Workshop der ›International TUSTEP User Group‹ (ITUG) zum Thema ›Lemmatisierung‹ Blaubeuren, 3.–6.10.1996

Vom 3. bis 6. Oktober 1996 hatte die International TUSTEP User Group zu einem Workshop zum Thema ›Lemmatisierung‹ ins Heinrich-Fabri-Institut nach Blaubeuren eingeladen.

›Lemmatisierung‹ umfaßt sowohl die Erstellung einer Lemmaliste, wie sie etwa einem Wörterbuch oder einem Werkindex zugrundeliegt, als auch die richtige Zuordnung von flektierten Wortformen und Homographen zu diesen Grundformen. Je nach Textcorpus und je nach Zielsetzung werden an die maschinelle Herangehensweise unterschiedliche Anforderungen gestellt. Ist die Herstellung eines Wortindex noch ein vergleichsweise einfaches Unterfangen, so muß der Lemmatisierung ein philologisches Spezialwissen zugrundeliegen, das es für die maschinelle Bearbeitung auf eine formale Ebene zu übersetzen gilt.

Die Lemmatisierung gewinnt Aktualität durch die wachsende Bedeutung des elektronischen Mediums für die Publikation und für die Archivierung von Texten. Mit steigenden Rechnerkapazitäten verbessern sich die Möglichkeiten für die Anlage größerer Textcorpora, gleichzeitig erhöhen sich die inhaltlichen Anforderungen. In vielen Fällen kommen Retrievalfunktionen und Linking nicht ohne lemmatisierte Indizes aus, wie sie elektronischen Textausgaben beigegeben werden müssen.

In Blaubeuren trafen sich insgesamt 22 Vertreter verschiedener Disziplinen. Die Teilnehmerinnen und Teilnehmer kamen von der Freien Universität Berlin (Institut für Judaistik; Institut für Rechtsgeschichte), aus dem Rechenzentrum der Universität Leipzig, aus München, von der Universität Trier (Mittelhochdeutsches Wörterbuch; Kant-Index; Ältere jiddische Editions- und Lemmatisierungsprojekte; Rechenzentrum; SFB Westmitteldeutsche Urkundensprache), von der Universität Tübingen (Deutsches Seminar; Zentrum für Datenverarbeitung), aus der Stiftung Weimarer Klassik in Weimar, von der Universität Würzburg (Institut für deutsche Philologie) und aus Spanien (Universitäten Valladolid und Burgos). Die Zusammensetzung dokumentiert das breite Spektrum der inhaltlichen Anforderungen, das von der Lemmatisierung im Rahmen wissenschaftlicher Einzelarbeiten über Großprojekte wie den Kant-Index und zu fortgeschrittenen Wörterbuch-Unternehmungen wie das ‘Mittelhochdeutsche Wörterbuch’ reichte. Lösungswege für die Lemmatisierung wurden an verschiedenen historischen Sprachstufen des Deutschen gezeigt, es wurden Methoden für das Hebräische, das ältere Jiddisch sowie für lateinische Beispieltexte vorgestellt.

Die Vorträge und Diskussionspunkte lassen sich in drei Schwerpunkte gliedern:

1. Philologische Anforderungen und Diskussion lexikographischer Einzelfragen
2. Strategien für die maschinelle Lemmatisierung und Sortierung
3. Auszeichnung und Markierung für die Herstellung maschinenlesbarer Textfassungen.

Neben der theoretischen Erörterung spezifischer philologischer Fragen stand die Vorstellung von praktischen Lösungen im Vordergrund. Im Rahmen der Erstellung eines neuen mittelhochdeutschen Wörterbuchs war eine Folge von TUSTEP-Prozeduren für die maschinelle Lemmatisierung mittelhochdeutscher Texte entwickelt worden, deren Realisierung die Programmautoren und die Trierer Arbeitsgruppe vorstellten (Paul Sappler, Tübingen; Ute Recker und Kurt Gärtner, Trier). Ansatzpunkte für ein regelbasiertes Verfahren bieten dagegen Texte in hebräischer Sprache, dessen Umsetzung von Gottfried Reeg (Berlin) als TUSTEP-Prozedur demonstriert wurde. Weitere vorgestellte Verfahren basieren auf der Auswertung von Markierungen und Markierungsfolgen in Texten (Wegstein, Würzburg) oder auf statistischen Methoden (Fiebig, Tübingen).

Als Werkzeug für die Arbeit in großen Projekten wird TUSTEP außer im Mittelhochdeutschen Wörterbuch (80.000 Lemmata) bei der Erstellung des Index für das Gesamtwerk Immanuel Kants eingesetzt (Heinrich Delfosse, Trier).

Zwei TUSTEP-Musterprozeduren zur Auszeichnung von Texten (Stahl, Würzburg) und zum Einstieg in die Lemmatisierung (Trauth, Trier) werden auf dem ITUG-Server in Würzburg allgemein zugänglich gemacht werden.

Außerhalb der Vorträge zur Lemmatisierung wurden die Neuerungen in TUSTEP vorgeführt. Einen ersten Schritt zur SGML-Unterstützung in TUSTEP bildet die Integration von Spitzklammer-Makros in das Satzprogramm (vgl. BI 96/7+8, S. 8–9). Außerdem wurde über ein Konzept berichtet, das bei der Weitergabe von

TUSTEP-Dateien an andere Systeme die Benutzung von TEI-Lite als Transportmedium vorsieht.

Im Rahmen des Workshops wurde außerdem eine Testversion der Windows 95/Windows NT-Version von TUSTEP präsentiert. In dieser 32–bit-Version für den PC entfallen die Beschränkungen der DOS-Fassung, so daß auch das Satzprogramm auf dem PC verfügbar ist. Gegenüber der bisherigen Fassung wurden die Kommandomakros wesentlich erweitert, die ein Werkzeug für die Steuerung von komplexen Abläufen darstellen und außerdem die Dateneingabe über Eingabemasken ermöglichen. Gezeigt wurde der derzeitige, schon testweise ablauffähige Planungsstand. Der Leistungsumfang der neuen UNIX- und die Windows-Versionen von TUSTEP, die voraussichtlich Ende 1996 freigegeben werden, ist identisch.

Ein ausführlicher Bericht über den Workshop ist auf dem Informationsserver der ITUG über
http://www.germanistik.uni-wuerzburg.de/itug.html
zugänglich.

Annegret Fiebig
fiebig@zdv.uni-tuebingen.de

BI 96/9 + 10, S. 12–13