Handreichung:

Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora

Inhalt

Vorbemerkungen

Die vorliegende Handreichung wurde in themenbezogenen Arbeitsgruppen aus Fachwissenschaftlern der Korpuslinguistik und der korpuslinguistisch arbeitenden Einzeldisziplinen erarbeitet. Die AGs konstituierten sich in zwei DFG-Rundgesprächen, die 2012 und 2013 auf Initiative des FK Sprachwissenschaften unter Federführung von Arnulf Deppermann und Mechthild Habermann sowie Helga Weyerts-Schweda von der DFG durchgeführt wurden.

Das Rundgespräch zu mündlichen Korpora wurde von Arnulf Deppermann und Thomas Schmidt (IDS Mannheim) organisiert und fand am 9. November 2012 in der DFG-Geschäftsstelle in Bonn statt. Die Mitglieder der AG, die die Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von mündlichen Korpora (Teil 1 der vorliegenden Handreichungen) erarbeitet haben, sind Bernt Ahrenholz, Arnulf Deppermann, Sebastian Drude, Christian Fandrych, Ulrike Gut, Stefan Pfänder und Thomas Schmidt. Die Koordination erfolgte durch Thomas Schmidt.

Das entsprechende Rundgespräch zu Schriftkorpora fand am 15. November 2013 ebenfalls bei der DFG statt und wurde von Alexander Geyken (BBAW Berlin) und Marc Kupietz (IDS Mannheim) organisiert. Die Mitglieder der drei AGs für die Formulierung der Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Schriftkorpora (Teil 2 dieser Handreichungen) waren Noah Bubenhofer, Dagmar Deuber, Eva-Maria Dickhaut, Thomas Gloning, Iryna Gurevych, Mechthild Habermann, Ulrich Heid, Gerhard Heyer, Erhard Hinrichs, Martin Huber, Wolf Peter Klein, Christian Mair, Alexander Mehler, Roland Meyer, Roland Schäfer, Ingrid Schröder, Silke Schwandt, Manfred Stede, Angelika Storrer, Elke Teich und Heike Zinsmeister. Textbeiträge oder Feedback wurden außerdem von Florian Barteld, Michael Beißwenger, Volker Boehlke, Kerstin Eckart, Thomas Eckart, Richard Eckart de Castilho, Judith Eckle-Kohler, Peter Fankhauser, Rüdiger Gleim und Jens Stegmann eingebracht. Die Koordinationsgruppe für Teil 2 bestand aus Alexander Geyken, Susanne Haaf und Christian Thomas (BBAW) sowie Marc Kupietz und Harald Lüngen (IDS).

Einleitung

Diese Handreichung gibt Empfehlungen für datentechnische Standards und Tools, die bei der Erstellung von Sprachkorpora verwendet werden sollten, um eine – nach jetzigem Kenntnisstand – optimale Archivierbarkeit und Nachnutzbarkeit der Daten zu gewährleisten. Sie geht von mehreren, vor allem in den letzten zehn Jahren formulierten, Vorschlägen für eine Vereinheitlichung verschiedener gebräuchlicher datentechnischer Methoden aus (siehe Referenzen) und leitet von diesen konkrete Empfehlungen ab, die Wissenschaftlern, die die Erstellung von Sprachkorpora planen oder mit der Bewertung entsprechender Vorhaben befasst sind, als Richtlinie dienen können. Da die Standards in vielen betroffenen Bereichen zurzeit erst im Entstehen begriffen sind und sich folglich aller Wahrscheinlichkeit nach in Zukunft weiter entwickeln werden, stellen wir den konkreten Empfehlungen jeweils die allgemeineren Überlegungen voran, aus denen sie sich motivieren.

Bei der Erstellung von *mündlichen Korpora* sind in der Regel andere Standards, Tools und Best Practices relevant als bei der Erstellung von *Schriftkorpora*, und in den meisten Projekten mit Korpusaufbau werden entweder rein mündliche oder rein schriftliche Korpora erstellt. Um zu gewährleisten, dass Benutzer dieser Handreichungen nur die für ihre Modalität relevanten Empfehlungen und Informationen zu rezipieren brauchen, besteht dieses Dokument aus zwei Teilen. Teil 1 (S. 4-16) befasst sich mit den Empfehlungen für den Aufbau von mündlichen Korpora, Teil 2 (S. 17-33) mit denen zum Aufbau von Schriftkorpora*,* wobei jeder Teil auch einen eigenen Abschnitt mit Referenzen aufweist.

Zu rechtlichen und ethischen Fragen, die sich in Zusammenhang mit der Erstellung, Nutzung und Archivierung von Sprachkorpora stellen, verweisen wir auf das ebenfalls aus den Rundgesprächen resultierende separate Dokument *Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora.*

Teil 1: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung mündlicher Korpora

Mündliche Korpora setzen sich aus Daten unterschiedlichen Typs zusammen: die eigentlichen Primärdaten werden als Audio- und/oder Videoaufnahmen erhoben; hinzu kommen gegebenenfalls Zusatzmaterialien (Handouts, Präsentationsfolien etc.), die für das Verständnis der Sprachdaten oder des kommunikativen Ereignisses relevant sind; auf der Grundlage der Aufnahmen werden Transkriptionen erstellt, die mit Abschnitten aus der Aufnahme verknüpft werden können; Transkriptionen können mit weiteren Annotationen angereichert werden; Metadaten schließlich dokumentieren die Aufnahmesituation und -umstände, Eigenschaften der beteiligten Sprecher sowie ggf. weitere für Analyse und Nachnutzung der Daten notwendige oder nützliche Informationen.
Die Abschnitte 1.1. bis 1.3 behandeln Audio-/Videoaufnahmen und Zusatzmaterialien, Transkriptionen und weitere Annotationen sowie Metadaten. Um zu garantieren, dass die Primärdaten, Annotationen und Metadaten mündlicher Korpora langfristig auch für eine Nachnutzung in anderen Projekten zur Verfügung stehen, ist ihre Archivierung zu gewährleisten. Dies wird in Abschnitt 1.4 angesprochen.

1.1. Primärdaten

1.1.1. Audio- und Videoaufnahmen

Allgemeine Empfehlungen

Audio- und Videoaufnahmen sollten heute in aller Regel **digital** (d. h. mit digitalen Aufnahmegeräten) erhoben werden, da erstens die Verwendung analoger Technik (z. B. VHS-Rekorder, Kompaktkassetten) weder günstiger ist noch zu qualitativ besseren Ergebnissen führt und zweitens eine nachträgliche Digitalisierung, die für Archivzwecke i. d. R. unerlässlich ist, üblicherweise zu Informations- und Qualitätsverlust führt.
Aufnahmen sollten mit der **bestmöglichen Aufnahmequalität**, die das eingesetzte Gerät bietet, gemacht werden. Dies bedeutet insbesondere, dass nach Möglichkeit bei der Aufnahme **keine verlustbehaftete Kompression** zum Einsatz kommen sollte. Beispielsweise ist bei einem Audio-Aufnahmegerät also ein Modus, der unkomprimierte Daten mit möglichst hoher Samplingrate speichert, einem Modus zum Speichern im MP3-Format unbedingt vorzuziehen. Weiterhin maßgeblich für die Aufnahmequalität sind die eingesetzten Mikrofone – so ist es z. B. meist empfehlenswert, eine Videokamera mit einem geeigneten externen Mikrofon auszustatten.[1] Bei der Wahl

[1] In Aufnahmesettings mit vielen Sprechern (z. B. bei Unterrichtskommunikation) kann es zum Erzielen einer ausreichenden Aufnahmequalität nötig sein, mehrere externe Mikrofone zu verwenden. Dies sollte bei der Anschaffung bedacht werden, da nicht alle Geräte im semiprofessionellen Bereich die Möglichkeit bieten, mehr als ein Mikrofon anzuschließen.

des Aufnahmegerätes sollte außerdem darauf geachtet werden, dass die zum Speichern verwendeten **Formate nicht proprietär** (d. h. an einen bestimmten Hersteller gebunden) sind oder sich zumindest problemlos und verlustfrei in nicht-proprietäre (d. h. offene) Formate überführen lassen.
Aufnahmen sollten möglichst unmittelbar nach der Erhebung vom Aufnahmegerät auf einen Rechner überspielt und ggf. in offene Standard-Formate konvertiert werden. Ein grundständiges Backup muss gewährleistet sein. Die Originalaufnahmen (Rohdaten) sollten zumindest für die Projektlaufzeit zu Archivzwecken behalten werden, auch wenn die eigentliche Transkription und Analyse anhand bearbeiteter (z. B. geschnittener oder konvertierter) Fassungen (Primärdaten) erfolgt. Erfolgt eine solche Bearbeitung, so ist es für Zwecke der Archivierung vorteilhaft, wenn der Bezug zur Originalaufnahme nachvollziehbar dokumentiert wird (z. B. “Die Primärdaten X wurden als Abschnitt 5:11 bis 20:45 von den Rohdaten Y herausgeschnitten”).

Konkrete Empfehlungen

a) Aufnahmegeräte (vgl. [2])

- Für Audio-Aufnahmen im Feld (zum Zwecke der Sprachdokumentation, Gesprächsanalyse etc.), bei denen eine einfache Handhabung vorrangig ist, haben sich **Flash-Mobilrekorder** mit wechselbaren Speicherkarten bewährt. Dabei ist darauf zu achten, dass das eingebaute Mikrofon von ausreichender Qualität ist (oder ggf. ein externes Mikrofon zu nutzen) und das Gerät ein unkomprimiertes Abspeichern der Aufnahme ermöglicht. Wenn die Größe des Aufnahmegeräts keine wichtige Rolle spielt, stellen Laptops mit entsprechender Peripherie (hochwertige Soundkarte, externes Mikrofon) und geeigneter Aufnahmesoftware (z. B. Audacity) eine sinnvolle Alternative dar, wenn ausgeschlossen werden kann, dass Eigengeräusche des Laptops (Lüftung, Festplatten u. Ä.) die Aufnahmequalität beeinträchtigen. Von der Verwendung von einfachen Diktiergeräten, Smartphones und ähnlichen Kleingeräten ist i. d. R. abzuraten, da mit solchen Geräten keine ausreichende Aufnahmequalität erzielt werden kann.
- Für Audio-Aufnahmen in Studioqualität (für Untersuchungen in der Phonetik etc.) können höherwertige Aufnahmegeräte notwendig sein.
- Es ist derzeit kaum möglich, entsprechende konkrete Empfehlungen für Video-Aufnahmegeräte auszusprechen, da hier erstens ein noch vielfältigeres Angebot besteht und sich dieses zweitens sehr schnell ändert.

b) Formate (vgl. [1], [3] und [12])

- Für Audio-Aufnahmen ist ein **unkomprimiertes lineares PCM-Format** (typischerweise als WAV gespeichert) mit dem Untersuchungszweck angemessenen Bit- und Samplingraten zu empfehlen. Für gesprochene Sprache in störungsfreien Umgebungen können 16bit/22kHz als Mindestanforderung gelten; 16bit/48kHz sind anzustreben, wenn auch die Aufnahmeumgebung angemessen repräsentiert werden soll (was bei Aufnahmen “im Feld” in der Regel der Fall ist).
- Für die meisten Zwecke sind die Video-Formate, die von den höherwertigen Amateur-Kameras auf dem Markt produziert werden (**MPEG2 oder MPEG4/H.264 mit hohen Bitraten**) von ausreichender Qualität. Sie können in ihrer Original-Fassung gespeichert und archiviert werden, bis kompressionsfreie Lösungen für die Langzeitarchivierung finanziell erschwinglich sind. Folgende Parameter können dabei als Richtwerte dienen:

    - Standard-Definition-Video (720px x 576px oder 704px x 480px) mit MPEG2-Kompression bis zu 9.8 MBit/s (typischerweise um 3.5 Mbit/s) ist geeignet für Aufnahmen im Feld;
    - High-Definition-Video (1280px x 720px oder 1920px x 1080px) mit H.264/MPEG-4 AVC-Kompression bis zu 48 MBit/s (typischerweise um 9 Mbit/s) ist geeignet für detaillierte Analysen von Gestik und Mimik.
- Unter gewissen Umständen kann auch von vorneherein eine Speicherung von Videos in unkomprimierter Form in Betracht gezogen werden. **MJPEG2000** (bei entsprechender Kodierung der Tonspur in PCM, s. o.) ist ein derzeit gängiges hierfür geeignetes Format.
- Ungeachtet dieser Empfehlungen kann es notwendig sein, Audio- oder Video-Daten für die Verwendung mit einem spezifischen Tool in einer spezifischen technischen Umgebung geeignet in andere (z. B. platzsparendere oder mit einem bestimmten Tool kompatible) Formate zu konvertieren. Solche konvertierten Daten können als Arbeitsdateien betrachtet werden. Die Versionen in den oben empfohlenen Formaten sollten als “Master-Kopien” in jedem Fall für die Archivierung und Nachnutzung aufbewahrt werden.

1.1.2. Zusatzmaterialien

Für bestimmte kommunikative Ereignisse ist es sinnvoll und empfehlenswert, für das bessere Verständnis und die Interpretation der Daten relevante Zusatzmaterialien zu sammeln, zu archivieren, mit den Sprachdaten zu verknüpfen und Korpusnutzern zugänglich zu machen (etwa Präsentationsfolien und Handouts bei Vorträgen; Tagesordnungen und Tischvorlagen bei Besprechungen; Fragebögen u.Ä., mit denen etwa Daten zu Sprachbiographien oder Sprachkompetenzen mehrsprachiger Sprecher erhoben wurden; Tafelanschrieb und andere schriftliche Materialien bei Unterrichtskommunikationen). Es wird empfohlen, bereits bei Projektbeginn die Relevanz etwaiger Zusatzmaterialien zu bedenken, entsprechende Einverständniserklärungen für die Autoren zu entwickeln, geeignete Formate für die Speicherung festzulegen und das Vorhandensein und die Art der erhobenen Zusatzmaterialien in den Metadaten zu dokumentieren. Gegebenenfalls sind spezielle rechtliche Fragen zu klären (etwa Copyright an Bildern, Graphiken).

1.2. Transkription und weitere Annotation

1.2.1. Tools und Formate

Allgemeine Empfehlungen

Für die Transkription und weitere Annotation mündlicher Daten ist in den letzten zwanzig Jahren eine Vielzahl spezialisierter Software-Tools entwickelt worden, die sowohl einer Effektivierung des Arbeitsablaufes als auch einer Verbesserung der Archivierbarkeit und Nachnutzbarkeit der entstehenden Daten dienen. Grundsätzlich sind solche spezialisierten Software-Tools der Verwendung allgemeiner Textverarbeitungssoftware für die Transkription und weitere Annotation vorzuziehen, da nur erstere die **Daten in strukturierter Form,** d. h. mit expliziter und damit computergestützt verwertbarer Auszeichnung relevanter Einheiten (Sprecher, Wörter, Äußerungen, Zeitbezüge etc.), speichern. Spezifischer können folgende Kriterien als Grundlage für die Beurteilung eines Tools bzw. eines Toolformats herangezogen werden (siehe [4]):

- **Unicode-basierte und offen dokumentierte Formate** sind anderen Formaten unbedingt vorzuziehen, da diese eine nachhaltige Datenhaltung und Archivierung ermöglichen.
- Formate, die auf einem expliziten **Datenmodell** oder auf einer expliziten **Dokumentgrammatik** basieren, vereinfachen das Verständnis und damit eine Weiterverarbeitung der Daten.
- Eine Nachnutzung der Daten wird weiterhin durch eine durchgängige Verknüpfung zwischen Transkriptionen/Annotationen und den zugrunde liegenden Audio- oder Videoaufnahmen wesentlich erleichtert. Formate, die ein solches **Alignment zwischen Aufnahme und Transkription** ermöglichen, sind daher vorzuziehen.
- Bei der Wahl eines Tools sollte möglichst auch darauf geachtet werden, dass dessen Format auch von anderen Tools gelesen werden kann bzw. **interoperabel** mit anderen gebräuchlichen Formaten ist. Bei Tools, deren Entwicklung noch aktiv ist, erhöhen sich die Chancen, dass das zugehörige Format kompatibel mit sich zukünftig herausbildenden Standards sein wird.
- Eine hohe Interoperabilität ist bei **XML-basierten Formaten** gegeben, da diese sich als allgemeine Standards für die Speicherung komplexerer Textdokumente flächendeckend etabliert haben.
- Rein oder vorwiegend präsentationsorientierte Formate (d. h. Formate, die statt der inhaltlichen Struktur der Daten nur deren visuelle Formatierung repräsentieren, wie z. B. HTML- oder MS Word-Formate) sind in aller Regel für Zwecke der Archivierung und Nachnutzung ungeeignet, weil sie sich nicht flexibel weiterverarbeiten lassen. Gleiches gilt für proprietäre Formate, d. h. solche, die nur von einem bestimmten Tool gelesen werden können.

Konkrete Empfehlungen (vgl. [4])[2]

[2] aktuell (Februar 2015) gültige URLs zu den hier empfohlenen Tools (Dokumentation und Download)

- Gemäß den obigen allgemeinen Empfehlungen können unter den derzeit weiter verbreiteten Tools zur Transkription und weiteren Annotation zumindest **ANVIL, ELAN, EXMARaLDA, FOLKER, Phon** und **Praat** als empfehlenswert gelten.
- Des Weiteren kann auch die Verwendung von **CLAN/CHAT** oder **Transcriber** mit nur wenigen Einschränkungen als empfehlenswert betrachtet werden.
- Tools wie **F4/F5** oder **Transana** erfüllen hingegen mehrere der oben aufgeführten Kriterien *nicht*. Von ihrer Verwendung ist daher abzuraten, sofern nicht durch zusätzliche Vorkehrungen im Arbeitsablauf sichergestellt wird, dass die Daten zuverlässig in ein anderes, für Nachnutzung und Archivierung besser geeignetes, Format überführt werden. Gleiches gilt für die Verwendung allgemeiner Textverarbeitungsformate wie **MS-Word** oder **OpenOffice**.
- Tools wie **atlas.ti** oder **MaxQDA** sollten in diesem Zusammenhang auf ihre Funktion als Instrumente zur qualitativen Datenanalyse beschränkt werden, da sie die oben formulierten Voraussetzungen höchstens zum Teil erfüllen; insbesondere treten beim Austausch von Daten dieser Tools mit anderen Editoren Informationsverluste auf. D. h. auch wenn die diese Werkzeuge für viele Analysezwecke notwendig und gebräuchlich sind, sollten sie im Hinblick auf eine optimale Archivier- und Nachnutzbarkeit möglichst nicht für die (Erst-)Transkription von Audio- oder Videodaten eingesetzt werden.
- Neben den Formaten selbst stellen mehrere der empfehlenswerten Tools zusätzliche **Mechanismen zur Konsistenzsicherung und transparenten Dokumentation** der Daten bereit – z. B. erlaubt ANVIL die Definition einer Spezifikationsdatei, ELAN ermöglicht eine Zuordnung der verwendeten “Linguistic Types” zu Einträgen in der ISOcat-Registry (siehe Abschnitt 3), und FOLKER beinhaltet Mechanismen zum Überprüfen der zeitlichen und syntaktischen Konsistenz von Transkriptionsdaten. Im Sinne einer Qualitätssicherung kann es, in Abhängigkeit von den angestrebten Arbeitsabläufen, sinnvoll sein, diese Mechanismen zu nutzen.

1.2.2. Transkriptionskonventionen und Annotationsschemata

Allgemeine Empfehlungen

Die Auswahl eines Transkriptionssystems und erst recht die Entscheidung, welche weiteren Annotationen für ein Korpus sinnvoll und notwendig sind, sind in besonderem Maße von projektspezifischen Gegebenheiten und Zielsetzungen abhängig. Es ist daher in diesem Bereich kaum möglich, einige wenige Lösungen als empfehlenswert herauszuheben. Dennoch können die Nachnutzbarkeit und Archivierbarkeit eines Korpus durch die Beachtung einiger allgemeiner Prinzipien auch im Hinblick auf die verwendeten Transkriptionskonventionen und Annotationsschemata deutlich verbessert werden. Insbesondere sollten projektspezifische Transkriptionskonventionen und Annotationsschemata möglichst immer in Bezug zu bereits etablierten und dokumentierten Verfahren gesetzt werden – d. h. vor der Entwicklung “projekteigener” Verfahren sollte eingehend geprüft werden, ob nicht bereits anderweitig eingesetzte Verfahren verwendet werden können. Ist dies nicht der Fall, sollten projektspezifische Konventionen und Schemata möglichst als Erweiterungen, Modifikationen oder Vereinfachungen existierender Verfahren entwickelt und entsprechend dokumentiert werden.

sind weiter unten unter “Referenzen” vermerkt

Konkrete Empfehlungen (vgl. [4])

- Für die orthographie-basierte Transkription von Spontansprache sind im deutschsprachigen Raum **GAT** [5] und **HIAT** [6] die am weitesten verbreiteten Verfahren. Daneben existiert mit **CHAT** [7] eine sehr weit verbreitete Konvention, die nicht auf den deutschsprachigen Raum beschränkt ist. Da diese Konventionen auch vergleichsweise gut (durch reguläre wissenschaftliche Publikationen) dokumentiert und in der Anwendung auf digitale Korpora erprobt sind, sollte ihre Eignung für die spezifischen Projektzwecke in jedem Falle geprüft werden. Falls dennoch ein eigenes Transkriptionssystem entwickelt wird, sollte dieses in einer für Außenstehende zugänglichen und nachvollziehbaren Form dokumentiert werden.
- Für phonetische Transkriptionen existiert mit **IPA** einer der wenigen "echten" Standards im Bereich der Linguistik. Bei der Verwendung von IPA sollte in jedem Fall eine Unicode-basierte Schriftart verwendet werden (insbesondere ist von der Verwendung von in den 1990er Jahren weit verbreiteten speziellen IPA-Schriftsätzen, bspw. vom SIL, abzuraten). Aus praktischen Gründen kann auch die Verwendung eines IPA-isomorphen ASCII-basierten Alphabets (**SAMPA** oder **X-SAMPA**) vorzuziehen sein – dies ist in datentechnischer Hinsicht unbedenklich.
- Für weiterführende Annotationen existieren teilweise De-Facto-Standards. Exemplarisch seien hier STTS (morphosyntaktische Annotation auf Wortebene), Tiger (syntaktische Annotation), SALSA (semantische Rollenannotation), GRAID (Grammatical Relations and Animacy in Discourse) und ToBi (Annotationsverfahren für Prosodie und Intonation) genannt. Sofern sich solche Verfahren für den jeweiligen Untersuchungszweck sinnvoll anwenden lassen, sollten sie ebenfalls mindestens als Ausgangsbasis herangezogen werden.
- In der Sprachdokumentation und -beschreibung sowie -typologie weit verbreitet sind Interlinearglossierungen. Der Standard hierfür sind die Leipzig Glossing Rules [LGR].
- Einen umfassenden Referenzrahmen für die Annotation mündlicher Daten zu Zwecken der Sprachdokumentation, der alle traditionellen strukturellen Ebenen (Phonetik, Phonologie, Morphologie, Syntax, Semantik) umfasst und in diesen jeweils die Einheiten, Strukturen und Relationen klar voneinander trennt, bietet Advanced Glossing (AG). Bei der Konzipierung projektspezifischer Annotationsebenen wird empfohlen, diese durch den Bezug auf AG (in ISOcat vorhanden) explizit und damit vergleichbar zu machen.

1.3. Metadaten

Allgemeine Empfehlungen

Eine sorgfältige und umfangreiche Dokumentation von Metadaten zu Gesprächsereignissen und den daran beteiligten Sprechern ist eine unablässige Voraussetzung für eine Integration mündlicher Korpora in digitale Infrastrukturen und für die Archivierbarkeit und Nachnutzbarkeit der Daten. Allgemein kann daher empfohlen werden, der systematischen Erhebung und Dokumentation von Metadaten bei der Erstellung mündlicher Korpora angemessenen Raum zuzugestehen. Dabei sollten insbesondere auch solche Metadaten von Vorneherein mit berücksichtigt werden, die für die unmittelbaren Untersuchungsinteresse nicht von Belang scheinen mögen, für die spätere Nachnutzung der Daten – auch und gerade durch Personen, die an der ursprünglichen Erhebung nicht beteiligt waren – aber unverzichtbar sind.

Für die **allgemeine Organisation von Metadaten** zu mündlichen Korpora existieren einige bewährte Datenmodelle, die in ihrer Grundstruktur weitestgehend identisch sind: sie organisieren ein Korpus als eine Menge von Sprechereignissen (Kommunikationen, Sessions), die zusammengehörige Aufnahmen, Transkriptionen und Annotationen zu einer Einheit bündeln, sowie eine Menge von Sprechern, die einem oder mehreren dieser Sprechereignisse zugeordnet sind. Teilweise wird die Erstellung und Verwaltung von Metadaten gemäß diesen Datenmodellen von zugehörigen Tools (z. B. ARBIL, EXMARaLDA Corpus Manager) unterstützt. Es empfiehlt sich, zumindest das Prinzip dieser Datenmodelle, wenn nicht das konkrete Datenmodell selbst, zu übernehmen, sofern nicht konkrete Eigenschaften des Korpus explizit dagegen sprechen.

Welche Metadaten innerhalb einer solchen Struktur konkret erhoben und dokumentiert werden, ist wiederum in hohem Maße vom spezifischen Korpusdesign und den damit verbundenen Untersuchungsinteressen abhängig. Dennoch gibt es eine Reihe von Metadaten, die generell erhebbar sein sollten und auch von generellem Interesse für die Nachnutzung eines Korpus sind. Dazu gehören Angaben zu Ort und Zeitpunkt der Aufnahmen, die zur Aufnahme verwendete Technik, Angaben zu gegebenenfalls relevanten Zusatzmaterialien, gewisse allgemeine soziobiographische (Geschlecht, Alter, Herkunft, Rolle im Gespräch) und soziolinguistische Angaben (z. B. welche Sprachen/Dialekte werden aktiv/passiv beherrscht und wann verwendet) zu den Sprechern, sowie Informationen zum datenschutzrechtlichen Status der Aufnahme.

Für die Frage, wie, d. h. mit welchen **Vokabularen**, solche Metadaten zu beschreiben sind, haben sich bislang erst ansatzweise Verfahren herausgebildet, die über individuelle Verwendungskontexte hinausgehen. Entsprechende Vorschläge wurden beispielsweise als Bestandteil von Transkriptionskonventionen (z. B. die Angaben zum Transkriptkopf in GAT oder in CHAT), als Bestandteil allgemeinerer Standards (z. B. DC und OLAC, aber auch in den Richtlinien der TEI, s. [13]) oder im Rahmen einer Vereinheitlichung von Metadaten an Datenzentren (z. B. IMDI) formuliert. Es empfiehlt sich, vorhandene Vokabulare, die bei der Dokumentation vergleichbarer Korpora bereits verwendet wurden, als Orientierung für die Dokumentation des eigenen Korpus heranzuziehen.

Mit der **Component MetaData Infrastructure (CMDI)** wurde in CLARIN ein Framework entwickelt, das es erlaubt, die (notwendigerweise) vorhandene Heterogenität im Bereich der Metadaten auf eine gemeinsame Grundlage zu stellen. Die ISOcat-Registry [9] wird dabei zur Definition

der verwendeten Kategorien und Termini verwendet. In der CLARIN Component-Registry [10] sind verschiedene CMDI-Profile registriert, die sich für die Beschreibung mündlicher Korpora eignen.

Konkrete Empfehlungen (vgl. [8])

- Metadaten zu mündlichen Korpora sollten **zum frühestmöglichen Zeitpunkt im Arbeitsablauf** erhoben werden.
- Metadaten sollten in **strukturierten Textformaten** (d. h. in der Regel als XML-Dateien, evtl. alternativ auch tabellarisch als CSV-Dateien) abgelegt werden. Die weit verbreitete Praxis, eine Dokumentation von Metadaten ausschließlich über Verzeichnisstrukturen und Dateibenennungen vorzunehmen, ist für Zwecke der Nachnutzung und Archivierung nicht ausreichend.
- Die Metadatenstandards der **Dublin Core (DC)** Initiative und der **Open Language Archives Community (OLAC)** können als Mindestanforderung für die Beschreibung und Katalogisierung eines Korpus genutzt werden, sind in aller Regel aber alleine nicht ausreichend, um eine adäquate Archivierung und Nachnutzung sicherzustellen.
- Metadaten sollten idealerweise unter Zuhilfenahme des **CMDI-Frameworks** und der **ISOcat-** und **CLARIN Component-Registries** beschrieben werden. Insbesondere sollte geprüft werden, ob ein bereits in der Component-Registry vorhandenes CMDI-Profil für die Metadaten-Beschreibung genutzt oder erweitert werden kann.
- CMDI-Profile für mündliche Korpora werden derzeit z. B. von verschiedenen CLARIN-Zentren (BAS München, HZSK Hamburg, IDS Mannheim, MPI Nijmegen) und vom DIPF entwickelt und bilden dort die Basis der Korpusarchivierung. Zur Beurteilung und ggf. Auswahl eines vorhandenen Profils, aber auch zu einer allgemeinen Beratung im Hinblick auf Metadaten zu mündlichen Korpora, kann es ratsam sein, diese Zentren im Zuge der Projektplanung zu kontaktieren.

1.4. Archivierung

Sinn der obigen Empfehlungen zur Standardisierung bei der Erstellung von Primärdaten (Aufnahmen), Sekundärdaten (Transkription und weiterer Annotation) und Metadaten ist es, die Daten für verschiedene Zwecke langfristig so leicht und breit wie möglich zugänglich zu machen, u. a. für ihre Überprüfung und insbesondere für eine Nachnutzung im Rahmen anderer Forschungsprojekte. Dazu ist eine Speicherung und Backup-Sicherung auf den Rechnern der eigenen Institution in den meisten Fällen nicht ausreichend.

Nicht nur ist es für den Fall eines Unglücks wichtig, Kopien an anderen Orten abgelegt zu haben und bei allen Orten die Hardware rechtzeitig zu erneuern, bevor die Sicherheit der Speicherung nicht mehr garantiert werden kann ("Bitstream-Erhalt"). Es ist auch bei einer Verwendung von offenen und gut dokumentierten Standard-Formaten voraussichtlich in Zukunft immer wieder notwendig, gewisse Daten von einem veraltenden in ein aktuelleres Format zu überführen ("Interpretierbarkeit"). Diese Aufgaben sollten nicht für jedes Projekt und jede Institution neu angegangen werden, sondern werden am besten von hierauf spezialisierten Zentren übernommen.

Welches Zentrum für die Archivierung der in einem bestimmten Projekt erhobenen und erstellten Daten infrage kommt, hängt von verschiedenen Faktoren (insbesondere den Datentypen) ab und kann hier nicht pauschal beantwortet werden. Es können hier aber einige Kriterien genannt werden, auf die bei der Archivierung geachtet werden sollte.

Allgemeine Empfehlungen

- Schon bei der Projektplanung, spätestens bei Projektbeginn, sollte ein geeignetes Zentrum identifiziert und Kontakt mit ihm aufgenommen werden.
- Das Zentrum sollte zumindest in einer bestimmten für das Projekt relevanten Community ein gutes Ansehen haben und allgemein als möglicher Ort für die Archivierung von Daten anerkannt sein.
- Das Zentrum muss in der Lage sein, den Erhalt der Daten hinsichtlich **Bitstream-Erhalt** und **Interpretierbarkeit** auf lange Zeit zu garantieren, und vor allem den einfachen Zugang zu den Daten, üblicherweise direkt über das Internet, ermöglichen können (Lösungen wie das Zusenden von Speichermedien per Post sind nicht mehr zeitgemäß und nur in Ausnahmefällen in Betracht zu ziehen). Idealerweise verfügt das Zentrum über eine explizite Politik der **Qualitätssicherung** etwa durch externe Organisationen. Dieser Prozess ist erst in den Anfängen, aber es soll hier beispielsweise das Data Seal of Approval als eine geeignete qualitätssichernde Instanz genannt werden.
- Möglichst früh sollte mit dem identifizierten Zentrum **Einigkeit über zulässige** und empfohlene **Formate und Standards** erzielt werden; vor allem hinsichtlich der Metadaten haben viele Zentren spezifische Anforderungen. Ein solcher Kontakt empfiehlt sich auch dann, wenn die Nachnutzung der Projektdaten nur sehr eingeschränkt oder gar nicht möglich ist.
- Ebenfalls sollte möglichst früh systematisch Klarheit darüber hergestellt werden, welche Daten welchen zukünftigen Nutzern zugänglich gemacht werden können (moderne Zentren erlauben meist einen konfigurierbaren Zugriff, individuell oder je nach Datentyp und Benutzergruppen).

- Dazu ist es notwendig, vor der Datenerhebung eine verlässliche informierte Einverständniserklärung seitens der aufgenommenen Sprecher zu erhalten (i. d. R. schriftlich, ggf. auch als Bestandteil der Aufnahmen). Standardformulare und Agreements sollten so abgefasst sein, dass sie, soweit wie möglich, die mögliche Nachnutzung auch mit zukünftigen, ggf. gegenwärtig noch nicht bekannten, Methoden oder Techniken klären.

Konkrete Empfehlungen

- Die Zentren, die am CLARIN (in Deutschland, **CLARIN-D**) Projekt teilnehmen und dessen Kriterien erfüllen (siehe [11]), können hier empfohlen werden. Möglicherweise entstehen ähnliche Qualitätsstandards in anderen Infrastrukturnetzwerken wie DARIAH oder META-NET.
- Für Daten aus dem Schulbereich kann auch das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) ein geeigneter Ansprechpartner sein.

Referenzen zu Teil 1

[1] Florian Schiel, Christoph Draxler, Angela Baumann, Tania Ellbogen, Alexander Steffen (2004): The Production of Speech Corpora ("BAS Cookbook"). Version 2.5 : June 1, 2004. [http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/]

[2] Gesprächsanalytisches Informationssystem. [http://prowiki.ids-mannheim.de/bin/view/GAIS/WebHome]

[3] Dieter van Uytvanck (2012): CLARIN-D User Guide, Chapter 6 "Types of Resources: Multimodal Corpora". [http://media.dwds.de/clarin/userguide/text/multimodal_corpora.xhtml]

[4] Schmidt, Thomas, Elenius, Kjell & Trilsbeek, Paul (2010): Multimedia Corpora (Media encoding and annotation). Draft submitted to CLARIN WG 5.7. as input to CLARIN deliverable D5.C-3 "Interoperability and Standards" [http://www1.uni-hamburg.de/exmaralda/files/CLARIN_Standards.pdf]

[5] Selting, M., Auer, P., Barth - Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper - Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., Uhmann, S. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung (10), S. 353 - 402. [http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf]

[6] Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004) Handbuch für das computergestützte Transkribieren nach HIAT. In: Arbeiten zur Mehrsprachigkeit, Folge B (56). [http://www1.uni-hamburg.de/exmaralda/files/azm_56.pdf]

[7] MacWhinney, B. (2000): The CHILDES project: tools for analyzing talk. Mahwah, NJ: Lawrence Erlbaum. [http://childes.talkbank.org/manuals/chat.pdf]

[8] Dieter van Uytvanck & Peter Wittenburg (2012): CLARIN-D User Guide, Chapter 2 "Metadata". [http://media.dwds.de/clarin/userguide/text/metadata.xhtml]

[9] ISOcat-Registry: [http://www.isocat.org/]

[10] CLARIN Component-Registry: [http://catalog.clarin.eu/ds/ComponentRegistry/]

[11] CLARIN-D-Zentren: [http://www.clarin-d.de/de/clarin-d-zentren.html]

[12] DFG-Praxisregeln "Digitalisierung": [http://www.dfg.de/formulare/12_151/]

[13] Guidelines of the Text Encoding Initiative (TEI) [http://www.tei-c.org/index.xml]

Tools

ANVIL	http://www.anvil-software.org/
ARBIL	http://tla.mpi.nl/tools/tla-tools/arbil/
CLAN	http://childes.psy.cmu.edu/
ELAN	http://tla.mpi.nl/tools/tla-tools/elan/
EXMARaLDA	http://www.exmaralda.org/
FOLKER	http://agd.ids-mannheim.de/folker.shtml
Phon	http://childes.psy.cmu.edu/phon/
Praat	http://www.fon.hum.uva.nl/praat/
Transcriber	http://trans.sourceforge.net/en/presentation.php

Datenzentren in Deutschland mit Kompetenzen für mündliche Korpora

Archiv für Gesprochenes Deutsch (AGD, IDS Mannheim) - http://agd.ids-mannheim.de
Bayerisches Archiv für Sprachsignale (BAS, LMU München) - http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html
Deutsches Institut für Internationale Pädagogische Forschung (DIPF Frankfurt) - http://www.dipf.de
Hamburger Zentrum für Sprachkorpora (HZSK, Universität Hamburg) - http://www.corpora.uni-hamburg.de/
The Language Archive (TLA, MPI Nijmegen) - http://www.mpi.nl/tla

DARIAH-Projekt: http://de.dariah.eu/
META-NET-Projekt: http://www.meta-net.eu/

Metadatenstandards und -registraturen

Component MetaData Infrastructure (CMDI) - http://www.clarin.eu/cmdi
Data Category Registry (ISOcat) - http://www.isocat.org/
CLARIN Component Registry http://catalog.clarin.eu/ds/ComponentRegistry/
Dublin Core Metadata Initiative (DC) - http://dublincore.org/
ISLE Meta Data Initiative (IMDI) - http://www.mpi.nl/imdi/
Open Language Archives Community (OLAC) - http://www.language-archives.org/documents.html

Annotationsstandards

[AG] – Advanced Glossing.
Hans-Heinrich Lieb and Sebastian Drude (2000): Advanced Glossing, a Language Documentation Format. (DOBES working paper 1). Berlin.
http://www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf

[GRAID] – Grammatical Relations and Animacy in Discourse.
Geoffrey Haig and Stefan Schnell (2011): Annotations using GRAID: Introduction and guidelines for annotators. Version 6.0. Bamberg u. Kiel.
http://www.linguistik.uni-kiel.de/GRAID_manual6.0_08sept.pdf

[LGR] – Leipzig Glossing Rules.
Bernard Comrie, Martin Haspelmath and Balthasar Bickel (2008): Leipzig Glossing Rules. Max Planck Institute for Evolutionary Anthropology and University of Leipzig.
http://www.eva.mpg.de/lingua/resources/glossing-rules.php

[STTS] – Stuttgart Tübingen Tagset.
Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).
http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

[Tiger] – Wolfgang Lezius (2002): Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), 2002, vol. 8, no.

4. IMS, Universität Stuttgart.
[ToBi] – Tones and Break Indices.
Mary E. Beckman, Gayle Ayers Elam: Guidelines for ToBI Labelling (version 3.0, March 1997). http://www.cs.columbia.edu/~agus/tobi/labelling_guide_v3.pdf
[X-SAMPA] – Computer-coding the IPA: a proposed extension of SAMPA.
http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

Teil 2: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Schriftkorpora

Korpora geschriebener Sprache liegen entweder bereits als elektronische Textressourcen vor (beispielsweise digitale Textarchive, Webseiten, Blogs) oder werden durch Digitalisierung von Schriftträgern gewonnen (vorrangig Manuskripte, Bücher, Zeitungen).
Abschnitt 2.1.befasst sich mit den Standards und Best Practices bei den verschiedenen Arten der Erfassung der Primärdaten. Abschnitt 2.2. gibt Empfehlungen für die Auswahl von nachhaltigen Standards, Formaten und Tools für die Bereitstellung von Sekundärdaten wie die textstrukturelle und linguistische Annotation von Korpustexten, Abschnitt 2.3 entsprechend für Metadatenstandards. Sinn der Verwendung von Standards ist es, die Korpusdaten auch langfristig für eine Nachnutzung in anderen Forschungszusammenhängen zugänglich zu machen. Abschnitt 2.4. gibt Empfehlungen zu Best Practices bei der nachhaltigen Verfügbarmachung und Langzeitarchivierung von Schriftkorpora. Die Referenzen zu Teil 2 enthalten eine Auflistung von Standards, Tools, Datenzentren und Infrastrukturprojekten, die derzeit (2014) den in den Abschnitten 2.1.-2.4. formulierten Anforderungen am besten genügen.

2.1 Digitalisierung und Korpuserfassung

Allgemeine Empfehlungen

Dieser Abschnitt befasst sich mit der Erstellung schriftsprachlicher Korpora durch Kompilierung bestehender Textressourcen (z. B. aus digitalen Textsammlungen, Webseiten, Blogs, Webforen u. ä.) oder durch die Digitalisierung von Trägern schriftlicher Überlieferung (d. h. vorrangig von Manuskripten, Zeitungen/Zeitschriften sowie gedruckt erschienener Literatur). Ziel ist es dabei, die Texte in standardkonformer Weise so aufzubereiten, dass sie in nachhaltigen Formaten vorliegen sowie interoperabel und daher im Kontext schriftsprachlicher Korpora vielfältig und flexibel nachnutzbar sind.

Die Entwicklung von Standards und Best Practices ist auch im Bereich der Erstellung schriftsprachlicher Korpora ein dynamischer, nicht abgeschlossener und stetig fortlaufender Prozess, in welchen projektübergreifende Anforderungen an die Aufbereitung und Auszeichnung, die Pflege sowie die Archivierung von Daten einbezogen werden. Die im Folgenden formulierten Hinweise und angegebenen Referenzen sollen als Orientierungspunkte dienen, vor deren Hintergrund projektspezifische Abweichungen von bereits etablierten Standards und verbreiteten Methoden vermieden bzw. beurteilt und hinterfragt werden können. Sollte sich im Projektverlauf nach Evaluation der etablierten Standards bzw. Best Practices eine Abweichung von diesen als notwendig erweisen, so sollten diese Abweichungen den jeweiligen Standardisierungsgremien bzw.

Best-Practice-Communities übermittelt werden, damit diese wiederum die Erweiterung des Standards bzw. ihrer Empfehlungen um die als fehlend oder unzureichend identifizierten Aspekte erwägen können.

Konkrete Empfehlungen

Empfehlungen zur Textauswahl

Am Beginn der Korpuserstellung sollte eine auf die Ziele des jeweiligen Projekts bezogene umfassende Bibliographie oder Quellenliste stehen, welche die zu digitalisierenden bzw. die bereits aus anderen Projektzusammenhängen digital vorliegenden und ggf. zu integrierenden Ressourcen umfasst. Eventuelle rechtliche Beschränkungen sollten dabei beachtet und dokumentiert werden. Die **Textauswahl** sollte dadurch nachvollziehbar erfolgen, d. h. **auf im Vorfeld festgelegten Kriterien beruhen**, welche hinsichtlich ihrer Zusammenstellung und Anwendung zu dokumentieren sind. Die daraus resultierende Nachvollziehbarkeit der Textauswahl ist eine unerlässliche Grundlage für die Beurteilung der am Korpus gewonnenen Forschungsergebnisse, deren Beurteilung hinsichtlich ihrer Generalisierbarkeit und Repräsentativität sowie, darauf aufbauend, die Beurteilung der Nachnutzbarkeit des bereitgestellten Korpus als Basis für weitere bzw. anderweitige Untersuchungen.

Für die Konzeption des angestrebten Korpus sollten die Faktoren der **Korpusgröße, Datenqualität und Tiefe der Erschließung** gegeneinander abgewogen werden. Beispielsweise können für bestimmte Fragestellungen eher kleinere, textsortenorientierte Korpora im Zentrum des Interesses stehen oder es kann vonnöten sein, die Qualitätssicherung und Erschließungs- bzw. Annotationstiefe höher zu bewerten als die Korpusgröße. Dies gilt auch in Szenarien, wo das Web als Quelle für ein zu erstellendes Korpus dient. Hier sollte die quantitative Gewinnung von Daten mit einer qualitativen Prüfung, Erschließung und Analyse der Daten einhergehen. Gleichzeitig können aber auch Größe und Dispersion wichtig für gesicherte Schlussfolgerungen sein. Diese Faktoren sind je nach Projekt abzuwägen; entsprechende Entscheidungen sind transparent zu machen und zu motivieren.

Primärdaten

Für schriftsprachliche Korpora sind drei Arten von Primärdaten denkbar:

1.) Die Quellen gedruckter oder handschriftlicher Texte (Lernertexte, Briefe u. ä., als Bilddigitalisate oder im Original), die der Textdigitalisierung für das jeweilige Korpus zugrunde gelegt werden;
2.) digitale Textdaten, die aus einer früheren Vervolltextung gedruckter oder handschriftlicher Quellen hervorgehen;
3.) genuin digitale Textdaten (‚born digital'), wie sie in der Regel für gegenwartssprachliche Texte vorliegen.

Zu 1.) Für den Aufbau eines Korpus auf Grundlage bestehender physischer Textquellen wird die Anfertigung von Bilddigitalisaten dieser Quellen empfohlen, um zum einen eine standortunabhängige Bearbeitung zu ermöglichen und zum anderen die unkomplizierte Überprüfbarkeit der Korrektheit der Erfassung zu gewährleisten. Die zugrundeliegenden Bilddigitalisate müssen in ausreichender Qualität (Mindestanforderung: unkomprimierte TIFF-Dateien oder JPEG2000 in seiner

verlustfreien Form als Format als Bildmaster, mind. 300 dpi, Farbscans) vorliegen. Abstriche bei der Bildqualität sind nachvollziehbar zu begründen. Im Vorhinein der Bilddigitalisierung ist zu prüfen, ob für die vorgesehenen Quellen bereits Bilddigitalisate in ausreichender Qualität zur Verfügung stehen.

Zu 2.) Liegen die vorgesehenen Quellen bereits als digitale Textdaten (originalgetreue Volltexte oder kritische Editionen) vor, so wird die Recherche und Bereitstellung der zugehörigen zugrundeliegenden Bilddateien empfohlen, um die Qualität der Transkription und deren Quellennähe überprüfbar zu machen. Abweichungen von dieser Empfehlung (z. B. aufgrund rechtlicher Beschränkungen für die Nachnutzung der Bilddateien) sind zu begründen. Darüber hinaus gilt in diesem Fall besondere Sorgfalt bzgl. der Zuverlässigkeit der Transkription gegenüber der (historischen) Vorlage. Die Zusammenstellung von Bilddigitalisaten ohne Zugabe ausreichend verlässlicher Transkriptionen genügt nicht für den Aufbau eines Korpus.

Zu 3.) Bei der Nachnutzung von ‚born digital' Texten sollte die Änderbarkeit der Quelldaten einkalkuliert werden. Einen Spezialfall stellen hier Webressourcen dar, die zu einem Korpus kompiliert werden sollen. Hier sollten eine Qualitätssicherung sowie die Replizierbarkeit von Ergebnissen im Vordergrund stehen. Die ist möglich, indem Korpora von Webtexten kompiliert werden, d. h. die notwendigen Sprachdaten aus dem Web nach vorher definierten und ausführlich dokumentierten Kriterien gewonnen werden und dieses Material sodann als Korpus erschlossen wird. Webtexte können auch mittels Crawling-Methoden kompiliert werden. Hier sollte eine ausführliche Dokumentation der zugrundeliegenden Crawling-Methoden und Algorithmen erstellt werden, im besten Fall der verwendete Crawler selbst auch nachnutzbar zur Verfügung gestellt werden, insbesondere wenn das Webkorpus selbst aus rechtlichen Gründen nicht weitergegeben werden darf (s. Kap. 2.4.).

Grundsätzlich sind bei der Nachnutzung digitaler Volltextdaten (Fall 1. und 2.) die Angaben zur Textquelle unerlässlich für die Nachprüfbarkeit der Forschungsergebnisse. Dabei sollte nicht nur die ursprüngliche Quelle und der Zeitpunkt der Übernahme angegeben, sondern auch die ggf. nachfolgenden Bearbeitungsschritte transparent (Kuration und Anreicherung von Daten) gemacht werden. Wichtig bei der Übernahme bestehender Ressourcen ist es selbstverständlich, die rechtlichen Rahmenbedingungen zu beachten (siehe dazu auch [17]).

Texterfassung

Der zentrale Schritt bei der Erstellung (historischer) Korpora ist (nach der bibliographischen Erfassung der Textauswahl und ggf. der Akquise bzw. Erstellung geeigneter Bilddigitalisate) die **Volltexterfassung**. Hierbei gibt es grundsätzlich zwei **Verfahren**: die (einfache, doppelte oder mehrfache) **manuelle Transkription** durch Projektmitarbeiter oder einen Dienstleister und die automatische Texterfassung per **Optical Character Recognition (OCR)**. Beide Verfahren unterscheiden sich hinsichtlich der entstehenden Kosten und der erwartbaren Textgenauigkeit je nach Textvorlage stark. Es wird empfohlen, schon bei der Transkription neben der Wiedergabe der Zeichen aus der Vorlage auch die den Text strukturierenden Merkmale mit zu erfassen (Absätze, Überschriften, Kapitelgrenzen etc.; siehe dazu Abschnitt Annotation). Beide Verfahren können zudem mit jeweils mehr oder weniger zeit- und damit kostenintensiven Maßnahmen der Vor- und

Nachbereitung kombiniert werden, um die Qualität der Texterfassung zu erhöhen bzw. sicherzustellen.[3] Es müssen daher die anfallenden Kosten und die angestrebte Qualität der Transkription mit Blick auf die Projektziele im Vorfeld gegeneinander abgewogen werden. Die Entscheidung für das jeweils für die Texterfassung gewählte Verfahren ist offenzulegen und zu begründen.
Generell gilt: Es sollte erstens eine möglichst hohe Textgenauigkeit angestrebt werden, um am Korpus erzielte Forschungsergebnisse nachvollziehbar zu machen. Daher sollte zweitens bei der Texterfassung mit automatischen Methoden (OCR) immer eine manuelle oder halbautomatische Überprüfung und ggf. Nachkorrektur vorgenommen werden. Der Verzicht auf eine solche Überprüfung/Nachkorrektur ist grundsätzlich zu begründen. Drittens sind die für die Texterfassung angewandten Verfahren, die Transkriptionsregeln, ggf. die eingesetzte OCR-Software, die erfolgten Schritte der Vor- bzw. Nachbereitung der Texte sowie die resultierende Erfassungsgenauigkeit ausführlich zu dokumentieren. Das Verfahren der Texterfassung und Hinweise zu Transkription sollten auch in den Metadaten zu jeder einzelnen Textressource vermerkt werden.

Eine **vollständige Texterfassung** gewährleistet unter anderem in der Regel eine größere Flexibilität der Nachnutzung in anderen Bereichen, d. h. auch für andere als die bei Korpuserstellung angedachten Forschungsfragen. Ist eine vollständige Texterfassung nicht realisierbar, kann je nach Forschungsfrage auch der Rückgriff auf eine **auszugsweise Texterfassung** hinreichend sein.

Die **Richtlinien zur Transkription** sollten immer mit angegeben werden. Vor allem Abweichungen von der Vorlage sollten begründet und somit möglichst nachvollziehbar gemacht werden. Dies beinhaltet **Angaben zum ‚Diplomatizitätsgrad'** der Transkription, die dokumentieren, inwieweit die Wiedergabe des originalen Sprachstandes und der Graphie gewährleistet ist.

Annotation
Die Annotation der Korpustexte sollte sowohl die Auszeichnung struktureller Merkmale (Überschriften, Paragraphen, Anmerkungen, Zitate etc.), linguistischer Eigenschaften (Lemmatisierung, Tokenisierung, morphologische Annotation/Part-of-Speech-Tagging etc.) sowie sprachlicher Spezifika (dialektale/regionale Zuordnung, medien- oder textsortenbezogene Besonderheiten) berücksichtigen. Entsprechend den Projektzielen ist der geeignete Skopus und die geeignete Tiefe der jeweiligen Annotation abzuwägen und begründet zu dokumentieren.
Die Annotation kann manuell, automatisch oder halbautomatisch erfolgen.

In allen Fällen ist das gewählte Tagset zu dokumentieren und die Annotation sämtlicher Korpustexte einheitlich nach diesem Tagset (und ggf. den weiteren dokumentierten Richtlinien zur Annotation) zu erstellen. Es sollten um willen der Interoperabilität möglichst bestehende Annotationsstandards zugrunde gelegt werden (s. unten Kap. 2.2). Um willen der Nachvollziehbarkeit

[3] Beispielsweise kann eine automatisch erstellte Transkription durch den Einsatz der für die Vorlage am besten geeigneten OCR-Software, die optimale Anpassung der Erkennungs-Parameter an die Vorlage, durch weiteres ‚Training' der Software, das Einbinden spezieller Wortlisten und Lexika etc. sowie durch manuelle oder (semi)automatische Nachkorrektur optimiert werden. Je nach angestrebter Qualität entsteht hierbei ein zeitlicher Mehraufwand, der bei der Abschätzung der Kosten für die ‚reine' Texterfassung nicht vernachlässigt werden darf.

getroffener Entscheidungen und der Nachnutzbarkeit der Daten sind darüber hinaus grundsätzlich die folgenden Parameter im Rahmen einer ausführlichen Projektdokumentation offenzulegen:

- das jeweils verwendete Tagset;
- die Strukturierungstiefe;
- im Falle der automatischen Annotation die eingesetzten Verfahren und Tools;
- die erwartbare Annotationsgenauigkeit.

Qualitätssicherung:
Für nachvollziehbare, unverfälschte und vollständige Rechercheergebnisse ist die Verlässlichkeit und Korrektheit der Korpusdaten von immanent wichtiger Bedeutung. Daher ist großer Wert auf die Qualitätssicherung im Vorhinein sowie im Anschluss an die Datenerfassung/Digitalisierung zu legen. Die Qualitätssicherung sollte die folgenden Aspekte umfassen:

- Textqualität, damit der Suchraum für Korpusrecherchen eindeutig ist (d. h. maximale Reduktion der false positives/negatives);
- Qualität der strukturellen Annotation;
- Qualität der linguistischen Annotation;
- Qualität der Metadaten

Dabei sind verlässliche Genauigkeitsmessungen sowie die Verbesserung der Erkennungsrate durch halbautomatische oder manuelle Nachkorrektur sowie durch Optimierung der eingesetzten Verfahren und Tools vonnöten.

2.2 Standards und Tools

Hinsichtlich der Textauszeichnung und -analyse wird die Orientierung an bestehenden Standards empfohlen. Dabei ist grundsätzlich zwischen der Auszeichnung textueller Merkmale (strukturelle Annotation) und der Auszeichnung sprachlicher Merkmale (linguistische Annotation) zu unterscheiden.

Grundsätzlich sind standardisierte/genormte Formate (z. B. **XML**) und Kodierungen (z. B. Unicode) proprietären Formaten vorzuziehen, da diese eine nachhaltige Datenhaltung und Archivierung ermöglichen. Für die Annotation der Forschungs- und Verarbeitungsdaten ist insbesondere die Verwendung eines XML-Formats als primäres Auszeichnungsformat zu empfehlen. Mit XML ist neben formalen Auszeichnungen auch eine inhaltliche/semantische Kategorisierung von Texteinheiten möglich. Darüber hinaus ermöglicht XML eine auf Standards basierte Weiterverarbeitung der Daten (mittels **XML-Technologien** wie XSLT, XPath, XQuery). Weiterhin lassen sich XML-Daten problemlos in andere (sekundäre) Formate (z. B. Repräsentationsformate wie HTML) konvertieren. Andersherum sehen reine Präsentationsformate wie HTML oder proprietäre Textverarbeitungsformate wie DOCX eine semantische Kategorisierung von Texteinheiten in der Regel nicht vor und sind darüber hinaus nicht speziell für die Langzeitarchivierung angepasst. Um willen der langfristigen **Interpretierbarkeit** der XML-Textauszeichnung ist schließlich die Festlegung und Offenlegung eines expliziten **Datenmodells** bzw. einer expliziten **Dokumentgrammatik (Schema)** vonnöten, diese vereinfachen die Interpretation sowie ggf. die Konvertierung und Integration der Daten und erleichtern damit auch langfristig deren Weiterverarbeitung und Nachnutzbarkeit.

Empfehlungen für die strukturelle Annotation

Als de facto-Standard für die **strukturelle Auszeichnung** von Textpassagen haben sich die P5-Richtlinien der **Text Encoding Initiative** (TEI) etabliert ([13]). Da der **TEI P5**-Regelsatz sehr umfassend ist, um für die verschiedensten editorischen Bedürfnisse verwendbar zu sein, sollte aus dem Gesamtregelsatz ein eingeschränktes Format, welches an die spezifischen Projektbedürfnisse angepasst ist, herausgelöst werden. Derlei Einschränkungen sind mithilfe des **ODD**-Formalismus, welchen die TEI bereitstellt, möglich. Die Entscheidung gegen alle existierenden TEI-Formate und für die Schaffung eines neuen, projekteigenen Formats ist besonders zu begründen. Ebenso sollte die Entscheidung für ein spezifisches TEI-Format motiviert werden.

Empfehlungen für die linguistische Annotation

Für die weitere, linguistische Annotation von Schriftdaten ist in den letzten zwanzig Jahren eine Vielzahl spezialisierter Software-Tools entwickelt worden, die sowohl einer Effektivierung des Arbeitsablaufs als auch einer Verbesserung der Archivierbarkeit und Nachnutzbarkeit der entstehenden Daten dienen. Anders als bei mündlichen Korpora handelt es sich bei Schriftkorpora oftmals um sehr große Textmengen, so dass eine vollständige linguistische Annotation oft nur (semi-)automatisch möglich ist. Andererseits können in Korpusprojekten auch manuelle Annotationen von (Teilen von) Korpora notwendig sein, beispielsweise, wenn noch keine Tools für eine bestimmte linguistische Beschreibungsebene vorliegen, wenn ein Goldstandard hergestellt werden soll, wenn grundsätzlich eine Annotationsgenauigkeit angestrebt wird, die mit automatischen Methoden nicht zu erreichen ist, oder wenn automatisch hergestellte Annotationen korrigiert werden sollen.[4] Auch für die Herstellung manueller linguistischer Annotationen sind spezialisierte Annotationstools grundsätzlich der Verwendung allgemeiner Textverarbeitungssoftware vorzuziehen, da nur erstere Primärdaten und Annotationen in einer konsistent strukturierten Form, die automatisch weiterverarbeitet und ausgewertet werden kann, speichern.

Werden einem Korpus viele verschiedene Annotationsschichten hinzugefügt, so empfiehlt sich meist eine Konzipierung als **Mehr-Ebenen-Annotationen**, in der die verschiedenen Annotationsschichten zunächst getrennt vorgehalten werden. Sie bleiben dabei aufeinander beziehbar, da sie per **Standoff**-Technik sämtlich auf dieselbe Textbasis referieren. In der Regel ist diese Basis bei Schriftkorpora die Tokenisierungsebene, seltener die Ebene der Character-Bytes. Standoff-Annotationen ermöglichen auch die Repräsentation konkurrierender Analysen auf derselben Annotationsebene und bieten sich damit beispielsweise für die Repräsentation alternativer Tokenisierungen an (wenn die Tokens nicht selbst die Standoff-Basis darstellen).

Annotationstools und Formate

Bei der Wahl geeigneter Tools zur linguistischen Annotation ist darauf zu achten, dass diese standardisierte/genormte Formate interpretieren und ausgeben können. Falls das Ausgabeformat eines gewählten Tools kein standardisiertes ist, sollte die Ausgabe zum Zwecke der Nachnutzbarkeit in ein Standardformat überführt werden, welches so einfach wie möglich und so komplex wie nötig sein sollte. Textbasierte **Spaltenformate**, wie sie von vielen Taggern gelesen und aus-

[4] s. CLARIN USER GUIDE – Ch. 7-5.1 / S. 92

gegeben werden und in einigen Shared Task-Wettbewerben vorausgesetzt werden, können einen Quasi-Standard darstellen.

Oftmals ist es erforderlich, dass während der Bearbeitung einer Annotationsebene **Korrekturen** an anderen Annotationsebenen (bis hin zur Tokenisierungsebene) vorgenommen werden müssen und dann auch eine Buchführung über derlei Korrekturen möglich sein soll, also die Eingabe von Kommentaren, Notizen oder bestimmten Metadaten (wie Label automatisch hinzugefügt vs. korrigiert) zu einzelnen Annotations-Items. Das Annotationstool und das Repräsentationsformat sollten dann entsprechend diesen Anforderungen ausgewählt werden. Sollte für die spezifischen Anforderungen eines Projekts **Anpassungen** an einem Tool oder an einem verwendeten Tag-Set vorgenommen werden müssen, so sollten die geplanten Modifikationen und der erforderliche Aufwand im Projektantrag dargelegt werden.

Die einem Korpus hinzugefügten Annotationen sollten dokumentiert werden hinsichtlich der Annotationskategorien (inklusive Tokenisierungs- und Segmentierungsprinzipien, Verweis auf das verwendete Tag-Set und Annotationsformat). Für automatisch hinzugefügte Annotationen sollten die eingesetzten Verfahren und Tools, d. h. ggf. auch Vorverarbeitungsschritte sowie auch die **Annotationsqualität** z. B. anhand der Angabe von Ergebnissen einer möglichst repräsentativen Evaluation dokumentiert werden. Auch für einen manuellen Annotationsprozess sollten die Annotationsqualität und die Verfahren ihrer Absicherung dokumentiert werden. Dazu gehört die Dokumentation der Annotationsrichtlinien, die alle verwendeten Tags und ihre Definitionen mit Beispielfällen aufführen, und der Annotatorenübereinstimmung (vgl. [14]).

Für die Untersuchung wenig beforschter Themen kann es angebracht sein, im Projekt eigene Tools zu entwickeln, dann sollten auch diese Tools entsprechend den hier beschriebenen Standards implementiert, dokumentiert und für eine Nachnutzung verfügbar gemacht werden, s. auch Kap. 2.4. Erreicht werden kann dies insbesondere durch die Veröffentlichung des Quellcodes unter einer permissiven Lizenz, welche die Weitergabe und Weiterentwicklung durch Dritte ausdrücklich gestattet (s. [17], Abschnitte 2.1.6 und 2.3.3). Eine skriptartige, nicht auf Nachhaltigkeit abzielende Implementierung kann auch ihre Berechtigung haben, wenn die Toolentwicklung kein Schwerpunkt des Projekts ist.

Analysetools

Nicht nur für den Annotationsprozess, sondern auch für viele Arten der **Abfrage und Analyse (Querying)** von annotierten Korpora stehen Tools zur Verfügung. Bei der Auswahl von Tools sollte auf standardisierte Ausgabeformate geachtet werden (z. B. CSV, JSON). Linguistische Korpusanalysetools basieren in der Regel auf einem Datenbanksystem, in dem Korpora und Annotationen strukturiert und schnell zugreifbar gespeichert werden. Bei einer Kooperation mit einem Datenzentrum besteht meist die Möglichkeit, während oder nach der Projektlaufzeit die zur Zentrumsinfrastruktur gehörigen Korpusdatenbanksysteme mit ihren Abfrage- und Analysetools zu verwenden. Für die nachhaltige Nutzbarmachung und Präsentation (z. B. über Web-Schnittstellen) bestimmter Korpusdaten, insbesondere im Bereich der Digital Humanities, wo Textdaten gemeinhin mit weiteren Datentypen verknüpft sind, werden Datenbanksysteme auch direkt, d. h. ohne eine spezielle Linguistik-Schicht, eingesetzt. Die verschiedenen Datenmodellierungs-, An-

frage- und Analysemöglichkeiten, die die verfügbaren Datenbankparadigmen (relationale Datenbanken, Dokumentdatenbanken, XML-Datenbanken oder Graphen-Datenbanken) bieten, haben einen unmittelbaren Einfluss auf die Klasse der Fragestellungen, die mit ihnen beantwortet werden können. Fragen dieser Art sollten im Projektzusammenhang möglichst interdisziplinär (Informatik und Geisteswissenschaften) erörtert werden.

Konkrete Empfehlungen für die strukturelle Annotation

- Um willen der **Interoperabilität** der projekteigenen Textdaten mit Korpusdaten aus bestehenden Projektkontexten ist die Nachnutzung eines bereits bestehenden **TEI**-Formats zu erwägen, z. B. der durch das Verbundprojekt CLARIN-D für die Auszeichnung gedruckter Texte empfohlenen TEI-basierten Formate **DTA-Basisformat** und **I5** (vgl. [16], Teil II, Kap. 6; [DTABf]; [I5]; [21],[22]).
- Strukturelle Informationen aus OCR können zunächst in den Formaten **hOCR**, **ALTO** oder **ABBYY XML** vorliegen. Diese Zwischenformate sollten in jedem Fall in ein finales TEI-Format überführt werden. Der Bezug zu den ursprünglichen OCR-Formaten sollte dabei erhalten bleiben.
- Die Auszeichnung internetbasierter Kommunikation (Computer-mediated communication, CMC) sollte sich an den Arbeiten der Special Interest Group (SIG) Computer-Mediated Communication der TEI orientieren (vgl. [15]).

Konkrete Empfehlungen für die linguistische Annotation

Für die Setzung, Speicherung, Bearbeitung und Abfrage linguistischer (Mehr-Ebenen-) Annotationen existieren spezialisierte Tools und Formate. Folgende konkrete Empfehlungen für die linguistische Annotation von Korpora können gegeben werden:

- Folgende XML-basierten Formate können als Austauschformate für als Standoff realisierte Mehr-Ebenen-Annotationen empfohlen werden: **PAULA** (Austauschformat für ANNIS), **LAF/GrAF** (LAF ist ISO-Standard, GrAF seine graph-basierte XML-Serialisierung, unterstützt auch Merkmalsstrukturen), **TIGER-XML**, das von CLARIN-D empfohlene Text Corpus Format (**TCF**) und **UIMA XMI** (ein OASIS-Standard für die Serialisierung von Merkmalsstrukturen mit internationaler Verbreitung in Forschung und Industrie) .
- Als Tagset für das POS-Tagging deutschsprachiger Texte ist das Stuttgart-Tübingen-Tagset (**STTS**) als Quasi-Standard etabliert.
- Für weitere linguistische Ebenen existieren noch keine derartig etablierten Standards. Es sollten grundsätzlich Tagsets und Taxonomien bevorzugt werden, für deren Kategorien bereits Eintragungen in der **ISOcat-Registry** [9] vorliegen, bzw. sollte bei Verwendung einer projektspezifischen Taxonomie erwogen werden, zum Zwecke der Konsistenzsicherung und Dokumentation die Kategorien der Taxonomie vorhandenen Einträgen in der ISOcat-Registry zuzuordnen.
- Für eine Orientierung über die Vielzahl an verfügbaren linguistischen Annotationstools für die verschiedensten Aufgaben, Analyseebenen und Sprachen wird auf den Überblick in Kapitel 7 „Linguistic tools“ im Teil II des CLARIN-Nutzerhandbuchs [16] verwiesen. Es sei hier nur der **TreeTagger** als weit verbreitetes Tool für Tokenisierung, Lemmatisierung und POS-Tagging deutscher Texte nach STTS genannt sowie **WebLicht** für die automatische linguistische Annotation in nutzerdefinierten Prozessketten über Webservices in der Cloud

(Daten bewegen sich zum Tool) oder DKPro Core für Prozessketten über lokale Daten (Tool bewegt sich zu den Daten). Die einzelnen Tools, die in WebLicht eingebunden sind, sind ebenfalls im CLARIN-Nutzerhandbuch aufgelistet.

- An Tools für die manuelle linguistische Annotation (verschiedene Funktionalitäten) werden empfohlen: **SALTO** (Annotation von Baumbanken), **annotate** (syntaktische Annotation), **WebAnno**, **TrED**, **MMAX2** oder **WordFreak**. Das CLARIN-Nutzerhandbuch [16] (Teil II, Abschnitt 5.1) macht ebenfalls Angaben zu Tools für die manuelle linguistische Annotation.
- Für die Hinzufügung von linguistischen Annotationen eignen sich auch die umfassenden Textprozessierungsarchitekturen **GATE**, **NLTK**, **Open NLP** und **UIMA**. GATE ist beispielsweise ein Framework, welches die manuelle Annotation von Textbereichen mit eigenen Kategorien erlaubt, aber auch die Einbindung und Konfiguration externer Tagger wie den TreeTagger.
- State-of-the art Tools für die Korpusabfrage, -analyse und -visualisierung sind **ANNIS** (insbesondere für Mehr-Ebenen-Annotationen), **ICARUS** (für Dependenz-Baumbanken), **TIGERSearch** (für Baumbanken) und **CWB/CQP**.
- Zur Evaluation der Qualität automatischer Annotationen bzw. zur Bestimmung der Inter-Annotatoren-Übereinstimmung stehen auch Implementierungen zur Verfügung, z.B. aus dem Kontext einschlägiger Shared Tasks oder in Anntotationstools oder -frameworks bzw. Programmierbibliotheken (z.B. **DKPro** Statistics oder **GATE** IAA Plugin)
- Bei Tools besteht grundsätzlich die Gefahr, dass sie nicht mehr weiter gepflegt werden, beispielsweise können java-basierte Tools wie MMAX2 oder SALTO inkompatibel mit neueren Java-Versionen werden. Es wird daher empfohlen, Tools zu bevorzugen, von denen bekannt ist, dass sie noch gepflegt oder sogar weiter entwickelt werden. Wenn die Wahl besteht, sollten quelloffene und permissiv-lizensierte Tools gegenüber proprietären Tools vorgezogen werden, da erste nach Bedarf weiter gepflegt werden können, selbst wenn die ursprünglichen Entwickler nicht mehr verfügbar sein sollten. Auch hier ist eine Orientierung daran, welche Tools in die CLARIN-Infrastruktur aufgenommen wurden, hilfreich [16].

2.3 Metadaten

Für jedes Dokument in einem Textkorpus sind Metadaten zu erheben, welche in homogener und standardkonformer Weise strukturiert und projektübergreifend nachvollziehbar sind. Die Metadaten sollten dabei möglichst ausführliche Informationen zu folgenden Aspekten enthalten:

- Angaben zur digitalen Ausgabe (Titel, Untertitel, Autor, Erscheinungsdatum, Herausgeber/Bearbeiter/verantwortliche Personen bzw. Organisationen);
- Angaben zur Textquelle (Titel, Autor, Herausgeber/Bearbeiter, Erscheinungsort, Erscheinungs- bzw. Entstehungsdatum, Verlag, Angabe zur Reihe bei unselbständigen Publikationen, Aufbewahrungsort und Signatur bzw. bei genuin digitalen Quellen Ort der Verfügbarkeit, ...);
- Angaben zum Projekthintergrund und ggf. Kontaktmöglichkeiten;
- Angaben zum Umfang, zur Annotationstiefe, zu den Richtlinien der Transkription bzw. Textzusammenstellung/Annotation;
- inhaltliche Angaben: Sprache; Klassifikation (Textsorte, Genre, ...); ...

- Angaben zu Nutzungsbedingungen, der rechtlichen und technischen Verfügbarkeit etc. der Korpora;
- Hinweise zur korrekten Zitierweise.

Nicht allein bei der Digitalisierung von (historischen) Werken, sondern auch bei der Übernahme bestehender Texte, z. B. durch Integration bestehender Texte/Korpora oder durch (Web)Crawling (nach definierten und explizierten Kriterien) sind die Angaben zur Textquelle unerlässlich für die Nachvollziehbarkeit der Forschungsergebnisse. Dabei sollten nicht nur die ursprüngliche Quelle und der Zeitpunkt der Übernahme angegeben, sondern auch die ggf. nachfolgenden Bearbeitungsschritte (z. B. Kuration und Anreicherung von Daten) sowie mögliche Verantwortlichkeiten transparent gemacht werden.

Möglichst früh im Projektverlauf, in jedem Fall jedoch zum Zeitpunkt der Veröffentlichung der einzelnen digitalisierten Ressourcen, sollten diese Metadaten über eine geeignete Schnittstelle (z. B. das Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)) zum **Harvesting** und **Einspeisen in geeignete Verzeichnisse** bereitgestellt werden.

Konkrete Empfehlungen
Als Primärformat für die Erfassung von Metadaten ist ein verbreitetes, standardisiertes Format zu wählen, welches die strukturierte Erfassung semantisch differenzierter Metadaten in dem beschriebenen Umfang erlaubt (z. B. TEI-Header, METS/MODS, CMDI, EAD). Aus dem gewählten Format sollten andere verbreitete Metadatenformate (möglichst verlustfrei) erzeugt und bereitgestellt werden.

Für weitere konkrete Empfehlungen zur Metadatenaufnahme vgl. [25] und [26].

2.4 Nachhaltigkeit, Zitierbarkeit, Nachnutzbarkeit und Langzeitarchivierung

Sinn der Empfehlungen zur Standardisierung bei der Erstellung von Primärdaten (Digitalisierung und Erfassung), Sekundärdaten (Transkriptionen und weitere Annotationen) und Metadaten ist es, die Daten für verschiedene Zwecke langfristig so leicht und breit wie möglich zugänglich zu machen, u. a. für ihre Überprüfung und insbesondere für eine Nachnutzung im Rahmen anderer Forschungsprojekte.

Bei der Planung eines Korpusprojekts sollte eine Nachnutzung der entstehenden Ressourcen durch ‚Sekundärnutzer' bereits konzeptionell berücksichtigt werden, und im Projektantrag sollte dargestellt werden, wie die spezifischen Projektzielsetzungen und die darauf bezogene Art der Korpusaufbereitung eine spätere Nachnutzung der Korpusdaten auch in anderen Projektzusammenhängen und für andere Fragestellungen erlauben, d. h. welche Möglichkeiten und welche Beschränkungen dabei bestehen.

Nachnutzbarkeit beinhaltet auch die Klärung rechtlicher Fragen, die Einrichtung der Zitierbarkeit

von Ressourcen und Ressourcenkomponenten sowie ihre Langzeitarchivierung. Für die Langzeitarchivierung und die Sicherung der Langzeitverfügbarkeit der im jeweiligen Projekt erzielten Forschungsdaten ist eine Speicherung und Backup-Sicherung auf den Rechnern der eigenen Institution in den meisten Fällen nicht ausreichend. Es empfiehlt sich die Kooperation mit einem/einer darauf spezialisierten Zentrum/Institution (s. [11] und konkrete Empfehlungen unten).

Allgemeine Empfehlungen

Nachhaltige Konzeptionierung von Transkriptionen und Annotationen:
In Anträgen für (historische) Korpusprojekte sollte dargelegt werden, inwiefern ein gewählter Transkriptionsmodus oder eine gewählte Annotationstiefe, der oder die für eine bestimmte Fragestellung geeignet ist, ggf. andere Nutzungsweisen ermöglicht bzw. ausschließt. Ggf. sollten alternative Transkriptions- bzw. Annotationsformate erwogen werden, die eine breitere Nachnutzung erlauben. Die getroffene Entscheidung sollte vor dem Hintergrund a.) projektspezifischer Ziele, b.) von Möglichkeiten der Nachnutzung und von c.) Kosten-Nutzen-Erwägungen nachvollziehbar begründet werden.

Klärung rechtlicher Fragen: In einem Antrag sollte angegeben werden, welche Korpora und welche Bestandteile (Teilkorpora, Annotationsschichten, Tools) welchen Nutzergruppen im Rahmen der Nachnutzung zur Verfügung gestellt werden können. In Kooperation mit einem spezialisierten Zentrum sollten die resultierenden **Nutzungsbedingungen** nach Möglichkeit im Rahmen einer **Lizenz**, welche die Nachnutzung regelt, formuliert werden. Voraussetzung dafür ist Klarheit über den rechtlichen Status der entwickelten Tools oder Ressourcen bzw. Ressourcenteile (s. dazu [17]).
Die Frage, inwiefern eine Nachnutzung der im Projekt aufgebauten Ressourcen möglich sein wird, sollte im Vorfeld der Antragstellung geklärt werden. Beschränkungen der Nachnutzung für bestimmte Ressourcenteile sollten rechtfertigt werden. Problematisch in Bezug auf Nutzungs- bzw. Nachnutzungsrechte gestalten sich z. B. **web-basierte Korpora**, die ohne die explizite Erlaubnis jedes Text-Urhebers bzw. Rechteinhabers nicht weitergegeben werden dürfen. Da die Einholung der entsprechenden Rechte angesichts der Größe und Diversität dieser Korpora in der Regel nicht (mit vertretbarem Aufwand) möglich ist und die Weitergabe der Korpora demzufolge widerrechtlich wäre, ist die wissenschaftliche Überprüfbarkeit der anhand solcher Korpora gewonnenen Forschungsergebnisse anderweitig zu garantieren. Hier empfiehlt sich die Beschränkung auf zitierfähige, frei verfügbare Webtexte, die längerfristig auch für andere Personen über ihre ursprünglichen Webpräsenzen zugänglich sind. Im Falle von Korpora internet-basierter Kommunikation (IBK, z. B. Chat-, Foren-, SMS- oder E-Mail-Korpora) kann es über die Klärung der Nutzungsrechte hinaus erforderlich sein, **Methoden zur Wahrung der Persönlichkeitsrechte** (Anonymisierung oder Pseudonymisierung) anzuwenden, bevor die Korpusdaten zur Nachnutzung freigegeben werden (s. [17]).
Im Falle von **Zweit- oder Parallelveröffentlichungen (**beispielsweise im Zusammenhang mit einer Verlagspublikation im Rahmen eines Editionsvorhabens) sollten die Nutzungs- und Nachnutzungsrechte vorab mit dem Rechteinhaber der zugrundeliegenden Publikation(en) geklärt werden.

Nachhaltige Verfügbarmachung und Langzeitarchivierung: Schon bei der Projektplanung, spätestens bei Projektbeginn, sollte ein geeignetes **Zentrum** identifiziert und Kontakt mit ihm aufgenommen werden (s. [11] und Liste der Zentren in den Referenzen). Das Zentrum muss in der Lage sein, den Bitstream-Erhalt der Daten und ihre Interpretierbarkeit auf lange Zeit zu garantieren und den einfachen Zugang zu den Daten, üblicherweise direkt über das Internet, zu ermöglichen. Idealerweise verfügt das Zentrum über eine explizite Politik der Qualitätssicherung etwa durch externe Organisationen (z. B. das Data Seal of Approval im Verbundprojekt CLARIN).
Mit dem Zentrum sollten konkrete verbindliche, belastbare Abstimmungen und Absprachen über die Datenformate, Standards und ggf. projektspezifische Anpassungen von vorhandenen Standards sowie eine **Mindestaufbewahrungszeit** getroffen werden. Ein solcher Kontakt empfiehlt sich auch dann, wenn die Nachnutzung der Projektdaten nur sehr eingeschränkt oder gar nicht möglich ist.
Sofern im Projekt eine spezifische **Präsentationsschicht** oder -infrastruktur für die Ressource implementiert wird, empfiehlt es sich, auch für diese Langzeitarchivierung und -pflege vorzusehen.

Zitierbarkeit: Sowohl Korpora als auch Tools liegen oftmals in verschiedenen Versionen vor, die sich zum Beispiel durch erweiterte Abdeckung, Korrekturen oder zusätzliche Funktionalitäten unterscheiden. Aus Gründen der **Zitierbarkeit** und der **Replizierbarkeit von Ergebnissen** wird empfohlen, die in einem Projekt erstellten Ressourcen zu versionieren und mit einer entsprechenden Versionsangabe zu veröffentlichen. Auch die in einem Tool ggf. verwendeten, aber **separaten, dynamischen Komponenten** wie Lexika, Tag-Sets, Parameterdateien oder Sprachmodelle sollten versioniert werden, und zwar getrennt von der eigentlichen Tool-Architektur. Für Annotationen bzw. annotierte Korpusversionen wird außerdem empfohlen, ihre Verarbeitungshistorie als **Prozessmetadaten** bzw. *provenance information* nachzuweisen. In diesem Zusammenhang kann es auch sinnvoll sein, Zwischenversionen zu archivieren.

Konkrete Empfehlungen

- Die **Zentren**, die am CLARIN- bzw., in Deutschland, **CLARIN-D**-Projekt teilnehmen und dessen Kriterien erfüllen (s. [11] und Liste der Zentren in den Referenzen), können hier empfohlen werden. Möglicherweise entstehen ähnliche Qualitätsstandards in anderen Infrastrukturnetzwerken wie DARIAH oder META-NET. Auch **internationale Zentren** kommen in Frage, beispielsweise für nicht-deutschsprachige Ressourcen.
- Für Daten aus dem Schulbereich kann auch das Deutsche Institut für Internationale Pädagogische Forschung (DIPF) ein geeigneter Ansprechpartner sein.
- Für die Nachnutzbarmachung von Ressourcen aus Editionsvorhaben empfiehlt es sich, mit den Verlagen bereits **Vorverhandlungen** über Zweit- bzw. Parallelveröffentlichungen zu führen, z. B. über eine **„moving wall“**-Lösung für die Zweitveröffentlichung in einer Korpusinfrastruktur.
- Ein besonders nachhaltiges Verfahren für Versionsangaben ist die Registrierung eines **PID** [19] für jede Ressourcenversion (s. [18] und [20] für die Lösung, die im Projekt CLARIN-D umgesetzt wird).
- Es wird empfohlen, für jede separate Toolkomponente eigene Metadaten zu erstellen, beispielsweise mittels des *ToolCompontentProfile* (einem Template in CMDI)

- Bei Standoff-Annotationen kann jede **Annotationsschicht** als eine separate Ressource betrachtet werden, die versioniert werden sollte und durch einen PID verweisbar gemacht werden sollte.
- Ein empfohlener Ort für die Dokumentation der Verwaltungshistorie von Prozessmetadaten für Korpora ist die Encoding Description des TEI-Headers, s. die entsprechenden Kapitel der TEI P5 Guidelines([23],[24]).

Referenzen zu Teil 2

[9] ISOcat-Registry: [http://www.isocat.org/]

[11] CLARIN-D-Zentren: [http://www.clarin-d.de/de/clarin-d-zentren.html]

[13] Guidelines of the Text Encoding Initiative (TEI) [http://www.tei-c.org/index.xml]

[14] Ron Artstein; Massimo Poesio (2008):. Inter-coder agreement for computational linguistics (survey article). Computational Linguistics 34(4): 555-596.

[15] Michael Beißwenger; Maria Ermakova; Alexander Geyken; Lothar Lemnitzer; Angelika Storrer (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative, issue 3 [http://jtei.revues.org/476]

[16] CLARIN-D-Nutzerhandbuch: [http://media.dwds.de/clarin/userguide]

[17] DFG-Handreichung: Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora

[18] Handle-System der Corporation for National Research Initiatives (CNRI) [http://www.handle.net]

[19] ISO/TC 37 /SC 4 (2011): ISO/FDIS 24619. Language Resource Management – Persistent identification and sustainable access (PISA)

[20] Service der Deutschen Nationalbibliothek zu Persistent Identifiers [http://www.persistent-identifier.de/english/204-examples.php]

[21] Alexander Geyken, Susanne Haaf, Frank Wiegand: *The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora.* In: 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference. Edited by Jeremy Jancsary. Wien, 2012 (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). [online-Version vom 13. September 2012]

[22] Harald Lüngen, Michael Sperberg-McQueen (2012): A TEI P5 Document Grammar for the IDS Text Model. In: Journal of the Text Encoding Initiative (2012), issue 3. [http://jtei.revues.org/508]

[23] TEI P5 Guidelines / Revision Description: [http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-revisionDesc.html]

[24] TEI P5 Guidelines / Encoding Description: [http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-encodingDesc.html].

[25] DTA Basisformat / Header: [http://www.deutschestextarchiv.de/doku/basisformat_header]

[26] Stefanie Rühle: *Kleines Handbuch Metadaten.* Kompetenzzentrum Interoperable Metadaten. [http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Hand-buch/metadaten.pdf?__blob=publicationFile]

[27] TEI CMC SIG: [http://www.tei-c.org/Activities/SIG/CMC/index.xml]

Tools

TreeTagger - http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

annotate - http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html
SALTO - http://www.coli.uni-saarland.de/projects/salsa/page.php?id=software
WebAnno - https://www.ukp.tu-darmstadt.de/software/webanno/

MMAX2 - http://mmax2.sourceforge.net/
TrED - http://ufal.mff.cuni.cz/tred/
WordFreak - http://wordfreak.sourceforge.net/

GATE - http://gate.ac.uk/
NLTK - http://www.nltk.org/
OpenNLP- http://opennlp.apache.org/
UIMA - https://uima.apache.org/

ANNIS - http://www.sfb632.uni-potsdam.de/annis/
TIGERSearch - http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html
CWB/CQP - http://cwb.sourceforge.net/
ICARUS - http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/icarus.en.html
DKPro Core - http://dkpro-core-asl.googlecode.com
DKPro Statistics - http://dkpro-statistics.googlecode.com

Datenzentren in Deutschland mit Kompetenzen für schriftsprachliche Korpora

CLARIN-D-Zentren

BBAW Berlin - Deutsche Sprache, Lexika und diachrone Korpora http://www.deutschestextarchiv.de/
IDS Mannheim - Deutsche Sprache, große Korpora des Deutschen http://www.ids-mannheim.de/kl/
Universität Tübingen - Annotierte Korpora, linguistische Wissenskomponenten und Webservices http://www.sfs.uni-tuebingen.de/
Universität Leipzig - Webservices und spezielle Referenzkorpora http://asv.informatik.uni-leipzig.de/

Universität Stuttgart - Korpora und Korpuswerkzeuge, parametrisierbare Tools und Webservices http://www.ims.uni-stuttgart.de/
MPI Nijmegen - Minoritätensprachen, Multimedia- und multimodale Daten [http://tla.mpi.nl/]
Universität des Saarlandes - Multilinguale Korpora und Korpuswerkzeuge [https://fedora.clarin-d.uni-saarland.de/]

Weitere Zentren und Projekte

DARIAH-Projekt: http://de.dariah.eu/
META-NET-Projekt: http://www.meta-net.eu/
LAUDATIO repository: http://www.laudatio-repository.org/

Metadatenstandards und -registraturen

CMDI - Component MetaData Infrastructure - http://www.clarin.eu/cmdi
CLARIN Component Registry http://catalog.clarin.eu/ds/ComponentRegistry/
IsoCAT - Data Category Registry - http://www.isocat.org/
EAD - Encoded Archival Description
EAC-CPF Encoded Archival Context - Corporate Bodies, Persons, and Families - http://www.loc.gov/ead/, http://eac.staatsbibliothek-berlin.de/
DC - Dublin Core Metadata Initiative - http://dublincore.org/
IMDI - ISLE Meta Data Initiative - http://www.mpi.nl/imdi/
Metadata Encoding & Transmission Standard/Metadata Object Description Schema (METS/MODS) - http://www.loc.gov/standards/mets/, http://www.loc.gov/standards/mods/
OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting - http://www.openarchives.org/pmh/
OLAC - Open Language Archives Community (OLAC) - http://www.language-archives.org/documents.html

Standards

[ALTO] – Analyzed Layout and Text Object.
http://www.loc.gov/standards/alto/

[DTABf] – DTA-Basisformat
http://www.deutschestextarchiv.de/basisformat

[hOCR] – hOCR Embedded OCR Workflow and Output Format.
https://docs.google.com/document/d/1QQnIQtvdAC_8n92-

LhwPcjtAUFwBlzE8EWnKAxlgVf0/preview

[I5] – Customisierung von TEI P5 für DeReKo
http://www1.ids-mannheim.de/kl/projekte/korpora/textmodell.html

[LAF/GrAF] – Ide, N. and Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007, Prague, June 28-29, 1-8. http://www.cs.vassar.edu/~ide/papers/LAW.pdf

[PAULA] – https://www.sfb632.uni-potsdam.de/en/paula.html
Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M. (2008), A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitment automatique des langues* 49, 271-293.

[STTS] – Stuttgart Tübingen Tagset.
Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).
http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

[TIGER-XML] –
http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html
Wolfgang Lezius (2002): Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), 2002, vol. 8, no. 4. IMS, Universität Stuttgart.

[TCF] – http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

Teilnehmerliste DFG-Rundgespräch „Mündliche Korpora“
9. November 2012
Geschäftsstelle der DFG, Kennedyallee 40, Bonn

Professor Dr. Bernt Ahrenholz	Friedrich-Schiller-Universität Jena
Dr. Jörg Bücker	Westfälische Wilhelms-Universität Münster
Professor Dr. Kristin Bührig	Universität Hamburg
Professor Dr. Arnulf Deppermann	Institut für deutsche Sprache, Mannheim
Dr. Sebastian Drude	Max Planck Institute for Psycholinguistics
Dr. Sigrun Eckelmann	DFG, Bonn
Dr. Oliver Ehmer	Freiburg
Professor Dr. Christian Fandrych	Universität Leipzig
Professor Dr. Caroline Féry	Goethe-Universität Frankfurt am Main
Professor Dr. Ulrike Gut	Westfälische Wilhelms-Universität Münster
Professor Dr. Rüdiger Harnisch	Universität Passau
Dr. Dagmar Jung	Universität zu Köln
Professor Dr. Roland Kehrein	Philipps-Universität Marburg
Dr. Kerstin Kucharczik	Ruhr-Universität Bochum
Dr. Christoph Kümmel	DFG, Bonn
Slawomir Messner	Philipps-Universität Marburg
Dr. Gaia di Lucio	Bonn, PT-DLR
Professor Dr. Bernd Meyer	Johannes Gutenberg-Universität Mainz
Ludger Paschen	Ruhr-Universität Bochum
Professor Dr. Stefan Pfänder	Albert-Ludwigs-Universität Freiburg
Dr. Christoph Purschke	Université du Luxembourg
Professor Dr. Uta M. Quasthoff	Technische Universität Dortmund
Professor Dr. Angelika Redder	Universität Hamburg
Dr. Ines Rehbein	Universität Potsdam
Professor Dr. Christian Sappok	Ruhr-Universität Bochum
PD Dr. Florian Schiel	Ludwig-Maximilians-Universität München
Dr. Thomas Schmidt	Institut für deutsche Sprache, Mannheim
Professor Dr. Stavros Skopeteas	Universität Bielefeld
Adriana Slavcheva	Universität Leipzig
Jan Strunk	Köln
Dr. Vera Szöllösi-Brenig	Volkswagen-Stiftung
Professor Dr. Doris Tophinke	Universität Paderborn
Dr. Helga Weyerts-Schweda	DFG Bonn
Professor Dr. Heike Wiese	Universität Potsdam
Dr. Kai Wörner	Universität Hamburg

Teilnehmerliste DFG-Rundgespräch „Textkorpora“
15. November 2013
Geschäftsstelle der DFG, Kennedyallee 40, Bonn

Dr. Noah Bubenhofer	Technische Universität Dresden
Professor Dr. Arnulf Deppermann	Institut für deutsche Sprache, Mannheim (IDS)
Professor Dr. Dagmar Deuber	Westfälische Wilhelms-Universität Münster
Dr. Eva-Maria Dickhaut	Akademie der Wissenschaften und der Literatur Mainz
Professor Dr. Mechthild Habermann	Friedrich-Alexander-Universität Erlangen-Nürnberg
Dr. Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Professor Dr. Thomas Gloning	Justus-Liebig-Universität Gießen
Professor Dr. Iryna Gurevych	Technische Universität Darmstadt
Professor Dr. Ulrich Heid	Stiftung Universität Hildesheim
Professor Dr. Gerhard Heyer	Universität Leipzig
Professor Dr. Erhard W. Hinrichs	Eberhard-Karls-Universität Tübingen
Professor Dr. Martin Huber	Universität Bayreuth
Professor Dr. Magnus Huber	Justus-Liebig-Universität Gießen
Professor Dr. Wolf Peter Klein	Julius-Maximilians-Universität Würzburg
Dr. Marc Kupietz	Institut für deutsche Sprache, Mannheim (IDS)
Professor Dr. Gerhard Lauer	Georg-August-Universität Göttingen
Professor Dr. Christian Mair	Albert-Ludwigs-Universität Freiburg
Professor Dr. Alexander Mehler	Goethe-Universität Frankfurt am Main
Professor Dr. Roland Meyer	Humboldt-Universität zu Berlin
Professor Dr. Manfred Pinkal	Universität des Saarlandes
Dr. Roland Schäfer	Freie Universität Berlin
Professor Dr. Ingrid Schröder	Universität Hamburg
Dr. Silke Schwandt	Goethe-Universität Frankfurt am Main
Professor Dr. Manfred Stede	Universität Potsdam
Professor Dr. Angelika Storrer	Universität Mannheim
Professor Dr. Elke Teich	Universität des Saarlandes
Dr. Helga Weyerts-Schweda	DFG, Bonn
Dr. Stefan Winkler-Nees	DFG, Bonn
Professor Dr. Heike Zinsmeister	Universität Hamburg