Achtung!
Dies ist eine Internet-Sonderausgabe des Aufsatzes
„TITUS. Das Projekt eines indogermanistischen Thesaurus“
von Jost Gippert (1995).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
LDV-Forum 12 / 1, 1995, 35-47
zu entnehmen.

Attention!
This is a special internet edition of the article
“TITUS. Das Projekt eines indogermanistischen Thesaurus”
[“TITUS. The project of an Indo-European thesaurus”]
by Jost Gippert (1995).
It should not be quoted as such. For quotations, please refer to the original edition in
LDV-Forum 12 / 1, 1995, 35-47.

TITUS

DAS PROJEKT EINES INDOGERMANISTISCHEN THESAURUS

Jost Gippert
Vergleichende Sprachwissenschaft
Universität Frankfurt
http://titus.uni-frankfurt.de

Auf der indogermanistischen Fachtagung in Leiden (September 1987) berieten einige der Teilnehmer darüber, inwieweit es möglich sei, im Hinblick auf die an verschiedenen Forschungsstätten des In- und Auslands angelaufenen oder laufenden Projekte, die die Einspeicherung von für die Vergleichende Sprachwissenschaft relevanten Texten auf Datenträger zum Ziel hatten, zu einer Zusammenarbeit zu kommen. Man war sich einig, daß es wenig sinnvoll sei, wenn jeder einzelne Forscher in diesem Bereich allein und für sich arbeiten würde, weil dies immer wieder zu einer überflüssigen Duplizierung des Aufwands führen müsse; statt dessen sei es zweckmäßig, die Eingabe von Texten von vornherein zu koordinieren.

Damit war die Idee zu einem — innerhalb der Indogermanistik beispiellosen — Gemeinschaftsprojekt geboren, das schon bald darauf, mit einem Aufruf in "Die Sprache" 32/2, 1987, unter dem Namen eines "Thesaurus indogermanischer Textmaterialien auf Datenträgern" ins Leben gerufen wurde. Nach einer nunmehr achtjährigen Laufzeit ist die primäre Zielsetzung des Projekts, nämlich über den gesamten die Grundlage der Vergleichenden Indogermanischen Sprachwissenschaft darstellenden Textbestand aus altüberlieferten Sprachen wie dem Altindischen (Sanskrit), dem Altiranischen (Avesta, Altpersisch), dem Altgriechischen, dem Lateinischen, altanatolischen Sprachen wie dem Hethitischen, altgermanischen Sprachen wie dem Althochdeutschen oder Altenglischen in einer für die elektronische Analyse zugänglichen Form zu verfügen, in greifbare Nähe gerückt. Im Zuge der ständig wachsenden Kapazitäten von Hard- und Software haben sich Zielsetzungen und Perspektiven des Projekts im gleichen Zeitraum jedoch ebenfalls erheblich ausgeweitet, was nicht zuletzt die neue Namensgebung "Thesaurus indogermanischer Sprach- und Textmaterialien (TITUS)" reflektieren soll, unter der es seit der 3. Tagung für Computereinsatz in der Historisch-Vergleichenden Sprachwissenschaft (Dresden 1994) geführt wird. Aus Anlaß dessen, daß sich auf der letztgenannten Tagung zugleich auch ein eigener Arbeitskreis der GLDV mit dem Titel "Historisch-vergleichende Sprachwissenschaft" konstituierte, dessen Mitglieder zum größten Teil am TITUS-Projekt beteiligt sind, erscheint es angebracht, die Entwicklung, den gegenwärtigen Stand und die nächsten Vorhaben des Projekts hier kurz darzustellen.

Das primäre Ziel des TITUS-Projekts bestand, wie bereits gesagt, in der koordinierten Erfassung der relevanten Originaltexte altüberlieferter indogermanischer Sprachen. Natürlich war die Idee, derartige Texte dem sich seinerzeit eher mühsam durchsetzenden neuen elektronischen Medium "anzuvertrauen", nicht wirklich neu; tatsächlich hatte es ja schon seit den 60er Jahren verschiedene Projekte (v.a. in USA) gegeben, die auf eine (vollständige oder teilweise) Erfassung von Textdaten auch altüberlieferter Sprachen abzielten. Unter ihnen ist zunächst das Projekt des "Thesaurus Linguae Graecae" zu nennen, das eine Erfassung sämtlicher klassisch- und spätklassisch-griechischer Textmaterialien seit dem Beginn der Überlieferung (Homer) bis in die frühmittelalterliche byzantinische Zeit anstrebt und inzwischen das gesteckte Ziel zu ca. 80% erreicht hat. Da das Griechische innerhalb der indogermanischen Sprachfamilie als einer der bedeutendsten Vertreter gelten kann, war mit dem TLG-Projekt von vornherein ein möglicher Kooperationspartner gegeben, der in mancherlei Hinsicht auch als Vorbild dienen konnte; im Bereich des Griechischen reduzierten sich die für den indogermanistischen Thesaurus verbleibenden Aufgaben auf die vorklassische Überlieferung (der mykenischen Epoche) sowie auf die Überlieferung von Textmaterialien aus den altgriechischen Dialekten, die freilich bis heute noch nicht in befriedigendem Umfang bearbeitet sind.

Den eigentlichen Grundstock der Textsammlung innerhalb des Thesaurus bildeten demgegenüber zwei Texte aus dem indoiranischen Sprachzweig, die von höchstem indogermanistischem Interesse sind und deren elektronische Bearbeitung der Konstituierung des Projektes vorausging. Es handelt sich zum einen um die altindische Ṛgveda-Saṃhitā, die bereits in den 70er Jahren unter der Leitung von W.P. LEHMANN an der University of Texas eingegeben worden war, zum anderen um das altiranische Avesta-Corpus, das im Hinblick auf die Erstellung einer Textkonkordanz im Rahmen eines von der DFG geförderten Projekts durch S. GIPPERT-FRITZ an der Freien Universität Berlin elektronisch aufbereitet wurde. An diesen beiden Einzelprojekten lassen sich bereits die wesentlichsten Probleme aufzeigen, die während der Anfangsphase des Thesaurus-Projekts zu lösen waren.

Der Ṛgveda-Text war, wie damals nicht anders möglich, ursprünglich auf einer Großrechneranlage eingegeben worden. An eine Eins-zu-eins-Wiedergabe in der schriftlichen Form, in der der Text selbst überliefert worden ist, d.h. in einem indischen Alphabet wie der Devanāgarī, war damals nicht zu denken; es wäre allerdings seitens der (sprachwissenschaftlich orientierten) Fachwelt auch gar nicht unbedingt erwünscht gewesen, da sich diese seit dem Vorliegen der Edition von Th. AUFRECHT ("Die Hymnen des Rigveda", ^{2}1877) an den transkribierten Text gewöhnt hat. Aber auch eine solche Transkription war nicht eins zu eins auf den Großrechner abbildbar, da sie von zahlreichen diakritischen Buchstabenkombinationen geprägt ist, die (bis heute) in keinem Codierungsstandard vorgesehen sind (z.B. Kombinationen von Vokalbuchstaben mit Makron und Akzent wie $\acute{\bar{a}}$ oder $\grave{\bar{u}}$ oder Kombinationen von Konsonantenbuchstaben mit subskribiertem Punkt oder Kringel wie $\dot{m}$, $\dot{n}$ oder $\mathring{r}$). Ein ganz gleich gelagertes Problem betraf auch das Avestacorpus, das in einer völlig einzigartigen, in der maßgeblichen Ausgabe K. GELDNERs ("Avesta. Die heiligen Bücher der Parsen", 1895) beibehaltenen Originalschrift ("Avestaschrift") überliefert ist: dessen Eingabe konnte zwar bereits mit einem PC begonnen werden, doch gab es auch hier a priori keine Möglichkeit, die Originalschrift oder auch nur die übliche Transkription, die mit diakritischen Kombinationen wie $\underset{\sim}{t}$ oder griechischen Buchstaben wie γ, δ arbeitet, auf Bildschirm und Drucker darzustellen. So mußte in beiden Fällen zunächst auf eine ersatzweise Darstellung auf der Grundlage des ASCII-Codes zurückgegriffen werden, bei der z.B. die Diakritika durch adskribierte Zeichen wie /, \, ~ repräsentiert wurden.

Auch wenn eine derartige Darstellung für die maschinelle Analyse kein schwerwiegendes Hindernis bedeutet, so wurde doch die Möglichkeit einer der wissenschaftlichen Gepflogenheit entsprechenden Wiedergabe zumindest transkribierter "Sonderzeichen" mit steigender Leistungsfähigkeit von Rechnern (PCs) und Druckern mehr und mehr als ein Desiderat empfunden. Tatsächlich bedeutete die Schaffung von Programmeinheiten, die eine Ausgabe von Transkriptionssystemen oder Originalschriften auf Bildschirm und Drucker ermöglichen sollten, seit dem Beginn des Projekts eine ständige Herausforderung. Nachdem sich erste "Gehversuche" in dieser Richtung noch auf das Drucken mit einem 24-Nadel-Drucker beschränkten[1], scheint in diesem Bereich jetzt, wo "maßgeschneiderte" Pakete vektorisierter Fonts (Postscript, True Type u.a. Formate) für alle in Frage kommenden Schriftsysteme erarbeitet sind, alles nötige getan. Dennoch kann im Zusammenhang mit der Darstellung und der mit ihr stets verknüpften rechnereischen Codierung der Zeichen aus mehreren Gründen noch keine "Entwarnung" gegeben werden: Zum einen sind die verschiedenen verbreiteten Rechnersysteme nach wie vor weit davon entfernt, eine einheitliche Zeichendarstellung zu benutzen. Gemeinsame Grundlage ist nach wie vor lediglich der sog. ASCII-Standard, der wohl die 26 Zeichen des latein. Alphabets, aber keinerlei diakritische Kombinationen wie *ä, á* oder *$\bar{a}$* enthält. Natürlich gibt es in jüngerer Zeit Versuche, die Codierungslücke durch die Schaffung einheitlicher, systemunabhängiger Standards zu schließen. Die bisher vorliegenden Lösungsversuche sind jedoch für die besonderen Aspekte der im Rahmen des TITUS-Projekts zu berücksichtigenden Schriftphänomene in keiner Weise ausreichend, ganz abgesehen davon, wieweit sie überhaupt schon von marktgängigen Betriebssystemen und Anwenderprogrammen unterstützt werden.

Das gilt zunächst für den Standard der sog. WordPerfect-Zeichensätze, der, seit er vor ca. sechs Jahren eingeführt (und über die Jahre nicht unerheblich erweitert) wurde, immerhin einen wichtigen ersten Schritt in die richtige Richtung darstellte. Das hinter dem WP-Standard stehende System baut zwar im wesentlichen auf der Grundlage einer 16-Bit-Codierung auf und ermöglicht damit theoretisch einen Zeichenvorrat von 65536 Zeichen, ausgenutzt ist es jedoch nur zu einem äußerst geringen Teil (ca. 2000 Zeichen), und daß es sich durchsetzen wird, erscheint aus marktpolitischen Gründen eher unwahrscheinlich.

Ähnliches gilt auch für den sog. "Unicode"[2], dessen (ebenfalls in 16 Bit codierte) 65536 Zeichen — wie schon zuvor die verschiedenen systeminternen 8-Bit-Zeichensätze und Codepages — in viel zu starkem Maße an heute gebräuchlichen Nationalalphabeten ausgerichtet sind, als daß sie zur Bearbeitung der — im TITUS-Projekt vorrangig interessierenden — schriftlichen Zeugnisse altüberlieferter Sprachstufen, sei es in Originalschriften, sei es in Transkription, geeignet wären.
Eine bessere Ausgangsposition würde demgegenüber die geplante ISO-Norm 10646 darstellen, die — bei einer 32-Bit-Codierung — mit insgesamt rund 3 Milliarden Zeichen tatsächlich einen für alle je von Menschen benutzten Schriftsymbole ausreichenden Vorrat darstellen könnten. Leider deutet im Moment aber nichts darauf hin, daß sich diese Norm irgendwann einmal zur Basis handelsüblicher Rechnersysteme entwickeln könnte.

Auch ein dritter Lösungsweg, der momentan an Aktualität gewinnt, ist derzeit noch nicht genügend ausgereift, um für die Verarbeitung von Sprachmaterialien der genannten Art eine geeignete Grundlage zu bilden. Es handelt sich um den sog. HTML-Standard ("HyperText Markup Language"), dessen Bedeutung im Zusammenhang mit dem "Weltweiten Netz" ständig zunimmt. Tatsächlich wird die Benutzbarkeit dieses Standards durch den in ihm vorgesehenen Zeichenvorrat,

---

[1] Den Generierungsmöglichkeiten war u.a. mein auf dem Internationalen Orientalistenkongreß (ICANAS) in Hamburg (28.8.1986) gehaltener Vortrag "Probleme der Textverarbeitung in der Orientalistik" gewidmet.

[2] Cf. dazu die von "Unicode Consortium" herausgegebene Darstellung "The Unicode Standard. Worldwide Character Encoding. Version 1.0.", Volume 1-2, 1991.

der neben den reinen "ASCII"-Zeichen, ähnlich wie die durch MS-Windows verbreitete ANSI-Norm, lediglich die in westeuropäischen Nationalalphabeten gebräuchlichen Zeichenkombinationen wie *Á, ç, Ñ* kennt, erheblich beeinträchtigt. Es ist eines der dringendsten Desiderate, daß die Entwickler beim ständigen weiteren Ausbau des HTML-Standards die Codierbarkeit beliebiger Akzentkombinationen sowie nichtlateinischer Schriften vorsehen. Da der HTML-Standard (entsprechend den Vorgaben des sog. SGML-Verfahrens ["Standard Generalized Markup Language"], als dessen Derivat er gelten kann) eine systemunabhängige 7-Bit-Grundlage verwendet, bei der nicht zur ASCII-Norm gehörende Zeichen explizit benannt werden (z.B. "á" für *á*, "ß" für *ß*), stellt er ein a priori offenes System dar, dessen Ausnutzung die Aufgabe des jeweiligen Anwendungsprogramms ist; somit wäre er für die avisierte Erweiterung bestens geeignet.

Es sei noch einmal ausdrücklich festgehalten, daß für eine sprachwissenschaftlich-philologische Auswertung gegebener Textmaterialien in beliebigen Sprachen die Eins-zu-Eins-Wiedergabe der Daten in Originalschrift bzw. Transkription gar nicht unbedingt die Voraussetzung ist. Entscheidend ist vielmehr, daß eine **umkehrbar-eindeutige** Codierung gegeben ist, die das gemeinte Zeichen jederzeit exakt abrufbar zu machen gestattet. Zu warnen ist in diesem Zusammenhang v.a. vor einem allzu leichtfertigen Umgang mit den Möglichkeiten einer freien Umdefinition von Zeichen innerhalb von 8-Bit-Systemen wie denjenigen von Windows-TrueType oder Macintosh-Postscript, denn hier ist gerade die Eindeutigkeit nicht immer gewährleistet (v.a. bei einem Transfer über die Systemgrenzen hinweg): Wenn in einem gegebenen TrueType-Font dasjenige Zeichen, das im zugrundeliegenden Windowssystem als *a* erscheint (Zeichen Nr. 97), durch ein griechisches *alpha* (α) ersetzt ist, dann hängt die Übermittlung der entscheidenden Information, daß eben ein griechisches *alpha* und kein lateinisches *a* gemeint ist, davon ab, ob der korrekte Font vorhanden und anwählbar ist oder nicht; eine Information, die beim Datentransfer zwischen verschiedenen Systemen, aber auch zwischen verschiedenen Anwenderprogrammen auf ein und demselben System viel zu leicht verloren geht. Besonders katastrophal wirkt sich dies aus, wenn innerhalb eines Textes mehrere derartige Fonts mit "überlappenden" Codierungen nebeneinander benutzt sind (wenn also z.B. sowohl griechisches *alpha* als auch lateinisches *a* vorkommen und beide den Bytewert 97 haben). Im Zweifelsfall erscheint demgegenüber jede "unelegant" oder sogar "unbeholfen" wirkende Mehrbytecodierung, wenn sie Eindeutigkeit mit sich bringt (z.B. $a für griechisches α), nach wie vor zweckmäßiger.

Solange wir also einer eindeutigen und systemunabhängig-einheitlichen Codierungsmöglichkeit für sämtliche in Frage kommenden Schriften und Transkriptionszeichen nicht näher gekommen sind, ist es für die im TITUS-Projekt erfaßten Textmaterialien noch nicht sinnvoll, ein bestimmtes "endgültiges" Format anzustreben. Statt dessen werden die Texte derzeit noch in verschiedenen Formaten "nebeneinander" gespeichert, wobei lediglich das Prinzip der eindeutigen Codierung obwaltet, durch das die Konvertierbarkeit bedingt ist. Unabhängig von dem jeweiligen "Ausgangsformat", d.h. demjenigen Format, in dem die Texte von den Beitragenden selbst erstellt werden, sollen die Texte in absehbarer Zeit v.a. in eine für ein sprachwissenschaftliches Retrieval geeignete Codierung gebracht werden. Von den hierfür zur Verfügung stehenden Systemen wird (auf DOS-Ebene) derzeit das von der Brigham-Young-University (Utah) entwickelte "Wordcruncher"-System bevorzugt, das — nach einmal erfolgter Durchindizierung auch großer Textmengen — einen enorm schnellen Zugriff auf einzeln oder im Verbund zu suchende Wortformen sowie eine bequeme Erstellung von Konkordanzen, Indizes etc. ermöglicht[3]. Ob das Wordcruncher-System das System

[3] Die umfangreichste Textsammlung, die im Rahmen von TITUS bisher mit dem Wordcruncher aufbereitet wurde, ist das altenglische Corpus der University of Toronto, das einen Gesamttext von 25 MB (25 Mio. Zeichen) repräsentiert.

der Zukunft sein wird, dürfte allerdings wieder von marktpolitischen Faktoren abhängen, die nicht zuletzt das zugrundeliegende Betriebssystem (DOS) betreffen.

Obwohl die Probleme der Codierung und, davon abhängend, der Konvertierbarkeit den Fortgang des Projekts bis heute immer wieder behindert haben, konnte der Bestand an zur Verfügung stehenden Texten und Textsammlungen seit den Anfängen doch kontinuierlich ausgeweitet werden, so daß heute die Zielsetzung nicht mehr unrealistisch erscheint, bis zum Jahre 2000 über sämtliche für die Vergleichende Indogermanische Sprachwissenschaft relevanten Textmaterialien in elektronisch codierter Form zu verfügen. Während die Textfiles, die für eine Integration gewonnen werden konnten, in der Anlaufphase des Projekts eher "zufällig" dadurch bestimmt waren, daß sich jemand — ein Institut, ein Projekt, eine Privatperson — die Mühe gemacht hatte, sie — meist von Hand — einzugeben, hat sich in der Zwischenzeit in viel größerem Maße der Gedanke der Koordination und Kooperation durchgesetzt; im Normalfall bedeutet dies, daß sich potentielle Beitragende vor dem Beginn ihrer Mitarbeit nach den offenstehenden "Desiderata" erkundigen und ihren Beitrag dementsprechend auswählen. Dies hat auch den Vorteil, daß bestimmte Prinzipien, die die Eingabe betreffen — insbesondere dasjenige der eindeutigen Codierung, aber auch Fragen der Formatierung — vorab geklärt werden können, um so eine nachträgliche Anpassung, die oft recht mühsam "von Hand" erfolgen muß, vermeidbar zu machen. Wie sinnvoll ein solches Verfahren ist, mag ein Beispiel illustrieren:
Bereits seit einiger Zeit wird der Fachwelt das umfangreiche Corpus des altindischen Epos, des Mahābhārata, das unter der Leitung von M. TOKUNAGA an der Universität von Kyōtō / Japan eingegeben wurde, frei über das Internet zur Verfügung gestellt. Die zugehörigen Textfiles (insgesamt ca. 10 MB) enthalten (in einem 16-Bit-Code) den transkribierten Text entsprechend der meistbenutzten kritischen Ausgabe (ed. V.S. SUKTHANKAR / S.K. BELVALKAR, Poona 1933-1959) ohne Variantenapparat. Abgesehen von zahlreichen Transkriptionsfehlern, die nicht ausbleiben können, wenn eine derartige Eingabe von Hand gemacht wird[4], hat die so verbreitete Fassung, die nunmehr auch in die TITUS-Sammlung integriert werden konnte, den schwerwiegenden Nachteil, daß in ihr nicht zwischen Wortgrenzen und den zwischen Kompositalgliedern bestehenden Morphemgrenzen unterschieden worden ist; man vgl. etwa die Notierung der Komposita *dhṛtarāṣṭro* (Eigenname, Nom.Sg.) und *mahārājaḥ* ("Großkönig", Nom.Sg.) in der folgenden Verszeile (mit gegenübergestellter "normaler" Transkription):

0110010013/.	dhRta.raaSTro.mahaa.raajah.zrutvaa.kim.akaron.mune.//	vs.
11,1,1c	*dhṛtarāṣṭro mahārājaḥ śrutvā kim akaron mune* //	

Diese Unterscheidung, die in verschiedenen anderen elektronisch bearbeiteten Sanskrittexten unterschiedlich durchgeführt worden ist (z.B. durch die Anwendung von Zeichen wie +, ^ oder _: *dhṛta+rāṣṭro, mahā+rājaḥ*), kann nun nicht mit automatischen Verfahren nachträglich appliziert werden; sie setzt statt dessen eine — bei der Größe des Gesamttexts höchst aufwendige — manuelle Weiterbearbeitung voraus, durch die zumindest alle vorkommenden Komposita als solche vorweg bestimmt werden müssen. Ein zweites Problem betrifft den sog. Sandhi, d.h. die in altindischen Texten grundsätzlich auftretende Erscheinung, daß Wortformen, wo sie im Satz aufeinander stoßen, lautlich aneinander angepaßt werden, wobei sich meist der Auslaut des ersten, seltener der Anlaut

[4] Für die Anwendung von optischen Scannern gibt es bei der in indischen Ausgaben verwendeten Devanāgarī-Schrift bisher m.W. noch keine zielführenden Verfahren, und selbst wenn dies so wäre, würde auch das automatische Lesen erfahrungsgemäß zu zahlreichen Eingabefehlern führen.

des zweiten Wortes ändert. Bei den bisher verfügbaren elektronisch verarbeiteten Sanskrittexten werden in dieser Hinsicht wiederum völlig unterschiedliche Verfahren angewendet: Teilweise wird, wie auch in manchen gedruckten Texteditionen, der Sandhi "aufgelöst", d.h. es wird eine Normalform eingesetzt, die die kontextuellen Veränderungen nicht aufweist und somit von der tatsächlich überlieferten Wortgestalt abweicht (im o.g. Fall: *dhṛtarāṣṭraḥ mahārājaḥ śrutvā kim akarot mune*); oder aber es wird der Sandhi "beibehalten", was dazu führt, daß ein und dieselbe morphologische Einheit (z.B. der Nom.Sg. *mahārājaḥ* "der Großkönig") in mehreren verschiedenen lautlichen Formen im Text erscheint (z.B. *-rājas, -rājo, -rāja*), die nicht ohne weiteres gemeinsam abrufbar sind. Da die Äquivalenz zwischen einer gegebenen Sandhivariante und der zugrundeliegenden Normalform ("Pausaform") nicht in allen Fällen umkehrbar-eindeutig ist, ist wiederum keine völlig automatische Transposition der einen in die andere Form möglich. Wünschenswert wäre es in dieser Hinsicht, eine doppelte Repräsentation der Texte (mit beibehaltenem **und** aufgelöstem Sandhi) zu haben. Eine solche Paralleladaptation liegt innerhalb der TITUS-Sammlung inzwischen für die älteste altindische Textsammlung überhaupt, die Ṛgveda-Saṃhitā, vor, wobei dem (inzwischen mehrfach überarbeiteten und korrigierten) eins-zu-eins transkribierten Ausgangstext (s.o.) der von A. LUBOTSKY im Hinblick auf eine zu publizierende Textstellenkonkordanz erarbeitete, dem sog. "Pada-Pāṭha" nahekommende interpretative Text Vers für Vers gegenübergestellt ist; man vgl. folgendes Beispiel:

RV 1,6	{a}	⟨yád aṅgá dāśúṣe tvám⟩
		[yát aṅgá dāśúṣe tvám]
	{b}	⟨ágne bhadrám kariṣyási /⟩
		[ágne bhadrám kariṣyási /]
	{c}	⟨távét tát satyám aṅgiraḥ⟩
		[táva ít tát satyám aṅgiraḥ //]

An diesem Beispiel dürfte bereits klar geworden sein, daß für die elektronische Textdatenbank, auf die das TITUS-Projekt abzielt, weit mehr als eine bloße elektronische Erfassung der Texte in ihrer kanonischen, z.B. in maßgeblichen Editionen schriftlich niedergelegten Form von Interesse ist. Eine solche Encodierung kann in den meisten Fällen allenfalls eine erste Arbeitsgrundlage sein, auf der weitere Bearbeitungsschritte, die eine sprachwissenschaftliche Auswertung vorbereiten, aufzubauen haben. Derartige Weiterbearbeitungen werden den zukünftigen Ausbau des TITUS-Thesaurus im wesentlichen bestimmen. Es wird z.B. vielfach, wie im Falle des Altindischen, darum gehen, auf verschiedenen Ebenen gelagerte, durch unterschiedliche Transkriptionsmodalitäten abbildbare Interpretationen schriftlich überlieferter Formen aufeinander beziehbar zu machen. Neben dem Altindischen betrifft das in hohem Maße z.B. die in Keilschrift überlieferten Textmaterialien der Sprachen Altanatoliens, des Hethitischen und seiner Schwestersprachen, die für eine sprachwissenschaftliche Auswertung in ganz erheblichem Maße einer — gegenüber der in den meisten gedruckten Texteditionen gepflogenen rein transliterierenden Wiedergabe — interpretativ-analytischen Transkription hinsichtlich der zugrundeliegenden Lautgestalt bedarf.

Des weiteren ist daran zu denken, daß elektronische Verfahren im gegebenen Zusammenhang auch dort sinnvoll zu einer Erkenntniserweiterung führen können, wo es um die Erschließung einer authentischen Textgestalt aus divergierenden handschriftlichen Traditionen selbst geht. So ist z.B. wahrscheinlich, daß eine vollständige Auswertung des Handschriftenmaterials mit elektronischen Verfahren bei Corpora wie dem (altiranischen) Avesta in zahlreichen Punkten zu einer Verbesserung des Kenntnisstands gegenüber der manuell erarbeiteten gedruckten Textedition führen wird, da es mit elektronischen Mitteln wesentlich leichter ist, sämtliche divergierenden Schreibweisen, die

in den Handschriften auftauchen, auf ihre Konsistenz hin zu überprüfen und damit zwischen singulären, für die Textherstellung relevanten Lesarten und etwaigen, weniger relevanten, Marotten oder Eigenheiten bestimmter Schreiber zu differenzieren.

Um gezielte Auswertungen im Hinblick auf morphosyntaktische Fragestellungen zu ermöglichen, wird es darüber hinaus erforderlich sein, parallel zu den eigentlichen Textmaterialien abrufbare grammatische Angaben zu integrieren. Im Hinblick auf derartige Angaben ("morphologisches Tagging"), wie sie außerhalb des TITUS-Projekts etwa zu verschiedenen Bibeltexten verfügbar sind, ist freilich zu berücksichtigen, daß verschiedene der hier interessierenden Sprachen als ausgesprochene Spezialgebiete anzusehen sind, bei denen eine communis opinio im Hinblick auf die Bewertung morphologischer oder syntaktischer Verhältnisse in vielen Fällen nicht zu erzielen sein wird. Hier wird wiederum vieles davon abhängen, ein möglichst flexibles Retrievalsystem zu haben, das auch alternative Auffassungen und Interpretationen zu verarbeiten zuläßt. So sollte z.B. ein morphologisches Tagging im Falle des Avesta die durch die Homographie gegebene Mehrdeutigkeit von Formen wie (altavest.) *vohū* (Adjektiv "gut" im Nom.Akk.Sg.ntr., Instr.Sg.mask./ntr. oder Nom./Akk.Pl.ntr. oder Substantiv ntr. "das Gut" im Nom.Sg., Instr.Sg. oder Nom./Akk.Pl.) nicht unterdrücken, sondern als solche zu erkennen geben (vgl. das in Tafel 1 wiedergegebene Beispiel). Als "automatisches" Zusatzresultat einer derart aufbereiteten Textsammlung sind umfassende elektronische Wörterbücher zu erwarten, die nicht nur die tatsächliche Beleglage aller morphologischen Varietäten dokumentieren, sondern zugleich auch — im Sinne von Indizes — Referenzen auf die Textstellen selbst enthalten. Damit werden sich zugleich völlig neue Perspektiven für die vergleichend-sprachwissenschaftliche Beschäftigung mit dem Material eröffnen, insofern z.B. die lautlichen Verhältnisse, die zwischen einzelnen der beteiligten (Corpus-)Sprachen bestehen, durch eine in phonologischer, morphologischer und syntaktischer Hinsicht vollständige Erfassung des Materials in ganz anderem Maße überprüfbar werden als bisher (etwa im Hinblick auf Lautgesetze und Ausnahmen davon sowie im Hinblick auf die relative Chronologie sprachlicher Veränderungen).

Von einer umfassenden Bearbeitung der Textmaterialien im genannten Sinne ist das TITUS-Projekt momentan natürlich noch weit entfernt. Was in absehbarer Zukunft zum Abschluß gelangen soll, ist, wie gesagt, als erster Schritt zunächst eine elektronische Erfassung der relevanten Texte selbst, wobei auch die oben als notwendig dargestellte Einheitlichkeit und Eindeutigkeit zunächst noch nicht in allen Fällen gewährleistet ist. Eine Aufstellung, die den derzeitigen Stand der Texterfassung dokumentieren soll, ist im Anhang beigegeben.

Daß eine erschöpfende elektronische Bearbeitung von philologisch problematischen Textmaterialien der genannten Art in ganz erheblichem Maße über das hinausgeht, was etwa im Falle von Textcorpora moderner, gesprochener Sprachen gilt, und im Normalfall nach wie vor als eine eigene wissenschaftliche Leistung anzusehen ist, dürfte außer Frage stehen. Damit ist auch der wichtigste Grund gegeben, warum die im Rahmen des TITUS-Projektes erarbeiteten Textmaterialien derzeit nicht frei und für jedermann zugänglich, etwa über das internationale Netz, abrufbar sind, sondern lediglich dem Kreis der Beiträger (im Sinne einer Arbeitsgruppe) für ihre wissenschaftliche Arbeit zur Verfügung stehen. Einer Publikation von Einzelauswertungen und -Ergebnissen steht damit natürlich nichts im Wege.

Gerade in jüngster Zeit hat das TITUS-Projekt noch eine erhebliche Weiterung erfahren. Bei einem informellen Treffen an der Universität Wien wurde Anfang dieses Jahres die Möglichkeit eruiert,

Y. 29,1c:

<Textherstellung:>

{Y.29,1c:} nōiṯ. mōi. vāstā. xšmaṯ. aniiō. aϑā. mōi. sąstā. vohū. vāstriiā.˙˙

{Y.29,1c: Gdn.} nōiṯ̃. mōi. vāstā.[12] xṣ̌maṯ̃. aniiō. aϑā. mōi. sąstā.[13] vohū. vāstriiā.[14]˙˙

<Lesarten (Ausw.):>

{Y.29,1c: K5} nōiṯ.mōi.vāstrā.xṣ̌maṯ./aniiō.aϑā.mōi.sąstrā.vohū.vāstr/iiā.˙˙

{Y.29,1c: J2} nōiṯ̃.mōi.vāstā.xṣ̌maṯ̃.aniiō.aϑā.mōi.sąstrā.vohū.vāstriiā.˙˙

{Y.29,1c: Pt4} nōiṯ̃.mōi.vāstʳā.xṣ̌maṯ̃.aniiō.aϑā.mōi.sąstā.vohū.vāstriiā.˙˙

{Y.29,1c: S1} nōiṯ̃.mōi. .xṣ̌maṯ̃.aniiō.aϑā.mōi.sąstā.vohū.vāstriiā.˙˙

<Textbearbeitungen:>

{Y.29,1c: Bthl.} nōiṯ. mōi. vāstā. xšmaṯ. anyō. aϑā. mōi. sąstā. vohū. vāstryā.

{Y.29,1c: Humb.1} nōiṯ mōi vāstā xšmaṯ anyō aϑā mōi sąstā vohū vāstryā 7+10

{Y.29,1c: Insl.} nōiṯ mōi vāstā xšmaṯ anyō aϑā mōi sąstā vohū vāstr[y]ā (7+9)

{Y.29,1c: Kell.} nōiṯ mōi vāstā xšmaṯ aniiō, aϑā mōi sąstā vohū vāstriiā/vāstrā? 7+10/9

{Y.29,1c: Humb.2} nōiṯ̃ mōi vāstā xšmaṯ̃ aniiō aϑā mōi sąstā vohū vāstríiā 7+10

{Y.29,1c: Monna} nait mai vāstā (x)šmat anyah āt mai sąsta vahū vāstriyā 7+9/10

	Formenbestimmung	phonolog.	Lemma
nōiṯ̃	C	naiṯ̂	nait
mōi	Eds	maiˆ	az-am
vāstā *vāstrā	Nnsm (Nism) (Nisn) (Nnpn) (Napn)	uāstā ˆuāstrā ˆuāstrā ˆuāstrā ˆuāstrā	ˆuāstār- ˆuāstra-
xšmaṯ̃	Pbp	xšmat	iˆūž-am
aniiō	Ansm	aniˆah	aniˆa-
aϑā	D	aϑā	aϑā
mōi	Eds	maiˆ	az-am
sąstā	V2pApA	sansta	√sand-
vohū	Aisn (Ansn) (Aasn) (Anpn) (Aapn) (Aism) (Nnsn) (Nasn) (Nisn) (Nnpn) (Napn)	ˆuahū ˆuahu ˆuahu ˆuahū ˆuahū ˆuahū ˆuahu ˆuahu ˆuahū ˆuahū ˆuahū	ˆuahu-
vāstriiā *vāstrā	Nisn (Nnpn) (Napn) (Nisn) (Nnpn) (Napn) (Nism)	uāstriiā ˆuāstriiˆā ˆuāstriiˆā ˆuāstrā ˆuāstrā ˆuāstrā ˆuāstrā	ˆuāstriiˆa- ˆuāstra-

Tafel 1

unter Ausnutzung der durch das Internationale Datennetz vorgegebenen Kapazitäten ein umfassendes Informationssystem einzurichten, das sich auf alle Bereiche aktueller Informationen aus der Indogermanistik bzw. Vergleichenden Sprachwissenschaft erstrecken soll. Über einige auf dem WWW-Server der Universität Frankfurt abgelegten HTML-Seiten (s. die o.g. URL) ist dieser Informationsdienst seit einigen Wochen erreichbar; in baldiger Zukunft werden sich sprachwissenschaftliche Institute der Universitäten Prag, Kopenhagen, Leiden, Wien u.a. beteiligen. Angeboten werden sollen dann — neben Mitteilungen über das eigentliche TITUS-Projekt — bibliographische, curriculare, personenbezogene u.ä. Informationen, die das gesamte Gebiet der Vergleichenden Indogermanischen Sprachwissenschaft wie auch angrenzende Gebiete betreffen. Das Gelingen dieses Vorhabens wird ganz wesentlich von der Bereitschaft der Benutzer zur Mitarbeit abhängen.

Anhang:

Das TITUS-Projekt:

Dokumentation der Texte (Stand: 1.6.95)

Alt-, mittel- und neuindisch:

vedisch:

Ṛgveda-Saṃhitā:
- Eingabe des Textes unter der Leitung von W.P. LEHMANN (Austin, Texas) durch H.S. ANANTHANARAYANA (Hyderabad); Überarbeitungen durch S.D. ATKINS (Pomona, California), G.E.DUNKEL (Zürich), J. GIPPERT und F.J. MARTÍNEZ (Frankfurt).

Ṛgveda-Pāṭha (sandhifreie Version):
- Eingabe durch A. LUBOTSKY (Leiden).

Ṛgveda-Khilani:
- Eingabe durch Carlos JORDÁN CÓLERA (Zaragoza; in Bearbeitung).

Aitareya-Brāhmaṇa:
- Eingabe durch J. GIPPERT und F.J. MARTÍNEZ (Frankfurt).

Atharvaveda-Saṃhitā:
- Eingabe durch V. PETR; Kollationierung durch P. VAVROUŠEK (Prag; in Bearbeitung)

Kāṭha-Saṃhitā:
- Eingabe durch Chl.H. WERBA (Wien; in Bearbeitung)

Taittirīya-Saṃhitā:
- Eingabe durch M. FUSHIMI (Ōsaka); Korrektur durch Y. IKARI (Kyōtō)

Taittirīya-Brāhmaṇa:
- Eingabe durch M. FUSHIMI (Ōsaka)

Taittirīya-Prātiśākhya:
- Eingabe durch M. FUSHIMI (Ōsaka)

Āpastamba-Śrauta-Sūtra:
- Eingabe durch M. FUSHIMI (Ōsaka)

Baudhāyana-Śrauta-Sūtra:
- Eingabe durch M. FUSHIMI (Ōsaka)

Vājasaneyi-Saṃhitā:
- Eingabe durch C.-M. BUNZ (Saarbrücken; in Bearbeitung)

Śatapatha-Brāhmaṇa (Mādhyaṃdina-Rezension):
- Eingabe durch G. KNOLL (Frankfurt; in Vorbereitung)

Śatapatha-Brāhmaṇa (Kāṇva-Rezension):
- Eingabe durch J. GIPPERT (Frankfurt; in Bearbeitung)

Gobhila-Gṛhya-Sūtra:
- Eingabe durch Carlos JORDÁN CÓLERA (Zaragoza); Überarbeitungen durch F.J. MARTÍNEZ (Frankfurt)

episches und klassisches Sanskrit:

Mahābhārata:
- Eingabe durch Muneo TOKUNAGA (Kyōtō)

Rāmāyaṇa:
- Eingabe durch Muneo TOKUNAGA (Kyōtō)

Tantrākhyāyika (Buch 1 und 2):
Eingabe durch L.M. FOSSE (Oslo)
Pañcatantra:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Hitopadeśa:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Kālidāsa, Kumārasambhava:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Kālidāsa, Meghadūta:
Eingabe durch J. GIPPERT (Frankfurt)
Kālidāsa, Ṛtusaṃhāra:
Eingabe durch J. GIPPERT (Frankfurt; in Bearbeitung)
Nalopakhyāna:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Dandin, Daśakumāracarita:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Harṣacarita:
Eingabe durch L.M. FOSSE (Oslo; in Bearbeitung)
Vikramacarita:
Eingabe durch P. OLIVIER (Frankfurt; in Bearbeitung)

mittelindisch:
Mahāvaṃśa:
Eingabe durch P. OLIVIER (Frankfurt; in Vorbereitung)

Alt-, mittel- und neuiranisch:

avestisch:
Avesta-Gesamtcorpus:
Eingabe durch S. GIPPERT-FRITZ und J. GIPPERT (Frankfurt)
Nirangistān:
Eingabe nach der Edition WAAG durch H. KUMAMOTO (Tōkyō)

altpersisch:
Altpersisches Gesamtcorpus:
Eingabe unter der Leitung von G.E. DUNKEL durch S. GINDRO, S. SCARLATA, P. WIDMER (alle Zürich); Ergänzungen und Korrekturen durch G. KEYDANA (Münster); Überarbeitung durch J. GIPPERT (Frankfurt);

parthisch:
Manichäische Texte:
Eingabe nach M. BOYCE, Reader durch J. GIPPERT; Überarbeitung durch D.N. MACKENZIE (Göttingen)

mittelpersisch:
Manichäische Texte:
Eingabe nach M. BOYCE, Reader durch J. GIPPERT; Überarbeitung durch D.N. MACKENZIE (Göttingen)
Kārnāmag-i Ardašīr-i Pābagān:
Eingabe durch D.N. MACKENZIE (Göttingen)
Mēnōg-i xrad:
Eingabe durch D.N. MACKENZIE (Göttingen)
Arda-virāf-nāmag:
Eingabe durch P. VAVROUŠEK (Prag)
Ayādgār-i Zarērān:
Eingabe durch A. CANTERA (Berlin; in Bearbeitung)
Vidēvdād-Pahlavī-Übersetzung:
Eingabe durch A. CANTERA (Berlin; in Bearbeitung)
Vizīdagīhā-i Zādspram:
Eingabe durch L. PAUL (Göttingen; in Bearbeitung)

khotan-sakisch:
Khotan-sakisches Corpus:
(Khotanese Buddhist Texts; Khotanese Texts 1-5; Book of Zambasta): Eingabe durch R.E. EMMERICK (Hamburg); Überarbeitung durch H. KUMAMOTO (Tōkyō)
Book of Zambasta:
Eingabe durch P. VAVROUŠEK (Prag)

sogdisch:
Sogdisches Corpus:
(Alles publizierte Material außer MÜLLER-LENTZ, Sogd.Texte II, Text 1-3): Eingabe durch N. SIMS-WILLIAMS (Cambridge)

Buddhistische und Manichänische Texte:
Eingabe durch Y. YOSHIDA (Tōkyō)
neupersisch:
Qabūlī, Ġazals
Eingabe durch M. GLÜNZ (Bern / Seattle)

Anatolisch:
hethitisch:
Hethitische Ritualtexte:
Eingabe durch Chr. ZINKO (Graz)
Hethitisches Corpus:
Eingabe durch P. VAVROUŠEK (Prag; in Bearbeitung)
luvisch:
Luvisches Corpus:
Eingabe durch J. TISCHLER (Dresden)
palaisch:
Palaisches Corpus:
Eingabe durch J. TISCHLER (Dresden)
lydisch:
Lydisches Corpus:
Eingabe durch J. TISCHLER (Dresden)
lykisch:
Lykisches Corpus:
Eingabe durch J. TISCHLER (Dresden)

Tocharisch:
A-tocharisch:
A-Tocharisches Corpus:
Eingabe durch O. HACKSTEIN (Halle; in Bearbeitung)
B-tocharisch:
B-Tocharisches Corpus:
Eingabe durch C. SCHAEFER (Berlin; in Bearbeitung)

Armenisch:
altarmenisch:
4 Evangelien und Psalter:
Eingabe nach der Zohrab-Bibel durch H. PALANDJIAN (Montreal)
4 Evangelien:
Eingabe nach der Edition KÜNZLE durch J. WEITENBERG (Leiden)
Agat‘angelos:
Eingabe durch J. GIPPERT und R.-P. RITTER (Frankfurt; in Bearbeitung)
Movsēs Xorenac‘i:
Eingabe durch H. PALANDJIAN (Montreal)
Patmowt‘iwn vrac‘:
Eingabe durch H. PALANDJIAN (Montreal)
Saraknoc‘:
Eingabe durch H. PALANDJIAN (Montreal)

Baltisch:
litauisch:
Duonelaitis, Metai:
Eingabe durch G. KEYDANA (Münster)

Germanisch:
gotisch:
Gotische Bibel:
Eingabe durch W. GRIEPENTROG (Nürnberg)
altenglisch:
Altenglisches Corpus:
Eingabe an der University of Toronto

althochdeutsch:

Isidor:

Eingabe durch Ma Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. BRYSCH (Dresden) und R. SCHUHMANN (Gießen)

Benediktiner-Regel:

Eingabe durch Ma Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. BRYSCH (Dresden); Korrekturlesungen durch A. POTTHOFF-KNOTH und Roland SCHUHMANN (Gießen)

Tatian:

Eingabe durch Ma Pilar FERNÁNDEZ ALVAREZ & M. M. GARCÍA-BERMEJO GINER (Salamanca); Konvertierungen durch J. KLINGER (Bochum) und J. TISCHLER (Dresden); Weiterbearbeitung unter der Leitung von R. LÜHR (Jena) durch J. Brysch und R. Schuhmann (Gießen); Korrekturlesungen durch S. ZEILFELDER (Jena) und A. POTTHOFF-KNOTH (Gießen)

Tatian: Præfatio Victoris Capuani / Übersicht der Capitelüberschriften / Pariser Fragmente:

Eingabe durch R. SCHUHMANN (Gießen); Weiterbearbeitung durch J. BRYSCH (Dresden)

Otfrid:

Eingabe unter der Leitung von R. LÜHR (Jena) durch R. SCHUHMANN und M. BAYER (Gießen); Weiterbearbeitung durch J. BRYSCH (Dresden); Korrekturlesungen durch A. POTTHOFF-KNOTH, R. SCHUHMANN, A. HOLZHAUER (Gießen) und S. ZEILFELDER (Jena)

Notker:

Eingabe unter der Leitung von R. LÜHR (Jena) durch K. LEPPER (Gießen) und S. ZEILFELDER (Jena)

Griechisch:

klassisch:

Septuaginta:

Eingabe an der University of Pennsylvania (CCAT)

Neues Testament:

Eingabe an der University of Pennsylvania (CCAT)

Italisch:

oskisch:

Tabula Bantina:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Cippus Abellanus:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Oskische Inschriften:

Eingabe durch V. SLUNEČKO (Prag; in Bearbeitung)

umbrisch:

Tabulae Iguvinae:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Umbrische Inschriften:

Eingabe durch J. GIPPERT (Frankfurt); Überarbeitung durch V. SLUNEČKO (Prag; in Bearbeitung)

Keltisch:

alt- und mittelirisch:

Würzburger Glossen:

Eingabe durch J. GIPPERT (Frankfurt) und D. DURKIN (Münster; in Bearbeitung)

Aided Oenfir Aífe

Eingabe durch D. DURKIN (Münster)

Compert Con Culainn

Eingabe durch D. DURKIN (Münster)

De chophur in da muccida

Eingabe durch D. DURKIN (Münster)

Esnada Tige Buched

Eingabe durch D. DURKIN (Münster)

Fled Dúin na nGéd

Eingabe durch D. DURKIN (Münster)

Fingal Rónáin

Eingabe durch D. DURKIN (Münster)

Orgain denna ríg

Eingabe durch D. DURKIN (Münster)

Orgguin trí Mac Diarmata Mic Cerbaill
Eingabe durch D. DURKIN (Münster)
Scéla Cano Meic Gartnáin
Eingabe durch D. DURKIN (Münster)
Scéla Mucce Meic Dathó
Eingabe durch D. DURKIN (Münster)
Serglige Con Culainn
Eingabe durch D. DURKIN (Münster)
Táin bó Froích
Eingabe durch D. DURKIN (Münster)
Togail Bruidne dá derga
Eingabe durch D. DURKIN (Münster)

Rest- und Trümmersprachen:

phrygisch:

Phrygisches Corpus:
Eingabe durch A. LUBOTSKY (Leiden)

Jost Gippert

Digital unterschrieben von Jost Gippert
DN: cn=Jost Gippert, o=Universität Frankfurt, ou=Vergleichende Sprachwissenschaft, email=gippert@em.uni-frankfurt.de, c=DE
Datum: 2011.12.28 16:35:05 +01'00'