Originalveröffentlichung in: Zeitschrift für germanistische Linguistik Jg. 50 (2022) Nr. 1, S. 36-81.
DOI: *https://doi.org/10.1515/zgl-2022-2048*

POSTPRINT

Hagen Hirschmann
Thomas Schmidt*

Gesprochene Lernerkorpora: Methodisch-technische Aspekte der Erhebung, Erschließung und Nutzung

Spoken learner corpora: Methodological and technical aspects of collection, indexing and use

Abstract: This article provides an overview of methodological and technical issues that arise in the collection, indexing and use of spoken learner corpora, i. e. corpora containing spoken utterances of learners of a target language. After an introductory discussion of the most important special features of this type of corpus that distinguish it from written language learner corpora and spoken corpora with L1 speakers, we will go into more detail on questions of corpus design. The main part of the paper is then an overview of the methodological and technical procedures of the individual steps of collecting, indexing, providing and using spoken learner corpora. The main aim of this overview is to highlight practices that can be considered best practices according to the current state of research. Finally, we outline the challenges that still exist for this type of corpus.

DOI: https://doi.org/10.1515/zgl-2022-2048

***Kontaktpersonen: Dr. Hagen Hirschmann**: Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik, Sprach- und literaturwissenschaftliche Fakultät, Unter den Linden 6, D-10099 Berlin, E-Mail: hagen.hirschmann@hu-berlin.de
Dr. Thomas Schmidt: Universität Basel, Research and Infrastructure Support (RISE), Spalenberg 65, CH-4051 Basel, E-Mail: th.schmidt@unibas.ch

1 Einleitung

1.1 Grundlegende Aspekte und Ziele

Dieser Beitrag beschäftigt sich mit methodischen und technischen Fragen, die sich bei der Erstellung und Verwendung gesprochener Lernerkorpora stellen.

Nach Granger (2008: 259) verstehen wir unter einem Lernerkorpus eine digitale Sammlung sprachlicher Produktionen von LernerInnen in einer Zweit- oder Fremdsprache. Spracherwerbskorpora, in denen es um den Erwerb einer L1 geht, schließen wir nicht in unsere Definition von „Lernerkorpus“ ein, auch wenn Korpora dieses Typs in den hier interessierenden Eigenschaften einige Gemeinsamkeiten mit L2-Korpora aufweisen und teilweise auch als Lernerkorpora bezeichnet werden (vgl. auch Wisniewski sowie Weiss et al., dieses Heft sowie Schmidt/Wörner 2012: XII). Gleiches gilt für L1-Korpora, also Korpora mit Äußerungen vollkompetenter Erstsprachler. Diese können gleichwohl als Referenz- oder Vergleichskorpora eine wichtige Rolle in der Arbeit mit Lernerkorpora spielen, z. B. wenn sie als Ressourcen der Sprachvermittlung zur DaF- oder DaZ-Lehre zum Einsatz kommen (vgl. z. B. Imo & Weidner 2018 oder Weiss et al., dieses Heft).

Gesprochene Lernerkorpora bestehen also aus mündlich produzierten Äußerungen von L2-Lernern, die als Audio- oder Videoaufnahmen erhoben und über eine Transkription und darauf aufbauenden Annotationen sowie zugehörige Metadaten für korpuslinguistische Analysen erschlossen werden.

Gesprochene Lernerkorpora weisen gegenüber schriftlichen Lernerkorpora zunächst all diejenigen Besonderheiten auf, die für gesprochene Korpora im Allgemeinen gelten. Erstens sind nicht die Primärdaten der unmittelbare Gegenstand der Analyse, sondern diese werden erst durch die wissenschaftliche Methode

der Transkription, und mit den entsprechenden theorie- und fragestellungsspezifischen modellhaften Verkürzungen (vgl. Ochs 1979, Schmidt 2005), analytisch zugänglich gemacht. Zweitens weisen die resultierenden Transkriptions-„Texte" gegenüber schriftsprachlichen Texten einige strukturelle Besonderheiten auf: Sie beinhalten zeitliche Verknüpfungen („Alignment") in die zugrunde liegenden Aufnahmen, Zuordnungen zu den verschiedenen an der Interaktion beteiligten SprecherInnen, zeitlich parallele Strukturen (Simultansprechen, Gleichzeitigkeit von verbalem und nonverbalem Verhalten), unvollständige Tokens (Fehlstarts, Abbrüche), nicht-lexikalische Tokens (Häsitationsmarker, Interjektionen) und andere Nicht-Wort-Tokens (Beschreibungen von Pausen, non-verbalem Verhalten). Diese strukturellen Besonderheiten sind der Grund dafür, dass gesprochene (Lerner-)Korpora nicht einfach mit Datenmodellen und Werkzeugen für schriftsprachliche Daten bearbeitet und analysiert werden können, sondern auf mündliche Daten angepasste Verfahren erfordern.

Weiterhin teilen gesprochene Lernerkorpora mit schriftlichen Lernerkorpora die besondere Eigenschaft, dass sie in vielerlei Hinsicht von der oder den entsprechenden Standard-Varietät(en) abweichen. Zu den Abweichungen gehören phonetisch, morphologisch oder syntaktisch unvollständige oder anderweitig abweichende Strukturen, aber auch Mehrsprachigkeits-Phänomene wie Code-Switches oder Struktur-Transfers von der L1 in die Zielsprache. Wegen dieser besonderen Phänomene sind Transkriptions- und Annotationsverfahren, die für mündliche Daten von L1-SprecherInnen entwickelt wurden, nicht unverändert auf Lernerkorpora übertragbar, sondern bedürfen geeigneter Anpassungen und Erweiterungen.

Vor diesem Hintergrund fokussiert der vorliegende Beitrag also Besonderheiten in der Erstellung, Erschließung und Nutzung gesprochener Lernerkorpora im Vergleich zu gesprochenen L1-Korpora und schriftsprachlichen Lernerkorpora. Dabei werden zunächst grundlegende Designkriterien dargestellt, die größtenteils nicht spezifisch für Lernerkorpora sind. Anschließend wird erläutert, welche Grundsatzentscheidungen hinsichtlich dieser Kriterien in vielen aktuellen Lernerkorpora gefällt und umgesetzt wurden. Hierbei wird auf die wesentlichen Schritte bei der Erhebung und Erschließung mündlicher Lernerkorpusdaten eingegangen: die Erstellung von Aufnahmen, die Erhebung, Speicherung und das Einpflegen von Metadaten, die Erstellung von Transkriptionen, Normalisierungsebenen, (weiterer) Annotationen sowie wichtige Aspekte der Bereitstellung von Korpusdaten. Auch rechtliche Aspekte finden Erwähnung. Anhand von derzeit verfügbaren frei zugänglichen gesprochenen Lernerkorpora des Deutschen werden der aktuelle Stand, aktuelle Defizite und Perspektiven herausgearbeitet.

1.2 Grundlegendes zum Korpusdesign

Unter Korpusdesign verstehen wir die Gesamtheit an Entscheidungen zur Datenzusammensetzung bzw. -verteilung im Korpus. Im Falle gesprochener Lernerkorpora sind dies Fragen zur Auswahl von SprecherInnen und Sprechanlässen, die ins Korpus aufgenommen werden, zu Parametern, mittels derer diese beschrieben und klassifiziert werden, und zu deren mengenmäßiger Repräsentation. Sofern es sich nicht um ein opportunistisches Korpus handelt, die Entscheidungen also bewusst nicht vor dem Korpusaufbau, sondern in dessen Verlauf (eben „opportun" nach den sich bietenden Gelegenheiten) getroffen werden, stellt das Korpusdesign den ersten Schritt beim Aufbau eines Korpus dar. Es wird üblicherweise geleitet von den wissenschaftlichen Fragestellungen, die mit Hilfe des Korpus bearbeitet werden sollen, und richtet sich auch an den zugehörigen theoretischen Hintergründen aus.

Die allermeisten existierenden Lernerkorpora haben ein sprecherzentriertes Design. D. h., bedingt durch die Forschungsziele des zugrunde liegenden Forschungsprojekts, sind es Eigenschaften der LernerInnen, die über die Auswahl der zu erhebenden Korpusdaten entscheiden. Es werden üblicherweise Daten von Gruppen von Lernenden erhoben, deren Zusammensetzung sich beispielsweise nach Erstsprache(n) (L1), Kompetenzniveaus oder nach Erwerbsverläufen ergibt. Das resultierende Korpus kann danach charakterisiert werden, welche dieser Gruppen in welchem Umfang repräsentiert sind. So wurden bspw. im MULTILIT-Korpus systematisch multilinguale in Deutschland lebende Kinder und junge Erwachsene mit durch die Eltern bedingter L1 (Heritage-Sprache) Türkisch und durch die umgebende Gesellschaft (Majoritätssprache) bedingte L2 (bzw. zweite L1) Deutsch aufgenommen. Zusätzlich wurde das sprachliche Verhalten dieser Probandinnen und Probanden in der Fremdsprache Englisch erhoben. In den Korpora HaMaTaC, HaMoTiC und BeMaTaC wurden systematisch Sprecherinnen und Sprecher des Deutschen in dialogischen Kontexten aufgenommen, bei denen L2- und L1-SprecherInnen aufeinandertreffen (HaMoTiC, BeMaTaC) bzw. bei denen L2-SprecherInnen mit unterschiedlichem Erwerbsstand kommunizieren (HaMaTaC). Im Leap-Korpus finden sich u. a. L2-SprecherInnen, die unmittelbar vor als auch unmittelbar nach einem mehrmonatigen Auslandsaufenthalt aufgenommen wurden. Diese Beispiele zeigen, dass die Auswahl der Probandinnen und Probanden in Lernerkorpora in aller Regel hochgradig gesteuert erfolgt und durch bestimmte Forschungsziele bedingt ist.

Bei den Sprechanlässen kann zwischen (eher) elizitierten und (eher) natürlichen Situationen (bzw. Daten) unterschieden werden (vgl. dazu auch den Beitrag von Karges et al. in diesem Heft; zur knappen Diskussion von Authentizität bei Korpusdaten vgl. auch Mukherjee 2009: 21). „Natürlichkeit" wird häufig auch mit

dem Begriff „Authentizität“ gefasst. Bei Elizitationen werden die sprachlichen Äußerungen in einem vorgegebenen, von den Forschenden arrangierten Setting erhoben. Typische Elizitationen zum Erheben von Spontandaten sind etwa Leitfadeninterviews oder aufgabenorientierte Kommunikation, bei denen ProbandInnen einzeln oder in Gruppen eine Aufgabe gestellt wird, die sprachlich gelöst werden muss. Beispiele hierfür sind die sog. Map Tasks, die in den Korpora HaMaTaC und BeMaTac verwendet wurden, oder die Nacherzählung einer Sequenz aus einem Charlie Chaplin-Film, die ursprünglich zur Erhebung der Daten des ESF-Korpus gehörte und auch in HaMoTiC Anwendung fand. Die IFCASL-Daten, die von den ProbandInnen ausschließlich vorgelesen wurden (vgl. auch Trouvain in diesem Heft), sind ein Beispiel für Lernerdaten, die besonders stark elizitiert und als nicht natürlich und nicht spontansprachlich einzustufen sind.

Demgegenüber sind authentische bzw. natürliche Daten solche, bei denen die Sprechanlässe alltägliche Kommunikationssituationen sind, also z. B. Tischgespräche, Service-Interaktionen usw., wobei zu beachten ist, dass das Merkmalspaar „elizitiert“-„authentisch“ nicht kategorial, sondern als fließender Übergang aufzufassen ist. Viele Erhebungsbedingungen (wie z. B. die Anwesenheit von WissenschaftlerInnen in der Erhebungssituation) können dazu beitragen, Erhebungen weniger authentisch zu machen; andere Bedingungen (z. B. die Verlagerung der Erhebungssituation in den häuslichen Bereich) bewirken das Gegenteil. Ein Sonderfall von authentischen Daten sind solche Primärdaten, die auf natürliche Weise produziert und ohne ein Eingreifen der ForscherInnen erstgespeichert wurden (z. B. Handy-Sprachnachrichten). Dies gilt für viele schriftliche Korpora und insbesondere Korpora, die auf Internet- und Telekommunikationsdaten zurückgehen. Die wissenschaftliche Arbeit der KorpuserstellerInnen besteht in diesen Fällen aus weiterverarbeitenden Schritten, die wir ab Abschnitt 3 behandeln. Leider existiert für die meisten Forschungsziele, die sich auf mündlich kommunizierte Lernersprache beziehen, dieser Datentyp schlichtweg nicht, weshalb auf die hier beschrieben Methoden zurückgegriffen werden muss.

Ein Korpus, das überwiegend den Gebrauch des Deutschen als Erstsprache bei verschiedenen Kommunikationsanlässen abbilden will und in dem somit solche Alltagssituationen erfasst werden, ist das FOLK-Korpus (Schmidt 2016). FOLK beinhaltet auch Redebeiträge von Lernenden des Deutschen als Fremd- oder Zweitsprache, jedoch nicht gezielt erhoben und nicht auf das Ziel spracherwerbstheoretischer Forschung hin aufbereitet, weshalb FOLK hier zwar als wichtiger Maßstab für die Erhebung und Aufbereitung mündlicher Sprachdaten, nicht aber als Vertreter unter der Gruppe der Lernerkorpora behandelt wird.

Die allermeisten Lernerkorpora arbeiten mit (stärker) elizitierten Daten – nicht nur, weil diese einfacher zu erheben sind (es ist kein Feldzugang nötig, die Aufnahmesituation ist besser plan- und kontrollierbar), sondern auch, weil

sie direkter vergleichbar sind. Eine Vergleichbarkeit kann ggf. korpusübergreifend zutreffen, wenn bspw. gleiche Aufgaben in verschiedenen Korpora verwendet werden – z. B. existieren zu den Map-Task-Daten aus HaMaTaC und BeMaTac (erstsprachliche) Vergleichsdaten im Korpus „Deutsch Heute“ (Kleiner et al. 2011). Natürliche Lernerdaten finden sich z. B. in den Prüfungsgesprächen und studentischen Vorträgen des GeWiss-Korpus, wobei auch hier durch die Beschränkung auf zwei spezifische Gesprächstypen eine gute Vergleichbarkeit der Daten gegeben ist.

Zum Korpusdesign gehört auch die Frage nach der Menge der zu erhebenden Daten. Nicht wenige korpuslinguistische Fragestellungen erfordern eine gewisse Menge an Daten für die analyserelevanten Kategorien, vor allem um belastbare quantitative Analysen zu ermöglichen. Während schriftsprachliche Referenzkorpora mittlerweile in Größenordnungen von mehreren Milliarden Tokens operieren, sind dem Umfang von gesprochenen Korpora – und Lernerkorpora insbesondere – wegen der aufwändigen Erhebungs- und Erschließungsmethoden generell deutlich engere Grenzen gesetzt (vgl. dazu Kupietz/Schmidt 2015). Die größten gesprochenen Korpora (z. B. das Korpus „Deutsch Heute“) bestehen aus ungefähr 1000 Stunden Aufnahmen, was etwa 10 Millionen transkribierter Wortformen entspricht. Die meisten gesprochenen Korpora sind jedoch deutlich kleiner, und öffentlich verfügbare gesprochene Lernerkorpora können wiederum deutlich kleiner ausfallen: HaMaTaC, BeMaTac, HaMoTiC und Leap bestehen aus weniger als 10 Stunden Aufnahmen und 100.000 transkribierter und weiterverarbeiteter (vgl. die Abschnitte 3.2.2–3.2.5 dieses Beitrags) Wortformen; die L2-Daten im GeWiss-Korpus umfassen knapp 50 Stunden Aufnahmen. Dass mit solchen Korpusgrößen gewisse korpuslinguistische, vor allem statistische Verfahren auf einer instabilen methodischen Basis stehen, thematisieren z. B. Goschler & Stefanowitsch (2014: 344) deutlich:

> Das sind Größenordnungen [gemeint sind gesprochene Korpora mit bis zu 10 Millionen Wörtern, Anm. der Verfasser], an die Zweitspracherwerbskorpora selten bis nie herankommen, mit der offensichtlichen Folge, dass sie häufig nur eine unzureichende Datengrundlage darstellen. Selbst für relativ häufige Phänomene liefern kleine Korpora meist keine ausreichend große Datenmenge für die Art von systematischer, statistisch auswertbarer Analyse, die ja überhaupt erst die Motivation für die Verwendung von Korpora darstellt. Interessante Beobachtungen in Zweitspracherwerbskorpora behalten deshalb oft zwangsläufig einen anekdotischen Charakter.

Ob die „systematische, statistisch auswertbare Analyse“ wirklich die einzige „Motivation für die Verwendung von Korpora darstellt“, kann allerdings bestritten werden: Bei einem gut durchdachten Korpusdesign und einer methodisch

fundierten Annotation der Daten können auch bzw. gerade kleinere Korpora erkenntnisfördernd sein. Vgl. hierzu Lüdeling et al. (2021) – hier wird argumentiert und zusammenfassend an drei Fallbeispielen erläutert, dass mittelgroße, händisch gerade noch kontrollierbare (also manuell annotierbare) Lernerkorpora für viele interessante Fragestellungen eine notwendige empirische Datengrundlage darstellen, weil diese Fragestellungen tiefe, komplexe Analysen erfordern, die mit ausschließlich automatisierten Verfahren nicht geleistet werden können. Zudem darf nicht vernachlässigt werden, dass die Aufbereitung von Korpusdaten selber linguistische Forschung bedeutet und sehr häufig die KorpuserstellerInnen diejenigen sind, die Einblicke in grundlegende Charakteristika der durch das Korpus abgebildeten Varietät(en) erfolgreich publizieren. Zuletzt sei erwähnt, dass die statistische Belastbarkeit einer quantitativen Auswertung nicht einfach von der Größe der Daten in Wortfrequenzen, sondern entschieden von der Häufigkeit des untersuchten Phänomens innerhalb der Daten sowie dem statistischen Verfahren selbst abhängt (vgl. zum Problem der Korpusgröße auch Wisniewski, dieses Heft).

Anhand einer Auswahl von bereits existierenden gesprochenen Lernerkorpora wollen wir genau herausarbeiten, wo die aktuellen Anforderungen, Grenzen und Probleme bei der Erstellung dieser Daten liegen, die sicherlich immer den Anspruch haben, möglichst umfangreich zu sein. Hieraus können ggf. methodische Innovationen zur Überwindung der dargestellten Schwierigkeiten bei statistischen Auswertungen erarbeitet werden.

1.3 Kriterien für die Korpusauswahl

Aus den bisherigen Betrachtungen und wenigen noch zu erläuternden Aspekten ergeben sich gewisse Kriterien für eine genauere Betrachtung von Korpusdaten im vorliegenden Kontext. Diese können wie folgt zusammengefasst werden:

- Mündliche Lernersprache Deutsch: Das Korpus sollte mit dem Ziel erstellt worden sein, Lernersprache abzubilden. Wir konzentrieren uns hier auf Korpora, die mündliche Äußerungen von Sprecherinnen und Sprechern des Deutschen als Fremd- oder Zweitsprache erhoben haben. Hierbei berücksichtigen wir nicht den kindlichen Erstspracherwerb, für dessen korpusbasierte Erforschung es etliche nennenswerte Quellen gibt, allen voran wahrscheinlich die CHILDES-Sammlung und die dazugehörige digitale Infrastruktur Talkbank (MacWhinney 2000).
- Korpus-Charakter: Wie bereits ausgeführt, ist die Frage, welche Datenerhebungen als Korpus zu fassen sind, strittig bzw. definitionsabhängig. Wir beschränken uns auf digital aufbereitete bzw. computerlesbare Daten, die

durch geeignete Datenformate strukturell geordnet und dadurch systematisch (mit Suchmaschinen und/oder Statistikprogrammen) auswertbar sind. Durch diese Anforderung blenden wir sehr viele an sich wertvolle Daten aus, die bspw. als Bild- oder unstrukturierte Textdatei gespeichert wurden. Wir sind uns auch darüber bewusst, dass es an vielen Orten Bestrebungen gibt, solche Daten in geeignete Zielformate zu bringen.

- Zur Anforderung eines Korpus, der Erforschung einer bestimmten Varietät dienlich zu sein, gehört unserer Ansicht auch, dass es Daten einer gewissen Sprecherpopulation beinhalten muss und nicht auf einem oder wenigen sprachlichen Einzelfällen beruhen sollte. Aus diesem Grund ziehen wir hier keine Datensammlungen in Betracht, die lediglich eine Sprecherin oder einen Sprecher bzw. Sprachdaten sehr weniger Personen beinhaltet (auch wenn es für solche Daten natürlich ebenso passende Fragestellungen geben mag).
- Verfügbarkeit: Wir wollen hier nur über Korpora sprechen, die der wissenschaftlichen Allgemeinheit zur Verfügung stehen. Bestimmte existierende Korpora können zwar den oben formulierten Ansprüchen entsprechen und eine an sich wertvolle Quelle sein, doch diese Quelle ist, wenn sie nicht verfügbar gemacht werden kann, mit Blick auf die Entwicklung des inhaltlichen und methodischen Forschungsdiskurses, mit Blick auf die Transparenz von Forschung und die Gestaltung einer Forschungsgemeinschaft eine Sackgasse. Wir wissen, dass viele Korpusressourcen nicht aus Egoismus oder anderen unlauteren Gründen privat bleiben, und kennen die ethischen, rechtlichen, technischen und infrastrukturellen Hürden, die einer Veröffentlichung von Forschungsdaten im Wege stehen können. Diese Probleme sind in unseren Augen jedoch solche, die bei der Planung eines Korpusprojektes zuallererst geklärt werden müssen. Ohne die Gewährleistung, dass Forschungsdaten verfügbar gemacht werden können, sollte kein entsprechendes Projekt gestartet werden.

Die nachfolgenden Korpora entsprechen im Wesentlichen diesen Anforderungen und werden in diesem Beitrag genauer betrachtet, um daraus eine Art Status quo abzuleiten. Die Auflistung der behandelten Ressourcen erfolgt in alphabetischer Reihenfolge (die meisten dieser Korpora befinden sich auch im Appendix des Beitrags von Katrin Wisniewski in diesem Heft).

- BeMaTaC (L2-Kohorte) (Berlin Map Task Corpus; Sauer & Lüdeling 2016; https://hu.berlin/bematac): ein Maptask-basiertes Dialogkorpus mit fortgeschrittenen Lernenden des DaF (heterogene L1) und deutschen ErstsprachlerInnen. Analysefokus: Disfluency-Phänomene, morphosyntaktische Phänomene.

- GeWiss (L2-Kohorte) (Gesprochene Wissenschaftssprache; Meißner & Slavcheva 2014 oder auch Fandrych et al. 2017; https://gewiss.uni-leipzig.de/):[1] monologische und dialogische Wissenschaftskommunikation (Sprachen: Deutsch, Englisch, Italienisch und Polnisch).
- HaMaTaC (Hamburg Map Task Corpus; Hedeland/Schmidt 2012); https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:hamatac-1.0.0):[2] ein Maptask-basiertes Dialogkorpus mit Lernenden des DaF (heterogener Erwerbsstand, heterogene L1).
- HaMoTiC (Hamburg Modern Times Corpus; https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:hamotic):[3] mündlich zusammengefasste Stummfilmsequenzen durch Lernende des DaF (heterogener Erwerbsstand, heterogene L1).
- IFCASL (Individualized Feedback in Computer-Assisted Spoken Language Learning; Trouvain et al. 2016; http://www.ifcasl.org/): Vorgelesene Sätze von Lernenden auf Anfänger- und Fortgeschrittenenniveau des Deutschen und Französischen als Fremdsprache sowie muttersprachlichen Vergleichsdaten („bidirektionales" Korpus) zur Untersuchung phonetisch-phonologischer spracherwerbsbedingter Abweichungen (siehe vor allem auch Trouvain, dieses Heft).
- Leap (Learning Prosody in a Foreign Language; Gut 2014; https://sourceforge.net/projects/leapcorpus/): Vorgelesene Wörter und Textdaten, mündliche Nacherzählungen und mündliche Interviews von Lernenden des DaF und EaF (bzw. ESL) mit verschiedenen Erwerbsständen sowie verschiedenen Erhebungszeitpunkten (vor sowie nach längeren Aufenthalten in Regionen der Zielsprache); Analysefokus: phonetisch-phonologische Annotationen, um prosodische Entwicklungen nachvollziehen zu können.
- MULTILIT (Schellhardt & Schroeder 2015, www.uni-potsdam.de/de/daf/projekte/multilit): Gesprochene Daten und schriftliche Texte multilingualer

1 Die deutschsprachigen Bestandteile von GeWiss sind außer über die Leipziger Plattform auch über die Datenbank für Gesprochenes Deutsch (DGD, dort mit zusätzlichen Annotationen) zugänglich. Die nicht-deutschsprachigen Bestandteile des Korpus werden aktuell für eine Veröffentlichung in der DGD (voraussichtlich in 2022) aufbereitet

2 HaMaTaC ist außerdem am Zentrum für Nachhaltiges Forschungsdatenmanagement der Universität Hamburg archiviert und zugänglich: https://doi.org/10.25592/uhhfdm.1481. Außerdem wir es aktuell für eine Integration in die Datenbank für Gesprochenes Deutsch aufbereitet und wird dort voraussichtlich ab Mitte 2021 zur Verfügung stehen.

3 HaMoTiC ist außerdem am Zentrum für Nachhaltiges Forschungsdatenmanagement der Universität Hamburg archiviert und zugänglich: https://doi.org/10.25592/uhhfdm.1483. Außerdem wir es aktuell für eine Integration in die Datenbank für Gesprochenes Deutsch aufbereitet und wird dort voraussichtlich ab Mitte 2021 zur Verfügung stehen.

Kinder und Erwachsener sowie monolingualer Vergleichsgruppen. Erhoben wurden narrative Beiträge (Weitererzählungen) sowie argumentative Texte.

Gerade mit Blick auf die noch folgenden Beiträge dieses Hefts müssen zwei weitere Korpora Erwähnung finden: SWIKO (Karges et al., dieses Heft) und WroDiaCo (Belz & Odebrecht, dieses Heft). Beide Korpora erfüllen die oben formulierten Anforderungen an computerauswertbare, verfügbare gesprochene Lernerkorpora. Sie waren bei der Erstellung dieses Beitrags noch im Entstehen begriffen und nur im Rahmen von Dokumentationen, nicht aber als Korpus selbst verfügbar, so dass viele der Designspezifika schwer zu ermitteln und die Fragen zur Bereitstellung (Abschnitt 3 und 4 dieses Beitrags) ggf. noch nicht beantwortbar gewesen wären.

Hier sei noch einmal ausdrücklich erwähnt: Wir räumen ein, dass wir uns mit der oben stehenden Auswahl (Liste) auf eine bestimmte Art von Forschungsdatentyp beschränken. Wir sind uns bewusst, dass etliche weitere Spracherwerbsdaten existieren, die sich zur Bearbeitung von Forschungsfragen im Fremd- und Zweitspracherwerbskontext eignen bzw. die den Gesamtdiskurs entscheidend geprägt haben. Exemplarisch verweisen wir auf die breite Datengrundlage, die in Clahsen et al. (1983) ausgewertet wurde, die aber unserem Erkenntnisstand nach nicht als digitale, systematisch durchsuchbare Datenquelle öffentlich verfügbar ist.

Da sich aber der hier näher dargestellte Datentyp für bestimmte qualitativ-quantitative Forschungszwecke (siehe Abschnitt 5) besonders eignet und die vorgestellten Korpora allgemein zugänglich gemacht werden oder wurden, erachten wir eine Fokussierung hierauf für angebracht.

2 Datenerhebung

Die Datenerhebung umfasst die Aufzeichnung des Sprechereignisses auf Audio und/oder Video, sowie die Erfassung weiterer Daten, insbesondere Metadaten zur Sprechsituation und zu den aufgenommenen SprecherInnen. Vor oder bei der Erhebung sollte außerdem direkt die informierte Einwilligung („Informed Consent“) der SprecherInnen eingeholt werden, um die beabsichtigte Verwendung der Daten (Analyse im Projekt, möglichst auch dauerhafte Archivierung und Weitergabe zu Forschungszwecken) datenschutzrechtlich zu autorisieren.[4]

4 In diesem Kontext können auch ethische Aspekte der Datenerhebung und ggf. die damit zusammenhängende Notwendigkeit, Datenerhebungen von einem Ethikrat oder ähnlichen Einrichtungen genehmigen zu lassen, einer Rolle spielen. Siehe dazu z. B. die Einträge unter „Ethikvotum“ auf der FAQ-Seite der DFG: https://www.dfg.de/foerderung/faq/geistes_sozialwissenschaften/

2.1 Aufnahmen

Hinsichtlich der Aufnahme stellen sich bei Lernerdaten keine prinzipiell anderen Herausforderungen als bei der Aufzeichnung anderer Arten mündlicher Kommunikation. Mindestanforderung ist eine Audio-Aufnahme in möglichst guter Qualität, die die Beiträge aller SprecherInnen einfängt und Störgeräusche so weit wie möglich vermeidet. Je nach Setting und Analyseinteresse kann es sinnvoll sein, ein oder mehrere Videoaufnahmen zu machen, insbesondere wenn auch nonverbale Interaktionsanteile interessieren oder – bei Mehrpersonengesprächen – eine Sprecherzuordnung allein auf Grundlage des Audios schwierig ist. Für viele phonetische Analysen ist eine sehr hohe Audioqualität unabdingbar. Diese wird vor allem durch das Vorhandensein mehrerer getrennter Audio-Aufnahmen für verschiedene Sprecher sowie Aufnahme-Hardware und Raumbedingungen, die eine hochwertige Aufnahmen garantieren, erzielt. Dies erleichtert auch das Transkribieren, insbesondere simultaner Passagen. Zu bedenken ist andererseits immer, dass komplexere Aufnahmesettings, in denen mehrere Geräte bedient werden müssen, die Natürlichkeit der Sprechsituation einschränken können. Wenn dasselbe Sprechereignis in mehreren Aufnahmen aufgezeichnet wird, muss sichergestellt werden, dass diese Aufnahmen miteinander synchron sind oder nachträglich synchronisiert werden können.

Audio- und Videoaufnahmen sollten grundsätzlich in der bestmöglichen Qualität gemacht werden, die das jeweilige Gerät bietet. Kompressionsfreie Formate (WAV bei Audio) sind in diesem Sinne verlustbehafteten komprimierten Formaten (MP3) vorzuziehen, und es sollten möglichst hohe Bildauflösungen und Samplingraten gewählt werden. Die Daten sollten in weit verbreiteten Standard-Formaten gespeichert werden. Das Archiv für Gesprochenes Deutsch hat z. B. die folgenden Parameter als gut geeignet für archiv-fähige audiovisuelle Sprachdaten festgelegt:

- Für Audio: PCM-WAV, Stereo, 48kHz Samplingrate, 16bit Auflösung
- Für Video: MPEG-4 mit Enkodierung H.264/MPEG-4 Part 10 AVC, Bildwiederholungsrate: 25fps, Einzelbildgröße 1980x1080, konstante Bitrate, Audio-Spur als AAC mit 48kHz und 384 Kbps

Der Speicherbedarf für Daten mit diesen Parametern beläuft sich auch etwa 1GB pro Stunde Audioaufnahme bzw. 4GB pro Stunde Videoaufnahme. Es ist dringend zu empfehlen, Daten direkt nach der Erhebung systematisch so abzulegen, dass sie gegen versehentlichen Verlust und unbefugten Zugriff geschützt sind – in aller Regel bedeutet dies eine Entscheidung für einen institutseigenen Serverplatz mit geeigneten Backup-Mechanismen.

2.2 Metadaten

Metadaten zur Sprechsituation und zu den SprecherInnen sind unabdingbar sowohl für viele Analysezwecke als auch für die Optimierung des Nachnutzungspotentials eines Korpus. Sie werden idealerweise vor oder unmittelbar nach der Aufnahme erhoben. Einige grundlegende Daten sollten unabhängig vom Korpusdesign und Untersuchungsinteresse immer erhoben werden, etwa Datum und Ort der Aufnahme, die verwendete Aufnahmetechnik, und zu jedem/r SprecherIn Alter, Geschlecht und Bildungsstand. Dazu können korpusspezifische Metadaten kommen, die Parameter des Korpusdesigns erfassen – beispielsweise bei einer Map Task die Information, ob ein Sprecher die Rolle des Instruktionsgebenden oder -nehmenden einnimmt. Eine besondere Rolle bei Lernerkorpora spielt sicherlich die Erhebung der Sprachbiografien der einzelnen SprecherInnen. Hier sollten umfassende Informationen zu L1 und L2, möglichst auch Angaben zu Erwerbsverläufen, sprachlich prägenden Regionen und zur Sprachverwendung systematisch erfasst werden.

Es gibt für die zu erhebenden Metadaten keine etablierten Standards (auch wenn hierzu seit einiger Zeit debattiert wird; vgl. Stemle et al. 2019), die etwa ein festes Vokabular für Metadaten-Parameter oder -Werte festlegen würden. Die bereits genannten Korpora können aber als Beispiele guter Praxis dienen.

Die folgenden Metadaten sind in den meisten Lernerkorpora dokumentiert. Sie sollten bei der Bereitstellung von Korpusdaten nicht fehlen, damit bei der Korpusauswertung die Daten hinsichtlich spracherwerbsbedingter und durch den Produktionskontext bedingter Faktoren interpretierbar sind:

- Biographische Angaben zu den KommunikationsteilnehmerInnen (Alter, Geschlecht, Bildungsstand, berufliche Angaben, ...)
- Spracherwerbsbiographische Angaben zu den KommunikationsteilnehmerInnen (Sprachstand, L1, erworbene Zweit- bzw. Fremdsprachen, Erwerbszeiten, Erwerbstypen, ...)
- Angaben zur Kommunikationssituation (Anzahl der TeilnehmerInnen, Gesprächstyp, Gesprächs- bzw. Erhebungszeit, ...)

Sollte das Korpus hinsichtlich einer bestimmten Variable homogen aufgebaut sein (z. B. ausschließlich SprecherInnen mit Erstsprache Englisch enthalten), kann es sein, dass diese Informationen nicht in den Metadaten zu den einzelnen Aufnahmen oder Transkripten im Korpus zugeordnet werden, sondern als Korpusmetadaten in der Korpusdokumentation oder innerhalb der Gesamtdaten festgehalten werden.

Für Informationen zur technischen Realisierung von Metadatenannotationen siehe Abschnitt 3.1 dieses Beitrags.

2.3 Rechtliche Autorisierung

Aufzeichnungen sprachlicher Interaktion fallen unter die Datenschutzgrundverordnung (DSGVO), denn Stimme und (bei Video) Bilder von Personen sind grundsätzlich als „personenbeziehbare Daten" schützenswert und dürfen nicht ohne Zustimmung der betroffenen Personen aufgezeichnet, verarbeitet oder weitergegeben werden. (Relevant für viele Erhebungen: Im Falle Minderjähriger bedarf es zusätzlich die Zustimmung von Erziehungsberechtigten). Dieser Verpflichtung seitens der Projektleitenden kann mit einer „Informierten Einwilligung" (Informed Consent) begegnet werden, bei der die SprecherInnen möglichst schriftlich darüber informiert werden, ...

- ... wer die Daten erhebt (z. B. Projekt XY);
- ... zu welchem Zweck die Daten erhoben werden (z. B. sprachwissenschaftliche Forschung);
- ... wer primär mit den Daten arbeitet (vor allem Projektmitglieder) und an wen die Daten potentiell weitergegeben werden (weitere ForscherInnen, die z. B. online Zugriff auf die Daten haben);
- ... wie die Daten gespeichert und verarbeitet werden (z. B. auf einem universitätsinternen Server, in einem Archiv oder Datenzentrum);
- ... welche Maßnahmen zum Schutz personenbeziehbarer Daten getroffen werden.

Die genaue Ausformulierung einer adäquaten Einwilligungserklärung lässt sich nur im konkreten Projektkontext klären, und dieser Beitrag kann auf die vielfältigen rechtlichen und forschungspraktischen Fragen, die sich in diesem Zusammenhang stellen, nicht im Detail eingehen. Allgemein ist festzuhalten, dass die informierte Einwilligungserklärung ein Element der Korpuserhebung ist, das wesentlich über Möglichkeiten und Beschränkungen der späteren Korpusnutzung mitentscheidet und deshalb mit der gebotenen Sorgfalt betrachtet werden sollte. Weitere Informationen zu diesem Thema finden sich beispielsweise in der Handreichung Datenschutz des RatSWD (RatSWD 2020). Beispiele für Einwilligungserklärungen, die sich in der Praxis der Datenerhebung beim Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) bewährt haben, finden sich im Gesprächsanalytischen Informationssystem (GAIS, http://gais.ids-mannheim.de/).

Siehe Abschnitt 3.3 dieses Beitrags für die technische Realisierung der Maskierung von Sprachdaten (bei persönlichkeitsrechtlich relevanten Inhalten) im Korpus.

3 Erschließung

Unter den Begriff „Erschließung“ fassen wir sämtliche Prozesse der Aufbereitung und Weiterverarbeitung derjenigen Daten, die primär in der Korpuserhebung erfasst wurden (man spricht auch von Primärdaten; als solche gelten häufig die dem Korpus zugrunde liegenden Audio- und/oder Videoaufnahmen). Für das eigentliche Ziel – die linguistische Analyse – sind diverse Verarbeitungsschritte erforderlich: Im Falle gesprochener Daten bedarf es insbesondere deren Transkription. Für jegliche digital auswertbare Korpusdaten sind korpusinterne Analysen – sog. Annotationen – erforderlich, aber auch die digitale Repräsentation der Metadaten zählen wir zur Korpuserschließung. Primäres Ziel ist es, Korpusdaten zu erhalten, die sich maschinell weiterverarbeiten und in Hinblick auf die interessierenden Forschungsfragen mit korpuslinguistischen Werkzeugen auswerten lassen. Darüber hinaus sollte bei der Erschließung auch auf Aspekte der Nachhaltigkeit geachtet werden, also auf die Frage, ob die gewählten Methoden und Formate eine langfristige Nutzbarkeit der Daten gewährleisten.

3.1 Metadaten

Metadaten werden im Feld oder in der Experimentsituation oft zunächst in unstrukturierter Form – z. B. auf Papierformularen – erhoben. Für die Arbeit mit einem Korpus müssen sie in strukturierte digitale Form überführt werden, die eine maschinelle Weiterverarbeitung ermöglicht. Die einfachste und für viele Zwecke auch ausreichende Methode hierfür ist die Eingabe der Daten in eine geeignete Tabellenstruktur (beispielsweise in einem oder mehreren Tabellenblättern in MS Excel). Ein systematisches Benennungssystem kann dabei sicherstellen, dass Aufnahmen, Transkripte, Sprecher und in den Transkripten verwendete Sprecherkürzel, auf die sich die Metadaten beziehen, einander eindeutig zuzuordnen sind.

Tab. 1: Metadaten zu Aufnahmen bzw. Sprechereignissen

Aufnahme	Transkript	Situation	Datum	Ort	Dauer	Sprecher …
MT_HH_001.WAV	MT_HH_001.exb	Map Task	15.06.2019	Hamburg	00:23:15	HG, VK
MT_HH_002.WAV	MT_HH_002.exb	Map Task	17.06.2019	Hamburg	00:19:45	FD, HG
CC_HB_003.WAV	CC_HB_003.exb	Nacherzählung	17.06.2019	Bremen	00:30:31	HP
…						

Tab. 2: Metadaten zu SprecherInnen

Sprecher-kürzel	Pseudonym	Alter	Geschlecht	Bildung	L1	L2-Niveau	...
HG	Hermione Granger	24	weiblich	Hochschul-abschluss	Englisch	B1	
VK	Viktor Krum	26	männlich	Abitur	Bulgarisch	A2	
FD	Fleur Delacour	25	weiblich	Fachabitur	Französisch	A2	
...							

Bei größeren Korpora und/oder umfangreichen oder komplexen Metadaten kann es empfehlenswert sein, ein spezialisiertes Tool zu verwenden. Für die nachhaltige Archivierung der Daten ist es darüber hinaus oft notwendig, Metadaten in ein Format zu überführen, das das jeweilige Archiv oder Datenzentrum vorgibt.

Ein spezialisiertes Tool, das z. B. bei HaMaTaC und HaMoTiC verwendet wurde, ist der Corpus Manager (CoMa) aus dem EXMARaLDA-System (Schmidt/Wörner 2014). CoMa stellt eine allgemeine Struktur für die Metadaten gesprochener Korpora bereit, in der Aufnahmen und Transkripte in sog. „Communications" gebündelt und mit den zugehörigen Sprechern verknüpft werden. Zu jeder Dateneinheit, also zu Communications, Aufnahmen, Transkripten und Sprechern können Metadaten als freie Attribut-Wert-Paare oder in vorgegebenen Strukturen (z. B. zu Orten oder Sprachen) festgehalten werden. Über die Metadaten-Dokumentation hinaus enthält CoMa Funktionen zur Wartung (z. B. Konsistenzprüfung) der Daten und einige grundlegende Analyse- und Annotationsfunktionen (z. B. zum Erstellen von Wortlisten oder zum Part-Of-Speech-Tagging eines Korpus).

Wenn CoMa für die Erschließung der Metadaten verwendet wird, kann das damit aufgebaute Korpus auch mit EXAKT, dem EXMARaLDA Analyse- und Konkordanztool (s. u.), durchsucht und ausgewertet werden.

Mit dem IMDI-Editor (www.mpi.nl/tools/imdieditor.html) und ARBIL (https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045) hat auch das MPI in Nijmegen Tools entwickelt, die, vergleichbar mit CoMa, die systematische Erfassung und Pflege von Metadaten unterstützen. Beide Tools werden jedoch nicht mehr weiter entwickelt. Aus einer neueren Initiative, die vor allem auf Daten aus der Sprachdokumentation abzielt, geht das Tool LaMeta (https://sites.google.com/site/metadatatooldiscussion/home) hervor.

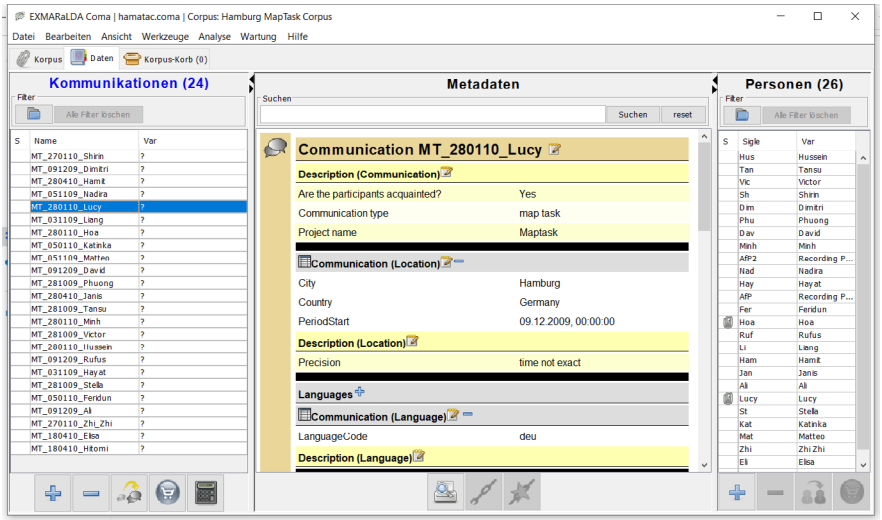

Abb. 1: Nutzerinterface des EXMARaLDA Corpus Managers (CoMa)

Im Kontext der Infrastrukturinitiative CLARIN wurde die „Component Metadata Infrastucture“ CMDI entwickelt. CMDI definiert kein konkretes Metadatenschema, sondern stellt ein Framework dar, in dem Metadaten-Beschreibungen aus vorhandenen und eigenen Komponenten zusammengestellt werden können. CMDI wird von allen CLARIN-Datenzentren gefordert und unterstützt.

3.2 Transkription, Normalisierung, Annotationen

In den folgenden Abschnitten werden die wichtigsten Schritte zur Weiterverarbeitung der Audio- oder Videoaufzeichnungen erläutert. Dies beinhaltet die Transkription (Verschriftung des Sprachsignals), ggf. detaillierte phonetische Analysen, eine Normalisierung (Vereinheitlichung von Wortformen gemäß der Standardorthographie) und verschiedene Annotationen (Hinzufügen weiterer linguistischer Analysen). Konzeptionell lassen sich diese Verarbeitungsschritte nicht strikt trennen, da es sich bei jedem Verfahren um eine Interpretation des zugrunde liegenden Sprachsignals handelt, womit die genannten Verarbeitungsschritte per Definition als bestimmte Typen von Annotation (der ursprünglich erhobenen Sprachdaten) einzuordnen sind. Dennoch sind Transkription, Normalisierung und Annotation(en) in aller Regel in der Praxis der Korpuserstellung getrennte Verarbeitungsschritte und werden dementsprechend in den folgenden Darstellungen auch getrennt behandelt.

3.2.1 Transkription und Annotation von Mündlichkeitsphänomenen

Einordnung

Bei der Erschließung gesprochener Korpora meint Transkription die Verschriftung des ursprünglich aufgenommenen Sprachsignals. Hierbei werden grundlegende Entscheidungen gefällt, nämlich welche sprachlichen Merkmale im Korpus abgebildet werden und somit systematisch untersucht werden können. Grundlegend zu beachten ist dabei, dass das akustische Sprachsignal selber kaum Möglichkeiten der systematischen Auswertung linguistischer Kategorien bietet, sondern dass Korpusauswertungen grundsätzlich auf verschrifteten Analysen stattfinden.

Grundsätzlich können sämtliche Merkmale des Sprachsignals transkribiert werden, z. B. Aussprachebesonderheiten in Form bestimmter Abweichungen von der Standardlautung, zu denen Merkmale wie Behauchung oder markierte Akzentstrukturen gehören können. In der Praxis der Korpusaufbereitung ist die Transkription bzw. die Annotation phonetischer Merkmale jedoch ein äußerst zeitaufwändiger Prozess, weil hierfür ein hohes Maß an manueller Analysearbeit aufgebracht werden muss. Deshalb wählen die meisten Projekte zur korpusbasierten Aufbereitung von gesprochener Sprache einen pragmatischen Ansatz zur Verschriftung des Sprachsignals, der sich durch die folgenden Merkmale auszeichnet:

- Auf eine lautschriftliche Repräsentation des Sprachsignals wird verzichtet. Standardgemäß artikulierte Formen und davon abweichende Lautungen werden mit normaler Alphabetschrift repräsentiert. Die hierfür notwendigen Regelungen sind in Transkriptionskonventionen formuliert (siehe nächster Abschnitt).
- Die Verschriftung phonetischer Merkmale sowie außersprachlicher Signale beschränkt sich auf wenige Aspekte und erfolgt häufig inline (gemeinsam mit dem transkribierten Sprachsignal).

Ausnahmen hierzu bilden die im phonetischen Analysewerkzeug PRAAT erstellten Korpora IFCASL und Leap (siehe hierzu Tab. 3 am Ende des Abschnitts): IFCASL enthält eine vollständige phonetische Transkription mit lautgenauer zeitlicher Alignierung zum akustisch dargestellten Sprachsignal mit Spektrogramm. In Leap wurde silbengenau transkribiert bzw. mit dem Sprachsignal aligniert. Alle anderen der hier genauer betrachteten Korpora im Grunde orthographisch transkribiert (ggf. mit leichten Abweichungen von der orthographischen Norm zur Kennzeichnung gewisser phonetischer Merkmale) und sehr viel gröber mit dem Audiosignal aligniert.

Richtlinien

Im Rahmen der Gesprächsanalyse (die ihre Datengrundlage traditionell nicht im modernen Sinn korpusbasiert aufbereitet) ist das Gesprächsanalytische Transkriptionssystem GAT (Selting et al. 1998) bzw. die weiterentwickelte Version GAT2 (Selting et al. 2009) entstanden, das ein breites Beschreibungsinventar für die Mündlichkeit in gesprochenen Dialogen bietet. Auf drei Detailstufen (Minimal-, Basis- und Feintranskript) können neben den artikulierten sprachlichen Einheiten in Form einer literarischen Umschrift (der Abbildung gesprochener Wörter in kleingeschriebener Alphabetschrift), die in Turns (sprecherbezogene Gesprächseinheiten) und Intonationsphrasen gegliedert werden, diverse Merkmale der Mündlichkeit, wie z. B. Pausen, Dehnungen, Akzente oder Tonhöhenverläufe, nach klaren Vorgaben abgebildet werden. Auch nicht hörbare Handlungen von Gesprächsteilnehmenden (z. B. gestisch-mimische Handlungen) sowie verschiedene Typen von Annotatorenkommentaren, etwa zur Deutlichkeit der Sprache oder (Un)Sicherheit der eigenen Analyse, können im Transkript festgehalten werden. GAT und GAT2 beinhalten keine Richtlinien für die Kennzeichnung erwerbsbedingter Phänomene.

Angelehnt an GAT2, aber speziell für die Nutzung dieser Konventionen in computergestützten Anwendungen entworfen, sind die Transkriptionsrichtlinien cGAT (Schmidt et al. 2015). Hier finden sich detaillierte Hinweise auch zur Handhabung von Transkriptionen in dem Programm FOLKER (s. u.).

Das Transkriptionssystem HIAT (Halbinterpretative Arbeitstranskription, Rehbein et al. 2004) beinhaltet dem GAT-System ähnliche Analysekategorien, allerdings wird hier auf die Möglichkeit von Mehrebenenannotationen zurückgegriffen, wodurch verschiedene Informationstypen auf separaten Spuren festgehalten werden können. Dies führt u. a. dazu, dass die phonetische Transkription der Lautsegmente orthographischer erfolgen kann: HIAT-Transkriptionen enthalten Groß- und Kleinschreibung sowie Interpunktion. Beide Merkmale sind in GAT der Markierung prosodischer Phänomene vorbehalten. Die HIAT-Richtlinien enthalten zusätzlich zu den Anweisungen zum Transkribieren Beschreibungen zum Realisieren der Transkriptionen im EXMARaLDA-Partitur-Editor. Zudem beinhalten die Richtlinien einen Abschnitt zur Behandlung von „Mehrsprachigkeit in der Transkription“ (S. 57 f.).

Die Transkription des Audiomaterials nach GAT2 führt dazu, dass grundsätzlich keine standardorthographische Groß-/Kleinschreibungsunterscheidung gemacht wird. Weiterhin wird bei solchen Wortformen von der Standardorthographie abgewichen, bei denen gewisse Abweichungen von der Standardlautung („zwo“ vs. „zwei“, „ick vs. „ich“, etc.) zu verzeichnen sind. Somit ist die Eignung der Transkriptionsebene für das systematische Suchen nach bestimmten Lexemen eingeschränkt (siehe 4.2.2. zur Erstellung einer Normalisierungsebene mit anti-

zipierbaren Wortformen). Umso mehr eignet sich die Transkriptionsebene zur Abbildung lautlicher Variation.

Würde man phonetische Aspekte priorisieren, müsste die Segmentierung der gesprochenen Einheiten konsequenterweise phonemgenau sein oder an phonotaktischen Merkmalen (Pausen) erfolgen. Dies läuft allerdings der grundlegenden Schreibgewohnheit der Transkribenden sowie der üblichen Token-Segmentierung im Korpus nach Worteinheiten zuwider. Aus diesem Grund wird in gesprochenen Korpora allgemein nach einzelnen Wörtern segmentiert, was im Grunde einen Normalisierungsschritt in Richtung schriftsprachlicher Norm vorwegnimmt (vgl. Abschnitt 4.2.2).

Die bereits erwähnten Korpora IFCASL und Leap basieren auf SAMPA-Transkriptionen (Wells 1997), die darauf ausgelegt sind, sprachübergreifend echte phonetische Transkriptionen mit einem standardkonformen (ASCII) Zeicheninventar zu erstellen.

Welche Merkmale des Sprachsignals und außersprachlicher Merkmale neben der Darstellung der artikulierten Lautfolgen mit der Transkriptionsebene (oder mehrerer Analyseebenen) abgebildet werden, unterscheidet sich von Korpus zu Korpus stark. Die folgende Liste fasst diejenigen Merkmale zusammen, die in unserer Korpusübersicht in mindestens einem Fall zu finden sind.

- Wellenform (Oszillogramm, das Intensitätsunterschiede des Sprachsignals zu jedem Zeitpunkt anzeigt, vgl. Abb. 3)
- Spektogramm (auch Sonagramm; Nutzung der möglichen Frequenzbereiche zu jedem Zeitpunkt, vgl. Abb 4)
- visuell ersichtliche Kommunikationssignale: Gestik, Mimik
  - verschriftet (annotiert)
  - durch Videoaufzeichnungen
- Akzente (Wort- und/oder Satzakzente)
- schnelle Anschlüsse zwischen Diskurseinheiten
- Überlappungen von Redebeiträgen
- Tonhöhenverläufe
- Pausen (stille und/oder gefüllte)
- hörbare nonverbale Redeanteile (Lachen, Seufzen usw.)
- extralinguistische Merkmale (Geräusche, die nicht als Teil einer sprachlichen Äußerung interpretierbar sind)
- Annotatorenkommentare (Unsicherheiten, bestimmte Eigenschaften des Aufnahmesignals)

Auch wenn einige Transkriptionsrichtlinien mehrere dieser Merkmale auf einer Beschreibungsebene vereinen, handelt es sich strenggenommen um die Annotation verschiedener phonetischer und/oder prosodischer Merkmale, die im

Korpus auch auf gesonderten Analyseebenen beschrieben werden können und z. T. auch separat analysiert werden.

Phu: äh also das erzähle ich dir nochmal ((holt Luft)) dann ähm vor den Büchern ähm gehst du geradeaus also Richtung obere Seite des Zettels ((lacht, 1,6s)) und dann ((0,4s)) ähm biegst du/ ähm ((0,2s)) biegst du nach/ ((0,8s)) äh (f/) wie heißt nochmal ((0,3s)) (j/) ähm biegst du nach rechts ((0,2s)) hinter dem Büchern ((0,5s)) dann ((0,3s)) gehst du geradeaus entlang diesem äh Zettelra/ also Zettelrand

(Transkriptausschnitt aus HaMaTaC, transkribiert nach HIAT mit Inline-Annotation von Pausen und hörbaren non-verbalen Handlungen in doppelten runden Klammern, Markierung von Reparaturen mit Schrägstrich, Markierung von Unsicherheit durch einfache runde Klammern)

730 [04:36.4]	731 [04:37.0]	732 [04:37.4]	733 [04:38.0]	734	735	736	737	738 [04:39.8]	739 [04:40.0]	740 [04:40.6]	741 [04:40.9]	742 [04:41.3]
((0,5s))	dann	((0,3s))	gehst	du	geradeaus	entlang	diesem	äh	Zettelra/	also	Zettelrand	ne
								EDIT PHASE	TROUBLE	EDIT PHASE	RESTART	

Abb. 2: Teil-Ausschnitt der gegebenen Transkription im EXMARaLDA Partitur-Editor (mit Ausweisung verschiedener Disfluency-Typen)

Werkzeuge

Transkriptionswerkzeuge dienen dazu, das Audiosignal (bzw. Videosignal) einzulesen, abzuspielen, zu transkribieren, wobei verschriftete Textpassagen mit dem jeweils zugehörigen Audioausschnitt aligniert werden.

Wie bereits angedeutet, kann der EXMARaLDA Partitur-Editor zur Erstellung von Transkriptionen verwendet werden. Eine etwas jüngere Lösung ist das für das bereits erwähnte gesprochene Korpus FOLK entwickelte Transkriptionswerkzeug FOLKER (https://exmaralda.org/de/folker-de/), das direkte Anbindung an das Normalisierungswerkzeug OrthoNormal (siehe 4.2.2) besitzt. Abbildung 3 zeigt die den Ausschnitt einer Transkription aus HaMaTaC im EXMARaLDA Partitur-Editor.

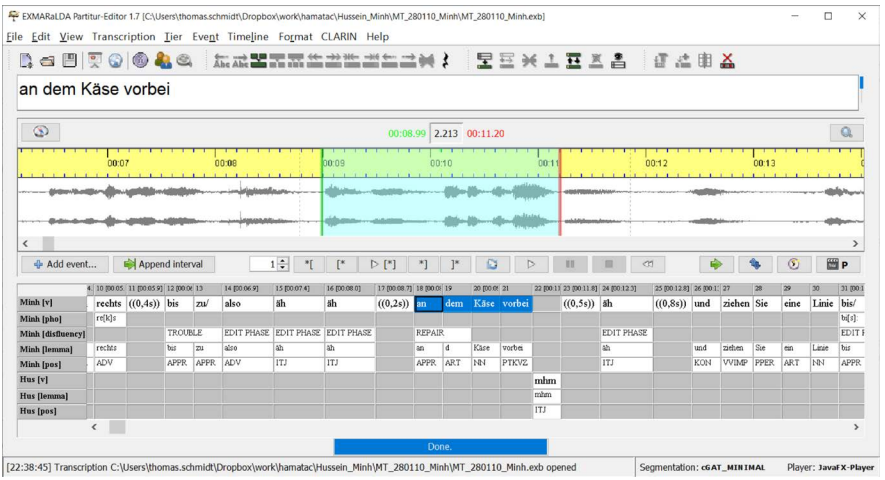

Abb. 3: Beispiel einer in EXMARaLDA durchgeführten Transkription: Die Spur (Zeile) „Minh [v]“ zeigt orthographisch transkribierte Wortformen, Filler und Pausen, die mit dem oberhalb abgebildeten Audiosignal (Wellenform) aligniert sind. Die Spur „Minh [pho]“ zeigt bei merklichen Abweichungen von der Standardlautung an. Die Spur „Minh [disfluency]“ macht bestimmte Disfluenzphänomene explizit.

Weitere weit verbreitete Transkriptionswerkzeuge (siehe auch Rohlfing et al. 2006) sind:

- ELAN: (Wittenburg et al. 2006, https://archive.mpi.nl/tla/elan) ist ein dem Partitur-Editor vergleichbares Werkzeug, das komplexe Mehrebenen-Annotationen erlaubt. Es ist weit verbreitet in der Sprachdokumentation und für die Annotation von Gebärdensprache und bietet besonders fortgeschrittene Funktionalität für die Videotranskription, z. B. die Möglichkeit, mehrere Videoaufnahmen synchron abzuspielen.
- Praat (Boersma 2001, www.fon.hum.uva.nl/praat/) ist ein Werkzeug mit umfangreicher Funktionalität für phonetische Auswertungen des Audio-Signals (z. B. Spektrogramm-Darstellung, Berechnung von Intonations-Konturen), kann aber auch einfach als Transkriptionswerkzeug verwendet werden. Dies bietet sich besonders dann für den Fall an, dass präzise Notation phonetisch/phonologischer Merkmale Teil der Transkriptionskonventionen sind.
- WebMAUS (Kisler et al. 2017, www.bas.uni-muenchen.de/Bas/BasMAUS.html#webmaus) liest Audiodaten und orthographische Transkriptionen ein und gibt dazu phonetische Transkriptionen (aligniert mit den eingegebenen Informationen) aus, z. B. im Praat- oder EXMARaLDA-Dateiformat.

– CLAN (MacWhinney 2000, https://dali.talkbank.org/clan/) ist ein älteres Werkzeug, das vornehmlich mit Blick auf Daten zum (kindlichen) Spracherwerb im CHILDES-Kontext entwickelt wurde, aber auch für L2-Daten Anwendung gefunden hat (vgl. z. B. die in Talkbank verfügbare Version des ESF-Korpus). CLAN bietet umfangreiche Annotations- und Auswertungsmöglichkeiten und ist international weit verbreitet. Im Vergleich zu EXMARaLDA, FOLKER, ELAN und Praat ist es allerdings von der Benutzeroberfläche her technologisch eher überholt und hat auch bezüglich der Präzision und Weiterverarbeitbarkeit der Daten einige Nachteile (z. B. Alignment nur äußerungsweise, Datenstrukturen werden nicht explizit in XML repräsentiert).

Vgl. Abb. 4 für eine in PRAAT erstellte voll phonematisch segmentierte und alignierte Transkription (nach SAMPA-Richtlinien) im IFCASL-Korpus.

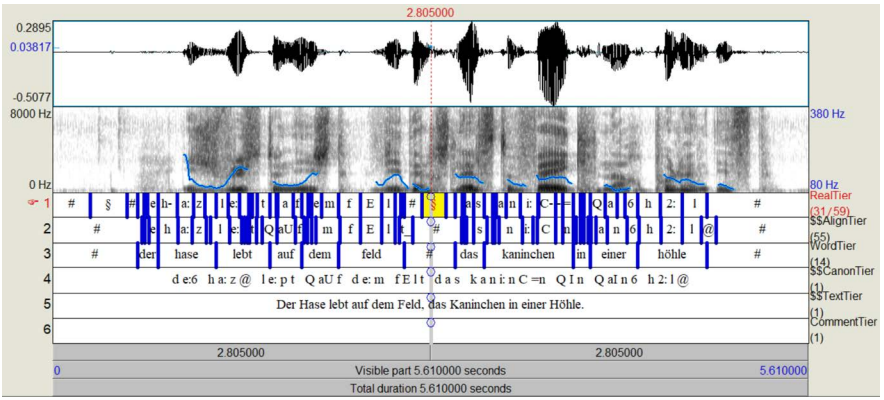

Abb. 4: Auf SAMPA-basierende Transkriptionen des Satzes *Der Hase lebt auf dem Feld, das Kaninchen in einer Höhle* im IFCASL-Korpus; oben abgebildet ist die Wellenform des Sprachsignals (die Lautstärke des Sprachsignals), darunter ein Spektogramm, in dem die Stärke der verschiedenen Frequenzbereiche des Sprachsignals durch Schwärzung angezeigt wird

Zwischen allen hier genannten Werkzeugen besteht eine weit reichende Interoperabilität (mit einigen Abstrichen für CLAN), d. h. es existieren Import- und Exportfunktionen, die es erlauben, Daten zwischen EXMARaLDA, FOLKER, ELAN und Praat auszutauschen (vgl. dazu auch Schmidt et al. 2009).

Die Übersicht (Tab. 3) zeigt, welche der erwähnten Transkriptionsschritte und -werkzeuge in den gesichteten Korpora Verwendung fanden.

Tab. 3: Übersicht zur Transkription in verschiedenen aktuellen gesprochenen Lernerkorpora

Korpus	Tools (Formate)	Richtlinien: Veröffentlichungen und Dokumentationsseiten	wesentliche Merkmale
BeMaTaC (L2-Kohorte)	EXMARaLDA (EXB-XML)	Zusammenfassende Webseite: https://hu.berlin/bematac-transkription	Videomitschnitte, wortweise aligniert mit Transkriptionsebene Umschrift nach individuellen Regeln, sehr orthographienah (mit Groß- und Kleinschreibung, ohne Interpunktion) Disfluency-Phänomene: Pausen, Filler, Abbrüche, Wiederholungen Extralinguistische Phänomene: nonverbale Ereignisse (Atmen, Lachen, …)
GeWiss (L2-Kohorte)	EXMARaLDA, FOLKER (EXB-XML/ FLN-XML)	Gräfe et al. 2015	Audiosignal, äußerungsweise aligniert mit Transkription nach GAT 2, Minimaltranskript
HaMaTaC	EXMARaLDA (EXB-XML)	Hedeland & Schmidt 2012 https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:hamatac-1.0.0	Audiosignal, äußerungsweise aligniert mit Transkription nach HIAT
HaMoTiC	EXMARaLDA (EXB-XML)	Angelehnt an HaMaTaC	Audiosignal, äußerungsweise aligniert mit Transkription nach HIAT
IFCASL	PRAAT (TextGrid)	Trouvain et al. 2016 http://www.ifcasl.org/	phonetische Segmente (Lautschrift nach SAMPA), aligniert mit Sprachsignal Darstellung mit Wellenform und Spektogramm
Leap	PRAAT EToBI, GToBI (modifiziert) (TextGrid)	Gut 2014 http://sourceforge.net/projects/leapcorpus/files/LeapCorpus_Manual.pdf/	silbische Segmentierung, aligniert mit Wellenform und Sonagramm Umschrift nach SAMPA und individuellen Regeln: Längungen Prosodie: Intonationsphrasen, unterbrochene Phrasen Akzentstrukturen Disfluency-Phänomene: (ungefüllte) Pausen Extralinguistische Phänomene: u. a. Geräusche, Atmung, Lachen
MULTILIT	EXMARaLDA (EXB-XML)	Schellhardt & Schroeder 2015	nach HIAT, aber GAT2-Segmentierung nach Intonationsphrasen o. satzbeendende Interpunktion

3.2.2 Normalisierung

Einordnung

Bei der Korpuserstellung bedeutet „Normalisierung“ die Vereinheitlichung von (heterogen repräsentierten) Formen. In der Praxis der späteren Korpusnutzung beziehen sich viele Auswertungsszenarien auf bestimmte Wortformen oder Grundformen, also lexikalische Formen. Um auf diese ganzheitlich zugreifen zu können, muss man sie einheitlich repräsentieren, d. h. bei der Erstellung gesprochener Korpora im Regelfall, dass Transkriptionsformen (Schriftformen, die die Lautung repräsentieren) gemäß der Standardorthographie repräsentiert werden. Auf diese Weise wissen KorpusnutzerInnen, wie bestimmte Lexeme gefunden werden, ohne antizipieren zu müssen, welche Varianten (Aussprachevarianten oder durch die Transkription bedingte Schreibungen) im Korpus vorliegen.

Die folgende Tabelle illustriert anhand von Beispielen aus FOLK (vgl. Winterscheid et al. 2019) verschiedene Aspekte der Normalisierung: im Falle von *gesagt* geben die transkribierten Formen einfach verschiedene Aussprachevarianten wieder; bei *hast Du* spielt zusätzlich teilweise Klitisierung, also das Zusammenziehen zweier (orthographischer) Wörter eine Rolle. Bei den verschiedenen Formen von *nein* wurde per Normalisierungsrichtlinie bewusst entschieden, auch die Form *nee*, die oft einen eigenen Wörterbucheintrag besitzt, unter die normalisierte Form *nein* zu fassen, eine mögliche Funktionsdifferenzierung zwischen *nee* und *nein* also auf der Ebene der Normalisierung nicht berücksichtigt. Bei *äh* schließlich wird die Formenvielfalt der Häsitationsmarker einfach generell auf die häufigste Form *äh* abgebildet, um eine systematische Recherche und Verarbeitung zu ermöglichen.

Tab. 4: Beispiele für Normalisierungen nach den FOLK-Richtlinien (Winterscheid et al. 2019)

Transkribierte Formen	Normalisierte Form
gesagt ; gesacht ; gsagt ; gsat ; gsacht ; gsaht ; gesach ; jesacht ; gesag ; gseit ; gesat ; gsogt ; gsa ; gesaht ; gesa ; gsag ; sat ; jesagt	**gesagt**
hast Du ; haste ; has Du ; hasch ; hasch Du ; hosch ; hast ; hasse ; hascht Du ; hascht ; hosch Du ; hasdu ; hoscht ; has ; host Du ; hasche ; hasde ; haschte ; hassu ; hastu	**hast Du**
nee ; nein ; nö ; na ; nä ; nei ; ne ; naa ; nää ; näh ; neihein ; noi ; neehee	**nein**
äh ; ähm ; hm ; öh ; mh ; öhm ; m ; ä ; n ; e ; ah ; a ; ahm ; eh ; w ; ehm ; ö	**äh**

Ein weiterer Nutzen, den die Normalisierung bringt, ist die bessere Verarbeitbarkeit der Daten: Programme, die bei der Annotation der Wortformen (vor allem

nach Wortarten und Lemmata, siehe 4.2.3) eingesetzt werden, sind normalerweise auf Standardtexten trainiert und funktionieren auf Nichtstandarddaten wie einer Transkriptionsebene im Korpus deutlich schlechter. In dieser Hinsicht ist die Normalisierung ein hilfreicher Schritt bei der Weiterverarbeitung der Korpusdaten mit automatischen Verfahren.

Die Normalisierung von Korpusdaten bedeutet keinesfalls, dass die der Normalisierung zugrunde liegenden Daten gelöscht oder überschrieben werden. In modernen Korpusarchitekturen können praktisch beliebig viele Analyseebenen hinzugefügt werden und nebeneinander bestehen. Auf diese Weise bedeutet Normalisierung das Hinzufügen einer zusätzlichen Betrachtungsebene im Korpus, die ergänzende Informationen liefert und unabhängig von oder zusätzlich zu den übrigen Ebenen im Korpus genutzt werden kann.

Bei der Erstellung gesprochener Korpora schließt sich der Verarbeitungsschritt der Normalisierung also in aller Regel an den Verarbeitungsschritt der Transkription an und ist hochgradig von ihm abhängig: Je enger sich die Transkription an der Abbildung der Lautlichkeit orientiert, desto mehr wird die Normalisierungsebene im Korpus von der Transkriptionsebene abweichen. Je mehr Transkriptionsformen bereits der standardorthographischen Repräsentation der Wörter entsprechen, umso weniger muss Normalisierungsschritte sind erforderlich.

Richtlinien

Normalisierung wird zwar als eine Operation aufgefasst, die sich auf nicht normierte Daten bezieht, sie lässt sich aber in verschiedene Aspekte zerlegen. Je nachdem, wie die Beschaffenheit der Textgrundlage (der transkribierten Textdaten) ist, können die folgenden Operationen unterschieden werden:

- Anpassung einer individuellen Lautung zu der durch die Standardschreibung repräsentierte Standardlautung (*ham* → *haben*), also Normalisierung der Wortschreibung
- Trennung klitisierter Formen gemäß der nicht klitisierten Standardvariante (*hastu* → *hast du*), also Normierung hinsichtlich der Zusammen- und Getrenntschreibung
- Hinzufügung von synkopierten oder apokopierten Lauten oder Lautfolgen (*machn* → *machen*, *schaffe* → *schaffen*)
- Löschung von im Redefluss wiederholten oder korrigierten Wortformen bzw. -fragmenten
- Realisierung der Substantivgroßschreibung (*meine hand* → *meine Hand*)
- Hinzufügen von Interpunktion (*ich glaube ich spinne* → *ich glaube, ich spinne!*)
- Realisierung der satzinitialen Großschreibung (*ich glaube, ich spinne!* → *Ich glaube, ich spinne!*)

Alle Punkte zusammengenommen bedeuten in der Praxis der Erstellung gesprochener Korpora, dass die Transkriptionsebene in einen Paralleltext überführt wird, der in allen Belangen der aktuell geltenden orthographischen Norm entspricht. In manchen Korpora wird aber auch auf bestimmte Normen verzichtet (so enthält z. B. das FOLK-Korpus an keiner Stelle standardorthographische Interpunktion).

Will man rein manuell (händisch) aus einer Transkriptionsebene eine Normalisierungsebene erstellen, so ist es sinnvoll, die Ebene mit transkribierten Wortformen zu duplizieren und Form für Form Korrekturen vorzunehmen, so dass gemäß den Richtlinien normalisierte Formen vorliegen.

Werkzeuge

Um die Normalisierung nicht rein händisch durchzuführen, wurden verschiedene Programme entwickelt, die den Normalisierungsprozess unterstützen sollen. Im Grunde ist die Funktionsweise der verschiedenen Werkzeuge ähnlich: Sie basieren auf einem Vollformenlexikon, signalisieren, wenn eine gegebene Wortform unbekannt ist und geben ggf. Vorschläge, welche Entsprechungen infrage kommen. OrthoNormal (https://exmaralda.org/de/orthonormal-de/; Schmidt 2017) baut auf einer mit dem Transkriptionswerkzeug Folker erstellten Transkription auf (liest also Dateien des Formats .flk ein). Ein Beispiel für die Überführung transkribierter Formen in normalisierte Formen zeigt Abb. 5.

Annotationen für GWSS_E_00003_SE_01_T_01_DF_01 / c10

ID	w402	w403	w404	w405	w406	w407	w408	w409	w410	w411	w412	w413	w414	w415	w416	w417
Transkription	eigentlich	bei	den	alltagsstra	äh	spra	sprachen	bei	den	dialekten	und	bei	ner	alltäglichen	sprache	anfangen
Normalisierung	eigentlich	bei	den	alltagsstra	äh	%	Sprachen	bei	den	Dialekten	und	bei	einer	alltäglichen	Sprache	anfangen

Abb. 5: Beispiel für eine Normalisierung aus dem GeWiss-Korpus (Quelle entstammt dem unter der Webadresse https://dgd.ids-mannheim.de/ verfügbaren DGD-Interface)

Die Abb. 5 zeigt wesentliche Normalisierungsschritte, die im FOLK-Korpus umgesetzt werden: Neben der Substantivgroßschreibung zeigt die Normalisierungsebene gegenüber der Transkriptionsebene eine (durch das %-Zeichen gekennzeichnete) Auslassung des wiederholten Wortfragments „spra“ sowie der normgemäßen Realisierung der verkürzten Form „ner“ des unbestimmten Artikels.

Werden Normalisierungen nicht mit dem dafür entwickelten Werkzeug OrthoNormal erstellt, bedeutet dies bei der Erstellung gesprochener Korpora in aller Regel, dass, basierend auf der Transkriptionsebene, den Korpusdaten eine Normalisierungsebene hinzugefügt wird (z. B. als ein entsprechender Tier in EXMARaLDA), die dann manuell und tokenbasiert bearbeitet wird.

Die Übersicht (Tab. 5) zeigt, welche der erwähnten Normalisierungsschritte und -werkzeuge in den gesichteten Korpora Verwendung finden.

Tab. 5: Übersicht zur Normalisierung in verschiedenen aktuellen gesprochenen Lernerkorpora

Korpus	Werkzeuge	Wesentliche Merkmale/Unterschiede zur Transkriptionsebene
BeMaTaC (L2-Kohorte)	EXMARaLDA	(Wortschreibungsregeln im Wesentlichen bereits auf Transkriptionsebene umgesetzt) Auflösung enklitischer Formen (Kontraktionen) Normierung der Zusammen- und Getrenntschreibung Löschung von Abbrüchen und wörtlichen Wiederholungen Vereinheitlichung von Interjektionen und Fillern keine Interpunktion
GeWiss (L2-Kohorte)	OrthoNormal	Im DGD-Version (https://dgd.ids-mannheim.de/) Normalisierung gemäß den FOLK-Richtlinien Normierung der Wortschreibung Vervollständigung von Auslassungen Umsetzung von Groß-/Kleinschreibungsregeln Löschung von Abbrüchen und wörtlichen Wiederholungen keine Interpunktion
HaMaTaC	(EXMARaLDA)	keine Normalisierung
HaMoTiC	(EXMARaLDA)	keine Normalisierung
IFCASL	PRAAT	Ebene „Lautsegment“: Hier wird bereits die realisierte Aussprache kanonisiert, was einem Normalisierungsvorgang entspricht Ebene „Word“: orthographische Wortformen in Kleinschreibung, keine Interpunktion Ebene „Text“: Orthographische Darstellung gesamter Äußerungen, inklusive Interpunktion
Leap	PRAAT	orthographische Wortformen in Kleinschreibung, keine Interpunktion
MULTILIT	(EXMARaLDA)	Segmentierung von Wortgrenzen, Formulierung von Zielhypothesen (vgl. Abschnitt 3.2.3)

Wie aus der Übersicht (Tab. 5) hervorgeht, besitzen nicht alle aufgeführten Korpora eine gesondert ausgewiesene Normalisierungsebene. Bei diesen Fällen handelt es sich um Korpora, die nach HIAT-Vorgaben transkribiert wurden, welche wesentliche Abweichungsmerkmale erfasst und somit eine standardsprachliche Norm nur implizit enthält.

3.2.3 Normalisierung vs. Annotation von Zielhypothesen

Einordnung

Der in Abschnitt 4.2.2 dargestellte Aufbereitungsschritt der Normalisierung gilt für alle Korpora gleichermaßen, die variierende und von dem schriftsprachlichen Standard abweichende Wortformen enthalten. In gesprochenen Korpora ist der Schritt der Normalisierung allgemein unabdingbar, weil im Mündlichen ganz regelmäßig Lautungen produziert werden, die, wenn man sie diplomatisch und nicht standardorthographisch transkribiert, entsprechend von der schriftsprachlichen Norm abweichen (vgl. z. B. die Standardlautung des Suffixes <-*ig*> als [ɪç], welches wiederum diplomatisch als „-ich" transkribiert werden könnte). Bei der Analyse von Lernersprache kommen weitere Normalisierungen ins Spiel, die aber für den erstsprachlichen Kontext nicht gelten, und zwar spracherwerbsbedingte Abweichungen von zielsprachlichen Formen. Vgl. hierzu Abb. 6.

dipl	un	dann	ich	bin	fast	bei	die	Ecke	wo
norm	und	dann	ich	bin	fast	bei	die	Ecke	wo

Abb. 6: Beispiel für eine Normalisierung (Ebene „norm") aus dem BeMaTaC-Korpus (Quelle entstammt dem unter der Webadresse https://hu-berlin.de/annis/ verfügbaren ANNIS-Interface)

Abb. 6 zeigt, dass die Normalisierung des auf der Ebene „dipl" transkribierten Redeausschnitts nur an einer Stelle einen Unterschied macht, und zwar bei der Komplettierung des nicht artikulierten letzten Konsonanten beim ersten Wort. Das Wortstellungsproblem (**dann ich bin* ...) sowie das Kasusproblem an dem Artikel (*die* statt *der*) werden nicht berücksichtigt. Dies ist korrekt so, weil laut den Normalisierungskonventionen die Normalisierungsebene nicht in (spracherwerbsbedingte) grammatische Probleme eingreifen soll. Bei der Analyse von Lernersprache sind nicht zielsprachliche Äußerungen bzw. Äußerungsteile jedoch interessant, weshalb es aus Sicht der Zweitspracherwerbsforschung sinnvoll wäre, solche Abweichungen systematisch zu finden. Hierfür wurde das Konzept der Zielhypothese (vgl. Reznicek et al. 2013 sowie Reznicek et al. 2012 für die Beschreibung der Zielhypothesenannotation im Falko-Korpus) etabliert. Vgl. hierzu die Abb. 7.

Das Beispiel in Abb. 7 entstammt dem Falko-Korpus, einem Lernerkorpus, basierend auf geschriebenen Essays fortgeschrittener DaF-Lernender. Die Zielhypothese dient dazu, jeder nicht zielsprachlichen Wortform eine grammatische Entsprechung zuzuweisen, so dass zum einen alle Abweichungen zwischen

tok	möglich		einem	klaren	Beitrag	zu	die	Gesellschaft	zu	sehen	.
ZH1	möglich	,	einen	klaren	Beitrag	zu	der	Gesellschaft	zu	sehen	.

Abb. 7: Beispiel für die wortgenaue Formulierung von Zielhypothesen (Ebene „ZH1“) aus dem Falko-Korpus (Quelle entstammt dem unter der Webadresse https://hu-berlin.de/annis/ verfügbaren ANNIS-Interface)

Lerneräußerung und Zielhypothese systematisch gefunden werden können und zum anderen ein anschließendes Fehlertagging auf der Interpretation des Unterschieds zwischen Lerneräußerung und Zielhypothese basiert und somit transparent ist.

Streng genommen sind Zielhypothesen eine spezifische Form der Normalisierung, denn auch hierbei werden Varianzen in den zugrunde liegenden Daten ausgeglichen und hinsichtlich einer Norm vereinheitlicht. Nur ist natürlich strikt zwischen einer Normalisierung, die auf der Beschaffenheit gesprochener Sprache im Allgemeinen beruht, und einer, die auf anderen Faktoren beruht (z. B. bestimmte grammatische Strukturen im Spracherwerb), zu trennen. Aus Sicht der Aufbereitungsprozedur bei gesprochenen Lernerkorpora wäre es günstig, nach der Normalisierung im Sinne des vorangegangenen Abschnitt 4.2.2 eine Erstellung von Zielhypothesen vorzunehmen, wie es in vielen schriftlichen Lernerkorpora der Fall ist. Doch bislang hat sich diese Herangehensweise bei der Aufbereitung gesprochener Lernerkorpora nicht durchgesetzt: Das einzige Korpus unter den hier genauer betrachteten, in welchem eine Annotation von Zielhypothesen und Fehlerkategorien umgesetzt wurde, ist MULTILIT. Ein wichtiger Grund für die Aussparung solcher Verarbeitungsschritte in den übrigen Korpora ist der Arbeitsaufwand für die Erstellung einer Transkriptions-, einer Normalisierungs- und zusätzlich einer Zielhypothesenebene. Dies schließt nicht aus, dass spracherwerbsbedingte Abweichungen von der Zielsprache mit bestimmten Fehlerkategorien gekennzeichnet werden (vgl. Abschnitt 4.2.5).

3.2.4 Annotation von Wortarten und Grundformen

Einordnung

Bei der Erstellung der meisten modernen Korpora erfolgen eine automatische Analyse von Wortarten – sog. POS-Tagging (POS: Part-of-Speech) – sowie die Zuweisung von Grundformen (Lemmata; der Prozess wird auch Lemmatisierung genannt). Da diese Analysen in der Regel in einem Verarbeitungsschritt durchgeführt werden (weil die Grundformanalyse eine wichtige Voraussetzung für die Zuweisung von Wortarten ist), werden sie hier auch gemeinsam behandelt.

Tagginggrundlage

Bei gesprochenen Korpora findet dieser zweiteilige Verarbeitungsschritt idealerweise auf der Ebene normalisierter Wortformen statt. Für die automatische Zuweisung von Wortarten und Lemmata existieren diverse Werkzeuge (sog. Tagger) mit hochakkuraten Analyseergebnissen bei normgerechten bzw. normalisierten Daten. Da die Akkuratheit dieser Werkzeuge maßgeblich an der Beschaffenheit der bearbeiteten Sprachdaten hängt und die meisten Tagger auf normgerechte Daten ausgelegt sind, ist die Normalisierungsebene eine ideale Grundlage für das Hinzufügen von Wortarten und Grundformen und wird im Normalfall auch als Input für automatische Tagger verwendet.

Im Fall von Lernerdaten, so wurde bereits erwähnt, bietet sich durch die Perspektive der Annotation von Zielhypothesen ein sehr spezifisches Tagging an. Die existierenden Lernerkorpora mit schriftsprachlicher Textgrundlage und Zielhypothesenannotation nutzen selbige als Input für automatische Taggingprozesse. Hierbei wird in gewisser Weise gleichermaßen die originale Textgrundlage (die Lerneräußerung selber) beschrieben, sofern man die Unterschiede zwischen Lerneräußerung und der Zielhypothese berücksichtigt. Hinsichtlich der hier näher behandelten gesprochenen Lernerkorpora ergibt sich ein einheitliches Bild: Sie beinhalten keine Zielhypothesen und das Tagging (zusammenfassend beschrieben in Tab. 6) findet auf einer allgemeinen Normalisierungsebene statt.

Richtlinien

Für die Annotation sowohl von Wortformen als auch von Lemmata existieren verschiedene Richtlinien. Bei Wortarten ist in erster Linie das Kategoriensystem (Tagset) entscheidend: Sowohl die Menge möglicher Kategorien als auch die Bezeichnungen (Terminologie) sowie die Kriterien zur Vergabe der einzelnen Kategorien (ihre Definitionen) können entschieden variieren. Was das Kategoriensystem (Tagset) angeht, so hat sich für das Deutsche im Wesentlichen das sog. STTS (Stuttgart-Tübingen-Tagset) durchgesetzt, welches in verschiedenen Versionen bzw. Adaptationen vorliegt (eine Zusammenfassung einiger STTS-Varianten sowie anderer Tagsets sind auf der Internetseite www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/) zu finden. Zugehörige Annotationsrichtlinien für die einzelnen Tags finden sich in Schiller et al. 1999. Eine Anpassung des STTS mit Blick auf Kategorien, die im mündlichen Sprachgebrauch relevant sind, ist das sog. STTS 2.0 (Westpfahl et al. 2017). Zusätzlich zum herkömmlichen STTS sind hier bestimmte Wortklassen ausdifferenziert (bspw. Enthält das Tagset verschiedene Partikeln, die das STTS nicht vorsieht), außerdem wird hier beschrieben, wie mit gewissen Performanzproblemen wie Wortabbrüchen umzugehen oder wie ähnliche, verwechselbare Wortklassen disambiguiert werden können.

Werkzeuge

Das am häufigsten verwendete Taggingprogramm für die automatische Annotation von Wortarten und Lemmata ist der TreeTagger (www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/; Schmid 1994). Der TreeTagger ist in seiner Hauptdistribution ein kommandozeilenbasiertes Stand-Alone-Programm. Da der TreeTagger auf Trainingsdaten mit beliebigen Tagsets trainiert werden kann, existieren verschiedene Versionen (unterschiedliche Parameterdateien, die spezifisch für bestimmte Tagsets und ein bestimmtes Tag-Vergabeverhalten sind).

Die ‚herkömmliche' Version des TreeTaggers für deutsche Textdaten – so gilt es auch für viele andere Tagger – verwendet als Tagset für die Klassifikation der Wortarten das STTS (Stuttgart-Tübingen Tagset; Schiller et al. 1999). Das Problem für mündliche Sprachdaten (wie auch für andere Nichtstandarddaten) besteht darin, dass das STTS für (bzw. anhand von) schriftliche(n) Standarddaten bzw. Zeitungstexte(n) entwickelt wurde: Das Tagset beinhaltet hier gewisse Lücken, z. B. dadurch, dass im STTS verschiedene Funktionen bzw. Klassen von Partikeln und Interjektionen nicht differenziert werden, die jedoch im Gesprochenen klar unterschiedliche Gebrauchskontexte besitzen. Deshalb wird das Standard-STTS in vielen Korpusprojekten um relevante Klassen erweitert. Zur taggergestützten Anwendung des bereits erwähnten „STTS 2.0" (Westpfahl et al. 2017) existiert eine TreeTagger-Parameterdatei (das ist eine auf STTS 2.0 trainierte Grammatik des Taggers; beziehbar unter www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), die für Inputdaten ohne standardorthographische Interpunktion optimiert wurde.

Ein wesentliches Problem beim Aufbau von Mehrebenenkorpora ist es, eine Ebene, die bereits an andere Ebenen geknüpft ist, mit einem Tagger zu bearbeiten und das Ergebnis des Taggingprozesses mit den bereits vorhandenen Ebenen zu alignieren. Deshalb wurden Wege geschaffen, innerhalb von Programmen und somit auch innerhalb komplexer Datenstrukturen bestimmte Ebenen taggen zu lassen. So ist es bspw. möglich, den TreeTagger im Normalisierungswerkzeug OrthoNormal (Schmidt 2017; https://exmaralda.org/de/orthonormal-de/) oder in einer angepassten Version von EXMARaLDA (Nolda 2019; https://hg.sr.ht/~nolda/exmaralda-dulko) zu verwenden.

Der Auszeichnung von Wortarten und Lemmata in Korpora allgemein wie in gesprochenen Lernerkorpora liegen also verschiedene Entscheidungen zugrunde: Zur Auswahl stehen verschiedene Tagger, verschiedene Tagging-Umgebungen, verschiedene Wortartentagsets und die Möglichkeit, eigene Tags hinzuzufügen sowie verschiedene Textebenen, die dem Tagging zugrunde liegen.

Die Übersicht (Tab. 6) zeigt, welche Annotationswerkzeuge zur Wortarten- und Lemma-Annotation in den gesichteten Korpora Verwendung finden.

Tab. 6: Übersicht zum pos-Tagging und zur Lemmatisierung in verschiedenen aktuellen gesprochenen Lernerkorpora

Korpus	Tools und ggf. Korrekturvorgänge	POS-Tagset (Richtlinien)
BeMaTaC (L2-Kohorte)	TreeTagger	STTS; Schiller et al. 1999
GeWiss (L2-Kohorte)	TreeTagger (DGD-Webinterface); teilweise manuelle Korrektur	STTS 2.0; Westpfahl et al. 2017
HaMaTaC	TreeTagger; manuelle Korrektur	STTS; ; Schiller et al. 1999
HaMoTiC	TreeTagger	STTS 2.0; Westpfahl et al. 2017
IFCASL	(kein Tagging durchgeführt)	–
Leap	keine Angabe zum Tagger; Evaluation der automatischen Annotation (Gut & Bayerl 2004)	STTS; ; Schiller et al. 1999
MULTILIT	TreeTagger; Anpassung des Tree-Tagger-Parameter-Files durch manuelle Annotationen und Trainingsprozess	STTS; ; Schiller et al. 1999, individuell erweitert

3.2.5 Annotation weiterer Merkmale

Einordnung

Mit der automatischen Annotation von Wortarten und Lemmata sind die Möglichkeiten, linguistische Merkmale in den Korpusdaten auszuweisen, keineswegs erschöpft. Moderne Korpusarchitekturen erlauben es, beliebig viele Informationen auf separaten, aber miteinander in Beziehung stehenden Analyseebenen hinzuzufügen, und prinzipiell kann jedes sprachliche (oder außersprachliche) Phänomen im Korpus annotiert werden. Im Folgenden sollen im Kontext der Spracherwerbsforschung interessante Phänomene, die bereits in bestehenden gesprochenen Lernerkorpora annotiert wurden, zusammengefasst werden.

Hier sei noch einmal darauf hingewiesen, dass es gewisse Überlappungen zwischen den Prozessen der Transkription und der Annotation gibt (strenggenommen ist die Transkription ein spezifischer Annotationsprozess). Deshalb kann es passieren, dass gewisse in Korpora annotierte Mündlichkeitsphänomene (wie z. B. Pausen) bereits beim Verarbeitungsschritt der Transkription berücksichtigt werden. Dies ist schlicht abhängig von den Transkriptionsrichtlinien. Obendrein können solche Phänomene spracherwerbsbedingte Merkmale aufweisen. Bspw. können Vokalfärbungen, Vokallängen oder -kürzen, bestimmte Wortakzentverhältnisse und viele andere Phänomene auf fremd- bzw. zweitsprachlichen Gebrauch zurückzuführen sein. Insofern sind an dieser Stelle viele Dopplungen mit den beim Transkriptionsprozess berücksichtigten Phänomenen möglich.

Schaut man auf die Auszeichnung von Merkmalen der grammatischen (zielsprachlichen) Anteile der Lerneräußerungen, so sind hier alle Merkmale annotierbar, die allgemein in Korpora annotiert werden können. Häufig annotierte Kategorien bzw. Phänomene in schriftsprachlichen Korpora sind:

- Morphologische Merkmale: wortbildungsmorphologische und/oder flexionsmorphologische Kategorien
- Phrasen und Phrasentypen
- Sätze und Satztypen
- Wortstellung, vor allem topologische Felder und Verbal- bzw. Satzklammern
- Syntax/Parsing: Vollständige Abbildung syntaktischer Strukturen durch syntaktische Phrasenstruktur- oder Dependenzbäume
- Informationsstrukturelle Kategorien: Koreferenz, Informationsstatus
- Diskursstrukturen, vor allem RST-Strukturen (RST=rhetorische Strukturtheorie)

In gesprochenen Korpora werden häufig bestimmte Mündlichkeitsphänomene, die an der Transkription erkennbar sind, aber gesondert markiert werden können, annotiert. Zu diesen gehören u. a.:

- Disfluency-Phänomene (auch als „Häsitationen“ oder „Performanzprobleme“ bezeichnet)
  - Abbrüche/Neuanfänge, Selbstkorrekturen
  - Wortwiederholungen
  - Auszeichnung von Pausen und Filler
- Rückkopplungssignale (Backchanneling)
- Überlappungen von dialogischen Redeanteilen
- Sprachwechsel (Codeswitching)

Will man spracherwerbsspezifische Abweichungen von der Zielsprache annotieren, so bietet sich an, zunächst die konkreten Abweichungen durch die Annotation von Zielhypothesen (siehe 3.2.3) explizit zu machen. Anschließend können bestimmte Typen von Abweichungen markiert werden. Zu diesem Zweck wurden verschiedene Fehlertagsets (dies sind Klassifikationen von Abweichungstypen) entwickelt. Siehe Lüdeling & Hirschmann (2015) für eine Zusammenfassung der Klassifikationsmöglichkeiten und existierenden Tagsets.

Die Übersicht (Tab. 7) zeigt, welche grammatischen und spracherwerbsspezifischen Merkmale in den gesichteten Korpora (zusätzlich zu den bislang behandelten Annotationen) annotiert wurden.

Tab. 7: Übersicht zur Annotation spracherwerbsspezifischer Merkmale in verschiedenen aktuellen gesprochenen Lernerkorpora

Korpus	Art der Annotation	Quelle/Beschreibung
BeMaTaC (L2-Kohorte)	Auszeichnung ungefüllter Pausen Äußerungsspannen Disfluenzen Backchannel-Signale	https://hu.berlin/bematac-annot https://hu.berlin/bematac-disc Belz 2014
GeWiss (L2-Kohorte)	Sprachwechsel Diskurskommentierungen (nur in L1-Daten) Zitate und Verweise (im L2-Subkorpus SV_DE_L2)	Gräfe et al. 2015 Baur et al. 2014
HaMaTaC	Disfluenzen; Pausen und Backchannel-Signale (Inline-Annotation auf der Transkriptionsebene)	Hedeland & Schmidt 2012
HaMoTiC	Pausen und Backchannel-Signale (Inline-Annotation auf der Transkriptionsebene)	–
IFCASL	Abweichungen auf der lautlichen Ebene Teilweise (in Kurztexten): Pausen und Disfluenzen	Trouvain et al. 2016 Trouvain, Fauth, Möbius 2016
Leap	Disfluenzen („hesitation phenomena“)	–
MULTILIT	(z. T. für Teile des Korpus, z. T. in Planung) – Satzspannen – Typen von Nominalphrasen – Textstrukturkategorien – Sprachwechsel – Markierung indirekter Rede – Korrektur von Norm-Abweichungen (Zielhypothesen) – Auszeichnung von Norm-Abweichungen (Fehlerklassen)	Schellhardt & Schroeder 2015

3.3 Maskierung

Oft beinhaltet die informierte Einwilligung der aufgenommenen SprecherInnen eine Zusage, dass Daten nur in einer Form weitergegeben werden, in der personenbeziehbare Bestandteile geeignet maskiert, anonymisiert oder pseudonymisiert wurden (siehe auch 2.3). Unter Maskierung verstehen wir, dass die (direkte) Personenbeziehbarkeit durch geeignete Methoden (die wir im Folgenden näher erläutern) vermieden wird. Was genau „personenbeziehbare“ Daten sind, hängt von den mit dem Korpus bereitgestellten Daten (vor allem Ton- und Bildaufnahmen) ab und muss in der Einwilligungserklärung festgehalten werden. Üblicherweise fallen hierunter:

- biographische Details wie Geburtsdatum oder Adresse;
- die Nennung von Eigennamen der aufgenommenen Personen oder von Menschen aus deren persönlichem Umfeld;
- weitere persönliche Details wie Telefonnummern, Kontodaten o. Ä., die im Verlauf eines aufgezeichneten Gesprächs genannt werden;
- „besonders schützenswerte Daten" im Sinne der DSGVO, also Äußerungen zu politischen, religiösen und philosophischen Überzeugungen, Gesundheit, Sexualität und Gewerkschaftszugehörigkeit.

Häufig ist es nicht sinnvoll (und manchmal auch schwer möglich), solche Daten komplett zu löschen. „Maskierung" meint dann häufig, dass die betreffenden Werte „verunschärft" werden. Im Falle von Metadaten können etwa statt des genauen Geburtsdatums nur das Geburtsjahr, statt des Wohnorts nur die Region festgehalten werden etc. In Transkripten können an den betreffenden Stellen statt Echtnamen geeignete Ersetzungen verwendet werden. Es empfiehlt sich, dafür eine konsistente Systematik anzuwenden, also z. B. immer Sprechercodes (wie W201 für „Nora Bayer") oder Pseudonyme (wie „Hans Müller" statt „Kurt Meyer") einzusetzen, so dass der gleiche Name immer durch die gleiche Ersetzung repräsentiert ist. In den Maskierungsrichtlinien zu FOLK wird zusätzlich versucht, „prosodische, insbesondere rhythmische Eigenschaften von Beiträgen zu erhalten, [...] Merkmale wie ethnische oder regionale Zugehörigkeit (z. B. türkische Namen), Kosenamen, Abkürzungen, [...] in die Ersetzung [zu] übertragen sowie [s]oziologisch relevante Merkmale wie z. B. Prestige, Status und Bildungsstand, die mit Berufen verbunden sind, [...] in der Maskierung [zu bewahren]" (vgl. Reineke et al. 2017 und Deppermann 2008: 31). Im konkreten Fall bedeutet dies beispielsweise, dass ein zu verwendendes Pseudonym die gleiche Silbenzahl wie der Echtname haben und sich auch bzgl. der (ggf. stereotypen) sozialen Zuschreibung nach dem zu ersetzenden Namen richten sollte – „Emil" ist demnach ein besseres Pseudonym für „Otto" als „Sebastian", „François" ein besseres Pseudonym für „Mathieu" als „Harry". In Audioaufnahmen werden die betreffenden Stellen üblicherweise durch eine Stille, ein Rauschen oder einen Piepton ersetzt. FOLKER und der EXMARaLDA Partitur-Editor unterstützen diesen Prozess technisch, indem sie erlauben, die betreffenden Stellen zu markieren und eine maskierte Version des Audios automatisch zu generieren.

Darüber hinaus gehende Maskierungen können in einer Verfremdung des Audiosignals und/oder im Verpixeln oder Grafisieren von Bestandteilen des Videobildes bestehen. Dadurch kann dem Umstand Rechnung getragen werden, dass auch Stimme und Gesicht eines Menschen grundsätzlich als personenbeziehbare Daten zu werten sind. Da durch eine solche Manipulation der Primärdaten der Nachnutzungswert eines Korpus aber deutlich eingeschränkt wird

(und die betreffenden Arbeitsschritte, vor allem bei Videos, auch recht aufwändig sind), sollte sie nach Möglichkeit vermieden werden, indem bei der informierten Einwilligung darauf hingewiesen wird, dass Stimme und Gesicht in den Daten erkennbar bleiben.

0768	**HB**	°h auf alle fälle ((schmatzt)) °hhh sacht die äh h° **johanna** °hhh ja hier in_e **lenne** da seh sieht (.) sieht man sich nich °h und die äh ((schmatzt)) ((stöhnt)) wie heißt der die heißen ja beide **jonas** °h unser **jonas** und hier **kesslers jonas** °hh äh die haben sich ja in berlin alle gesehn mit **jürgen** und so un [da sach ich] ach da musst ich aber lach[en ne]

Abb. 8: Beispiel aus FOLK mit Maskierungen von direkt personenbeziehbaren Daten (fett gedruckt). Dazu gehören Vor- und Nachnamen von Personen (hier mit den Pseudonymen johanna, jonas, jürgen und kesslers maskiert) und der Name einer kleineren Ortschaft (hier mit dem Pseudonym lenne maskiert), nicht aber Berlin als Ortsname. In der zugehörigen Audiodatei wurden die betreffenden Passagen mit einem Braunschen Rauschen maskiert. [https://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00342_SE_01_T_01_DF_01&cID=c768&wID=c768]

4 Bereitstellung

Gesprochene Lernerkorpora werden in aller Regel in zeitlich befristeten Einzelprojekten erhoben und erschlossen und dort in Bezug auf die projektspezifischen Fragestellungen ausgewertet. Dieselben Daten lassen sich aber auch für andere Fragestellungen fruchtbar machen, auf andere Arten erschließen oder mit anderen Korpora kombinieren. Es ist daher wünschenswert und wird von der wissenschaftlichen Community und von Forschungsförderern auch zunehmend als gute Praxis gefordert, dass einmal erhobene Daten nach Abschluss eines Projekts der Forschungsgemeinschaft zur Nachnutzung bereitgestellt werden. Für das Erhebungsprojekt kann die Bereitstellung der Daten als eigenständige Forschungsleistung gewertet werden, und es ist oft auch der Verbreitung der Forschungsergebnisse zuträglich, wenn neben regulären Publikationen auch deren Datengrundlage für andere WissenschaftlerInnen zugänglich ist.

Die Bereitstellung und dauerhafte Archivierung von Forschungsdaten sind Aufgaben, die von Einrichtungen übernommen werden müssen, die über dafür notwendige technische Expertise, personelle Kapazitäten und längerfristige Finanzierungsperspektiven verfügen. In den letzten Jahren sind, u. a. in Forschungsdateninfrastruktur-Initiativen wie CLARIN, eine Reihe von Datenzentren

entstanden, die diese Aufgaben für unterschiedliche Forschungsdaten aus den Sprachwissenschaften übernehmen. In Deutschland sind das Bayerische Archiv für Sprachsignale (BAS) in München, das Hamburger Zentrum für Sprachkorpora (HZSK) und das Archiv für Gesprochenes Deutsch (AGD) am Leibniz-Institut für Deutsche Sprache in Mannheim die wichtigsten Datenzentren mit einer Spezialisierung auf mündliche Daten.[5]

Um ein Korpus aus einem Projekt an eines dieser Datenzentren zu übergeben, empfehlen sich eine frühzeitige Kontaktaufnahme und eine Vereinbarung, die Modalitäten der Übergabe für alle Beteiligten verbindlich festhält. Das AGD bietet zum Beispiel ein Kooperationsmodell an, in dem Projekte schon bei der Antragstellung zu Fragen des Datenmanagements beraten und im Projektverlauf durch Schulungen und technischen Support begleitet werden. Damit wird erreicht, dass Daten konform zu den oben ausgeführten Regeln guter Praxis erhoben und erschlossen, der Aufwand, sie für eine Bereitstellung aufzubereiten minimiert, und ihr Nachnutzungswert optimiert werden.

Im einfachsten Falle werden Korpora von Datenzentren als Dateien zum Download bereitgestellt. Darüber hinausgehende Möglichkeiten entstehen, wenn ein Korpus zusätzlich in eine geeignete Web-Plattform integriert wird, wie es z. B. für viele im AGD archivierte Korpora der Fall ist, die nach entsprechender Aufbereitung auch über die Datenbank für Gesprochenes Deutsch (DGD) recherchiert werden können.

Die Übersicht (Tab. 8) zeigt, in welchen Suchinterfaces die gesichteten Korpora durchsuchbar sind sowie ob (und ggf. wo) die Korpusdaten herunterladbar sind.

5 Auswertungen

Für die Auswertung von gesprochenen Lernerkorpora gelten dieselben methodischen Herausforderungen und Möglichkeiten wie für Lernerkorpora im Allgemeinen. Vgl. Granger (2008) für eine Einführung in die Verwendung von Lernerkorpora in der Forschung, aber unbedingt auch die diesbezüglichen Ausführungen in Wisniewski aus diesem Heft (hier wird auch auf neuere Entwicklungen in der Methodologie eingegangen).

5 Mit dem Ende der Finanzierung der CLARIN-D- und CLARIAH-DE-Projekte und der anlaufenden Ausgestaltung der Nationalen Forschungsdateninfrastruktur (NFDI) kann sich diese Situation aktuell schnell und grundlegend ändern, also z. B. existierende Datenzentren ihre Dienste einstellen und an andere Zentren übertragen oder sich neue Zentren formieren.

Tab. 8: Übersicht zur Verfügbarkeit verschiedener aktueller gesprochener Lernerkorpora

Korpus	Suchinterface	Offline-Verfügbarkeit der Korpusdaten?
BeMaTaC (L2-Kohorte)	ANNIS (https://hu-berlin.de/annis; freier Zugang)	https://hu.berlin/bematac-download
GeWiss (L2-Kohorte)	Gewiss-Plattform Leipzig DGD (https://dgd.ids-mannheim.de/; Zugang nach Registrierung)	über AGD (http://agd.ids-mannheim.de/korpus_index.shtml#2) oder DGD (https://dgd.ids-mannheim.de/)
HaMaTaC	Mit EXAKT nach Download Geplant auch für DGD	Über HZSK (https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:hamatac) bzw. FDM der Universität Hamburg (www.fdr.uni-hamburg.de/)
HaMoTiC	Mit EXAKT nach Download Geplant auch für DGD	Über HZSK (https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:hamotic) bzw. FDM der Universität Hamburg (www.fdr.uni-hamburg.de/)
IFCASL	derzeit keine Angabe	In Planung: Bayerisches Archiv für Sprachsignale (BAS; www.phonetik.uni-muenchen.de/Bas/) Samples: www.ifcasl.org/corpus.html
Leap	derzeit keine Angabe	https://sourceforge.net/projects/leapcorpus/
MULTILIT	derzeit keine Angabe	In Planung

Die erste Anforderung für die Nutzerinnen eines Korpus bei der Verwendung von Korpora für die eigene Forschung ist, den Aufbau der zur Verfügung stehenden Korpusdaten nachzuvollziehen. Hierzu dienen Beschreibungen (Handbücher, Annotationsrichtlinien, Dokumentationsseiten usw.), die mit Korpora veröffentlicht werden müssen, sofern das Korpus nutzbar sein soll. Im Fall von online Verfügbaren Korpora gilt es, die Anfragesprache des jeweiligen Suchinterfaces, in dem das Korpus verfügbar ist, kennenzulernen und die für das jeweilige Forschungsanliegen passenden Suchanfragen zu formulieren. In Hirschmann (2019) sind die Anfragesprachen der in Tab. 8 genannten Online-Suchsysteme (ANNIS und DGD) detailliert beschrieben.

Um zu zeigen, wie die bislang behandelten Verarbeitungsschritte in der Praxis empirisch fundierter Spracherwerbsforschung Anwendung finden, bietet sich ein methodischer Blick auf die Forschungsbeiträge in diesem Heft an. Da in Wisniewski (dieses Heft) bereits eine inhaltliche Zusammenfassung der Forschungsbeiträge erfolgt, wollen wir hier auf Doppelungen verzichten und uns auf die methodischen Aspekte konzentrieren, die entweder bestimmte Beiträge besonders herausstellen, durch die sich die Beiträge grundlegend unterscheiden oder die als allgemeine Trends zu interpretieren sind. Wir polarisieren dabei jeweils auf bestimmte grundlegende methodische Entscheidungen, die einander

ausschließen. Grundsätzlich lässt sich festhalten, dass die einzelnen Erhebungen und die daraus resultierenden Sprechertypen, die die Korpora beinhalten bzw. abbilden, sowie das Korpusdesign einen enormen Einfluss auf methodische Möglichkeiten besitzen. In den Projekten, die hinter den Erhebungen und veröffentlichten Korpora stehen, wurden die Korpusdaten in vorbildlicher Weise auf bestimmte Forschungsziele ausgerichtet. Dies spiegelt sich auch in den folgenden Entscheidungen bei der Auswertung der Korpora wider. Wir fassen auf diese Weise die Beiträge Trouvain, Karges et al., Belz & Odebrecht, Weiss et al. und Fandrych & Wallner aus diesem Heft vergleichend zusammen.

CIA (Contrastive Interlanguage Analysis: kontrastive Analysen zwischen Kohorten) vs. EA (Error Anaysis: Analyse von Abweichungen)

In der Methodologie für die Auswertung von Lernerkorpora (vgl. z. B. Granger 2008) wird zwischen diesen beiden Perspektiven polarisiert: Lernerkorpusdaten können hinsichtlich beliebiger linguistischer Merkmale quantitativ mit erstsprachlichen Vergleichsdaten oder anderen Lernerdaten verglichen werden (CIA) oder hinsichtlich Abweichungen von gewissen Standards oder Normen (EA; traditionell: „Fehleranalyse“) analysiert werden. In den Forschungsbeiträgen werden beide Perspektiven eingenommen und in einem Fall (Trouvain) kombiniert, den wir deshalb genauer betrachten wollen: Die von Trouvain beschriebenen Auswertungen stützen sich (u. a.) auf die Annotation von phonetischen Abweichungen im IFCASL-Korpus, so z. B. auf nicht zielsprachliche Realisierungen von Wörtern mit [h] im Silbenonset bei französischen Lernenden des Deutschen. Hier gibt es zwei Realisierungsvarianten: stumm und mit Glottalplosiv. Auch Abweichungen in der Vokalqualität und -quantität werden gemessen. Durch das Korpusdesign (wie bereits beschrieben, enthält das Korpus bidirektionale, parallele Artikulationsdaten von Deutsch- und FranzösischsprecherInnen mit jeweils demselben Inhalt) lassen sich die Abweichungen auf zweierlei Weise beschreiben: Im Fall der [h]-Alternation können alle Fälle berücksichtigt werden, in denen das [h]-Zielwort ohne [h] oder mit Glottalplosiv artikuliert wurden. Dies entspricht einem klassischen EA-Verfahren. Man kann aber auch mit dem CIA-Verfahren feststellen, dass die französischen ErstsprachlerInnen weniger Wörter mit silbeninitialen [h] verwenden bzw. dass bei Stellen, an denen deutsche ErstsprachlerInnen im Silbenanlaut [h] artikulieren, die französischen DeutschlernerInnen dies teilweise nicht tun. Eine weitere CIA-Perspektive wird durch Analysen eröffnet, die Aussprachereализierungen von Lernenden unterschiedlichen Sprachstands miteinander vergleichen und somit Aussagen zu Erwerbsverläufen zulassen.

Intra- vs. interindividuelle Variation und Gruppentrends vs. Gruppenvarianz

Je nach Forschungsziel steht in Spracherwerbsstudien entweder die sprachliche Variation bei denselben Personen (intraindividuelle Variation) oder aber die sprachliche Variation über verschiedene Personen hinweg (interindividuelle Variation) im Vordergrund. Klassischerweise bevorzugen Studien zur Sprachentwicklung die Analyse intraindividueller Variation an verschiedenen Zeitpunkten, um eben gerade interindividuelle Effekte auszuschließen. Hierbei wird mit Bezug auf die Korpuszusammensetzung auch von (echt) longitudinalen Daten gesprochen. CIA-Studien, die gerade systematische Abweichungen im Verhalten von verschiedenen Sprechergruppen aufzeigen wollen, schauen auf interindividuelle Variation.

Gruppentrends werden durch Mittelwerte ermittelt: Die gemittelte Frequenz eines Merkmals bzw. Phänomens über eine gesamte Gruppe hinweg, zeigt eine bestimmte Tendenz für diese Gruppe auf. Häufig wird vernachlässigt, dass aber die Frequenz des gemessenen Phänomens in den einzelnen in der Gruppe enthaltenen Texten bzw. bei den verschiedenen ProbandInnen extrem von dem Mittelwert abweichen kann, und zwar in beide Richtungen. Dem kann abgeholfen werden, indem die Varianz innerhalb der Gruppe (hier als Gruppenvarianz bezeichnet) ermittelt und dargestellt wird. Dass es sehr sinnvoll und linguistisch überaus informativ sein kann, alle diese Perspektiven zu mischen, zeigen verschiedene Beiträge; wir heben hier Karges et al. hervor: Hier werden (u. a.) verschiedene Maße bei Lernenden des Deutschen als Fremdsprache angewendet, die Aufschluss über das Sprachniveau der Lernenden geben können. Die vordergründige Fragestellung ist, inwieweit außersprachliche Faktoren – die Sprachmodalität (mündliche vs. schriftliche Produktionen) und verschiedene Aufgabentypen – Effekte bei denselben Lernenden bei denselben Erwerbsniveaustufen hervorrufen. Es wird gezeigt, dass die Modalität und der Aufgabentyp einen wesentlichen Einfluss auf die Messergebnisse haben, dass also die durch Aufgaben- und Modalitätseffekte bedingte intraindividuelle Variation bei Sprachstanduntersuchungen unbedingt mitberücksichtigt muss. Gleichzeitig wird auch immer die interindividuelle Varianz, nämlich die Streuung der Messergebnisse innerhalb der Gruppe von Individuen bei derselben Verteilung der übrigen Parameter (gleiche Aufgabe, gleiche Modalität) aufgezeigt. Im Ergebnis herrscht auch hier häufig eine starke Varianz. Hierbei wird ersichtlich, dass zusätzlich zu Gruppentrends auch immer die Heterogenität der Gruppe mitbeleuchtet werden sollte.

Bezugseinheiten der Messung

Wenn Korpora auf linguistische Phänomene hin ausgewertet werden, ist je nach Forschungsfrage nicht unerheblich, welche Einheiten im Korpus den Rahmen

für die Messung ausmachen. Im einfachsten Fall wird einfach über das gesamte Korpus gemittelt, womit das Korpus selbst der Rahmen ist. Moderne Korpora beinhalten aber nicht einfach fortlaufende Sprachdaten, sondern verschiedene, mehr oder weniger linguistisch motivierte Abschnitte. So sind bspw. üblich, dass Korpora nach einzelnen Dokumenten, Texten, Gesprächen usw. organisiert sind, die gemeinsam, aber auch relativ unabhängig voneinander berücksichtigt werden können. Dass man darüber hinaus die in der Auswertung berücksichtigten Korpuseinheiten durch Annotationen im Grunde beliebig flexibel organisieren kann und dass dies für die Auswertung selbst entscheidend relevant sein kann, beleuchten Belz und Odebrecht: In ihrer Studie zu registerabhängigen Verteilungen von Flüssigkeitsmaßen in Dialogdaten von Deutschlernenden zeigen sie u. a., dass es einen entscheidenden Unterschied bewirken kann, ob die gesamten Dialoge oder bestimmte Dialogausschnitte bei der Auswertung berücksichtigt werden.

Distanz- vs. textnahe Auswertungen (distant reading vs. close reading)
Korpora sind darauf ausgelegt, über die Gesamtheit gesammelter Stichproben (ganze Korpora) hinweg Aussagen zu generieren, die im Idealfall über die Stichprobe (das Korpus) hinaus Gültigkeit haben. Dies kann in der Auswertung und ihrer Darstellung in Publikationen dazu führen, dass einzelne Texte, Textpassagen usw. nicht mehr in Erscheinung treten. Diese Analysemethode kann man als Distanzauswertung bzw. distant reading bezeichnen. Natürlich können Korpora (ggf. in Abhängigkeit von ihrer Größe) aber auch ganz gegensätzlich verwendet und im Extremfall „durchgelesen“ werden. Dies kann man als textnahe Auswertung oder close reading bezeichnen. Dass beide Perspektiven absolut sinn- und wertvoll sein können, wird durch die Gegenüberstellung der Beiträge Weiss et al. – einem Beispiel, das zum Distanz-Ansatz passt – und Fandrych & Wallner – einem Beispiel für textnahe Auswertungen – ersichtlich. Obwohl im erstgenannten Beitrag zur Untersuchung sprachlicher Komplexität bei Lehrenden und Lernenden in verschiedenen deutschen Schultypen ausschließlich Kategorien sprachlicher Komplexität behandelt werden, ohne sprachliche Belege anzuführen, und auf der anderen Seite in Fandrych & Wallner bei einer Analyse stilistischer und funktionaler Merkmale in der Wissenschaftssprache Deutsch bei fortgeschrittenen Lernenden sehr exemplarisch Textpassagen behandeln und genau diese Beispiele besprechen (ohne weitergehend zu quantifizieren), erachten wir beide Beiträge linguistisch gleichermaßen aufschlussreich.

6 Herausforderungen

Noch mehr als bei ihren erstsprachlichen Pendants muss für gesprochene Lernerkorpora des Deutschen festgestellt werden, dass aktuell nur relativ wenige Korpora öffentlich verfügbar sind und diese mehrheitlich auch nur geringe Datenmengen umfassen, teilweise auch aus technischer Perspektive nicht auf dem aktuellen Stand oder unzureichend dokumentiert sind. Aktuelle Entwicklungen zeigen aber, dass die hier angesprochenen technisch/methodischen Aspekte zunehmend (wieder) ernstgenommen werden und die Verbreitung guter Praktiken der Datenerhebung, -erschließung, -archivierung und -weitergabe das Potential besitzt, die empirische Basis für die Lernersprachforschung mittelfristig deutlich zu verbessern. Wünschenswert wäre, dass beim Aufbau entsprechender Datenbasis zukünftig mehr auf die technische und methodische Vergleichbarkeit einzelner Korpora geachtet würde. Wenn Lernerkorpora sich aus einem gemeinsamen Inventar von Korpusdesign-Elementen (z. B. vergleichbare Elizitations-Aufgaben) bedienen und die für die Erschließung verwendeten Werkzeuge technisch interoperabel sind, ergibt sich daraus auch eine Möglichkeit, Korpora aus verschiedenen Quellen zu „poolen“ und damit der von Goschler & Stefanowitsch (2014) bemängelten „Datenarmut“ (s. o.) entgegenzutreten. Langfristig könnte auf diese Weise ein Referenzkorpus der mündlichen Lernersprache entstehen, das gänzlich neue Perspektiven für die korpusbasierte Lernersprachenforschung eröffnet. Trotz der Fokussierung auf mündliche Lernerkorpora in diesem Heft darf nicht vergessen werden, dass viele Fragestellungen zur Mündlichkeit die Vergleichsgröße schriftlicher Daten erfordern, die im Fall von Lernerkorpora in deutlich höherem Maße verfügbar sind. Eine Zusammenführung der bereits vorliegenden schriftlichen Lernerkorpora mit den allmählich wachsenden mündlichen Datenmengen ist gewiss eine eigene Herausforderung und könnte sich zumindest teilweise am Referenzkorpus des (erstsprachlichen) Deutschen „DeReKo“ (Kupietz et al. 2018), das ursprünglich ausschließlich schriftliche und seit einigen Jahren zunehmend mündliche Daten beinhaltet, orientieren. Gleichzeitig sollten die Vorteile mittelgroßer Korpora, manuelle Analysen erstellen und die Korpusdaten händisch überprüfen und im Gesamten sichten zu können, nicht vernachlässigt und die Möglichkeiten statistischer Analysen auf solchen Korpusdaten weiter ausgelotet werden.

Literatur

Baur, Benedikt/Gräfe, Karen/Schmidt, Julia (2014): Dokumentation zur Annotation der Diskurskommentierungen. (https://gewiss.uni-leipzig.de/fileadmin/documents/Annotationsdokumentation_GeWiss.pdf)

Belz, Malte (2014): Richtlinien zur Annotation von Reparaturen in BeMaTaC. Technischer Bericht. Humboldt-Universität zu Berlin. (https://hu.berlin/bematac-guidelines)

Boersma, Paul (2001): Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341–345.

Deppermann, Arnulf (2008): Gespräche analysieren. Eine Einführung. 4. Aufl. Wiesbaden: Verlag für Sozialwissenschaften.

Clahsen, Harald/Meisel, Jürgen M. & Pienemann, Manfred (1983): Deutsch als Zweitsprache: Der Spracherwerb ausländischer Arbeiter. Tübingen: Narr

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (Hg.; 2017): Gesprochene Wissenschaftssprache – digital: Verfahren zur Annotation und Analyse mündlicher Korpora. Tübingen: Stauffenburg.

Goschler, Juliana & Stefanowitsch, Anatol (2014): Korpora in der Zweitspracherwerbsforschung: Sieben Probleme aus korpuslinguistischer Sicht

Gräfe, Karen/Lange, Daisy/Sieradz, Magda/Meißner, Cordula/Slavcheva, Adriana (2015): Gewiss. Handbuch zum Korpus (https://gewiss.uni-leipzig.de/fileadmin/documents/Handbuch.pdf)

Granger, Sylvaine (2008): Learner corpora. In: Lüdeling, A. & Kytö, M. (Hg.): Corpus Linguistics. An International 5 Handbook. Volume 1. Berlin & New York: De Gruyter, 259–275

Gut, Ulrike (2014): The Leap Corpus. In: Durand, Jacques; Gut, Ulrike; Kristoffersen, Gjert (Hg.) The Oxford Handbook of Corpus Phonology. Oxford; Oxford University Press.

Gut, Ulrike & Bayerl, Petra S. (2004): Measuring the Reliability of Manual Annotations of Speech Corpora. In: *Proceedings of Speech Prosody 2004*, Nara, Japan.

Hedeland, Hanna & Schmidt, Thomas (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. In: Schmidt, Thomas & Wörner, Kai (eds.): Multilingual Corpora and Multilingual Corpus Analysis. Hamburg Studies on Multilingualism (14). Amsterdam: Benjamins, 25–46. https://doi.org/10.1075/hsm.14.04hed

Hirschmann, Hagen (2019): Korpuslinguistik. Eine Einführung. Stuttgart; Metzler.

Imo, Wolfgang & Weidner, Beate (2018): Mündliche Korpora im DaF- und DaZ-Unterricht. In: Kupietz, M. & Schmidt, T. (Hg.): Korpuslinguistik. Band 5 der Reihe Germanistische Sprachwissenschaft um 2020. Berlin & Boston: De Gruyter, 231–253.

Kisler, Thomas; Reichel, Uwe D.; Schiel, Florian (2017): Multilingual processing of speech via web services. In: *Computer Speech & Language* (45), 326–347.

Kleiner, Stefan; Berend, Nina; Brinckmann, Caren; Knöbl, Ralf (2011): „Deutsch Heute". Ein sprachgebietsweites Forschungsprojekt zur regionalen Variation in der gesprochenen deutschen Standardsprache. In: *Klagenfurter Beiträge zur Sprachwissenschaft* 34–36, S. 179193. https://ids-pub.bsz-bw.de/files/2874/Kleiner_Berend_Brinckmann_Kn%C3%B6bl-Deutsch_heute_2011.pdf

Kupietz, Marc; Lüngen, Harald; Kamocki, Paweł; Witt, Andreas (2018): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis,

Stelios/Tokunaga, Takenobu (eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki: European Language Resources Association (ELRA), 2018. S. 4353–4360. www.lrec-conf.org/proceedings/lrec2018/pdf/737.pdf

Lüdeling, Anke & Hirschmann, Hagen (2015): Error annotation systems. In: Granger, Sylviane; Gilquin, Gaëtanelle; Meunier, Fanny (Hg.): The Cambridge Handbook of Learner Corpus Research. Cambridge; Cambridge University Press, 135–158.

Lüdeling, Anke; Hirschmann, Hagen; Shadrova, Anna & Wan, Shujun (2021): Tiefe Analyse von Lernerkorpora. In: IDS Jahrbuch 2020.

MacWhinney, Brian (2000): The CHILDES project: Tools for analyzing talk. 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Meißner, Cordula & Adriana Slavcheva (2014): Ein Vergleichskorpus der gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen. Zum Design und Aufbau des GeWiss-Korpus. In: Fandrych, Christian; Meißner, Cordula; Slavcheva, Adriana (Hg.): Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen. Heidelberg: Synchron, 15–38.

Mukherjee, Joybrato (2009): Anglistische Korpuslinguistik. Eine Einführung. Berlin; Erich Schmidt Verlag.

Nolda, Andreas (2019): Annotation von Lernerdaten mit EXMARaLDA (Dulko). Technischer Bericht. (https://andreas.nolda.org/publications/nolda_2019_annotation_lernerdaten.pdf)

Ochs, Elinor (1979): Transcription as theory. In: Ochs, E. & Schieffelin, B. (Hg.) Developmental Pragmatics, New York: Academic Press, 43–72.

RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2020): Handreichung Datenschutz. 2. vollständig überarbeitete Auflage. RatSWD Output 8 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.50

Rehbein, Jochen/Schmidt, Thomas/Meyer, Bernd/Watzke, Franziska/Herkenrath, Annette (2004): Handbuch für das computergestützte Transkribieren nach HIAT. Arbeiten zur Mehrsprachigkeit: Folge B, Sonderforschungsbereich 538 (56). https://nbn-resolving.org/urn:nbn:de:bsz:mh39-23681

Reineke, Silke/Schmidt, Thomas/Schedl, Evi/Kaiser, Julia (2017): Maskierung von Audio- und Videoaufnahmen. Version 2.1, Gesprächsanalytisches Informationssystem: Überarbeitung und Ergänzung. (http://prowiki.ids-mannheim.de/pub/GAIS/MasKierung/Maskierung_von_Audio_und_Videoaufnahmen_2.1_GAIS.pdf)

Reznicek, Marc/Anke Lüdeling/Hagen Hirschmann (2013): Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. Díaz-Negrillo, Ana; Ballier Nicolas; Thompson, Paul (Hg.): Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: Benjamins, 101123.

Reznicek, Marc/Lüdeling, Anke/Krummes, Cedric/Schwantuschke, Franziska/Walter, Maik/Schmidt, Karin/Hirschmann, Hagen/Andreas, Torsten (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Technischer Bericht. Humboldt-Universität zu Berlin. (https://hu.berlin/falkohandbuch)

Rohlfing Katharina/Loehr Daniel/Duncan Susan/Brown Amanda/Franklin Amy/Kimbara Irene/Milde, Jan-Torsten/Parrill, Fay/Rose, Travis/Schmidt, Thomas/Sloetjes Han (2006): Comparison of multimodal annotation tools. In: *Gesprächsforschung* (7), 99–123.

Sauer, Simon & Lüdeling, Anke (2016): Flexible Multi-Layer Spoken Dialogue Corpora. In: *International Journal of Corpus Linguistics* 21, 419–438.

Schellhardt, Christin & Schroeder, Christoph (Hg., 2015): MULTILIT. Manual, criteria of transcription and analysis for German, Turkish and English. Tech. Report. Univertät Potsdam. https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/8039/file/multilit_manual.pdf

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht. Institut für maschinelle Sprachverarbeitung, Stuttgart. (www.sfs.uni-tuebingen.de/resources/stts-1999.pdf)

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using Decision Trees.In: Proceedings of the International Conference on New Methods in Language Processing. (/www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf)

Schmidt, Thomas (2005): Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a.M.: Peter Lang.

Schmidt, Thomas (2016): Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. International Journal of Corpus Linguistics (21/3), 396–418. https://doi.org/10.1075/ijcl.21.3.05sch

Schmidt, Thomas (2017): Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. In: *Corpus Linguistic Software Tools, Journal for Language Technology and Computational Linguistics* (JLCL 31/1), by Kupietz, Marc & Geyken, Alexander (Hrsg.), S. 127–154.

Schmidt, Thomas/Duncan, Susan/Ehmer, Oliver/Hoyt, Jeffrey/Kipp, Michael/Loehr, Dan/Magnusson, Magnus/Rose, Travis/Sloetjes, Han (2009): An exchange format for multimodal annotations. In: Kipp, Michael/Martin, Jean-Claude/Paggio, Patrizia/Heylen, Dirk (Hg.): Multimodal corpora: from models of natural interaction to systems and applications. Berlin/Heidelberg: Springer, 2009. S. 207–221.

Schmidt, Thomas/Schütte, Wilfried (2010): FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odjik, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10), may 19–21, 2010, Valletta, Malta. European Language Resources Association (ELRA), 2010. S. 2091–2096.

Schmidt, Thomas/Wörner, Kai (2014): EXMARaLDA. In: Jacques Durand, Ulrike Gut, and Gjert Kristoffersen (Hg.): The Oxford Handbook of Corpus Phonology. Oxford: OUP 2014, S. 402–419.

Schmidt, Thomas/Winterscheid, Jenny/Schütte, Wilfried (2015): cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2). Mannheim: Institut für Deutsche Sprache. [https://nbn-resolving.org/urn:nbn:de:bsz:mh39-46169]

Selting, Margret/Auer, Peter/Barden, Birgit/Bergmann, Jörg R./Couper-Kuhlen, Elizabeth/Günthner, Susanne/Meier, Christoph/Quasthoff, Uta M./Schlobinski, Peter/Uhmann, Susanne (1998): Gesprächsanalytisches Transkriptionssystem (GAT). In: *Linguistische Berichte* 173. S. 91–122. (http://www.mediensprache.net/de/medienanalyse/transcription/gat/gat.pdf)

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg R./Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian / Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stukenbrock, Anja/Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT

2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, S. 353–402. (http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf)

Stemle, Egon, Boyd, Adriane, Janssen, Maarten, Lindström Tiedemann, Therese, Mikelić Preradović, Nives, Rosen, Alexandr, Rosén, Dan & Volodina, Elena (2019):: Working together towards an ideal infrastructure for language learner corpora. In: Andrea Abel, Aivars Glaznieks, Verena Lyding & Lionel Nicolas (Hg.) *Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference*. Corpora and Language in Use – Proceedings 5, Louvain-la-Neuve: Presses universitaires de Louvain, 427–468.

Trouvain, Jürgen; Bonneau, Anne; Colotte, Vincent; Fauth, Camille; Fohr, Dominique et al. (2016): The IFCASL Corpus of French and German Non-native and Native Read Speech. In: *Proceedings of LREC'2016*, May 2016, Portorož, Slovenia, 1333–1338. www.coli.uni-saarland.de/~juegler/Publications/trouvain_etal_lrec2016.pdf

Trouvain, Jürgen; Fauth, Camille; Möbius, Bernd (2016): Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In: *Proceedings of Speech Prosody* (SP8), Boston, 31–35.

Wells, John C. (1997): SAMPA computer readable phonetic alphabet. In: Gibbon, D.; Moore, R.; Winski, R. (Hg.) *Handbook of Standards and Resources for Spoken Language Systems*. Berlin/New York; Mouton de Gruyter.

Westpfahl, Swantje; Schmidt, Thomas; Jonietz, Jasmin; Borlinghaus, Anton (2017): STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Arbeitspapier. Mannheim: Institut für Deutsche Sprache. (http://nbn-resolving.de/urn:nbn:de:bsz:mh39-60634)

Winterscheid, Jenny; Deppermann, Arnulf; Schmidt, Thomas; Schütte, Wilfried; Schedl, Evi; Kaiser, Julia (2019): Normalisieren mit OrthoNormal. Konventionen und Bedienungshinweise für die orthografische Normalisierung von FOLKER-Transkripten. Mannheim: Leibniz-Institut für Deutsche Sprache. (https://doi.org/10.14618/ids-pub-9326)

Wittenburg, Peter/Brugman, Hennie/Russel, Albert/Klassmann, Alex/Sloetjes, Han (2006): ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.