Leitfaden für die Konvertierung von Legacy Data

Thomas Schmidt, 25.07.2008

Einleitung

Dieses Dokument gibt einige Empfehlungen für das Konvertieren älterer Bestände linguistischer Korpora (sog. „Legacy Data“) in eine moderne, nachhaltig (wieder)verwendbare Form. Wir haben selbst mehrere solcher Datenbestände bearbeitet[1], und sind uns dabei bewusst geworden, an wie vielen Stellen in diesem Prozess sich verschiedene Varianten von Murphy’s Gesetz („what can go wrong will go wrong“) manifestieren, zum Beispiel:

- Niemand weiß etwas über die Originaldaten
- Die Originaldaten sind höchstens halb so gut (konsistent, umfangreich, …), wie ihr Urheber glaubt
- Alles dauert mindestens doppelt so lange wie geplant
- Wenn Hilfskräfte gut eingearbeitet sind, kündigen sie
- Wenn alles fertig ist, wird es versehentlich gelöscht

Wir haben diese Empfehlungen formuliert, um anderen, die ähnliche Aufgaben bearbeiten wollen, den mühsamen Lernprozess zumindest teilweise zu ersparen.

Die einzelnen Abschnitte dieses Leitfadens sind in der Reihenfolge angeordnet, die wir nach unseren bisherigen Erfahrungen für die richtige halten:

- Als allererstes sollten die Modalitäten für eine Publikation der Daten festgelegt werden.
- Anschließend sollte eine Inventur der vorhandenen Daten gemacht werden.
- Darauf aufbauend kann die Konvertierungsarbeit geplant …
- … und schließlich, in mehreren verschiedenen Arbeitspaketen durchgeführt werden.

Publikationsmodalitäten festlegen

Das Ziel der Konvertierung ist es, Daten zu erstellen, die von anderen weitergenutzt werden können. Dazu müssen die Daten nach Abschluss der Konvertierung in irgendeiner Form veröffentlicht werden. Die allermeisten Daten gesprochener Sprache können aber, aus datenschutzrechtlichen und/oder aus urheberschutzrechtlichen Gründen, nicht einfach frei verfügbar gemacht werden, sondern bedürfen irgendeiner Art der Zugriffskontrolle. Wie genau diese Zugriffskontrolle aussieht, hängt von den individuellen Eigenschaften der Daten und von Vereinbarungen, die mit den aufgenommenen Personen getroffen wurden, ab. Beispielsweise kann es bei sensiblen Daten sinnvoll sein, pseudonymisierte Versionen der Transkriptionen über einen Passwortschutz im Netz zugänglich zu machen, die Aufnahmen selbst aber gar nicht zu veröffentlichen.

Weil die Publikationsmodalitäten auch die Aufgaben bei der Konvertierung mitbestimmen, müssen sie bereits zu Beginn der Arbeiten weitestgehend geklärt sein. Insbesondere muss zu diesem Zeitpunkt bereits klar sein,

1) ob Transkriptionen pseudonymisiert werden müssen (oder bereits sind), d.h. ob Sprechernamen (und ggf. auch Ortsnamen) durch Pseudonyme ersetzt werden müssen. Dies hat insbesondere auch Auswirkungen auf das Anlegen einer Ordnerstruktur und auf die Bearbeitung der Metadaten (d.h. auch dort müssen dann ggf. jeweils Pseudonyme statt Echtnamen stehen)

[1] Siehe dazu T.Schmidt / J. Bennoehr (2008): Rescuing Legacy Data. In: Language Documentation and Conservation. http://www.nflrc.hawaii.edu/ldc/

2) ob die Aufnahmen veröffentlicht werden können. Wenn nicht, bedeutet z.B. die Alignierung u.U. einen nicht zu rechtfertigenden Aufwand.
3) ob Aufnahmen analog zur Pseudonymisierung anonymisiert werden müssen, d.h. ob dort Vorkommen von Eigennamen durch einen Piepton oder eine Stille ersetzt werden müssen.

Datenurheber tun sich mit diesem Schritt erfahrungsgemäß schwer und schieben Entscheidungen über die Publikationsmodalitäten gerne so lange wie möglich hinaus. Die Publikationsmodalitäten bestimmen aber ganz wesentlich die unvermeidliche Kosten-Nutzen-Rechnung, die man an vielen Stellen im Konvertierungsprozess anstellen muss. Daten, die nicht veröffentlicht werden können, sind nur für ihre Urheber nützlich – der Aufwand der zur Konvertierung betrieben werden muss, lässt sich damit i.d.R. nicht rechtfertigen.

Inventur

Die Inventur ist der erste Schritt und sollte vollständig abgeschlossen sein, bevor irgendwelche weiteren Schritte in Angriff genommen werden. In der Inventur wird ein Überblick über alle vorhandenen Daten (Aufnahmen, Transkriptionen, Metadaten, jeweils digital oder nicht-digital) erstellt, anhand dessen der Arbeitsablauf geplant und Aufwand abgeschätzt werden kann. Es ist sehr wichtig, bei diesem Schritt sorgfältig vorzugehen, also z.B. sicherzustellen, dass eine Aufnahme, die laut einer Liste vorhanden sein müsste, auch wirklich existiert, und sich im Falle, dass von einer Transkription mehrere Version bestehen, zu entscheiden, welche davon die aktuellste und damit gültige ist. Als Ergebnis der Inventur sollte folgendes stehen:

1) Eine Verzeichnisstruktur auf einem backupfähigen Rechner, in dem alle vorhandenen digitalen Daten in einer transparent organisierten Weise abgelegt sind. Dieser Rechner sollte für alle an der Konvertierung beteiligten Personen zugänglich sein. Wir haben hierfür den Novell-Server des Hamburger Rechenzentrums genutzt. Folgendes ist dabei zu beachten:

   - Wenn die Dateien von einer MAC OS 9.x-Anwendung (also z.B. syncWriter) geschrieben wurden und damit (potentiell) aus „resource fork“ und „data fork“ bestehen, muss der Rechner ein Dateisystem haben, das mit solchen Dateien umgehen kann. Man erkennt dies einfach daran, dass sich Dateien von einem MAC OS 9.x-Rechner aus erfolgreich schreiben und wieder öffnen lassen.
   - Es empfiehlt sich, bereits in diesem Schritt zu bedenken, dass zu einem Diskurs mehrere Dateien existieren können (z.B. zwei Aufnahmen, drei Transkriptionen, ein Protokoll). Wir haben daher grundsätzlich für jeden Diskurs einen eigenen Ordner angelegt. Dies mag anfangs als eine übertrieben kleinteilige Ordnerstruktur erscheinen, ist aber für die Planung und Durchführung aller weiteren Schritte sehr nützlich.
   - Teilweise gibt es in diesem Schritt zu manchen Diskursen noch gar keine digitalen Daten (z.B. wenn die Aufnahme auf Kassette vorliegt und nicht transkribiert wurde). Ein entsprechender Ordner sollte trotzdem angelegt werden.
   - Der Rechner sollte von Anfang an ausreichend Speicherplatz zur Verfügung haben, um alle entstehenden digitalen Daten, insbesondere die digitalisierten Aufnahmen, beherbergen zu können. Eine Audio-Aufnahme von 45 Minuten (also eine Kassetten-Seite) braucht als digitale WAV-Datei etwa 500 MB Speicherplatz, für Video kann in einer vergleichbaren Größenordnung gerechnet werden. Für 100 Stunden Aufnahme ist so bereits etwa 1 Terrabyte Speicherplatz erforderlich.
   - Datei- und Ordnernamen sollten so gewählt werden, dass sie über verschiedene Betriebs- und Dateisysteme hinweg erhalten bleiben. Dies dient nicht zuletzt dazu sicher-

zustellen, dass zuverlässig Backups gemacht werden können. Im einzelnen bedeutet dies:

- Nur die Buchstaben A-Z, a-z, die Ziffern 0-9, den Unterstrich _ und den Bindestrich - verwenden. Der Punkt sollte nur verwendet werden, um das Dateityp-Suffix (z.B. „.doc“) anzuhängen.
- Also: In Datei- und Ordnernamen keine Umlaute, keine anderen „Sonderzeichen“, keine anderen Interpunktionszeichen (Fragezeichen, Ausrufezeichen, etc.) und keine Leerzeichen verwenden![2]
- Datei- und Ordnernamen nicht anhand von Groß- und Kleinschreibung unterscheiden.
- Datei- und Ordnernamen sollten nicht länger als 32 Zeichen sein.

- Datei- und Ordnernamen sollten *systematisch* aufgebaut sein, z.B. nach einem Schema wie „Sprachkürzel + Bindestrich + Gesprächstypskürzel + Bindestrich + Diskursnummer (+ Dateitypsuffix)“ funktionieren. Im DiK-Korpus bekommt beispielsweise das deutsche Anamnesegespräch mit der Diskursnummer 49 den Ordner „D-ANA-49“, die zugehörigen Dateien heißen „D-ANA-49.txt“, „D-ANA-49.wav“ etc.

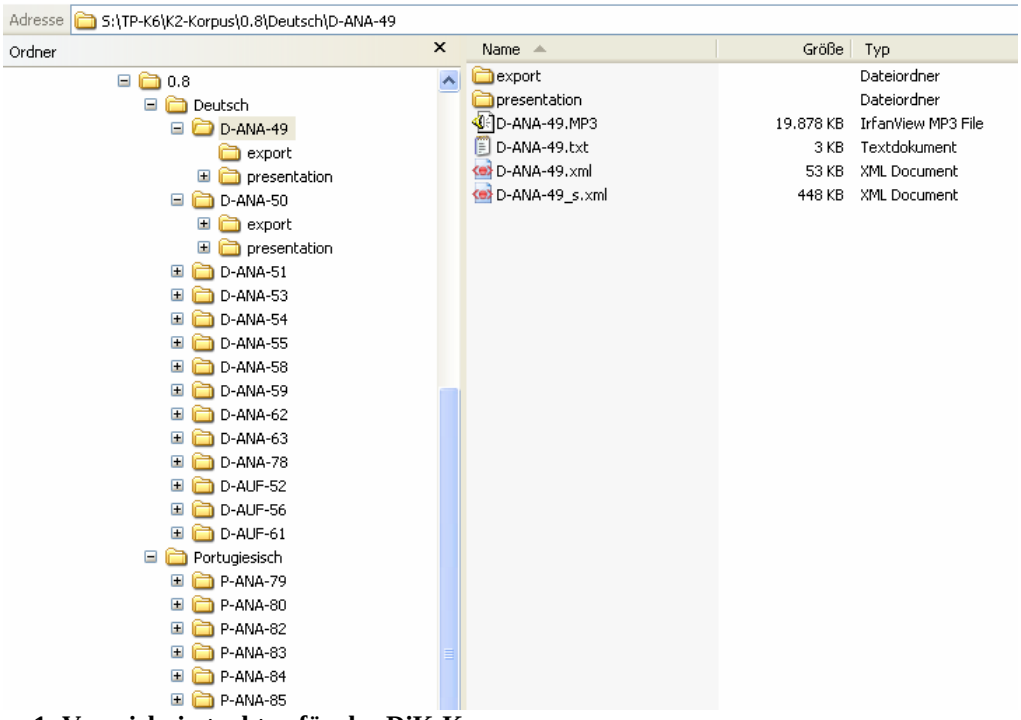


Abbildung 1: Verzeichnisstruktur für das DiK-Korpus

2) Ein Schrank (sic!), in dem alle nicht-digitalen Daten (z.B. Audio- und oder Videokassetten, handschriftlich festgehaltene Metadaten etc.) in einer transparent organisierten Weise aufbewahrt werden. Dieser Schrank sollte für alle an der Konvertierung beteiligten Personen zugänglich sein. Wenn dort Material entnommen wird, sollte das immer entsprechend vermerkt werden.
3) Eine erste Dokumentation des Korpusinhalts, z.B. in Form einer Tabelle, in der diskursweise alle vorhandenen digitalen und nicht-digitalen Daten aufgelistet sind. Es empfiehlt sich, bereits an dieser Stelle Angaben zum Umfang einzelner Datensätze zu machen, denn nur mit Hilfe solcher Angaben lassen sich für die Planung Berechnungen zum Aufwand anstellen.

[2] Die Dateibenennung in vielen syncWriter-Projekten folgte *nicht* diesen Vorgaben – dort wurde z.B. oft ein Häkchen als Bestandteil von Dateinamen benutzt, um anzuzeigen, dass die betreffende Datei fertig bearbeitet war. Dies ist unbedingt zu ändern.

4) Eine vollständige Liste mit der verfügbaren Literatur über das Korpus und dessen Erstellung. Ganz besonders wichtig sind hier projektinterne Dokumente wie Transkriptionsvorgaben, Bearbeitungslisten etc.

Diskurs	Aufnahme(n)	Dauer	Transkription(en)	Größe	Sonstiges
D-ANA-49	D-ANA-49.wav (digital)	00:21:03	D-ANA-49 (syncWriter)	44 kb	Gesprächsprotokoll in Ordner „Protokolle“
PD-ANA-25	Kassette 25	00:40:01	PD-ANA-25_1 (syncWriter) PD-ANA-25_2 (syncWriter) PD-ANA-25_3 (syncWriter)	12 kb 20 kb 13 kb	Anamnesebogen in Ordner “Protokolle”
T-AUF-312	Kassette 312	00:25:02	*Nicht transkribiert*		Gesprächsprotokoll im Ordner “Protokolle”
…	…	…	…	…	…

Abbildung 2: Dokumentation der Inventur eines Korpus

Planung

Anhand des Ergebnisses der Inventur kann die eigentliche Konvertierung geplant werden. Dies beinhaltet:

1) einzelne Konvertierungsschritte auswählen und ihre Reihenfolge festlegen,
2) den (zeitlichen und personellen) Aufwand für einzelne Konvertierungsschritte abschätzen und entsprechende Arbeitspakete planen,
3) Risiken („was kann schief gehen?“) identifizieren

Wir schlagen im folgenden Abschnitt eine Planung vor, in der zuerst die wichtigsten Daten (Aufnahmen und Metadaten) „gerettet“ werden, was mit vergleichsweise wenig und gut abzuschätzendem Aufwand bewerkstelligt werden kann. Erst danach wird nach dieser Planung die Konvertierung von Transkriptionen in Angriff genommen. Diese Reihenfolge hat den Vorteil, dass aufwändige und schwer abzuschätzende Schritte erst dann in Angriff genommen werden, wenn bereits eine solide Basis von wieder verwendbaren Daten vorhanden ist.[3]

Bei der Planung spielen Erwägungen der folgenden Art eine wichtige Rolle:

1) „Breite vor Tiefe“? Sollen einzelne Schritte erst vollständig für alle Datensätze abgeschlossen werden, bevor der nächste Schritt in Angriff genommen wird? Oder geht man eher datensatzweise vorher, führt also alle Schritte (von der Digitalisierung der Aufnahme bis zur Nachbearbeitung der Transkriptionsdaten) für einen Datensatz durch, bevor man sich den nächsten Datensatz vornimmt? Nach unserer Erfahrung ist „Breite zuerst“ grundsätzlich vorzuziehen, weil die Organisation dadurch wesentlich überschaubarer wird.
2) „Wie lange dauert Schritt X“? Die Antwort hängt einerseits von der Aufgabe selbst ab, andererseits aber auch davon, wie schnell, zuverlässig und regelmäßig die damit betrauten Personen arbeiten. Wir können folgende Anhaltspunkte geben:
    a. Für die Digitalisierung von Audio-Aufnahmen (s.u.) kann man ungefähr die Gesamtlänge der Aufnahmen + 20% veranschlagen.
    b. Für die Digitalisierung von Video-Aufnahmen gilt ähnliches. Allerdings ist dabei zu beachten, dass nach der Aufnahme in der Regel noch ein Komprimierungsschritt erforderlich ist, der zusätzliche Zeit in Anspruch nimmt.

[3] Tatsächlich sind wir selbst bisher meist umgekehrt vorgegangen, haben also zunächst Transkriptionen konvertiert und uns erst später um Aufnahmen und Metadaten gekümmert. Wir haben dabei aber festgestellt, dass die Konvertierung und Nachbearbeitung von Transkriptionen eine eigentlich nie endgültig abzuschließende Aufgabe ist. Für die Digitalisierung von Aufnahmen und Metadaten gilt dies nicht. Weiterhin hat sich herausgestellt, dass sich viele Nachbearbeitungsschritte für Transkriptionen wesentlich besser bewerkstelligen lassen, wenn auf die digitalisierte Aufnahmen und Metadaten zugegriffen werden kann. Wir werden in Zukunft daher die Reihenfolge der Bearbeitungsschritte entsprechend umkehren.

c. Das Alignieren von Transkription und Aufnahme hängt stark vom Geschick und der Übung der ausführenden Person ab. Manche geübten Hilfskräfte konnten diesen Schritt ungefähr in „doppelter Echtzeit“ ausführen, haben also zum vollständigen Alignieren einer halbstündigen Aufnahme etwa eine Stunde benötigt. Die SKOBI- und ENDFAS-Daten (ca. 800 Transkriptionen von variabler Länge zwischen 5 und 40 Minuten Dauer) haben zwei sehr fähige Hilfskräfte in insgesamt 450 Stunden fast vollständig mit Zeitmarken im Abstand von etwa 30 Sekunden aligniert bekommen.
d. Der dramatischste Zeitverlust tritt ein, wenn Hilfskräfte nicht regelmäßig an einer Aufgabe arbeiten, sondern es z.B. unangekündigte lange Unterbrechungen des Arbeitsverhältnisses (Praktikum etc.) oder plötzliche zusätzliche Arbeitsaufträge („Transkribieren Sie doch mal schnell noch Aufnahme xy“) gibt.
e. Der Zeitaufwand für die Nachbearbeitung von Transkriptionen lässt sich kaum unabhängig von einem konkreten Datensatz abschätzen. Die Nachbearbeitung von Transkriptionen hat bei uns in allen Fällen mindestens die Hälfte der gesamten Arbeitszeit ausgemacht.

3) „Wer führt Schritt X aus“? Für die Digitalisierung der Aufnahmen und das Eingeben von Metadaten braucht man im Prinzip außer gewissen Computer-Grundfertigkeiten keine speziellen Kenntnisse (Grundkenntnisse in den beteiligten Sprachen sind aber wahrscheinlich auch hier unabdingbar). Das Nachbearbeiten von Transkriptionen erfordert hingegen ein fundiertes linguistisches Verständnis, eine gewisse Vertrautheit mit dem betreffenden Transkriptionssystem, gute Kenntnisse der betreffenden Sprachen und ggf. noch weitere spezielle Kenntnisse. Teilweise erwerben Hilfskräfte solche Kenntnisse erst im Rahmen ihrer Tätigkeit. Wir sind immer dann am erfolgreichsten vorangekommen, wenn wir eine Hilfskraft über längere Zeit mit denselben Daten betrauen konnten. Eine Aufgabe von einer Person an die nächste zu übergeben, nimmt sehr viel Zeit für Einweisung, Einarbeitung etc. in Anspruch.

Aufgrund dieser Erwägungen empfiehlt es sich bei größeren Korpora auf jeden Fall, einzelne Schritte (z.B. „Digitalisierung der Aufnahme“) noch einmal in kleinere Pakete (z.B. „Digitalisierung der deutschen / portugiesischen / türkischen Aufnahmen“) zu zerlegen und für jedes Paket eine eigene Abschätzung zu machen.

Ein erster Plan für ein Korpus mittlerer Größe könnte z.B. so aussehen:

Digitalisierung		
100 Audio-Aufnahmen à ca. 30 Minuten mit Audacity digitalisieren		
Wann? Nov - Dez 2007	**Aufwand?** <100 Stunden	**Wer?** Daisy
Metadaten		
102 handschriftliche Protokolle à 1-2 Seiten in Coma eingeben		
Wann? Dez 2007 - Jan 2008	**Aufwand?** <100 Stunden	**Wer?** Klarabella
Transkriptionen vorbereiten		
31 finnisch einsprachige Transkriptionen in syncWriter vorbereiten		
Wann? Feb 2008 – April 2008	**Aufwand**? ca. 50 Stunden	**Wer**? Daisy
25 ungarisch einsprachige Transkriptionen in syncWriter vorbereiten		
Wann? Feb 2008 – April 2008	**Aufwand**? ca. 50 Stunden	**Wer**? Donald
35 ungarisch/finnisch zweisprachige Transkriptionen in syncWriter vorbereiten		
Wann? Feb 2008 – April 2008	**Aufwand**? ca. 80 Stunden	**Wer**? Klarabella
Konvertieren der Originaltranskriptionen		
91 Transkriptionen von syncWriter nach EXMARaLDA überführen		
Wann? Mai 2008	**Aufwand**? ca. 30 Stunden	**Wer**? Daisy
Definition des Zielformats		
91 konvertierte Transkriptionen analysieren, Zielformat festlegen		
Wann? Juni 2008	**Aufwand**? z.Z. nicht abzuschätzen	**Wer**? Daniel Düstentrieb

Alignierung der konvertierten Transkriptionen 31 finnisch einsprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Juli-September 2008 **Aufwand**? ca. 70 Stunden **Wer**? Daisy 25 ungarisch einsprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Juli-September 2008 **Aufwand**? ca. 70 Stunden **Wer**? Donald & Daisy 35 ungarisch/finnisch zweisprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Juli-September 2008 **Aufwand**? ca. 100 Stunden **Wer**? Klarabella
Nachbearbeiten der konvertierten Transkriptionen 31 finnisch einsprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Ab September 2008 **Aufwand**? z.Z. nicht abzuschätzen **Wer**? Daisy 25 ungarisch einsprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Ab September 2008 **Aufwand**? z.Z. nicht abzuschätzen **Wer**? Donald & Daisy 35 ungarisch/finnisch zweisprachige Transkriptionen in EXMARaLDA alignieren **Wann**? Ab September 2008 **Aufwand**? z.Z. nicht abzuschätzen **Wer**? Klarabella

Abbildung 3: Beispiel für eine Planung

Konvertierung

Digitalisierung der Aufnahmen

Wenn Aufnahmen in nicht-digitaler Form (also i.d.R. als Audio- oder VHS-Kassetten) vorliegen, müssen sie digitalisiert werden. Dies ist ein unabdingbarer Schritt, um die Nachhaltigkeit des Korpus zu gewährleisten, denn Transkriptionen, deren Qualität sich nicht zumindest stichprobenartig anhand der zugehörigen Aufnahme überprüfen lässt, werden für die meisten Wiederverwendungszwecke nicht in Frage kommen.

Für **Audio-Aufnahmen** können wir eine sehr einfache Empfehlung geben: Audio-Kassetten sollten mit Hilfe einer geeigneten Software (wir benutzen Audacity, http://www.audacity.de/) in der bestmöglichen Qualität (44 kHz) als Dateien im WAV-Format digitalisiert werden. Dazu verbindet man ein normales HiFi-Tapedeck über den Audio-Eingang mit dem Rechner, spielt die Aufnahme ab und nimmt sie gleichzeitig mit der Software auf. Für die Qualität des Resultats sind entscheidend:

- die Abspielqualität des Tapedecks
- die Qualität der Soundkarte des Rechners
- die korrekte Aussteuerung der Lautstärke beim Aufnehmen

Für **Video-Aufnahmen** ist eine solche Empfehlung schwieriger, weil sich in diesem Bereich Technologien und Standards noch viel und schnell ändern. Wir haben für das Digitalisieren von Video-Aufnahmen folgendes Setup verwendet: ein VHS-Rekorder wurde über eine externe digitale Videoschnittstelle (Dazzle) mit einem USB-Port des Rechners (Windows PC) verbunden. Mit Hilfe dieser Schnittstelle und der Software „Pinncacle Studio“ wurde eine unkomprimierte digitale Fassung des Videos erstellt. Diese wurde anschließend mit Hilfe der Software „Auto Gordian Knot“ komprimiert. Als Zielformat haben wir dabei AVI mit einem XVID-Codec verwendet.

Metadaten

Metadaten, also Daten über Gespräche und Sprecher, sind wesentlich für die Wiederverwendbarkeit eines Korpus. Wer lange mit einem Korpus gearbeitet hat und an seiner Erstellung selbst mitgewirkt hat, mag diese Daten so verinnerlicht haben, dass er sie als nebensächlich betrachtet – für Dritte, die sich mit einem Korpus auseinandersetzen, das sie nicht selbst erstellt haben, sind solche Metadaten jedoch der erste und damit wichtigste Zugang zu den Daten.

Metadaten sollten daher erstens so gestaltet sein, dass sie ohne weitere Informationen und ohne die Transkriptionen selbst Außenstehenden den wesentlichen Inhalt des Korpus vermitteln können. Sie sollten weiterhin zweitens soweit systematisiert sein, dass sie sich bei der computergestützten Analyse des Korpus konstruktiv verwenden lassen. Folgendes gilt es in diesem Zusammenhang zu beachten:

- Metadaten sollten in verständlich benannten Kategorien beschrieben werden. Kürzel sind zu vermeiden (also nicht „GT: FI, BG“, sondern „Gesprächstyp: Freies Interview, Bildergschichte“).
- Metadaten sollten nicht bzw. nicht nur in Dateinamen kodiert sein (dies ist eine gängige Praxis bei syncWriter-Korpora). M.a.W.: Informationen, die im Dateinamen stecken, müssen noch einmal gesondert festgehalten werden.
- Gleiche Kategorien müssen gleiche Namen bekommen (also nicht einmal „Gesprächstyp“, einmal „Diskursart“) und nach einer konsistenten Logik besetzt werden (also z.B. keine sich überschneidenden Werte wie „Nacherzählung“ und „Nacherzählung Bildergeschichte“ verwenden).
- Metadaten müssen kontextfrei interpretierbar sein. Z.B. ist eine Eigenschaft wie „Alter“ eines Sprechers nicht kontextfrei interpretierbar – sie ist abhängig von einem Datum (z.B. einer Aufnahme). Statt dem „Alter“ sollte daher besser das „Geburtsdatum“ festgehalten werden.

Für die Organisation der Metadaten hat sich das von Coma vorgegebene Schema bewährt: auf der obersten Ebene werden einerseits **Kommunikationen**, andererseits **Sprecher** gelistet. **Transkriptionen** und **Aufnahmen** werden genau einer Kommunikation zugeordnet (wobei es zu einer Kommunikation durchaus mehrere Aufnahmen und/oder Transkriptionen geben kann), zwischen Sprechern und Kommunikationen besteht eine n:m-Zuordnung (d.h. ein Sprecher kann an mehreren Kommunikationen beteiligt sein, jeder Kommunikation sind i.d.R. mehrere Sprecher zugeordnet). Für die Benennung von Sprechern und Kommunikationen sollte ein einheitliches Schema gelten, das sicherstellt, dass jede Kommunikation und jeder Sprecher auf das Gesamtkorpus gesehen eine eindeutige Bezeichnung hat. In unseren Daten ist Verwirrung z.B. oft dadurch entstanden, dass Sprecher mit Rollen wie „Interviewer“ oder Verwandtschaftsbezeichnungen wie „Mutter“, „Tante“ bezeichnet wurden, obwohl in verschiedenen Kommunikationen verschiedene Sprecher die Interviewerrolle innehatten bzw. verschiedene Mütter, Tanten etc. an den Gesprächen beteiligt waren (und darüber hinaus ein- und dieselbe Sprecherin einmal als „Mutter“, einmal als „Tante“ bezeichnet war). Sprecher sollten daher besser durch ein eindeutiges Kürzel (bestehend z.B. aus Initialien des Vor und Nachnamens bzw. deren pseudonymisierten Varianten) identifiziert werden. Ihre Rolle(n) und Verwandtschaftsbeziehung(en) sollten ggf. in untergeordneten Metadaten-Feldern festgehalten werden. Zu jeder Kommunikation, zu jedem Sprecher, zu jeder Aufnahme und zu jeder Transkription kann es (im Prinzip beliebig viele) Metadaten-Sätze geben. Diese bestehen entweder aus einfachen Attribut-Wert-Paaren (z.B. Rolle: Interviewer) oder aus (von Coma) vorgegebenen komplexeren Datentypen wie „Location“ (bestehend aus einem Datum, einem Ort etc.).

Abbildung 4: Beispiel für einen Metadatensatz zu einer Kommunikation (Projekt K5)

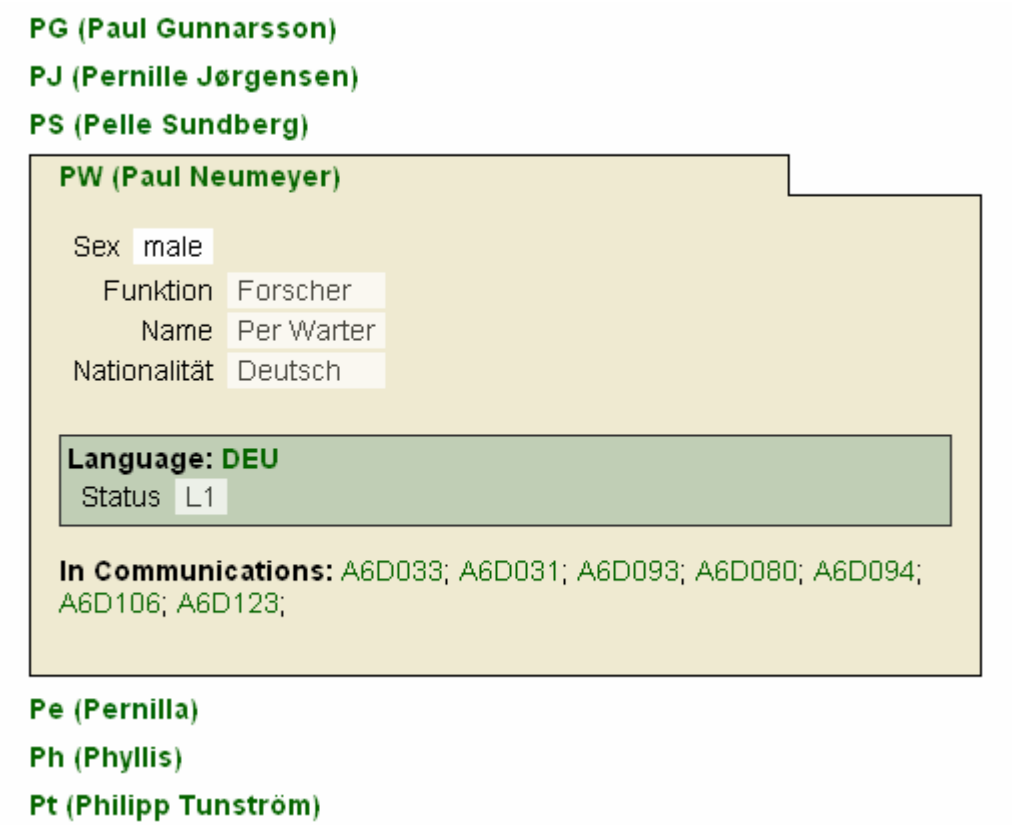


Abbildung 5: Beispiel für einen Metadatensatz zu einem Sprecher (Projekt K5)

Wenn die Metadaten bereits in irgendeiner Form (Excel-Tabelle, Filemaker-Datenbank etc.) vorliegen, können sie unter Umständen automatisch in eine Coma-Datei überführt werden und müssen dort dann nur nachbearbeitet werden. Metadaten in nicht-digitaler Form können, falls das für die betreffenden Personen einfacher ist, ggf. ebenfalls zunächst in Tabellenform (z.B. in Excel) eingegeben und später in eine Coma-Datei konvertiert werden.

Vorbereiten der Originaltranskriptionen

Konvertieren der Originaltranskriptionen

Definition des Zielformats

Nachbearbeiten der konvertierten Transkriptionen

Alignierung
Vereinheitlichung

Anhang A: Importieren von syncWRITER-Daten

Technische Voraussetzungen

- Mac OS X (10.2. – Jaguar) mit Java 1.3.1. (Dies sind die Versionen, mit denen erfolgreich getestet wurde. Unter nachfolgenden Versionen sollte der Export ebenfalls möglich sein.)
- Lauffähige Version des syncWRITERs (Getestet wurde mit Version D1-2.0.2. Ob andere Versionen sich anders verhalten, ist nicht bekannt.)
- EXMARaLDA Partitur-Editor in der jeweils aktuellsten Version
- Apple-Skript ExSync (auf der EXMARaLDA-Website unter „Zubehör“)
- Apple-Skript CountEntries (auf der EXMARaLDA-Website unter „Zubehör“)

Anleitung

Vorbereiten des zu konvertierenden syncWRITER-Dokuments

1. Öffnen Sie das zu konvertierende Dokument im syncWRITER und schließen Sie alle anderen Dokumente im syncWRITER.
2. Löschen Sie ferner alle Bild-, Film- und Skriptspuren, da diese nicht überführt werden können.
3. Falls nicht schon vorhanden: Tragen Sie einen syncTab am Anfang des Dokuments ein und synchronisieren Sie den Beginn jeder Spur mit diesem syncTab.
4. Wenn eine Spur in einer nicht standardmäßig kodierten Schriftart formatiert ist (also z. B. und insbesondere „HIAT Times“), markieren Sie die gesamte Spur (⌘ + A) und weisen Sie ihr über den Menüpunkt „Schrift“ im Menü „Text“ diese Schriftart zu. So stellen Sie sicher, dass die gesamte Spur einheitlich formatiert ist.
5. Starten Sie das Skript „CountEntries“. Sie erhalten einen Dialog, der Ihnen anzeigt, wie viele Einträge die einzelnen Spuren enthalten:

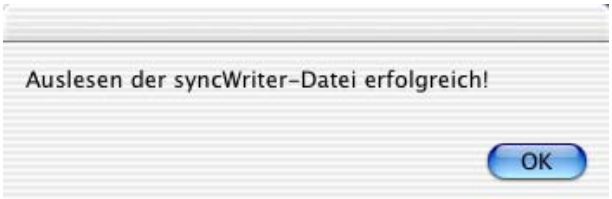
1. Schließen Sie den Dialog, indem Sie auf *OK* klicken.
2. Tragen Sie in der Spur mit den meisten Einträgen (im Beispiel oben also die Spur „A“) so viele syncTabs nach wie möglich. Gehen Sie zu diesem Zweck alle syncTabs durch. Falls die betreffende Spur an einem syncTab keinen Eintrag aufweist,
   - markieren Sie diesen syncTab,
   - setzen Sie den Cursor an das Ende des nächstgelegenen, vorherigen Eintrags in der Spur und

- wählen Sie *Sync* > *Mit Ziel-syncTab verbinden* (oder drücken Sie die Tabulator-Taste).

3. Speichern Sie das syncWRITER-Dokument unter einem neuen Namen und schließen Sie es, ohne den syncWRITER zu beenden.

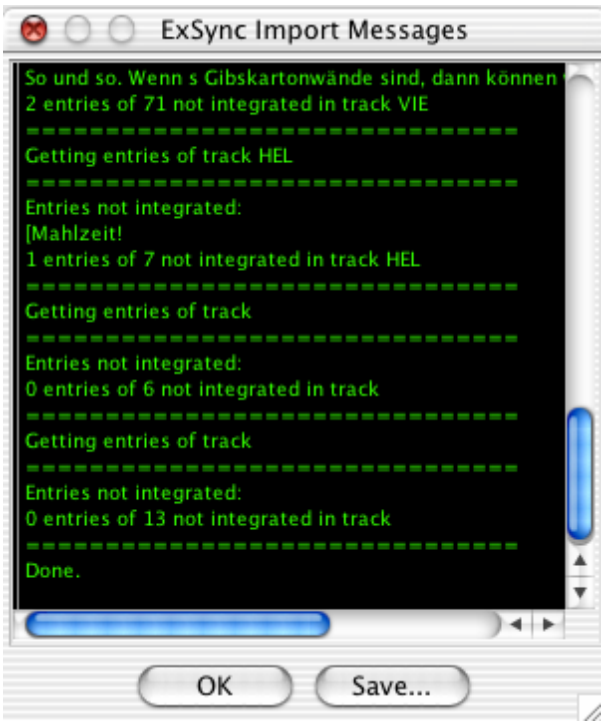
Auslesen des syncWRITER-Dokuments

4. Starten Sie das Skript „ExSync“.
5. Sie werden aufgefordert, das auszulesende syncWRITER-Dokument festzulegen. Suchen Sie das eben bearbeitete Dokument.
6. Sie werden aufgefordert, den Namen der Ausgabedatei festzulegen. Suchen Sie den gewünschten Ordner und geben Sie einen Namen ein. Hängen Sie diesem das Suffix „.xml“ an.
7. Das Dokument wird ausgelesen. Dies kann mehrere Minuten dauern. Während des Auslesens wird ein runder schwarz-weißer Cursor angezeigt – solange dieser zu sehen ist, sollten keine anderen Aktionen am Rechner durchgeführt werden, da dies das Auslesen u. U. zum Absturz bringt. Wenn das Auslesen erfolgreich war, erhalten Sie folgende Nachricht:



Importieren des ausgelesenen Dokuments in EXMARaLDA

8. Starten Sie den EXMARaLDA Partitur-Editor
9. Wählen Sie *File* > *Import* > *“Import” ExSync Data… .*
10. Suchen Sie das in den Schritten 9-12 ausgelesene Dokument und klicken Sie auf *Öffnen*. Das Dokument wird importiert, und Sie erhalten einen Dialog mit Nachrichten über den Verlauf des Imports:



Im Idealfall steht dort für jede Spur („Track“): „0 entries of n not integrated in Track xxx“. Das bedeutet: Alle Einträge aus dem syncWRITER-Dokument konnten in die EXMARaLDA-Basic-Transcription integriert werden.

Sollte dies nicht der Fall sein, werden die Einträge, die nicht integriert werden konnten, aufgeführt. In diesem Fall:

- Wenn nur wenige Einträge nicht integriert werden konnten, klicken Sie auf *Save…* und speichern Sie die Nachrichten in einer Text-Datei. Sie können diese dann mit einem (Unicode-fähigen) Text-Editor öffnen und die nicht integrierten Einträge per „Copy & Paste“ in die importierte Transkription an den betreffenden Stellen einfügen.
- Wenn hingegen viele Einträge nicht integriert werden konnten, ist dies ein Anzeichen dafür, dass bei der Vorbereitung des auszulesenden syncWriter-Dokuments Fehler unterlaufen sind. Wiederholen Sie in diesem Fall den betreffenden Schritt.

11. Beenden Sie den Dialog mit *OK*. Die importierte Transkription wird nun im Partitur-Editor angezeigt und kann dort nachbearbeitet werden.

Nachbearbeiten

Sprechertabelle

Der syncWRITER kennt keine Sprechertabelle. Beim Import wird jedoch versucht, eine solche aus den Spurbenennungen zu konstruieren. Überprüfen Sie das Resultat und korrigieren Sie es, sofern erforderlich, über *File* > *Edit Speakertable.*

Zuordnen von Spuren zu Sprechern, Typen und Kategorien

Spuren in EXMARaLDA müssen einem Sprecher, einem Typ und einer Kategorie zugeordnet werden. Nehmen Sie diese Zuordnung über *Tier* > *Edit tier properties* vor. Beachten Sie dabei folgende Grundsätze:

- Verbale Spuren erhalten den Typ „T(ranscription)“.
- Non-verbale Spuren erhalten den Typ „D(escription)“.
- Übersetzungsspuren o. Ä. erhalten den Typ „A(nnotation)“.
- Die externe Kommentarspur ist eigentlich mit der EXMARaLDA-Logik nicht vereinbar, da sie mehreren Sprechern und mehreren Kategorien gleichzeitig zugeordnet ist. Weisen Sie ihr den Typ „U(ser) D(efined)“ und keinen Sprecher zu.

ExSync Event Shrinker

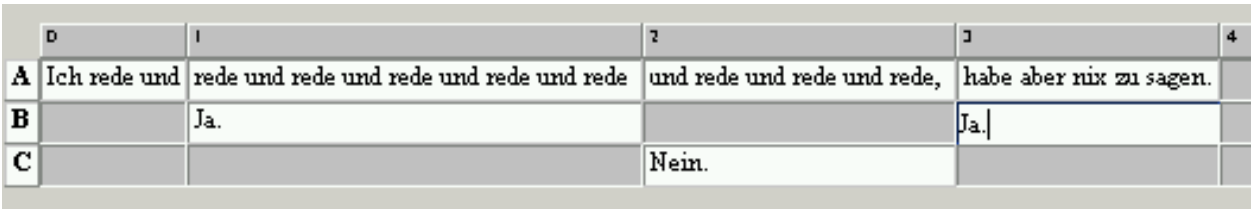
Die syncTabs im syncWRITER markieren jeweils den Beginn einer Synchronpassage. Es ist theoretisch möglich, auch das Ende von Synchronpassagen mit einem syncTab zu markieren. In der Praxis wird dies aber nicht immer gemacht. EXMARaLDA-Transkriptionen brauchen jedoch zwingend einen solchen Endpunkt. Beim Konvertieren wird deshalb bei Abwesenheit eines End-syncTabs einfach der nächste syncTab in der betreffenden Spur verwendet. Folgende Struktur im syncWRITER (syncTabs sind durch rote Striche angedeutet) ...

	0	1	2	3
A	Ich rede und	rede und rede und rede und rede und rede	und rede und rede und rede,	habe aber nix zu sagen.
B		Ja.		Ja.
C			Nein.	

... führt deshalb zu folgender Struktur im EXMARaLDA Partitur-Editor:

	0	1	2	3	4
A	Ich rede und	rede und rede und rede und rede und rede	und rede und rede und rede,	habe aber nix zu sagen.	
B		Ja.		Ja.	
C			Nein.		

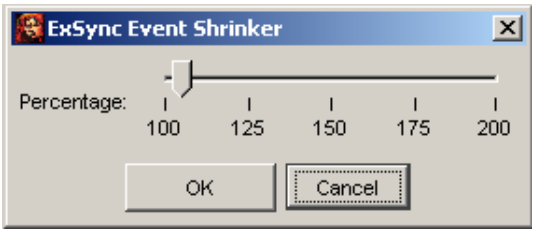
Der erste Eintrag in der Spur von Sprecher B erstreckt sich über mehrere Zeitpunkte bis zum nächsten Eintrag in der gleichen Spur. Der „ExSync Event Shrinker“ kann dies teilweise beheben. Er berechnet auf der Grundlage der typographischen Ausdehnung solcher Spureinträge, ob das betreffende Ereignis mit einem früheren Endpunkt versehen werden kann. Das obige Beispiel würde nach Aufruf des „ExSync Event Shrinkers“ so aussehen:



	0	1	2	3	4
A	Ich rede und	rede und rede und rede und rede und rede	und rede und rede und rede,	habe aber nix zu sagen.	
B		Ja.		Ja.	
C			Nein.		

Dies entspricht immer noch nicht der angestrebten Struktur. Da dieser Schritt aber auf die gesamte Partitur vollautomatisch angewendet werden kann, wird der Nachbearbeitungsaufwand auf diese Weise bereits deutlich reduziert. Gehen Sie wie folgt vor:

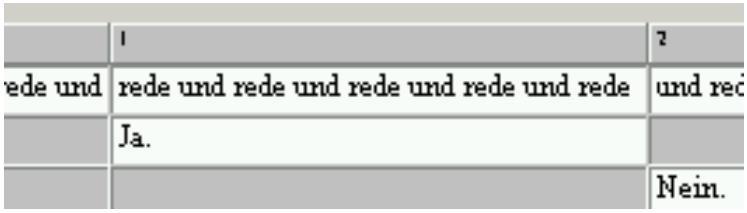
12. Formatieren Sie zunächst die EXMARaLDA-Partitur so, dass die Schriftgrößen in den Spuren annähernd den Schriftgrößen im Original-syncWRITER-Dokument entsprechen.
13. Wählen Sie dann *Edit* > *Extras* > *ExSync Event Shrinker…* . Sie erhalten folgenden Dialog:



14. Der Dialog fragt nach einem Wert der festlegt, ab wann ein Ereignis nicht weiter geschrumpft werden soll. Beispiel: wenn der vorhandene Platz für ein Ereignis kleiner ist als 110 % der typographischen Ausdehnung des betreffenden Eintrages, wird das Ereignis nicht weiter geschrumpft. Normalerweise sollte der voreingestellte Wert von 105 % adäquat sein.
15. Beenden Sie den Dialog mit *OK*. Die Partitur wird zunächst einmal neu formatiert. Anschließend wird der „ExSync Event Shrinker angewandt, und die Partitur wird ein weiteres Mal formatiert. Bei umfangreichen Transkriptionen nimmt dieser Vorgang eine Weile in Anspruch.

Weiteres Nachbearbeiten von Endpunkten

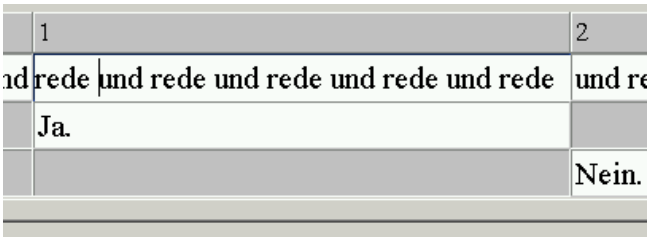
Im obigen Beispiel gibt es immer noch Ereignisse, deren Endpunkte nicht mit den tatsächlichen zeitlichen Verhältnissen korrespondieren. Beispielsweise endet die Äußerung „Ja“ des Sprechers B mit Sicherheit lange Zeit bevor Sprecher C mit der Äußerung „Nein“ einsetzt:



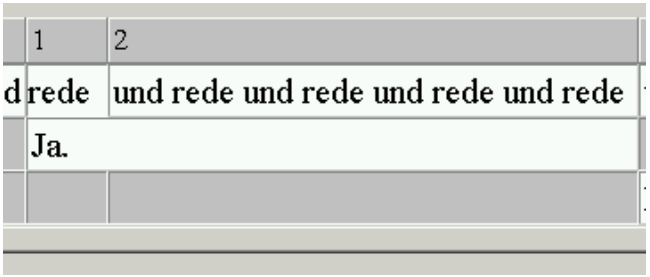
	1	2
ede und	rede und rede und rede und rede und rede	und rec
	Ja.	
		Nein.

Um dies zu korrigieren müssten Sie im Idealfall die Originalaufnahme abhören, um den Endpunkt der Äußerung von Sprecher B relativ zum zeitlichen Verlauf der Äußerung von Sprecher A zu bestimmen. Oft wird dies nicht möglich sein. In diesem Falle können Sie wie folgt eine ungefähre Korrektur vornehmen:

1. Platzieren Sie den Cursor in die Äußerung von Sprecher A, und zwar ungefähr dort, wo – der typographischen Ausdehnung nach – das Ereignis von Sprecher B endet (im Beispiel also z. B. vor das erste „und“):



2. Wählen Sie *Event* > *Split Event* (oder den entsprechenden Button in der Toolbar). Das Ereignis wird an der Cursor-Position geteilt:



3. Setzen Sie den Cursor in das Ereignis von Sprecher B und wählen Sie *Event* > *Shrink event on the right.* Die rechte Ereignisgrenze wird an den neu eingefügten Zeitpunkt verschoben: