Achtung!

Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Der TITUS-Server:
Grundlagen eines multilingualen Online-Retrieval-Systems“
von Jost Gippert (2001).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Computerlinguistik. Was geht, was kommt? /
Computational Linguistics. Achievements and Perspectives.
Festschrift für Wilhelm Lenders,
ed. G. Willée / B. Schröder / H.-Chr. Schmitz,
(Sprachwissenschaft, Computerlinguistik, Neue Medien, 4),
Bonn: gardez! 2002, 81-85
zu entnehmen.

Attention!

This is a special internet edition of the article
“Der TITUS-Server:
Grundlagen eines multilingualen Online-Retrieval-Systems”
[“The TITUS Server:
Basics of A Multilingual Online Retrieval System”]
by Jost Gippert (2001).
It should not be quoted as such. For quotations, please refer to the original
edition in *Computerlinguistik. Was geht, was kommt? /*
Computational Linguistics. Achievements and Perspectives.
Festschrift für Wilhelm Lenders,
ed. G. Willée / B. Schröder / H.-Chr. Schmitz,
(Sprachwissenschaft, Computerlinguistik, Neue Medien, 4),
Bonn: gardez! 2002, 81-85.

Jost Gippert

Der TITUS-Server
Grundlagen eines multilingualen Online-Retrieval-Systems

1 Das TITUS-Projekt

Das Projekt eines „Thesaurus indogermanischer Text- und Sprachmaterialien“ wurde 1987 mit einem in der Zeitschrift „Die Sprache“ veröffentlichten Aufruf initiiert, der darauf abzielte, die Bemühungen um eine digitale Erfassung von Textmaterialien in altindogermanischen Sprachen zu koordinieren und somit einen für die fachlich interessierte Öffentlichkeit gemeinsam zur Verfügung stehenden Datenpool zu schaffen.[1] Das ehrgeizige Ziel, die textuale Bezeugung der Indogermania, die sich vom ältesten (vedischen) Altindischen über altiranische Sprachen wie das Avestische oder das Altpersische, altanatolische Sprachen wie das Hethitische oder das Luvische, die Sprachen des Klassischen Altertums (Griechisch und Latein) bis hin zu den Überlieferungen der germanischen und keltischen Völker (z.B. in Form des Altirischen) erstreckt, in elektronischer Form für die wissenschaftliche Analyse zugänglich zu machen, hat in den seither vergangenen knapp 15 Jahren mehr und mehr konkrete Konturen angenommen, wobei insbesondere die im Jahre 1995 erfolgte Errichtung des „TITUS“-Web-Servers,[2] von dem bereits verfügbare Materialien abgerufen werden können, zu einer Intensivierung der gemeinschaftlichen Datenaufbereitung führte.[3] Von den mehr als 150 Beiträgern, die sich teils in Europa, teils auch in Übersee an dem Projekt beteiligt haben, ist ein Datenbestand erarbeitet worden, der inzwischen weit mehr als 1 GB reiner Textdaten umfasst, wobei durch vergleichbare Projekte abgedeckte Bereiche wie derjenige der altgriechischen Überlieferung, für den der „Thesaurus Linguae Graecae“ der Universität von Kalifornien[4] und das „Perseus-“Projekt der Tufts-University[5] zuständig sind, weitgehend ausgeklammert bleiben.

Die somit umrissene Textdatenbank bildet nach wie vor das Kernstück des TITUS-Servers, der unter den Webadressen <http://titus.uni-frankfurt.de> und <http://titus.

---

[1] Das Projekt wurde seinerzeit unter dem Namen „Thesaurus altindogermanischer Textcorpora auf datenträgern“ angekündigt (Gippert 1987:151t).

[2] „TITUS“ = Thesaurus Indogermanischer Text- und Sprachmaterialien.

[3] Vgl. hierzu die Berichte in Gippert (1995) und (1996).

[4] Cf. den URL <http://www.tlg.uci.edu/~tlg/>.

[5] Cf. den URL <http://www.perseus.tufts.edu/Texts.html>.

fkidg1.uni-frankfurt.de> zugänglich ist und von dem außer den Textmaterialien selbst verschiedene allgemeine Informationen aus dem Bereich der (indogermanischen und allgemeinen) Sprachwissenschaft, Lehrmaterialien wie Sprachenkarten oder Audioaufnahmen gesprochener Sprachen und anderes mehr abrufbar sind. Als Online-Textdatenbank will der TITUS-Thesaurus heute nicht mehr nur die Überlieferungen in alten indogermanischen Sprachen erschließen, sondern auch Textcorpora in „benachbarten" nicht-indogermanischen Sprachen, Texte in weniger verbreiteten modernen indogermanischen Sprachen, in kaukasischen Sprachen u.a.m. Der Einstieg erfolgt über den URL <http://titus.uni-frankfurt.de/texte/texte.htm>, der aktuelle Stand der verfügbaren Textmaterialien ist auf der Seite <http://titus.uni-frankfurt.de/texte/texte2.htm] dokumentiert.

2 Das TITUS-Retrievalsystem

Begünstigt durch die rasche Entwicklung der Webtechnik in jüngster Zeit ist es möglich geworden, über den einfachen Austausch von Textdateien (via *Up-* und *Download*) hinauszukommen und anstelle eines statischen Datenpools ein interaktives Retrieval-System zu erstellen, das für die genannten Überlieferungen zahlreiche zusätzliche Informationsfelder erschließt. Dabei handelt es sich zunächst um eine tiefergehende Quellendokumentation, die neben den eigentlichen Textdaten die digitale graphische Erfassung der Überlieferungsträger, meist Handschriften oder Inschriften, umfasst. Zwei weitere Kernpunkte, an denen zur Zeit intensiv gearbeitet wird, betreffen die automatische Auswertung der Textdaten, nämlich einmal im Sinne einer textübergreifenden Eruierung und Ausgabe von Belegstellen (Konkordanzerstellung), zum anderen im Sinne einer linguistischen Analyse der in den Texten enthaltenen Wortformen. Die hierbei zu beachtenden sprach- und schriftspezifischen Probleme sowie die vom TITUS-Projekt eingeschlagenen Lösungswege seien an einem Beispiel aufgezeigt.

Mit dem Namen „Tocharisch" bezeichnet man zwei eng miteinander verwandte, als Ost- und West- oder „A-" und „B-Tocharisch" differenzierte indogermanische Sprachen, die etwa in der Zeit des 4.-8. Jh. n.Chr. in Chinesisch-Turkestan (der heutigen chinesischen Provinz Hsinkiang-Uygur) gesprochen und geschrieben wurden. Das uns verfügbare Material besteht aus ca. 4000 Handschriften, die im Zuge von drei Expeditionen der Preußischen Akademie der Wissenscahften zu Beginn des 20.Jhs. rund um die Oase Turfan zutage gefördert wurden und heute von der Berlin-Brandenburgischen Akademie verwaltet werden (sog. Berliner Sammlung); ca. 1000 weitere Handschriften finden sich in Paris, London, St. Petersburg und an anderen Orten. Geschrieben

sind die Dokumente in einer lokalen Abart der indischen Brāhmī-Schrift; sie enthalten zumeist aus dem Sanskrit oder anderen Sprachen übersetzte Textstücke buddhistischen Inhalts.

Im Rahmen des TITUS-Projekts, das in diesem Zusammenhang seit 1995 intensiv mit der Stiftung Preußischer Kulturbesitz zusammenarbeitet, werden die tocharischen Materialien sukzessive in folgender Weise aufgearbeitet: Zunächst erfolgt eine Digitalisierung bereits in Druckform ediert vorliegender Textmaterialien, wobei die Texte wie in den gedruckten Editionen selbst in Umschrift eingegeben werden. Der zweite Schritt besteht in der Digitalisierung der Originalhandschriften, wobei aus Qualitätsgründen, aber auch zum Zwecke der Konservierung, der Umweg über die Erstellung von Farbdias gewählt wurde, die dann mit einem hochauflösenden Dia-Scanner eingelesen werden. Mit den so digitalisierten Handschriften werden die elektronischen Texte neu kollationiert und, wo erforderlich, überarbeitet.[6] Das eigentlich Retrieval wird durch eine Vorindizierung der digitalen Texte vorbereitet, für die das WordCruncher Retrieval System eingesetzt wird.[7]

Das erste Ergebnis dieser Bearbeitung ist die über <http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/thtframe.htm> verfügbare Ausgabe der edierten Teile der Berliner Sammlung (1100 Handschriften). Die auf HTML-Basis in Frames angelegte Ausgabe erschließt die Texte nach den Nummern des aktuellen Berliner Katalogs,[8] aber auch nach der Numerierung in den Druckausgaben;[9] gemeinsam mit den Texten, die in einer vereinfachten, in Unicode (UTF-8) codierten Umschrift gehalten sind, werden, soweit verfügbar, Bilder der betreffenden Handschriftenblätter bereitgestellt (ca. 260 Handschriften sind bisher nicht auffindbar gewesen; es handelt sich vermutlich um Kriegsverluste).

Für das eigentliche linguistische Retrieval (Suche nach Wörtern, Wortformen etc. und ihren Belegstellen) hat das TITUS-Projekt 1996 (mit Unterstützung der WordCruncher Company) einen WordCruncher Web-Server einrichten können, mit dem die umfangreichen Leistungsmerkmale dieses (von der Brigham Young University entwickelten) Retrievalsystems online verfügbar gemacht werden konnten <http://titus.fkidg1.uni-frankfurt.de/texte/tituswc2.htm>. Dies beinhaltet insbesondere die Möglichkeit einer sprachspezifischen Indizierung auch gemischtsprachiger Texte, die die Besonderheiten der jeweiligen Sprach- und Schriftstrukturen (z.B. alphabetische

---

[6] Cf. hierzu die ausführliche Darstellung in Gippert (1998).

[7] Für die Grundlagen dieses Verfahrens cf. Gippert (1999).

[8] „THT“ = Tocharischen Handschriften aus Turfan.

[9] Cf. Sieg, Siegling (1921 und 1949-1953) sowie Thomas (1983).

Anordnung, Gleichwertigkeit unterschiedlicher Zeichen etc.) zu berücksichtigen gestattet. Die Nachteile des WordCruncher Web-Servers bestehen im Wesentlichen darin, dass die Codierung der Texte nach wie vor auf einer 8-Bit-Basis erfolgen muss, was die Notwendigkeit sich überlappender Spezialfonts mit sich bringt ("font mapping"), und dass die Nutzbarkeit des Systems plattformabhängig ist, da es bisher lediglich für MS-Windows-Oberflächen verfügbar ist.[10] Es kommt hinzu, dass für die Online-Kommunikation ein spezifisches Protokoll verwendet wird, was vielfach zu Übertragungsproblemen führt.

Da es zudem unsicher ist, ob und in welcher Form das WordCruncher-System weitergepflegt wird,[11] wird im Rahmen des TITUS-Projekts derzeit ein eigenständiges Retrievalsystem entwickelt, das entsprechende Analysemöglichkeiten in einer HTTP-basierten Online-Umgebung bietet. Grundlage dieses Systems ist auf HTML-Grundlage strukturierter Text, der in Unicode codiert und, entsprechend dem jeweils üblichen Referenzierungssystem, in Einheiten (z.B. Text, Buch, Kapitel, Absatz, Satz oder Handschrift, Seite, Zeile) zerlegt ist. Im Laufe einer Vorindizierung werden sämtliche Wortformen, die in dem so aufbereiteten Text enthalten sind, mit ihrer Referenzangabe in einer Datenbank abgelegt; der Zugriff auf diesen Index kann dann durch manuelle Eingabe in einem Suchformular erfolgen – textunabhängig oder textbezogen –, aber auch über Hyperlinks, mit denen jedes Wort im Text versehen ist. Ein Prototyp dieses Ausgabemodus ist für das A-tocharische Corpus über <http://titus.uni-frankfurt.de/texte/texte2.htm#toch> bzw. <http://titus.fkidg1.uni-frankfurt.de/texte/etcc/toch/tocha/tocha.htm> verfügbar.

Durch die zweiteilige Grundkonzeption – Verknüpfung reiner Textdaten mit einer referentiellen Datenbank – lässt sich nicht nur der Aufwand des Taggings in den eigentlichen Textfiles auf ein Minimum reduzieren, was der Übertragungsgeschwindigkeit im Netz zugute kommt, sondern es können mit geringem Aufwand auch weitergehende Retrievalelemente hinzugegeben werden. Bereits jetzt ist das System in der Lage, beliebige tocharische Verbalformen, auch ungeachtet eventueller Schreibvariationen, korrekt zu bestimmen. So wird z.B. bei Aktivierung des betreffenden Hyperlinks durch Anklicken der Verbalformen *nämseñc* „sie verneigen sich" oder *kumseñc* „sie kommen" in dem A-tocharischen Text THT 634 auf dem TITUS-NT-Server ein ASP-Script aufgerufen, das nicht nur sämtliche Belegstellen dieser Formen in den Texten aus der Datenbank extrahiert, sondern auch ihre Bestimmung als 3. Person Plural Präsens Aktiv; in der Ausgabeseite selbst ist weiter ein Hyperlink auf die betreffen-

[10] Cf. hierzu bereits Gippert (1997: 75-93 und 1999: 374-379).

[11] Der Web-Server der WordCruncher Company <http://www.wordcruncher.com> existiert offenbar nicht mehr.

de Verbalwurzel (*näm-* „(sich) verneigen“ bzw. *käm-* „kommen“) enthalten, der seinerseits über einen ASP-Aufruf zur Ausgabe der kompletten Paradigmen dieser Verben führt. Es sei angemerkt, dass die Verwendung von Unicode bei der erforderlichen Übergabe von Parametern in den URL-Kommandozeilen heute noch nicht möglich ist, weshalb die Angaben in eine gemischte ASCII-Notierung umgesetzt werden müssen (z.B. <kE400m2D00> für *käm-*).

3 Ausblick

Die Grundlagen des TITUS-Retrievalsystems, das derzeit noch in der Erprobungsphase steckt, seien abschließend zusammengestellt. Sie umfassen:

— eine Aufarbeitung der Texte in Unicode-Codierung mit einheitlicher HTML-Struktur (über *Style-Sheet*-Vorgaben);
— eine eindeutige Sprachzuweisung für alle in den Texten enthaltenen Wortformen;
— eine Vorindizierung des Textmaterials mit exakter Stellenreferenzierung;
— die Erfassung der auf die Wortformen bezogenen Daten in einer Datenbank (wegen seiner Unicode-Kompatibilität wird derzeit MS-Access verwendet), wobei durch Rückgriff auf sprachspezifische Vorgaben Schreibvarianten etc. einander zugeordnet werden;
— einen SQL-basierten Zugriff auf die Datenbank via ASP-Seiten (Windows-NT-Server), die über Javascript-basierte Hyperlinks direkt aus dem HTML-Text oder über Eingabeformulare aufgerufen werden.

Die jeweils aktuellen Retrievalmöglichkeiten werden auf der Seite <http://titus.uni-frankfurt.de/texte/textex.htm> beschrieben.
An folgenden Weiterentwicklungen wird derzeit gearbeitet:

— Ausweitung der Suchmöglichkeiten (verbundene Suche mehrerer Wortformen im Kontext etc.; sprachspezifische Wildcards, z.B. für „alle Vokale“, „alle Konsonanten“ etc.);
— Ausweitung der Wiedergabe in Originalschriften (Voraussetzungen: Unicode-Ergänzung, z.B. für Avestisch, Mittelpersisch etc.; Ausbau vorhandener Unicode-Implementierungen, z.B. für Syrisch.
— Geprüft wird der Übergang von HTML-Strukturen zu XML-Strukturen.

Literatur

Gippert, Jost (1987): „Mitteilung über einen Thesaurus altindogermanischer Textcorpora auf Datenträgern“. – Die Sprache 32/1, 151 t.

— (1995): „TITUS. Das Projekt eines indogermanistischen Thesaurus“. – LDV-Forum 12/2, 35-47.

— (1996): „TITUS – Alte und neue Perspektiven eines indogermanistischen Thesaurus“. – Studia Iranica, Mesopotamica et Anatolica 2, 49-76.

— (1997): “Multilingual text retrieval: Requirements and solutions”. – Studia Iranica, Mesopotamica et Anatolica 3, 75-93.

— (1998): “Digitization of Tocharian Manuscripts from the Berlin Turfan Collection”. – Manuscripta Orientalia 4/1, 49-57.

— (1999): “Language specific encoding in multilingual corpora: Requirements and solutions”. – In: J. Gippert, P. Olivier (eds.), Multilinguale Corpora: Codierung, Strukturierung,. Analyse. 11. Jahrestagung der GEsellschaft für Linguistische Datenverarbeitung, 371-384. Praha: enigma.

Sieg, Emil, Walter Siegling (1921): Tocharische Sprachreste. I. Band: Die Texte. A: Transcription; B: Tafeln. – Berlin, Leipzig: de Gruyter.

— (1949-1953): Tocharische Sprachreste. Sprache B, Heft 1-2. – Göttingen: Vandenhoeck & Ruprecht.

Thomas, Werner (1983): Tocharische Sprachreste, Sprache B. Teil I: Die Texte. Bd. 1: Fragmente Nr. 1-116 der Berliner Sammlung, neubearb. und mit einem Kommentar nebst Register versehen. – Göttingen: Vandenhoeck & Ruprecht.

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: 2012.01.02 21:14:50 +01'00'