Diversitätsorientiert lehren mit einer Open Source-Lösung für Medientranskripte (astAV)

Patrick Löw, Marie Westerdick, Klara Groß-Elixmann, Dirk Burdinski

Zusammenfassung/Abstract *Das Programm astAV (automatic speech recognition toolkit for Audio and Video) ist eine Open Source-Lösung zur Transkript- und Untertitelerstellung für Audio- und Videodateien. astAV wurde im Jahr 2020 als studentisches Projekt erarbeitet und wird seit 2021 kooperativ weiterentwickelt. Im Beitrag wird diskutiert, wie eine umfassende Digitialisierungsinitiative für Untertitelung diversitätsgerecht sowie ressourceneffizient umgesetzt werden kann. Die Genese des Programms verdeutlicht, wie die Gestaltung digitaler Kulturen an Hochschulen maßgeblich auch durch globale Schlüsselereignisse beeinflusst wird. So hat die Veröffentlichung der Sprachbibliothek Whisper durch OpenAI eine breite Implementierung von Spracherkennung weltweit ermöglicht.*

The program astAV (automatic speech recognition toolkit for Audio and Video) is an open source solution for transcript and subtitle creation for audio and video files. astAV was developed in 2020 as a student project and has been undergoing further cooperative development since 2021. The article discusses how a comprehensive digitization initiative for subtitling can be implemented in a diversity-friendly and resource-efficient manner. The genesis of the program illustrates how the design of digital cultures at universities also is significantly influenced by key global events. For instance, the publication of the Whisper language library by OpenAI enables a broad international implementation of speech recognition.

Schlüsselwörter/Keywords *Untertitel; Spracherkennung; Lehrmedien; Diversität; Interkulturalität; Barrierefreiheit; subtitles; speech recognition; educational media; diversity; interculturality; accessibility*

1. Einleitung

Das in diesem Beitrag vorgestellte Open Source Programm *astAV (Automatic speech recognition toolkit for Audio and Video)* basiert auf operativen Bedarfen in der Hochschullehre. Digitale Lehrmedien sind essenziell für studierendenzentrierte und abwechslungsreiche Lehrformate. Insbesondere für die Gestaltung von Inverted- bzw. Flipped-Class-

room-Lehrveranstaltungen erweisen sie sich häufig als unverzichtbar (Burdinski & Glaeser, 2020). Dabei ist es entscheidend, die nachfolgend diskutierten Diversitätsbedarfe der Zielgruppe zu berücksichtigen und Zugänglichkeit zu den Lehrinhalten zu schaffen. Dies ist durch die Erstellung von Transkripten (bspw. für Podcasts) und Untertiteln für Videodateien möglich.

Auf dieser Basis untersucht der Beitrag, wie im Hochschulkontext eine auf offenen Ressourcen basierte Softwarelösung für die Generierung von Untertiteln bzw. Transkripten für bestehende und neu produzierte Lehrmedien (insbesondere Videos, aber auch Podcasts etc.) im Rahmen eines studentischen Projekts entwickelt werden und diese die Bereitstellung und Nutzung diversitätsorientierter Lehrmedien fördern kann. Dazu wurde das Programm *astAV* gestaltet, um Transkription für alle Nutzenden bzw. in der Hochschule für alle Lehrenden leicht, automatisch, datenschutzkonform durch offline Nutzungsmöglichkeit, kostenlos und plattformunabhängig anzubieten. *astAV* wurde als Open Source Programm kreiert und ist auf der Plattform *GitHub* frei zugänglich.

Die freie Verfügbarkeit der Sprachbibliothek *Whisper* ab ihrer Veröffentlichung im Jahr 2022 nimmt maßgeblich Einfluss auf die rasante Entwicklung von kommerziellen Transkriptionsprogrammen. Die nun weitweite Verfügbarkeit von *Whisper* wird in diesem Beitrag als Schlüsselereignis definiert und im Folgenden wird diskutiert, was dieses Schlüsselereignis für die Weiterentwicklung der studentischen Initiative *astAV* bedeutet.

2. Diversitätsorientierte Lehre

Zur umfänglichen Ausschöpfung der Potenziale einer divers zusammengesetzten studentischen Gruppe ist die Zusammenarbeit von großer Bedeutung (Buß, 2018). Damit Lehrformate die Interaktion zwischen Studierenden fördern können, sind jedoch Rahmenbedingungen notwendig, die allen Lernenden gleichermaßen Zugang zu den Medieninhalten und zum anschließenden Austausch ermöglichen. Es gilt festzustellen, welche individuellen Bedarfe berücksichtigt werden müssen (Aye, Dahmen & Karaaslan, 2017). Für die praktische Umsetzung diversitätsorientierter Lehre ist näher einzugrenzen, welche Aspekte der Vielfalt in der Hochschule, der Fakultät oder in einer spezifischen Lehrveranstaltung relevant sind. Das sogenannte *HEAD Wheel* (*Higher Education Awareness for Diversity*, Gaisch & Aichinger, 2016) bietet einen wichtigen Ansatzpunkt. Gaisch und Aichinger (2016) gliedern hochschulrelevante Vielfalt in fünf Bereiche: demografische, kognitive, fachliche, funktionale und institutionelle Diversität. In Bezug auf die individuell unterschiedlichen Bedingungen der Studierenden werden im vorliegenden Beitrag zwei der fünf Diversitätsbereiche betrachtet. Zum einen wird die Bedeutung der demografischen Diversität in Bezug auf die familiäre Situation, physische und psychische Beeinträchtigungen, Internationalisierung und Bildungssozialisierung, zum anderen die kognitive Diversität bezüglich Lernzugängen, Denkweisen, Problemlösestrategien und Informationsverarbeitung diskutiert (Gaisch & Aichinger, 2016).

Durch eine differenzierte Eingrenzung können individuelle Lebenszusammenhänge der Studierenden identifiziert werden, die Einfluss auf Lehr- und Lernkontexte nehmen

(Aye, Dahmen & Karaaslan, 2017). Für die Entwicklung der Untertitelungssoftware *astAV* wurden Diversitätsaspekte in Betracht gezogen, die sich auf den konkreten Lernprozess der Studierenden im Umgang mit audiovisuellen Lehrinhalten auswirken können. Als Programm in der Hochschullehre ist *astAV* darauf ausgerichtet, Lehrende dabei zu unterstützen mediale Lehrinhalte für Studierende bestmöglich zugänglich zu machen. Zugleich gelten die im Folgenden dargestellten Aspekte für alle Akteur:innen der Hochschule. Dabei stehen drei Dimensionen im Fokus: Interkulturalität, Wahrnehmbarkeit und Flexibilität.

2.1 Interkulturalität

Als Teil einer multikulturellen Gesellschaft begegnen Hochschulen einer zunehmend diversen Studierendenschaft. Die Ergebnisse der 22. Sozialerhebung zeigen, dass 17,3 % der Studierenden aus Familien mit Einwanderungsgeschichte kommen, wobei 28,8 % dieser Gruppe selbst im Ausland geboren sind. Zugleich liegt der Anteil internationaler Studierender an deutschen Hochschulen bei 14,8 % (Kroher et al., 2023). Interkulturalität als Diversitätsdimension ist jedoch weitaus umfangreicher und umfasst den kulturellen und ethnischen Hintergrund, die geografische Lage der Hochschule im Verhältnis zum Wohnort sowie die sprachlichen Kompetenzen auf individueller Ebene.

Das Sprachverständnis im Lernkontext bezieht sich nicht allein auf eine Abweichung der sprachlichen Kompetenzen im Vergleich zwischen Erstsprache und der in der Lehrveranstaltung genutzten Sprache, sondern meint ebenso unterschiedliche Voraussetzungen in der Fachsprache. Unter den Studierenden ohne internationale Geschichte kommen 32,7 % aus einem nicht-akademischen Elternhaus (Kroher et al., 2023). Die jeweilige kulturelle Ausgangslage kann zu unterschiedlichen Startbedingungen unter den Erstsemester-Studierenden auch in Bezug auf den Zugang zu akademischer Sprache führen. Daher sollte das Angebot digitaler Lehr- und Lernmedien erweitert werden. Durch die Auseinandersetzung mit vordergründig fachlichen Inhalten auf einem bekannten Niveau können zusätzlich auch sprachliche und kommunikative Kompetenzen gefördert und somit die Studierfähigkeit für alle Studierenden nachhaltig verbessert werden. Aus Medien erlernte Inhalte und Zusammenhänge können in anderen Kontexten, z.B. beim selbstständigen Lehrbuchstudium (textuell) oder in Gruppendiskussionen (auditiv) leichter zugeordnet und angewendet werden. Das Lernen wird dadurch nachhaltiger (Danan, 2004; Harji et al., 2010). Diese Erkenntnisse unterstreichen den Bedarf, Bildungsressourcen interkulturell anzupassen, um inklusive und gleichberechtigte Bildungsformate zu schaffen.

2.2 Wahrnehmbarkeit

Wahrnehmbarkeit als Diversitätsdimension wird hier definiert als das Erfassen der Inhalte und Informationen über die Sinne. Die Dimension bezieht sich auf die individuelle Physis und betont Differenzen in der Wahrnehmung aufgrund körperlicher Bedingungen wie Beeinträchtigungen im Seh- und Hörvermögen sowie in der neuronalen Verarbeitung (bspw. ADHS). Zusätzlich können äußere Faktoren wie die Verfügbarkeit von Hilfsmitteln betrachtet werden: Wenn Studierende eine ähnliche

Sinneseinschränkung haben, jedoch über unterschiedliche adaptive Technologien zur Bewältigung verfügen, entstehen weitere Differenzen. In Umfragen geben fast 16 % der Studierenden in Deutschland an, mindestens eine gesundheitliche studienerschwerende Beeinträchtigung zu besitzen (Kroher et al., 2023). Bei circa 3 % dieser Gruppe handelt es sich um Einschränkungen des Hör- oder Sehvermögens (Kroher et al., 2023). Die Behindertenrechtskonvention der Vereinten Nationen stellt ebenso wie die nationale Rechtslage (GG, 1994, Art. 3, Abs. 3; BGG, §4, 2018) die Anforderung an Bildungseinrichtungen, nicht zu diskriminieren. Daher sind Informationsquellen so zu gestalten, dass sie für alle Studierenden ohne besondere Erschwernis und grundsätzlich ohne fremde Hilfe auffindbar, zugänglich und nutzbar sind (BGG, §4, 2018). Dies kann die Bereitstellung von barrierefreien Lehrmaterialien, sowie alternativen Veranstaltungs- und Prüfungsformaten umfassen.

Hinsichtlich auditiver Lehrmedien bietet das Programm *astAV* Möglichkeiten zur Überwindung von Barrieren. Beispielsweise trägt *astAV* zur Umsetzung des sog. Zwei-Sinne-Prinzips bei (Beauchamp-Gauvin & Groß-Elixmann, 2023). Nach diesem Prinzip werden Informationen parallel über zwei Sinne übermittelt – beispielsweise zugleich visuell und auditiv – um sicherzustellen, dass sowohl Menschen mit Seh- als auch solche mit Höreinschränkungen die Inhalte verstehen können. Podcasts bieten, als ein Beispiel, eine hohe Zugänglichkeit für Menschen, die blind sind, schränken jedoch den Zugang zu den Inhalten für Menschen mit Hörbeeinträchtigungen ein. Durch die Transkription der Inhalte mit *astAV* können auch solche Lehrinhalte über zwei Informationskanäle erfasst werden. Neben Personen mit Sinneseinschränkungen können weitere Studierende von der Softwarelösung *astAV* profitieren. Die Studierendenbefragung in Deutschland ermittelte, dass von den 16 % Studierender mit studienerschwerender Beeinträchtigung circa 5,1 % von einer anderen Beeinträchtigung/Erkrankung, wie beispielsweise einer Autismus-Spektrum-Störung (ASS), Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung (AD(H)S) oder Migräne, betroffen sind (Kroher et al., 2023). Personen, die mit einer ASS, ADHS oder Migräne Lehrinhalte erfassen wollen, profitieren häufig von einer Reizreduktion der Inhalte. Lehrvideos mit grellen farblichen Elementen, schnellen Übergängen oder Hintergrundmusik können ihnen den Zugang erschweren.

2.3 Flexibilität

Die dritte Diversitätsdimension Flexibilität umfasst hier angewandte Lernstrategien und Lernstile sowie individuelle Lernzugänge und Präferenzen der Studierenden. In diesem Beitrag wird die Kategorie um weitere Umstände der Lebenssituation, wie beispielsweise zu leistende Fürsorge- und Betreuungsarbeit im familiären Umfeld, erweitert. Umfrageergebnissen nach bewältigen rund 12 % der Studierenden in Deutschland neben ihrem Studium Care-Arbeit, wie beispielsweise eine Pflegetätigkeit oder Haushaltsbetreuung. Circa 8 % der Studierenden sind zum Zeitpunkt der Befragung bereits Eltern (Kroher et al., 2023). Diese Studierenden benötigen eine hohe Flexibilität, was die Planung und Teilnahme an Vorlesungen und die eigenständige Erarbeitung von Inhalten erschweren kann. Zusätzlich kann die Kombination von Pflegetätigkeiten und Studium die akademischen Leistungen beeinträchtigen, da Studierende weniger Zeit und Energie für ihr Studium aufbringen können. In der Statistik korrespondiert die

Übernahme von Pflegeaufgaben im privaten Bereich mit einem überdurchschnittlichen Anteil an Studierenden, die nicht in Vollzeit studieren (Kroher et al., 2023). Um dennoch gleiche Zugangschancen zu den Inhalten zu ermöglichen, sollten Zeitpunkt und Ort des Lernens flexibel gestaltbar sein (Dinmore, 2019). Heruntergeladene Transkripte sind beispielsweise offline verfügbar und bieten eine gute Ergänzung zu den Videos. Zusätzlich können Studierende durch die Suchfunktion von Schlagwörtern in Textdokumenten gezielter Informationen finden und schneller Verknüpfungen zwischen Inhalten herstellen. Mit der Transkriptionsfunktion von astAV können Lernmaterialien individueller angepasst und leichter zugänglich gemacht werden.

Darüber hinaus beeinflussen weitere Faktoren die Lernvoraussetzungen der Studierenden. Laut Buß (2013) kann der Erfolg im Lernprozess ebenso durch individuelle Lernstrategien und Lernstile geprägt sein. Forschungsergebnisse der Kognitionswissenschaft, Psychologie und Neurowissenschaft unterstreichen die Bedeutung der Individualisierung des Lernens, woraus sich ein Plädoyer für die Bereitstellung gleicher Informationen über verschiedene Modalitäten ergibt (CAST, 2018).

3. *astAV*: Forschungsfrage und Methode

Wie eingangs erläutert, untersucht dieser Beitrag wie im Hochschulkontext eine auf offenen Ressourcen basierte Softwarelösung für die Generierung von Untertiteln bzw. Transkripten für Lehrmedien im Rahmen eines studentischen Projekts entwickelt werden und diese die Bereitstellung und Nutzung diversitätsorientierter Lehrmedien fördern kann. In diesem Kontext wurden an der Fakultät für Angewandte Naturwissenschaften der TH Köln in den Jahren 2015–2022 jährlich Teaching Analysis Polls (TAP) durchgeführt (Frank et al., 2011). Hierin betonten insbesondere die internationalen Studierenden die Bedeutung der Lernvideos für den eigenen Lernerfolg und wünschten sich eine durchgehende Untertitelung. In den Jahren 2017–2019 wurden daher im Rahmen eines im Programm »Studienstart MINTernational« geförderten Projekts (»EACh – Erfolgreich ankommen im Chemiestudium«) alle an der Fakultät in deutscher Sprache produzierten Lehrvideos deutsch untertitelt. Dieses Vorgehen wurde anschließend für alle neu erstellten Videos bis heute weitergeführt (Burdinski & Rausch, 2021; Burdinski, 2022, 2023a). Diese Lehrvideos sind als offene Bildungsressourcen auf der Videostreaming-Plattform YouTube zugänglich und werden intensiv genutzt (Burdinski, 2018a; 2018b; 2019). Auffällig war im Wintersemester 2021–2022 und damit dem zweiten Studienjahrbeginn unter den Bedingungen der Corona-Pandemie eine stark zunehmende Zahl der Videoaufrufe unter Nutzung der erstellten Untertitel (Burdinski, 2023b).

Die manuelle Erstellung von Untertiteln ist ressourcenaufwendig. Für die vollständige Untertitelung eines ca. fünfzehnminütigen Videos war zu Projektbeginn inklusive sprachlicher Korrekturen und passender Formatierung ein Zeitaufwand von insgesamt etwa einem Arbeitstag erforderlich. Bei vorliegendem Produktionsskript verringerte sich dieser Aufwand auf etwa einen halben Tag. Dies war vergleichbar mit der eigentlichen Produktion des Videos. Daher sollte im Sinne der Forschungsfrage eine technische Lösung sowohl für die inhaltliche Erstellung der Untertitel als auch für deren Formatierung mit Zeitmarkierungen entwickelt werden.

Im Jahr 2016 wurde der für die Untertitelerstellung erforderliche Workflow analysiert und dokumentiert. Hieraus wurden Funktionsanforderungen an die technische Lösung sowie rechtliche Anforderungen an den eigentlichen Betrieb und die nachhaltige Nutzung abgeleitet. Dazu gehörten insbesondere die freie Verfügbarkeit des Programms und die freie Weiterentwicklung des Softwarecodes. Bis 2020 gab es nur im Videoschnittprogramm *Kdenlive* eine integrierte Untertitelfunktion sowie separate Onlinedienste wie z.B. *IBM Watson*, *Google Speech-to-Text Services* oder *YouTube*. Im gleichen Zeitraum wurden mit *VOSK* und *Mozilla DeepSpeech* Open Source-Spracherkennungsbibliotheken veröffentlicht. Auf dieser Basis wurde ein exploratives Bachelorprojekt definiert (Löw, 2021), welches eine Konzeption für ein Software-Programm mit den folgenden Anforderungen entwickelte. Die Benutzeroberfläche soll insbesondere für Anwender:innen ohne technischen Hintergrund einfach zu bedienen und die Untertitelung von erstellten Videos ohne Skriptvorlage möglich sein. Des Weiteren ist die Fähigkeit zur Transkription bereits bestehender Audio- und Videodateien essentiell. Die Software strebt eine präzise Sprachausgabe an, die neben korrekter Rechtschreibung und Zeichensetzung auch fachsprachliche Inhalte erfasst. Ihr modularer Aufbau ermöglicht die unkomplizierte Implementierung zusätzlicher Funktionalitäten, Sprachen und Sprachmodelle. Durch einen offenen Programmcode wird die freie Weiterentwicklung, Ergänzung und Anpassung der Software ermöglicht.

Eine erste Programm-Rohversion (*astAV* 2022) wurde zunächst mit unterschiedlichen Medien getestet. Die Anforderungen wurden weiter spezifiziert und das Programm im MediaLab der TH Köln am Campus Leverkusen weiterentwickelt. Ab Ende 2022 wurde die Softwarepakete *NeMo*, ab Mai 2023 dann auch *Whisper* als Spracherkennungsmodul implementiert. Seit 2016 wurden von insgesamt 207 erstellten Videos 31 Videos mithilfe von *astAV* untertitelt.

3.1 Funktionsweise und Limitationen des Programms

astAV wurde als Open Source Programm entwickelt und steht auf der Plattform *GitHub* zur Verfügung. Die Transkription findet direkt auf dem Computer der nutzenden Person statt und ist somit in Bezug auf datenschutzrechtliche Vorgaben unbedenklich. Es soll allen Benutzer:innen die Möglichkeit geben, ohne technisches Vorwissen Untertitel oder Transkripte aus Video- und Audiodateien zu generieren, und ist plattformunabhängig einsetzbar. Über eine grafische Benutzeroberfläche lassen sich die gewünschten Video- oder Audiodateien in eine Warteschlange ziehen und verarbeiten, womit die Anforderung einer automatischen Verarbeitung erfüllt ist. Das fertige Transkript oder die Untertitel werden anschließend als einfach zu bearbeitende Textdateien neben den Video- oder Audiodateien im zugehörigen Projektordner erstellt. Das Programm *astAV* kann die meisten gängigen Video- und Audioformate verarbeiten, um den Benutzer:innen weitere Arbeitsschritte für eine Übertragung in ein eventuell vorgegebenes Dateiformat zu ersparen. Die in der aktuellen Version verwendete Spracherkennungs-Software *Whisper* unterstützt nahezu alle Sprachen. Als Ausgabeformate stehen die Untertitelformate SRT und WebVTT sowie TXT zur Auswahl, wobei es sich um strukturierte, unformatierte Textdateien handelt, aus denen Videoprogramme Untertitel darstellen können.

3.2 Technischer Aufbau

Das in der Programmiersprache *Python* geschriebene Programm *astAV* kann alle Schritte einer Untertitelung oder Transkription automatisieren. Dazu gehören erstens die Audioextraktion, zweitens die Spracherkennung des Audios und drittens die Formatierung des erkannten Textes. Jeder Prozessschritt sowie die Benutzeroberfläche von *astAV* wurden in verschiedene Komponenten aufgeteilt, die untereinander durch definierte Schnittstellen verbunden sind. Die Aufteilung ermöglicht eine einfache Änderung von einzelnen Programmteilen. Für die Spracherkennung und die Audioextraktion wird auf bestehende Spracherkennungsbibliotheken zurückgegriffen. Diese Bibliotheken bieten Funktionen, die häufig in Programmen gebraucht werden und so einfach wiederverwendet werden können. Hierzu gehören mathematische Funktionen oder die Anzeige einer Schaltfläche. Für die Anwender:innen sind diese Bibliotheken nicht zugänglich.

Eine wichtige Komponente ist die Benutzeroberfläche von *astAV*: Diese kann alle verfügbaren Spracherkennungsbibliotheken auslesen und den Nutzer:innen zur Auswahl bereitstellen. Aktuell können *VOSK* von alphacephei, *NeMo* von Nvidia und *Whisper* von OpenAI als Spracherkennungsbibliothek genutzt werden. Um eine andere Spracherkennung hinzuzufügen, ist aufgrund der Architektur des Programms keine Änderung der Benutzeroberfläche notwendig. So wurden auch *NeMo* und *Whisper* im späteren Entwicklungszeitraum hinzugefügt. Eine weitere Komponente, neben der vorgeschalteten Audioextraktion, ist die Speicherung in verschiedene Untertitel- oder Textformate. Die Benutzeroberfläche unterstützt aktuell die deutsche und englische Sprache, die sich nach dem Start des Programms, entsprechend der im Betriebssystem eingestellten Sprache, anpasst. Weitere Sprachen können durch die Erstellung einer Übersetzungsdatei in das Programm integriert werden.

3.3 Limitationen

Da das Programm auf den einfachen Austausch der zugrundeliegenden Spracherkennung ausgelegt war, wurden nur sehr grundlegende Funktionen in der Architektur von *astAV* festgelegt. Dies hat gewisse Einschränkungen in der Funktionalität des Programms zur Folge. Erstens ist es nicht möglich, für die Transkription ein direktes Eingabegerät wie ein Mikrofon zu nutzen. Es muss also immer eine Video- oder Audiodatei vorliegen. Die fehlende individuelle Identifizierung der Sprecher:innen stellt eine zusätzliche Einschränkung dar. So ist es *astAV* nicht möglich, die einzelnen Personen in einem Video voneinander zu unterscheiden. Eine weitere Limitation betrifft den Ressourcenverbrauch. Die Spracherkennungssoftware braucht, je nach eingesetzter Methode, Wortschatz und Präzision des Systems, einen vergleichsweise hohen Rechenaufwand. So können schwache Computer ohne Grafikkarte nur kleine Modelle der Spracherkennung *Whisper* nutzen. Diese Modelle sind schneller in der Verarbeitung, weisen aber eine höhere Wortfehlerrate auf. Um die mittleren bis großen Modelle zu nutzen, ist eine Grafikkarte des Herstellers Nvidia mit 6 bis 12 Gigabyte Speicher erforderlich.

4. Schlüsselereignis Veröffentlichung von Whisper

4.1 Entwicklung von Spracherkennungssystemen seit den 1970er Jahren

Zur Kontextualisierung der als Schlüsselereignis definierten Veröffentlichung der Spracherkennungsbibliothek Whisper in 2022 wird im Folgenden erläutert, wie sich automatisierte Spracherkennung insgesamt entwickelt hat. Durch diesen Rückblick wird der enorme Schritt in der technologischen Entwicklung von Spracherkennung deutlicher und die Bezeichnung als Schlüsselereignis begründet. Spracherkennungssoftware zum Erkennen von Wörtern wurde zwar schon seit 1972 entwickelt, musste aber zunächst auf die sprechende Person angepasst werden. In den ersten Versionen war der Wortschatz der Systeme mit einem Umfang von ca. 200 Wörtern sehr beschränkt (Scott & Nj, 1975).

Trotz der schnell voranschreitenden Entwicklung wurden erst in den späten 1990er Jahren die Diktierprogramme *Dragon Dictate* von Dragon Systems und *ViaVoice* von IBM für den Einsatz auf üblichen Desktop-Computern vorgestellt (Scannell, 1997). Zur Transkription wurde die Video- oder Audiodatei bei *ViaVoice* durch eine:n Sprecher:in rezipiert, die den Inhalt für die Diktiersoftware nachgesprochen hat. Ein großes Problem war dabei weiterhin der von der Software genutzte Wortschatz und die Geschwindigkeit. 2003 war die Entwicklung so weit fortgeschritten, dass die britische BBC erstmals Nachrichtensendungen mit Diktiersoftware ergänzte, um damit Untertitel sowohl für live ausgestrahlte als auch für aufgezeichnete Sendungen zu generieren (Evans, 2003).

In den letzten 20 Jahren hat sich die Spracherkennung kontinuierlich weiterentwickelt. Wesentlich ist hierbei die wachsende Leistung der Computerhardware, durch die heute auch die Nutzung von neuronalen Netzen zur Spracherkennung möglich wird (Pfister & Kaufmann, 2017). Diese Netze können mit immer größeren Datensätzen vieler verschiedener Sprecher:innen trainiert werden, wodurch sie von diesen unabhängig werden. Einige dieser neuen Ansätze wurden im Rahmen wissenschaftlicher Veröffentlichungen als sogenannte Spracherkennungsbibliotheken kostenlos zur Verfügung gestellt (Kuchaiev et al., 2019; Radford et al., 2022). Diese Bibliotheken lassen sich zwar für die Softwareentwicklung nutzen, sind in der Ursprungsform für nicht spezialisierte Anwender:innen allerdings nicht verwendbar. Das Programm *astAV* kann die technischen Möglichkeiten dieser Spracherkennungsbibliotheken auf eine benutzungsfreundliche Weise zugänglich machen, insbesondere für Personen ohne Vorkenntnisse.

4.2 Entwicklung im Zeitraum nach 2020

Die Weiterentwicklung der Spracherkennung ist im Jahr 2023 vor allem mit *Whisper* von OpenAI verknüpft. Die Open Source-Spracherkennungsbibliothek ist 2022 veröffentlicht worden und wird fortlaufend weiterentwickelt. *Whisper* ist einfach zu integrieren und kann nahezu alle Sprachen erkennen und transkribieren, ohne dass diese vorher festgelegt werden müssen (Radford et al., 2022). Die Zugänglichkeit, die einfache Nutzung und die hohe Sprachgenauigkeit von *Whisper* führten unter anderem dazu, dass eine große Zahl unterschiedlicher Programme bereits diese oder ähnliche lokale

Spracherkennungen in den Funktionsumfang aufgenommen hat. Mit der Verfügbarkeit hochwertiger automatischer Spracherkennung ist die darüber hinaus notwendige Audioextraktion und die Formatierung des erkannten Textes leicht zu implementieren. So bieten zwei der großen Schnittprogramme, *Davinci Resolve* und *Adobe Premiere Pro* nun eine eigene automatische und lokal ausgeführte Untertitelfunktion (Blackmagic Design, 2023; Adobe, 2022). Auch *Kdenlive*, ein Open Source-Schnittprogramm, verfügt inzwischen über eine solche Funktion (Mohr, 2023). Gleichermaßen sind verschiedene Open Source-Untertitelprogramme inzwischen mit einer automatischen Transkription verfügbar (Olsson, 2022). Viele dieser Programme bieten damit einen größeren Funktionsumfang als *astAV* in der Version 1.0 im November 2023.

5. Bedeutet digitale Verfügbarkeit mehr Diversitätsorientierung?

Durch die Veröffentlichung der frei zugänglichen Spracherkennungsbibliotheken und ihre Adaption in kommerziellen Anwendungsprogrammen kann die Untertitelung und Transkription inzwischen einfach auf den Geräten der Nutzer:innen und zum Teil auch direkt in den von ihnen genutzten Programmen durchgeführt werden. Dies reduziert die Einstiegshürde zur Erzeugung von Untertiteln und Transkripten erheblich. Verglichen mit einer manuellen Transkription sinkt der Arbeitsaufwand nach den bisherigen Erfahrungen im Projekt bei fünfzehnminütigen Lehrvideos von ca. acht Stunden auf etwa 30 Minuten, inklusive Nachbearbeitung und finaler Korrektur. Damit ist für den deutsch- und englischsprachlichen Bereich die Herausforderung der Transkription, sowohl im eigenen Projektkontext als auch weltweit, technologisch gelöst.

Damit wird hier argumentiert, dass die Veröffentlichung von *Whisper* ein Schlüsselereignis ist, das die Kommunikation, den Umgang und die weitere Nutzung von digitaler Spracherkennung dauerhaft beeinflussen wird (Rauchenzauner, 2008). Zugleich ist die Verfügbarkeit damit auf einem Niveau, das im Jahr 2016, zu Beginn des Projektes, undenkbar war. Es zeigt sich, dass hochschuleigene Entwicklungen oft zwar schneller auf akute Bedarfe reagieren, auf längere Sicht aber nicht immer mit dem rasanten Fortschritt kommerzieller Anbieter Schritt halten können.

Auch wenn die Transkription von audiovisuellen Lehr- und Lernmedien einfacher geworden ist, bleibt deren Nutzung ein Problem. Die in den Hochschulen genutzten Lernplattformen bzw. Lernmanagementsysteme (LMS) unterstützen grundsätzlich die Einbettung von Untertiteln. Diese Funktionen können aber je nach Software und Version nicht standardmäßig aktiviert oder abgeschaltet sein. Obwohl sie diese Funktionen bieten, handelt es sich zudem nicht um dezidierte Videostreaming-Plattformen, sie können also meist keine hohen und damit lernförderlichen Übertragungsraten gewährleisten. Streaming-Plattformen wiederum stellen leider nicht viele Hochschulen ihren Lehrenden zur Verfügung. Inwieweit hochschulübergreifende Plattform-Angebote hier Abhilfe schaffen können, ist derzeit schwer absehbar. Die weiterhin übliche Nutzung kommerzieller Plattformen, wie YouTube, bleibt ein unbefriedigender Kompromiss.

Im Hinblick auf die eingangs definierten Anforderungen kann das an der TH Köln entwickelte Programm *astAV* die Transkription sowie die Untertitelung audiovisueller Medien für alle Nutzenden leicht, automatisch, datenschutzkonform durch offline Nut-

zungsmöglichkeit, kostenlos und plattformunabhängig anbieten und damit die geforderten diversitätsförderlichen Impulse geben. *astAV* adressiert die drei Dimensionen diversitätsorientierter Lehre, indem es deutschen und internationalen Studierenden mit Herausforderungen im sprachlich-kognitiven Bereich gleichermaßen neue Zugänge zu digitalen Lernmaterialien eröffnet und allen Studierenden mehr Wahlmöglichkeiten für die individuelle Gestaltung ihrer Lernaktivitäten ermöglicht. Obwohl diese Anforderungen nun auch durch die Dienste kommerzieller Anbieter:innen gelöst werden können, kann *astAV* als Good-Practice-Projekt auch für andere Lehrentwicklungsinitiativen im Bereich digitaler Medien wichtige Impulse geben. Die Entwicklung von *astAV* trug dazu bei, Diversitätsbedarfe der digitalen Lehre an der TH Köln herauszustellen und hochschulinterne Prozesse zur Verbesserung herauszuarbeiten. Zudem hat der Forschungsprozess, der zu *astAV* geführt hat, den Studierenden der Fakultät für Angewandte Naturwissenschaften der TH Köln bereits deutlich vor dem Schlüsselereignis *Whisper* Untertitelung angeboten und damit diversitätsorientierte Lehre gefördert.

Literatur

Adobe (2022). Funktionszusammenfassung | Premiere Pro (Version Februar 2022). https://helpx.adobe.com/de/premiere-pro/using/whats-new/2022-2.html

Aye, M., Dahmen B., & Karaaslan, N. (2017). Diversity-Kompetenz in der Hochschullehre: Ein E-Learning-Tool für Hochschullehrende. In B. Berendt et al. (Hg.), *Neues Handbuch Hochschullehre* (F4.6, S. 1–18). DUZ Verlags- und Medienhaus GmbH.

Beauchamp-Gauvin, F. R., & Groß-Elixmann, K. (2023). Barrieren in der Lehre erkennen und Sensibilität für Barrierefreiheit fördern. In B. Berendt et al. (Hg.). *Neues Handbuch Hochschullehre* (F4.8, S. 1–16) DUZ Verlags- und Medienhaus GmbH.

BGG (2018). https://www.gesetze-im-internet.de/bgg/index.html

Blackmagic Design (2023). Media | BlackMagic Design. https://www.blackmagicdesign.com/media/release/20230416-03

Burdinski, D. [Chemie Grundlagen] (2018a, 06.09). Kanaltrailer – Chemie Grundlagen [Video]. YouTube. https://www.youtube.com/watch?v=wIFPmlCSUIQ

Burdinski, D. [Anorganische Chemie]. (2018b, 19.09). Kanaltrailer Anorganische Chemie [Video]. YouTube. https://www.youtube.com/watch?v=vLI3A-W_vOE

Burdinski, D. [Praktikum Anorganische Chemie]. (2019, 13.11.). Kanaltrailer Praktikum Anorganische Chemie [Video]. YouTube. https://www.youtube.com/watch?v=u3oiokpKAL8

Burdinski, D., & Glaeser, S. (2020). Flipped Lab – Effektiver lernen in einem naturwissenschaftlichen Grundlagenpraktikum mit großer Teilnehmerzahl. In M. Deimann et al. (Hg.), *Digitalisierung der Hochschullehre* (S. 145–169) DUZ open.

Burdinski, D., & Rausch, E. (2021). Teilvirtuelle Umgestaltung eines Chemie-Laborpraktikums – Maßnahmen und Wirkungen, In M. Barnat, E. Bosse, B. Szczyrba (Hg.), *Forschungsimpulse für hybrides Lehren und Lernen an Hochschulen* (S. 193–212). Cologne Open Science.

Burdinski, D. (2022). Problemfeld Laborpraktika – Wie Studierende durch eine multimedial unterstützte Vorbereitungsphase in ihrer Handlungskompetenz gefördert

werden können. In N. Leben et al. (Hg.), *Hochschullehre als Gemeinschaftsaufgabe – Akteur:innen und Fachkulturen in der lernenden Organisation* (S. 33–39). wbv Publikation. DOI: 10.3278/6004857w

Burdinski, D. (2023a). Ausprägungen und Wirkungen eines teilvirtualisierten Flipped Lab. In N. Vöing et al. (Hg.), *Aktive Teilhabe fördern: ICM und Student Engagement in der Hochschullehre* (S. 83–102). Visual Ink Publishing.

Burdinski, D. (2023b). Lehrvideos und virtuelle Lernumgebungen in der Studieneingangsphase: Anforderungen und Wirkungen im Grenzbereich Schule, Hochschule und Gesellschaft. In L. Mrohs, J. Franz, D. Herrmann, K. Lindner & T. Staake (Hg.), *Digitale Kulturen der Lehre entwickeln: Rahmenbedingungen, Konzepte und Werkzeuge* (S. 369–392). Springer VS.

Buß, I. (2013). Diversity im Kontext von Organisationsentwicklung: Lernprozesse in den Mittelpunkt stellen. In B. Berendt et al. (Hg.), *Neues Handbuch Hochschullehre* (S. 1–30). DUZ Verlags- und Medienhaus GmbH.

Buß, I. (2018). Erfolgreich studieren mit Beeinträchtigung durch Interaktionen im Studium. *Beiträge zur Hochschulforschung, 40*(3), 56–77. https://www.bzh.bayern.de/uploads/media/3_2018_Buss.pdf

CAST (2018). Universal Design for Learning Guidelines version 2.2. Retrieved from http://udlguidelines.cast.org

Danan, M. (2004). Captioning and Subtitling. Undervalued Language Learning Strategies. *Meta: Journal des traducteurs, 49*(1), 67–77.

Dinmore, S. (2019). Beyond lecture capture: Creating digital video content for online learning – a case study. *Journal of University Teaching & Learning Practice, 16*(1). https://doi.org/10.53761/1.16.1.7

Frank, A., Fröhlich, M., & Lahm, S. (2011). Zwischenauswertung im Semester: Lehrveranstaltungen gemeinsam verändern. *Zeitschrift für Hochschulentwicklung, 6*(3), 310–318.

Gaisch M., & Aichinger R. (2016). Das Diversity Wheel der FH OÖ: Wie die Umsetzung einer ganzheitlichen Diversitätskultur an der Fachhochschule gelingen kann. http://ffhoarep.fh-ooe.at/bitstream/123456789/637/1/114_215_Gaisch_FullPaper_Final.pdf

Harji, M. B., Woods, P. C., & Alavi, Z. K. (2010). The Effect Of Viewing Subtitled Videos On Vocabulary Learning. *Journal of College Teaching & Learning (TLC), 7*(9), 37–42.

Kroher, M. et al. (2023). Die Studierendenbefragung in Deutschland: 22. Sozialerhebung. Die wirtschaftliche und soziale Lage der Studierenden in Deutschland 2021. Bundesministerium für Bildung und Forschung. https://www.studierendenwerke.de/fileadmin/api/files/Soz22_Hauptbericht.pdf

Löw, P. (2022). Automatische offline Untertitelerstellung, mithilfe KI gestützter Spracherkennung. (Unveröffentlichte BA-Arbeit)

Mohr, E. (2023). Kdenlive 23.04.0. https://kdenlive.org/de/2023/04/kdenlive-23-04-0-2/

Olsson, N. L. (2022). Nikse.dk. https://nikse.dk/subtitleedit

Radford, A. et al. (2022). Robust speech recognition via Large-Scale Weak Supervision. *arXiv*, 1–28. https://doi.org/10.48550/arxiv.2212.04356

Rauchenzauner, E. (2008). *Schlüsselereignisse in der Medienberichterstattung*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90951-6_2

Scannell, E. (1997). IBM dictation software package gives computers a voice. *InfoWorld 24*, 36.