

# TECA-Projekt: Mehrschichtig annotierte Texte für interpretative Textanalyse

Melina ALEXA, Alfons GEIS und Ingrid SCHMIDT<sup>1</sup>

## Abstract

In unserem Beitrag berichten wir von dem interdisziplinären Projekt TECA, einer Machbarkeitsstudie, deren Ziel es ist, mehrschichtig linguistisch annotierte Texte zu erstellen und zu nutzen. Eine weitere Perspektive des Projekts ist es herauszufinden, inwieweit linguistische Informationen in der sozialwissenschaftlichen Textanalyse – für die automatische wie die intellektuelle Inhaltsanalyse – eine Hilfe darstellen können. Der Beitrag skizziert zunächst das Projekt, seine Zielsetzung und die Art der Textkorpora, die dafür benutzt werden, um danach das modulare DTD-Modell, das für die Strukturierung und Archivierung der Rohtexte und der codierten Texte entwickelt wurde, vorzustellen.

## 1. Kontext und Zielsetzung des TECA-Projekts

TECA steht für **T**owards **E**xtending **C**ontent **A**nalysis und ist ein Grundlagenforschungsprojekt (ALEXA & GEIS 1998). Aufgrund seines explorativen Charakters ist das Projekt als Machbarkeitsstudie angelegt. Der Schwerpunkt des Projekts ist ein methodischer und besteht darin, das Methodenspektrum für die Analyse sozialwissenschaftlich relevanter Textdaten zu erweitern. Dabei werden Erfahrungen und Techniken aus der Computerlinguistik und dem Anwendungsbereich standardisierter Textformate und Informationsmodellierungssprachen mit einbezogen und für die Sozialwissenschaften nutzbar gemacht.

Im sozialwissenschaftlichen Kontext ist Textanalyse eine Methode, die für die Interpretation von qualitativen, nicht numerischen Daten, d.h.

---

<sup>1</sup> Melina ALEXA, Alfons GEIS: Zentrum für Umfragen, Methoden und Analysen, Postfach 122155, D-68072 Mannheim; {alexa,geis}@zuma-mannheim.de; Ingrid SCHMIDT, Informationsarchitekturen, Heidelberg; schmidt@epc.de.

Textdaten, eingesetzt wird. Dabei kommen verschiedene Texttypen in Frage: Dialogtexte wie etwa Interviews oder politische Debatten, aber auch Zeitungsartikel, Antworten aus Umfragen, Parteiprogramme, Werbetexte oder literarische Texte. Das Untersuchungsmaterial für TECA besteht aus Antworten auf offene Fragen, einem Texttyp, der in Umfragen häufig vorkommt. Im Gegensatz zu den geschlossenen Fragen, die von den Befragten nach dem Multiple-Choice-Prinzip beantwortet werden, handelt es sich bei den Antworten auf offene Fragen um frei formulierte Texte, die von dem Interviewer handschriftlich im Fragebogen oder am Computer festgehalten werden. Die handschriftlichen Texte müssen verschriftet und als Textsammlung in elektronischer Form erfaßt werden; anschließend werden sie inhaltlich analysiert.

Für TECA wurden zwei Textkorpora erstellt, die jeweils Antworten auf offene Fragen aus einer repräsentativen Stichprobe enthalten:

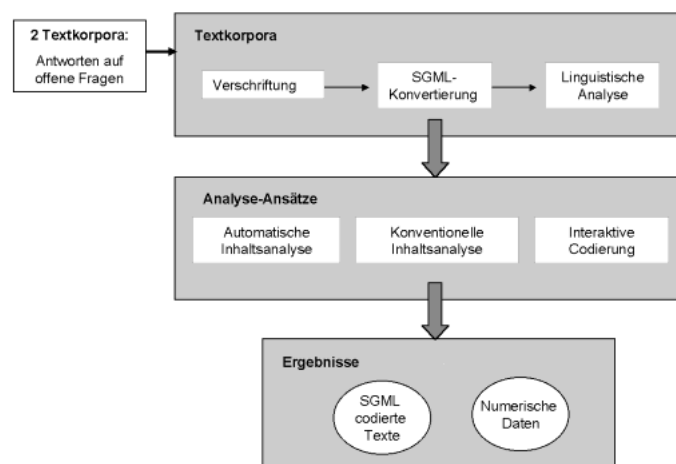
- Das *Stolz-Korpus* besteht aus den Antworten einer Nachwahl-Studie von 1994. Die Frage lautete: “Wenn Sie an die ehemalige DDR zurückdenken, gibt es Dinge, auf die die Menschen dort / Sie (West-/ Ostversion) stolz sein können?” Wenn mit ja geantwortet wurde, kam die Nachfrage: “Und auf was sind die / Sie stolz?”
- Das *Links-/Rechts-Korpus* sammelt die Antworten auf die offene Frage aus einem Forschungsprojekt, das sich mit der Bedeutung der politischen Begriffe “links” und “rechts” seit 1974 befaßt (BAUER-KAASE & GEIS 1998). Die Frage lautete: “Können Sie mir bitte nun noch sagen, was Sie persönlich unter den Begriffen LINKS und RECHTS verstehen, wenn es um Politik geht? LINKS bedeutet: ... RECHTS bedeutet: ...”

Zielsetzung des TECA-Projekts ist es,

- eine Textstandardisierung zu erreichen, das heißt, eine einheitliche Form der Beschreibung und Archivierung zu entwickeln, welche die sozialwissenschaftlichen textanalytischen Arbeitsschritte und verschiedenen Analysemethoden berücksichtigt,
- sowie einen Methodenvergleich durchzuführen zwischen der automatischen, der intellektuellen und der interaktiven Codierung, welche zusätzlich morphosyntaktische Informationen nutzt.

Darüber hinaus sollen die linguistischen Informationen spezifiziert werden, die für die inhaltliche Analyse genutzt werden können.

Eine graphische Darstellung verdeutlicht den Projektablauf:



Zunächst werden auf der Basis eines Regelwerks die im Fragebogen notierten Antworten verschriftet. Die maschinenlesbaren Texte werden nach SGML konvertiert, wobei die semantischen Bestandteile ihrer Struktur identifiziert und ausgezeichnet werden. Die zwei SGML-Korpora werden mit Hilfe eines morphosyntaktischen Parsers (FIRZLAFF & KÖNYVES-TÓTH 1999) analysiert und mit linguistischen Angaben wie Wortstämmen, Wortartkategorien, Verbphrasen, Nominalphrasen etc. angereichert. Im weiteren werden drei Analyseansätze angewendet: die automatische Inhaltsanalyse, die konventionelle (intellektuelle) Inhaltsanalyse und die interaktive Codierung. Letztere ist eine Kombination der Verfahren, die darüber hinaus das mit allen verfügbaren Informationen angereicherte Textkorpus nutzt. Damit werden *mehrschichtige Analysen* möglich (ALEXA & SCHMIDT 1999), das heißt, es können gezielt verschiedene Informationsebenen ausgewertet werden. Die drei Ansätze werden miteinander verglichen und bewertet. Am Ende dieses Arbeitsschrittes stehen SGML-Textdaten, die verschiedene Arten von Informationen enthalten können, beispielsweise Code-Zuweisungen und linguistische Annotationen. Da-

neben entstehen numerische Dateien mit den Codierungen, die mit üblichen Statistikprogrammen ausgewertet werden können.

## 2. Mehrschichtige Annotation

Jedes Korpus enthält drei Annotationsebenen, nämlich Verschriftung, Morphosyntax und *problemorientierte* Codierung (*problem-oriented tagging*, s. DE HAAN 1984), eine interpretative Codierung, die sich an der inhaltlichen Fragestellung eines Projekts oder einer Studie orientiert. Die problemorientierte Codierung erfolgt zweimal, automatisch und intellektuell, mit einer jeweils eigenen Annotation. So ergeben sich für jedes Korpus vier Annotationseben, die den vier verschiedenen Stufen der Textbearbeitung entsprechen: (1) Verschriftung, (2) morphosyntaktische Annotation, (3) automatische Inhaltsanalyse und (4) konventionelle Inhaltsanalyse.

Mit Ausnahme der Verschriftung wird für jede Ebene ein Kategorienschema benutzt. Das Tagset für das morphosyntaktische Tagging ist für beide Korpora identisch. Für jedes der beiden Textkorpora und jeden Codiermodus (automatisch und intellektuell) liegt ein eigenes Kategorienschema vor. Der Umfang der Kategorien für die Kategorienschemata bzw. das Tagset variiert stark. Das Schema für die automatische Inhaltsanalyse des Stolz-Korpus enthält zum Beispiel nur zwölf Kategorien, das Schema für die konventionelle Inhaltsanalyse des Links-/Rechts-Korpus 180.

Die Anreicherung des Textes mit umfassender Information und die Verwendung für unterschiedliche Analyseansätze erfordern ein entsprechendes Textformat, das es ermöglicht, neben dem eigentlichen Text alle vorhandenen Informationen verfügbar zu haben und wahlweise abzurufen. Damit das Textformat über das derzeitige Projekt hinaus verwendbar wird, muß es den Ansprüchen eines Standards genügen. Daher haben wir uns für SGML entschieden.

## 3. Modulares DTD-Konzept

Aufgrund der Problemstellung lag es nahe, ein modulares Konzept für die Antworten auf offene Fragen zu entwickeln. Dabei haben sich fünf inhaltlich konzipierte Module herauskristallisiert, die fünf verschiedene Stufen der Bearbeitung der Antworten auf offene Fragen widerspiegeln,

nämlich die Verschriftung, die linguistische Analyse, die konventionelle Inhaltsanalyse, die automatische Inhaltsanalyse und die Archivierung.

Zunächst wurde für jedes der fünf Module eine Dokument-Typ-Definition (DTD) erstellt. Dabei zeigte sich, daß jede dieser DTDs im wesentlichen aus derselben Grundstruktur bestand, die nur an bestimmten Stellen variierte. Änderte sich etwas an der gemeinsamen Grundstruktur, was leicht vorkommen kann, dann mußten an allen fünf DTDs die Änderungen vorgenommen werden. Deshalb wurde eine *Rahmenstruktur-DTD* entwickelt, auf die von den *Haupt-DTDs* aus als *DTD-Subset* zugegriffen wird. Zusammen mit veränderbaren Platzhaltern in der Rahmen-DTD ergab sich so eine hohe Flexibilität, und gleichzeitig waren die Einzelmodule jeweils einem bestimmten Arbeitsschritt zuzuordnen.

Dieses Prinzip der DTD-Organisation wurde für die linguistische Analyse ausgeweitet. Die linguistischen Annotationen basieren auf denen der Text Encoding Initiative (TEI, SPERBERG-MCQUEEN & BURNARD 1994). Da nur ein Teil der von der TEI vorgeschlagenen Annotationen gebraucht wurde, und um auch morphosyntaktische Ambiguitäten abbilden zu können, wurden inhaltliche Modifikationen nach den Regeln der TEI durchgeführt, die auch den softwarespezifischen Erfordernissen des morphosyntaktischen Parsers Rechnung trugen.

In technischer Hinsicht wurden die Regeln der TEI nicht eingehalten. Das sogenannte *pizza model* der TEI mit Rahmenstruktur (*Teig*), Struktur (*Grundbelag*) und rein textbezogenen Strukturen (*zusätzlicher Belag*) konnte nicht verwendet werden, da für das TECA-Projekt nur die linguistischen Annotationen der TEI gebraucht wurden.

Für das TECA-DTD-Modell wurde also ein DTD-Subset der TEI-Definitionen hinzugefügt, das festlegte, welche TEI-Elemente übernommen, welche variiert und welche weggelassen wurden, sowie das TEI-Subset selbst mit den TECA-Modifikationen. Zur linguistischen Annotation und Archivierung wird auf diese beiden DTD-Subsets neben dem DTD-Subset für die Rahmenstruktur in der Haupt-DTD Bezug genommen.

Kurz gesagt, inhaltlich werden fünf Module unterschieden. Diese fünf Module sind SGML-technisch als fünf Haupt-DTDs organisiert, die sich alle auf das Rahmenstruktur-DTD-Subset beziehen. Die Haupt-DTD für

die linguistische Annotation und die für die Archivierung referenzieren darüber hinaus noch die beiden anderen DTD-Subsets (für weitere Details zum Modell siehe SCHMIDT & ALEXA 1998).

#### 4. Analyseablauf: ein Beispiel

Im folgenden wird das Modell anhand eines Antworttextes aus dem Stolz-Korpus verdeutlicht.

##### 4.1 Verschriftung und SGML-Konvertierung

Bei der Verschriftung umfaßt der Antworttext die Antwort und die Fragebogennummer:

**3288** Die Kameradschaft unter den Kollegen , die zwischenmenschlichen Beziehungen ; was unter schwierigen Bedingungen geschaffen wurde .

Nach der Konvertierung des Texts in SGML wird der Antworttext so ausgezeichnet:

```
<!DOCTYPE OFFENEFRAGEN SYSTEM "O-FRAGE.DTD">
<offeneFragen>
<projektdaten>
<titel>Stolz-Texte</titel>
<umfragezeitraum von="1.4.95" bis="30.4.95">
<betreuer>Alfons Geis</betreuer>
<sprache sprache="DE">
<fragetext id-frage="f1">Wenn Sie an die ehemalige
DDR zur&uuml;ckdenken, gibt es Dinge, auf die die
Menschen dort/Sie (West-/Ostversion) stolz sein
k&ouml;nnen? Wenn mit ja geantwortet wurde, kam die
Nachfrage: Und auf was sind die/Sie stolz?
</fragetext>
[... ]
<fragebogen ref-fragebogen="3288">
<antwort ref-frage="f1">Die Kameradschaft unter den
Kollegen , die zwischenmenschlichen Beziehungen ;
was unter schwierigen Bedingungen geschaffen wurde
</antwort>
</fragebogen>
```

Jeder Antworttext wird durch die Fragebogennummer (<ref-fragebogen="3288">) und die Fragennummer (<ref-frage="f1">) eindeutig identifizierbar gemacht und die Gesamtheit der Fragen wird durch Informationen zum Projekt (<Projektdaten>) ergänzt.

## 4.2 Morphosyntaktische Analyse

Der Parser nimmt als Input das verschriftete und nach SGML konvertierte Korpus, das morphosyntaktisch analysiert und entsprechend in SGML annotiert wird. Bei der Analyse wird jedes Wort einer Wortartkategorie zugeordnet, sein Lemma wird hinzugefügt und jeder Antworttext wird syntaktisch analysiert:

```
<fragebogen ref-fragebogen="3288">
<antwort ref-frage="f1">
<S>
[... ]
<PHR ANA="NP">
<W ANA="ART" LEMMA="d-" LEMMA-ANZAHL="1"> Die </W>
<W ANA="NOMEN" LEMMA="Kameradschaft" LEMMA-
ANZAHL="1"> Kameradschaft </W> </PHR>
<PHR ANA="PP"> <W ANA="PRAEP" LEMMA="unter" LEMMA-
ANZAHL="1"> unter </W>
<PHR ANA="NP"> <W ANA="ART" LEMMA="d-" LEMMA-
ANZAHL="1"> den </W>
<W ANA="NOMEN" LEMMA="Kollege" LEMMA-ANZAHL="1">
Kollegen </W> </PHR> </PHR>
[... ]
</S>
</antwort>
</fragebogen>
```

Der Parser wurde für die Analyse “normaler” deutscher Texte entwickelt. Die Analyse der Antworten auf offene Fragen ist eine sehr spezielle Aufgabe, da häufig elliptische, fehlerhafte und unvollständige Sätze vorkommen. Der Parser produziert entweder eine komplette syntaktische Struktur für jeden Satz oder führt eine partielle Analyse durch, wenn bei grammatikalisch unvollständigen Sätzen ein sogenanntes “Rettungs-Modul” greift (FIRZLAFF & KÖNYVES-TÓTH 1999).

### 4.3 Konventionelle Inhaltsanalyse

Der SGML-Text auf der Grundlage der konventionellen Inhaltsanalyse enthält neben sonstigen projektspezifischen Informationen auch das Kategorienschema mit Kategorienbezeichnung und Kategorienkennziffer – die Kategorie “soziale Dienste” hat zum Beispiel die Ziffer 12. Im Text selbst werden die manuell vergebenen Kategorien in Form der sie repräsentierenden Ziffern nur für die gesamte Antwort (nicht für jede einzelne codierte Textstelle) angezeigt, und zwar als Attributwert zu dem Attribut Kategorien:

```
<fragebogen ref-fragebogen="0196">
<antwort ref-frage="f1" kategorien="12 02 11"> Gutes
Gesundheitssystem , alles kostenlos ; Kin-
derg&auml;rten waren kostenlos . </antwort>
</fragebogen>.
```

### 4.4 Automatische Inhaltsanalyse

Das SGML-codierte Korpus enthält ebenfalls projektspezifische Informationen, wie zum Beispiel den Namen der für die automatische Inhaltsanalyse eingesetzten Textanalyse-Software und den Dateinamen des verwendeten Kategorienschemas. Die Kategorien werden als Attributwerte zum Attribut CODENR den jeweiligen Textstellen genau zugeordnet.

```
<fragebogen ref-fragebogen="0196">
<antwort ref-frage="f1"> Gutes Gesundheitssystem
<CODE CODENR="12">, alles kostenlos <CODE CO-
<DENR="02"> ; Kinderg&auml;rten <CODE CODENR="11">
waren kostenlos <CODE CODENR="02"> . </antwort>
</fragebogen>.
```

Wichtig zu erwähnen ist, daß es sowohl für die konventionelle Analyse als auch für die automatische keine überlappenden Markierungen gibt.

## 5. Schlußfolgerung

Mit dem DTD-Modell können, entsprechend einem der Projektziele, alle Arbeitsschritte der verwendeten Analyseansätze in einer einheitlichen Form beschrieben und archiviert werden. Weitere Vorteile ergeben sich dadurch, daß die verschiedenen Analyseebenen unabhängig voneinander ansprechbar sind und Texte so für unterschiedliche Zwecke genutzt werden können; darüber hinaus können die Metadaten nicht verloren gehen, wie es bei üblichen Notations- und Archivierungsverfahren leicht geschehen kann.

Von den Fragen und Aspekten, die über die Aufgabenstellung des Projekts und den Kernbereich der sozialwissenschaftlichen Textanalyse hinausgehen, seien beispielsweise die Einbeziehung von linguistischen Informationen, die mehrschichtige Analyse und die Entwicklung von geeigneter Textanalyse-Software genannt:

- *Linguistische Informationen*: Wie modelliert man linguistische Annotationen? Welches Modell bietet sich für welche Zwecke an? So wurde zum Beispiel bei der Entwicklung des modularen Konzepts im TECA-Projekt auch das Textmodell des *Corpus Encoding Standards* (CES 1998) in die Überlegungen einbezogen. Trotz des Vorteils der flexiblen Erweiterbarkeit wurde darauf verzichtet, unter anderem deshalb, weil wir bislang keine Software ausfindig machen konnten, welche die Adaption des CES-Modells für das TECA-Projekt unterstützen würde.
- *Mehrschichtige Analyse*: Worin besteht im einzelnen der Vorteil, die unterschiedlichen Informationsebenen zu trennen? Ist ein mehrschichtig annotierter Text besser verwendbar? Es ist zu erwarten, daß mehrschichtig annotierte Texte für einen größeren Nutzerkreis interessant werden.
- *Textanalyse-Software*: Welche Konsequenzen ergeben sich aus solchen modularen DTDs für das Datenmodell von Textanalyse-Software? Insbesondere muß eine Software, die für die manuelle oder semi-automatische Textanalyse konzipiert ist, derartige Modelle berücksichtigen, um eine mehrschichtige Textanalyse zu ermöglichen.

## Literatur

- ALEXA, M. und GEIS, A. (1998): Towards Extending Content Analysis (TECA): ein ZUMA-Grundlagenforschungsprojekt. In: *ZUMA Nachrichten* 43, 150-152.
- ALEXA, M. und SCHMIDT, I. (1999): Modell einer mehrschichtigen Textannotation für die computerunterstützte Textanalyse. In: W. MÖHR und I. SCHMIDT (Hrsg.): *SGML und XML – Anwendungen und Perspektiven*. Heidelberg: Springer-Verlag.
- BAUER-KAASE, P. und GEIS, A. (1998): Towards Extending Content Analysis (TECA). Schlußbericht zu Arbeitspaket 2: Coderbasierte Inhaltsanalyse. In: *ZUMA-Technischer Bericht T98/19*. Mannheim: ZUMA.
- CES (1998): Corpus Encoding Standard (CES). <http://www.cs.vassar.edu/~ide/CES> [Stand: August 1999].
- GEIS, A. (1998): Towards Extending Content Analysis (TECA). Schlußbericht zu Arbeitspaket 1: Verschriftung. In: *ZUMA-Technischer Bericht T98/18*. Mannheim: ZUMA.
- DE HAAN, P. (1984): Problem-oriented tagging of English corpus data. In: J. AARTS and W. MEIJIS (eds.): *Corpus Linguistics*, 123-139. Amsterdam: Rodopi.
- FIRZLAFF, B. und KÖNYVES-TÓTH, M. (1999): Towards Extending Content Analysis (TECA): Schlußbericht zu Arbeitspaket 5: Morpho-syntaktische Analyse von deutschsprachigen Antworttexten. In: *ZUMA-Technischer Bericht T99/01*. Mannheim: ZUMA.
- SCHMIDT, I. und ALEXA, M. (1998): Towards Extending Content Analysis (TECA): Schlußbericht zu den Arbeitspaketen 4 und 6: Umsetzung in SGML-Format. In: *ZUMA-Technischer Bericht T98/16*. Mannheim: ZUMA.
- SPERBERG-MCQUEEN, M.C. and BURNARD, L. (eds., 1994): Guidelines for the encoding and interchange of machine-readable texts (TEI P3). Chicago & Oxford, April 1994.