D-Spin/DTA-Workshop "Kumulatives Arbeiten mit Textarchiven"

26. – 27. März 2009

Berlin-Brandenburgische Akademie der Wissenschaften

Jägerstr. 22/23

10117 Berlin

Konferenzraum 1

Motivation

Die Projekte D-SPIN (Deutsche Sprachressourcen-Infrastruktur) und DTA (Deutsches Textarchiv) am Zentrum Sprache der BBAW führten gemeinsam einen Workshop zum Thema kumulatives Arbeiten mit Textarchiven durch.

Für das D-SPIN Projekt ging es darum zu erfahren, wie ein technischer, rechtlicher und organisatorischer Rahmen bereitgestellt werden kann, in welchem Textkorpora für die gemeinsame Arbeit mehrerer Personen oder Institutionen bereitgestellt werden können, wobei diese Personen oder Institutionen zu einer Wertschöpfung beitragen. Die Wertschöpfung kann in zusätzlichen Annotationen auf den Primärdaten, in der durch Forschungsfragen geleiteteten Erstellung von virtuellen Kollektionen auf Grundlage der exisitierenden Primärdaten und Annotationen oder in der Ableitung und Anreicherung sekundärer Projekte wie Konkordenzen und Wortlisten liegen. Für D-SPIN stand darüber hinaus die Nutzerperspektive im Vordergrund, da die Analyse von Benutzeranforderung ein zentrales Thema des von der BBAW geleiteten Arbeitspaktes 3 ist.

Für das DTA war es wichtig, in Kontakt mit Personen und Institutionen zu treten, die in ähnlicher Weise mit dem Aufbau oder der Archivierung von historischen Korpora befasst sind. Der Erfahrungsaustausch über verschiedene Aspekte der Ressourcenerstellung, von der Digitalisierung über die Erstellung von Metadaten bis hin zur Bereitstellung von Werkzeugen für die Verwendung der Daten ermöglicht es uns, eine gemeinsame wissenschaftliche Praxis auf diesem Gebiet zu etablieren.

Diese Interessen bestimmten die Auswahl der eingeladenen Referenten.

Programm

Donnerstag, 26. März 2009

14:00 – 14:15 A. *Geyken (Berlin):* Begrüßung

14:15 – 15:00 *L. Lemnitzer / J. Didakowski / A. Herold (Berlin):* Sprachressourcen vernetzen: Das Projekt D-Spin

15:00 – 15:45 *F. Jannidis / G. Lauer (Darmstadt, Göttingen):* Literatur rechnen. Für eine korpusbasierte Narratologie

15:45 – 16:00 Kaffeepause

16:00 – 16:45 *T. Roth (Basel):* TEI beim Schweizer Textkorpus

16:45 – 17:30 *E. Breiteneder (Wien):* AAC-CONTAINER DESIGN: Überlegungen zur Strukturierung von Textobjekten im AAC – Austrian Academy Corpus

17:30 – 17:45 Kaffeepause

17:45 – 18:30 *M. Kupietz (Mannheim):* Das Deutsche Referenzkorpus DeReKo? – Konzept, Textmodell, Annotationen und Repräsentationen

19:15 Gemeinsames Abendessen

Freitag, 27. März 2009

09:00 – 10:30 *O. Duntze / M. Drotschmann / C. Fritze / A. Geyken / B. Jurish / A. Siebert (Berlin):* Deutsches Textarchiv kodieren – denkbare Perspektiven kumulativen Arbeitens

10:30 – 10:45 Kaffeepause

10:45 – 11:30 *H. Lobin (Gießen):* Der Einsatz von Korpora beim Computer-Assisted Language Learning

11:30 – 12:15 *A. Rapp (Trier):* Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen (BMBF)

12:15 – 13:00 *M. Wynne (Oxford):* Oxford Text Archive: local, national and international collaborations

13:00 – 14:30 Mittagspause

14:30 – 15:15 *M. Durrell (Manchester):* Das GerManC? Projekt - Aufbau und Konzept eines repräsentativen Korpus der deutschen Sprache 1650–1800

15:15 – 16:00 *T. Gloning (Gießen):* Lexikologische Markierungen in einem digitalen lexikalischen System: Ältere Fach- und Gebrauchstexte

Protokoll

Donnerstag:

14:15 – 15:00 ***L. Lemnitzer / A. Herold (Berlin):*** **Sprachressourcen vernetzen: Das Projekt D-Spin**

Die Referenten stellten das Projekt vor, sowohl den allgemeinen und administrativen Rahmen, die Partner, Finanzierung etc., als auch den spezifischen Beitrag des Zentrums Sprache. Dieser liegt in der Aufbereitung der am Zentrum Sprache vorhandenen Ressourcen - Korpora und Lexika - incl. der Bereitstellung von standardkonformen Metadaten sowie in der Erhebung von Benutzeranforderungen hinsichtlich der Ausgestaltung des geisteswissenschaftlichen Arbeitsplatzes, welcher ein Ziel des Gesamtprojekts ist. Außerdem wird das Zentrum Sprache als ein Servicezentrum bestimmte Dienste nach außen hin anbieten.

Diksussion / Fragen: Diskutiert wurden insbesondere verschiedene Erfahrungen und Anregungen zur Erhebung der Benutzeranforderungen. Eine wichtige Anregung ist der Hinweis auf die Konstruktion von Szenarien in Zusammenarbeit mit einschlägigen Experten diverser geisteswissenschaftlicher Fächer, bzw. die Analyse/Einbeziehung bereits erstellter Szenarien benachbarter Projekte.

15:00 – 15:45 ***G. Lauer (Darmstadt, Göttingen):*** **Literatur rechnen. Für eine korpusbasierte Narratologie**

Die Referenten stellten das Konzept der Narratologie als eine literaturwissenschafltliche Methode vor, die in ihrer empirischen, korpusbasierten und quantitativen Ausrichtung im Feld der Geisteswissenschaften eher am Rande steht. Sie betonten, dass Literaturwissenschaftlern heute noch in der Regel das Rüstzeug für eine solche Arbeitsweise fehlt. Dies sei ein Mangel, dem in der zukünftigen Ausbildung abgeholfen werden müsse, auch wenn diese Art des Methodenverständnisses in der Literaturwissenschaft im Moment schwach ausgeprägt sei.

Ausgangspunkt ihrer Überlegung ist die Annahme, ähnlich wie bei korpuslinguistischen Fragestellungen auf der Basis sehr großer Textmengen zu arbeiten. Es geht den Autoren um Mengen von Texten, die ein menschlicher Leser (auf Grund beschränkter Lebenszeit) gar nicht durch *close reading* erfassen kann. Würde man tausende Romane vollständig elektronisch im Volltext zugreifbar haben, könnte man beispielsweise Fragestellungen wie der historischen Entwicklung der Wiedergabe direkter Rede nachgehen - bei entsprechender Abdeckung auch im internationalen Vergleich. Damit wird in Bezug auf die Literaturwissenschaft ein radikal anderes Vorgehen als traditionell üblich ist gefordert.

Die Autoren sagten zu Beginn, sie würde noch gar nicht produktiv arbeiten, außerdem fehle noch ein geeigneter Annotierungstandard, der im Zuge des Projekts entwickelt werden solle. Ebenso fehlen naturgemäß Annotationsbeispiele und algorithmische Annotationsverfahren zur Markierung von narratologischen Strukturen. Ebenso gibt es noch keine konkreten Vorstellungen darüber, was als kanonische Sammlung sinnvollerweise ein literaturwissenschaftlich orientiertes Korpus ausmachen sollte.

Diskussion / Fragen: ???, Lobin wies auf Probleme bei der erhofften automatischen Erkennung narratologischer Strukturen hin; Diskussion, ob Narratologie wirklich so sehr 'hard science' ist, wie Lauer/Jannidis es annehmen, bzw. wie stark der interpretatorische Anteil bei der Bestimmung narratologischer Strukturen ist

15:45 – 16:00 Kaffeepause

16:00 – 16:45 **T. Roth (Basel): TEI beim Schweizer Textkorpus**

Der Autor vertritt die Arbeitsgruppe, die z.Z. ein Korpus des geschriebenen Schweizerdeutschen aufbaut. Der Korpusaufbau orientiert sich an den Arbeiten der BBAW beim Aufbau des Kerncorpus. Das Schweizer Textkorpus ist allerdings mit ca. 20 Mill. Textwörtern deutlich kleiner. Die Daten werden gemäß dem TEI P4 Standard annotiert. Im Hauptteil seines Vortrags ging der Referent auf Probleme ein, die eine Anpassung eines TEI P4 Schemas an bestimmte Idiosynkrasien des konkreten Korpus bereiten kann. Er wählte hierzu den Texttyp des Formulars mit verschiedenen Zonen und einem schematischen Aufbau mit Leerstellen, die später hand- oder maschinenschriftlich oder gar mit Stempel ausgefüllt werden. Der Referent berichtete, dass es viele derartige Idiosynkrasien gibt und es einer ständigen sorgfältigen Abwägung bedarf zwischen dem Prinzip der texttreuen Nachbildung und dem damit verbunden Aufwand, der womöglich für sehr wenige Spezialfälle getrieben werden müsste. Die Entscheidung für oder wider texttreue Wiedergabe muss deshalb eine pragmatische sein. Der Schweizer Ansatz geht von einem strengen Schema aus, das jeweils im Fall von bislang noch nicht aufgetretenen Besonderheiten erweitert worden ist.

Fragen / Diskussion: Diskussion über bestimmte Kodierungsentscheidungen (warum add-Element für hss. Eintragungen in Formularen); Nachfragen zu Kodierungsrichtlinien (wird handschriftliches immer mit transkribiert? Wird zwischen hss. und gedrucken Teilen unterschieden) Wie wird nach Schema- und Richtlinienanpassungen die Konsistenz der bisher erfolgten Annotationen sichergestellt? Es folgt eine kurze Diskussion über die Grenze zwischen exakter Abbildung/Annotation und aufwendig zu kodierenden Spezialfällen, die unter Umständen singulär auftreten (hier bestimmen beschränkte Ressourcen den trade-off).

16:45 – 17:30 **_E. Breiteneder (Wien):_ AAC-CONTAINER DESIGN: Überlegungen zur Strukturierung von Textobjekten im AAC – Austrian Academy Corpus**

Die Referentin stellte die verschiedenen Teilprojekte des texttechnologischen Zentrums an der AAC vor, wobei hier vor allem das Fackel-Korpus und das Korpus der Weltbühne hervorgehoben werden sollen.

Angesichts der digitalen Erfassung von Zeitschriften wie der Weltbühne mit Tausenden von Autoren ist die Klärung der Urheberrechte eine besondere Herausforderung, der man an der Akademie mit einer pragmatischen Herangehensweise begegnet.

Die Container-Metapher bezieht sich dabei auf die Integration von Teilkorpora (Containern), die für sich genommen vollständige Werke enthalten (bspw. Fackel oder Sammlung der Weltbühne), separat durchsuchbar und abfragbar sind, aber ebenso im Verbund als Gesamtkorpus verwendbar bleiben. Darstellung von Problemen im Zusammenspiel der einzelnen Korpuskomponenten

Die Diskussion fokussierte auf rechtliche Aspekte der Veröffentlichung digitaler Editionen, sowie auf die organisatorische Einbettung der Arbeitsgruppe an der Akademie, kaum auf technologische Fragen.

17:45 – 18:30 **M. Kupietz (Mannheim): Das Deutsche Referenzkorpus DeReKo? – Konzept, Textmodell, Annotationen und Repräsentationen**

Der Referent, Mitarbeiter in der Arbeitsgruppe Korpustechnologie am Institut für deutsche Sprache und D-SPIN Konsortialpartner, stellte die Korpora - vor allem das Deutsche

Referenzkorpus - sowie die Korpustechnologie vor. Er ging ausführlich auf die rechtliche Situation ein, die die Interessen der Autoren und anderen Rechteinhaber berücksichtigen müsse und gleichzeitig den größtmöglichen Nutzungsspielraum für IDS-interne und externe Nutzer der Daten ermöglichen sollte. Das IDS als vermittelnde Stelle muss hinsichtlich der Wahrung von Autorenrechten besonders genau sein, um das Vertrauenverhältnis, das zu vielen Textlieferanten aufgebaut wurde, nicht zu gefährden.

Die von den jeweiligen Rechteinhabern auferlegten Beschränkungen stellen eine ernstzunehmende Hürde für Szenarien des kumulativen Arbeitens dar. Von diesen stellte der Referent zwei vor: a) Benutzer kreieren virtuelle Korpora auf der Basis der existierenden Daten; b) Benutzer extrahieren Daten wie Konkordanzen und Wortlisten und reichern diese mit zusätzlichen Informationen an. Während das erste Szenario in der Regel aus rechtlicher Sicht kein Problem darstellt, ergeben sich im zweiten Szenario Konflikte zwischen möglichen Beschränkungen der Veröffentlichung solcher Daten und der wissenschaftlichen Pflicht, genau dies zu tun.

Der Referent stellte dar, wie dieses Problem sowie Probleme der unabhängigen Referenzierbarkeit von virtuellen Kollektionen im Rahmen von D-SPIN mit technischen, administrativen und rechtlichen Mitteln gelöst werden sollen.

Fragen / Diskussion: Nachfragen zu den verwendeten Annotations-Tools

Freitag:

09:00 – 10:30 ***O. Duntze / C. Fritze :*** **Deutsches Textarchiv kodieren – denkbare Perspektiven kumulativen Arbeitens**

Referat: s. Folien

Diskussion:
* Vorschlag: Korpus mit der Krünitz Enzyklopädie verlinken
* TEI-Elemente als Stand-off verwalten
* klarer herausarbeiten: Zoning vs. Zonen in der tei-stand-off Annotation

10:45 – 11:30 **_H. Lobin (Gießen):_ Der Einsatz von Korpora beim Computer-Assisted Language Learning**

Der Referent ist ein Partner des D-SPIN Konsortiums und vor allem Partner des Arbeitspakets 3, das sich mit geisteswissenschaftlichen Anwendungen befasst. Sein Referat stützte sich auf drei Gutachten, die er, seine Kollgen und Mitarbeiter im Rahmen des D-SPIN AP 3 erstellt haben.

1. Korpora für die sprachdidaktische Verwendung
1.1 Beispiele (Intersect, Desiderat)
1.2 Lernerkorpora (FALKO, Lindsei, ICLE)
1.3 Fehler-Datenbanken (Augsburger Fehler-DB)
2. Korpora im Unterricht
* Data driven learning vs rule driven learning (Ziel des DDL: Motivation erhöhen)
* nicht idiomat. Form der Sprache aus dem Unterricht verbannen
* serendipity Methode
* genre Methode, d.h. Arbeiten mit Texten aus einer best. Textsorte, dann Produktion von

eigenen Texten
* korpusbasierte Wortschatzarbeit (Arbeiten von Belicza et al.)
* Erstellungsaspekte bei der Unterrichtsvorbereitung
- Lernerfortschritte / Grad der Idiomatisierung feststellen
- Korpuserstellung soll für "jedermann" möglich sein
3. Korpora und ICALL
NN-System, eGramD, System ALLES (Wirtschaftssprache, E, D, Katal.); Ergebnisse jeweils nicht überzeugend. System ALLES verweist aufs DWDS.

Diskussion:
* Lernerkorpora sinnvoll, für die Unterrichtsvorbereitung wäre es wichtig, wenn ad-hoc-Corpora erstellt werden könnten.
* Problem: bestehende Korpora (vom Typ 1.1) sind schlecht dokumentiert, Oberfläche Stand der 90er Jahre weiterführen.
* Aufgabe D-SPIN: Referenzkorpora in Form von Schnittstellen für didakt. Zwecke zur Verfügung stellen.
* Aufgabe der Großkorpora: Texte für didakt. Gesichtspunkte auswählen (Best Practice); Verweis auf Referat Kupietz: Virtuelle Corpora aus Referenzkorpora bilden - insbesondere Nachnutzung ermöglichen

11:30 – 12:15 **_A. Rapp (Trier):_ Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen (BMBF)**

Teilvorhaben:

TV 1: Basis-Lemmaliste (IDS)
Lemmaliste der nhd. Standardsprache (DeReWo? -Liste)
als Verweisnetz strukturiert (erlaubt semasiolog. Verweise)

TV 2: Varietäten Lemmaliste
Basis: Wörterbuchnetz
Bsp. Brombeere mit allen Varianten in den anderen Wb. (elsäss. Wb, DWB, lothring. WB etc.)
sollen auf Metalemmata verweisen

Probleme: wie werden ausgestorbene Wörter/Dialektwörter abgebildet?

TV 3. Biologie-Seite
TV 3.1 Grammatik der Varianz in Genomen
TV 3.2 Technik
Bezugsystem zu Metalemmata frei verfügbar in Genomdb; Vergleich zwischen Sprache und Biologie

Kooperationen:
www.regionalsprache.de (REDE)
Überführung in Textgrid

Anwendungen & Fazit:
Sinn der Varietäten-Lemmatisierung: flexibler auf Daten zugreifen (verbessert Rechercheergebnisse; über Varianzlisten)

Diskussion:

Gloning: die Darstellung von Sprache als Code, die zumindest implizit dem Ansatz zugrundliegt, ist veraltet und wird der heutigen Sichtweise, die Sprache im Gebrauch betrachtet, in keiner Weise gerecht.

Gloning: Varianz auf welcher Ebene: graphematisch ist zu wenig?!

12:15 – 13:00 **_M. Wynne (Oxford):_ Oxford Text Archive: local, national and international collaborations**

Der Referent ist Mitglied des CLARIN Konsortiums und Executive Boards. S. auch seine Folien zum Vortrag.

Historisches:
Erfahrungen aus internationaler Sicht OTA:
* 1976: Start des OTA Motivation Archiv, um Mehrfachdigitalisierungen zu vermeiden
* 1996: Arts and Humanities Research Council (JISC: Funder)

Collaborations (Institutions):
Other new text archives (e.g. Virginia) - Übergabe als Kopie
Oxford libraries
Emerging standards initiatives

Early Collaborations: Digitalisierung spielte eine große Rolle (Archiv hilft anderen Forschern, Bücher zu digitalisieren, zu annotieren)
Mit Verfügbarkeit des Internets:
* Distribution von Digitalisierung einfacher,
* Digitalisierung beschleunigt sich,
* Fokus auf Arbeit an Standards (xml, tei -- Lou Burnard, S. Rahtz),
* reuse of resources.

1996: AHDS (Arts and Humanities Data Service)
organized around 5 centres: OTA, History Data Service, Arch. Data Services, Visual Arts DS, Performing Arts DS + AHDS of King's College
changes in organization and textual format (common adisory, access, preservation, ingest procedures...)

problems of AHDS:
* Is there one or are there several Arts and Humanities communities? (Zentralisierung oder Dezentralisierung von access points)
* satisfactory metadata sets turned out to be difficult to produce - no satisfying common Tiefe/Metadatenbeschreibung could be produced);
Beispiel:
- difference of data;
- level of ganularity of describing resources
* common access mechanism
biggest problem of all: AHDS couldn't produce enough arguments to show which is the added

value of creating and maintaining electronic resources?
As a consequences, AHDS was dissolved

Post-AHDS-collaborations (UK)
* Reseach Data management landscape (each university has its own digitization)
* discipline-specific initiatives
* e-Science and Grid
* UK Research Data Survey
Alle fühlen sich zuständig für elektronische Daten (OTA zieht sich auf consulting (meta-data, text structure) zurück

Schon in Oxford ist die "Digitalisierungslandschaft" zersplittert.

OTA ist Partner in CLARIN:
Zu den Zielen s. CLARIN short guides
Probleme:
TEI to * via XSLT (=Dublin Core, OLAC, CLARIN-CMF, IMDI, Other -- expose it to OAI-PMH

Conclusions:
the end of the DIY age
* relationships are the problem
- find a place in a competitive environment
- embedding in local ... international infrastructure
- embedding in initiatives
- relationships with stakeholders

14:30 – 15:15 **_M. Durrell (Manchester):_ Das GerManC? Projekt - Aufbau und Konzept eines repräsentativen Korpus der deutschen Sprache 1650–1800**

Motivation: kontrastive Arbeiten zum Standardisierungsprozess beider Sprachen.

Vorbild ARCHER
Repräsentativität:
- keine Volltexte: je Dokument 2000 Wörter aus 9 Textsorten
- Textsorten: Dramen, Zeitungen, Briefe, Predigten, Prosa, Tagebücher, Geistseswss. Texte, nat.wiss Texte, jurist. Texte
- Periodisierung (vg. Bonner Korpus des Frühnhdt.)
- Regionen (norddt, westmitteld, ostmitteld., estoberd. ostoberd.)
- Gesamtgröße: ca. 900.000 Tokens

Pilotprojekt GerManC (bis März 2007):
Versuch auf einer Textsorte: Zeitungen (1650-1800)
Digitalisierung: double keying
Struktur: TEI P4 (tiefe Struktur, einschl. NE, Fremdwörter, Nasalisierung: Problem zu aufwändig)
Tools:
* Variant detector (VARD, Rayson, u.a., 2005)
* Mercurius Baumbank (Demske 2004, 2006)
* RSNSR - Regelbasierte Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung (Pilz u.a. 2008).

Anwendungen:
Untersuchungen zum Standardisierungsprozess
z.B. Flexion des schwachen Adjektivs
* keine Adj. Subst. Kongruenz
* zeitliche Unterschiede

erweitertes GerManC II (seit 2008)

Diskussion: Kooperation mit DTA: Werkzeuge des DTA nachnutzen (Kontakt nimmt Bryan auf)

15:15 – 16:00 **_T. Gloning (Gießen):_ Lexikologische Markierungen in einem digitalen lexikalischen System: Ältere Fach- und Gebrauchstexte**

Ältere Fach-, Gebrauchs- und Alltagstexte (Zeitraum ca. 1460-1900/1920)
Beispiele:
ältere Zeitungstexte u.a. Streitschriften
Kochbücher
Kräuterbücher
medizinische Texte
balneologische Texte
Technik-Texte (Bergbau), Alchemie

Beispiel:
* Kepler (Tertius interveniens)
* 1543: Kräuterbuch
* Kochbuch (1780)

A) Fragestellungen

1 .Wortgebrauch und Wortschatzorganisation
* Welche Wörter stehen in einem besonderen Bezug zu Themen
* Welche Wörter tragen zu Textfunktion bei (Querverweis, Redeeinleitung)
* Lehneinflüsse, Worbildung, Regionalismen

2. Textorganisation und Textaufbau
Texttypen und ihre Dynamik
funktionale Textbausteine

3. Syntakt. Muster

B) Dokumentationssystem

1. Bestandteile
Digitales Texcorpus
Sprachbez. Untersuchungen
Dynam. Dokumentation

2. Vernetzung zwischen Bestandteilen

3. Dynam. Elem. und komplexe Abfragen
komplexe Kreuzklassifikation

veranschauliche Frequenz/Gebrauchsprofil von Querverweis-Ausdrücken

C) Lexikologische Markierungen
Verwendungen
Verwendungsweise
Wortgebrauch = a+b
Parameter

Beispiel: Passevolant (Markierung: Thema Militär, Personenangabe)

Diskussion:

Lexikologische Markierungen werden aus der Zeit heraus markiert, nicht als Projektion über aktuelle Ontologien.

Anregungen zu weiteren Textsorten (Bilbo)