Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 2, Nummer 1 (Juli 2022), ISSN 2749-4411

KORPUSLINGUISTIK UND MÜNDLICHKEIT
Methodische und technische Herausforderungen bei der Erstellung eines fachspezifischen Korpus zur Verständigung über Literatur im Deutschunterrichtsdiskurs auf der Grundlage archivierter Transkripte

Christina Schmidt
Georg-August-Universität Göttingen

Abstract

Der Beitrag setzt sich mit den methodischen und technischen Herausforderungen auseinander, die im Rahmen des Dissertationsprojektes *Musterhafter Sprachgebrauch beim Sprechen über Literatur im Deutschunterrichtsdiskurs. Eine korpuslinguistische Untersuchung von Deutschunterrichtstranskripten* bei der Erstellung eines fachspezifischen Korpus auf der Grundlage archivierter Transkripte aus unterschiedlichen Projektkontexten aufgetreten sind. Dabei wird zum einen auf die Anforderungen der Erstellung eines fachspezifischen Korpus mit gesprochensprachlichen Daten eingegangen, indem auf einer theoretischen Ebene Merkmale der Verständigung über Literatur im Deutschunterricht erläutert werden. Auf dieser Grundlage wird zum anderen aufgezeigt, welche Aufbereitungsschritte in dem Projekt die maschinelle, korpuslinguistische Analyse der Daten ermöglichen. In einem Fazit werden die Grenzen und Potenziale dieses Zugriffs diskutiert.

Keywords: Korpuslinguistik; Sprachgebrauchsmuster; Spracherwerb; Mündlichkeit; Deutschunterricht; Verständigung; Literatur

Abstract

This article deals with the methodological and technical challenges that arose in the course of the dissertation project *Musterhafter Sprachgebrauch beim Sprechen über Literatur im Deutschunterrichtsdiskurs. Eine korpuslinguistische Untersuchung von Deutschunterrichtstranskripten* during the creation of a subject-specific corpus on the basis of archived transcripts from different project contexts. On the one hand, the requirements of creating a subject-specific corpus with spoken language data are addressed by explaining, on a theoretical level, characteristics of communication about literature in German classes. On this basis, it will be shown on the other hand, which processing steps in the project enable the machine-based, corpus-linguistic analysis of the data. In a conclusion, the limits and potentials of this approach will be discussed.

Keywords: corpus linguistics; language usage patterns; language acquisition; orality; German teaching; communication; literature

Es gehört zum Menschsein dazu, sich auszutauschen über jene Angebote zur Selbstdeutung einer Kultur, die die Kunst, und damit auch die Literatur, uns macht.
(Ulf Abraham 2011: 49).

1. Einleitung

Die Verständigung über Unterrichtsgegenstände stellt eine Grundlage des Lernens und Verstehens in schulischen Lehr-Lern-Situationen dar. Im Fokus des Dissertationsprojektes *Musterhafter Sprachgebrauch beim Sprechen über Literatur im Deutschunterrichtsdiskurs. Eine korpuslinguistische Untersuchung von Deutschunterrichtstranskripten* (vgl. Schmidt i. Vorb.) steht die Analyse musterhafter

Strukturen in der Verständigung über Literatur im Deutschunterrichtsdiskurs, die in ihrem Anbahnungspotenzial für Spracherwerbsprozesse im Unterricht und ihrer Spezifik herausgestellt werden. Mit Blick auf die Fähigkeit, über Literatur in der Sozialisationsinstanz Schule kommunizieren zu können, wird unter einer spracherwerbsbezogenen Perspektive davon ausgegangen, dass in der Verständigung zugleich der Einsatz systematischer sprachlicher Formen zur Verständlichkeit der Äußerungen vorausgesetzt und auch musterhafte sprachliche Strukturen in der Interaktion gemeinsam entwickelt werden. Insbesondere im Deutschunterricht, in dem die Sprache sowohl Lerngegenstand als auch Lernmedium ist, stellt sich die Frage, inwiefern sprachliche Muster zur Verständigung über Literatur implizit genutzt oder expliziert thematisiert werden und ob eine Entwicklung des „musterhaften Sprachgebrauchs“ (Bubenhofer 2009: 5) schüler*innen- und lehrer*innenseitig von der Unterstufe bis zur Oberstufe beobachtet werden kann.

Über einen korpuslinguistischen Zugriff können diese Muster herausgestellt und interpretiert werden. Ausgehend von der fachspezifischen Forschungsfrage, welche Sprachgebrauchsmuster sich beim Sprechen über Literatur im Deutschunterricht auf der Sprachoberfläche zeigen, wird ein Korpus zusammengestellt, das eine korpuslinguistische Untersuchung ermöglicht. Um auch in quantitativer Hinsicht eine Basis für diese Untersuchung zu schaffen und um bestehende Daten nachhaltig zu nutzen, wird das Korpus aus archivierten Unterrichtstranskripten zusammengestellt. Die Grundlage der Korpusanalyse stellen somit Daten gesprochener Sprache dar, die einen besonderen methodischen Zugang voraussetzen. Da es sich um ein fachspezifisches Korpus handelt, welches auf der Basis einer theoretischen, deutschdidaktischen Fragestellung zusammengestellt wurde, bedingt diese zugleich die weitere Korpusarbeit. Daher ist es notwendig, vor der Explikation der Herausforderungen, die sich für die Aufarbeitung der archivierten Transkripte des Korpus und für die korpuslinguistischen Analysen stellen, die theoretischen Grundlagen des Projektes zu erläutern. Obwohl die Korpuserstellung und Transkriptaufbereitung anhand einer konkreten deutschdidaktischen Fragestellung an der Schnittstelle zwischen Literatur- und Sprachdidaktik vorgestellt wird, stellen sich viele dieser Herausforderungen ebenso in korpuslinguistischen Arbeiten, die den Fokus auf die Untersuchung sprachlicher Phänomene in anderen Sprachaneignungskontexten legen.

Ausgehend von der Fragestellung des Dissertationsprojektes wird in diesem Beitrag zunächst herausgestellt, warum das Sprechen über Literatur im Deutschunterricht eine besondere Form der Verständigung zwischen dem „begrifflichen“ und dem „begrifflosen“ Verstehen (Brandstätter 2011: 36-37) darstellt und welche Funktion der Gebrauch von sprachlichen Mustern in diesen Verständigungsprozessen einnimmt. In einem zweiten Schritt wird erläutert, ob mithilfe eines korpuslinguistischen Zugriffs diese Muster aufgedeckt werden können und inwiefern die theoretische Fundierung der Arbeit den Aufbau und die Strukturierung des Korpus und der Korpusdaten beeinflusst. Darüber hinaus wird gefragt, welche konkreten methodischen Zugänge für dieses spezifische Korpus unter der Fragestellung des Projektes zielführend sein können. Der Fokus des Beitrags wird dann drittens darauf liegen, wie die vorliegenden Transkripte aufbereitet werden können, um eine korpuslinguistische Untersuchung archivierter, gesprochensprachlicher Daten überhaupt zu ermöglichen und welche Herausforderungen sich dabei ergeben. In einem Fazit werden die aufgezeigten Herausforderungen bei der Korpuserstellung und Datenaufbereitung resümiert und Chancen des Zugriffs für die Erstellung fachspezifischer Korpora erläutert.

2. Sprechen über Literatur im Deutschunterrichtsdiskurs

Ein zentrales Ziel des Deutschunterrichts ist es, die Schüler*innen beim literarischen Lernen und Verstehen zu unterstützen. Literarisches Lernen umfasst nach Petra Büker die Gesamtheit aller Prozesse, die „zum Erwerb von Einstellungen, Fähigkeiten, Kenntnissen und Fertigkeiten […] nötig

sind, um literarisch-ästhetische Texte […] zu erschließen, zu genießen und mit Hilfe des produktiven und kommunikativen Auseinandersetzungsprozesses zu verstehen“ (Büker 2002: 121). An dieser Stelle wird bereits die Spezifik literarischer Texte hervorgehoben, die auch Horst-Jürgen Gerigk beschreibt. Seinen Ausführungen zum Lesen und Interpretieren zufolge stellt das „Denken“ der „poetologischen Differenz“ die Grundlage literarischen Verstehens dar (vgl. Gerigk 2013: 30). Die poetologische Differenz kann als ein zu erkennender Aspekt des literarischen Gegenstandes beschrieben werden, dessen Wahrnehmung über die Einnahme einer bestimmten Perspektive auf den Gegenstand im Rahmen des Deutschunterrichts vermittelt werden kann und die Art der Rezeption eines Textes in den Fokus rückt (vgl. Gerigk 2013: 30-33). Der Deutschunterricht zeichnet sich in Hinblick auf den Umgang mit Literatur demnach dadurch aus, dass Schüler*innen ausgewählte Texte, die insbesondere aufgrund der literaturwissenschaftlichen Kriterien der „Literarizität“, der „Poetizität“ und der „Ästhetizität“ bestimmt wurden (Heydebrand / Winko 1996: 23), „autonom-ästhetisch“ und auch „formal-ästhetisch“ rezipieren und dabei fachspezifische Konventionen, wie die Ästhetikkonvention, (er)kennen und einhalten können (vgl. Heydebrand / Winko 1996: 23).

Dabei wird dem Sprechen über die Rezeption eines literarischen Textes eine besondere Bedeutung zugeschrieben. Das Ziel der Verständigung über subjektive Zugänge und Erfahrungen gründet in dem anthropologischen Bedürfnis nach „geteilter Intentionalität“ (vgl. Thomasello / Carpenter 2007: 121-122): Ich möchte verstanden werden; ich möchte mein Gegenüber verstehen; und zeigt bereits in der von Jérôme Bruner untersuchten Interaktion von Bezugspersonen und Kleinkindern kooperative, systematische und musterhafte Strukturen an der Sprachoberfläche beim gemeinsamen „Lesen“ eines Buches auf (vgl. Bruner et al. 1997: 65-67). Das Sprechen über die subjektiven Vorstellungen in Verbindung mit dem literarischen Text erfordert eine möglichst konkrete Explikation, um teilbar zu werden, und stellt neben der kognitiven Verfügbarkeit eines mentalen Modells vom Text eine enorme Herausforderung auf der sprachlichen Ebene dar. Dabei ermöglicht der Austausch über die Leseerfahrungen, andere Perspektiven auf den vermeintlich gleichen literarischen Text zu eröffnen, die als Angebot einer Erweiterung der eigenen Vorstellungen aufgefasst werden können und neue ‚Sehepunkte‘ auf den literarischen Gegenstand zulassen (vgl. Köller 2004: 9-10). Die Sichtweisen der Interaktanden können sowohl assimiliert als auch akkommodiert, als gleichberechtigt zugelassen oder begründet abgelehnt werden (vgl. Abraham 2015: 10; Maiwald 2015: 93). Darüber hinaus geht es auch darum, die eigene Perspektive auf den literarischen Gegenstand bewusst wahrzunehmen, um an dem Diskurs über diesen Gegenstand teilnehmen und teilhaben zu können. Die besondere Herausforderung liegt dabei darin, dass die subjektiven literarischen Erfahrungen – die zumeist auch an Emotionen gebunden sind und an die eigene Lebenswelt anschließen – auf der einen Seite nur bedingt mitteilbar sind (auf der sprachlichen Ebene, aber auch bereits auf der kognitiven und sozialen Ebene) und auf der anderen Seite nur Ausschnitte der literarischen Erfahrung sein können (vgl. Mitterer / Wintersteiner 2015: 103). Zudem ist ein Merkmal literarischer Texte, bedeutungsoffen und vielsinnig zu sein und keine eindeutige Lesart zu implizieren, die dann lediglich ‚entdeckt‘ werden muss.

Literarisches Lernen bedeutet dann, mit den Eigenschaften des Textes und den Anforderungen an die Rezeption umgehen zu können und sich unter anderem auf „die Unabschließbarkeit des Sinnbildungsprozesses ein[zu]lassen“ (Spinner 2006: 12). Dieser Umgang mit Literatur erfordert und bedingt eine bestimmte Art und Weise, wie über literarische Texte im Unterricht gesprochen wird. Kaspar H. Spinner (vgl. Spinner 2011) stellt neun Aspekte heraus, die eine literarische Kommunikationskompetenz auszeichnen. Dazu zählen der Umgang mit der Zeit, die man sich für den Text nehmen soll (vgl. Spinner 2011: 63), das Sich-Einlassen auf die Gesprächsatmosphäre (vgl. Spinner 2011: 65) und das Zuhören, wenn andere ihre subjektiven Textrezeptionen mitteilen (vgl. Spinner 2011: 69). Für die Artikulation des eigenen Textverständnisses sei es insbesondere wichtig, subjektive Eindrücke zu äußern, sich über die entwickelten Imaginationen sowie genaue

Beobachtungen auszutauschen und sich dabei sowohl auf Differenzerfahrungen als auch auf die Unabschließbarkeit der Sinnbildung einzulassen (vgl. Spinner 2011: 65-71). Im Sprechen über den literarischen Text könne sich, so Spinner, dann gerade das Austarieren begrifflichen und begrifflosen Verstehens zeigen (vgl. Spinner 2011: 67-68; vgl. dazu auch Brandstätter 2011): „Gerade die ästhetische Erfahrung wird oft dadurch charakterisiert, dass sie einen unabschließbaren Wechselprozess zwischen Allgemeinem und Konkret-Einmaligem, zwischen Bestimmbarem und Unbestimmbaren auslöst“ (Brandstätter 2011: 31). Das begrifflose Verstehen sei auf die Einzigartigkeit des Einzelnen fokussiert und zeichne sich durch die Fülle an Sinnlichkeit aus, die auf das Unbestimmbare abzielt (vgl. Brandstätter 2011). Das begriffliche Verstehen hingegen ziele auf die Abstraktion eines Einzelfalls und somit auf das Verallgemeinerbare ab und könne über konventionalisierte sprachliche Einheiten verbalisiert werden, über die subjektive Zugänge bestimmt und mit anderen verglichen werden können (vgl. Brandstätter 2011).

Gerade im Prozess des Sprechens könne die Zusammenführung begrifflichen und begrifflosen Verstehens gelingen, wobei dieses Wechselspiel einer besonderen Form der Sprache bedürfe, die „Raum […] für das Begrifflose, für die ‚Irrläufe im Kopf‘“ (Brandstätter 2011: 39) lässt. Spinner betont, dass Schüler*innen lernen sollten, „dass die begriffliche Sprache immer nur eine Annäherung an den Text und seine vorstellungsbildende, symbolisch-bedeutungserweiternde und emotionsauslösende Wirkung sein kann“ (Spinner 2011: 67) und gerade der fragmentarische Charakter von Gesprächsbeiträgen zeige, „dass sowohl das begrifflose Verstehen wie die Suche nach einer begrifflichen Sprache ernst genommen werden“ (Spinner 2011: 68). Das Sprechen über Literatur im Deutschunterrichtsdiskurs könnte sich dann gerade durch die Artikulation von kognitiven Dissonanzen (Irritationen), Zögern, Abbrüchen, Gedankensprüngen, Metaphern und Symbolen, Offenlassen von Interpretationsspielräumen etc. auszeichnen, die im Laufe des Aneignungsprozesses der literarischen Gesprächskompetenz sogar vermehrt auftreten könnten.

Obwohl sich die Sprache im Deutschunterricht über das Wechselspiel zwischen „instrumentellen“ und „annähernden“ Sprachgebrauch (Schmidt 2016: 93) auszeichnen könnte, bleibt das Ziel dennoch die Verständigung, die über bereits konventionalisierte aber auch neu konstruierte und gemeinsam genutzte sprachliche Strukturen gelingen kann. Somit könnten konventionalisierte sprachliche Strukturen beispielsweise als Unterstützungsverfahren eingesetzt werden, um annähernden Sprachgebrauch zu markieren. Ein bewusster Umgang mit sprachlichen Mustern im Unterricht wirkt sich dann auf die gesamte Verständigung aus und ermöglicht einen sprachsensiblen Unterricht, der das Wissen über die semantische und pragmatische Funktion dieser Muster auf produktiver und rezeptiver Ebene sowie den kompetenten Umgang mit diesen nicht voraussetzt, sondern explizit zum Thema macht. Dabei spielt es nicht primär eine Rolle, ob es sich um einen Unterricht handelt, in dem Deutsch als Erst-, Zweit- oder Fremdsprache vorausgesetzt werden kann, da das Wissen über und das Können im Umgang mit diesen Mustern an subjektive Erfahrungen des Sprachgebrauchs anknüpft. Die Reflexion des Gebrauchs musterhafter Sprache ist insbesondere dann von Relevanz, wenn die Muster spezifisch für die sprachliche Verständigung in einem Schulfach sind. So hat das Nomen *Figur* (literarisch) beispielweise eine andere Bedeutung und steht in Kombination mit anderen Wörtern als das Nomen *Figur* (geometrisch) im Mathematikunterricht. Ein anderes Beispiel wäre der Gebrauch des Nomens *Geschichte* im Deutschunterricht, das sich auf die literarische Narration im Sinne einer Erzählung, auf die historische Dimension der Erzählung und auf das Schulfach *Geschichte* beziehen kann. Darüber hinaus wird angenommen, dass sprachliche Muster in Abhängigkeit von kulturellen und schulpraktischen Hintergründen sowie in ihrer historischen Dimension variieren können. So könnte der Einsatz von Sprachgebrauchsmustern je nach Erfahrungskontext unterschiedlich interpretiert werden. Ein Literaturunterricht beispielsweise, in dem Nicht-Verstehen, Irritationen und der Umgang mit Unbestimmtheit auch auf der sprachlichen Ebene als Spezifika des Fachs verstanden werden, könnte für Schüler*innen, deren Bild von Schule

durch die Selektionsfunktion und Leistungsorientierung der Institution geprägt ist, zunächst befremdlich wirken und auf Unverständnis stoßen.

Im Folgenden wird herausgestellt, inwiefern durch einen korpuslinguistischen Zugriff sprachliche Muster analysiert werden können und welche Anforderungen in Bezug auf die Forschungsdaten damit einhergehen. Diese rekurrenten sprachlichen Strukturen, die auf der Sprachoberfläche beobachtet werden können, werden im Folgenden als *Sprachgebrauchsmuster* bezeichnet (vgl. Bubenhofer 2009; Brommer 2018: 61).

3. Der korpuslinguistische Zugriff

Um diese Sprachgebrauchsmuster mit korpuslinguistischen Tools herausstellen und interpretieren zu können, ist der Aufbau eines Korpus und die Aufbereitung der im Korpus zusammengefassten Daten notwendig. Bevor der Korpusaufbau und die Transkriptaufbereitung des Projektes vorgestellt werden und auf technische und methodische Herausforderungen näher eingegangen wird, erfolgt eine kurze Herausstellung der korpuslinguistischen Perspektive, indem diese für die Bearbeitung fachdidaktischer Forschungsfragen beleuchtet wird.

Ein korpuslinguistischer Zugriff ermöglicht „die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theorienbildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind“ (Lemnitzer / Zinsmeister 2015: 14). In der weitesten Definition ist ein ‚Korpus‘ eine Sammlung, die sowohl schriftliche als auch gesprochene Äußerungen enthalten kann. In der Korpuslinguistik kommen jedoch noch weitere Kriterien hinzu. So umfasst ein Korpus typischerweise digitalisierte Daten, die auf Rechnern gespeichert und maschinenlesbar sind. Die Bestandteile eines Korpus sind dann die Sprach- oder Textdaten, beschreibende Metadaten (bspw. Informationen zur Aufnahmezeit oder zu den Sprecher*innen) und linguistische Annotationen (bspw. die Markierung von Adverbien oder Appellativa über Part-of-Speech-Tags) (vgl. Lemnitzer / Zinsmeister 2015: 13).

Der Datenzugriff kann auf zwei Arten erfolgen: Zum einen über den deduktiven corpus-based- bzw. korpusbasierten Zugriff, in dem das Korpus eher funktionalisiert wird, um zugrundeliegende Thesen mithilfe des Korpus zu überprüfen (vgl. Bubenhofer 2009: 17; Brommer 2018: 97). Zum anderen über den induktiven corpus-driven- bzw. korpusgesteuerten Ansatz, in dem das Korpus im Fokus steht und möglichst ohne theoretische Vorannahmen statistische Auffälligkeiten herausgestellt werden. Die Theorienbildung erfolgt hier auf der Grundlage des Korpus (vgl. Bubenhofer 2009: 17; Brommer 2018: 97). Rainer Perkuhn und Cyril Belica sprechen mit Blick auf das induktive Vorgehen von einer „eigene[n] Perspektive“ (Perkuhn / Belica 2006: 2), die die Korpuslinguistik auf der Grundlage einer entsprechenden Korpusgröße und entsprechender Analysemethoden eröffnen kann. Joachim Scharloth hebt hervor, dass diese Perspektive auf ein Korpus „nicht theoretisch begründete Hypothesen mittels Korpusdaten [überprüft], sondern […] induktiv nach Mustern in großen Sprachdatenmengen [sucht], um so zu neuen Einsichten über Sprache zu gelangen und neue Beschreibungskategorien zu entwickeln“ (Scharloth 2018: 65).

In den neueren korpuslinguistischen Arbeiten und theoretischen Modellierungsansätzen werden deduktive und induktive Zugänge miteinander verknüpft und nicht konträr gegenübergestellt (vgl. Brommer 2018: 104). Sarah Brommer hebt hervor, dass „das korpuslinguistische Potential […] sich nicht allein in induktiven, quantitativen Analysen [erschöpft], die für sich genommen wenig Aussagekraft haben“ (Brommer 2018: 104). Vielmehr sollten die Ergebnisse eines ersten induktiven Zugriffs in einem zweiten Schritt über „weitere Analyseschritte selektiert, klassifiziert und interpretiert werden“ (Brommer 2018: 104), die dann wiederum deduktive Zugänge nahelegen (vgl. Brommer

2018: 98, 105). Die Ergebnisse dieser Analysen können dann als „Sprachgebrauchsmuster“ (Bubenhofer 2009) beschrieben werden. Noah Bubenhofer fasst ein solches (sprachliches) Muster als „eine Wortform, eine Verbindung von Wortformen oder eine Kombination von Wortformen und nichtsprachlichen Elementen“ (Bubenhofer 2009: 23) im Sinne eines Zeichenkomplexes zusammen. Dieser Zeichenkomplex diene ferner „als Vorlage für die Produktion weiterer Zeichenkomplexe“ und weise die gleiche „Materialität“ „wie die darauf entstehenden Zeichenkomplexe“ auf (Bubenhofer 2009: 23). Diese Definition wird von Brommer um die Kriterien der „Rekurrenz“, der „Signifikanz“ und der „Typizität“ erweitert (Brommer 2018: 54).

Für die Fachdidaktik eröffnet dieser Zugang das Potenzial, über einen ersten induktiven Zugriff auf Unterrichtsdaten, der dann durch qualitative Analysen und durch die Berücksichtigung des Kontextes ergänzt wird, musterhafte Strukturen auf der Sprachoberfläche zu erkennen und theoriegeleitet Schlüsse über Spracherwerbs- und Professionalisierungsprozesse zu ziehen. Gerade mit Blick auf die Verständigung in Lehr- und Lernsituationen spielt der Kontext der Sprachgebrauchsmuster und deren Funktion für den Gesprächsverlauf eine entscheidende Rolle, die einen qualitativen, interpretierenden zweiten Zugriff unabdingbar macht. Darüber hinaus könnten sprachliche Phänomene mit Einfluss auf Lehr- und Lernprozesse in den Fokus gerückt werden, die in ihrer Spezifik bislang nicht wahrgenommen wurden und über einen quantitativen Zugriff erfasst werden können. Gerade diese musterhaften Strukturen, die in der Sprachpraxis kaum zu erkennen sind, werden durch den Zugriff auf einer analytischen Ebene an der Sprachoberfläche durch korpuslinguistische Verfahren zum Vorschein gebracht (vgl. Bubenhofer 2009: 24). Noah Bubenhofer hebt hervor, „dass Sprachgebrauchsmuster ein Ausdruck von Konventionalität sind, ohne die sprachliches, und damit soziales Handeln nicht möglich wäre“ (Bubenhofer 2009: 52). Mit Blick auf die Lesesozialisationsforschung kann jedoch nicht davon ausgegangen werden, dass selbst in einem monolingualen Deutschunterricht alle Schüler*innen über dieselben sprachlichen Konventionen beim Sprechen über Literatur verfügen (vgl. Gölitzer 2007: 206-208 zu „Schemata des Lesens und Verstehens“). Das gilt insbesondere auch für einen Unterricht, in dem Deutsch als Zweit- oder Fremdsprache unterrichtet und ein literarischer Text behandelt wird. An dieser Stelle könnte die Explikation dieser konventionalisierten Sprachgebrauchsmuster einen großen Beitrag zur Verständigung in Lehr- und Lernprozessen leisten.

4. Methodische und technische Herausforderungen

Um musterhaften Sprachgebrauch beim Sprechen über Literatur im Deutschunterrichtsdiskurs untersuchen zu können, gilt es zunächst, ein Korpus zu erstellen, das Analysen ermöglicht, in denen Kombinationen von induktiven und deduktiven Zugriffen realisiert werden können. Dazu muss das Korpus – und somit die Daten in dem Korpus – quantitative, qualitative und technische Kriterien erfüllen (vgl. dazu auch Schmidt 2021). Dabei erfordert ein induktiver Zugang die Maschinenlesbarkeit der Transkriptdaten. Um Musteranalysen durchführen zu können, in denen beispielweise Frequenzen und Signifikanzen untersucht werden, müssen Analysetools gefunden werden, die die Transkriptdaten auswerten können. An dieser Stelle spielt das Medium der Mündlichkeit eine entscheidende Rolle, da der Großteil aller Analysetools auf der Grundlage von schriftsprachlichen Daten entwickelt wurde und sich ohne weitere Anpassungen nicht für die Analyse gesprochensprachlicher Daten eignet (vgl. Schmidt 2018: 212 in Bezug auf Gesprächskorpora). Thomas Schmidt stellt die automatische Auswertung gesprochensprachlicher Daten derzeit in einem ernüchternden Fazit als einen „Wunschtraum“ dar (Schmidt 2018: 222). Das führt dazu, dass die Untersuchung gesprochensprachlicher Daten bisher nur mit bestehenden korpuslinguistischen Verfahren gelingen kann, wenn diese schriftlich repräsentiert werden (vgl. Lehmann 2007: 17). Dabei muss bedacht werden, dass es sich insbesondere bei Korpora gesprochensprachlicher Daten trotz aller Bemühungen nach Repräsentativität nicht um

einen Ausschnitt aus einem Phänomenbereich, hier Deutschunterricht, handelt, „sondern [um] eine Repräsentation davon“ (Lehmann 2007: 17), da die Rohdaten durch Abstraktionen und Interpretationen bereits durch die Verschriftlichung und die Perspektiven der Transkribierenden verändert und beeinflusst werden (vgl. Lehmann 2007: 17).

Dieser Umgang mit gesprochensprachlichen Daten, den Christian Mair als „[d]as gravierendste Defizit in der bisherigen Geschichte der Korpuslinguistik“ (Mair 2018: 12) bezeichnet, wirft eine Vielzahl an methodischen Problemen auf. Forschende stehen an dieser Stelle vor der Herausforderung, eine Balance zu finden zwischen der Nutzung pragmatischer und technischer Möglichkeiten der Datenanalyse und der Entwicklung neuer Verfahren, die den Eigenschaften mündlicher Sprache (eher) gerecht werden und über die bereits bestehenden ‚Schablonen‘ hinausdenken (vgl. Mair 2018: 15). Eine derzeit ideale Vorstellung von der Erstellung und zur Verfügung-Stellung multimodaler gesprochensprachlicher Korpora „als Video- und Tondatei, mit maschineller Unterstützung mit einer orthographischen Transkription aligniert, die in zusätzlichen Versionen mit weiterer grammatischer und prosodischer Information angereichert ist […]“ (Mair 2018: 15), kann zwar als Zielvorstellung bei der Erfassung neuer Daten angestrebt und gesetzt werden, versperrt jedoch den Zugang zu bereits bestehenden Daten in Archiven.

An dieser Stelle könnte radikal geschlussfolgert werden, Daten, die nicht mit Video- und Audiomaterial angereichert sind, für korpuslinguistische Untersuchungen als unbrauchbar zu erklären. Ebenso könnte jedoch auch gefragt werden, wie die bestehenden Archivdaten aufbereitet und nutzbar gemacht werden können, um korpuslinguistische Verfahren auch an teils historischen Daten zu ermöglichen und nachhaltig zu arbeiten. Hier steht es außer Frage, dass es dabei nicht darum gehen kann, das gesamte Potenzial gesprochensprachlicher Daten zu erschöpfen.

Gerade mit Blick auf Mehrpersoneninteraktionen in schulischen Lehr- und Lernsituationen müssten jedoch auch unter einer idealisierten Perspektive an dieser Stelle erst Verfahren gefunden werden, die Audio- und Videoaufnahmen mit sachlichen Hinweisen über das Geschehen ermöglichen. Dabei könnten beispielweise die folgenden Fragen thematisiert werden: Wie wird mit Kameraperspektiven umgegangen? Wie können parallellaufende Gruppenarbeitsphasen aufgezeichnet und untersucht werden? Wie wird mit schriftlichen Medien wie Tafelanschrieben umgegangen, über die gesprochen wird?

Diese Diskrepanz zwischen dem Umgang mit gesprochensprachlichen und schriftsprachlichen Daten in der Korpuslinguistik zeigt sich auch beim maschinellen Einlesen der Transkripte, da es sich bei einem Transkript nicht um einen Text handelt, der als Ganzes linear eingelesen und ausgewertet werden kann. Bei der Analyse der gesprochenen Sprache, die in Transkripten verschriftlicht wird, werden viele Informationen zur Lesbarkeit, zum Nachvollzug und zur Strukturierung von den Transkribierenden eingebracht, die bei einer Analyse gesprochener Sprache gefiltert und als Metadaten gekennzeichnet und extrahiert werden müssen. Das betrifft unter anderem die Markierung, wer den Sprecher*innenbeitrag geäußert hat (bspw. die Lehrerin oder der Lehrer), Informationen darüber, auf welche Art und Weise ein Sprecher*innenbeitrag geäußert wurde (bspw. leise, laut, langsam) oder auch Informationen zum Kontext (bspw. Schüler*in XY liest aus der Lektüre vor). An dieser Stelle müssen auf der Grundlage der Forschungsfrage projektinterne Entscheidungen getroffen werden, wie die Transkripte aufbereitet werden können und welche Informationen als Metadaten beibehalten oder hinzugefügt werden, um über diese weitere Analyseschritte einleiten zu können (bspw. die Filterung der Daten nach der Jahrgangsstufe, in der die Sprecher*innenbeiträge getätigt werden) (vgl. dazu auch Schmidt 2018). Dabei muss jedoch bedacht werden, dass sich viele Metadaten nur auf einzelne Elemente des Transkripts beziehen (vgl. Schmidt 2018: 223-224).

Die Aufbereitung der Transkripte zur Realisierung eines induktiven quantitativen Zugriffs beeinflusst jedoch zugleich, inwiefern ein qualitativer Zugriff in einem zweiten Schritt gelingen kann. Damit stellt sich die Frage, wie das Transkript konkret eingelesen werden soll und ob beispielweise

einzelne Sprecher*innenbeiträge oder ganze Turns berücksichtigt werden. Bei dieser Entscheidung ist jedoch zu beachten, dass die Identifizierung einzelner Turns bereits ein Tagging im Vorfeld erfordert, in dessen Rahmen bestimmt wird, wann ein Turn beginnt und endet. Sofern der Kontext eines sprachlichen Musters bei der Interpretation mit eingebunden werden soll, ist es notwendig, ein Tool zu wählen, das eine Rückbindung ermöglicht und bestenfalls den Kontext direkt mit ausgeben kann. Der Kontext kann sich dann sowohl auf die Tokens, die dem zu untersuchenden Muster vorausgehen und folgen als auch auf Sprecher*innenbeiträge beziehen, die vor und nach dem Sprecher*innenbeitrag getätigt wurden (vgl. Schmidt 2018).

Des Weiteren erfordert die Annotation der Daten ein Tagset, das auf der Grundlage von gesprochensprachlichen Daten entwickelt wurde und Eigenschaften mündlicher Sprache wie Abbrüche, Interjektionen, Hesitationssignale oder Diskursmarker miteinschließt. Thomas Schmidt merkt an dieser Stelle an, dass eine Entscheidung gegen die Erfassung gesprochensprachlicher Elemente, die in Tagsets, die auf den Eigenschaften schriftsprachlicher Texte gründen, nicht erfasst werden können, bereits Auswirkungen auf die Berechnung von Token-Frequenzen hat (vgl. Schmidt 2018). Zudem muss ein Entschluss darüber gefasst werden, ob sich die Daten für ein solches Tagging überhaupt eignen und inwiefern an dieser Stelle Anpassungen der Transkriptdaten und Analysetools notwendig sind. Im Folgenden wird skizziert, wie ein Korpus erstellt und die Transkripte aufbereitet wurden, um korpusanalytisch Sprachgebrauchsmuster beim Sprechen über Literatur im Deutschunterrichtsdiskurs zu untersuchen und welche konkreten Herausforderungen bei der Aufbereitung archivierter Transkripte aufgetreten sind.

5. Korpuserstellung und Transkriptaufbereitung

Zur Untersuchung musterhaften Sprachgebrauchs beim Sprechen über Literatur im Deutschunterrichtsdiskurs wurde ein Korpus erstellt, das Transkripte umfasst, in denen Deutschunterrichtsstunden an Gymnasien und Gesamtschulen aufgezeichnet wurden, deren Unterrichtsgegenstand ein literarischer Text ist. Somit setzt bereits die Auswahl der Transkripte für das Korpus Kriterien voraus, die sich aus der Forschungsfrage ergeben und weitere Entscheidungen bedingen, die beispielweise die Jahrgangsstufe oder auch die Gattung des literarischen Textes betreffen. Neben Entscheidungen, die durch die theoretische Fundierung der Arbeit begründet werden können, ergeben sich jedoch auch forschungspragmatische Auswahlkriterien. An dieser Stelle mussten Datenbanken und Archive gefunden werden, über die auf Transkripte zugegriffen werden kann, die sowohl die technischen als auch die rechtlichen Voraussetzungen erfüllen, um sie für eine korpuslinguistische Untersuchung nutzen zu können. Das Archiv für pädagogische Kasuistik (ApaeK) an der Goethe-Universität in Frankfurt verfügt über eine Vielzahl archivierter Daten mit einem Schwerpunkt in der Schul- und Unterrichtsforschung und ermöglicht über eine Suchmaske eine Eingrenzung der Archivdaten nach Kriterien wie „Schulform“, „Unterrichtsfach“ und „Jahrgangsstufe“. Über die Beschreibung der einzelnen Transkripte ist zudem eine kurze Zusammenfassung des Inhalts des Transkripts und eine Sichtung zentraler Metadaten gegeben, worüber eine weitere Eingrenzung und Auswahl der Transkripte gewährleistet werden kann.

Mithilfe der Strukturierung des Archivs konnten Transkripte ausgewählt werden, in denen ein epischer Text im Rahmen des Deutschunterrichts als Unterrichtsgegenstand behandelt wird. Unter einer Entwicklungsperspektive auf die zu untersuchenden Sprachgebrauchsmuster sind sowohl Stunden aus der Unter-, der Mittel- als auch aus der Oberstufe zu möglichst ausgeglichenen Anteilen im Korpus abgebildet. Da die Transkripte aus unterschiedlichen qualitativen Forschungsprojekten und kasuistischen Seminaren gewonnen wurden, erfolgte eine Prüfung dieser ausgewählten Transkripte.

Um in das Korpus aufgenommen zu werden, müssen die Transkripte über Daten verfügen, die maschinell eingelesen und formell bearbeitet werden können. Das Korpus wurde nach der Prüfung dieser Kriterien aus dreiundzwanzig Transkripten zusammengestellt.

Jedoch fiel auch bei diesen ausgewählten Transkripten auf, dass aufgrund der unterschiedlichen Projektkontexte und der qualitativen Ausrichtung der Forschungsprojekte kein einheitliches Transkriptionsschema und keine einheitliche Formatierung vorliegt. Um eine korpuslinguistische Untersuchung der Daten realisieren zu können, müssen diese sowohl in Hinblick auf die Formatierung als auch auf die Transkriptionskonventionen vereinheitlicht werden. Bei dieser Aufbereitung spielt die Sichtung geeigneter Analysetools eine entscheidende Rolle, da diese zugleich Möglichkeiten und Grenzen an die Aufbereitung der Daten herantragen. Ein nicht zu unterschätzender Punkt ist an dieser Stelle sowohl das Überangebot und die schnelle Entwicklung der technischen Hilfsmittel für die Analysen (insb. der schriftsprachlichen Daten), als auch die Möglichkeit, „ein Verständnis dafür zu entwickeln, wie die Werkzeuge im Detail funktionieren und was ihre Potenziale und Grenzen sind“ (Mair 2018: 23). Dies liegt insbesondere dann nahe, wenn Analyseplattformen und -tools entwickelt werden, die bei aller Anwenderfreundlichkeit jedoch kaum noch einen Einblick in die dahinterliegenden, statistischen und technischen Prozesse sowie keine Anpassung und Erweiterung der bestehenden Analyseschritte ermöglichen.

Das Tool *CorpusExplorer* von Jan Oliver Rüdiger (vgl. Rüdiger 2018) ist eine OpenSource Software, die eine Vielzahl verschiedener Auswertungsmöglichkeiten zur Verfügung stellt und sowohl das POS-Tagging sowie Möglichkeiten der grafischen Ergebnispräsentation beinhaltet. Darüber hinaus ist die Software so angelegt, dass sie verschiedene Dateiformate einlesen kann und Anregungen für neue Funktionen eingebracht werden können. Durch diese Flexibilität und Anpassungsfähigkeit konnte über den *CorpusExplorer* ein Dateiformat ausgewählt werden, über das die aufbereiteten Transkripte eingelesen und ein POS-Tagging sowie korpuslinguistische Analysen ermöglicht werden können.

Die Aufbereitung der Transkripte erfolgte in mehreren Arbeitsschritten, die sorgfältig dokumentiert werden müssen, um einerseits einen Nachvollzug zu ermöglichen und andererseits immer die Option eines Rückbezugs auf die ursprünglichen Transkripte sicherzustellen (vgl. dazu auch Schmidt 2021).

1. Formatierung der Transkripte und Metadatenaufbereitung

Die Anlage einer Tabelle in dem Programm „MS Excel“ (Version 2108) ermöglicht die Erfassung einzelner Sprecher*innenbeiträge, strukturiert in Datenfeldern. Zur Untersuchung der Transkripte werden die Sprecher*innenbeiträge einzeln eingelesen und mit Metadateninformationen (bspw. Jahrgangsstufe, Sprecher*in) angereichert. Über die Zuordnung der Metadaten zu den Sprecher*innenbeiträgen kann eine gefilterte Suchanfrage auf Metadatenbasis erfolgen. So kann beispielsweise eine Abfrage zu Sprecher*innenbeiträgen der Lehrer*innen der achten Jahrgangsstufe gestellt werden.

Zur Identifikation eines Datensatzes, bestehend aus Metadaten und Sprecher*innenbeitrag, ist die Vergabe eines Schlüsselfeldes notwendig. In der Datenbasis des Dissertationsprojektes ist die Identifikationserkennung beispielweise 1apaek>1 für den ersten Sprecher*innenbeitrag im ersten Transkript. Der Schlüssel für den zweiten Sprecher*innenbeitrag lautet 1apaek>2. Bei dem Import der einzelnen Sprecher*innenbeiträge können bereits erste Fehler, die durch das inkorrekte Einlesen der Sprecher*innenbeiträge entstehen, getilgt werden. Um keine Metadateninformationen, die über das Archiv ersichtlich sind, zu verlieren, wird ein Großteil der Metadaten, die innerhalb der einzelnen Transkriptdateien und in den Suchmasken des Archivs aufgeführt sind, übernommen. Diese Metadaten geben beispielweise Auskunft über das Aufnahmejahr, die Aufnahmezeit, die Schulform, die Jahrgangsstufe, den Titel der im Unterricht behandelten Lektüre etc. Des Weiteren werden manuell Informationen zu den Belegstellen des jeweiligen Sprecher*innenbeitrags ergänzt (z.B. über den Zeilenbeginn des jeweiligen Sprecher*innenbeitrags), um einen qualitativen Rückgriff auf die Transkripte und die Analyse des Sprecher*innenbeitragkontextes zu ermöglichen. Über diese Information könnten dann im weiteren Analyseverlauf einzelne Sprecher*innenbeiträge innerhalb eines Turns untersucht werden. Neben diesen bereits vorhandenen Metadaten werden

aus der Theorie abgeleitete Metadaten ergänzt, die für die Untersuchung der Sprachgebrauchsmuster relevant sein könnten. Zu diesen Metadaten zählen beispielweise die Sozialform und der Modus der Äußerung, der bestimmt, ob in dem Sprecher*innenbeitrag von einem schriftlichen Medium vorgelesen wird oder ob es sich um eine spontansprachliche Äußerung handelt.

2. Vereinheitlichung und Dokumentation der Transkriptionskonventionen und Formatierungen

Der Datensatz, bestehend aus Sprecher*innenbeiträgen und Metadaten, wird in einem zweiten Schritt aufbereitet (erste Datenaufbereitung), indem die Sprecher*innenbeiträge hinsichtlich der Transkriptionskonventionen und Formatierungen vereinheitlicht werden.

Diese Aufbereitung erfolgt manuell und betrifft unter Anderem den Umgang mit und die Differenzierung von Metatextinformationen (bspw. „der Lehrer betritt die Klasse“ oder „die Schülerin lacht“), den Umgang mit Hervorhebungen (Fettsetzungen, Kursivsetzungen, Großschreibungen etc.), den Einsatz von Sonderzeichen (bspw. < >, [ ]) und deren Bedeutung, den Umgang mit Dialekten und Korrekturen (bspw. offensichtliche Rechtschreib- bzw. Tippfehler bei der Transkription). Alle Aufbereitungsschritte wurden für jedes Transkript einzeln dokumentiert und in einer Gesamtdokumentation zusammengefasst. Mit Blick auf die folgende Normalisierung der Daten (zweite Datenaufbereitung) können die cGAT-Konventionen hier bereits eine Orientierung geben.

3. Normalisierung der Daten

Da das ursprüngliche Stuttgart-Tübingen-Tagset (STTS) (vgl. Schiller et al. 1999) auf der Grundlage von schriftsprachlichen Daten entwickelt wurde, zeigte sich in einem ersten Testdurchlauf, dass die Trefferquote des POS-Taggings sehr ungenau und fehlerhaft für gesprochensprachliche Daten ausfällt. Am Institut für Deutsche Sprache in Mannheim wurde jedoch ein Tagset entwickelt, das das Tagging dieser Daten ermöglicht und auf dem Stuttgart-Tübingen-Tagset aufbaut: das STTS 2.0 (vgl. Westphal et al. 2017). Um das STTS 2.0 nutzen zu können, müssen die aufbereiteten Sprecher*innenbeiträge aus der ersten Datenaufbereitung jedoch normalisiert werden. Die Normalisierung erfolgte in einer zweiten Datenaufbereitung manuell nach den Normalisierungs-Konventionen von OrthoNormal (vgl. Winterscheid et al. 2019), die wiederum auf den cGAT-Konventionen (vgl. Schmidt / Schütte / Winterscheid 2015) aufbauen.

Erste Datenaufbereitung																	
Ja	,	also	ich	hab	jetzt	also	ähm	so	ähm	[…]	dass	es	eine	(gosar)	Erzählung	ist	.
Zweite Datenaufbereitung (Normalisierung)																	
ja	/	also	ich	habe	jetzt	also	äh	so	äh	[…]	dass	es	eine	%	Erzählung	ist	/

Abbildung 1
2apaek02488>125; Enders / Hundsdorf (2008: Z. 211-213); auch in Schmidt (2021: 153.).

In der Abbildung 1 wird beispielhaft gezeigt, wie die erste und die zweite Datenaufbereitung erfolgt. Im Rahmen der ersten Datenaufbereitung werden die Sprecher*innenbeiträge hinsichtlich der Transkriptionskonventionen und Formatierungen vereinheitlicht (s. 2 „Vereinheitlichung und Dokumentation der Transkriptionskonventionen und Formatierungen“). Die Konventionen der ApaeK-Transkripte orientieren sich zum großen Teil an der schriftsprachlichen Orthografie. An dieser Stelle muss beispielweise bereits bedacht werden, dass ein orthographischer Satz eine höhere Abstraktionsstufe darstellt als eine Äußerung (vgl. Lehmann 2007: 14). In der zweiten Datenaufbereitung ist in diesem Zusammenhang beispielsweise zu erkennen, dass im Zuge der Normalisierung die Interpunktion aufgehoben wird, Apokopen wieder orthographisch korrigiert, Häsitationsphänomene vereinheitlicht werden und Unverständliches mit Dummys (bspw. mit %) gemappt wird (vgl. Winterscheid et al. 2019).

Die Sprecher*innenbeiträge der zweiten Datenaufbereitung werden in den Analysen auf Sprachgebrauchsmuster beim Sprechen über Literatur im Deutschunterrichtsdiskurs untersucht. Dabei stehen lediglich die gesprochensprachlichen Äußerungen im Fokus. Eine Herausforderung ergibt sich dadurch an

dieser Stelle, da die (Kontext)-Informationen in den Transkripten, die nicht direkt zur Äußerung des Sprecher*innenbeitrages gehören, durch eine Ersetzung mit Dummys weiter aufbereitet oder exkludiert und für die quantitative Analyse getilgt werden müssen. Die Beibehaltung dieser Informationen und auch die Ersetzung durch einen Dummy haben einen Einfluss auf die weiteren Korpusanalysen, indem diese bei der Berechnung von Frequenzen als Token des Sprecher*innenbeitrags mit aufgenommen werden. Sofern ein Dummy für diese Informationen gesetzt wird, muss das Tagging der betroffenen Stellen insbesondere kontrolliert werden.

In dem vorliegenden Korpus wurde der Dummy *%* erweitert, indem dieser auch Informationen ersetzt, die eindeutig einem Sprecher oder einer Sprecherin zugeordnet werden können und bei denen es sich um eine verbalisierte, aber unverständliche Äußerung, wie beispielsweise ‚Gemurmel', handelt.
Die so entstandene Datenbasis kann in unterschiedlichen Dateiformaten exportiert werden, die einen Import in andere Tools (hier den *CorpusExplorer*; vgl. Rüdiger 2018) zulassen. Die aufbereitete Version der Sprecher*innenbeiträge wird dann vom *CorpusExplorer* unter dem Feldnamen ‚Text'[1] eingelesen.

4. Annotation der Daten

Die normalisierten Daten können im *CorpusExplorer* (vgl. Rüdiger 2018) maschinell segmentiert und im Rahmen der Annotation über das STTS 2.0 getaggt werden. Dabei sollte zunächst geprüft werden, inwiefern das Tagging korrekt verläuft.

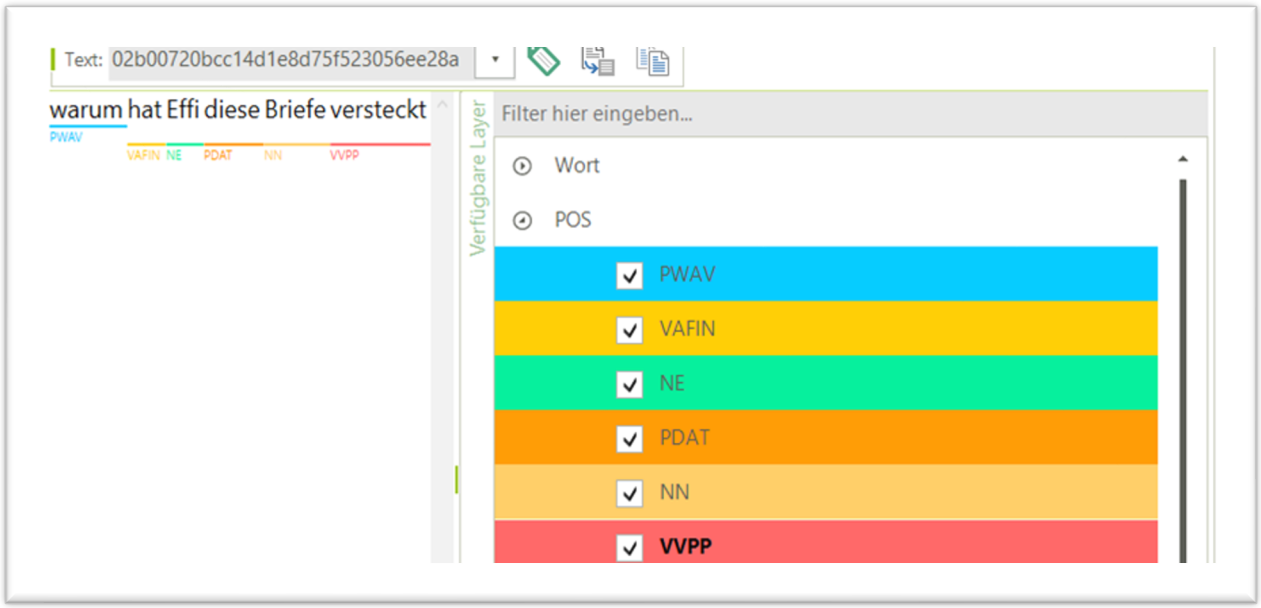

Abbildung 2
POS-Tagging im CorpusExplorer (vgl. Rüdiger 2018).

In Abbildung 2 kann über die Analysefunktion ‚Volltextzugriff > Texte annotieren' im *CorpusExplorer* geprüft werden, ob die POS-Tags den einzelnen Token korrekt zugeordnet werden. In der beispielhaften Äußerung „warum hat Effi diese Briefe versteckt" wird deutlich, dass „Effi" als Eigenname und „Briefe" als Appellativa eingelesen wird. Eine solche Prüfung sollte über mehrere Sprecher*innenbeiträge erfolgen. An dieser Stelle können dann rekursive Schritte anfallen, die eine erneute Normalisierung und Aufbereitung bedingen. So zeigte sich bei der Prüfung im Rahmen dieses Projektes beispielsweise, dass anonymisierende Formen für Schüler*innen wie bspw. „Sw18" nicht als Eigennamen erkannt werden.

Für das vorliegende Korpus wurde das POS-Tagging von 43 Sprecher*innenbeiträgen untersucht und eine Fehlerquote von 4,58% bei 1004 Tokens ermittelt (vgl. Schmidt i. Vorb.). Auf der Grundlage dieser Annotationen kann ein induktiver Zugriff auf die Korpusdaten erfolgen.

[1] Die Bezeichnung des Metadatums als ‚Text' wirkt hier irreführend. Es handelt sich bei den Daten um die aufbereiteten und vereinheitlichten Sprecher*innenbeiträge.

6. Fazit

Die Erstellung und Analyse von Korpora gesprochener Sprache stellen die Korpuslinguistik bislang vor große Herausforderungen. Neben den Eigenschaften mündlicher Sprache, denen nur zu Teilen Transkriptionen – in sowohl qualitativer als auch in quantitativer Hinsicht – gerecht werden können, sind auch die meisten Analysetools auf der Grundlage von schriftsprachlichen Daten entwickelt worden. Die Untersuchung der Verständigung über Literatur im Deutschunterrichtsdiskurs erfordert die Erstellung eines Korpus, das mehrere Unterrichtsstunden abbildet. Über ein solches Korpus könnten über korpuslinguistische Verfahren, in denen quantitative und qualitative Zugriffe kombiniert werden, Anbahnungs- und Sprachentwicklungsprozesse über die Herausstellung musterhaften Sprachgebrauchs herausgearbeitet werden. Insbesondere das Sprechen über Literatur im Sinne einer Kunstform verdeutlicht, dass die Sprache keineswegs satzförmig und linear dargestellt werden kann. Vielmehr gehört zu der Untersuchung dieser Sprache auch die semantische und pragmatische Funktion von Abbrüchen, Versprechern und sprachlich markierten Irritationen in den Blick zu nehmen und auch über maschinelle Analysetools mit zu berücksichtigen. Über das STTS 2.0 können Transkripte, die bestimmten Transkriptionskonventionen und Normalisierungen sowie Dateiformaten entsprechen, so annotiert werden, dass eine Vielzahl von Phänomenen gesprochener Sprache berücksichtigt werden kann.

Doch gerade bestehende, archivierte Transkripte, die in zumeist qualitativen Forschungsprojekten erstellt wurden, sind nicht an Standards und Kriterien für quantitative Untersuchungen von Daten gesprochener Sprache ausgerichtet und bedürfen einer intensiven Aufbereitung. Im Rahmen dieses Beitrags wurde auf einige Herausforderungen eingegangen, die diese Aufbereitung archivierter Transkriptdaten mit sich bringen und Lösungsansätze aufgezeigt, die sich diesen Herausforderungen mit Blick auf die derzeitigen technischen Möglichkeiten stellen. Dabei wurde deutlich, dass die Aufbereitung ein hohes Maß an manueller Arbeit voraussetzt, die bisher nicht maschinell geleistet werden kann, dass eine Kooperation mit Softwareentwickler*innen, Archiven, weiteren Forschungsprojekten und somit eine disziplinenübergreifende Zusammenarbeit unumgänglich ist und dass gerade die Arbeit mit archivierten Daten nicht dem gesamten Potenzial gesprochener Sprache gerecht werden kann. Nichtsdestotrotz ist es möglich, auch archivierte, auf qualitative Zugriffe angelegte Transkripte korpuslinguistisch zu untersuchen und musterhafte Strukturen im Sprachgebrauch auf der Sprachoberfläche aufzudecken.

Literatur und Ressourcen

Abraham, Ulf (2015): Literarisches Lernen in kulturwissenschaftlicher Sicht. In: *Leseräume. Zeitschrift für Literalität in Schule und Forschung* 2, 6-15. http://leseräume.de/wp-content/uploads/2015/10/lr-2015-1-abraham.pdf (15.07.2021).

Brandstätter, Ursula (2011): »In jeder Sprache sitzen andere Augen« - Herta Müller. Grundsätzliche Überlegungen zum »Reden über Kunst«. In: Kirschenmann, Johannes / Richter, Christoph / Spinner, Kaspar H. (Hrsg.): *Reden über Kunst*. München: kopaed (Kontext Kunstpädagogik, Bd. 28), 29–43.

Brommer, Sarah (2018): *Sprachliche Muster. Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte*. Berlin, Boston: de Gruyter (Empirische Linguistik / Empirical Linguistics, 10).

Bruner, Jerome S. et al. (1997): *Wie das Kind sprechen lernt*. 1. Aufl., 2. Nachdr. Bern: Huber (Huber-Psychologie-Sachbuch).

Bubenhofer, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin / Boston: de Gruyter (Sprache und Wissen (SuW), 4).

Büker, Petra (2002): Literarisches Lernen in der Primar- und Orientierungsstufe. In: Bogdal, Klaus-Michael / Korte, Hermann (Hrsg.): *Grundzüge der Literaturdidaktik*. Orig.-Ausg. München: Dt. Taschenbuch-Verl. (dtv, 30798), 120–133.

Enders, Stephanie / Hundsdorf, Sascha (2008): *Unterrichtstranskript einer Deutschstunde an einem Gymnasium (8. Klasse). Stundenthema: "Der Schimmelreiter"*. PDF-Dokument (1 Datei), 47 Seiten. https://archiv.apaek.uni-frankfurt.de/2488 (01.09.2021).

Gerigk, Horst-Jürgen (2013): *Lesen und Interpretieren*. 3. Aufl. Heidelberg: Mattes.

Gölitzer, Susanne (2007): Lesesozialisation. In: Lange, Günter / Weinhold, Swantje (Hg.): *Grundlagen der Deutschdidaktik : Sprachdidaktik - Mediendidaktik - Literaturdidaktik*. 3. Aufl. Baltmannsweiler: Schneider Verl. Hohengehren, 202-225.

Heydebrand, Renate von / Winko, Simone (1996): *Einführung in die Wertung von Literatur. Systematik - Geschichte - Legitimation*. Paderborn, Zürich etc.: Schöningh (UTB, 1953).

Köller, Wilhelm (2004): *Perspektivität und Sprache. Zur Struktur von Objektivierungsformen in Bildern, im Denken und in der Sprache*. Berlin, New York: de Gruyter.

Lehmann, Christian (2007): Daten – Korpora – Dokumentation. In: Kallmeyer, Werner / Zifonun, Gisela (Hrsg.): *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*. Berlin: de Gruyter (Jahrbuch / Institut für Deutsche Sprache, 2006), 9-27.

Lemnitzer, Lothar / Zinsmeister, Heike (2015): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr Francke Attempto Verlag (narr studienbücher).

Mair, Christian (2018): Erfolgsgeschichte Korpuslinguistik? In: Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpuslinguistik*. Berlin, Boston: de Gruyter (Germanistische Sprachwissenschaft um 2020, 5), 5-25.

Maiwald, Klaus (2015): Literarisches Lernen als didaktischer Integrationsbegriff. Spinners „Elf Aspekte“ als Struktur- und Denkrahmen für weiterführende Modellierung(en). In: *Leseräume. Zeitschrift für Literalität in Schule und Forschung* 2, 85-95. http://leseräume.de/wp-content/uploads/2015/10/lr-2015-1-maiwald.pdf (15.07.2021).

Mitterer, Nicola / Wintersteiner, Werner (2015): Literarische Erfahrung. Ästhetischer Modus und literarisches Lernen. In: *Leseräume. Zeitschrift für Literalität in Schule und Forschung* 2, 96-108. http://leseräume.de/wp-content/uploads/2015/10/lr-2015-1-mitterer-wintersteiner.pdf (15.07.2021).

Perkuhn, Rainer / Belica, Cyril (2006): Korpuslinguistik – Das unbekannte Wesen oder Mythen über Korpora und Korpuslinguistik. In: *Sprachreport* 22: 1, 2-8.

Rüdiger, Jan Oliver (2018): CorpusExplorer [Software]. Universität Kassel / Universität Siegen, 01.01.2018. http://www.CorpusExplorer.de (01.09.2021).

Scharloth, Joachim (2018): Korpuslinguistik für sozial- und kulturanalytische Fragestellungen. Grounded theory im datengeleiteten Paradigma. In: Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpuslinguistik*. Berlin, Boston: de Gruyter (Germanistische Sprachwissenschaft um 2020, 5), 61-80.

Schiller, Anne et al. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset).* Institut für maschinelle Sprachverarbeitung (Stuttgart); Universität Tübingen Seminar für Sprachwissenschaft (Tübingen). http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (02.09.2021).

Schmidt, Annika (2016): *Ästhetische Erfahrung in Gesprächen über Kunst. Eine empirische Studie mit Fünft- und Sechstklässlern*. München: kopaed (KREAplus, Band 9).

Schmidt, Christina (2021): Die Verständigung über Literatur im Deutschunterricht – Potenziale und Herausforderungen eines korpuslinguistischen Zugangs. In: Maurer, Christian / Rincke, Karsten / Hemmer, Michael

(Hrsg.): *Fachliche Bildung und digitale Transformation – Fachdidaktische Forschung und Diskurse.* Fachtagung der Gesellschaft für Fachdidaktik 2020. Regensburg: Universität, 151-154.

Schmidt, Christina (i. Vorb.): *Musterhafter Sprachgebrauch beim Sprechen über Literatur im Deutschunterrichtsdiskurs. Eine korpuslinguistische Untersuchung von Deutschunterrichtstranskripten*. Unveröffentlichte Dissertation, Universität Göttingen.

Schmidt, Thomas (2018): Gesprächskorpora. In: Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpuslinguistik*. Berlin, Boston: de Gruyter (Germanistische Sprachwissenschaft um 2020, 5), 209-230.

Schmidt, Thomas / Schütte, Wilfried / Winterscheid, Jenny (2015): cGAT. Konventionen für das computergstützte Transkribieren in Anlehnung an des Gesprächsanalytische Transkriptionssystem 2. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4616/file/Schmidt_Schuette_Winterscheid_cGAT_2015.pdf (01.09.2020).

Spinner, Kaspar H. (2006): Literarisches Lernen. In: *Praxis Deutsch* 200, 6-16.

Spinner, Kaspar H. (2011): Gespräche über Literatur: Was Schülerinnen und Schüler lernen sollen. In: Kirschenmann, Johannes / Richter, Christoph / Spinner, Kaspar H. (Hrsg.): *Reden über Kunst*. München: kopaed (Kontext Kunstpädagogik, Bd. 28), 63-72.

Thomasello, Michael / Carpenter, Melissa (2007): Shared intentionally. In: *Developmental Science* 10: 1, 121-125.

Westphal, Swantje et al. (2017): STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf (21.02.2020).

Winterscheid, Jenny et al. (2019): Normalisieren mit OrthoNormal. Konventionen und Bedienungshinweise für die orthographische Normalisierung von FOLKER-Transkripten. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/9326/file/Winterscheid_etal._Normalisierungskonventionen_2019.pdf (21.02.2020).

Biographische Notiz: Christina Schmidt ist Promotionsstudentin in der Abteilung Didaktik der deutschen Sprache und Literatur an der Georg-August-Universität in Göttingen. Ihre Forschungsschwerpunkte liegen in der Korpuslinguistik und in der Analyse des Sprachgebrauchs im Literaturunterricht.

Kontaktanschrift:
Christina Schmidt
c.schmidt.deutschdidaktik@web.de