Nachhaltigkeit linguistischer Daten

Kooperationsprojekt der SFBs 441, 538 und 632
angegliedert an den SFB 441

Leiter:
Prof. Dr. Marga Reis
Prof. Dr. Erhard Hinrichs

Tübingen, Hamburg und Potsdam
Juni 2005

3.1. Allgemeine Angaben zum Teilprojekt C2

3.1.1. Titel

Nachhaltigkeit linguistischer Daten

3.1.2. Fachgebiet und Arbeitsrichtung

Annotationsstandards, Korpusabfragesysteme, Metadaten und Textklassifikation, Terminologien, nachhaltige Datenarchivierung

3.1.3. Leiter/in

Reis, Marga, Prof. Dr.	Hinrichs, Erhard, Prof. Dr.
geb. am 18.06.1941	geb. am 17.08.1954
Deutsches Seminar	Seminar für Sprachwissenschaft
Universität Tübingen	Universität Tübingen
Wilhelmstr. 50	Wilhelmstr. 19
72074 Tübingen	72074 Tübingen
Tel.: 07071-2976741	Tel.: 07071-2975446
Fax: 07071-295321	Fax: 07071-551335
E-mail: mer@uni-tuebingen.de	E-mail: eh@sfs.uni-tuebingen.de

Sind die Stellen der Leiterin/des Leiters des Projektes befristet?

[x] nein [ ] ja, befristet bis zum _____

Vorgesehener Beginn:
01.10.2005

Vorgesehene Laufzeit:
zunächst 3 1/4 Jahre (bis 31.12.2008)

3.1.4. In dem Teilprojekt sind vorgesehen:

• Untersuchungen am Menschen oder am menschlichen Material	[ ] ja	[x] nein
• klinische Studien im Bereich der somatischen Gentherapie	[ ] ja	[x] nein
• Tierversuche	[ ] ja	[x] nein
• gentechnische Untersuchungen	[ ] ja	[x] nein
• Untersuchungen an humanen embryonalen Stammzellen	[ ] ja	[x] nein

3.1.5. Beantragte Förderung des Teilprojektes im Rahmen des Sonderforschungsbereichs (Ergänzungsausstattung)

Haushaltsjahr	Personalmittel	Sachmittel	Investitionsmittel	Gesamt
2005/4	**37,2**	**18,05**	**0**	**55,25**
2006	**148,8**	**10,2**	**0**	**159,0**
2007	**176,4**	**10,2**	**0**	**186,6**
2008	**176,4**	**10,2**	**0**	**186,6**

(Beträge in Tausend EUR)

3.2. Zusammenfassung

In den Sonderforschungsbereichen 441, 538 und 632 werden elektronische Kollektionen linguistischer Daten erstellt, die für die Untersuchung sprachwissenschaftlicher Fragestellungen genutzt werden. Diese empirischen Ressourcen sind über die jeweiligen SFBs hinaus für die linguistische und philologische Forschung insgesamt von hohem Nutzen. Das Projekt C2, das im SFB 441 als Kooperationsprojekt der drei SFBs beantragt wird, hat das Ziel, die Voraussetzungen für die nachhaltige allgemeine Verfügbarkeit dieser Daten auch nach der Beendigung der SFBs zu schaffen.

Die in den drei beteiligten SFBs vorhandenen Daten zeichnen sich durch ein hohes Maß an Heterogenität aus. Bereits innerhalb der einzelnen SFBs ist eine signifikante Diversität der Ressourcen zu konstatieren. Betrachtet man die Daten-Kollektionen aller drei SFBs zusammen, zeigt sich diese Diversität umso ausgeprägter. Insgesamt decken diese Ressourcen ein breites Spektrum an zentralen Datentypen und typischen Daten ab (Korpora geschriebener und gesprochener Sprache; synchrone und diachrone Daten; hierarchische und Zeitachsen-basierte Annotationen auf verschiedenen Ebenen; lexikalische Ressourcen und andere Sekundärdaten etc.). Daraus ergibt sich, dass das Ziel der nachhaltigen Verfügbarkeit dieser Daten grundsätzliche Herausforderungen stellt, die exemplarischen Charakter für die Nachhaltigkeit linguistischer Daten-Kollektionen insgesamt haben. In diesem Sinn soll das beantragte Projekt generische Lösungen entwickeln, die auf andere linguistische Daten-Kollektionen übertragbar sind. Insbesondere soll ein genereller infrastruktureller Rahmen erarbeitet werden, der für linguistische Daten-Ressourcen im Allgemeinen offen ist. Dieser Rahmen soll mit anderen Nachhaltigkeits-Initiativen kompatibel sein.

3.3. Ausgangssituation des Teilprojekts

3.3.1. Stand der Forschung

In den letzten Jahren ist die Frage der nachhaltigen Verfügbarkeit von empirisch er-

hobenen Daten immer stärker in den Blick gekommen. Diese Verfügbarkeit soll gewährleisten, dass die in einem Forschungsprojekt (oft mit hohem Zeit- und Personalaufwand) gesammelten und aufbereiteten Daten auch nach Abschluss dieses Projekts generell für andere Forschungsaktivitäten wiederverwendbar sind, d.h. nicht als sog. „Datenfriedhöfe“ weiter existieren und anderen Forschern unzugänglich bleiben. Dies setzt zum einen voraus, dass die Archivierung und Distribution solcher Daten langfristig institutionell gesichert ist. Zum anderen ist es notwendig, dass die Daten in einem Format vorliegen, das unabhängig von speziellen Hard- und Software-Plattformen ist, damit die abzusehende Veralterung dieser Plattformen nicht zum faktischen Verlust der Daten führt. Dieses Problem ist umso drängender, als angesichts der derzeitigen schnellen Entwicklungszyklen im IT-Bereich sowohl Hardware-Technologien als auch Software-Systeme bereits nach wenigen Jahren von neuen Produkten abgelöst und dadurch obsolet werden.

Forschungsinitiativen zur nachhaltigen Datenarchivierung Um für die Problematik der nachhaltigen Datenverfügbarkeit grundsätzliche Lösungen zu erarbeiten, sind zahlreiche Forschungsinitiativen ins Leben gerufen worden. In Deutschland gehören dazu das DFG-Förderprogramm „Leistungszentren für Forschungsinformation“, das u.a. die exemplarische Umsetzung innovativer Konzepte für den Aufbau digitaler Text- und Datenzentren zur Archivierung und Bereitstellung relevanter Daten für Forschung und Lehre fördert, sowie das BMBF-Projekt „nestor“[1], das ein breites Informations- und Kommunikations-Netzwerk von Institutionen, Organisationen und Unternehmen etabliert, die mit diesen Fragen befasst sind, und das verwandte BMBF-Projekt „KOPAL“[2], das eine generelle Lösung für technische Herausforderungen der Langzeitarchivierung und -verfügbarkeit erarbeitet und implementiert. Diese Initiativen behandeln generelle (d.h. fächerunabhängige) infrastrukturelle und technische Aspekte der Nachhaltigkeitsproblematik. Diese Aspekte sind auch für das hier beantragte Projekt C2 von hoher Relevanz. In Komplementarität zu den genannten Initiativen verfolgt C2 jedoch eine fachspezifische Perspektive: Hier stehen die speziellen Herausforderungen der nachhaltigen Aufbereitung linguistischer Ressourcen im Vordergrund.

Zentren für Sprach-Ressourcen Die Etablierung eines umfassenden Zentrums für Sprach-Ressourcen wird durch das vom DFKI und der Universität des Saarlandes getragene Projekt COLLATE[3] verfolgt. Diesem Projekt obliegt der Auf- und Ausbau des Deutschen Kompetenzzentrums für Sprachtechnologie in Saarbrücken. Der

[1]Network of Expertise in long-term STorage and availability of digital Resources in Germany — Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen in Deutschland

[2]Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen

[3]Computational Linguistics and Language Technology for Real Life Applications: http://collate.dfki.de

Zweck dieses Kompetenzzentrums weist gewisse Berührungspunkte zur prospektiven infrastrukturellen Plattform für nachhaltige linguistische Daten-Kollektionen auf, für die das beantragte SFB-Projekt C2 grundsätzliche Leitlinien erarbeiten soll. In beiden Fällen geht es um die breite Verfügbarkeit sprachbezogener Forschungsergebnisse. Jedoch sind in mehrfacher Hinsicht unterschiedliche Schwerpunktsetzungen zu konstatieren. Zum einen stehen bei COLLATE Sprach*technologien* im Vordergrund, während sich C2 auf Kollektionen von Sprach*daten* konzentriert. Des Weiteren sind die Leistungen des durch COLLATE etablierten Kompetenzzentrums vornehmlich auf sprachtechnologisch orientierte Wissenschaftler sowie industrielle Hersteller und Anwender von Sprachtechnologien zugeschnitten. Dagegen wendet sich eine Infrastruktur, deren Prinzipien in C2 entwickelt werden sollen, an Forscher im Bereich der Linguistik und der Philologien. Schließlich verfolgt das Kompetenzzentrum in Saarbrücken das Grundanliegen, umfassende Informationen über eine möglichst breite Palette existierender Ressourcen und Systeme bereitzustellen und dadurch potenzielle Nutzer sowie die interessierte Öffentlichkeit über den aktuellen Entwicklungsstand der Sprachtechnologie zu informieren und einen kurzfristigen Transfer von Forschungsergebnissen in die industrielle Anwendung zu ermöglichen. Demgegenüber hat die in C2 verfolgte Initiative das Ziel, existierende Sprachdaten gezielt aufzubereiten, um ihre langfristige Nutzung zu gewährleisten. Beide Initiativen verhalten sich somit auf vielfache Weise komplementär zueinander.

Linguistische Ressourcen und ihre Kodierung In den letzten Jahren sind zahlreiche linguistische Daten-Ressourcen (annotierte Korpora, Lexika etc.) erstellt worden, die eine empirische Grundlage für die Entwicklung von maschinellen Sprachverarbeitungssystemen und gleichzeitig für die linguistische Forschung bilden. Dabei werden in den verschiedenen Ressourcen i.d.R. unterschiedliche Annotationsschemata und Verarbeitungsmechanismen (Abfragesprachen, Tools) eingesetzt (vgl. (Sasaki, Witt, Gibbon & Trippel 2004) für eine Übersicht über gängige Formate und Query-Sprachen). Bei aller Diversität ist jedoch zu konstatieren, dass sich als generelles Kodierungs-Format für linguistische Ressourcen die Auszeichnungssprache XML inzwischen weitgehend durchgesetzt hat. Bei den meisten Ressourcen ist XML (oder SGML) entweder das originäre Kodierungs-Format, oder es existiert eine zum originären Kodierungs-Format äquivalente (d.h. informationsgleiche) XML-Version. Von der Nachhaltigkeits-Problematik her gesehen birgt diese Entwicklung einige Vorteile. XML ist ein offener Standard, der vom World Wide Web Consortium (W3C)[4] gepflegt wird. Neben dem eigentlichen XML-Standard wurden und werden weitere Standards entwickelt, die auf XML basieren oder mit XML interagieren, z.B. Unicode zur Kodierung von Zeichensätzen, oder XPath, XSLT und XQuery zur Suche nach XML-Strukturen bzw. zur Transformation von XML-Dokumenten. Es existieren zahlreiche kommerzielle und freie Software-Systeme, die diese Standards implementieren

[4] http://www.w3.org

bzw. unterstützen (z.B. XSLT-Prozessoren). Da sich XML als grundlegende Technologie zur Repräsentation von Daten und Dokumenten im WWW und zur Verwaltung von Informationsbeständen allgemein immer stärker durchsetzt, ist davon auszugehen, dass derartige Software für die entsprechenden Standards auch langfristig zur Verfügung stehen wird.

Korpus-Annotationsstandards Die nachhaltige Pflege und Zugänglichkeit linguistischer Ressourcen würde wesentlich vereinfacht, wenn anstatt der oben konstatierten Diversität der Annotationsschemata diese Ressourcen nach gemeinsamen Annotationsstandards kodiert wären. In der Vergangenheit sind mehrere generelle Annotationsstandards vorgeschlagen worden, z.B. TEI[5] oder XCES[6]. In den letzten Jahren hat sich jedoch in der Praxis der Korpuserstellung gezeigt, dass diese Standards projektspezifische Annotationsformate nicht ersetzen können. Vielmehr muss ein konkretes Annotationsformat Besonderheiten der zu kodierenden Phänomene, die zugrunde liegende linguistische Theorie sowie spezielle Erfordernisse im Annotations- und Verarbeitungsprozess berücksichtigen, was ein genereller Standard naturgemäß nicht leisten kann. Um trotz der Diversität der eingesetzten Annotationsschemata ein Höchstmaß an Vergleichbarkeit und Wiederverwendbarkeit von linguistischen Ressourcen herzustellen, wird innerhalb des ISO Technical Commitee 37 / sub-commitee 4 (ISO TC37/SC4)[7] eine Infrastruktur, das *Linguistic Annotation Framework (LAF)*, erarbeitet, die es ermöglicht, Korpus-spezifische Annotationen auf ein generisches Format abzubilden (Ide, Romary & de la Clergerie 2003). Dieses generische Format beinhaltet zwei Komponenten, die Struktur und Inhalt der Annotation voneinander trennen. Eine Komponente bildet ein generisches Meta-Modell linguistischer Strukturen. Dieses Modell erfasst Konstituentenstrukturen und sekundäre Relationen ebenso wie Mengen, Listen und Merkmalstrukturen. Zu diesem Modell existiert eine isomorphes XML-Schema, das sog. *dump format*. Auf die zweite Komponente, das *Data Category Registry (DCR)*, kommen wir unten zu sprechen.

Datenmodelle für Korpus-Annotationen Den für linguistische Ressourcen verwendeten Annotationsschemata liegen auf einer abstrakteren Ebene unterschiedliche Datenmodelle zugrunde, die nicht zuletzt durch die jeweils zu erfassenden Phänomene und deren Struktur determiniert werden. So erfordert die Annotation syntaktischer Strukturen, z.B. in Baumbanken, ein hierarchisches Datenmodell, während für die Transkription von Video-Aufzeichnungen (verschiedene Sprecher, Gestik, situativer Kontext) ein Zeitachsen-basiertes Mehr-Ebenen-Modell adäquat ist, d.h. ein Modell bestehend aus mehreren flachen (nicht-hierarchischen) Ebenen, die durch eine gemeinsame Zeitachse aligniert sind. Im EU-Projekt NITE[8] wurde mit dem NITE

[5]Text Encoding Initiative: http://www.tei-c.org

[6]XML Corpus Encoding Standard: http://www.xml-ces.org

[7]http://www.tc37sc4.org

[8]Natural Interactivity Tools Engineering: http://nite.nis.sdu.dk

Object Model (NOM) ein Meta-Modell entwickelt, das beide genannten Datenmodelle subsumiert (Carletta, Kilgour, O'Donnell, Evert & Voormann 2003). Es ist zu beachten, dass NOM und LAF komplementäre Schwerpunkte setzen: Während NOM ein Framework für die explizite Kodierung von Korpus-Strukturen (Annotationsebenen und ihre Beziehungen zueinander) darstellt, spezifiziert LAF Normierungen für die Kodierung der individuellen linguistischen Phänomene. Zum NOM wurde eine Query-Sprache entwickelt, deren Mächtigkeit der Prädikaten-Logik erster Stufe (einschließlich All-Quantor) entspricht und die die Abfrage von sowohl hierarchischen (z.B. Dominanz) also auch temporalen Relationen (z.B. Präzedenz, Überlappung oder Inklusion) ermöglicht. Ein Prototyp eines entsprechenden Abfrage-Tools wurde implementiert (Voormann, Evert, Kilgour & Carletta 2003). Wie in (Heid, Voormann, Milde, Gut, Erk & Pado 2004) demonstriert wurde, können sowohl hierarchisch strukturierte Annotationen als auch Mehr-Ebenen-Annotationen in ein NOM-konformes Format überführt werden und sind mit Hilfe der NITE-Query-Sprache abfragbar. Das NITE Object Model eignet sich also als generelles Datenmodell für linguistische Korpora, das eine umfassende Grundlage für die Realisierung genereller Formate und Zugriffsmechanismen darstellt.

Metadaten Mit der rapide wachsenden Zahl verfügbarer Sprachressourcen steigt die Notwendigkeit, gezielt nach Ressourcen suchen zu können, die bestimmte, für die eigene Verwendung erforderliche Kriterien erfüllen. Für die nachhaltige Nutzbarkeit solcher Ressourcen ist es essenziell, dass sie in der Vielzahl existierender Daten-Kollektionen auffindbar sind und ihre jeweilige potenzielle Relevanz für individuelle Forschungsvorhaben beurteilt werden kann. Eine notwendige Voraussetzung für eine Infrastruktur, die derartige Suchen ermöglicht, ist eine formale Spezifikation solcher Kriterien, d.h. von Metadaten (Informationen) über linguistische Ressourcen wie Art der Ressource, Entstehungszeit, Sprache(n), Größe, kodierte Informationen etc. Z.Zt. existieren zwei internationale Initiativen zur Entwicklung einheitlicher Standards zur Metadaten-Spezifikation: IMDI[9] und OLAC[10]. Diese Initiativen haben außerdem das Ziel, allgemein zugängliche Repositorien von Metadaten über möglichst viele Ressourcen aufzubauen, sodass eine umfassende Suche nach Ressourcen mit bestimmten Eigenschaften möglich ist.

Integration linguistischer Terminologien Im Abschnitt *Korpus-Annotationsstandards* wurde schon auf das *Linguistic Annotation Framework (LAF)* verwiesen, das darauf abzielt, Annotationsschemata vergleichbar und damit wiederverwendbar zu machen. Während oben der Aspekt der *Struktur* von Annotationen im Vordergrund stand, geht es bei der zweiten Komponente von LAF, dem im Aufbau befindlichen *Data Category Registry (DCR)*, um den *Inhalt* der Annotationen. Das DCR wird eine Kollektion von linguistischen Referenz-Kategorien (z.B. GENDER) bereitstellen, in-

[9] ISLE Metadata Initiative: http://www.mpi.nl/IMDI

[10] Open Language Archive Community: http://www.language-archives.org

klusive generellen (theorie- und sprachunabhängigen) Definitionen, Verwendungsbeispielen sowie ggf. Wertebereichen (z.B. MASC, FEM, NEUT). Die Tags spezifischer Annotationsschemata können dann auf diese Referenzkategorien abgebildet werden. In eine ähnliche Richtung gehen die typologisch ausgerichteten Datenbankprojekte wie das *Typological Database Project* (Monachesi, Dimitriadis, Goedemans, Mineur & Pinto 2001), AUTOTYP[11] (Bickels & Nichols 2002) oder das *Cross-linguistic Reference Grammar Project*[12].
Daneben sind die Ontologie-orientierten Ansätze im Umfeld des Semantic Web zu nennen, deren Referenzkategorien in Ontologien organisiert sind, die auch Beziehungen zwischen Kategorien modellieren und das Ziehen von Schlüssen erlauben. Zu diesen Ansätzen zählen z.B. *DOLCE: a Descriptive Ontology for Linguistic and Cognitive Engineering*[13] und die im Rahmen des E-MELD-Projekts[14] entstandene *Generalized Ontology for Linguistic Description (GOLD)* (Farrar & Langendoen 2003).
Für alle Initiativen gilt, dass sie flexibel und modifizierbar konzipiert sind, so dass sie für spezifische Anforderungen, die von den z.T. eher generell formulierten Referenzkategorien nicht abgedeckt werden, erweitert werden können. Um die Erstellung und Erweiterung der Ontologien zu erleichtern, wird derzeit im Rahmen eines Nachfolgeprojekts von GOLD (*Data Driven Ontology Development, DDLOD)*[15] untersucht, wie linguistische Ontologien mit Hilfe von im Internet verfügbaren glossierten Daten datengetrieben und halb-automatisch erstellt werden können.

Rules of Best Practice Für systematische Überlegungen zur Nachhaltigkeits-Problematik bei linguistischen Daten-Kollektionen ist insbesondere (Bird & Simons 2003) richtungsweisend. In dieser Arbeit werden sog. *Rules of Best Practice* für die Erstellung, Archivierung und Distribution von Sprachressourcen motiviert und formuliert. Diese Regeln erfassen die Problematik anhand von sieben Dimensionen (Aspekten): „content, format, discovery, access, citation, preservation, and rights“. So wird beispielsweise hinsichtlich der Dimension Format empfohlen, die Daten mit Hilfe von offenen, nicht-proprietären Standards zu kodieren, die durch Software verschiedener Anbieter unterstützt werden, also z.B. MP3 für Audio-Aufzeichnungen, Unicode für Schriftzeichen und XML für Annotationen. Die Nutzung dieser Formate vermeidet die Abhängigkeit von einer bestimmten Software. Außerdem wird empfohlen, neben maschinellen Formaten die Daten auch in einem für Menschen lesbaren Format (z.B. Papierdruck) verfügbar zu halten. Die in (Bird & Simons 2003) spezifizierten Best-Practice-Richtlinien sind im Rahmen zweier Initiativen entstanden: OLAC (s.o.), ein offenes internationales Netzwerk von Partnern, die sich mit der Archivie-

[11] http://www.uni-leipzig.de/∼autotyp
[12] http://www.crg.lmu.de
[13] http://www.loa-cnr.it/DOLCE.html
[14] Electronic Metadata for Endangered Languages Data: http://emeld.org
[15] http://zimmer.csufresno.edu/∼wlewis/projects/DDLOD.html

rung von Sprachressourcen befassen, und EMELD (s.o.), ein Projekt zur Entwicklung von Standards zur Archivierung von Daten und Dokumentationen über Sprachen, die vom Aussterben bedroht sind. Bird und Simons betonen, dass die Erarbeitung adäquater Best-Practice-Richtlinien keineswegs abgeschlossen ist, und fordern zu einer breiten Diskussion über solche Regeln auf. Dementsprechend bietet OLAC eine offene technische Plattform (in Form von Mailing-Listen), die eine derartige Diskussion ermöglicht.

Zugriffsbeschränkungen Obwohl die Existenz und Relevanz von rechtlichen und ethischen Problemen, die sich im Zusammenhang mit der Veröffentlichung von sprachlichen Daten stellen, kaum bestritten werden, sind — zumindest in der Linguistik im deutschsprachigen Raum — bislang kaum Publikationen zu diesem Thema entstanden. Einen Überblick über die bestehenden Schwierigkeiten sowie erste Empfehlungen zu einer diesbezüglichen Best-Practice geben sowohl (Bird & Simons 2003) als auch (Baude, Blanche-Benveniste, Calas, Cordereix, De Lambertierie, Goury, Jacobson, Marchello-Nizia & Mondada 2005). Im Rahmen der Talkbank-Initiative (MacWhinney, Bird, Cieri & Martell 2004) wurden diese Empfehlungen in Form eines 9-Stufen-Modells für den Zugriff auf linguistische Datenarchive konkretisiert. Während diese Arbeit vornehmlich auf die Wahrung der Persönlichkeitsrechte von aufgenommenen Personen abzielt, wurde in der CODATA-Initiative ein Konzept zur „Zitierfähigkeit wissenschaftlicher Primärdaten“ (Lautenschlager & Sens 2002) erarbeitet, das sich vor allem mit Belangen des wissenschaftlichen Wettbewerbs auseinander setzt.
Weitere Expertise im Bereich rechtlicher und ethischer Probleme im Zusammenhang mit Datenarchiven ist zum einen in publizierten Arbeiten aus anderen Wissenschaftsgebieten dokumentiert (z.B. für die Sozialforschung: diverse Beiträge aus (Corti, Kluge, Mruck & Opitz 2000), (Corti, Witzel & Bishop 2005), für die Medizin: (NIH 2003)), zum anderen in Einrichtungen, die bereits Archive betreiben (wie den Max-Planck-Instituten in Nijmegen und Leipzig und dem IDS), aber auch in Diskussionsforen der betroffenen linguistischen Forschergemeinschaften (z.B. Linguist List[16], Corpora-List[17], Mailingliste Gesprächsforschung[18]) zu suchen.

3.3.2. Eigene Vorarbeiten

Die Sonderforschungsbereiche 441 „Linguistische Datenstrukturen“ (Tübingen), 538 „Mehrsprachigkeit“ (Hamburg) und 632 „Informationsstruktur“ (Potsdam/Berlin) haben die grundlegende Gemeinsamkeit, dass im Rahmen ihrer Forschungsaktivitäten elektronische Sammlungen linguistischer Daten aufgebaut werden, die als empirische Grundlage für die jeweils untersuchten linguistischen bzw. philologischen Fragestel-

[16] http://www.linguistlist.org
[17] http://nora.hd.uib.no/fileserv.html
[18] http://www.gespraechsforschung.de

lungen dienen. Den größten Teil dieser Kollektionen bilden Korpora geschriebener oder gesprochener Sprache, wobei letztere jeweils einen nicht-textuellen Bestandteil (Audio- bzw. Videoaufnahmen von Interaktionen) und einen textuellen Bestandteil (deren Transkriptionen mit zugehörigen Annotationen) haben. Darüber hinaus werden jedoch auch andere Typen von Daten-Kollektionen erstellt, z.B. Sammlungen von Materialien für die Datenerhebung (Audios, Videos oder Graphiken, mit deren Hilfe Daten elizitiert werden), Daten aus bildgebenden Verfahren, Dokumentationen von Experimenten für die Elizitation von Daten gesprochener Sprache, Datenbanken mit grammatischen Merkmalen der in den Korpora dokumentierten Sprachen und lexikalische Ressourcen. In den genannten SFBs erfolgt der Aufbau der verschiedenen Daten-Ressourcen (Transkription von Audio- bzw. Video-Aufzeichnungen; Annotation von Texten mit Informationen auf verschiedenen Ebenen; Kompilation lexikalischer Daten etc.) dezentral in den einzelnen SFB-Teilprojekten, unter besonderer Berücksichtigung des jeweiligen Forschungsgegenstandes. Entsprechend unterscheiden sich die in den jeweiligen Projekten erstellten Daten in mehrfacher Hinsicht, z.B den Datentypen bzw. Textsorten (gesprochene Sprache, Zeitungstexte, historische Texte etc.), Sprachen, Annotationsebenen, d.h. den annotierten Informationen (Text- und Dialogstruktur, (morpho-)syntaktische Informationen, Diskursinformationen, situativer Kontext etc.), oder den zugrunde liegenden linguistischen Theorien. In jedem der drei SFBs existieren Projekte, die sich mit der Integration und zentralen Aufbereitung der verschiedenen Daten-Kollektionen befassen (Projekt C1 im SFB 441, Projekt Z2 im SFB 538, D-Projekte im SFB 632). Diese Projekte haben die Aufgabe, einheitliche Annotationsschemata und Datenformate für die im jeweiligen SFB erstellten Daten zu entwickeln sowie, darauf aufbauend, Werkzeuge für die Annotation, Verwaltung und Abfrage der Daten zu implementieren oder anzupassen.

Aufgrund der jeweiligen Ausrichtung der einzelnen SFBs ergeben sich unterschiedliche Fragestellungen und Schwerpunkte, die wiederum unterschiedliche Lösungen bei der Struktur des Annotationsschemas sowie der Auswahl und Implementierung geeigneter Annotations- und Abfrage-Werkzeuge erfordern. So werden im SFB 538 in großem Umfang Aufnahmen gesprochener Sprache transkribiert. Bei derartigen Daten ist es adäquat, für das gemeinsame Annotationsschema ein Mehr-Ebenen-Datenmodell zugrunde zu legen, bei dem sich verschiedene parallele Ebenen (*tiers*) an einer gemeinsamen Zeitachse orientieren. Dagegen stehen in den im SFB 441 annotierten Korpora komplexe textuelle und syntaktische Strukturen im Vordergrund. Hierfür bietet sich ein Annotationsschema an, das auf einem hierarchischen Datenmodell beruht. Die im SFB 632 erstellten Daten zeichnen sich durch eine besonders große Varietät der erfassten Annotationsebenen aus. Alle annotierten Daten dienen dort jedoch der Erforschung von Informationsstrukturen. Dies ermöglicht eine größere inhaltliche Kohärenz des Annotationsschemas für die verschiedenen Teilkorpora als in den beiden anderen SFBs, in denen eine große Bandbreite linguistischer Phänomene untersucht wird. Zusammen genommen repräsentieren die in den drei SFBs

vorhandenen Ressourcen ein breites Spektrum von zentralen Datentypen und generischen Datenstrukturen, die exemplarisch für linguistische Daten-Ressourcen im Allgemeinen sind.

Insgesamt erfordert die Verwaltung und Erschließung der in den drei SFBs erstellten Daten angesichts ihrer Komplexität und Heterogenität umfassende und detaillierte Kenntnisse im Hinblick auf Strukturen, Formate und Zugriffsmechanismen. Diese Expertise — und damit letztlich die nachhaltige Nutzbarkeit der Daten — droht nach Beendigung der SFBs verloren zu gehen, sofern keine geeigneten Maßnahmen getroffen werden.

Im Folgenden werden die in den einzelnen SFBs geleisteten Vorarbeiten kurz skizziert. Eine detaillierte Übersicht über die erstellten Daten-Kollektionen findet sich im Anhang.

SFB 441 (Tübingen)

Im SFB 441 wurden bzw. werden Korpora und andere Daten-Kollektionen von den Projekten A1, A3, B1, B3, B6, B8, B9, B10, B11 und B16 erstellt. Diese Daten werden zum gemeinsamen Repositorium TUSNELDA (= TUebinger Sammlung Nutzbarer Empirischer Linguistischer DAtenstrukturen) zusammengefasst. TUSNELDA umfasst größtenteils Korpora geschriebener Sprache; in geringerem Umfang sind Transkriptionen gesprochener Sprache vertreten. Diese Korpora unterscheiden sich nicht nur hinsichtlich der erfassten Textsorten, Sprachen und der jeweils kodierten (Kombinationen von) Annotationsebenen; auch die Größe der einzelnen Korpora variiert erheblich.[19] Neben den Korpora bilden einige relationale Datenbanken weitere Komponenten von TUSNELDA.

Für die Korpora in TUSNELDA wurde im Projekt C1 ein gemeinsames XML-Annotationsschema ausgearbeitet. Als Ausgangspunkt dienten die existierenden Annotationsstandards XCES und TEI. Dieses Schema beruht, wie bereits erwähnt, auf einem hierarchischen Datenmodell, da im SFB 441 die Erfassung hierarchischer Strukturen (Text, Syntax) im Vordergrund stehen. Nicht-hierarchische Phänomene, z.B. Koreferenz-Beziehungen, werden durch Pointer in der XML-Struktur kodiert. Das TUSNELDA-Annotationsschema ist in (Wagner & Kallmeyer 2001), (Kallmeyer & Wagner 2000), (Wagner & Zeisler 2004) und (Wagner 2005) beschrieben. Detaillierte Guidelines finden sich in (Kallmeyer, Meyer & Wagner 2001) und (Wagner 2004). Speziell für das POS-Tagging des Russischen haben C1 und B1 ein im Rahmen von MULTEXT-East[20] entwickeltes Tagset verfeinert (vgl. (Meyer 2003)).

[19] Das größte Korpus ist die automatisch generierte deutsche Baumbank „TüPP-D/Z“ (A1) mit ca. 200 Millionen Wörtern.

[20] http://nl.ijs.si/ME

Für die Annotation der einzelnen Korpora werden, je nach optimaler Eignung verfügbarer Annotations-Werkzeuge, zwei Strategien verfolgt. Einige Korpora werden direkt im TUSNELDA-XML-Format mit Hilfe allgemeiner XML-Annotations-Tools kodiert. Insbesondere ist hier das CLaRK-System (Simov, Peev, Kouylekov, Simov, Dimitrov & Kiryakov 2001) zu nennen, das am Linguistic Modelling Laboratory der bulgarischen Akademie der Wissenschaften im Rahmen einer Kooperation mit dem Seminar für Sprachwissenschaft (Lehrstuhl Hinrichs) der Universität Tübingen realisiert wurde. Dieses System ist ein XML-Editor, der speziell für linguistische Annotation entwickelt wurde. Neben komfortablen Editierfunktionen und flexiblen Darstellungsoptionen enthält es mannigfaltige, flexibel konfigurierbare Tools zur Automatisierung von Annotationsschritten. Des Weiteren erlaubt das CLaRK-System die Suche in Dokumenten und deren Transformation mit XPath bzw. XSLT sowie Möglichkeiten zur Generierung von Statistiken über annotierten Werten und Strukturen. Für die Annotation des tibetischen Korpus (B11), das sich durch eine besonders reichhaltige Annotation auf vielen Ebenen auszeichnet (vgl. (Zeisler 2004)), wurden von C1 entsprechende XPath-basierte Tool-Konfigurationen sowie XSLT-Stylesheets für komplexe Such- und Statistikfunktionen spezifiziert (vgl. (Wagner & Zeisler 2004)). Für andere Korpora erfolgt die Annotation mit Hilfe einschlägiger Annotations-Tools, die mit einem proprietären Format arbeiten. So wurden z.B. die deutschen Baumbanken TüBa-D/Z (A1) (vgl. (Telljohann, Hinrichs & Kübler 2003), (Hinrichs, Kübler, Naumann, Telljohann & Trushkina 2004)) und SINBAD (A3) (vgl. (Kepser, Steiner & Sternefeld 2004)) mit *annotate* (Plaehn 1998) annotiert. In solchen Fällen müssen diese Formate in das TUSNELDA-XML-Schema überführt werden. Entsprechende Werkzeuge wurden in C1 implementiert.

Für die meisten Daten-Kollektionen[21] sind inzwischen Internet-Abfrage-Schnittstellen implementiert worden, welche unter `www.sfb441.uni-tuebingen.de/tusnelda.html` zugänglich sind. Diese Schnittstellen operieren teilweise auf Daten im TUSNELDA-Format, wobei XML-spezifische, aber proprietäre Zugriffssoftware eingesetzt wird. (Z.Zt. existiert eine Schnittstelle für Korpora mit vorwiegend textstruktureller und einzelphänomenorientierter Annotation sowie der Prototyp einer Abfrageoberfläche für Baumbanken im TUSNELDA-Format.) Teilweise wurden sie, um eine höhere Effizienz zu erreichen, mit Hilfe spezieller Abfrage-Software[22] realisiert, die ein spezifisches, für Abfragen optimiertes Datenformat voraussetzen.[23] (Dies gilt für die Abfrageoberfläche für russische Kor-

[21] D.h. für diejenigen Kollektionen, deren Aufbau hinreichend fortgeschritten ist und bei denen keine Copyright-Restriktionen vorliegen.

[22] Im Falle der Baumbank SINBAD (A3) wurde diese Software am SFB selbst, im Projekt A2, entwickelt, vgl. (Kepser 2003).

[23] Die Erhöhung der Abfrage-Effizienz bei direkter Verwendung von XML-Strukturen ist ein wesentlicher Untersuchungsgegenstand von C1 in der laufenden Phase.

pora (B1) sowie das Interface zur Beispielsatz-Sammlung SINBAD). Ferner existiert eine Web-Anbindung für die im SFB 441 erstellten relationalen Datenbanken. Neben diesen Internet-Schnittstellen ist im Projekt A1 das Offline-Abfragetool VIQTORYA für Baumbanken entwickelt worden, das über eine komfortable Benutzeroberfläche verfügt (Steiner & Kallmeyer 2002).

Neben der Kodierung linguistischer Phänomene selbst umfasst das TUSNELDA-Schema auch Spezifikationen für Metadaten. Durch den Kontakt des SFB mit dem EU-Projekt INTERA[24], das den Aufbau eines Metadaten-Repositoriums für linguistische Ressourcen auf der Basis des IMDI-Schemas zum Ziel hat, wird dieses Schema erweitert und ggf. modifiziert, um die Kompatibilität mit IMDI zu gewährleisten.

Insgesamt verfügt der SFB 441 über erhebliche Kompetenzen auf dem Gebiet der Texttechnologie. Neben den hier skizzierten (hauptsächlich von C1 getragenen) Aktivitäten im Zusammenhang mit TUSNELDA zeigt sich dies nicht zuletzt an diversen Aktivitäten der Projektleiter von A1 (Hinrichs/Kübler) und A2 (Mönnich) und ihrer Mitarbeiter. Genannt seien hier die langjährigen Erfahrungen am Lehrstuhl Hinrichs mit dem Aufbau und der Distribution unterschiedlicher linguistischer Ressourcen[25] im Rahmen verschiedener Projekte (Verbmobil, GermaNet, MiLCA, DEREKO, KIT etc.) sowie die Mitwirkung der Arbeitsgruppe Theoretische Computerlinguistik (Mönnich) bei der DFG-Forschergruppe „Texttechnologische Informationsmodellierung“ an der Universität Bielefeld, die mit dem Projekt C2 („Adaptive Ontologien auf extremen Auszeichnungsstrukturen“) in dieser Forschergruppe vertreten ist.

SFB 538 (Hamburg)

Am SFB 538 wird seit Juli 2000 im Projekt Zb „Mehrsprachige Datenbank“ (ab Juli 2005: Projekt Z2 „Computergestützte Erfassungs- und Analysemethoden multilingualer Daten“) an Konzepten und Werkzeugen für die Verarbeitung von Korpora gesprochener und geschriebener Sprache gearbeitet. Im Mittelpunkt steht dabei die Entwicklung eines Verfahrens, mit dem sich Transkriptionen gesprochener Sprache über spezifische Verwendungszusammenhänge (wie Einzelsprachen, linguistische Theorien, technische Umgebungen) hinweg erstellen, austauschen und bewahren lassen. Zu diesem Zwecke wurde das XML-basierte System EXMARaLDA (Extensible Markup Language for Discourse Annotation) entworfen und in Form von mehreren aufeinander aufbauenden Datenformaten (vgl. (Schmidt 2002) und (Schmidt 2005*d*)) und miteinander interagierenden Software-Werkzeugen (hauptsächlich: ein Transkriptionseditor, ein Korpus-Manager und ein Suchwerkzeug, vgl. (Schmidt & Wörner 2005) implementiert.

Ältere am SFB 538 vorhandene Datenbestände, die sich in ihrer ursprünglichen Form

[24] INTegrated European language data Repository Area

[25] vgl. http://www.sfs.uni-tuebingen.de/de_nf_asc_resources.shtml

nicht für eine Weiterverarbeitung oder einen Austausch eigneten, wurden und werden nach EXMARaLDA konvertiert, so dass zum gegenwärtigen Zeitpunkt fünf Projektkorpora (Projekte E3, E5, K1, K2 und K5) in diesem Format vorliegen, während drei weitere (Projekte E1, E2 und E4) noch ihre ursprüngliche Form (dBase bzw. andere proprietäre Formate) haben. Von den ab Juli 2005 neu hinzukommenden Projekten werden drei (H6, K6 und K8) ihre Datenverarbeitung von Beginn an mit EXMARaLDA durchführen. Des Weiteren wurden drei schriftsprachliche Korpora (Projekte K4, H3 und H4) in TEI-konformes XML überführt, um auch hier bessere Möglichkeiten einer langfristigen Erhaltung zu eröffnen.

Durch die Nutzung von EXMARaLDA für die Korpora gesprochener Sprache und von anderen XML-basierten Technologien für die Korpora geschriebener Sprache ist damit am SFB 538 eine relativ homogene Ausgangsbasis vorhanden, von der ausgehend Lösungen für eine nachhaltige Datenarchivierung erarbeitet werden können. Von zusätzlichem Interesse für das hier beantragte Vorhaben dürfte zum einen sein, dass die EXMARaLDA-Formate und -Werkzeuge auch über den SFB 538 hinaus genutzt werden[26] und somit erstens auch andernorts entsprechende Korpora in der Entstehung begriffen sind, zweitens EXMARaLDA als Anknüpfungspunkt für eine breiter angelegte Diskussion über nachhaltige Datenarchivierung genutzt werden kann (vgl. (Schmidt 2005*b*)). Zum anderen stellt die Einbindung von EXMARaLDA in die aktuelle texttechnologische Forschung (u.a. in Form einer Kooperation mit der Forschergruppe „Texttechnologische Informationsmodellierung“) sicher, dass die am SFB 538 entwickelten Lösungen nicht isoliert von anderen Entwicklungen in der Korpustechnologie bleiben[27] (vgl. dazu z.B. (Schmidt 2004*c*) und (MacWhinney, Schmidt, Martell, Wagner, Wittenburg, Brugman, Broeder & Hoffert 2004)).

SFB 632 (Potsdam/Berlin)

Im SFB 632 „Informationsstruktur“ werden auf der Basis von empirischen linguistischen Daten, die auf mehreren Ebenen annotiert sind, integrative Modelle der Informationsstruktur entwickelt. Die einzelnen Teilprojekte führen dazu selbst Datenerhebungen durch oder nutzen und annotieren bereits bestehende Korpora.
Die Mehrzahl der Daten im SFB 632 wird mit einem SFB-weiten Annotationsschema (dem „SFB-Annotationsstandard“) annotiert, das Merkmale der linguistischen Ebe-

[26] Z.B. am SFB 632 in Potsdam (vgl. (Dipper, Götze & Stede 2004), in DFG-geförderten Projekten an den Universitäten Dortmund, Mannheim, Flensburg, im Graduiertenkolleg „Bildungsgangforschung“ der Universität Hamburg, am Institut für Deutsche Sprache in Mannheim (vgl. (Schütte 2004)), sowie in zahlreichen Lehrveranstaltungen an deutschen und ausländischen Universitäten.

[27] Konkret äußert sich das vor allem in der Kompatibilität der EXMARaLDA-Datenformate mit ELAN, TASX, Praat, AIF, etc, die auch durch entsprechende Import- und Exportfilter unterstützt wird.

nen Phonetik, Phonologie, Morphologie, Syntax, Semantik, Diskurs- und Informationsstruktur umfasst und sich an bestehenden Annotationsstandards orientiert. Für dieses Annotationsschema werden derzeit in projektübergreifenden, interdisziplinären Arbeitsgruppen unter Beteiligung von Allgemeinen Sprachwissenschaftlern, Historischen Linguisten, Psycholinguisten, Typologen und Computerlinguisten weitgehend sprachunabhängige Annotationsrichtlinien entwickelt, so dass die annotierten Daten für verschiedene Forschergemeinden von Interesse sein werden.
Durch die Mitarbeit in den Arbeitsgruppen zu standardisierter Annotation bringen die D-Projekte eine reichhaltige Erfahrung mit in der Erfassung und Annotation sprachspezifischer Eigenheiten mit Hilfe standardisierter, genereller Tagsets. Dazu gehören sowohl das Definieren von Tags, die abstrakt und generell genug sind, um sprachenübergreifend eingesetzt werden zu können, wie auch das Erstellen der entsprechenden sprachenübergreifenden Annotationsrichtlinien.

Die Daten werden mit frei erhältlichen, spezialisierten Annotationswerkzeugen (annotate, EXMARaLDA, RSTTool, MMAX) annotiert (Dipper, Götze & Stede 2004) und dann in ein generisches, XML-basiertes Standoff-Format (Dipper & Götze 2005) überführt. Dieses bildet das Eingangsformat für die im SFB entwickelte linguistische Datenbank ANNIS ((Dipper, Götze, Stede & Wegst 2004), (Dipper & Götze 2005)), die die Visualisierung und Recherche in komplexen Mehrebenen-Annotationen über eine nutzerfreundliche Internetschnittstelle erlaubt. ANNIS selbst wird laufend weiterentwickelt und soll der Forschergemeinde als Open-Source-Projekt zur Verfügung gestellt werden.

3.3.3. Liste der publizierten einschlägigen Vorarbeiten

Dipper, S. & M. Götze (2005): „Accessing Heterogeneous Linguistic Data — Generic XML-based Representation and Flexible Visualization“, in *Proceedings of the 2nd Language & Technology Conference 2005*.

Dipper, S., M. Götze & M. Stede (Hrsg.) (2005): *Heterogeneity in Focus: Creating and Using Linguistic Databases*, Band 2 von *Interdisciplinary Studies on Information Structure (ISIS)*, Universitätsverlag Potsdam, Potsdam, Germany.

Dipper, S., M. Götze & M. Stede (2004): „Simple Annotation Tools for Complex Annotation Tasks: an Evaluation“, in *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, Lisboa, S. 54–62.

Dipper, S., M. Götze, M. Stede & T. Wegst (2004): „ANNIS: A Linguistic Database for Exploring Information Structure“, *in* S. Ishihara, M. Schmitz & A. Schwarz (Hrsg.), *Interdisciplinary Studies on Information Structure (ISIS)*, Band 1, Universitätsverlag Potsdam, Potsdam, Germany, S. 245–279.

Dipper, S., M. Götze & S. Skopeteas (2004): „Towards User-Adaptive Annotation Guidelines“, in *Proceedings of the COLING Workshop on Linguistically Interpre-*

ted Corpora LINC-2004, Geneva, Switzerland, S. 23–30.

Gévaudan, P. & D. Wiebel (2004): „Dynamic Lexicographic Data Modelling. A Diachronic Dictionary Development Report“, in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisboa, S. 263–266.

Hinrichs, E., S. Kübler, K. Naumann, H. Telljohann & J. Trushkina (2004): „Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank“, in *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, S. 51–62.

Kallmeyer, L. & A. Wagner (2000): „The TUSNELDA annotation standard: An XML encoding standard for multilingual corpora supporting various aspects of linguistic research“, in *Proceedings of Digital Resources for the Humanities*, Sheffield, S. 78–81.

Kallmeyer, L., R. Meyer & A. Wagner (2001): *Guidelines for the TUSNELDA Corpus Annotation Standard.*, SFB 441, Tübingen.
URL: *http://www.sfb441.uni-tuebingen.de/c1/TUSNELDA-guidelines.html*

Kepser, S. (2003): „Finite Structure Query — A Tool for Querying Syntactically Annotated Corpora“, in *Proceedings of EACL 2003*, S. 179–186.

Kepser, S., I. Steiner & W. Sternefeld (2004): „Annotating and Querying a Treebank of Suboptimal Structures“, in *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, S. 63–74.

MacWhinney, B., T. Schmidt, C. Martell, J. Wagner, P. Wittenburg, H. Brugman, D. Broeder & E. Hoffert (2004): „Collaborative Commentary: Opening Up Spoken Language Databases“, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Meyer, R. (2003): „Halbautomatische morphosyntaktische Annotation russischer Texte“, *in* R. Hammel & L. Geist (Hrsg.), *Linguistische Beiträge zur Slavistik. X. JungslavistInnen-Treffen in Berlin, 24.05.–27.05.2001*, Band 139 von *Specimina Philologiae Slavicae*, Verlag Otto Sagner, München, S. 192–205.

Penn, G., D. Meurers, K. De Kuthy, M. Haji-Abdolhosseini, V. Metcalf, S. Müller & H. Wunsch (2003): *Trale Milca Environment v. 2.5.0 User's Manual*.

Rehbein, J., T. Schmidt, B. Meyer, F. Watzke & A. Herkenrath (2004): „Handbuch für das Computergestützte Transkribieren nach HIAT“, *Working Papers in Multilingualism, Series B* 56.

Schmidt, T. (2001): „The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse“, in *Proceedings of the IRCS Workshop On Linguistic Databases*, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, S. 219–227.

Schmidt, T. (2002): „EXMARaLDA - ein System zur Diskurstranskription auf dem Computer“, *Working Papers in Multilingualism, Series B* 34.

Schmidt, T. (2003): „Korpus „Skandinavische Semikommunikation“ — ein mehrsprachiges Diskurskorpus auf XML-Basis“, in *Sprachtechnologie für die multilinguale Kommunikation - Textproduktion, Recherche, Übersetzung, Lokalisierung (Beiträge der GLDV-Frühjahrstagung 2003)*, gardez!, Sankt Augustin, S. 421–427.

Schmidt, T. (2004*a*): „EXMARaLDA — ein System zur computergestützten Diskurstranskription“, *in* A. Mehler & H. Libin (Hrsg.), *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, Verlag für Sozialwissenschaften, Wiesbaden, S. 203–218.

Schmidt, T. (2004*b*): „EXMARaLDA Partitur-Editor. Software and Manual“, *in* C. Stocker, D. Macher, R. Studler, N. Bubenhofer, D. Crvelin, R. Liniger & M. Volk (Hrsg.), *Studien-CD Linguistik — Multimediale Einführungen und interaktive Übungen zur germanistischen Sprachwissenschaft*, Niemeyer, Tübingen.

Schmidt, T. (2004*c*): „Transcribing and annotating spoken language with EXMARaLDA“, in *Proceedings of LREC 2004 Workshop on XML-based Richly Annotated Corpora*, Lisboa, S. 69–74.

Schmidt, T. (2005*a*): *Computergestützte Transkription — Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*, Peter Lang, Frankfurt a.M. (Zgl. Dissertation, Universität Dortmund).

Schmidt, T. (2005*b*): „Datenarchive für die Gesprächsforschung: Perspektiven, Probleme und Lösungsansätze“, *Gesprächsforschung (Online-Zeitschrift zur verbalen Interaktion)* . Im Erscheinen.

Schmidt, T. (2005*c*): „EXMARaLDA und die Datenbank „Mehrsprachigkeit“ — Konzepte und praktische Erfahrungen“, *in* S. Dipper, M. Götze & M. Stede (Hrsg.), *Heterogeneity in Focus: Creating and Using Linguistic Databases*, Band 2 von *ISIS (Interdisciplinary Studies on Information Structure), Working Papers of the SFB 632*, Potsdam, S. 21–42.

Schmidt, T. (2005*d*): „Time based data models and the Text Encoding Initiative's guidelines for transcriptions of speech“, *Working Papers in Multilingualism, Series B* 62. (Handed in to the P5 Initiative of the TEI).

Schmidt, T. & K. Wörner (2005): „Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA“, *Gesprächsforschung (Online-Zeitschrift zur verbalen Interaktion)* 6. Im Erscheinen.

Stede, M. (2004): „The Potsdam Commentary Corpus“, *in* B. Webber & D. K. Byron (Hrsg.), *ACL 2004 Workshop on Discourse Annotation*, Association for Computational Linguistics, Barcelona, Spain, S. 96–102.

Steiner, I. & L. Kallmeyer (2002): „VIQTORYA — A Visual Query Tool for Syn-

tactically Annotated Corpora“, in *Proceedings of the 3rd Language Resources and Evaluation Conference (LREC)*, Las Palmas, S. 1704–1711.

Telljohann, H., E. Hinrichs & S. Kübler (2003): *Stylebook for the Tübingen Treebank of Written German*, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen.

Wagner, A. (2004): *Guidelines for the TUSNELDA Corpus Annotation Standard. Part II: Syntactic Structures*, SFB 441, Tübingen.
URL: *http://www.sfb441.uni-tuebingen.de/c1/TUSNELDA-guidelines-II.html*

Wagner, A. (2005): „Unity in Diversity: Integrating Different Linguistic Data in TUSNELDA“, *in* S. Dipper, M. Götze & M. Stede (Hrsg.), *Heterogeneity in Focus: Creating and Using Linguistic Databases*, Band 2 von *ISIS (Interdisciplinary Studies on Information Structure), Working Papers of the SFB 632*, Potsdam, S. 1–20.

Wagner, A. & B. Zeisler (2004): „A syntactically annotated corpus of Tibetan“, in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisboa, S. 1141–1144.

Wagner, A. & L. Kallmeyer (2001): „Der TUSNELDA-Standard: Ein Korpusannotierungsstandard zur Unterstützung linguistischer Forschung“, in *Proceedings der GLDV-Frühjahrstagung*, Gießen, S. 253–262.

Zeisler, B. (2004): „An annotation of what is not there: Empty arguments and cross-clausal reference in spoken and written Tibetan texts“, in *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, S. 161–174.

3.4. Planung des Teilprojekts (Ziele, Methoden, Arbeitsprogramm)

Die in den drei SFBs aufgebauten Daten-Kollektionen sind nicht nur für die SFB-interne Forschung relevant. Vielmehr bilden sie empirische Ressourcen, von denen auch für andere Forschungsvorhaben ein hoher Nutzen zu erwarten ist und die ferner einen exemplarischen Querschnitt linguistischer Daten-Ressourcen repräsentieren. Aus diesem Grund ist es ein wichtiges Anliegen, diese Ressourcen für die Wissenschaftsgemeinschaft dauerhaft zur Verfügung zu stellen. Während der Laufzeit der SFBs obliegt diese Aufgabe den jeweiligen zentralen Daten-Projekten. Danach jedoch ist der institutionelle und prozedurale Rahmen für die Verwaltung und Distribution der Daten bislang nicht geklärt. Diese offene Frage soll durch das beantragte Projekt systematisch angegangen werden. Das Projekt C2 verfolgt drei Hauptziele:

1. die nachhaltige Verfügbarkeit der Daten nach Beendigung der SFBs sicherzustellen, und dies in einer mit anderen Nachhaltigkeitsinitiativen kompatiblen Form

2. benutzerfreundliche Oberflächen für den komfortablen Datenzugriff (Abfrage)

zu entwickeln, die Nachhaltigkeits-Anforderungen in maximaler Weise entsprechen

3. anhand dieser Lösungen exemplarische Best-Practice-Richtlinien für zukünftige, mit maschineller Erhebung, Erschließung und Verwaltung linguistischer Daten befasste Projekte zu erarbeiten (was letztlich Forschungszeit und Fördergelder in erheblichem Umfang sparen könnte)

4. durch diese Bemühungen zur Einrichtung einer offenen, für alle Linguisten und Philologen nutzbaren Plattform beizutragen.

Diese Ziele sollen durch die im Folgenden beschriebenen Arbeitspakete umgesetzt werden. Jeder der beteiligten SFBs soll in allen Paketen mitarbeiten und das Wissen über die Beschaffenheit der eigenen Daten sowie die sich daraus ergebenden Kompetenzen einbringen. Jedes Arbeitspaket soll von einem der drei SFBs koordiniert werden; hierbei fällt die Koordination von fünf Arbeitspaketen an den SFB 441 als dem federführenden SFB sowie die Koordination jeweils eines Arbeitspakets an die SFBs 538 und 632. Der jeweils koordinierende SFB hat die Aufgabe, in Kooperation mit den anderen SFBs die konkreten Aktivitäten im Arbeitspaket zu einer Agenda zu bündeln, Informationsaustausch und Diskussionen anzustoßen und zu organisieren sowie Ergebnisse zusammenzuführen. Um die persönliche Kommunikation hinsichtlich der vielfältigen Arbeitsschwerpunkte in ausreichendem Maß zu gewährleisten, sind regelmäßige Projekt-Workshops geplant, die viermal pro Jahr stattfinden sollen. Hierbei sollen i.d.R. zwei dieser jährlichen Workshops auch unter Beteiligung externer Experten[28] stattfinden, die zu bestimmten thematischen Schwerpunkten wertvolle Anregungen geben können. Die übrigen Treffen dienen ausschließlich dem internen Gedankenaustausch und der Koordination.

Die Arbeitspakete legen den Schwerpunkt bewusst auf den XML-kodierten textuellen Bestandteil der linguistischen Korpora. Was die nachhaltige Sicherung nicht-textueller Daten (insbesondere Audio- und Video-Daten) betrifft, stellen sich keine Probleme, die für die Linguistik spezifisch sind. Insofern können hier Lösungen, die im Rahmen allgemeiner Nachhaltigkeits-Initiativen für diese Datenformate entwickelt werden, voraussichtlich ohne Adaption übernommen werden. Ähnliches gilt für weitere nicht XML-basierte Datenformate (z.B. relationale Datenbanken), sofern sie für dieses Vorhaben relevant sind.

[28] Beispielsweise bestehen gute Arbeitsbeziehungen aller drei SFBs zur Forschergruppe Texttechnologische Informationsmodellierung in Bielefeld und zur Abteilung Korpuslinguistik des Instituts für deutsche Sprache und Linguistik an der Humboldt-Universität Berlin. Internationale Kontakte bestehen u.a. mit dem MPI in Nijmegen, dem Institute for Research in Cognitive Science an der University of Pennsylvania sowie dem Linguistic Data Consortium (Steven Bird).

3.4.1. Technische Aufbereitung der Daten

AP1 Entwicklung nachhaltiger Datenformate (Koordination: SFB 441)

Die in den SFBs erstellten Daten-Kollektionen sind in ein Format zu überführen, das allgemein zugänglichen und akzeptierten Standards entspricht und unabhängig von proprietärer Software ist. Wie in Abschnitt 3.3.1 erläutert, bietet sich aus Nachhaltigkeitsgründen eine Kodierung in XML an, zumal in den einzelnen SFBs bereits jetzt XML-Schemata für die Kodierung der Korpora entwickelt werden. In AP1 sollen die SFB-spezifischen Korpus-Annotationsschemata in einem übergreifenden Schema integriert und die annotierten Korpora in dieses Schema transformiert werden, sodass sie neben dem jeweiligen SFB-eigenen XML-Format auch in einem gemeinsamen Format vorliegen. Ein einheitliches Format erleichtert die Interoperabilität und Handhabbarkeit der Daten (gemeinsame Verwaltung und Durchsuchung) erheblich. Als zugrunde liegendes Datenmodell für dieses umfassende Schema bietet sich das NITE Object Model an, das die den spezifischen Schemata zugrunde liegenden (hierarchischen bzw. Zeitachsen-basierten) Datenmodelle subsumiert.

Dieses Arbeitspaket ist eng mit den Paketen AP2 (Abfragewerkzeuge), AP3 (Benutzeroberflächen) und AP4 (Metadaten) verbunden. Die in AP2 und AP3 zu implementierenden Werkzeuge für den Datenzugriff greifen auf die in AP1 erarbeiteten Datenformate zu. Werkzeuge und Formate sind also so aufeinander abzustimmen, dass, unter Berücksichtigung der unter AP2 formulierten Desiderate, ein möglichst effizienter Zugriff auf die Daten gewährleistet ist. Das in AP4 zu entwickelnde gemeinsame Metadaten-Schema soll u.a. eine Spezifikation der verschiedenen Annotationsebenen umfassen, die in den Korpora annotiert sind (z.B. Text-Ebene, Syntax-Ebene, Token-Ebene etc.). Das NITE Object Model sieht vor, diese Ebenen (inklusive der Kodierungs-Elemente, die sie umfassen) sowie ihre strukturellen Beziehungen zueinander explizit zu deklarieren. Daraus ergibt sich die AP1 wie auch AP4 betreffende Aufgabe, ein gemeinsames Inventar von Annotationsebenen festzulegen und für jede Ebene zu bestimmen, welche Phänomene sie umfasst. So könnte man beispielsweise POS-Tags der Token- oder alternativ der Syntax-Ebene zuordnen oder hierfür eine eigene Morphosyntax-Ebene einführen. Derartige Entscheidungen sind essenziell für beide Arbeitspakete. Aus diesen Überlegungen ergibt sich, dass es zweckmäßig ist, wenn, wie vorgesehen, die Koordination von AP1–AP4 in einer Hand liegt.

Das in diesem Projekt zu entwickelnde gemeinsame Annotationsschema hat mit dem im Linguistic Annotation Framework (LAF) spezifizierten *dump format* gemeinsam, dass es ein generisches Schema darstellen wird, in das verschiedene Korpora mit unterschiedlichen Kodierungsschemata überführt werden können. Aus diesem Grund bietet es sich an, die in diesem *dump format* angebotenen Lösungen für die Kodierung einzelner Phänomene (z.B. Syntax-Bäume) als Ausgangspunkt für das hier zu erarbeitende Schema heranzuziehen. Begründete Änderungen gegenüber dem LAF,

z.B. zur Erhöhung der Zugriffs-Effizienz der in AP2 implementierten Tools, sollen jedoch möglich sein. In jedem Fall ist (mit Hilfe von XSLT-Stylesheets) eine Abbildung des erarbeiteten Schemas nach LAF zu spezifizieren.

Insgesamt wird zum einen durch die Diversität der von dem gemeinsamen Annotationsschema zu erfassenden Daten-Kollektionen der beteiligten SFBs und zum anderen durch die Orientierung an generischen Datenmodellen und Kodierungsschemata wie NOM und LAF gewährleistet, dass dieses Schema keine idiosynkratische Lösung für die drei SFBs darstellt, sondern generell offen ist für die Subsumierung anderer Korpora.

Wegen der oben genannten Abhängigkeit der Pakete AP2–AP4 von den Ergebnissen von AP1 soll das gemeinsame Annotationsschema im Wesentlichen bereits Ende 2006 vorliegen. Im zweiten und dritten Jahr soll geprüft werden, inwieweit die Ergebnisse der anderen Arbeitspakete — etwa Effizienz-Erwägungen — entsprechende Anpassungen erfordern.

AP2 Entwicklung generischer Werkzeuge für die Distribution und den Datenzugriff (Koordination: SFB 441)

Dieses Arbeitspaket befasst sich mit technischen Aufgaben, die zum Zweck der nachhaltigen Verfügbarkeit der Daten erforderlich sind. Konkret geht es um die Entwicklung von Werkzeugen, die einerseits die Ressourcen für ihre Distribution aufbereiten und andererseits die Durchsuchung (Abfrage) der Daten in adäquater Weise ermöglichen. Grundsätzlich sind hierbei verschiedene Distributionsformen sinnvoll:

- Eine Internet-Abfrageschnittstelle ermöglicht interessierten Forschern einen direkten Zugriff sowie eine schnelle Orientierung über die Beschaffenheit der Daten.
- Eine elektronische Offline-Distribution (z.B. auf DVD) erlaubt umfassende Bearbeitungsmöglichkeiten am lokalen Rechner.
- Eine (von Menschen lesbare) Druckversion ist die nachhaltigste Form der Archivierung, da sie völlig unabhängig von jedweder Hard- und Software ist.

Es ist Ziel des beantragten Projekts, diese drei Möglichkeiten technisch zu realisieren. Im AP2 sollen generische Werkzeuge entwickelt werden, die die Daten in den drei genannten Formen verfügbar machen. Für die elektronischen Distributionen (Internet und offline) sollen Tools zur linguistischen Suche in den Daten implementiert werden. Für die dritte Distributions-Form sind Tools zu realisieren, die eine Druckversion der Daten (soweit für diese angemessen) zur medienunabhängigen Archivierung generieren. Mit solchen Tools sollen dann druckbare Fassungen der Korpora erstellt werden,

die — vergleichbar digitalen Publikationen im PDF-Format — von Bibliotheken archiviert und angeboten werden können.

Wie in Abschnitt 3.3.2 erläutert, werden in den beteiligten SFBs bereits Abfrage-Mechanismen für die jeweiligen Daten-Kollektionen realisiert. Es ist jedoch zu betonen, dass die Entwicklung jener Mechanismen unter anderen Voraussetzungen erfolgt als die Tool-Entwicklung im beantragten Projekt. Für die in den SFBs realisierten bzw. geplanten Implementationen zur Daten-Abfrage gelten im Wesentlichen drei Desiderate: Mächtigkeit (die Möglichkeit, komplexe Suchkriterien zu spezifizieren), Effizienz (kurze Suchzeiten) und Komfortabilität (Eingabeschnittstellen und Ausgabeformate, die für Linguisten und Philologen ohne fortgeschrittene computertechnische Kenntnisse nachvollziehbar sind). Im Rahmen ihrer Möglichkeiten nutzen die SFBs jegliche zur Verfügung stehende Technologie (leistungsfähige Hardware, spezielle Datenbank-Software etc.), um diese Desiderate umzusetzen. Prinzipiell sind diese Kriterien, die unter dem Begriff der Benutzerfreundlichkeit zusammengefasst werden können, auch unter Nachhaltigkeitsaspekten von hoher Relevanz, da Benutzerfreundlichkeit ein wichtiges Ziel sein muss, um „Datenfriedhöfe“ zu vermeiden. Jedoch ist, im Unterschied zu den bereits laufenden Aktivitäten, in diesem Projekt das Kriterium der Langlebigkeit von entscheidender Bedeutung. Daraus folgt, dass die in AP2 erstellten Tools mit Hilfe allgemeiner Standards zu realisieren sind, um ihre Unabhängigkeit von spezieller Hardware und proprietären Software-Systemen zu gewährleisten. Für die XML-kodierten Korpora ist es daher sinnvoll, diese Tools mit XSLT-Stylesheets bzw. XQuery-Spezifikationen zu realisieren, da eine Unterstützung dieser Standards durch geeignete Software auch langfristig zu erwarten sein wird (vgl. Abschnitt 3.3.1).

In diesem Rahmen sollen die oben genannten Desiderate soweit wie möglich berücksichtigt werden. Das Mächtigkeits-Kriterium erfordert, dass die Tools Abfragen von der Ausdrucksstärke der NITE Query Language (vgl. Abschnitt 3.3.1) erlauben. XQuery und XSLT verfügen prinzipiell über diese Ausdrucksstärke; in AP2 sind kombinierbare Templates zur Implementierung der einzelnen Sprachkonstrukte zu entwerfen.[29] Die Effizienz kann durch sorgfältige Programmierung (z.B. dem Einsatz des *key*-Mechanismus in XSLT) erhöht werden. Wie bereits erwähnt, ist nicht zuletzt das Annotationsformat, in dem die Daten vorliegen, für die Abfrage-Effizienz bedeutsam. Dies soll bei der Erarbeitung des gemeinsamen Annotationsschemas in AP1 berücksichtigt werden. Außerdem sollen in AP2 Such-Tools sowohl für das gemeinsame Annotationsschema als auch für die spezifischen Schemata von TUSNELDA, EXMARaLDA und ANNIS erstellt werden. Es ist zu erwarten, dass letztere Schemata effizienter durchsucht werden können, da sie weniger komplex sind und jeweils homogenere Datenmodellierungen repräsentieren als das gemeinsame Schema. Um dem Komfortabilitäts-Kriterium zu genügen, sind intuitive Benutzeroberflächen zu imple-

[29] Beispielsweise hat (Smith 2005) die Realisierung des Allquantors mit XSLT illustriert.

mentieren. Diese wichtige Aufgabe soll in AP3 gesondert angegangen werden, da hierfür auf andere Technologien zurückgegriffen werden muss (s.d.).

Da zu Beginn des Projekts das gemeinsame Annotationsschema noch nicht vorliegt, soll sich die Entwicklung nachhaltiger Werkzeuge bis Ende 2006 auf die bestehenden Annotationsschemata von TUSNELDA, EXMARaLDA und ANNIS konzentrieren. Dies soll primär an den jeweiligen Standorten erfolgen, jedoch in enger Abstimmung, damit gemeinsam verwendbare Module erarbeitet werden können, was Synergieeffekte bringt und die langfristige Wartbarkeit der Tools erhöht. Ab 2007 wird die Realisierung der Werkzeuge für das gemeinsame Schema hinzukommen. Diese soll im Wesentlichen in Tübingen erfolgen, wobei die anderen Standorte beratend mitwirken.

AP3 Entwicklung komfortabler Benutzeroberflächen (Koordination: SFB 441)

Eine wesentliche Voraussetzung für den nachhaltigen Nutzen von Daten-Kollektionen für die linguistische Forschung ist die *effektive Verwendbarkeit* der zugehörigen Abfragewerkzeuge durch die intendierte Zielgruppe. Um diese effektive Verwendbarkeit zu ermöglichen, sind intuitive, auf Linguisten und Philologen zugeschnittene Benutzeroberflächen unabdingbar. In AP3 sollen generische Module für gemeinsame Web- und Offline-Benutzeroberflächen entwickelt werden, die als Front-End zu den in AP2 entwickelten Zugriffs-Tools dienen sollen. Mit Hilfe dieser Oberflächen sollen Abfragen auf einfache Weise (z.B. menügesteuert oder durch graphische Elemente) spezifiziert werden können. Ferner sollen „gängige“ Beispiel-Abfragen bzw. Templates direkt auswählbar sein; zugleich soll fortgeschrittenen Benutzern die Formulierung beliebiger Query-Formeln möglich sein. Suchergebnisse sollen in intuitiven, den jeweiligen Daten angemessenen Formaten (z.B. Bäume für syntaktische Strukturen oder Partitur-Notation für zeitliche Alignierungen) ausgegeben werden, wobei ggf. verschiedene Formate (z.B. Baumstruktur vs. indizierte Klammerstruktur, Zeitachsen- vs. Dialog-Turn-orientierte Notation) auswählbar sein sollen. Grundsätzlich soll die zu realisierende Internet-Schnittstelle, deren primärer Zweck es ist, einen Einblick in die Daten zu geben, lediglich elementare Suchfunktionen (z.B. Suche nach Wörtern und/oder POS-Tags) bereitstellen. Dagegen soll eine Offline-Distribution mächtige Werkzeuge zur umfassenden Abfrage komplexer Suchkriterien enthalten. Diese Strategie wurde auch für andere Korpora umgesetzt, z.B. für das British National Corpus (Burnard 1995), für das ein einfaches Web-Interface mit eingeschränkten Suchmöglichkeiten[30] sowie eine CD-Distribution mit komplexer Abfrage-Software zur Verfügung steht.

Die Implementierung der Benutzeroberflächen soll in HTML und Java erfolgen. Unter den derzeit verfügbaren Software-Plattformen, die sich für die Realisierung komfortabler Schnittstellen (inklusive Menüs oder graphischer Elemente) eignen, gewährlei-

[30] http://sara.natcorp.ox.ac.uk/lookup.html

sten diese beiden Optionen am ehesten eine maximale Abwärts-Kompatibilität über einen längeren Zeitraum hinweg, d.h. eine maximale Langlebigkeit der Implementationen. Dennoch kann v.a. für Java eine hundertprozentige Abwärtskompatibilität der verwendeten Bibliotheken nicht garantiert werden. Dies bedeutet, dass durch die Institution, welche die Daten in Zukunft verwalten wird, ggf. gewisse Anpassungen an den Schnittstellen-Implementationen vorgenommen werden müssen, um ihre Funktionalität sicherzustellen. Bei der Implementation soll darauf geachtet werden, diese kommende Aufgabe soweit wie möglich zu erleichtern, u.a.. durch die Erstellung eines gut nachvollziehbaren und ausführlich dokumentierten Programmcodes und durch die Vermeidung „peripherer“ Klassen-Bibliotheken. Grundsätzlich ist jedoch anzumerken, dass die Realisierung der Benutzeroberflächen das Langlebigkeits-Desiderat aus technischen Gründen nicht in dem hohen Umfang erfüllen kann wie die in AP2 entwickelten „Kern-Tools“ für den Datenzugriff. Die hier verfolgte Beschränkung der Verwendung von Java auf die Bereiche, die nicht mit XSLT und XQuery realisierbar sind, minimiert den Aufwand möglicher späterer Anpassungen und ist somit im Sinne der Nachhaltigkeit geboten.

Die Arbeit in AP3 soll 2007 beginnen. Zu diesem Zeitpunkt liegt das gemeinsame Annotationsschema im Wesentlichen fest; ferner sind erste Erfahrungen mit der Realisierung der Zugriffs-Tools gemacht, mit denen die Nutzeroberflächen interagieren. Als Ausgangspunkt für die Oberflächen-Entwicklung sollen Design und Funktionalität der in den einzelnen SFBs entwickelten Abfrageschnittstellen (vgl. Abschnitt 3.3.2) herangezogen werden. Ferner sind die in der laufenden Phase geplanten Aktivitäten des Projekts C1 im SFB 441 zur Erarbeitung benutzerfreundlicher Korpus-Schnittstellen (vgl. Antrag C1) von großer Bedeutung, sodass eine intensive Kooperation zwischen C1 und C2 auf diesem Gebiet geplant ist. Die Implementierung der Benutzeroberflächen soll am SFB 441 erfolgen; die anderen SFBs sollen beratend mitwirken.

Wie bereits angemerkt, decken die in den beteiligten SFBs erstellten Daten-Kollektionen aufgrund ihrer Diversität ein breites Spektrum an linguistischen Annotationen und Datenstrukturen ab, was an die Entwicklung von integrierbaren Benutzeroberflächen-Modulen vielfältige Herausforderungen stellt. Dieses breite Spektrum bürgt für die Einhaltung des Ziels, dass das in C2 entwickelte Ensemble von Annotationsschema (AP1), Zugriffswerkzeugen (AP2) und Benutzeroberflächen (AP3) übergreifender, generischer Natur und damit prinzipiell auch auf andere linguistische Korpus-Initiativen anwendbar ist.

3.4.2. Konzeptionelle Integration der Daten

AP4 Metadaten und Textklassifikation (Koordination: SFB 441)

Damit der mögliche Nutzen einzelner Teil-Korpora bzw. Daten-Sammlungen für konkrete Forschungsvorhaben adäquat eingeschätzt werden kann, ist es erforderlich, dass

ihre Merkmale, z.B. Textsorte/Datentyp, Sprache oder Annotationsebenen (vgl. AP1), umfassend spezifiziert sind. Aus diesem Grund soll im Projekt ein umfassendes Metadaten- und Textklassifikations-Schema erarbeitet werden, das alle in den SFBs erstellten Daten angemessen beschreibt. Dies bedeutet, dass ein solches Schema die Metadaten, die in den beteiligten SFBs erstellt wurden, integriert und ggf. erweitert. Die einzelnen Daten-Kollektionen sind dann mit Hilfe dieses Schemas zu klassifizieren. Dieses Schema soll mit bestehenden linguistischen Metadaten-Schemata, IMDI und OLAC, kompatibel sein, wird jedoch in verschiedenen Aspekten spezifischer sein. So sieht beispielsweise das IMDI-Schema für Korpora geschriebener Sprache vor, die Größe eines Textes bzw. Korpus in Wörtern anzugeben. Dies ist ein gängiges und generell sinnvolles Maß. Jedoch ist für manche (z.B. agglutinierende) Sprachen die Anzahl von Buchstaben in einem Text/Korpus ebenso informativ oder vielleicht sogar informativer als die Anzahl der Wörter. Aus diesem Grund empfiehlt es sich, dieses Maß zusätzlich In ein Metadaten-Schema aufzunehmen, wie dies für TUSNELDA bereits der Fall ist (vgl. (Wagner & Kallmeyer 2001)). Die vergleichende Sichtung der Daten-Kollektionen in TUSNELDA, EXMARaLDA und ANNIS (in den ersten 15 Monaten) soll wichtige Aspekte für die Definition eines übergreifenden Metadatenschemas (im zweiten Jahr) aufzeigen. Dieses Schema soll (im dritten Jahr) in die Metadaten-Repositorien von IMDI und OLAC eingetragen werden, damit die Daten-Kollektionen der drei SFBs auch als gemeinsame Ressource auffindbar sein werden.

AP5 Integration der Terminologien (Koordination: SFB 632)

Um die gemeinsame Verwendbarkeit der Daten in den beteiligten Projekten auch auf konzeptueller Ebene zu sichern, soll ein Ontologie-artiges Repositorium linguistischer Terme erstellt werden, auf das die in den Annotationsschemata verwendeten Begriffe dann abgebildet werden.

In der ersten Projektphase sollen dazu linguistische Ontologie-Initiativen (wie GOLD, DOLCE etc.) auf ihre Eignung für die Integration der SFB-Terminologien untersucht werden. Dabei spielt, neben den SFB-spezifischen Anforderungen, der Aspekt der Nachhaltigkeit der verwendeten Verfahren und Formate eine besondere Rolle. Die dann ausgewählte Ontologie soll dann in einem ersten Testlauf für die Wortarten-Tagsets, die innerhalb der SFBs Anwendung finden, angewendet werden. Dazu werden zum einen Erweiterungen und/oder Modifikationen der Ontologie notwendig sein, zum anderen muss die Abbildung der Wortarten-Tags in diese Ontologie definiert werden. Um eine nachhaltige Modellierung der Beziehungen zwischen den Tagsets und der Ontologie zu garantieren, ist, ähnlich zu Simons „Semantic Interpretation Language“ (Simons 2003), der Einsatz von verbreiteten XML-Standards wie XSLT, OWL und RDF vorgesehen.

In der nächsten Projektphase folgen sukzessive die Ontologien und Abbildungen für weitere Annotationsebenen, die sich z.T. durch eine höhere Komplexität auszeichnen

(Syntax, Textstruktur etc.). Hierbei soll ein ständiger, enger Austausch mit den Standardisierungsinitiativen für linguistische Ontologien gepflegt werden. Anzustreben ist auch eine aktive Beteiligung an der Entwicklung von GOLD und ähnlichen Initiativen, was für eine Propagierung der Ergebnisse des Projekts und für die Verbreitung und Wiederverwendbarkeit der SFB-Daten natürlich sehr förderlich wäre.

Um den Aufwand der manuellen Ontologieerstellung zu verringern und gleichzeitig die Integration SFB-fremder Terminologien zu unterstützen, sollen schließlich Methoden der datengetriebenen Ontologieerstellung (vgl. das bereits erwähnte Projekt DDLOD) aufgenommen und weiterentwickelt werden. Ebenso gilt es, Methoden zur (semi-)automatischen Evaluation der erstellten Ontologie und den Abbildungen aus den einzelnen Tagsets zu entwickeln. Die im Arbeitsprozess entstehenden Ressourcen sollen dokumentiert und im Rahmen der Best-Practice-Richtlinien (AP 6) fortlaufend veröffentlicht werden.

3.4.3. Rules of Best Practice

AP6 Allgemeine Best-Practice-Richtlinien (Koordination: SFB 441, SFB 538, SFB 632)

Obgleich die hier skizzierten Arbeitspakete konkret auf die in den drei SFBs erstellten Daten ausgerichtet sind, ist es ein grundsätzliches Anliegen des Projekts, keine „Insel-Lösung“ zu generieren. Vielmehr sollen die erarbeiteten Lösungen exemplarischen Charakter für die nachhaltige Verfügbarkeit linguistischer Daten-Kollektionen im Allgemeinen haben. Insbesondere sollen die entwickelten Schemata, Werkzeuge, Ontologien und Richtlinien offen sein für die Integration anderer linguistischer Daten-Ressourcen. Diese allgemeine, offene Ausrichtung wird durch zwei Prinzipien untermauert, die in allen beschriebenen Arbeitspaketen konkretisiert sind: zum einen die Aufgabe der Integration der heterogenen Daten-Kollektionen der drei SFBs, die in weiten Teilen komplementäre Schwerpunkte aufweisen und zusammen ein exemplarisches Spektrum linguistischer Daten bilden, und zum anderen die Ausrichtung an allgemeinen Standards.

Die Übertragbarkeit der Projektergebnisse auf andere Ressourcen erfordert nicht zuletzt eine ausführliche Dokumentation dieser Konzepte. Darüber hinaus sollen allgemeine „Rules of Best Practice“ für die Nachhaltigkeit linguistischer Daten formuliert werden. (Bird & Simons 2003) verstehen die in ihrem Beitrag formulierten Best-Practice-Regeln ausdrücklich als eine Diskussionsgrundlage und fordern zu einem breiten Diskurs über solche Regeln auf. Ihre Überlegungen wurden insbesondere (wenn auch nicht ausschließlich) im Hinblick auf Daten und Dokumentationen über bedrohte Sprachen formuliert. Demgegenüber kann das beantragte Projekt seine Erfahrungen mit reichhaltig annotierten Korpora in den Diskurs einbringen.

Neben der Daten-Aufbereitung und -Integration sollen die in diesem Projekt erarbei-

teten Regeln auch Vorschläge für einen infrastrukturellen Rahmen der Archivierung und Distribution der Daten umfassen. Idealerweise könnte in einem solchen Rahmen ein offenes Repositorium für linguistische Daten etabliert werden, inklusive eines Internet-Portals für die Suche nach bzw. in diesen Daten. Eine solche Infrastruktur soll nach Möglichkeit in enger Abstimmung mit bestehenden Initiativen zur Archivierung wissenschaftlicher Daten der DFG (z.B. die Förderinitiative „Leistungszentren für Forschungsinformation“) und des BMBF (z.B. nestor oder KOPAL) angeregt werden. Das Projekt wird im Bewilligungsfall entsprechende Kontakte aufnehmen.

Mit der Erarbeitung von Best-Practice-Richtlinien sollen umfassende Publikations-Aktivitäten auf nationalen und internationalen Konferenzen einhergehen. Dadurch können derartige Richtlinien wie auch die Nachhaltigkeits-Initiative als Ganzes einem einschlägigen Fachpublikum bekannt gemacht und zur Diskussion gestellt werden, was nicht nur dem Projekt fruchtbare Impulse geben, sondern auch entscheidend zur allgemeinen Akzeptanz der zu etablierenden Datenarchivierungs-Infrastruktur beitragen kann. Neben Konferenzen bieten einschlägige Mailing-Listen, z.B. von OLAC oder nestor, eine gute Diskussions-Plattform. Nicht zuletzt soll der interne Diskurs mit den Projekten der SFBs gesucht werden, die in ihrer Vielfalt nützliche Hinweise geben können.

Dieses Arbeitspaket soll im zeitlichen Wechsel von jeweils einem der beteiligten SFBs koordiniert werden. Bis Ende 2006 soll der SFB 441, 2007 der SFB 538 und 2008 der SFB 632 diese Aufgabe übernehmen. Auf diese Weise ist gewährleistet, dass die spezifischen Erfahrungen, Kompetenzen und Kontakte aller drei SFBs in die Koordinations-Aktivitäten eingebracht werden und damit die zu formulierenden Richtlinien, die eine zentrale Stellung in der durch dieses Projekt verfolgten Nachhaltigkeits-Initiative einnehmen, möglichst vielfältige Aspekte linguistischer Daten-Kollektionen einbeziehen.

Die in AP6 zu erarbeitenden Regeln ergeben sich wesentlich aus den Resultaten der anderen Arbeitspakete. Aufbauend auf bestehenden Vorschlägen für Best-Practice-Richtlinien (z.B. (Bird & Simons 2003)) soll ein Katalog solcher Regeln inkrementell auf der Basis dieser Resultate spezifiziert werden. Dieser Katalog soll jeweils im Zuge der viermal jährlich stattfindenden Projekt-Treffen erörtert und weiterentwickelt werden. Der jeweils koordinierende SFB wird die entsprechenden Ergebnisse zusammenfassen, eine aktualisierte Version des Regel-Katalogs erstellen und in Zusammenarbeit mit den anderen SFBs die interne und externe Diskussion hierüber führen.

AP7 Richtlinien für Zugriffsbeschränkungen (Koordination: SFB 538)

Zugriffsbeschränkungen auf linguistische Daten können aus drei Gründen notwendig sein:

1. Bei Audio- oder Videoaufnahmen im privaten Bereich (und deren Transkriptionen) sowie bei privaten schriftlichen Dokumenten (z.B. Briefe) müssen die Persönlichkeitsrechte der aufgenommenen Personen (Recht am eigenen Bild etc.) bzw. der Autoren gewahrt bleiben. Dabei bedürfen besonders sensible Daten (z.B. Aufnahmen von Arzt-Patienten-Gesprächen) einer erhöhten Aufmerksamkeit.

2. Bei Audio- oder Videoaufnahmen aus dem öffentlichen Bereich (Rundfunk, Fernsehen) sowie bei geschriebenen Texten, die unter das Urheberrecht fallen (z.B. Zeitungsberichte), müssen Copyright- und Urheberrechtsfragen berücksichtigt werden.

3. Bei allen Datentypen muss — im Interesse eines fairen wissenschaftlichen Wettbewerbs — sichergestellt werden, dass die Herkunft und die Urheberschaft von Korpora bei deren Nutzung angemessen (z.B. in Form einer Zitierung) vermerkt wird.

Geeignete Zugriffsbeschränkungen können in allen drei Punkten helfen, einen Missbrauch oder einen Verstoß gegen bei der Datenerhebung getroffene Vereinbarungen zu vermeiden, indem sie erstens die Nutzungsbedingungen für einzelne Korpora in juristisch fundierter Weise festlegen und zweitens die Einhaltung solcher Bedingungen über geeignete Kontrollmechanismen sicherstellen.

Dieses Arbeitspaket soll vom SFB 538 in Hamburg aus koordiniert werden, da dort aufgrund der vorhandenen Daten-Kollektionen vor allem Fragen des Schutzes von Persönlichkeitsrechten von herausragender Bedeutung sind. Die Arbeit in diesem Paket gliedert sich in drei Phasen: In der ersten Phase (ca. ein 3/4 Jahr) soll aus den in Abschnitt 3.3.1 genannten Vorarbeiten (insbesondere (MacWhinney, Bird, Cieri & Martell 2004), (Lautenschlager & Sens 2002) und (Baude et al. 2005)) eine Grundlage für eine Diskussion der rechtlichen und ethischen Probleme, die sich bei der Arbeit mit linguistischen Korpora stellen, geschaffen werden. Dabei soll bereits in dieser Phase auch Expertise aus Einrichtungen, die bereits Datenarchive betreiben, herangezogen werden. Das Ergebnis dieser Phase wird in einem schriftlichen Bericht, der der interessierten Forschergemeinde zur Verfügung gestellt wird, festgehalten. Inhalt der zweiten Phase (ca. ein Jahr) wird dann sein, die Diskussion über die sich stellenden Probleme zu führen. Dies soll in drei Kontexten geschehen: erstens in regelmäßigen Treffen mit MitarbeiterInnen der drei beteiligten SFBs, zweitens mit interessierten linguistischen Communities (z.B. über die Mailingliste Gesprächsforschung oder die Corpora-List), drittens mit einem juristischen Experten auf dem betroffenen Gebiet. Ziel dieser Diskussion soll nicht nur sein, ein möglichst vollständiges Bild von Problemen und Lösungsansätzen in diesem Bereich zu erarbeiten, sondern gleichzeitig Vorbehalte unter Forschern, die auf einer Unsicherheit bzgl. rechtlicher Fragen beruhen,

so weit wie möglich abzubauen, um so für eine breite Akzeptanz der zu erstellenden Best-Practice-Richtlinien zu sorgen. In der dritten Phase (ca. 1,5 Jahre) sollen diese Best-Practice-Richtlinien dann zunächst auf der Grundlage der in Phase II geführten Diskussionen ausformuliert werden. Eine unmittelbare Überprüfung und ggf. Adaption erfahren die Richtlinien, indem die Einzelprojekte der beteiligten SFBs in ihnen geschult und bei ihrer Anwendung auf die einzelnen Korpora beraten werden. Gleichzeitig soll eine entsprechende Veröffentlichung auch die Diskussion der Richtlinien in einer breiteren wissenschaftlichen Öffentlichkeit ermöglichen.

Schließlich ist während der gesamten Laufzeit des Projektes zu prüfen, wie sich die in diesem Paket erarbeiteten Lösungen auf die übrigen Arbeitspakete auswirken. Insbesondere wird die technische Realisierung von Zugriffsbeschränkungen bei der Entwicklung von Werkzeugen für die Distribution (AP2) sowie Abfrage-Oberflächen (AP3) eine Rolle spielen, und auch das zu erarbeitende Metadaten-Schema (AP4) wird die Ergebnisse dieses Arbeitspakets berücksichtigen müssen.

3.4.4. Zeitplan

01.10.2005 – 31.12.2006

- AP1: Erarbeitung eines übergreifenden Annotationsschemas
- AP2: Implementierung nachhaltiger Zugriffswerkzeuge für die SFB-spezifischen Annotationsschemata
- AP3: —
- AP4: systematische Zusammenstellung der Eigenschaften der in den SFBs vorhandenen Daten
- AP5: Evaluation von Initiativen zur Erstellung linguistischer Ontologien; Definition einer Wortarten-Ontologie; Abbildung der Wortarten-Annotationsschemata
- AP6: Formulierung von Best-Practice-Richtlinien sowie deren interne und externe Diskussion
- AP7: Sichtung und Zusammenfassung vorliegender Arbeiten zu Zugriffsbeschränkungen; interne und externe Diskussion der verschiedenen Aspekte

2007

- AP1: Optimierung des übergreifenden Annotationsschemas
- AP2: Implementierung nachhaltiger Zugriffswerkzeuge für die SFB-spezifischen Annotationsschemata sowie das übergreifende Annotationsschema

AP3: Implementierung komfortabler Benutzeroberflächen

AP4: Definition eines übergreifenden Metadaten- und Textklassifikationsschemas; Erfassung der Daten in diesem Schema

AP5: Evaluation der Wortarten-Ontologie und -Abbildung; Integration weiterer Annotationsebenen; Untersuchung datengetriebener Methoden zur Ontologieerstellung

AP6: Formulierung von Best-Practice-Richtlinien sowie deren interne und externe Diskussion

AP7: interne und externe Diskussion der verschiedenen Aspekte von Zugriffsbeschränkungen; Formulierung von Best-Practice-Regeln und Diskussion derselben

2008

AP1: Optimierung des übergreifenden Annotationsschemas; abschließende Dokumentation

AP2: Implementierung nachhaltiger Zugriffswerkzeuge für das übergreifende Annotationsschema; abschließende Dokumentation

AP3: Implementierung komfortabler Benutzeroberflächen; abschließende Dokumentation

AP4: Einfügen der Metadaten über die SFB-Ressourcen in die Repositorien von IMDI und OLAC; abschließende Dokumentation

AP5: Integration weiterer Annotationsebenen; Nutzung und Weiterentwicklung datengetriebener Methoden zur Ontologieerstellung; abschließende Dokumentation

AP6: (vorläufig) abschließende Formulierung von Best-Practice-Richtlinien sowie deren interne und externe Diskussion

AP7: interne und externe Diskussion sowie (vorläufig) abschließende Formulierung von Best-Practice-Regeln zu Zugriffsbeschränkungen

3.5. Stellung innerhalb des Sonderforschungsbereichs

Das Projekt C2 nimmt im SFB 441 in zweifacher Hinsicht eine Sonderstellung ein. Zum einen ist es ein Kooperationsprojekt zwischen diesem SFB und den SFBs 538 und 632. Zum anderen befasst es sich mit der langfristigen Verfügbarkeit der in diesen SFBs aufgebauten Daten, ist also in seinen Zielen über das Ende der SFBs hinaus gerichtet. Bei der Diskussion von Best-Practice-Richtlinien wird C2 alle Projekte der drei SFBs einbeziehen und deren Expertise nutzen. Insbesondere wird das Projekt jedoch eng mit den Projekten kooperieren, die in den jeweiligen SFBs die Erstellung

der Ressourcen unterstützen und die Daten in ein SFB-weites Repositorium zusammenführen. Dies sind, wie in Abschnitt 3.3.2 erwähnt, das Projekt C1 im SFB 441, das Projekt Z2 im SFB 538 sowie die D-Projekte im SFB 632. Die Kooperation mit diesen Projekten wird sich auf alle Arbeitspakete erstrecken. Diese Projekte befassen sich ebenfalls mit der Konstruktion von Annotationsschemata, der Implementierung von Zugriffswerkzeugen, der Definition von Metadaten-Schemata sowie — im Falle des Projekts D2 im SFB 632 — der Erarbeitung eines gemeinsamen Terminologie-Inventars. Jedoch unterscheidet sich die in diesen Projekten verfolgte Zielrichtung signifikant von der Zielrichtung des Projekts C2. Die zentralen Datenprojekte in den einzelnen SFBs entwickeln vereinheitlichte Schemata, welche jeweils die in diesen SFBs erstellten Daten integrieren. Dagegen zielen die in C2 kooperativ zu erarbeitenden Schemata auf eine Integration der drei SFB-spezifischen Schemata ab, und damit auf die Integration aller in den SFBs erstellten Daten. Darüber hinaus sollen diese Schemata offen sein für andere linguistische Daten-Kollektionen. Bei der Entwicklung von Abfragewerkzeugen wird C2 aufgrund des Nachhaltigkeitsgebots, das für dieses Projekt bestimmend ist, andere Strategien verfolgen als die zentralen SFB-Datenprojekte. Dort werden alle verfügbaren spezifischen Hard- und Software-Möglichkeiten genutzt, um Abfrage-Mechanismen mit maximaler Effizienz zu realisieren, wobei diese Lösungen voraussichtlich nur eine kurz- bzw. mittelfristige Lebensdauer haben werden. Dagegen werden in C2 mit XSLT bzw. XQuery unabhängige und langlebige Standards eingesetzt, was eine nachhaltige Lebensdauer erwarten lässt, jedoch Einbußen bei der Effizienz mit sich bringt. Dieses zweigleisige Vorgehen kombiniert die Vorteile beider Strategien. Insgesamt steht C2 in einem komplementären Verhältnis zu den existierenden Datenprojekten, das einerseits vielfältige Kooperationen erlaubt (und erfordert), und andererseits das eigenständige Profil von C2 klar erkennbar macht.

Die für den SFB 538 beantragten Stellen sollen organisatorisch an das dortige Projekt Z2 angegliedert werden. Die für den SFB 632 beantragten Stellen sollen organisatorisch an das dortige Projekt D1 angegliedert werden.

3.6. Abgrenzung gegenüber anderen geförderten Projekten des/der Teilprojektleiter/Teilprojektleiterinnen

Die ProjektleiterInnen von C2 leiten gleichzeitig das Projekt C1. Die Gemeinsamkeiten und Unterschiede dieser Projekte sind in Abschnitt 3.5 erläutert. Weitere Angaben zu geförderten Projekten der Teilprojektleiterin und des Teilprojektleiters sind dem Finanzierungsantrag des SFB 441 für die laufende Förderperiode zu entnehmen, insbesondere den entsprechenden Abschnitten der Teilprojekte B15 (Reis) und A1 (Hinrichs).

Zitierte Literatur

Baude, O., C. Blanche-Benveniste, M.-F. Calas, P. Cordereix, I. De Lambertierie, L. Goury, M. Jacobson, C. Marchello-Nizia & L. Mondada (2005): *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*, Centre National de la Recherche Scientifique, Paris.

Baumgarten, N., A. Herkenrath, T. Schmidt, K. Wörner & L. Zeevaert (2005): „Studying Connectivity with the Help of Computer-Readable Corpora: Some Exemplary Analyses from Modern and Historical, Written and Spoken Corpora", *in* C. Hohenstein, L. Pietsch & J. Rehbein (Hrsg.), *Connectivity in Grammar and Discourse*, Hamburg Studies in Multilingualism, Benjamins, Amsterdam.

Bickels, B. & J. Nichols (2002): „Autotypologizing databases and their use in fieldwork", in *Proceedings of the LREC Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain.

Bird, S. & G. Simons (2003): „Seven dimensions of portability for language documentation and description", *Language* 79, 557–582.

Burnard, L. (1995): *Users Reference Guide for the British National Corpus. Version 1.0*, Oxford University Computing Services.

Carletta, J., J. Kilgour, T. O'Donnell, S. Evert & H. Voormann (2003): „The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets", in *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003)*, Budapest.

Corti, L., A. Witzel & L. Bishop (Hrsg.) (2005): *Sekundäranalyse qualitativer Daten*, Forum Qualitative Sozialforschung (6,1).

Corti, L., S. Kluge, K. Mruck & D. Opitz (Hrsg.) (2000): *Text, Archiv, Re-Analyse*, Forum Qualitative Sozialforschung (1,3).

Farrar, S. & D. T. Langendoen (2003): „A linguistic ontology for the Semantic Web", *GLOT International* 7(3), 97–100.

Heid, U., H. Voormann, J.-T. Milde, U. Gut, K. Erk & S. Pado (2004): „Querying both time-aligned and hierarchical corpora with NXT Search", in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisbon, S. 1455–1458.

Ide, N. & L. Romary (2004): „A Registry of Standard Data Categories for Linguistic Annotation", in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisboa, S. 135–138.

Ide, N., L. Romary & E. de la Clergerie (2003): „International Standard for a Linguistic Annotation Framework", in *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*, Edmonton.

Lautenschlager, M. & I. Sens (2002): „Konzept zur Zitierfähigkeit wissenschaftlicher

Primärdaten“, *Information — Wissenschaft und Praxis*.

MacWhinney, B., S. Bird, C. Cieri & C. Martell (2004): „TalkBank — Building an Open Unified Multimodal Database of Communicative Interaction“, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, S. 525–528.

Monachesi, P., A. Dimitriadis, R. Goedemans, A.-M. Mineur & M. Pinto (2001): „The typological database project“, in *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia.

NIH (2003): *Final NIH Statement on Sharing Research Data*.
URL: *http://grants.nih.gov/grants/policy/data_sharing/index.htm*

Plaehn, O. (1998): *Annotate Bedienungsanleitung*, Universität des Saarlandes, Sonderforschungsbereich 378, Projekt C3, Saarbrücken, Germany.

Sasaki, F., A. Witt, D. Gibbon & T. Trippel (2004): „Concept-based Queries: Combining and Reusing Linguistic Corpus Formats and Query Languages“, in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisboa, S. 259–262.

Schütte, W. (2004): *Transkriptionsrichlinien für die Eingabe in EXMARaLDA nach DIDA-Konventionen*, Institut für Deutsche Sprache, Mannheim.
URL: *http://www.ids-mannheim.de/ksgd/kt/*

Simons, G. (2003): *SIL: A Metaschema Language for the Semantic Interpretation of XML Markup in Documents*. Draft.

Simons, G., W. Lewis, S. Farrar, T. Langendoen, B. Fitzsimons & H. Gonzalez (2004): „The Semantics of Markup: Mapping Legacy Markup Schemas to a Common Semantics“, *in* N. I. Graham Wilcock & L. Romary (Hrsg.), *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004)*, Association for Computational Linguistics, Barcelona, Spain, S. 25–32.

Simov, K., Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov & A. Kiryakov (2001): „CLaRK - an XML-based System for Corpora Development“, in *Proceedings of the Corpus Linguistics 2001 Conference*, S. 558–560.

Smith, G. (2005): „Refining Queries on a Treebank with XSLT Filters. Approaching the Universal Quantifier“, *in* S. Dipper, M. Götze & M. Stede (Hrsg.), *Heterogenity in Focus: Creating and Using Linguistic Databases*, Band 2 von *ISIS (Interdisciplinary Studies on Information Structure), Working Papers of the SFB 632*, Potsdam, S. 117–128.

Voormann, H., S. Evert, J. Kilgour & J. Carletta (2003): *NXT Search User's Manual (Draft)*.
URL: *http://www.ims.uni-stuttgart.de/projekte/nite/manual.pdf*