"KODIERUNG UND NORMUNG MASCHINENLESBARER TEXTE" - BERICHT AUS DEM GLDV-ARBEITSKREIS

Peter Scherber
Göttingen

Zum Arbeitskreis

Dieser Arbeitskreis wurde im Herbst 1991 auf der GLDV-Tagung in Trier gegründet. Er verstand sich als ein lokales deutschsprachiges Forum für die weltweite Text Encoding Initiative (TEI), die damals gerade den ersten Teil ihrer Arbeit mit der Herausgabe von "Guidelines" abgeschlossen hatte1. Da die zweite Phase des TEI-Projekts zeitlich sehr knapp kalkuliert worden war, - man plante, im Sommer 1992 sowohl die 2. Projektphase (P2) zum Abschluß zu bringen und direkt anschließend daran die endgültigen Guidelines als P3 zu veröffentlichen - glaubten wir damals, der Arbeitskreis werde schon sehr bald vor allem mit der Evaluierung und praktischen Ausfüllung dieses neuen Normwerkes vollauf zu tun haben. Doch es ist dann anders gekommen.

Zwar haben wir von Anfang an die einzelnen Publikationen der P2-Phase mitgehalten und kurz nach Erscheinen auch auf dem Göttinger Listserver bereitgestellt, doch schwand auch mit dem geduldigen Warten auf die einzelnen Teile des Normenwerks die Bereitschaft zu ungeduldiger Aktivität im Arbeitskreis. Hinzu kam, daß auch die Motivation der AK - Teilnehmer, das angebotene elektronische Diskussionsforum zu nutzen, äußerst gering war. Man kann das Interesse der Teilnehmer am Arbeitskreis sicher ganz zutreffend als überwiegend rezeptiv charakterisieren, d. h. es besteht zwar ein breites und intensives Interesse daran, diese Entwicklung im Auge zu behalten, aber bevor die Phase P2 abgeschlossen ist, wartet man ab bzw. widmet man sich den Dingen des wissenschaftlichen Alltags, die keinen Aufschub gestatten.

Der Zeitverzug um mittlerweile eineinhalb Jahre und auch das Anwachsen der ursprünglich noch übersichtlichen Guidelines auf das mittlerweile zu erwartende Achtfache ihres Umfanges in der Phase P2 hat uns bewogen, Zielsetzung und Aufgaben des Arbeitskreises neu zu überdenken.

Nachdem auf der GLDV-Tagung in Kiel im März 1993 das Interesse am Arbeitskreis erneut bestätigt worden ist, trafen wir uns am 22. Oktober in Göttingen zu einem relativ kleinen Treffen, bei dem sich zeigte, daß das Vorhaben der TEI mittlerweile auch durchaus skeptisch gesehen werden muß. Hierzu soll weiter unten noch einiges berichtet werden.

Was ist die Text Encoding Initiative?

Seit 1987 haben sich in der Text Encoding Initiative, an der auch mehrere

1 Guidelines For the Encoding and Interchange of Machine-readable Texts TEI PI, hg. von C. M. Sperberg-McQueen u. Lou Burnard, Chicago u. Oxford 1990/91,289 S.; Träger der 1987 gegründeten Initiative waren die drei Organisationen ACH, ACL und ALLC.

deutsche Wissenschaftler beteiligt waren, zahlreiche Kommissionen mit der Erstellung von Kodierrichtlinien (Guidelines) beschäftigt. Dabei ist schon in einer sehr frühen Phase die Entscheidung gefallen, die bestehende ISO-Norm 8879 (SGMLY zur Grundlage der Verhandlungen zu machen. Daraus resultierte eine zweigeteilte Aufgabenstellung. Es mußten einerseits Kodierrichtlinien entwickelt werden, die es gestatteten, möglichst jede vorhandene Textsorte kontrolliert in maschinenlesbare Form zu überführen, und es waren andererseits SGML-konforme Dokument-Typdefinitionen (die sogenannten DTDs) zu entwickeln, die einen weltweiten Austausch derartiger Dokumente zu garantieren im Stande waren.

Zweckbestimmung dieser Richtlinien waren neben dem Austausch und damit der möglichen Wiederverwertung einmal erfaßter Ressourcen außerdem die Unterstützung von applikations-unabhängiger Dokumenterstellung und eine Auszeichnung der Texte nach standardisierten Regeln, die es ermöglichen sollten, so gut wie alle relevanten Textmerkmale der Analyse zugänglich zu machen 3.

Es hat sich gezeigt, daß die Guidelines von 1990/91 auch quantitativ nur ein erster Schritt waren. Die Phase P2, die noch andauert, hat sowohl den ersten Entwurf der Guidelines, als auch die für die erste Phase gültigen DTDs einer gründlichen Revision unterzogen. Der Umfang aller 42 Kapitel von P2 wird voraussichtlich einen Umfang von über 2000 Seiten besitzen.

2 International Organization for Standardization, ISO 8879: Information processing - Text and office systems - Standard Generalized Markup Language (SGML), 1986.

3 ausführlicher dazu vgl. Winfried Lenders: Fragen der Standardisierung, in: Computereinsatz in der Angewandten Linguistik, hg. v. W. Lenders, Frankfurt u. a. 1993, S. 63-74.

Derzeitiger Stand des P2-Projekts

In der definitiven und expliziten Form eines P2-Dokuments4 sind bislang nur die ersten beiden (von acht) Teilen erschienen. Der einleitende Teil I enthält eine kompakte Einführung in die SGML-Philosophie und eine Beschreibung der TEI-DTDs. Teil II (Core Tags and General Rules) äußert sich zu Zeichen und Zeichensätzen, beschreibt den TEI-Reader, in dem Informationen über das erfaßte Dokument mitgeführt werden können und enthält diejenigen Kodierelemente (Tags), die allen TEI-konformen Texten gemeinsam sein sollen. In diesem Zusammenhang wird eine Standard- Textstruktur konstituiert, die sozusagen die gemeinsamen Merkmale aller zu erfassenden Textsorten enthält.

Die Teile III und VI, die aus den Kodierrichtlinien (Tag sets) für alle textsortenspezifischen Merkmale bestehen, sind noch immer lückenhaft. Bei den Teilen V bis VIII sind bislang erst vier von 14 Kapiteln erschienen.

Lou Burnard, der Mitherausgeber der Guidelines hat im Juli 1993 festgestellt, daß damals ca. 25 % des für P2 vorgesehenen Materials erschienen sei, dies läßt den Schluß zu, daß wir heute, d. h. Ende 1993 ca. 40 - 50 % erreicht haben, und uns wohl noch auf eine längere Durststrecke einzustellen haben.

Das Treffen des Arbeitskreises am 22. Oktober

Am 22. Oktober 1993 trafen sich in Göttingen 9 Teilnehmer (incl. der beiden Moderatoren) und diskutierten die aktuelle Situation des Arbeitskreises. Nach Berichten der beiden Moderatoren zu Erfahrungen mit zwei SGML-Parsern *(G.* Koch) und zum Stand des P2-Projekts (P. Scherber)

4 Die Phase 3 (P3) läßt, im Gegensatz zu P2, nur mehr geringfügige redaktionelle Veränderungen des Normenwerkes zu, so daß man davon ausgehen kann, daß P2 tatsächlich den definitiven Status der Guidelines widerspiegelt.

referierte aus seiner Arbeit Martin Volk (Universität Koblenz-Landau), der ein Korpus von deutschen Sätzen (zum Einsatz für Zwecke sowohl in der Forschung als auch in der Lehre) mit SGML-Markierung aufgebaut hat. Bruno Schulze aus Stuttgart (Institut für maschinelle Sprachverarbeitung, Projekt Textkorpora und Erschließungswerkzeuge) berichtete von den dortigen Arbeiten zu Feature-Strukturen. Bei beiden Referenten wurde offensichtlich, daß die Arbeit der TEI, wenn sie überhaupt "ankommen" will und nicht schon zu spät kommt, noch eine geraume Weile als konkurrierendes System gesehen werden muß, das sich im Wettbewerb mit anderen Taggingsystemen erst noch bewähren muß. Insbesondere der bisherige erfolgreiche Einsatz SGML-konformer, speziell auf eine bestimmte Anwendung optimierter Systeme, die sich in der Praxis bewährt haben, wird eine skeptische und abwartende Haltung gegenüber dem TEI-Normenwerk zur Folge haben.

Dies war auch das Fazit der anschließenden Diskussion, die sich mit der Thematik eines nach dem Abschluß der P2-Phase vorgesehenen Status-Workshops auseinandersetzte. Es wurde festgestellt, daß ein derartiges Unternehmen sich unter dem Arbeitstitel: "Standardisierung bei der Erfassung maschinenlesbarer Texte" vor allem drei Aufgaben widmen sollte

1. dem Für und Wider der TEI-Guidelines als Kodiervorschrift

2. Erfahrungsberichten von Anwendern

3. den Softwareinstrumenten (Tools, Parser, Editoren).

Was ist zu tun, wie geht es weiter

Auch in den mehr als zwei Monaten seit Oktober ist der Stand des P2-Projekts noch fast unverändert geblieben[5]. Hinzu traten organisatorische Probleme, mit denen die beiden Moderatoren des Arbeitskreises konfrontiert wurden und die es erforderlich machten, den auf dem Treffen diskutierten Termin eines Status-Workshop zu den TEI-Guidelines im Mai 1994 vorerst fallen zu lassen.

In den nächsten Wochen werden wir zu allererst die Distribution der TEI-Dokumente auf eine neue Basis stellen. Dem Trend der Zeit folgend werden wir die Diskussionsliste auf anderer Plattform, aber mit demselben bewährten Namen , 'MARKUP-L" weiterführen und die Verteilung der Dokumente und DTD-Dateien über FTP vornehmen. Wir hoffen, daß diese Arbeiten Ende Januar 1994 abgeschlossen sein werden.

Erfahrungen und Kenntnisse über Softwareprodukte (Parser, Editoren, Tools usw.), die die Arbeit mit SGML unterstützen, sind immer noch sehr dünn gesät, dies liegt nicht zuletzt auch daran, daß für die meisten dieser Produkte nur kommerziell kalkulierte Preise verlangt werden. Aus diesem Grunde möchten wir an dieser Stelle an diejenigen appellieren, die praktische Erfahrung mit derartigen SGML-Produkten haben, sich uns anzuschließen und/oder auf einem der nächsten AK- Treffen zu berichten.

Einige Ideen, wie man das umfangreiche Material, das uns durch die P2Dokumentation geboten wird, für den praktischen Gebrauch des "Corpusarbeiters vor Ort" erschließen kann, wurden auf dem vergangenen Treffen diskutiert. Dazu gehört das Projekt eines Leitfadens (dies ist vorerst nur der Arbeitstitel: ein Leitfaden für die Guidelines!), der es dem Wissenschaftler entbehrlich machen soll, vor der Erfassung eines Corpus mittlerer Reichweite und Größe zuerst das gesamte Normwerk von

[5]im Dezember 1993 erschienen endlich die Kodierrichtlinien für Verstexte, so daß wenigstens die TEI-konforme Erfassung von allen drei großen belletristischen Textsorten abgeschlossen ist.

dann vielleicht 2000 Seiten durchzustudieren.

Diese eben genannten Themen und die letzthin aufgeworfene "Gretchenfrage": sollte man das ganze TEI-Projekt bereits jetzt schon als vorerst gescheitert betrachten und nach anderen Lösungswegen (natürlich auf der Basis der rundum akzeptierten SGML-Norm) suchen, werden uns auch weiter beschäftigen, vielleicht sogar Zündstoff bieten für die weitere Diskussion. Wir beabsichtigen deshalb, für 1994 mindestens zwei Treffen vorzusehen, eines im April oder Mai und eines im Zusammenhang mit der GLDV-Tagung auf der KONYENS in Wien (28.-30.9.1994)

Zum Abschluß möchte ich diesen Bericht auch verbinden mit dem Aufruf zur tätigen Mitarbeit an unserem Arbeitskreis, besonders wertvoll wären uns natürlich Kollegen, die bereits mit den TEI-Richtlinien oder anderen SGML-konformen Kodierungen arbeiten und darüber berichten könnten.

Bitte wenden Sie sich an:

Peter Scherber
bzw. Günter Koch GWDG
Am Faßberg
D-37077 Göttingen Tel.
0551/201559 bzw.
0551/201550
FAX: 0551/21119
e-mail: pscherb@gwdg.de
bzw. gkoch@gwdg.de