DHd 2014, 25.-28.März 2014, Universität Passau

Abstract zum Vortrag:

Was ich nicht weiß, … macht mich heiß: Zum Mehrwert der Anwendung informatischer Methoden bei der Analyse von Textkorpora am Beispiel des Projektes „Biblia Hebraica transcripta“

Christian Riepl, IT-Gruppe Geisteswissenschaften, LMU München

Der Vortrag betrachtet das Mitte der 1980er Jahre von Wolfgang Richter an der LMU München initiierte Projekt „Biblia Hebraica *transcripta*“ (BHt) im Licht der aufkommenden „Digital Humanities“ und versucht unter den Aspekten a) Interdisziplinarität, b) Theoriebildung und Methodik sowie c) Gegenstand und Kollaboration, den Mehrwert der Anwendung informatischer Methoden in einem geisteswissenschaftlichen Langzeitprojekt herauszustellen.

a) Interdisziplinarität:

Das Projekt ist von Beginn an wesentlich geprägt durch eine enge und langjährige Kooperation einer geisteswissenschaftlichen mit informatischen Disziplinen. Über die einzelnen Projektförderphasen zwischen 1986 bis 1998 hinaus waren die entwickelten Systeme zum Teil bis zum Jahr 2010 im Einsatz. Der Datenbestand ist weiterhin system-, plattform- und programmunabhängig der Forschung zugänglich. Er umfasst Texte, systematisiertes Grammatikwissen und Ergebnisse der sprachwissenschaftlichen Analyse.

Die Zusammenarbeit führte auf beiden Seiten zu Aus- und Rückwirkungen. Das Forschungsinteresse seitens der Informatik war zunächst begründet in den großen Mengen an Text- und Metadaten sowie den darauf anzuwendenden komplexen Regeln der Grammatik, die sich in den Schwerpunktbereichen „Logikprogrammierung“, „Deduktive Datenbanken“ und „Expertensysteme“ z.B. mit Auswertungsstrategien von Logikprogrammen und der Analysemethodik befasste. Im Rahmen der Schwerpunkte „Informationretrieval“ und „Netzzugang zu multimedialen Informationssystemen“ war zunächst die Suche in Feature-Baum-Datenbanken, später der webbasierte Zugang zu Datenbanken interessant.

Auf der anderen Seite war in der althebraistischen Sprachwissenschaft die Übernahme der informatischen, streng formalisierten Denk- und Herangehensweise an einen Gegenstand grundlegend. Die Auswirkungen sind sichtbar z.B. bei der Kodierung der Transkriptionszeichen, der Wahl des Zeicheninventars zur Kodierung linguistischer Annotationen, der logischen und eindeutigen Strukturierung aller Daten, dem Entwurf von Datenbanken und der für die Entwicklung von Analyseprogrammen erforderlichen Formalisierung von Grammatikwissen in Regeln. Erkennbar ist der Einfluss der Informatik weiter an allen Arbeitsschritten der sprachwissenschaftlichen Analyse auf allen methodischen Ebenen, von der Anwendung der Analyseprogramme in einem halbautomatischen Verfahren auf Wort-

und Wortfügungsebene ausgehend bis hin zur datenbankgestützten manuellen Analyse der Satzebene. Schließlich umfasst der Einfluss der Informatik auch das Gebiet der Publikation der Daten, zunächst durch die automatische Konvertierung der Daten zur Verwendung des Drucksatzprogrammes TeX, sodann durch die Entwicklung und den Einsatz von webbasierten Datenbanken auf der Grundlage einer Server-Client-Architektur.
Das Projekt BHt gilt damit in vielerlei Hinsicht und für viele Projekte in den Geisteswissenschaften der LMU als modellhaft für den Einsatz informatischer Methoden und Technologien in geisteswissenschaftlichen Disziplinen.

b) Theoriebildung und Methodik:
Der althebräische Text des Alten Testamentes sollte zunächst transkribiert und in den Computer eingegeben werden, um anschließend durch Computerprogramme grammatisch analysiert zu werden. Die theoretische und methodische Grundlegung für eine orthographiebezogene, morphologisch-syntaktische Transkription war durch W. RICHTER, Transliteration und Transkription: ATSAT 19 (1983) geschaffen. Die Wahl der Datenstrukturen mit Referenzsystem, Segmentierung und Tokenisierung geschah in Abstimmung mit der Informatik, um zum einen größtmögliche Kompatibilität zwischen den Projektpartnern zu erreichen und zum anderen die Daten für eine automatische Analyse vorzubereiten. Basierend auf einem ebenenspezifischen Grammatikmodell, das durch W. RICHTER, Grundlagen einer althebräischen Grammatik, Band 1-3: ATSAT 8 (1978), ATSAT 10 (1979), ATSAT 13 (1980) systematisch begründet war, wurde das dort an einem Ausschnitt des Alten Testamentes gewonnene Grammatikwissen in formale Regeln überführt, die wiederum als Computerprogramme formuliert wurden. Dabei entspricht RICHTER (1978) dem Programm SALOMO zur Analyse der Morphologie, und RICHTER (1979) dem Programm AMOS zur Analyse der Morphosyntax. Beide Analyseprogramme arbeiten kontextunabhängig, streng auf die jeweilige methodische Ebene (Wort bzw. Wortfügung) bezogen und ohne Lexikon. Methodisch setzt die Analyse auf der Wortebene an und schreitet dann über die Wortfügungsebene zu den höheren Ebenen (Satz und Satzfügung) fort. Ein grammatisches, ebenenbezogenes Lexikon entsteht bei der sukzessiv wortweise vorgehenden Analyse der Texte.
Beachtenswert sind nun gerade die unerwarteten Mehrdeutigkeiten, die sich aus der automatischen Analyse auf jeder Beschreibungsebene ergeben. Sie zeigen alle Deutungen auf, die entweder tatsächlich zutreffen können, oder die der Präzisierung, Differenzierung bzw. Einführung weiterer grammatischer und/oder semantischer Regeln bedürfen. Der Analyseprozess legt damit den Erkenntnisweg bei der Sprachbeschreibung offen und zwingt den Experten zu einer Entscheidung, die er auf Grund seiner Kenntnis der jeweils höheren

Beschreibungsebenen reflektiert vornimmt und über ein den Analyseprogrammen nachgeschaltetes Dialogsystem eingibt.
Weiter ist beachtenswert, dass alle in SALOMO und AMOS implementierten Grammatikregeln im Expertendialog konsequent auf den gesamten Textkorpus angewendet worden sind. Die Methodik sah vor, alle im Expertendialog entstandenen Analyseergebnisse aufzuheben und eine Neuberechnung durch die Analyseprogramme nur im Bedarfsfall, z.B. für den Test einer grammatischen Regel oder für ein Experiment durchzuführen. Alle Analyseergebnisse befinden sich neben dem transkribierten Textkorpus in einer relationalen Datenbank, deren Schema versucht, das ebenenspezifische Grammatikmodell in je einer Relation für die Wort-, Wortfügungs-, Satz- und Satzfügungsebene abzubilden.
Als Ergebnis allein dieser Projektphase liegen vor: Ein vollständig transkribierter, morphologisch und morphosyntaktisch analysierter bzw. annotierter Textkorpus des gesamten hebräischen Alten Testamentes. Aus dem Datenmaterial ist eine vollständige Theorie der Morphologie (Bauformen, Wortarten, Kernseme) und der Morphosyntax (Wortfügungsarten, rekursive Wortverbindungen) herleitbar. Zugleich bildet das Datenmaterial ein grammatisches Lexikon. Konzeptionell sind die Analyseprogramme modifizierbar und somit der Datenbestand unter geänderten Voraussetzungen neu berechenbar. Daneben bietet die Datenbank Möglichkeiten, den Datenbestand unter bestimmten Bedingungen abzufragen, Regeln zu verifizieren bzw. zu falsifizieren, Experimente durchzuführen und neues Wissen zu deduzieren.
Die Anwendung informatischer Methoden führt erfahrungsgemäß zu Überraschungen und Nebeneffekten, die wiederum ein neues Licht auf den Gegenstand werfen und Erklärungen verlangen.

c) Gegenstand und Kollaboration:
Während seiner knapp 30-jährigen Dauer hat das Projekt BHt verschiedene Phasen mit jeweiligen Schwerpunkten durchlaufen. Beispielsweise wurden nach Abschluss der Analysearbeiten die Programme SALOMO und AMOS nicht weiter gewartet. Der Versuch einer automatischen Analyse der Satzebene wurde bislang nicht weiter verfolgt. Ebenso steht das Informationretrievalsystem für Feature-Baum-Strukturen nicht mehr zur Verfügung. Andererseits wurde der gesamte Datenbestand in einer relationalen Datenbank mehrmals reorganisiert. Die Datenbank bhtdb2 konnte schon in einer Frühphase der Entwicklung von Webtechnologien über das webbasierte Informationssystem MultiBHT in erster Linie für Recherchen genutzt werden. Eine soziale Komponente war insofern enthalten, als einfache Benutzerkommentare zu Tokens angebracht werden konnten. Die veraltete und seit 2010 nicht mehr lauffähige Systemtechnologie der Datenbank- und Webschnittstelle wird zurzeit einem umfangreichen Reengineering, das auch kollaborative Aspekte berücksichtigt,

unterzogen. Das Projekt BHt befindet sich mit der nahezu abgeschlossenen rechnergestützt-manuell durchgeführten syntaktischen und semantischen Analyse der Satzebene am Übergang zur dritten Forschergeneration. Ein weiteres Projekt zum althebräischen/semitischen Onomastikon kann ab Frühjahr 2014 auf dem Datenbestand aufbauen.

Als Essenz eines genuinen Langzeit-DH-Projektes lässt sich beobachten:

Den primären Kern digitaler Projekte bildet der gesamte in logisch eindeutig strukturiertem Format vorliegende Datenbestand. Sekundär, weil von der jeweiligen Fragestellung abhängig und daher austauschbar, sind alle digitalen Analyse-, Präsentations- und Recherchewerkzeuge, wobei sich der für diese erforderliche Entwicklungsaufwand durch die rasch fortschreitenden Technologien reduziert, sich Standards herausbilden und digitale Werkzeuge auf andere Problemstellungen übertragbar werden.

Neben der Veröffentlichung von Forschungsergebnissen in Buchform, mit denen Projekte in der Regel ihren Abschluss finden, stehen die Primärdaten eines digitalen Projektes über die eigentliche Projektlaufzeit hinaus weiter zur Nachnutzung zur Verfügung. Digitale Projekte sind somit nie abgeschlossen und entfalten eine ihnen eigene Dynamik.

Auf dem Weg der Entwicklung hin zu einer kollaborativen Forschungsumgebung stehen der Gegenstand und die Analyseergebnisse objektiviert, transpersonalisiert und universal transformierbar, damit unabhängig von Projekt, Plattform und Forschern im Mittelpunkt. Die Daten können von verschiedenen Forschern verändert und z.B. auch um konkurrierende Meinungen ergänzt, aus anderen Perspektiven, mit anderen Theorien und Methoden, aber auch durch andere Disziplinen untersucht, oder mit digitalen Daten anderer Disziplinen verknüpft werden. Zudem bietet ein Transfer von Projektdaten in Formate und Infrastrukturen von CLARIN und DARIAH hervorragende Möglichkeiten für den Datenaustausch und fördert damit nationale und internationale Kooperation. Der digitale Forschungsgegenstand wird kollaborativ, interdisziplinär und multidimensional erfasst und aus pluraler Sicht betrachtet werden können.