung 3 zeigt die Annotationskategorien, die den jeweiligen Ebenen zugeordnet sind.

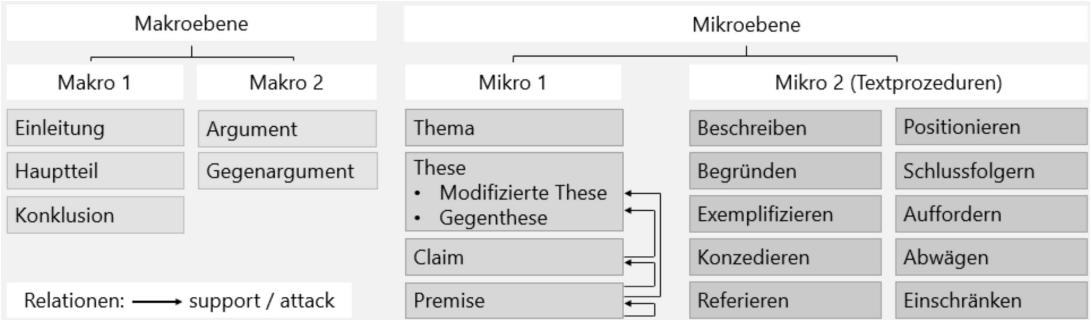


Abbildung 3: Annotationskategorien auf Makro- und Mikrostrukturebene

Auf der Makroebene 1 wird die Grobstruktur des Textes in Einleitung, Hauptteil und Konklusion annotiert. Das Vorgehen basiert auf dem Verfahren von Persing et al. (2010). In Anlehnung an Walton et al. (2008) wurden auf Makroebene 2 Argumente und Gegenargumente identifiziert. Es handelt sich hier in der Regel um größere Textbausteine, die mindestens satzwertig sind.

Das Vorgehen auf Mikroebene 1 erfolgte in Anlehnung an Stab und Gurevych (2017). Auf dieser Ebene werden die Nennung des Themas, zu dem argumentiert wird, die These – wobei hier zusätzlich noch Gegenthese und modifizierte Thesen unterschieden werden – sowie Claims und Premises annotiert. Claims stellen Hauptbegründungen für die These (bei Stab/Gurevych 2017 *major claim*) dar. Premises können Claims näher spezifizieren oder begründen. Es handelt sich hier nicht um Prämissen im formallogischen Verständnis. Claims können Thesen bzw. Premises Claims sowohl unterstützen als auch attackieren. Diese Relationen (in Abbildung 3 symbolisiert durch Pfeile) wurden ebenfalls annotiert.

Die Mikroebene 2 wurde in die Mehrebenenannotation integriert, um auch schreibdidaktisch relevante Ansatzpunkte, und zwar Textprozeduren (vgl. Feilke/Rezat 2021; Rezat et al. 2024), für das spätere Feedback zu berücksichtigen.

> Textprozeduren sind aus einer systemisch-funktionalen Sicht Handlungskomponenten im Aufbau von Texten […]. Sie werden ausdrucksseitig in variabel gefügten Ausdrucksmustern fassbar, die pragmatische Gebrauchsinformationen anreichern und Handlungsschemata der Textbildung semantisch verfügbar halten. (Feilke 2024, 9)

Im obigen Fall werden zehn solcher für argumentative Texte zentrale Textprozeduren untersucht. Im Vergleich zu den anderen Ebenen wird mit der Mikroebene 2 durch die Annotation der Textprozeduren eine handlungsorientierte Perspektive auf Texte einbezogen und es werden kleinere Texteinheiten annotiert, zu denen das System später differenziert Feedback ausgeben kann.

## 3.3 | Inter-Annotator-Agreement

Nach einer Schulung der Annotatorinnen erfolgte im Herbst 2022 zunächst die Pilotstudie mit 120 Texten, später die Hauptstudie mit 1200 Texten. An der Pilotierung waren drei Annotatorinnen beteiligt, die in die Nachschärfung der Annotationsrichtlinien involviert wurden, an der Hauptstudie noch zwei von diesen, die jeweils 600 Texte annotierten. Dafür war ein hohes Inter-Annotator-Agreement (IAA) in der Pilotstudie nötig. Nachfolgend werden die Ergebnisse dieses IAA basierend auf Krippendorff's α gezeigt, wobei 1,0 vollständige Übereinstimmung bedeutet, 0 eine systematische Nichtübereinstimmung. 0,8 gilt als Richtwert für ein ausreichend hohes Agreement (vgl. Krippendorff 2004). Kursiv markiert ist jeweils der niedrigste Kategorie-Wert je Ebene:

| *Makro 1: Annotationskategorie* | α |
|---|---|
| Einleitung | 0,99 |
| Hauptteil | *0,89* |
| Schluss | 0,94 |

Tabelle 1: Inter-Annotator-Agreement (IAA) der Makrostrukturebene 1 nach Krippendorff's α

| *Makro 2: Annotationskategorie* | α |
|---|---|
| Argument | *0,86* |
| Gegenargument | 1,00 |

Tabelle 2: Inter-Annotator-Agreement (IAA) der Makrostrukturebene 2 nach Krippendorff's α

| *Mikro 1: Annotationskategorie* | α |
|---|---|
| Thema | 0,98 |
| These | 0,88 |
| Gegenthese | 0,98 |
| Modifizierte These | 0,89 |
| Claim | 0,81 |
| Premise | *0,78* |

Tabelle 3: Inter-Annotator-Agreement (IAA) der Mikrostrukturebene 1 nach Krippendorff's α

| *Mikro 2: Annotationskategorie* | α |
|---|---|
| Positionieren | 0,87 |
| Beschreiben | *0,71* |
| Exemplifizieren | 0,94 |
| Begründen | 0,75 |
| Konzedieren | 0,95 |
| Referieren | 0,97 |
| Abwägen | 0,96 |
| Auffordern | 0,90 |
| Einschränken | 0,88 |
| Schlussfolgern | 0,74 |

Tabelle 4: Inter-Annotator-Agreement (IAA) der Mikrostrukturebene 2 nach Krippendorff's α

Für eine Diskussion und Interpretation der Ergebnisse sei an dieser Stelle auf Kilsbach et al. (2025) verwiesen. Der Richtwert von 0,8 wurde aber bei allen Kategorien erreicht, mit Ausnahme von der Premise auf Mikrostrukturebene 1 sowie den Textprozeduren Beschreiben, Begründen und Schlussfolgern auf Mikrostrukturebene 2. Damit war es möglich, in die Hauptstudie überzugehen.

## 3.4 | Text-Rating

Die Bestimmung der Textqualität ist ein herausforderndes Problem innerhalb der Forschung an NLP und entsprechenden Verfahren (vgl. Stede/Schneider 2018, 113). Da im vorliegenden Projekt nur die Zielsetzung verfolgt wird, Feedback zur Struktur (und nicht zur inhaltlichen

Qualität) argumentativer Texte zu geben, war es für die Formulierung von Rating-Richtlinien wichtig, die Rating-Items anhand der Textstruktur, nicht dem semantischen Inhalt auszurichten.

Bei dem Rating wurden die Texte in einem ersten Schritt holistisch eingeschätzt, um generelle Aussagen über die Textqualität und den Gesamtcharakter des Textes zu erhalten (vgl. Kruse et al. 2012, 93; Lindauer 2024, 93). Hier galt es einzuschätzen, ob der Text eine globale Textstruktur (Einleitung, Hauptteil, Konklusion) aufweist, eine Antizipation der Leser*in zu erkennen ist und ob Gegenargumente entkräftet bzw. Argumente abgewägt werden.

Darüber hinaus wurden die Texte in einem weiteren Schritt bezogen auf vier Textdimensionen eingeschätzt, um zusätzlich zum Gesamturteil differenziertere Aussagen zu den einzelnen Dimensionen zu erhalten. Dieses semi-holistische Rating (vgl. Lindauer 2024, 93) bezog sich auf folgende Dimensionen:

1 | Die *Textfunktion* betrifft die Angemessenheit der Texte bezogen auf das Schreibziel und die Adressat*innen.
2 | Die *Textstruktur* betrifft die Gesamtstruktur des Textes und die schlüssige Beziehung der argumentativen Einheiten im Text. Im Fokus steht die Frage, ob die einzelnen Komponenten in einer kohärenten Beziehung zueinanderstehen und textstrukturell ein Gesamtbild über die jeweilige Schreibaufgabe zulassen.
3 | Bei der *inhaltlichen Ausgestaltung* wurde eingeschätzt, inwiefern die gewählten Argumente und Gegenargumente angemessen bezogen auf die Stützung der im Text vertretenen Position sind und zum Standpunkt passen.
4 | Bei der *sprachlichen Gestaltung* geht es um die Angemessenheit der Wortwahl, des Satzbaus und der verwendeten argumentationstypischen Sprachmittel, d. h. der Prozedurenausdrücke.

In Anlehnung an das Vorgehen bei Persing et al. (2010) wurde bei beiden Ratingschritten das gesamte Korpus auf einer 7-Punkte-Skala von 1 (nicht gelungen), 2 (eher nicht gelungen), 3 (eher gelungen) bis 4 (vollständig gelungen) geratet. Wie bei Persing et al. (2010) wurde dabei mit Zwischenstufen (1,5; 2,5; 3,5) gearbeitet, die genutzt wurden, wenn keine eindeutige Zuordnung zu den Stufen 1 bis 4 möglich war. Beim semi-holistischen Rating wurden entsprechende Deskriptoren für die vier Haupt-Qualitätsstufen entwickelt.

Zu Beginn eines jeden Text-Ratings wurde zunächst auch die sprachliche Richtigkeit bewertet, um zu minimieren, dass morphosyntaktische Fehler die Bewertung der anderen Indikatoren beeinflussen; das Scriptoria-Korpus ist orthografisch bereinigt. Diese Werte flossen als Null-Rating aber nicht in die Datenanalyse ein. Als Bewertende fungierten die beiden Annotatorinnen der Hauptstudie, die gezielt jene 600 Texte bekamen, die während der Annotation von der jeweils anderen Annotatorin bearbeitet wurden.

In Tabelle 5 werden exemplarisch nur die Ergebnisse des Gesamteindrucks aus dem holistischen Rating präsentiert, da dieser im Folgenden bei der Modellierung des Feedbacks die Grundlage bildet. Die erste Zeile gibt die Qualitätsstufe mit Zwischenschritten an, die zweite Zeile die Anzahl der Texte, die der jeweiligen Stufe zugeordnet wurde.

| *Ratingstufe* | 1,0 | 1,5 | 2,0 | 2,5 | 3,0 | 3,5 | 4,0 |
|---|---|---|---|---|---|---|---|
| *Textanzahl* | 90 | 142 | 478 | 390 | 195 | 23 | 2 |

Tabelle 5: Text-Rating des Gesamteindrucks

Ein zentrales Ergebnis des holistischen Ratings ist darin zu sehen, dass nur sehr wenige Texte in die Qualitätsstufen 3,5 und 4, und damit als „vollständig gelungen", eingestuft wurden. Der Großteil der Texte wurde in den Bereich „eher nicht gelungen" (Stufe 2 und 2,5) eingeordnet. Der Mittelwert war bei Texten der 9. Klassenstufe mit 2,42 wesentlich höher als bei denen der 5. Klasse (1,98).

## 3.5 | Ableitung von Feedback auf Grundlage von Textannotation und Rating

Es wurde bereits deutlich, dass die Strukturannotation und das Rating des Textkorpus die zentralen Bezugspunkte für die Generierung des automatisierten Feedbacks darstellen. Bereits anhand dieser Daten ist es möglich, Feedback zur Struktur der argumentativen Texte abzuleiten. Bevor dies anhand eines Beispiels exemplarisch gezeigt wird, ist es wichtig zu verstehen, welches Verständnis von einem lernförderlichen Feedback wir haben.

Lipnevich und Panadero (2021) und Graham et al. (2015) verstehen unter Feedback einen Prozess, in dem ein*e Feedbackgeber*in einer*einem Feedbackempfänger*in verschiedene Informationen bereitstellt, und zwar, wo der*die Lernende steht, was die Ziele des Feedbacks sind und wie diese Ziele erreicht werden können. Ziel von Feedback ist es, dass der*die Empfänger*in durch die Informationen die Diskrepanz zwischen der aktuell erbrachten Leistung und der erwünschten Leistung minimiert (vgl. Lipnevich/Panadero 2021, 24 f.).

In der Feedbackforschung sind zentrale Merkmale eines lernförderlichen Feedbacks herausgearbeitet worden (vgl. Busse/Siekmann 2023). Feedback sollte formativ ausgerichtet sein und auf Feed-Forward-Strategien beruhen (vgl. Wisniewski et al. 2020; Narciss/Zumbach 2022). Außerdem sollte es Informationen zur Unterstützung von Prozesswissen, Sprachwissen und metakognitivem Wissen enthalten (vgl. Narciss/Zumbach 2022; Sturm 2024). Uns ist hier insbesondere das Sprachwissen wichtig, genauer gesagt eine Unterstützung auf der Ebene von Textprozeduren, da in schreibdidaktischen Untersuchungen gezeigt wurde, dass sich eine Überarbeitung von Texten mithilfe von Textprozeduren als Feedbackkriterien positiv auf die Qualität der Textüberarbeitung auswirkt (vgl. Anskeit 2019; Schicker 2020). Des Weiteren ist die Adaptivität des Feedbacks, d. h. die Ausrichtung des Feedbacks an den Lernstand des jeweiligen Feedbacknehmenden, zentral. Adaptives Feedback zielt gleichzeitig auf eine nächste Entwicklungsstufe im Sinne des Feed Forward ab (vgl. Hattie/Timperley 2007). Diese Merkmale würden nicht erfüllt, wenn man Lernenden alle Informationen zum Lernstand gibt, die ein System automatisiert erkennt. Dies möchten wir im Folgenden anhand eines Beispiels exemplarisch zeigen.

Im folgenden Text sollte sich die Schülerin zur Frage positionieren, ob offene Schulgelder lieber für Sessel und Sofas oder für Sport- und Spielgeräte ausgegeben werden sollten. Die Aufgabe ist persuasiv ausgerichtet und an eine Lehrerin gerichtet.

> Liebe Frau Grotje, ich nehme den zweiten Vorschlag, weil nach den Unterricht ist man in Stress und soll wieder runterkommen und dann gibt es keine Prügelei zu gegen den den anderen Vorschlag. Stell Ihnen mal vor in einer Ganztagsschule Sie werden nicht überleben nich mal 6. Stunden. Also mein Vorschlag war das. Liebe Grüße XXX
> Textbeispiel (1) (FD-LEX-ID: 0H56AA16, IGS 5)

Der Text wurde eher schlecht bewertet (Gesamt: 1,5; Textfunktion: 1,5; inhaltliche Ausgestaltung: 2; Struktur: 1,5; sprachliche Ausgestaltung: 2). Auf Basis des Ratings wäre bereits ein nummerisches Feedback möglich. Ebenso wäre ein Feedback zur Textstruktur auf Grundlage der im Projekt entwickelten NLP-Verfahren, mit deren Hilfe Strukturen des Textes automatisch ermittelt werden, möglich.[2] Tabelle 6 zeigt diese Strukturen für Beispiel 1. In den vier Spalten rechts ist die annotierte Makro- und Mikrostruktur verzeichnet.

2 Im Projekt ist bereits ein Prototyp entwickelt worden, mit dem es möglich ist, die Textstrukturen argumentativer Texte automatisiert zu ermitteln.

| *Beispiel (1)* | *Makro 1* | *Makro 2* | *Mikro 1* | *Mikro 2* |
|---|---|---|---|---|
| Liebe Frau Grotje | – | – | – | – |
| ich nehme den zweiten Vorschlag | Hauptteil | – | These | Positionieren |
| weil nach den Unterricht ist man in Stress und soll wieder runterkommen | Hauptteil | Argument 1 | Claim | Begründen |
| und dann gibt es keine Prügelei zu gegen den den anderen Vorschlag. | Hauptteil | Argument 2 | Claim | Schlussfolgern |
| Stelle Ihnen mal vor in einer Ganztagsschule. Sie werden nicht überleben nich mal 6. Stunden. | Hauptteil | Argument 2 | Premise | Beschreiben |
| Also mein Vorschlag war das. | – | – | – | – |
| Liebe Grüße ... | – | – | – | – |

Tabelle 6: Textbeispiel (1) mit Annotationsangaben (FD-LEX, 0H56AA16, IGS 5)

Dieser Text besteht ausschließlich aus einem Hauptteil. Aufgrund dieser Informationen kann passgenaues Feedback gegeben werden, indem etwa angeführt wird, dass der Text schon einen Hauptteil enthält, aber noch Einleitung und Konklusion ergänzt werden müssen. Die Formulierung „Also mein Vorschlag war das" weist zwar den Ansatz oder die Absicht einer Konklusion auf, wurde gemäß der engen Annotationsrichtlinien aber nicht als Konklusion annotiert, da weder die These wiederholt noch eine abschließende Bewertung vorgenommen wird. Des Weiteren finden sich in dem Text zwei Argumente, aber kein Gegenargument. Hier ließe sich rückmelden, dass gelungen ist, dass schon Pro-Argumente genannt werden. Anregen könnte man, den Text überzeugender zu gestalten, indem Gegenargumente angeführt und entkräftet werden. Ähnlich könnte man auf Mikroebene mit These, Claim, Premise und den Textprozeduren verfahren. Dann würde man ein sehr umfassendes Feedback zu allen annotierten Textebenen und -kategorien geben, um aufzuzeigen, wo die Lernende steht. Dieses Feedback wäre allerdings noch nicht adaptiv und entspräche damit nicht dem Feedbackverständnis, das diesem Beitrag zugrunde liegt und oben dargelegt wurde.

# 4 | Varianzanalysen: Ableitung von Niveaustufen als Grundlage für adaptives Feedback

## 4.1 | Methodisches Vorgehen

Um ein lernförderliches und adaptives Feedback zu generieren, ist es erforderlich, das Feedback an bestimmten Niveaustufen zu orientieren. Für die Ableitung von Niveaustufen ist es wiederum notwendig, auf Grundlage des annotierten und mit Ratings versehenen Korpus entsprechende textuelle Merkmale für bestimmte Niveaustufen festzumachen. Im vorliegenden Projekt wurden Varianzanalysen durchgeführt, um entsprechende Niveaustufen und niveaustufenbezogene Textmerkmale abzuleiten. Auf diese Weise kann bestimmt werden, welche der annotierten Textstrukturen charakteristisch für eine bewertete Textqualität sind. Bei einer solchen ANOVA (*analysis of variance*) wird geprüft, ob signifikante Unterschiede zwischen mehr als zwei Gruppen vorliegen (vgl. Koster/Albert 2002, 129). In unserem Fall geht es um die Frage, ob das Vorkommen einer annotierten Textstruktur (unabhängige Variable), z. B. die Häufigkeit von Argumenten, signifikant unterschiedlich ist zum Vorkommen in allen anderen Texten anderer Textqualität (abhängige Variable). Dafür wurde der Mann-Whitney U-Test (vgl. Rasch et al. 2021) angewendet. Ergebnis dieser Signifikanzanalyse ist der p-Wert. Liegt dieser unter 0,05, ist das ein Indikator, dass das Vorkommen einer Kategorie sich signifikant vom

Vorkommen in anderen Qualitätsstufen unterscheidet. Bei $p<0{,}05$ ist „die Wahrscheinlichkeit, dass der Zufall die Ursache für unser Resultat ist“, weniger als 1:20; bei $p<0{,}01$ nur 1:100 und so weiter (Koster/Albert 2002, 135 f.). In Kombination mit der durchschnittlichen Häufigkeit der Kategorie pro Text lässt sich ermitteln, ob und wie sich das Vorkommen einer Kategorie innerhalb der Qualitätsstufe verhält. Dies ist schließlich ein Indikator dafür, welche strukturellen Unterschiede zwischen den Texten der einzelnen Ratingwerte bestehen, womit man einem adaptiven Feedback näherkommt.

Im Kontext des Projekts wurden Varianzanalysen zu allen annotierten Kategorien durchgeführt. Im Folgenden beschränken wir uns exemplarisch auf die Ergebnisse der Varianzanalysen für zwei ausgewählte annotierte Textstrukturen: die (Anzahl der) *Argumente* in Texten (Makroebene 2) sowie die Textprozedur *Konzedieren* (Mikroebene 2).

Vorab ist ein Hinweis zum methodischen Vorgehen wichtig: Aufgrund der geringen Textmenge auf den höchsten Qualitätsstufen (vgl. Tab. 5) wurden die Qualitätsstufen des Ratings an den Rändern zusammengezogen,[3] so dass insgesamt nur fünf anstatt der ursprünglich sieben Niveaustufen unterschieden werden.

## 4.2 | Ergebnisse der Varianzanalysen für die Kategorie *Argument*

In Arbeiten zum Erwerb argumentativen Schreibens stellt die Anzahl von Argumenten in einem Text eine relevante Größe dar, um Erwerbsstufen voneinander abzugrenzen (vgl. z. B. Augst/Faigel 1986). Tabelle 7 zeigt die Ergebnisse der ANOVA für die Kategorie *Argument*.

| *Qualitätsstufe (Gesamtrating)* | *Anzahl Texte* | *Anzahl Wörter pro Text Ø* | *Argument (p-Wert)* | *Argument (Vorkommen ⌀ pro Text)* |
|---|---|---|---|---|
| 1–1,5 | 232 | 35 | <0,00001 | 1,2500 |
| 2 | 478 | 47 | <0,00001 | 1,6925 |
| 2,5 | 390 | 68 | <0,00001 | 2,3718 |
| 3 | 195 | 93 | <0,00001 | 2,9744 |
| 3,5–4 | 25 | 119 | <0,00001 | 3,5200 |

Tabelle 7: ANOVA zur Kategorie *Argument* (bezogen auf das Rating der Gesamtqualität der Texte)

In Spalte eins werden die Qualitätsstufen der Texte gezeigt. Die Textqualität steigt von Stufe 1 zu Stufe 4. In der zweiten Spalte ist die Gesamtzahl der Texte pro Qualitätsstufe angegeben. Spalte drei gibt die Anzahl der Wörter pro Text an und Spalte 4 den ermittelten p-Wert nach dem Mann-Whitney U-Test wieder. Die fünfte Spalte zeigt schließlich das durchschnittliche Vorkommen der Annotationskategorie pro Qualitätsstufe, d. h. macht Aussagen darüber, wie viele Argumente pro Qualitätsstufe im Mittel vorliegen.

Das durchschnittliche Vorkommen der Argumenteinheiten pro Text steigt sukzessive von 1,25 (auf der Qualitätsstufe 1–1,5) auf 3,52 (auf Qualitätsstufe 3,5–4). Der p-Wert ist auf allen Qualitätsstufen niedrig und zeigt, dass die Unterschiede zwischen jeweils einer Qualitätsstufe im Vergleich zu den anderen Qualitätsstufen statistisch signifikant und nicht zufällig sind. Anhand dieser Daten lassen sich Rückschlüsse auf den Zusammenhang von Argument-Anzahl und Qualität von Texten ziehen: Eine höhere Zahl an Argumenten geht im Durchschnitt mit einer höheren durchschnittlichen Wortanzahl sowie einer höheren Qualität eines argumentativen Textes einher, und der Anteil der Argumente pro Qualitätsstufe unterscheidet sich signifikant von den anderen Qualitätsstufen.

Das adaptive Feedback kann man nun auf dieser Grundlage modellieren. Bei Texten, die auf der Qualitätsstufe 3 eingestuft werden, könnte das Feedback z. B. dahingehend sein, dass Aussagen zum Lernstand (Feed Back) gemacht werden, z. B. „Du hast bereits zwei Argumente in deine Argumentation eingebaut“. Und es können Aussagen zum nächsten Lernniveau

3 Dies betrifft die Qualitätsstufen 1 und 1,5 sowie die Qualitätsstufen 3,5 und 4 (vgl. Abschnitt 3.5).

(Qualitätsstufe 3,5–4), das erreicht werden soll (Feed Forward), gemacht werden: „Baue noch ein weiteres, drittes Argument in den Text ein."

### 4.3 | Ergebnisse der Varianzanalysen zur Kategorie *Konzedieren*

Das Konzedieren gilt in der Forschungsliteratur als Gradmesser für die Qualität elaborierter argumentativer Texte (Coirier et al. 1999, 8). Schreibdidaktische Untersuchungen zeigen, dass sich die Einbeziehung von Gegenargumenten erst in Texten älterer Schreibender (ca. ab dem 14. Lebensjahr) zeigt (vgl. Augst/Faigel 1986; Coirier/Golder 1993) oder bestimmte Aufgabenkontexte vorhanden sein müssen, um das Konzedieren schon bei jüngeren Schreibenden anzuregen (vgl. Rezat 2011). Die Ergebnisse der ANOVA für die Kategorie *Konzedieren* bestätigen diese Befunde.

| *Qualitätsstufe Gesamtrating* | *Anzahl Texte* | *Anzahl Wörter pro Text Ø* | *Konzedieren (p-Wert)* | *Konzedieren (Anteil pro Text)* |
|---|---|---|---|---|
| 1–1,5 | 232 | 35 | 0,000653 | 0,0345 |
| 2 | 478 | 47 | 0,000240 | 0,0607 |
| 2,5 | 390 | 68 | 0,045222 | 0,1359 |
| 3 | 195 | 93 | 0,000005 | 0,2205 |
| 3,5–4 | 25 | 119 | 0,000085 | 0,3600 |

Tabelle 8: ANOVA zur Kategorie *Konzedieren* (bezogen auf das Rating der Gesamtqualität der Texte)

Der Anteil des Konzedierens pro Text ist in allen Qualitätsstufen sehr gering. Er steigt aber mit zunehmender Textqualität an: von 0,035 in Texten der Qualitätsstufe 1–1,5 auf 0,36 in Texten der Qualitätsstufe 3,5–4. Der p-Wert liegt stets unter dem Signifikanzniveau von 0,05; auf der Qualitätsstufe 2,5 ist die Signifikanz am niedrigsten. Grundsätzlich zeigen die Daten, dass ein steigender Anteil des Konzedierens in den Texten mit spezifischen Qualitätsstufen korreliert. Für die Frage, wie man sinnvoll Feedback geben kann, ist es wichtig, die Mittelwerte (Anteil des Konzedierens pro Text) einzubeziehen. Dieser liegt selbst bei der höchsten Qualitätsstufe unter 1. Das Konzedieren kommt also auch auf dieser Stufe nur in wenigen Texten vor. Feedback bezogen auf die Kategorie *Konzedieren* ist demnach erst bei Texten sinnvoll und adaptiv, die eine hohe Qualitätsstufe (3 bis 4) aufweisen. Dafür sprechen auch Ergebnisse aus Erwerbsstudien zum Konzedieren (vgl. Leitao 2003; Rezat 2011). Texte, die auf der Qualitätsstufe 1 bis 2,5 sind, würden demnach kein Feedback zum Konzedieren erhalten.

## 5 | Didaktische Modellierung und computerlinguistische Generierung automatisierten adaptiven Feedbacks

Abschließend wird skizziert, wie bei der konkreten Modellierung eines später automatisierten Feedbacks vorgegangen wird. Hier sind die computerlinguistische Seite und die schreibdidaktische Vorgehensweise mit den Informationen, die im Zug der Annotation, des Text-Ratings und der Varianzanalysen vorliegen, in Einklang zu bringen. Das Kapitel zeigt, wie im Projekt methodisch ein Feedback-Korpus erstellt wird, und gibt Beispiele aus dem aktuellen Vorgehen.

Es wurde bereits erwähnt, dass zu allen Strukturannotationen und Ratingergebnissen, die in den voranstehenden Kapiteln erläutert wurden, Feedback gegeben werden könnte. Ein solches umfassendes Feedback wäre allerdings aus didaktischer Perspektive nicht lernförderlich und adaptiv. In der praktischen Anwendung könnte es viel eher demotivierend oder überfordernd für Lernende sein, wenn zu viele Feedback-Aspekte angeführt werden und nicht bezogen auf den Lernstand Feedback gegeben wird (vgl. Deeva et al. 2021).

Computerlinguistisch verhält es sich aber so, dass für die spätere automatisierte Generierung zunächst alle Feedbackbausteine vorliegen müssen. Aus diesem „Feedback-

Baukasten“ wählt das System nach einer Überprüfung der Niveaustufe nur diejenigen relevanten Bausteine aus, die didaktisch angemessen bezogen auf den Lernstand und die folgende Niveaustufe sind. Bei der manuellen Modellierung des „Feedback-Baukastens“ werden Entscheidungsbäume herangezogen, die sich auf die annotierten Strukturen beziehen. Die Baumstruktur wird mit entsprechenden Leitfragen zum Vorliegen bzw. Nicht-Vorliegen einer spezifischen Textstruktur verbunden. Auf diese Weise wird eruiert, welche verfügbaren Informationen zur Textstruktur sich auf welcher Ebene lokalisieren lassen.

Schreibdidaktisch zentral ist für die Modellierung des Feedbacks, zwischen starken und schwachen Schreibenden zu differenzieren (vgl. Busse/Siekmann 2023) und unterschiedliches Feedback zu formulieren. Stärkere Schreibende sollten eher metakognitiv und auf der Ebene der Selbstregulation angeregt werden (z. B. durch Fragen). Schwächere benötigen konkrete Hinweise bezogen auf die Aufgabenebene und das Sprachwissen. Beim Sprachwissen unterscheiden wir Feedback zu textprozeduralen Handlungsschemata (z. B. „Ergänze in deinem Text eine Einleitung, in der du das Thema der Argumentation nennst und warum du die Argumentation schreibst.“) und Feedback zu Textprozedurausdrücken, d. h. zu konkreten Formulierungen (z. B. „Ich möchte mich zur Frage äußern, ob...“, „Ich nehme heute Stellung zu...“). Diese methodischen Vorüberlegungen gelten für alle Strukturebenen.

Nachfolgend wird dieses Vorgehen bezogen auf die Annotationskategorie *Einleitung* etwas differenzierter erläutert. Abbildung 4 zeigt den entsprechenden Entscheidungsbaum.

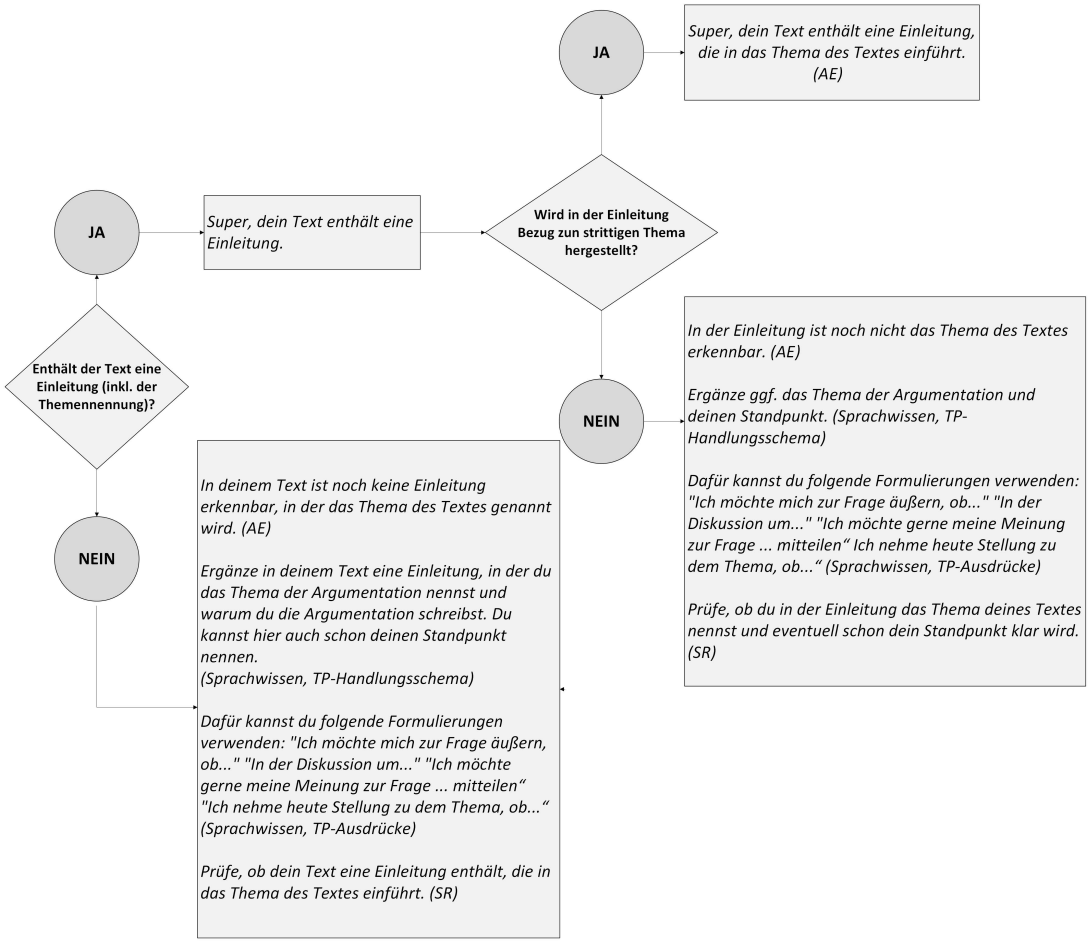


Abbildung 4: Entscheidungsbaum für die Annotationskategorie Einleitung (inkl. Themennennung) auf Makrostrukturebene 1 mit Leitfrage sowie Ja- und Nein-Ableitungen zu Feedback-Bausteinen.

Das konkrete Feedback bei einer Bejahung zur Frage nach dem Vorhandensein einer Einleitung könnte lauten: „Super, dein Text enthält eine Einleitung, die in das Thema des Textes einführt.“ Ist keine Einleitung vorhanden, kann das Feedback je nach Feedback-Ebene unterschiedlich ausgestaltet und formuliert werden. In dem Feedback-Baukasten in Abbildung 4 werden

Feedback-Bausteine zur Aufgabenebene (AE), zur Selbstregulation (SR), zum Sprachwissen bezogen auf Handlungsschemata von Textprozeduren sowie zum Sprachwissen bezogen auf Textprozedurausdrücke unterschieden und angeboten.

Welche Feedback-Ebene im konkreten Fall bei einer fehlenden Einleitung ausgewählt wird, hängt vom automatisierten Rating des Textes, dem (Nicht-)Vorhandensein weiterer spezifischer Textstrukturen und den Varianzanalysen ab.

Textbeispiel (1) (vgl. Kapitel 4) enthält keine Einleitung, wurde eher schwach bewertet und hat auf sehr vielen Ebenen strukturelle Lücken. In diesem Fall könnte der Lernerin Feedback zur Aufgabenebene in Kombination mit Feedback zum Sprachwissen gegeben werden. Für schwache Schreibende ist es wichtig, dass sie möglichst konkrete Hinweise zum Formulieren bekommen, gleichzeitig sollten sie auch Hinweise zu den relevanten Handlungsschemata erhalten (vgl. Rüßmann et al. 2016). Es würde also eine Kombination der entsprechenden Feedbackbausteine ausgewählt. Feedback auf der Ebene der Selbstregulation würde in diesem Fall nicht gegeben, weil diese Feedbackebene eher für starke Schreibende geeignet ist (vgl. Busse/Siekmann 2023).

Welches weitere Feedback zu Textbeispiel (1) gegeben würde, hängt von den Varianzanalysen ab (vgl. Kapitel 4.1). Bezogen auf Textbeispiel (1) ergibt sich aufgrund der Varianzanalysen, dass auf dieser Niveaustufe Feedback zum Argumentausbau (Anzahl der Argumente) gegeben wird, aber kein Feedback zur fehlenden Einleitung und Konklusion.

## 6 | Fazit

Im vorliegenden Artikel wurde am Beispiel argumentativer Lernendentexte demonstriert, wie ein automatisiertes adaptives Feedback aus schreibdidaktischer und computerlinguistischer Perspektive modelliert werden kann. Voraussetzung für eine solche Modellierung ist im ersten Schritt eine Annotation eines umfassenden Textkorpus sowie ein Rating der Texte. Diese Schritte bilden die Grundlage für ein Mining der Texte und eine Ableitung von Niveaustufen der Textqualität auf Grundlage von Varianzanalysen. Im Beitrag wurden Ergebnisse der Varianzanalysen vorgestellt, die zeigen, dass das Auftreten bestimmter argumentativer Textstrukturen (z. B. die Anzahl von Argumenten, das Auftreten des Konzedierens) mit bestimmten Niveaustufen in Zusammenhang steht. Davon ausgehend wurde die konkrete Modellierung adaptiven Feedbacks skizziert. In weiteren Projektschritten wird es darum gehen, die Varianzanalysen bezogen auf die verschiedenen Schreibaufgaben des Korpus auszuwerten. Geplant ist auch, in qualitativen Settings zu untersuchen, wie sich adaptives Feedback auf die Qualität von Texten auswirkt.

Das vorgestellte Vorgehen versteht sich als ein Beitrag zur Entwicklung domänenspezifischer Feedbacktools. Für den schulischen Kontext kann zwar auf ChatGPT oder spezifische Feedback-Tools (vgl. Haverkamp et al. 2024) zurückgegriffen werden, doch diese Tools sind nicht mit spezifischen Schüler*innendaten trainiert worden, so dass die Adaptivität des Feedbacks bezogen auf eine spezifische Textsorte und Kompetenzniveaus kaum gegeben ist. Welche spezifischen schreibdidaktischen Konzepte den Tools zugrunde liegen, ist auch unklar. Die Ausführungen in diesem Beitrag zeigen freilich, wie komplex und zeitaufwändig eine Modellierung und Umsetzung adaptiven automatisierten Feedbacks ist, weil eine entsprechende Annotation und ein Rating eines umfangreichen Textkorpus erfolgen müssen. Anzunehmen ist aber, dass diese Schritte in Zukunft durch entsprechende KIs übernommen werden können. Damit gehen aber zugleich verschiedene Herausforderungen einher. Für weitere Forschungsvorhaben in diesem Bereich sehen wir eine zentrale Herausforderung darin, dass kaum öffentlich zugängliche umfangreiche Textkorpora mit deutschsprachigen Lernendentexten für andere Textformen (z. B. Beschreiben, Erklären, Erzählen) und damit zusammenhängende Textsorten zur Verfügung stehen. Dies wäre aber wiederum eine zentrale Voraussetzung, um domänenspezifische Feedback-Tools zu entwickeln. Hinzu kommt, dass derartige Vorhaben zwangsläufig einen interdisziplinären Zugang erfordern, in dem Expertise zum Schreiberwerb und zum Textfeedback mit computerlinguistischer Expertise verbunden werden. Trotz der genannten

Herausforderungen halten wir die im Beitrag dargelegten Schritte für ähnliche Vorhaben aber gut adaptierbar.

## 7 | Literaturverzeichnis


Anskeit, Nadine (2019): Schreibarrangements in der Primarstufe. Eine empirische Untersuchung zum Einfluss der Schreibaufgabe und des Schreibmediums auf Texte und Schreibprozesse in der 4. Klasse. Münster: Waxmann.

Augst, Gerhard / Faigel, Peter (1986): Von der Reihung zur Gestaltung. Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13–23 Jahren. Frankfurt am Main: Lang.

Becker-Mrotzek, Michael / Grabowski, Joachim (2018): Textkorpus Scriptoria. In: Becker-Mrotzek, Michael / Grabowski, Joachim (Hg.): FD-LEX (Forschungsdatenbank Lernertexte). Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache. Online unter https://fd-lex.uni-koeln.de (letzter Aufruf 25. Oktober 2024).

Benetos, Kalliopi (2023): Digital Tools for Written Argumentation. In: Kruse, Otto / Rapp, Christian / Anson, Chris M. / Benetos, Kalliopi / Cotos, Elena / Devitt, Ann / Shibani, Antonette (Hg.): Digital Writing Technologies in Higher Education. Theory, Research, and Practice. Cham: Springer, S. 81–99. https://doi.org/10.1007/978-3-031-36033-6

Benetos, Kalliopi / Bétrancourt, Mireille (2020): Digital Authoring Support for Argumentative Writing. What does it change? In: Journal of Writing Research, 12, H. 1, S. 263–290. https://doi.org/10.17239/jowr-2020.12.01.09

Busse, Vera / Siekmann, Lea (2023): Process-oriented writing and formative feedback in EFL classes. A comparison of teachers' and learners' perceptions. In: Wilden, Eva / Alfes, Luisa / Cantone, Katja F. / Çıkrıkçı, Sevgi / Reimann, Daniel (Hg.): Standortbestimmungen in der Fremdsprachenforschung. Bielefeld: wbv Media, S. 274–290.

Coirier, Pierre / Andriessen, Jerry / Chanquoy, Lucile (1999): From Planning to Translating. The Specificity of Argumentative Writing. In: Andriessen, Jerry / Coirier, Pierre (Hg.): Studies in Writing: Vol. 5. Foundations of Argumentative Text Processing. Amsterdam: Amsterdam University Press, S. 1–28.

Coirier, Pierre / Golder, Caroline (1993): Writing Argumentative Text. A Developmental Study of the Acquisition of Supporting Structures. In: European Journal of Psychology of Education, 8, H. 2, S. 169–181.

Deeva, Galina / Bogdanova, Daria / Serral, Estefania / Snoeck, Monique / De Weerdt, Jochen (2021): A review of automated feedback systems for learners. Classification framework, challenges and opportunities. In: Computers & Education, 162, Art. 104094. https://doi.org/10.1016/j.compedu.2020.104094

Feilke, Helmuth (2024): Textprozeduren: erkennen, erwerben, fördern. In: Rezat, Sara / Grundler, Elke / Feilke, Helmuth / Schmölzer-Eibinger, Sabine (Hg.): Textprozeduren in Spannungsfeldern. Tübingen: Stauffenburg, S. 9–52.

Feilke, Helmuth / Rezat, Sara (2021): Textprozeduren und der Erwerb literaler Kompetenz. In: Der Deutschunterricht, 73, H. 5, S. 69–79.

Graham, Steve / Hebert, Michael / Harris, Karen R. (2015): Formative Assessment and Writing. A Meta-Analysis. In: The Elementary School Journal, 15, H. 4, S. 523–547.

Hattie, John / Timperley, Helen (2007): The Power of Feedback. In: Review of Educational Research, 77, H. 1, S. 81–112.

Haverkamp, Henrik / Hecht, Malte / Schindler, Kirsten (2024): Feedback KI-basiert vermitteln. In: Der Deutschunterricht, 76, H. 5, S. 60–71.

Kilsbach, Sebastian / Rezat, Sara / Michel, Nadine / Karabey, Rabia / Stahl, Maja / Wachsmuth, Henning (2025): Mehrebenenannotationen argumentativer Lerner*innentexte für die automatische Textauswertung. In: Zeitschrift für Angewandte Linguistik, H. 82, S. 102–129.

KMK (2024): Large Language Models und ihre Potenziale im Bildungssystem. Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz. Online unter https://www.kmk.org/fileadmin/Dateien/pdf/KMK/SWK/2024/SWK-2024-Impulspapier_LargeLanguageModels.pdf (letzter Aufruf 25. Oktober 2024).

Koster, Cor J. / Albert, Ruth (2002): Empirie in Linguistik und Sprachlehrforschung. Ein methodologisches Arbeitsbuch. Tübingen: Narr.

Krippendorff, Klaus (2004): Measuring the Reliability of Qualitative Text Analysis Data. In: Quality & Quantity, 38, H. 6, S. 787–800.

Kruse, Norbert / Reichardt, Anke / Herrmann, Maik / Heinzel, Friederike / Lipowsky, Frank (2012): Zur Qualität von Kindertexten. Entwicklung eines Bewertungsinstruments in der Grundschule. In: Didaktik Deutsch, 17, H. 32, S. 87–110.

Lehnen, Katrin (2023): Kooperatives digitales Schreiben. Ko-Konstruktion, Feedback und Kommentar zwischen sozialer und automatisierter Textproduktion. In: Der Deutschunterricht, 75, H. 5, S. 18–28.

Leitao, Selma (2003): Evaluating and Selecting Counterarguments. Studies of Children's Rhetorical Awareness. In: Written Communication, 20, H. 3, S. 269–306.

Lindauer, Nadja (2024): Effektivität und Effizienz von holistischen Benchmarkratings. In: Petersen, Inger / Reble, Raja / Kilian, Jörg (Hg.): Texte schreiben in allen Unterrichtsfächern. Textbeurteilung als Grundlage für Schreibförderung und Leistungsbewertung. Münster: Waxmann, S. 91–111.

Lipnevich, Anastasiya A. / Panadero, Ernesto (2021): A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. In: Frontiers in Education, 6, Art. 720195, S. 1–29. https://doi.org/10.3389/feduc.2021.720195

Narciss, Susanne / Zumbach, Jörg (2022): Formative Assessment and Feedback Strategies. In: Zumbach, Jörg / Bernstein, Douglas / Narciss, Susanne / Marsico, Giuseppina (Hg.): International Handbook of Psychology Learning and Teaching. Cham: Springer, S. 1359–1386.

Persing, Isaac / Davis, Alan / Ng, Vincent (2010): Modeling Organization in Student Essays. In: Li, Hang / Màrquez, Lluís (Hg.): Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA: Association for Computational Linguistics, S. 229–239. Online unter: https://aclanthology.org/D10-1023.pdf (letzter Aufruf 25. Oktober 2024).

Rasch, Björn / Friese, Malte / Hofmann, Wilhelm / Naumann, Ewald (2021): Quantitative Methoden 2. Einführung in die Statistik für Psychologie, Sozial- & Erziehungswissenschaften. Heidelberg: Springer.

Rezat, Sara (2011): Schriftliches Argumentieren. Zur Ontogenese konzessiver Argumentationskompetenz. In: Didaktik Deutsch, 16, H. 31, S. 50–67.

Rezat, Sara / Grundler, Elke / Schmölzer-Eibinger, Sabine / Feilke, Helmuth (Hg.) (2024): Textprozeduren in Spannungsfeldern. Tübingen: Stauffenburg.

Rüßmann, Lars / Steinhoff, Torsten / Marx, Nicole / Wenk, Anne K. (2016): Schreibförderung durch Sprachförderung? Zur Wirksamkeit sprachlich profilierter Schreibarrangements in der mehrsprachigen Sekundarstufe I unterschiedlicher Schulformen. In: Didaktik Deutsch, 21, H. 40, S. 41–59.

Schicker, Stephan (2020): Förderung der Textbeurteilungskompetenz von Lernenden. Eine Interventionsstudie in sprachlich heterogenen Klassen. Münster: Waxmann.

Stab, Christian / Gurevych, Iryna (2017): Parsing Argumentation Structures in Persuasive Essays. In: Computational Linguistics, 43, H. 3, S. 619–659. Online unter: https://direct.mit.edu/coli/article/43/3/619/1573/Parsing-Argumentation-Structures-in-Persuasive (letzter Aufruf 25. Oktober 2024).

Stahl, Maja / Michel, Nadine / Kilsbach, Sebastian / Schmidtke, Julian / Rezat, Sara / Wachsmuth, Henning (2024): A School Student Essay Corpus for Analyzing Interactions of Argumentative Structure and Quality. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, S. 2661–2674. https://doi.org/10.48550/arXiv.2404.02529

Stede, Manfred / Schneider, Jodi (2018): Argumentation Mining. San Rafael: Morgan & Claypool.

Sturm, Afra (2024): Lernförderliches Feedback im Bereich Schreiben. In: Petersen, Inger / Reble, Raja / Kilian, Jörg (Hg.): Texte schreiben in allen Unterrichtsfächern. Textbeurteilung als Grundlage für Schreibförderung und Leistungsbewertung. Münster: Waxmann, S. 19–37.

Wachsmuth, Henning / Khatib, Khalid Al / Stein, Benno (2016): Using Argument Mining to Assess the Argumentation Quality of Essays. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, S. 1680–1691.

Walton, Douglas / Reed, Christopher / Macagno, Fabrizio (2008): Argumentation Schemes. New York: Cambridge University Press.

Wambsganss, Thiemo / Küng, Tobias / Söllner, Matthias / Leimeister, Jan Marco (2021): ArgueTutor. An Adaptive Dialog-Based Learning System for Argumentation Skills. In: CHI Conference on Human Factors in Computing Systems (CHI '21). New York: Association for Computing Machinery, S. 1–13. https://doi.org/10.1145/3411764.3445781

Wisniewski, Benedikt / Zierer, Klaus / Hattie, John (2020): The Power of Feedback Revisited. A Meta-Analysis of Educational Feedback Research. In: Frontiers in Psychology, 10, Art. 3087, S. 1–14. https://doi.org/10.3389/fpsyg.2019.03087

Sara Rezat

Universität Paderborn
sara.rezat@uni-paderborn.de

Sebastian Kilsbach

Universität Paderborn
sebastian.kilsbach@uni-paderborn.de

Rabia Karabey

Universität Paderborn
rabia.karabey@uni-paderborn.de

Nadine Michel

Universität Paderborn
nadine.michel@uni-paderborn.de

Maja Stahl

Leibniz Universität Hannover
m.stahl@ai.uni-hannover.de

Henning Wachsmuth

Leibniz Universität Hannover
h.wachsmuth@ai.uni-hannover.de