Exposé zur Diplomarbeit

Anfragen an komplexe Korpora

Thorsten Vitt*

13. April 2005

Betreuung:
Dr. Lukas Faulstich, Prof. Ulf Leser,
Forschungsverbund Linguistik – Informatik – Bioinformatik

1 Hintergrund

Das Projekt *Deutsch Diachron Digital (DDD)*[1] entwickelt ein umfangreiches diachrones Korpus des Deutschen: Eine Sammlung zahlreicher historischer und zeitgenössischer deutscher Texte, teils in unterschiedlichen, miteinander alignierten Versionen und mit verschiedenen Annotationen versehen.

Deutsch Diachron Digital hebt sich durch einige Eigenschaften von anderen Korpusprojekten ab:

- Texte werden mit Annotationen in *mehreren, zueinander orthogonalen Annotationsebenen* versehen.

  So kann ein Text etwa gleichzeitig mit seiner physischen (Zeilen, Seiten, . . .) und seiner logischen (Wörter, Sätze, . . .) Struktur annotiert sein.

- Texte können in verschiedenen Ausgaben vorliegen, die *miteinander aligniert* sind.

  Beispielsweise kann ein Text in einer engdiplomatischen (d. h. nahe an der physischen Gestalt orientierten) und einer weiter normalisierten Fassung vorliegen, die aneinander orientiert sind; oder es ist denkbar, einen Text und seine Übersetzung in eine andere Sprache miteinander zu alignieren.

- Das Korpus soll in einer (objekt-) relationalen Datenbank mit XML- und Volltextunterstützung verwaltet werden.

*vitt@informatik.hu-berlin.de
[1] http://www.deutschdiachrondigital.de

- Die Ergebnisse von Anfragen sollen maschinelle Auswertungsmöglichkeiten eröffnen, aber auch in einer für Menschen sinnvoll rezipierbaren Form dargestellt werden können.

  So sind bei einer Anfrage z. B. nach Wörtern mit bestimmten Eigenschaften nicht nur die passenden Wörter, sondern die sie enthaltenden Sätze darzustellen und die relevanten Wörter darin hervorzuheben.

2 Zielsetzung

Die von Dipper et al. (2004) vorgestellte grundlegende Systemarchitektur sieht vor, dass Benutzer über lokale Clients oder über ihren Webbrowser mit einer *Middleware* agieren, die Anfragen an den Datenbankserver stellt und deren Ergebnisse sinnvoll für die Benutzer aufbereitet.

Das System soll dabei soweit wie möglich auf Standard-XML-Techniken und -Werkzeuge zurückgreifen: Insbesondere bieten sich zur Erzeugung der diversen Ausgabeformate XML-Transformationstechniken wie XSLT an. Dazu müssen die Anfrageergebnisse zunächst in ein *Datenschema* (auf XML-Basis) gebracht werden, das eine geeignete Basis für die erwünschten Ausgabeformate ist, indem es für alle unterstützten Ausgaben die benötigten Informationen enthält.

Obschon die Anfragen an das RDBMS in SQL erfolgen werden und als Benutzungsschnittstelle für die Endanwender des Korpus aus den Sprachwissenschaften Webformulare unterschiedlicher Komplexität vorgesehen sind, ist die Entwicklung einer *Anfragesprache* sinnvoll: Zum einen kann diese Sprache als Schnittstelle für Experten oder das Korpus automatisch bearbeitende Anwendungen dienen, zum anderen als Interface zwischen formularartigen oder grafischen Anfrageschnittstellen und dem RDBMS. Dies ist auch dadurch motiviert, dass bereits einfache Anfragen auf der Korpusstruktur recht komplexen und schwer verständlichen SQL-Code erfordern (vgl. Vitt, 2004).

Ziele der Diplomarbeit sind Entwurf und Implementierung einer Anfragesprache für DDD sowie der Entwurf des XML-Datenschemas für die Ergebnisse des Anfragemoduls.

3 Herangehensweise

3.1 Datenschema zur Ergebnisausgabe

Für den Entwurf des XML-Datenschemas für die Ergebnisse muss zunächst analysiert werden, welche Informationen für die Ergebnispräsentation benötigt werden. Dazu muss zunächst festgelegt werden, was die künftigen Benutzungsschnittstellen leisten werden, um dann das hierfür nötige Vorwissen zu ermitteln.

Benutzungsschnittstellen
Hier sind zunächst Eigenschaften zu spezifizieren, die die Repräsentation der Spezifika des Korpus selbst betreffen. So müssen etwa verschiedene Aspekte (aus unterschiedlichen Annotationsebenen) eines Objektes, aggregierte Informationen, alignierte Texte und ggf. alignierte Bilder (Faksimiles) angezeigt werden können.

Darüberhinaus müssen Eigenschaften der Benutzerschnittstelle berücksichtigt werden, die dem Anwender eine auf vorherigen Suchergebnissen aufbauende Navigation ermöglichen. Dies betrifft etwa:

- (weitere) Navigation
- Ein- / Ausblenden (weiterer) Annotationen
- Kontext bzw. Umgebung von Treffern
- Expansion aggregierter Daten
- Suche in den Ergebnissen

Zur Festlegung der Fähigkeiten der Benutzungsschnittstellen können vorhandene Tools aus dem Bereich der diachronen Korpora (vgl. Kroymann et al., 2004, Abschnitt 4.3) und Anforderungen an das Korpus etwa von Lüdeling et al. (2004) betrachtet sowie Linguisten interviewt werden.

Beim Entwurf des Ergebnisschemas ist darüberhinaus zu berücksichtigen, wie Abbildungen auf und von der relationalen Datenstruktur des DBMS, der XML-Datenstruktur der weiterverarbeitenden Tools und das interne Datenschema des Korpus möglich sind.

3.2 Anfragesprache

Für den Entwurf einer Anfragesprache, die sich in das Anfragemodell einfügt, müssen diverse Aspekte berücksichtigt werden:

- *Allgemeine Operatoren* für die Anfragesprache wurden bereits von Faulstich et al. (2005) spezifiziert.
- Das wie im vorigen Abschnitt beschrieben entwickelte *Ergebnisschema* sowie das *Korpus-Datenmodell* müssen geeignet einbezogen werden. Es muss ein Datenmodell spezifiziert werden, auf dem die Sprache arbeitet.
- Zur *Integration mit dem Ergebnisschema* ...
  - sind ggf. weitere Operatoren nötig, um Spezifika wie etwa Hervorhebungen bzw. Kontext anfragen zu können;
  - muss in Bezug auf Fähigkeiten der Benutzungsschnittstellen zum Verfeinern einer bereits gestellten Frage geklärt werden, welche Eigenschaften nicht nur des Datenmodells, sondern auch der Anfragesprache notwendig sind;
  - muss geklärt werden, ob es sinnvoll ist, Kontext- und ähnliche Informationen unmittelbar mit der ersten Abfrage zurückzuliefern oder ob die Benutzungsschnittstellen diese mit zusätzlichen Anfragen (zunächst ohne weitere Interaktion mit dem Benutzer) ermitteln sollten.

- Existierende Sprachen wie XPath und dessen Erweiterungen – etwa LPath (Bird et al., 2005, für linguistische Anforderungen) oder EXPath (Iacob, 2005, für nebenläufige Hierarchien) – oder das aus dem linguistischen Umfeld stammende TIGER (Lezius, 2002) werden zum Vergleich – und um den Benutzern ggf. eine vertraute Syntax anbieten zu können – betrachtet.
- Zur Implementierung der Sprache wird auf die vorhandene prädikatenlogische Spezifikation (Faulstich et al., 2005) und auf die Sprachfeatures der zugrundeliegenden Datenbank einzugehen sein.
- Anfragen an das Datenbank-Backend liefern üblicherweise Tupelsequenzen oder XML-Fragmente. Es ist zu klären, wie diese in das Baum- bzw. Graphmodell des Ergebnisschemas zu überführen sind.

Literatur

Bird, Steven, Chen, Yi, Davidson, Susan, Lee, Haejoong und Zheng, Yifeng (2005): Extending XPath to Support Linguistic Queries. In *Workshop on Programming Language Technologies for XML (PLAN-X)*. Long Beach, California.

Dipper, Stefanie, Faulstich, Lukas, Leser, Ulf und Lüdeling, Anke (2004): Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In *Workshop on XML-based richly annotated corpora*. Lisbon, Portugal. URL http://www.informatik.hu-berlin.de/Forschung_Lehre/wbi/publications/2004/xbrac04_final.pdf.

Faulstich, Lukas C., Leser, Ulf und Lüdeling, Anke (2005): Storing and Querying Historical Texts in a Relational Database. Informatik-Bericht 176, Institut für Informatik der Humboldt-Universität, Berlin.

Iacob, Emil (2005): The Extended XPath language (EXPath) for querying Concurrent Markup Hierarchies. URL http://dblab.csr.uky.edu/~eiacoo/docs/expath/.

Kroymann, Emil, Thiebes, Sebastian, Lüdeling, Anke und Leser, Ulf (2004): Eine vergleichende Analyse von historischen und diachronen digitalen Korpora. Technical Report 174, Institut für Informatik, Humboldt-Universität, Berlin. URL http://www.deutschdiachrondigital.de/publikationen/TRHistorischeKorpora.pdf.

Lezius, Wolfgang (2002): *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart.

Lüdeling, Anke, Poschenrieder, Thorwald und Faulstich, Lukas (2004): DeutschDiachronDigital – Ein diachrones Korpus des Deutschen. *Jahrbuch für Computerphilologie*. URL http://www.deutschdiachrondigital.de/publikationen/ddd-computerphilologie.pdf.

Vitt, Thorsten (2004): *Speicherung linguistischer Korpora in Datenbanken*. Studienarbeit, Humboldt-Universität zu Berlin. URL http://www.informatik.hu-berlin.de/~vitt/stud/studienarbeit/slk.pdf.