Jan Měškank / Astrid Schmiedel

Zur Vorlesefunktion für die nieder- und obersorbische Schriftsprache – Einführung und Überblick

In der digitalisierten Welt sind viele Inhalte und Informationen schon oder teilweise sogar nur noch im Internet zu finden – meist in Textform. Für Personen mit Sehbehinderungen oder anderen Einschränkungen beim Lesen stellt dies eine besondere Herausforderung dar. Für die meisten großen Sprachen existieren mittlerweile technische Lösungen in Form von Sprachausgaben bzw. Vorleseprogrammen zur Einbindung in browserintegrierte ScreenReader oder als eigenständige Software, die den gewünschten Text auditiv zugänglich machen. Gebräuchliche Produkte sind beispielsweise der Voice Reader von Linguatec, der webReader von ReadSpeaker oder die Browser Erweiterung Read Aloud.[1] Eine Vorlesefunktion erfasst und analysiert den vorzulesenden Text im Internetbrowser und gibt ihn dann zumeist mit einer synthetisierten Stimme wieder. Diese Software kann darüber hinaus auch für weitere Zwecke genutzt werden, beispielsweise für Navigationsgeräte, die Kommunikation auf Bahnhöfen oder in öffentlichen Verkehrsmitteln, zum Lernen und Lehren von Sprachen, für Lernspiele oder allgemein Computerspiele.

Für die nieder- und obersorbische Sprache gibt es solch ein technisches Hilfsmittel bisher nicht. Bei der Implementierung einer Vorlesefunktion für die deutschsprachigen Texte auf der Website des Sorbischen Instituts e. V. mittels des Linguatec Voice Readers wurde deutlich, dass für die Barrierefreiheit notwendige Maßnahmen für die sorbischen Sprachen aktuell nur schwer umsetzbar sind.

1. Ein Projekt des Sorbischen Instituts

Im Zusammenhang mit der Umsetzung eines Aktionsplans zur Inklusion gemäß der UN-Behindertenrechtskonvention wurde Anfang 2018 im Sorbischen Institut e. V. (SI) sondiert, ob die Entwicklung einer Vorlesefunktion mit den vorhandenen technischen Mitteln, dem aktuellen Forschungsstand und dem verfügbaren Fachpersonal in einem angemessenen Zeitrahmen durchführbar wäre. Da die sorbischen Sprachen als Minderheitensprachen digital unterversorgt sind, stellt ein solches Vorhaben eine besondere Herausforderung dar. Viele technische und wissenschaftliche Grundlagen, die für große Sprachen bereits allgemein und leicht zugänglich sind, müssen zuerst erarbeitet werden, um eine hohe Produktqualität und breite Akzeptanz bei den potenziellen Endnutzern sicherzustellen. Außerdem gilt es zu beachten, dass trotz naher Verwandtschaft beider Sprachen für Nieder- und Obersorbisch jeweils eine eigenständige Vorlesestimme geschaffen werden muss. Einzelne Arbeitsschritte können zwar parallel durchgeführt werden, doch der Arbeitsaufwand insgesamt wird dadurch nur geringfügig kleiner.

Nach der Sondierung mit positivem Ergebnis wurden ab Mitte 2018 vom Sächsischen Staatsministerium für Wissenschaft und Kultur (SMWK) im Rahmen der SMWK-Richtlinie Inklusion Fördermittel für eine sechsmonatige „Konzeptionsphase für eine Vorlesefunktion Sorbisch“ als Maßnahme zur Verbesserung der kommunikati-

---

[1] Momentan verfügbar für Google Chrome, Mozilla Firefox und Microsoft Edge [18.01.2022].

ven Barrierefreiheit bewilligt. Für die Umsetzung wurde eine Arbeitsgruppe aus Phonetikern, Sprachwissenschaftlern und Computerlinguisten gebildet. Das Ziel des Projekts ist eine Anwendung zur Ausgabe in erster Linie akustisch gut verständlich gesprochenen Textes. Die synthetische Aussprache wird dabei als bestmögliche Annäherung an eine „gute natürliche Aussprache“ definiert, da eine durchweg perfekte Aussprache, auch im Sinne einer Orthoepie, im technischen Kontext ausgeschlossen ist.

1.1 Konzeptionsphase – Juli bis Dezember 2018

Die Entwicklung einer Vorlesefunktion für Nieder- und Obersorbisch in Eigenregie bedurfte der Klärung einer Reihe von Fragen sowohl im linguistischen als auch im technischen Bereich. Die Konzeptionsphase gliederte sich dazu in drei Hauptarbeitspakete.

Zunächst wurde der konzeptionelle Rahmen erarbeitet, d. h. grundlegend erforderliche Module und Teilbereiche wurden identifiziert, geprüft, zusammengestellt sowie konkrete Anforderungen definiert. Aus dem Bereich der klassischen Linguistik und insbesondere der phonetischen Linguistik wurden die Teildisziplinen Phonologie, Phonetik und Prosodie als besonders relevant eingestuft, nachrangig sind auch Morphologie und Syntax gelistet. Speziell computerlinguistisch sind Verfahren zur Verarbeitung natürlicher Sprache – *natural language processing tools* (NLP-Tools) – von hoher Wichtigkeit, ebenso die Struktur und Funktionsweise von Sprachsynthese in sogenannten TTS-Systemen (*text-to-speech*). Hieran schließt sich direkt der Komplex der Programmierung an, in dem Kriterien für die Evaluation möglicher Sprachsynthese-, aber auch Signalverarbeitungssoftware zusammengestellt wurden. Schließlich fanden auch sprachtechnologische Überlegungen in Bezug auf Verarbeitung, Segmentierung und Annotation gesprochener Sprache Eingang in die Konzeption.

Im zweiten Paket wurde geprüft, welche Komponenten der erforderlichen Grundlagenmodule bereits vorliegen und damit nutzbar sind, welche in Ansätzen vorhanden und adaptier- und erweiterbar sind bzw. welche gänzlich fehlen, also ggf. neu entwickelt werden müssen. Dazu wurden im linguistischen Bereich die bisherigen Forschungen zur Phonetik des Nieder- und Obersorbischen zusammengetragen, der Forschungsstand mit besonderem Augenmerk auf der Beschreibung des Lautinventars, distinktiver Merkmale, Graphem-Phonem-Beziehungen, Ausspracheregeln, Akzent und Betonung erfasst und Desiderate aufgezeigt. Im technischen Bereich wurden verfügbare TTS-Systeme bezüglich ihrer spezifischen Anforderungen, Systemeinbindung, Anwendung, Funktionsumfang etc. sondiert. Nach der Evaluation anhand der im ersten Paket erstellten Kriterien fiel die Entscheidung für die Open-Source-Software MaryTTS. Das System wurde als Gemeinschaftsprojekt des Forschungsbereichs Sprachtechnologie am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) Saarbrücken und dem Institut für Phonetik der Universität des Saarlandes entwickelt, Näheres siehe 3. Weiterhin wurden der technische Stand vorhandener sprachlicher Ressourcen und Werkzeuge der Text- und Korpusanalyse sowie Möglichkeiten zur Einbindung bzw. zielgerichteten Weiterentwicklung erfasst.

Im dritten Schritt wurden schließlich auf Grundlage der bisherigen Pakete verschiedene Tests einzelner Arbeitsschritte und Komponenten durchgeführt, um einen validen Arbeits- und Zeitaufwand abschätzen zu können. Darauf basierend entstand ein Ar-

beitsplan für die Umsetzung des Vorhabens ab 2019. (Für die Hauptphase wurden bereits Änderungen berücksichtigt, die wegen der Covid-19-Pandemie notwendig waren.)

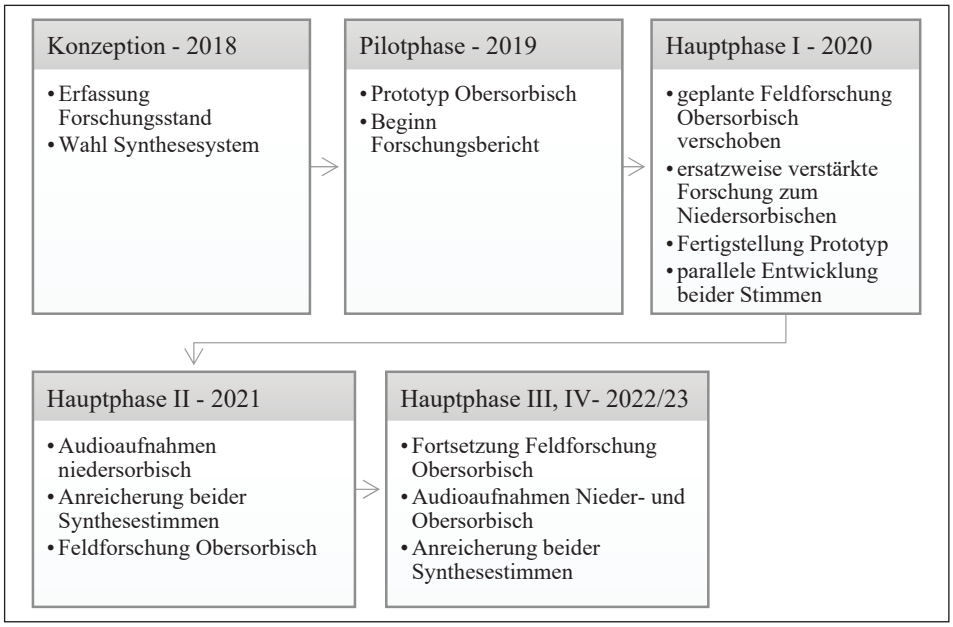

Abbildung 1: Arbeitsplan zum Projekt Vorlesefunktion

1.2 Pilotphase – 2019

Der Konzeptionsphase schloss sich 2019 eine Pilotphase an, in der u. a. die ausführliche Beschreibung des aktuellen Stands der Forschung im Bereich der nieder- und obersorbischen Phonetik und Phonologie begonnen und ein darauf aufbauender Forschungsplan entwickelt wurde. Die Beschreibung ist im Laufe des Projekts zu einem umfangreichen Forschungsbericht angewachsen, der kontinuierlich erweitert wird.

Auf technischer Seite wurde die gewählte Synthesesoftware auf einem virtuellen Linuxsystem installiert. Sprachsynthesesoftware ist ein komplexes System voneinander abhängiger Einzelkomponenten (siehe 3. und Abb. 3), die im Zusammenspiel die Synthesequalität bestimmen. Bei der Erarbeitung dieser Einzelkomponenten handelt es sich um einen kontinuierlichen und iterativen Prozess, bei dem wiederholt vom aktuellen Synthesestand ausgehend alle Komponenten geprüft und verbessert werden, um die Qualität hinsichtlich Verständlichkeit, Wohlklang und der Nähe zur angestrebten Aussprache zu erhöhen. Um mit dem gewählten System vertraut zu werden, sollten mit einem Minimum an erforderlichen Komponenten die Abläufe von MaryTTS durchgespielt, Probleme identifiziert und Anforderungen für die Hauptumsetzungsphase (ab 2020) spezifiziert werden. Aufgrund der besseren Daten- und Forschungslage wurde für diesen internen Prototyp die obersorbische Sprache ausgewählt. Die technische Umsetzung erforderte wider Erwarten aufwendige Softwareanpassungen, die innerhalb der Pilotphase nicht abgeschlossen werden konnten. In Zusammenarbeit mit den MaryTTS-

Entwicklern erfolgten die nötigen Anpassungen abweichend vom Plan innerhalb des ersten Quartals des Folgejahres.

1.3 Hauptphase I – 2020

Für das erste Jahr der Hauptphase waren zunächst Feldforschungen zum Obersorbischen vorgesehen, um die im Forschungsbericht identifizierten Lücken zu schließen und für die Synthese wichtige Ausspracheentscheidungen fundiert treffen zu können. Aufgrund der COVID-19-Pandemie konnte das Vorhaben jedoch 2020 nicht angegangen werden. Daher wurde verstärkt das niedersorbische Forschungsvorhaben vorangetrieben. Hier konnte auf bereits vorhandenes, phonetisch annotiertes Audiomaterial aus einem vorangegangenen Projekt des SI zur „Dokumentation bedrohter Sprachen“[2] und aus externen Quellen wie z. B. das Korpus „Gesprochenes Niedersorbisch“[3] oder Rundfunkaufnahmen zurückgegriffen werden.

Der im Vorjahr entwickelte Prototyp der obersorbischen Stimme wurde getestet und konnte als Modell für das entsprechende niedersorbische Programm herangezogen werden.

1.4 Hauptphase II – 2021

Im zweiten Jahr der Hauptphase konnten mehrere Studioaufnahmen mit dem für die niedersorbische Stimme ausgewählten Sprecher realisiert werden. Das dabei gewonnene Aufnahmematerial von etwa 50 Minuten wurde aufbereitet (siehe 3.3) und ist die Grundlage einer funktionierenden niedersorbischen Synthesestimme, die am Institutstag 2021 des SI im November 2021 erstmals der Öffentlichkeit vorgestellt wurde. Weitere Aufnahmen finden kontinuierlich statt, ebenso die Anreicherung der anderen Komponenten (siehe 3.), um die niedersorbische Synthesestimme zu verbessern.

Die Planungen zur Feldforschung zum Obersorbischen wurden angesichts der Verzögerung im Vorjahr angepasst. Statt der ursprünglich geplanten drei Phasen wird nun in zwei Phasen das freie Gespräch mit den Informant:innen mit dem Einsatz bereits fertig entwickelter Fragebögen zur gezielten Ansteuerung sprachlich relevanter Phänomene kombiniert. So konnten im Herbst 2021 erste Erhebungen stattfinden.

1.5 Ausblick Hauptphasen III und IV – 2022/2023

In den beiden kommenden Jahren soll die Qualität beider Stimmen verbessert und schließlich auf das angestrebte Niveau in erster Linie gut verständlicher Stimmen angehoben werden. Dazu sind für das Obersorbische im kommenden Jahr 2022 die Fortführung der Feldforschung und erste Aufnahmen im Tonstudio für das erforderliche Audio-

---

[2] „Maminorěcna dolnoserbšćina“ – ein mittels Webinterface durchsuchbares Korpus unter https://dolnoserbski.de/dobes/?rěc=de [28.01.2022], wo auch ausführliche Informationen zum Projekt zu finden sind.

[3] Internet: https://genie.coli.uni-saarland.de/cgi-bin/korpus.html [26.01.2022]. Weitere Informationen zum Projekt: Roland Marti/Bistra Andreeva/William Barry: Korpora bedrohter Sprachen als eierlegende Wollmilchsau? Das Beispiel GENIE, in: Linguistik Online 39/3 (2009), https://doi.org/10.13092/lo.39.482 [26.01.2022].

korpus unerlässlich. Auch für Niedersorbisch stehen noch einige Sitzungen im Tonstudio aus. Parallel werden auch die lexikalischen und linguistischen Ressourcen (siehe 3.1 und 3.2) beider Sprachen aufgrund neuer Erkenntnisse stetig überarbeitet und ergänzt. Außerdem sollen die Erkenntnisse aus der obersorbischen Feldforschung und dem niedersorbischen Forschungsvorhaben zukünftig in die Entwicklung einer Orthoepie für die jeweilige Sprache einfließen.

2. Sprachsynthese – TTS

Eine Vorlesefunktion ist eine konkrete Anwendung eines Teilbereichs der maschinellen Sprachverarbeitung: der Sprachsynthese. Mittels Sprachsynthese – häufig auch als *text-to-speech* (*TTS*) bezeichnet – wird ein geschriebener Text in gesprochene Sprache umgewandelt. So trivial das zunächst klingen mag, so komplex ist es im Detail, bedenkt man, dass hierbei maschinell auch kognitive Aspekte der menschlichen Sprachproduktion (z. B. in Bezug auf die korrekte Aussprache von Symbolen/Daten/Abkürzungen) nachgeahmt werden.

Das Lesen eines Textes in einer bekannten Sprache erfordert neben der korrekten Erkennung der geschriebenen Zeichen als Buchstaben und deren Verknüpfung zu Wörtern auch Kenntnis von Symbolen, Daten, Zahlen, Abkürzungen, fremdsprachlichen Ausdrücken usw. und natürlich deren regulärer Aussprache im grammatikalischen Kontext. Zur richtigen Aussprache beim Vorlesen gehören auch korrekte Betonung, Satzmelodie und Tempo. Dem Menschen gelingt der richtige Wortakzent bei Homographen in der Regel anhand seines Kontext- und Weltwissens, während die Intonation durch Interpretation von Interpunktionszeichen gesteuert wird. Ein Beispielsatz verdeutlicht die Komplexität:

> *07.02.2020 pochowa so w Kaliforniskej 41lětny Kobe Bryant, sławny basketballist, kotryž docpě 2006 w dobyćerskej hrě 81 dypkow, potajkim ca. 66 % dypkow swojeho mustwa Los Angeles Lakers.*[4]

Soll dieser Satz innerhalb eines Sprachsynthesesystems verarbeitet werden, müssen zunächst die Spezifika der Sprache – hier also obersorbisch – hinterlegt werden. Die Grundlage besteht in der Zuordnung des Allophoninventars der Sprache zu seiner (ortho)grafischen Repräsentation. Dabei werden sowohl einzelnen Buchstaben als auch Buchstabenkombinationen Aussprachеregeln zugewiesen. Hinzu kommen Informationen, wie Akronyme, Zahlen und Symbole in welchem Kontext aufgelöst werden müssen, damit sie korrekt vorgelesen werden.

Der Vorgang der Sprachsynthese lässt sich grob in drei Komplexe unterteilen, die in Abbildung 2 dargestellt sind. Bevor ein Text vorgelesen werden kann, muss er analysiert werden. Ein Tokenisierer zerteilt den Text anhand der Leerzeichen in seine Bestandteile. Als Interpunktionszeichen deutet ein Punkt auf ein Satzende hin. Doch der Beispielsatz zeigt, dass der Punkt darüber hinaus als Teil von Ordnungszahlen in Datumsangaben und als Kennzeichnung von Abkürzungen fungieren kann. Solche besonderen Formen müssen normalisiert, d. h. in Wortformen konvertiert werden. Ist der Eingabetext voll

---

[4] Deutsch: Am 07.02.2020 wurde in Kalifornien der 41jährige Kobe Bryant beigesetzt, der berühmte Basketballspieler, der 2006 in einem siegreichen Spiel 81 Punkte erzielte, also ca. 66 % der Punkte seiner Mannschaft Los Angeles Lakers.

ausgeschrieben, kann das System die zugewiesenen Ausspracheregeln auf diese regulären Wortformen anwenden.

> *Sedmeho druheho dwaj tysac a dwaceći pochowa so w Kaliforniskej jedynaštyrcećilětny Kobe Bryant, sławny basketballist, kotryž docpě dwaj tysac šěsć w dobyćerskej hrě jedynawosomdźesat dypkow, potajkim cirka šěsćašěsćdźesat procentow dypkow swojeho mustwa Los Angeles Lakers.*

Um die richtige Aussprache zu gewährleisten, gilt es Wörter wie *pochowa* richtig zu segmentieren, hier also in *po|chowa* und nicht etwa *pochow|a*. Einem Sprecher der jeweiligen Sprache wird das in der Regel intuitiv gelingen, und er wird die Buchstabenkombination <chow> richtigerweise [k^{h}o] aussprechen. Anders als beispielsweise in *pochowy* ‚Torf-', wo die gleiche Folge korrekt [xo] lautet. Eine weitere Schwierigkeit enthält der Beispielsatz mit den englischen Eigennamen *Kobe Bryant* und *Los Angeles Lakers*. Hier sollte die Aussprache gerade nicht anhand der obersorbischen Ausspracheregeln umgesetzt werden, sondern englisch erfolgen.

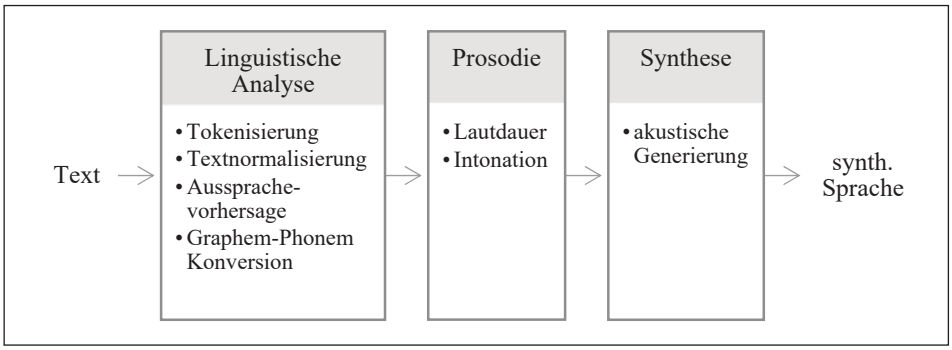

Abbildung 2: Schematische Darstellung der Sprachsynthese

Im zweiten Modul wird dem zu synthetisierenden Text eine lautliche Struktur zugewiesen. Bei der Prosodiemodellierung geht es im Wesentlichen um die zeitliche Struktur – Syllabifizierung und Dauer der Einzellaute – und um die Phonetik der Intonation – Rhythmus von Heben und Senken der „Stimme" über mehrere Äußerungseinheiten hinweg und Pausensteuerung. Aus dieser phonetischen Repräsentation wird im letzten Schritt der Synthese das akustische Sprachsignal erstellt.

3. MaryTTS: Unit-Selection-Synthese in Modulen

MaryTTS[5] lautet der Name der während der Konzeptionsphase ausgewählten Sprachsynthesesoftware. Die Abkürzung steht für M*odular* A*rchitecture for* R*esearch on speech s*Ynthesis T*ext-to-Speech System* – es handelt sich um ein Open-Source-Programm, welches gemeinschaftlich vom Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) und der Universität des Saarlandes entwickelt wurde. Neben der Quelloffenheit, die es ermöglicht, den Programmcode in allen Details unseren Zwecken

[5] Internet: http://mary.dfki.de/ [abgerufen 04.01.2022].

anzupassen, besteht ein weiterer Vorteil darin, dass das „Erlernen“ völlig neuer Sprachen bereits als grundlegender Baustein implementiert ist. Seit Beginn des Projektes gewährleisten zudem persönliche Kontakte mit den Entwicklern gute Beratung und Unterstützung bei auftretenden Problemen.

MaryTTS unterstützt mit der Unit-Selection-Synthese ein konkatenatives Syntheseverfahren. Dabei entsteht das künstliche Sprachsignal durch die Verkettung natürlichsprachlichen Audiomaterials. Die Datenbasis besteht aus aufgezeichneten kurzen Texten, Sätzen und Phrasen, die anschließend segmentiert und orthografisch (ggf. zusätzlich phonetisch) annotiert werden. Aus diesem Inventar unterschiedlich großer Einheiten von Diphonen über Halbsilben, Wörtern, Wortgruppen bis hin zu Phrasen werden bei der Unit-Selection-Methode die größtmöglichen passenden Segmente ausgewählt und zu der Audioausgabe verknüpft, die dem Eingabetext entspricht. Je mehr Sprachaufzeichnungen vorhanden sind, desto größere Segmente lassen sich verwenden und desto weniger Verkettungsstellen sind zu erwarten. Und wenige dieser Konkatenationsstellen bedeuten gleichzeitig wenige potenziell diskontinuierliche Signalübergänge. Im Gegensatz zu einer vollständig computerbasierten Synthese müssen prosodische Elemente nicht vollständig modelliert werden, man erzielt mit dieser Methode bei entsprechend sorgfältig erstellter Datengrundlage also eine deutlich natürlicher klingende synthetische Sprache.

MaryTTS ist modular aufgebaut (siehe Abbildung 3). Die Module können unabhängig voneinander bearbeitet und kontinuierlich angereichert werden. Dennoch bauen sie aufeinander auf, sodass man zu Beginn bei der Erarbeitung einer neuen Sprache bzw. Stimme der Reihe nach vorgehen sollte. Im Folgenden werden die einzelnen Komponenten näher erläutert.

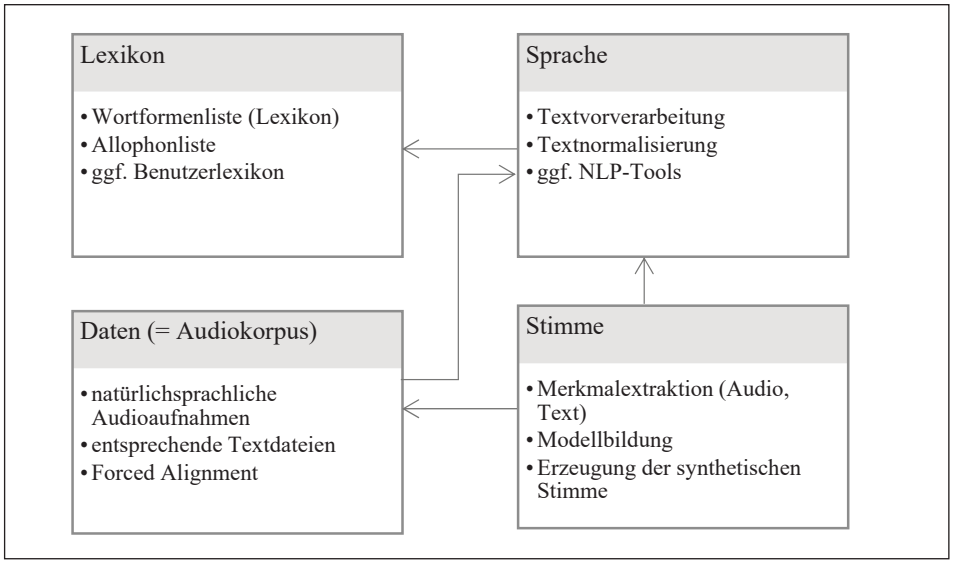

Abbildung 3: Module und Abhängigkeiten MaryTTS: Das Sprachmodul greift auf die Daten der Lexikonkomponente zurück. Innerhalb des Datenmoduls sind für das Forced Alignment die Ausgaben der Sprachkomponente erforderlich. Das Modul Stimme, als eigentliches Synthesemodul, erfordert gleichermaßen die Ausgaben der Sprach- und Datenkomponente. (Die Pfeile symbolisieren demnach Zugriff auf bzw. Nutzung von Daten eines anderen Moduls.)

3.1 Modul „Lexikon“

Das Lexikonmodul beinhaltet sprachspezifische Datengrundlagen der zu synthetisierenden Sprache. In einer Allophonliste werden sämtliche relevanten Laute der Sprache inklusive ihrer phonologischen Standardmerkmale erfasst. Die Notation erfolgt auf zweierlei Weise: einmal gemäß dem internationalen phonetischen Standard, d. h. nach den Transkriptionsregeln des Internationalen Phonetischen Alphabets (IPA) mittels Unicode und einmal nach einer für eigene Zwecke angepassten SAMPA-Transkription. SAMPA steht für *Speech Assessment Methods Phonetic Alphabet* und kann als maschinenlesbare Version des IPA verstanden werden. Es sind nicht alle Aussprachedetails wie in IPA darstellbar, aber die SAMPA-Notation ist benutzerfreundlicher, da sie auf die Verwendung einfacher Tastaturzeichen (ASCII) ausgelegt ist. Durch die Eins-zu-eins-Zuordnung in der Allophonliste sind Dopplungen ausgeschlossen.

Die lexikalische Ressource besteht in einer listenartigen Gegenüberstellung von Wörtern der Sprache in ihrer orthografischen Repräsentation und deren Aussprache in der angepassten phonetischen SAMPA-Transkription. Innerhalb des Lexikon-Moduls werden diese Daten kompiliert und formal in einer Datenkomponente abgelegt, die im nächsten Modul herangezogen wird. Aus der Allophonliste ergeben sich so Graphem-zu-Phonem-Regeln (G2P), aus der Wortliste werden die beiden Zeichensätze der jeweiligen Repräsentation herausgezogen und formal zueinander in Beziehung gesetzt. Daraus wird in MaryTTS die Aussprache des zu verarbeitenden Texts abgeleitet. Sind die Tokens als solche im Lexikon (der Wortformenliste) enthalten, werden die Phoneme auf dieser Grundlage zurückgegeben; ist kein passender Eintrag vorhanden, erfolgt die Vorhersage aufgrund der G2P-Regeln.

Je größer das vorgegebene Lexikon ist, desto besser und genauer kann das Programm die Aussprache wiedergeben. Zu bedenken sind allerdings zwei Probleme. Die Transkription erfordert einen nicht unerheblichen Zeitaufwand. Im Projektablauf hat sich, wie erwartet, auch gezeigt, dass Allophonliste und Transkription an neuere Erkenntnisse angepasst werden müssen und dann entsprechend auch das Lexikon überarbeitet werden muss. Daher wurde als ungefähre Zielsetzung in beiden Sprachen eine Lexikongröße von etwa 40 000 Einträgen gewählt. Inzwischen wurden technische Mittel entwickelt, um die Transkription zeitsparend halbautomatisch durchzuführen. Die Wörter für beide Lexika wurden aus automatisch erzeugten Wortformenlisten ausgewählt. Dafür konnten morphologische Generatoren genutzt werden, die mit Beteiligung des SI bzw. am Institut entwickelt wurden und in der Lage sind, zu jeder erfassten Wortgrundform alle korrekten grammatischen Formen der jeweiligen Sprache zu bilden.[6] Die Auswahl wurde mithilfe eines automatisch ausgeführten Algorithmus so gesteuert, dass möglichst jede in der Sprache vorhandene Lautkombination abgebildet werden kann.

Ein erhebliches Problem für das Lexikon-Modul stellen Fremdwörter dar. Sowohl Obersorbisch als auch Niedersorbisch haben in ihrer Entwicklungsgeschichte, meist über die Vermittlung des Deutschen als Kontaktsprache, eine große Zahl von Lehnwör-

---

[6] Die Programme zur automatischen Rechtschreibkontrolle für Nieder- und Obersorbisch basieren auf den genannten Generatoren, siehe https://dolnoserbski.de/ortografija/kontrola?rěc=de (niedersorbisch) bzw. https://soblex.de/download/download.html (obersorbisch) [Stand beide: 21.01.2022].

tern aufgenommen und nehmen auch heute noch stetig neue Fremdwörter in den Sprachgebrauch auf, so aus dem Französischen und aktuell vor allem aus dem Englischen, also aus Sprachen mit deutlich anderen Graphem-Phonem-Zuordnungen. Dies führt zu einem potenziellen Dilemma: Um die korrekte Aussprache dieser Fremd- und Lehnwörter zu gewährleisten, benötigt MaryTTS entsprechende Einträge in der Allophonliste und im Lexikon. Bei zunehmender Menge solcher Einträge kann es jedoch passieren, dass die im Programm hinterlegten fremdsprachlichen Ausspracheregeln auf indigene Wörter übertragen werden. Wenn ein Wort fremder Herkunft in der schriftlichen Form zufällig einem indigenen Wort entspricht, kann dies nicht ohne Weiteres (z. B. durch eine umfangreiche Kontextanalyse) unterschieden werden. Ein Beispiel dafür sind die Wörter *dom* [dɔm] (Haus) und *dom* [doːm] (Dom) – mit geschlossenem [o] wie im Deutschen. In solchen Konfliktfällen wird grundsätzlich die indigene Aussprachevariante bevorzugt.

Der grundsätzliche Lösungsansatz für diese Probleme sind Ausnahmelisten, die MaryTTS innerhalb eines Benutzerlexikons ermöglicht. Dort können von der Regel (der eigentlichen Synthesesprache) abweichende Wortformen zusammengefasst und bereitgestellt werden. So haben sie keinen Einfluss auf die allgemeine durch das Programm erzeugte Aussprache. Weiterhin sind zum aktuellen Stand auch eine Reihe spezifischer Laute aus dem Deutschen (z. B. Langvokale) und einzelne Laute aus dem Englischen und Französischen (z. B. Nasalvokale) in der Allophonliste mitdefiniert. Es ist aber davon auszugehen, dass letztlich nicht jedes Fremdwort und jeder fremdsprachliche Name völlig korrekt in der fremden Form vorgelesen werden kann. Dies ist jedoch für die Verständlichkeit auch nicht zwingend notwendig.

3.2 Modul „Sprache“

Das zweite Modul wird innerhalb der MaryTTS-Architektur als *language* – Sprache bezeichnet, weil das Ergebnis dieses Moduls die Sprachkomponente ist, die der Verarbeitung des Eingabetextes dient. In Listenform werden Regeln zur Analyse und Aufschlüsselung von Zahlen, Symbolen, Akronymen, Datumsangaben u. ä. bereitgestellt. Für die richtige Aufschlüsselung ist zu beachten, dass beide sorbischen Sprachen stark flektierend sind. Es muss also mitunter auch eine Kontextanalyse stattfinden, damit grammatisch korrekt nach Kasus, Numerus und Genus aufgelöst und vorgelesen werden kann. Es ist allerdings möglich, dass manche Problemstellungen mit angemessenem Aufwand in Bezug auf investierte Arbeitszeit und erforderliche Rechenleistung nicht gänzlich gelöst werden können. In solchen Fällen muss auf eine vorab ausgewählte Standardvariante zurückgegriffen werden. Mit den Elementen der Textnormalisierung können weitere sprachspezifische Werkzeuge zur Verarbeitung natürlicher Sprache oder wahlweise auch generische Tools, wie der Tokenisierer, kombiniert werden, um die Qualität der Sprachkomponente auszuweiten.

Außerdem sollte es zwischen der Texteingabe durch den Endnutzer und der Audioausgabe durch das Programm keine zu lange Verzögerung geben, komplexere Rechenvorgänge können aber eben das hervorrufen.

3.3 Modul „Daten“

Das Datenmodul beinhaltet das Audiokorpus. Es enthält umfangreiche natürlichsprachliche Aufnahmen eines Sprechers in der Zielsprache. Ein ausgewogenes Korpus sollte sprachlich möglichst alle Lautkombinationen mehrfach mit unterschiedlicher Intonation abbilden, um damit sämtliche koartikulatorischen Phänomene zu erfassen. Es sollte inhaltlich breit angelegt sein, das heißt verschiedene Textsorten berücksichtigen, sodass die Synthese universell einsetzbar ist. Da die Aufnahmequalität in direktem Zusammenhang mit der Ausgabequalität der Synthese steht, empfiehlt es sich, hochwertige Aufnahmen z. B. im Tonstudio zu machen. Die Audiodateien werden als WAV-Datei (.wav) und somit verlustfrei im Datenmodul gespeichert. Zu jeder Audiodatei wird eine namensgleiche Textdatei (.txt) abgelegt, die den genauen Wortlaut der jeweiligen Aufzeichnung orthografisch transkribiert enthält.

Die finale Stimme der Synthese ähnelt aufgrund der natürlichsprachlichen Basis in Stimmfarbe und Stimmqualität[7] der originalen Sprechstimme. Bei der Auswahl eines Sprechers/einer Sprecherin ist zunächst darauf zu achten, dass es sich um Muttersprachler[8] mit einer gesunden und belastbaren Stimme handelt. Üblicherweise würde man für solche Zwecke auf professionelle Sprecher:innen aus dem Schauspiel- oder Medienbereich zurückgreifen. Bei diesem Projekt wäre das allerdings kontraproduktiv. Die Ausbildung in diesen Berufen erfolgt in der Regel am Deutschen, sodass Einflüsse auf das muttersprachliche Nieder- bzw. Obersorbisch anzunehmen sind. Der Muttersprachler oder die Muttersprachlerin der Wahl sollte nicht nur eine exemplarische Aussprache vorweisen, sondern auch in der Lage sein, möglichst homogen über einen längeren Zeitraum an verschiedenen Terminen zu sprechen. Je homogener das Audiokorpus, desto besser die Synthesequalität, da MaryTTS nicht nur ganze im Korpus vorhandene Wörter ausgeben, sondern auch aus deren Einzelteilen neue Wörter synthetisieren kann. Je homogener diese Einzelteile sind, desto besser und weicher klingen die Übergänge, was die Verständlichkeit erleichtert und letztlich zu einem angenehmeren Hörerlebnis führt. Doch auch hier ist zu beachten, dass dem Umfang der Aufnahmen durch den erforderlichen Arbeitsaufwand Grenzen gesetzt sind. Damit die Aufnahmen von MaryTTS zur Synthese verwendet werden können, müssen sie phonetisch annotiert werden. Der Inhalt jeder Aufnahme ist durch die zugehörige txt-Datei bekannt, und in einem weiteren Aufbereitungsschritt wird anhand des Textes die Abfolge der einzelnen phonetischen Einheiten mit dem Zeitbereich des Audiosignals in Übereinstimmung gebracht (*forced alignment*). Es entsteht ein programminternes Benutzerwörterbuch, worin anhand der Daten akustische Modelle zum Zweck der Aussprachevorhersage trainiert werden. Die Genauigkeit der Zuordnung und damit der Aussprachevorhersage erhöht sich mit zunehmender Datenmenge. Der ganze Prozess ist Teil des MaryTTS-Workflows und läuft im Grunde automatisch ab, ist jedoch zeitintensiv. Im Anschluss ist in der Regel, insbesondere nach einer Erweiterung der Datengrundlage, eine intensive manuelle Überprüfung

---

[7] Hier bezogen auf die individuellen Merkmale der menschlichen Stimme und nicht auf die Qualität der Synthese.

[8] Im Niedersorbischen ist der Begriff Muttersprachler i. w. S. zu verstehen. Für dieses Projekt wurde ein Sprecher mit sehr guter niedersorbischer Sprachkenntnis und Aussprache ausgewählt. Die typischen/ursprünglichen Muttersprachler des Niedersorbischen sind aufgrund ihres hohen Alters und der daraus resultierenden Probleme nicht für Sprecheraufnahmen geeignet.

und Nachbearbeitung erforderlich. Der eigentliche Stimmerzeugungsprozess im Synthesemodul hängt von diesem sorgfältig orthografisch und phonetisch annotierten Korpus ab.

Ähnlich wie beim Lexikon-Modul stellen auch hier Fremdwörter eine besondere Herausforderung dar. Zunächst ist es wichtig, dass sie möglichst homogen, d. h. möglichst immer mit derselben Aussprache, aufgenommen werden. Gerade bei teilintegrierten Lehnwörtern oder bei Fremdwörtern existieren oft parallel mehrere schwankende Aussprachevarianten. Eine zu große Varianz führt zu Problemen beim Sprachlernprozess von MaryTTS und bei der Synthese, im schlimmsten Fall sogar zu Fehlinterpretationen bei indigenen Wörtern. Hierbei wird allerdings grundsätzlich wieder die indigene Aussprache bevorzugt, auch wenn das zu nicht dem Usus entsprechender phonetisch adaptierter Aussprache von Fremdwörtern führen könnte (z. B. [ʃɛma] statt [ʃeːma] für orthografisches *šema* ‚Schema').

3.4 Modul „Stimme"

Im Stimm-Modul laufen alle Fäden zusammen. Hier befindet sich die verarbeitende Software, die aus den Informationen des Sprach- und Daten-Moduls die synthetisierte Sprachausgabe erstellt. Zur Laufzeit muss das TTS-System in der Lage sein, Prosodie und Stimmparameter anhand von Text verzögerungsfrei auszugeben. Durch maschinelles Lernen trainiert das System an Vorhersagemodellen die Zuweisung dieser Parameter. Die Modelle entstehen automatisch im Stimm-Modul und basieren auf der Merkmalsextraktion von Informationen der akustischen Analyse des Audiokorpus und auf den Informationen der linguistischen Analyse des Sprachmoduls. Da das Datenmodul durch Annotation aufbereitet wurde, können auch die zeitlich ausgerichteten Transkriptionen in eine Merkmalsdarstellung überführt und so mit den linguistischen und akustischen Merkmalen zu Vektoren kombiniert werden. Bei der hier verwendeten Unit-Selection-Synthesemethode werden die Merkmalsrepräsentation und die damit verknüpften Metadaten für jede Einheit gespeichert. Dasselbe gilt für die statistischen Modelle zur Prosodievorhersage, obgleich die Anwendung durch die natürlichsprachliche Basis auf die Modifizierung diskontinuierlicher Schnittstellen begrenzt ist. Mit dem erforderlichen Zugriff auf die tatsächlichen Audiodaten ist die Unit-Selection-Synthese ein speicherintensives Verfahren.

Zusammenfassung

Die Entwicklung einer Vorlesefunktion für die nieder- und obersorbische Schriftsprache ist ein auf fünf Jahre angelegtes Forschungsprojekt. Das angestrebte Ziel, einen anwenderfreundlichen Dienst zu schaffen, der es seinen Nutzer:innen ermöglicht, sich sorbischsprachige Texte auf der Webseite des Sorbischen Instituts sowie auf anderen Webseiten verständlich und – soweit technisch umsetzbar – mit weitgehend „guter" Aussprache vorlesen zu lassen, erfordert grundständige Forschungsarbeit zur Phonetik und Phonologie des Sorbischen. Das Projekt erlaubt die Verknüpfung des Serviceaspekts mit dem Schließen einer großen Forschungslücke in diesem Bereich. Eine fundierte Beschreibung der Orthoepie beider sorbischer Sprachen kann zum Erhalt des Nieder- und Obersorbischen beitragen. Die Ergebnisse der Forschungen und technischen Entwicklungen, die im Rahmen des vorgestellten Projekts entstehen, sollen in einer

Reihe weiterer Beiträge publiziert werden. Den unmittelbaren Auftakt macht ein Aufsatz zum Lautsystem des Niedersorbischen (s. Jocz/Měškank im vorliegenden Heft).

Literatur

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Ralf; Langer, Hagen (Hg.) 2001: Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg-Berlin.

Steiner, Ingmar; Le Maguer, Sébastien 2018: Creating New Language and Voice Components for the Updated MaryTTS. Text-to-Speech Synthesis Platform, in: Proceedings of the 11th Language Resources and Evaluation Conference (LREC). Miyazaki, S. 3171–3175.