TECHNISCHE UNIVERSITÄT DRESDEN
FAKULTÄT ELEKTROTECHNIK UND INFORMATIONSTECHNIK
Institut für Akustik und Sprachkommunikation
Professur für Sprachtechnologie und Kognitive Systeme

# DIPLOMARBEIT

zum Thema

*„Aufbau und Evaluation einer Datenbasis von annotierten Sprachdaten des Deutschen“*

von

*Hannes Kath*
geboren am 01.06.1994 in Hamburg

zur Erlangung des akademischen Grades

DIPLOMINGENIEUR

(Dipl.-Ing.)

Tag der Einreichung: 11. Mai 2021
Betreuer: Dipl.-Ing. Simon Stone
Erstgutachter: Prof. Dr.-Ing. Peter Birkholz
Zweitgutachter: Prof. Dr.-Ing. habil. Ercan Altinsoy

# Selbständigkeitserklärung

Hiermit erkläre ich, Hannes Kath, dass die heute beim Prüfungsausschuss der Fakultät Elektrotechnik und Informationstechnik eingereichte Diplomarbeit zum Thema

„Aufbau und Evaluation einer Datenbasis von annotierten Sprachdaten des Deutschen“

vollkommen selbständig von mir verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet und Zitate kenntlich gemacht wurden.

Dresden, den 11. Mai 2021

Hannes Kath

# Kurzfassung

Gegenstand der vorliegenden Ausarbeitung ist die Erstellung und Dokumentation einer Datenbasis gesprochener deutscher Sprache inklusive Annotationen auf unterschiedlichen Ebenen.

Im ersten Kapitel wird der Begriff Sprachkorpus eingeführt. Hierbei wird genauer auf die unterschiedlichen Ebenen der Annotation eingegangen. Im zweiten Kapitel werden gesprochene Korpora aufgeführt und im Hinblick auf die Annotationen und den Umfang diskutiert. In den Kapiteln drei bis sieben wird der automatisch erstellte *Corpus of Aligned Read speech Including Annotations* (CARInA) vorgestellt, welcher im Rahmen dieser Ausarbeitung entstanden ist. Das dritte Kapitel beschreibt die Beschaffung des Sprachmaterials und die orthographische und phonetische Transkription. Das vierte Kapitel beschreibt die kanonische und morphosyntaktische Transkription mithilfe erstellter Wörterbücher. Im fünften Kapitel wird die prosodische Annotation des CARInA beschrieben und im sechsten Kapitel die Struktur des Korpus. Das siebte Kapitel beinhaltet die Auswertung und Validierung des CARInA. Die Validierung wurde mittels Referenzen und durch das Training einer Spracherkennung durchgeführt. Das achte Kapitel beinhaltet die Zusammenfassung und einen Ausblick der vorliegenden Arbeit.

Für die vorliegende Arbeit wurde die Programmierumgebung MATLAB R2019b genutzt.

Sprachliche Eigenschaften wurden mit dem Programm Praat version 6.1.39 berechnet.

# Abstract

The subject of this thesis is the creation and documentation of a database of spoken German, including annotations at different levels.

In the first chapter, the term speech corpus is introduced with regard to its creation and application. Here, the different levels of annotation are discussed in more detail. In the second chapter, spoken corpora are listed and discussed in terms of annotations and scope. Chapters three to seven are about the automatically created *Corpus of Aligned Read speech Including Annotations* (CARInA), which was developed for this paper. The third chapter describes the acquisition of the speech material and the orthographic and phonetic transcription. The fourth chapter describes the canonical and morphosyntactic transcription with the help of dictionaries. The fifth chapter describes the prosodic annotation of CARInA and the sixth chapter the structure of the corpus. The seventh chapter contains the evaluation and validation of the CARInA. The validation was carried out by means of references and by training a speech recogniser. The eighth chapter contains the summary and an outlook of the present work.

The programming environment MATLAB R2019b was used for the present work.

Linguistic properties were calculated with the program Praat version 6.1.39.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| ARPABET | engl. *Advanced Research Projects Agenca alphabet* |
| ASCII | engl. *American Standard Code for Information Interchange* |
| AuToBI | engl. *Automatic Tones and Break Indices annotation* |
| BAS | Bayrisches Archiv für Sprachsignale |
| BITS-US | engl. *BITS Unit Selection synthese corpus* |
| BURSC | engl. *Boston University Radio Speech Corpus* |
| CARInA | *Corpus of Aligned Read speech Including Annotations* |
| CAU | Christian-Albrechts-Universität |
| CNN | engl. *Convolutional Neural Network* |
| CORPRES | engl. *Corpus Of Russian Professionally Read Speech* |
| D | Determinante |
| DGD | Datenbank für Gesprochenes Deutsch |
| DNN | engl. *Deep Neural Network* |
| FOLK | Forschungs- und Lehrkorpus gesprochenes Deutsch |
| GSWC | engl. *German Spoken Wikipedia Corpus* |
| GToBI | engl. *German Tones and Break Indices* |
| GToBI light | engl. *German Tones and Break Indices ’light’* |
| G2P | Graphem-zu-Phonem |
| HKCSE | engl. *Hong Kong Corpus of Spoken English* |
| HMM | engl. *Hidden Markov Model* |
| HTML | engl. *Hypertext Markup Language* |
| hzsk | Hamburger Zentrum für Sprachkorpora |
| IDS | Institut für Deutsche Sprache |
| IPA | Internationales Phonetisches Alphabet |
| KCSG | engl. *Kiel Corpus of Spoken German* |
| KCSGrs | engl. *Kiel Corpus of Spoken German read speech* |
| KCSGss | engl. *Kiel Corpus of Spoken German spontaneous speech* |
| KIM | Kieler Intonationsmodell |

| | |
|---|---|
| LDC | engl. *Linguistic Data Consortium* |
| MAE_ToBI | engl. *Mainstream American English Tones and Break Indices* |
| MAUS | engl. *Munich Automatic Segmentation* |
| N | Nomen |
| NP | Nominalphrase |
| PD2 | PhonDat 2 |
| POS | engl. *Part-Of-Speech* |
| PRR | engl. *Prosody Recognition Revisited* |
| PRS | engl. *Prosody Recognition System* |
| PyToBI | engl. *Python Tones and Break Indices* |
| RCPCE | engl. *Research Centre for Professional Communication in English* |
| ReLU | *Rectified Linear Unit* |
| RSA | engl. *Relative Symmetric Accuracy* |
| S | Satz |
| SAMPA | engl. *Speech Assessment Methods Phonetic Alphabet* |
| SI100 | Siemens 100 |
| SI1000P | engl. *Siemens Synthese Korpus* |
| SNR | engl. *Signal-to-Noise Ratio* |
| SPSU | engl. *Saint Petersburg State University* |
| SRNC | engl. *Stuttgart Radio News Corpus* |
| STTS | Stuttgard-Tübingen-Tagset |
| SWC | engl. *Spoken Wikipedia Corpus* |
| TIMIT | engl. *Texas Instruments/Massachusetts Institute of Technology* |
| ToBI | engl. *Tones and Break Indices* |
| TTS | engl. *Text-To-Speech* |
| TUD | Technische Universität Dresden |
| UNIX | engl. *Uniplexed Information and Computing Service* |
| UTF-8 | engl. *Universal Coded Character Set Transformation Format* |
| V | Verb |
| VP | Verbalphrase |
| XML | engl. *Extensible Markup Language* |

# Danksagung

An dieser Stelle möchte ich mich bei denjenigen Personen bedanken, die mich während der Anfertigung dieser Diplomarbeit unterstützt und motiviert haben.

An erster Stelle bedanke ich mich bei Dipl.-Ing. Simon Stone, der meine Arbeit betreut hat. Die regelmäßigen Treffen und Gedankenanstöße haben wesentlich die Entwicklung meiner Arbeit beeinflusst. Für die zielführende und freundliche Unterstützung in jedem Teilbereich meiner Diplomarbeit möchte ich meinen Dank aussprechen. Auch bin ich dankbar, dass ich trotz der außergewöhnlichen gesellschaftlichen Situation in den Räumlichkeiten der Universität arbeiten konnte.

Ich bedanke mich bei Prof. Dr. Stephan Rapp für die vielen langen Nachmittage, welche er für mich aufbrachte, um das Programm seiner Dissertation zur automatischen Vorhersage prosodischer Etiketten zu installieren und zu überarbeiten. Auch für die entgegengebrachte Geduld und die ausführlichen Erklärungen bin ich äußerst dankbar.

Ich bedanke mich bei allen Personen, die mir fachlich geholfen haben und für Fragen offen waren. Ich danke Dr. Benno Peters für seine Hilfe bei der Konvertierung der prosodischen Etiketten. Ich danke Dr. Mónica Domínguez-Bajo für den regen Austausch über das Programm PyToBI. Ich danke Dr. Timo Baumann dafür, dass er mir die Validierungsdaten des *German Spoken Wikipedia Corpus* zur Verfügung gestellt hat.

Ich danke Dipl.-Ing. Carina Schmidt für die großartige Unterstützung und die kleinen Überraschungen in jeder einzelnen Woche meiner Diplomarbeit. Ich bin dankbar für die vielen Gespräche, die gemeinsame Zeit und nicht zuletzt für die Kritik und das Interesse an meiner Diplomarbeit.

Abschließend bedanke ich mich bei meinen Eltern. Ich bedanke mich für das Vertrauen, das sie in mich haben und für das Gefühl, mit keinem Problem alleine bleiben zu müssen.

# Einführung

Die Verarbeitung gesprochener Sprache ist in vielen modernen Applikationen enthalten. Programme zur Sprachsignalverarbeitung basieren zu einem Großteil auf tiefen neuronalen Netzen [53], [97]. Für das Training dieser Programme sind umfangreiche Sprachkorpora von teilweise mehreren Hundert Stunden Sprachmaterial notwendig [97].

Die manuelle Erstellung dieses Sprachmaterials und der zugehörigen Annotationen auf unterschiedlichen sprachlichen Ebenen ist eine komplexe und zeitaufwändige Aufgabe [73]. Umfangreich annotierte Korpora sind zu einem überwiegenden Anteil englischsprachig. Deutschsprachige Sprachressourcen stammen aus einer vergleichsweise geringen Anzahl an Datenquellen und sind weniger umfangreich.

Für die Erstellung deutschsprachiger Korpora werden zu einem großen Teil halbautomatische oder automatische Verfahren eingesetzt. Die Qualität automatisch erstellter Annotationen ist für keine Annotationsebene vergleichbar mit der Qualität manuell erstellter Annotationen (siehe Abschnitt 1.2). Dennoch erzielen Programme, welche mit automatisch erstellten Sprachkorpora trainiert wurden, hohe Genauigkeiten.

Thema dieser Arbeit ist die Erstellung und Dokumentation eines umfangreichen deutschsprachigen Korpus mit Annotationen auf orthographischer, phonetischer und prosodischer Ebene.

## *Kapitel 1*

# 1 Sprachkorpora

Mit der Möglichkeit der maschinellen Datenverarbeitung entstand die Korpuslinguistik [7], [73]. Diese sprachwissenschaftliche Teildisziplin beinhaltet die Erstellung, Aufbereitung und Verwendung maschinell lesbarer Sprachkorpora (im Folgenden auch Korpora genannt) [7].

Als Korpus wird in der Sprachwissenschaft eine Sammlung geschriebener und/oder gesprochener Sprache bezeichnet [7], [58]. Diese Definition beinhaltet einen Stapel Zeitschriften oder Kassetten genauso wie digital vorliegende Korpora [73]. Durch die größtenteils maschinelle Verarbeitung von Korpora impliziert der Begriff „Korpus“ seit einiger Zeit die Möglichkeit der digitalen Datenverarbeitung [73]. In der vorliegenden Arbeit wird der Begriff Korpus ausschließlich für digitale Datensätze verwendet.

## 1.1 Korpus-Typologien

Korpora werden nach unterschiedlichen Typologien charakterisiert [7], [73]. Eine Übersicht dieser Typologien ist im Folgenden dargestellt.

**Sprachauswahl**
Unterschieden werden monolinguale (einsprachige), bilinguale (zweisprachige) und multilinguale (mehrsprachige) Korpora.

**Medium**
Unterschieden werden geschriebene, gesprochene und multimodale Korpora. Multimodale Korpora bestehen aus Verbindungen zwischen Text, Audiodateien und visuellen Aufnahmen.

**Größe**
Unterschieden werden Korpora anhand der vorhandenen Datenmenge.

**Annotation**
Unterschieden werden Korpora anhand der beigefügten Informationen. Annotation ist auf unterschiedlichen Ebenen möglich (siehe Abschnitt 1.2).

**Persistenz**
Unterschieden werden statische (abgeschlossene) Korpora und Monitorkorpora (welche stetig erweitert werden).

**Zielsetzung**

Unterschieden wird anwendungsspezifisch in Kategorien. Ein Referenzkorpus beispielsweise ist die Repräsentation einer bestimmten Sprache zu einem bestimmten Zeitpunkt.

Gesprochene Korpora werden zusätzlich nach der Sprechart kategorisiert. Unterschieden werden frei formulierte, ungeplante Aufnahmen (Spontansprache) und vorgelesene Aufnahmen (gelesene Sprache) [102].

## 1.2 Annotation

Annotationen sind dem Korpus hinzugefügte Anmerkungen. Sie enthalten Informationen der Daten auf unterschiedlichen sprachlichen Ebenen [7], [73].

Annotationen können manuell, semiautomatisch oder automatisch erstellt werden [7], [73]. Manuelles Annotieren erfordert viel Zeit und linguistisch geschultes Personal zur Bearbeitung [73]. Die Zeit für die manuelle Annotation einer Audiodatei auf phonetischer Ebene beträgt das 130-fache bis 800-fache der Audiolänge [14]. Für semiautomatische oder automatische Annotation werden spezielle Algorithmen benötigt. Durch die Komplexität der Sprache existiert für keine sprachliche Ebene ein Algorithmus, welcher qualitativ gleichwertige Ergebnisse wie die manuelle Annotation liefert [73]. Vor allem bei Spontansprache ist eine automatische Annotation schwierig, da diese oftmals von den grammatikalischen Regeln abweicht. Dennoch sind automatische Werkzeuge zur Annotation ein wesentlicher Bestandteil der Korpuslinguistik. Dies liegt insbesondere an der immensen Menge benötigter Daten (vergleiche Abschnitt 1.3).

Gesprochene Korpora enthalten zwei Besonderheiten. Zum einen ist für die Annotationen eine zeitliche Zuweisung (Alignment) zu der entsprechenden Audiodatei notwendig. Zum anderen sind im Allgemeinen Informationen über die aufgenommenen Personen vorhanden. Diese Metadaten beinhalten persönliche Angaben wie Alter, Geschlecht, Beruf, Herkunft, sprachliche Besonderheiten et cetera [7], [58].

Im Folgenden werden die sprachlichen Ebenen vorgestellt. Zu jeder Ebene wird ein Überblick über den technischen Stand der automatischen Annotation gegeben.

### 1.2.1 Orthographie

Annotation auf orthographischer Ebene beschreibt die Verschriftlichung des Korpusinhaltes und die Segmentierung in einzelne sprachliche Äußerungen (engl. *Tokens*). Diese wird bei gesprochenen Korpora als Transkript bezeichnet. Es können nicht alle Informationen einer Audioaufnahme verschriftlicht werden (zum Beispiel Husten, Lachen, gleichzeitiges Sprechen). Deswegen sind die Aufnahme und das Transkript keine isomorphen Abbilder. Jeder Transkriptionsprozess ist gleichzeitig ein Selektionsprozess [86].

Orthographische Annotation ist in geschriebenen Korpora vorhanden. Das Annotieren auf orthographischer Ebene ist somit ausschließlich für gesprochene Korpora notwendig. Aufgaben sind das Transkribieren einer Audiodatei (Spracherkennung) und das Alignieren eines Transkripts zu einer Sprachaufnahme.

Automatische Spracherkennung wird unter anderem mittels faltenden neuronalen Netzen (engl. *Convolutional Neural Network* (CNN)) durchgeführt. Hierbei werden für englische Sprache Wortfehlerraten von 1.9 % erreicht [35].

Für das Alignieren von Text zu Sprachaufnahmen existiert eine Vielzahl von Algorithmen. Für lange Aufnahmen wird unter anderem der *SailAlign*-Algorithmus genutzt [40]. Hierbei werden Markierungen (engl. *Landmarks*) gesetzt, wenn eine bestimmte Anzahl an Wörtern als ununterbrochene Folge erkannt wurden. Mithilfe der Landmarks wird das Alignment der verbleibenden Wörter durchgeführt. Für einen Toleranzbereich von 50 ms liegt die Alignmentrate bei ca. 75 % [40].

### 1.2.2 Morphosyntax

Morphosyntaktische Informationen beinhalten Angaben zu der Wortart (engl. *Part-Of-Speech* (POS)). Jedem Token wird eine Wortart zugewiesen. Es können Tempus, Numerus, Genus, Kasus und Person angegeben werden. Dies kann zum Beispiel mit dem Stuttgard-Tübingen-Tagset (STTS) erfolgen [84]. Daneben ist das Hinzufügen von Lemmata (Grundformen) möglich.

Morphosyntaktisches Annotieren wird als *POS-Tagging* bezeichnet [7]. Die Wortart eines Tokens lässt sich oftmals nur mithilfe der Kontextworte bestimmen. POS-Tagger nutzen hierfür in vielen Fällen ein verborgenes Markovmodell (engl. *Hidden Markov Model* (HMM)). Durch die Markoveigenschaft

$$P(q_i = a|q_1...q_{i-1}) = P(q_i = a|q_{i-1}) \tag{1.1}$$

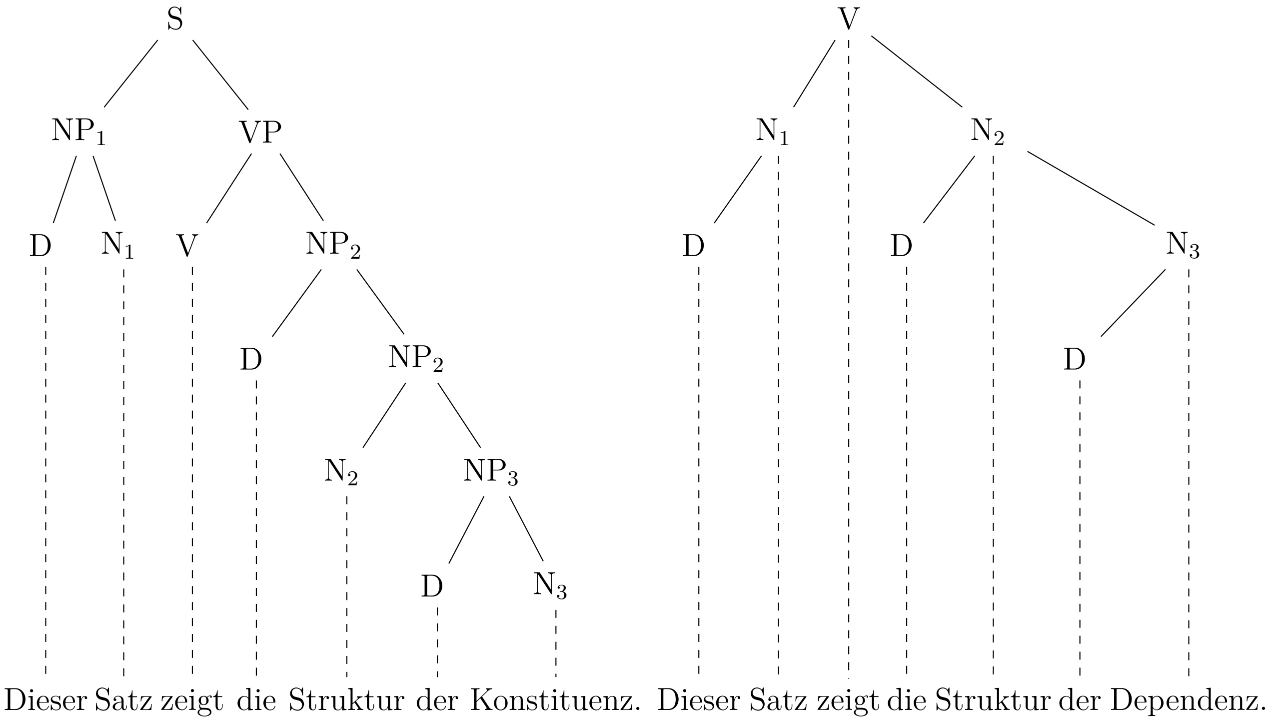
wird die Komplexität des Modells erheblich verringert. Ein HMM-basierter POS-Tagger bestimmt die wahrscheinlichste Folge von Wortarten zu einer gegebenen Folge von Token.

Ein Beispiel für einen POS-Tagger deutscher Sprache ist Trigrams'n'Tags [13]. Dieser Algorithmus erreicht mit dem STTS eine Erkennungsrate von 96.7 % für den Negra-Zeitungskorpus [13].

### 1.2.3 Syntax

Informationen zur Syntax (Grammatik) können als Konstituenten- oder als Dependenzstruktur angegeben werden. Als Konstituentenstruktur wird die Zerlegung eines Satzes in schrittweise kleinere Einheiten (Konstituenten) bezeichnet [7]. Abbildung 1.1a zeigt einen nach Nominalphrasen (NP) und Verbalphrasen (VP) gegliederten Beispielsatz (S).

Die Enden eines Pfades entsprechen den Wortarten (hier: Nomen (N), Determinante (D), Verb (V)). Als Dependenzstruktur wird die hierarchische Abhängigkeit eines Wortes zu einem anderen Wort bezeichnet [7]. Abbildung 1.1b zeigt ein Beispiel dieser Struktur.



(a) Konstituentenstruktur

(b) Dependenzstruktur

Abbildung 1.1: Syntaktische Zerlegung eines Satzes nach der Konstituenten- und der Dependenzgrammatik

Syntaktisches Annotieren wird zerteilen (engl. *Parsing*) genannt. Aufgrund umständlicher und mehrdeutiger Satzstrukturen und dem Abweichen der spontanen Sprache von syntaktischen Regeln ist zuverlässiges automatisches Parsing bisher nicht möglich und wird auch in modernen Systemen nicht verwendet [9].

Aus diesem Grund wird das Parsing oftmals durch eine flache linguistische Analyse (engl. *Chunk Parsing*) ersetzt [9]. Der Algorithmus *Chunk Tagger* erreicht mit drei möglichen Phrasentypen eine korrekte syntaktische Zuweisung von von ca. 89 % [94].

### 1.2.4 Aussprache

Annotationen zur Aussprache erfolgen auf segmentaler Ebene. Es wird die breite (kanonische) und die enge (phonetische) Transkription unterschieden. Die kanonische Transkription besteht aus einer Konvertierung von Graphemen zu Phonemen (G2P). Jedem

Graphem wird ein Phonem zugewiesen. Die kanonische Transkription eines Wortes beinhaltet dessen lexikalische Aussprache [70]. Die phonetische Transkription ist ausschließlich für gesprochene Korpora relevant. Transkribiert wird eine Sprachaufnahme inklusive der sprachlichen Besonderheiten wie Dialekt, Akzent et cetera [70].

Die G2P-Konvertierung kann auf unterschiedliche Weise erfolgen. Im einfachsten Fall wird ein Wörterbuch genutzt, in welchem die Lautschrift für jedes Token nachgeschlagen wird [10]. Eine weitere Möglichkeit ist die regelbasierte Zuweisung von Phonemen zu Graphemen. Aufgrund der vielen Regeln und sprachlichen Ausnahmen ist dies jedoch nur begrenzt möglich [10]. Mit neuronalen Netzen können komplexe Regelwerke implementiert werden. Durch diese Technik lassen sich G2P-Konvertierungen mit kleinen Fehlerraten durchführen. Ein Beispielalgorithmus hierfür ist NetTalk [90]. Mit 1000 Trainingswörtern erreichte NetTalk eine Aussprachegenauigkeit von 77 % für ein Wörterbuch mit 20 012 Einträgen [90].

Für phonetisches Transkribieren und das Alignieren von Phonemen zu einer Audiodatei existieren verschiedene Ansätze. Der Algorithmus *Munich Automatic Segmentation* (MAUS) berechnet mithilfe von HMM einen Wahrscheinlichkeitsgraphen für mögliche Aussprachevarianten und gleicht diese an das akustische Sprachsignal an [78], [83]. Phonetische Transkription ist nicht in jedem Fall eindeutig, weswegen auch manuelle Transkripte desselben Satzes Unterschiede aufweisen können. Aus diesem Grund wird die Qualität der automatischen phonetischen Transkription durch die relative symmetrische Genauigkeit (engl. *Relative Symmetric Accuracy* (RSA)) beschrieben. Die RSA gibt das Verhältnis des Mittelwertes von System-zu-Mensch Übereinstimmung zu dem Mittelwert von Mensch-zu-Mensch Übereinstimmung an. Für MAUS liegt ein RSA von 97.43 % vor [83].

### 1.2.5 Prosodie

Prosodie beschreibt diejenigen Eigenschaften der Sprache, welche nicht durch einzelne Segmente beschrieben werden können [52]. Prosodische Annotation erfolgt auf suprasegmentaler Ebene. Eigenschaften der Prosodie sind unter anderem Intonation, Wortbetonung, Phrasen- und Wortakzent, Pausen und Lautstärke [52]. Prosodische Eigenschaften sind insbesondere für den natürlichen Klang der Stimme bei der Sprachsynthese von Bedeutung [96].

Prosodische Merkmale werden durch Silbenlänge, Intensität und Grundfrequenzverlauf klassifiziert [96]. Aufgrund des hohen Aufwandes manueller prosodischer Annotation sind keine umfangreich prosodisch annotierten Korpora verfügbar [20]. Außerdem existiert keine einheitliche Etikettierung für prosodische Eigenschaften. Mit am weitesten verbreitet sind die *Tones and Break Indices* (ToBI) [92].

Der erste Algorithmus zur prosodischen Annotation mit ToBI für englische Sprache wurde 2010 entwickelt und *Automatic Tones and Break Indices annotation* (AuToBI) genannt [69]. Unzureichend an AuToBI ist, dass die Etiketten Wort für Wort berechnet werden. Hierdurch werden höhere Ebenen wie Phrasen nicht erkannt. Außerdem wurde der Algorithmus mit Radionachrichten (*Boston direction corpus* [51]) trainiert, wodurch die Ergebnisse domänen- und sprachabhängig sind [20].

Ein weiterer Algorithmus für die automatische ToBI-Etikettierung ist *Python Tones and Break Indices* (PyToBI) [21]. Der in Python implementierte Algorithmus erstellt akustische Parameter mit Praat [29] und weist einzelnen Phonemen ToBI-Etiketten zu. Für die englische Sprache beträgt das F-Maß [31], [95] für die Lokalisierung von Tonakzenten 0.77 und von Pausen 0.97 [21].

Für die Erstellung von Grundfrequenzverläufen existieren mehrere semiautomatische und automatische Algorithmen. Zwei praatbasierte Beispiele hierfür sind *Prosogram* [48] und *ProsodyPro* [106].

## 1.3 Anwendungsgebiete

Es existieren viele Anwendungsgebiete für Korpora. Geschriebene Korpora werden unter anderem für die Erstellung von Lexika und Wörterbüchern, aber auch für Statistiken von Wortverteilungen genutzt [7].

Der Fokus in der vorliegenden Ausarbeitung liegt auf gesprochenen Korpora, welche in der Sprachverarbeitung verwendet werden. Weitere Anwendungen gesprochener Korpora sind beispielsweise das Erstellen von Lehrmaterial, Übersetzungen, Aussprachevarianten und das Festhalten von Mundarten [73].

Die Sprachverarbeitung umfasst drei Schwerpunktthemen: Die Spracherkennung/ Erkennung der sprechenden Person, die Sprachsynthese und die Sprachanalyse.

### 1.3.1 Korpora der Sprachverarbeitung

Algorithmen zur Spracherkennung und zur Erkennung der sprechenden Person benötigen teilweise viele Trainingsdaten. Für ein Vortraining neuronaler Netze wurden 5870 Stunden Sprachmaterial von *Voice Search* und 1400 Stunden Sprachmaterial von *YouTube* verwendet [39]. Für das Training eines rekurrenten neuralen Modells für automatische Spracherkennung wurden ein Korpus mit 300 Stunden Sprachmaterial und ein Korpus mit 22 Millionen Äußerungen verwendet [97].

Der Algorithmus *WaveNet* besteht aus einem tiefen neuronalen Netz (engl. *Deep Neural Network* (DNN)) [53]. WaveNet wird sowohl zur Spracherkennung als auch zur Sprachsynthese genutzt. Die Spracherkennung wurde mit dem 5:24 Stunden großen Korpus *Texas Instruments/Massachusetts Institute of Technology* (TIMIT) [27] trainiert, welcher auf orthographischer und phonetischer Ebene aligniert ist [53]. Für die Sprachsynthese wurden ein Algorithmus zur Erstellung von realistischen Intonationskurven und ein Vorleseautomat (engl. *Text-To-Speech* (TTS)) trainiert [53]. Für die Generierung der

Intonationskurven wurde 44 Stunden Sprachmaterial von 109 Personen verwendet. Für TTS wurden ein englischsprachiger Korpus mit 24:36 Stunden und ein chinesischesprachiger Korpus mit 34:48 Stunden Sprachmaterial genutzt [53].

Die Sprachanalyse unterteilt sich in viele weitere Bereiche. Es wird unter anderem die Erkennung und Vorhersage von Phonemlängen untersucht. Hierfür wurde ein Korpus mit 509 399 Phonemen verwendet [30]. Außerdem wird der Zusammenhang von Sprache und Gesundheit untersucht, da eine Korrelation von prosodischen und gesundheitlichen Eigenschaften vermutet wird. Für eine umfangreiche Untersuchung steht zur Zeit keine ausreichende Menge an prosodisch annotiertem Sprachmaterial zur Verfügung [19].

### 1.3.2 Eigenschaften eines Korpus zur Sprachverarbeitung

Aus den in Unterabschnitt 1.3.1 beschriebenen Anwendungsbeispielen lassen sich die Anforderungen an einen Korpus zur Verarbeitung deutscher Sprache extrahieren. Eine Präzisierung dieser Eigenschaften ist in Tabelle 1.1 gegeben.

Tabelle 1.1: Anforderungen an einen Korpus zur Verarbeitung deutscher Sprache

| Typologie | Eigenschaften |
|---|---|
| Sprachauswahl | – deutsch |
| Medium | – gesprochen |
| Größe | – größtmöglich |
| Annotation | – orthographisch<br>– phonetisch<br>– prosodisch |
| Persistenz | – statischer Korpus oder Monitorkorpus |
| Zielsetzung | – viel Sprachmaterial von vielen Personen<br>– akzentfreies Hochdeutsch<br>– große Themenvielfalt |

# 2 Gesprochene Korpora

Die in Unterabschnitt 1.3.2 genannten Anforderungen an einen Korpus zur deutschen Sprachverarbeitung treffen auf keinen existierenden Korpus vollständig zu. Vor allem eine umfangreiche prosodische Annotation ist für deutschsprachige Korpora nicht vorhanden [20]. Im Folgenden werden Korpora vorgestellt, welche für die Sprachverarbeitung genutzt werden.

## 2.1 Fremdsprachige Korpora

In einigen anderen Sprachen als Deutsch sind deutlich umfangreichere Korpora zur Sprachverarbeitung erstellt worden. Nachfolgend werden ausgewählte Korpora englischer und russischer Sprache vorgestellt. Eine tabellarische Übersicht der fremdsprachigen Korpora ist am Ende dieses Abschnittes in Tabelle 2.1 gegeben.

Einer der am meisten genutzten Korpora zur englischen Sprachverarbeitung ist der *Boston University Radio Speech Corpus* (BURSC), Eigentum des *Linguistic Data Consortium* (LDC) [56]. Dieser für TTS erstellte Korpus entstand 1995 durch den steigenden Bedarf an prosodisch annotiertem Sprachmaterial [55]. Der BURSC ist manuell auf orthographischer, phonetischer, morphosyntaktischer und teilweise auf prosodischer Ebene annotiert, beinhaltet Sprachaufnahmen von sieben Personen und besteht aus sieben Stunden Sprachmaterial [55].

Der umfassendste prosodisch annotierte Korpus (2005) ist der *Hong Kong Corpus of Spoken English* (HKCSE), Eigentum des *Research Centre for Professional Communication in English* (RCPCE) [17]. Dieser aus vier Subkorpora bestehende Korpus umfasst 200 Stunden englisches Sprachmaterial, von denen 106 Stunden manuell prosodisch annotiert wurden [17]. Inhalt dieses Korpus sind akademische, geschäftliche und öffentliche Gespräche von Chinesen aus Hong Kong und Personen mit englischer Erstsprache. Die Verwendung des HKCSE ist online möglich [17].

Ein Beispiel für einen umfangreichen prosodisch annotierten Korpus russischer Sprache ist der *Corpus Of Russian Professionally Read Speech* (CORPRES), Eigentum der *Saint Petersburg State University* (SPSU) [52]. CORPRES wurde für TTS über drei Jahre erstellt und beinhaltet 60 Stunden Sprachmaterial, von denen 24 Stunden manuell prosodisch annotiert wurden [52]. Die akzentfreien Sprachaufnahmen stammen von vier professionellen Sprechern und vier professionellen Sprecherinnen [52].

Tabelle 2.1: Übersicht ausgewählter gesprochener Korpora englischer und russischer Sprache

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| LDC | BURSC[56] | gelesene Sprache; Radionachrichten | 7 | 7:00 | – orthographisch<br>– morphosyntaktisch<br>– phonetisch<br>– prosodisch | 1200 $ | – englischsprachig<br>– 3:30 h prosodisch annotiert |
|  | TIMIT[27] | gelesene Sprache | 630 | ca. 5:15 | – orthographisch<br>– phonetisch | 0 € | – englischsprachig |
| RCPCE | HKCSE[17] | spontane Sprache; akademische, geschäftliche, öffentliche Gespräche | k.A. | 200:00 | – orthographisch<br>– prosodisch | 0 € | – englischsprachig<br>– 106 h prosodisch annotiert<br>– ausschließlich online nutzbar |
| SPSU | CoRuSS[98] | spontane Sprache | 60 | 14:00 | – orthographisch<br>– prosodisch | k.A. | – russischsprachig |
|  | CORPRES[52] | gelesene Sprache | 8 | 60:00 | – orthographisch<br>– phonetisch<br>– prosodisch | k.A. | – russischsprachig<br>– 24 h prosodisch annotiert |

## 2.2 Deutschsprachige Korpora

Es existiert eine Vielzahl gesprochener Korpora deutscher Sprache. Nur eine kleine Teilmenge davon enthält prosodisch annotiertes Material. Im Folgenden werden zwei Plattformen vorgestellt, welche eine Vielzahl von gesprochenen Korpora besitzen. Für die Sprachsynthese besonders interessante Korpora werden in Unterabschnitt 2.2.3 ausgeführt. Eine Übersicht der gesprochenen Korpora deutscher Sprache ist am Ende dieses Abschnittes in Tabelle 2.3 gegeben.

### 2.2.1 Bayrisches Archiv für Sprachsignale

Das Bayrische Archiv für Sprachsignale (BAS) wurde 1995 für die Verwaltung deutscher Sprachressourcen an der Ludwig-Maximilians-Universität München gegründet [22]. Aufgabe des BAS ist das zentrale Sammeln, Pflegen, Standardisieren und Distribuieren von deutschen Sprachressourcen [22].

Das Standardisieren von Korpora ist eine notwendige Aufgabe, um diese maschinell zu verarbeiten. Die meisten ad hoc erstellten Dateiformate weisen signifikante Defizite auf [82]. Diese drücken sich beispielsweise dadurch aus, dass das Dateiformat nicht erweiterbar ist, nicht mit Standardwerkzeugen für *Uniplexed Information and Computing Service* (UNIX) bearbeitet werden kann oder dass sprachliche Ebenen vermischt werden [82].

#### Partitur

Das BAS hat für die Standardisierung von Sprachkorpora ein Partiturformat entworfen. Diese Datei mit der Endung *.par* gliedert sich in einen Kopf und einen Rumpf [82]. Jede Zeile der Partitur beginnt mit einem drei Byte langen Etikett und einem Doppelpunkt.

```
LHD: Partitur Version
REP: Aufnahmeort
SNB: Anzahl Bytes pro Abtastwert
SAM: Abtastfrequenz
SBF: Byte-Reihenfolge (big-endian 01, little-endian 10)
SSB: Bit Auflösung
NCH: Anzahl Audiokanäle
SPN: Sprecher Identifikation
LBD:
```

Abbildung 2.1: Aufbau des Kopfes der BAS Partitur mit allen obligatorischen Angaben

Der Kopf beinhaltet Informationen der Audiodatei. Eine Übersicht über alle notwendigen Einträge des Kopfes sind in Abbildung 2.1 dargestellt. Zusätzlich können weitere

Informationen mit bestimmten Etiketten angegeben werden [82]. Das Etikett `LBD:` zeigt das Ende des Kopfes an.

Der Rumpf beinhaltet die Annotationen der Audiodatei. Für die verschiedenen Annotationen werden unterschiedliche Etiketten verwendet [82]. Ein Beispiel für einen Rumpf ist in Abbildung 2.2 dargestellt. In dem Beispiel sind die orthographischen (`ORT:`), kanonischen (`KAN:`), prosodischen (`PRM:`) und mit MAUS erstellten phonetischen (`MAU:`) Informationen einer Aufnahme von /Sie hören/ vorhanden. Alle Zeiten werden in Abtastwerten der zugehörigen Audiodatei angegeben.

```
ORT:   0     Sie
ORT:   1     h"oren
KAN:   0     z i:
KAN:   1     " h 2: R @ n
WOR:   0     7938   0  Sie
WOR:   7938  13230  1  h"oren
MAU:   0     2205   0  z
MAU:   2205  5733   0  i:
MAU:   7938  2646   1  h
MAU:   10584 3087   1  2:
MAU:   13671 3087   1  r
MAU:   16758 1323   1  @
MAU:   18081 3087   1  n
PRM:   4208  0      L*H
PRM:   11378 1      H*L
PRM:   16837 1      L*
```

Abbildung 2.2: Aufbau des Rumpfes der BAS Partitur (Beispiel einer Realisierung von /Sie hören/)

Die Etiketten des Rumpfes werden in fünf Kategorien eingeteilt. Eine Übersicht mit Beispielen dieser Kategorien ist nachfolgend aufgelistet.

**Informationen mit Relation**

Enthalten ist nach dem Etikett die Referenz auf das nummerierte Wort (Nummerierung startet bei 0) und die zugehörige Information.

```
KAN:   0     z i:
```

**Informationen für Zeitspannen**

Enthalten ist nach dem Etikett der Startzeitpunkt, die Dauer und die Information.

```
WOR:   0     7938   Sie
```

**Informationen für Zeitpunkte**

Enthalten ist nach dem Etikett der Zeitpunkt und die Information.

```
PRM:   4208  L*H
```

**Informationen für Zeitspannen mit Relation**

Enthalten ist nach dem Etikett der Startzeitpunkt, die Dauer, die Referenz auf das nummerierte Wort und die Information.

```
WOR:   0     7938   0  Sie
```

**Informationen für Zeitpunkte mit Relation**

Enthalten ist nach dem Etikett der Zeitpunkt, die Referenz auf das nummerierte Wort und die Information.

```
PRM:   4208  0      L*H
```

### Auswahl an Korpora des Bayrischen Archivs für Sprachsignale

Das BAS ist im Besitz von 34 gesprochenen Korpora, welche für unterschiedlichste Zwecke erstellt wurden.

Aufgrund der umfangreichen prosodischen Annotation ist der *BITS Unit Selection synthese corpus* (BITS-US) der am besten geeignete Korpus zur Sprachsynthese [79]. In Abschnitt 2.2.3 wird ausführlicher auf diesen Korpus eingegangen.

Ein weiterer prosodisch annotierter Korpus ist der PhonDat 2 (PD2) [81]. PD2 enthält orthographisch, kanonisch, phonetisch und prosodisch annotierte Daten von 16 Personen, welche je 200 Sätze lesen. Nachteile für die Sprachsynthese sind, dass die prosodischen Annotationen ausschließlich segmental vorliegen, der Korpus nur ca. 4:25 Stunden Sprachmaterial umfasst, die Sätze zu einem großen Anteil aus Reisebüros stammen und nicht alle Personen hochdeutsch sprechen [81].

Ein mit 31:30 Stunden Sprachmaterial umfangreicher Korpus ist der Siemens 100 (SI100) [74]. Die Aufnahmen des SI100 besteht aus jeweils ca. 100 Sätzen von 101 Personen. Dieser Korpus wurde für die Erstellung von Diktiersoftware erstellt. Aus diesem Grund werden Satzzeichen laut ausgesprochen und es liegen flache Intonationsverläufe vor [74].

Prosodisch annotiertes Material von einem professionellen Sprecher enthält der *Siemens Synthese Korpus* (SI1000P) [76]. Dieser liest 1000 Sätze aus Zeitungsartikeln, welche orthographisch, kanonisch und prosodisch annotiert wurden. Der SI1000P beinhaltet 1000 Sätze eines weiteren professionellen Sprechers ohne prosodische Annotation [76].

Aus dem Projekt Verbmobil sind in den Jahren 1993–2000 die beiden Korpora Verbmobil I (VM1) und Verbmobil II (VM2) entstanden [75], [77]. Diese Korpora wurden für die automatische Übersetzung von deutsch, englisch und japanisch geschaffen. Gemeinsam beinhalten sie ca. 27 Gigabyte Datenmaterial von mehreren Hundert Personen. VM1 und VM2 bestehen aus Dialogen zur Terminplanung, Reiseplanung und Hotelreservierung [75], [77].

### 2.2.2 Datenbank für Gesprochenes Deutsch

Die Datenbank für Gesprochenes Deutsch (DGD) ist eine Internetseite des Instituts für Deutsche Sprache Mannheim (IDS) [88]. Über diese Plattform stellt das IDS gesprochene Korpora deutscher Sprache zur Verfügung. Insgesamt sind 36 Korpora mit über 3100 Stunden Audiomaterial vorhanden [58]. Der überwiegende Anteil der Korpora entstand aus linguistischen Untersuchungen von Sprachvariationen (Jugenddeutsch, Mundarten, Emigrantendeutsch et cetera) [58]. Nach einer kostenlosen Registrierung können die Sprachressourcen online verwendet werden. Die Korpora sind größtenteils orthographisch aligniert und enthalten teilweise weitere Annotationen wie POS-Tags und Lemmata [58].

Die DGD bietet zwei Möglichkeiten der Verwendung [87]. Im *Browsing*-Modus ist das explorative Ansehen und -hören der Sprechereignisse und Transkripte sowie eine Einzelanalyse möglich [87]. Im *Recherche*-Modus können die Korpora gezielt durchsucht werden. Das Erstellen und Speichern von virtuellen Teilkorpora (auch über mehrere Korpora hinweg) sowie eine Token-Suche auf allen Annotationsebenen sind möglich [87].

Ein Vorzeigeprojekt in der DGD ist der seit 2008 entstehende Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK) [88]. FOLK ist ein Monitorkorpus und umfasst 285:39 Stunden Audiomaterial in 377 Aufnahmen von 1131 Sprecherinnen und Sprechern (2021) [88]. FOLK ist orthographisch in Zeitabschnitte mit einer maximalen Länge von 5 Sekunden aligniert und wurde mit POS-Tags und Lemmata annotiert [88]. Der gesamte Korpus kann online verwendet werden und 22 Ausschnitte können heruntergeladen werden [88]. FOLK ist ein Referenzkorpus und beinhaltet Gespräche aus privaten, institutionellen und öffentlichen Bereichen. Anwendungen für FOLK sind unter anderem die Gesprächsforschung und die Korpuslinguistik [58].

### 2.2.3 Korpora für die deutschsprachige Sprachsynthese

In diesem Abschnitt werden drei Korpora vorgestellt, welche die Anforderungen aus Tabelle 1.1 zu einem Großteil erfüllen und somit für die Sprachsynthese deutscher Sprache geeignet sind.

#### BITS Unit Selection synthese corpus

Der BITS-US ist mit 13:33 Stunden der größte prosodisch annotierte Korpus des BAS [22]. Das Sprachmaterial des BITS-US stammt von zwei ausgewählten Sprechern und zwei ausgewählten Sprecherinnen. Jede Person liest 1683 Sätze. Die Sätze enthalten Diphon-Kombinationen in verschiedenen prosodischen Kontexten, welche größtenteils keinen Sinn ergeben (Beispiel: „Das Dunkel war ein kurzes und leiser werdendes Verschwinden.“) [79].

Für jeden gesprochenen Satz ist für jede Person eine Aufnahme mit einem Kopfbügelmikrophon, eine Aufnahme mit einem Elektroglottographen und eine Aufnahme mit einem Raummikrophon vorhanden [79].

Der gesamte Korpus wurde manuell auf orthographischer, kanonischer, phonetischer und prosodischer Ebene annotiert. Die kanonische und die phonetische Annotation besteht aus Etiketten des erweiterten deutschen *Speech Assessment Methods Phonetic Alphabet* (SAMPA). Die prosodische Annotation besteht aus den Etiketten der *German Tones and Break Indices 'light'* (GToBI light) [79] (siehe Unterabschnitt 5.1.1).

Für jeden gesprochenen Satz ist für jede Person eine Partitur-Datei nach dem BAS Standard (vergleiche Abschnitt 2.2.1), eine TextGrid-Datei mit den phonetischen und den prosodischen Etiketten und eine *Extensible Markup Language*-Datei (XML) mit allen Informationen der Partitur-Datei vorhanden.

### Kiel Corpus of Spoken German

Der *Kiel Corpus of Spoken German* (KCSG) wurde an der Christian-Albrechts-Universität (CAU) in Kiel erstellt und unterteilt sich in zwei Subkorpora: den Korpus gelesener Sprache *Kiel Corpus of Spoken German read speech* (KCSGrs) und den spontansprachlichen Korpus *Kiel Corpus of Spoken German spontaneous speech* (KCSGss) [41].

Der KCSG beinhaltet orthographische, kanonische, phonetische und prosodische Annotation. Die kanonische und die phonetische Annotation besteht aus Etiketten von modifiziertem SAMPA [41]. Für die prosodische Etikettierung wurde das Kieler Intonationsmodell (KIM) [59] genutzt.

Nicht jede Sprachaufnahme ist auf allen vier Ebenen annotiert. Die Menge der vollständig annotierten Sprachaufnahmen bildet den Kern [41]. Die Endung der Annotationsdatei zeigt an, welche Ebenen annotiert wurden. Diese Zuordnung ist in Tabelle 2.2 aufgelistet. Für Dateien, welche ausschließlich orthographisch annotiert sind, existiert keine Annotationsdatei.

Tabelle 2.2: Dateiendungen des *Kiel Corpus of Spoken German* in Abhängigkeit der vorhandenen Annotationsebenen

| | **Annotationen** | | | |
|---|---|---|---|---|
| **Dateiendung** | orthographisch | kanonisch | phonetisch | prosodisch |
| .s0 | ✓ | ✓ | | |
| .s1 | ✓ | ✓ | ✓ | |
| .s2 | ✓ | ✓ | ✓ | ✓ |

Für jeden Sprecher und jede Sprecherin im KCSG sind Metadaten verfügbar. Diese beinhalten das Geschlecht, das Alter, den regionalen Akzent und die Gesamtlänge der gesprochenen Aufnahmen [41].

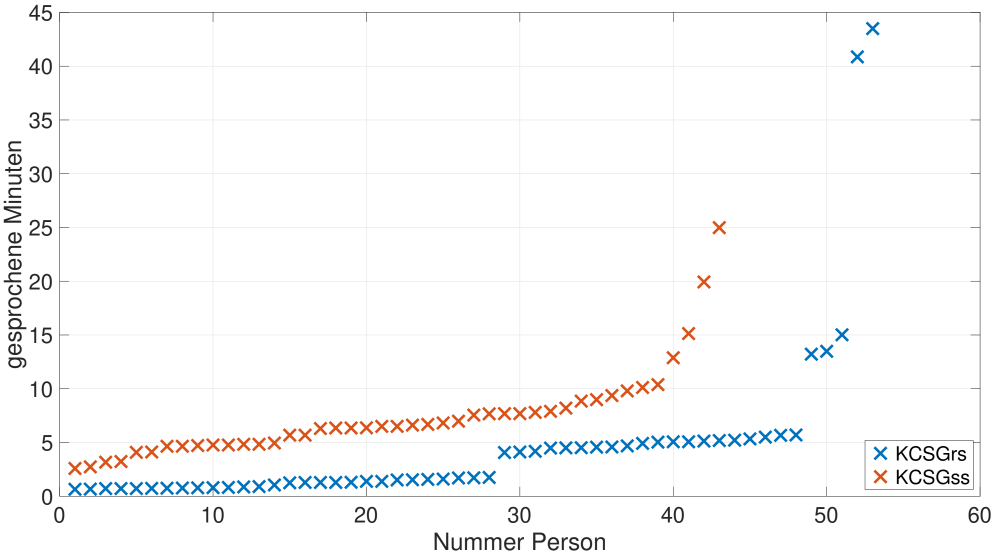


Abbildung 2.3: Verteilung des Sprachmaterial der Kerne vom *Kiel Corpus of Spoken German read speech* und vom *Kiel Corpus of Spoken German spontaneous speech*

Der KCSGrs beinhaltet Aufnahmen der Korpora PhonDat90 und PhonDat92, welche 1994 veröffentlicht wurden [41]. Der Inhalt besteht aus unterschiedlichen Sätzen und zwei Kurzgeschichten. Der KCSGrs enthält 5:41 Stunden Sprachmaterial von 26 Sprecherinnen und 27 Sprechern. Der Kern des KCSGrs beinhaltet 4:15 Stunden Sprachmaterial derselben Personen [41]. Die Verteilung des Sprachmaterials für den Kern ist in Abbildung 2.3 dargestellt. Es ist zu erkennen, dass keine homogene Verteilung vorliegt. Die Aufnahmelänge pro Person variiert zwischen 39 Sekunden und 43:30 Minuten.

Der KCSGss besteht aus zwei Teilen. Der erste Teil enthält Aufnahmen des Korpus VerbMobil, welche zwischen 1995 und 1997 veröffentlicht wurden. Der Inhalt dieser Aufnahmen besteht aus Dialogen über Terminanfragen [59]. Der zweite Teil enthält Gespräche über die Fernsehserie „Lindenstraße“. Der KCSGss enthält 8:34 Stunden Sprachmaterial von 27 Sprecherinnen und 37 Sprechern. Der Kern des KCSGss beinhaltet 5:04 Stunden Sprachmaterial von 13 Sprecherinnen und 18 Sprechern [41]. Die Verteilung des Sprachmaterials für den Kern ist in Abbildung 2.3 dargestellt. Es ist zu erkennen, dass keine homogene Verteilung vorliegt. Die Aufnahmelänge pro Person variiert zwischen 2:35 Minuten und 24:59 Minuten.

### Spoken Wikipedia Corpus

Der *Spoken Wikipedia Corpus* (SWC) ist an der Universität Hamburg entstanden und im Besitz des Hamburger Zentrums für Sprachkorpora (hzsk).

Der SWC wurde vollständig automatisch erstellt. Als Sprachressource wurden gelesene Artikel der Internetplattform Wikipedia [99] genutzt. Alle Artikel von Wikipedia stehen unter der Lizenz *Creative Commons Attribution-ShareAlike 3.0 Unported* und unter der GNU-Lizenz für freie Dokumentation, wodurch jegliches Nutzen und Verarbeiten gestattet ist [5].

Bei dem SWC handelt es sich um einen Monitorkorpus. Dieser kann zu jedem Zeitpunkt mit den aktuellen Daten von Wikipedia erstellt werden. In den letzten 10 Jahren ist das Sprachmaterial des deutschen und des englischen Wikipedias mit einer Rate von je 32-33 Stunden pro Jahr angestiegen [5].

Der Algorithmus für die Aufbereitung der Daten besteht aus mehreren Schritten. Zunächst werden die Dateien für Audio, Text und Metadaten der gesprochenen Artikel heruntergeladen [5]. Es folgt ein Alignment auf orthographischer Ebene mit einem modifizierten *SailAlign*-Algorithmus (vergleiche Unterabschnitt 1.2.1) [5]. Mithilfe dieser Segmentierung auf Wortebene werden Sätze mit Start- und Endzeitpunkt extrahiert und mit MAUS (vergleiche Unterabschnitt 1.2.4) auf phonetischer Ebene aligniert [5].

Der SWC ist nach Artikeln strukturiert. Für jeden Artikel liegen eine Audiodatei, eine Metadatendatei und eine XML-Datei mit den Annotationen vor. Zusätzlich existiert der Inhalt des Artikels als *Hypertext Markup Language*-Datei (HTML), XML-Datei und Textdatei [5]. Der SWC kann kostenlos auf der Internetseite des hzsk heruntergeladen werden [5].

Wikipedia stellt Artikel in unterschiedlichen Sprachen zur Verfügung. Der Algorithmus für die Erstellung des SWC wurde für deutsche, englische und niederländische Sprache getestet und kann für weitere Sprachen verwendet werden [5]. Der SWC umfasst die Gesamtmenge aller bearbeiteten Artikel. Die Artikel einer einzelnen Sprachen sind als Teilkorpus verfügbar.

Der *German Spoken Wikipedia Corpus* (GSWC) besteht aus 1014 deutschsprachigen Artikeln mit einer Gesamtlänge von 386 Stunden. Diese wurden von 350 Personen gelesen [5]. Ob diese Personen professionell sprechende Personen sind, ist nicht angegeben. Der GSWC enthält 249 Stunden orthographisch und 129 Stunden phonetisch aligniertes Audiomaterial (Stand 2016) [5]. Die Verteilung des Sprachmaterials ist in Abbildung 2.4 dargestellt. Es ist eine stark inhomogene Verteilung zu erkennen.

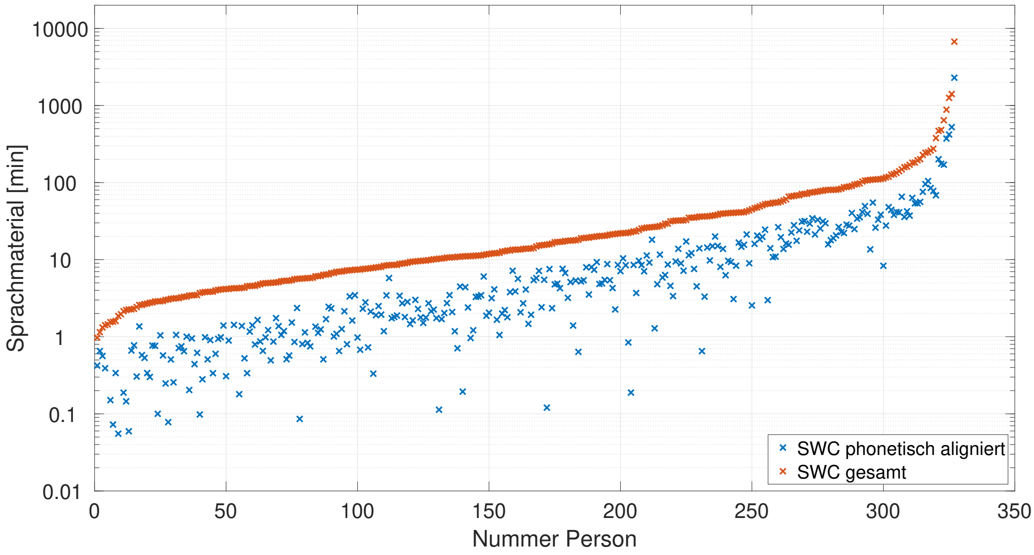


Abbildung 2.4: Sprachmaterial des *German Spoken Wikipedia Corpus*

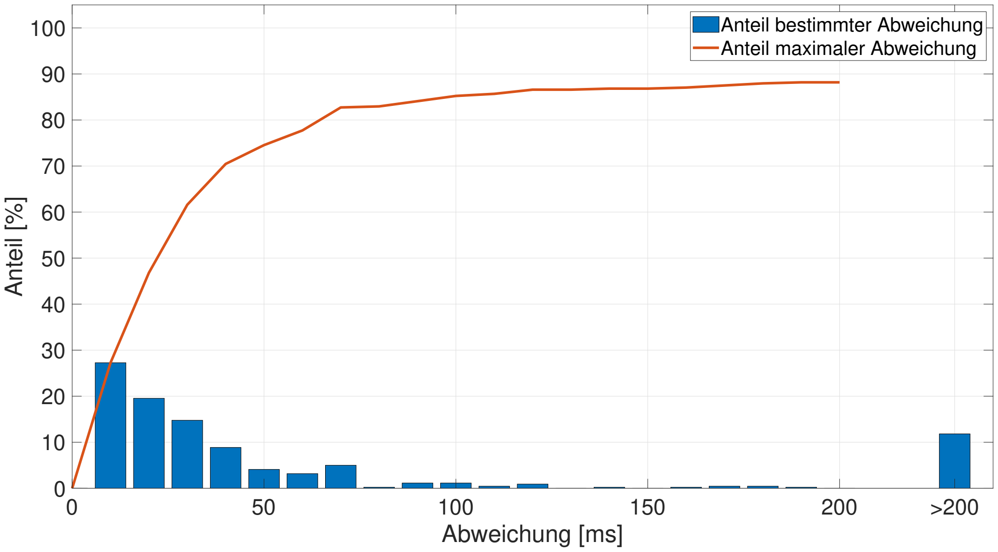


Abbildung 2.5: Anteil der automatisch erstellten Wortgrenzen mit bestimmten und maximalen zeitlichen Abweichungen zu den manuell erstellten Wortgrenzen des Artikels „Photodiode“ aus dem *German Spoken Wikipedia Corpus*

Gründe für den geringen Anteil an aligniertem Audiomaterial sind schlechte Audioqualität, stark akzentuierte Sprache, synthetische Sprache, Hintergrundgeräusche, Besonderheiten der Aussprache (z.B. Papst Pius XI) oder falsche Zuordnung von Audiodatei und Artikel [5].

Evaluiert wurde die orthographische Ebene des GSWC mittels des manuell annotierten Artikels „Photodiode“ [5]. Dieser Artikel wurde zu 97.5 % orthographisch aligniert [5]. Der Anteil an automatisch erstellten Wortgrenzen mit bestimmten zeitlichen Abweichungen zu manuell erstellten Wortgrenzen ist in Abbildung 2.5 dargestellt. 75 % der automatisch erstellten Wortgrenzen weichen maximal 50 ms von den manuell erstellten Wortgrenzen ab.

Tabelle 2.3: Übersicht ausgewählter gesprochener Korpora deutscher Sprache

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| BAS[22] | BITS-US | gelesene Sprache; Diphon-Kombinationen in verschiedenen prosodischen Kontexten | 4 | 13:33 | – orthographisch<br>– kanonisch<br>– phonetisch<br>– prosodisch | 754.35 € | – an der TUD vorhanden<br>– beinhaltet keine sinnvollen Sätze<br>– inklusive Elektroglottographensignal |
|  | BITS-LG | gelesene Sprache; Logatome mit allen deutschen Diphon-Kombinationen | 4 | 3:07 | – orthographisch<br>– kanonisch<br>– phonetisch | 754.35 € | – Kombination mit BITS-US möglich, da gleiche Sprecherinnen und Sprecher |
|  | PD2 | gelesene Sprache; Zugauskunft | 16 | k.A. | – orthographisch<br>– kanonisch<br>– phonetisch<br>– prosodisch | 255.65 € | – jede Person liest 200 Sätze<br>– ausschließlich prosodische Segmentierung vorhanden<br>– regionale Vielfalt der Mundarten |

Tabelle 2.3: Übersicht ausgewählter gesprochener Korpora deutscher Sprache (Fortsetzung)

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| | SI100 | gelesene Sprache; Sätze aus Zeitungsartikeln | 101 | 31:30 | – orthographisch<br>– kanonisch<br>– phonetisch | 0 € | – ca. 100 gelesene Sätze pro Person<br>– für Diktiersoftware entwickelt |
| | SI1000P | gelesene Sprache; Sätze aus Zeitungsartikeln | 2 | k.A. | – orthographisch<br>– kanonisch<br>– prosodisch | 5010.66 € | – 1000 gelesene Sätze pro Person<br>– Ein Sprecher (ca. 1:22 h) prosodisch annotiert |
| | VM1/ VM2 | spontane Sprache; Dialoge über Terminplanung, Reiseplanung, Hotelreservierung | 1261/ 445 | k.A./ k.A. | – orthographisch<br>– morphosyntaktisch<br>– kanonisch<br>– phonetisch<br>– prosodisch | 3834.75 €/ 9970.35 € | – deutsch-, englisch- und japanischsprachig<br>– 3002 Dialogaufnahmen auf 54 CDs (26.6 GB)<br>– Teile erhältlich (225.65 €/CD)<br>– für Übersetzungssoftware |

Tabelle 2.3: Übersicht ausgewählter gesprochener Korpora deutscher Sprache (Fortsetzung)

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| CAU | KCSGrs[41] | gelesene Sprache; Sätze und Kurzgeschichten | 53 | 5:41 | – orthographisch<br>– kanonisch<br>– phonetisch<br>– prosodisch | 300 € | – 4:15 h prosodisch annotiert<br>– Sprachmaterial ungleich verteilt |
|  | KCSGss[41] | spontane Sprache; Dialoge über Terminvergaben | 64 | 8:34 | – orthographisch<br>– kanonisch<br>– phonetisch<br>– prosodisch | 600 € | – 5:04 h prosodisch annotiert<br>– Sprachmaterial ungleich verteilt |
| DGD[88] | FOLK | spontane Sprache; Dialoge | 1131 | 285:39 | – orthographisch<br>– morphosyntaktisch | 0 € | – Referenzkorpus<br>– Monitorkorpus |
|  | PF | gelesene und spontane Sprache | 402 | 79:11 | – orthographisch<br>– morphosyntaktisch | 0 € | – Sprache der BRD, der DDR, Österreich und der Schweiz |
| HU Berlin | BeMaTaC[58] | spontane Sprache; Dialoge über Stadtpläne | 34 | 2:23 | – orthographisch<br>– morphosyntaktisch<br>– syntaktisch | 0 € | – Gliederung in L1- und L2-Korpus |

Tabelle 2.3: Übersicht ausgewählter gesprochener Korpora deutscher Sprache (Fortsetzung)

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| hzsk | GSWC[5] | gelesene Sprache; Wikipedia-Artikel | 350 | 386:00 | – orthographisch<br>– phonetisch | 0 € | – Monitorkorpus<br>– Korpus wurde automatisch annotiert |
| KgSR[43] | KgSR | spontane Sprache; Gespräche von Studierenden und Gartenarbeitskräften | k.A. | 120:00 | – orthographisch<br>– kanonisch<br>– phonetisch<br>– morpho-syntaktisch<br>– prosodisch | k.A. | – 10 h prosodisch annotiert<br>– Akzente und Grenztonmuster prosodisch annotiert<br>– Mundart Ruhrgebiet |
| Universität Bielefeld | Leap[58] | gelesene und spontane Sprache; Geschichten, Interviews | 131 | 12:00 | – orthographisch<br>– phonetisch<br>– morpho-syntaktisch<br>– syntaktisch<br>– prosodisch | 0 € | – multilingual<br>– unvollständige Annotationen |

Tabelle 2.3: Übersicht ausgewählter gesprochener Korpora deutscher Sprache (Fortsetzung)

| Quelle | Bezeichnung | Inhalt | Sprecher und Sprecherinnen | Dauer [hh:mm] | Annotation | Kosten | Anmerkungen |
|---|---|---|---|---|---|---|---|
| Universität Leipzig | GeWiss[58] | gelesene und spontane Sprache; akademischer Kontext | 462 | 126:00 | – orthographisch | 0 € | – multilingual<br>– Informationen zu Sprachwechseln, Diskurskommentierungen, Zitaten, Verweisen vorhanden |
| Universität Stuttgart | DIRNDL[24] | spontane Sprache; Radionachrichten | 9 | 5:00 | – orthographisch<br>– morpho-syntaktisch<br>– prosodisch | k.A. | – Satzstruktur im XML-Format vorhanden |
|  | GRAIN[23] | spontane Sprache; Radionachrichten | >100 | 23:00 | – orthographisch<br>– phonetisch<br>– morpho-syntaktisch | k.A. | – automatisch erstellt<br>– Wortarten händisch annotiert |

## 2.3 *Corpus of Aligned Read speech Including Annotations* (CARInA)

Ziel der vorliegenden Arbeit ist die Zusammenstellung und Validierung einer Datenbasis qualitativ hochwertiger Sprachdaten und der dazugehörigen Annotationen für die Technische Universität Dresden (TUD). Kein einzelner Korpus erfüllt jeden der in Tabelle 1.1 aufgeführten Punkte. Aus der Aufgabenstellung entstand ein aus drei Sprachkorpora bestehender Datensatz.

Der BITS-US ist im Besitz der TUD und bildete bisher die gesamte Datenbasis. Aufgrund der Unnatürlichkeit des Sprachmaterials war eine Ergänzung beziehungsweise ein Ersetzen notwendig.

Im Rahmen der vorliegenden Arbeit wurde der KCSGrs beschafft. Dieser liegt in originaler Form an der TUD vor. Für die Vereinigung des BITS-US und des KCSGrs ist eine Konvertierung der prosodischen Etiketten notwendig. Aufgrund des Informationsgehaltes ist eine automatische Konvertierung ausschließlich vom KIM zum System GToBI light möglich. Ein Vorschlag für diese Konvertierung ist in Unterabschnitt 5.1.3 beschrieben.

Der größte Teil der neuen Datenbasis besteht aus dem *Corpus of Aligned Read speech Including Annotations* (CARInA). Der CARInA wurde vollständig automatisch generiert. Der Korpus beinhaltet das orthographisch und phonetisch alignierte Sprachmaterial des GSWC, bestehend aus 1015 Artikeln (Stand 26.01.2021). Diesen Daten wurden in mehreren Schritten kanonische, morphosyntaktische, silbische und prosodische Informationen hinzugefügt.

Da der GSWC ein Monitorkorpus ist, kann auch der CARInA als Monitorkorpus betrachtet werden. Für die Erstellung des aktuellen CARInA wird der GSWC inklusive Audiomaterial von der Internetseite des hzsk [15] heruntergeladen und entpackt. Der Ordner `de-with-audio` wird umbenannt zu `SpokenWikipediaCorpus` und mit dem Skript `createCARInA.m` und den Ordnern `CARInAfunction` und `Dictionary` in ein Verzeichnis abgelegt. Die Arbeitsschritte in `createCARInA.m` werden gesetzt und der CARInA wird erstellt. Die Vorverarbeitung des Ordners `Dictionary` wird in Abschnitt 4.2 beschrieben. Das Skript `createCARInA.m` erstellt den CARInA mit orthographischen, phonetischen, kanonischen, silbischen und morphosyntaktischen Informationen. Die Annotation prosodischer Informationen wird in Kapitel 5 beschrieben.

Für die Erstellung des CARInA sind mehrere Arbeitsschritte notwendig. Diese werden in den folgenden Kapiteln detailliert beschrieben. Eine Übersicht der Arbeitsschritte ist in Abbildung 2.6 dargestellt.

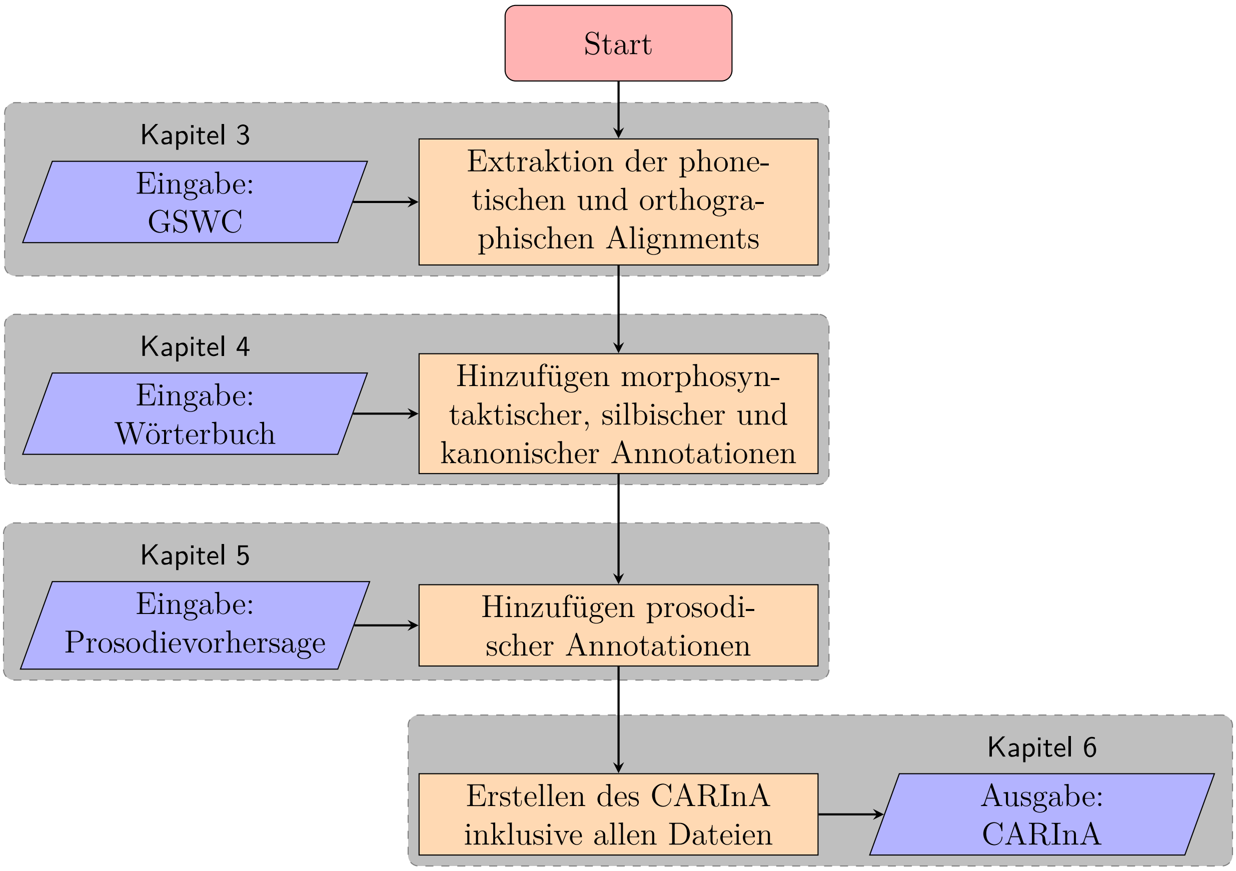


Abbildung 2.6: Programmablaufplan für die Erstellung des *Corpus of Aligned Read speech Including Annotations*. Die vier Arbeitsschritte werden in den jeweiligen Kapiteln ausführlich beschrieben.

Im Gegensatz zum GSWC ist der CARInA nicht nach Artikeln, sondern nach Personen strukturiert. Durch die automatische Erstellung des CARInA sind nicht alle enthaltenen Daten vollständig. Aus diesem Grund wurde der Korpus in zwei Teildatensätze gegliedert. Der Teildatensatz *Complete* beinhaltet alle Daten, welche zuverlässig und vollständig annotiert wurden. Der Teildatensatz *WorkInProgress* beinhaltet alle unvollständigen Daten. Die Struktur des CARInA wird in Abschnitt 6.1 beschrieben.

# *Kapitel 3*

# 3 CARInA – Orthographische und phonetische Transkription

## 3.1 Artikelübersicht

Der GSWC ist nach den gelesenen Artikeln strukturiert. Die Dateistruktur ist in Abbildung 3.1 dargestellt. Die Informationen über das Thema der Artikel ist für die meisten Anwendungen eines Korpus nicht relevant. Aus diesem Grund wurde die Tabelle `ListFolders.mat` erstellt. In dieser Tabelle ist jedem Artikel eine Nummer zugeordnet. Die Nummern wurden für die Identifikation der Artikel und die Speicherung von Variablen verwendet.

```
+ ---German Spoken Wikipedia Corpus
| + ---%c9cber_Sinn_und_Bedeutung
| | + ---aligned.swc
| | + ---audio.ogg
| | + ---audiometa.txt
| | + ---info.json
| | + ---wiki.html
| | + ---wiki.txt
| | + ---wiki.xml
| +---%c3%9f
.
.
|+---Zwinger_(Dresden)
```

Abbildung 3.1: Dateistruktur des *German Spoken Wikipedia Corpus*

## 3.2 Artikelzuweisung

Jeder Artikel enthält eine Datei mit Metadaten (`audiometa.txt`), in welcher die Sprecherinnen und Sprecher unter anderem ihren Benutzernamen und ihr Geschlecht angeben können. Mithilfe der Benutzernamen wurden die Artikel einer Person zugeordnet. Der GSWC beinhaltet Artikel von 337 Personen. Die Anzahl der gelesenen Artikel pro Person variiert zwischen 1 und 161, wobei 128 Personen mindestens 2 Artikel gelesen haben.

Für eine übersichtliche Struktur wurden die Benutzernamen zu Identifikationsnamen konvertiert. Tabelle 3.1 stellt diese Konvertierung in Abhängigkeit des angegebenen Geschlechtes dar.

Tabelle 3.1: Zuordnung der Benutzernamen zu den Identifikationsnamen in Abhängigkeit des angegebenen Geschlechts. '****' steht für eine fortlaufende Nummerierung von 0000 bis 0337.

| Geschlecht | Anzahl | Identifikationsname |
|---|---|---|
| feminin | 36 | SpeakerID****_f |
| maskulin | 267 | SpeakerID****_m |
| unbekannt | 34 | SpeakerID****_u |

Die Vorverarbeitung des Programms `createCARInA.m` besteht unter anderem aus der Erstellung der Tabelle `ListAuthors.mat`. Die Tabelle `ListAuthors.mat` enthält Informationen über die Benutzernamen, angegebenen Geschlechter, Identifikationsnamen, Anzahl gelesener Artikel und Artikelnamen für jede Person.

Die Angabe des Benutzernamens und des Geschlechts ist nicht obligatorisch für das Hochladen einer Audiodatei. Es ist somit nicht ausgeschlossen, dass eine Person Artikel unter verschiedenen Benutzernamen erstellt hat und somit mehrere Identifikationsnamen erhält. Einige offensichtliche Schreibfehler wurden überprüft und korrigiert. Eine Person besitzt beispielsweise die Benutzernamen `Stollii` und `Stolliii`. Jedem Artikel ohne angegebenen Benutzernamen wird ein neuer Identifikationsname zugewiesen.

Die Angabe des Geschlechts wurde exemplarisch für jede Person mit einem subjektiven Hörtest überprüft. Bei Unstimmigkeiten zwischen dem subjektiven Hörtest und der Angabe aus der Metadatendatei wurde der Identifikationsname nicht korrigiert. Für geschlechtsspezifische Anwendungen werden bis auf einige Ausnahmen die Geschlechter vorgeschlagen, welche nach Tabelle 3.1 im letzten Zeichen des Identifikationsnamen codiert sind. Die Ausnahmen sind in Tabelle 3.2 aufgelistet.

Tabelle 3.2: Personen, bei welchen Unstimmigkeiten zwischen dem angegebenen Geschlecht aus den Metadaten und dem subjektiven Höreindruck vorliegen

| Identifikationsname | Geschlecht (angegeben) | Geschlecht (Höreindruck) |
|---|---|---|
| SpeakerID0023_f | feminin | maskulin |
| SpeakerID0304_u | – | maskulin |
| SpeakerID0305_u | – | feminin |
| SpeakerID0306_u | – | maskulin |
| SpeakerID0307_u | – | maskulin |
| SpeakerID0308_u | – | feminin |
| SpeakerID0309_u | – | maskulin |
| SpeakerID0310_u | – | maskulin |
| SpeakerID0311_u | – | feminin |
| SpeakerID0312_u | – | maskulin |
| SpeakerID0313_u | – | maskulin |
| SpeakerID0314_u | – | maskulin |
| SpeakerID0315_u | – | maskulin |
| SpeakerID0316_u | – | feminin |
| SpeakerID0317_u | – | maskulin |
| SpeakerID0318_u | – | maskulin |
| SpeakerID0319_u | – | maskulin |
| SpeakerID0320_u | – | maskulin |
| SpeakerID0321_u | – | maskulin |
| SpeakerID0322_u | – | maskulin |
| SpeakerID0323_u | – | maskulin |
| SpeakerID0324_u | – | maskulin |
| SpeakerID0325_u | – | maskulin |
| SpeakerID0326_u | – | maskulin |
| SpeakerID0327_u | – | maskulin |
| SpeakerID0328_u | – | maskulin |
| SpeakerID0329_u | – | maskulin |
| SpeakerID0330_u | – | maskulin |
| SpeakerID0331_u | – | maskulin |
| SpeakerID0332_u | – | maskulin |
| SpeakerID0333_u | – | maskulin |
| SpeakerID0334_u | – | maskulin |
| SpeakerID0335_u | – | maskulin |
| SpeakerID0336_u | – | feminin |
| SpeakerID0337_u | – | maskulin |

## 3.3 Orthographische und phonetische Alignments

Für das Auslesen der orthographischen und phonetischen Alignments wurden die XML-Dateien `aligned.swc` mithilfe der Funktion `xml2struct()` [25] eingelesen. Iterativ wurden alle Sätze extrahiert und die Alignments der Wörter und der Phoneme gespeichert.

Für jeden Artikel wurde eine Liste mit dem Namen `AnnotationSWC_*.mat` gespeichert. Das '`*`' steht für die jeweilige Artikelnummer. Für jeden Satz eines Artikels enthält die zugehörige Liste einen Eintrag. Die Namen der Sätze lauten:
`SpeakerID****_*_article****_sentence****`

Der erste Teil (`SpeakerID****_*`) beinhaltet den Identifikationsnamen der Person, der zweite Teil (`_article****`) beinhaltet die Nummerierung des Artikels und der dritte Teil (`_sentence****`) beinhaltet die Nummerierung der Sätze innerhalb des Artikels. Alle Nummerierungen bestehen aus vier Ziffern und enthalten eventuell vorangestellte Nullen.

In der Liste ist für jeden Satz des zugehörigen Artikels eine Tabelle hinterlegt. Die Tabelle beinhaltet die geschriebenen Wörter, ihre Start- und Endabtastwerte, die zugehörigen Phoneme und deren Start- und Endabtastwerte. Wie in Abschnitt 2.2.3 ausgeführt, sind nicht alle Sätze vollständig auf orthographischer und phonetischer Ebene aligniert. Aus diesem Grund sind einige Tabellen unvollständig. Die Tabelle 3.3 zeigt ein Beispiel für diese Tabelle mit allen möglichen Alignmentzuständen, welche für ein Wort innerhalb eines Satzes vorkommen können.

Tabelle 3.3: Beispielsatz „Die Sonne dreht durch.“ mit allen im *German Spoken Wikipedia Corpus* vorkommenden Alignmentzuständen. Vollständiges Alignment des Wortes „Die“, ausschließlich orthographisches Alignment des Wortes „Sonne“, kein Alignment des Wortes „dreht“ und ausschließlich phonetisches Alignment des Wortes „durch“. Satzzeichen werden nicht aligniert.

| **Worteigenschaften** | | | **Phonemeigenschaften** | | |
|---|---|---|---|---|---|
| Wort | Startabtastwert | Endabtastwert | Phonem | Startabtastwert | Endabtastwert |
| Die | 186810 | 186910 | d | 186800 | 186850 |
| | | | i: | 186850 | 187050 |
| Sonne | 186910 | 187800 | | | |
| dreht | | | | | |
| durch | | | d | 201120 | 201150 |
| | | | U | 201150 | 201180 |
| | | | 6 | 201180 | 201350 |
| | | | C | 201350 | 201480 |
| . | | | | | |

Wie in Abschnitt 2.2.3 beschrieben, wurden die orthographischen und die phonetischen Alignments mit unterschiedlichen Programmen erstellt. Hierdurch kommt es zu zeitlichen Unterschieden der Wortgrenzen. Nach dem subjektiven Eindruck der Ersteller des GSWC sind die phonetischen Alignments des Korpus genauer als die orthographischen Alignments [5]. Da außerdem unterschiedliche Wortgrenzen in der Anwendung zu Problemen führen können, wurden die orthographischen Alignments nicht in den CARInA übernommen.

Die orthographischen Wortgrenzen wurden dennoch zur Selektion der qualitativ hochwertigen phonetischen Alignments genutzt. Aus dem GSWC wurden 55 473 vollständig phonetisch alignierte Sätze extrahiert. Dies entspricht 129:47 Stunden Sprachmaterial. In Abbildung 3.2 sind die maximalen Differenzen der bestimmten Wortgrenzen auf orthographischer und phonetischer Ebene dieses Teildatensatzes dargestellt.

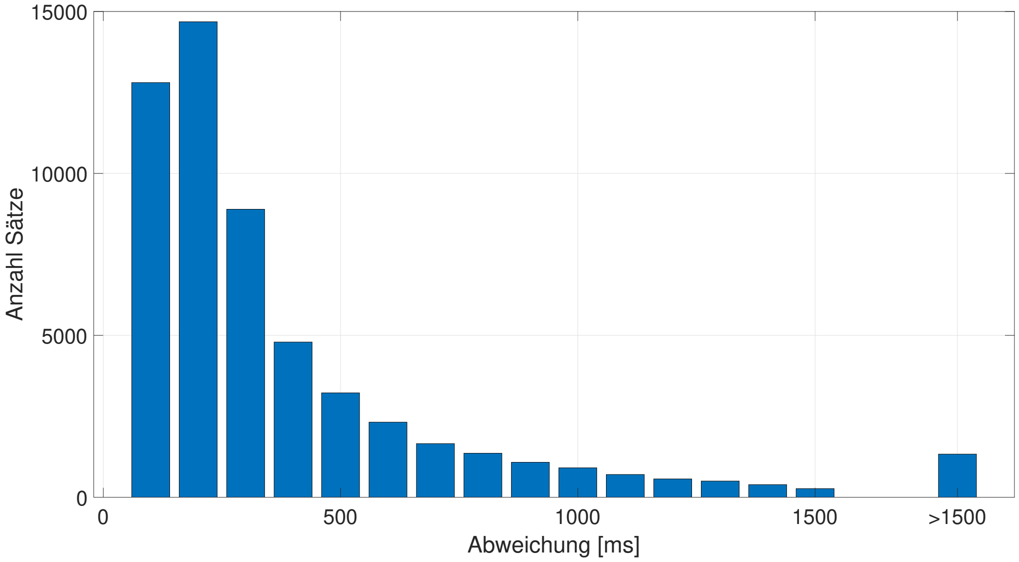


Abbildung 3.2: Anzahl der vollständig phonetisch alignierten Sätze mit einer bestimmten maximalen Differenz zwischen den errechneten Wortgrenzen. Die Wortgrenzen wurden mit dem *SailAlign*-Algorithmus und mit MAUS bestimmt.

Eine subjektive Untersuchung ergab eine deutlich erhöhte Fehlerrate der phonetischen Alignments in Sätzen mit Differenzen der Wortgrenzen von mehr als 1500 ms. Diese großen Differenzen sind zu einem signifikanten Anteil auf unterschiedliche Inhalte der Sprachaufnahme und des zugehörigen Artikels zurückzuführen.

Das Ende des fünften Kapitels aus dem Artikel `Geschichte_Aubings` ist ein Beispiel hierfür. Der Inhalt des geschriebenen Artikels ist „Beide emigrierten später (siehe Chemische Fabrik Aubing)“. Der Inhalt der Audioaufnahme ist „Beide emigrierten später. Der Artikel Geschichte Aubings Kapitel 1803/1818 bis 1942 selbstständige Gemeinde wurde vom Benutzer Blik am achten Februar 2015 gesprochen. Eine unverbindliche Zusammenfassung der Lizenz *Creative Commons Attribution-ShareAlike 3.0 Unported*".

Aus diesem Grund wurden Sätze mit einer Differenz der Wortgrenzen von mehr als 1500 ms als phonetisch unvollständig markiert.

Der verbleibende phonetisch vollständig alignierte Anteil beinhaltet 54 141 Sätze mit 124:15 Stunden Sprachmaterial. Dieser Anteil bildet die Grundlage des Teilkorpus *Complete*.

# 4 CARInA – Kanonische und morphosyntaktische Transkription

Eine unkomplizierte Möglichkeit, wortbezogene Informationen zu erhalten, ist die Nutzung von Wörterbüchern (vergleiche Abschnitt 1.2).

Wörterbücher wurden bis in die 1990er-Jahre nahezu ausschließlich von einer kleinen Personengruppe mit lexikographischen Fachkenntnissen erstellt [49]. Diese Art der Anfertigung eines Wörterbuches ist sehr zeit- und kostenintensiv. Mit der Entwicklung des Internets entstand die Möglichkeit, Wörterbücher in Gemeinschaftsprojekten zu erstellen [49].

Der Vorteil eines Gemeinschaftsprojektes sind die niedrigen Produktions- und Nutzungskosten. Oftmals stehen die Projekte kostenlos zur Verfügung. Auch sind gemeinschaftlich erstellte Wörterbücher häufig deutlich umfangreicher als kommerzielle Wörterbücher [72]. Der Nachteil gemeinschaftlich erstellter Wörterbücher ist die deutliche höhere Wahrscheinlichkeit, Fehlinformationen zu beinhalten [49], [72].

Da die Konvertierung eines graphemischen Wortes in die kanonische Realisierung sprachwissenschaftliche Kompetenzen erfordert, ist die Erstellung von Aussprachewörterbüchern eine besondere Herausforderung. Es existieren verschiedene digitale deutschsprachige Wörterbücher.

Beispiele für kommerziell nutzbare deutschsprachige Aussprachewörterbücher sind der Duden [16] und EasyPronounciation [6]. Beispiele für gemeinschaftlich entstandene und kostenlos zugängliche deutschsprachige Aussprachewörterbücher sind das Wiktionary [2] und HowToPronounce [67].

Das umfangreichste kostenlos erhältliche Aussprachewörterbuch deutscher Sprache ist das Wiktionary [45], [49]. Mithilfe des Wörterbuches Wiktionary wurden morphosyntaktische, kanonische und silbische Informationen dem CARInA hinzugefügt.

## 4.1 Das Wiktionary

Das Wiktionary wird von der Wikimedia Foundation betrieben und ist ein kostenlos zugängliches Wörterbuch für mehrere Sprachen [107], [108]. Das Wiktionary wurde als Begleiter von Wikipedia entwickelt. Es ist ein Gemeinschaftsprojekt und kann von jeder Person bearbeitet und erweitert werden [108]. Das Wort „Wiktionary“ ist ein Kofferwort aus den Wörtern *wiki* (schnell) und *dictionary* (Wörterbuch) [49].

Das Projekt Wiktionary wurde 2002 als englischsprachiges Wörterbuch gestartet [49]. 2003 wurde das Projekt auf die Sprachen Französisch, Polnisch, Finnisch, Deutsch und Italienisch ausgeweitet. Im Jahr 2010 existierten 104 000 deutschsprachige Einträge im Wiktionary [49].

Zum jetzigen Zeitpunkt (2021) enthält das deutschsprachige Wiktionary 126 503 Einträge mit Grundformen (flektierte Formen nicht mitgezählt) [2]. Durch das Einbinden fremdsprachlichen Materials kommt es zu Überschneidungen in den Wiktionarys [107]. Im deutschsprachigen Wiktionary befinden sich 76 % deutsche, 6 % englische und 3 % tschechische Artikel [100]. Die verbleibenden 15 % verteilen sich auf über 200 weitere Sprachen (Dezember 2020) [100].

Ein Artikel aus dem Wiktionary kann verschiedene Informationen zu einem Wort aufweisen. Der Artikel kann beispielsweise die Wortart, die Worttrennung, die Aussprache des Internationalen Phonetischen Alphabets (IPA), Wortbedeutungen, Synonyme, sinnverwandte Wörter, Gegenwörter und Beispiele enthalten [72].

Artikel aus dem Wiktionary können über die offizielle Internetseite [2] heruntergeladen werden. Von dem Herunterladen großer Datenmengen über diese Seite wird aufgrund des entstehenden Datenverkehrs abgeraten. Über einen Server der Umeå-Universität in Schweden [103] kann eine vollständige Kopie des Wiktionarys im XML-Format heruntergeladen werden. Die dort gespeicherten Kopien werden monatlich aktualisiert [103].

Durch die kontinuierliche Bearbeitung und Erweiterung ist eine aussagekräftige Validierung des Wiktionarys schwierig. Im Jahr 2010 wurde die deutschsprachige Lautschrift des Wiktionarys mit einem aus dem GlobalPhon [89] extrahierten Datensatz verglichen. Hierbei waren nur 28 % der kanonischen Realisierungen identisch [85]. Seit dem hat sich der Umfang des deutschsprachigen Wiktionarys jedoch verzehnfacht [2].

Für das Wiktionary existiert keine umfangreiche Evaluation [68]. Dennoch ist das Wiktionary in aktuellen Sprachforschungen präsent. Beispiele hierfür sind der Wiktionary Matcher [62], welcher das Wiktionary als lexikalische Wissensquelle nutzt und der Yawipa Wiktionary Parser [104].

Der subjektive Eindruck einer stichprobenhaften Untersuchung des Wiktionarys deutet auf eine geringe Fehlerrate hin. Ein Beispiel für ein Wort mit fehlerhaften Informationen ist die Lautschrift des Wortes „dreiundzwanzigstes“, im Wiktionary als [ˈdʁaɪ̯ʔʊntˈt͡svant͡sɪçstəs] angegeben. Auf der ersten Silbe des Wortes liegt jedoch eine sekundäre, und keine primäre Betonung. Die richtige Schreibweise des Wortes in Lautschrift ist [ˌdʁaɪ̯ʔʊntˈt͡svant͡sɪçstəs]. Ein weiteres Beispiel für eine Fehlinformation ist die Silbentrennung des Wortes „darauf“. Diese ist als „da · r · auf“ angegeben. Die richtige Silbentrennung ist „da · rauf“.

## 4.2 Wörterbücher – Erstellung

Für das Hinzufügen von morphosyntaktischen, kanonischen und silbischen Informationen zum CARInA wurden drei Wörterbücher erstellt. Diese sind aus der Kopie des Wiktionarys der Umeå-Universität vom 20.01.2021 entstanden.

Vor der Erstellung des CARInA wurden die Wörterbücher generiert. Hierfür wurde die Kopie des Wiktionarys heruntergeladen [103] und in dem Ordner `Dictionary` abgelegt. Zuerst wurde das Skript `AddToDict.m` ausgeführt, welches die manuelle Erweiterung der Wörterbücher formatiert. Als zweites wurde das Skript `createDict.m` ausgeführt, welches die Informationen aus der Kopie des Wiktionarys ausliest und gemeinsam mit den manuell erstellten Informationen drei Wörterbücher generiert. Der Programmablaufplan für das Skript `createDict.m` ist in Abbildung 4.2 dargestellt und ist nachfolgend erläutert.

### 4.2.1 Wortbezogenen Informationen des Wiktionarys

Die Lautschrift der Artikel des Wiktionarys basiert auf dem IPA. Für eine einfachere maschinelle Verarbeitung wurde eine Konvertierung der Symbole in das SAMPA geschaffen. Es wurde die Textdatei `InputIpa2Sampa.txt` angefertigt. Diese enthält für alle deutschen und einige fremdsprachlichen Phoneme und Zeichen (Diakritika, Supresegmentalia et cetera) die Symbole des IPA und des SAMPA.

Zunächst wurde die Datei `InputIpa2Sampa.txt` und die Kopie des Wiktionarys geöffnet. Mit der MATLAB-Funktion `containers.Map` wurde ein IPA zu SAMPA Wörterbuch erstellt. Die Kopie des Wiktionarys ist mit 1.45 Gigabyte zu groß zum Einlesen. Deswegen wurden die Artikel einzeln eingelesen. Von den 1 061 360 Artikeln besteht ein Teil aus Meta-Artikeln (Hilfsseiten, Vorlageseiten, Benutzerseiten et cetera) [72]. Aus jedem nutzbaren Artikel wurde das Wort ausgelesen. Soweit vorhanden wurden zu dem Wort dessen graphemische Silbifizierung, die Lautschrift und die Wortart ausgelesen und gespeichert.

Aus der Kopie des Wiktionary sind drei Wörterbücher entstanden. In diesen kann jeweils für ein Wort die Lautschrift (SAMPA), die Wortartenzugehörigkeit oder die graphemische Silbifizierung ausgelesen werden. Die Anzahlen der Wörter in den jeweiligen Wörterbüchern können der zweiten Spalte der Tabelle 4.1 entnommen werden. Aus dem Wiktionary wurden 755 728 Wörter ausgelesen, welche in allen drei Wörterbüchern enthalten sind.

Tabelle 4.1: Anzahl der Wörter in den einzelnen Wörterbüchern, extrahiert aus dem Wiktionary und insgesamt mit den manuell hinzugefügten Wörtern.

| Wörterbuch | Wörter Wiktionary | Wörter gesamt |
|---|---|---|
| Lautschrift | 764 185 | 765 847 |
| Wortart | 915 648 | 917 303 |
| Silbifizierung | 825 888 | 827 536 |

### 4.2.2 Manuelle Erweiterung

Die wortbezogenen Informationen aus dem Wiktionary wurden dem CARInA hinzugefügt. Diese decken jedoch nicht den Inhalt des gesamten Korpus ab. Insgesamt enthält der CARInA 102 766 verschiedene Wörter und Zeichen, welche in mindestens einer der drei Kategorien nicht im Wiktionary enthalten sind. Komposita, Eigennamen und Zahlen stellen die drei großen Gruppen der fehlenden Wörter dar.

Von den 54 141 Sätzen der Grundlage des Teilkorpus *Complete* wurden 17 027 Sätze vollständig mit kanonischen, silbischen und wortartenbezogenen Informationen annotiert. Für eine Erweiterung des Teilkorpus *Complete* wurden die Wörterbücher ergänzt.

Die ausgeschriebenen Zahlen von „eins“ bis „neunundneunzig“ wurden erzeugt. Für jede Zahl wurden zusätzlich die Endungen „-st, -ste, -stem, -sten, -stens, -ster, -stes“ hinzugefügt (Beispiel: zweiundsechzigstes). Des Weiteren wurden die Jahreszahlen von „1100“ bis „1999“ hinzugefügt.

Die Erweiterung des Wörterbuches durch Eigennamen und Komposita ist manuell möglich. Aufgrund der großen Anzahl fehlenden Wörter wurde eine Auswahl getroffen. Für jeden Artikel wurde eine Liste mit dem Namen `NotInG2P_*.mat` erstellt, wobei ’`*`’ der Artikelnummer entspricht. In dieser Liste ist enthalten, wie viele und welche Wörter eines Satz nicht in allen Wörterbüchern enthalten sind.

Aus diesen Informationen wurden alle Wörter extrahiert, welche als einziges Wort in einem Satz mit ansonsten vollständiger Information fehlen. Diese Wörter und ihre auftretenden Häufigkeiten wurden in `NotInG2Pfa.txt` gespeichert.

Das am häufigsten auftretende Wort in `NotInG2Pfa.txt` ist somit das profitabelste. Es wird direkt die korrespondierende Anzahl an Sätzen zu dem Teilkorpus *Complete* hinzugefügt. Durch das Hinzufügen des Wortes zu den Wörterbüchern ändert sich die Liste `NotInG2Pfa.txt`, da ohne das Wort andere Wörter die einzigen fehlenden Wörter in einem Satz sein können.

Das effizienteste Vorgehen ohne Berücksichtigung der Rechenzeit ist das iterative Hinzufügen des am häufigsten auftretenden Wortes in `NotInG2Pfa.txt` zu den Wörterbüchern mit einer Neuberechnung von `NotInG2Pfa.txt` nach jeder Ergänzung.

Die Zeit für eine Neuberechnung liegt bei ca. 120 Minuten. Aus diesem Grund wurde nicht nach jedem hinzugefügten Wort eine Neuberechnung durchgeführt. Die Abbildung 4.1 zeigt die Anzahl an entstehenden Sätzen mit vollständiger Information in Abhängigkeit der Anzahl zuzufügender Wörter. Diese Kurve kann sich durch jedes hinzugefügte Wort massiv verändern. Sie ist hier für den initialen Zustand und für den aktuellen Zustand dargestellt. Zur Erreichung des aktuellen Zustands wurden über mehrere Neuberechnungen von `NotInG2Pfa.txt` insgesamt 237 Wörter manuell zu dem Wörterbuch hinzugefügt.

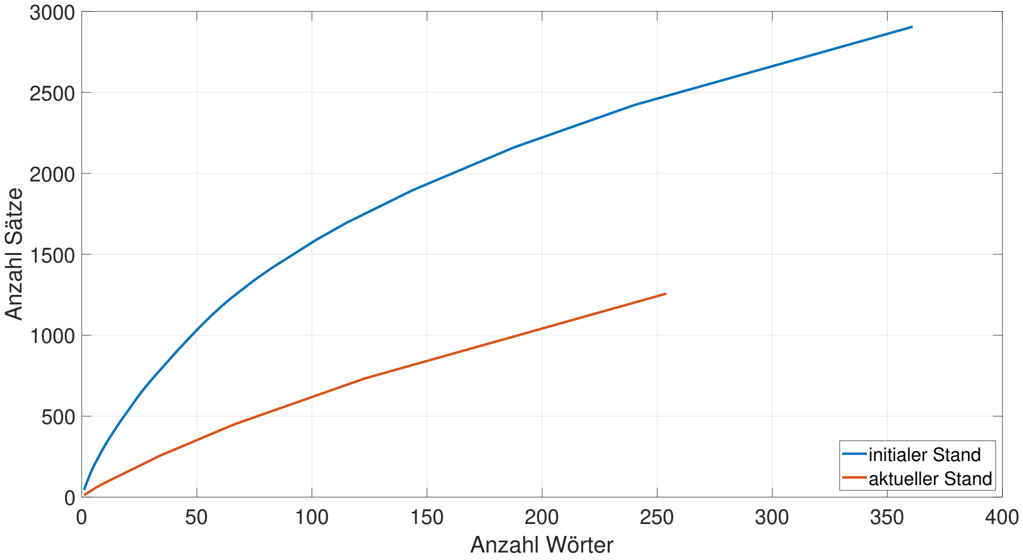


Abbildung 4.1: Anzahl an Sätzen, welche durch das Hinzufügen der Anzahl an Wörtern zu dem Teilkorpus *Complete* hinzugefügt werden. Die Abbildung wurde auf denjenigen Anteil an Wörtern beschränkt, welche in mindestens vier Sätzen als einziges Wort in einem Satz mit ansonsten vollständiger Information fehlen.

Insgesamt wurden 1846 Wörter erstellt und den Wörterbüchern hinzugefügt. Bereits enthaltene Wörter wurden überschrieben. Die finalen Anzahlen der Wörter in den Wörterbüchern können der dritten Spalte der Tabelle 4.1 entnommen werden.

Durch das Hinzufügen der Zahlenworte und Wörter in die Wörterbücher wurde die Grundlage des Teilkorpus *Complete* mit vollständigen orthographischen, phonetischen, kanonischen, silbischen und wortartbezogenen Informationen auf 19 613 Sätze erhöht.

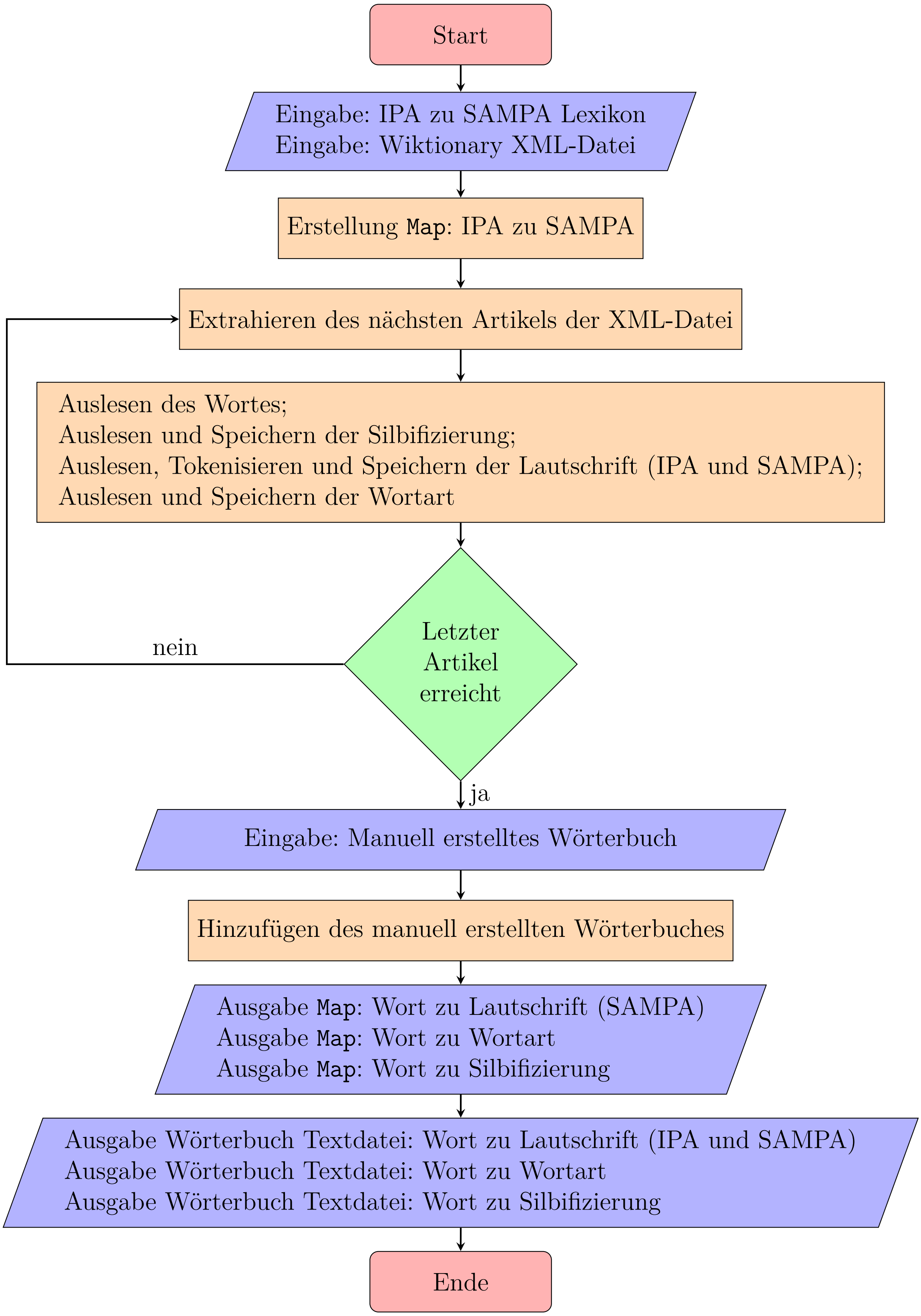


Abbildung 4.2: Programmablaufplan `createDict.m` für die Erstellung der Wörterbücher für Lautschrift, Wortart und Silbifizierung

### 4.2.3 Struktur

Aus den Informationen des Wiktionarys und den nachträglich hinzugefügten Erweiterungen sind drei Wörterbücher entstanden. Die Wörterbücher liegen als Textdatei vor, sind alphabetisch sortiert und mit simplen Methoden maschinell einlesbar. Alle Wörterbücher liegen im *Universal Coded Character Set Transformation Format* (UTF-8) vor.

Das Wörterbuch `Dict_graphem2phonem.txt` besteht aus drei Spalten, welche durch einen Tabulator getrennt werden. Die erste Spalte beinhaltet das graphemische Wort. Die zweite Spalte beinhaltet die kanonische Realisierung des Wortes, codiert in IPA. Die dritte Spalte beinhaltet die kanonische Realisierung des Wortes, codiert in SAMPA. Die Lautschrift des Wortes liegt in tokenisierter Form vor. Einzelne Phoneme der zweiten und dritten Spalte sind durch Leerzeichen voneinander getrennt.

Das Wörterbuch `Dict_syllabification.txt` besteht aus zwei Spalten, welche durch einen Tabulator getrennt werden. Die erste Spalte beinhaltet das graphemische Wort. Die zweite Spalte beinhaltet die graphemische Silbentrennung. Ein Wörterbucheintrag kann wie bei der Wortverbindung „et cetera“ aus mehreren Wörtern bestehen. Aus diesem Grund werden die einzelnen Wörter durch ein Leerzeichen und die Silben innerhalb eines Wortes durch einen ’ · ’ getrennt.

Das Wörterbuch `Dict_PartOfSpeech.txt` besteht aus zwei Spalten, welche durch einen Tabulator getrennt werden. Die erste Spalte beinhaltet das graphemische Wort. Die zweite Spalte beinhaltet die Wortart. Tabelle 4.2 beinhaltet eine Liste der Wortarten und der verwendeten Etiketten, welche den englischen Übersetzungen entsprechen.

Das Bestimmen der Wortart ohne Kontext ist für einen Teil deutscher Wörter nicht möglich. Das Wort „sein“ kann beispielsweise ein Verb oder ein Possessivpronomen sein. Und das Wort „weiß“ kann ein Adjektiv oder ein flektiertes Verb sein. In solchen Fällen sind im Wiktionary beide Wortarten zu finden. In das Wörterbuch `Dict_PartOfSpeech.txt` wurde die erste im Wiktionary genannte Wortart übernommen. Hierdurch kann es zu fehlerhaften Bestimmungen der Wortart kommen.

Tabelle 4.2: Wortarten des Wörterbuches `Dict_graphem2phonem.txt` mit zugehörigen Etiketten

| Wortart | Etikett |
|---|---|
| Abkürzung | abbreviation |
| Adjektiv | adjective |
| Adverb | adverb |
| Affix | affix |
| Artikel | article |
| Buchstabe | letter |
| Demonstrativpronomen | demonstrative pronoun |
| Eigenname | proper noun |
| Indefinitpronomen | indefinite pronoun |
| Interjektion | interjection |
| Interrogativadverb | interrogative adverb |
| Interrogativpronomen | interrogative pronoun |
| Konfix | konfix |
| Konjunktion | conjunction |
| Konjunktionaladverb | conjunctive adverb |
| Kontraktion | contraction |
| Lokaladverb | local adverb |
| Modaladverb | modal adverb |
| Numerale | number |
| Onomatopoetikum | onomatopoeia |
| Partikel | particle |
| Personalpronomen | personal pronoun |
| Possessivpronomen | possessive pronoun |
| Präposition | preposition |
| Pronomen | pronoun |
| Pronominaladverb | pronominal adverb |
| Redewendung | phrase |
| Reflexivpronomen | reflexive pronoun |
| Relativpronomen | relative pronoun |
| Reziprokpronomen | reciprocal pronoun |
| Schriftzeichen | characters |
| Sprichwort | saying |
| Subjunktion | subjunction |
| Substantiv | noun |
| Symbol | symbol |
| Temporaladverb | temporal adverb |
| Umschrift | transcription |
| Verb | verb |
| Wortverbindung | compound |
| Zahlzeichen | number |
| Zirkumposition | adposition |

## 4.3 Wörterbücher – Nutzung

Ein systematisch auftretendes Problem bei der Bestimmung von Wortarten ohne Kontext ist das Anfangswort eines jeden Satzes. Durch die Großschreibung wird signalisiert, dass dieses Wort ein Substantiv ist. Dies trifft nicht für jedes Anfangswort zu. Außerdem existieren Wörter, welche sowohl ein Substantiv als auch eine andere Wortart sein können. Ein Beispiel hierfür ist das Wort „Heilige“. In „Heilige sind die gottesfürchtigen ...“ ist es ein Substantiv, in „Heilige Schriften ...“ ist es ein Adjektiv.

Zur Dezimierung dieses Problems wurde eine Wortartenanalyse mit den MATLAB-Funktionen `tokenizedDocument` und `addPartOfSpeechDetails` durchgeführt. Mithilfe dieser Funktionen wurde die Wortart des ersten Wortes ausgewertet. Wird das erste Wort des Satzes als Substantiv (engl. *noun*) oder Eigenname (engl. *proper-noun*) erkannt, wird es großgeschrieben. Ansonsten wird es kleingeschrieben. Die Wortarten jedes Wortes werden gespeichert.

Diese Funktionen sind systematisch fehlerbehaftet. So werden beispielsweise die Wortarten des Satzes „Wie geht es dir?“ als „Wie [*proper-noun*] geht [*proper-noun*] es [*proper-noun*] dir [*noun*] ? [*punctuation*]“ erkannt. Außerdem werden Kontraktionen als zwei Wörter gespeichert und mit zwei Wortarten etikettiert (siehe Tabelle 4.3). Aus diesem Grund wird ausschließlich die Information der Großschreibung des ersten Wortes für den CARInA genutzt.

Für die wortbezogenen Informationen der kanonischen Aussprache, der Wortartenzugehörigkeit und der graphemischen Silbifizierung werden die drei Wörterbücher genutzt, welche in Abschnitt 4.2 ausführlich beschrieben wurden.

Die kanonische Realisierung enthält unter anderem Informationen zu der Akzentuierung eines Wortes. Weist ein Wort mehrere Silben auf, wird die primär akzentuierte Silbe gekennzeichnet. Existiert eine sekundär akzentuierte Silbe, wird auch diese gekennzeichnet. Diese Akzentuierungen existieren auch in der phonetischen Realisierung des Wortes. Die Informationen hierfür sind im GSWC nicht enthalten.

Aus diesem Grund wurden die Informationen für die Akzentuierung der phonetischen Realisierung aus der kanonischen Realisierung abgeleitet. Akzentuierungen liegen immer auf dem ersten Vokal oder Diphthong nach einem Akzentzeichen. Für die Zuweisung der Akzente wurden die entsprechenden Vokale aus der kanonischen Realisierung mit dem vorherigen und dem nachfolgenden Phonem extrahiert. Die entstehende Kombination aus drei Phonemen wurde in der phonetischen Realisierung gesucht. Da diese von der kanonischen Realisierung abweichen kann, wurde eine Liste mit sprachlichen Ausnahmen hinzugefügt.

Beispiel:

Die kanonische Realisierung des Wortes „Artikel“ ist [ˌa ʁ ˈt iː k l̩], eine mögliche phonetische Realisierung ist [aː r t ɪ k ə l].

Der primäre Akzent der kanonischen Realisierung liegt auf der zweiten Silbe. Der nachfolgende Vokal ist das [iː]. Es wird die Phonemfolge [t iː k] gesucht. In der phonetischen Realisierung ist diese nicht vorhanden. Mit der Ausnahme, dass ein [iː] auch als [ɪ] gesprochen werden kann, wird die Folge gefunden. Der primäre Akzent liegt auf dem [ɪ].

Für den sekundären Akzent wird analog vorgegangen. Eine Besonderheit ist, dass der akzentuierte Vokal [a] das erste Phonem ist. Das vorherige Phonem ist somit leer, der Akzent kann nur dem ersten Phonem zugeordnet werden.

Die entstehende Akzentuierung der phonetischen Realisierung ist [ˌaː r t ˈɪ k ə l].

Es konnten nicht alle Akzentuierungen von der kanonischen Realisierung auf die phonetische Realisierung übertragen werden. Gründe hierfür sind hauptsächlich inkorrekte phonetische Alignments und starke Akzentuierungen. Von der Grundlage des Teilkorpus *Complete* mit 19 613 Sätzen wurden 17 229 Sätze vollständig mit Akzenten versehen. Dieser Anteil des CARInA bildet den Teilkorpus *Complete*.

Für jeden Artikel wird eine Liste mit dem Namen `Annotation_*.mat` gespeichert. Das '*' steht für die jeweilige Artikelnummer. Diese Liste ist eine Erweiterung der Liste `AnnotationSWC_*.mat`. Beide Listen enthalten für jeden Satz des Artikels einen Eintrag. Die Namen der Sätze sind in in beiden Listen identisch (vergleiche Abschnitt 3.3).

Für jeden Satz ist eine Tabelle hinterlegt. Die ersten sechs Spalten der Tabelle entsprechen großteils den Spalten aus Tabelle 3.3. Unterschiedlich ist die Groß- beziehungsweise Kleinschreibung des Anfangswortes und die Akzentuierung der Phoneme. Die Spalten sieben und acht beinhalten die Wort- und Wortartinformationen der Wortartenanalyse mittels MATLAB. In den letzten drei Spalten sind die kanonische Realisierung, die Wortart und die Silbifizierung aus den Wörterbüchern enthalten. Beispielhaft ist dieser Aufbau in Tabelle 4.3 dargestellt.

Tabelle 4.3: Beispielsatz „Die Sonne dreht durch vom Scheinen.“ mit hinzugefügten Informationen. Die Ergebnisse der Wortartenanalyse von MATLAB wurden ausgewertet und gespeichert. Die Informationen aus den Wörterbüchern wurden für jedes Wort gespeichert.

| **Worteigenschaften** | | | **Phonemeigenschaften** | | | **MATLAB** | | **Wörterbücher** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Wort | Start-abtast-wert | End-abtast-wert | Pho-nem | Start-abtast-wert | End-abtast-wert | Wort | Wortart | kanonische Realisierung | Wortart | Silbifizierung |
| die | 186810 | 186910 | d | 186800 | 186850 | die | determiner | d i: | article | die |
| | | | i: | 186850 | 187050 | | | | | |
| Sonne | 186910 | 187800 | | | | Sonne | noun | " z O n @ | noun | Son · ne |
| dreht | | | | | | dreht | verb | d R e: t | verb | dreht |
| durch | | | d | 201120 | 201150 | durch | adposition | d U R C | preposition | durch |
| | | | U | 201150 | 201180 | | | | | |
| | | | 6 | 201180 | 201350 | | | | | |
| | | | C | 201350 | 201480 | | | | | |
| vom | | | f | 201480 | 201520 | von \| dem | adposition \| determiner | f O m | contraction | vom |
| | | | o: | 201580 | 201670 | | | | | |
| | | | m | 201670 | 201710 | | | | | |
| Scheinen | | | S | 201710 | 201790 | Scheinen | noun | " S aI n @ n | noun | Schei · nen |
| | | | "aI | 201790 | 201830 | | | | | |
| | | | n | 201830 | 201900 | | | | | |
| | | | @ | 201900 | 201980 | | | | | |
| | | | n | 201980 | 202050 | | | | | |
| . | | | | | | | | | | |

# 5 CARInA – Prosodische Transkription

Das Hinzufügen prosodischer Informationen zum CARInA ist aufgrund der Datenmenge nur mit automatischen Hilfsmitteln möglich.

Der Teilkorpus *Complete* wurde mit prosodischen Etiketten aus zwei Quellen versehen. Zum einen wurden Etiketten mit PyToBI generiert (vergleiche Unterabschnitt 1.2.5). Zum anderen wurde das Programm *Prosody Recognition Revisited* (PRR) genutzt, welches für deutsches Sprachmaterial erstellt wurde. Das Programm PRR wird in Abschnitt 5.3 detailliert beschrieben.

## 5.1 Systeme zur prosodischen Etikettierung

Wie in Unterabschnitt 1.2.5 beschrieben, existieren keine einheitlichen Etiketten für prosodische Informationen. Es existieren zwei Modelle, welche auf unterschiedlichen Ansätzen basieren. Ein Modell beschreibt prosodische Eigenschaften durch verschiedene unabhängige Ebenen. Die Verläufe der Intonation werden hierbei als Tonstufen aufgefasst [32]. Das System ToBI ist ein auf Ebenen basierendes System. Das andere Modell beschreibt prosodische Eigenschaften durch die Intonationskontur. Die Konturen beschreiben die Tonbewegungen in bestimmte Richtungen [32]. Das KIM ist ein auf Konturen basierendes System [59].

Im Folgenden werden das ToBI System und das KIM beschrieben und es wird ein Konvertierungsvorschlag vom KIM zu ToBI Etiketten unterbreitet.

### 5.1.1 *Tones and Break Indices*

Der Begriff ToBI wird in zwei unterschiedlichen Weisen gebraucht. Ursprünglich bezeichnete ToBI ein System zur prosodischen Annotation von amerikanischem Englisch [8]. Aufgrund des Erfolgs wurde das System an weitere Sprachen angepasst, sodass der Begriff ToBI weitgehend für das entstandene *Framework* steht [8]. In der vorliegenden Ausarbeitung wird der Terminus ToBI für das Framework verwendet, Umsetzungen für spezielle Sprachen werden genauer benannt.

### ***Mainstream American English Tones and Break Indices***

Das *Mainstream American English Tones and Break Indices* (MAE_ToBI) wurde in vier Treffen von 1991 bis 1994 entwickelt [8]. An den Treffen waren mit 25 bis 30 Personen verschiedene Interessengruppen wie Ingenieure, Psychologen und Linguisten vertreten [1], [8]. Die Intention war die Erstellung eines einfach zu bedienenden Frameworks. [8]. Das MAE_ToBI basiert auf fünf grundsätzlichen Annahmen [8]:

1. Das prosodische Muster einer Äußerung ist durch separate Ebenen darstellbar. Die Intonationskontur ist durch die lineare Interpolation von Akzenten darstellbar, während die Wortgrenzen durch hierarchische Indizierung unterscheidbar sind.
2. Die Intonationskontur ist in relative Akzenthöhen zerlegbar. H-Akzente sind höher, L-Akzente niedriger als die lokale Umgebung.
3. Der lokale Tonhöhenbereich wird durch eine Vielzahl von Effekten bestimmt, wie beispielsweise phrasale Prominenzbeziehungen. Diese Effekte werden unabhängig von der absoluten Tonhöhe angegeben, sodass ein H-Akzent in einem Teil der Äußerung niedriger sein kann als ein L-Akzent in einem anderen Teil.
4. Die Akzente für jede Phrase werden funktionell unterschieden. Die absolute Tonhöhe hängt von der Position und von der Funktion ab. Die Ausrichtung des Akzentes wird durch seine Funktion bestimmt. Ein Tonakzent wird an den betonen Segmenten der Silbe ausgerichtet, ein Grenzton an den Phrasengrenzen.
5. Der Grenzton einer Phrase beeinflusst die Intonationskontur nach dem letzten vorherigen Tonakzent bis zum Phrasenende.

Aus diesen Annahmen resultierten die beiden hauptsächlichen Kategorien des ToBI Systems: Die Akzentuierung und Phrasierung [1]. Transkripte von Sätzen mit MAE_ToBI Etiketten bestehen aus vier Ebenen [1].

**Orthographischen Ebene**
Transkription der Orthographie

**Tonale Ebene**
Transkription tonaler Ereignisse mit Tonakzenten und Grenztönen

**Phrasale Ebene**
Beschreibung der hierarchisch abgestuften Wortgrenzen

**Sonstiges**
Transkription zusätzlicher Ereignisse wie Lachen, Husten et cetera

Die MAE_ToBI Etiketten der tonalen Ebene bestehen aus zwei monotonalen Akzenten (H* und L*), fünf bitonalen Akzenten (L+H*, L*+H, L+!H*, L*+!H, H+!H) und fünf Grenztönen (L-L%, H-H%, L-H%, H-L%, H*L-H%) [1].

Die MAE_ToBI Etiketten für Wortgrenzen bestehen aus fünf hierarchisch abgestuften Etiketten von 0 bis 4 [60]. Die Funktion der Etiketten ist in Tabelle 5.1 beschrieben.

Tabelle 5.1: Indizes für Wortgrenzen des Systems MAE_ToBI

| Index | Bedeutung |
|---|---|
| 0 | ausgelöschte Wortgrenze |
| 1 | standardmäßige Wortgrenze |
| 3 | Ende einer intermediären Phrase |
| 4 | Ende einer Intonationsphrase |
| 2 | Index für zwei Ausnahmen:<br>– Rhythmischen Bruch bei tonaler Kontinuität (die Melodie setzt sich über eine Unterbrechung hinweg fort)<br>– Tonaler Bruch bei rythmischer Kontinuität (Ende einer intermediären oder Intonationsphrase ohne starke Unterbrechung) |

### *German Tones and Break Indices*

Aus den MAE_ToBI Etiketten wurde für die deutsche Sprache das System *German Tones and Break Indices* (GToBI) abgeleitet [4].

Transkripte von Sätzen mit GToBI Etiketten bestehen aus mindestens drei Ebenen: der orthographischen, der tonalen und der phrasalen Ebene. Die Ebene „Sonstiges“ des Systems MAE_ToBI ist nicht obligatorisch [4].

Die Etiketten des Systems GToBI beinhalten zwei monotonale Akzente (L* und H*) und vier bitonale Akzente (L*+H, L+H*, H+L* und H+!H*) [32]. Eine individuelle Modifizierung der sechs Akzente ist möglich [32]. Die Modifizierung erfolgt durch Diakritika (vergleiche Tabelle 5.2). Das System GToBI beinhaltet drei Grenztöne für intermediäre Phrasengrenzen (L-, H- und !H-), vier Grenztöne für Intonationsphrasengrenzen (L-%, H-%, L-H% und H-^H%) und ein initialer Grenzton (%H) [32]. Eine Übersicht der Intonationsverläufe ausgewählter Tonakzente des Systems GToBI ist in Abbildung 5.1 dargestellt.

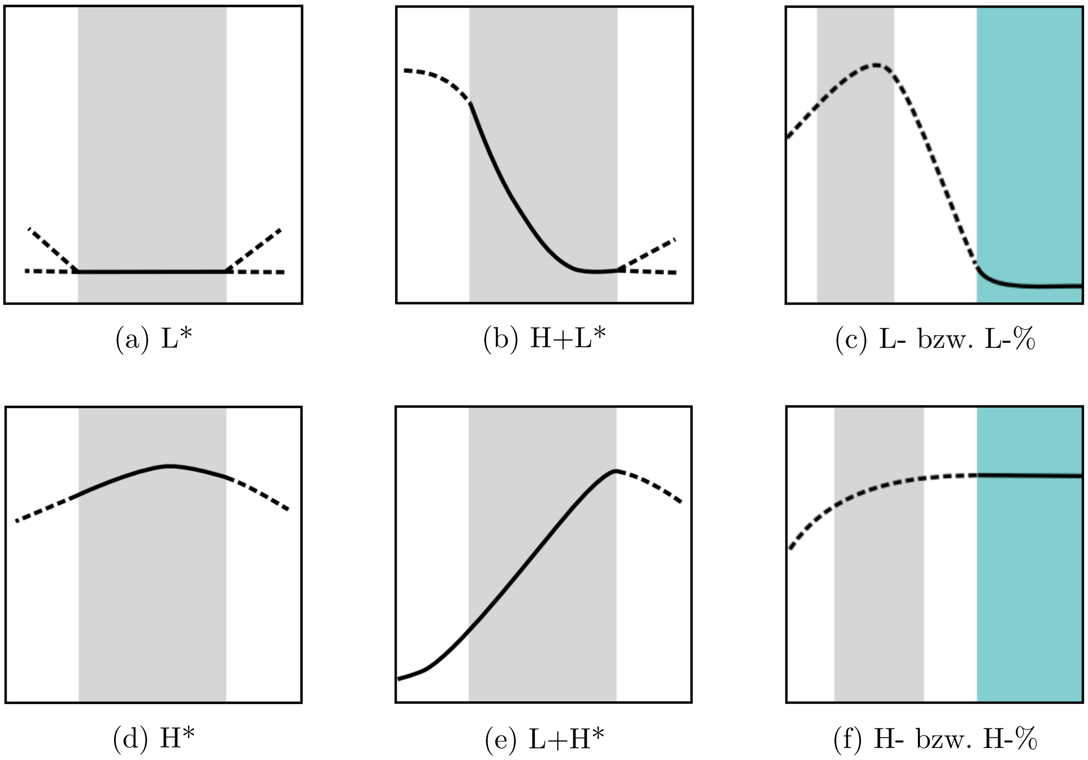

(a) L*
(b) H+L*
(c) L- bzw. L-%
(d) H*
(e) L+H*
(f) H- bzw. H-%

Abbildung 5.1: Ausgewählte Tonakzente und Grenztöne des Systems GToBI, entnommen aus [33]. Zu den Tonakzenten ist die Grundfrequenz über der Zeit aufgetragen. Die akzentuierte Silbe ist grau, der Grenzton ist blau markiert.

Mithilfe von Diakritika können zusätzliche Informationen wie der Fall von einer hohen in eine mittlere Stimmlage (H+!H*) beschrieben werden. Eine Übersicht der im System GToBI verwendeten Diakritika ist in Tabelle 5.2 gegeben.

Tabelle 5.2: Diakritika des Systems GToBI

| Diakritikum | Bedeutung |
|---|---|
| * | Zielpunkt der akzentuierten Silbe |
| + | Verbindung von zwei Zielpunkten |
| - | Grenzton einer intermediären Phrase |
| % | Grenzton einer Intonationsphrase |
| ! | Abfall (engl. *Downstep*) eines H Tones bei Akzenten und Grenztönen |
| ˆ | Anstieg (engl. *Upstep*) eines H Tones bei Akzenten und Grenztönen |
| ? | Unsicherheit eines Etiketts (Anwendung nach dem Etikett) |
| $ | Musterbeispiel eines Etiketts (Anwendung nach dem Etikett) |

Im System GToBI existieren vier Etiketten für phrasale Grenzen. Die Indizes 3 und 4 entsprechen in ihrer Funktion den Etiketten des MAE_ToBI [32]. Der Index 2 aus dem MAE_ToBI wurde unterteilt zu 2r und 2t. Eine tabellarische Übersicht der Etiketten für phrasale Grenzen ist in Tabelle 5.3 dargestellt. Die Indizes 0 und 1 aus dem MAE_ToBI System finden im GToBI System keine Anwendung [32].

Tabelle 5.3: Pausenindizes des Systems GToBI

| Index | Bedeutung |
|---|---|
| 3 | intermediäre Phrasengrenze |
| 4 | Intonationsphrasengrenze |
| 2r | rhythmischer Bruch bei tonaler Kontinuität |
| 2t | tonale Unterbrechung bei rythmischer Kontinuität |

Abgeleitet aus dem System GToBI wurde das System GToBI light. Dieses Modell entstand mit der Motivation, ein primär phonologisches Symbolinventar zu schaffen [47]. Das System GToBI light beinhaltet die Zusammenlegung der tonalen und der phrasalen Ebene [47]. Alle Etiketten des Systems GToBI light werden auf derselben Ebene annotiert. Eine Übersicht aller Etiketten ist in Tabelle 5.4 gegeben. Das Diakritikum für die Verbindung zweier Zielpunkte (+) wird nicht verwendet.

Tabelle 5.4: Etiketten des Systems GToBI light

| Symbol | Bedeutung |
|---|---|
| - | intermediäre Phrasengrenze |
| % | Intonationsphrasengrenze |
| H% | hoher Grenzton |
| -? | unsicher, ob intermediäre Phrasengrenze vorliegt |
| %? | unsicher, ob intermediäre oder Intonationsphrasengrenze vorliegt |
| H%? | unsicher, ob ein hoher Grenzton oder eine Intonationsphrasengrenze vorliegt |
| H*L | hoher Zielpunkt, fallend |
| L*H | niedriger Zielpunkt, steigend |
| H* | hoher Zielpunkt |
| ..H | hoher gleichbleibender Ton |
| L* | tiefer Zielpunkt |
| ..L | tiefer gleichbleibender Ton |
| *? | unsicher, ob Akzentuierung vorliegt |
| x? | unsicher, welches Etikett x vorliegt |

### 5.1.2 Kieler Intonationsmodell

Konturbasierte Systeme erstellen eine möglichst detaillierte Darstellung des Intonationsverlaufes über alle Silben hinweg [32]. Auf diese Art werden seit den 1960er-Jahren prosodische Merkmale im englischsprachigen Raum beschrieben [18].

Das KIM wurde Anfang der 1990er-Jahre in Kiel entwickelt [59]. Das Modell wurde auf der Basis gelesener Sprache erarbeitet und später auf spontansprachliche Äußerungen erweitert [59]. Das KIM verfolgt drei Ziele:

1. Die adäquate Beschreibung von mikro- und makroprosodischen Ereignissen. Makroprosodische Ereignisse sind von der sprechenden Person beabsichtigte Melodiemuster. Mikroprosodische Ereignisse werden nicht von der sprechenden Person intendiert und oftmals nicht wahrgenommen. Aus diesem Grund müssen sie zur Erstellung der Intonationskontur herausgefiltert werden [59].

2. Die Einordnung der Melodiemuster in ein phonologisches System. Die Melodiemuster basieren auf konkreten Signaleigenschaften. Für einen Zuhörer werden Intonationsmuster jedoch unabhängig von den lokalen Ausprägungen der Konturen als suprasegmentale Information wahrgenommen [59].

3. Die Einbindung syntaktischer, semantischer und pragmatischer Informationen, welche in Zusammenhang mit der prosodischen Erscheinung stehen. Beispielsweise liegt am Ende eines klaren, nüchternen Aussagesatzes ein Absinken des Intonationsverlaufes vor, während bei einem Fragesatz die Grundfrequenz zum Ende hin ansteigt [59].

#### Aufbau

Der Aufbau einer Kontur des KIM gliedert sich in drei Teile. Eine Übersicht des Aufbaus ist in Abbildung 5.2 dargestellt. Der erste Teil beinhaltet eine phraseninitiale Kontur bis zur ersten Akzentkontur. Eine Phrase beginnt meist mit einer Reihe von unakzentuierten Silben. In dieser Zeit kann die Grundfrequenz der Person ansteigen, abfallen oder konstant bleiben [59]. Der zweite Teil besteht aus den Akzentkonturen, welche durch Konkatenationskonturen verbunden werden. Eine Akzentkontur beinhaltet einen frühen, mittleren oder späten Gipfel, ein frühes oder spätes Tal oder eine ebene Kontur. Konkatenationskonturen können ohne Einbuchtung, mit leichter bis mittelstarker Einbuchtung oder mit Einbuchtung bis an die untere Grenze der Sprechstimme vorliegen [59]. Der dritte Teil besteht aus der phrasenfinalen Kontur nach der letzten Akzentkontur. Phrasenfinale Konturen unterscheiden sich durch die vorhergehende Akzentkontur. Nach Gipfelkonturen und ebenen Konturen können stetige und pseudoterminale Konturen auftreten. Nach Talkonturen können verschieden stark ansteigende Konturen auftreten [59].

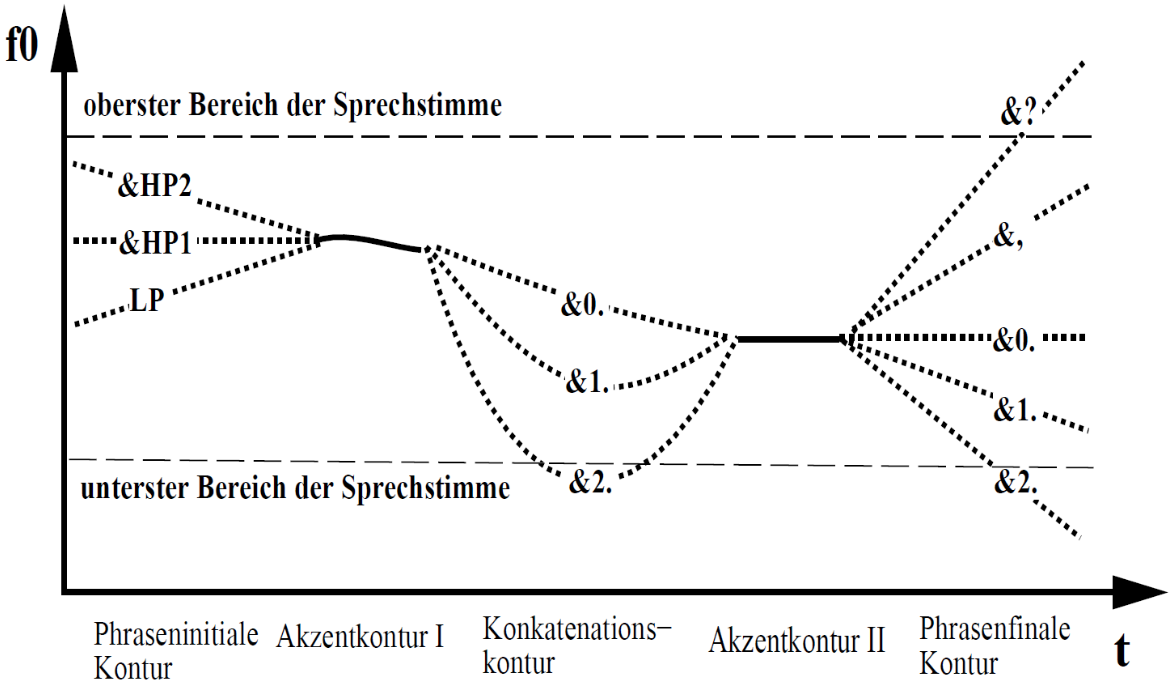


Abbildung 5.2: Stilisierte Intonationsmuster über einer prosodischen Phrase mit zugehörigen Etiketten des Kieler Intonationsmodells, entnommen aus [59]

Phrasengrenzen treten im Normalfall nach der phrasenfinalen Kontur auf. Bedingt durch technische oder syntaktische Abbrüche ist das auftreten von Phrasengrenzen auch an anderen Stellen möglich [59].

Eine Übersicht der im KIM verwendeten prosodischen Etiketten ist in Tabelle 5.5 aufgelistet. Mit diesem vergleichsweise großen Satz an Etiketten ist eine detaillierte Beschreibung des Intonationsverlaufes möglich [59].

Tabelle 5.5: Etiketten des Kieler Intonationsmodells

| Art der Kontur | Etikett | Bedeutung |
|---|---|---|
| phraseninitial | #&HP2 | hoch einsetzend, fallend |
| | #&HP1 | hoch einsetzend, eben |
| | #LP | tief einsetzend, steigend |
| Wortakzent | #&0) | früher Gipfel, völlig deakzentuiert |
| | #&1) | früher Gipfel, partiell deakzentuiert |
| | #&2) | früher Gipfel, Standardakzent |
| | #&3) | früher Gipfel, emphatisch verstärkt |
| | #&0^ | zentraler Gipfel, völlig deakzentuiert |
| | #&1^ | zentraler Gipfel, partiell deakzentuiert |
| | #&2^ | zentraler Gipfel, Standardakzent |
| | #&3^ | zentraler Gipfel, emphatisch verstärkt |
| | #&0( | später Gipfel, völlig deakzentuiert |
| | #&1( | später Gipfel, partiell deakzentuiert |
| | #&2( | später Gipfel, Standardakzent |
| | #&3( | später Gipfel, emphatisch verstärkt |
| | #&0] | frühes Tal, völlig deakzentuiert |
| | #&1] | frühes Tal, partiell deakzentuiert |
| | #&2] | frühes Tal, Standardakzent |
| | #&3] | frühes Tal, emphatisch verstärkt |
| | #&0[ | spätes Tal, völlig deakzentuiert |
| | #&1[ | spätes Tal, partiell deakzentuiert |
| | #&2[ | spätes Tal, Standardakzent |
| | #&3[ | spätes Tal, emphatisch verstärkt |
| | #&0- | eben, völlig deakzentuiert |
| | #&1- | eben, partiell deakzentuiert |
| | #&2- | eben, Standardakzent |
| | #&3- | eben, emphatisch verstärkt |
| | #&0 | vollständig deakzentuiertes Inhalts- oder Funktionswort |
| Konkatenation | #&0. | keine Einbuchtung |
| | #&1. | leichte bis mittelstarke Einbuchtung |
| | #&2. | Einbuchtung zu unterer $F_0$-Grenze |

Tabelle 5.5: Etiketten des Kieler Intonationsmodells (Fortsetzung)

| Art der Kontur | Etikett | Bedeutung |
|---|---|---|
| Phrasenfinal | #&0. | eben |
| (nach Gipel/ | #&1. | leicht bis mittelstark fallend |
| ebener Kontur) | #&2. | fallend zu unterer $F_0$-Grenze |
| | #&1., | leicht bis mittelstark fallend + Anstieg |
| | #&1.? | leicht bis mittelstark fallend + Anstieg zu oberer $F_0$-Grenze |
| | #&2., | fallend zu unterer $F_0$-Grenze + Anstieg |
| | #&2.? | fallend zu unterer $F_0$-Grenze + Anstieg zu oberer $F_0$-Grenze |
| | #&0; | eben + sehr leichter Anstieg |
| | #&1; | leicht bis mittelstark fallend + sehr leichter Anstieg |
| | #&2; | fallend zu unterer $F_0$-Grenze + sehr leichter Anstieg |
| Phrasenfinal | #&, | leichter bis mittlerer Anstieg |
| (nach Tal) | #&? | Anstieg zu oberer $F_0$-Grenze |
| Phrasengrenze | #&PGn | intendierter Abbruch oder Häsitationspartikel |
| | #&=PGn | wortinterner Abbruch |
| | #&PG/ | syntaktischer Abbruch |
| | #&PG; | technischer Abbruch |

### 5.1.3 Konvertierung des Kieler Intonationsmodells zum System *German Tones and Break Indices 'light'*

Zur Ergänzung der sprachlichen Datenbasis der TUD wurde im Rahmen der vorliegenden Arbeit der KCSGrs beschafft. Für eine Ergänzung der Datenbasis ist ein einheitliches Format notwendig. Der vorhandene Datensatz BITS-US beinhaltet Etiketten aus dem System GToBI light [47].

Das KIM besteht aus deutlich mehr Etiketten und detaillierteren Informationen als das System GToBI light. Aus diesem Grund ist eine Konvertierung ausschließlich vom KIM zum System GToBI light möglich.

In Rücksprache mit Benno Peters, dem Verwalter des KCSG, wurden fünf Regeln für eine Konvertierung vom KIM zu GToBI light Etiketten erstellt.

1. Ein Gipfel (früh, zentral oder spät) mit nachfolgender eingebuchteter Konkatenationskontur wird auf ein H*L abgebildet.
   `#&[23][)(^] mit rechts #&[12].    ->    H*L`
2. Ein Gipfel (früh, zentral oder spät) mit nachfolgender linearer Konkatenationskontur und nachfolgendem Gipfel oder Tal wird auf ein H* abgebildet.
   `#&[23][)(^] mit rechts #&0. und #&[23][)([]^]    ->    H*`

3. Ein Tal (früh oder spät) wird auf ein L*H abgebildet.
   ```
   #&[23][][]    ->    L*H
   ```
4. Eine Phrasengrenze mit steigender phrasenfinalen Kontur wird auf ein H% abgebildet.
   ```
   #&PG[n/;]  und #&[,?]    ->    H%
   #&=PGn     und #&[,?]    ->    H%
   ```
5. Eine Phrasengrenze mit flacher oder fallender phrasenfinalen Kontur wird auf ein % abgebildet.
   ```
   #&PG[n/;] und #&[012][.;]    ->    %
   #&=PGn    und #&[012][.;]    ->    %
   ```

Mit diesen fünf Regeln lassen sich die Etiketten des KIM auf eine Untermenge des Systems GToBI light abbilden. Da im KIM keine intermediären und Intonationsphrasen unterschieden werden, ist eine Konvertierung dieses Unterschiedes nicht möglich. Auch der L* kann nicht eindeutig aus den Etiketten des KIM ausgelesen werden. Aufgrund seiner Seltenheit wurde der Akzent vernachlässigt. Die gleichbleibenden Töne wurden aufgrund ihrer Komplexität und Seltenheit nicht konvertiert. Zu unsicheren Etiketten kann ebenso nicht konvertiert werden.

Das KIM lässt sich durch massiven Informationsverlust auf die Etiketten H*, H*L, L*H, H% und % abbilden.

## 5.2 *Python Tones and Break Indices*

Für die Nutzung von PyToBI sind das Sprachverarbeitungsprogramm Praat [11] und ein Programm zur Verarbeitung von Python-Skripten notwendig [21].

Die Eingabe von PyToBI besteht aus einer Audiodatei und der zugehörigen gleichnamigen *TextGrid*-Datei. Die TextGrid-Datei muss einem speziellen Aufbau folgen. Die Namen der ersten beiden Ebenen müssen `words` und `phones` sein. In der Ebene `words` sind die alignierten Worte enthalten. Diese dürfen keine Umlaute beinhalten. In der Ebene `phones` sind die alignierten Phoneme enthalten. Für die zeitliche Zuweisung der prosodischen Etiketten werden die primären Akzente der Phoneme genutzt. Die Akzente müssen mit dem *Advanced Research Projects Agenca alphabet* (ARPABET) kodiert sein.

Die Ausgabe von PyToBI besteht aus einer TextGrid-Datei, welche den Namen der ursprünglichen Datei mit dem Zusatz `_result.TextGrid` trägt. Die Datei enthält zwei Ebenen. Die Ebene `tones` enthält die Tonakzente einer Untermenge der ToBI-Etiketten. Die Ebene `breaks` enthält die Pausen einer Untermenge der ToBI-Etiketten.

In PyToBI sind für Tonakzente die Etiketten L*+H, L+H*, H*+L, L*, H* und !H* vorhanden. Für die Grenztöne sind die Etiketten LH-, L-H%, HL-, H-L%, L-, LL- und L-L%

vorhanden. Für Wortgrenzen sind die Etiketten 1, 2, 3 und 4 vorhanden. Diese entsprechen in ihrer Funktion den in Tabelle 5.1 beschrieben Bedeutungen. Die Detektion des Etiketts 0 ist sehr schwierig und wurde nicht berücksichtigt [21].

### 5.2.1 Validierung

Für die Validierung von PyToBI wurde ein ungefähr fünfminütiger Ausschnitt eines Vortrages der Konferenz *Technology, Entertainment, Design* von zwei linguistisch geschulten Personen annotiert. Das Material stammt von einem amerikanischen Sprecher und beinhaltete 830 Wörter [21].

Die prosodische Annotation beinhaltete 459 beziehungsweise 473 Etiketten. Die Ergebnisse der Evaluation sind in Tabelle 5.6 dargestellt. Für das Mess-System Mensch wurde die Ähnlichkeit der manuell erstellten Etiketten zueinander bestimmt. Für das System PyToBI wurde die Ähnlichkeit der automatisch erstellten Etiketten zu den manuell erstellten Etiketten bestimmt. Ein automatisch erstelltes Etikett zählte als richtig, wenn es mindestens einem der manuell erstellten Etiketten entsprach [21]. Für beide Systeme wurde die Lage und die genaue Übereinstimmung der Etiketten evaluiert.

Tabelle 5.6: Validierung des Systems *Python Tones and Break Indices*

| Mess-System | Mess-Objekt | Übereinstimmung Akzente [%] | Übereinstimmung Pausen [%] |
|---|---|---|---|
| Mensch | Lage | 91 | 100 |
| | Etikett | 78 | 85 |
| PyToBI | Lage | 77 | 97 |
| | Etikett | 47 | 90 |

## 5.3 *Prosody Recognition Revisited*

Das System *Prosody Recognition Revisited* (PRR) ist die aktuelle Überarbeitung des *Prosody Recognition System* (PRS). Das PRS ist ein Programm zur automatischen prosodischen Etikettierung von Sprachmaterial. Das PRS wurde auf Basis deutschsprachiger Aufnahmen entwickelt und 1998 veröffentlicht [65].

Die Eingabe des PRS besteht aus einer Audiodatei und der gleichnamigen zugehörigen graphemischen Textdatei [64]. Die Audiodatei muss eine Abtastrate von 16 kHz aufweisen. Die Vorverarbeitung besteht aus einer Merkmalsextraktion der Daten. Die Zuweisung prosodischer Etiketten erfolgt durch einen trainierten Entscheidungsbaum [65].

### 5.3.1 *Prosody Recognition System* – Datenaufbereitung

Die Aufbereitung der Daten des PRS erfolgt in zwei Schritten.

Der erste Schritt besteht aus der Segmentierung und dem Alignieren der Textdatei. Für das Alignment wird ein explizites HMM-basiertes Segmentierungsverfahren genutzt. Die Konvertierung von G2P erfolgt mithilfe des CELEX-Lexikons [3]. Nicht im Lexikon enthaltene Wörter werden regelbasiert konvertiert. Die Funktionen `Alignwords`, `Alignphones` und `phonemic2syllables` erzeugen die alignierte phonemische Umschrift in SAMPA mit Silbengrenzen [65].

Im zweiten Schritt wird der Grundfrequenzverlauf des Sprachsignals parametrisiert. Der Grundfrequenzverlauf wird automatisch berechnet. Für Zeitfenster mit je zwei Silben werden für jede Silbe sieben Parameter berechnet, welche die spektralen Informationen der Grundfrequenz beinhalten [65]. Die gewählte Funktion zur Parametrisierung lautet

$$f(t) = \alpha \tanh(\beta(t-\gamma)) + \delta e^{-(\varepsilon(t-\zeta))^2} + \eta. \tag{5.1}$$

Die hyperbolische Tangensfunktion wurde zur Detektion von Anstiegen oder Abfällen gewählt (Unterschied H*+L und L*+H). Die Exponentialfunktion unterscheidet zwischen Gipfeln und Tälern. Die Konstante beinhaltet das Niveau der Grundfrequenz [65].

### 5.3.2 *Prosody Recognition System* – Training und Validierung

Aus den vorverarbeiteten Daten wurden 14 Parameter für die automatische prosodische Etikettierung extrahiert.

Sieben dieser Parameter wurden aus der Segmentierung und dem Alignment der Silben bestimmt. Diese Parameter beinhalten Angaben zu der Silbenakzentuierung, der Anzahl von Silben bis zum Wortende, der erwarteten, gemessenen und relativen Länge der Silbe, der Zeit bis zur nächsten Sprechpause und der Länge der nächsten Sprechpause.

Die anderen sieben Parameter wurden aus der Grundfrequenzparametrisierung bestimmt. Eine Übersicht der einzelnen Funktionen dieser Parameter ist in Tabelle 5.7 dargestellt.

Tabelle 5.7: Interpretation der Grundfrequenzparameter des *Prosody Recognition System*

| Parameter | Interpretation für den Grundfrequenzverlauf |
|---|---|
| $\alpha$ | Halbe Höhe des Anstiegs oder Abfalls |
| $\beta$ | Steilheit des Anstiegs oder Abfalls |
| $\gamma$ | Zeitpunkt des Anstiegs oder Abfalls |
| $\delta$ | Höhe eines Gipfels oder Tals |
| $\varepsilon$ | Steilheit eines Gipfels oder Tals |
| $\zeta$ | Zeitpunkt eines Gipfels oder Tals |
| $\eta$ | $F_0$-Niveau |

Für das Training und die Validierung sind prosodisch etikettierte Daten notwendig. Das PRS wurde auf der Basis des *Stuttgart Radio News Corpus* (SRNC) [65] trainiert. Der SRNC besteht aus einer Stunde Sprachmaterial von einem Sprecher und beinhaltet gelesene Sprache in Form von Radionachrichten. Der SRNC wurde manuell prosodisch etikettiert [64]. Die Etikettierung beinhaltet 22 Etiketten für Akzente (!H*, !H*L, !H*L?, !HH*L, *?, 0, H*, H*?, H*L, H*L?, H*M, H*M?, HH*L, HH*L?, L*, L*!H, L*!H?, L*?, L*H, L*H?, L*HL, L*HL?) und 13 Etiketten für Silbengrenzen (%, %?, %H, -, -?, ..H, ..H?, ..L, ..L?, 0, H%, H%?, L%). Das jeweils enthaltene Etikett 0 bedeutet, dass kein Akzent beziehungsweise keine Pause vorliegt. Da der SRNC mit Unsicherheiten markierte Etiketten beinhaltet (?), wurden auch Etiketten mit Unsicherheiten vorhergesagt.

Für die Validierung des PRS wurde der SRNC in einen Trainingsdatensatz mit 80 % und einen Validierungsdatensatz mit 20 % der Daten unterteilt. Aus den 14 extrahierten Parametern wurde mit dem Algorithmus C4.5 [63] ein Entscheidungsbaum für die automatische prosodische Etikettierung erstellt.

Die Genauigkeit der richtig klassifizierten Etiketten lag mit dem PRS bei 78.7 % [65] für einen einzelnen Durchlauf des C4.5. Im Programm C4.5 ist die Möglichkeit iterativer Neuberechnungen zur Verbesserung des Entscheidungsbaumes möglich. Hierdurch wurde die Genauigkeit auf 79.6 % erhöht [65].

### 5.3.3 *Prosody Recognition System* – Neuerungen

In Zusammenarbeit mit Stephan Rapp, dem Ersteller des PRS, wurde das Programm aufbereitet und auf den CARInA angewendet. Das PRS konnte aus unterschiedlichen Gründen nicht im originalen Zustand verwendet werden. Das aktuelle Programm wird PRR genannt. Im Folgenden sind die Unterschiede und deren Notwendigkeit zwischen dem PRS und dem PRR ausgeführt.

Das PRS nutzt für die G2P-Konvertierung das Lexikon CELEX. CELEX ist ein kommerzielles Lexikon und nicht an der TUD vorhanden. Aus diesem Grund wurde für die G2P-Konvertierung das Wörterbuch `Dict_graphem2phonem.txt` verwendet, welches aus den Daten des Wiktionary entstand und manuell erweitert wurde (vergleiche Abschnitt 4.2).

Der Algorithmus des Entscheidungsbaumes C4.5 wurde überarbeitet. Das PRR verwendet die neue Version C5.0 [57]. Der Entscheidungsbaum C5.0 generiert verbesserte und vereinfachte Entscheidungsbäume [57].

Es wurden zusätzlich einige Änderungen an dem Programm selbst durchgeführt. Die Grundfrequenzparametrisierung des PRS war für eine spezielle Person ausgelegt. Da der CARInA aus dem Sprachmaterial vieler Personen besteht, ist im System PRR eine Normalisierung der Grundfrequenz implementiert worden. Für eine einfachere phonetische Lesbarkeit wurde die phonemische Segmentierung von SAMPA zu IPA umgestellt und die Textdateien von *American Standard Code for Information Interchange* (ASCII) zu UTF-8. Verschiedene andere Änderungen ergaben sich hinsichtlich der Vorverarbeitung, wie beispielsweise die Behandlung von Zahlen, Bindestrichen, Abkürzungen und Sonderzeichen.

Es wurde eine Validierung des PRR mit dem SRNC durchgeführt. Hierfür wurde dieselbe Zusammenstellung von Trainings- und Validierungsdatensatz gewählt, mit welcher das PRS validiert wurde. Die Validierung erfolgte durch dasselbe Verfahren wie bei dem PRS. Für die Vorhersage prosodischer Etiketten wurde eine Genauigkeit von 78.5 % erreicht. Durch iterative Neuberechnungen des Entscheidungsbaumes wurde die Genauigkeit auf 80.6 % erhöht. Die erreichte Genauigkeit entspricht somit circa der Genauigkeit des PRS.

Für die Vorhersage prosodischer Etiketten wurde die Etikettierung des SRNC überarbeitet. Die 22 Etiketten für Akzente wurden auf fünf Etiketten (0, H*, H*L, L*H, L*) abgebildet. Die 13 Etiketten für Silbengrenzen wurden auf fünf Etiketten (0, -, %, L%, H%) abgebildet.

### 5.3.4 Anwendung des Programms

Das Programm PRR wurde für die automatische Etikettierung des Teilkorpus *Complete* verwendet. Die Vorverarbeitung der Daten erfolgte nach dem in Unterabschnitt 5.3.1 beschriebenen Schema. Mithilfe des Systems PRR wurde der Teilkorpus *Complete* auf drei Arten mit prosodischen Etiketten versehen.

Der erste Entscheidungsbaum für die Generierung prosodischer Etiketten wurde mit dem modifizierten SRNC erstellt. Dieser beinhaltet fünf Etiketten für Akzente (0, H*, H*L, L*H, L*) und fünf Etiketten für Wortgrenzen (0, -, %, L%, H%).

Der zweite Entscheidungsbaum für die Generierung prosodischer Etiketten wurde mit dem BITS-US erstellt. Aus dem BITS-US wurden die Etiketten mit Unsicherheiten und für gleichbleibende Töne entfernt. Es verbleiben fünf Etiketten für Akzente (0, H*, H*L, L*, L*H) und vier Etiketten für Wortgrenzen (0, -, %, H%).

Der dritte Entscheidungsbaum für die Generierung prosodischer Etiketten wurde mit dem KCSGrs erstellt. Die Etiketten des KIM wurden nach den in Unterabschnitt 5.1.3 beschrieben Regeln zu einer Untermenge der GToBI light Etiketten konvertiert. Diese Untermenge besteht aus vier Tonakzenten (0, H*, H*L, L*H) und drei Etiketten für Wortgrenzen (0, %, H%).

## 5.4 Validierung

Die in Unterabschnitt 5.2.1 und Unterabschnitt 5.3.2 beschriebenen Validierungen von PyToBI und dem PRS sind aufgrund unterschiedlicher Verfahren nicht vergleichbar. Aus diesem Grund wurde für PyToBI und das PRR eine umfangreiche einheitliche Validierung durchgeführt. Im Allgemeinen sind bei der Validierung prosodischer Etiketten zwei Aspekte zu berücksichtigen.

1. Es gibt keine eindeutigen Etiketten als Referenz. Die Übereinstimmung manuell erstellter prosodischer Etiketten für das System GToBI liegt für Tonakzente bei 71 % und für Grenztöne bei 86 % [34].

2. Ein Großteil der gesprochenen Einheiten ist unbetont. Der modifizierte SRNC wurde mit 80 % der Daten trainiert und mit 20 % validiert. Die Ergebnisse des Validierungsdatensatzes sind in Tabelle 5.8 dargestellt. Unbetonte Silben (0/0) weisen einen Anteil von 76 % auf. Ein System zur automatischen Prosodievorhersage, welches ausschließlich unbetonte Silben vorhersagt, weist für diesen Validierungsdatensatz eine Genauigkeit von 76 % auf. Das verwendete System weist eine Genauigkeit von 82.6 % auf.

Die Validierung von PyToBI erfolgte durch den Vergleich der automatisch vorhergesagten Etiketten mit zwei manuell erstellten Etikettierungen. Die gewählte Anzahl von zwei Referenzen ist willkürlich. Weitere Referenzen würden ausschließlich zur Verbesserung der Übereinstimmung beitragen können. Zusätzlich ist nicht ersichtlich, ob der Vergleich der Etiketten Wort- oder Silbenweise erfolgte.

Das PRS wurde mit einem manuell annotierten Datensatz validiert. Die Validierung beinhaltete einen Vergleich der Etiketten pro Silbe. Bei der Validierung ist ausschließlich angegeben, ob die Etiketten der Silbe für den Tonakzent **und** die Grenze richtig erkannt wurden. Eine getrennte Auswertung wurde nicht vorgenommen. Zusätzlich ist die Konfusionsmatrix nicht bekannt, sodass die Lage eines Akzentes nicht validiert werden kann. Auch ist für das PRS ausschließlich die Genauigkeit angegeben. Die als „falsch negativ“ klassifizierten Elemente bleiben in der Validierung unberücksichtigt.

Tabelle 5.8: Auswertung der automatisch erstellten prosodischen Etikettierung des Validierungsdatendatzes vom *Stuttgart Radio News Corpus*. Die prosodische Etikettierung erfolgte mit System *Prosody Recognition Revisited*.

| Etikett (Ton/Grenze) | Anzahl | falsch positiv | falsch negativ |
|---|---|---|---|
| 0/- | 127 | 54 | 84 |
| 0/% | 105 | 47 | 49 |
| 0/0 | 3063 | 430 | 99 |
| 0/H% | 38 | 14 | 30 |
| 0/L% | 25 | 29 | 15 |
| H*/0 | 75 | 0 | 74 |
| H*L/- | 2 | 0 | 2 |
| H*L/% | 12 | 5 | 7 |
| H*L/0 | 246 | 53 | 159 |
| H*L/L% | 0 | 2 | 0 |
| L*/0 | 6 | 0 | 6 |
| L*H/- | 32 | 10 | 23 |
| L*H/% | 9 | 7 | 7 |
| L*H/0 | 269 | 46 | 144 |
| L*H/H% | 2 | 3 | 1 |

### 5.4.1 Übereinstimmungsgrad prosodischer Etikettierungen

Die Validierung von PyToBI und des Programms PRR erfolgte mithilfe des Übereinstimmungsgrades prosodischer Etikettierungen. Der Übereinstimmungsgrad kann für Pausen und Akzente berechnet werden.

Der Übereinstimmungsgrad verschiedener prosodischer Etikettierungen für die gleiche Sprachaufnahme wird durch den Mittelwert paarweise berechneter Hamming-Abstände bestimmt [34], [60]. Der Algorithmus wird beispielhaft an den 4 Etikettierungen in Tabelle 5.9 erklärt. Für die Bestimmung eines Hamming-Abstandes werden jeweils 2 Etikettierungen miteinander verglichen. Für $N$ Etikettierungen ergeben sich somit $\frac{N^2-N}{2}$ Paare. Für jedes Paar werden die Etiketten für jedes Wort miteinander verglichen. Für $W$ Wörter ergeben sich somit $W \cdot \frac{N^2-N}{2}$ Vergleiche. Der Anteil übereinstimmender Vergleiche an der Gesamtanzahl der Vergleiche ergibt den Übereinstimmungsgrad. Der Übereinstimmungsgrad kann für die Etiketten oder die Lagen der Etiketten berechnet werden.

In Tabelle 5.9 sind 4 exemplarische Etikettierungen von Tonakzenten mit dem System GToBI dargestellt. Für jedes Wort werden jeweils 2 Etiketten verglichen (Nichtakzentuierungen sind als eigenes Etikett zu betrachten). Mit 4 Etikettierungen können 6 Paare gebildet werden. Da 6 Wörter verglichen werden, ergeben sich 36 Vergleiche. Für das erste, zweite und dritte Wort sind die Etiketten jeder Etikettierungen identisch. Hierdurch ergeben sich 18 übereinstimmende Vergleiche. Für das vierte, fünfte und sechste Wort ist jeweils ein Etikett unterschiedlich. Somit stimmen jeweils 3 der 6 Vergleiche nicht überein. Für die letzten 3 Wörter ergeben sich 9 übereinstimmende Vergleiche. Insgesamt sind 27 Vergleiche positiv ausgefallen. Der Übereinstimmungsgrad der Etiketten beträgt $\frac{27}{36}$ beziehungsweise 75 %.

Der Übereinstimmungsgrad der Akzentlagen wird durch den paarweisen Vergleich der Akzente bestimmt. In Tabelle 5.9 ist das erste, zweite, dritte, vierte und sechste Wort in jeder Etikettierung akzentuiert. Es resultieren 30 übereinstimmende Vergleiche. Das fünfte Wort ist einmal akzentuiert und dreimal nicht akzentuiert. Es ergeben sich 3 übereinstimmende Vergleiche. Der Übereinstimmungsgrad der Akzentlagen beträgt $\frac{33}{36}$ beziehungsweise 91.7 %.

Tabelle 5.9: Prosodische Etikettierungen des Satzes „Die Sonne dreht durch vom Scheinen.“ aus vier unterschiedlichen Quellen für dieselbe Sprachaufnahme

| Orthographie: | Die | Sonne | dreht | durch | vom | Scheinen. |
|---|---|---|---|---|---|---|
| Etikettierung 1: | H* | | H* | H* | | L-% |
| Etikettierung 2: | H* | | H* | H* | | L-% |
| Etikettierung 3: | H* | | H* | !H* | | L-% |
| Etikettierung 4: | H* | | H* | H* | L- | L-H% |

### 5.4.2 Beurteilung der automatisch erstellten Etiketten

Zur Validierung der automatisch generierten prosodischen Etiketten wurde für die Algorithmen der Übereinstimmungsgrad bestimmt. Zu validierende Programme sind PyToBI, PRR SRNC, PRR BITS-US und PRR KCSGrs.

Als Referenz wurde die zu GToBI light konvertierte Version des KCSGrs genutzt. Der KCSGrs beinhaltet Aufnahmen zu 603 unterschiedlichen Inhalten mit je zwei bis 16 Realisierungen von insgesamt 53 Personen. Um den Grad der Personenabhänggigkeit der prosodischen Etiketten zu bestimmen, wurde der Übereinstimmungsgrad der jeweiligen Realisierungen für den gesamten Korpus berechnet. Die Übereinstimmung der Etiketten für Tonakzente liegt bei 83.4 %, die Übereinstimmung der Akzentlagen bei 86.3 %. Für Pausen liegt die Übereinstimmung der Etiketten bei 93.3 % und die Übereinstimmung der Lagen bei 95.7 %. Der Übereinstimmungsgrad manuell erstellter prosodischer Etiketten mit dem System GToBI liegt für Tonakzente bei 71 % und für Pausen bei 86 % [34]. Der KCSGrs kann somit für das Arbeiten mit prosodischen Etiketten als personenunabhängig angenommen werden.

Für eine 10-fache Kreuzvalidierung wurden die 603 unterschiedlichen Inhalte aufgeteilt, sodass die Summe des Sprachmaterials der jeweiligen Realisierungen gleich ist. Eine Übersicht der einzelnen Partitionen ist in Tabelle 5.10 aufgelistet. Die einzelnen Inhalte sind in der Tabelle A.1 in Abschnitt A.1 aufgeführt. Durch unterschiedlich viele Realisierungen und unterschiedlich umfangreiche Inhalte entsteht durch eine Einteilung des Korpus in gleiche Längen eine Differenz in der Anzahl verschiedener Inhalte.

Tabelle 5.10: Zusammensetzung der Partitionen des *Kiel Corpus of Spoken German read speech* für die 10-fache Kreuzvalidierung der automatisch erstellten prosodischen Etiketten

| Teil | Länge [mm:ss] | Anzahl verschiedener Inhalte |
|---|---|---|
| 1 | 25:28 | 56 |
| 2 | 25:34 | 22 |
| 3 | 25:31 | 56 |
| 4 | 25:26 | 63 |
| 5 | 25:32 | 70 |
| 6 | 25:32 | 72 |
| 7 | 25:27 | 68 |
| 8 | 25:34 | 74 |
| 9 | 25:30 | 45 |
| 10 | 25:33 | 77 |

Der KCSGrs wurde für die Validierung mit je neun Partitionen trainiert und auf die zehnte angewendet. Die anderen Systeme des PRR wurden mit dem jeweiligen gesamten Datenmaterial trainiert.

Der Satz verwendeter Etiketten unterscheidet sich zwischen den Systemen. Für eine aussagekräftige Validierung wurden die vorhergesagten Etiketten auf die Etiketten der zu GToBI light konvertierten Version des KCSGrs abgebildet. Die Abbildung der Etiketten ist in Tabelle 5.11 aufgelistet.

Tabelle 5.11: Abbildung der vorhergesagten prosodischen Etiketten zu den Etiketten der zu GToBI light konvertierten Version des *Kiel Corpus of Spoken German read speech*.

| Etikett KCSGrs | Auf das Etikett des KCSGrs abgebildete Etiketten |
|---|---|
| H* | H*, !H*, L+H* |
| H*L | H*L, H*+L |
| L*H | L*H, L*+H |
| 0 (Akzent) | 0, L*, LH-, L-H%, HL-, H-L%, L-, LL-, L-L% |
| % | %, -, L%, 3, 4 |
| H% | H%, 4 mit LH- oder L-H% |
| 0 (Pause) | 0, 1, 2 |

Für die 10 Teile des KCSGrs wurde die Übereinstimmung von den vorhergesagten Etiketten zu den manuell erstellten, zu GToBI light konvertierten Etiketten berechnet. Der Übereinstimmungsgrad wurde jeweils für die Etiketten der Akzente, die Akzentlagen, die Etiketten der Pausen und die Pausenlagen berechnet.

Die Ergebnisse sind in Tabelle 5.12 und Tabelle 5.13 dargestellt. Die Übereinstimmungsgrade für ein System sind für die unterschiedlichen Teile des KCSGrs sehr ähnlich. Dies bestätigt die Personenunabhängigkeit der prosodischen Etiketten.

Aus der Tabelle 5.12a ist zu entnehmen, dass das System PRR KCSGrs mit 81.3 % die höchte Übereinstimmung der Etiketten der Tonakzente aufweist. Unter ausschließlicher Berücksichtigung der Akzentlagen wird die Übereinstimmung auf 84 % erhöht (vergleiche Tabelle 5.12b). Die geringste Übereinstimmung der Etiketten der Tonakzente weist mit 59.5 % das System PyToBI auf. Unter ausschließlicher Berücksichtigung der Akzentlagen wird diese auf 62.14 % erhöht.

Der Übereinstimmungsgrad für die Vorhersage von Pausen ist für jedes System höher als der von Akzenten. Nach Tabelle 5.13a liegt die beste Vorhersage der Etiketten bei 93 % und wurde durch das System PyToBI erreicht. Durch ausschließliche Berücksichtigung der Pausenlagen wurde der Übereinstimmungsgrad auf 94.6 % erhöht (vergleiche Tabelle 5.13b). Die geringste Übereinstimmung erreichte mit 84.8 % das System PRR BITS-US. Diese wurde durch ausschließliche Berücksichtigung der Pausenlagen auf 87.1 % erhöht. Die mit dem PRR trainierten Systeme weisen für den Teil 2 einen deutlich geringeren Übereinstimmungsgrad für Pausen auf. Der Teil 2 weist mit 22 unterschiedlichen Inhalten die geringste Anzahl verschiedener Inhalte auf. Unter anderem sind zwei Kurzgeschichten enthalten. Das Programm PyToBI weist für diesen Teil jedoch keine Auffälligkeiten auf.

Tabelle 5.12: Übereinstimmungsgrad der Etiketten und Lagen für Tonakzente der Systeme zur automatischen prosodischen Annotation. Die Berechnung des Übereinstimmungsgrades erfolgte durch den Vergleich der vorhergesagten Etiketten mit den zu GToBI light konvertierten Etiketten des in 10 Teile eingeteilten *Kiel Corpus of Spoken German read speech*. Das PRR KCSGrs wurde jeweils mit allen Teilen bis auf den Validierungsteil trainiert.

(a) Etiketten Tonakzente

| | **Übereinstimmungsgrad** [%] | | | |
|---|---|---|---|---|
| **Teil** | PRR KCSGrs | PRR SRNC | PRR BITS-US | PyToBI |
| 1 | 79.7 | 65.5 | 65.2 | 58.4 |
| 2 | 79.4 | 65.6 | 61.6 | 58.3 |
| 3 | 82.7 | 70 | 68.7 | 59.7 |
| 4 | 82.4 | 67.6 | 65.9 | 59.6 |
| 5 | 81.1 | 64.5 | 63.6 | 61.1 |
| 6 | 81.2 | 66.8 | 67.9 | 58.8 |
| 7 | 80.7 | 67.1 | 64.6 | 59.9 |
| 8 | 82.1 | 66.5 | 65.2 | 59.2 |
| 9 | 80.4 | 65.2 | 65 | 60 |
| 10 | 83.2 | 68.6 | 67 | 59.8 |
| Durchschnitt | 81.29 | 66.74 | 65.47 | 59.48 |

(b) Lagen Tonakzente

| | **Übereinstimmungsgrad** [%] | | | |
|---|---|---|---|---|
| **Teil** | PRR KCSGrs | PRR SRNC | PRR BITS-US | PyToBI |
| 1 | 80.6 | 72.6 | 70.4 | 60.7 |
| 2 | 80.3 | 71.5 | 66.9 | 61.9 |
| 3 | 83.4 | 74.8 | 73.1 | 62.4 |
| 4 | 83.3 | 73.9 | 71.3 | 62.5 |
| 5 | 81.9 | 71.1 | 69.3 | 63.5 |
| 6 | 81.9 | 72.5 | 73.4 | 61.5 |
| 7 | 81.9 | 71.9 | 69.8 | 62.9 |
| 8 | 82.8 | 72.2 | 70.1 | 60.8 |
| 9 | 81.3 | 71 | 70.1 | 63 |
| 10 | 84 | 73.1 | 70.8 | 62.2 |
| Durchschnitt | 82.14 | 72.46 | 70.52 | 62.14 |

Tabelle 5.13: Übereinstimmungsgrad der Etiketten und Lagen für Pausen der Systeme zur automatischen prosodischen Annotation. Die Berechnung des Übereinstimmungsgrades erfolgte durch den Vergleich der vorhergesagten Etiketten mit den zu GToBI light konvertierten Etiketten des in 10 Teile eingeteilten *Kiel Corpus of Spoken German read speech*. Das PRR KCSGrs wurde jeweils mit allen Teilen bis auf den Validierungsteil trainiert.

(a) Etiketten Pausen

| | **Übereinstimmungsgrad** [%] | | | |
|---|---|---|---|---|
| **Teil** | PRR KCSGrs | PRR SRNC | PRR BITS-US | PyToBI |
| 1 | 87.9 | 84.6 | 83.9 | 93.8 |
| 2 | 82.5 | 79.8 | 78.8 | 92.9 |
| 3 | 88.5 | 86.2 | 84.5 | 94.3 |
| 4 | 90.6 | 87.6 | 86.3 | 92.8 |
| 5 | 91.2 | 88.5 | 88.4 | 92.8 |
| 6 | 90.1 | 87.6 | 86 | 92.3 |
| 7 | 90 | 86.9 | 85 | 92.4 |
| 8 | 90.5 | 87.6 | 85.8 | 92.1 |
| 9 | 85.6 | 81.6 | 81.3 | 92.7 |
| 10 | 92.7 | 88.9 | 87.5 | 93.5 |
| Durchschnitt | 88.96 | 85.93 | 84.75 | 92.96 |

(b) Lagen Pausen

| | **Übereinstimmungsgrad** [%] | | | |
|---|---|---|---|---|
| **Teil** | PRR KCSGrs | PRR SRNC | PRR BITS-US | PyToBI |
| 1 | 89.4 | 87.4 | 87 | 95.5 |
| 2 | 83.6 | 81.8 | 81 | 95.9 |
| 3 | 89.3 | 88.1 | 86.4 | 95.9 |
| 4 | 91.1 | 89.9 | 88.1 | 93.8 |
| 5 | 91.9 | 90.7 | 90.3 | 94 |
| 6 | 90.6 | 89.7 | 88.1 | 93.6 |
| 7 | 91.1 | 89.7 | 87.6 | 94.1 |
| 8 | 91.5 | 89.9 | 88.2 | 93.5 |
| 9 | 86.7 | 84.4 | 83.9 | 95 |
| 10 | 93.6 | 91.7 | 90.4 | 95.2 |
| Durschnitt | 89.88 | 88.33 | 87.1 | 94.65 |

Der Anteil an Wörtern ohne zugewiesenen Tonakzent im KCSGrs beträgt 76.6 %. Ein System, welches ausschließlich Nichtakzentuierungen vorhersagt, weist somit einen durchschnittlichen Übereinstimmungsgrad von 76.6 % auf. Dieser Übereinstimmungsgrad liegt über den Übereinstimmungsgraden des PRR SRNC, des PRR BITS-US und von PyToBI.

Der Anteil an Wörtern ohne zugewiesene Pause im KCSGrs beträgt 79.4 %. Ein System, welches zu keinem Zeitpunkt eine Pause vorhersagt, weist somit einen durchschnittlichen Übereinstimmungsgrad von 79.4 % auf. Dieser Übereinstimmungsgrad wird von allen Systemen übertroffen.

Für eine differenziertere Auswertung wurden die Konfusionsmatrizen für jedes System erstellt. Die Menge der vorhergesagten Etiketten entspricht nicht für jedes System der Menge der Referenzetiketten.

In Tabelle 5.14 sind die Konfusionsmatrizen für Tonakzente und Pausen des PRR KCSGrs dargestellt. Die Akzente H* und L*H wurden zu weniger als 24 % richtig etikettiert und zu weniger als 31 % einem Akzent zugewiesen. Der Akzent H*L wurde zu 46 % als Akzent erkannt und zu 43 % richtig zugewiesen. Von den erkannten H*L liegen 75 % zu Zeitpunkten eines Akzentes. Das Pausenetikett % wurde zu 71 % als Pause erkannt und das Pausenetikett H% zu 40 %.

Tabelle 5.14: Konfusionsmatrizen der automatisch erstellten Akzente und Pausen mit dem PRR KCSGrs zu den Referenzetiketten des *Kiel Corpus of Spoken German read speech.*

(a) Tonakzente

| | **Vorhersage** | | | |
|---|---|---|---|---|
| **Referenz** | 0 | H* | H*L | L*H |
| 0 | 23 009 | 116 | 687 | 205 |
| H* | 1053 | 81 | 28 | 30 |
| H*L | 2485 | 25 | 2018 | 84 |
| L*H | 1054 | 38 | 60 | 357 |

(b) Pausen

| | **Vorhersage** | | |
|---|---|---|---|
| **Referenz** | % | 0 | H% |
| % | 3414 | 1465 | 120 |
| 0 | 562 | 23 998 | 305 |
| H% | 166 | 873 | 427 |

Die Konfusionsmatrizen für Tonakzente und Pausen des PRR SRNC sind in Tabelle 5.15 dargestellt. Die Akzente H*, H*L und L*H wurden zu 54 % als Akzent erkannt. Im Vergleich zum PRR KCSGrs wurde ein großer Anteil an nichtakzentuierten Wörtern nicht richtig erkannt. Aus diesem Grund liegt der durchschnittliche Übereinstimmungsgrad mit 66.74 % deutlich unterhalb des durchschnittlichen Übereinstimmungsgrades des PRR KCSGrs. Der Akzent L* wurde für kein Wort vorhergesagt und ist nicht mit aufgeführt.

Die vorhergesagten Pausen mit dem Etikett % entsprechen zu 63 % dem Referenzetikett. Die richtige Vorhersage des Etiketts 0 bei Pausen liegt bei 91 %. Zu der Richtigkeit des

Pausenetiketts - kann keine Aussage getroffen werden, da diese nicht in den Referenzetiketten enthalten ist.

Tabelle 5.15: Konfusionsmatrizen der automatisch erstellten Akzente und Pausen mit dem PRR SRNC zu den Referenzetiketten des *Kiel Corpus of Spoken German read speech.*

(a) Tonakzente

| | **Vorhersage** | | | |
|---|---|---|---|---|
| **Referenz** | 0 | H* | H*L | L*H |
| 0 | 18 746 | 12 | 2243 | 3016 |
| H* | 539 | 1 | 209 | 443 |
| H*L | 2129 | 0 | 1674 | 809 |
| L*H | 692 | 0 | 331 | 486 |

(b) Pausen

| | **Vorhersage** | | | | |
|---|---|---|---|---|---|
| **Referenz** | % | - | 0 | H% | L% |
| % | 1388 | 309 | 1377 | 201 | 1724 |
| 0 | 521 | 606 | 23 371 | 190 | 177 |
| H% | 296 | 115 | 819 | 96 | 140 |

Die Konfusionsmatrizen für Tonakzente und Pausen des PRR BITS-US sind in Tabelle 5.16 dargestellt. Für die Akzente H* und L*H wurde zu 67 % ein Akzent vorhergesagt, für den Akzent H*L zu 78 %. Das PRR KCSGrs weist über 6000 richtige Vorhersagen mehr für nichtakzentuierte Wörter auf als das PRR BITS-US. Aus diesem Grund liegt der Übereinstimmungsgrad des PRR BITS-US mit 65.47 % deutlich unterhalb des Übereinstimmungsgrades des PRR KCSGrs.

Die vorhergesagten Pausen mit dem Etikett % entsprechen zu 83 % einer Grenze. Die Vorhersagen des Etiketts - kann nicht beurteilt werden.

Tabelle 5.16: Konfusionsmatrizen der automatisch erstellten Akzente und Pausen mit dem PRR BITS-US zu den Referenzetiketten des *Kiel Corpus of Spoken German read speech.*

(a) Tonakzente

| | **Vorhersage** | | | |
|---|---|---|---|---|
| **Referenz** | 0 | H* | H*L | L*H |
| 0 | 16 637 | 47 | 4991 | 2342 |
| H* | 387 | 6 | 439 | 360 |
| H*L | 992 | 11 | 3145 | 464 |
| L*H | 476 | 1 | 316 | 716 |

(b) Pausen

| | **Vorhersage** | | | |
|---|---|---|---|---|
| **Referenz** | % | - | 0 | H% |
| % | 3291 | 208 | 1315 | 185 |
| 0 | 734 | 1096 | 22 903 | 132 |
| H% | 409 | 140 | 802 | 115 |

Die Konfusionsmatrizen für Tonakzente und Pausen des Systems PyToBI sind in Tabelle 5.16 dargestellt. Aufgrund der vielen unterschiedlichen Grenztöne ist ein Großteil der nichtakzentuierten Wörter mit einem Etikett versehen worden. Die Etiketten HL-, LH- und LL- werden mit einer sehr geringen Wahrscheinlichkeit vorhergesagt.

Der Übereinstimmungsgrad der Pausen liegt mit 94.65 % über dem der anderen Systeme. Die Vorhersage des Referenzetiketts H% kann mithilfe der Tabelle nicht beurteilt werden, da für das System PyToBI die Grenztöne in der Ebene der Tonakzente sind. Das Etikett 2 wurde 261 mal häufiger vorhergesagt als das Etikett 3. Dies ist erstaunlich, da das Etikett 2 für eine Unstimmigkeit der tonalen und rhythmischen Eigenschaften steht (vergleiche Unterabschnitt 5.1.1).

Tabelle 5.17: Konfusionsmatrizen der automatisch erstellten Akzente und Pausen mit PyToBI zu den Referenzetiketten des *Kiel Corpus of Spoken German read speech*.

(a) Tonakzente

| | **Vorhersage** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Referenz** | !H* | 0 | H* | H*+L | H-L% | HL- | L* | L*+H | L+H* | L-H% | L-L% | LH- | LL- |
| 0 | 2338 | 9920 | 1581 | 1020 | 256 | 0 | 6320 | 897 | 279 | 269 | 1133 | 0 | 4 |
| H* | 133 | 494 | 72 | 19 | 0 | 0 | 290 | 139 | 23 | 10 | 12 | 0 | 0 |
| H*L | 184 | 754 | 142 | 281 | 1341 | 3 | 301 | 206 | 25 | 186 | 1186 | 1 | 2 |
| L*H | 36 | 565 | 18 | 28 | 12 | 0 | 147 | 221 | 42 | 322 | 109 | 7 | 2 |

(b) Pausen

| | **Vorhersage** | | | |
|---|---|---|---|---|
| **Referenz** | 1 | 2 | 3 | 4 |
| % | 645 | 343 | 6 | 4005 |
| 0 | 20 416 | 4419 | 2 | 26 |
| H% | 450 | 200 | 11 | 805 |

*Kapitel 6*

# 6 CARInA – Aufbau und Struktur

Die Listen `Annotation_*.mat` enthalten alle vorhandenen orthographischen, phonetischen, kanonischen, wortartenbezogenen und silbischen Informationen des entsprechenden Artikels (vergleiche Abschnitt 4.3).

Viele Sätze weisen unvollständige Informationen auf. Mögliche Ursachen hierfür sind zum Beispiel, dass ein Anteil des SWC nicht aligniert wurde (siehe Abschnitt 2.2.3) oder dass die Wörterbücher nicht jedes Wort enthalten (siehe Abschnitt 4.2).

Für viele Anwendungen der Sprachverarbeitung werden nicht alle Informationen benötigt. Für die anwendungsspezifische Erstellung eines Teilkorpus ist eine Übersicht aller zur Verfügung stehender Informationen notwendig.

Für die Extraktion eines Satzes aus einem Artikel des GSWC muss der Start- und Endzeitpunkt des Satzes bekannt sein. Jeder Satz mit einem Start- und Endzeitpunkt wurde extrahiert. Informationen zur Vollständigkeit der Annotationen wurden in der Liste `ContentStatus.mat` gespeichert.

Die Liste `ContentStatus.mat` enthält 12 Spalten. Die erste Spalte enthält die Information über den Identifikationsnamen und den Namen des entsprechenden Satzes. Die Spalten 2–9 enthalten Informationen über die Vollständigkeit des Alignments auf Wortebene, des Alignments auf phonetischer Ebene, der Wortartenbestimmung durch MATLAB, der kanonischen Realisierung, der Wortartenbestimmung durch das Wörterbuch, der Silbifizierung, der phonetischen Akzentsetzung und der Prosodie. Sind die entsprechenden Informationen vollständig, wird eine ’1’ hinterlegt, ansonsten eine ’0’. Die Spalte 10 enthält die maximale Differenz der Wortgrenzen zwischen den orthographischen und phonetischen Alignments. Die Spalten 11 und 12 enthalten den Start- und den Endabtastwert des Satzes in der zugehörigen Audiodatei.

## 6.1 Struktur

Der CARInA besitzt eine hierarchische Struktur. Eine graphische Darstellung dieser Struktur ist in Abbildung 6.1 dargestellt. In dem Ordner `CARInA` befinden sich die Textdateien `README.txt`, `ContentStatus.txt` und `MissingSentences.txt` sowie die Unterordner `Complete` und `WorkInProgress`.

```
+---CARInA
| +---README.txt
| +---ContentStatus.txt
| +---MissingSentences.txt
| +---Complete
| | +---SpeakerID0001_f
| | | +---article0004_sentence0022.par
| | | +---article0004_sentence0022.TextGrid
| | | +---article0004_sentence0022.wav
   .
   .
| | +---SpeakerID0002_f
 .
 .
| | +---SpeakerID0337_u
| +---WorkInProgress
| | +---SpeakerID0001_f
| | | +---article0004_sentence0004.par
| | | +---article0004_sentence0004.TextGrid
| | | +---article0004_sentence0004.wav
 .
 .
| | | +---article0004_sentence0022.par.PrrBITSUS.snippet
| | | +---article0004_sentence0022.par.PrrKCSGrs.snippet
| | | +---article0004_sentence0022.par.PrrSRNC.snippet
| | | +---article0004_sentence0022.par.PyToBI.snippet
| | | +---article0004_sentence0022.TextGrid.PrrBITSUS.snippet
| | | +---article0004_sentence0022.TextGrid.PrrKCSGrs.snippet
| | | +---article0004_sentence0022.TextGrid.PrrSRNC.snippet
| | | +---article0004_sentence0022.TextGrid.PyToBI.snippet
   .
   .
| | +---SpeakerID0002_f
 .
 .
| | +---SpeakerID0337_u
```

Abbildung 6.1: Dateistruktur des CARInA

Die Datei `README.txt` enthält alle Informationen, welche zur Nutzung des Korpus benötigt werden.

Die Datei `ContentStatus.txt` wurde mithilfe der Liste `ContentStatus.mat` erstellt und enthält acht Spalten. Die Spalten sind durch einen Tabulator getrennt. Ein Auszug dieser Datei ist in Abbildung 6.2 dargestellt. Die erste Zeile erklärt die acht Spalten. Die erste Spalte enthält den Identifikationsnamen und den Namen des entsprechenden Satzes. Die Spalten 2–7 enthalten Angaben zu dem Alignment auf Wortebene, den Wortarten, der Silbifizierung, der kanonischen Aussprache, dem phonetischen Alignment und den Wortakzenten. Diese Spalten beinhalten eine ’0’ für eine unvollständige Annotation dieser Ebene und eine ’1’ für eine vollständige Annotation. Die achte Spalte enthält Angaben zur prosodischen Annotation. Jedes für den Satz verwendete Annotationssystem ist aufgeführt. Die verschiedenen Systeme werden mit einem ’|’ getrennt. Wurde der Satz nicht prosodisch annotiert, beinhaltet die Spalte eine ’0’.

```
File WordAlignment PartOfSpeech syllables canonical PhonesAlignment
stress ProsodyLevel
.
.
SpeakerID0001_f\article0004_sentence0020.wav 1 0 0 0 1 1 0
.
.
SpeakerID0001_f\article0343_sentence0004.wav 1 1 1 1 1 1
PrrBITSUS|PrrKCSGrs|PrrSRNC|PyToBI
.
.
```

Abbildung 6.2: Auszug aus der Datei `ContentStatus.txt`. Dargestellt ist die erste Zeile mit der Beschreibung der Spalten, eine Zeile mit unvollständiger Information und eine Zeile mit vollständiger Information. Aufgrund der Zeilenlänge wurden diese in der Abbildung umgebrochen.

Die Datei `MissingSentences.txt` beinhaltet alle Sätze, welche nicht extrahiert werden konnten und besteht aus drei Spalten. Die Spalten sind durch einen Tabulator getrennt. Die erste Spalte beinhaltet den Identifikationsnamen der Person und die Artikelnummer. Die zweite Spalte beinhaltet die Satznummer innerhalb des jeweiligen Artikels. Die dritte Spalte beinhaltet den Inhalt des Satzes.

Die Ordner `Complete` und `WorkInProgress` sind identisch aufgebaut. Sie enthalten für jede Sprecherin und jeden Sprecher einen Unterordner mit dem zugehörigen Identifikationsnamen. In diesen Unterordnern ist für jeden gesprochenen Satz eine Partitur-, eine TextGrid- und eine Audiodatei vorhanden.

Der Ordner `Complete` enthält den Teilkorpus *Complete*, welcher vollständige phonetische, kanonische, silbische und wortartenbezogene Information enthält. Der Ordner

`WorkInProgress` enthält alle Sätze, welche aus dem GSWC extrahiert werden konnten und unvollständig annotiert sind. Zusätzlich enthält der Ordner `WorkInProgress` für jede Audiodatei aus dem Ordner `Complete` acht Dateien mit der Endung `.snippet`, welche die generierten prosodischen Informationen beinhalten.

## 6.2 Inhalt der Dateien

Für jeden Satz aus der Liste `ContentStatus.txt` existieren drei Dateien. Die Dateinamen eines Satzes unterscheiden sich ausschließlich in der Endung. Die Namen der Sätze wurden aus den Listen `Annotation_*.mat` extrahiert und entsprechen dem Format `article****_sentence****` (vergleiche Abschnitt 3.3).

Aus den in Abschnitt 3.3 erläuterten Gründen werden sowohl für die orthographischen als auch für die phonetischen Alignments die Zeitpunkte der phonetischen Alignments genutzt, welche mit MAUS erstellt wurden.

Aufgrund der niedrigen Qualität automatisch erstellter prosodischer Etiketten wurden diese nicht zu den Dateien des Ordners `Complete` hinzugefügt. Die zu den Audiodateien aus dem Ordner `Complete` zugehörigen Dateien sind in dem Ordner `WorkInPogress` unter dem Identifikationsnamen der jeweiligen Person zu finden. Für die Audiodateien aus dem Ordner `WorkInProgress` wurden keine prosodischen Etiketten bestimmt.

### 6.2.1 Audio

Die Audiodatei eines Satzes wird aus dem Audio des gesprochenen Artikels extrahiert. Die Endung der Audiodatei ist `.wav`. Alle Audiodateien weisen einen Kanal, eine Abtastrate von 44 100 Hz und eine Auflösung von 16 Bit pro Abtastwert auf.

### 6.2.2 Partitur

Die Partitur eines Satzes wurde nach dem Partiturformat des BAS erstellt (vergleiche Abschnitt 2.2.1). Ein Beispiel für eine Partitur aus dem Teilkorpus *Complete* ist in Abbildung 6.3 dargestellt.

Der Kopf einer Partiturdatei besteht aus den in Abschnitt 2.2.1 beschriebenen obligatorischen Etiketten. Zusätzlich wurden der Name des Korpus (`DBN:`) und die Dauer der zugehörigen Audiodatei in Abtastwerten (`RET:`) angegeben.

Der Rumpf beinhaltet Angaben zur Orthographie (`ORT:`), zur kanonischen Realisierung (`KAN:`), zu den Wortarten (`POS:`), zu den Alignments auf Wortebene (`WOR:`) und zu den Alignments auf phonetischer Ebene (`MAU:`). Für die Dateien des Ordners `WorkInProgress` sind alle vorhandenen Informationen in der Partiturdatei angegeben.

Nach den Empfehlungen des BAS sind die Angaben zur Wortart aus dem STTS für deutsche Sprache zu entnehmen [80]. Die vorliegenden Wortarten lassen sich nicht in das STTS überführen. Aus diesem Grund wurden die in Tabelle 4.2 aufgelisteten Etiketten für die entsprechenden Wortarten genutzt.

Für die graphemische Silbentrennung existiert kein eigenes Etikett. Die Wörter der Etiketten `ORT:` sind Bezugsgrößen und nach Möglichkeit nicht zu verändern [80]. Die Information der Silbentrennung wurde in den Etiketten `WOR:` hinterlegt. Einzelne Silben sind durch ein Leerzeichen getrennt.

In den Etiketten `ORT:` sind neben den Wörtern auch die Satzzeichen hinterlegt. Diese wurden je nach Satzzeichen voran- oder nachgestellt.

```
LHD: Partitur 1.2
DBN: CARInA
REP: unknown
SNB: 2
SAM: 44100
SBF: 01
SSB: 16
NCH: 1
SPN: SpeakerID0008_f
RET: 94816
LBD:
ORT: 0 der
ORT: 1 Kopf
ORT: 2 der
ORT: 3 Figur
ORT: 4 wurde
ORT: 5 besch"adigt.
KAN: 0 d e:6
KAN: 1 k O pf
KAN: 2 d e:6
KAN: 3 f i " g u:6
KAN: 4 " v U R d @
KAN: 5 b @ " S E: d I C t
POS: 0 article
POS: 1 noun
POS: 2 article
POS: 3 noun
POS: 4 verb
POS: 5 verb
WOR: 0 11466 0 der
WOR: 11466 11466 1 Kopf
WOR: 22932 7938 2 der
WOR: 30870 23373 3 Fi gur
WOR: 54243 13671 4 wur de
WOR: 67914 26902 5 be sch"a digt
.
.
MAU: 30870 1323 3 f
MAU: 32193 3528 3 i
MAU: 35721 3969 3 g
MAU: 39690 7056 3 "u:
MAU: 46746 7497 3 6
.
.
```

Abbildung 6.3: Auszug aus der Partitur `article0211_sentence0032.par` der Sprecherin `SpeakerID0008_f`

### 6.2.3 TextGrid

Das verbreitetste Programm zur Bearbeitung von Sprachdaten ist Praat [11]. Für das Einlesen und Annotieren ist das Datenformat *TextGrid* notwendig. Für die Erstellung der `.TextGrid` Dateien wurde das Programm mPraat genutzt [12].

Die Informationen der TextGrid-Datei entsprechen weitestgehend der zugehörigen Partitur. Ein Beispiel der zu Abbildung 6.3 zugehörigen TextGrid-Datei ist in Abbildung 6.4 gegeben.

Die TextGrid-Dateien unterscheiden sich von der Partitur in der Darstellung der Umlaute. In der Partitur werden Umlaute mit einem `"` codiert. Beispielsweise wird ein ’Ä’ als ’`"`A’ dargestellt. In der TextGrid-Datei wurde die Darstellung der Umlaute nicht verändert.

Jede TextGrid-Datei enthält fünf Ebenen. Die erste Ebene mit dem Namen *phrase* beinhaltet die vollständige graphemische Aussage inklusive Satzzeichen. Die zweite Ebene mit dem Namen *words* beinhaltet die graphemischen Wörter inklusive Silbentrennung. Die dritte Ebene mit dem Namen *phones* beinhaltet die phonetische Realisierung des Satzes. Die vierte Ebene mit dem Namen *canonic* beinhaltet die kanonische Realisierung der Wörter. Die fünfte Ebene mit dem Namen *part of speech* beinhaltet die Wortarten der Wörter.

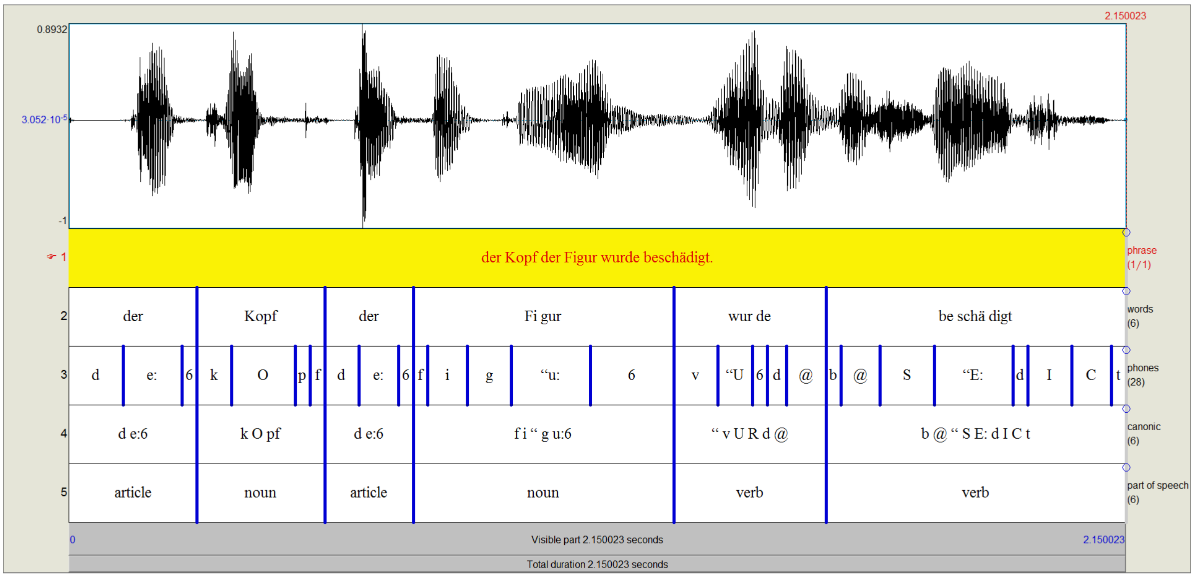


Abbildung 6.4: TextGrid-Datei des Satzes `article0211_sentence0032.par` der Sprecherin `SpeakerID0008_f`

### 6.2.4 Snippet

Mit den Systemen PRR BITS-US, PRR KCSGrs, PRR SRNC und PyToBI wurden prosodische Etiketten generiert. Die Informationen wurden im Partiturformat des BAS und im TextGrid-Format abgespeichert. Da für beide Dateien vier Systeme zur Prosodievorhersage genutzt wurden, ergeben sich acht `.snippet` Dateien pro Audiodatei.

Die Namen der Partiturdateien setzen sich aus dem Namen der zugehörigen Partitur und der Endung `.*.snippet` zusammen, wobei '*' für das jeweilige Annotationssystem steht. Die Datei beinhaltet die prosodischen Etiketten und kann an die zugehörige Partiturdatei angehängt werden.

Das System PyToBI erstellt separate prosodische Etiketten für Pausen und Tonakzente. Für diese Informationen existiert das Etikett `PRB:`, welches die Zuweisung einer Pause und eines Grenztones zu einem Zeitpunkt erlaubt. Ein Beispiel für eine `par.PyToBI.snippet` Datei ist in Abbildung 6.5 gezeigt.

```
PRB: 13671 1 TON: L*+H
PRB: 20286 1 BRE: 2
PRB: 37485 -1 BRE: 4; TON: L-L%
PRB: 49832 3 TON: L*
PRB: 52038 3 BRE: 2
PRB: 55566 -1 BRE: 4; TON: L-L%
.
.
```

Abbildung 6.5: Auszug aus der Snippet-Datei `article0004_sentence0022.par.PyToBI.snippet` der Sprecherin `SpeakerID0001_f`

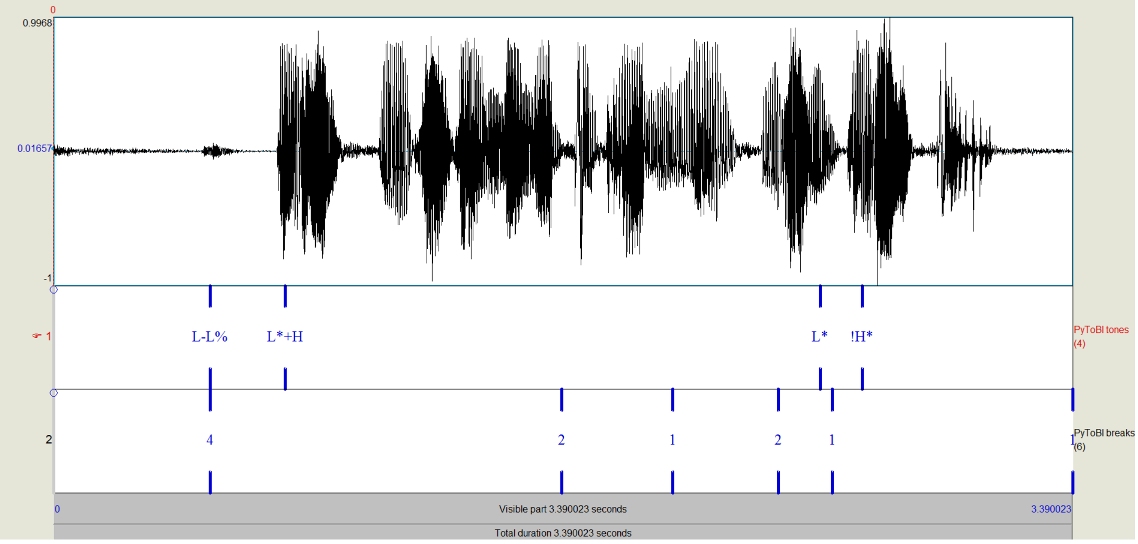
Die mit dem PRR trainierten Systeme erzeugen die Etiketten für Tonakzente und Pausen in einer Ebene. Hierfür eignet sich das Etikett `PRM:`, welches genau ein prosodisches Etikett einem Zeitpunkt zuweist. Ein Beispiel für eine `par.Prr*.snippet` Datei ist in Abbildung 6.6 gezeigt.

```
PRM: 23814 H*L
PRM: 43879 H*L
PRM: 63724 L*H
PRM: 124803 L*H
PRM: 150822 L*H
PRM: 171328 L*H
PRM: 197127 H*L
PRM: 218074 H*L
PRM: 225351 %
PRM: 231084 %
```

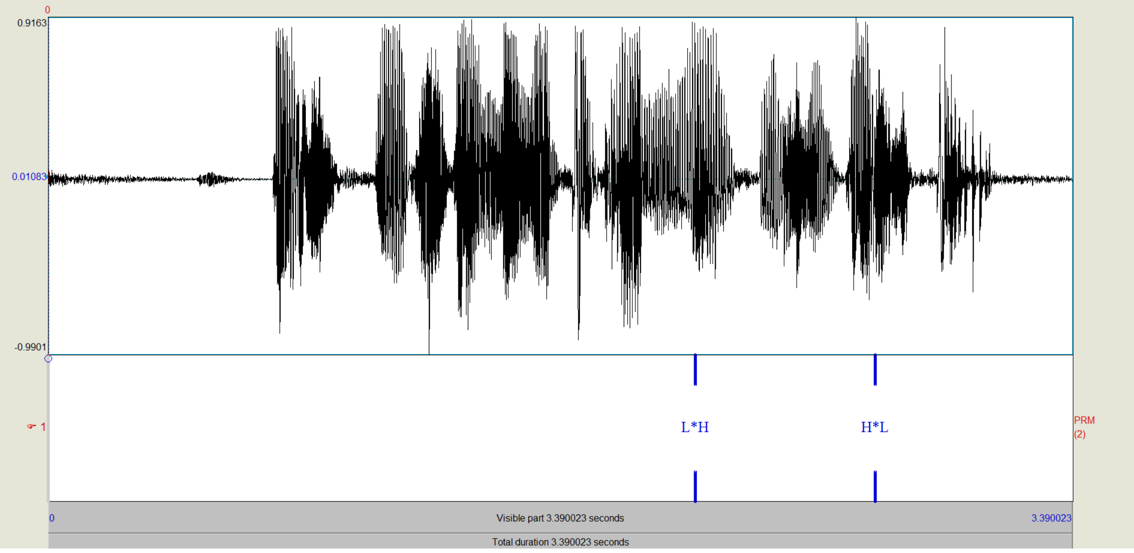
Abbildung 6.6: Inhalt der Snippet-Datei `article0004_sentence0022.par.PrrBITSUS.snippet` der Sprecherin `SpeakerID0001_f`

Die Namen der TextGrid-Dateien setzt sich aus dem Namen der zugehörigen TextGrid-Datei und der Endung `.*.snippet` zusammen, wobei '`*`' für das jeweilige Annotationssystem steht. Die Datei beinhaltet die prosodischen Etiketten und kann einzeln geöffnet werden beziehungsweise zu der zugehörigen TextGrid-Datei hinzugefügt werden. Das Hinzufügen ist beispielsweise in MATLAB mit dem Programm mPraat möglich.

Für das Programm PyToBI weist die TextGrid-Datei zwei Ebenen mit den Namen `PyToBI tones` und `PyToBI breaks` auf. Ein Beispiel hierfür ist in Abbildung 6.7a dargestellt. Für die Programme, welche mit dem PRR trainiert wurden, weist die TextGrid-Datei eine Ebene mit dem Namen `PRM` auf. Ein Beispiel hierfür ist in Abbildung 6.7b dargestellt.



(a) TextGrid-Datei `article0014_sentence0020.TextGrid.PyToBI.snippet` des Sprechers `SpeakerID0046_m`



(b) TextGrid-Datei `article0014_sentence0020.TextGrid.PrrBITSUS.snippet` des Sprechers `SpeakerID0046_m`

Abbildung 6.7: TextGrid-Datei mit prosodischen Informationen des Systems PyToBI und des Systems PRR BITS-US

# 7 CARInA – Auswertung und Validierung

Ein Überblick über den Umfang des CARInA ist in Abbildung 7.1 gegeben. Der Korpus beinhaltet 194:20 Stunden Sprachmaterial von 327 Personen. Unter den 327 Personen sind 34 Sprecherinnen, 259 Sprecher und 34 Personen unbekannten Geschlechts. Für 10 Personen konnte aufgrund der fehlenden Alignments kein Satz aus einem entsprechenden Artikel extrahiert werden.

Der vollständig phonetisch alignierte Anteil des CARInA beinhaltet 124:15 Stunden Sprachmaterial von 323 Personen. Unter den 323 Personen sind 34 Sprecherinnen, 256 Sprecher und 33 Personen unbekannten Geschlechts.

Der Teilkorpus *Complete* beinhaltet 29:47 Stunden Sprachmaterial von 291 Personen. Unter den 291 Personen sind 30 Sprecherinnen, 230 Sprecher und 31 Personen unbekannten Geschlechts.

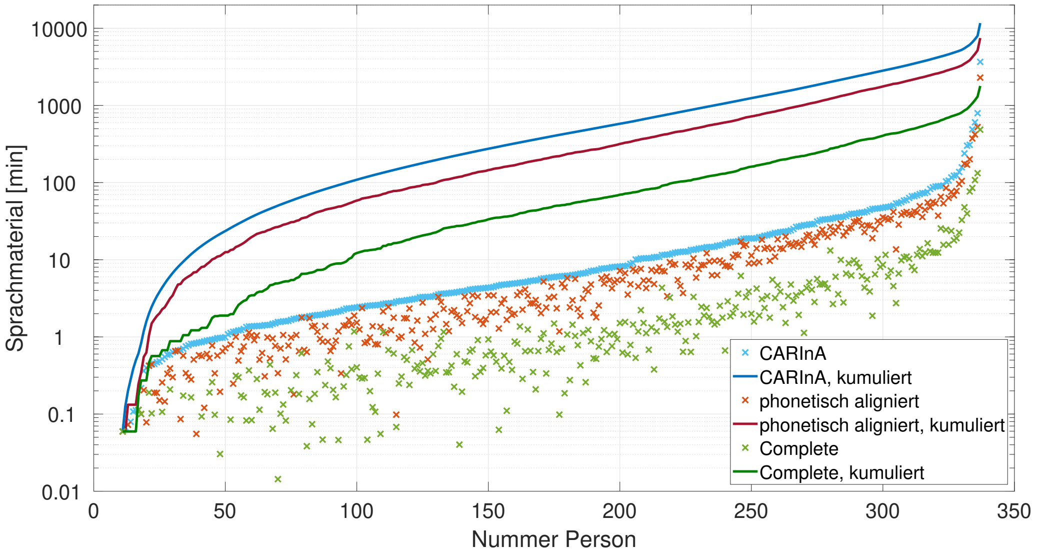


Abbildung 7.1: Sprachmaterial des CARInA, des vollständig phonetisch alignierten Anteils des CARInA und des Teilkorpus *Complete*. Das Sprachmaterial ist jeweils pro Person und kumuliert aufgetragen.

## 7.1 *Complete* – Personenbezogene Statistiken

Der Teilkorpus *Complete* bildet den wichtigsten Teil des CARInA, da er zu jeder Untermenge des CARInA hinzugefügt werden kann. Im Folgenden werden die Eigenschaften des *Complete* aufgeführt.

In Abbildung 7.2 ist das Sprachmaterial des Teilkorpus *Complete* pro Person und kumuliert aufgetragen. Aus der halblogarithmischen Abbildung ist zu erkennen, dass das meiste Sprachmaterial von wenigen Personen stammt. Die mittlere Dauer des Sprachmaterials der 291 Personen liegt bei 6:08 Minuten, der Median bei 56 Sekunden.

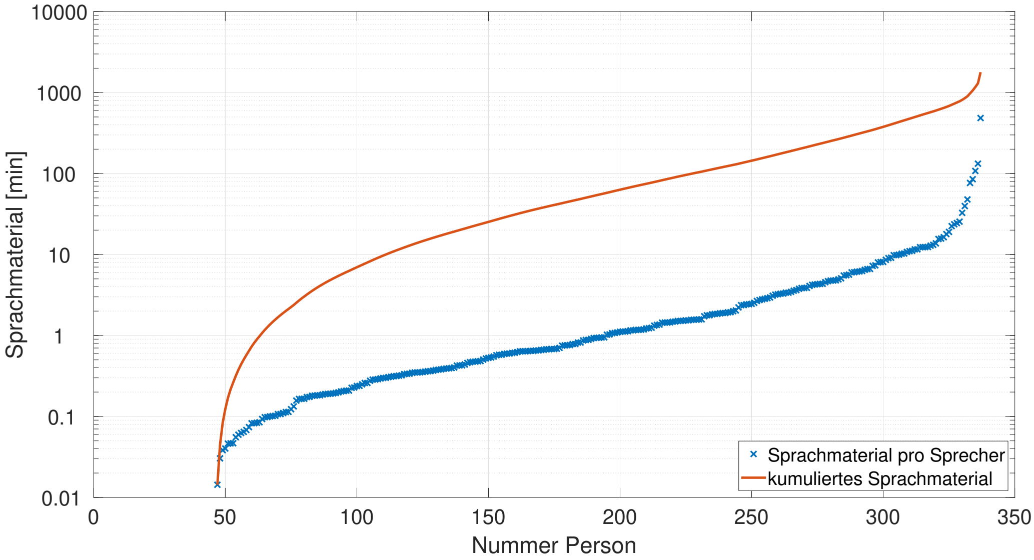


Abbildung 7.2: Sprachmaterial des Teilkorpus *Complete*. Das Sprachmaterial ist pro Person und kumuliert augetragen

In Abbildung 7.3 ist das Sprachmaterial des Teilkorpus *Complete* aller Personen mit mindestens 20 Minuten Sprachmaterial aufgetragen. 12 Personen erfüllen dieses Kriterium. Der Sprecher mit dem Identifikationsnamen `SpeakerID0041_m` weist mit 8 Stunden Sprachmaterial den größten Anteil auf. Insgesamt liegt das Sprachmaterial dieser 12 Personen bei 18 Stunden. 61 % des Sprachmaterials aus dem Teilkorpus *Complete* werden somit von 4 % der Personen gesprochen.

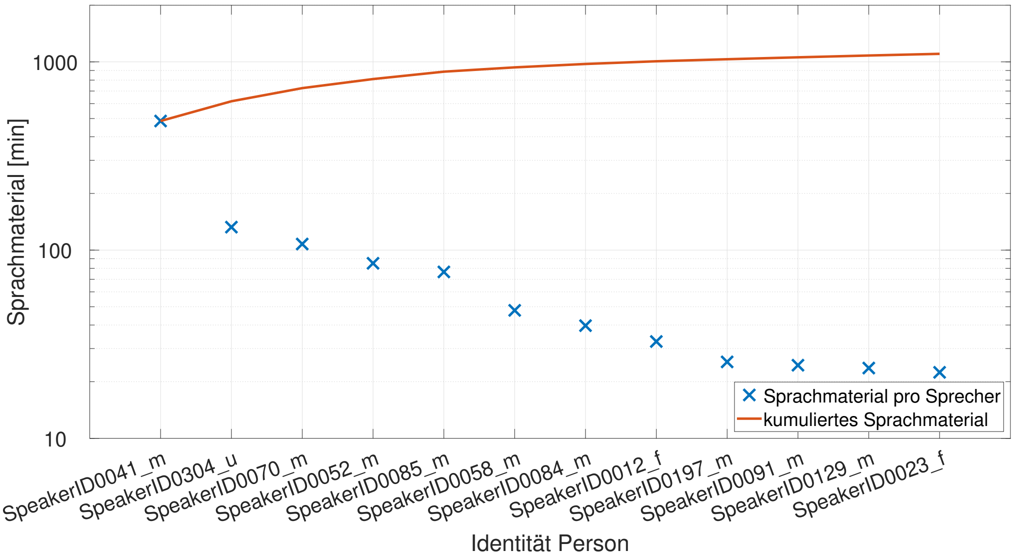


Abbildung 7.3: Sprachmaterial des Teilkorpus *Complete* von allen Personen mit mehr als 20 Minuten Sprachmaterial. Das Sprachmaterial ist pro Person und kumuliert augetragen

Für Anwendungen, in welchen Sprachaufnahmen von wenigen Personen benötigt werden, steht somit umfangreiches Trainingsmaterial zur Verfügung.

Über alle gesprochenen Phoneme hinweg wurden die Sprechgeschwindigkeiten der Personen bestimmt. Pausen bleiben in dieser Statistik unberücksichtigt. In Abbildung 7.4 sind die Sprechgeschwindigkeiten der ersten 12 Personen angegeben.

Die Personen des KCSGss weisen eine durchschnittliche Sprechgeschwindigkeit von 13.48 bis 15.24 Phonemen pro Sekunde auf [71]. Die mittleren Sprechgeschwindigkeiten des Teilkorpus *Complete* liegen zwischen 14.28 und 16.67 Phonemen pro Sekunde. Somit sind die Sprechgeschwindigkeiten der gelesenen Sprache höher als die der spontanen Sprache.

Im Allgemeinen ist die Sprechgeschwindigkeit bei spontanen Äußerungen deutlich höher als bei gelesener Sprache. Im amerikanischen liegt die Sprechgeschwindigkeit spontaner Sprache um ca. 50 % über der Sprechgeschwindigkeit gelesener Sprache [38].

Eine mögliche Erklärung dieser Diskrepanz kann in der Beschaffenheit des KCSGss liegen. Die Sprachaufnahmen des Korpus beinhalten unter anderem Dialoge zur Terminabstimmung [41]. Diese Dialoge bestehen zum Teil aus Häsitationspartikeln. Auch werden einige Wörter langsam ausgesprochenen, da beim Sprechen überlegt wird. Eine genauere Untersuchung dieses Umstandes ist an dieser Stelle nicht möglich, da aus den Daten des KCSGss nur Beispieldateien vorliegen.

Nach einer amerikanischen Studie korreliert die Sprechgeschwindigkeit für gelesene Sprache nicht mit dem Geschlecht [38]. Für spontane Sprache liegt die Sprechgeschwindigkeit für männliche Personen deutlich oberhalb der von weiblichen Personen [38]. Dies deckt sich mit den Ergebnissen. Ein Zusammenhang der Sprechgeschwindigkeiten mit dem Geschlecht ist nicht zu erkennen.

Die Varianz des Sprechtempos ist bei den Personen `SpeakerID0304_u`, `SpeakerID0052_m` und `SpeakerID0085_m` am Geringsten.

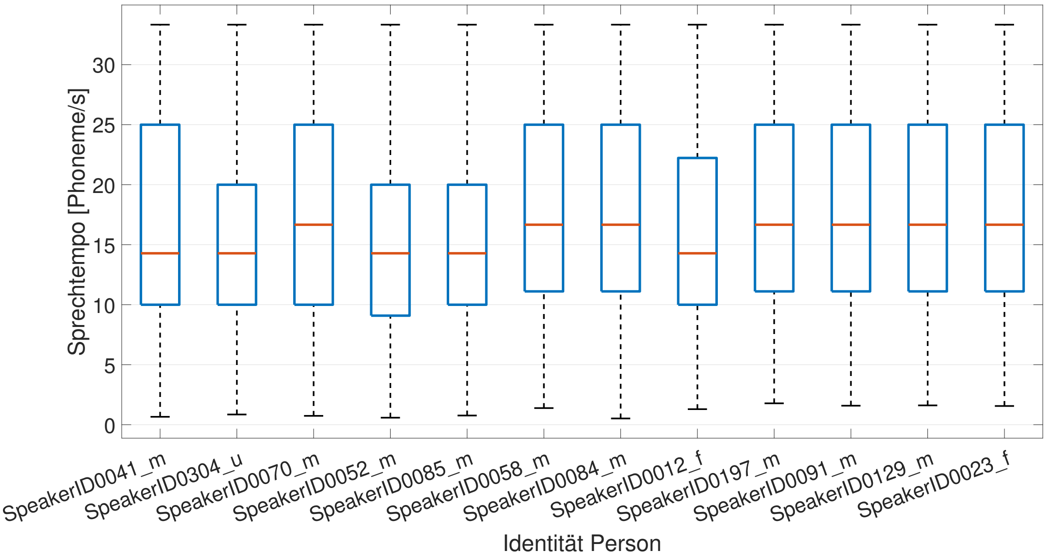


Abbildung 7.4: Sprechgeschwindigkeit von allen Personen, welche mehr als 20 Minuten Sprachmaterial im Teilkorpus *Complete* aufweisen. Das Sprechtempo wurde über alle im *Complete* gesprochenen Phoneme bestimmt.

Für den Teilkorpus *Complete* wurde eine Grundfrequenzanalyse durchgeführt. Die Grundfrequenzen jeder Audiodatei wurden mithilfe der Funktion `To Pitch` mittels Praat bestimmt. Es wurden die Standardeinstellungen der Funktion gewählt, welche Frequenzen von 75 Hz bis 600 Hz zulassen [11]. In Abbildung 7.5 sind die Grundfrequenzen der ersten 12 Personen dargestellt. Durch die Wahl der Standardeinstellungen in Praat entstanden Ausreißer bis 600 Hz. Für eine bessere Übersichtlichkeit wurden diese nicht dargestellt.

Der Grundfrequenzbereich für Sprecher liegt bei 70 Hz bis 160 Hz mit einer Mittelfrequenz von ca. 120 Hz [66]. Für Sprecherinnen liegt der Grundfrequenzbereich bei 180 Hz bis 330 Hz mit einer Mittelfrequenz von ca. 240 Hz [66].

Die Grundfrequenz eines Sprecher liegt somit im Regelfall deutlich unterhalb der Grundfrequenz einer Sprecherin. Dies ist für die meisten Personen des Korpus *Complete* zutreffend.

Die Person mit der Identitätsnummer `SpeakerID0129_m` weist mit einer Mittelfrequenz von 165 Hz für eine männliche Person eine sehr hohe Grundfrequenz auf. Eine nähere Analyse der Audiodateien bestätigte die hohe Grundfrequenz.

Die Person mit der Identitätsnummer `SpeakerID0023_f` weist mit einer Mittelfrequenz von 98 Hz für eine weibliche Person eine sehr niedrige Grundfrequenz auf. Wie aus Tabelle 3.2 hervorgeht, ist das subjektiv zugeordnete Geschlecht dieser Person männlich.

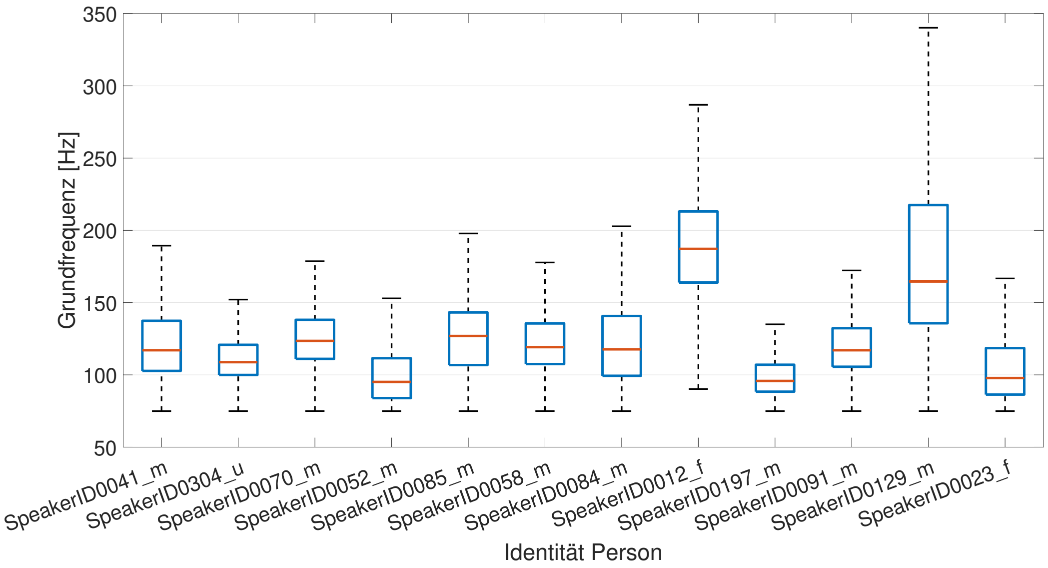


Abbildung 7.5: Grundfrequenz von allen Personen, welche mehr als 20 Minuten Sprachmaterial im Teilkorpus *Complete* aufweisen. Die Grundfrequenz wurde über das gesamte zugehörige Sprachmaterial des Teilkorpus *Complete* bestimmt. **Die Ausreißer wurden der Übersichtlichkeit halber ausgeblendet.**

Für nahezu jede Anwendung ist die Qualität der Sprachdateien von Bedeutung. Maßgebend für die Qualität sind die verwendeten Aufnahmegeräten und die Intensität des Umgebungsrauschens während der Aufnahme. Diese Informationen sind für die Aufnahmen des CARInA nicht vorhanden. Ein weiteres Maß für die Qualität ist die Angabe des Signal-Rausch-Verhältnisses (engl. *Signal-to-Noise Ratio* (SNR)). Damit das SNR einer Audiodatei nicht von der Länge der enthaltenden Stille abhängig ist, werden Pausen herausgefiltert [36]. Für die Berechnung des SNR wird der Effektivwert der Sprachaufnahmen ohne Pausen ($S_{rms}$) und der Effektivwert der Pausen ($N_{rms}$) berechnet. Das SNR der Sprachaufnahme berechnet sich durch

$$SNR = 20 \cdot \lg \left( \frac{S_{rms}}{N_{rms}} \right) \tag{7.1}$$

und wird in dB angegeben [36].

In den Dateien des Teilkorpus *Complete* ist Stille durch das Etikett '<p:>' gekennzeichnet. Zeitspannen mit diesem Etikett wurden zur Berechnung des $N_{rms}$ verwendet. Das $S_{rms}$ wurde durch alle anderen Zeitspannen berechnet. Da eine Kontinuität der Störeinflüsse über einen Artikel hinweg nicht angenommen werden kann, wurde für jeden Satz des Teilkorpus *Complete* das SNR bestimmt.

Die berechneten SNR sind als untere Schranke zu betrachten. Während Pausen erzeugte Geräusche wie lautes Atmen verringern das SNR des entsprechenden Satzes. SNR von 30 dB sind als einwandfreie Aufnahme zu betrachten. SNR von 20 dB werden von zuhörenden Personen kaum wahrgenommen und bei SNR von 0 dB ist eine gute Verständlichkeit gewährleistet [36].

Das mittlere SNR des Teilkorpus *Complete* liegt bei 26.8 dB und weist eine Standardabweichung von 8 dB auf. In Abbildung 7.6 ist die Verteilung des SNR für den Teilkorpus *Complete* dargestellt. 16 % der Sätze weisen ein SNR von weniger als 20 dB auf, 29 % ein SNR von mehr als 30 dB.

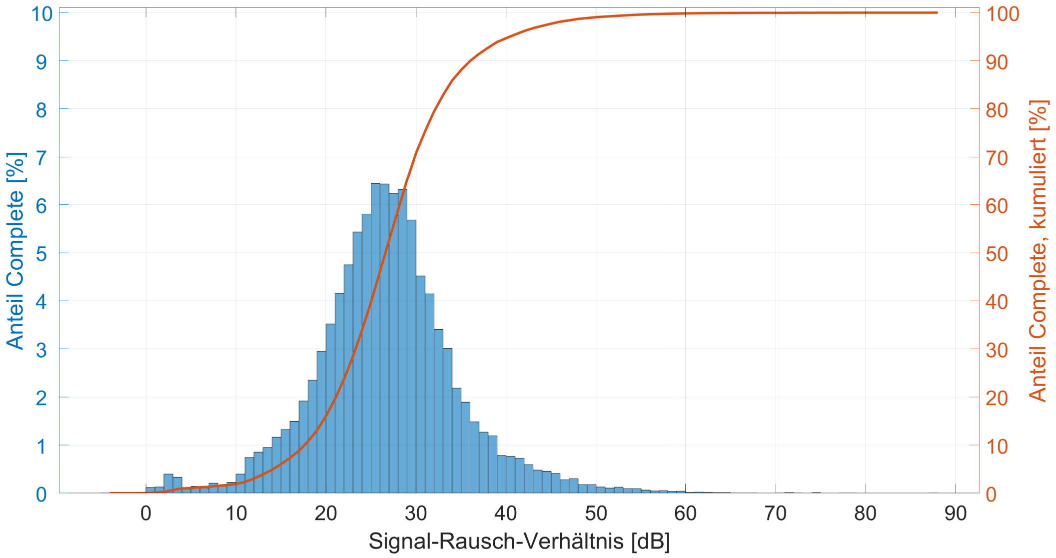


Abbildung 7.6: Verteilung der Signal-Rausch-Verhältnisse aller Sätze des Teilkorpus *Complete*. Die Verteilung ist als Histogramm und kumuliert aufgetragen.

In Abbildung 7.7 sind die SNR der ersten 12 Personen dargestellt. Bis auf drei Ausnahmen liegen die durchschnittlichen SNR zwischen 20 dB und 30 dB. Der Sprecher mit der Identifikationsnummer `SpeakerID0070_m` weist mit einem mittleren SNR von 38.5 dB das höchste SNR auf. Die Aufnahmen dieses Sprechers sind sehr rauscharm. Das mittlere SNR des Sprechers mit der Identifikationsnummer `SpeakerID0129_m` beträgt 17 dB. Eine subjektive exemplarische Überprüfung mehrerer Dateien von jedem gesprochenen

Artikel dieses Sprechers ergab, dass die Qualität der Aufnahmen hoch ist. Das geringe SNR ist die Folge geräuschvollen Luftholens während der Pausen.

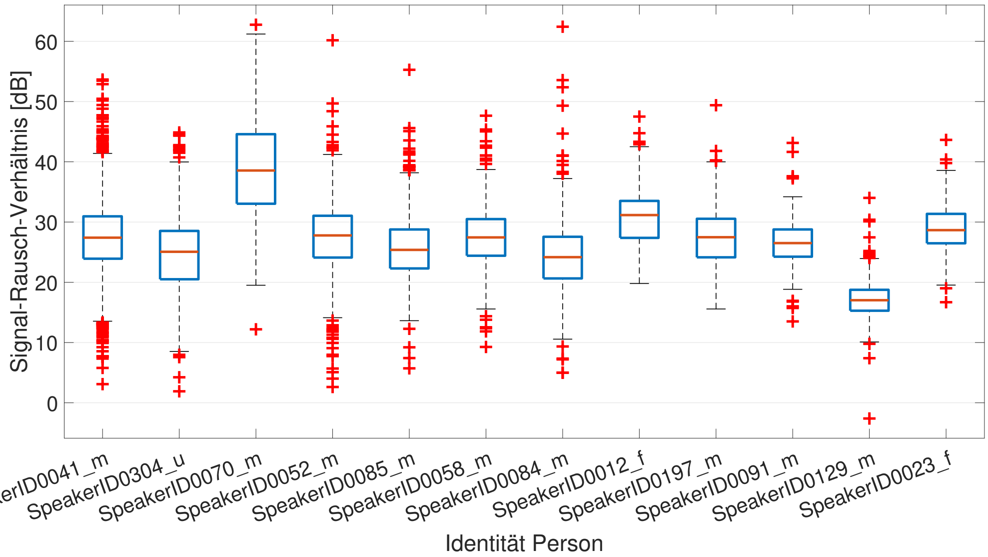


Abbildung 7.7: Signal-Rausch-Verhältnis von allen Sätzen aus dem Teilkorpus *Complete* der Personen, welche mehr als 20 Minuten Sprachmaterial im Teilkorpus *Complete* aufweisen.

## 7.2 *Complete* – Phonetische Statistiken

Zur Auswertung des Teilkorpus *Complete* wurde die Anzahl der enthaltenen Phoneme bestimmt. Diese sind in Abbildung 7.8 dargestellt. Die relative Häufigkeit der Phoneme deckt sich weitestgehend mit der durchschnittlich auftretenden Phonemhäufigkeit deutscher Sprache [42]. Die letzten fünf Phoneme sind weniger als 350-mal im Teilkorpus *Complete* enthalten.

Es sind drei fremdsprachliche Laute vorhanden. Das Phonem [ã] entstammt dem französischen (beispielsweise in *chant* [ʃã] (Gesang)). Der Diphthong [əʊ] kommt im englischen vor (beispielsweise in *go* [gəʊ] (gehen)), genauso wie das Phonem [w] (beispielsweise in *wind* [wɪnd] (Wind)).

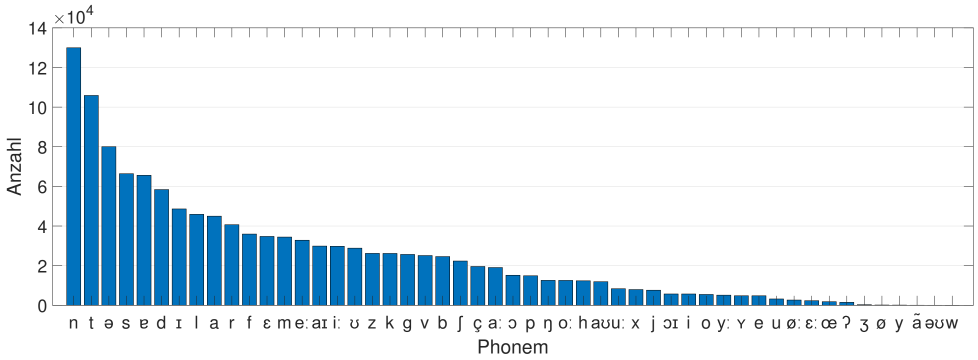


Abbildung 7.8: Verteilung der Phoneme im Teilkorpus *Complete*

Für jedes Phonem wurde über alle Ereignisse aus dem Teilkorpus *Complete* die mittlere Länge bestimmt. Diese sind in Abbildung 7.9 dargestellt. Die mittleren Phonemlängen des Teilkorpus *Complete* entsprechen denen des KCSGrs mit einer mittleren Abweichung von ca. 10 % [50].

Zehn Phoneme weisen eine mittlere Dauer von mindestens 100 ms auf. Unter diesen Phonemen sind die vier enthaltenen Diphthonge, fünf Vokale und der glottale Plosiv. Diphthonge und Vokale gehören zu den langen Phonemen [50]. Die durchschnittliche Dauer des glottalen Plosivs liegt mit 151 ms oberhalb der allgemeinen durchschnittlichen Dauer von ca. 120 ms [44]. Dies ist darauf zurückzuführen, dass der Startzeitpunkt des glottalen Plosivs an Wortanfängen schwer zu detektieren ist.

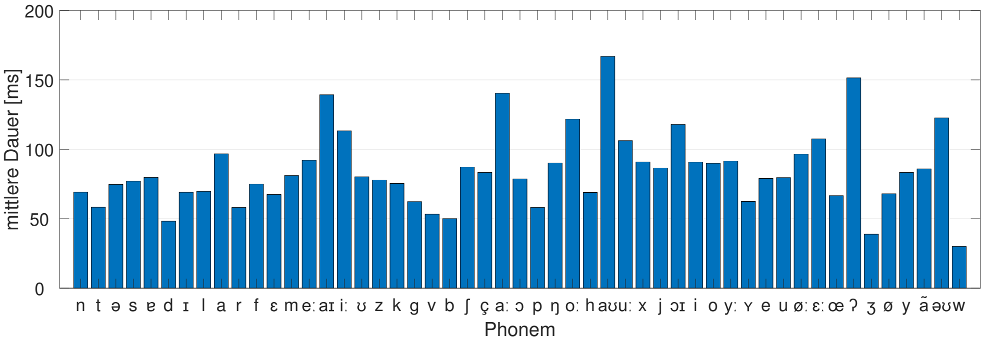


Abbildung 7.9: Durchschnittliche Länge der Phoneme im Teilkorpus *Complete*

In Abbildung 7.10 sind die Durchschnittswerte der ersten beiden Formanten der deutschen Vokale aufgetragen. Die Vokale [uː], [aː] und [iː] bilden das Vokaldreieck. Innerhalb dieses Dreiecks befinden sich die anderen Vokale [61]. Die Länge des Vokaltraktes von femininen Personen ist durchschnittlich kürzer als bei maskulinen Personen. Aus diesem Grund sind die Formantfrequenzen für feminine Personen höher als für maskuline Personen [91].

Mithilfe der Funktion `To Formant (burg)` wurden die Formanten für jeden Vokal aus dem Teilkorpus *Complete* mittels Praat bestimmt. Für die Funktion wurden die Standardeinstellungen verwendet, nach welchen für jeden Zeitbereich von 25 ms die ersten fünf Formanten bestimmt werden [11].

Für jeden Vokal wurden die Werte der ersten beiden Formanten im mittleren Drittel des Vokals bestimmt. Die extrahierten Formantfrequenzen entsprechen dem Median dieser Werte. Für die Vokale wurde der Mittelwert und die Standardabweichung bestimmt. Die nach Geschlechtern sortierten tabellarischen Ergebnisse sind der Tabelle A.2 in Abschnitt A.2 aufgelistet. Die Ergebnisse sind graphisch in Abbildung 7.11 dargestellt.

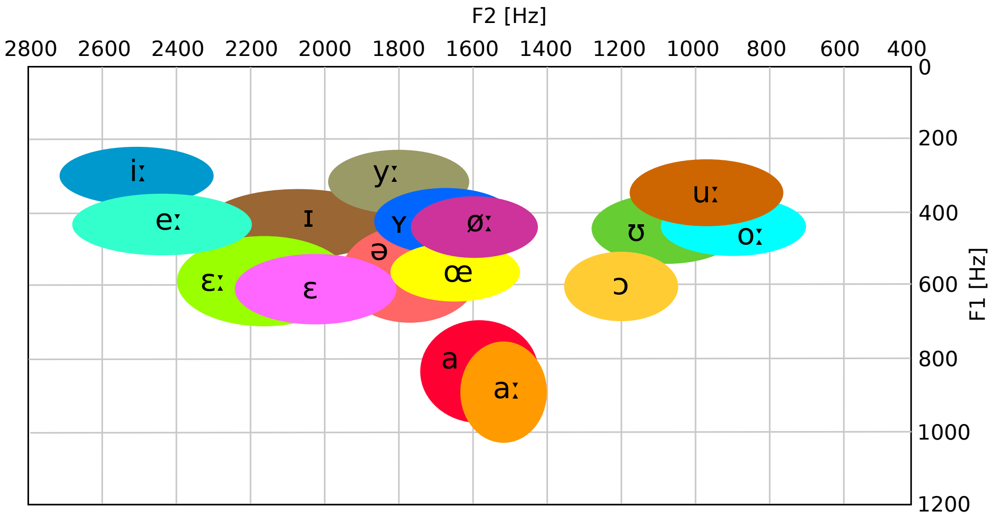
Aus dem Vergleich der Abbildung 7.11a mit der Abbildung 7.11b ist zu erkennen, dass die weiblichen Personen für nahezu jeden Vokal höhere Formantenfrequenzen aufweisen als männliche Personen. Der Grund hierfür ist die durchschnittlich kürzere Länge des Vokaltraktes bei weiblichen Personen.

Es sind einige Differenzen zwischen der Abbildung 7.11 und den Referenzen in Abbildung 7.10 zu erkennen.

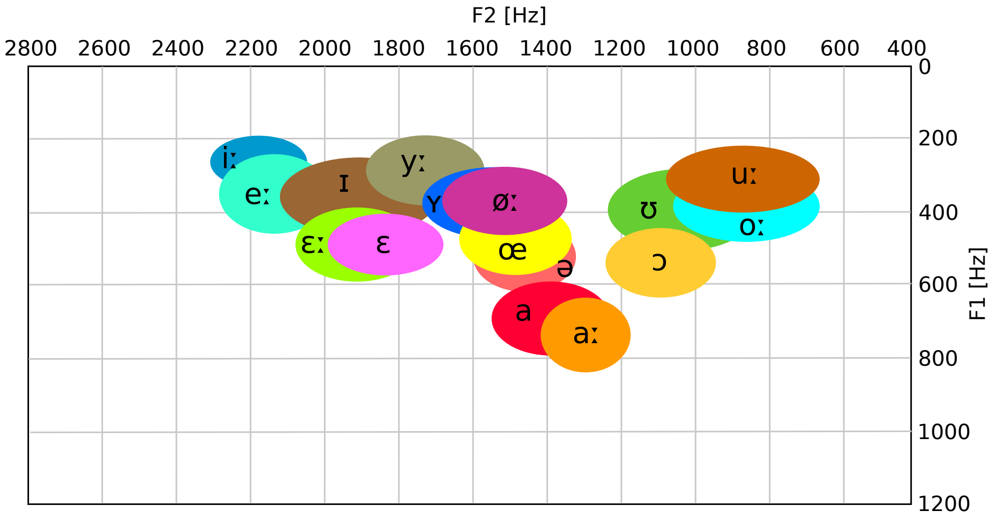
Der zweite Formant des Vokals [eː] weist eine deutlich geringere Frequenz auf als die Referenz. Der Grund hierfür liegt darin, dass der gesprochene Vokal [ɛ] zu einem signifikanten Anteil als Vokal [eː] etikettiert wurde. Wörter wie „der“ [deːɐ] und „er“ [eːɐ] werden oftmals mit einem [ɛ] realisiert ([dɛɐ] beziehungsweise [ɛɐ]). Hierdurch verschiebt sich der Vokal [eː] zu dem Vokal [ɛ].

Der erste Formant des Vokals [ə] weist eine deutlich geringere Frequenz auf als die Referenz. Dieser Unterschied entsteht aus zwei Gründen. Zum einen liegt der Vokal [ə] oftmals am Ende eines Wortes vor einer Pause. An diesen Übergangsstellen ist die Genauigkeit der Formantenmessung mittels Praat geringer als zwischen anderen stimmhaften Phonemen. Zum anderen wird der Vokal [ə] zwischen einem Plosiv und einem Nasal oft nicht ausgesprochen. In diesem Fall wurde der Vokal [ə] von MAUS aligniert, obwohl er nicht realisiert wurde.

Für die männlichen Personen weist der Vokal [uː] einen deutlich größeren Bereich für den zweiten Formanten auf als die Referenz. Die Ursache hierfür liegt darin, dass der Vokal [uː] oft in Zusammenhang stimmloser Frikative vorkommt. Dies sorgt für dieselbe Ungenauigkeit in der Formantenmessung von Praat wie eine Wortgrenze.

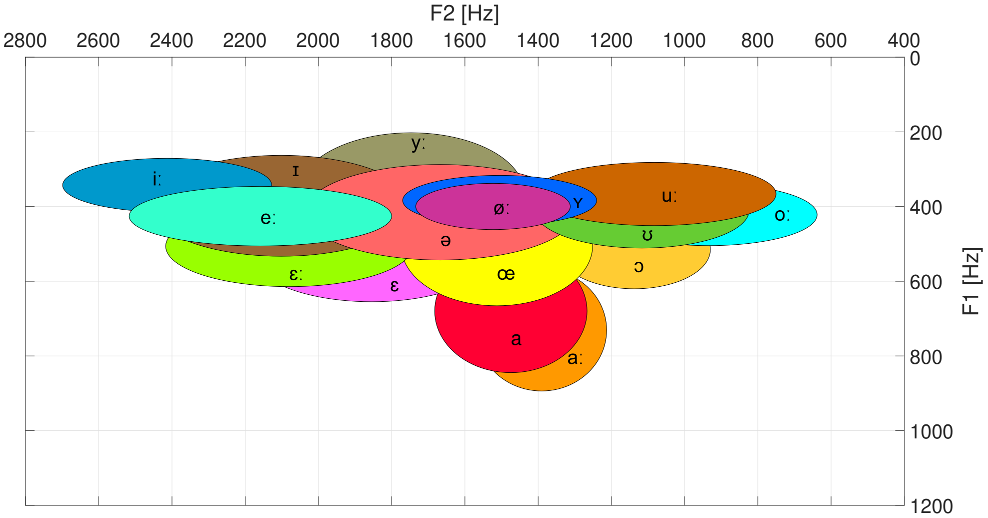


(a) Mittlere Formantwerte von 58 femininen Personen

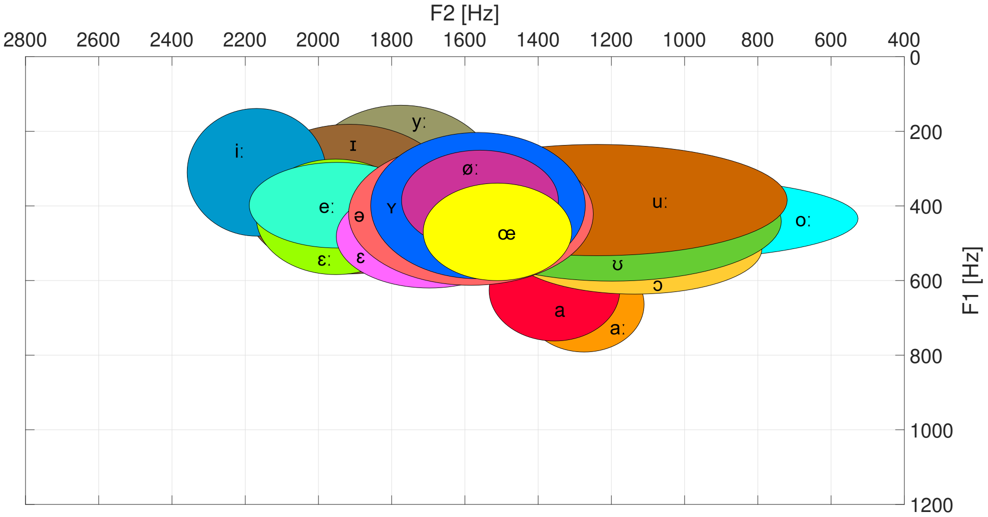


(b) Mittlere Formantwerte von 69 maskulinen Personen

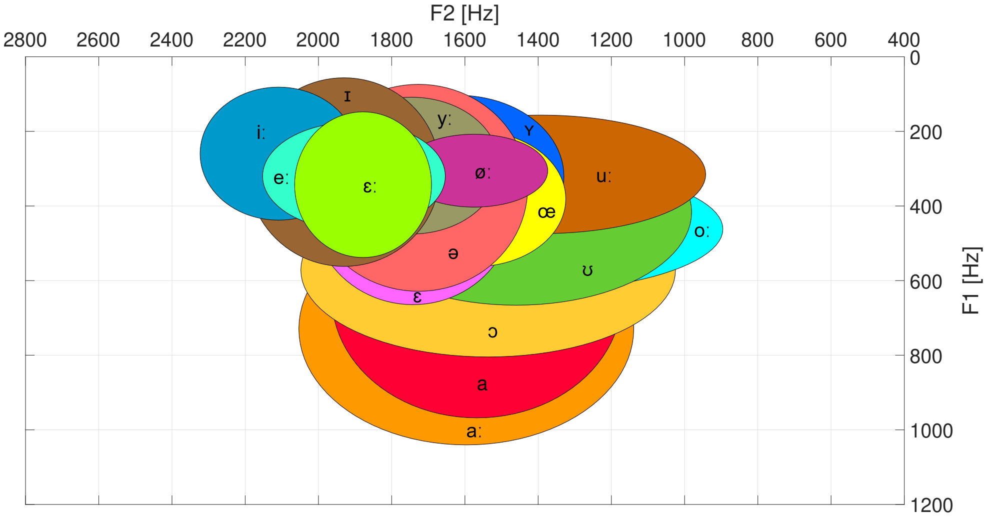
Abbildung 7.10: Durchschnittliche Formantwerte deutscher Vokale, konstruiert nach [91]. Die Frequenzwerte der Formanten wurden mit dem Programm Praat bestimmt. Die Standardabweichungen sind durch die ellipsenförmigen Umrandungen dargestellt.



(a) Formantwerte der femininen Personen



(b) Formantwerte der maskulinen Personen



(c) Formantwerte der Personen unbekannten Geschlechts

Abbildung 7.11: Formantwerte mit Standardabweichungen des Teilkorpus *Complete*

Im Allgemeinen sind die Varianzen der Formantfrequenzen größer als bei der Referenz. Da auch für die Erstellung der Referenz Daten von je mehr als 50 Personen aufgenommen wurden, liegt der Grund hierfür nicht an der umfangreichen Personenanzahl. Die größeren Varianzen entstehen daraus, dass Praat oft den ersten Formanten nicht erkannt hat. In diesen Fällen wurde der zweite Formant dem ersten zugeteilt. Der Grund für diese Verwechslung kann in der schnellen Sprechweise liegen, wodurch Vokale teilweise nicht stimmhaft realisiert werden. Ein anderer Grund kann die Qualität der Audiodateien sein. Wie in Abschnitt 7.1 beschrieben, liegt das mittlere SNR des Teilkorpus *Complete* bei 26.8 dB. Die Formantenerkennung wird durch größere SNR ungenauer. In Tabelle 7.1 sind die durchschnittlichen Werte der SNR angegeben. Der Vergleich dieser Tabelle mit den Varianzen zeigt, dass die Varianzen für niedrigere SNR ansteigen.

Tabelle 7.1: Signal-Rausch-Verhältnis der Daten aus dem Teilkorpus *Complete* nach Geschlechtern sortiert

| | **Signal-Rausch-Verhältnis [dB]** | |
|---|---|---|
| **Geschlecht** | Mittelwert | Standardabweichung |
| feminin | 28.4 | 6.5 |
| maskulin | 26.9 | 8.1 |
| unbekannt | 24.3 | 8 |

## 7.3 Training und Evaluation einer Spracherkennung

Für die Validierung des CARInA wurde ein CNN zur Spracherkennung trainiert. Der verwendete Algorithmus basiert auf dem in MATLAB implementierten Beispiel zur Spracherkennung *Speech Command Recognition Using Deep Learning* [46]. CNN erkennen im Gegensatz zu vorwärts gerichteten neuronalen Netzen Zusammenhänge zwischen benachbarten Datenpunkten. Aus diesem Grund eignet sich ein CNN besser für die Erkennung mehrdimensionaler Eingaben als ein einfaches vorwärts gerichtetes neuronales Netzwerk [54].

### 7.3.1 Beschreibung des Algorithmus

Die Eingabe des Algorithmus besteht aus einem in Trainings- und Validierungsdaten eingeteilten Datensatz und einer Liste mit zu erkennenden Wörtern. Zu dem Datensatz werden verschiedene Rauschkategorien hinzugefügt, welche für die Erkennung von Stille notwendig sind [46]. Die Audiodateien bestehen jeweils aus einem Wort und weisen eine Abtastrate von 16 kHz auf [46].

Das CNN arbeitet mit den Infomationen aus Spektrogrammen. Die Berechnung der Spektrogramme erfolgt mit einer Fensterlänge von 512 Abtastwerten, einer Überlappung von 99 % und 50 Frequenzgruppen. Das Trainieren und Validieren des CNN erfordert Eingaben gleicher Länge, weswegen die Dateien auf das längste Wort angepasst werden müssen. Die Anpassung erfolgt durch das Erweitern der Audiodatei, auch *padding* genannt. Die Audiodatei wird mit Nullen erweitern (Zero-Padding) [46].

Das Spektrogramm dient zur Erstellung eines Bark Spektrogramms, welches die Energie den wahrgenommene Tonhöhen zuordnet. Es wird ein vierdimensionales Objekt erstellt. Die erste Dimension repräsentiert das zeitliche Raster in Einheiten von 10 ms. Die zweite Dimension repräsentiert die 50 Frequenzgruppen. Die dritte Dimension repräsentiert die Anzahl der Eingangskanäle, welche für die Spektrogramme 1 ist. Die Spektrogramme werden entlang der vierten Dimension aufgelistet [46]. Für die Trainings- und Validierungsdaten wird je ein vierdimensionales Objekt und eine Liste mit den zugehörigen Etiketten entlang der vierten Dimension erstellt.

Die Architektur des Netzwerkes besteht aus 24 Schichten. Eine Übersicht des Aufbaus der Architektur ist in Tabelle 7.2 aufgelistet. Nachfolgend werden die verschiedenen Schichten kurz beschrieben.

**Eingabeschicht**

Die Eingabeschicht extrahiert die Daten für eine Klassifikation. Die Daten des zugehörigen Spektrogramms werden ausgelesen und an die nächste Schicht überreicht [54].

**Faltungsschicht**

Die Faltungsschicht projiziert die Eingabematrix mithilfe eines Filterkernels auf eine Ausgabematrix [54]. Die im CNN enthaltenen Faltungsschichten beinhalten Filterkernel der Größe 3x3 und die Schrittweite beträgt 1. Die erste Faltungsschicht beinhaltet 12, die zweite 24 und die folgenden 48 Filterkernel. Das Padding ist so gewählt, dass die Ausgabematrix die gleiche Größe wie die Eingabematrix aufweist [46].

**Batch-Normalisierungsschicht**

Die Batch-Normalisierungsschichten gleichen die Ausgangsdaten des aktuellen Datenpakets (engl. *Batch*) der vorherigen Schicht an. Der Ausgabewert des Neurons $\hat{x}_i$ berechnet sich mit dem Eingabewert $x_i$, dem Mittelwert $\mu_B$ und der Standardabweichung $\sigma_B$ des Batches und einer Stabilitätskonstante $\varepsilon$ zu

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}. \tag{7.2}$$

Durch die Batch-Normalisierungsschichten wird der Verarbeitungsprozess beschleunigt, da die Wertebereiche konstant bleiben. Außerdem hat die Batch-Normalisierungsschicht einen regularisierenden Faktor. Da für die Normalisierung nur die Daten des Batches genutzt werden, entsteht ein Rauschen in den Daten. Dies wirkt sich positiv auf die Überanpassung aus [37]

### ReLU-Schicht

Die *Rectified Linear Unit*-Schicht (ReLU-Schicht) dient als Gleichrichtungsschicht. Alle negativen Werte der neuronalen Gewichte werden auf 0 gesetzt, die positiven Werte bleiben unverändert. Das Training von neuronalen Netzen ist erfolgreicher mit der rektifizierenden Aktivierungsfunktion als beispielsweise mit der Sigmoid-Funktion, welche Anfang der 2000er-Jahre weitestgehen genutzt wurde [28].

### Pooling-Schicht

Die *Pooling*-Schicht dient zur Datenreduktion. Es wird das Max-Pooling für einen 3x3 Filter mit einer Schrittweite von 2 genutzt. Das Padding ist so gewählt, dass die Ausgabematrix die gleiche Größe wie die Eingabematrix aufweist. Durch das Pooling erhöht sich die Berechnungsgeschwindigkeit. Zusätzlich entsteht ein Rauschen in den Daten, wodurch Überanpassung reduziert wird [46].

### Dropout-Schicht

Die *Dropout*-Schicht dient zur Regularisierung und zur Vermeidung von Überanpassung. Beim Training werden zufällig ausgewählte Eingabepunkte mit 0 initialisiert. Die Wahrscheinlichkeit des Ausfalls eines Eingabepunktes beträgt 20 % [46].

### vollständig vernetzte Schicht

Die vollständig vernetzte Schicht besitzt für jede Klasse ein Ausgangsneuron. Mithilfe dieser Schicht erfolgt die eindeutige und ortsunabhängige Zuweisung der Eingabedaten zu einer Klasse [54].

### Softmax-Schicht

Die *Softmax*-Schicht projiziert die Eingabe auf eine Wahrscheinlichkeitsverteilung. Jeder Eingabe wird ein Wert zwischen 0 und 1 zugeordnet, wobei die Gesamtsumme der Ausgabe den Wert 1 aufweist. Ist für jede Klasse eine Eingabe vorhanden, wird mithilfe der Softmax-Schicht die Wahrscheinlichkeit jeder Klasse berechnet.

### gewichtete Klassifikationsschicht

Die gewichtete Klassifikationsschicht minimiert den Kreuzentropieverlust zwischen dem Klassenvektor und dem Ausgabevektor der Softmax-Schicht. Der Klassenvektor beinhaltet genau eine 1. Die anderen Einträge sind 0, da ein Objekt genau einer Klasse angehört. Der Ausgabevektor beinhaltet eine Wahrscheinlichkeitsfunktion. Die Kreuzentropie ist ein Maß für die Übereinstimmung zweier Vektoren. Die Annäherung der Wahrscheinlichkeitsfunktion an den Klassenvektor erfolgt mithilfe der gewichteten Klassifikationsschicht [46].

Tabelle 7.2: Zusammensetzung der Schichten des *Convolutional Neural Network* zum Training der Spracherkennung

| Schicht Nummer | Art der Schicht |
|---|---|
| 1 | Eingabeschicht |
| 2 | Faltungsschicht |
| 3 | Batch-Normalisierungsschicht |
| 4 | ReLU-Schicht |
| 5 | Pooling-Schicht |
| 6 | Faltungsschicht |
| 7 | Batch-Normalisierungsschicht |
| 8 | ReLU-Schicht |
| 9 | Pooling-Schicht |
| 10 | Faltungsschicht |
| 11 | Batch-Normalisierungsschicht |
| 12 | ReLU-Schicht |
| 13 | Pooling-Schicht |
| 14 | Faltungsschicht |
| 15 | Batch-Normalisierungsschicht |
| 16 | ReLU-Schicht |
| 17 | Faltungsschicht |
| 18 | Batch-Normalisierungsschicht |
| 19 | ReLU-Schicht |
| 20 | Pooling-Schicht |
| 21 | Dropout-Schicht |
| 22 | vollständig vernetzte Schicht |
| 23 | Softmax-Schicht |
| 24 | gewichtete Klassifikationsschicht |

Die Tiefe des Netzwerkes definiert sich anhand der verwendeten Faltungsschichten. Durch das Hintereinanderschalten mehrerer Faltungsschichten mit zugehörigen Batch-Normalisierungsschichten, ReLU-Schichten und Pooling-Schichten steigt die Komplexität der erkennbaren Muster [54].

Das CNN wird über 25 Epochen trainiert. Die Lernrate beträgt zu Anfang $\eta = 10^{-4}$ und wird nach 20 Epochen um den Faktor 10 gesenkt. Die Größe eines Batches beträgt 128 [46].

Der Algorithmus wurde mit dem Datensatz *Speech Commands* [101] validiert. Für 10 Wörter mit je ca. 1850 Realisierungen im Trainingsanteil und ca. 260 Realisierungen im Validierungsanteil wird für den Validierungsanteil eine Vorhersagegenauigkeit von 94.5 % erreicht [46].

### 7.3.2 Erstellung des Trainingsmaterials

Die Grundlage des Trainings- und Validierungsdatensatzes bildet der Teilkorpus *Complete*. Mithilfe der Informationen zu den Wortarten wurden aus dem Teilkorpus *Complete* alle Adjektive, Substantive und Verben mit mindestens 25 Realisierungen ausgelesen. Eine Übersicht der Anzahlen an Wörtern aus dem Teilkorpus *Complete* ist in Tabelle 7.3 aufgelistet.

Insgesamt 757 Adjektive, Substantive oder Verben mit mindestens 25 Realisierungen sind im Teilkorpus *Complete* enthalten. Für das Training des CNN wurden ausschließlich Wörter in der Grundform genutzt. 373 Wörter erfüllen dieses Kriterium. Aus den Realisierungen der 373 Wörter wurden vier Datensätze generiert. Für jeden Datensatz gilt, dass die Realisierungen eines Wortes von derselben Person entweder im Trainingsanteil oder im Validierungsanteil enthalten sind. Hierdurch wird das Programm personenunabhängig validiert. Die Zusammensetzungen der Datensätze sind der Tabelle 7.3 zu entnehmen. Eine Übersicht der Wörter ist in Tabelle A.3 in Abschnitt A.3 aufgelistet.

Die Wörter wurden anhand der im Korpus enthaltenen Wortgrenzen ausgeschnitten. Die separaten Wörter wurden mit einem Tukey-Fenster mit einem Kosinusanteil von 50 % multipliziert.

Tabelle 7.3: Anzahl an Adjektiven, Substantiven und Verben, welche mit einer bestimmten Mindestanzahl im Teilkorpus *Complete* vorkommen und in einer bestimmten Wortform vorliegen. Berücksichtigt wurden nur Wörter, für welche von separaten Personen ein Validierungsanteil von 10 %–20 % erstellt werden konnte.

| | **Anzahl Wörter** | | | | |
|---|---|---|---|---|---|
| | alle Formen | Grundformen | | | |
| **Wortart** | min. 25 Realisierungen im *Complete* | min. 25 Realisierungen im *Complete* (Datensatz 1) | min. 29 Realisierungen im *Complete* (Datensatz 2) | min. 57 Realisierungen im *Complete* (Datensatz 3) | min. 115 Realisierungen im *Complete* (Datensatz 4) |
| Adjektiv | 184 | 59 | 45 | 15 | 0 |
| Substantiv | 406 | 279 | 209 | 66 | 15 |
| Verb | 167 | 35 | 28 | 11 | 5 |
| Gesamt: | 757 | 373 | 282 | 92 | 20 |

Der Datensatz 1 beinhaltet alle 373 Wörter mit allen vorhandenen Realisierungen. Von jedem Wort wurden 10 %–20 % der Realisierungen zur Validierung verwendet. Die Nutzung aller Realisierungen führt zu einer ungleichen Repräsentation der Wörter in dem Datensatz. Die mittlere Anzahl an Realisierungen pro Wort liegt bei 56.5, der Median bei 35.5. Die drei häufigsten Wörter im Trainingsdatensatz sind 'Hundert' (1530 Realisierungen), 'werden' (1155 Realisierungen) und 'Tausend' (593 Realisierungen). Die seltensten Wörter im Trainingsdatensatz weisen 21 Realisierungen auf. Zu dem Datensatz 1 wurden 100 verschiedene Realisierungen von Rauschen hinzugefügt.

Die Nutzung eines Teildatensatzes mit einer homogenen Verteilung verkürzt die Trainingszeit eines Modells und kann zu äquivalenten Ergebnissen führen [26], [105]. Aus diesem Grund wurden die drei weiteren Datensätze mit einer homogenen Verteilung generiert.

Der Datensatz 2 beinhaltet alle Wörter, welche sich in ein Trainingsanteil mit 25 Realisierungen und einen Validierungsanteil mit 4 Realisierungen einteilen lassen. Es verbleiben 282 Wörter. Zu dem Datensatz 2 wurden 100 verschiedene Realisierungen von Rauschen hinzugefügt.

Der Datensatz 3 beinhaltet alle Wörter, welche sich in ein Trainingsanteil mit 50 Realisierungen und einen Validierungsanteil mit 7 Realisierungen einteilen lassen. Es verbleiben 92 Wörter. Zu dem Datensatz 3 wurden 200 verschiedene Realisierungen von Rauschen hinzugefügt.

Der Datensatz 4 beinhaltet alle Wörter, welche sich in ein Trainingsanteil mit 100 Realisierungen und einen Validierungsanteil mit 15 Realisierungen einteilen lassen. Es verbleiben 20 Wörter. Zu dem Datensatz 4 wurden 200 verschiedene Realisierungen von Rauschen hinzugefügt.

### 7.3.3 Auswertung der Spracherkennung

Für die in Unterabschnitt 7.3.2 beschriebenen Datensätze wurde ein CNN zur Erkennung der jeweiligen Wörter trainiert und mit dem entsprechenden Validierungsanteil validiert. Die Ergebnisse zeigen, dass der CARInA für die Sprachverarbeitung verwendet werden kann. Nachfolgend werden die Ergebnisse der einzelnen Datensätze ausführlich diskutiert.

Die Hyperparameter wurden nicht auf die einzelnen Datensätze angepasst. Für die Verbesserung der Erkennungsraten ist ein Optimieren der Anzahl an Epochen, der Lernrate, der Batch-Größe und der Tiefe des CNN notwendig. Des Weiteren sind die Hyperparameter für die Erstellung der Spektrogramme zu optimieren. Auf diese Optimierung wurde im Rahmen der vorliegenden Ausarbeitung verzichtet.

Zur subjektiven Validierung der trainierten neuronalen Netze wurde das Programm `TestNetwork.m` geschrieben. Das Programm liest anfangs eines der trainierten CNN ein. Das Mikrophon dient als Eingabe. Die erkannte Klasse wird als Überschrift des berechneten Spektrogrammes ausgegeben. Der subjektive Eindruck dieses Programms ist eine starke Korrelation der Genauigkeit mit dem Geschlecht (männliche Personen werden besser erkannt), der Sprechgeschwindigkeit und der Länge des Wortes (lange Wörter werden besser erkannt). Umgebungsgeräusche und die Qualität des Mikrophons beeinflussen die Erkennung erheblich. Durch das Zero-Padding der Trainingsdaten ist für die Erkennung eine sehr rauscharme Umgebung notwendig. Viele gesprochene Wörter werden als Hintergrundrauschen klassifiziert (dies wird nicht angezeigt). Eine Möglichkeit zur Verbesserung des Programms ist das Ersetzen des Zero-Paddings durch Rauschen.

### Datensatz 1

Der Anteil richtig erkannter Realisierungen zu allen vorhandenen Realisierungen für den Validierungsanteil des Datensatzes 1 beträgt 81 %. Der Mittelwert der Genauigkeiten aller Klassen, unabhängig von den Anzahlen der Realisierungen, liegt bei 79 %. Somit wird die Genauigkeit nicht ausschließlich durch die Klassen mit vielen Realisierungen erreicht. Dennoch ist ein Zusammenhang zwischen der Vorhersagegenauigkeit und den Anzahlen an Realisierungen erkennbar.

In Abbildung 7.12 ist dieser Zusammenhang dargestellt. Die Anzahlen der Realisierungen im Validierungsanteil entsprechen 10 %-20 % der Anzahlen im Trainingsanteil. Es ist eine Tendenz zur stabileren Vorhersagegenauigkeit für mehr Realisierungen erkennbar. Dennoch wird auch für viele Wörter mit nur drei Realisierungen im Validierungsanteil eine Erkennungsrate von 100 % erreicht, sodass der Mittelwert nahe 100 % liegt. Das Wort mit 20 Realisierungen weist eine starke Abweichung dieser Tendenz auf. Das entsprechende Wort lautet 'München' und wurde sechsmal als das phonetisch ähnliche Wort 'Männchen' klassifiziert, welches seinerseits keinmal als 'München' klassifiziert wurde.

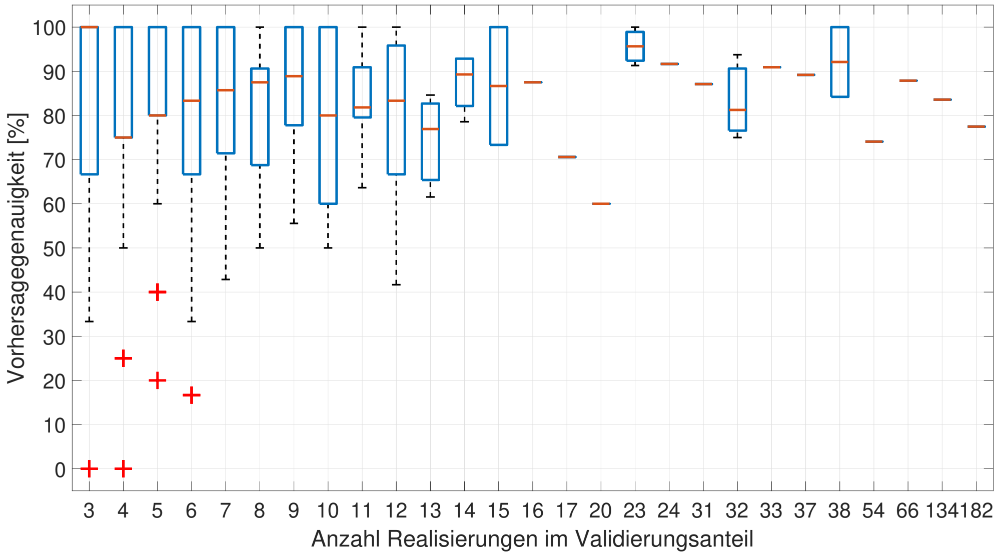


Abbildung 7.12: Vorhersagegenauigkeiten des Datensatz 1 in Abhängigkeit der Anzahl an Realisierungen im Validierungsanteil.

Die durchschnittliche Anzahl von Personen pro Wort beträgt für den Trainingsanteil 18.4 mit einer Standardabweichung von 14.4. Für den Validierungsanteil beträgt diese 3.9 mit einer Standardabweichung von 2.8.

In Tabelle 7.4 sind alle Wörter aus dem Datensatz 1 aufgelistet, welche eine Genauigkeit von maximal 25 % aufweisen. Auch hier ist die Korrelation weniger Daten zu niedrigeren Vorhersagegenauigkeiten zu erkennen. Die maximale Häufigkeit ist 6. Die erkannten Wörter weisen zu einem großen Anteil signifikante phonetische Unterschiede zu dem gesagten Wort auf, wie beispielsweise 'England' und 'Text'. Einige Verwechslungen sind sich phonetisch ähnlich, wie 'Staat' und 'stark'. Es ist kein Zusammenhang zwischen der Anzahl an Personen im Trainings- beziehungsweise Validierungsanteil zu der Genauigkeit ersichtlich.

Tabelle 7.4: Wörter des Datensatzes 1, welche in der Validierung eine Genauigkeit von maximal 25 % aufweisen. Zu den Wörtern sind die erreichte Genauigkeit, die Anzahl der Realisierungen im Validierungsdatensatz, die Anzahl der Personen und die häufigsten Verwechslungen angegeben.

| | | | **Anzahl Personen** | | **Häufigste Verwechslungen** | | |
|---|---|---|---|---|---|---|---|
| **Wort** | **Genauigkeit [%]** | **Anzahl** | Training | Validierung | 1 | 2 | 3 |
| Bürger | 0 | 4 | 9 | 3 | November: 1 | Tausend: 1 | Wirkung: 1 |
| England | 0 | 4 | 14 | 1 | Kinder: 3 | Text: 1 | |
| Lied | 0 | 3 | 5 | 2 | Weg: 3 | | |
| Planck | 0 | 4 | 2 | 2 | Brandt: 2 | Erde: 1 | Kampf: 1 |
| Serie | 0 | 3 | 6 | 1 | wenig: 2 | schwer: 1 | |
| Angriff | 16.7 | 6 | 4 | 3 | Anteil: 3 | Amt: 1 | Anfang: 1 |
| Hand | 16.7 | 6 | 20 | 1 | Amt: 1 | Hundert: 1 | Land: 1 |
| nehmen | 16.7 | 6 | 16 | 2 | Juni: 1 | dienen: 1 | geben: 1 |
| liegen | 20 | 5 | 14 | 2 | Leben: 3 | Liebe: 1 | |
| Brücke | 25 | 4 | 5 | 3 | Gruppe: 1 | Mutter: 1 | Programm: 1 |
| Heinrich | 25 | 4 | 6 | 3 | Einfluss: 1 | Frankreich: 1 | Recht: 1 |
| Napoleon | 25 | 4 | 2 | 4 | April: 1 | Natur: 1 | Probleme: 1 |
| Staat | 25 | 4 | 10 | 2 | stark: 3 | | |
| Text | 25 | 4 | 12 | 2 | Sitz: 1 | Stadt: 1 | fest: 1 |

### Datensatz 2

Der Datensatz 2 weist eine Erkennungsrate von 61 % auf. Diese deutlich geringere Genauigkeit gegenüber dem Datensatz 1 ist durch die vergleichsweise vielen Klassen mit wenigen Realisierungen zu erklären. Auch wurden die Hyperparameter wie die Lernrate und die Anzahl der Epochen für das Training nicht angepasst. In Abbildung 7.13 ist die Verteilung der Vorhersagegenauigkeit für den Datensatz 2 dargestellt.

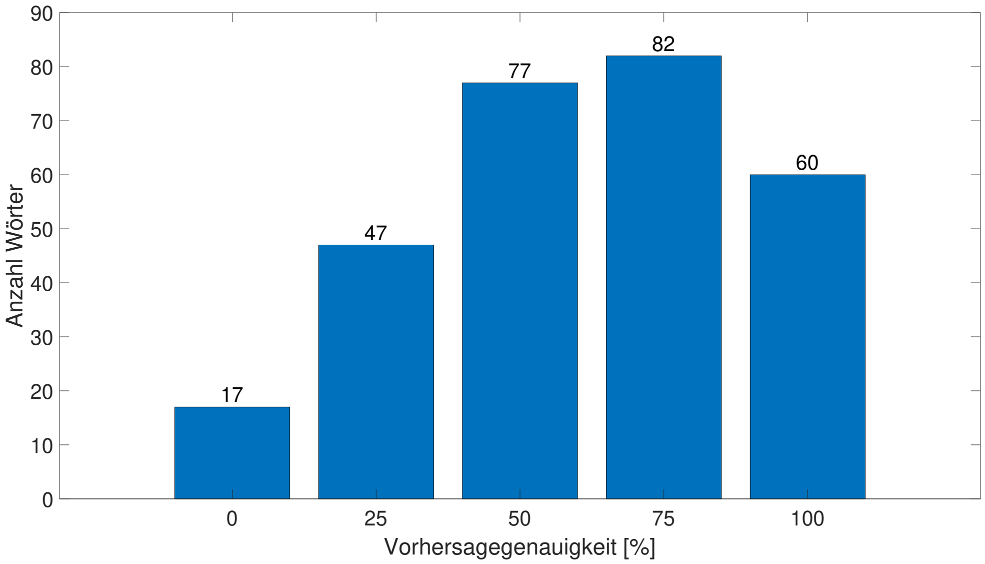


Abbildung 7.13: Verteilung der Vorhersagegenauigkeit über alle Wörter aus dem Datensatz 2

Die durchschnittliche Anzahl von Personen pro Wort beträgt für den Trainingsanteil 12.1 mit einer Standardabweichung von 4.1. Für den Validierungsanteil beträgt diese 2.9 mit einer Standardabweichung von 0.9.

In Tabelle 7.5 sind die 17 Wörter aus Abbildung 7.13 aufgelistet, für welche keine Realisierung erkannt wurde. Es ist kein offensichtlicher Grund wie phonetische Ähnlichkeit erkennbar, weshalb diese Wörter für keine Realisierung richtig erkannt wurden. Die Audiodateien sind nicht fehlerbehaftet. Auch ein Zusammenhang der Genauigkeit zu der Personenanzahl im Trainings- oder Validierungsanteil ist nicht ersichtlich. Vermutlich entsteht dieser Anteil an nicht erkannten Wörtern, da der Algorithmus nicht an den Datensatz angepasst wurde.

Tabelle 7.5: Wörter des Datensatzes 2, welche in der Validierung eine Genauigkeit von 0 % aufweisen. Zu den Wörtern sind die Anzahl der Personen und die häufigsten Verwechslungen angegeben.

| | | **Anzahl Personen** | | **Häufigste Verwechslungen** | | |
|---|---|---|---|---|---|---|
| **Wort** | **Genauigkeit** [%] | Training | Validierung | 1 | 2 | 3 |
| Angriff | 0 | 3 | 2 | Anteil: 2 | Art: 1 | halten: 1 |
| Bau | 0 | 10 | 3 | Art: 1 | Ort: 1 | Raum: 1 |
| Brücke | 0 | 5 | 3 | Gruppe: 2 | Partei: 1 | Punkt: 1 |
| Hand | 0 | 20 | 1 | Art: 1 | Boden: 1 | Mann: 1 |
| Hilfe | 0 | 17 | 2 | Brücke: 1 | Insel: 1 | Richtung: 1 |
| Lage | 0 | 13 | 2 | Deutschland: 1 | Elemente: 1 | Norden: 1 |
| Land | 0 | 15 | 3 | Jahr: 1 | Mensch: 1 | Raum: 1 |
| Liebe | 0 | 11 | 2 | Meter: 2 | Bewegung: 1 | Österreich: 1 |
| Weg | 0 | 12 | 3 | Leben: 1 | Welt: 1 | gehen: 1 |
| Windows | 0 | 3 | 2 | März: 2 | Buch: 1 | genau: 1 |
| bilden | 0 | 16 | 2 | Berlin: 1 | Krieg: 1 | Leben: 1 |
| gerade | 0 | 16 | 1 | Frage: 1 | Italien: 1 | erneut: 1 |
| hören | 0 | 17 | 3 | Jahr: 1 | Mann: 1 | Reihe: 1 |
| nehmen | 0 | 13 | 2 | Juni: 2 | Weg: 1 | liegen: 1 |
| sehen | 0 | 18 | 4 | Süden: 2 | Film: 1 | liegen: 1 |
| sollen | 0 | 14 | 4 | Seite: 1 | Zeit: 1 | sein: 1 |
| werden | 0 | 18 | 2 | bilden: 2 | Ende: 1 | nehmen: 1 |

### Datensatz 3

Der Datensatz 3 weist eine Erkennungsrate von 59 % auf. Die Erkennungsrate ist somit geringer als bei dem Datensatz 2, obwohl doppelt so viel Trainingsmaterial pro Wort vorhanden war und nur ein knappes Drittel der Wörter verwendet wurde. Dies ist darauf zurückzuführen, dass die Hyperparameter im Training nicht an das zu trainierende System angepasst wurden. In Abbildung 7.14 ist die Verteilung der Vorhersagegenauigkeit für den Datensatz 3 dargestellt.

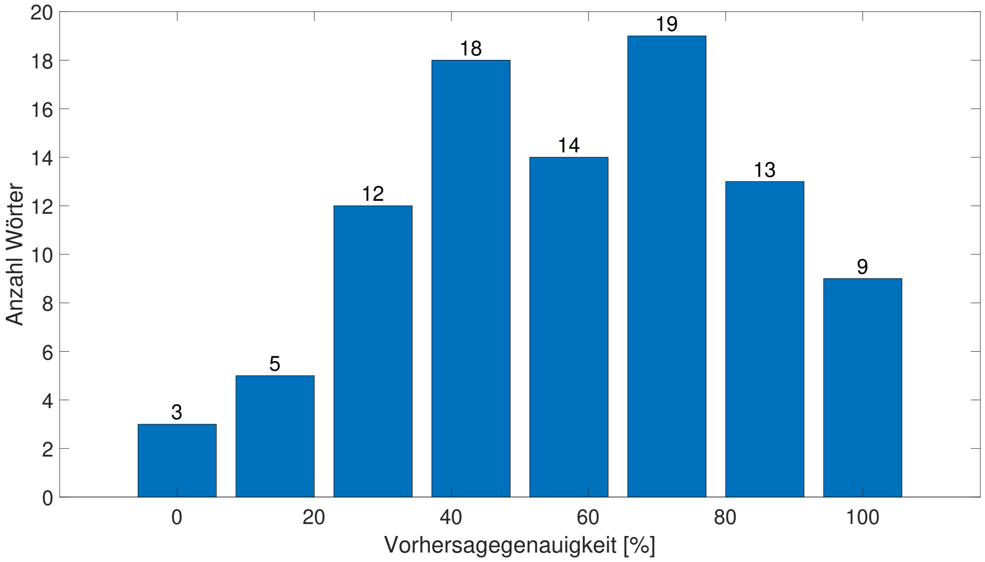


Abbildung 7.14: Verteilung der Vorhersagegenauigkeit über alle Wörter aus dem Datensatz 3

Die durchschnittliche Anzahl von Personen pro Wort beträgt für den Trainingsanteil 19.8 mit einer Standardabweichung von 6.7. Für den Validierungsanteil beträgt diese 4.4 mit einer Standardabweichung von 1.5.

In Tabelle 7.6 sind die Wörter mit einer Vorhersagegenauigkeit von weniger als 29 % angegeben. Bei vielen Wörtern ist keine offensichtliche phonetische Korrelation zu den Verwechslungen erkennbar. Dies unterstützt die Vermutung, dass der Algorithmus auf die Datensätze abzustimmen ist. Eine Korrelation der Personenanzahl des Trainings- beziehungsweise Validierungsanteils mit der Genauigkeit ist nicht erkennbar.

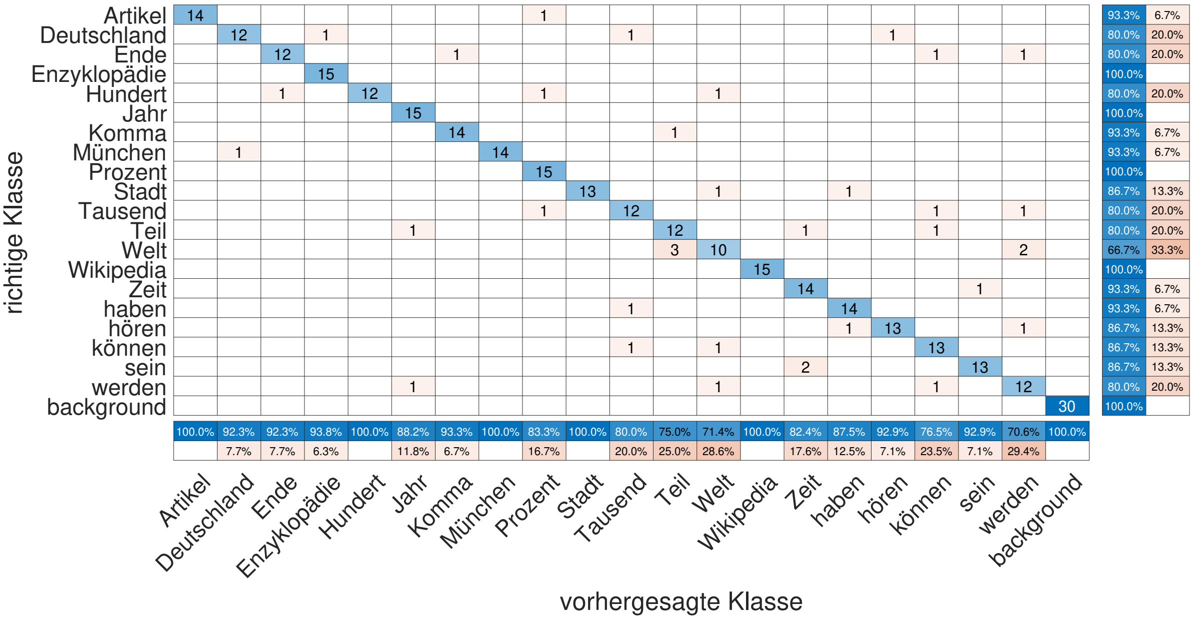
Tabelle 7.6: Wörter des Datensatzes 3, welche in der Validierung eine Genauigkeit von weniger als 29 % aufweisen. Zu den Wörtern sind die Anzahl der Personen und die häufigsten Verwechslungen angegeben.

| | | **Anzahl Personen** | | **Häufigste Verwechslungen** | | |
|---|---|---|---|---|---|---|
| **Wort** | **Genauigkeit** [%] | Training | Validierung | 1 | 2 | 3 |
| Grund | 0 | 20 | 4 | Bedeutung: 1 | Jahr: 1 | Komma: 1 |
| Welt | 0 | 25 | 6 | Jahr: 1 | Krieg: 1 | Regel: 1 |
| werden | 0 | 22 | 7 | finden: 2 | lassen: 2 | Bahnhof: 1 |
| Begriff | 14.3 | 28 | 3 | Beginn: 4 | Gebäude: 2 | |
| Form | 14.3 | 22 | 2 | Grund: 4 | Fall: 1 | Komma: 1 |
| Land | 14.3 | 22 | 7 | Jahr: 2 | Art: 1 | Teil: 1 |
| hören | 14.3 | 23 | 5 | Jahr: 1 | Tausend: 1 | Welt: 1 |
| müssen | 14.3 | 21 | 7 | lassen: 3 | München: 2 | hören: 1 |
| Bayern | 28.6 | 6 | 5 | Bahnhof: 1 | Fall: 1 | Form: 1 |
| Bevölkerung | 28.6 | 17 | 6 | Bedeutung: 1 | Enzyklopädie: 1 | Europa: 1 |
| Geschichte | 28.6 | 26 | 3 | Entwicklung: 4 | Leben: 1 | |
| Lage | 28.6 | 23 | 4 | Mai: 1 | Mehrheit: 1 | München: 1 |
| Regel | 28.6 | 17 | 5 | Ende: 3 | hören: 2 | |
| Reich | 28.6 | 13 | 5 | Reihe: 2 | Mai: 1 | können: 1 |
| Wasser | 28.6 | 17 | 4 | Prozent: 2 | April: 1 | Art: 1 |
| deutlich | 28.6 | 19 | 6 | Deutschland: 1 | gleichzeitig: 1 | hören: 1 |
| gut | 28.6 | 23 | 6 | rund: 3 | deutlich: 1 | hören: 1 |
| lang | 28.6 | 22 | 3 | Art: 1 | Lage: 1 | München: 1 |
| machen | 28.6 | 29 | 7 | Frankreich: 1 | Mehrheit: 1 | haben: 1 |
| neu | 28.6 | 17 | 6 | Mai: 2 | Hundert: 1 | Jahr: 1 |

### Datensatz 4

Der Datensatz 4 weist eine Erkennungsrate von 89 % auf. In Abbildung 7.15 ist die Konfusionsmatrix des Datensatzes 4 abgebildet. Für diesen vergleichsweise kleinen Datensatz mit 20 Wörtern wird jedes Wort mit einer Genauigkeit von mindestens 66 % erkannt. Verhältnismäßig lange Wörter wie 'Wikipedia' oder 'Enzyklopädie' werden sehr gut erkannt. Die Klassifikation des Hintergrundrauschens (mit *background* bezeichnet) funktioniert für jede Realisierung fehlerfrei.

Die durchschnittliche Anzahl von Personen pro Wort beträgt für den Trainingsanteil 35.2 mit einer Standardabweichung von 10.9. Für den Validierungsanteil beträgt diese 8.4 mit einer Standardabweichung von 2.7. Wie der Tabelle 7.7 zu entnehmen ist, ist keine Korrelation der Personenanzahl im Trainings- beziehungsweise Validierungsanteil zu der Genauigkeit erkennbar.



| | Artikel | Deutschland | Ende | Enzyklopädie | Hundert | Jahr | Komma | München | Prozent | Stadt | Tausend | Teil | Welt | Wikipedia | Zeit | haben | hören | können | sein | werden | background | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Artikel | 14 | | | | | | | | 1 | | | | | | | | | | | | | 93.3% | 6.7% |
| Deutschland | | 12 | | 1 | | | | | | | 1 | | | | | | 1 | | | | | 80.0% | 20.0% |
| Ende | | | 12 | | | | 1 | | | | | | | | | | | 1 | | 1 | | 80.0% | 20.0% |
| Enzyklopädie | | | | 15 | | | | | | | | | | | | | | | | | | 100.0% | |
| Hundert | | | 1 | | 12 | | | | 1 | | | | 1 | | | | | | | | | 80.0% | 20.0% |
| Jahr | | | | | | 15 | | | | | | | | | | | | | | | | 100.0% | |
| Komma | | | | | | | 14 | | | | | 1 | | | | | | | | | | 93.3% | 6.7% |
| München | | 1 | | | | | | 14 | | | | | | | | | | | | | | 93.3% | 6.7% |
| Prozent | | | | | | | | | 15 | | | | | | | | | | | | | 100.0% | |
| Stadt | | | | | | | | | | 13 | | | 1 | | | 1 | | | | | | 86.7% | 13.3% |
| Tausend | | | | | | | | | 1 | | 12 | | | | | | | 1 | | 1 | | 80.0% | 20.0% |
| Teil | | | | | | 1 | | | | | | 12 | | | 1 | | | 1 | | | | 80.0% | 20.0% |
| Welt | | | | | | | | | | | | 3 | 10 | | | | | | | 2 | | 66.7% | 33.3% |
| Wikipedia | | | | | | | | | | | | | | 15 | | | | | | | | 100.0% | |
| Zeit | | | | | | | | | | | | | | | 14 | | | | 1 | | | 93.3% | 6.7% |
| haben | | | | | | | | | | | 1 | | | | | 14 | | | | | | 93.3% | 6.7% |
| hören | | | | | | | | | | | | | | | | 1 | 13 | | | 1 | | 86.7% | 13.3% |
| können | | | | | | | | | | | 1 | | 1 | | | | | 13 | | | | 86.7% | 13.3% |
| sein | | | | | | | | | | | | | | | 2 | | | | 13 | | | 86.7% | 13.3% |
| werden | | | | | | 1 | | | | | | | 1 | | | | | 1 | | 12 | | 80.0% | 20.0% |
| background | | | | | | | | | | | | | | | | | | | | | 30 | 100.0% | |
| | 100.0% | 92.3% | 92.3% | 93.8% | 100.0% | 88.2% | 93.3% | 100.0% | 83.3% | 100.0% | 80.0% | 75.0% | 71.4% | 100.0% | 82.4% | 87.5% | 92.9% | 76.5% | 92.9% | 70.6% | 100.0% | | |
| | | 7.7% | 7.7% | 6.3% | | 11.8% | 6.7% | | 16.7% | | 20.0% | 25.0% | 28.6% | | 17.6% | 12.5% | 7.1% | 23.5% | 7.1% | 29.4% | | | |

Abbildung 7.15: Konfusionsmatrix des Datensatz 4

Tabelle 7.7: Wörter des Datensatzes 4 mit den jeweiligen Anzahlen an Personen im Trainings- und Validierungsanteil

| | | **Anzahl Personen** | |
|---|---|---|---|
| **Wort** | **Genauigkeit** [%] | Training | Validierung |
| Welt | 66.7 | 33 | 9 |
| Deutschland | 80 | 26 | 10 |
| Ende | 80 | 30 | 9 |
| Hundert | 80 | 30 | 8 |
| Tausend | 80 | 35 | 9 |
| Teil | 80 | 32 | 11 |
| werden | 80 | 44 | 9 |
| Stadt | 86.7 | 22 | 8 |
| hören | 86.7 | 49 | 8 |
| können | 86.7 | 43 | 7 |
| sein | 86.7 | 37 | 10 |
| Artikel | 93.3 | 43 | 9 |
| Komma | 93.3 | 27 | 7 |
| München | 93.3 | 3 | 10 |
| Zeit | 93.3 | 44 | 13 |
| haben | 93.3 | 49 | 1 |
| Enzyklopädie | 100 | 46 | 9 |
| Jahr | 100 | 35 | 11 |
| Prozent | 100 | 34 | 5 |
| Wikipedia | 100 | 42 | 4 |

Von den generierten vier Datensätzen weist der Datensatz 4 die größte Übereinstimmung zu dem original verwendeten Datensatz *Speech Commands* auf. Aus diesem Grund sind die eingestellten Hyperparameter für den Datensatz 4 günstiger als für die anderen Datensätze. Der Datensatz 4 weist mit 89 % die höchste Erkennungsrate auf.

## *Kapitel 8*

# 8 Zusammenfassung und Ausblick

Die Zielstellung der vorliegenden Arbeit ist die Erstellung und Dokumentation einer Datenbasis gesprochener deutscher Sprache inklusive Annotationen auf unterschiedlichen Ebenen.

Sprachkorpora bilden die Grundlage der Sprachverarbeitung und werden für nahezu jedes System der Spracherkennung, Sprachsynthese und Sprachanalyse verwendet. Korpora werden durch unterschiedliche Merkmale wie die Sprachauswahl, die Größe oder den Umfang der Annotationen klassifiziert. Für viele Anwendungen ist ein umfangreicher Korpus notwendig.

In anderen Sprachen wie Englisch oder Russisch existieren umfangreich annotierte Korpora mit teilweise mehreren Hundert Stunden Sprachmaterial. Für die deutsche Sprache ist die Größe der Korpora mit detaillierten Annotationen deutlich geringer. Es wurde eine tabellarische Auflistung ausgewählter deutschsprachiger Korpora und der zugehörigen Annotationsarten erstellt.

Der entstandene Datensatz beinhaltet drei Korpora. Der *BITS Unit Selection synthese corpus* ist im Besitz des BAS und weist einen Umfang von 13:33 Stunden mit manuell erstellten orthographischen, kanonischen, phonetischen und prosodischen Annotationen auf. Der *Kiel Corpus of Spoken German read speech* beinhaltet 4:15 Stunden, welche manuell auf orthographischer, kanonischer, phonetischer und prosodischer Ebene annotiert sind.

Der Hauptteil der vorliegenden Arbeit beinhaltet die Beschreibung der Erstellung des *Corpus of Aligned Read speech Including Annotations* (CARInA).

Das Sprachmaterial des CARInA stammt aus dem *German Spoken Wikipedia Corpus* (GSWC). Der GSWC beinhaltet automatisch erstellte Alignments von gesprochenen Artikeln der Enzyklopädie Wikipedia zu den zugehörigen Texten. Die Alignments sind auf orthographischer und phonetischer Ebene vorhanden. Durch die Nutzung dieser Sprachressourcen wird ein thematisch vielfältiger Datensatz erzeugt.

Für die Weiterverarbeitung der Daten wurde das Sprachmaterial in Sätze unterteilt. Der Anteil phonetisch vollständig alignierter Sätze beinhaltet 129 Stunden Sprachmaterial. Durch das Entfernen der Sätze mit ungenauen Phonemgrenzen verbleiben 124 Stunden Sprachmaterial von 323 Personen.

Dem CARInA wurden automatisch kanonische und morphosyntaktische Informationen hinzugefügt. Hierfür wurden drei Wörterbücher erstellt, in welchen für ein Wort die kanonische Aussprache, die Silbentrennung beziehungsweise die Wortart enthalten ist. Die

Quelle der Informationen für die Wörterbücher ist das Wiktionary, welches im Internet kostenlos zur Verfügung steht. Insgesamt sind 102 766 Wörter im CARInA enthalten, welche nicht im Wiktionary aufgeführt sind. Diese beinhalten zu einem Großteil Zahlworte, Eigennamen und Komposita. Eine Auswahl von 1846 Wörtern wurde den Wörterbüchern hinzugefügt. Der Anteil des CARInA mit vollständiger orthographischer, phonetischer, kanonischer und morphosyntaktischer Information beinhaltet 29:47 Stunden Sprachmaterial.

Durch das Hinzufügen weiterer Wörter zu den Wörterbüchern kann der Umfang des CARInA mit vollständiger Information vervierfacht werden. Speziell Komposita können automatisch durch die einzelnen (im Wiktionary enthaltenden) Bestandteile beigefügt werden. Bei der manuellen Bildung von Komposita ist die Zuweisung der Wortakzente zu den Silben zu beachten. Eine weitere Möglichkeit ist das Implementieren von regelbasierten Programmen, welche für nicht im Wörterbuch enthaltene Wörter die jeweiligen Informationen generiert. Speziell für die Informationen der Wortart ist ein regelbasiertes System zuverlässiger, da für die eindeutige Erkennung der Wortart der jeweilige Kontext notwendig ist.

Der Anteil des CARInA mit vollständiger Information bildet den Teilkorpus *Complete*. Diesem Korpus wurden automatisch prosodische Informationen hinzugefügt. Die prosodischen Informationen wurden mithilfe der Programme *Phython Tones and Break Indices* (PyToBI) und *Prosody Recognition Revisited* (PRR) erstellt. Die Validierung prosodischer Etiketten ist aufgrund der hohen Anzahl unbetonter Silben und der niedrigen Übereinstimmung manuell erstellter Etikettierungen schwierig. Eine umfangreiche theoretische Validierung der Etiketten wurde durchgeführt.

Für eine Qualitätsanalyse der prosodischen Etiketten ist eine Studie notwendig. Mithilfe der Informationen des Teilkorpus *Complete* kann synthetische Sprache erstellt werden. Der Intonationsverlauf kann mithilfe der prosodischen Etiketten generiert werden. Durch die subjektive Beurteilung der synthetisierten Sprache kann die Qualität der prosodischen Etiketten abschließend beurteilt werden.

Der CARInA gliedert sich in den beschriebenen Teilkorpus *Complete* und den Teilkorpus *WorkInProgress*, welcher die Sätze mit unvollständiger Information enthält. Für jeden Satz ist eine Audiodatei mit einer Abtastrate von 44 100 Hz vorhanden. Zu jeder Audiodatei existiert eine dem BAS-Standard entsprechende Partitur-Datei und eine TextGrid-Datei. Die beiden Dateien beinhalten dieselben Informationen in unterschiedlichen Formaten. Die prosodischen Informationen des Teilkorpus *Complete* sind als Snippet-Dateien vorhanden und können in die Partitur- beziehungsweise die TextGrid-Dateien integriert werden. Eine Übersicht über die Vollständigkeit der Informationen jedes Satzes ist dem Korpus beigefügt.

Für die Auswertung des Teilkorpus *Complete* wurden die Verteilung des Sprachmaterials, die Sprechgeschwindigkeiten, die Grundfrequenzen, die Qualität der Sprachaufnahmen und die Verteilung der Formantfrequenzen für Vokale ermittelt. Die Ergebnisse dieser

Auswertung entsprechen weitestgehend den jeweiligen Referenzen. Der Teilkorpus *Complete* kann somit für Anwendungen der Sprachverarbeitung genutzt werden.

Aus dem Teilkorpus *Complete* wurden vier Datensätze erstellt. Mit diesen nach Wörtern strukturierten Datensätzen wurde ein tiefes neuronales Netz zur Spracherkennung trainiert. Die Erkennungsrate eines homogen verteilten Datensatzes mit 282 verschiedenen Wörtern liegt bei 61 %, mit 20 verschiedenen Wörtern bei 89 %. Diese Erkennungsraten wurden ohne Anpassung der Hyperparameter erzielt und dienen der Evaluation des CARInA. Zur Verbesserung des Trainings sind die Anzahl der Epochen, die Lernrate, die Batch-Größe und die Tiefe des verwendeten neuronalen Netzes anzupassen.

Die Ergebnisse der Spracherkennung zeigen, dass sich der CARInA als Trainingsmaterial für die Sprachverarbeitung eignet. Die Ergebnisse der Spracherkennung sind in Form eines trainierten Programms vorhanden.

Im Folgenden sind die Schwerpunkte der Aufgabenstellung und der Erfolg der Bearbeitung zusammenfassend aufgeführt.

**Recherche & Analyse von bestehenden freien und kommerziellen Datenbasen mit deutschem Sprachmaterial**

Eine größere Anzahl deutscher Sprachkorpora werden von dem Bayrischen Archiv für Sprachsignale und der Datenbank für Gesprochenes Deutsch gehalten. Weitere Einrichtungen besitzen Korpora unterschiedlicher Größe und Qualität. Tabelle 2.3 beinhaltet eine Übersicht deutscher Sprachkorpora, welche zur Sprachverarbeitung geeignet sind. Zu den Korpora sind die Inhalte, die Anzahl der Sprecherinnen und Sprecher, der Umfang, die vorhandenen Ebenen der Annotation, der Preis und eventuelle korpusspezifische Anmerkungen vorhanden.

**Identifikation und Beschaffung geeigneter Kandidaten**

Die Tabelle 1.1 enthält die Anforderungen an einen Korpus zur Verarbeitung deutscher Sprache. Kein Korpus erfüllt diese Anforderungen vollständig. In Unterabschnitt 2.2.3 wurden die drei Korpora vorgestellt, welche die Anforderungen zu einem Großteil erfüllen. Der BITS-US ist bereits an der TUD vorhanden gewesen. Der KCSGrs wurde im Rahmen der vorliegenden Arbeit beschafft. Beide Korpora sind manuell auf orthographischer, kanonischer, phonetischer und prosodischer Ebene annotiert. Der GSWC beinhaltet automatisch erstellte orthographische und phonetische Alignments. Dieser Korpus wurde aufgrund seines Umfangs von mehreren Hundert Stunden Sprachmaterial ausgewählt.

**Ergänzung fehlender Annotationsformen durch geeignete (halb-)automatische Verfahren**

Das Sprachmaterial und die Alignments des GSWC wurden für die Erstellung des CARInA genutzt. Aufgrund des Umfangs wurden fehlende Annotationsformen automatisch hinzugefügt. Für die Annotation kanonischer und morphosyntaktischer

Informationen wurden Wörterbücher erstellt. Diese Wörterbücher beinhalten Informationen zu der Silbifizierung, der kanonischen Aussprache und der Wortart eines bestimmten Wortes. Die Wörterbücher wurden mithilfe des digital verfügbaren Wörterbuchs Wiktionary erstellt und manuell ergänzt. Prosodische Informationen wurden mit den Programmen PyToBI und PRR hinzugefügt. Das PRR wurde mit dem SRNC, dem BITS-US und dem KCSGrs trainiert. Es entstanden prosodische Etiketten aus vier Quellen, welche dem CARInA hinzugefügt wurden.

**Zusammenstellung einer finalen Datenbasis, ggf. durch Verbindung von Daten mehrerer Kandidaten**

Die finale Datenbasis besteht aus dem BITS-US, dem KCSGrs und dem CARInA. Die Korpora liegen getrennt vor. Für die Nutzung der orthographischen, kanonischen und phonetischen Informationen können die Korpora gemeinsam genutzt werden. Der KCSGrs beinhaltet prosodische Etiketten des KIM, der BITS-US und der CARInA beinhalten ToBI-Etiketten. Eine Konvertierung der Etiketten von KIM zu ToBI ist unter großem Informationsverlust möglich und wurde in Unterabschnitt 5.1.3 beschrieben. Mithilfe dieser Konvertierung können die Korpora gemeinsam verwendet werden.

**Validierung der Datenbasis durch Training und Evaluation eines tiefen neuronalen Netzes zur Spracherkennung**

Für den BITS-US und den KCSGrs liegen Validierungen vor. Die Auswertung des CARInA ist in dem Kapitel 7 ausführlich beschrieben. Das Sprachmaterial des vollständig annotierten Anteils wurde im Hinblick auf die Verteilung, die Sprechgeschwindigkeiten, die Sprachgrundfrequenzen, die Qualität, die Phonemeigenschaften und die Formantwerte untersucht. Es wurden vier Datensätze mit unterschiedlich umfangreichem Trainingsmaterial erstellt. Mit den Datensätzen wurden CNN zur Spracherkennung trainiert. Eine Anpassung der gewählten Architektur erfolgte nicht. Die Ergebnisse zeigen, dass der CARInA zur Sprachverarbeitung verwendet werden kann.

**Dokumentation der Datenbasis und der Annotationsformate**

Die vorliegende Ausarbeitung beinhaltet eine umfangreiche Dokumentation des CARInA. Der BITS-US, der KCSGrs und der CARInA enthalten jeweils eine `README` Datei mit Informationen zur Nutzung und zum Aufbau des jeweiligen Korpus.

# *Anhang A*

## A Ergänzende Ausarbeitungen

### A.1 Validierung Prosodie

Tabelle A.1: Zusammenstellung der Validierungsanteile aus dem *Kiel corpus of spoken german read speech* für die Validierung der automatischen prosodischen Annotationssysteme

| Teil | Inhalte |
|---|---|
| 1 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059 |
| 2 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021 |
| 3 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059 |
| 4 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064 |
| 5 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064, be068, be067, be070, be065, be069, be062, be066 |

Tabelle A.1: Zusammenstellung der Validierungsanteile aus dem *Kiel corpus of spoken german read speech* für die Validierung der automatischen prosodischen Annotationssysteme (Fortsetzung)

| Teil | Inhalte |
|---|---|
| 6 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064, be068, be067, be070, be065, be069, be062, be066, be075, be072 |
| 7 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064, be068, be067, be070, be065, be069 |
| 8 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064, be068, be067, be070, be065, be069, be062, be066, be075, be072, be073, be078 |
| 9 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048 |
| 10 | be001, be002, be003, be004, be005, be006, be007, be008, be009, be010, be017, be011, be019, be015, be013, be018, be014, be020, be016, be012, be028, be021, be024, be029, be026, be023, be030, be027, be022, be025, be040, be032, be033, be037, be035, be036, be039, be038, be031, be034, be047, be042, be044, be046, be048, be049, be045, be050, be041, be043, be056, be052, be053, be057, be058, be059, be055, be060, be051, be054, be063, be061, be064, be068, be067, be070, be065, be069, be062, be066, be075, be072, be073, be078, be076, be080, be077 |

## A.2 Auswertung Formanten

Tabelle A.2: Mittelwerte und Standardabweichungen der ersten beiden Formanten der Vokale des Teilkorpus *Complete*

| | **Mittelwert F1/F2 [Hz]** | | | **Standardabweichung F1/F2 [Hz]** | | |
|---|---|---|---|---|---|---|
| **Vokal** | feminin | maskulin | unbekannt | feminin | maskulin | unbekannt |
| øː | 400/1523 | 470/1511 | 343/1878 | 62/211 | 131/203 | 195/187 |
| ɐ | 515/1547 | 505/1462 | 483/1614 | 158/348 | 156/290 | 324/343 |
| œ | 505/1513 | 310/2169 | 416/1460 | 161/261 | 171/190 | 250/479 |
| ə | 415/1667 | 515/1136 | 351/1727 | 128/369 | 121/347 | 277/298 |
| ɛː | 506/2080 | 434/1078 | 315/1392 | 108/337 | 110/551 | 159/449 |
| ɛ | 526/1854 | 382/1915 | 462/1454 | 129/329 | 201/289 | 169/558 |
| ɪ | 398/2102 | 398/1949 | 321/1618 | 135/358 | 115/240 | 218/289 |
| ɔ | 515/1138 | 360/1775 | 572/1537 | 105/208 | 230/278 | 233/512 |
| ʊ | 410/1117 | 384/1239 | 309/1930 | 101/293 | 149/519 | 253/262 |
| ʏ | 384/1505 | 399/1564 | 305/1574 | 67/265 | 196/293 | 98/201 |
| aː | 731/1390 | 664/1274 | 729/1596 | 163/177 | 128/164 | 311/457 |
| a | 680/1475 | 626/1355 | 642/1567 | 165/208 | 136/178 | 326/397 |
| eː | 426/2159 | 385/1559 | 320/1903 | 80/359 | 135/214 | 145/249 |
| iː | 343/2413 | 482/1697 | 381/1572 | 72/286 | 138/254 | 182/247 |
| oː | 422/934 | 441/1198 | 292/1746 | 83/296 | 160/462 | 183/240 |
| uː | 366/1083 | 422/1584 | 260/2109 | 85/333 | 191/334 | 178/214 |
| yː | 364/1747 | 429/1953 | 408/1742 | 162/309 | 155/217 | 257/274 |

## A.3 Spracherkennung Datensätze

Tabelle A.3: Wörter der Datensätzen für die Spracherkennung

| Datensatz | Inhalte |
|---|---|
| Datensatz 1 | Amt, Anfang, Angriff, Anteil, Antrag, Anzahl, April, Arbeit, Armee, Art, Artikel, Aufgabe, Augen, Augsburg, August, Ausland, Autoren, allein, angesehen, ausschließlich, ähnlich<br>Bahn, Bahnhof, Bau, Bayern, Bedeutung, Beginn, Begriff, Beispiel, Bereich, Berlin, Besitz, Besucher, Betrieb, Bevölkerung, Bewegung, Bewusstsein, Blut, Boden, Brandt, Brücke, Buch, Bundeskanzler, Bundesregierung, Bundestag, Bürger, befinden, bekannt, bilden<br>China, Christus<br>Darstellung, Daten, Deutschland, Dezember, Donau, Dresden, Druck, deutlich, dienen, direkt |

Tabelle A.3: Wörter der Datensätzen für die Spracherkennung (Fortsetzung)

| Datensatz | Inhalte |
|---|---|
| | Einfluss, Einsatz, Einwohner, Elemente, Ende, England, Entscheidung, Entwicklung, Enzyklopädie, Erde, Ergebnis, Europa, einfach, einigen, eng, erfolgreich, erhalten, erneut, erreichen, existieren |
| | Fall, Familie, Farbe, Februar, Film, Folge, Form, Frage, Frankreich, Frau, Friedrich, Funktion, fallen, fest, finden, frei, führen |
| | Gebiet, Gebäude, Gefahr, Gegensatz, Gegner, Geist, Geld, Geschichte, Gesellschaft, Gesetz, Gott, Grenze, Großbritannien, Grund, Gruppe, Größe, ganz, gar, geben, gehen, gelegentlich, gelten, gemeinsam, genau, gerade, gleich, gleichzeitig, groß, gut |
| | Hamburg, Hand, Hauptstadt, Heinrich, Herbst, Herkunft, Herrschaft, Hilfe, Hintergrund, Hitler, Holz, Hundert, Hälfte, haben, halten, hauptsächlich, hoch, häufig, hören |
| | Idee, Insel, Italien |
| | Jahr, Januar, Japan, Juli, Juni, jährlich |
| | Kaiser, Kampf, Kanal, Karl, Kilometer, Kinder, Kirche, Komma, Kopf, Krieg, Kritik, Kultur, Kunst, König, Körper, klar, knapp, kommen, kurz, können |
| | Lage, Land, Leben, Liebe, Lied, Linie, Luft, Luther, lang, lassen, laut, leicht, liegen |
| | Mai, Mann, Markt, Mathematik, Maßnahmen, Mehrheit, Meinung, Mensch, Meter, Mitte, Mittelalter, Museum, Musik, Mutter, Männchen, März, München, Münster, machen, mehrfach, möglich, müssen |
| | Nachfolger, Nachricht, Nacht, Napoleon, Natur, Niederlage, Norden, November, Nutzung, Nähe, nehmen, neu |
| | Oktober, Opfer, Ort, Osten, Otto, Öffentlichkeit, Österreich |
| | Papier, Partei, Person, Peter, Philosophie, Planck, Platz, Politik, Position, Preußen, Probleme, Produktion, Programm, Projekt, Prozent, Punkt, politisch |
| | Raum, Recht, Regel, Regensburg, Regierung, Region, Reich, Reihe, Republik, Richtung, Rolle, Russland, regelmäßig, relativ, rund |
| | Sachsen, Schiffe, Schloss, Schlüssel, Schutz, Schweiz, Seite, September, Serie, Service, Sicht, Sinn, Situation, Sitz, Software, Sohn, Sommer, Sowjetunion, Spieler, Sprache, Staat, Stadion, Stadt, Stein, Stellung, Straße, Strecke, Stück, System, Süden, schnell, schwer, sehen, sein, selten, sollen, stark, stehen, stellen |
| | Tag, Tausend, Technik, Teil, Text, Thema, Theorie, Titel, Tod, tatsächlich, traditionell, tragen |
| | Umgebung, Universität, Unternehmen, Unterstützung, ungefähr, ursprünglich |

Tabelle A.3: Wörter der Datensätzen für die Spracherkennung (Fortsetzung)

| Datensatz | Inhalte |
|---|---|
| | Vater, Verbindung, Verbreitung, Verfahren, Verfassung, Verfügung, Verhalten, Verhältnis, Verlauf, Version, Vertrauen, Verwendung, Viren, Volk, verboten, verbunden, verhindern, vertreten, viel, vollständig, vorhanden<br>Wagen, Wahl, Wasser, Weg, Weibchen, Wels, Welt, Werk, Westen, Widerstand, Wikipedia, Windows, Wirkung, Wirtschaft, Wort, wahrscheinlich, weit, wenig, werden, wesentlich<br>Zahl, Zeit, Zeitpunkt, Zentimeter, Zentrum, Ziel, Zug, Zusammenhang, zeigen, ziehen, zusätzlich |
| Datensatz 2 | Anfang, Angriff, Anteil, Anzahl, April, Arbeit, Armee, Art, Artikel, Augen, August, Autoren, allein, angesehen, ausschließlich, ähnlich<br>Bahn, Bahnhof, Bau, Bayern, Bedeutung, Beginn, Begriff, Beispiel, Bereich, Berlin, Besucher, Betrieb, Bevölkerung, Bewegung, Bewusstsein, Boden, Brücke, Buch, Bundeskanzler, Bundestag, Bürger, befinden, bekannt, bilden<br>Christus<br>Deutschland, Dezember, deutlich, direkt<br>Einfluss, Einsatz, Elemente, Ende, Entscheidung, Entwicklung, Enzyklopädie, Erde, Europa, einfach, einigen, erfolgreich, erhalten, erneut, erreichen, existieren<br>Fall, Familie, Februar, Film, Folge, Form, Frage, Frankreich, Frau, Friedrich, fest, finden, frei, führen<br>Gebiet, Gebäude, Gefahr, Gegensatz, Gegner, Geist, Geld, Geschichte, Gesellschaft, Gott, Großbritannien, Grund, Gruppe, Größe, ganz, gar, geben, gehen, gelegentlich, genau, gerade, gleich, gleichzeitig, gut<br>Hamburg, Hand, Heinrich, Herbst, Herrschaft, Hilfe, Hitler, Hundert, Hälfte, haben, halten, hauptsächlich, hoch, häufig, hören<br>Insel, Italien<br>Jahr, Januar, Japan, Juli, Juni, jährlich<br>Kaiser, Kilometer, Kinder, Kirche, Komma, Krieg, Kritik, Kultur, Kunst, König, Körper, kommen, kurz, können<br>Lage, Land, Leben, Liebe, Linie, Luther, lang, lassen, leicht, liegen<br>Mai, Mann, Mehrheit, Mensch, Meter, Mitte, Museum, Musik, Mutter, Männchen, März, München, machen, möglich, müssen<br>Napoleon, Natur, Norden, November, Nähe, nehmen, neu<br>Oktober, Opfer, Ort, Otto, Öffentlichkeit, Österreich<br>Partei, Peter, Philosophie, Planck, Platz, Politik, Position, Preußen, Probleme, Prozent, Punkt<br>Raum, Recht, Regel, Regensburg, Regierung, Region, Reich, Reihe, Republik, Richtung, Rolle, Russland, relativ, rund |

Tabelle A.3: Wörter der Datensätzen für die Spracherkennung (Fortsetzung)

| Datensatz | Inhalte |
|---|---|
| | Sachsen, Schiffe, Schloss, Schlüssel, Schweiz, Seite, September, Sicht, Situation, Sitz, Software, Sohn, Sommer, Sowjetunion, Spieler, Stadt, Straße, Strecke, System, Süden, schnell, schwer, sehen, sein, selten, sollen, stark, stehen, stellen<br>Tag, Tausend, Teil, Text, Thema, Theorie, Titel, Tod, tatsächlich<br>Universität, Unternehmen, Unterstützung, ursprünglich<br>Vater, Verbreitung, Verfahren, Verfassung, Verfügung, Verhältnis, Version, Verwendung, Volk, verbunden, vertreten, viel, vollständig, vorhanden<br>Wahl, Wasser, Weg, Weibchen, Welt, Werk, Westen, Widerstand, Wikipedia, Windows, Wirkung, Wirtschaft, Wort, weit, wenig, werden, wesentlich<br>Zahl, Zeit, Zeitpunkt, Zentimeter, Zentrum, Ziel, Zug, Zusammenhang, zeigen, zusätzlich |
| Datensatz 3 | April, Art, Artikel, August<br>Bahnhof, Bayern, Bedeutung, Beginn, Begriff, Beispiel, Berlin, Bevölkerung, Bundeskanzler, Bundestag, bekannt<br>Deutschland, deutlich, direkt<br>Ende, Entwicklung, Enzyklopädie, Europa, erhalten<br>Fall, Film, Form, Frage, Frankreich, finden<br>Gebäude, Geschichte, Grund, ganz, gleichzeitig, gut<br>Hundert, haben, häufig, hören<br>Jahr<br>Kinder, Kirche, Komma, Krieg, König, kommen, können<br>Lage, Land, Leben, lang, lassen<br>Mai, Mehrheit, Meter, März, München, machen, möglich, müssen<br>neu<br>Oktober, Österreich<br>Prozent<br>Regel, Regierung, Reich, Reihe, Rolle, relativ, rund<br>Schloss, Seite, September, Spieler, Stadt, sein, stark<br>Tag, Tausend, Teil, Tod<br>viel<br>Wasser, Weg, Welt, Wikipedia, weit, werden<br>Zahl, Zeit, Zug |
| Datensatz 4 | Artikel<br>Deutschland<br>Ende, Enzyklopädie<br>Hundert, haben, hören<br>Jahr<br>Komma, können |

Tabelle A.3: Wörter der Datensätzen für die Spracherkennung (Fortsetzung)

| Datensatz | Inhalte |
|---|---|
| | München<br>Prozent<br>Stadt, sein<br>Tausend, Teil<br>Welt, Wikipedia, werden<br>Zeit |

# Literatur

[1] A. Agarwal und A. Jain, „Tones and Break Indices for speech processing – A review,“ in *Proceedings of the 4th national conference: Computation for nation development*, New Delhi, 2010, S. 319–324.

[2] D. Alston. (2002). Wiktionary, Adresse: `https://de.wiktionary.org/wiki/Wiktionary:Hauptseite` (besucht am 07. 04. 2021).

[3] R. Baayen, R. Piepenbrock und H. van Rijn. (1993). The CELEX lexical database (CD-ROM), Adresse: `https://pure.mpg.de/pubman/faces/ViewItemFullPage.jsp?itemId=item_2339741_4` (besucht am 21. 04. 2021).

[4] S. Baumann, M. Grice und R. Benzmüller, „GToBI – a phonological system for the transcription of German intonation,“ in *Prosody 2000: Speech recognition and synthesis*, Poznań, 2001, S. 21–28.

[5] T. Baumann und A. Köhn, „The Spoken Wikipedia Corpus collection: harvesting, alignment and an application to hyperlistening,“ *Language resources and evaluation*, Jg. 53, S. 303–329, 2016.

[6] T. Baytukalov. (2013). EasyPronounciation, Adresse: `https://easypronunciation.com/de/` (besucht am 13. 04. 2021).

[7] M. Becker, „Korpuslinguistik,“ in *Einführung in die spanische Sprachwissenschaft.* J. Metzler, 2013, S. 192–205.

[8] M. Beckmann, J. Hirschberg und S. Schattuck-Hufnagel, „The original ToBI system and the evolution of the ToBI framework,“ in *Prosodiy typology: the phonology of intonation and phrasing.* Oxford university press, 2005, S. 9–54.

[9] P. Birkholz, „Textnormalisierung und linguistische Analyse,“ Vorlesungsskript, 2019.

[10] M. Bisani und H. Ney, „Joint-sequence models for grapheme-to-phoneme conversion,“ *Speech communication*, Jg. 5, S. 434–451, 2008.

[11] P. Boersma und D. Weenink. (2020). Praat, Adresse: `http://www.praat.org/` (besucht am 21. 04. 2021).

[12] T. Bořil und R. Skarnitzl, „Tools rPraat and mPraat,“ in *Text, speech and dialogue*, Springer international publishing, 2016, S. 367–374.

[13] T. Brants, „TnT – a statistical Part-Of-Speech tagger,“ in *6th applied natural language processing conference*, association for computational lingustics, 2000, S. 224–231.

[14] S. Brognaux, S. Roekhaut, T. Drugman und R. Beaufort, „Automatic phone alignment,“ in *Advances in natural language processing*, H. Isahara und K. Kanzaki, Hrsg., Springer Berlin Heidelberg, 2012, S. 300–311.

[15] K. Bührig. (2021). The spoken wikipedia corpora, Adresse: `https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:swc-2.0#additional-files` (besucht am 26. 01. 2021).

[16] O. Carstens und D. Zentgraf. (2011). Duden, Adresse: `https://www.duden.de/` (besucht am 07. 04. 2021).

[17] W. Cheng, C. Greaves und M. Warren, „The creation of prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic),“ *International computer archive of modern English*, Jg. 29, S. 47–68, 2005.

[18] D. Crystal, *Prosodic systems and intonation in english.* Cambridge university press, 1969.

[19] N. Cummins, A. Baird und B. Schuller, „Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning,“ *Methods*, Jg. 151, S. 41–54, 2018.

[20] M. Domínguez, M. Farrús und L. Wanner, „An automatic prosody tagger for spontaneous speech,“ in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, The COLING 2016 organizing comittee, 2016, S. 377–387.

[21] M. Domínguez, P. Rohrer und J. Soler-Company, „PyToBI: a toolkit for ToBI labeling under python,“ in *Interspeech 2019*, ISCA, 2019, S. 3675–3676.

[22] C. Draxler und F. Schiel. (1995). Bayerisches Archiv für Sprachsignale, Adresse: `https://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html` (besucht am 16. 02. 2021).

[23] K. Eckart und M. Gärtner, „Creating silver standart annotations for a corpus of non-standard data,“ in *13th conference on natural language processing*, Natural language processing association of india, 2016, S. 90–96.

[24] K Eckart, A. Riester und K. Schweitzer, „A discourse information radio news database for linguistic analysis,“ in *Linked data in linguistics.* Springer Berlin Heidelberg, 2012, S. 65–75.

[25] W. Falkena. (2021). xml2struct, Adresse: `https://de.mathworks.com/matlabcentral/fileexchange/28518-xml2struct` (besucht am 06. 04. 2021).

[26] A. Garcia-Moral, R. Solera-Ureña, C. Peláez-Moreno und F. Díaz-de María, „Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems,“ *IEEE transactions on audio speech and language processing*, Jg. 19, S. 468–481, 2011.

[27] J. Garofolo, L. Lamel, W. Fisher u. a., „DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1,“ *NASA STI/Recon technical report N*, Jg. 93, S. 1–79, 1993.

[28] X. Glorot, A. Bordes und Y. Bengio, „Deep sparse rectifier neural networks,“ in *Proceedings of the 14th international conference on artificial intelligence and statistics*, PMLR, 2011, S. 315–323.

[29] J. Goldman, „Easyalign: An automatic phonetic alignment tool under praat,“ in *Interspeech 2011*, ISCA, 2011, S. 3233–3236.

[30] O. Goubanova und S. King, „Bayesian networks for phone duration prediction,“ *Speech communication*, Jg. 50, S. 301–311, 2008.

[31] C. Goutte und E. Gaussier, „A probablistic interpretation of precision, recall and F-Score, with implication for evaluation,“ in *Advances in information retrival*, Springer Berlin Heidelberg, 2005, S. 345–359.

[32] M. Grice und S. Baumann, „Deutsche Intonation und GToBI,“ *Linguistische Berichte*, Jg. 191, S. 267–298, 2002.

[33] M. Grice, S. Baumann, S. Ritter und C. Röhr. (2006). GToBI, Adresse: `http://www.gtobi.uni-koeln.de/` (besucht am 21. 05. 2021).

[34] M. Grice, M. Reyelt, R. Benzmuller u. a., „Consistency in transcription and labelling of German intonation with GToBI,“ in *International conference on spoken language processing 1996*, ISCA, 1996, S. 1716–1719.

[35] W. Han, Z. Zhang, Y. Zhang u. a., „ContextNet: Improving convolutional neural networks for automatic speech recognition with global context,“ in *Interspeech 2020*, H. Meng und B. Xu, Hrsg., ISCA, 2020, S. 3610–3614.

[36] H. Hirsch, „Estimation of noise spectrum and its application to SNR-Estimation and speech enhancement,“ International computer science institute, Techn. Ber., 1993.

[37] S. Ioffe und C. Szegedy. (2015). Batch normalisation: Accelerating deep network training by reducing internal covariate shift, Adresse: `https://arxiv.org/abs/1502.03167` (besucht am 21. 04. 2021).

[38] E. Jacewicz, R. Fox und L. Wei, „Between-speaker and within-speaker variation in speech tempo of american english,“ *The journal of the acoustical society of america*, Jg. 128, S. 839–850, 2010.

[39] N. Jaitly1, P. Nguyen, A. Senior und V. Vanhoucke, „Application of pretrained Deep Neural Networks to large vocabulary speech recognition,“ in *Interspeech 2012*, ISCA, 2012, S. 2578–2581.

[40] A. Katsamanis, M. Black, P. Georgiou u. a., „SailAlign: Robust long speech-text alignment,“ in *Workshop on new tools and methods for very-large scale phonetics research*, 2011.

[41] K. Kohler, B. Peters und M. Scheffers, „The Kiel Corpus of Spoken German read and spontaneous speech,“ Christian-Albrechts-Universität, Techn. Ber., 2017.

[42] B. Kollmeier und M. Wesselkamp, „Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment,“ *The journal of the acoustical society of america*, Jg. 102, S. 2412–2421, 1997.

[43] K. Kucharczlik und S. Hessler. (2016). Bochumer Korpus der gesprochenen Sprache im Ruhrgebiet, Adresse: `https://www.ruhr-uni-bochum.de/kgsr/` (besucht am 18. 02. 2021).

[44] M. Lennes, M. Toivola, L. Wahlberg und E. Aho, „On the use of the glottal stop in finnish conversational speech," in *The phonetics symposium 2006*. Helsingin yliopisto, 2006, S. 93–102.

[45] F. Lin und A. Krizhanovsky, „Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint," *CEUR workshop proceedings*, Jg. 803, S. 1–8, 2011.

[46] MATLAB. (2019). Speech command recognition using deep learning, Adresse: `https://de.mathworks.com/help/deeplearning/ug/deep-learning-speech-recognition.html` (besucht am 30. 04. 2021).

[47] J Mayer, „Transcription of German intonation, the munich system," Universität Stuttgart, Techn. Ber., 2005.

[48] P. Mertens, „The prosogram: semi-automatic transcription of prosody based on a tonal perception model," in *Speech prosody 2004*, ISCA, 2004, S. 549–552.

[49] C. Meyer und I. Gurevych, „Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography," in *Electronic lexicography*, Oxford University Press, 2012, S. 259–291.

[50] B. Möbius und J. van Saten, „Modeling segmental duration in German text-to-speech synthesis," in *International conference on spoken language processing 1996*, ISCA, 1996, S. 2395–2398.

[51] C. Nakatani, J. Hirschberg und B. Grosz, „Discourse structure in spoken language: Studies on speech corpora," 1995 AAAI spring symposium, Techn. Ber., 1995.

[52] S. Nooteboom, „The prosody of speech: Melody and rhythm," *The handbook of phonetic sciences*, Jg. 5, S. 640–673, 1997.

[53] A. Oord, S. Dieleman, H. Zen u. a. (2016). WaveNet: A generative model for raw audio, Adresse: `https://arxiv.org/abs/1609.03499` (besucht am 09. 03. 0021).

[54] K. O'Shea und R. Nash. (2015). An introduction to convolutional neural networks, Adresse: `https://arxiv.org/abs/1511.08458` (besucht am 21. 04. 0021).

[55] M. Ostendorf, P. Price und S. Shattuck-Hufnagel, „The Boston university radio news corpus," Boston university, Techn. Ber., 1995.

[56] ——, (1996). Boston University Radio Speech Corpus, Adresse: `https://catalog.ldc.upenn.edu/LDC96S36` (besucht am 21. 03. 2021).

[57] R. Pandya und J. Pandya, „C5.0 algorithm to improved decision tree with feature selection and reduced error pruning," *International journal of computer applications*, Jg. 117, S. 18–21, 2015.

[58] P. Paschke, „Korpora gesprochener Sprache von/für DaF-LernerInnen,“ in *Gesprochene (Fremd-)Sprache als Forschungs- und Lehrgegenstand.* B. Vogt, 2018, S. 21–51.

[59] B. Peters und K. Kohler, „Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM,“ Christian-Albrechts-Universität, Techn. Ber., 2004.

[60] J. Pitrelli, M. Beckmann und J. Hirschberg, „Evaluation of prosodic transcription labeling reliability in the ToBI framework,“ in *International conference on spoken language processing 1994*, ISCA, 1994, S. 123–126.

[61] B. Pompino-Marschall, *Einführung in die Phonetik.* De Gruyter, 2009.

[62] J. Portisch, M. Hladik und H. Paulheim, „Wiktionary matcher,“ in *Proceedings of the 14th international workshop on ontology matching co*, RWTH, 2020, S. 181–188.

[63] J. Quinlan, *C4.5: Programs for machine learning.* Morgan Kaufmann publishers, 1993.

[64] S. Rapp, „Automatic labelling of German prosody,“ in *International conference on spoken language processing 1998*, ISCA, 1998, S. 1267–1270.

[65] ——, „Automatisierte Erstellung von Korpora für die Prosodieforschung,“ Diss., Universität Stuttgart, 1998.

[66] W. Rappaport, „Über Messungen der Tonhöhenverteilung in der deutschen Sprache,“ *Acustica*, Jg. 8, S. 220–225, 1958.

[67] B. Richter. (2021). HowToPronounce, Adresse: `https://de.howtopronounce.com` (besucht am 07. 04. 2021).

[68] A. Roa-Valverde, S. Sanchez-Alonso, M. Sicilia und D. Fensel, „An approach to measuring and annotating the confidence of Wiktionary translations,“ *Language resources and evaluation*, Jg. 51, S. 319–349, 2017.

[69] A. Rosenberg, „AuToBI – a tool for automatic ToBI annotation,“ in *Interspeech 2010*, ISCA, 2010, S. 146–149.

[70] B. Rues, B. Redecker, E. Koch u. a., *Phonetische Transkription des Deutschen.* narr studienbücher, 2009.

[71] M. Russo und W. Barry, „Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo,“ in *Speech prosody 2008*, ISCA, 2008, S. 419–422.

[72] F. Sajous, E. Navarro, B. Gaume u. a., „Semi-automatic enrichment of crowdsourced synonymy networks: The WISIGOTH system applied to Wiktionary,“ *Language resources and evaluation*, Jg. 47, S. 63–96, 2013.

[73] F. Sasaki und A. Witt, „Linguistische Korpora,“ in *Texttechnologie. Perspektiven und Anwendungen*, H. Lobin und L. Lemnitzer, Hrsg., Stauffenburg, 2004, S. 195 –216.

[74] F. Schiel. (1995). Siemens 100, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasSI100deu.html` (besucht am 16. 02. 2021).

[75] ——, (1996). Verbmobil I, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasVM1deu.html` (besucht am 16. 02. 2021).

[76] ——, (1998). Siemens Synthese Korpus, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasSI1000Pdeu.html` (besucht am 16. 02. 2021).

[77] ——, (2000). Verbmobil II, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasVM2deu.html` (besucht am 16. 02. 2021).

[78] ——, „MAUS goes iterative,“ in *Proceedings of the 4th international conference on language resources and evaluation*, ELRA, 2004.

[79] ——, (2007). BITS Unit Selection synthese corpus, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasBITSUSdeu.html` (besucht am 16. 02. 2021).

[80] ——, (2020). Bayerisches Archiv für Sprachsignale File-Formate, Adresse: `https://www.phonetik.uni-muenchen.de/Bas/BasFormatsdeu.html` (besucht am 09. 04. 2021).

[81] F. Schiel und A. Baumann. (2013). PhonDat 2, Adresse: `https://www.bas.uni-muenchen.de/forschung/Bas/BasPD2deu.html` (besucht am 16. 02. 2021).

[82] F. Schiel, S. Burger, A. Geumann und K. Weilhammer, „The partitur format at BAS,“ Ludwig-Maximilians-Universität München, Techn. Ber., 1997.

[83] F. Schiel, C. Draxler und J. Harrington, „Phonemic segmentation and labelling using the MAUS technique,“ University of Pennsylvania, Techn. Ber., 2011, Workshop on new tools and methods for very-large scale phonetics research.

[84] A. Schiller, S. Teufel und C. Stöckert, „Guidelines für das Tagging deutscher Textcorpora mit STTS,“ IMS Stuttgart/Seminar für Sprachwissenschaften Tübingen, Techn. Ber., 1999.

[85] S. Schlippe T. Ochs und T. Schultz, „Wiktionary as a source for automatic pronunciation extraction,“ in *Interspeech 2010*, ISCA, 2010, S. 2290–2293.

[86] T. Schmidt, „EXMARaLDA–ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache,“ Österreichische Gesellschaft für Artificial Intelligence, 2014.

[87] ——, „DGD – die Datenbank für gesprochenes Deutsch,“ *Zeitschrift für germanistische Linguistik*, Jg. 45, S. 451–463, 2017.

[88] T. Schmidt, S. Dickgießer und J. Gasch. (2012). Datenbank für gesprochenes Deutsch, Adresse: `https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome` (besucht am 17. 02. 2021).

[89] T. Schultz, N. Vu und T. Schlippe, „GlobalPhone: A multilingual text & speech database in 20 languages,“ in *2013 International conference on acoustics, speech and signal processing*, IEEE, 2013, S. 8126–8130.

[90] T. Sejnowski und C. Rosenberg, „Parallel networks that learn to pronounce english text,“ *Complex systems*, Jg. 1, S. 145–168, 1987.

[91] W. Sendlmeier und J. Seebode, „Formantkarten des deutschen Vokalsystems,“ Technische Universität Berlin, Techn. Ber., 2007.

[92] K. Silverman, M. Beckman, J. Pitrelli u. a., „ToBI: a standard for labeling english prosody,“ in *International conference on spoken language processing 1992*, ISCA, 1992, S. 867–870.

[93] P. Skrelin, N. Volskaya, D. Kocharov u. a., „CORPRES,“ in *Text, speech and dialogue*, Springer Berlin Heidelberg, 2010, S. 392–399.

[94] W. Skut und T. Brants. (1998). Chunk tagger – statistical recognition of noun phrases, Adresse: `https://arxiv.org/abs/cmp-lg/9807007` (besucht am 21. 03. 2021).

[95] M. Sokolova, N. Japkowicz und S. Szpakowicz, „Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation,“ in *AI 2006: Advances in artificial intelligence*, Springer Berlin Heidelberg, 2006, S. 1015–1021.

[96] V. Sridhar, S. Bangalore und S. Narayanan, „Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework,“ *IEEE transactions on audio, speech, and language processing*, Jg. 16, S. 797–811, 2008.

[97] S. Toshniwal1, A. Kannan, C. Chiu u. a. (2018). A comparison of techniques for language model integration in encoder-decoder speech recognition, Adresse: `https://arxiv.org/abs/1807.10857` (besucht am 21. 03. 2021).

[98] N. Volskaya und T. Kachkovskaia, „Prosodic annotation in the new corpus of russian spontaneous speech CoRuSS,“ in *Speech prosody 2016*, 2016, S. 917–921.

[99] J. Wales. (2001). Wikipedia, Adresse: `https://www.wikipedia.de/` (besucht am 18. 02. 2021).

[100] ——, (2001). Wikipedia: Wiktionary, Adresse: `https://de.wikipedia.org/wiki/Wiktionary` (besucht am 13. 04. 2021).

[101] P. Warden. (2018). Speech commands: A dataset for limited-vocabulary speech recognition, Adresse: `https://arxiv.org/abs/1804.03209` (besucht am 21. 04. 2021).

[102] C. Widera, „Gelesene und spontane Sprache – Ihre Lebendigkeit und ihre prosodische Realisierung,“ *IKP-Arbeitsbericht NF*, Jg. 6, S. 1–29, 2018.

[103] H. Wiklund. (2021). Index of /mirror/wikimedia.org/dumps/dewiktionary, Adresse: `http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/dewiktionary/` (besucht am 22. 01. 2021).

[104] W. Wu und D. Yarowsky, „Computational etymology and word emergence,“ in *Proceedings of the 12th international conference on language resources and evaluation*, ELRA, 2020.

[105] Y. Wu, R. Zhang und A. Rudnicky, „Data selection for speech recognition,“ in *2007 IEEE workshop on automatic speech recognition understanding*, IEEE, 2007, S. 562–565.

[106] Y. Xu, „ProsodyPro – A tool for large-scale systematic prosody analysis,“ *Tools and resources for the analysis of speech prosody 2013*, S. 7–10, 2013.

[107] T. Zesch, C. Müller und I. Gurevych, „Extracting lexical semantic knowledge from wikipedia and wiktionary,“ in *Proceedings of the 6th international conference on language resources and evaluation*, 2008, S. 1646–1652.

[108] ——, „Using wiktionary for computing semantic relatedness,“ in *Proceedings of the 23rd national conference on artificial intelligence*, 2008, S. 861–866.