Towards Extending Content Analysis (TECA) - Schlußbericht zu den Arbeitspaketen 4 und 6, Umsetzung in SGML-Format

Schmidt, Ingrid; Alexa, Melina

Veröffentlichungsversion / Published Version
Abschlussbericht / final report

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Schmidt, I., & Alexa, M. (1998). *Towards Extending Content Analysis (TECA) - Schlußbericht zu den Arbeitspaketen 4 und 6, Umsetzung in SGML-Format.* (ZUMA-Technischer Bericht, 98/16). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-48752-1

ZUMA-Technischer Bericht T98/16
ISSN 1437-4129

Towards Extending Content Analysis (TECA)

Schlußbericht zu den Arbeitspaketen 4 und 6
Umsetzung in SGML-Format

Ingrid Schmidt
Melina Alexa

ZUMA
Postfach 12 21 55
68072 Mannheim

Telefon: (06 21) 1246 - 222
Telefax: (06 21) 1246 - 100
E-Mail: alexa@zuma-mannheim.de

ZUMA-Grundlagenforschungsprojekt TECA

TECA ist eine Pilotstudie, die das Methodenspektrum der Analyse von sozialwissenschaftlichen Texten erweitern soll. Dabei werden die Erfahrungen und Techniken aus der Linguistik, insbesondere der Computerlinguistik, sowie der angrenzenden Wissenschaften einbezogen und für die Sozialwissenschaften nutzbar gemacht.

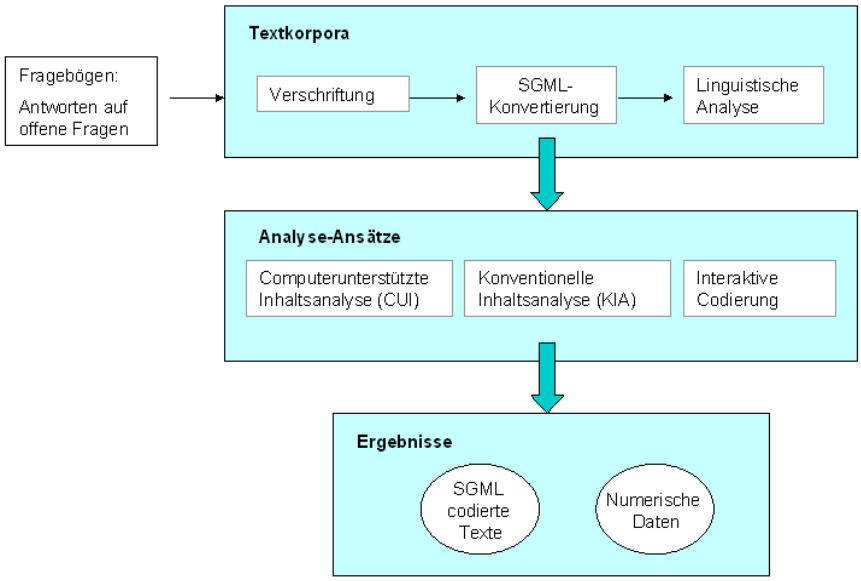
Eine Schwäche der Methode der *computerunterstützten Inhaltsanalyse* besteht darin, daß die Analyse überwiegend auf Einwort-Codierungen basiert, ohne den Kontext zu berücksichtigen. Diese fehlende Kontextsensitivität führt zu ungenauen Ergebnissen bzw. beschränkt die Anwendung der Methode der computerunterstützten Inhaltsanalyse auf bestimmte Themen und Textsorten. Ein wesentlicher Nachteil der *coderbasierten* (oder konventionellen) *Inhaltsanalyse* besteht u.a. in dem hohen Zeitaufwand und den aufwendigen Arbeitstechniken/-abläufen.

Lösungsmöglichkeiten sehen wir darin, daß einerseits zusätzliche linguistische Informationen bereitgestellt werden, die der computerunterstützten Inhaltsanalyse als „intelligente“ Komponente und der Codierkraft als leicht zugängliche Zusatzinformation dienen, andererseits versprechen wir uns in der Integration und gemeinsamen Anwendung der coderbasierten wie der computerunterstützten Inhaltsanalyse synergetische Effekte.

Als Untersuchungsmaterial wurde ein Texttyp gewählt, wie er in der Alltagspraxis von ZUMA am häufigsten vorkommt: Antworten auf offene Fragen aus zwei repräsentativen Stichproben. Sowohl kurze stichwortartige Antworten als auch längere argumentative sind vertreten. Für jede Stichprobe liegt ein Textkorpus vor. Die zwei Textkorpora werden mit Hilfe eines linguistischen Analysesystems (Parsers) bearbeitet. Dadurch wird jedes Textkorpus mit linguistischen Angaben angereichert, wie z.B. Wortstämmen, Wortartkategorien und syntaktischen Komponenten wie Verbphrasen, Nominalphrasen etc.

Im weiteren werden drei Analyse-Ansätze angewendet: die coderbasierte Inhaltsanalyse, die computerunterstützte Inhaltsanalyse und eine interaktive Codierung, die neben der Kombination der beiden ersten Verfahren auch das mit allen verfügbaren Informationen angereicherte Textkorpus nutzt. Die drei Ansätze werden miteinander verglichen und bewertet.

Ausgangspunkt für diese Analysen sind also Texte, die neben dem eigentlichen Text die Beschreibung der Textstruktur (inkl. projektspezifischer Daten), die maschinell und coderbasiert generierten Codes, die zugrunde liegenden Kategorienschemata und linguistischen (morpho-syntaktische) Merkmale enthalten wie z.B. Wortkategorien, Verbphrasen, Nominalphrasen u.a. Damit werden „Multilevel-Analysen“ möglich, d.h. es können parallel verschiedene Informationsebenen ausgewertet werden. Außerdem versprechen wir uns davon eine Erweiterung des bisherigen Kategorisierungsspektrums sowie die Überprüfbarkeit und Präzisierung der bisherigen Codierung. In der folgenden Grafik wird der Analyseablauf für das Projekt TECA skizziert:



Die Anreicherung des Textes mit umfassender Information und die Verwendung für die unterschiedlichen Ansätze erfordern ein entsprechendes Textformat, das es erlaubt, neben dem eigentlichen Text alle vorhandenen und generierten Informationen verfügbar zu haben und wahlweise abrufen zu können. Das Textformat muß den Ansprüchen eines Standards genügen; nur so ist es über das derzeitige Projekt hinaus verwendbar. Seit einiger Zeit zeichnet sich eine Standardisierung auf dem Gebiet der Textdokumentation und -archivierung ab, nämlich SGML (Standard Generalized Markup Language). SGML ist plattform- und maschinenunabhängig, frei verfügbar (nicht-proprietär) und bietet eine genormte Vorgehensweise, eine Textstruktur zu spezifizieren: eine Document Type Definition (DTD) definiert mit Hilfe einer Reihe von Regeln die Struktur des Textes. Inhalt und Struktur bilden ein Ganzes, wobei die Strukturelemente und ihre Beziehungen zueinander durch die jeweiligen Dokumenttypen eindeutig beschrieben und definiert werden; eine solche DTD ist bereits für den in diesem Projekt verwendeten Texttyp erstellt und für die linguistischen Merkmale erweitert worden. Auch bezüglich der Texterfassung (Verschriftung) unmittelbar im SGML-Format liegen erste vielversprechende Erfahrungen vor.

Mit der Einführung eines Standards für Textformate wird gleichzeitig ein Hindernis beseitigt, indem die Austauschbarkeit von Texten zwischen verschiedenen Programmen und die Weitergabe von Texten zwischen Wissenschaftlern erleichtert wird.

Bei dem Projekt TECA handelt es sich um eine Machbarkeitsstudie, in der geprüft wird, mit welchem Aufwand welche Ergebnisse erzielt werden können und wie die verschiedenen Analyse-Ansätze zusammenwirken können (Abfolge, Unterstützung, Ergebnis). Die derzeit noch unbeantwortete Frage, inwieweit sich die gewonnenen Erfahrungen auch auf andere Texttypen (z.B. Leitfadengespräche oder Medientexte) übertragen lassen, wird am Ende zu diskutieren sein.

Für weitere Informationen beziehungsweise Anregungen wenden Sie sich an:
Dr. Melina Alexa (alexa@zuma-mannheim.de) und
Alfons Geis (geis@zuma-mannheim.de)

Inhaltsverzeichnis

1. Abstract

Der hier vorliegende Schlußbericht zu den Arbeitspaketen 4 und 6 des TECA-Projekts gibt zunächst einen Überblick über das entwickelte modulare Konzept (Abschnitt 1), um anschließend die einzelnen Module unter inhaltlichen Gesichtspunkten zu diskutieren (Abschnitt 2). Dannach wird auf die zugrundegelegte SGML-Modellierung eingegangen und die damit zusammenhängende Aspekte werden erörtert (Abschnitt 3). Die abschließenden Schlußbemerkungen fassen die Schwierigkeiten und die weiterführenden Zielsetzungen zusammen (Abschnitt 4).

2. Beschreibung des modularen Konzepts

Aufgrund der Ergebnisse des Zwischenberichts vom 23. Juni 1998 und der sich daran anschließenden Überlegungen wurde ein modulares Konzept für die Antworten auf offene Fragen entwickelt. Im Zuge dieser Diskussionen haben sich fünf inhaltlich motivierte Module herauskristallisiert, in denen fünf verschiedene Stufen in der Bearbeitung der Antworten auf offene Fragen zum Ausdruck kommen. Im Arbeitspaket 4 wurden Module für die Verschriftung, die konventionelle Inhaltsanalyse (KIA), die computerunterstützte Inhaltsanalyse (CUI) und die Archivierung erstellt. Im Arbeitspaket 6 entstand ein weiteres Modul für linguistische Annotationen, und das schon bestehende Archivierungsmodul wurde in diesem Sinne erweitert.

Aus SGML-technischer Sicht stellte sich im Arbeitspaket 4 die Situation so dar, daß für jedes der fünf inhaltlich verankerten Module eine Dokumenttypdefinition (DTD) erstellt wurde. Dabei zeigte sich, daß jede dieser Haupt-DTDs im wesentlichen aus derselben Grundstruktur bestand, die an bestimmten Stellen variierte. Das bedeutet, daß eventuell anfallende Änderungen an dieser Grundstruktur jedes Mal in allen fünf Haupt-DTDs nachgeführt werden müßten, damit die Struktur für alle Module über die Zeit konsistent bleibt. Die Wahrscheinlichkeit von Inkonsistenzen ist in einem solchen Modell sehr hoch. Um dem entgegenzuwirken, wurde die Grundstruktur aus den fünf Haupt-DTDs extrahiert und in einer separaten Datei als Rahmenstruktur-DTD abgelegt. Diese Rahmenstruktur ist nun nicht mehr physischer Bestandteil der einzelnen Haupt-DTDs, sondern sie wird als sogenanntes DTD-Subset von den Haupt-DTDs aus referenziert. Um die modulspezifischen Modifikationen berücksichtigen zu können, besteht die Rahmenstruktur-DTD neben den festen Definitionen aus veränderbaren Platzhaltern. Ein solches Modell bewirkt einerseits, daß die Rahmenstruktur nur an einer Stelle gepflegt werden muß, um eine konsistente Grundstruktur in allen Modulen sicherzustellen, andererseits erlaubt es, daß sich die Struktur innerhalb der Module an bestimmten Stellen unterscheiden kann.

Dieses Prinzip der DTD-Organisation wurde für die linguistischen Annotationen im Rahmen des Arbeitspakets 6 übernommen. Da die linguistischen Annotationen auf denen der Text Encoding Initiative (TEI) basieren, entstanden ein weiteres DTD-Subset für die TEI-Definitionen und ein zweites für die TECA-Modifikationen der TEI-Definitionen. Diese beiden DTD-Subsets werden neben dem DTD-Subset für die Rahmenstruktur in der Haupt-DTD für die linguistischen Annotationen und für die Archivierung referenziert.

Das modulare Konzept bezieht sich somit auf zwei unterschiedliche, miteinander verschränkte Ebenen: Inhaltlich werden fünf Module unterschieden. Diese fünf Module sind SGML-technisch als fünf Haupt-DTDs organisiert, die alle das Rahmenstruktur-DTD-Subset referenzieren. Die Haupt-DTD für die linguistischen Annotationen und die für die Archivierung referenzieren darüber hinaus noch die beiden anderen DTD-Subsets.

Die Module/Haupt-DTDs sind:

Modul 1: Verschriftung

Modul 2: Linguistische Annotation

Modul 3a: Konventionelle Inhaltsanalyse (KIA)

Modul 3b: Computerunterstützte Inhaltsanalyse (CUI)

Modul 4: Archivierung

Die DTD-Subsets sind:

Subset 1: Rahmenstruktur für Antworten auf offene Fragen

Subset 2: TECA-spezifische Elemente und Attribute für die linguistische Analyse

Subset 3: Modifizierte Version der TEI-DTD teiana2.dtd

Bei der Entwicklung dieses modularen Konzepts wurde im Vorfeld auch das Textmodell des *Corpus Encoding Standards* (CES) (→ *2.2 Modul 2: Linguistische Annotation*) in die Überlegungen mit einbezogen. Es hält den Originaltext und die Annotationen voneinander getrennt. Die Annotationen werden somit nicht, wie allgemein üblich, direkt in den Text eingebracht, sondern sind in einer separaten Datei gespeichert. Annotation und Originaltext werden mittels eines HyTime-basierten, von der TEI entwickelten Adressierungsmechanismus miteinander verknüpft. Ein solches Textmodell wäre durchaus auch auf das TECA-Projekt übertragbar. Es hätte den Vorteil, daß neben KIA, CUI und Linguistik problemlos und flexibel weitere Annotationsaspekte hinzugefügt und auch wieder weggenommen werden könnten, indem einfach Annotationsdateien entweder berücksichtigt werden oder entfallen. Bei direkt im Originaltext vorgenommenen Annotationen führt das Hinzufügen immer weiterer Aspekte zu einer "Überladenheit" des Originaltextes, häufig auch zu Überschneidungen der Kodierungen und damit zu großer Unübersichtlichkeit. Darüber hinaus ist das Entfernen von Annotationsaspekten aus einem Korpus sehr zeitaufwendig, da jede einzelne Kodierung angefaßt werden muß. Die Adaption des CES-Modells für das TECA-Projekt scheitert jedoch derzeit daran, daß am Markt keine Software verfügbar ist, die die Handhabung dieses Textmodells unterstützen würde.

3. Inhaltliche Beschreibung der Module

3.1 Modul 1: Verschriftung

Die Verschriftung ist die erste Stufe der elektronischen Verarbeitung der Antworten auf offene Fragen einer Umfrage. Dabei werden die zwei großen Blöcke Projektdaten und Fragebögen unterschieden. Dazwischen wird in einem separaten Informationsblock der Name des Verschrifters erfaßt.

Projektdaten

Die Projektdaten sind:

- Projekttitel
- Umfragezeitraum
- Projektbetreuer
- Sprache der Umfrage
  Die Sprachen Deutsch, Schweizerdeutsch, Englisch und Französisch werden von anderen Sprachen unterschieden.
- Wortlaut der offenen Frage(n)
  Jede Frage kann als Frage, Unterfrage und Angabe typisiert werden.
- Allgemeine Informationen
  In diesem Pufferelement können arbiträre projektspezifische Informationen eingegeben werden.

Fragebögen

Unter Fragebögen ist die Gesamtheit der verschrifteten Antworten der Fragebögen der Umfrage zu verstehen. Dabei wird zunächst unterschieden, ob in der Umfrage eine oder mehrere offene Fragen vorkommen. Daran schließen sich die einzelnen Fragebögen mit dem Antworttext bzw. den Antworttexten an. Außerdem ist jedem Fragebogen ein Memofeld zugeordnet, in dem Anmerkungen aller Art zu den Antworten gemacht werden können.

3.2 Modul 2: Linguistische Annotation

Die linguistischen Annotationen basieren auf der oben beschriebenen Struktur des Verschriftungsmoduls. Sie werden in einem automatisieren Prozeß von einer linguistischen Analysesoftware vorgenommen. Nach dem Namen des Verschrifters wird der Name der Analysesoftware eingefügt.

Das Analyseprogramm nimmt in der Regel eine vollständige Analyse einer Antwort vor. Dabei werden zunächst die einzelnen Sätze ausgewiesen, innerhalb welcher dann weitere syntaktische Kategorien und Wortartkategorien markiert sind. Ambige Strukturen werden auf jeder Ebene ausgezeichnet und können hinsichtlich syntaktischer Strukturen auch geschachtelt vorkommen. In Ausnahmefällen sind auf der Satzebene auch Teilanalysen möglich, die dann von einem vollständig analysierten Satz eindeutig unterschieden sind.

Syntaktische Kategorien

Syntaktische Kategorien auf der Satzebene sind:
- Satz
- unvollständig analysierter Satz

Allgemeine syntaktische Kategorien sind:
- Adverbphrase
- Apposition
- Adjektiv-Attribut-Phrase
- Genitivattribut
- Hilfsverbphrase
- Infinitiv
- Interrogativsatz
- Komparativphrase
- Konjunktionalsatz
- Koordination
- Kopulaverbphrase
- Lokalsatz
- Nebensatz
- Nominalphrase
- Präpositionalphrase
- Prädikat
- Folge von Präpositionen
- Relativsatz
- Verbphrase
- Zahlausdrücke, Ordinalzahlen mit Punkt

Spezielle syntaktische Kategorien sind:
- Abkürzung mit Punkt
- Datum
- geklammerter Textteil
- Maßanagabe
- komplexer Name
- Zitat
- Zitatnachfeld

Syntaktische Kategorien speziell für die Antworten auf offene Fragen sind:
- Koordination, die spezifisch für die Antworttexte ist
- grammatikalisch unvollständiger Satz, der spezifisch für die Antworttexte ist

Wortartkategorien

Wortarkategorien für Verben sind:
- Verb
- Hilfsverb
- Kopulaverb
- Verb im Infinitiv

Wortartkategorien für Pronomen sind:

- Pronomen
- Fragepronomen
- Interrogativpronomen
- Relativpronomen

Andere allgemeine Wortartkategorien sind:
- Adjektiv
- Adverb
- Artikelwort
- Konjunktion
- Nomen
- Präposition
- Präfix
- Quantor

Wortartkategorien für Zahlen sind:
- Kardinalzahl
- Ordnungszahl

Wortartkategorien für Zeichen sind:
- Abkürzungs- und Ordinalzahlpunkt
- halbe Wörter mit Bindestrich
- Satzzeichen

Spezielle Wortartkategorien sind:
- Abkürzung
- Datums- und Jahresangaben
- Name
- Sonderzeichen

Bei der Entscheidung, wie die linguistischen Annotationen kodiert werden sollen, stellte sich zunächst die Frage, ob dafür schon bestehende, frei verfügbare Strukturen genutzt werden können, oder ob im Rahmen des TECA-Projekts eine eigene Struktur entwickelt werden soll. Dabei war zu bedenken, daß eine eigene Struktur zwar einen hohen Entwicklungsaufwand bedeuten würde, andererseits aber auch die Prüfung bestehender Strukturen auf ihre Verwendbarkeit zeitaufwendig ist.

Da der eingesetzte Syntaxparser Ambiguität sowohl auf der Lemma-Ebene als auch bezogen auf größere syntaktische Einheiten ausweist, war es wichtig, diese Information nicht verloren zu geben, sondern sie bei der Strukturierung zu berücksichtigen.

Gerade in der Linguistik gibt es Bestrebungen, eine gemeinsame Basis für die Kodierung linguistisch annotierter Korpora zu schaffen. Die Idee dahinter ist, einen Quasi-Standard zu schaffen, der es einem Wissenschaftler ermöglicht, Texte verschiedenen Ursprungs mit nur geringem Aufwand zu einem für seine Forschungsziele geeigneten Korpus zusammenzuführen.

In den Geisteswissenschaften am weitesten verbreitet sind die von der *Text Encoding Initiative* (TEI) entwickelten Strukturen, die sich als Richtlinien verstanden wissen wollen. Diese sind so angelegt, daß die Basisstruktur unverändert übernommen oder anwendungsbezogen modifiziert werden kann. Bei der Prüfung der linguistischen Annotationen stellte sich heraus, daß von der TEI keine Möglichkeit vorgesehen ist, morpho-syntaktische Ambiguitäten abzubilden. Daher wurden im Hinblick darauf zwei weitere einschlägige linguistische Projekte mit in Betracht gezogen, von denen jedoch keines einen geeigneten Lösungsvorschlag bot:

Bei der *Expert Advisory Group on Language Engineering Standards* (EAGLES) heißt es: "The encoding of ambiguity in morphosyntactic annotation has so far received little attention, and we make no recommendations except to propose that in principle, all the kinds of ambiguity listed above should be distinguishable by different mark-up."

Der *Corpus Encoding Standard* (CES) – einerseits TEI-basiert, andererseits an den Ergebnisse des EAGLES-Projekts orientiert – schien auf den ersten Blick über die Möglichkeiten der TEI hinauszugehen und einen Vorschlag zur Auszeichnung ambiger Strukturen zu machen. Bei näherem Hinsehen wurde jedoch deutlich, daß es dabei nicht um Ambiguität geht, die auf den Satzkontextes bezogen ist. Es werden vielmehr nur alle lexikalisch verzeichneten Lesarten eines Wortes gelistet und eine davon als die für den Satz relevante herausgestellt. Ambiguität auf syntaktischer Ebene wird dabei nicht berücksichtigt.

Diese Ergebnisse führten dazu, daß der linguistischen Annotation im TECA-Projekt die TEI-Richtlinien zugrundegelegt und projektbezogen modifiziert wurden. Die inhaltlichen Modifikationen wurden nach den Regeln der TEI durchgeführt und bestehen im Wesentlichen darin, daß zum einen nur eine Untermenge der von der TEI vorgeschlagenen Annotationen verwendet wird, die zum anderen durch die Möglichkeit, ambige Strukturen abzubilden, erweitert wurde. Darüber hinaus wurden die einzelnen Strukturelemente in der Regel spezifischer definiert als in der TEI und tragen damit den softwarespezifischen Ausgaben des Syntaxparsers Rechnung.

In technischer Hinsicht wurden die Regeln der TEI nicht eingehalten. Das sogenannte *pizza model* der TEI schreibt vor, daß zunächst die Rahmenstruktur (*Teig*) mit einer auf die Textgattung, beispielsweise Drama oder Lexikon, ausgerichteten Struktur (*Grundbelag*) ergänzt wird, auf die dann rein textbezogene Strukturen, wie linguistischen Annotationen (*zusätzlicher Belag*), angewandt werden können. In diesem Modell ist die linguistische Annotation quasi die dritte Schicht in einem eng miteinander verwobenen Verbund von Modulen.

Für das TECA-Projekt sollten nur die linguistischen Annotationen der TEI verwendet werden. Dazu mußten sie aus diesem Verbund herausgelöst und in die Struktur der Antworten auf offene Fragen eingebunden werden. Um das tun zu können, mußte das TEI-Modul für die linguistischen Annotationen formal soweit verändert werden, daß es als eigenständige Einheit fungieren kann. Diese Vorgehensweise weicht von den Benutzungsvorgaben der TEI ab, jedoch konnte im Rahmen dieser Projektstufe keine befriedigendere, d.h. vollständig TEI-konforme Lösung gefunden werden. Ein weiterführendes Projekt sollte eine erneute Auseinandersetzung mit diesem Aspekt vorsehen.

3.3 Modul 3a: Konventionelle Inhaltsanalyse (KIA)

Die KIA basiert auf der oben beschriebenen Struktur des Verschriftungsmoduls. Dabei werden die Projektdaten durch das verwendete Kategorienschema ergänzt, und in dem sich daran anschließenden Informationsblock wird nach dem Namen des Verschrifters der des Codierers eingefügt. Bei den daraufflogenden Fragebögen können jeder Antwort manuell beliebig vielen Kategorien aus dem Kategorienschema zugeordnet werden. Diese Modellierung der Struktur trägt der bisherigen Vorgehensweise bei ZUMA Rechnung. Wünschenswert wäre, automatisch zu kontrollieren, ob die manuell vergebenen Kategorien auch im Kategorienschema vorkommen. Wie dies mit den Möglichkeiten von SGML umzusetzen wäre ist unter → *3.1.3 Haupt-DTD 3a: Konventionelle Inhaltsanalyse (KIA)* beschrieben

3.4 Modul 3b: Computerunterstützte Inhaltsanalyse (CUI)

Die CUI basiert ebenfalls auf der oben beschriebenen Struktur des Verschriftungsmoduls. Dabei werden die Projektdaten durch einen Verweis auf die Datei ergänzt, in der sich das von der Textanalyse-Software verwendete Diktionär befindet. In dem sich daran anschließenden Informationsblock wird nach dem Namen des Verschrifters die Textanalyse-Software genannt, die, basierend auf dem Diktionär, Codes in die Antworttexte eingefügt hat.

3.5 Modul 4: Archivierung

Im Archivierungsmodul können alle oben beschriebenen Analyseansätze, die für eine Umfrage durchgeführt wurden, in beliebiger Kombination zusammengeführt und abgespeichert werden. Wie dies geschehen soll ist derzeit noch nicht spezifiziert, es sollte sich aber in jedem Fall um einen automatisierten Prozeß handeln.

4. SGML-Modellierung

4.1 Die Haupt-DTDs

Alle Haupt-DTDs, die den Modulen der inhaltlichen Beschreibung entsprechen, basieren auf *Subset 1: Rahmenstruktur für Antworten auf offene Fragen*.

4.1.1 Haupt-DTD 1: Verschriftung
Datei: vschrift.dtd

Bei der Verschriftung wird in der Rahmenstruktur als Metainformation (`%metainfo;`) der Name des Verschrifters genannt. Der Antworttext (`%m.antwort;`) ist nicht weiter strukturiert.

4.1.2 Haupt-DTD 2: Linguistische Annotation
Datei: lingx.dtd

Um die linguistischen Annotationen TEI-kompatibel zu gestaltetem, ist eine Erweiterung der Rahmenstruktur-DTD notwendig, die durch die Submodule 2 und 3 erreicht wird. In der Rahmenstruktur-DTD wird als Metainformation (`%metainfo;`) der Name des Verschrifters und die linguistische Analysesoftware genannt. Der Antworttext (`%m.antwort;`) wird nach den unter Submodul 2 beschriebenen morpho-syntaktischen Kategorien aufgegliedert.

4.1.3 Haupt-DTD 3a: Konventionelle Inhaltsanalyse (KIA)
Datei: kia.dtd

In der Rahmenstruktur-DTD wird als Lexikoninformation (`%lexikon;`) das verwendete Kategorienschema aufgeführt. Im Kategorienschema können die Kategorien einfach gelistet sein oder zu Gruppen zusammengefaßt werden. Zu jeder Kategorie wird ein eindeutiger Schlüssel vergeben, der diese logisch repräsentiert. Bei den Gruppenbenennungen ist die Nennung eines Schlüssels (Code) optional. Als Metainformation (`%metainfo;`) werden die Namen des Verschrifters und des Codierers genannt. Der Antworttext (`%m.antwort;`) wird nicht weiter unterstrukturiert und erhält als Attribut eine Kategorienliste (`%kategorien;`). Die Kategorien werden manuell vergeben und sind dem den Projektdaten zugeordneten Kategorienschema entnommen. Alle einer Antwort zugeordneten Kategorien bilden als Gesamtheit einen Attributwert. Würde jede Kategorie für sich einen Attributwert bilden, könnte so über den ID-IDREF-Mechanismus in SGML zusätzlich eine formale Kontrolle durchgeführt werden, die sicherstellen würde, daß nur Kategorien vergeben werden, die auch im Kategorienschema gelistet sind.

4.1. 4 Haupt-DTD 3b: Computerunterstützte Inhaltsanalyse (CUI)
Datei: cui.dtd

In der Rahmenstruktur-DTD wird als als Lexikoninformation (`%lexikon;`) auf die Datei verwiesen, in der sich das von der Textanalyse-Software verwendete Diktionär befindet. Als Metainformation (`%metainfo;`) wird dann der Namen des Verschrifters und die Textanalyse-Software genannt. Im Antworttext (`%m.antwort;`) werden

die von der Textanalyse-Software vergebenen, auf dem Diktionär basierenden Codes eingefügt.

4.1.5 Haupt-DTD 4: Archivierung
Datei: archiv.dtd

Im Archivierungsmodul können alle in Modul 2, 3a und 3b gemachten Analyseansätze in beliebiger Kombination zusammengeführt werden. Wie dies geschehen soll ist derzeit noch nicht spezifiziert, es sollte sich aber in jedem Fall um einen automatisierten Prozeß handeln.

4.2 Die DTD-Subsets

Die DTD-Subsets sind abhängige DTDs, die in die Haupt-DTDs eingebunden sind.

4.2.1 Subset 1: Rahmenstruktur für Antworten auf offene Fragen
Datei: rahmen.dtd

Die Rahmenstruktur-DTD definiert die allen Modulen zugrundeliegende Struktur. Dabei werden zunächst Projektdaten, Metainformationen und Fragebögen inhaltlich unterschieden. Auf der formalen Ebene wurden Entities mit Platzhalterfunktion geschaffen, die je nach Modul unterschiedlich gefüllt werden.

Projektdaten

Die Projektdaten sind:
- Projekttitel
- Umfragezeitraum
- Projektbetreuer
- Sprache der Umfrage
  Die gängigen Sprachen Deutsch, Schweizerdeutsch, Englisch und Französisch können explizit ausgewählt werden, sonstigen Sprachen wird eine summarische Klassifikation zugewiesen.
- Wortlaut der offenen Frage(n)
  Hier kann jede Frage als Frage, Unterfrage und Angabe typisiert werden und erhält eine eindeutige Identifikation, der die Funktion einer logischen Repräsentation des Fragetextes zukommt. Wegen der einfacheren Handhabung für den Verschrifter sind die Inhaltsmodelle als *mixed content* definiert.
- Lexikoninformationen
  Hier steht, je nachdem in welches Modul die Rahmenstruktur-DTD eingebunden ist, das Kategorienschema der konventionellen Inahaltsanalyse oder das Diktionär der computerunterstützten Inhaltsanalyse
- Allgemeine Informationen

Metainformationen

Über die Metainformationen ist ein Platzhalter geschaffen worden, der je nach Modul unterschiedlich ausgefüllt ist. Beispielweise wird hier der Name des Verschrifters, des Codierers oder der verwendeten Analyse-Software genannt.

Fragebögen

Unter Fragebögen ist die Gesamtheit der verschrifteten Antworten der Fragebögen des Projektes zu verstehen. Dabei wird im nächsten Schritt unterschieden, ob in der Umfrage eine oder mehrere offene Fragen vorkommen. So kann bei nur einer Frage die formale Referenz der Antwort auf die in den *Projektdaten* genannte Frage automatisch gesetzt werden, wohingegen bei mehreren Fragen die Antworten der jeweils zugehörigen Frage zugeordnet werden müssen. Jede Antwort kann, abhängig vom Modul, weiter strukturiert sein oder es können ihr Kategorien zugeordnet werden. Durch die Nennung der Fragebogennummer eindeutig identifizierbar gemacht, kann neben der Antwort bzw. den Antworten am Ende noch ein Memo mit möglichen Kommentaren und Anmerkungen zu den Antworten aus den verschiedenen Bearbeitungsstufen enthalten sein.

Platzhalter

Die sogenannten Platzhalter werden, je nachdem in welchem Modul die Rahmenstruktur-DTD eingebunden ist, unterschiedlich gefüllt. Die Platzhalter sind in der Rahmenstrukur als SGML-Entities angelegt, die dann im einzelnen Modul (wichtig: *vor* dem Einbinden der Rahmenstruktur-DTD) neu definiert werden können und somit die Entity der Rahmenstruktur-DTD überschreiben.

Entities mit Platzhalterfunktion sind:

- `%metainfo;`
  Metainformationen, die als Ergänzung zwischen den *Projektdaten* und den *Fragebögen* stehen. Diese Entity umfaßt den gesamten Block *Metainformationen.*
- `%lexikon;`
  Lexikoninformationen, die innerhalb der *Projektdaten* vorkommen.
- `%kategorien;`
  Attributliste zu einer Antwort, welche die der Antwort zugeordneten Kategorien aufnimmt. Die Antwort ist ein Element innerhalb der *Fragebögen*.
- `%m.antwort;`
  Inhaltsmodell einer Antwort, die Teil von *Fragebögen* ist.

4.2.2 Subset 2: TECA-spezifische Elemente und Attribute für die linguistische Analyse
Datei: tecaling.dtd

Die TECA-spezifischen Strukturen basieren auf der Text Encoding Initiative (TEI). Grundlage ist die `teiana2.dtd` vom 9. September 1994. Soweit sich die Elemente von der TEI-DTD ableiten sind sie englischsprachig. Dabei wurden neue Elemente hinzugefügt, einige Elemente wurden modifiziert, andere Elemente wurden gar nicht berücksichtigt.

Die neuen Elemente sind:

- Sentence rescued `<sResc>`
  Darunter sind Sätze zu verstehen, die nur unvollständig syntaktisch analysiert werden konnten.

- Ambiguouse structure <ambig>
  Ganze Sätze oder Phrasen, auf die mehr als eine morpho-syntaktische Analyse zutrifft, werden als ambige Strukturen markiert. Ambige Strukturen bestehen aus mehreren Parses.
- Parse <parse>
  Unter Parse ist *eine* morpho-syntaktische Interpretation innerhalb einer ambigen Struktur zu verstehen, die sich entweder auf einen ganzen Satz oder auf eine Phrase bezieht.

Die Modifikationen sind:
- Sentence <s>
  Das Inhaltsmodell wurde geändert und die Attributliste entfiel.
- Phrase <phr>
  Das Inhaltsmodell wurde geändert und die Attributliste wurde auf das Attribut ana reduziert, dem eine endliche Menge von Attributwerten zugeordnet wurde, aus der stets einer ausgewählt werden muß.
- Word <w>
  Das Inhaltsmodell wurde geändert und aus der Attributliste wurde nur das Attribut ana übernommen, dem eine endliche Menge von Attributwerten zugeordnet wurde, aus der stets einer ausgewählt werden muß. Die Attriubtliste wurde ergängzt durch zehn Lemma-Attribute, um Ambiguität auf der Lemma-Ebene sichtbar zu machen, die Gesamtzahl der genannten Lemmata und ein Typ-Attribut, um die Normalform eines Wortes von seinem Komparativ bzw. Superlativ zu unterscheiden. Das Typ-Attribut findet nur bei Adverbien und Adjektiven Anwendung.

Nicht berücksichtigt wurden:
- Clause <cl>
- Morpheme <m>
- Character <c>

4.2.3 Subset 3: Modifizierte Version der TEI-DTD teiana2.dtd
Datei: teiling.dtd

Um bei der morpho-syntaktischen Analyse TEI-kompatibel zu sein, müssen die TEI-Richtlinien bei der Zusammensetzung und Modifikation der TEI-Module berücksichtigt werden. Dannach muß diese TEI-DTD in die bestehende DTD des TECA-Projekts eingebunden werden. Dieser letzte Schritt, das Einbinden der TEI-DTD in die DTD des TECA-Projektes, warf die Frage auf, ob dies prinzipiell überhaupt möglich ist, oder ob die Schwierigkeiten softwarebedingt waren. Da dies im Rahmen dieser Projektstufe nicht geklärt werden kann, wurde eine pragmatische Lösung gewählt, die es ermöglicht, jetzt TEI-kompatibel zu sein, ohne die gesamte TEI-Struktur einzubinden, dies zu einem späteren Zeitpunkt tun zu können, ohne die jetztige DTD umstrukturieren zu müssen. Zu diesem Zweck wurde die teiana2.dtd aus dem TEI-Verbund einzeln herausgenommen und in die TECA-DTD integriert. Die Modifikationen an der teiana2.dtd wurden vorgenommen, um möglichst wenig andere TEI-DTDs, oder Teile daraus, mitintegrieren zu müssen.

5. Die SGML-Deklaration

Die für das TECA-Projekt verwendete SGML-Deklaration erlaubt Elementnamen, die bis zu 32 Zeichen lang sind.

6. Die Catalog-Datei

In der Catalog-Datei werden alle in den DTDs verwendeten Entity-Dateien und DTD-Module mit ihren PUBLIC- oder SYSTEM-Identifiers aufgelistet und mit den entsprechenden Dateinamen und -pfaden verknüpft. Die Catalog-Datei dient dazu, die SGML-DTDs zwischen den verschiedenen Computerplatformen einfach handhabbar zu machen.

7. Schlußbemerkungen

Im Projektverlauf haben sich drei Problemfelder herauskristallisiert:

1. Arbeitsprozesse

Dadurch, daß die Strukturierung der Antworten auf offene Fragen nicht nur an den inhaltlichen Anforderungen ausgerichtet war, sondern auch die bei ZUMA bislang üblichen Vorgehensweisen mitberücksichtigte, konnten nicht alle Möglichkeiten der SGML-basierten Strukturkontrolle ausgeschöpft werden. Dies wurde im Rahmen des Projektes bei der KIA deutlich, da hier ein Abgleich der manuell vergebenen Kategorien mit denen, die im Kategorienschema aufgelistetet sind so nicht möglich ist. Würden bei der Strukturierung konsequent nur die inhaltlichen Aspekte berücksichtigt, ergäben sich möglicherweise noch andere, bislang nicht bedachte Informationen aus der Struktur. Würde man beispielsweise bei der KIA festhalten, wieviele und eventuell auch welche Anhaltspunkte in der Antwort zur Vergabe einer Kategorie geführt haben, könnte man die einer Antwort zugeordneten Kategorien im Verhältnis zueinander gewichten.

2. Konformität mit schon bestehenden Richtlinien/Standards

Gerade bei der linguistischen Annotation gibt es eine Reihe von Projekten, auf die – zumindest auf den ersten Blick – zurückgegriffen werden konnte (TEI, EAGLES, CES). Dabei erwies sich allerdings, daß bei keiner dieser Arbeiten die Darstellung ambiger Strukturen vorgesehen war. Durch eine im Sinne der TEI vorgenommene Modifikation konnte die inhaltliche Konformität mit den TEI-Richtlinien sichergestellt werden. Die technische Einbindung des in der TEI unselbständigen Moduls für die linguistischen Annotationen mußte aus dem gesamten Verband herausgelöst und nicht-TEI-konform modifiziert werden, damit es im Sinne des TECA-Projekts eingesetzt werden konnte.

3. Software

Bei der für das Projekt ausgewählte SGML-Software der Firma Softquad Inc. erwies sich der SGML-Editor Author/Editor insofern als problematisch, als er Sonderzeichen in den Attributwerten nicht als Entity-Referenzen exportiert, sondern als Sonderzeichen stehen läßt. Daher muß ein nachgeschalteter Konvertierungsprozeß diese Aufgabe übernehmen. Diese Schwierigkeit besteht jedoch auch bei anderen in Augenschein genommenen SGML-Editoren. Ob es am Markt überhaupt eine Software gibt, die diesen Anforderungen genügt, bliebt zu prüfen.

Als Perspektive für das TECA-Projekt wären zwei Dinge anzustreben:

1. XML-Kompatibilität

Die vorliegenden SGML-DTDs sollten auf ihre XML-Kompatibilität geprüft und gegebenenfalls modifiziert werden, um die Textbasis künftig in einer internet-tauglichen Form zur Verfügung stellen zu können.

2. Textmodell des *Corpus Encoding Standards* (CES)

Die Adaption des CES-Modells für das TECA-Projekt, die im Moment noch daran scheitert, daß am Markt keine Software verfügbar ist, die die Handhabung dieses Textmodell unterstützt, sollte im Auge behalten werden. Damit könnte Annotationen ein flexibel ausbaubares Konzept zugrundegelegt werden, das auch in der Zukunft tragfähig bleibt.

Anhang

1. Zwischenbericht zu Arbeitspaket 4 Umsetzung in SGML-Format

Im Rahmen dieses Arbeitspaketes wurde in einem ersten Schritt eine SGML-Dokumenttypdefinition (DTD) für die Antworten auf offene Fragen erstellt. Diese Rahmenstruktur war als komplexe Archivstruktur gedacht, die nach der Verkodung der Antworten durch automatische Konvertierung erstellt wird. Inhaltlich wurden zunächst Projektdaten und Fragebögen unterschieden.

Projektdaten

Den Projektdaten sind dabei folgende Informationseinheiten zuzurechnen:

- Projekttitel
- Zeitpunkt der Umfrage
- Projektbetreuer
- Sprache der Umfrage
  Die gängigen Sprachen Deutsch, Schweizerdeutsch, Englisch und Französisch können explizit ausgewählt werden, sonstigen Sprachen wird eine summarische Klassifikation zugewiesen.
- Informationen zur statistischen Auswertung
  Darunter fallen die Studiennummer und das verwendete Statistikprogramm, ergänzt durch Angaben zum und dem vergebenen Dateinamen.
- Wortlaut der offenen Frage(n)
  Hier kann jede Frage als Frage, Unterfrage und Angabe typisiert werden und erhält eine eindeutige Identifikation, der die Funktion einer logischen Repräsentation des Fragetextes zukommt.
- Kategorienschema
  Es muß eindeutig Stellung bezogen werden, ob das Kategorienschema in die Projektdaten aufgenommen wird oder nicht. Wird es aufgenommen, so können die Kategorien einfach gelistet sein oder zu Gruppen zusammengefaßt werden. Zu jeder Kategorie wird ein eindeutiger Schlüssel vergeben, der diese logisch repräsentiert. Bei den Gruppenbenennungen ist die Nennung eines Schlüssels optional.
- Andere Daten
  In diesem Pufferelement können arbiträre projektspezifische Informationen eingegeben werden.

Fragebögen

Unter Fragebögen ist die Gesamtheit der verschrifteten Antworten der Fragebögen des Projektes zu verstehen. Jeder Fragebogen wird durch die Nennung der Fragebogennummer eindeutig identifizierbar. Ebenso wird jede Antwort über eine logische Referenz eindeutig auf die in den Projektdaten genannte zugehörige Frage bezogen. Darüber hinaus kann jeder Antwort sowohl das Ergebnis einer konventionellen Inhaltsanalyse (KIA) zugeordnet werden oder es können die Codes der computergestützten Inhaltsanalyse (CIA) integriert sein. In beiden Fällen ist eine eindeutige Verknüpfung der vergebenen Kategorien mit dem Kategorienschema gegeben.

Im Verlauf des Projekts haben sich Überlegungen herauskristallisiert, die auf eine Zerlegung der oben beschriebenen allumfassenden Rahmenstruktur in ein *modulares* Konzept zielen. Auslöser waren Tests, die ergaben, daß eine Verschriftung der Antworten in SGML nur eine leichte Verlangsamung der Erfassung nach sich zog und gleichzeitig von den Testpersonen als Unterstützung ihrer Arbeit angesehen wurde, da durch die Strukturführung des Editors formale Eingabefehler reduziert werden konnten. Für eine Verschriftung in SGML wäre die Komplexität der derzeitigen Rahmenstruktur nicht notwendig, sie fordert im Gegenteil Informationseinheiten ein, die bei diesem Arbeitsschritt gar nicht genannt werden können.

Es wäre also denkbar, die für die Verschriftung relevanten Teile in einer separaten DTD zusammenzufassen. Diese Basis-DTD könnte im nächsten Bearbeitungsschritt von einem zweiten *Layer* überlagert werden, so daß eine komplexere Struktur entstünde, die nun erlaubte, alle zur Verkodung notwendigen Informationen einzugeben. Dabei würden sich die Anforderungen von KIA und CUI in unterschiedlichen Strukturen niederschlagen, die in einer letzten Schicht zur Archivierung eine Vereinheitlichung erfahren müßten.

Darauf aufbauend werden die im Arbeitspaket 6 definierten linguistischen Informationen in die SGML eingebunden.

23. Juni 1998

2. Dokument-Typ-Definitionen (DTDs)

2.1 vschrift.dtd

```
<!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Modul 1: Verschriftung
               Rahmenstruktur fuer die Verschriftung offener Fragen

               Datei:     vschrift.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt

-->

<!-- ************************ modulspezifische Element-Entities
*************** -->

<!ENTITY  %  metainfo            "vschrifter," >


<!-- ************************ modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort           "(#PCDATA)" >


<!-- ************************ Einbindung der Rahmenstruktur
******************* -->

<!ENTITY  %  rahmen  SYSTEM  "rahmen.dtd" >
%rahmen;
```

2.2 rahmen.dtd

```
 <!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Subset 1:  Rahmenstruktur fuer Antworten auf offene Fragen

               Datei:     rahmen.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Alfons J. Geis
               Cornelia Zuell
               Ingrid Schmidt

-->

<!-- ********************** Alphabetische Zeichen westeuropaeischer
Sprachen * -->
<!ENTITY % ISOlat1 PUBLIC "ISO 8879-1986//ENTITIES Added Latin 1//EN"
"isolat1.ent">
%ISOlat1;

<!ENTITY % ISOlat2 PUBLIC "ISO 8879-1986//ENTITIES Added Latin 2//EN"
"isolat2.ent">
%ISOlat2;


<!-- ************************ modulspezifische Element-Entities
*************** -->

<!ENTITY  %  metainfo           "" >

<!ENTITY  %  lexikon            "" >


<!-- ************************ modulspezifische Attribut-Entities
*************** -->

<!ENTITY  %  kategorien         "" >


<!-- ************************ modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort          "" >


<!-- ************************ modulspezifische Deklarationen
****************** -->

<!-- VerschrifterIn -->
<!ELEMENT  vschrifter        - -  (#PCDATA) >

<!-- KodiererIn -->
<!ELEMENT  kodierer          - -  (#PCDATA) >
```

```
<!-- Textanalyse-Software -->
<!ELEMENT ta-software        - -  (#PCDATA) >

<!-- Linguistische-Analyse-Software -->
<!ELEMENT ling-software       - -  (#PCDATA) >


<!-- Kategorienschema -->
<!ELEMENT k-schema           - -  (kategorie+ | k-gruppe+) >

<!ELEMENT k-gruppe           - -  (name?, (kategorie+ | k-gruppe+)) >
<!ATTLIST k-gruppe           id-kgruppe  NUTOKEN                #IMPLIED >

<!ELEMENT name               - -  (#PCDATA) >

<!ELEMENT kategorie          - -  (#PCDATA) >
<!ATTLIST kategorie          schluessel  NUMBER                #REQUIRED >


<!-- Diktionaer -->
<!ELEMENT diktionaer         - -  (#PCDATA) >
<!ATTLIST diktionaer         datei   CDATA                 #REQUIRED >

<!-- zugeordneter Code -->
<!ELEMENT code               - o  EMPTY >
<!ATTLIST code               codenr   NUMBER               #REQUIRED >


<!-- ************************ Rahmenstruktur-Entities
************************ -->

<!ENTITY  % offeneFragen         "(projektdaten, %metainfo; frageboegen)" >

<!ENTITY  % projektdaten         "(titel, umfragezeitraum, betreuer+,
sprache,
                                   (fragetxt | unterfragetxt | angabetxt)+,
                                    %lexikon allginfo?)" >


<!-- ************************ Hauptstruktur
********************************** -->

<!ELEMENT offeneFragen     - -   %offeneFragen; >


<!-- ************************ Angaben zum Projekt
**************************** -->

<!ELEMENT projektdaten     - -  %projektdaten; >

<!-- Projekttitel -->
<!ELEMENT titel            - -  (#PCDATA) >

<!-- Umfragezeitraum -->
<!ELEMENT umfragezeitraum  - o  EMPTY >
<!ATTLIST umfragezeitraum  von  NUTOKEN                  #REQUIRED
                           bis  NUTOKEN                  #REQUIRED >

<!-- Betreuer des Projekts -->
<!ELEMENT betreuer         - -  (#PCDATA) >
```

```
<!-- Sprache -->
<!ELEMENT sprache          - o  EMPTY >
<!ATTLIST sprache          sprache (DE | CH | EN | FR | andere)   #REQUIRED
>

<!-- Wortlaut der offenen Frage -->
<!ELEMENT fragetxt         - -  (#PCDATA, (unterfragetxt* | angabetxt*)) >
<!ATTLIST fragetxt         id-frage       ID             #REQUIRED >

<!ELEMENT unterfragetxt    - -  (#PCDATA, angabetxt*) >
<!ATTLIST unterfragetxt    id-unterfrage  ID             #REQUIRED >

<!ELEMENT angabetxt        - -  (#PCDATA) >
<!ATTLIST angabetxt        id-angabe      ID             #REQUIRED >

<!-- darueberhinausgehende andere Informationen -->
<!ELEMENT allginfo          - -  (#PCDATA) >


<!-- ************************ Verschriftung der Frageboegen
****************** -->

<!ELEMENT frageboegen       - -  (eineFrage  | mehrereFragen) >

<!ELEMENT eineFrage         - -  (fragebogen+) >

<!ELEMENT fragebogen        - -  (antwort, memo?) >
<!ATTLIST fragebogen        ref-fragebogen  NUMBER        #REQUIRED >

<!ELEMENT antwort           - -  %m.antwort; >
<!ATTLIST antwort           ref-frage       IDREF         #CURRENT
                            %kategorien; >

<!ELEMENT mehrereFragen     - -  (fragebgen+) >

<!ELEMENT fragebgen         - -  (antwrt+, memo?) >
<!ATTLIST fragebgen         ref-fragebogen  NUMBER        #REQUIRED >

<!ELEMENT antwrt            - -  %m.antwort; >
<!ATTLIST antwrt            ref-frage       IDREF         #REQUIRED
                            %kategorien; >

<!ELEMENT memo              - -  (#PCDATA) >
```

2.3 kia.dtd

```
 <!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Modul 3a: Konventionelle Inhaltsanalyse (KIA)
               Rahmenstruktur fuer die Verschriftung offener Fragen
               mit konventioneller Kodierung der Antworten

               Datei:     kia.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt

-->


<!-- ************************ modulspezifische Element-Entities
*************** -->

<!ENTITY  %  metainfo        "vschrifter, kodierer," >

<!ENTITY  %  lexikon         "k-schema," >


<!-- ************************ modulspezifische Attribut-Entities
*************** -->

<!ENTITY  %  kategorien      "kategorien   NUMBERS      #IMPLIED">


<!-- ************************ modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort       "(#PCDATA)" >


<!-- ************************ Einbindung der Rahmenstruktur
******************* -->

<!ENTITY  %  rahmen  SYSTEM  "rahmen.dtd" >
%rahmen;
```

2.4 cui.dtd

```
 <!--
                Towards Extending Content Analysis (TECA)
                ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

                Modul 3b: Computergestuetzte Inhaltsanalyse (CUI)
                Rahmenstruktur fuer die Verschriftung offener Fragen
                mit computergestuetzter Kodierung der Antworten

                Datei:     cui.dtd
                Version:   1.0
                Datum:     28. Dezember 1998


                SGML-Arbeitsgruppe:
                Melina Alexa
                Ingrid Schmidt

-->

<!-- ************************ modulspezifische Element-Entities
*************** -->

<!ENTITY  %  metainfo        "vschrifter, ta-software," >

<!ENTITY  %  lexikon         "diktionaer," >


<!-- ************************ modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort       "(#PCDATA | code)+" >


<!-- ************************ Einbindung der Rahmenstruktur
******************* -->

<!ENTITY  %  rahmen  SYSTEM  "rahmen.dtd" >
%rahmen;
```

2.5 lingx.dtd

```
 <!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Modul 2: Linguistische Annotation
               Rahmenstruktur fuer die Verschriftung offener Fragen
               mit linguistisch annotierten Antworten

               Datei:     lingx.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt

-->


<!-- ************************* Entities
************************************** -->

<!ENTITY  %  metainfo        "vschrifter, ling-software," >


<!-- ************************* modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort       "(s|sResc)+" >


<!-- ************************* Einbindung linguistischer DTD-Module
************ -->

<!-- TECA specific linguistic categories -->
<!ENTITY  % TECALing    SYSTEM  "tecaling.dtd" >
%TECALing;

<!-- Linguistic segment categories (TEI, 15.1) -->
<!ENTITY  % TeiLing     SYSTEM  "teiling.dtd" >
%TeiLing;


<!-- ************************* Einbindung der Rahmenstruktur
******************* -->

<!ENTITY  %  rahmen  SYSTEM  "rahmen.dtd" >
%rahmen;
```

2.6 tecaling.dtd

```
<!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Subset 2:  TECA-spezifische Elemente und Attribute
                          fuer die linguistische Analyse

               Datei:     tecaling.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt
-->

<!-- TECA specific entities -->
<!-- Wortarten -->
<!ENTITY  %   verb            "verb|h-verb|k-verb|infv" >

<!ENTITY  %   nomen           "nomen" >

<!ENTITY  %   pronomen        "pron|fra-pron|int-pron|rel-pron" >

<!ENTITY  %   zahlen          "kza|oza" >

<!ENTITY  %   zeichen         "dot|wdash|sz" >

<!ENTITY  %   wortart-spez    "abk|date|name|soz" >

<!ENTITY  %   wortart         "%verb;|%nomen;|%pronomen;|adje|adv|art|konj|
                               praep|%zahlen;|%zeichen;|prfx|quantor|
                               %wortart-spez;" >

<!-- Phrasen -->
<!ENTITY  %   phrase-allg     "advp|appos|attr|gen-attr|h-verbp|inf|
                               int-satz|kompara|konj-satz|koord-ende|
                               k-verbp|loc-satz|neben-satz|np|pp|praed|
                               praepp|rel-satz|verbp|zahlp" >

<!ENTITY  %   phrase-spez
"abkd|datum|klammer|massang|namen|zitat|zitat-nf" >

<!ENTITY  %   phrase-atw      "Atw-Koord|Atw-Koord-Teil|atw-satz|satz" >

<!ENTITY  %   phrase          "%phrase-allg;|%phrase-spez;|%phrase-atw;" >


<!-- Entities with TECA extentions -->
<!-- Sentence -->
<!ENTITY  %  x.m.s     "(ambig|phr|w)+" >

<!-- Phrase -->
<!ENTITY  %  x.m.phr   "(ambig|phr|w)+" >
<!ENTITY  %  x.a.phr   "ana  (%phrase;)  #REQUIRED" >

<!-- Word -->
```

```
<!ENTITY  %  x.m.w     "(#PCDATA)" >
<!ENTITY  %  x.a.w     "ana (%wortart;)                          #REQUIRED
                        typ      (normal|komparativ|superlativ) #IMPLIED
                        lemma1         CDATA                    #REQUIRED
                        lemma2         CDATA                    #IMPLIED
                        lemma3         CDATA                    #IMPLIED
                        lemma4         CDATA                    #IMPLIED
                        lemma5         CDATA                    #IMPLIED
                        lemma6         CDATA                    #IMPLIED
                        lemma7         CDATA                    #IMPLIED
                        lemma8         CDATA                    #IMPLIED
                        lemma9         CDATA                    #IMPLIED
                        lemma10        CDATA                    #IMPLIED
                        lemma-anzahl   NUMBER                   #REQUIRED"
>

<!-- TECA-specific elements -->
<!ELEMENT  sResc         - -  (#PCDATA|phr|w)+>

<!ELEMENT  ambig          - -  (parse+) >

<!ELEMENT  parse          - -  (ambig|phr|w)+ >
```

2.7 teiling.dtd

```
<!--

               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Subset 3:  Modifizierte Version der TEI-DTD
                          teiana2.dtd

               Datei:     teiling.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt



      This is a modified version of:

      teiana2.dtd:  written by OddDTD 1994-09-09

      Modifications are:

      - extracting the Linguistic Segment Categories (15.1)
      - expanding the entities
      - deleting the marked sections
      - modifications of the content models (see there)
      - %m.phrase; which is part of %phrase.seq was reduced to
        %m.seg;

      Modifications are due to taking one "toping" out of the whole TEI
system,
      and only using the entities essential to our task.
-->

<!-- Content model entities were taken from teiclas2.ent -->
<!ENTITY % x.seg '' >
<!ENTITY % m.seg '%x.seg  anchor | c | cl | m | phr | s | seg | w' >

<!-- Attribute value entity was taken from teikey2.ent -->
<!ENTITY % INHERITED '#IMPLIED'  >


<!-- Attribute entities were taken from: teiana2.ent, teiclas2.ent -->
<!ENTITY % a.analysis 'ana       IDREFS      #IMPLIED' >

<!ENTITY % a.linking     ''                        -- not invoked --  >

<!ENTITY % a.terminology ''                        -- not invoked --  >

<!ENTITY % a.global      '%a.analysis;
                          %a.linking;
                          %a.terminology;
                          id       ID                 #IMPLIED
                          n        CDATA              #IMPLIED
```

```
                         lang     IDREF              %INHERITED
                         rend     CDATA              #IMPLIED' >

<!ENTITY % a.seg         'type     CDATA              #IMPLIED
                         function CDATA              #IMPLIED'  >


<!-- Entities created to prepare for TECA extentions -->
<!-- Sentence -->
<!ENTITY  %  x.m.s       '(#PCDATA |  %m.seg; )*    -(s)' >


<!-- Phrase -->
<!ENTITY  %  x.m.phr     '(#PCDATA |  %m.seg; )*' >

<!ENTITY  %  x.a.phr     '%a.global;
                          %a.seg;
                          TEIform        CDATA      "phr"' >
<!-- Word -->
<!ENTITY  %  x.m.w       '((#PCDATA | seg | w | m | c)*)' >

<!ENTITY  %  x.a.w       '%a.global;
                          %a.seg;
                          lemma           CDATA      #IMPLIED
                          TEIform         CDATA        "w"' >


<!-- 15.1:  Linguistic Segment Categories -->
<!ELEMENT s           - -   %x.m.s; >
<!ATTLIST s          %a.global;
                     %a.seg;
                     TEIform        CDATA      "s" >

<!ELEMENT cl          - -  (#PCDATA | %m.seg)* >
<!ATTLIST cl         %a.global;
                     %a.seg;
                     TEIform      CDATA         'cl'   >

<!ELEMENT phr         - -   %x.m.phr; >
<!ATTLIST phr        %x.a.phr; >

<!ELEMENT w           - -   %x.m.w; >
<!ATTLIST w          %x.a.w;>

<!ELEMENT m           - -  ((#PCDATA | seg | c)*)     >
<!ATTLIST m          %a.global;
                     %a.seg;
                     baseform      CDATA       #IMPLIED
                     TEIform       CDATA          'm' >

<!ELEMENT c           - -  (#PCDATA) >
<!ATTLIST c          %a.global;
                     %a.seg;
                     TEIform       CDATA      'c'  >



<!-- SEG and ANCHOR are taken from teilink2.dtd -->

<!-- %m.inter; was deletet in the content model of SEG -->
<!ELEMENT seg         - -  (#PCDATA | %m.seg)*  >
<!ATTLIST seg        %a.global;
                     %a.seg;
```

```
                    subtype     CDATA               #IMPLIED
                    part        (Y | N | I | M | F) N
                    TEIform     CDATA               'seg'   >


<!ELEMENT anchor    - O  EMPTY  >
<!ATTLIST anchor    n       CDATA         #IMPLIED
                    lang     IDREF        %INHERITED;
                    rend     CDATA        #IMPLIED
                    %a.seg;
                    id        ID          #REQUIRED
                    TEIform   CDATA       'anchor'  >
```

2.8 archiv.dtd

```
<!--
               Towards Extending Content Analysis (TECA)
               ZUMA, Arbeitsgruppe Sozialwissenschaftliche Textanalyse

               Modul 4: Archivierung
               Rahmenstruktur fuer die Verschriftung offener Fragen
               mit allen vorgenommenen Kodierungen

               Datei:     archiv.dtd
               Version:   1.0
               Datum:     28. Dezember 1998


               SGML-Arbeitsgruppe:
               Melina Alexa
               Ingrid Schmidt

-->

<!-- ************************ Aenderungen gegenueber tecaling.dtd
*************** -->

<!-- Entities with TECA extentions -->
<!-- Sentence -->
<!ENTITY  %  x.m.s     '(ambig|phr|w|code)+' >

<!-- Phrase -->
<!ENTITY  %  x.m.phr   '(ambig|phr|w|code)+' >

<!-- Word -->
<!ENTITY  %  x.m.w     '(#PCDATA|code)+' >


<!-- ***********  *********** Einbindung linguistischer DTD-Module
************* -->

<!-- TECA specific linguistic categories -->
<!ENTITY  % TECALing    SYSTEM  "tecaling.dtd" >
%TECALing;

<!-- Linguistic segment categories (TEI, 15.1) -->
<!ENTITY  % TeiLing     SYSTEM  "teiling.dtd" >
%TeiLing;

<!-- ************************ modulspezifische Element-Entities
*************** -->

<!ENTITY  %  metainfo        "vschrifter, ling-software?, kodierer?, ta-
software?," >

<!ENTITY  %  lexikon         "k-schema?, diktionaer?," >


<!-- ************************ modulspezifische Attribut-Entities
*************** -->

<!ENTITY  %  kategorien      "kategorien   NUMBERS      #IMPLIED">
```

```
<!-- ************************ modulspezifische Inhaltsmodelle
***************** -->

<!ENTITY  %  m.antwort       "(#PCDATA|code|s|sResc)+" >


<!-- ************************ Einbindung der Rahmenstruktur
****************** -->

<!ENTITY  %  rahmen  SYSTEM  "rahmen.dtd" >
%rahmen;
```

3. SGML-Deklaration

```
<!SGML"ISO 8879:1986"
      CHARSET
            BASESET     "ISO 646-1983//CHARSET International Reference
Version (IRV)//ESC 2/5 4/0"
            DESCSET       0   9 UNUSED
                          9   2 9
                         11   2 UNUSED
                         13   1 13
                         14  18 UNUSED
                         32  95 32
                        127 129 UNUSED
      CAPACITY    SGMLREF
                  TOTALCAP    60000
      SCOPE       DOCUMENT
      SYNTAX
            SHUNCHAR NONE
            BASESET "ISO 646-1983//CHARSET International Reference Version
(IRV)//ESC 2/5 4/0"
            DESCSET
                  0  128   0
            FUNCTION
                  RE            13
                  RS            10
                  SPACE         32
                  TAB SEPCHAR    9
            NAMING
                  LCNMSTRT ""
                  UCNMSTRT ""
                  LCNMCHAR "-."
                  UCNMCHAR "-."
                  NAMECASE    GENERAL YES
                              ENTITY  NO
            DELIM
                  GENERAL           SGMLREF
                  SHORTREF    SGMLREF
            NAMES SGMLREF
            QUANTITY SGMLREF
                  NAMELEN           32
                  LITLEN      2000
                  ATTCNT      256
                  GRPCNT      128
      FEATURES
            MINIMIZE
                  DATATAG           NO
                  OMITTAG           NO
                  RANK        NO
                  SHORTTAG    YES
            LINK
                  SIMPLE      NO
                  IMPLICIT    NO
                  EXPLICIT    NO
            OTHER
                  CONCUR      NO
                  SUBDOC      YES   99999999
                  FORMAL      YES
      APPINFO           NONE
>
```

4. Catalog-Datei

```
-- catalog:  Towards Extending Content Analysis (TECA) ZUMA, Arbeitsgruppe
Sozialwissenschaftliche Textanalyse / 28-12-98 --

-- Public identifiers --

PUBLIC "ISO 8879-1986//ENTITIES Added Latin 1//EN"
"c:\ae\dtds\entities\iso-lat1.ent"
PUBLIC "ISO 8879-1986//ENTITIES Added Latin 2//EN"
c:\ae\dtds\entities\iso-lat2.ent"


-- System identifiers --

SYSTEM "rahmen.dtd"   "c:\zuma\dtds\rahmen.dtd"
SYSTEM "tecaling.dtd" "c:\zuma\dtds\tecaling.dtd"
SYSTEM "teiling.dtd"  "c:\zuma\dtds\teiling.dtd"
```

5. Quick Reference

Element
Attribut
Attributwert

---

allginfo	allgemeine Information
ambig	linguistisch ambige Struktur (Satz oder Phrase)
angabetxt	Text der Angabe
id-angabe	eindeutige Identifikation der Angabe
antwort	Text der Antwort (bei genau einer offenen Frage in der Umfrage)
ref-frage	Referenz auf die zugehörige Frage
kategorien	Kategorien, die von der KIA zugeordnet werden
antwrt	Text der Antwort (bei mehreren offenen Fragen in der Umfrage)
ref-frage	Referenz auf die zugehörige Frage
kategorien	Kategorien, die von der KIA zugeordnet werden
betreuer	Betreuer der Verschriftung und Codierung der Fragebögen
code	Code, der von der CUI zugeordnet wird
codenr	Codenummer aus dem für die CUI relevanten Diktionär
diktionaer	Diktionär, das der CUI zugrundegelegt ist
datei	Name der Diktionär-Datei
eineFrage	Rahmenelement, wenn die Umfrage nur eine offene Frage beinhaltet
frageboegen	Rahmenelement für die Gesamtheit aller Fragebögen
fragebgen	Fragebogen (bei mehreren offenen Frage in der Umfrage)
ref-fragebogen	Referenz auf die Fragebogennummer des Fragebogens
fragebogen	Fragebogen (bei genau einer offenen Frage in der Umfrage)
ref-fragebogen	Referenz auf die Fragebogennummer des Fragebogens
fragetxt	Text der Frage
id-frage	eindeutige Identifikation der Frage

Element
Attribut
Attributwert

kategorie	Kategorie im Kategorienschema (KIA)
schluessel	numerische Bezeichnung der Kategorie
k-gruppe	Bezeichnung für eine Gruppe von Kategorien (KIA)
id-kgruppe	eindeutige Identifikation dieser Kategoriengruppe
kodierer	Name des Kodierers
k-schema	Kategorienschema
ling-software	Name der linguistischen Analyse-Software
mehrereFragen	Rahmenelement, wenn die Umfrage mehrere offene Fragen beinhaltet
memo	Memofeld für alle Arten von Kommentaren, Notizen etc. zu einer Antwort
offeneFragen	oberstes Element, das die Projektdaten, den Verschrifter, Kodierer sowie die Namen der verwendeten Software und die Gesamtheit der Fragebögen umfaßt
parse	eine Lesart einer linguistisch ambigen Struktur
phr	Phrase
ana	Annotation
abkd	Abkürzung mit Punkt
advp	Adverbpharase
appos	Apposition
attr	adjektivische Attributphrase
atw-koord	für die Antworttexte spezifische Art der Koordination
atw-koord-teil	die erste Koordinante in der `atw-koord`
atw-satz	unvollständiger Satz, spezifisch für die Antworttexte
datum	möglicherweise komplexe Datumsangabe
gen-attr	Genitivattribut
h-verbp	ein Hilfsverb oder eine Folge von Hilfsverben
inf	einfache oder komplexe Infinitivkonstruktion
int-satz	durch Fragepronomen eingeleiteter Nebensatz mit Verbendstellung
klammer	mit Klammern umgebener Teil einer Antwort
kompara	mit *als* oder *wie* eingeleitete Phrase, die einen Vergleich ausdrückt
konj-satz	mit subordinierender Konjunktion eingeleiteter Nebensatz mit Verbendstellung

Element Attribut Attributwert	
phr	Phrase (Forts.)
ana	Annotation
koord-ende	die mit koordinierender Konjunktion eingeleitete letzte Koordinante in einer Koordination
k-verbp	ein Hilfsverb oder eine Folge von Kopulaverben
loc-satz	durch lokatives Fragepronomen eingeleiteter Nebensatz mit Verbendstellung
massang	Maßangabe
namen	möglicherweise komplexe Namen
neben-satz	Nebensatz, außer Relativ-, Lokal- und Interrogativsätzen sowie Infinitive
np	Nominalphrase
pp	Präpositionalphrase
praep	Verben als Satzprädikat
praepp	mehrere Präpositionen am Anfang einer Präpositionalphrase
rel-satz	Relativsatz
satz	grammatikalisch vollständiger und unvollständiger Satz
verbp	komplexe Gruppe von Vollverben
zahlp	einfache Ordinalzahlen oder komplexere Gruppe von Zahlen
zitat	mit Anführungszeichen gekennzeichnetes Zitat
zitat-nf	Sätze, deren Subjekt oder Objekt eine direkte Rede, d.h. ein mit Anführungszeichen gekennzeichnetes Zitat sind
projektdaten	Projektdaten
s	Satz
sprache	Sprache der Umfrage
sprache	Abkürzung der Umfragesprache
DE	deutsch
CH	schweizerdeutsch
EN	englisch
FR	französisch
andere	andere Sprache, als eine der oben genannten
sResc	unvollständig analysierter Satz (sentence rescued)
ta-software	Name der Textanalyse-Software
titel	Titel der Umfrage

Element
Attribut
Attributwert

umfragezeitraum	Zeitraum der Umfage
von	Beginn des Zeitraums, z.B. Monat und Jahr
bis	Ende des Umfragezeitraums, z.B. Datum
unterfragetxt	Text der Unterfrage
id-unterfrage	eindeutige Identifikation der Unterfrage
vschrifter	Name des Verschrifters
w	Wort
ana	Annotation
abk	Abkürzung, die möglicherweise einen Abkürzungspunkt erwartet
adje	Adjektiv
adv	Adverb
art	Artikelwort
date	Jahres- und Datumsangabe
dot	Ordinalzahl- und Abkürzungspunkt
fra-pron	Fragepronomen
h-verb	Hilfsverb
infv	Verb im Infinitiv
int-pron	Fragepronomen
konj	Konjunktion
k-verb	Kopulaverb
kza	Kardinalzahl
name	Eigenname, Namensbestandteile(de, von etc.) und unbekannte Wörter
nomen	Nomen
oza	Ordinalzahl
praep	Präposition
prfx	von einem Verb abgetrennntes Präfix
pron	Personalpronomen und pronominal verwendete Artikelwörter
quantor	Quantor
rel-pron	Relativpronomen
soz	Sonderzeichen
sz	Satzzeichen
verb	Verb
wdash	"halbe" Wörter mit Bindestrich

Element	Attribut	Attributwert	
w			Wort (Forts.)
	typ		Typ der Vergleichsform
		positiv	Positiv (Grundstufe, gleicher Grad)
		komparativ	Komparativ (Hoherstufe, ungleicher Grad)
		superlativ	Superlativ (Höchststufe, höchster Grad)
		elativ	Elativ (absoluter Superlativ, sehr hoher Grad)
	lemma1		erstes Lemma
	lemma2		zweites Lemma
	lemma3		drittes Lemma
	lemma4		viertes Lemma
	lemma5		fünftes Lemma
	lemma6		sechstes Lemma
	lemma7		siebtes Lemma
	lemma8		achtes Lemma
	lemma9		neuntes Lemma
	lemma10		zehntes Lemma
	lemma-anzahl		Gesamtzahl der Lemmata

6. Das DTD-Modulsystem im TECA-Projekt – Benutzungsanleitung

Modul 1: Verschriftung

1. Der Verschriftung der offenen Fragen wird die `vschrift.dtd` zugrundegelegt. Sie gliedert sich in die drei Teile: Projektdaten, Verschrifter und Fragebögen.

   Author/Editor 3.5:
   Die vom Betreuer vorbereitete Datei wird im Author/Editor geöffnet (`File → Open`). Die Projektdaten sind bereits ausgefüllt. Nach den Projektdaten muß der Name der Verschrifterin eingetragen werden und darauf folgen die Fragebögen.

2. Nach der Verschriftung der Fragebögen, vor dem endgültigen Abspeichern, muß geprüft werden, ob die SGML-Auszeichnungen korrekt und vollständig eingebracht wurden. Je nach verwendetem SGML-Editor wird die Datei zusätzlich im SGML-Format exportiert.

   Author/Editor 3.5:
   Die Datei muß zunächst validiert werden (`Special → Validate Document`), um die Richtigkeit und Vollständigkeit der SGML-Auszeichnungen zu prüfen. Ist die Datei korrekt, so erscheint in einem Fenster die Meldung "Validation has completed. The file has been successfully validated." Diese Meldung kann mit "OK" weggeklickt werden. Danach wird die Datei gespeichert (`File→ Save`), die automatisch die Dateierweiterung `.ae` erhält. Anschließend wird sie im SGML-Format exportiert (`File→ Export`); dabei wird die Dateierweiterung `.sgm` vergeben. Nur die SGML-Datei mit der Dateierweiterung `.sgm` kann mit anderen Programm weiterbearbeitet werden.

Modul 2: Linguistische Annotation

Die Linguistische Annotation wird von einem Analyseprogramm durchgeführt. Ausgangsbasis ist die SGML-Struktur der Verschriftung (Modul 1). Nach der Analyse wird die Richtigkeit und Vollständigkeit der SGML-Auszeichnungen gegen die TEI-konforme `lingx.dtd` geprüft.

Modul 3a: Konventionelle Inhaltsanalyse (KIA)

1. Die im SGML-Format verschrifteten Fragebögen werden in einen SGML-Editor geladen und dort mit der `kia.dtd` weiter bearbeitet.

   Author/Editor 3.5:
   (<u>Wichtig:</u> Der Pfad darf nicht so gesetzt sein (`Special → Options: Extentions/Paths`), daß Author/Editor automatisch die zugehörigen Styles findet.)
   Für die KIA können die im Author/Editor verschrifteten Fragebögen auf zwei verschiedene Arten im Author/Editor aufgerufen werden.

   1. Möglichkeit:
   Die Datei mit der Dateierweiterung `.ae` wird geöffnet (`File → Open`). Ein

Fenster erscheint, das dazu auffordert, die gewünschte Rules-Datei auszuwählen. Das Verzeichnis, in dem sich die Regeln für die KIA befinden, kann durch Anklicken ausgewählt werden, ebenso die Datei selbst (`kia.rls`). Es erscheint ein weiteres Fenster mit der Meldung "The rules file has been changed. Do you want to convert the document to the new rules?" Das Fenster wird mit "Convert to new rules" weggeklickt.

2. Möglichkeit:
Die Datei mit der Dateierweiterung `.sgm` wird importiert (`File → Import`). Ein Fenster erscheint, das dazu auffordert, die gewünschte Rules-Datei auszuwählen. Das weitere Vorgehen ist entsprechend dem unter der 1. Möglichkeit beschriebenen. Jedoch erscheint hier am Ende zusätzlich die Meldung "Rules checking cannot be turned on: An invalid element, 'FRAGEBOEGEN', was found." Diese Meldung kann mit "OK" weggeklickt werden.

2. Dazu müssen zunächst die für die KIA notwendigen zusätzlichen Elemente eingefügt werden:
   das Kategorienschema (`k-schema`) am Ende der Projektdaten (`projektdaten`) bzw. vor der Allgemeinen Info (`allginfo`) und die Kodiererin (`kodierer`) muß nach der Verschrifterin (`vschrifter`) genannt werden.

   Author/Editor 3.5:
   Der Cursor wird vor `</projektdaten>` bzw. vor `<allginfo>` plaziert. Durch `F7` wird eine Liste der möglichen Elemente angezeigt, die, je nachdem ob eine Allgemeine Info bereits vorhanden ist oder nicht, nur das Kategorienschema-Element (`k-schema`) oder eine Liste von Elementen anzeigt, unter denen sich das Kategorienschema befindet. Das Kategorienschema-Element wird ausgewählt und das Kategorienschema mit der vorgegebenen SGML-Struktur eingegeben. Dannach wird der Cursor vor `<vschrifter>` plaziert und wie oben beschrieben wird das Element für den Kodierer ausgewählt und der Name des Kodierers eingefüllt. Wurde die Verschriftungsdatei importiert (Punkt 2, 2. Möglichkeit), so müssen jetzt noch die Regeln eingeschaltet werden (`Special → Turn Rules Checking On`).

   Die Datei sollte nun unter neuem Namen gespeichert werden (`File→ Save As`).

3. Die konventionelle Inhaltsanalyse kann nun nach dem zugehörigen Kategorienschema vorgenommen werden. Die für jede Antwort zutreffenden Kategorien werden als Attribute angegeben. Dabei ist es nicht zwingend erforderlich, daß jede Antwort Kategorien zugeordnet bekommt.

   Author/Editor 3.5:
   Um den Antworten Kategorien aus dem Kategorienschema zuzuordnen, wird der Cursor vor den zu bearbeitenden Antworttext plaziert. Mit der Taste `F6` wird die Attributliste zur Antwort angezeigt. Im Kategorien-Attribut werden eine oder mehrere Kategorien eingetragen, gegebenenfalls durch einen Leerschritt voneinander getrennt.

4. Nach Abschluß der konventionellen Inhaltsanalyse wird das Dokument auf Richtigkeit und Vollständigkeit der SGML-Auszeichnungen geprüft, abgespeichert und, je nach verwendetem Editor, zusätzlich im SGML-Format exportiert.

   Author/Editor 3.5:
   Die Vorgehensweise entspricht der unter Modul1, Punkt 2 beschriebenen.

Modul 3b: Computergestützte Inhaltsanalyse (CUI)

1. Erfolgt die Inhaltsanalyse der Antworttexte computergestützt, so wird die `cui.dtd` zugrunde gelegt.

2. Dazu müssen zunächst die für die CUI notwendigen zusätzlichen Elemente eingefügt werden:
   das Diktionär (`dictionaer`) am Ende der Projektdaten (`projektdaten`) bzw. vor der Allgemeinen Info (`allginfo`) und der Name der verwendeten Textanalyse-Software (`ta-software`) muß nach der Verschrifterin (`vschrifter`) genannt werden.

3. Das Diktonär, auf dem die CUI beruht, steht in einer separaten Datei, die als Attributwert zum Datei-Attribut angegeben wird. Die Kategorien-Codes (`code`) werden hinter das Wort bzw. die Phrase gestellt, die für die Zuordung der Kategorie ausschlaggebend ist. Der eigentliche Code wird als Wert des Codenr-Attributs eingetragen.

4. Nach Abschluß der computergestützten Inhaltsanalyse wird das Dokument auf Richtigkeit und Vollständigkeit der SGML-Auszeichnungen geprüft, abgespeichert und, je nach verwendetem Editor, zusätzlich im SGML-Format exportiert.

Author/Editor 3.5:
Soll die der CUI zugrundegelegte Struktur manuell getestet werden, so ist die Vorgehensweise entsprechend der für die KIA beschriebenen.

Modul 4: Archivierung

Im Archivierungsmodul können alle in Modul 2, 3a und 3b gemachten Analyseansätze in beliebiger Kombination zusammengeführt werden. Wie dies geschehen soll ist derzeit noch nicht spezifiziert, es sollte sich aber in jedem Fall um einen automatisierten Prozeß handeln.