

Christoph Draxler und Florian Schiel

8 Moderne phonetische Datenbanken

Erstellung und Datenaufbereitung

Abstract: Dieser Beitrag befasst sich mit Korpora und Datenbanken, die als empirische Grundlage phonetischer Analysen verwendet werden. Moderne phonetische Datenbanken stehen in einer starken Wechselwirkung zwischen Phonetik und Sprachtechnologie: die Sprachtechnologie liefert technische Verfahren zur Speicherung und Analyse gesprochener Sprache und ermöglicht damit überhaupt erst die Verarbeitung auch großer Mengen an Sprachdaten. Im Gegenzug sind Erkenntnisse der Phonetik die Basis vieler sprachtechnologischer Verfahren bzw. tragen zu ihrer Verbesserung bei. Der Beitrag gliedert sich in zwei Teile: im ersten Teil werden die Datenarten phonetischer und sprachtechnologischer Datenbanken beschrieben. Im zweiten Teil wird der in der Praxis relevante Prozess der Erstellung und Nutzung einer phonetischen Datenbank anhand eines konkreten Projekts präsentiert; der Fokus liegt hierbei auf den fachlichen Aspekten: Forschungsfrage, Datensammlung, -aufbereitung und -auswahl sowie Analyse und Interpretation.

Keywords: Datenbanken, Phonetik, Sprachtechnologie

1 Gegenstand und Motivation

Untersuchungsgegenstand der Phonetik ist nach Pompino-Marschall (1995: 3) der „lautliche Aspekt der sprachlichen Kommunikation“. Becker (2012) schreibt in Abgrenzung zur Phonologie:

Die *Phonetik* [...] beschreibt die materielle Seite der Laute sprachlicher Äußerungen, die Abläufe der Sprachproduktion und -wahrnehmung durch die Sprecher, einschließlich der kognitiven und neuronalen Aspekte, mit naturwissenschaftlichen Methoden, etwa mit Experimenten oder Messungen, ohne unmittelbare Berücksichtigung des Sprachsystems. (Becker 2012: 13)

Christoph Draxler, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: draxler@phonetik.uni-muenchen.de

Florian Schiel, Institut für Phonetik und Sprachverarbeitung, LMU München, Schellingstr. 3, D-80799 München, E-Mail: schiel@phonetik.uni-muenchen.de

Reetz & Longman (2009) unterteilen die Phonetik in die drei traditionellen Teilbereiche Produktion, Akustik und Perzeption gesprochener Sprache und sie fügen den Teilbereich Transkription von Sprache hinzu.

- In der *Sprachproduktion* steuern kognitive und neuronale Prozesse Muskelbewegungen im Artikulationstrakt, z. B. Atmung, Zungen- und Lippenbewegungen.
- Die *Akustik* umfasst die physikalischen Grundlagen des Sprachschalls und seiner Übertragung.
- Bei der *Perzeption* wird der Sprachschall beim Hörer wieder in neuronale und kognitive Prozesse umgewandelt und der im Sprachsignal enthaltene Gehalt extrahiert.
- Die *Transkription* befasst sich mit der komplexen Beziehung zwischen wahrnehmbaren Sprachlauten und den zu ihrer Beschreibung verwendeten Symbolen, z. B. dem phonetischen Alphabet der IPA (*International Phonetic Association*) (IPA 1999).

Die moderne phonetische Grundlagenforschung ist häufig korpusbasiert. Harrington (2010) schreibt gleich zu Beginn seiner Einführung:

A speech corpus is a collection of one or more digitized utterances usually containing acoustical data and often marked for annotations. The task in this book is to discuss some ways that a corpus can be analyzed to test hypotheses about how speech sounds are communicated. But why is a corpus needed for this at all? (Harrington 2010: 1)

Er beantwortet diese Frage so, dass Intuition, Introspektion und Transkription die notwendigen Voraussetzungen für phonetische Hypothesen und darauf aufbauend neue Erkenntnis seien, dass aber nur das gemessene Sprachsignal eine objektive Basis für die empirische Überprüfung dieser Hypothesen darstelle.

Im Folgenden stellt er fest, dass es auch weiterhin notwendig sei, eigene phonetische Korpora zu erstellen:

Unfortunately, most kinds of phonetic analysis still require building a speech corpus that is designed to address a specific research question. In fact, existing large-scale corpora ... are very rarely used in basic phonetic research, partly because, no matter how extensive they are, a researcher inevitably finds that one or more aspects of the speech corpus ... are insufficiently covered for the research question to be completed. (Harrington 2010: 6)

Die erwähnten großen Korpora sind Resultat unterschiedlicher Entwicklungen:

- In der phonetischen Grundlagenforschung werden neben dem akustischen Signal auch artikulatorische Messwerte erfasst, was sehr datenintensiv ist. Dazu werden Mess- und bildgebende Verfahren aus der Medizin eingesetzt, die auf nicht-invasive Weise Vorgänge im Inneren des Körpers sichtbar machen.

- In der Sprachtechnologie-Entwicklung, insbesondere der Spracherkennung und -synthese, haben sich statistische Verfahren durchgesetzt. Diese müssen trainiert werden, wozu große und den Anwendungsbereich möglichst vollständig abdeckende Sprachdatensammlungen notwendig sind.
- In der Sprachdokumentation werden bedrohte Sprachen – oftmals einhergehend mit ethnologischer Dokumentation – phonetisch und linguistisch systematisch erfasst, dokumentiert und der Forschung zugänglich gemacht.

Diese Entwicklungen führten zu einem enormen Zuwachs an Daten, der neue Formen der Datenorganisation, -speicherung und -verfügbarkeit notwendig machte. Die zeitgleiche Entwicklung des World Wide Web zu einem weltumspannenden Kommunikationsnetz hat diese Notwendigkeit noch verstärkt.

Darüberhinaus besteht eine starke Wechselwirkung zwischen Phonetik und Sprachtechnologie: die Sprachtechnologie liefert technische Verfahren zur Speicherung und Analyse gesprochener Sprache und ermöglicht damit überhaupt erst die Verarbeitung auch großer Mengen an Sprachdaten, gerade in der Grundlagenforschung. Im Gegenzug sind Erkenntnisse der Phonetik die Basis vieler sprachtechnologischer Verfahren bzw. tragen zu deren Verbesserung bei.¹

Weitere Aspekte sind die Forderung vieler Förderinstitutionen nach langfristigem Datenmanagement, damit aufwendig erstellte Datensammlungen auch nach Abschluss der Projektförderung verfügbar bleiben,² sowie die Frage nach dem kollegialen Datenaustausch und der Reproduzierbarkeit von Studien. Gut dokumentierte und nach dem Stand der Technik erstellte Korpora werden häufig ganz oder in wesentlichen Teilen mehrfach und auch zur Untersuchung verschiedener Fragestellungen genutzt. Die Bereitstellung von Rohdaten, Skripten und Auswertungsroutinen erlaubt es, Studien zu replizieren oder durch das Hinzufügen neuer Annotationen oder Daten den Nutzen oder Wert eines Korpus zu erhöhen. Die Sprachdatenbank TIMIT (Garofolo et al. 1986) ist dafür ein gutes Beispiel: ursprünglich zur Entwicklung von Spracherkennungssystemen erstellt, wurde das Datenbankdesign auf viele Sprachen und technische Kommunikationsmittel und sogar auf artikulatorische Datenbanken übertragen, die ursprünglichen Transkriptionen wurden vielfach korrigiert sowie um zusätzliche Annotationen ergänzt.

¹ So kann z. B. die Wortfehlerrate bei der Spracherkennung durch phonetisches Wissen über die Artikulationsvorgänge um relativ 10–20 % verbessert werden (Richardson, Bilmes & Diorio 2003).

² Siehe dazu die Handreichungen der DFG mit dem Titel „Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora“.

Dieser Beitrag gliedert sich in zwei Teile: im ersten Teil werden die Datenarten phonetischer und sprachtechnologischer Datenbanken beschrieben. Im zweiten Teil wird der in der Praxis relevante Prozess der Erstellung und Nutzung einer phonetischen Datenbank anhand eines konkreten Projekts präsentiert; der Fokus hierbei wird auf den fachlichen Aspekten liegen: Forschungsfrage, Datensammlung, -aufbereitung und -auswahl sowie Analyse und Interpretation.

2 Sprachdatenbanken

In der linguistischen und phonetischen Literatur wird häufig die Bezeichnung *Korpus* verwendet. Damit werden ganz allgemein entweder natürlich vorgefundene oder explizit zusammengestellte Sammlungen sprachlicher Daten bezeichnet (siehe z. B. Gippert, Himmelmann & Mosel 2006; Lemnitzer & Zinsmeister 2006). Um die enge Beziehung zwischen Phonetik und Sprachtechnologie auch terminologisch deutlich werden zu lassen, bevorzugen wir den Begriff *Sprachdatenbank* (engl. *speech database*) und verwenden ihn für eine wohlstrukturierte, auf Dauer angelegte Sammlung von digitalen Daten gesprochener Sprache, dazugehörigen Annotationen und Metadaten.

Sprachdatenbanken bestehen aus *Primär*-, *Sekundär*- und *Metadaten*. Primärdaten sind die bei der Erfassung von gesprochener Sprache erhobenen Rohdaten, Sekundärdaten davon automatisch oder manuell abgeleitete oder erstellte Mess- und Annotationsdaten. Metadaten umfassen die sonstigen erhobenen Daten bzw. beschreiben Aufbau und Struktur der Datenbank.

Primärdaten sind, abgesehen von technischer Konvertierung, prinzipiell unveränderlich. Sekundärdaten können wiederholt und mit verschiedenen Verfahren berechnet oder durch Annotation der Primärdaten erstellt werden, sie können korrigiert und um neue Daten erweitert werden. Metadaten erlauben die Beschreibung und Katalogisierung von sowie die Suche nach Datenbeständen.

Die Annotation von Primärdaten ist ein Kategorisierungsprozess: einem Signalabschnitt werden kategoriale Einheiten eines Symbolinventars zugeordnet. In der Phonetik wird dieser Vorgang als *Transkription* bezeichnet (IPA 1999: 3). Dieser Prozess ist stets mit einer Unsicherheit behaftet und somit niemals *korrekt*, sondern höchstens *plausibel*.

Erst mit der *Digitalisierung* sind das systematische Zusammenführen und die informationserhaltende Konvertierung der unterschiedlichen Datenarten technisch möglich geworden. In digitaler Form können unterschiedliche Transkriptionen miteinander verknüpft, Transkriptionen mit Zeitsignalen aligniert

und Zeitsignale miteinander synchronisiert werden. Dies eröffnet der Phonetik, der Sprachtechnologie und anderen sprachverarbeitenden Gebieten ganz neue Möglichkeiten des Zugriffs und des Erkenntnisgewinns.

2.1 Datenarten

Grundvoraussetzung phonetischer Datenbanken sind die eigentlichen Daten in digitaler Form. Jede Datenart, jedes Messverfahren hat je eigene Datenmodelle und -formate. Beschreibungen dieser Datenmodelle und -formate müssen öffentlich verfügbar sein, damit sie auch unabhängig von den Tools, mit denen sie erstellt wurden, langfristig zugänglich sind.

Zur Beschreibung der in phonetischen Datenbanken gespeicherten Daten eignet sich eine erste Unterteilung in *zeitunabhängige* und *zeitbezogene* Daten. Zeitunabhängig sind Daten, die außer der Tatsache, dass sie eine gesprochene Äußerung wiedergeben, keine Zeitinformation enthalten, z. B. der orthographische Wortlaut einer Äußerung oder eine phonetische Transkription.

Zeitbezogene Daten sind entweder Ereignis- oder Intervalldaten. Ein Ereignis hat nur einen Zeitpunkt und keine Dauer, ein Intervall hat einen Anfangszeitpunkt und eine Dauer. Beispiel für ein Ereignis ist ein Wendepunkt in einer Intonationskurve, Beispiel für ein Intervall der einem Wort, Phonem oder

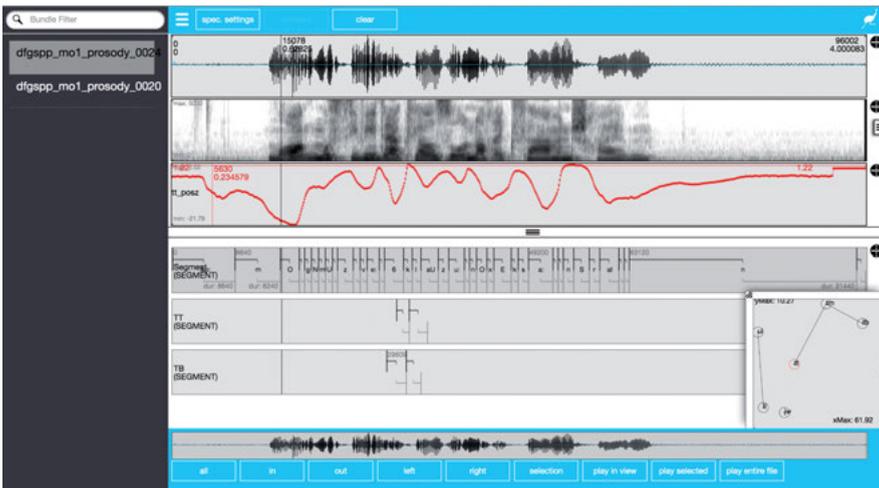


Abb. 8.1: Emu WebApp Labeler mit drei Signaldarstellungen (Oszillogramm, Sonagramm, EMA-Sensormessdaten), drei Annotationsebenen (Segment, TT [*tongue tip*] und TB [*tongue back*]) sowie einer Visualisierung der Bewegungen der EMA-Sensoren in einem eigenen Fenster unten rechts.

Allophon zugeordnete Signalabschnitt einer Aufnahme (siehe Abb. 8.1). Neben dem Zeitbezug unterscheidet man die Daten nach den jeweiligen Medien. In Sprachdatenbanken sind dies in der Regel Text-, Audio- und Video- sowie Sensordaten.

Diese Daten werden soweit möglich getrennt erfasst, um sie sowohl isoliert betrachten als auch miteinander in Beziehung setzen zu können. Die eigentliche Aufgabe von Sprachdatenbanken ist, die Daten so miteinander zu verknüpfen, dass sie die Beantwortung phonetischer und sprachtechnologischer Fragestellungen erlauben. Diese an die Anforderungen der Sprachforschung und -verarbeitung angepasste Datenmodellierung unterscheidet phonetische Datenbanken von Datenbanken für andere Anwendungsgebiete oder Universaldatenbanken.

2.2 Datenmodelle für phonetische Daten

Bird & Liberman (2001) entwickelten *Annotation Graphs* als allgemeines formales Modell für Annotationen gesprochener Sprache. Annotation Graphs sind im Wesentlichen Sammlungen gerichteter azyklischer binärer Graphen. Die Knoten tragen einen eindeutigen Bezeichner und einen optionalen Zeitstempel, mit dem sie einen Zeitpunkt in einer Signaldatei angeben. Die Kanten haben einen Annotationstyp und ein Annotationslabel. Pfade innerhalb eines Annotationsgraphen sind die transitive Hülle von Kanten eines Annotationstyps; die Knoten eines Pfades sind nach Zeitstempel geordnet, eine Kante darf nicht zu einem Knoten mit einem früheren Zeitstempel führen.

Die Autoren zeigen, dass sich die in der Literatur beschriebenen oder in Annotationseditoren implementierten Annotationsnotationen für Aufnahmen gesprochener Sprache mit Annotation Graphs darstellen lassen.

Annotation Graphs eignen sich aber nur eingeschränkt für phonetische Datenbanken. Auf theoretischer Ebene ist die Beschränkung auf gerichtete azyklische Graphen in der Praxis zu eng, denn es gibt Phänomene in gesprochener Sprache, die als diskontinuierliche Elemente oder in Form zyklischer Strukturen beschrieben werden. In der Praxis ist das Fehlen eines Datenbankschemas ein Problem, da deshalb eine automatische Integritätskontrolle der Datenbank nicht möglich ist.

Ein erweitertes Datenmodell, das sowohl Annotationsebenen als auch ein explizites Datenbankschema vorsieht, ist im Emu-System realisiert (Harrington et al. 1993; Cassidy & Harrington 1996, 2001). In Emu enthält eine Annotations-ebene nur Daten eines bestimmten Annotationstyps – phonetische Segmentation, phonemische Transkription, Silbe, orthographischer Wortlaut usw. Die

Annotationsebenen sind hierarchisch angeordnet, wobei es innerhalb einer Datenbank mehrere Hierarchien geben kann. Innerhalb einer Annotations-ebene sind die Elemente sequenziell angeordnet, zwischen den Elementen benachbarter Ebenen gibt es Dominanzbeziehungen. Ein Schema beschreibt die in einer Sprachdatenbank vorgesehenen Annotationsebenen und in welcher quantitativen Beziehung Elemente einer Ebene zu denen benachbarter Ebenen stehen.

Das Sprachdatenbanksystem Emu-SDMS ist die von Winkelmann, Harrington & Jänsch (2017) vollständig in R implementierte neue Fassung von Emu. Emu-SDMS besteht aus der Emu WebApp zur graphischen Darstellung von Sprachdaten in einem Webbrowser sowie der eigentlichen EmuR-Sprachdatenbank und der Signalverarbeitung *wrassp*, die beide auf das lokale Dateiverzeichnis mit Audio- und Annotationsdateien zugreifen (siehe Abb. 8.1).

EmuR unterscheidet drei Klassen von Annotationselementen:

- *Item*: Element ohne Zeitbezug
- *Event*: Ereignis-Element mit einem Zeitpunkt
- *Segment*: Intervall-Element mit Anfangszeitpunkt und Dauer

Innerhalb einer Ebene sind nur Elemente eines Typs erlaubt und sie sind sequenziell angeordnet. Der Inhalt eines Annotationselements ist in seinen Labels gespeichert, jedes Annotationselement ist durch einen eindeutigen, vom System vergebenen Bezeichner identifiziert.

Das Schema legt fest, welche Ebenen miteinander verknüpft sind. Dabei sind 1 : 1-, 1 : *n*- und auch *n* : *m*-Dominanzbeziehungen möglich, außerdem kann eine Ebene in mehr als einer Hierarchie vorkommen (siehe Abb. 8.2).

Die Abfragesprache EQL (*Emu Query Language*) erlaubt die kompakte Formulierung von Abfragen von Dominanz- und Sequenzbeziehungen. Eine Abfrage in der Datenbank AE nach Silben, die ein Segment [n] auf der Ebene *Phonetic* enthalten, wird wie folgt formuliert:

```
[Phonetic == n ^ #Syllable = ~.*]
```

Der Operator ^ ist der Dominanzoperator, d. h. die Ebene *Phonetic* wird von der Ebene *Syllable* dominiert. # zeigt an, dass die Zeitinformation für die Ebene *Syllable* angezeigt werden soll – in diesem Fall bedeutet dies, dass die Zeitinformation aus der Ebene *Phonetic* in die Ebene *Syllable* propagiert wird, d. h. Anfang und Dauer der ganzen Silbe zurückgegeben werden sollen; für die Silben gelten keine weiteren Einschränkungen.

Der Dominanzoperator ^ ist rekursiv, d. h. er kann über mehrere Ebenen einer Hierarchie berechnet werden.

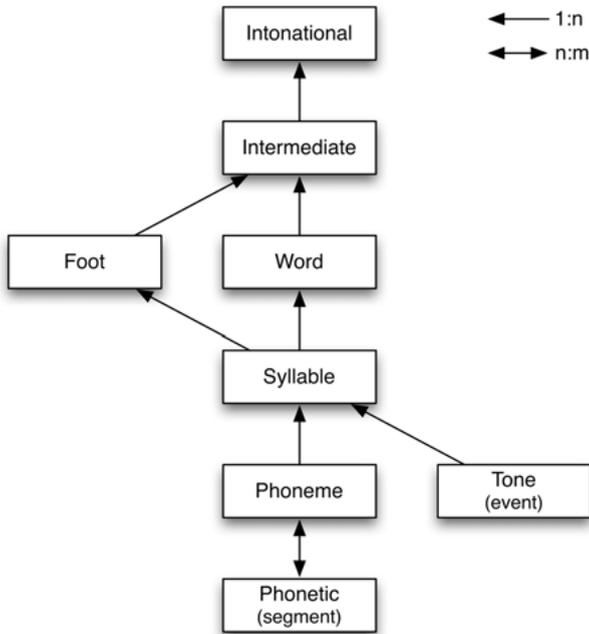


Abb. 8.2: Schema der Datenbank AE (*Australian English*) mit mehreren Hierarchien und den beiden zeitbasierten Annotationsebenen *Phonetic* der Klasse *Segment* und *Tone* der Klasse *Event* aus Harrington (2010: 99).

Abfragen in der Emu-Datenbank ergeben Segmentlisten der Form

```
AUDIOFILE ELEMENT BEGIN DURATION
```

wobei AUDIOFILE die Audiodatei der Äußerung ist, ELEMENT das Annotations-element der gewählten Ebene, und BEGIN und DURATION die Zeitangaben der zeitbezogenen Annotationsebene der aktuell ausgewählten Hierarchie.

Die direkte Einbindung in die Statistiksoftware R erlaubt in Kombination mit dem Signalverarbeitungspaket *wrassp* statistische Auswertungen des Datenbestands sowie die Aufbereitung der Daten für die Visualisierung.

Neben Emu gibt es weitere Ansätze für phonetische Datenbanken. So beschreiben Draxler & Kleiner (2015) eine phonetische Datenbank auf der Basis eines relationalen Datenbanksystems. Damit sind auf einfache Weise Abfragen auch über mehrere Sprachdatenbanken möglich und auch nichtsprachliche Daten wie z. B. Orts- und Signaldateiangaben können redundanzfrei gespeichert werden. Allerdings müssen die Abfragen in SQL formuliert werden und

Signalverarbeitung, statistische Auswertung und Visualisierung müssen außerhalb des Datenbanksystems erfolgen.

2.3 Textdaten

Textdaten sind Lesetexte, Annotationen, orthographische, breite phonemische und enge phonetische Transkriptionen, Datentabellen und statistische Rohdaten, frei formatiert oder mit definierter Struktur. Sie sind vorgegeben bzw. wurden durch manuelle oder automatische Annotation erzeugt. Technisch sollten Textdaten in Unicode und UTF-8-Kodierung vorliegen, die Struktur von Textdokumenten sollte öffentlich definiert oder selbstbeschreibend sein.^{3,4}

Die IPA empfiehlt Transkriptionen auf mindestens zwei Ebenen: breite phonemische Transkription der Wörter in Zitierform und eine enge phonetische Transkription (IPA 1989: 81), dazu kommt üblicherweise noch eine orthographische Transkription (Gibbon, Moore & Winski 1997: 152). Je nach Transkriptionstyp werden verschiedene, an die jeweilige Aufgabe angepasste Editoren verwendet. Die orthographische Transkription soll möglichst rasch erstellt werden, phonetische Fachkenntnisse sind in der Regel nicht notwendig. Die Organisation der Transkriptionsarbeit via Crowdsourcing und das Verwenden von webbasierten Editoren mit entweder einem Standard-Audioplayer oder einem einfachen Oszillogramm haben sich bewährt.

Abbildung 8.3 zeigt die 2D-Ansicht für lange Signaldateien sowie das Editierfenster des webbasierten Editors OCTRA (Pömp & Draxler 2017).

Editoren für die phonetische Mehr-Ebenen-Annotation bieten in der Regel eine partiturartige graphische Darstellung. Bekannte Editoren sind Praat (Boersma & Weenink 1996), EXMARaLDA (Schmidt & Wörner 2005), ELAN⁵ (Sloetjes, Russel & Klassmann 2007) und Emu Webapp (Winkelmann, Harrington & Jänsch 2017). Abbildung 8.1 zeigt als Beispiel den Emu WebApp Labeller.

Diese Editoren unterscheiden sich z. T. erheblich in ihrer Bedienung und ihrem Funktionsumfang. Die damit erzeugten Annotationsdaten sind nur teilweise kompatibel, so dass beim Datentransfer Informationsverlust auftreten kann. Für eine Diskussion dieses Themas siehe z. B. Schmidt et al. (2009) oder Draxler et al. (2011).

³ Das IPA-Alphabet zur Wiedergabe phonetischer Zeichen ist integraler Bestandteil der Unicode-Zeichentabelle und kann somit in Unicode-kompatibler Software verwendet werden.

⁴ Die *Text Encoding Initiative* (TEI) hat einen Standard zur Transkription gesprochener Sprache vorgeschlagen, siehe <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html> (letzter Zugriff: 7.11.2017).

⁵ ELAN ist primär für die Annotation von Videos gedacht.

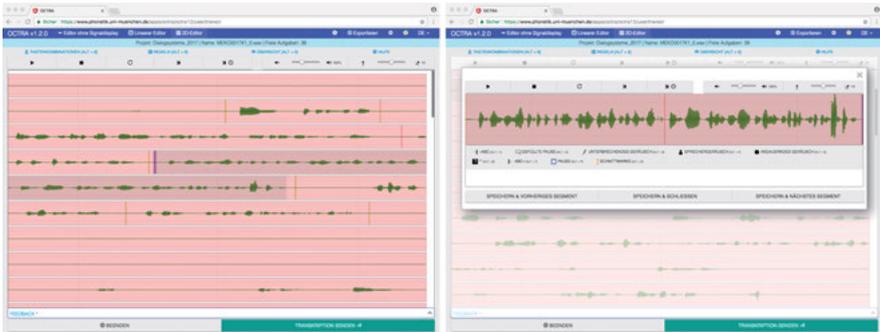


Abb. 8.3: OCTRA Annotationseditor mit einem 2D-Oszillogramm zur Darstellung langer Audiodateien (links) und mit überlagertem Transkriptionseditor für Segmente (rechts).

2.4 Audiodaten

Digitale Audiodaten erfassen das akustische Signal. Die Abtastrate (engl. *sampling rate*) gibt an, wie viele diskrete Werte pro Zeiteinheit erfasst werden, die Quantisierung, wie viele Werte unterschieden werden können. Üblicherweise verwendet man für Sprachdaten Abtastraten von mindestens 16.000 Messpunkten pro Sekunde (abgekürzt 16 kHz) und eine Quantisierung von mindestens 16 Bit. Sprachaufnahmen erfolgen über ein oder mehrere Mikrofone, die über einen Analog/Digital-Konverter an ein Aufnahmegerät oder einen Rechner angeschlossen sind. Ziel von Sprachaufnahmen ist in der Regel ein unter den gegebenen Umständen und für die geplante Untersuchung optimales Sprachsignal zu bekommen. Dazu stehen verschiedene Mikrofontypen zur Verfügung, die Aufnahmen können in kontrollierten Bedingungen im Studio oder im Feld erfolgen.⁶

Für standardisierte Sprachaufnahmen, in denen z. B. vorbereitete Sätze vorgelesen werden, ist der Einsatz von spezieller Software sinnvoll. Die Software SpeechRecorder (Draxler & Jänsch 2004) führt skriptgesteuert Sprachaufnahmen durch und schreibt jede Aufnahme automatisch in eine separate Datei. Damit sind teil- oder vollautomatisch ablaufende Aufnahmesitzungen möglich, ein nachträgliches Schneiden von Signalen kann weitgehend vermieden werden (siehe Abb. 8.4).

Audiodaten sollten entweder gar nicht oder verlustfrei komprimiert werden. Verlustbehaftete Kompression entfernt aus dem Sprachsignal für den Menschen nicht wahrnehmbare Anteile, diese können aber in automatischen Verfahren durchaus relevant sein.

⁶ Eine Einführung in Aufnahmetechnik und -situationen gibt Draxler (2008: 132–169).



Abb. 8.4: SpeechRecorder Sprecheransicht (links) und Aufnahmeleiteransicht (rechts). In der Aufnahmeleiteransicht sind zusätzlich ein Oszillogramm zur Beurteilung des Mikrofonpegels sowie die Liste aller schon aufgenommenen bzw. noch ausstehenden Aufnahme-Items zu sehen.

2.5 Videodaten

Videoaufnahmen erfassen sichtbare Aspekte gesprochener Sprache. Das reicht von Gesamtaufnahmen einer Aufnahmesituation über Halbtotale und Gesichtsaufnahmen bis hin zu Nahaufnahmen der Lippen. Üblich sind aktuell Videoaufnahmen mit einer Bildgröße von mindestens HD-Auflösung (d. h. 1920×1080 Pixel) mit einer Bildwiederholrate von 30 Bildern pro Sekunde; für spezielle Anwendungen sind auch Bildwiederholraten von weit über 100 Bildern pro Sekunde notwendig.

Digitales Video wird in der Regel verlustbehaftet komprimiert, um die Datenrate und den damit verbundenen Speicherbedarf zu begrenzen.

2.6 Sensordaten

Sensordaten werden vor allem im Bereich der Sprachproduktion erhoben. Damit werden Strukturen und Bewegungen im Körperinneren erfasst und teilweise auch direkt sichtbar gemacht. Aus den komplexen Muskel- und Artikulatorbewegungen sind Rückschlüsse auf die neuronale Ansteuerung und muskuläre Koordination sowie auf Art und Organisation des mentalen Lexikons möglich.

Die im Folgenden beschriebenen Sensordaten sind von besonderer Relevanz für moderne Sprachdatenbanken.

2.6.1 Elektropalatographie

Bei der Elektropalatographie erfassen in einer Matrix angeordnete Elektroden in einem künstlichen Gaumen den Kontakt mit der Zunge. Damit lassen sich

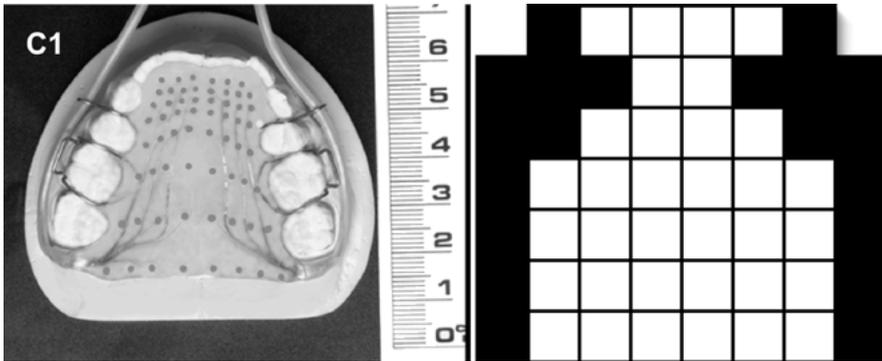


Abb. 8.5: Künstlicher Gaumen mit 62 Elektroden für die Elektropalatographie [links, aus Gibbon & Crampin (2001: 98)] und schematisches Elektropalatogramm des /s/ in der Äußerung /a s a/ (rechts).

insbesondere Konsonanten gut erfassen, weil diese durch Enge- oder Verschlussbildung im Artikulationstrakt gebildet werden.

Die Abtastrate beträgt bis zu 200 Hz, die Quantisierung mindestens 64 Bit. Mit dem künstlichen Gaumen kann nur der Bereich des harten Gaumens abgedeckt werden. Der Zungenkontakt ganz vorne im Artikulationstrakt, an Lippen oder Zähnen, oder ganz hinten im Bereich des Velums und des Rachens kann nicht erfasst werden (siehe Abb. 8.5).

2.6.2 Laryngographie und Laryngoskopie

Bei der Laryngographie werden die Schwingungen der Stimmlippen im Kehlkopf mit Elektroden auf der Haut gemessen (siehe Abb. 8.6), bei der Laryngoskopie werden sie gefilmt. Dazu wird eine kleine Videokamera mit Lichtquelle vom Rachenraum aus auf die Stimmlippen gerichtet. Erwachsene Männer haben eine Grundfrequenz von 50 bis 150 Hz, Frauen von 200 bis 300 Hz, Kinder deutlich darüber. Daher sind sehr hohe Bildwiederholraten von 100 bis über 2.000 Hz notwendig.

Häufig werden in Laryngographie-Videos die Konturen der Stimmlippen automatisch ermittelt oder manuell erfasst und in Form von Vektordaten gespeichert. Damit ist eine erhebliche Reduktion des Datenumfangs und eine präzise Messung der Geschwindigkeit und Beschleunigung einzelner Punkte auf den Stimmlippen möglich.

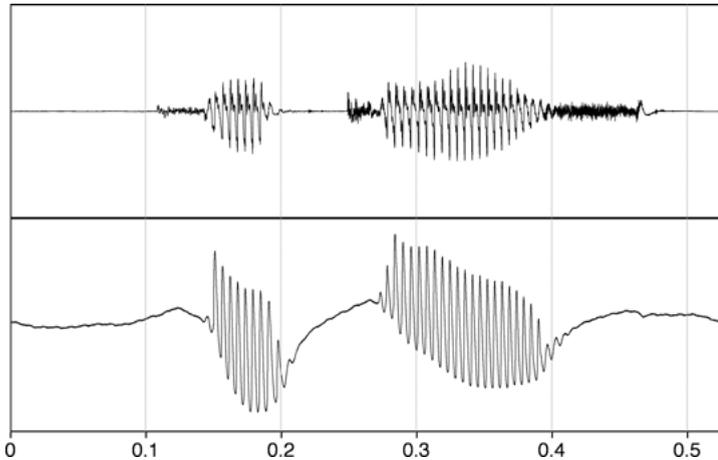


Abb. 8.6: Oszillogramm und Laryngographensignal der Äußerung /p a t a x/. Die stimmlosen Plosive /p/ und /t/ und der stimmlose Frikativ /x/ sind im Oszillogramm deutlich zu erkennen, im Laryngogramm nicht. Die beiden /a/-Vokale sind, da sie stimmhaft sind, als synchrone Schwingungen sowohl im Oszillogramm als auch im Laryngogramm zu sehen (aus Draxler 2008: 71).

2.6.3 Echtzeit Magnet-Resonanz-Tomographie

Magnet-Resonanz-Tomographie (MRT, engl. *magnet resonance imaging MRI*) ist ein bildgebendes Verfahren zur Darstellung des weichen Gewebes im Körper. Dabei wird der Körper schrittweise in einzelnen Schichten erfasst, aus denen zweidimensionale Schnittbilder oder auch dreidimensionale Darstellungen berechnet werden können.

In der Phonetik wird MRT zur Darstellung des Artikulationstrakts beim Sprechen verwendet. Damit lassen sich die Positionen der Zunge, des Velums, der Lippen und des Kiefers visualisieren. In vielen MRT-Geräten muss der Sprecher liegen. Das beeinflusst die Geometrie des Artikulationstrakts.

Werden mehrere MRT-Bilder in so kurzer Zeit nacheinander aufgenommen, dass sie als flüssig ablaufendes Video betrachtet werden können, spricht man von Echtzeit-MRT. Aktuell sind Bildwiederholraten von über 50 Bildern pro Sekunde möglich. In der Regel sind in MRT-Filmen Details mit einer Kantenlänge von knapp 2 mm zu erkennen.

MRT-Aufnahmen sind sehr aufwendig und teuer, sie können nur an wenigen Labors weltweit durchgeführt werden. Da die Geräte sehr laut sind und ein starkes Magnetfeld erzeugen, sind synchrone Sprachaufnahmen nur mit nichtmetallischen Mikrofonen und starken Störgeräuschen möglich.

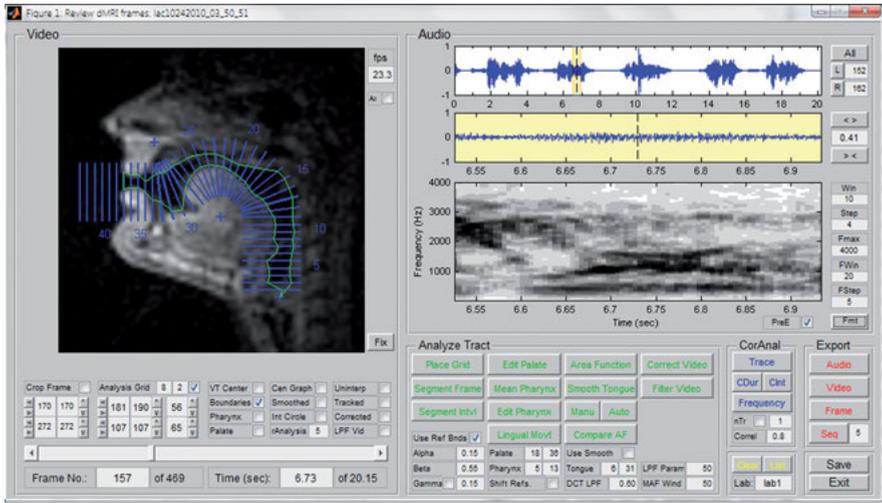


Abb. 8.7: MRT-Video-Standbild mit überlagerter berechneter Kontur des Artikulationstrakts sowie Oszillogramm und Sonogramm (Narayanan et al. 2014: 1310).

Wie bei anderen bildgebenden Verfahren kann durch Tracken der Konturen in den einzelnen Videobildern eine datenreduzierte Vektordarstellung erzeugt werden, die kinetische Daten einzelner Punkte bei der Artikulation liefert (siehe Abb. 8.7).

2.6.4 Ultraschall

Mit Ultraschall oder *Sonographie* kann man Schallreflexionen an den Übergängen zwischen Gewebe und Luft messen. In der Phonetik wird Ultraschall zur nicht-invasiven Erfassung der Zungenbewegungen verwendet; dazu wird die Sonde auf der Haut an der Unterseite des Unterkiefers angesetzt. Ultraschallaufnahmen gelten als ungefährlich. Das korrekte Positionieren der Sonde ist für ein klares Bild wichtig. Die Zunge kann bis fast zur Zungenspitze erfasst werden (siehe Abb. 8.8). Die Abtastrate kann bei reduzierter räumlicher Auflösung über 100 Bilder pro Sekunde betragen, wird aber häufig auf die Bildrate beim Videoexport, d. h. 25–30 Bilder pro Sekunde, beschränkt. Der geringe technische Aufwand, die gute Softwareunterstützung sowie die einfache Handhabung haben zu einer raschen Verbreitung von Ultraschall geführt. Synchron Audioaufnahmen sind möglich, tragbare Ultraschallgeräte erlauben Aufnahmen im Feld.

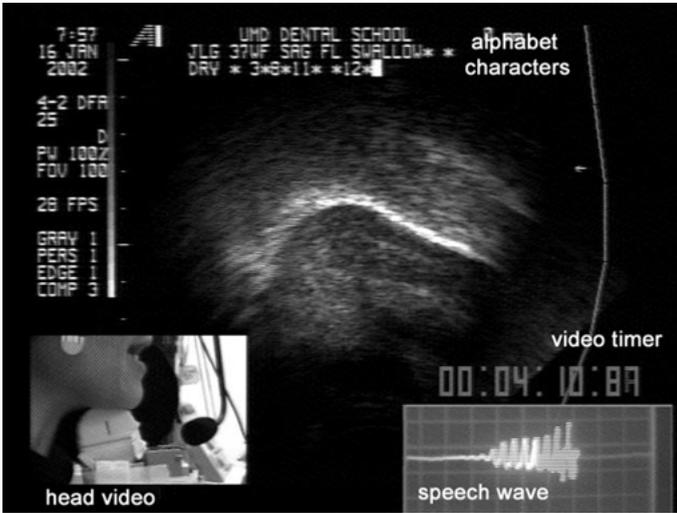


Abb. 8.8: Ultraschallaufnahme der Zungenbewegung. Die Zungenkontur ist durch die helle Linie deutlich sichtbar. Unten links ist die Sensorposition eingeblendet, unten rechts das Oszillogramm der Äußerung (aus Stone 2005: 25).

2.6.5 Elektromagnetische Artikulation

Bei der Bewegung von Spulen in einem Magnetfeld wird in den Spulen ein Strom induziert. Bei der elektromagnetischen Artikulographie (EMA) befinden sich kleine Spulen auf Zunge und Lippen des Sprechers. Moderne EMA-Geräte erlauben eine freie Bewegung des Kopfes im magnetischen Feld und eine Erfas-

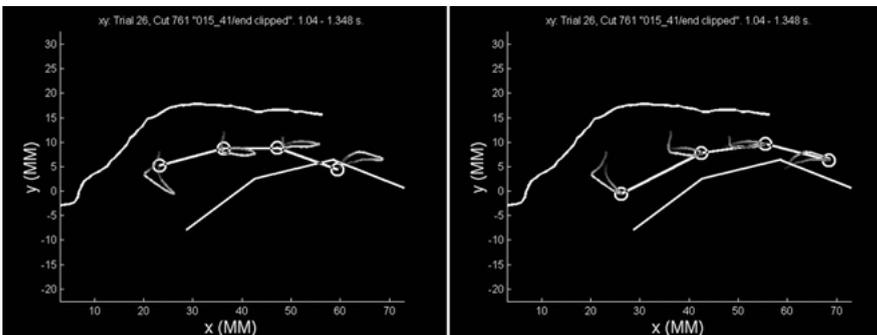


Abb. 8.9: EMA-Zungenkonturen in der Äußerung „tote“. Die obere Kurve ist der harte Gaumen, die Mundöffnung ist links. Die untere viergliedrige Linie ist die Zungenkontur bei gehaltenem /o:/, die mittlere Linie die Zungenkontur für die Phoneme /t/ (links) und /o/ (rechts).

sung der Bewegung der Spulen in fünf Dimensionen (x,y,z -Koordinaten sowie Rotation in zwei Ebenen). Die EMA liefert Positionsdaten der einzelnen Spulen, aus denen zweidimensionale (siehe Abb. 8.9) oder räumliche Darstellungen der Artikulation berechnet werden. Mit der EMA können Details im Bereich eines Millimeters gemessen werden, die Abtastrate beträgt bis zu 1.250 Hz. EMA-Aufnahmen sind aufwendig und erfordern geschultes Personal.

2.6.6 Eyetracking

Beim Eyetracking werden Blickrichtung, -bewegung und Veränderung der Pupille gemessen. Die Alignierung der Bewegung mit dem akustischen Sprachsignal erlaubt Rückschlüsse auf mentale Prozesse bei der Sprachverarbeitung, sowohl bei der Produktion als auch der Perzeption. Huettig, Rommers & Meyer (2011) geben einen Überblick der Studien mit Eyetracking, Holmqvist, Nyström & Mulvey (2012) diskutieren die Zuverlässigkeit von Eyetracking-Messdaten.

Die Sensoren für das Eyetracking sind entweder in der Nähe des zu erfassenden Bildschirms angeordnet, oder sie befinden sich in einer speziellen Brille. Typische Abtastraten liegen im Bereich von 30 bis 300 Hz, der anvisierte Punkt wird mit x,y -Koordinaten angegeben, woraus sich bei bekannter Entfernung des gemessenen Auges von der betrachteten Bildfläche der Winkel der Bewegung berechnen lässt. Wichtige Angaben zur Qualität der Messung sind die Abweichung (engl. *accuracy* oder *offset*), die angibt, wie weit der anvisierte Punkt vom Ziel abweicht, und die Streuung (engl. *precision*), die angibt, um welchen Betrag die Messwerte bei unveränderter Blickrichtung streuen; beide werden in Winkelgraden angegeben.

2.6.7 Röntgenbilder und -filme

Wegen ihrer Gesundheitsgefährdung werden Röntgenaufnahmen bzw. das damit verwandte Verfahren X-Ray Microbeam (XRMB) nur noch in speziellen Situationen, z. B. zur Vorbereitung und Nachsorge von Operationen, durchgeführt. Es gibt allerdings viele historische Röntgenaufnahmen, sowohl Stand- als auch Bewegtbilder, die digitalisiert wurden und nun zu Lehr- und Forschungszwecken genutzt werden (Abb. 8.10).⁷

⁷ Diese Abbildung sowie die EMA-Darstellungen in Abbildung 8.9 wurden freundlicherweise von Phil Hoole vom Institut für Phonetik und Sprachverarbeitung der LMU München zur Verfügung gestellt.



Abb. 8.10: Röntgenaufnahme des Mundraums eines Sprechers bei der Artikulation des Diphthongs /a/ und Sonagramm der Äußerung *It's ten below outside*.

XRMB ist ein Verfahren, bei dem ein extrem schmaler Röntgenstrahl verwendet wird, um 2–3 mm große Goldkügelchen, die in Längsrichtung mittig auf die Zunge, die Lippen, den Kiefer und den weichen Gaumen geklebt werden, in ihrer Bewegung zu erfassen. Durch den sehr schmalen Röntgenstrahl kann die Belastung der aufgenommenen Person stark reduziert werden. Die Kügelchen sind im Röntgenbild deutlich erkennbar, so dass durch Interpolation zwischen den Messpunkten die Kontur der Zunge berechnet werden kann.⁸

⁸ XRMB-Aufnahmen werden seit 1993 praktisch nicht mehr durchgeführt.

2.7 Beispiele phonetischer Datenbanken

Seit der Veröffentlichung von TIMIT wurden laufend weitere Sprachdatenbanken für die phonetische Grundlagenforschung und die Entwicklung von Sprachtechnologie erstellt. Um die Sicht- und Verfügbarkeit dieser Sprachdatenbanken zu verbessern, wurden Zentren wie das *Linguistic Data Consortium* (LDC) in den USA, die *European Language Resources Association* (ELRA) in Europa, das *Bayerische Archiv für Sprachsignale* (BAS) und das *Hamburger Zentrum für Sprachkorpora* (HZSK) in Deutschland und viele weitere ähnliche weltweit gegründet. In sog. Repositories katalogisieren sie die Sprachdatenbanken, bieten Such- und Blätterfunktionen im Datenbestand an und erlauben das Herunterladen bzw. Lizenzieren von Sprachdatenbanken. Viele Zentren bieten darüberhinaus web-basierte sprachtechnologische Dienste an. Ein Beispiel dafür ist die automatische phonetische Segmentation von Sprachaufnahmen auf den Webseiten des BAS (<http://clarin.phonetik.uni-muenchen.de/BASWebServices/> [letzter Zugriff: 7. 11. 2017]).

Die weitaus meisten der in Repositories verfügbaren Datenbanken sind akustische Sprachdatenbanken. Abbildung 8.11 zeigt beispielhaft die Webseite des BAS Repositories. Dort sind mehr als 40 Sprachdatenbanken aufgelistet, die größtenteils am Institut für Phonetik und Sprachverarbeitung der LMU München erstellt wurden. Zunehmend kommen aber auch von Dritten erstellte Sprachdatenbanken hinzu, denn das Repository hält diese Daten auch nach Auslaufen der jeweiligen Projektfinanzierung vor.

The screenshot shows the BAS CLARIN Repository website. At the top, there is a navigation bar with the CLARIN CENTRE B logo, a 2014 DSA 2017 seal, and a HELPDESK button. Below the navigation bar, there is a login prompt: "You are not yet authenticated to have access to the BAS repository. Click here to login either via your academic institution or via your CLARIN IDP account. If you are not an academic, or if your academic institution is not part of the DEF-AAI, you can register here to get a CLARIN IDP account. Please read our privacy policies for AAI authentication." On the left, there is a "Menu" section with links to Repository, ENG, Search, and Links. The main content area displays a list of language data banks, each with details such as Owner, Title, Modality, Recorded language(s), and Access. The list includes:

- ASD**: Owner: Institut für deutsche Sprache, RS 6-13, 68161 Mannheim; Title: Audiotapes Siebenbürgisch-Sächsischer Dialekte; Modality: Spoken; Recorded language(s): Bavarian, German, Romanian, Undefined; Access: restricted (contact bas@bas.uni-muenchen.de to obtain a license)
- ASCa**: Owner: Institut für Romanische Philologie, Ludwig-Maximilians-Universität München; Title: LMU AsCa; Modality: Spoken; Recorded language(s): Italian; Access: restricted (contact bas@bas.uni-muenchen.de to obtain a license)
- BROTHERS**: Owner: Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München; Title: Untersuchung audiotiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Bruderpaaren unter forensischen Aspekten; Modality: Spoken; Recorded language(s): German; Access: free if you are a scientist, otherwise contact bas@bas.uni-muenchen.de to obtain a license
- CL1**: Owner: Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München; Title: BAS Thesis data Veronika Neumeyer: CI Articulation; Modality: Spoken; Recorded language(s): German; Access: free if you are a scientist, otherwise contact bas@bas.uni-muenchen.de to obtain a license
- Ch-Jugendforscher**: Owner: Fabrice Trépoigt/SNF-DSPR; Title: Schweizer Jugendsprache; Modality: Spoken

Abb. 8.11: Webseite des BAS Repository mit Sprachdatenbanken (<http://clarin.phonetik.uni-muenchen.de/BASRepository/> [letzter Zugriff: 7. 11. 2017]).

Im Gegensatz zu den akustischen Sprachdatenbanken gibt es nur sehr wenige verfügbare Sprachdatenbanken mit artikulatorischen Daten. Die im Folgenden aufgeführten vier Sprachdatenbanken sind daher in der Forschung vielgenutzte und trotz ihres teilweise schon weit zurückliegenden Entstehungsdatums immer noch aktuelle Ressourcen. So nennt z. B. Richmond, Hoole & King (2011) in einem Überblick neun Phonetik- und Sprachtechnologiefelder, in denen Arbeiten auf der Basis der 1999 erstellten MOCHA EMA-Sprachdatenbank publiziert wurden.

Diese vier Sprachdatenbanken enthalten hauptsächlich englische Sprachdaten. Vergleichbare öffentlich verfügbare Sprachdatenbanken mit primär deutschen Sprachaufnahmen sind uns nicht bekannt.

2.7.1 X-Ray Microbeam

Für die *X-Ray Microbeam Speech Production*-Sprachdatenbank (Westbury, Turner & Dembrowski 1994) wurden in hauptsächlich zwei Phasen von 30 bzw. 10 Monaten in den Jahren 1987 bis 1991 XRMB-Aufnahmen durchgeführt.⁹ In die Datenbank aufgenommen wurden die Daten von insgesamt 57 Personen und je ca. 18 Minuten Sprachmaterial. Neben den Messdaten sind die gelesenen Sätze sowie ihre phonetischen Umschriften verfügbar.

Die XRMB-Sprachdatenbank wurde von Anfang an als offene Ressource für die Wissenschaft erstellt. Forscherinnen und Forscher sollten diese sehr aufwendigen Aufnahmen ohne Einschränkungen nutzen können. Auch war es explizites Ziel, vielen verschiedenen Forschergruppen Zugang zum Aufnahmegerät zu ermöglichen, um die Verbreitung und gemeinsame Nutzung der Daten zu fördern.

2.7.2 MOCHA EMA

Die MOCHA EMA-Sprachdatenbank wurde im November 1999 am Queen Margaret University College Edinburgh aufgenommen (Wrench & Hardcastle 2000). Sie enthält synchrone Audio-, Elektropalatographie-, Laryngographie- und EMA-Aufnahmen von 460 Sätzen aus der britischen TIMIT-Sprachdatenbank. Diese

⁹ Das Kapitel „A Short History of the UW XRMB Facility“ im Bericht von Westbury, Turner & Dembrowski (1994) zeigt eindrucksvoll, dass schon Entwurf und Realisierung des XRMB-Geräts mehrjährige Projekte waren, mit denen technologisches Neuland betreten wurde.

Datenbank ist auf den Seiten der Universität Edinburgh unter <http://data.cstr.ed.ac.uk/mocha/> (letzter Zugriff: 7. 11. 2017) frei verfügbar.

Von je einer Sprecherin und einem Sprecher gibt es alle 460 gelesenen Sätze mit allen Messdaten; diese wurden auch manuell überprüft und wo notwendig korrigiert. Von 10 weiteren Personen gibt es ausreichend viele Daten, so dass sie ebenfalls in die Datenbank aufgenommen wurden. Geplant war eine weitaus größere Anzahl an Aufnahmen. Diese wurde aber nicht erreicht, weil nur wenige Personen bereit waren, die unangenehme Prozedur des Anpassens eines künstlichen Gaumens zu durchlaufen, und weil die Organisation der Aufnahmen zeitlich deutlich aufwendiger war als ursprünglich geplant.

2.7.3 mngu0

Die mngu0-Sprachdatenbank ist eine Ergänzung und Verbesserung der MOCHA EMA-Sprachdatenbank. Ausgangspunkt ist die Feststellung in Richmond, Hoole & King (2011), dass a) für empirische Analysen zusätzliche Daten benötigt werden, b) die Vergleichbarkeit von Daten verschiedener Quellen nur bedingt möglich ist, und dass c) bekanntgewordene technische Fehler korrigiert werden und neue Messverfahren zum Einsatz kommen können.

Die mngu0-Sprachdatenbank wird EMA-Sensormessdaten mit synchronen Audioaufnahmen und Videoaufnahmen der unteren Gesichtshälfte, volumetrische MRT-Scans des Lautinventars der Sprecher sowie 3D Modellierungen der Zähne im Unter- und Oberkiefer umfassen. Die EMA-Aufnahmen wurden am Institut für Phonetik und Sprachverarbeitung der LMU München durchgeführt, und sie bestehen aus 2.000 an zwei aufeinanderfolgenden Tagen gelesenen Sätzen. Die 1.354 Aufnahmen des ersten Tages, bestehend aus aufbereiteten EMA-Spuren und den Audioaufnahmen, sind als erste Version der mngu0-Sprachdatenbank im Internet unter <http://www.mngu0.org> (letzter Zugriff: 7. 11. 2017) frei verfügbar.

2.7.4 USC-TIMIT

Die USC-TIMIT ist eine noch in Aufbau befindliche Sprachdatenbank mit MRT-Bewegtbildern, EMA-Daten und Audioaufnahmen (Narayanan et al. 2011, 2014). Um die Vergleichbarkeit zu den bestehenden artikulatorischen Datenbanken zu gewährleisten wurden die 460 Sätze der TIMIT-Sprachdatenbank verwendet. Laut Narayanan et al. (2014) wurden bislang je fünf Sprecherinnen und Sprecher aufgenommen und ihre Daten für die Nutzung aufbereitet. Ge-

plant sind weitere Aufnahmen mit Sprechern von anderen Muttersprachen als amerikanischem Englisch.

Daten und Auswertungssoftware sind nach einer Online-Registrierung frei verfügbar (<http://sail.usc.edu/span/usc-timit/>).

3 Workflow bei der Erstellung einer Sprachdatenbank

Dieser Abschnitt beschreibt die Erstellung einer Sprachdatenbank anhand des konkreten Beispiels der Sprachdatenbank *Brothers*. Diese Sprachdatenbank entstand im Rahmen der Dissertation von Hanna Feiser, in der untersucht wurde, inwieweit sich die Stimmen von Brüdern voneinander unterscheiden (Feiser 2015). Es gibt viele Untersuchungen zur Stimmähnlichkeit von Zwillingen, jedoch bislang noch keine zu Brüderpaaren. Bei Zwillingsuntersuchungen ist die genetische Ausstattung und damit verbunden die Physiognomie innerhalb eines Paares weitgehend identisch – Unterschiede in der Stimme sind also auf soziale Faktoren zurückzuführen. Bei Brüderpaaren ist die genetische Ausstattung verschieden, ebenso die Physiognomie, aber dennoch werden – das zeigt die Alltagserfahrung – die Stimmen von Familienangehörigen häufig miteinander verwechselt, gerade unter akustisch ungünstigen Bedingungen wie z. B. am Mobiltelefon. Ziel der Dissertation war daher, mögliche sprachrelevante soziale Faktoren konstant zu halten und die Beziehung zwischen akustischen Merkmalen und Sprecheridentifikation zu untersuchen. Diese Frage ist gerade in der forensischen Praxis relevant, da hier eine möglichst sichere Zuordnung einer Sprachaufnahme zu einer Person wünschenswert ist und Brüderpaare deutlich häufiger sind als Zwillingspaare.

Die Erstellung einer Sprachdatenbank gliedert sich nach Schiel et al. (2003) in die Phasen *Spezifikation, Vorbereitung, Datensammlung, -aufbereitung, Annotation, Dokumentation, Validierung* und *Distribution*. Diese Phasen sind in Abbildung 8.12 dargestellt und werden in entsprechender Abfolge in diesem Abschnitt beschrieben. Die Sprachdatenbank *Brothers* eignet sich besonders gut als Fallbeispiel, weil sie alle vier in Abschnitt 1 genannten Teilgebiete der Phonetik berührt:

- *Produktion*: Es wird sowohl gelesene als auch Spontansprache von Brüderpaaren aufgezeichnet.
- *Akustik*: Die Aufnahmen erfolgen in zwei akustischen Qualitäten im Studio bzw. über Mobiltelefon.
- *Perzeption*: Die Stimmähnlichkeit wird mit einem akustischen Perzeptionsexperiment ermittelt.

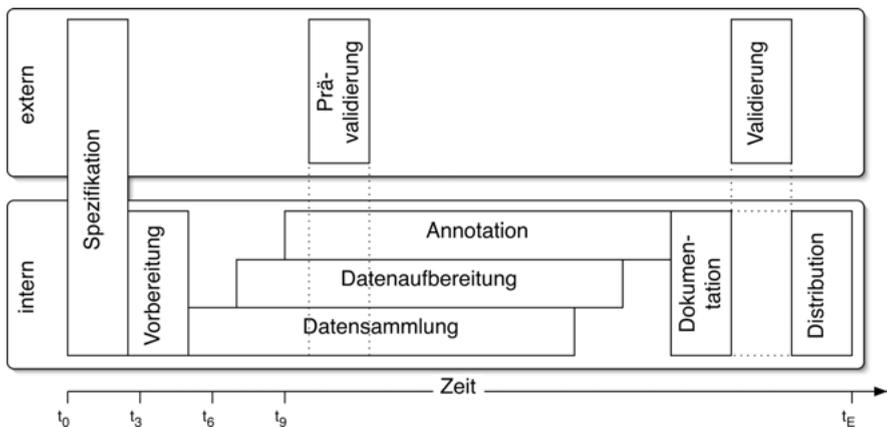


Abb. 8.12: Phasenmodell zur Erstellung von Sprachdatenbanken nach Schiel et al. (2003).

- *Transkription:* Die automatisch erstellte Segmentation bzw. die berechneten Merkmale werden manuell überprüft und ggf. angepasst.

Außerdem ist die Sprachdatenbank über ein Repository für Forschungszwecke frei verfügbar.

3.1 Spezifikation

Für die Sprachdatenbank *Brothers* soll sowohl gelesene als auch spontane Dialogsprache von mindestens zehn Brüderpaaren in zwei technischen Qualitäten aufgenommen werden. Die folgenden demographischen Faktoren sollen soweit möglich konstant gehalten werden:

- *regionale Herkunft:* Die Brüderpaare sollen aus der Region München stammen.
- *örtliche Konstanz:* Die Familien der Brüder sollen möglichst seit mindestens einer Generation im selben Wohnort leben.
- *gemeinsames Aufwachsen:* Die Brüder sollen im selben Haushalt aufgewachsen sein und immer noch in engem Kontakt stehen.
- *Alter und Altersabstand:* Die Brüder sollen mindestens 18 Jahre alt sein, der Altersabstand solle zwischen 2 und 10 Jahren betragen.

Die Sprecher werden über persönliche Kontakte sowie Teilnahmeaufrufe über soziale Medien und Mailinglisten geworben.

Die Aufnahmen sollen im Tonstudio des Instituts für Phonetik erfolgen. Zur akustischen Trennung werden die Sprecher in getrennten Räumen ohne Sichtkontakt aufgenommen. Die Aufnahmen erfolgen synchron in hoher Signalqualität (44,1 kHz Abtastrate, 16 Bit lineare Quantisierung) mit einem Großmembranmikrofon (Neumann TLM 103 P48) sowie in Telefonqualität über Mobiltelefon (Nokia 1680C-2), wobei das Telefonsignal nach der Aufnahme auf 8 kHz Abtastrate und 16 Bit lineare Quantisierung konvertiert wird. Zusätzlich werden Formant- und F0-Tracks mit der Software ASSP berechnet.

Das Sprachmaterial besteht aus:

- 80 Minimalpaaren in Trägersätzen der Form *Anna hat ... gesagt*.
- 100 zu lesenden Sätzen, den sog. *Berliner Sätzen*, die alle Phonemkombinationen mit Vokalen des Deutschen enthalten.
- einem spontanen Dialog zu Ausschnitten aus einer Folge der Fernsehkrimiserie *Tatort*. Die Brüder bekommen je unterschiedliche Ausschnitte zu sehen, damit sich ein Dialog entwickeln kann.

Die Aufnahmen erfolgen mit der Software SpeechRecorder.

Die Sprecher hören einander während des Lesens der Sätze nicht, so dass hier keine gegenseitige Beeinflussung gegeben ist.

3.2 Vorbereitung

Die Vorbereitung umfasst die Punkte Einverständniserklärung, Einrichtung der Tontechnik sowie die Vorbereitung des Aufnahmемaterials.

Die Sprecher wurden vor Beginn der Aufnahmen über den Zweck der Aufnahmen sowie die Verwendung der Sprachdaten aufgeklärt, insbesondere auch darüber, dass die Aufnahmen in Form einer Sprachdatenbank auch anderen für Forschungszwecke zugänglich gemacht werden.

Das Tonstudio des Instituts für Phonetik und Sprachverarbeitung hat eine schallgedämmte Aufnahmekabine, einen reflexionsarmen Raum sowie den Kontrollraum. Für die Aufnahmen wurden die Kabine und der Kontrollraum verwendet. Im Fenster der Aufnahmekabine ist ein Monitor angebracht, auf dem die Sprecheransicht eines SpeechRecorder-Skripts sichtbar ist. Die Steuerung der SpeechRecorder-Aufnahme erfolgt von der Aufnahmeleiterin im Kontrollraum.

Für die Aufnahmen über Mobiltelefon wurde ein ISDN-Server eingerichtet, der für eine Sitzung von beiden Telefonen angerufen wurde. Für diesen Anruf wurden die beiden ISDN-Kanäle miteinander verbunden.

Das Aufnahmемaterial wurde in ein SpeechRecorder-Aufnahmeskript importiert; dieses Skript umfasst die drei Abschnitte Einleitung, Minimalpaare

und Sätze in dieser Reihenfolge. Innerhalb der Abschnitte Minimalpaare und Sätze werden die einzelnen zu lesenden Prompts in zufälliger Reihenfolge präsentiert.

3.3 Datensammlung

Die Aufnahmen fanden im Zeitraum Oktober 2012 bis November 2013 statt.

Die Aufnahmesitzungen waren dreigeteilt: zuerst saß ein Bruder in der Aufnahmekabine und hat die Sätze gelesen, während der zweite sich im Kontrollraum verschiedene Ausschnitte aus einer *Tatort*-Folge angeschaut hat. Danach haben beide ihren Platz getauscht. Im abschließenden Dialog haben sie sich maximal zehn Minuten über die gesehenen Filmausschnitte unterhalten. Während der Aufnahmen hatten sie keinen Sichtkontakt.

Die Sprachaufnahmen der gelesenen Sprache erfolgten in der Kabine des Tonstudios. Die Aufnahmeleiterin hat den Fortgang der Aufnahmen über den Aufnahme-PC gesteuert. Der Sprecher hat die zu sprechenden Sätze von einem Monitor im Fenster der Kabine abgelesen. Versprecher usw. wurden sofort korrigiert, indem die Aufnahme wiederholt wurde.

Die SpeechRecorder-Aufnahmesoftware hat die mit dem Studiomikrofon aufgezeichneten gelesenen Sätze in je eigene Dateien geschrieben. Der Dialog wurde mit der Software Audacity aufgezeichnet. Auf dem ISDN-Server wurde die gesamte Sitzung als eine lange Audiodatei gespeichert.

3.4 Datenaufbereitung

Die Sprachaufnahmen auf dem ISDN-Server wurden manuell entsprechend der gelesenen Sätze geschnitten und verlustfrei in das WAV-Format mit 8 kHz Abtastrate mit 16 Bit Quantisierung konvertiert. Die mit Audacity erstellten Dialogaufnahmen wurden manuell in je zwei Aufnahmekanäle geteilt, um die jeweiligen Beiträge der beiden Sprecher getrennt voneinander bearbeiten zu können.

Für jede Audiodatei eines gelesenen Satzes wurde eine separate Textdatei mit gleichem Dateinamen und der Extension `.txt` angelegt, die den Wortlaut der Äußerung in normalisierter Schreibweise enthält.

Die Dialogaufnahmen wurden nachträglich geringfügig überarbeitet, um z. B. Passagen, in denen die Aufnahmeleiterin zu hören war, zu entfernen.

Für die akustische Analyse der gelesenen Sätze wurden sowohl für die Studio- als auch die Mobiltelefonaufnahmen mittels der Emu-Signalverarbeitung die ersten vier bzw. drei Formanten sowie die Grundfrequenz f_0 berech-

net. Die Formantwerte für die Telefonaufnahmen wurden anschließend in Emu manuell korrigiert.

Alle Daten wurden anschließend in das Aufnahmeverzeichnis auf dem Projektrechner im Institutsnetzwerk kopiert.

3.5 Annotation

Die Annotation erfolgte in zwei Schritten: zunächst wurden die Sätze mit dem Webdienst WebMAUS automatisch segmentiert. Ergebnis der Segmentation waren zum einen TextGrid-Dateien mit den drei Annotationsebenen ORT, KAN und MAU für die normalisierte Orthographie, die kanonische Form und die phonemische Segmentation.

Anschließend wurde die automatisch erstellte Segmentation mit der Software Praat manuell überprüft und ggf. angepasst. Die Segmentation der Aufnahmen über das Studiomikrofon erforderte nur wenige Korrekturen, die Segmentation der Telefonaufnahmen war dagegen deutlich schlechter. Der Grund dafür ist, dass die akustischen Modelle von WebMAUS nur mit Studioaufnahmen trainiert wurden. Die akustischen Eigenschaften von Mobiltelefonaten unterscheiden sich deutlich von denen von Studioaufnahmen, mit der Folge, dass die automatische Segmentation für Mobiltelefonataufnahmen sowohl für das Setzen der Labels als auch für die Segmentgrenzen schlechtere Ergebnisse liefert.

Neben der Segmentation wurden auch zwei Perceptionsexperimente durchgeführt. Im ersten Experiment wurden phonetisch geschulte Hörer in einem ABX-Diskriminationstest befragt, im zweiten phonetisch nicht geschulte Hörer

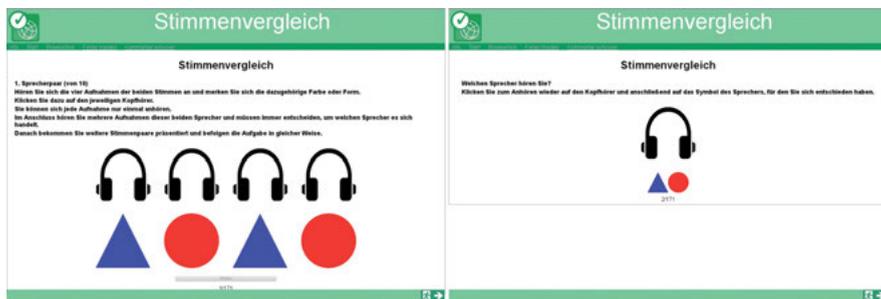


Abb. 8.13: Trainingsphase (links) und Testphase (rechts) des zweiten Perceptionsexperiments. Beim Training assoziiert die Teilnehmerin eine Stimme mit einer Farbe oder einer Form, beim Test muss sie die gehörte Stimme der entsprechenden Farbe oder Form zuordnen (aus Feiser 2015).

in einem Zuordnungstest; dieses zweite Experiment wurde parallel sowohl in der kontrollierten Umgebung des Tonstudios als auch in frei wählbarer Umgebung durchgeführt.

Die Sprachaufnahmen für das erste Perzeptionsexperiment stammen von Sprechern aus dem ripuarischen (mittelfränkischen) Sprachraum aus früheren Aufnahmen; diese sind ebenfalls Bestandteil der Sprachdatenbank. Für das zweite Perzeptionsexperiment wurden die am Institut für Phonetik und Sprachverarbeitung durchgeführten Aufnahmen mit bairischen Sprechern verwendet.

Das erste Perzeptionsexperiment wurde mit Praat durchgeführt, das zweite mit der Online-Experiment-Software percy (Draxler 2011). Abbildung 8.13 zeigt die Eingabemasken der Trainings- bzw. der Testphase des Online-Experiments.

Auch die Stimulusauswahl sowie die Eingaben der Teilnehmerinnen und Teilnehmer sind Bestandteil der Sprachdatenbank.

3.6 Dokumentation

Der gesamte Prozess der Erstellung der *Brothers*-Sprachdatenbank wurde detailliert dokumentiert, u. a. mit Fotos der Aufnahmekabinen und genauer Angaben der bei den Aufnahmen verwendeten Geräte. Diese Dokumentation bildet ein eigenes Kapitel der Dissertation.

3.7 Validierung

Nach Abschluss der Dissertation wurde die *Brothers* Sprachdatenbank vor der Veröffentlichung im BAS Repository im Mai 2015 validiert. Der Validierungsbericht ist als Teil der Sprachdatenbank dort verfügbar.

Bei dieser Validierung wurde festgestellt, dass die Qualität der automatischen Segmentierung für die Studioaufnahmen deutlich besser ist als für die Telefonaufnahmen. Gemessen wurde der prozentuale Anteil an falsch alignierten Wörtern im Text. Dieser Anteil lag bei 9 % für die Studioaufnahmen und 24 % für die Telefonaufnahmen, d. h. dass fast jedes vierte Wort im automatischen Verfahren nicht korrekt segmentiert wurde.

3.8 Verfügbarkeit der Datenbank

Die Datenbank *Brothers* ist über das Repository des BAS über den *persistent identifier* (PID) <http://hdl.handle.net/11022/1009-0000-0001-55C3-3> für akademische Nutzer frei verfügbar, unter anderem im Emu-Datenbankformat.

4 Zusammenfassung

Brothers ist ein Beispiel für eine phonetische Sprachdatenbank, die zunächst konkret für die Untersuchung der Stimmähnlichkeit von Brüdern im Rahmen einer Dissertation entwickelt und ausgewertet wurde, und die danach so aufbereitet wurde, dass sie in das CLARIN Repository des Bayerischen Archivs für Sprachsignale aufgenommen werden konnte.

Im Repository stehen nun nicht nur die in der Dissertation ausgewerteten Roh- und abgeleiteten Messdaten, sondern auch noch weiteres Sprachmaterial wie z. B. die spontansprachlichen Dialoge, die noch nicht transkribiert oder ausgewertet wurden, sowie ein Validierungsbericht zur Sprachdatenbank mit Angaben zur Qualität der automatischen Segmentation.

Die wesentlichen Ergebnisse der Dissertation sind, dass

1. Brüderpaare in Perzeptionsexperimenten anhand der Stimme signifikant häufiger als solche erkannt wurden als Nichtbrüder,
2. die Stimmen von Brüderpaaren *auditiv* häufiger miteinander verwechselt werden als die Stimmen von nicht-verwandten Sprechern,
3. diese Verwechslung beim Telefon höher ist als bei Studioaufnahmen und dass
4. die *akustischen* Messungen für einzelne Sprecher jeweils charakteristisch sind, aber mit Ausnahme der Lesebedingung nicht zur Trennung von Brüderpaaren und nicht-verwandten Sprechern geeignet sind.

Feiser (2015) zieht daraus den Schluss, dass die perzeptive Ähnlichkeit von Brüderstimmen kaum von den akustischen Merkmalen abhängt, sondern „eher das erworbene Sprecherverhalten der Geschwister“ widerspiegelt (Feiser 2015: 140). Das Beispiel der Sprachdatenbank *Brothers* zeigt zum einen, dass Sprachdatenbanken eine unverzichtbare Grundlage der systematischen Untersuchung phonetischer Fragestellungen sind, zum anderen, dass damit Sprachressourcen, bestehend aus Primär-, Sekundär- und Metadaten in nachhaltiger Form vorgehalten und für weitere Untersuchungen genutzt werden können.

Literatur

- Becker, Thomas (2012): *Einführung in die Phonetik und Phonologie des Deutschen*. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Bird, Steven & Mark Liberman (2001): A formal framework for linguistic annotation. *Speech Communication* 33 (1,2), 23–60.
- Boersma, Paul & David Weenink (1996): Praat, a system for doing phonetics by computer. Tech. Rep. 132 Institute of Phonetic Sciences of the University of Amsterdam.

- Cassidy, Steve & Jonathan Harrington (1996): Emu: An enhanced hierarchical speech data management system. In *Proc. SST*, 361–366. Adelaide.
- Cassidy, Steve & Jonathan Harrington (2001): Multi-level annotation in the emu speech database management system. *Speech Communication* 33, 61–77.
- Draxler, Christoph (2008): *Korpusbasierte Sprachverarbeitung – eine Einführung*. Tübingen: Gunter Narr.
- Draxler, Christoph (2011): Percy – An HTML5 framework for media rich web experiments on mobile devices. In *Proc. Interspeech*, 3339–3340. Florence, Italy.
- Draxler, Christoph, Toomas Allosaar, Sadaaki Furui, Mark Liberman & Peter Wittenburg (2011): Speech processing tools – an introduction to interoperability. In *Proc. Interspeech*, 3229–3232. Florence, Italy.
- Draxler, Christoph & Klaus Jänsch (2004): SpeechRecorder – a universal platform independent multi-channel audio recording software. In *Proc. LREC*, 559–562. Lisbon, Portugal.
- Draxler, Christoph & Stefan Kleiner (2015): A cross-database comparison of two large german speech databases. In *Proceedings ICPhS*, Glasgow.
- Feiser, Hanna (2015): *Untersuchung auditiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Brüderpaaren unter forensischen Aspekten*. Frankfurt am Main: Verlag für Polizeiwissenschaft.
- Garofolo, John, Lori Lamel, William Fisher, Jonathan Fiscus, David S. Pallett & Nancy Dahlgren (1986): *The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM*. NIST.
- Gibbon, Dafydd, Roger Moore & Richard Winski (1997): *Handbook of standards and resources for spoken language systems*. Berlin: Mouton de Gruyter.
- Gibbon, Fiona & Lisa Crampin (2001): An electropalatographic investigation of middorsum palatal stops in an adult with repaired cleft palate. *Cleft Palate Craniofacial Journal* 38, 96–105.
- Gippert, Jost, Nikolaus P. Himmelmann & Ulrike Mosel (Hrsg.) (2006): *Essentials of language documentation*. Mouton de Gruyter.
- Harrington, Jonathan (2010): *Phonetic analysis of speech corpora*. Oxford: Wiley-Blackwell.
- Harrington, Jonathan, Steve Cassidy, Janet Fletcher & Andrew McVeigh (1993): The MU+ system for corpus based speech research. *Computer Speech and Language* 7, 305–331.
- Holmqvist, Kenneth, Marcus Nyström & Fiona Mulvey (2012): Eye tracker data quality: what it is and how to measure it. *Proceedings acm Symposium on Eye Tracking Research and Applications* 45–52.
- Huettig, Falk, Joost Rommers & Antje Meyer (2011): Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica* 137, 151–171.
- IPA (1989): IPA Kiel Convention Workgroup 9 Report. *Journal of the IPA* 19 (2), 81–82.
- IPA (1999): *Handbook of the IPA*. Cambridge: Cambridge University Press.
- Lemnitzer, Lothar & Heike Zinsmeister (2006): *Korpuslinguistik – eine Einführung*. Tübingen: Narr Francke Attempto.
- Narayanan, Shrikanth, Erik Bresch, Prasanta Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon Kim, Adam Lammert, Michael Proctor, Vikram Ramanarayanan & Yinghua Zhu (2011): A multimodal real-time MRI articulatory corpus for speech research. In *Proceedings Interspeech*, Florence.
- Narayanan, Shrikanth, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis & Michael Proctor (2014): Real-

- time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *Journal of the Acoustical Society of America* 136(3), 1307–1311.
- Pömp, Julian & Christoph Draxler (2017): OCTRA – A configurable browser-based editor for orthographic transcription. In *Proceedings Phonetik und Phonologie*, Berlin.
- Pompino-Marschall, Bernd (1995): *Einführung in die Phonetik*. Berlin: de Gruyter Mouton.
- Reetz, Henning & Allard Longman (2009): *Phonetics – transcription, production, acoustics and perception*. Blackwell.
- Richardson, Matt, Jeff Bilmes & Chris Diorio (2003): Hidden-articulator markov models for speech recognition. *Speech Communication* 41 (2–3), 511–529.
- Richmond, Korin, Phil Hoole & Simon King (2011): Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech*, 1505–1508.
- Schiel, Florian, Christoph Draxler, Angela Baumann, Tania Ellbogen & Alexander Steffen (2003): *The production of speech corpora*. Institut für Phonetik und Sprachliche Kommunikation, Universität München.
- Schmidt, Thomas, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Dan Loehr, Magnus Magnusson, Travis Rose & Han Sloetjes (2009): An exchange format for multimodal annotations. In *Multimodal Corpora* (Lecture Notes in Computer Science 5509), 207–221. Springer.
- Schmidt, Thomas & Kai Wörner (2005): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. *Gesprächsforschung* 6, 171–195.
- Sloetjes, Han, Albert Russel & Alex Klassmann (2007): ELAN: a free and open-source multimedia annotation tool. In *Proc. Interspeech*, 4015–4016. Antwerp.
- Stone, Maureen (2005): A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics* 19 (6–7), 455–501.
- Westbury, John R., Greg Turner & Jim Dembrovski (1994): X-ray microbeam speech production database user's handbook. Tech. rep. Waisman Center, Washington University.
- Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch (2017): Emu-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech and Language*.
- Wrench, Alan A. & William J. Hardcastle (2000): A multichannel articulatory speech database and its application for automatic speech recognition. In *Proc. 5th Seminar on Speech Production*, 305–308.

