

Phonetische Annotation spontansprachlichen Materials

Sonja LATTNER, Sonja LÖHR, Cristina SARTI, Kerstin SEHNERT,
Brigitte ZIPFEL, Fred ENGLERT

Einleitung

Die Anwendung digitaler Aufnahme- und Speichertechniken ermöglicht seit ca. 15 Jahren die relativ kostengünstige Herstellung und Vervielfältigung von Sprachdatensammlungen. Insbesondere im Bereich der Sprachtechnologie (Sprach- und Sprechererkennung, Sprachsynthese) werden umfangreiche Korpora benötigt, die einen möglichst repräsentativen Ausschnitt der "sprachlichen Wirklichkeit" bieten. Eine der ersten großen Sprachdatensammlungen, die für Arbeiten im Bereich der Spracherkennung veröffentlicht wurde, ist das TIMIT-Korpus (GAROFALO et al. 1986), das Aufzeichnungen gelesener Sprache enthält. Bei aktuellen Sprachdatensammlungen, z.B. für das VERBMOBIL-Projekt (www.dfki.uni-sb.de/verbmobil/), wird der Versuch unternommen, Sprachaufnahmen zu gewinnen, die eher "Spontansprache" als gelesene "Laborsprache" repräsentieren sollen. Auch in phonetischen Untersuchungen werden zunehmend Sprachdatensammlungen verwendet, z.B. um Aussprachevarianten (KOHLER 1994) aufzufinden oder Intonationsmuster (ENGLERT 1999) festzustellen.

Ein für derartige Zwecke geeignetes Korpus enthält in der Regel nicht nur die katalogisierten Sprachaufnahmen, sondern vor allem auch Annotationen, die die aufgezeichneten Äußerungen unter verschiedensten Aspekten beschreiben. Die Art dieser Beschreibungen kann von einer Darstellung der Dialogstruktur über eine orthographische Verschriftung bis hin zur detaillierten phonetischen Transkription und Einzellautanalyse reichen. Ein wesentliches Merkmal der Annotationen ist ihre zeitliche Zuordnung zu den entsprechenden Abschnitten im Sprachsignal.

Da die Mehrzahl dieser Korpora an verschiedenen Orten für spezielle Untersuchungsziele hergestellt wurde, unterscheiden sich nicht nur die Beschreibungskategorien (z.B. orthographische vs. phonetische Trans-

kription) sondern auch die Annotationsformate voneinander. Um ein Korpus für eine Untersuchung zu verwenden, verschiedene Korpora zu kombinieren oder ein bestehendes Korpus zu erweitern, sind häufig aufwendige Formatierungsarbeiten durchzuführen. Das Fehlen eines einheitlichen und offenen Annotationsformates wurde von den Distributoren von Sprachdatensammlungen (morph.ldc.upenn.edu, www.icp.inpg.fr/ELRA/, www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html) mittlerweile als ein dringend zu lösendes Problem erkannt. Ein Überblick zu bestehenden Formaten und ein Ansatz für ein einheitliches Rahmenkonzept wurde kürzlich von Bird und Liberman publiziert (BIRD & LIBERMAN 1999). Neben eher traditionellen Formaten (SCHIEL et al. 1998) finden sich hier auch Hinweise auf SGML-basierte Annotationsformate. Eine weitere Initiative zur Gestaltung eines standardisierten Annotationsformats und -werkzeugs ist das EC-Telematics Projekt MATE (mate.nis.sdu.dk).

Um praktische Erfahrungen mit der Annotation eines Korpus für Spontansprache zu gewinnen, wurde im Rahmen der Veranstaltung "Experimentalphonetik II" am Frankfurter Institut für Phonetik eine phonetische Beschreibung spontansprachlichen Materials anhand der Aufzeichnung einer Diskussionssendung aus dem Rundfunk durchgeführt. Ein Ziel des Projektes war es, ein Annotationsformat zu finden, das die Beschreibung von Spontansprache auf mehreren Ebenen – von der orthographischen Verschriftung bis hin zu phonetischen Merkmalen – ermöglicht. Auf jeder dieser Ebenen sollten Zeitpunkte im Sprachsignal mit Beschreibungen verknüpfbar sein, so daß sich eine Struktur parallel verlaufender Spuren ergibt, deren jeweilige Elemente sich zeitlich beinhalten oder überschneiden können. Außerdem sollte im Rahmen der Veranstaltung anhand konkreten Sprachmaterials eine Grundlage für weitere Arbeiten auf diesem Gebiet geschaffen werden.

Vorbereitende Arbeiten

Am Beginn der Arbeit stand die Sammlung von Informationen über bestehende Korpora phonetischer Sprachdaten, bisher verwendete Annotationsformate und existierende Software zur Erleichterung der Annotation. Außerdem mußte eine Entscheidung über das zu verwendende Sprachma-

terial getroffen werden. Da es Spontansprache in Form einer Konversation enthalten sollte, erschien eine Radioaufzeichnung als geeignet. Beim ausgewählten Material handelt es sich um ein ca. 24-minütiges Interview¹ zweier Journalisten (männlich und weiblich) mit dem derzeitigen Staatsminister für Kultur, Michael Naumann.

Der einfachste Weg der Annotation linguistischer Daten ist die Verwendung bereits verfügbarer Programme, mit deren Hilfe auf einer graphischen Oberfläche Annotationen direkt an die entsprechenden Abschnitte im Audiosignal geknüpft werden können. Für phonetische Zwecke vielversprechend erschienen z.B. *Emu* (www.shlrc.mq.edu.au/emu/), *Praat* (www.fon.hum.uva.nl/praat) oder das *Speech Filing System* (www.phon.ucl.ac.uk/resource/sfs.html). Allerdings wäre es mit keinem dieser Werkzeuge möglich gewesen, die Annotation nach eigenen Vorstellungen zu strukturieren, denn die Einbindung programmfremder Formate in Form einer SGML-*document-type-definition* (DTD) erlaubte zu diesem Zeitpunkt keines dieser Programme². Gegen die Verwendung dieser Annotationsprogramme sprachen außerdem auftretende technische Probleme, die zum Teil mit dem jeweiligen Programmaufbau bzw. dem frühen Entwicklungsstadium der Programme zusammenhingen.

Unabhängig von der Frage des Werkzeugs wurde nach einem Annotationsformat gesucht, das ein breites Spektrum an Möglichkeiten für die Annotation phonetischer Daten bietet und gleichzeitig durch eigene Strukturelemente erweiterbar ist. Um Kompatibilität mit einem allgemein akzeptierten Standard zu gewährleisten, wurden hauptsächlich SGML-konforme Annotationsformate in Betracht gezogen. Geeignet für das Projekt erschienen zunächst vor allem die Empfehlungen der *Text Encoding Initiative* (TEI 1994) und das *Universal Transcription Format* (UTF) des *National Institute of Standards and Technology* (NIST 1998). Beide Formate unterstützen die Beschreibung der Dialogstruktur auf der Ebene

¹ "Tacheles – Das Streitgespräch", Deutschlandfunk Berlin, gesendet am 14.5.1999

² Eine Ausnahme bildet das Annotationsprogramm "Transcriber" (www.upenn.edu/~mirror/Transcriber), das seit der Version 1.3 mit einer externen XML-DTD arbeitet, die aber noch nicht durch selbst definierte DTDs ersetzt werden kann.

einer orthographischen Transkription und erlauben durch das Setzen von Zeitmarken eine zeitliche Zuordnung von symbolischer Beschreibung und akustischem Signal. Eine phonemische oder phonetische Transkription allerdings spezifiziert keines dieser Formate. Die relativ einfache Handhabung sowohl bei der Transkription als auch bei der Modifikation des Formates führte zur Entscheidung für das UTF.

Aufgrund der fehlenden Möglichkeit des Imports externer Formate in verfügbare Annotationsprogramme erfolgte die erste Umsetzung der Annotation manuell. Mit Hilfe eines Audio-Editors wurden die zu annotierenden Einheiten im Zeitverlauf fixiert, während für die orthographische Verschriftung und Annotierung nach UTF ein Texteditor ausreichte. Bei der Transkription verzichteten wir auf die genaue Wiedergabe von Aussprachevarianten, d.h. es wurden nur vollständige Lexeme transkribiert (etwa *ich habe* anstelle von *ich hab'*). Auf diese Weise können solche Lexeme einheitlich erkannt werden. Aus demselben Grund wurden Nicht-Lexeme wie Häsitationslaute lediglich drei Kategorien (*äh*, *mh*, *f*) zugeordnet, anstatt für jede der Realisierungen eine orthographische Entsprechung zu suchen.

Die Annotation im UTF-Format

Die Entwicklung des UTF zielte auf die Verschriftung sowohl von Nachrichtensendungen als auch von spontansprachlichen Konversationen ab. Da für den Zweck des Projektes ausschließlich die Transkription von Konversationen von Bedeutung war, beschränkt sich die folgende kurze Gliederung des UTF auf diesen Bereich.

Dem SGML-Standard entsprechend identifiziert UTF zu beschreibende Einheiten im Sprachsignal durch die Einteilung der orthographischen Verschriftung in hierarchisch gegliederte Elemente, die jeweils durch Anfangs- und Endmarkierungen (*Tags*) begrenzt sind. Den Wurzelknoten der UTF-Struktur bildet stets das Element `<utf>`, das eine Konversation umspannende Element ist `<conversation_trans>`. Die in der Hierarchie folgende Komponente innerhalb der Konversation ist der Sprecher-*Turn*, die von einer Person geäußerte Sprachsequenz bis zum nächsten Sprecher-Wechsel.

Die so markierten Elemente können näher beschrieben werden, indem ihren Attributen bestimmte Werte zugewiesen werden, wie z.B. den Wert "Michael Naumann" für das "Sprecher"-Attribut eines <turn>-Elementes. Besonders wichtige Attribute sind – insbesondere bei Elementen wie <turn> oder <overlap> – die Start- und Endzeitpunkte des annotierten Abschnitts. Außer einigen Elementen, die z.B. Hintergrundgeräusche markieren, betreffen alle weiteren annotierbaren Elemente Ereignisse innerhalb eines solchen <turn>, ohne weitere hierarchische Untergliederung. Diese Elemente lassen sich grob in drei Kategorien einteilen:

Sogenannte *pseudo-bracketing-elements* kennzeichnen eine bestimmte Eigenschaft einer Sprachsequenz. Einige dieser *Tags* markieren z.B. feststehende Namen von Personen, Organisationen oder Orten, Zeitangaben etc., andere kennzeichnen eine Sequenz als unklar gesprochen, in einer fremden Sprache gesprochen oder, was für die Transkription einer Konversation besonders wichtig ist, als mit der Äußerung eines anderen Sprechers überlappend gesprochen. Die Bezeichnung "pseudo-bracketing" rührt daher, daß UTF für die Kennzeichnung dieser Sequenzen nicht die übliche SGML-Syntax für klammernde Elemente verwendet, sondern z.B. mit <b_overlap> und <e_overlap> (für Beginn und Ende) zwei allein-stehende, leere Elemente zu ihrer Begrenzung benutzt. Eine weitere Gruppe ebenfalls leerer Elemente bezieht sich auf einzelne Wörter im Konversationstext. *Lexical tags* markieren das unmittelbar folgende Wort z.B. als fehlerhaft oder fragmentarisch ausgesprochen, als Eigennamen, idiosynkratischen Ausdruck oder als Nonlexem (z.B. Häsitationslaute). Eine Unterklasse hiervon bilden leere Elemente, die nicht auf transkribierten Text referieren, sondern auf nichtsprachliche Ereignisse im Signal wie beispielsweise Atmen oder Husten verweisen.

Um UTF-konform zu bleiben, aber dennoch eine weitere Hierarchieebene unterhalb des <turn> zu differenzieren, wurde die Möglichkeit, willkürlich Zeitpunkte im Sprachsignal durch sog. <time>-*Tags* im Dokument festzuhalten, in ein festgelegtes Schema verwandelt: die Folge <time sec="xx:xx.xxx"> {*breath* gibt jeweils den Zeitpunkt von neuen Atemabschnitten (*breath groups*) an.

Die Erweiterung des UTF-Formats

Um neben der orthographischen Transkription eine phonetische notieren zu können, war es notwendig, den UTF-Standard um ein Element zu erweitern, das einzelne Wörter umspannt. Das neue Element `<word>` wurde in der UTF-DTD mit zwei Attributen - "normtrans" für eine normative Transkription nach Ausspracheduden und "emprans" für eine empirische Transkription - deklariert. Erstere wurde automatisch erzeugt (ENGLERT 1993) und manuell korrigiert, während die weitaus aufwendigere empirische Transkription noch aussteht. Weitere Attribute sollen auch hier Start- und Endzeitpunkt des einzelnen Wortes sein.

Präsentation der Daten: SGML, HTML, XML

Eines der Ziele des Projektes war es, den annotierten Text zusammen mit dem Audiosignal zur interaktiven Nutzung und Weiterentwicklung verfügbar zu machen. Das UTF-Dokument sollte also nicht nur intern der Datenabfrage und Analyse dienen, sondern auch in ein übersichtliches, visuelles Format gebracht werden, das es dem Nutzer erlaubt, auf verschiedene Darstellungsarten zuzugreifen und diverse Eigenschaften von Signalsequenzen abzurufen. Langfristig ist geplant, durch eine komfortable Menüführung neben orthographischem Text und Transkription auch Spektrogramm, Grundfrequenzverlauf oder weitere phonetisch oder anderweitig linguistisch relevante Darstellungen interaktiv zu präsentieren.

Diese Ziele können auf verschiedenen Wegen realisiert werden. Einer ist die externe Umwandlung der UTF-Annotationen in HTML. Auf diese Weise konnte eine Darstellung erreicht werden, die ein interaktives Arbeiten³ mit dem Korpus ermöglicht. Dem Konzept von SGML als Markup-Language wird dieses Vorgehen allerdings nicht gerecht, da es die explizit markierten Elemente nicht direkt zur Visualisierung der Daten nutzt. Da SGML mit all seinen Möglichkeiten aber von keinem Browser direkt unterstützt wird, war die externe Transformation der einzige Weg, das wenig veränderte UTF-Format beizubehalten.

³ <http://www.informatik.uni-frankfurt.de/~ifb/exphon/exphon.html> .

Große Fortschritte bei der Entwicklung der Browser sind dagegen aktuell dort zu registrieren, wo es um die Verarbeitung von XML, der einfacheren, restringierten Unterklasse von SGML, geht. Zusätzlich zur externen Konvertierung wurde daher eine XML-konforme Variante des UTF-Dokumentes erstellt. Eine Reihe leerer Elemente des UTF-Formates, die dennoch auf Inhalte des orthographischen Texts referieren, mußten dazu der XML Syntax angepaßt werden, was neben der Korrektur des Markups die Erstellung einer XML-DTD erforderte. Dabei wurde auch die Kennzeichnung der Atemeinheiten an die bestehende Hierarchie angepaßt, so daß ein weiteres umspannendes Element `<breathunit>` entstand, das mit den Attributen Start- und Endzeit analog dem Element `<turn>` die Konversation in kleinere Abschnitte untergliedert. Die ehemals unabhängig voneinander einzelne Wörter bezeichnenden lexikalischen *Tags* wie `<fragment>`, `<mispronounced>` oder `<nonlexeme>` wurden nicht als eigene, umspannende Elemente, sondern als Ausprägungen des `<word>`-Attributs "type" definiert.

Dieser UTF-basierten XML Notation konnte nun ein interaktives Darstellungsformat zugewiesen werden. Dabei bot sich zunächst, einzig aufgrund der einfachen Aneignung und schnellen Generierung, das System der *Cascading Style Sheets* (CSS2) an, das es ermöglichte, sowohl die einzelnen Elemente des Dokuments visuell voneinander abzuheben als auch die in den Attributwerten enthaltenen Informationen, wie z.B. Start- und Endzeitpunkte, darzustellen. Für weitere Arbeiten ist vorgesehen, statt dessen auf eine andere Formatierungssprache, etwa die *Extensible Style Language* (XSL), zurückzugreifen, mit deren Hilfe sich die Ergebnisse von Datenbankabfragen, wie in Abbildung 1 dargestellt, visualisieren lassen.

Startzeit	orthogr.	normativ	empirisch
8:59.740	und	?Unt	?Unt
8:59.846	sie	zi:	si:
8:59.936	meinen	m'aI-n@n	m'aI-n
9:00.313	daß	das	d@s
9:00.464	sie	zi:	si:
9:00.604	das	das	d@s
9:00.875	auch	?aUx	?aUx
9:01.157	in	?In	?In
9:01.255	dieser	d'i:-z6	d'i-z6
9:01.468	Funktion	fUNk-ts'io:n	fUNk-ts'ion
9:01.901	die	di:	di:
9:01.975	sie	zi:	zi:
9:02.114	jetzt	jEtst	jEts
9:02.324	haben	h'a:-bn	h'am
9:02.602	tun	tu:n	tu:n
9:02.918	können	k'9-n@n	k'9-n

Abbildung 1: Unter Verwendung von XSL formatiertes Resultat einer Datenbankabfrage

Ausblick

Die beschriebene Datenbank wird als eine geeignete Ausgangsbasis für weiterführende Analysen und Beschreibungen anhand spontansprachlichen Materials angesehen. Ein Ziel zukünftiger Arbeiten an diesem Material ist die Verknüpfung von symbolischer Annotation mit den aus dem Sprachsignal ableitbaren Meßdaten. Zu nennen sind hier etwa die Verbindung zwischen Transkriptionssymbolen und lautlichen Segmentgrenzen oder die Kopplung von prosodischer Notation mit dem Verlauf der Grundfrequenz.

In der Fortführung des Projektes soll zunächst eine detailliertere symbolische Beschreibung des vorhandenen Sprachmaterials erreicht werden. Hierfür wird unter anderem eine verbesserte Darstellung der Dialogstruktur angestrebt, die es beispielsweise erlaubt, den Modus des Sprecherwechsels sowie den Themenwechsel in einer Konversation zu beschreiben. Als eine der schwierigsten Aufgaben kann die empirische phonetische Transkription des Sprachmaterials angesehen werden, die zusammen mit einer lautlichen Segmentierung vorgenommen werden soll. Um eine möglichst hohe Konsistenz der Transkription zu erreichen, soll diese über ein sogenanntes *Forced Alignment* (BALSS et al. 1997) gewonnen werden.

Das gesamte Material dieses Projekts, inklusive einer interaktiven Darstellung der bearbeiteten Konversation ist unter <http://www.informatik.uni-frankfurt.de/~ifb/exphon/ss99/overview.html> zu finden.

Literatur

- BALSS et al. (1997): U. B., F. ENGLERT, H. REININGER und M. SCHLOTHAUER, Automatische Extraktion von Intonationsparametern aus Sprachsignalen. In: *Fortschritte der Akustik – DAGA '97*, 549-550.
- BIRD, S., LIBERMAN, M. (1999): A Formal Framework for Linguistic Annotation. *Linguistic Data Consortium, University of Pennsylvania, Technical Report MS-CIS-99-01*, Department of Computer and Information Science, March 1999.
- ENGLERT, F. (1993): Normative Transkription mit Künstlichen Neuronalen Netzen. In: *Beiträge zur Symbol- und Signalphonetik* (Phonetica Francofortensia, 6), 1-13. Frankfurt a/M.
- (1999): Feature patterns of sentence accent in German interrogative sentences. In: *14th International Congress of Phonetic Sciences*, San Francisco, Aug. 1999, 1553-1556.
- GAROFOLO et al. (1986): John S. G., Lori F. LAMEL, William M. FISHER, Jonathon G. FISCUS, David S. PALLETT and Nancy L. DAHLGREN, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST. www.ldc.upenn.edu/ldc/docs/TIMIT.html.
- KOHLER, K.J. (1994): Lexica of the Kiel PHONDAT Korpus, Read Speech, Vol. I (Arbeitsberichte des Instituts für Phonetik und digitale Sprachsignalverarbeitung, Universität Kiel, 27).

- NIST (1998): National Institute of Standards and Technology, The Universal Transcription Format (UTF) Annotation Specification. www.nist.gov/speech/hub4_98/utf-1.0-v2.ps .
- SCHIEL et al. (1998): Florian SCH., Susanne BURGER, Anja GEUMANN and Karl WEILHAMMER, The Partitur format at BAS. In: *Proceedings of the First International Conference on Language Resources and Evaluation*, 1295-1301. www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html .
- TEI (1994): Text Encoding Initiative, Guidelines for Electronic Text Encoding and Interchange (TEI P3). Oxford University Computing Services. www.uic.edu/orgs/tei/ .

Web-Seiten

- VERBMOBIL: <http://www.dfki.uni-sb.de/verbmobil/>
- Linguistic Data Consortium: <http://morph.ldc.upenn.edu>
- European Language Resources Association: <http://www.icp.inpg.fr/ELRA/>
- Bayerisches Archiv für Sprachsignale: <http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html>
- MATE EC-Telematics Projekt: <http://mate.nis.sdu.dk>
- Emu: <http://www.shlrc.mq.edu.au/emu/>
- Praat: <http://www.fon.hum.uva.nl/praat>
- SFS: <http://www.phon.ucl.ac.uk/resource/sfs.html>
- Transcriber: <http://www.upenn.edu/mirror/Transcriber>