Künftige Software zur Textanalyse

Das Center for Electronic Texts in the Humanities (CETH) hatte vom 17.–19. Mai 1996 in Princeton ein Treffen organisiert, bei dem Anforderungen an künftige Software zur Analyse von Texten in den Geisteswissenschaften besprochen wurden. Rund 30 namentlich eingeladene Teilnehmer aus den USA, Kanada, Deutschland, England, Italien, Norwegen und Japan waren der Einladung gefolgt.

Folgende Fragen standen auf der Tagesordnung:
- Welches sind die geisteswissenschaftlichen Nutzer von Software zur Textanalyse?
- Welche Funktionalität zur Textanalyse gibt es in gegenwärtig verfügbaren Werkzeugen (TACT, OCP, TUSTEP, Monoconc, Opentext, SARA, LEXA etc.)?
- Welche Funktionalität wird künftig von Geisteswissenschaftlern benötigt?
- Sollte SGML das grundlegende Kodierungsschema darstellen, von dem alle Entwicklung künftiger Textanalyse-Software ausgehen sollte?
- Welches sind mögliche Architekturen für eine Paket von Textanalyse-Werkzeugen?
- Welche andere Software, die in den Geisteswissenschaften benötigt wird, sollte mit der Textanalyse-Software interagieren können?

Die Tagung war natürlich zu kurz, um diese Fragen wirklich zu klären – bis auf die eine vielleicht, daß an der Standard Generalized Markup Language (SGML) bzw. den TEI-Guidelines künftig kein Weg vorbeiführt. Nach der kurzen Zusammenfassung, die Susan Hockey, die Direktorin des CETH, als Ergebnis der Tagung in der Humanist Discussion Group veröffentlicht hat, einigte man sich darauf, daß in der Tat Handlungsbedarf besteht. Die nächsten Schritte sollen sein:
- Analyse des Bedarfs der Geisteswissenschaften
- detaillierteres Studium und Analyse der existierenden Software
- Aufstellen von Richtlinien, um die Interoperabilität eines Satzes von plattform-unabhängigen, modularen und erweiterbaren Werkzeugen zu gewährleisten.

Diesem – sich eher mager darstellenden – Ergebnis sollen ein kurzer Bericht und einige Beobachtungen aus (meiner) Tübinger Sicht hinzugefügt werden. Ein sehr viel ausführlicherer Bericht von Michael Sperberg-McQueen, der an der Organisation dieses Treffens beteiligt war, ist unter
http://www.uic.edu/~cmsmcq/trips/ceth9505.html
zugänglich.

1. Vorhandene Software

Der erste Vormittag war der Vorstellung existierender Software gewidmet, die in zwei Gruppen, »Pre-SGML Tools« und »SGML Tools«, präsentiert wurde.

Von den »Pre-SGML Tools« wurden in je 15 Minuten vorgestellt:
- Lexa, eine Sammlung von DOS-Programmen zur Verwaltung und zur Analyse von Textcorpora (vorgestellt von Espen Ore vom Norwegian Computing Centre for the Humanities);
- Monoconc, ein unter Windows laufendes Konkordanz-Programm mit Schwerpunkt auf einfacher Bedienbarkeit (vorgestellt von seinem Entwickler, Michael Barlow, Rice-University und Fa. Athelstan);
- OCP, das bekannte und weit verbreitete Oxford Concordance Program, das ursprünglich für Großrechner entwickelt wurde und vor allem in der von Oxford University Press vertriebenen PC-Version große Verbreitung erfahren hat (vorgestellt von Susan Hockey, unter deren Leitung es entwickelt wurde);
- TACT (»Text Analysis Computing Tools«), ein DOS-basiertes interaktives Konkordanz-Programm (vorgestellt von seinem Entwickler, John Bradley aus Toronto);
- TextPack, das am ZUMA in Mannheim zum Gebrauch vor allem in den Sozialwissenschaften entwickelte Paket zur Inhaltsanalyse von Texten (vorgestellt von Cornelia Züll aus Mannheim);
- TUSTEP, das »Tübinger System von Textverarbeitungs-Programmen« (vorgestellt von Wilhelm Ott).

Von diesen Paketen fand TUSTEP die meiste Aufmerksamkeit. Einer der Gründe war sicher, daß es in diesem Kreis außer vom Namen her wenig bekannt war (und es viele überraschte, daß es neben der deutschen auch mit englischer Oberfläche bedienbar ist); andererseits hat es von den vorgestellten Systemen den größten Funktionsumfang. Die knappe Vorstellung der Leistung und vor allem des konsequent modularen Aufbaus hatte die Teilnehmer so neugierig gemacht, daß sie auf einer zusätzlichen Vorführung bestanden.

Von den »SGML-Tools« wurden vorgestellt: Dynatext (Hersteller: EBT) und Explorer (Hersteller: SoftQuad), beides Systeme zur (elektronischen) Publikation von Dokumenten, die mit SGML ausgezeichnet sind; PAT (Hersteller: Open Text), eine leistungsfähige Suchmaschine, die ursprünglich für das New Oxford English Dictionary entwickelt wurde, und SARA (»SGML Aware Retrieval Application«), das an der Universität Oxford für das (ebenfalls mit SGML ausgezeichnete) British National Corpus entwickelt wurde. Bis auf SARA handelt es sich um kommerzielle Produkte, die nicht als Text-Analyse-Software, sondern als »browser« konzipiert sind; SARA ist auf die Fragestellungen des zugrundeliegenden Corpus zugeschnitten.

2. Anforderungen der Nutzer

Die Frage, welche Anforderungen an künftige Textanalyse-Software zu stellen sind, wurde in 5 Kurzvorträgen aus der Sicht verschiedener Nutzergruppen eröffnet: Klassische Philologie und Bibelwissenschaften (Winfried Bader, Deutsche Bibelgesellschaft, Stuttgart); Handschriften (Michael Neumann, Georgetown Univ., Washington); Literarkritik (John Burrows, Univ. Newcastle); Ostasien-Wissenschaften (Shoichiro Hara, Nat. Inst. of Japanese Literature, Tokio); Edition von historischen Dokumenten (David Chesnutt, Univ. of South Carolina, Columbia). Die gleiche Frage wurde anschließend in vier getrennten Gruppen diskutiert; die – bis auf sehr generelle Forderungen erwartungsgemäß eher mageren – Ergebnisse dieser Diskussion wurden dann im Plenum vorgestellt.

Gute Suchfunktionen (einschließlich pattern matching), SGML-Unterstützung, Plattform-Unabhängigkeit, Modularität und (neben interaktivem Zugang über graphische Benutzeroberfläche) ein »fully scriptable user interface« waren Hauptforderungen, die in praktisch allen Gruppen formuliert wurden.

3. Implementierungsfragen

Der zweite Tag war mit »Implementierungsfragen« überschrieben, wobei in den Vorträgen und Diskussionen ein deutlicher Schwerpunkt auf SGML (und seinen Grenzen) und den speziell für die Geisteswissenschaften erarbeiteten TEI-Guidelines (vgl. BI 96/3+4, S. 9–10) lag (und durchaus auch betont wurde, daß auch die bisherigen, nicht nach SGML ausgezeichneten Texte und Corpora weiterhin analysierbar sein sollten – und sei es mit existierender Software). Die mögliche Architektur eines künftigen Systems war Thema eines gemeinsamen Beitrags von John Bradley, dem Entwickler von TACT, und Geoffrey Rockwell von der McMaster University, die am Beispiel von Eye-ConTact das Modell einer graphischen Benutzeroberfläche für ein modulares, offenes, ausbaufähiges System von Textanalyse-Werkzeugen vorstellten. Im gegenwärtigen Zustand ist Eye-ConTact eine solche Oberfläche für BTACT, eine batch-Version des schon erwähnten Text-Analyse-Systems aus Toronto.

Noch einmal teilten sich anschließend die Teilnehmer in drei Gruppen, um Fragen der notwendigen Suchfunktionen (Suchen im »Text an sich«, in Markierungen und Anmerkungen, in externen Objekten wie Bildern, Audio, Video), weiterer Nutzer-Anforderungen (Dokumentation; Training; Support) und möglicher Architektur eines künftigen Systems (Benutzer-Schnittstelle; Modul-Management; Arbeiten im Netz und lokal; Basis: Texte, Datenbanken, andere Daten) zu diskutieren. Die Forderung nach einer gemeinsame Such-Sprache wurde von Antonio Zampolli aus Pisa mit dem Hinweis untermauert, daß ein Europäisches Projekt in den nächsten Monaten starten wird, das eine »common query language« für Textcorpora entwickeln soll. Als nach außen formulierbares Ergebnis einigte man sich schließlich auf die von Susan Hockey formulierten und oben zitierten drei Punkte.

Den letzten dieser drei Punkte soll ein Zitat aus dem Bericht von M. Sperberg-McQueen etwas näher erläutern, das er freilich ausdrücklich als seine persönliche Sicht verstanden wissen will: »What is needed is a commitment to cooperative work among developers in a chaotic environment of experimentation and communication«; »systematic top-down definition of architecture« hält er nach dem Treffen nicht mehr für »realistic, or even necessarily desirable. ... We are not building a building; blueprints will get us nowhere. We are trying to cultivate a coral reef ... of cooperating programs«, wobei es gilt, durch »regular communication among developers« für »interoperability among their programs« zu sorgen: »all we can do is try to give the polyps something to attach themselves to, and watch over their growth«.

4. Die Tübinger Perspektive

Mir selbst hat dieses Treffen gezeigt, daß wir mit unserem Tübinger Konzept und mit TUSTEP als Software eine gute Ausgangsbasis auch für künftige Anforderungen haben. Die

Stichworte Modularität, Professionalität, Integration und Portabilität, mit denen ich das TUSTEP zugrunde liegende Konzept umrissen hatte, sind in der oben erwähnten Zusammenfassung als »perhaps the most important points for thinking about the next generation of text analysis tools« bezeichnet, die bei der Vorstellung existierender Werkzeuge vorgebracht wurden.

Obwohl TUSTEP auch mit SGML- bzw. TEI-kodierten Texten besser zurechtkommt als viele andere Software, besteht hier Nachholbedarf. Mit der komfortablen Unterstützung nicht-linearer Text-Strukturen tut sich TUSTEP durchaus noch schwer und kann deshalb dem Vorwurf »None of the current generation of text-analysis tools support the significantly richer structural model of SGML« noch nicht adäquat begegnen.

Nicht akzeptiert an TUSTEP wird, zumindest für den Anfänger bzw. den breiteren Anwenderkreis, die bestehende Benutzeroberfläche als alleiniger Zugang zu diesem Werkzeug; außer dem »expert mode« (wie die bestehende Oberfläche bezeichnet wurde) erscheint – trotz der Forderung nach scriptability – eine graphische Benutzeroberfläche als unumgänglich.

Wilhelm Ott
ott@zdv.uni-tuebingen.de

BI 96/5 + 6, S. 24–26