Leitfaden zur
Beurteilung von

Aufbereitungsaufwand und Nachnutzungswert von Korpora gesprochener Sprache

Thomas Schmidt
Hamburger Zentrum für Sprachkorpora /
Archiv für Gesprochenes Deutsch

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]

Geschichte: SFB-Daten (Schmidt/Bennöhr), AGD-Daten

Zweck des Dokuments: Checkliste für die Bearbeitung alter Korpora, auch: Orientierungshilfe für die Planung von neuen Korpora

Begriffe:

- wissenschafts-öffentliche Publikation
- Datenschutz
- Urheberrecht
- Datengeber / Datennehmer

Legende

 Idealfall

 positiv

 negativ

 Schwerwiegendes Problem / Ausschlusskriterium

1. Inventarisierung

				Aufwand	Nutzen
Verfügbarkeit der Daten					
1.1.	Gibt es eine verbindliche, vollständige Liste aller Original-Daten?		**Ja**		
			Nein		
1.1.1.	**Falls ja**	Sind alle in dieser Liste vorhandenen Daten für den Datennehmer verfügbar und werden alle Original-Daten, die der Datennehmer erhält, in der Liste aufgeführt?	**Ja**		
			Nein		
Lesbarkeit der Daten					
1.2.	Ist für das Lesen der Daten eine spezielle Software oder ein spezielles Gerät notwendig?		**Ja**		
			Nein		
1.2.1.	**Falls ja**	Kann ein funktionsfähiges Exemplar der Software und/oder des Geräts vom Datennehmer genutzt werden?	**Ja**		
			Nein		
Abschluss der Arbeit an den Daten					
1.3.	Sind die Arbeiten an den Daten von Seiten des Datengebers definitiv abgeschlossen?		**Ja**		
			Nein		
1.3.1.	**Falls nein**	Sollen zu einem späteren Zeitpunkt weitere oder geänderte Daten in das aufzubereitende Korpus integriert werden?	**Ja**		
			Nein		

Erläuterungen zu 1.1.:

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)

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

Erläuterungen zu 1.2.:

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 ist.
Für Daten, die sich nur mit Hilfe solcher spezieller Software oder Geräte 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 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 2xChinch-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).

Erläuterungen zu 1.3.:

Es ist in der Regel sehr schwierig, 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 erweitert (z.B. durch neue Aufnahmen und Transkriptionen) werden soll, macht dies deutlich mehr Planungsarbeit erforderlich. Idealerweise sollte der Datengeber also 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 das aufbereitete Korpus einfließen zu lassen.

2. Rechtliche Aspekte

				Aufwand	Nutzen
Datenschutz					
2.1.	Sind für die Veröffentlichung der Daten datenschutzrechtliche Aspekte zu beachten?		**Ja**		
			Nein	+	
2.1.1.	**Falls ja**	Wurden schriftliche Vereinbarungen mit den aufgenommenen Personen getroffen?	**Ja**		
			Nein		
2.1.1.1.	**Falls ja**	Liegen die Vereinbarungen für alle aufgenommenen Personen vor?	**Ja**		
			Nein		
2.1.1.2.		Erlauben die Vereinbarungen eine wissenschafts-öffentliche Bereitstellung der Daten?	**Ja**		
			Nein		
2.1.1.3.		Gelten die Vereinbarungen für Aufnahmen, Transkriptionen und Metadaten	**Ja**		
			Nein		
2.1.1.4.		Ist eine Maskierung (Anonymisierung bzw. Pseudonymisierung) der Aufnahmen bzw. Metadaten und/oder Transkriptionen Gegenstand der Vereinbarung?	**Ja**		
			Nein		
2.1.1.5.		Ergeben sich aus der Vereinbarung weitere oder andere Beschränkungen bzgl. der wissenschafts-öffentlichen Publikation der Daten?	**Ja**		
			Nein		
2.1.1.6.		Gibt es die Möglichkeit, die aufgenommen Personen nachträglich zu kontaktieren die Vereinbarung nachträglich zu ändern oder zu erweitern?	**Ja**		
			Nein		
2.1.1.7.	**Falls nein**	Gibt es die Möglichkeit, die aufgenommen Personen nachträglich zu kontaktieren und eine solche Vereinbarung nachträglich zu treffen?	**Ja**		
			Nein		
Urheberrecht					
2.2.	Sind für die Veröffentlichung der Daten urheberschutzrechtliche Aspekte zu beachten?		**Ja**		
			Nein		
2.2.1.	**Falls ja**	Wurden mit dem Inhaber dieser Urheberrechte Vereinbarungen über eine wissenschafts-öffentliche Publikation der Daten getroffen?	**Ja**		
			Nein		
2.2.1.1.	**Falls ja**	Erlauben die Vereinbarungen eine wissenschafts-öffentliche Bereitstellung der Daten?	**Ja**		
			Nein		
2.2.1.2.		Ergeben sich aus der Vereinbarung weitere oder andere Beschränkungen bzgl. der wissenschafts-öffentlichen Publikation der Daten?	**Ja**		
			Nein		
2.2.1.3.	**Falls nein**	Gibt es die Möglichkeit, eine solche Vereinbarung nachträglich zu treffen?	**Ja**		
			Nein		
Anforderungen des Datengebers					
2.3.	Gibt es von Seiten des Datengebers besondere Bedingungen / Anforderungen für eine wissenschafts-öffentliche Publikation der Daten?		**Ja**		
			Nein		
2.3.1.	**Falls ja**	Gibt es eine schriftliche Formulierung dieser Bedingungen / Anforderungen, die dem Datennehmer vorliegt?	**Ja**		
			Nein		

Erläuterungen zu 2.1.:

Ob und welche datenschutzrechtliche Aspekte bei der Veröffentlichung des Korpus zu beachten sind, entscheidet wesentlich über dessen Nachnutzungswert. Der Idealfall, dass es keine solchen Aspekte zu beachten gibt, ist eine Ausnahme, die nach aller Erfahrung ausschließlich dann auftritt, wenn die Aufnahmen des Korpus aus öffentlichen Veranstaltungen stammen (in diesem Fall gibt es dann aber i.d.R. dennoch urheberschutzrechtliche Aspekte zu beachten).

Der Normalfall ist also der, dass bei der Veröffentlichung der Daten datenschutzrechtliche Aspekte zu beachten sind. Grundlage hierfür ist üblicherweise eine schriftliche Vereinbarung mit den aufgenommenen Personen, die regelt, wer die Daten in welcher Form für welche Zwecke nutzen darf. Typischerweise wird vereinbart, dass Daten nur für die nicht-kommerzielle Nutzung in Wissenschaft und Lehre, nur auf Antrag und nur in einer Form, die keine unmittelbaren Rückschlüsse auf die Identität der aufgenommenen Personen zulässt, veröffentlicht werden dürfen. Folgende Problemfälle treten häufiger auf:

- Es liegen generell keine schriftlichen Vereinbarungen 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 schriftlichen Vereinbarungen sind ausdrücklich mit Bezug auf das ursprüngliche Erhebungsprojekt getroffen worden, d.h. der Fall, dass die Daten auch in einem anderen Zusammenhang genutzt werden bzw. zur Nutzung angeboten werden, ist nicht erwähnt oder sogar explizit ausgeschlossen.

Wenn solche Problemfälle auftauchen, ist der angestrebte Nutzungszweck der Daten i.d.R. nicht ausreichend legitimiert. Eine Aufbereitung der Daten sollte erst begonnen werden, falls und wenn sich die Probleme durch nachträgliches Kontaktieren der aufgenommenen Personen ausräumen lassen.

Erläuterungen zu 2.3.:

Idealerweise überträgt der Datengeber dem Datennehmer alle Nutzungsrechte, die sich aus den beiden vorgenannten Punkten ergeben und überlässt diesem dann die Entscheidung, ob die Daten in einem konkreten Fall zur Nutzung durch Dritte freigegeben werden können oder nicht.

Oft oder sogar typischerweise knüpfen Datengeber die Weitergabe der Daten aber an zusätzliche Bedingungen, die weder datenschutz- noch urheberrechtlichen Gründen geschuldet 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 ausdrücklichen 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 fast immer 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 dann oft nur sehr restriktiv, z.B. nur oder bevorzugt an Personen, die dem Datengeber persönlich bekannt sind, erteilt. Ersteres erhöht also den Aufwand zur Verwaltung, letzteres verringert den Nutzen des betreffenden Korpus.

Wenn möglich, sollten solche zusätzlichen Bedingungen daher vermieden werden. Lassen sie sich nicht vermeiden, ist unbedingt darauf zu achten, dass sie in möglichst konkreter und expliziter Form schriftlich festgehalten werden, bevor die Aufbereitung begonnen wird.

3. Metadaten

Korpusdesign						
3.1.	Ist das Korpusdesign schriftlich dokumentiert?			**Ja**		
				Nein		
3.1.1.	**Falls ja**	Ist die schriftliche Dokumentation des Korpusdesigns veröffentlicht?		**Ja**		
				Nein		
3.1.2.	**Falls nein**	Lässt sich eine schriftliche Dokumentation des Korpusdesigns nachträglich erstellen?		**Ja**		
				Nein		
Gesprächsereignisse						
3.2.	Liegen Metadaten zu den einzelnen Gesprächen und deren Aufnahmen in schriftlicher Form vor?			**Ja**		
				Nein		
3.2.1.	**Falls ja**	Sind diese Metadaten aus sich heraus, d.h. ohne weitere Dokumentation verständlich und nutzbar?		**Ja**		
				Nein		
3.2.1.1.	**Falls ja**	Liegen die Metadaten für alle Gespräche und Aufnahmen vor?		**Ja**		
				Nein		
3.2.1.2.		Folgen die Metadaten einem dokumentierten Schema?		**Ja**		
3.2.1.3.		Liegen die Metadaten in elektronischer Form vor?		**Ja**		
				Nein		
3.2.1.3.1		**Falls ja**	Handelt es sich um ein strukturiertes Dateiformat?	**Ja**		
				Nein		
3.2.1.4.	**Falls nein**	Können die zusätzlichen Informationen, die zum Verständnis und zur Nutzung dieser Metadaten notwendig wären, nachträglich bereitgestellt werden?		**Ja**		
Sprecher						
3.3.	Liegen Metadaten zu den einzelnen Sprechern in schriftlicher Form vor?			**Ja**		
				Nein		
3.3.1.	**Falls ja**	Sind diese Metadaten aus sich heraus, d.h. ohne weitere Dokumentation verständlich und nutzbar?		**Ja**		
				Nein		
3.3.1.1.	**Falls ja**	Liegen diese Metadaten für alle Sprecher vor?		**Ja**		
				Nein		
3.3.1.2.		Folgen die Metadaten einem dokumentierten Schema?		**Ja**		
3.3.1.3.		Liegen die Metadaten in elektronischer Form vor?		**Ja**		
				Nein		
3.3.1.3.1		**Falls ja**	Handelt es sich um ein strukturiertes Dateiformat?	**Ja**		
				Nein		
3.3.1.4.	**Falls nein**	Können die zusätzlichen Informationen, die zum Verständnis und zur Nutzung dieser Metadaten notwendig wären, nachträglich bereitgestellt werden?		**Ja**		
				Nein		

Erläuterungen zu 3:

Eine sinnvolle Nachnutzung der Daten ist nicht ohne eine grundständige Dokumentation des Korpusdesigns, der Gesprächsereignisse und der daran beteiligten Sprecher möglich. Idealerweise liegen solche 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 Dokumentation liegt nicht elektronisch und/oder nicht in strukturierter Form (z.B. als Notizen auf Karteikarten oder als unstrukturiertes Word-Dokument) vor. In diesem Falle muss eine Strukturierung nachträglich vom Datennehmer vorgenommen werden. Die ist aufwändig, in aller Regel aber möglich.

Aufnahmen

4. Transkriptionen

Literatur

Beal, Joan C. (2010): “Creating corpora from spoken legacy materials: variation and change meet corpus linguistics”, in: ????

Schmidt, Thomas & Bennöhr, Jasmine (2008): “Rescuing Legacy Data”, in: Language Documentation and Conservation 2(1), 109–129.

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