Leitfaden zur
Beurteilung von

Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache

Thomas Schmidt, Kai Wörner, Hanna Hedeland, Timm Lehmberg,
Hamburger Zentrum für Sprachkorpora, Universität Hamburg /
Archiv für Gesprochenes Deutsch, IDS Mannheim

Version 1 : September 2013

Inhalt

Einleitung

> „[M]uch [spoken language material] remains in often widely dispersed and inaccessible locations in departmental collections, or, we must admit to our shame, kept in inadequate storage conditions in our own offices, or even at home, gathering dust, wow and flutter, print-through and meltdown, silently shedding the hard-won sounds of twentieth century speech in the constantly dispersing particles of ferric oxide of an obsolete recording system“
>
> [Widdowson 2003:84, zitiert nach Beal 2010:34f]

Korpora gesprochener Sprache werden mindestens seit den 1950er Jahren von Sprachwissenschaftlern und Forschern anderer Disziplinen mit verschiedensten Forschungsinteressen aufgebaut. Die technischen Möglichkeiten für die Erhebung und Bereitstellung solcher Daten haben sich seitdem fortwährend und grundlegend gewandelt. Heute kann es als Normalfall angesehen werden, dass ein Korpus gesprochener Sprache digital erhoben wird. Die wissenschaftliche Community ist außerdem auf dem Wege, sich auf gewisse Mindeststandards zu einigen, die bei der Erhebung bezüglich Dokumentation, Strukturierung und Enkodierung der Daten eingehalten werden sollten, um eine möglichst nachhaltige Nutzung der Korpora zu ermöglichen. Verschiedene Datenzentren schließlich haben sich zum Ziel gesetzt, Korpora gesprochener Sprache zu einer eben solchen Nachnutzung dauerhaft zu archivieren und in digitalen Infrastrukturen bereitzustellen.

Eine der wichtigsten Aufgaben solcher Zentren ist es, Korpora aus abgeschlossenen Projekten zu übernehmen und sie so aufzubereiten, dass eine dauerhafte Archivierung und Bereitstellung überhaupt möglich wird. Dieser Leitfaden basiert auf Erfahrungen, die hinsichtlich dieser Aufgabe an zwei Standorten – dem Sonderforschungsbereich 538 ‚Mehrsprachigkeit' bzw. dem Zentrum für Sprachkorpora (HZSK) an der Universität Hamburg, sowie dem Archiv für gesprochenes Deutsch (AGD) am Institut für Deutsche Sprache in Mannheim – gesammelt wurden.[1]

Am SFB 538 (Laufzeit: 1999-2011) hatte das Projekt Z2 „Computergestützte Erfassungs- und Analysemethoden“ die Aufgabe übernommen, Korpora aus den Teilprojekten des SFB nach deren Abschluss für eine Archivierung und Nachnutzung vorzubereiten (siehe dazu Schmidt/Bennöhr 2007). Die Archivierung und Bereitstellung der Daten im Gesamtumfang von 30 Korpora erfolgt nun im zum Abschluss des SFB (2011) gegründeten HZSK (Hedeland/Lehmberg/Schmidt/Wörner 2011). Das Archiv für Gesprochenes Deutsch bzw. dessen Vorläufer, das Deutsche Spracharchiv (Stift/Schmidt 2014), fungiert bereits seit den 1960er Jahren als eine zentrale Sammelstelle für Korpora des gesprochenen Deutsch. Im Laufe der Jahre hat es aus IDS-internen und -externen Projekten knapp 50 Korpora übernommen, die verschiedene Stadien der Aufbereitung erfahren haben und der wissenschaftlichen Gemeinschaft nun u.a. über die Datenbank für Gesprochenes Deutsch (DGD2, Schmidt/Dickgießer/Gasch 2013) zur Verfügung gestellt werden.

Das derzeitige Angebot dieser beiden Einrichtungen zeigt, dass es prinzipiell möglich ist, von den im einleitenden Zitat beschriebenen Sammlungen zu dauerhaft nachnutzbaren digitalen

[1] Die Konzeption dieses Leitfadens war Gegenstand eines Arbeitspakets im Projekt „Etablierung eines Schwerpunkts ‚Mehrsprachigkeit und Gesprochene Sprache' am Hamburger Zentrum für Sprachkorpora“, das von der Deutschen Forschungsgemeinschaft im Rahmen des Förderprogramms „Literaturversorgungs- und Informationssysteme (LIS)“ gefördert wurde. An der Umsetzung haben sich die genannten MitarbeiterInnen des HZSK und des AGD beteiligt.

Ressourcen zu gelangen. Die Erfahrung zeigt aber auch, dass dies oft ein langwieriger Prozess mit vielen unvorhergesehenen Hindernissen ist, an dessen Ende man sich zumindest gelegentlich die Frage stellen kann, ob Aufwand und Nutzen der Datenaufbereitung in einem angemessenen Verhältnis zueinander stehen.
Zweck dieses Leitfadens ist es, Kriterien für die Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache zu definieren, mittels derer bereits bei der Planung eines entsprechenden Projektes eine Abschätzung der Kosten und Nutzen getroffen werden kann. Kosten bezeichnen in diesem Kontext insbesondere den zeitlichen Arbeitsaufwand, der sich nicht immer leicht in monetäre Kosten umrechnen lässt. Die Nachnutzbarkeit definiert sich vor allem darüber, wie offen oder restriktiv der Zugang zum Korpus gestaltet wird und über die Quantität und Qualität der Korpusbestandteile.
Der Leitfaden gliedert sich in sechs Abschnitte, die in Form von strukturierten Fragebäumen die wichtigsten Eigenschaften einer aufzubereitenden Ressource abfragen. Den Fragebäumen sind Erläuterungen zum besseren Verständnis der einzelnen Fragen vorangestellt. Die Pfade in den Fragebäumen führen jeweils zu einem „Ampelsymbol“, anhand dessen über das weitere Vorgehen bei der Aufbereitung entschieden werden kann. Die Bedeutung der einzelnen Ampelsymbole ist in der folgenden Legende erläutert.

Dies ist der Idealfall, bei dem die Datenübernahme ohne weitere Einschränkungen erfolgen kann.

Hier liegt ein schwerwiegendes Problem bzw. Ausschlusskriterium für die Datenübernahme vor. Kann das Problem vom Datengeber behoben werden, kann später eine erneute Beurteilung stattfinden.

In diesem Fall liegt ein Problem vor, das aber nicht unbedingt eine Datenübernahme ausschließt. Die betreffenden Ressourcen müssen aber ***vor*** der Datenübernahme korrigiert, ergänzt oder aussortiert werden.

Auch wenn auf diese Weise versucht wird, den Prozess der Datenübernahme in einem systematisch strukturierten Ablauf abzubilden, so soll dennoch nicht der Eindruck erweckt werden, dass damit notwendigerweise alle Fragen, die sich bei einer solchen Übernahme stellen, abschließend und erschöpfend geklärt werden. Wir verstehen den Leitfaden als eine Orientierungshilfe für Datengeber und Datennehmer, die helfen soll, den Übernahmeprozess für beide Seiten transparent und planbar zu gestalten. Dies schließt jedoch keinesfalls aus, dass im Einzelfall zusätzliche Kriterien zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit herangezogen werden können.

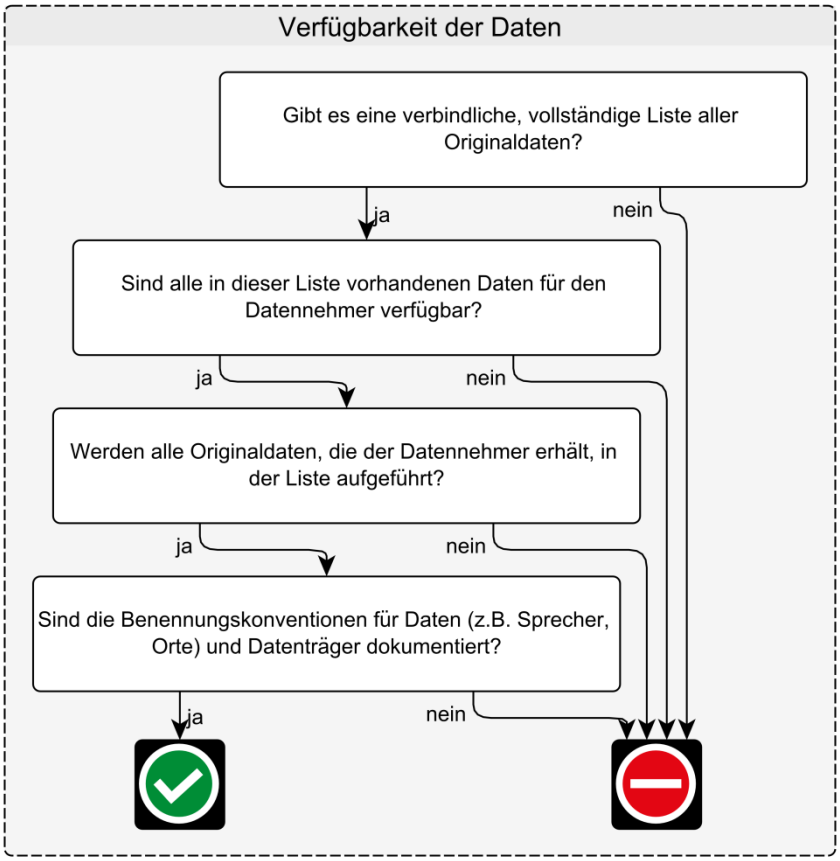
1. Inventarisierung

1.1. Verfügbarkeit der Daten

Um den Aufbereitungsprozess planen zu können, ist es unerlässlich, dass die vorhandenen Original-Daten zunächst vollständig aufgelistet und dem Datennehmer zur Verfügung gestellt werden. Idealerweise erhält der Datennehmer vom Datengeber eine Inventarliste, die mit den tatsächlich vorhandenen Originaldaten abgeglichen ist. Folgende Fälle verursachen hingegen Probleme für die Planung der Aufbereitung:

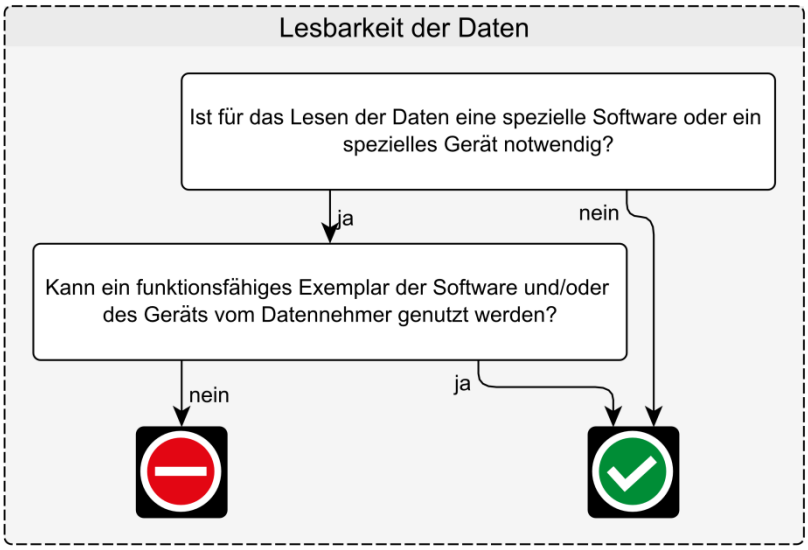
- Es gibt keine Liste der Original-Daten, d.h. der Datennehmer erhält nur die Original-Daten (z.B. als Kiste mit Aufnahmen und Ordnern oder als Dateien auf einer Festplatte) ohne weitere Hilfen zu deren Verständnis
- Ein Datensatz „müsste eigentlich da sein“ (d.h. er befindet sich in der Liste der Original-Daten), lässt sich aber nicht auffinden (d.h. er ist nicht für den Datennehmer verfügbar) oder „muss nachgereicht werden“
- Ein Datensatz „wurde vergessen“ (d.h. er befindet sich nicht in der Liste der Original-Daten), soll aber dennoch bei der Aufbereitung berücksichtigt werden (d.h. er befindet sich unter den Daten, die der Datennehmer erhält)
- Wichtige Metadaten sind ausschließlich den Namen von Dateien zu entnehmen, die Benennungskonventionen sind jedoch nirgendwo dokumentiert.

Wenn solche Probleme bestehen, sollten sie vom Datengeber behoben werden, bevor weitere Schritte in der Aufbereitung in Angriff genommen werden.



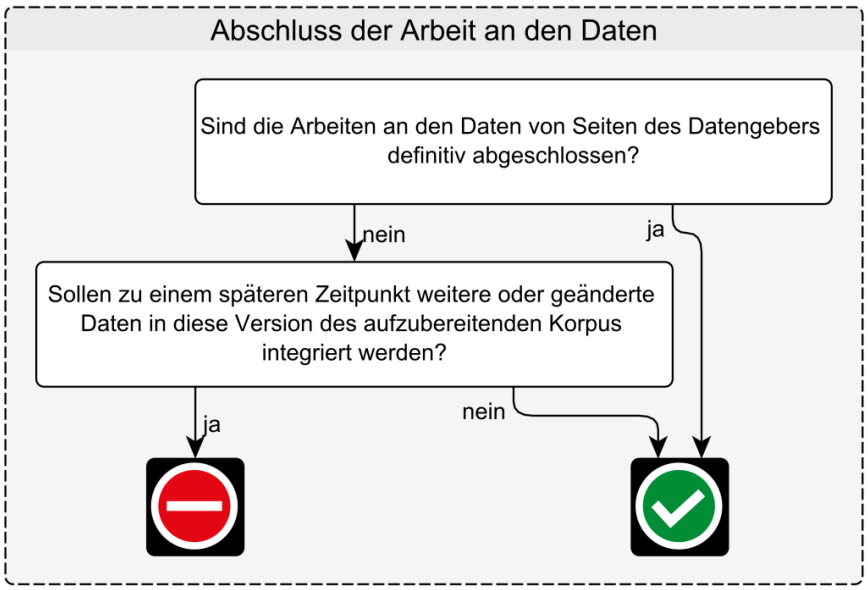
1.2. Lesbarkeit der Daten

Um ältere Korpora verarbeiten zu können, können spezielle Software oder spezielle Geräte notwendig sein.
Unter „spezieller Software“ wird Software zum Lesen von digitalen Metadaten, Transkriptionsdaten und/oder Audio- und Videodaten verstanden, die aktuell nicht mehr (z.B. syncWriter, HIAT-DOS) oder nur kostenpflichtig erhältlich ist (z.B. atlas.ti, Adobe Premiere). Dazu gehören auch spezielle Schriftsätze (z.B. HIAT-Times oder ältere phonetische Schriftsätze). Falls die Daten mit einer solchen Software erstellt wurden, sich aber verlustfrei auch von aktuell verfügbarer, kostenfreier oder allgemein vorhandener Software lesen lassen (z.B. Lesen von Transana-Daten mit MS Word, Lesen von dBase-Daten mit MS Access), kann die Frage 1.2 mit „Nein“ beantwortet werden.
Unter speziellen Geräten werden alle Geräte verstanden, die zum Lesen von analogen Audio-Datenträgern (z.B. Kompaktkassetten, Reel-to-Reel-Tonbänder), analogen Video-Datenträgern (z.B. VHS-Kassetten, Reel-to-Reel-Videobänder, Super-8-Filme), aber auch nicht mehr geläufigen digitalen Audio-Datenträgern (z.B. DAT-Bänder, MiniDiscs) oder Video-Datenträgern (z.B. MiniDV) notwendig sind.
Für Daten, die sich nur mit Hilfe solcher spezieller Software oder Geräte lesen lassen, muss der Datennehmer über mindestens ein funktionsfähiges Exemplar der Software bzw. des Gerätes verfügen, um die Daten bearbeiten zu können.
Bei Software muss für die Funktionsfähigkeit neben einer Kopie und Lizenz der Software selbst auch eine Umgebung zur Verfügung stehen, in der die Software ausgeführt werden kann – z.B. wird für das Lesen von syncWriter-Daten ein Apple-Rechner benötigt, auf dem entweder das Betriebssystem Mac OS 9 läuft oder dieses in der sog. „Classic“-Umgebung in frühen Versionen von Mac OS X emuliert werden kann.
Bei Geräten ist neben der Funktionsfähigkeit als solcher auch zu klären, ob es Möglichkeiten gibt, das Gerät mit einem Computer zu verbinden. Bei Audio-Geräten genügt dafür in der Regel ein geeignetes Kabel, das den Ausgang des Geräts mit dem Eingang einer Soundcard am Rechner verbindet (z.B. ein 2xCinch-auf-kleine-Klinke-Kabel zum Anschluss eines Tapedecks). Bei Video-Geräten können zusätzlich ein Analog-Digital-Wandler und eine zugehörige Software nötig sein (z.B. Pinnacle Studio mit zugehöriger Video-Capture-Hardware zum Anschluss eines VHS-Players).



1.3. Abschluss der Arbeit an den Daten

Es ist in der Regel nicht möglich, die Aufbereitung eines (Teil)-Korpus mit laufenden Erweiterungs- oder Änderungsarbeiten an eben diesem Korpus zu koordinieren. Wenn also von Seiten des Datengebers weitere Bearbeitungen (z.B. Annotation, Qualitätskontrolle) des vorhandenen Datenbestandes geplant sind und/oder der Datenbestand von Seiten des Datengebers in Zukunft noch (z.B. durch neue Aufnahmen und Transkriptionen) erweitert werden soll, macht dies deutlich mehr Planungsarbeit erforderlich. Der Datengeber sollte daher zusagen, dass entweder von seiner Seite keine weiteren Arbeiten an den Daten mehr geplant sind, oder dass ausdrücklich nicht beabsichtig ist, die Ergebnisse der noch geplanten Arbeiten in diese Version des aufbereiteten Korpus einfließen zu lassen.



2. Rechtliche Aspekte

Grundsätzlich ist festzustellen, dass aufgrund der Vielschichtigkeit linguistischer Daten keine allgemeingültige Klärung aller rechtlichen Aspekte erfolgen kann. Diese muss grundsätzlich im Einzelfall erfolgen. Der folgende Abschnitt gibt daher nur grundlegende Empfehlungen für die Evaluation der rechtlichen Aspekte bei der Nutzung und Weitergabe von Korpora gesprochener Sprache. Dabei soll zwischen datenschutzrechtlichen (2.1) und urheberrechtlichen (2.2) Bestimmungen unterschieden werden.

2.1. Datenschutz

Fragen des Datenschutzes werden in Deutschland durch die europäische Gesetzgebung, das Bundesdatenschutzgesetz (BDSG) und die jeweiligen Landesdatengesetze (LDSG) geregelt (siehe Bücker et a. 2013). Die darin enthaltenen Bestimmungen sind grundsätzlich zu beachten, wenn ein Korpus *personenbezogene Daten* enthält, also Daten, die eindeutig einer bestimmten natürlichen Person zugeordnet sind oder bei denen diese Zuordnung zumindest

mittelbar erfolgen kann. Eine *Nutzung* und *Weitergabe* dieser Daten ohne Zustimmung der betroffenen Personen (zumeist Probanden) ist in der Regel nicht zulässig. Eine Regelung dieses Umstandes im Einzelfall entscheidet somit wesentlich über die Nachnutzbarkeit der jeweiligen Ressource.

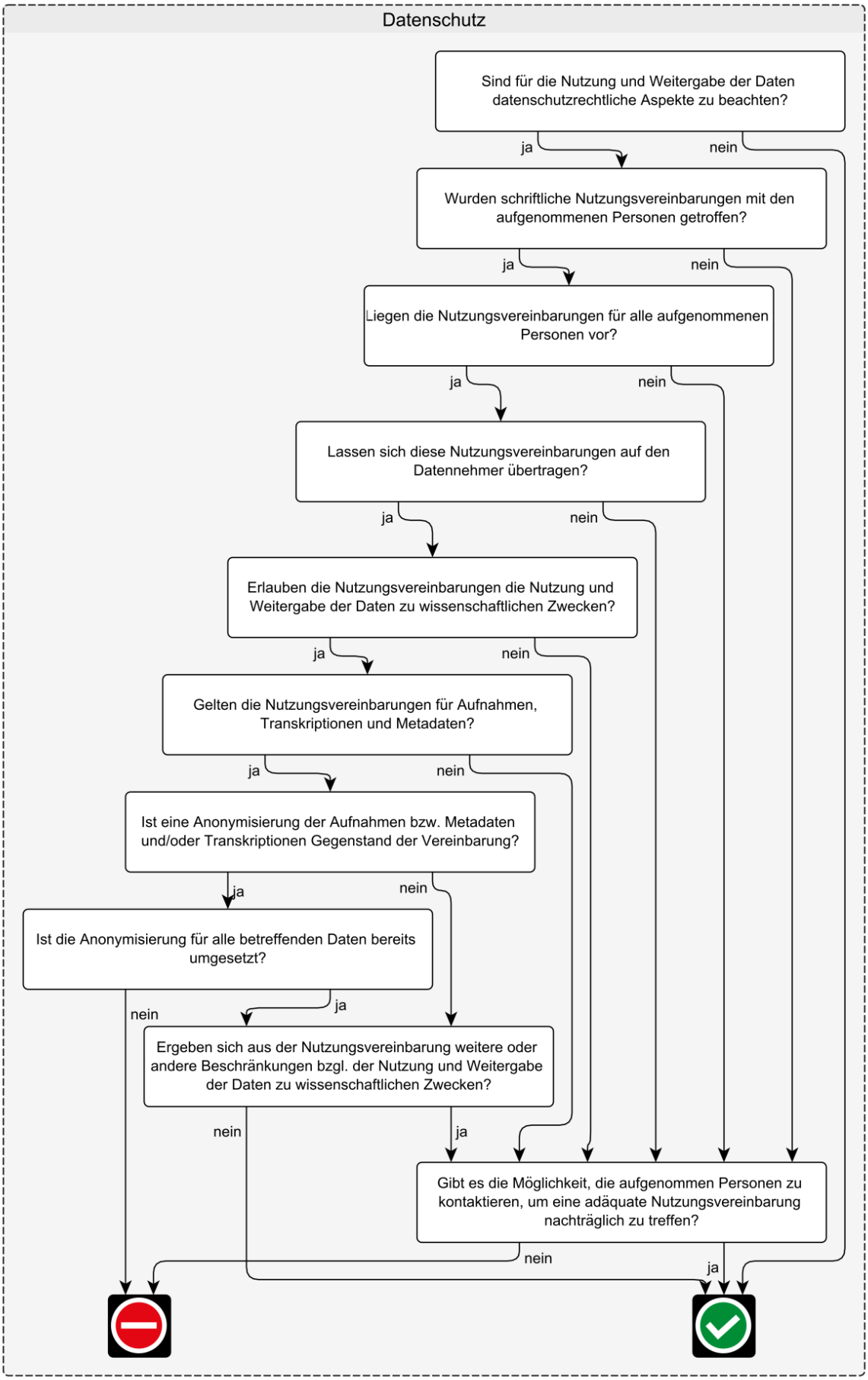
Eine Regelung erfolgt zumeist mithilfe von Nutzungsvereinbarungen zwischen den wissenschaftlichen Einrichtungen, die die Erhebung vorgenommen haben und den jeweiligen Probanden. In den Vereinbarungen, die den erforderlichen formalen Kriterien (Unterschrift eines Vertreters (ggf. des Leiters) der wissenschaftlichen Einrichtung, Widerrufsklausel etc.) genügen, sind Umfang und Zweck der Nutzung und Weitergabe[2], sowie ggf. Art und Umfang einer Anonymisierung beschrieben. Im Fall linguistischer Korpora ist es grundsätzlich empfehlenswert, eine *nicht-kommerzielle* Nutzung und Weitergabe zu ausschließlich *wissenschaftlichen Zwecken* zu vereinbaren.

In der Praxis treten häufig die folgenden Problemfälle auf:

- Es liegen generell keine schriftlichen Nutzungsvereinbarungen vor.
- Schriftliche Vereinbarungen liegen nur für einen Teil der aufgenommenen Personen (z.B. für Probanden oder Gewährspersonen, nicht aber für zufällig anwesende Dritte, die am Gespräch teilhaben) vor.
- Die Vereinbarungen sind ausdrücklich mit Bezug auf die Nutzung im Rahmen des Erhebungsprojektes getroffen worden, d.h. der Fall, dass die Daten auch in einem anderen Zusammenhang genutzt bzw. weitergegeben werden, ist nicht geregelt oder sogar explizit ausgeschlossen.

In diesen Fällen ist eine Nutzung und Weitergabe der Daten durch und an Dritte in nicht-anonymisierter Form (s.u.) nicht möglich. Eine Aufbereitung der Daten sollte erst beginnen, wenn (ggf. durch nachträgliches Einholen von Nutzungsvereinbarungen) eine Klärung herbeigeführt wurde. Eine Möglichkeit des Umgehens datenschutzrechtlicher Probleme stellt die vollständige Anonymisierung, also das Löschen/Maskieren aller personenbezogenen Inhalte (bei gesprochen sprachlichen Daten also ggf. ganzer Gesprächsabschnitte und streng genommen auch aller Audioaufnahmen) dar. Es ist jedoch zu erwägen, ob die jeweiligen Ressourcen dann noch Attraktivität für eine wissenschaftliche Analyse besitzen.

[2] Die *Nutzung* und *Weitergabe* kann auch eine *Veröffentlichung* der Daten bedeuten, ist jedoch nicht damit gleichzusetzen.

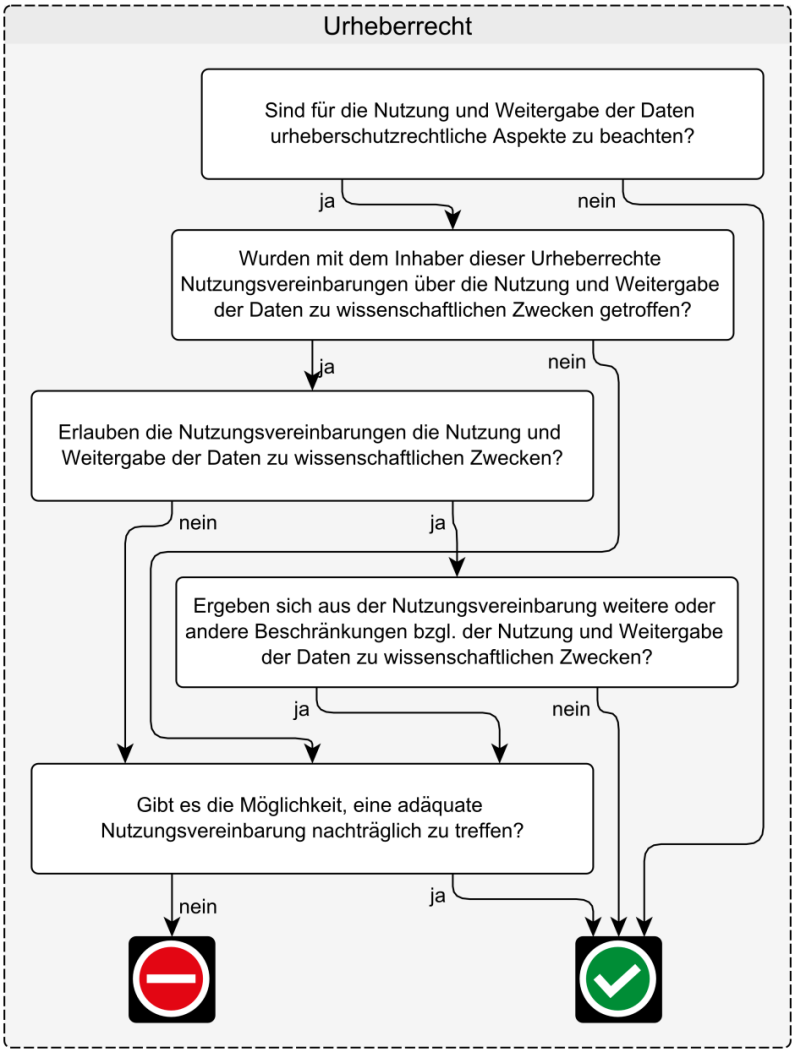

Datenschutz
Sind für die Nutzung und Weitergabe der Daten datenschutzrechtliche Aspekte zu beachten?
ja
nein
Wurden schriftliche Nutzungsvereinbarungen mit den aufgenommenen Personen getroffen?
ja
nein
Liegen die Nutzungsvereinbarungen für alle aufgenommenen Personen vor?
ja
nein
Lassen sich diese Nutzungsvereinbarungen auf den Datennehmer übertragen?
ja
nein
Erlauben die Nutzungsvereinbarungen die Nutzung und Weitergabe der Daten zu wissenschaftlichen Zwecken?
ja
nein
Gelten die Nutzungsvereinbarungen für Aufnahmen, Transkriptionen und Metadaten?
ja
nein
Ist eine Anonymisierung der Aufnahmen bzw. Metadaten und/oder Transkriptionen Gegenstand der Vereinbarung?
ja
nein
Ist die Anonymisierung für alle betreffenden Daten bereits umgesetzt?
nein
ja
Ergeben sich aus der Nutzungsvereinbarung weitere oder andere Beschränkungen bzgl. der Nutzung und Weitergabe der Daten zu wissenschaftlichen Zwecken?
nein
ja
Gibt es die Möglichkeit, die aufgenommen Personen zu kontaktieren, um eine adäquate Nutzungsvereinbarung nachträglich zu treffen?
nein
ja

2.2. Urheberrecht

Fragen des Urheberrechts werden in Deutschland durch das Urheberrechtsgesetz (UrhG) geregelt. Als Urheberrecht wird allgemein das Recht eines Urhebers (Autors, Künstlers, Softwareentwicklers etc.) an dem von ihm geschaffenen *Werk* verstanden. Es kann nicht übertragen oder veräußert werden und erlischt erst 70 Jahre nach dem Tod des Urhebers. Ein urheberrechtlich geschütztes Werk liegt jedoch erst vor, wenn eine gewisse (nicht näher vom Gesetzgeber definierte) *Schöpfungshöhe* in Form von kreativer Eigenleistung in die Entstehung eines Werks eingeflossen ist.
Im Falle der Nutzung und Weitergabe von Korpora gesprochener Sprache können diesbezüglich die folgenden Problemfälle auftreten:

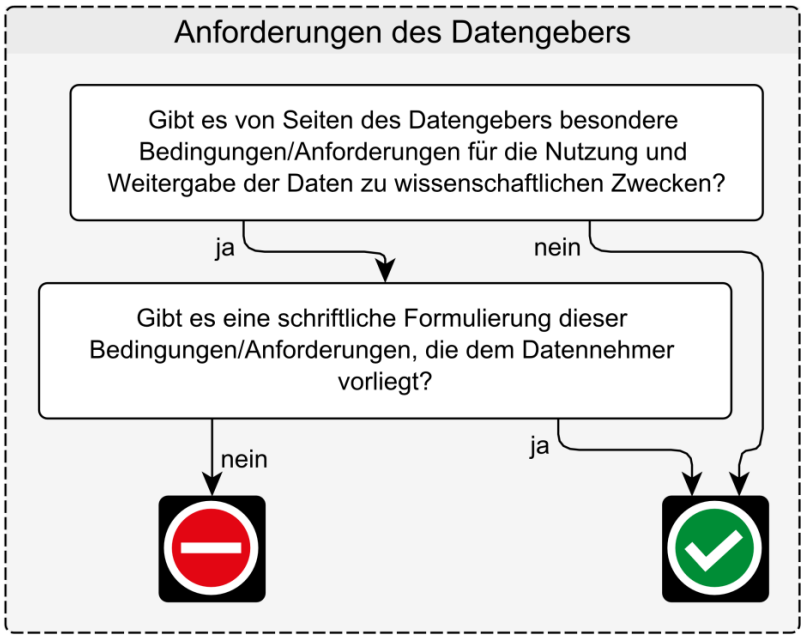
- Im Rahmen einer Erhebung wird urheberrechtlich geschütztes Material (Vortragsfolien, Musik, Videosequenzen u.Ä.) als Stimulus verwendet oder (auch beiläufig) mit aufgezeichnet.
- Die Schöpfer eines Korpus machen eigene Urheberrechte an den Datenstrukturen oder für den Datenzugriff erforderlichen Werkzeugen geltend, sodass keine Nutzung und Weitergabe der Daten durch Dritte möglich ist.

Beide Fälle können ebenfalls durch Nutzungsvereinbarungen, in denen eine Zweckbindung vereinbart wird (vgl. 2.1), oder den Erwerb von Nutzungsrechten mithilfe einer Lizensierung gelöst werden.



2.3. Anforderungen des Datengebers

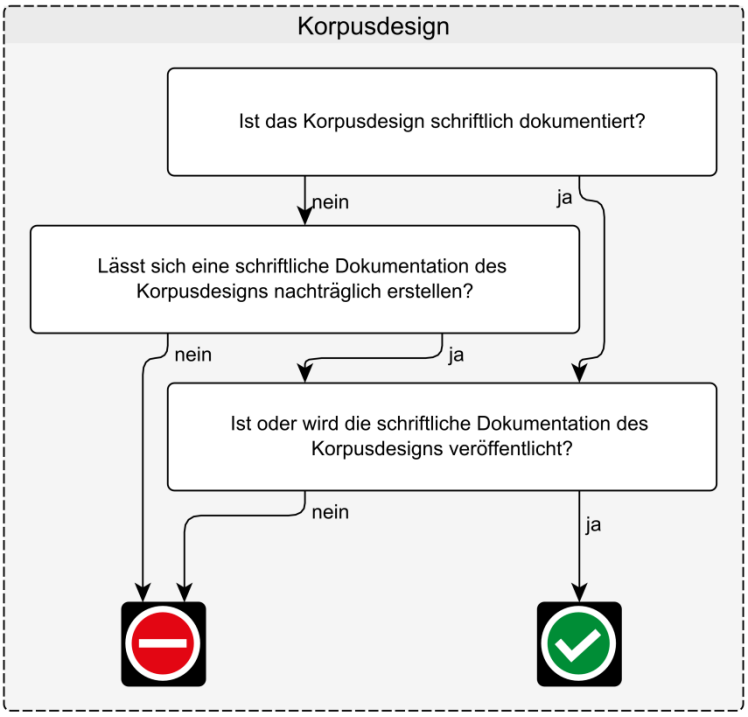
Idealerweise überträgt der Datengeber dem Datennehmer alle Nutzungsrechte, die sich aus den beiden vorgenannten Punkten ergeben und befugt ihn darüber hinaus, diese auch an Dritte weiterzugeben. Dieses Vorgehen muss jedoch durch geltendes Recht abgedeckt sein.
Oft knüpfen Datengeber die Weitergabe der Daten aber an zusätzliche Bedingungen, die weder durch datenschutz- noch durch urheberrechtliche Erwägungen motiviert sind. Das häufigste Beispiel hierfür ist, dass der Datengeber verlangt, bei jeder Anfrage zur Nutzung der Daten informiert zu werden, und die Daten vom Datennehmer nur mit ausdrücklicher Zustimmung des Datengebers Dritten zur Verfügung gestellt werden dürfen. Als häufigste Gründe hierfür werden eine besondere Sensibilität der Daten und/oder der Wunsch angeführt, keine Konkurrenz zu eigenen noch ausstehenden Forschungsarbeiten an den Daten zu schaffen.
Die Erfahrung zeigt, dass solche zusätzlichen Bedingungen sich in der Praxis oft als problematisch erweisen. Erstens verursacht das Weiterleiten von Anfragen an den Datengeber einen nicht unerheblichen organisatorischen Aufwand. Zweitens werden Zugänge zu den Daten unter solchen Bedingungen oft nur sehr restriktiv, z.B. nur oder bevorzugt an Personen, die dem Datengeber persönlich bekannt sind, erteilt. Ersteres erhöht also den Verwaltungsaufwand, letzteres verringert die Nachnutzbarkeit des betreffenden Korpus.
Wenn möglich, sollten solche zusätzlichen Bedingungen daher vermieden werden. Ist dies nicht möglich, ist unbedingt darauf zu achten, dass sie in möglichst konkreter und expliziter Form schriftlich festgehalten werden, bevor die Aufbereitung begonnen wird, und dass Regelungen für den Fall getroffen werden, dass der Datengeber Anfragen zu einem bestimmten Zeitpunkt nicht (mehr) beantworten kann.

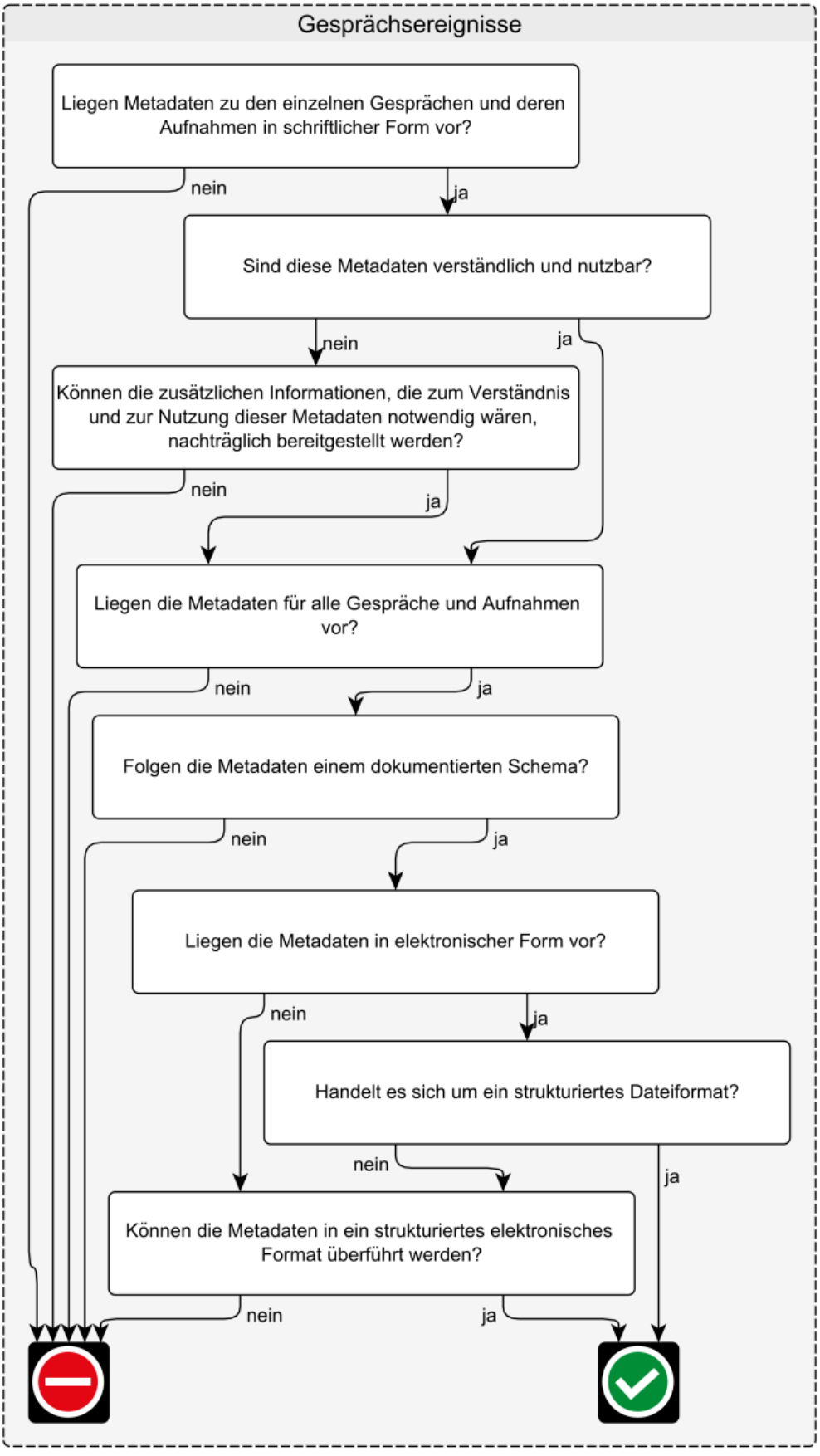


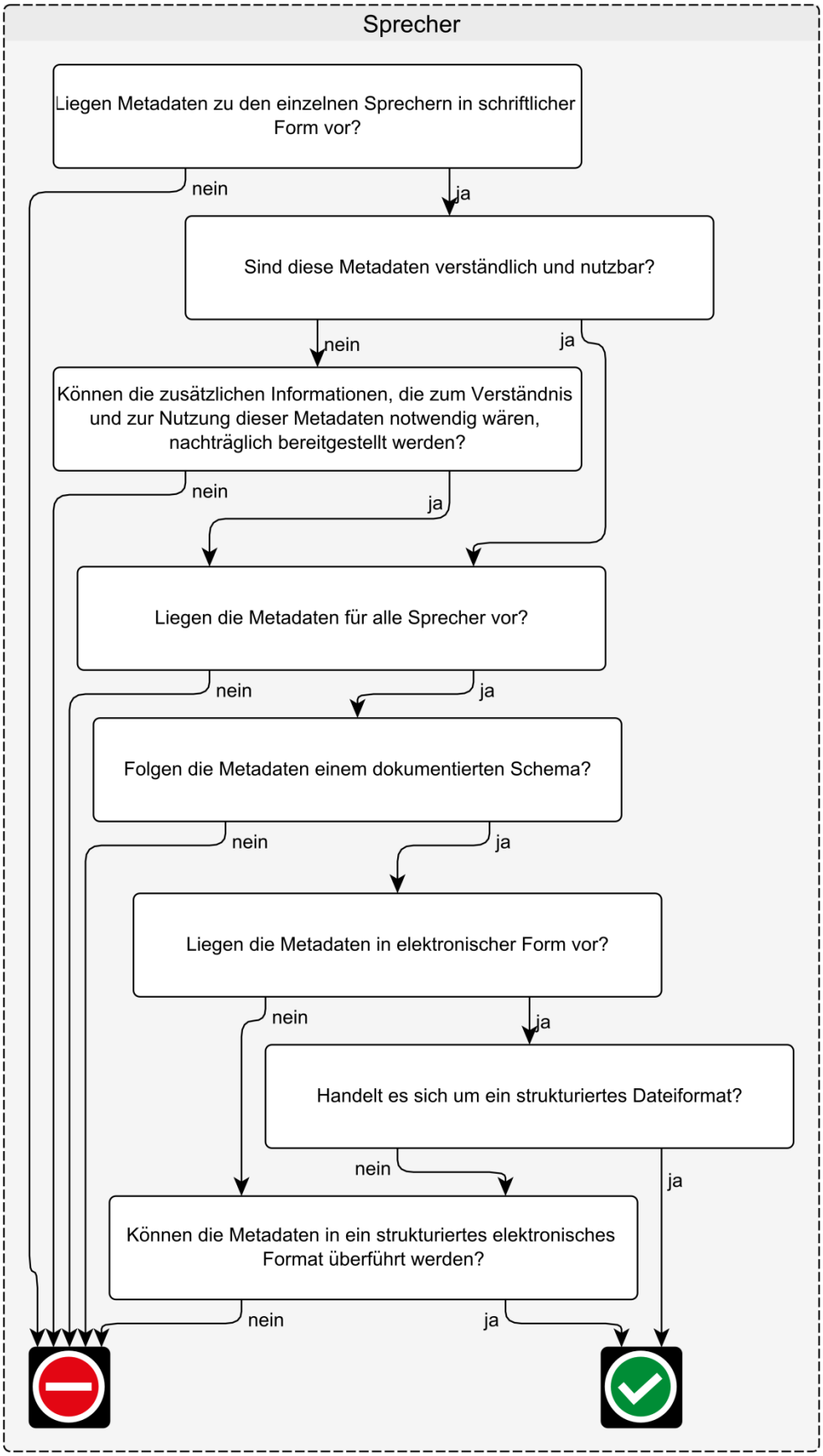
3. Metadaten

Eine sinnvolle Nachnutzung der Daten ist nicht ohne eine grundständige Dokumentation des Korpusdesigns, der Gesprächsereignisse und der daran beteiligten Sprecher möglich. Idealerweise liegen eine Beschreibung des Korpusdesigns in Form einer regulären wissenschaftlichen Veröffentlichung und detaillierte Metadaten in strukturierter elektronischer Form (bspw. als Excel-oder XML-Dateien) vor. Häufige Problemfälle sind:

- Es gibt keine oder keine ausreichende Dokumentation des Korpus, der Gesprächsereignisse und/oder der Sprecher. In diesem Falle muss die Dokumentation entweder nachträglich vom Datengeber vorgenommen werden, oder der Datennehmer versucht, die Dokumentation aus den vorliegenden Daten zu rekonstruieren (bspw. Daten zu Sprechern aus den Gesprächsinhalten abzuleiten). Beides ist mit hohem Aufwand verbunden.
- Die im Dokumentationsschema vorgesehene Information ist ausreichend, wurde bei der tatsächlichen Dokumentation aber nur lückenhaft festgehalten. Auch in diesem Falle ist die notwendige nachträgliche Ergänzung der Metadaten mit hohem Aufwand verbunden.
- Ausreichende und vollständige Metadaten liegen vor, wurden aber nur in nicht-strukturierter Form (z.B. in Word-Dateien) oder in nicht-elektronischer Form (d.h. auf Papier) festgehalten. In diesen Fällen muss eine strukturierte, digitale Fassung der Metadaten konzipiert und erstellt werden, was in der Regel möglich, aber auch mit nicht zu vernachlässigendem Aufwand verbunden ist.
- Es liegen ausreichende, vollständige, strukturierte Metadaten vor, diese sind aber nicht mehr vollständig interpretierbar, weil z.B. gewisse Eigenschaften in Zahlencodes o.Ä. festgehalten wurden und der zugehörige Kodierungsschlüssel nicht mehr verfügbar ist. Auch in diesem Fall muss eine mit entsprechendem Aufwand verbundene Rekonstruktion der betreffenden Daten versucht werden.



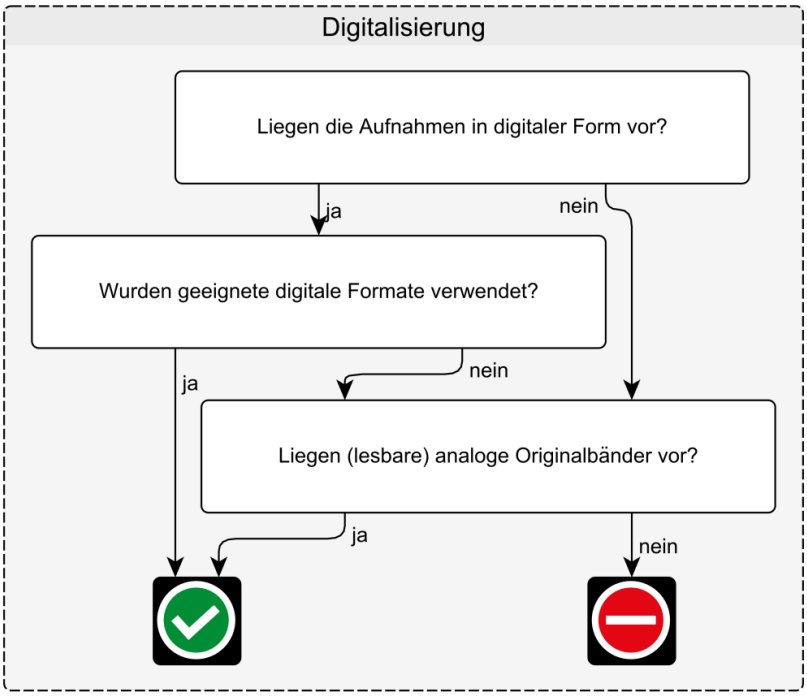

Gesprächsereignisse
Liegen Metadaten zu den einzelnen Gesprächen und deren Aufnahmen in schriftlicher Form vor?
nein
ja
Sind diese Metadaten verständlich und nutzbar?
nein
ja
Können die zusätzlichen Informationen, die zum Verständnis und zur Nutzung dieser Metadaten notwendig wären, nachträglich bereitgestellt werden?
nein
ja
Liegen die Metadaten für alle Gespräche und Aufnahmen vor?
nein
ja
Folgen die Metadaten einem dokumentierten Schema?
nein
ja
Liegen die Metadaten in elektronischer Form vor?
nein
ja
Handelt es sich um ein strukturiertes Dateiformat?
nein
ja
Können die Metadaten in ein strukturiertes elektronisches Format überführt werden?
nein
ja


Sprecher
Liegen Metadaten zu den einzelnen Sprechern in schriftlicher Form vor?
nein
ja
Sind diese Metadaten verständlich und nutzbar?
nein
ja
Können die zusätzlichen Informationen, die zum Verständnis und zur Nutzung dieser Metadaten notwendig wären, nachträglich bereitgestellt werden?
nein
ja
Liegen die Metadaten für alle Sprecher vor?
nein
ja
Folgen die Metadaten einem dokumentierten Schema?
nein
ja
Liegen die Metadaten in elektronischer Form vor?
nein
ja
Handelt es sich um ein strukturiertes Dateiformat?
nein
ja
Können die Metadaten in ein strukturiertes elektronisches Format überführt werden?
nein
ja

4. Aufnahmen

4.1. Digitalisierung

Eine Archivierung und Bereitstellung von Audio- oder Videoaufnahmen erfolgt heute grundsätzlich in digitaler Form. Für den Fall, dass die Aufnahmen nur in analoger Form (beispielsweise Audiodaten auf Kompaktkassetten oder Tonbändern, Videodaten auf VHS-Bändern) vorliegen, müssen sie daher für eine Nachnutzung digitalisiert werden. Eine (erneute) Digitalisierung auf Seiten des Datennehmers kann auch notwendig sein, wenn die Qualität einer vom Datengeber bereits durchgeführten Digitalisierung nicht ausreichend ist. Für die Archivierung werden in aller Regel unkomprimierte Formate (z.B. PCM WAV bei Audio) verwendet. Für digitale Daten, die (nur) in komprimierter Form vorliegen (z.B. MP3 bei Audio) reduziert sich möglicherweise die Nachnutzbarkeit der Ressource, beispielsweise sind bestimmte phonetische Untersuchungen an verlustbehaftet komprimierten Audiodaten nicht möglich.

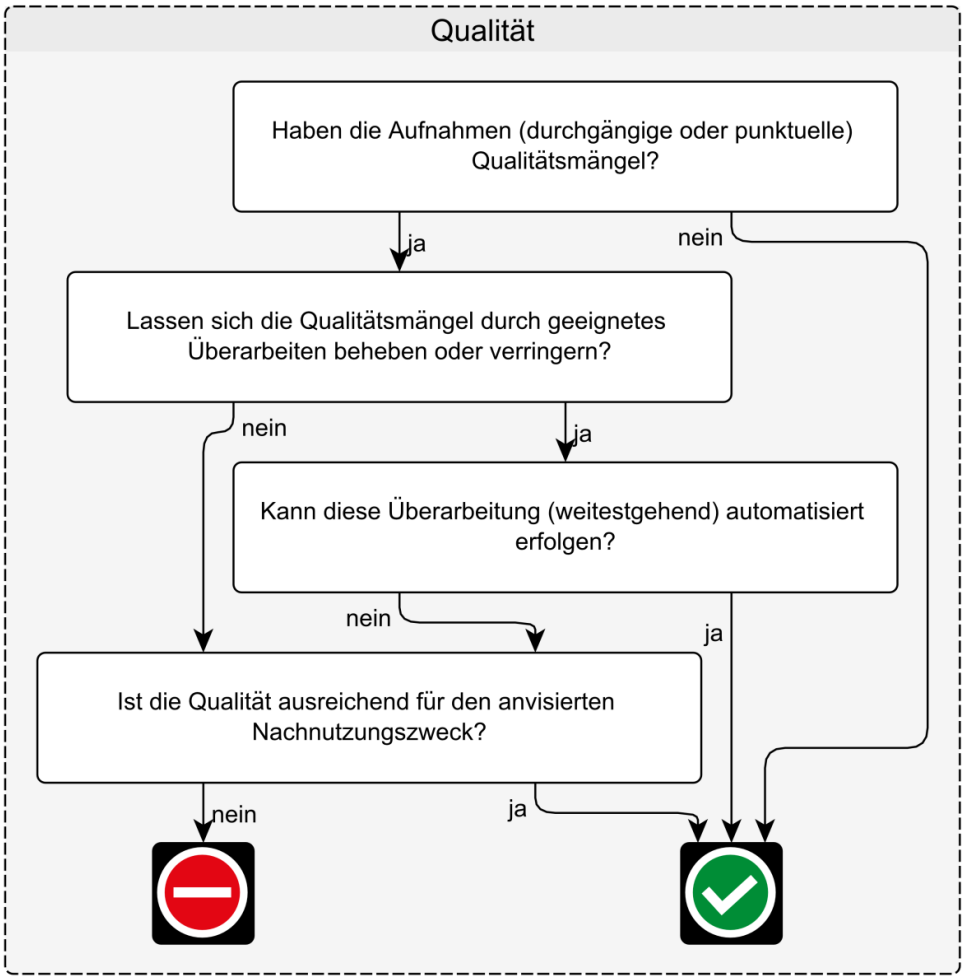


4.2. Qualität

Insbesondere bei Aufnahmen unter nicht vollständig kontrollierten Bedingungen („im Feld“) kommt es – unabhängig von den technischen Grundparametern der Aufnahme – häufig zu Qualitätsmängeln wie Über- oder Untersteuerung, Rauschen, Störgeräuschen (im Audio) bzw. Fehlbelichtungen, Verwackeln oder Fehlfokussierungen (im Video). Der Nachnutzungswert einer Ressource hängt auch davon ab, wie häufig und ausgeprägt solche Qualitätsmängel in den Aufnahmen sind. Gewisse (durchgängige) Qualitätsmängel (wie eine generelle Untersteuerung von Audioaufnahmen) lassen sich weitestgehend automatisiert

ermitteln und beheben. Ist eine manuelle Beurteilung und Behebung notwendig, so erhöht dies den Aufwand für die Aufbereitung i.d.R. beträchtlich.
Eine definitive Beurteilung und ggf. Bearbeitung solcher Mängel kann i.d.R. erst durch spezialisiertes Personal beim Datennehmer erfolgen. Zur Einschätzung des zu erwartenden Aufwandes ist eine detaillierte Einschätzung des Datengebers zu diesem Punkt dennoch sehr nützlich.

Falls zu einem Korpus keine Aufnahmen (mehr) vorliegen oder die vorhandenen Aufnahmen sich nicht auf einen ausreichenden Qualitätsstandard bringen lassen, reduziert das den Nachnutzungswert der Ressource i.d.R. erheblich. Dennoch kann es in Einzelfällen sinnvoll sein, eine Ressource auch ohne die zugehörigen Audio- oder Videodaten für eine Nachnutzung zugänglich zu machen.



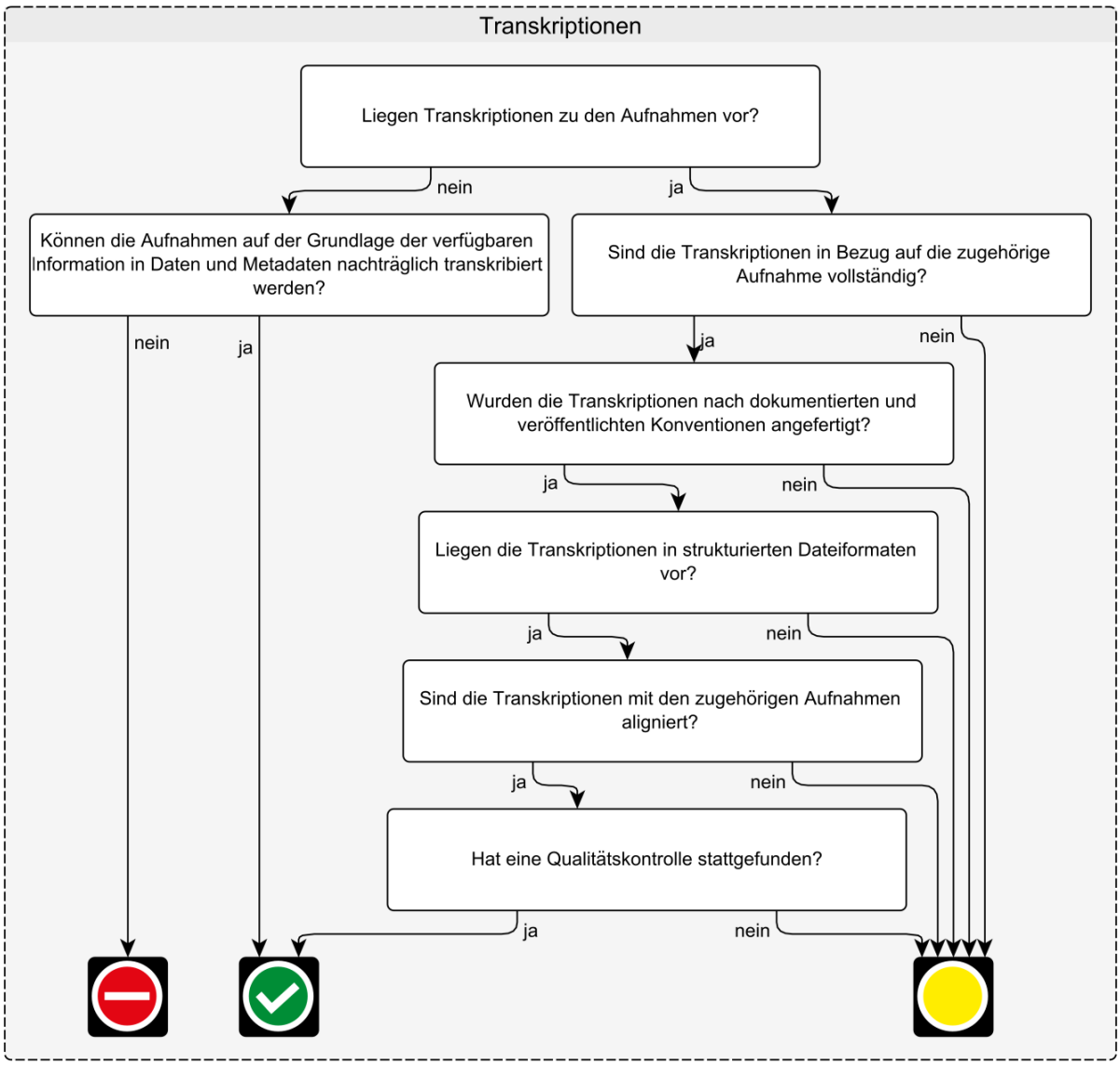
5. Transkriptionen und Annotationen

5.1. Transkriptionen

Transkriptionen haben als Sekundärdaten prinzipiell einen anderen Status als die in den vorhergehenden Abschnitten behandelten Primärdaten. Die Archivierung und Bereitstellung einer Ressource kann grundsätzlich auch sinnvoll sein, wenn solche Sekundärdaten gar nicht vorliegen, da sie im Prinzip auf Grundlage der Primärdaten auch noch nachträglich erstellt werden können.
Dennoch können Transkriptionen als essentiell für den Nachnutzungswert einer Ressource angesehen werden, denn nur durch sie wird eine detaillierte inhaltliche Erschließung des Korpus möglich. Den Wert einer Transkription kann man nicht zuletzt an der in sie investierten Zeit bemessen – typischerweise macht die Transkription den größten Anteil der Arbeit bei der Korpuserstellung aus. Idealerweise sind die zu einem Korpus gehörigen Aufnahmen also zum Abschluss eines Projekts vollständig transkribiert. Damit die Transkription durch den Datennutzer (oder auch für computerlinguistische Tools) interpretierbar ist, müssen die verwendeten Transkriptionskonventionen angemessen dokumentiert sein. Die Nachnutzung wird weiterhin deutlich erleichtert, wenn Transkription und Aufnahme durch geeignete Zeitmarken im Transkript verknüpft („aligniert“) sind. Dies ist in der Regel der Fall, wenn die Transkription mit entsprechenden spezialisierten Tools (Praat, FOLKER, EXMARaLDA, ELAN etc.) vorgenommen wurde.
Folgende Problemfälle treten häufig auf:

- Es liegen Transkriptionen vor, diese sind aber in Bezug auf die jeweiligen Aufnahmen nicht vollständig, weil nur auszugsweise oder nur Redebeiträge bestimmter Sprecher transkribiert wurde.
- Die Transkriptionen folgen einer „ad-hoc“-Konvention, die nicht oder nicht ausreichend dokumentiert und folglich auch nicht konsistent angewendet wurde.
- Die Transkriptionen liegen in unstrukturierten und/oder präsentationsorientierten Dateiformaten (z.B. RTF oder DOC) vor.

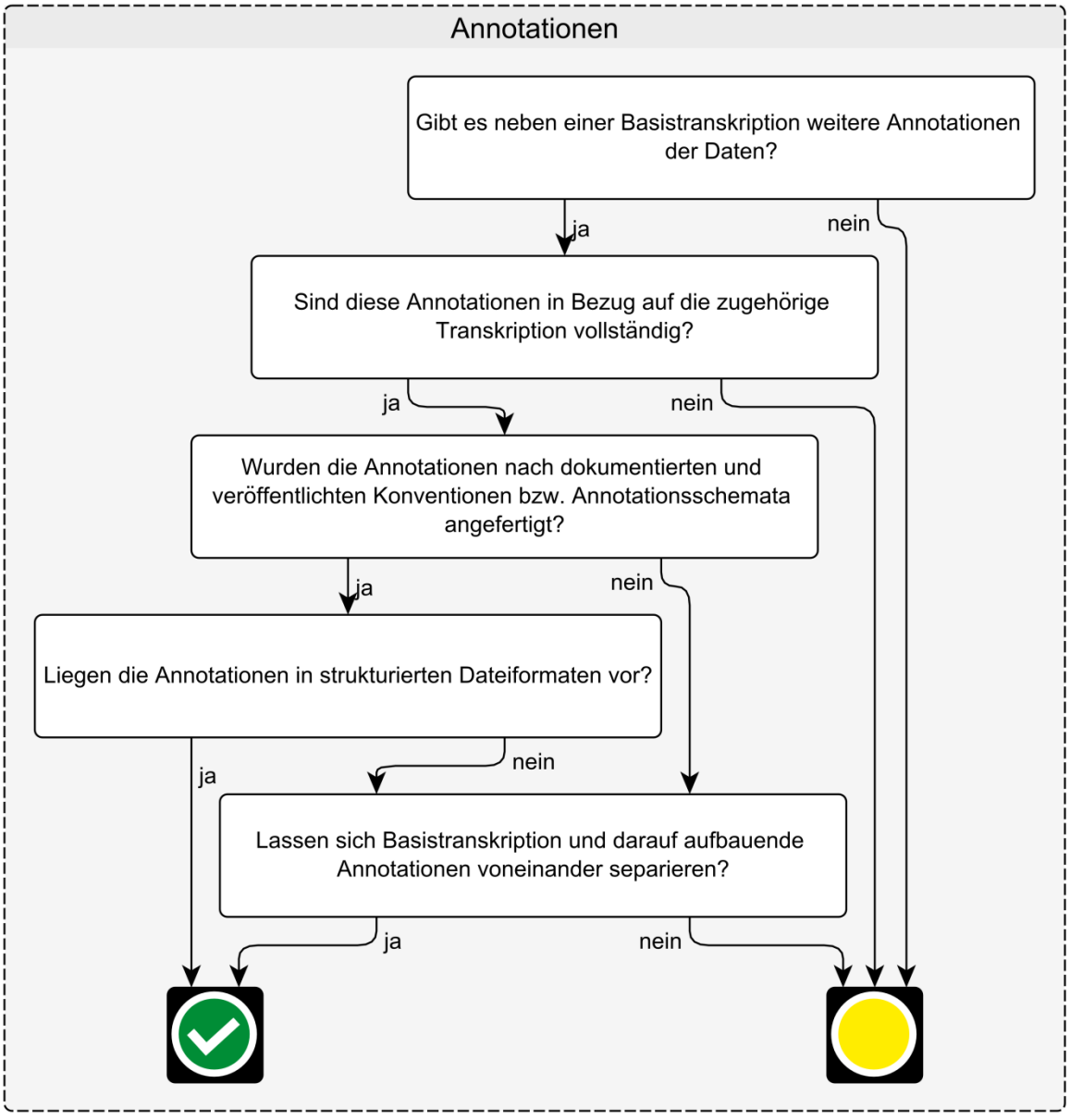
In vielen Fällen kann die Bereitstellung nicht oder nur zum Teil transkribierter Aufnahmen im Übrigen ein erfolgversprechender Weg sein, ein Korpus nachträglich durch – von Datennutzern bereitgestellte – Transkriptionen zu ergänzen. Dabei ist allerdings zu beachten, dass für eine Nach-Transkription teilweise Kenntnisse nötig sein können, die für nicht unmittelbar am Erstellungsprojekt beteiligte Personen nur schwer zu erlangen sind. So ist etwa die Zuordnung von Redeteilen zu bestimmten Sprechern in Mehrpersonengesprächen oft nur für Personen möglich, die bei der Aufnahme anwesend waren oder mit den beteiligten Sprechern anderweitig ausreichend bekannt sind, um sie anhand ihrer Stimmen durchgängig zu identifizieren. Weitere Hindernisse für eine Nach-Transkription außerhalb des Erstellungsprojekts können stark dialektale oder ideolektale Sprechweisen sein.



5.2. Annotationen

Unter Annotationen verstehen wir über die basale Transkription hinausgehende analytische Information, die den Daten hinzugefügt wurde. Annotationen können äußerst vielfältiger Natur sein – sie reichen von prosodischen Markierungen über morphologische Annotationen (z.B. POS-Tagging) bis zu semantisch-pragmatischen Informationen (z.B. Sprechaktannotationen). Mehr noch als die Transkription sind Annotationen oft an spezifische Fragestellungen und/oder spezifische theoretische Herangehensweisen gebunden. Insofern können sie einerseits äußerst nützlich sein, wenn eine Nachnutzung angestrebt wird, die sich nahe an den Fragestellungen und Herangehensweisen des Ursprungsprojekts befindet. Andererseits können sie im Fall, dass bei einer Nachnutzung deutlich andere Fragestellungen und Herangehensweisen verfolgt werden, auch hinderlich sein. Bei der Datenaufbereitung wird daher in der Regel versucht, Basistranskription und darauf aufbauende Annotationen so zu separieren, dass sie unabhängig voneinander für eine Nachnutzung zur Verfügung stehen.

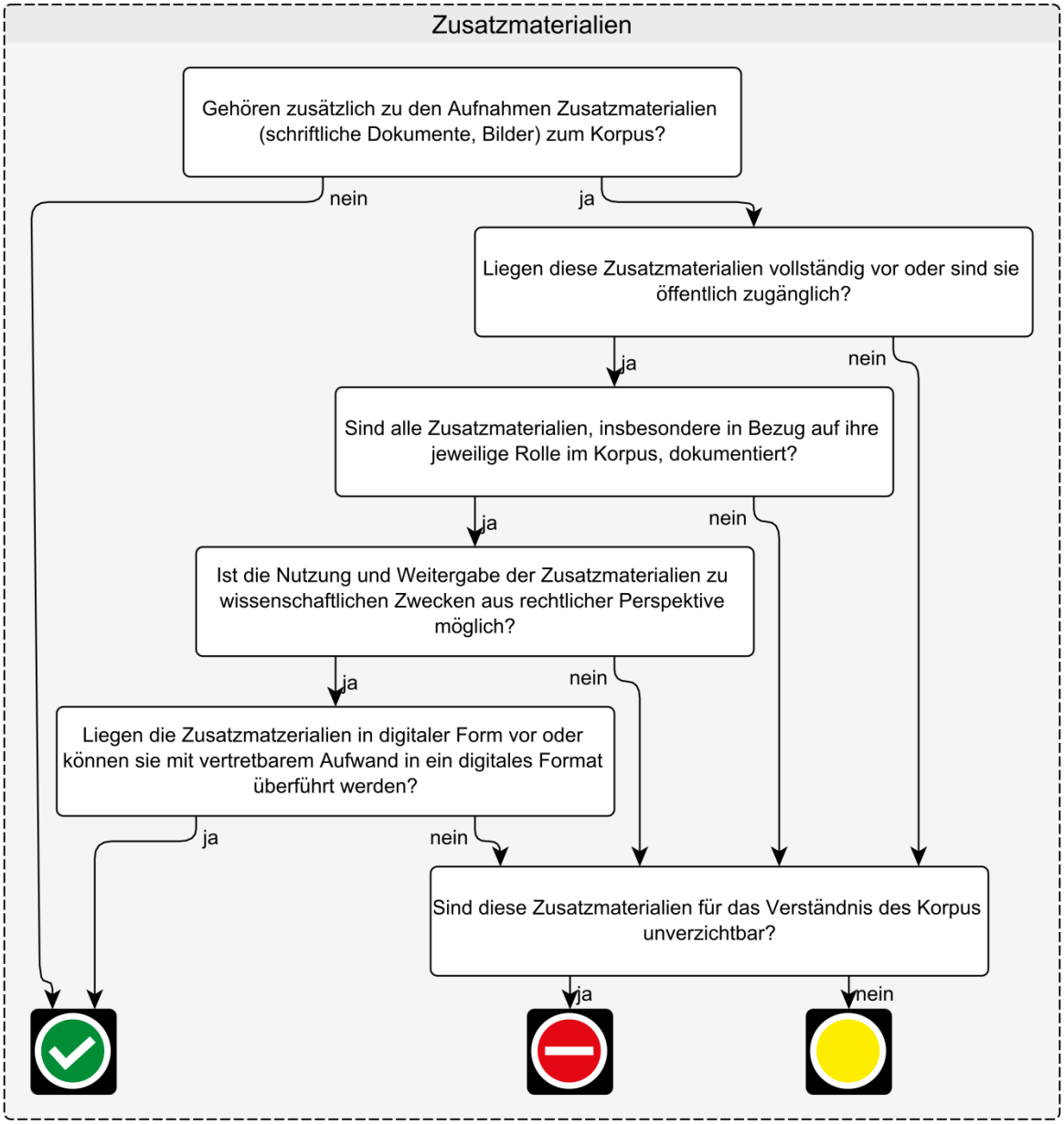
Im Übrigen stellen sich für die Annotationen vergleichbare Fragen wie für die Basistranskription, also insbesondere die Frage der Vollständigkeit, der Dokumentation und der strukturierten Speicherung. Liegen für ein Korpus nur stark unvollständige oder nur unzureichend dokumentierte oder nicht strukturiert gespeicherte Annotationen vor, kann es aus Kostengründen sinnvoll sein, diese bei der Aufbereitung nicht zu berücksichtigen, also aus dem aufbereiteten Korpus zu entfernen. Die Nachnutzbarkeit des Korpus wird dadurch in der Regel nicht so weit beeinträchtigt, dass dies den Wert der Ressource als solche in Frage stellen würde.



6. Zusatzmaterialien

Häufig gehören zum Korpus außer Korpus- und Metadaten auch weitere Dokumente, die quasi als ergänzende Metadaten zum Verständnis der Korpusdaten beitragen. Sie sind daher für die Nachnutzung meist von großer Bedeutung. Diese sogenannten Zusatzmaterialien können der Korpusebene, der Kommunikationsebene oder der Sprecherebene zugeordnet werden. Zusatzmaterialien auf der Korpusebene sind für das gesamte Korpus relevant, wie beispielsweise Dokumentation einer spezifischen Sprechergemeinschaft, aus der die Sprecher im Korpus stammen. Auf der Kommunikationsebene liegen vielleicht Fotos oder Sitzpläne von der Kommunikationssituation, Vorgaben und Lösungen einer aufgenommenen Aufgabe, Stimulibilder für elizitierte Aufnahmen oder Mitschriften aus dem aufgenommenen Unterricht vor. Verschiedene (linguistische) Tests und Fragebögen, die Sprachkenntnisse, Verwendungsmuster, Einstellungen usw. erfassen sollen, kommen besonders bei Korpora im Bereich Spracherwerb oder Mehrsprachigkeit häufig als Zusatzmaterialien auf Sprecherebene vor. Wenn diese Daten nicht als Metadaten erfasst werden können, müssen sie als Zusatzmaterialien ins Korpus integriert werden.

Problematisch sind Fälle, in denen Zusatzmaterialien tatsächlich eine große Rolle für das Verständnis und somit für die Nachnutzung des Korpus spielen, die Zusatzmaterialien aber gar nicht oder nicht in einem für eine Nachnutzung geeigneten Zustand vorliegen. Hier müssen der potentielle Nachnutzungswert des Korpus mit bzw. ohne Zusatzmaterialien sowie der Umfang der notwendigen Aufbereitungsarbeiten bei einer Vereinbarung zwischen Datengeber und Datennehmer berücksichtigt werden. Für eine möglichst hohe Nachnutzbarkeit sollten alle vorhandenen Materialien digitalisiert, mit Metadaten versehen und ins Korpus integriert werden.


Zusatzmaterialien
Gehören zusätzlich zu den Aufnahmen Zusatzmaterialien (schriftliche Dokumente, Bilder) zum Korpus?
nein
ja
Liegen diese Zusatzmaterialien vollständig vor oder sind sie öffentlich zugänglich?
ja
nein
Sind alle Zusatzmaterialien, insbesondere in Bezug auf ihre jeweilige Rolle im Korpus, dokumentiert?
ja
nein
Ist die Nutzung und Weitergabe der Zusatzmaterialien zu wissenschaftlichen Zwecken aus rechtlicher Perspektive möglich?
ja
nein
Liegen die Zusatzmatzerialien in digitaler Form vor oder können sie mit vertretbarem Aufwand in ein digitales Format überführt werden?
ja
nein
Sind diese Zusatzmaterialien für das Verständnis des Korpus unverzichtbar?
ja
nein

Literatur

Beal, Joan C. (2009): *Creating corpora from spoken legacy materials: variation and change meet corpus linguistics*, in: Renouf, Antoinette & Kehoe, Andrew (Hrsg.), Corpus linguistics: Refinements and reassessments (Vol. 69, S. 33-47). Amsterdam: Rodopi.

Bücker, Jörg; Drude, Sebastian; Jung, Dagmar; Kamocki, Pawel: Ketzan, Erik; Purschke, Christoph; Redder, Angelika & Schmidt, Thomas (2013): *Handreichung: Informationen zu rechtlichen Aspekten bei der Erhebung mündlicher Korpora.* Bonn: DFG.

Hedeland, Hanna; Lehmberg, Timm; Schmidt, Thomas & Wörner, Kai (2011): *Multilingual Corpora at the Hamburg Centre for Language Corpora*, in: Hedeland, Hanna; Schmidt, Thomas & Wörner, Kai (Hrsg.): Multilingual Resources and Multilingual Applications. Proceedings of the GSCL conference, Hamburg. Arbeiten zur Mehrsprachigkeit (Working Papers in Multilingualism), Serie B (96). Hamburg, 227-233.
[http://www1.uni-hamburg.de/exmaralda/files/Corpora_HZSK_GSCL2011.pdf]

Schmidt, Thomas & Bennöhr, Jasmine (2008): *Rescuing Legacy Data*, in: Language Documentation and Conservation 2(1), 109–129.
[http://scholarspace.manoa.hawaii.edu/handle/10125/1803]

Schmidt, Thomas; Dickgießer, Sylvia & Gasch, Joachim (2013): *Die Datenbank für Gesprochenes Deutsch - DGD2.* Mannheim: Institut für Deutsche Sprache.
[http://ids-pub.bsz-bw.de/frontdoor/index/index/docId/1274]

Stift, Ulf-Michael & Schmidt, Thomas (2014): *Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch.* erscheint in: Eichinger, Ludwig et al. (Hrsg.): Festschrift zum 50. Geburtstag des Instituts für Deutsche Sprache.

Widdowson, John (2003): *Hidden depths: Exploiting archival resources of spoken English*, in: Lore and Language, 17(1&2): 81-92.