Fotis Jannidis

WAS IST COMPUTERPHILOLOGIE?

Im Zuge seiner weltweiten Verbreitung konnte sich der PC gegen anfängliche Bedenken und Widerstände auch in der Literaturwissenschaft als Werkzeug der täglichen Arbeit etablieren. Anfangs waren es vor allem die Vorteile der Textverarbeitung und deren Entlastung vom mechanischen Aspekt des Schreibens und Wiederschreibens, die den Rechnern den Weg auf die Schreibtische ebneten. Ist aber die Maschine einmal vorhanden, man sich mit geringem Aufwand Zugang zum Internet verschaffen. E-Mail und das World Wide Web eröffnen einfachere Kommunikationswege, dazu kommen die Vorteile des Intranets, also eines universitätseigenen Netzes mit Zugriff auf elektronische Bibliographien und die Bibliothekskataloge einschließlich der Bestellmöglichkeiten vor Ort. Nicht wenige Literaturwissenschaftler haben sich inzwischen auch mit den neueren elektronischen Texten angefreundet, deren einfachen Benutzeroberflächen althergebrachte philologische Tätigkeiten sehr beschleunigen, zum Beispiel die Klärung von Wortbedeutungen mittels der Suche nach Parallelstellen beim selben Autor oder in derselben Epoche.[1]

Teilbereiche der Literaturwissenschaft wurden noch sehr viel eingreifender von der Arbeit mit dem PC umgestaltet, aber scheinbar waren diese Veränderungen für die übrigen ohne Belang. Die Editionsphilologie wurde etwa sehr schnell auf die Möglichkeiten aufmerksam, ihre Tätigkeit mittels des Rechners zu erleichtern; aber angesichts des sehr gepflegten Äußeren historisch-kritischer Ausgaben kommt man ohnehin ins Grübeln, wie es um die Akzeptanz der Arbeitsergebnisse dieses Fachteils bestellt ist. Wer elektronische Editionen verwendet, wird bei genauerer Betrachtung nun jedoch feststellen, daß die Entscheidungen der Editionsphilologen über das Design der jeweiligen Ausgabe großen Einfluß darauf haben, wonach man überhaupt suchen kann und wie man Suchen formulieren muß, damit sie die gewünschte Information auch finden. Die Kenntnis solcher Abhängigkeiten gehört zum neuen kulturellen Wissen im Umgang mit den neuen Medien.

Die neuen Arbeitsmöglichkeiten werden auch Einfluß auf die Einschätzung der primären fachlichen Kompetenzen haben. Die einfache Gedächtnisleistung verliert an Wichtigkeit, wenn schon ein Erstsemester alle Vorkommen des Wortes ›Licht‹ in Goethes Werken ermitteln kann. Dadurch aber wird die Fähigkeit immer wichtiger, Daten zu finden und schon bei der Suche relevante von irrelevanten Daten zu trennen. Die Veränderungen in der Literaturwissenschaft aufgrund der neuen Medien sind insgesamt so zahlreich, daß es an der Zeit ist, auch innerhalb des Faches eine Struktur zu schaffen, dieses Wissen auszutauschen und zu tradieren.[2]

Anknüpfen wird man dazu an die Tradition des professionellen Computer-Einsatzes in den Geisteswissenschaften, die übrigens überraschend lange zurückreicht. Bereits 1949 begann Roberto Busa mit seiner Arbeit an der computergestützten Erstellung einer Konkordanz zu den Werken Thomas von Aquins, eine Arbeit, die erst Jahrzehnte später abgeschlossen

wurde.[3] Inzwischen gibt es zahlreiche und – selbst wenn man nur die Textwissenschaften betrachtet – sehr unterschiedliche Projekte: computergestützter Druck, computergenerierte Indices, elektronische Editionen, computergestützte Stil-, Inhalts- und Rezeptionsanalyse, Spracherkennung und -generierung und vieles mehr. Dieses Forschungsfeld wird zusammenfassend als ›Humanities Computing‹ bezeichnet, und ein wesentliches Forschungsorgan, die Zeitschrift Computers and the Humanities erhebt diesen umfassenden Anspruch bereits im Titel. Da einige Fächer von Haus aus mehr Erfahrung mit quantitativen Methoden hatten, lag es nahe, daß diese die Führung übernahmen. In Deutschland hat sich inzwischen etwa ein eigener Teilbereich der Linguistik etabliert, die ›Computerlinguistik‹, die mit einigem Recht beanspruchen konnte, für zahlreiche Belange des textbezogenen Humanities Computing Forschungsmethoden und Problemlösungen bereitzustellen.[4]
Es gab allerdings schon immer Bereiche, in denen Literaturwissenschaftler eher eigenständig gearbeitet haben, etwa in der computergestützten Editionsphilologie. Erst in den letzten Jahren scheint sich aber eine Arbeitsteilung herauszubilden, die sich an den bereits etablierten Fachgrenzen orientiert: Computerlinguistik untersucht in erster Linie Gegenwartssprache auf den verschiedenen Sprachebenen mit einem besonderen Interesse an gesprochener Sprache, während die ›Computerphilologie‹ historische Sprachstufen untersucht mit einem deutlichen Schwerpunkt im Bereich der Edition und Interpretation fiktionaler Texte. Interessiert sich die Computerlinguistik zum Beispiel für Verfahren der automatisierten Disambiguierung und Lemmatisierung, so können die Philologen aufgrund der sehr viel größeren Variabilität ihres Materials auf solche Verfahren in absehbarer Zeit nicht hoffen und suchen nach anderen Formen intelligenten Text Retrievals. Diese Ausdifferenzierung auch institutionell in einem Jahrbuch für Computerphilologie[5] festzuschreiben, soll keine Trennung herbeibeschwören, sondern lediglich für Fachwissenschaftler mit gemeinsamen Interessen ein eigenes Diskussionsforum bereitstellen – mit Schnittstellen einerseits zum Humanities Computing und andererseits zur allgemeinen Literaturwissenschaft.

Unter dem Etikett ›Computerphilologie‹ soll also das Wissen um die Einsatzmöglichkeiten des Computers in der Literaturwissenschaft gesammelt werden. Insbesondere gehören dazu das (2) Erstellen und (3) Verwenden elektronischer Texte, einschließlich (3.1) der Lektüre und des (3.2) Information Retrievals, (4) die Hypertexttheorie und -praxis mit Berücksichtigung von Hyperfiction und (5) das Programmieren von Anwendungen für Literaturwissenschaftler. Ganz an den Anfang aber soll eine methodische Überlegung gestellt werden, wie das Wissen der Computerphilologie bestmöglich zu formulieren und tradieren ist.

1. ABSTRAKTION

In einem der ersten deutschsprachigen Bücher, das versucht, das neue Land ›Computer und Geisteswissenschaften‹ zu vermessen, findet sich ein Hinweis darauf, wie die »Idealausstattung« des PCs für solche Arbeiten auszusehen hat: »16-Bit-Gerät mit mindestens 512 KB Arbeitsspeicher, einem Disketten-Laufwerk, einer Festplatte von 20 MB und

einem monochromen oder einem Graphik-Bildschirm.«[6] Angesichts von Festplatten, deren Größe inzwischen in Gigabyte angegeben wird, und durchschnittlichem Arbeitsspeicher, dessen Menge die damalige Festplattengröße um einiges übertrifft, veranschaulichen diese Angaben besonders eindrücklich, wie schnell das aktuelle Wissen über Computer veraltet. Ähnlich flüchtig scheint alles Wissen dieses Arbeitsfelds zu sein: Die Kenntnisse, wie Programme zu verwenden sind, scheinen überflüssig zu werden, sobald man sie richtig erworben hat, da die nächste Generation des Programms mit neuen Menüs und Befehlen aufwartet.

Tatsächlich aber zeigt ein zweiter Blick, daß es zwar Wissen gibt, das sehr schnell veraltet, zum Beispiel in welchem spezifischen Menü eines bestimmten Programms der Befehl verborgen ist, mit dem man eine Datei öffnet. Daneben aber gibt es Wissen, das sehr viel widerständiger gegen den Lauf der Zeit ist, etwa die Kenntnis der Tatsache, daß im Computer Informationen langfristig in Dateien gespeichert werden und man diese öffnen muß, wenn man an die Informationen gelangen will. Dieses Wissen ist keineswegs resistent gegen Entwicklungen, es verändert sich nur sehr viel langsamer. Ein wesentlicher Bestandteil aller wissenschaftlichen Beschäftigung mit dem Computer ist die Suche nach den dauerhaften Prinzipien der Computerarbeit. Allerdings ist oft nur rückblickend zu ermitteln, welches Wissen langfristig stabil bleibt und welches schneller verfällt.[7] Die Informatik hat für die Speicherung und Manipulation von Daten eine Reihe von abstrakten Prinzipien beschrieben, die zwar nicht konstant bleiben, aber eine relativ langsame Entwicklungsgeschwindigkeit aufweisen und gleichzeitig bei der Lösung von Problemen erprobte Werkzeuge zur Verfügung stellen. Analog dazu ist es die Aufgabe einer Computerphilologie, solche dauerhaften Prinzipien zu ermitteln, zusammenzustellen und zu tradieren.

Ein wichtiger Schlüssel zu dauerhafteren Prinzipien des computerphilologischen Wissens ist Abstraktion. Abstrahiert werden muß von den kontingenten Elementen, um die stabileren, dauerhafteren Elemente zu ermitteln, doch, wie schon erwähnt, ist diese Unterscheidung keineswegs einfach zu treffen. Andererseits wird der Computer auch in den Geisteswissenschaften inzwischen einige Jahrzehnte verwendet und ein Blick auf die Entwicklung dieses Wissens macht die Arbeit heute leichter als noch vor zehn Jahren.

Die Informatik kann gleich auf zwei formalisierte Sprachsysteme zurückgreifen, nämlich Mathematik und Programmiersprachen einschließlich abstrakterer Notationsweisen.[8] Dieses Modell kann in einigen Teilgebieten auch die Computerphilologie übernehmen: Die automatisierte oder computergestützte inhaltliche und stilistische Textanalyse verwendet statistische Methoden, das Suchen in oder das Sortieren von Texten verwendet ebenso wie die Konvertierung elektronischer Texte gut erforschte Programmier-Algorithmen. Größere Teile der Computerphilologie lassen sich jedoch nicht so formalisiert erfassen, etwa die Frage, welche Aspekte eines Textes in welcher Weise ausgezeichnet werden müssen; dennoch handelt es sich dabei um sprachlich formulierbare Regeln oder etwas vorsichtiger formuliert, um Regelmäßigkeiten. Die Erfassung dieser Regelmäßigkeiten in so konziser Weise, daß man über

sie diskutieren und sie eventuell auch falsifizieren kann, ist ein besonderes Ziel des neuen Fachteils.

Neben der Suche nach den computerphilologischen Prinzipien bildet die Auseinandersetzung mit den einschlägigen Standards einen weiterer Schlüssel zu beständigerem Wissen. Mit ›Standard‹ ist in diesem Kontext die Verabredung gemeint, daß etwas so und nicht anders sein soll. In der Welt des Computers, in der ständige Innovationen an der Tagesordnung sind, müssen auch ständig neue Verabredungen getroffen werden, die das Zusammenspiel von neuer und alter Hard- und Software gewährleisten. Standards werden vor allem durch zwei Instanzen gesetzt: durch Firmen bzw. andere Institutionen in einem dezisionistischen Akt (man spricht dann etwas verharmlosend von Industrie- oder de facto Standard) oder durch unabhängige Standardisierungskomitees, zum Beispiel dasjenige das für den internationalen ISO Standard verantwortlich ist. Offene Standards, deren Gestaltung im Idealfall sogar durch die Anwender, etwa die scientific community, selbst vorgenommen wird, haben den Nachteil, daß die Entscheidungsprozesse über Veränderungen lange dauern und Optimierungen aufgrund neuer technischer Möglichkeiten daher auch nur langsam umgesetzt werden. Dieser Nachteil ist jedoch zugleich ihr Vorteil, da sie langfristig stabil sind und ohne Schielen auf die jeweilige Jahresbilanz gepflegt werden können. Eine Orientierung an solchen Standards zur Textauszeichnung, zum Bildformat, Zeichensatz, Hyperlink-Mechanismus ist der Computerphilologie nicht nur geboten, weil die Erstellung von Produkten, zum Beispiel einer aufwendig kodierten digitalen Edition, recht zeitintensiv ist und mit den kurzatmigen Soft- und Hardwarezyklen nicht mithalten kann, sondern auch, weil sie so Wissen erzeugt, verwendet und vermittelt, das dauerhafter ist.

2. ERSTELLUNG ELEKTRONISCHER TEXTE

Der elektronische Text ist Grundlage fast aller computerphilologischen Tätigkeiten. Anfangs war der digitale Text lediglich eine Stufe auf dem Weg zur Erstellung eines Drucktextes, inzwischen ist er auch Endprodukt in Form von elektronischen Editionen und fachspezifischen Informationssystemen und Datenbanken. Das Versprechen eines schnellen und unermüdbaren Zugriffs auf beliebig große Texte hat dazu geführt, daß die umfangreichsten und auch der literaturwissenschaftlichen Öffentlichkeit bekannteren Leistungen aus diesem Feld stammen. Um einen Text oder einen Informationsbestand zu digitalisieren, muß die Buchstaben-Information vom analogen aufs digitale Medium verlagert werden und die Metainformation eingetragen werden. Metainformationen in elektronischen Texten[9] sind alle zusätzlichen Angaben: von der Unterstreichung und dem Fettdruck über Kapitelgrenzen und Akt- und Versende sowie Strophenbegrenzungen bis zum Autornamen und den Werktitel, um nur einige zu nennen. Der übliche Produktionsweg bei der Digitalisierung eines gedruckt vorliegenden Werks besteht (1) im Erfassen des Texts und (2) im Anreichern des Texts mit zusätzlichen Metainformationen. Diese Schritte sind, das sei gleich vorweggeschickt, nur analytisch getrennt, in der Praxis aber oft Teil eines Arbeitsgangs.

(1) Für die Texterfassung haben sich zwei Wege etabliert: Das manuelle Erfassen und das Scannen mit anschließender

automatisierter Zeichenerkennung.[10] Vorteil der manuellen Eingabe ist, daß bereits bei der Eingabe der Text (je nach Vorlage) mit basalen Auszeichnungen versehen werden kann und außerdem eine relativ hohe Fehlerfreiheit gewährleistet ist. Nachteil ist der Aufwand an Arbeitskraft. Für sehr aufwendige Projekte wird der Text unabhängig zweimal eingegeben und in einem anschließenden Vergleichslauf mittels Computer werden alle Differenzen herausgefiltert, da sie wahrscheinlich Fehler anzeigen.

Das Scannen eines Texts erzeugt ein digitales Bild. Bilder werden im Computer prinzipiell anders kodiert als Buchstaben. Buchstaben werden mittels eines Zahlenschlüssels kodiert, der vom verwendeten Zeichensatz abhängig ist.[11] Bilder werden in Bildpunkten mit Angaben zu Farben und Helligkeit abgespeichert. Das gescannte Bild muß also für die Weiterverarbeitung erst mittels eines OCR-Programms (Optical Character Recognition) in einen Text umgewandelt werden.[12] Problematisch ist die immer noch relativ große Fehlerdichte von gescannten Texten, die eine aufwendige manuelle Nachbearbeitung erfordert.

(2) Die Textauszeichnung (markup), also das Eintragen von Metainformationen in den Text, kann teilweise direkt oder indirekt bei der Texterfassung geschehen oder automatisch aufgrund von vorhandenen Texteigenschaften. Wenn es sich jedoch nicht um stark strukturierte Texte wie Wörterbücher oder Lexika handelt, dann wird eine mehr oder weniger aufwendige Bearbeitung durch einen Philologen notwendig sein, der Anmerkungen einträgt, verschiedene Textteile mittels Hyperlinks verbindet oder einen kritischen Apparat erstellt. Elektronische Editionen werden zumeist in Autorensystemen oder mit Textbearbeitungsprogrammen bearbeitet, die einzelne Arbeitsschritte unterstützen, zum Beispiel die Kollationierung mehrerer Texte. Zur Publikation wird ein elektronischer Text zumeist noch dem Programm angepaßt, mit dem man den Text lesen und auch in ihm suchen kann.

Als wesentliches Problem hat sich die langfristige Speicherung eines elektronischen Texts erwiesen: Die meisten kommerziell vertriebenen Editionen sind zur Zeit aufgrund ihrer proprietären Auszeichnung eng an das jeweilige Darstellungs- und Retrievalprogramm und damit an dessen Lebensdauer gekoppelt. Eine weitgehend betriebssystem- und softwareunabhängige Kodierung, die elektronischen Texten eine mit Drucktexten vergleichbare Lebensdauer ermöglichen soll, kann mit dem philologischen Textauszeichnungssystem der Text Encoding Initiative (TEI) erreicht werden. TEI setzt auf dem internationalen Standard für Auszeichnungssysteme SGML (Standard General Markup Language)[13] auf und ermöglicht die Notierung gattungsspezifischer Merkmale von Prosa, Lyrik und Drama sowie die Auszeichnung von Transkriptionen gesprochener Sprache, von Wörterbüchern und terminologischen Datenbanken; es stellt außerdem einen Mechanismus zur Implementierung auch komplexer Hypertextverknüpfungen und zur Kodierung beliebiger Zeichen zur Verfügung.[14] SGML und damit auch TEI basieren auf dem Konzept einer semantischen Auszeichnung, die möglichst sauber von jeder typographischen Information zu unterscheiden ist. Für die typographische Umsetzung werden dann Angaben in einer eigenen Formatbeschreibungssprache verwendet, die auf

die jeweiligen Besonderheiten des Ausgabemediums Rücksicht nehmen können.

Der Vorteil von TEI besteht vor allem darin, daß Philologen, die eine Edition konservieren möchten, einen Standard verwenden können, der seit zehn Jahren in Entwicklung und Erprobung ist und der aus einer weltweiten Kooperation von Fachwissenschaftlern hervorgegangen ist. Er wird inzwischen in zahlreichen, teilweise sehr umfangreichen Editionsprojekten eingesetzt, was auch bedeutet, daß man im Fall von Problemen eine große Zahl möglicher Ansprechpartner hat. TEI hat einige Nachteile: Es bietet bislang kaum Möglichkeiten, die materialen Aspekte von Texten (Buchäußeres usw.) präzise zu beschreiben. Es verlangt eine hierarchische Strukturierung der Daten. Die klare Trennung von Typographie und Semantik, die alle SGML konformen Systeme voraussetzen, ist nicht in allen Fällen zu verwirklichen und noch nicht einmal immer wünschenswert. Außerdem sind die zugehörigen Formatbeschreibungssprachen keineswegs einfach zu handhaben. Allerdings gilt für diese Nachteile, daß sie alle erkannt und diskutiert werden, das heißt Lösungen für diese Probleme sind entweder bereits beschrieben oder es wird an ihnen gearbeitet.

Die Auszeichnung eines historischen Texts ist stets eine philologische Arbeit. Das gilt schon für die richtige Auswahl und Dokumentation der Textgrundlage und gilt insbesondere, wenn dabei der Text mit anderen Informationen verknüpft wird und Kommentare und Erläuterungen eingetragen werden. Wie weiter unten noch ausgeführt wird, entstehen einige typische Probleme des Text Retrievals dadurch, daß im Normalfall nur nach Zeichenketten und nicht nach Sinneinheiten gesucht werden kann. Dem läßt sich bereits bei der Textauszeichnung dadurch begegnen, daß in den Text Normalisierungen der Schreibung und die Grundformen der Wörter eingetragen werden. Auch die Disambiguierung von Homographen, seien es nun Worte oder Satzzeichen,[15] macht den Text für spätere Such- und Auswertungsoperationen brauchbarer.[16]

Jede Auszeichnung ist eine Interpretation des Textes. Einige können sich auf allgemein akzeptierte Standards stützen, andere kodieren aufgrund neuer Auffassungen, welche Textaspekte wesentlich sind, alle aber notieren eine bestimmte Sichtweise des Texts. Das ist aber kein größeres Problem: Zum einen gilt dies auch für jede gedruckte Edition, zum anderen haben Textauszeichnungssysteme wie TEI nicht nur die Möglichkeit, mehrere Sichtweisen auf den Text parallel einzutragen, sondern stellen auch das Instrument bereit, die gewählte Auszeichnung zu dokumentieren und damit zur Diskussion zu stellen.

3. VERWENDUNG ELEKTRONISCHER TEXTE

Kaum ein Umstand hat so sehr zu dem Frieden beigetragen, den viele Geisteswissenschaftler inzwischen mit dem Computer geschlossen haben, wie die Erfahrung, daß die Lektüre von Bildschirm-Texten deutlich unangenehmer ist als das Lesen von Gedrucktem. Die Augen scheuen auf Dauer den Monitor, es fehlen die bekannten Informationssignale über die eigene Position im Text, die überschaute Textmenge ist geringer als auf einem Blatt Papier, und das Medium ist zudem kaum

transportabel und vergleichsweise störanfällig. Selbst wenn sich durch die technische Entwicklung das eine oder andere Manko beheben lassen sollte, so besteht auf absehbare Zeit keine Gefahr für das geliebte Buch. Seitdem die begeisterten oder kulturkritischen Totsprechungen der Gutenberg-Epoche nicht mehr die Diskussion beherrschen, konnte sich die Einsicht durchsetzen, daß auch in diesem Fall ein neues Medium nicht zum Verschwinden des alten führt, sondern zu einer Ausdifferenzierung der Verwendungsweisen. Den Computer wird man auf absehbare Zeit nicht ins Bett nehmen, im Buch kann man nicht schnell einmal etwas suchen lassen. Die folgende Skizze beschreibt die zwei Verwendungsweisen elektronischer Texte, die inzwischen üblich sind: (3.1) Die direkte sinnliche Wahrnehmung des Texts, (3.2) das Information Retrieval, das (3.2.1) die Suche im Text und (3.2.2) die quantitative Analyse umfaßt.

Solange man davon ausgehen mußte, daß Geisteswissenschaftler keinen Rechner haben, um auf elektronische Texte zugreifen zu können, waren diese nur ein Zwischenformat, um daraus entweder eine Druckedition zu erzeugen oder um daran Informationen zu gewinnen, die dann in einem wiederum gedruckten wissenschaftlichen Text mitgeteilt wurden, und es gibt heute wohl kaum noch Editionen, Lexika, Wörterbücher, die nicht mit einer elektronischen Zwischenstufe arbeiten. Seitdem PCs so verbreitet sind, ist es eine Wahl des Editors bzw. des Verlags, welches Ausgabemedium man wählt. Elektronische Texte können weiterverwendet werden. Ohne größeren Aufwand können Zitate fehlerfrei aus einem Korpus entnommen und in die eigene wissenschaftliche Arbeit eingesetzt werden. Die Wiederverwendbarkeit betrifft aber vor allem diejenigen, die Texte herstellen und vertreiben: Sie können den einmal erstellten Text mit relativ geringem Aufwand wieder verwenden, eine Auswahl daraus treffen oder ihn in einen umfangreicheren Korpus eingliedern. Eine Editionsform, die den verschiedenen Verwendungsweisen von Texten gerecht zu werden versucht, ist die Hybridedition, die einen Drucktext mit einer begleitenden elektronischen Edition umfaßt.

3.1 WAHRNEHMUNG DES TEXTS[17]

Wie schon erwähnt, ist die längere Lektüre am Monitor eine eher unerfreuliche Erfahrung und wird nur dann gewählt, wenn auf diese Weise besondere Vorteile erreicht werden können. Vier solcher Vorteile haben sich als besonders zugkräftig erwiesen: Die schnelle, aktuelle Information, die Verbindung mehrerer, herkömmlicherweise getrennter Medien zu einem, das automatisierte Verfolgen von Verknüpfungen innerhalb elektronischer Texte und die Möglichkeit, verschiedene Ansichten auf denselben Text zu haben.

Elektronische Texte können über relativ billige Speichermedien, sei es die CD-ROM oder das Internet, vertrieben werden. Eine regelmäßige Aktualisierung der Daten ist auf diesem Weg kein großes Problem mehr. Entsprechend haben sich vor allem Bibliographien, Lexika und andere Nachschlagewerke als besonders digitalisierungstauglich erwiesen.

In modernen Leseprogrammen, sogenannten ›Browsern‹, können Texte mit Bildern, Tönen und Filmen zu einer neuen

Einheit verknüpft werden, was vorher nur mit großem finanziellen Aufwand oder gar nicht möglich war. Das Ergebnis wird zur Zeit insbesondere für didaktische Zwecke genutzt,[18] da einer gleichwertigen Verbindung der verschiedenen Medien für Forschungszwecke, also zum Beispiel die Verknüpfung einer Textdatenbank mit einer Datenbank von Ton- und Filmdokumenten, oft das Copyright und der notwendige Aufwand entgegensteht.[19]

Die Möglichkeit, Text- und andere Informationselemente mittels Hyperlinks miteinander zu verknüpfen, hat eine Fülle von theoretischen Überlegungen und Experimenten nach sich gezogen. Sie wird deshalb weiter unten in einem eigenen Abschnitt behandelt. Insbesondere durch die Leseprogramme für das World Wide Web haben sich in kurzer Zeit schon einige Standards, wie man sich in solch einem Medium bewegt und orientiert, herausgebildet.

Jeder Benutzer einer Textverarbeitung kennt inzwischen die Möglichkeit, zwischen einem Eingabemodus, der für das Lesen am Bildschirm optimiert ist, und einer Druckvorschau zu wechseln. Dies ist nur ein Beispiel für einen prinzipiellen Vorteil elektronischer Texte: Sie erlauben mehrere Ansichten auf dieselben Daten. Ein bereits realisiertes Beispiel bildet eine historisch-kritische Ausgabe, in der der Benutzer die Wahl hat, welche Fassung er als Basistext sehen möchte und welche Textstufen demzufolge in den Apparat gesetzt werden.[20]

3.2 INFORMATION RETRIEVAL

In elektronischen Editionen kann man suchen und das Gefundene zählen lassen: Das ist ihr wesentlicher Vorteil gegenüber herkömmlichen Editionen. Suche und Zählen können mit vorher nicht zu erreichender Genauigkeit und Schnelligkeit durchgeführt werden. Allerdings kann immer nur nach ›Zeichenketten‹ gesucht werden, also nach einer Reihe von Zeichen, ganz unabhängig davon, ob sich daraus Sinneinheiten wie Worte ergeben. Zwei Einsichten aus den letzten 30 Jahren computergestützter Textanalyse, die aus dieser Tatsache resultieren, müssen berücksichtigt werden: 1. Soll eine Zeichenkette als Sinneinheit erkannt werden, muß dem Computer entweder mittels eines Algorithmus oder aufgrund der Textauszeichnung mitgeteilt werden, wie er aus den Zeichenketten Sinneinheiten gewinnen kann. Die Entwicklung von produktionsreifen Algorithmen stellt, wenn es nicht um triviale Anwendungen geht, immer noch ein großes Problem dar[21] und ist für sehr variantes historisches Textmaterial sobald nicht zu erwarten. Sowohl die Algorithmen als auch die manuell eingetragenen Auszeichnungen sind eine Interpretation des Textes aufgrund eines Sprachmodell und Kontextwissens.
2. Elektronische Texte generieren nicht automatisch Daten über sich, sondern erst, wenn man höflich darum bittet.[22] Die Anfrage aber ist wiederum eingebettet in einen interpretatorischen Kontext, in dem sie erst sinnvoll wird. Erst diese Vorannahmen und Hypothesen erlauben auch eine angemessene Interpretation der Daten, die vom Computer geliefert werden. Ganz anders als der Mythos vom objektiven und präzisen Computers erwarten läßt, werden hier Interpretationen aufgrund von Daten gewonnen, die wiederum auf Interpretationen basieren. Dies bedeutet keineswegs den fröhlichen Einstieg in die Beliebigkeit. Ein Text kann nur

aufgrund von Weltwissen angemessen verstanden werden und sinnvolle Fragestellungen sind daher stets in einen entsprechenden Deutungshorizont eingebettet. Computerphilologen sind daher nur besonders verpflichtet, ihre Vorannahmen und Hypothesen möglichst explizit darzulegen. Und noch eine zweite Folgerung ist daraus zu ziehen: Jeder, der eine elektronische Edition verwenden will, muß lernen, seine inhaltliche Frage so zu formulieren, daß sie den Eigenheiten seines Recherche-Instruments gerecht wird.

3.2.1 Suchen

Der Unterschied zwischen der Suche nach Worten und nach Zeichenketten wird bereits angesichts sehr einfacher Dinge sichtbar: ›Sein‹, ›seyn‹ und ›sein‹ sind für ein Suchprogramm drei unterschiedliche Zeichenketten. Wer über die zeitlichen Grenzen von Sprachstandardisierungsprozessen und Rechtschreibreformen hinweg fündig werden will, muß seine Suche entsprechend gestalten.

Man kann nach einer beliebigen Kombination von Buchstaben, Zahlen, Satz- und Sonderzeichen suchen, tatsächlich ist es aber oft notwendig oder effektiver, die Suche allgemeiner zu formulieren. Eine erste Stufe der Verallgemeinerung bietet die Verwendung von Platzhaltern; bekannt sind aus der Arbeit mit dem Betriebssystem zwei Platzhalter: Einer steht für ein einzelnes beliebiges Zeichen, der andere für eine beliebige Menge beliebiger Zeichen. Aus der Welt der Unix-Rechner stammt ein Verfahren, mit ›regulären Ausdrücken‹ sehr präzise Zeichenmuster bzw. Klassen von Zeichenketten zu beschreiben.[23]

Mittels der Booleschen Operatoren UND, ODER und NICHT[24] können Suchen noch genauer spezifiziert werden, zum Beispiel erzielt die Suche nach »Herz UND Schmerz« nur einen Treffer, wenn beide Zeichenketten im Textabschnitt enthalten sind. Einzelne Zeichenketten können auf diese Weise zu komplexen Abfragen kombiniert werden. Spätestens dann stellt man fest, daß es eine Rolle spielt, in welcher Einheit die gesuchten Worte vorkommen müssen, damit ein Treffer gemeldet wird. Internet-Suchmaschinen finden stets ganze Dokumente, Volltextdatenbanken untergliedern den Text in kleinere Einheiten, oder der Anwender kann selbst festlegen, wie weit die Zeichenketten höchstens voneinander entfernt sein dürfen. Außerdem kann man die Struktur des Textes, insoweit sie als Textauszeichnung eingetragen ist, ebenfalls für die Suche verwenden und so nur in den Überschriften oder nur den handschriftlichen Ergänzungen in roter Tinte suchen. Neuere Suchmaschinen führen die Ergebnisse in einer gewichteten Liste auf; Texte oder Textabschnitte, die besonders viele der gesuchten Zeichenketten enthalten, werden zuerst aufgelistet.[25] Ein Problem haben alle Suchstrategien: Man weiß nicht, was man nicht findet. Zwar kann man bei der Durchsicht einer Treffermenge gut erkennen, wie präzise die Suchanfrage war, also wieviele der gemeldeten Treffer tatsächlich im inhaltlichen Sinne als Treffer gezählt werden können, aber man weiß nicht, wieviele einschlägige Textstellen man gar nicht erst sieht. Das findet man erst heraus, wenn man den Text durchliest – was immer seltener eine tatsächliche Option ist.[26]

Die hier geschilderten einfacheren Verfahren des Information

Retrieval werden durch die Verbreitung des Internets inzwischen Teil dessen, was als Computer Literacy bezeichnet wird, also einer neuen Kulturtechnik, die die notwendige Voraussetzung für einen kompetenten Umgang mit den neuen Medien darstellt.

3.2.2 Quantitative Analyse

Der Computer als number cruncher ist geradezu prädestiniert dazu, zentrales Werkzeug für quantitative Verfahren der Textuntersuchung zu sein. Solche Verfahren sind zwar keineswegs erst für den Computer erfunden worden, aber Umfang des verarbeiteten Materials und Komplexität der Zugriffe können durch ihn deutlich gesteigert werden. Statistische Verfahren sind, das sei gleich vorweg gesagt, lediglich ein Werkzeug, sind also an keine besondere Form der Fragestellung gebunden. Sie führen auch nicht automatisch zu inhaltlichen Ergebnissen, sondern sind stets wiederum Basis einer Interpretation.

Bereits eine längere Geschichte, die im Methodischen sogar vor den Computer zurückreicht, haben zwei Felder der quantitativen Textuntersuchung: die computergestützte Stilanalyse (Stylometrie), die auch für die Autorattribution verwendet wird, und die Inhaltsanalyse.[27] Betrachtet man zum Beispiel die Stylometrie, stellt man fest, daß auf der Suche nach dem digitalen Fingerabdruck ganz verschiedene Textmerkmale herangezogen werden – je nach der zugrundeliegenden Text- und Stiltheorie: charakteristische Unterschiede in der Satzlänge, die Häufigkeit ausgewählter Funktionswörter, Identität und Ort von hapax legomena, also Wörtern die in einem Textkorpus nur einmal vorkommen, das gemeinsame Auftreten von Wörtern oder auch ganze Batterien unauffälligerer Merkmale. Besonders beliebt, da in mehreren Arbeiten erfolgreich erprobt, ist die Untersuchung der häufigsten Wörter eines Textes.

Statistische Verfahren haben mit einer ganzen Reihe von typischen Problemen zu kämpfen. Die Korrektheit der Daten, also die einheitliche Normalisierung, ist eine wesentliche Voraussetzung, um überhaupt Vergleiche anstellen zu können. Die Homogenität und hinreichende Größe der Gesamtmenge und der Stichproben müssen je nach der angezielten Reichweite der Aussagen gewährleistet sein. Die Autor-Zuschreibung von aussagekräftigen Differenzen setzt zum Beispiel voraus, daß man Datenunterschiede berücksichtigt, die durch Epochen- und Textsortenspezifika bedingt sind.[28]

Quantitative Verfahren werden in der Literaturwissenschaft nicht immer gern gesehen. Bestätigen sie gängige Einsichten, stehen sie im Verdacht, überflüssig zu sein. Widersprechen sie aber den üblichen Ansichten, schafft man sie sich mit dem leisen Hinweis vom Hals, daß man Statistiken ohnehin nicht trauen könne. Letztendlich, so kann man mit kaum verhohlener Erleichterung hören, entziehen sich die wesentlichen literaturwissenschaftlichen Fragestellungen der Quantifizierung. Solche Vorbehalte übersehen die Chance, die sowohl in der Bestätigung des Bekannten als auch in der Problematisierung liebgewonnener Vorurteile durch neue und andere Forschungsmethoden liegt; man entzieht sich so einem Diskussionsangebot und dessen gewichtigen Argumenten.

Solche Vorbehalte verfehlen wohl auch den wirklich problematischen Aspekt der quantifizierenden Literaturwissenschaft: Das Zerbrechen einer Forschungsgemeinschaft in eine kleine Gruppe, die statistische Aussagen überprüfen und somit in eine wissenschaftliche Diskussion eintreten kann, und eine sehr viel größere Gruppe, die entweder gläubig annimmt oder ungläubig leugnet. Schon deshalb ist es wichtig, daß computergestützte Text- und Stilanalyse stets ihre Daten offenlegt und ihre interpretatorischen Prämissen und Kontextannahmen sowie ihre Auswertungsverfahren expliziert und somit zur Diskussion stellt.

4. HYPERTEXTTHEORIE UND -PRAXIS

Es existieren sehr verschiedene Definitionsvorschläge für den Begriff ›Hypertext‹;[29] an dieser Stelle soll damit ein Korpus elektronischer Texte bezeichnet werden, die durch Links miteinander verknüpft sind.[30] Links sind Stellen eines elektronischen Text, die durch eine Aktion des Lesenden aktiviert werden können und dann einen anderen Text bzw. eine andere Stelle desselben Texts auf dem Bildschirm erscheinen lassen.[31] Da Links an jeder beliebigen Stelle eines Hypertexts gesetzt werden können und in Hypertexten üblicherweise zahlreiche enthalten sind, wird als ihr besonderes Kennzeichen des öfteren die nicht-sequentielle Lektüre gesehen, die aufgrund der Links möglich ist; manchmal erscheint das Argument auch in verkürzter Form: der Hypertext als nicht-sequentielles Medium.

Zentrale Konzepte des Hypertexts, etwa die beliebige, auch assoziative Verknüpfung von Informationseinheiten durch den Autor oder Leser, gehen auf Überlegungen aus den 40er und 70er Jahren zurück,[32] konnten aber lange Zeit aufgrund der technischen Schwierigkeiten nicht realisiert werden. Erst die leistungsfähigen Personal Computer, die im Laufe der 80er Jahre in größerer Zahl verfügbar wurden, ermöglichten die Präsentation multimedialer Informationen. Erfolgreichstes Beispiel eines Hypertexts ist - seit Anfang der 90er Jahre - das World Wide Web. Anwender können in den Texten und audiovisuellen Informationen, die – weltweit auf Computern verteilt – mit der Auszeichnungssprache HTML (Hypertext Markup Language) kodiert sind, mittels WWW-Browsern navigieren.

Besonderes Interesse hat der Hypertext gefunden, weil er als Realisierung eines Textideals verstanden wurde, das poststrukturalistische Texttheoretiker bereits in den sechziger Jahren entworfen hatten.[33] Neben der Nichtlinearität und Netzwerkstruktur des Textes wird die Verlagerung der Sinnkonstitution vom Autor zum Leser als wichtige Entsprechung gesehen. Inzwischen haben insbesondere empirisch arbeitende Kritiker diese Position mit einer Reihe von Untersuchungen in Frage gestellt, zum Beispiel betonen sie die praktische Relevanz von Orientierungsmitteln, die der Autor eines Hypertexts zur Verfügung stellt.[34] Hypertexte und herkömmliche Texte sind demzufolge weniger als Gegensätze aufzufassen, vielmehr weisen Hypertexte neben den Eigenschaften früherer Textformen auch neue auf und stellen somit eine Herausforderung für moderne Texttheorien dar.[35]

Tatsächlich erweist sich die vielbeschworene Offenheit des

Hypertexts als besonderes Problem für den Autor eines solchen Texts. Für den Leser bedeutet die Konfrontation mit Hypertexten nämlich zuerst einmal eine Steigerung der zu verarbeitenden Komplexität; die daraus entstehenden Orientierungsprobleme müssen einsichtig gelöst werden, damit die Lektüre nicht vorzeitig abgebrochen wird. Die Präsentationsprogramme für Hypertexte enthalten inzwischen zahlreiche Vorrichtungen, die den Überblick im Informationsmeer gewährleisten sollen, zum Beispiel ein oder mehrere Inhaltsverzeichnisse, zum schnellen Wiederfinden bereits besuchter Informationsknoten die History- bzw. Bookmark-Funktion. Auch das Annotieren des elektronischen Texts und die Suche nach solchen Zusätzen, ja überhaupt das gesamte Information Retrieval kann zur Orientierung dienen. Aber neben der Abstimmung des Texts auf diese Navigationsmöglichkeiten hin muß der Autor auch durch den Einsatz von Links eine sequentielle Lektüre organisieren und zugleich für die inhaltliche Konsistenz der nicht-sequentiellen Informationswege sorgen, die mittels der Links angeboten werden. Besonders problematisch – und für den Anwender ein besonderes Orientierungsproblem – ist der Umstand, daß Links eine rein mechanische Verknüpfung zweier Informationseinheiten darstellen. Welche Bedeutung diese Verknüpfung hat, ob es sich einfach um eine Weiterführung der Lektüre, eine Anmerkung, einen Exkurs, eine Erläuterung, einen Beleg oder anderes mehr handelt, ist offen und muß stets erst erschlossen werden, wenn nicht das Design des Hypertexts diese Beliebigkeit organisiert.

Ein besonderes Arbeitsfeld für die Philologie ist die Untersuchung von fiktionalen Hypertexten (Hyperfictions), die sich als eigenständige Kunstform etabliert haben und die neuen Möglichkeiten des Mediums experimentell explorieren.[36] Man sollte übrigens den Begriff Hyperfiction weit genug fassen, um darunter nicht nur die eher esoterische digitale Experimentalliteratur, sondern auch die sehr weit verbreiteten fiktionalen Rollen- und Abenteuerspiele zu begreifen, die inzwischen mit einem Aufwand hergestellt werden, wie er sonst nur aus Filmproduktionen bekannt ist.

5. SOFTWARE FüR PHILOLOGEN UND LITERATURWISSENSCHAFTLICHES PROGRAMMIEREN

Wie bereits in Abschnitt 1 über Abstraktion ausgeführt, ist die Beschreibung konkreter Programme als wissenschaftliches Unternehmen nur wenig sinnvoll, da die Software sich zumeist schon bis zur Veröffentlichung wieder verändert hat. An dieser Stelle soll auch nur eine Typologie der philologischen Verwendung von Software versucht werden.

Die meisten Philologen benutzen lediglich fertige Programmpakete, etwa die Retrievalprogramme, mit denen elektronische Texte üblicherweise ausgeliefert werden. Eine nicht geringe Zahl verwendet zwar keine Programmier-,[37] aber doch Skript-[38] und Makrosprachen.[39] Die Anpassung vorliegender Programme an die eigenen Wünsche mittels Makros oder komplexerer programmspezifischer Konfigurationen bestimmt wohl den größten Teil des sonstigen computerphilologischen Arbeitens. Nur sehr wenige aber schreiben selbst Programme, noch weniger in einem solchen

Umfang, daß sie diese als eigenständige Applikationen anderen zugänglich machen können.[40]

Der Markt für philologische Software ist zu klein, um für größere kommerzielle Unternehmungen attraktiv zu sein. Die erfolgreicheren Anwendungen in diesem Arbeitsfeld sind daher auch zumeist an Hochschulen entwickelt worden, zum Beispiel das deutsche Satzprogramm und Autorensystem zur Textbearbeitung TUSTEP in Tübingen,[41] das Textanalyseprogramm TACT[42] oder das Kollationierungsprogramm Collate.[43] Ebenso häufig ist die Verwendung kommerzieller Software für philologische Belange; so wird etwa zur Zeit Information Management Software wie FolioViews für elektronische Editionen verwendet, da solche kommerziellen Systeme den Datenimport aus der vertrauten Textbearbeitung leicht machen und sehr leistungsfähige Information Retrieval Funktionen bieten. Hierbei zeigt sich ein prinzipielles Problem: Kommerzielle Software entspricht dem Stand der Technik, insbesondere was die Benutzerschnittstelle, Arbeitsmöglichkeiten und -geschwindigkeit betrifft. Allerdings ist sie fast immer auf die Bedürfnisse einer anderen Benutzergruppe zugeschnitten und an philologische Sonderwünsche, etwa die Visualisierung von Handschrift mit Transkription und Apparat oder die parallele Bewegung in zwei Textfenstern, kaum anzupassen. Die Software von Philologen dagegen ist sehr gut auf die spezifischen Bedürfnisse des Faches abgestimmt, bleibt aber aufgrund der sehr viel langsameren Entwicklung des Programms hinter den Möglichkeiten und Usancen der Gegenwart zurück – insbesondere was die Gestaltung der Benutzerschnittstelle und die Verwendung von Standards zur Textauszeichnung betrifft. [44]

Da kommerzielle Programme nur selten von vornherein zur Lösung philologischer Probleme taugen, verwenden viele die ihnen bekannte Software und passen sie mittels Makros an ihre Wünsche an.[45] Der unbestreitbare Vorteil liegt darin, daß die Anwender in einer ihnen vertrauten Arbeitsumgebung arbeiten können. Mindestens ebenso gewichtig sind jedoch die Nachteile: Die Programme stoßen sehr schnell an die Grenzen ihrer Leistungsfähigkeit; dann bestimmt oft die Software, was gemacht wird, und nicht die Frage, was philologisch wünschenswert ist. Zudem können die Makros schon mit der nächsten Version der verwendeten Software unbrauchbar werden.

Nur die wenigsten Philologen programmieren ihre Werkzeuge zur Problemlösung selbst. Versuche, literaturwissenschaftliches Programmieren zu etablieren, hatten bislang kaum Erfolg. [46] Ursache dafür ist nicht zuletzt der Entwicklungsstand der Software-Industrie, die lange Zeit den Programmierern lediglich die Sprache, den Compiler und Funktionsbibliotheken zur Verfügung stellte, sie aber letztendlich zwang, ihre Aufgabenstellung relativ maschinennah zu formulieren. Schon ein kurzer Blick in die Lehrbücher zu Programmieralgorithmen zeigt, welchen Umfang Überlegungen zu effizienten Such- und Sortierroutinen dort einnehmen. Wer als Programmierer aber Probleme auf dieser Ebene lösen muß, wird nur, wenn er über viel Zeit und Arbeitskraft verfügt, zu den Problemen kommen, die für Geisteswissenschaftler eigentlich interessant sind. Erst die Entwicklung der letzten fünf Jahre hat einem Paradigma

zum Durchbruch verholfen, das diese Situation auch für das philologische Programmieren ändern kann: die Komponententechnik.[47] Sie erlaubt die Kapselung komplexer Bearbeitungsroutinen und Datenbehälter in einem ›Fertigbauteil‹, das dem Anwender seine Funktionalität zur Verfügung stellt, zum Beispiel Sortierroutinen für alle Zeichen aller philologisch wichtigen Sprachen. Selbst ein wenig erfahrener Programmierer kann dann mehrere solcher Komponenten für seine spezifische Problemlösung zusammensetzen. Inzwischen sind mehrere Überlegungen im Gange, wie Philologen international kooperieren können, um Werkzeuge für das literaturwissenschaftliche Programmieren bereitzustellen. [48]

6. PERSPEKTIVEN DER COMPUTERPHILOLOGIE FüR DIE LITERATURWISSENSCHAFT

Die voranstehende Skizze hat die Tätigkeitsfelder des Fachteils ›Computerphilologie‹ umrissen; bleibt also nur noch zusammenfassend zu fragen, wie der Geschäftsverkehr zwischen der allgemeinen Literaturwissenschaft und der Computerphilologie aussieht beziehungsweise aussehen sollte.

Ein wichtiger Service der Computerphilologie ist die Bereitstellung von Anwendungen und Werkzeugen für die Literaturwissenschaft, wie es mustergültig die Programmierer von TACT und TUSTEP geleistet haben. Die Herstellung von philologischer Software ist aber nur ein Teil computerphilologischer Dienstleistungen: Textcenter stellen der Forschung gesicherte Texte zur Verfügung, Kommunikationsforen wie die Diskussionsgruppe HUMANIST und die Website der Computerphilologie ermöglichen nicht nur die Diskussion zwischen den Fachwissenschaftlern, sondern auch über die disziplinären Grenzen hinaus. Noch wichtiger als solche Sachleistungen ist aber die Wissensvermittlung, damit Literaturwissenschaftlern ein kritischer Umgang mit den neuen Medien ermöglicht wird; sachlich, gleich weit von Euphorie und Verteufelung entfernt, soll über die Möglichkeiten und Limitierungen der neuen Werkzeuge informiert werden. Jede Anleitung zur Verwendung des Internets und fast jedes Handbuch einer besseren elektronischen Edition enthält inzwischen das Grundwissen des Information Retrievals, aber es ist eine wichtige Aufgabe der Computerphilologie, fortgeschrittenere Verwendungsweisen zu vermitteln, Kriterien der Evaluation elektronischer Texte zu bestimmen und auf typische Probleme und Stolpersteine aufmerksam zu machen.

Nicht zuletzt ist die Computerphilologie aber auch eine theoretische Herausforderung für die Literaturwissenschaft, da sie klassische Fragestellungen mit anderen Methoden angehen kann. So wurde die moderne Texttheorie durch den Hypertext und seine Theoretiker veranlaßt, ihre Bestimmungen zu überprüfen: Sind die neuen Medien tatsächlich Belege für die gewünschte Ermächtigung des Lesers und die mediale Wiedergabe der Netzstruktur unseres Wissens und Denkens? Auch die literaturtheoretische Diskussion um den Autor mit der zentralen These, der ›Autor‹ sei lediglich ein Begriffskonstrukt des Lesers, wird von computerphilologischer Seite weitergeführt. Quantitative Vergleichsuntersuchungen konnten zeigen, daß Texte vom selben Autor Merkmale gemein haben, die eine deutliche Unterscheidung von Texten anderer Autoren

ermöglicht, der Autorbegriff also keine bloße Konstruktion auf seiten des Lesers darstellt.[49]

Die Computerphilologie ist insgesamt also ein Arbeitsfeld, das auf interdisziplinäre Kooperation hin ausgerichtet ist: Verwurzelt im Wissensstand einer Philologie integriert sie Arbeitsmethoden der quantitativen Forschung und der angewandten Informatik. Zugleich kooperiert sie mit anderen Philologien und geisteswissenschaftlichen Fächern im übergreifenden Diskussionsforum des Humanities Computing. Die Diskussion dient nicht nur dem Austausch des technischen Wissens, sondern vor allem auch der Information darüber, wie die beteiligten Disziplinen ihre Fragestellungen für die Bearbeitung mit dem Computer operationalisieren. So stellt das Humanities Computing und – darin enthalten – auch die Computerphilologie eine der wenigen langfristigen interdisziplinären Kooperationen dar.

ANHANG: INFORMATIONSQUELLEN ZUR COMPUTERPHILOLOGIE

Die Computerphilologie beginnt sich eben erst als eigenständiger Teilbereich des Humanities Computing auszudifferenzieren und nutzt dessen Kommunikationswege und Informationsquellen. Das sind zum einen die Zeitschriften Computers and the Humanities (seit 1966) und Literary and Linguistic Computing (seit 1973) und die gemeinsamen Tagungen der zugehörigen Verbände Association for Computers and the Humanities[50] (ACH) und Association for Literary and Linguistic Computing[51] (ALLC). Unverzichtbar für die aktuelle Information ist außerdem die Diskussionsliste HUMANIST.[52] Einen guten Überblick über die Forschung bietet die kumulative Auswahlbibliographie zur einschlägigen (insbesondere englischsprachigen) Literatur bis 1996.[53]

Forschungszentren, von denen einige inzwischen auch Aufbaustudiengänge anbieten, institutionalisieren das Humanities Computing. Zu erwähnen sind hier insbesondere das Centre for Computing in the Humanities des King's College London,[54] die Humanities Computing Unit in Oxford mit ihren verschiedenen Komponenten[55] und das Humanities Advanced Technologies and Information Institute in Glasgow.[56] Ein gewichtiges Forschungszentrum außerhalb der angelsächsischen Welt ist ›The Humanities Information Technologies Research Programme at the University of Bergen‹.[57]

Für zahlreiche Belange sind auch die elektronischen Textcenter gute Anlaufadressen; besonders bekannt sind das der University of Virginia Library,[58] das Center for Electronic Texts in the Humanities der Rutgers University in New Jersey[59] und das Oxford Text Archive. [60]

In Deutschland ist die wichtigste Adresse die Abteilung Literarische und Dokumentarische Datenverarbeitung des Zentrums für Datenverarbeitung an der Universität Tübingen, wo unter der Leitung von Wilhelm Ott seit 1973 Kolloquien über die Anwendung der EDV in den Geisteswissenschaften stattfinden.[61] Die Website Computerphilologie,[62] eine elektronische Ergänzung zum vorliegenden Jahrbuch, gibt Hinweise auf die einschlägigen Webinformationen und stellt

zudem ein Diskussionsforum für praktische und theoretische Probleme zur Verfügung.

---

[1] Die Verwendung des Computers bleibt in diesen Fällen im wissenschaftlichen Text zumeist unsichtbar, da die Ergebnisse solcher Überprüfungen entweder verworfene Arbeitshypothesen oder einzelne Belegstellen sind.
[2] Bislang fangen nicht wenige in diesem Arbeitsfeld – ganz entgegen den sonstigen wissenschaftlichen Gepflogenheiten – immer wieder von vorne an und erfinden sozusagen das virtuelle Rad immer wieder neu.
[3] Vergleiche Roberto Busa: The Annals of Humanities Computing. The Index Thomisticus. In: Computers and the Humanities 14 (1980), S. 83-90. – Der vorliegende Text beschreibt ein Forschungsfeld und ist kein Forschungsbericht; Literaturhinweise werden daher nur relativ sparsam gegeben. Weiterführende Angaben findet man unten im Anhang ›Informationsquellen zur Computerphilologie‹ und in der Bibliographie am Ende dieses Jahrbuchs.
[4] So immer noch Winfried Lenders/Gerd Willée: Linguistische Datenverarbeitung. Ein Lehrbuch. Opladen, Wiesbaden: Westdeutscher Verlag 1998, S. 11f.
[5] Der Begriff ›Computerphilologie‹ als Analogbildung zu ›Computerlinguistik‹ wurde meines Wissens erstmals verwendet von Rolf Bräuer: Zu den Aufgaben und Möglichkeiten der Computerlinguistik und Computerphilologie an den Universitäten und Hochschulen der DDR. In: Ernst-Moritz-Arndt-Universität Greifswald (Hg.): Wissenschaftliche Beiträge der Ernst-Moritz-Arndt-Universität. Sprache und Computer 1. Neue Forschungsperspektiven durch Computereinsatz in den Gesellschaftswissenschaften. 1. Tagung ›Sprache und Computer‹ am 18. und 19. November 1986 in Greifswald. Greifswald: 1988, S. 9-12.
[6] Manfred Krifka/Bernd Gregor: Einleitung. In: M.K./B.G. (Hg.): Computerfibel für die Geisteswissenschaften. Einsatzmöglichkeiten des Personal Computers und Beispiele aus der Praxis. München: Beck 1986, S. 43.
[7] Vergleiche Thomas A. Standish: Data Structures in Java. Reading u.a.: Addison-Wesley 1998, S. 6.
[8] Mit ›abstrakteren Notationsweisen‹ sind Vereinbarungen zur Verschriftlichung von Programmen gemeint, die nicht an eine bestimmte Programmiersprache gebunden sind; für prozedurales Programmieren Pseudo-Code (ein nicht formalisiertes Verfahren) und für objektorientiertes Programmieren UML (Unified Modeling Language).
[9] Im Falle von relationalen Fachdatenbanken sind dies die Angaben zur Struktur der Datenbank. Tatsächlich ist dieser Typ Datenbank mit einer genau vorgegebenen Struktur für Literaturwissenschaftler eher selten nützlich, da ihre Daten, die literarischen Texte, sich nur selten in solch eine Ordnung bringen lassen und daher zumeist in Freitextdatenbanken aufbewahrt werden. Es gibt noch andere strukturierte Datenbankformate, aber die relationale Datenbank ist im Augenblick am weitesten verbreitet; vergleiche zu Datenbankmodellen Abraham Silberschatz/Henry F. Korth/S. Sudarshan: Database System Concepts. New York u.a.: McGraw Hill ^{3}1996.
[10] Ein dritter Weg besteht in der Konvertierung von bereits bestehenden Daten, etwa von Satzbändern oder proprietär ausgezeichneten Texten.
[11] Übliche Zeichensätze sind ASCII (128 Zeichen) oder ISO 8879-1 (256 Zeichen). Diese kleineren Zeichensätze, die für Literaturwissenschaftler, die griechisch, hebräisch, kyrillisch schreiben wollten, meist unnötige Umwege bedeuteten, werden nun abgelöst durch den Standard Unicode bzw. ISO/IEC 10646 (in Version 2.1 sind 38, 887 Zeichen definiert), vgl. <http://www.unicode.org> (12.6.1999).
[12] Das Bild wird in kleinere Bild-Einheiten zerlegt, in Textblöcke, Zeilen und Grapheme. Die Identifikation des Buchstabens geschieht dann zumeist aufgrund von Regelwissen über den Aufbau aller Buchstaben. Solche Software hat den Vorteil sofort einsetzbar zu sein, kann aber nur für eine bestimmte Schriftfamilie eingesetzt werden. Man kann so alle Antiqua-Texte bearbeiten, nicht aber Texte, die in Fraktur gedruckt sind. Das Gegenstück bildet OCR-Software, der der Anwender die Zuordnung eines Graphems zu einem Buchstaben für jedes Graphem erst antrainieren muß, die aber dafür prinzipiell jede Schrift lernen kann.
[13] SGML ist ein Standard, mit dem Auszeichnungssprachen definiert werden können, zum Beispiel HTML oder eben TEI. Der Vorteil von SGML ist, daß jeder Text, der mit einem so definierten System ausgezeichnet wurde, mit allen SGML-konformen Programmen weiterverarbeitet werden kann. Zu SGML vgl.

Wolfgang Rieger: SGML für die Praxis. Ansatz und Einsatz von ISO 8879. Berlin, Heidelberg, New York: Springer 1995. Der Nachteil der SGML ist ihre ausgesprochen hohe Komplexität, deshalb wurde inzwischen ein Subset von SGML namens XML (eXtensible Markup Language) definiert, das sehr viel einfacher anzuwenden ist. Die TEI Spezifikation wird zur Zeit nach XML portiert; vgl. zu XML die Spezifikation des World Wide Web Consortium <http://www.w3.org/> (10.6.1999) und außerdem <http://www.xml.com/> (10.6.1999).

[14] Vergleiche das Handbuch zur TEI von Lou Burnard und Michael Sperberg-McQueen: Guidelines for Electronic Text Encoding and Interchange. 1995. <http://www-tei.uic.edu/orgs/tei/p3/elect.html> (12.6.1999). Neben den Handbüchern und den Grammatikdateien findet man auf der TEI-Homepage auch eine Liste von Editionsprojekten, die mit TEI arbeiten: <http://www.uic.edu:80/orgs/tei/> (12.6.1999). Die organisatorische Leitung der TEI hat ein Konsortium aus Universitäten übernommen: <http://www.tei-c.org/> (12.6.1999). Eine Einführung in den Teilbereich von TEI, der zur Auszeichnung von Editionen dient, findet man bei Peter Robinson: The Transcription of Primary Textual Sources. Oxford: Office for Humanities Communication Publications 1994. Einen knappen Überblick über die Funktionsweise von TEI gibt Fotis Jannidis: Wider das Altern elektronischer Texte. Philologische Textauszeichnung mit TEI. In: editio 11 (1997), S. 152-177. *Computers and the Humanities* hat anläßlich des 10. Geburtstages von TEI ein zweites Sonderheft mit einschlägigen Aufsätzen veröffentlicht, in dem auch schon der Übergang von SGML zum Subset XML diskutiert wird: vgl. Computers and the Humanities 33,1 (1999). Das erste Sonderheft zu TEI erschien 1995 als Band 29,3.

[15] Satzzeichen werden zumeist als Trennzeichen verwendet. Wird etwa die mittlere Satzlänge eines Texts untersucht, dann muß das Programm zwischen Punkten als Satzbegrenzer und als Abkürzungssignal unterscheiden können.

[16] Zu einem Überblick über diese Probleme am Beispiel der Erstellung von Wörterbüchern und Indices vgl. Kurt Gärtner/Peter Kühn: Indices und Konkordanzen zu historischen Texten des Deutschen: Bestandsaufnahmen, Typen, Herstellungsprobleme, Benutzungsmöglichkeiten. In: Werner Besch/Anne Betten/Oskar Reichmann/Stefan Sonderegger (Hg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Berlin, New York: de Gruyter ^{2}1998, S. 715-742.

[17] Diese Wahrnehmung wird im folgenden besonders am Beispiel der Bildschirmlektüre erörtert, es gibt jedoch inzwischen Möglichkeiten, den Text vorlesen zu lassen.

[18] Vergleiche den Beitrag von Rainer Baasner in diesem Band. Als Beispiel für eine gelungene Integration von Text- und Bild vgl. das virtuelle Seminar zur englischen Lyrik des 1. Weltkriegs <http://info.ox.ac.uk/jtap/> (1.6.1999).

[19] Etwas häufiger sieht man inzwischen die Verknüpfung von Bildern der Handschriften mit elektronischen Texten.

[20] Vergleiche den Beitrag von Walter Morgenthaler in diesem Band.

[21] Vergleiche Nancy Ide/J. Véronis: Word Sense Disambiguation: The State of the Art. Computational Linguistics 24 (1998), S. 1-40. Jan de Vuyst: Knowledge Representation for Text Interpretation. In: Literary and Linguistic Computing 5 (1990), S. 296-302. Außerdem:

[22] So eine treffende Formulierung von John F. Burrows.

[23] Zu regulären Ausdrücken vgl. Jeffrey E.F. Friedl: Mastering Regular Expressions. Cambridge u.a.: O'Reilly 1997.

[24] Die Notation der Operatoren ist selbstverständlich programmspezifisch, hier wird nur eine verdeutlichende Schreibweise gewählt.

[25] Die genannten Verfahren sind lediglich eine kleine und eher triviale Teilmenge dessen, was in der Forschung diskutiert wird. Vergleiche dazu: Norbert Fuhrs sehr umfangreiches Skript von 1996 zur Vorlesung ›Information Retrieval‹ <http://ls6-www.informatik.uni-dortmund .de/ir/teaching/courses/ir/> (1.6.1999)

[26] Siehe Catherine N. Ball: Automated Text Analysis. Cautionary Tales. In: Literary and Linguistic Computing 9 (1994), S. 295-302.

[27] Nur erwähnt sei von den zahlreichen anderen Computeranwendungen in der Literaturwissenschaft die statistische Auswertung von Leserreaktionen; vgl. etwa Christoph Meister im vorliegenden Band. Oder auch Studien zur Metrik und zur Prosodie. – Zur Anwendung quantitativer Verfahren in der Literaturwissenschaft gibt es mehrere empfehlenswerte Überblicksdarstellungen: Rosanne G. Potter: Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966-1990. In: Computers and the Humanities

25 (1991), S. 401-29. John F. Burrows: Computers and the Study of Literature. In: Christopher S. Butler (Hg.): Computers and Written Texts. Oxford: Blackwell 1992, S. 167-204. Thomas Rommel: And Trace it in this Poem Every Line. Methoden und Verfahren computerunterstützter Textanalyse am Beispiel von Lord Byrons Don Juan. Tübingen: Gunter Narr 1995, insbesondere S. 26-67.

[28] Vergleiche Joseph Rudman: The State of Authorship Attribution Studies. Some Problems and Solutions. In: Computers and the Humanities 31 (1998), S. 351-365.

[29] Die Sekundärliteratur zum Thema Hypertext ist so umfangreich, daß die Feststellung dieser Menge ebenfalls schon topisch geworden ist. Eine bibliographische Übersicht der englischsprachigen Literatur ist unter <http://www.gwu.edu/~gelman/train/hyperbib.htm> (10.5.1999) zu finden (Stand 12/97). Sucht man in der elektronischen Bibliographie der *Modern Language Association* jahrweise nach dem Stichwort ›Hypertext‹, dann wird deutlich, daß die Diskussion 1996 ihren Höhepunkt erreicht hatte und inzwischen deutlich zurückgegangen ist.

[30] Üblich sind zur Zeit nur 1 -> 1 Links, also Verbindungen von einem Punkt zu einem anderen Punkt. Nur teilweise realisiert sind die anderen Möglichkeiten n -> 1, 1 -> n (wobei n für eine beliebige Zahl steht), sowie 1 <-> 1 usw.

[31] Wenn in dem Korpus auch Bilder, Filme oder Töne enthalten sind, spricht man auch von ›Hypermedia‹. Wenn im weiteren von Hypertext die Rede ist, dann sollen Hypermedia mitgemeint sein.

[32] Das erste Modell eines solchen Hypertexts hat gleich nach dem Ende des Zweiten Weltkriegs Vannevar Bush entworfen; vgl. Vannevar Bush: As We May Think. In: Atlantic Monthly 176,1 (Juli 1945), S. 101-108. Jetzt auch: <http://www.theatlantic.com/unbound /flashbks/ computer/bushf.htm> (2.5.1999). Ein wichtiger Vordenker ist außerdem Theodor H. Nelson: Literary Machines. Swarthmore: o.A. 1981.

[33] Locus classicus dieser Auffassung ist George P. Landow: Hypertext. The Convergence of Contemporary Critical Theory and Technology. Baltimore, London: Johns Hopkins University Press 1992. Überarbeitet und erweitert: Hypertext 2.0. Baltimore u.a.: Johns Hopkins University Press 1997.

[34] Vergleiche Jean-François Rouet u.a. (Hg.): Hypertext and Cognition. Mahwah, N.J.: Lawrence Erlbaum 1996, und darin insbesondere den Beitrag von Andrew Dillon: Myths, Misconceptions, and an Alternative Perspective on Information Usage and the Electronic Medium, S. 25-42.

[35] Vergleiche Simone Winko: Lost in hypertext? Autorkonzepte und neue Medien. In: Fotis Jannidis u.a. (Hg.) Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs. Tübingen: Niemeyer 1999, S. 509-532.

[36] Vergleiche Jürgen Daiber im vorliegenden Band und die Hinweise in *Hyperizons. Hypertext Fiction*: <http://www.duke.edu/~mshumate/hyperfic.html> (10.6.1999).

[37] Programmiersprachen erlauben in sehr hardwarenaher (zum Beispiel Assembler) oder eher problemorientierter Weise (zum Beispiel C++, Java) die Erstellung von Programmen, die anschließend in einen Binärcode übersetzt werden (Kompilierung), der einem Prozessor (oder, im Falle von Java, einer virtuellen Maschine) angepaßt ist.

[38] Skriptsprachen, zum Beispiel Perl oder Python, werden nicht kompiliert, sondern als Klartext einem Interpreter übergeben. Sie werden insbesondere für kleinere, einmalig anfallende Problemlösungen bevorzugt.

[39] Im Gegensatz zu Programmen und Skripten, die direkt innerhalb des Betriebssystems ablaufen, laufen Makros innerhalb eines Kontext-Programms ab. Ihr besonderer Vorteil ist die Nutzung der speziellen Leistungsfähigkeit des Programms für die Lösung des anstehenden Problems.

[40] Bekanntlich ist es weniger ein Problem, einen arbeitenden Prototypen zu programmieren, als ein zur Weitergabe taugliches Programm zu schreiben, das bei allen möglichen Benutzereingaben sinnvoll reagiert.

[41] Das Tübinger Textbearbeitungssystem TUSTEP, das bereits seit über 20 Jahren im Einsatz ist, wird unter der Leitung von Wilhelm Ott ständig weiterentwickelt und auf neue Plattformen portiert. Es stellt dem Philologen für zahlreiche Vorarbeiten bei der Erstellung einer elektronischen Edition Bearbeitungsmodule zur Verfügung und kann auch als Satzprogramm verwendet werden. Der Komplexität der Anwendungsmöglichkeiten entspricht allerdings auch die sehr komplexe Benutzerschnittstelle, daher wählen nicht wenige Philologen weniger leistungsfähige, aber zugänglichere Autorensysteme. Zu TUSTEP vgl. <http:// www.uni-tuebingen.de/zdv/zrlinfo/tustep-des.html> (1.6.1999). Siehe auch Wilhelm Ott: Edition und Datenverarbeitung. In: Herbert

Kraft (Hg.): Editionsphilologie. Darmstadt: Wissenschaftliche Buchgesellschaft 1990, S. 59-70, und Winfried Bader (Bearb.): Lernbuch TUSTEP. Einführung in das Tübinger System von Textverarbeitungsprogrammen. Tübingen: Niemeyer 1995.

[42] TACT ist frei erhältlich unter: <http://www.cch.epas.utoronto.ca:8080/cch/1001h/06soft. html> (1.6.1999).

[43] Zu Collate vgl. <http://www.dlib.dmu.ac.uk/projects/Collate/> (1.6.1999).

[44] Vergleiche dazu im vorliegenden Band Walter Morgenthaler.

[45] Für ein recht ausgefallenes Beispiel für solche Anpassung vergleiche Eckhardt Meyer-Krentler: Edition & EDV. Elektronische Arbeitshilfen für Editoren, Philologen, Bücherschreiber mit dem WORD-Zusatzpaket ECCE. München: Fink 1992.

[46] Es gibt dennoch immer wieder Vermittlungsversuche, zum Beispiel Hans-Werner Ludwig: EDV für Literaturwissenschaftler. Arbeits- und Programmiertechniken für den PC. Tübingen: Gunter Narr 1991.

[47] Zwei erprobte Standards zur Komponenten stehen zur Verfügung: Plattformunabhängig JavaBeans und – nur für MS Windows – ActiveX Controls.

[48] Vergleiche die Hinweise bei Matthias Kopp, Marc Wilhelm Küster, Wilhelm Ott: TUSTEP im WWW-Zeitalter. Werkzeug für Anwender und Programmierer. <http://www.uni-tuebingen.de/zdv/zrlinfo/prot/prot74-lddv.html> (1.6.1999) und auf der Website der *Computerphilologie*.

[49] Vergleiche John F. Burrows: Computers and the Idea of Authorship. In: Fotis Jannidis u.a. (Hg.): Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs. Tübingen: Niemeyer 1999, S. 165-180; und Colin Martindale: What Can Texts Tell Us About Authors and What Can Authors Tell Us About Texts? Ebda., S. 181-206.

[50] <http://ach.org/> (10.6.1999).

[51] <http://www.kcl.ac.uk/humanities/cch/allc/> (10.6.1999).

[52] <http://www.princeton.edu/~mccarty/humanist/> (10.6.1999).

[53] <http://www.kcl.ac.uk/humanities/cch/bib/> (10.6.1999). Zum wissenschaftlichen Publizieren im Internet siehe <http://info.lib.uh.edu/sepb/sepb.html> (10.6.1999).

[54] <http://www.kcl.ac.uk/humanities/cch/> (10.6.1999).

[55] <http://www.oucs.ox.ac.uk/humanities/> (10.6.1999).

[56] <http://www.arts.gla.ac.uk:80/HATII/> (10.6.1999).

[57] <http://www.hit.uib.no/english/> (10.6.1999).

[58] <http://etext.lib.virginia.edu/> (10.6.1999).

[59] <http://www.ceth.rutgers.edu/> (10.6.1999).

[60] <http://ota.ahds.ac.uk/> (10.6.1999).

[61] Eine Liste aller stattgefundenen Kolloquien mit den Vortragstexten der letzten Jahre findet man unter <http://www.uni-tuebingen.de/zdv/zrlinfo/kolloq.html> (10.6.1999).

[62] <http://computerphilologie.uni-muenchen.de>
