Achtung!
Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Sprachliche Morphologie digitalisiert“
von Jost Gippert (2007).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Morphology and Digitisation.
Proceedings of the International Conference Berlin, July 7-8, 2006. =
Chatreššar 2009, 35–62
zu entnehmen.

Attention!
This is a special internet edition of the article
“Sprachliche Morphologie digitalisiert”
by Jost Gippert (2007).
It should not be quoted as such. For quotations, please refer to the original edition in
Morphology and Digitisation.
Proceedings of the International Conference Berlin, July 7-8, 2006. =
Chatreššar 2009, 35–62.

Sprachliche Morphologie digitalisiert

Jost Gippert (Frankfurt)

The article discusses the question how and to what extent the digitisation of morphological features of natural languages is able to support linguists in both teaching and research. Within the scope of applied linguistics, it focusses on the impact of digitised morphological data on optical character recognition, speech recognition, spell and grammar checking, machine translation and teaching scenarios. The implementation of digitised means in linguistic research is thematised with respect to lexicography and grammar writing as well as diachronic investigations of various kinds. The perspectives of applying digitisation to linguistic morphology can be summarised in two theses: a) the more linguistic data are provided with a thorough morphological (and beyond that, syntactical, semantical, pragmatical, metrical etc.) analysis, the more reliable the basis for theoretical investigations will be both in synchronical-descriptive and in diachronical-comparative linguistics; b) the digital preparation of large data sets is an urgent methodological prerequisite for the verification and falsification of linguistic hypotheses.

Für manch einen Außenstehenden hat die Sprachwissenschaft spätestens mit dem Strukturalismus eine Ausrichtung angenommen, die Goethe und Schiller in ihren Xenien schon lange vorher andeuteten, als sie den „Sprachforscher“ so apostrophierten[1]:

> *Anatomieren magst du die Sprache, doch nur ihr Cadaver,*
> *Geist und Leben entschlüpft flüchtig dem groben Scalpell.*

Mit dem Einsatz digitaler Verarbeitungsverfahren dürfte sich das Schreckensbild einer immer tiefergehenden „anatomistischen“ Zerlegung sprachlicher Äußerungen in den Augen vieler eher noch weiter verfestigt haben. Im folgenden soll gezeigt werden, daß eine digital basierte Morphologie natürlicher Sprachen, im Gegensatz zur Einschätzung unserer Klassiker, durchaus nicht mit Geist- und Leblosigkeit einherzugehen braucht, sondern

1 Xenie Nr. 141; Goethe, *Werke* 1889-1896, I. Abtlg., Bd. V, 1. Abtlg., 1893, S. 225; Schiller, *Werke* 1943–, Bd. I, 1943, S. 326. Ob das Distichon von Goethe oder Schiller verfaßt wurde, scheint nicht sicher geklärt; Corkhill 1991, S. 248 schreibt es offensichtlich Goethe zu, der damit „die klinisch anmutende fachliche Sprachbetrachtung ... persiflieren wollte“.

gerade dazu dienen kann, sprachliche Inhalte genauer zu beleuchten und den Umgang mit Sprache in unterschiedlichen Bereichen zu unterstützen.

1. Anwendungsbezogene Zielsetzungen digital basierter Morphologie

Für eine digital basierte sprachwissenschaftliche Morphologie kommen im wesentlichen fünf anwendungsbezogene Zielsetzungen in Betracht, nämlich die Unterstützung von Texterkennung („OCR" = „Optical Character Recognition"), Spracherkennung, Rechtschreibung („Spell-checking"), automatische Übersetzung sowie Sprach- und sprachwissenschaftlicher Unterricht. In verschiedenen dieser Anwendungsbereiche wird digitalisierte Morphologie seit längerem bereits mit stetig steigendem Erfolg eingesetzt. Einige wenige Grundüberlegungen verdienen es dennoch, hier kurz abgehandelt zu werden.

1.1. Texterkennung

Digitale Texterkennung, d.h. die retrograde Umwandlung bereits gedruckter oder geschriebener Textmaterialien in eine digitale Form zum Zwecke der Weiterverarbeitung, setzt typischerweise den folgenden Ablauf voraus: Nach dem Einscannen einer Druckseite als Rasterbild müssen in diesem zunächst die groben Einheiten des Schriftbilds herausisoliert werden; dies geschieht durch die Erfassung von Zonen ohne schwarze Bildpunkte („Pixel"), durch die Abstände zwischen Zeilen, Wörtern, Buchstaben, Satzzeichen etc. gekennzeichnet sind, sowie, umgekehrt, durch die Erfassung zusammenhängender Pixelstrukturen, die Buchstaben, Satzzeichen etc. repräsentieren. Letztere müssen dann mit Mustern verglichen werden, um die repräsentierten Zeichen im einzelnen zu bestimmen. Hierbei ist immer mit einer gewissen Fehlerquote zu rechnen, die z.B. von der drucktechnischen Qualität der Vorlage oder den in ihr benutzten Zeichenformen und -sätzen abhängt. Eine Reduzierung der Fehlerquote, die auch heute, nach rund 20 Jahren fortschreitender Entwicklung von OCR-Verfahren[2]

2 Seit 1987 habe ich kontinuierlich kommerzielle OCR-Programme getestet und mit mehr oder weniger großem Erfolg sowohl auf lateinschriftliche als auch auf andere Quellen angewendet. Erwähnen möchte ich SPOT 3.1 (Flagstaff Engineering, 1989), ein auf Zeichenformen „trainierbares", DOS-basiertes Programm, das ich sehr erfolgreich an die georgische Schrift anpassen konnte und dessen Leseergebnisse die Grundlage für zahlreiche über den TITUS-Server (s.u. Fn. 9) verfügbare elektronische Texteditionen darstellen; das Kurzweil Scanner-System, in den frühen 90er Jahren das Non-Plus-Ultra der PC-basierten OCR, mit dem verschiedene lateinschriftlich gedruckte TITUS-Texte eingelesen wurden; sowie den Fine Reader (ABBYY), der sich in den letzten Jahren als die leistungsfähigste OCR-Software für multilinguale Anwendungen herausgestellt hat (derzeit verfügbar in Version

noch immer beträchtlich sein kann[3], setzt verschiedene Arten von Plausibilitätsprüfungen voraus, bei denen das Leseresultat auf seine sprachliche Korrektheit hin getestet wird. Ein typisches Beispiel wäre die „Verlesung" der deutschen Wortform *Mündern* (Dat.Pl.) als *Mundem:* sowohl die Ignorierung des Diakritikums, die auf der Diskontinuität des Buchstabenbildes <ü> beruht, als auch die irrige Zusammenlesung der Buchstabenfolge *rn* als *m* sind charakteristische Probleme von OCR-Anwendungen. Um die gelesene Wortform *Mundem* nun auf ihre Plausibilität hin zu überprüfen, bieten sich zwei Verfahren an. Das in kommerziellen OCR-Programmen meist genutzte Verfahren ist rein „lexikalisch": Zum Abgleich wird eine (mehr oder weniger umfangreiche) Wortformenliste der betr. Sprache benutzt, in der natürlich neben Lemmaformen auch flektierte Formen aller Art enthalten sein müssen. Die der gelesenen Form am nächsten kommende wird dann vorschlagsweise eingesetzt; im gegebenen Fall dürfte dies die Form *Munden* sein, die sich von *Mundem* nur in einem Buchstaben unterscheidet, damit freilich aber noch nicht „korrekt" ist und somit bereits die Grenzen eines rein lexikalischen Abgleichs aufzeigt. Es kommt hinzu, daß keine Wortformenliste des Deutschen jemals vollständig sein kann, da insbesondere die Möglichkeiten der Komposition das deutsche Lexikon unendlich groß gestalten.

Um eine bessere Lesegenauigkeit zu erzielen, ist deshalb ein anderes, mehrstufiges Verfahren vonnöten, das neben einer lexikalischen Datenbasis auch andere Vorinformationen, darunter morphologische, verarbeiten kann. Ein solches Verfahren, das ich „gemischt" nennen würde, müßte neben einem Basislexikon der Wortstämme zunächst die grammatischen Regeln der Morphologie, d.h. der Formen- und Wortbildung der betr. Sprache berücksichtigen, darüber hinaus „graphische" und „grapho-

9). Die spezifischen Bedürfnisse von Sprachwissenschaftlern (v.a. in Bezug auf transkriptive Sonderzeichen) sind allerdings bis heute nirgends zufriedenstellend berücksichtigt.

3 Eine Fehlerquote von 2% bedeutet z.B. zwei Fehler innerhalb von zwei normalen Druckzeilen eines Buches. Welche Schwächen oberflächlich angewandte OCR-Verfahren nach wie vor haben und zu welchen Irrtümern sie führen können, zeigt deutlich das ehrgeizige Projekt „Google Books" (http://books.google.com): Sucht man hier nach SCHLEICHERs Compendium, so werden verschiedene Einträge aus Kuhns Zeitschrift für Vergleichende Sprachwissenschaft ausgegeben, in denen das Werk u.a. als „Compendium tier vergleichenden grammatik" und sein Band I als ein „Kurzer abril" einer lautlehre der indogermanischen ..." erscheint; entsprechend liefert eine Suche nach „Compendium" und „Tier" genau den erstgenannten Eintrag zuerst, noch vor zoologischen Werken (Stand: 25.3.2007, 17:51h). [Korrekturzusatz: Der gleiche Eintrag lautet am 8.10.2009, 13:19h wie folgt: "August Schleicher, compendium der vergleichenden grammatik der indogermanischen sprachen. Bd. I. Kurzer abrifs einer lautlehrc der...". Die Ergebnisse der (fortschreitenden) Texterkennung bei "Google Books" bleiben also unzuverlässig.]

taktische“ Regeln. Gehen wir wieder von unserem Beispiel, der Wortform *Mündern*, aus, die sich als Dativ Plural von *Mund* in drei Morpheme, das Lexem {Mund}, das Pluralsuffix {er} und die Dativendung {n} zerlegen läßt und gleichzeitig das Phänomen des „Umlauts“ von *u* zu *ü* aufweist, das im morphologischen Sinne als ein Ablautphänomen aufzufassen ist. Ein „gemischtes“ Analyseverfahren müßte die eingelesene Form *Mundem* nun, von hinten beginnend, schrittweise verkürzen, bis eine Übereinstimmung mit einem im Basislexikon enthaltenen Wortstamm gegeben ist; dies wäre der vierbuchstabige Stamm *Mund*. Daraufhin wäre zu überprüfen, ob der verbleibende Rest, *em*, im Morpheminventar verankert ist – dies ist zwar der Fall, da die deutsche Nominalflexion ein solches postlexematisches Morphem kennt (vgl. *gutem*, *schönem*), es ist jedoch nicht mit dem gefundenen Basisstamm *Mund* kompatibel, da dieser kein Adjektiv darstellt. Dasselbe Verfahren müßte dann iterativ weiter durchgeführt werden (*Mun –dem* etc.), käme aber allein noch zu keinem Ergebnis. An dieser Stelle, nach negativem erstem Durchlauf, wäre das Leseresultat (*Mundem*) dann auf seine graphische Plausibilität hin zu überprüfen, wobei zunächst „untypische“ Zeichenfolgen zu eruieren wären; solche sind im gegebenen Beispiel nicht vorhanden, lägen aber z.B. in einer Verlesung *IVIund* vor, die aufgrund allgemeinen Vorwissens über Zeichenformen der Lateinschrift und ihre Verteilung in Wortformen unmittelbar plausibel in *Mund* zu korrigieren wäre. Bei „gelesenem“ *Mundem* müßten umgekehrt die graphotaktisch unproblematischen, aufgrund desselben Vorwissens für Verlesungen jedoch als besonders anfällig geltenden Buchstaben, <m> und <u>, versuchsweise durch ihre nächstähnlichen „Partner“, <rn> und <ü>, ersetzt und die sich so ergebenden Wortformen wieder der iterativen Analyse unterzogen werden. Dies würde bereits für *Mundern* die mögliche Verknüpfung eines Lexems (*Mund*) mit „passenden“ Flexionsmorphemen {er-n} ergeben, jedoch noch keine korrekte Wortform, da der Plural von *Mund* eben Umlaut aufweist. Erst ein dritter Durchlauf, mit der Abänderung von <u> zu <ü>, ergibt dann das richtige Resultat. Vorausgesetzt ist dabei, daß auch die Information über das Umlaut- bzw. Ablautverhalten der Stämme im lexikalischen und morphologischen Basiswissen verankert sein muß.

Hinsichtlich der „graphotaktischen“ Wissensbasis, auf die ein solches Verfahren zwingend zurückgreifen muß, sei noch einmal festgehalten, daß diese sowohl schriftspezifische, d.h. durch die Ähnlichkeiten von Buchstabenformen gegebene Probleme als auch sprachspezifische Regelungen umfassen muß, die jeweils statistisch erfaßt werden können. Zu den ersteren gehören z.B. in der Lateinschrift neben den bereits behandel-

ten Verwechslungspaaren solche, bei denen sich Buchstaben und Ziffern gegenüberstehen, wie z.B. <S> und <5> oder <O> und <0> (null). Eine Plausibilitätsprüfung kann in diesen Fällen ohne weiteres davon ausgehen, daß eine <l> (eins), die mitten in einem „Wort" und von sicher gelesenen Kleinbuchstaben umrankt zu stehen scheint, eher ein kleines <L> repräsentieren dürfte, während eine Folge von <7> und dreimaligem Buchstaben <O> eher „siebentausend" meinen wird.[4] Entsprechende Vorhersagen werden für andere Schriften, wie z.B. die Kyrillische, nicht unbedingt zutreffen; hier wird z.B. eher eine Verwechslung der Buchstaben <л> (≈ <l>), <п> (≈ <p>), <н> (≈ <n>) und <и> (≈ <i>) untereinander anzunehmen sein.

Der sprachspezifische Anteil der Wissensbasis kann z.B. allgemeine statistische Erkenntnisse über die Verteilung von Graphemen in Wortformen der betr. Sprache verwenden; danach würde etwa ein gelesenes <sehön> im Deutschen ohne weiteres zu *schön,* nicht aber zu *sehen* zu korrigieren nahegelegt, da das <ö> aufgrund seiner Seltenheit gewissermaßen eine *lectio difficilior* gegenüber <e> darstellt (und schwerlich für dieses verlesen sein kann), während von den beiden Zeichenfolgen <seh> und <sch> die letztere eine klare statistische Präponderanz für sich hat. Im gleichen Sinne könnten Regelungen der Groß- und Kleinschreibung berücksichtigt werden, die ihrerseits mit der morphologischen Bestimmung einhergehen sollten; so hängt z.B. bei einer Verlesung von *Gütern* als *Gutem* die Plausibilität der letzteren Form davon ab, ob sie – als Adjektiv – in einer Umgebung erscheint, die einen initialen Großbuchstaben erwarten läßt, oder nicht.

Damit ist bereits angedeutet, daß eine wirklich zuverlässige Texterkennung bei der Plausibilitätsanalyse noch einen Schritt weiter gehen muß, nämlich bis hin zu einer syntakto-semantischen Überprüfung. *Gutem* ist ja eine vollkommen korrekte Wortform des Deutschen, kommt allerdings im Vergleich zu *Gütern*, obwohl ebenfalls ein Dativ, mit großem <G> in nur sehr reduzierten syntaktischen Konstellationen vor, nämlich entweder satzeinleitend, als Adjektiv ohne vorangehenden Artikel (etwa: *Gutem Brauch folgend, schließt sich der Landrat...*)[5], oder als substantiviertes Adjektiv, ebenfalls ohne Artikel (etwa: „*...ob er nach Gutem oder Bösem*

4 Erstaunlicherweise werden in den meisten kommerziellen OCR-Programmen nicht einmal derartig banal anmutende Regelungen berücksichtigt. Auch lassen sie sich nicht in die von den Programmen verwendete Wissensbasis integrieren, da selbst „trainierbare" Programme normalerweise nur Zeichenformen zu speichern erlauben, nicht jedoch spezifische distinktive Merkmale von Zeichen oder graphotaktische Regeln.

5 S. `http://www.odenwaldkreis.de/index.p?funktion=presseartikel.php&selected=1&id=848` (19.1.2007).

jagt“)[6]. Eine Entscheidung zwischen *Gütern* und *Gutem* setzt also eine syntaktische Analyse voraus, die zumindest die umgebenden Wortformen mit berücksichtigen muß; hierbei können statistische Angaben über sog. „Kollokationen“ Entscheidungshilfen bilden.[7] In anderen Fällen, wie etwa bei dem – graphisch minimalen – Unterschied zwischen *Tochter* und *Töchter,* wird nicht einmal durch Informationen aus der syntaktischen Umgebung immer eindeutig geklärt werden können, ob eine gegebene „Lesung“ richtig ist; man vgl. z.B. die Lesungen <*die jüngere Tochter*> und <*die jüngeren Tochter*>, wo die Endungen der attributiven Adjektive im Verbund mit den Artikelformen Eindeutigkeit herstellen, während in gelesenem <*Ich sehe seine Tochter*> (gegenüber <*Ich sehe seine Töchter*>) ohne Kenntnis des Kontextes keine Entscheidung für die Singular- oder die Pluralform möglich ist. Wie sehr die morphologische Bestimmung von syntakto-semantischen Vorgaben abhängen kann, zeigt sich nicht zuletzt bei homonymen Wortformen, die unterschiedlichen morphologischen Klassen zugehören; man vgl. z.B. das Verb *zugreifen* gegenüber dem Kompositum *Zugreifen* = {Zug} + {Reifen} oder das Partizipialadjektiv *vereinzelt* gegenüber dem Kompositum *Vereinzelt* = {Verein} + {Zelt}.

Auch das „gemischt-morphologische“ Verfahren zur Unterstützung der Texterkennung hat also Grenzen, die auch mit großen Datenmengen nicht leicht zu überbrücken sein werden. Hier dürften auf längere Sicht Verfahren der statistischen Untermauerung durch Kollokationen eine entscheidende Rolle spielen. Morphologisches Wissen bleibt dabei aber mit Sicherheit ein unverzichtbarer Bestandteil.

1.2. Spracherkennung

Der Einsatz morphologischer Verfahren unterliegt bei der Spracherkennung ganz ähnlichen Bedingungen wie bei der Texterkennung: Auch hier geht es um eine eindeutige Bestimmung sprachlicher Formen bei der digitalen

6 Goethe, *Werke*, I. Abtlg., Bd. II, 1888, S. 246. – Theoretisch sollte ein gut fundiertes „gemischtes“ System auch in der Lage sein, Druckfehler stillschweigend zu korrigieren; dies dürfte jedoch immer dann ausgeschlossen sein, wenn das verdruckte Wort eine morphologisch „korrekte“ Form darstellt wie etwa im Falle von für <Verballhornungen> verdrucktem <Verbalhornungen> = {Verb}{al}{Hornung}{en}. Hier mag selbst eine kollokationsbasierte semantische Analyse scheitern, wenn der Druckfehler in einem sprachwissenschaftlichen Kontext auftritt, der u.a. auch Wörter wie *Verbalflexion* aufweist (so bei Hoffmann, *Altiranisch*, S. 10 = Aufsätze I, S. 67, Z. 6 v.u. bzw. S. 16 / 73, Z. 11).

7 Vgl. hierzu das Projekt „Kollokationen im Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften; s. `http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/kollokationen/de/ueberblick`.

Erfassung von Äußerungen. Ein Unterschied zur Texterkennung besteht nur darin, daß die graphische Komponente entfällt, da die zugrundeliegenden Daten nicht schriftlich, d.h. zeichenkettenbasiert, sondern mündlich (und als akustisches Signal digitalisiert) vorliegen. An die Stelle der „graphotaktischen" Entscheidungshilfen müßten hier, wenn man ein „gemischtes" Verfahren anstrebt, deshalb phonotaktische Statistiken treten. Die Notwendigkeit einer Einbindung morphologischer Analyseverfahren bleibt davon unberührt.

1.3. Rechtschreib- und Grammatikprüfung

Im Gegensatz zu den beiden vorgenannten Anwendungsbereichen setzt die Rechtschreib- und Grammatikprüfung eine Stufe später an, nämlich bei einem schon existierenden, digital vorliegenden Text. Unabhängig davon, ob dieser von Hand eingegeben oder via OCR oder Spracherkennung generiert wurde, geht es hier jedoch wieder um genau dieselbe Art von Plausibilitätsprüfung, wie sie oben als Bestandteil des Erkennungsprozesses beschrieben wurde. Ein klarer Unterschied besteht gegenüber der Texterkennung lediglich in der Hinsicht, daß bei von Hand eingegebenen Texten nicht dieselben „graphotaktischen" Vorgaben gelten, da beim „Tippen" systematisch andere Buchstaben verwechselt werden als beim optischen Einlesen, nämlich typischerweise die auf der Tastatur benachbarten Zeichen (etwa <i> und <o> oder <b> und <v>). Die Einsetzbarkeit morphologischer Verfahren der o.g. Art ist von diesem Unterschied nicht betroffen, und solche Verfahren werden von gängigen Textverarbeitungsprogrammen heute bereits weitgehend, wenn auch nicht immer zufriedenstellend, ausgenutzt[8].

1.4. Automatische Übersetzung

Noch eine Stufe später setzt der Einsatz digitaler Morphologie bei der automatischen Übersetzung an. Anders als bei den Plausibilitätsprüfungen zur Etablierung eines „korrekten" Texts wird hier ein solcher bereits vorausgesetzt; bei der morphologischen Analyse geht es dann statt dessen um die Herstellung der Basis für eine adäquate Übersetzung. Es entfallen somit grapho- und phonotaktische Begleitverfahren; die Probleme rund um

8 So ergibt z.B. die Überprüfung des fehlerhaften Satzes „Das Wortform ‚Mundem' ist unsinig." in MS-Word 2003 lediglich die Zurückweisung von „Mundem" sowie von „unsinig"; der Artikel *das* wird nicht als falsch erkannt.

Homonymien etc. bleiben jedoch bestehen, und auch hier müssen syntaktosemantische Verfahren vorrangig einbezogen werden.

1.5. Sprach- und sprachwissenschaftlicher Unterricht

Daß die Morphologie ein unabdingbarer Bestandteil des Fremdsprachenunterrichtsund, mehr noch, der sprachwissenschaftlichen Ausbildung ist, dürfte keinem Zweifel unterliegen. Digitale Verfahren können hier in zweierlei Weise unterstützend eingesetzt werden, nämlich bei der Analyse fremdsprachiger Textmaterialien im Hinblick auf gegebene Wortformen und, umgekehrt, bei der Erzeugung von Wortformen. Die hierfür zu schaffenden Datenbasen sind im Prinzip dieselben wie die bereits oben beschriebenen: Für weniger anspruchsvolle Anwendungen wird man zunächst mit einem „statischen", rein lexikalischen Verfahren operieren können, das lediglich eine bestimmte (aber erweiterbare) Menge von manuell vordefinierten Wortformen umfaßt; um eine morphologische Analyse beliebiger Wortformen zu ermöglichen, wird hingegen wieder ein „gemischtes" Verfahren eingesetzt werden müssen, bei dem neben Listen lexematischer und nicht-lexematischer Morpheme auch deren Verknüpfungsregeln (gegebenenfalls unter Einschluss von Um- und Ablautsphänomenen) verarbeitet sind. Beide Verfahren seien kurz an Implementationsbeispielen aus der TITUS-Textdatenbank[9] illustriert.

1.5.1. Tocharisch

Wie auch andere TITUS-Texte sind diejenigen des A-tocharischen Corpus[10] mit einer relationalen Datenbank verknüpft, die die Belegstellen jeder einzelnen Wortform umfaßt und somit einen unmittelbaren Zugriff auf die Beleglage derselben ermöglicht; dies geschieht durch einfaches Anklicken der betreffenden Wortform im Text oder durch Eingabe in eine Suchmaske. Darüber hinaus ist das A-tocharische Corpus bereits mit morphologischen Informationen versehen, die die Bestimmung von Verbalformen betreffen. Klickt man z.B. auf die in Abb. 1 enthaltene Verbalform *kumseñcä*,[11] so

9 Die Textdatenbank des TITUS-Projekts („Thesaurus Indogermanischer Text- und Sprachmaterialien") ist online über `http://titus.uni-frankfurt.de/texte/texte2.htm` verfügbar.

10 S. `http://titus.uni-frankfurt.de/texte/etcs/toch/tocha/tocha.htm`.

11 Da der Unicode-Standard noch nicht alle für die traditionelle Transkription des Tocharischen erforderlichen Zeichen umfaßt, sind die toch. Texte in der TITUS-Datenbank vorerst in einer Behelfsumschrift wiedergegeben; hier vertritt z.B. ein Majuskel-<Ä> das „stumme" <ä> neben Virāma im Auslaut.

erfolgt ein Datenbankaufruf, der die Analyse der Form ausgibt (s. Abb. 2). Die Einbettung der Form in ihren paradigmatischen Zusammenhang (sowie ihr Verhältnis zu den entsprechenden B-tocharischen Formen) kann durch ein weiteres Anklicken der Wurzel (*käm*) abgerufen werden (s. Abb. 3). Das dahinter stehende Verfahren ist, im Sinne der obigen Definition, rein „lexikalisch", da jede Verbalform mit ihrer Bestimmung sowie mit dem Verweis auf das jeweilige Lemma (Wurzel) für sich bereits in der Datenbank abgelegt ist; eine eigentliche morphologische Analyse findet also nicht beim Aufruf statt, sondern ist bereits vorher (manuell) in die Datenbank eingeflossen (lediglich gewisse graphische Alternationen werden beim Aufruf automatisch „abgefangen")[12]. Dieses Verfahren empfiehlt sich bei einer Corpussprache mit begrenztem Wortformenvorrat wie dem Tocharischen durchaus, zumal wenn die verbale Formenbildung wie in dieser Sprache in erheblichem Maße durch ablautbedingte und andere Stammvariationen gekennzeichnet ist und viele Irregularitäten aufweist.

1.5.2. Vedisch

Während die „halbautomatische" Analyse im Falle des Tocharischen noch auf das Verbum beschränkt ist, steht für die vedischen Texte in der TITUS-Kollektion bereits ein weitergehendes morphologisches Retrievalsystem zur Verfügung, das zumindest die Wortformen der Ṛgveda-Saṃhitā vollständig erfaßt. Auch diese Datenbank ist „lexikalisch", indem sämtliche Wortformen in ihr vorab bestimmt und auf ihr jeweiliges Lemma bezogen enthalten sind;[13] lediglich die durch Sandhi entstehende Variation wird durch einen eigenen Regelmechanismus beim Aufruf abgeglichen. Dieser kann wiederum durch einfaches Anklicken einer im Text gegebenen Wortform erfolgen (vgl. Abb. 4 mit RV 1,1,1a–c); ausgegeben wird dann zunächst eine morphologische Bestimmung mit Bedeutungsangabe wie in Abb. 5 für *agnim*[14]. Darüber hinaus steht bereits eine lemmabezogene Suchmaschine zur Verfügung, die für ein Lexem wie *agni-* „Feuer" sämtliche bezeugten

12 Dies betrifft generell den Wechsel zwischen den sog. „Fremdzeichen" und den eigentlichen Brāhmī-Akṣaras, aber auch z.B. das o.e. „stumme" <ä> im In- und Auslaut.

13 Grundlage ist die RV-Konkordanz von A. Lubotsky; Bedeutungsangaben greifen zusätzlich auf ein von P. Schreiner (Zürich) erarbeitetes Digitalisat des Sanskrit-Wörterbuchs von K. Mylius (mit Ergänzungen des Autors) zurück. Die Programmierarbeit wurde im Rahmen des Projekts „Avesta und Ṛgveda: Elektronische Analyse" (AUREA, 1995-1998; cf. `http://titus.uni-frankfurt.de/curric/aurea/aurea.htm`) von R. Gehrke erbracht.

14 Die morphologische Bestimmung ist (auch abgesehen von unklaren Wortformen) noch nicht in allen Einzelfällen endgültig verifiziert; die Retrievalergebnisse sind deshalb als noch nicht verläßlich gekennzeichnet.

Formen abzurufen gestattet; vgl. Abb. 6, die den Aufruf des „Thesaurus-Suche“ ausgehend von der Akkusativ-Form *agnim* zeigt,[15] und Abb. 7, die das mit der Nominativ-Form *agniḥ* beginnende Suchresultat darstellt. Es versteht sich von selbst, daß eine Datenbank, die allein die in der Ṛgveda-Saṃhitā auftretenden Wortformen umfaßt, den kompletten vedischen Formenbestand nicht erschließen kann; eine Ausweitung im Sinne eines gemischten „lexikalisch-morphologischen” Verfahrens ist also unumgänglich, wenn man die altindische Überlieferung in größerem Umfang abdecken will. Dies gilt umso mehr, als eine Ausdehnung auf das epische, buddhistische und klassische Sanskrit durch die Möglichkeiten der Wortkomposition praktisch wieder lexikalische Unendlichkeit mit sich bringt.

1.5.3. Alt- und Neugeorgisch

Ein „gemischtes“ lexikalisch-morphologisches Verfahren ist in der TITUS-Datenbank z.B. für das umfangreiche altgeorgische Textmaterial implementiert worden, und zwar für dessen nominale Bestandteile. Hierzu wurde zunächst der Wortbestand der Sammlungen von I. Abuladze, Z. Sarjveladze und H. Fähnrich[16] als Lemmaliste mit deutschen Übersetzungen digitalisiert; dann wurden sämtliche in der nominalen Formenbildung auftretenden Morpheme (Suffixe, Endungen) mit ihren Kombinationsmöglichkeiten in einer eigenen Datentabelle erfaßt. Ruft man diesen Datenverbund nun über das Anklicken einer Form wie *ġaġadebisay* in Mk. 1,3 innerhalb der altgeorg. „Protovulgata“ ab (s. Abb. 8),[17] so erhält man die korrekte Analyse wie in Abb. 9 dargestellt ausgegeben. Dabei ist zu beachten, daß bei der Rückführung auf das Lemma *ġaġadeba-y* „Rufen” zunächst die Nominativ-Endung *-y* und das stammauslautende *-a-* dieses Verbalnomens („Masdar”) durch den Endungskomplex *-isay* verdrängt sind und letzterer seinerseits in drei Morpheme zerfällt, nämlich die Genetivendung {is}, deren sog. „emphatische” Erweiterung {a} sowie die Nominativendung {i}, die nach {a} als <y> = [j] repräsentiert ist. Daß die Genetivendung ihrerseits um eine Nominativendung erweitert ist, ist im Altgeorgischen durch die Regeln der sog. „Suffixaufnahme” bedingt: der Genetiv „des Rufens” ist an der gegebenen Stelle von dem im Nominativ

15 S. http://titus.fkidg1.uni-frankfurt.de/search/query.htm.

16 Abulaze *Masalebi* 1973; Sarǯvelaze *Masalebi* 1995; Sardschweladse /Fähnrich *Wörterbuch* 1999. Zugrunde liegt eine elektronische Fassung des letzteren Wörterbuchs, das die Materialien der beiden erstgenannten mit umfaßt; für die Bereitstellung sei H. Fähnrich und Z. Sarǯvelaze auch an dieser Stelle herzlich gedankt.

17 S. http://titus.uni-frankfurt.de/texte/etca/cauc/ageo/nt/cinant/cinan.htm.

erscheinenden *qmaɣy* „die Stimme" als Attribut abhängig[18] und diesem nachgestellt und erfordert deshalb die zusätzliche Markierung für den Kasus seines Bezugsworts. Die einzelnen Analyseelemente seien noch einmal zusammengefaßt:

- Lemmaform <ġaġadebay>: {ġaġad}{eb}{a}{i}, bestehend aus {Verbal-Wz.} +{Präs.-Stammbildungssuffix}+{Masdar-Suffix}+{Nom.-Endung}
- zu analysierende Form <ġaġadebisay>: {ġaġad}{eb}{a}{is}{a}{i}, bestehend aus {Verbal-Wz.}+{Präs.-St.}+{Masdar-S.}+{Gen.-E.}+{„emph." El.}+{Nom.-E.}
- Regel: {Masdar-Suffix} > Ø/_{Gen.-Endung} | {Instr.-Endung}

Daß ein entsprechendes Analyseverfahren für das altgeorg. Verbum noch nicht entwickelt wurde, liegt an der enormen Formenfülle, durch die dieses gekennzeichnet ist, wobei Präfixe, Suffixe, Infixe und Ablaut involviert sind. Auch ein solch komplexes System kann jedoch als Regelapparat programmiert werden, wie die von Paul Meurer (Bergen) implementierte Verbalformenanalyse des Neugeorgischen beweist.[19] Man vgl. z.B. die in Abb. 10 wiedergegebene Ausgabeseite, die die paradigmatische Einbettung der Wortform *daumalavs* zeigt. Die Form ist in korrekter Weise doppelt erfaßt, nämlich a) als Futurform „er/sie/es wird ihn/sie/es//sie vor ihm/ihr verbergen" und b) als Perfektform „er/sie/es soll ihn/sie/es// sie (Sachen) verborgen haben". Dabei ist zu berücksichtigen, daß sich die beiden Formen allein dadurch unterscheiden, daß die Markierung des handelnden Subjekts (Agens) in ersterer in der Endung *-s* steckt, in letzterer jedoch in dem präradikalen *-u-*, das in der Futurform ein indirektes („versionales") Objekt „vor ihm/ihr/ihnen" bezeichnet:

- Futur {da}{u}{mal}{av}{s}:
  {Präv.}+{ind.Obj.V.3.Ps.}+{Wz.}+{Präs./Perf.-Stammbildungssuffix} +{Subj.3.Sg.}
- Perfekt {da}{u}{mal}{av}{s}:
  {Präv.}+{Subj.3.Ps.}+{Wz.}+{Präs./Perf.-Stammbildungssuffix} +{dir.Obj.3.Sg.}

18 Mit der „Stimme des Rufens in der Wüste" (statt „des Rufers", griech. τοῦ βοῶντος) zeigt die altgeorg. Bibel an der gegebenen Stelle, einem Zitat aus Is. 40,3, im übrigen eine bemerkenswerte Übereinstimmung mit der armen. Bibel (*jayn barbarōy,* wtl. „Stimme des Sprechens"), die auf eine syrische Vorlage weist.

19 Die Analyse beruht auf einer Implementierung der morphologischen Angaben in K. Tschenkelis *Wörterbuch*; s. `http://maximos.aksis.uib.no/Aksis-wiki/Georgian _ Grammar _ Project`

Es sei noch hinzugefügt, daß auf der Basis eines derartigen, umfassenden morphologischen Formengenerators auch eine tiefgehende syntaktische Analyse programmierbar ist; ein Prototyp für das Georgische ist von Paul Meurer als Finite-State-Anwendung auf der Basis der „Lexical-Functional Grammar" (LFG)[20] entwickelt worden (vgl. Abb. 11).

2. Linguistische Zielsetzungen digital basierter Morphologie

Alle bisher behandelten Anwendungsbeispiele setzen linguistisches Wissen voraus, führen jedoch selbst allenfalls in begrenztem Maße zu einer Weiterentwicklung unseres Wissens über einzelne Sprachen oder menschliche Sprache im allgemeinen. Dennoch kann die Digitalisierung morphologischer Strukturen auch ihrerseits in erheblichem Maße zu linguistischem Kenntnisgewinn beitragen. Die betreffenden Einsatzbereiche seien im folgenden kurz umrissen.

2.1. Einsatzbereich Lexikographie

Elektronische Textcorpora schaffen eine hervorragende Basis für lexikographische Untersuchungen aller Art bis hin zur Erstellung vollständiger Konkordanzen. Um das Textcorpus einer Sprache mit reichhaltiger Flexionsmorphologie lexikographisch in konsistenter Weise auszuwerten, bedarf es einer Lemmatisierung, die nur dann automatisch erfolgen kann, wenn eine korrekte Bestimmung sämtlicher Wortformen gewährleistet ist.[21] Hierfür ist es erforderlich, Verfahren zu implementieren, wie sie oben am Vedischen und, weitergehend, am Georgischen illustriert wurden.

2.2. Einsatzbereich Grammatikographie

Auf der Basis von Textcorpora kann eine Digititalisierung morphologischer Strukturen praktisch in allen Bereichen zur Unterstützung einer grammatischen Beschreibung der betr. Sprache eingesetzt werden. Dies betrifft zum einen die Morphologie der Sprache selbst, insofern, je nach Umfang des Corpus, mehr oder weniger erschöpfende Aussagen zur Gestaltung und Distribution von Morphemen möglich werden. Darüber hinaus ist

20 S. dazu z.B. Kroeger *Approach* 2004.

21 Vgl. den 1994 publizierten Wortformenindex zu den Schriften Galens (Gippert, *Index Galenicus*), bei dessen Erarbeitung noch nicht auf eine automatische Lemmatisierung des Altgriechischen zurückgegriffen werden konnte. Heute, 15 Jahre nach der Erarbeitung dieser Wortformenkonkordanz, würden geeignete Mittel für eine automatische Lemmatisierung bereitstehen.

die Syntax betroffen, insofern sie (als Morphosyntax) die funktionale Verwendung von Morphemen zum Gegenstand hat. Auch für die Phonologie kann digitalisierte Morphologie Erkenntnisse bringen, soweit sie z.B. die lautliche Ausprägung von Allomorphen und phonotaktische Phänomene an Morphemgrenzen zu untersuchen erlaubt.

Dabei ist es zweckmäßig, zwei Verfahren zu unterscheiden. Das erste, das ich „intrinsisch" nennen würde, basiert wieder auf der digitalen Grundstruktur von Text, nämlich der „Zeichenkette" (String); ein darauf aufbauendes Auswertungsverfahren („zeichenkettenbasierte Suche") wird allerdings nur bei (nahezu) eindeutiger Form, d.h. variantenloser Schreibung, und bei (nahezu) eindeutigem, klar definierbaren Zeichenkontext zufriedenstellende Ergebnisse zeitigen. Dies sei am Beispiel gotischer Passivformen illustriert.

2.2.1. Gotisches Passiv

Die (wenigen) synthetischen Passivformen des Gotischen sind bekanntlich durch spezifische Endungen gekennzeichnet, nämlich, im Präsens Indikativ, *-za* für die 2.Ps.Sg., *-da* für die 1. und 3.Ps.Sg. und *-nda* für die Pluralpersonen. Will man diese Endungen als solche abfragen, so muß zunächst sichergestellt werden, daß sich die Abfrage nur auf wortfinales *-za* bzw. *-(n)da* bezieht; diese Einschränkung kann in der TITUS-Suchmaschine durch die Eingabe eines Sternchens als Stellvertreterzeichen für eine beliebige Zeichenfolge markiert werden, das den Wortkörper repräsentiert („Wildcard"; vgl. Abb. 12 mit der Eingabe für die Endung *-za* in der Form **za*). Die daraufhin ausgegebene Wortformenliste zeigt auf den ersten Blick (s. den Ausschnitt in Abb. 13) die Unzulänglichkeit dieses Verfahrens auf; denn neben den (zwei) gewünschten Passivformen (*haitaza* „du wirst genannt (werden)", Lk. 1,76,[22] und *us-maitaza* „du wirst abgehauen (werden)", Röm. 11,22) finden sich hier in weitaus größerer Zahl andere Formen, die auf *-za* auslauten, wie z.B. die Komparative *maiza* und *minniza* „größer, mehr" und „kleiner, minder" (Mt. 11,11 u.ö.) oder die substantivischen Dativformen *riqiza* „der Finsternis" (Mt. 10,27 u.ö.) und *Moseza* „(dem) Mose" (2.Tim. 3,8). Noch weitaus diffuser ist die Liste, die bei der Abfrage nach **da* ausgegeben wird (s. den Ausschnitt in Abb. 14); denn hier erscheinen insbesondere Präteritalformen auf *-da* (wie z.B. *gaweihaida* „er hat geheiligt", Jo. 10,36) in großer Zahl.

22 Im intr. *heißen* = „genannt werden" (gegenüber trans. *jdn. (etw. tun) heißen*) manifestiert sich im Deutschen offensichtlich noch eine Reminiszenz an die frühere Existenz eines (Medio-)Passivs des gotischen Typs.

2.2.2. Annotative Verfahren

Angesichts solcher Ergebnisse ergibt sich die Notwendigkeit, ein besser geeignetes Verfahren anzuwenden, nahezu von selbst. Als ein „extrinsisches" Verfahren, das bereits weithin genutzt wird, bietet sich die durchgehende Annotation von Textdaten an. Hierbei werden durch ein sog. „Tagging", vorzugsweise auf der Basis der sog. „Extended Markup Language" (XML), die für eine gezielte Abfrage erforderlichen Daten (Formenbestimmungen, Lemmaangaben) in den Text eingebunden, so daß sie für gezielte Abfragen verfügbar sind. Der Vorteil eines derartigen Verfahrens besteht in der rel. frei definierbaren und anpassungsfähigen Informationsstruktur: Das Tagging kann in nahezu beliebiger Weise explizit (d.h. als lesbarer „Klartext") oder implizit (in Form von Abkürzungen, Siglen etc.) gehalten werden, nur Konsistenz muß gewahrt bleiben. Abb. 15–16 zeigen zur Illustration eines maximal expliziten Taggings einen Ausschnitt aus dem Bonner Frühneuhochdeutschcorpus[23] in verschiedenen Darstellungsformen[24]. Als Beispiel für ein besonders wenig explizites Annotationsverfahren kann man denselben Ausschnitt aus der ursprünglichen Fassung des Corpus (Abb. 17) oder den von Han Steenwijk aufbereiteten Slovenischen (Resischen) Katechismus (Abb. 18) vergleichen. Eines muß natürlich klargestellt werden: Wie explizit auch immer das Tagging ist – es reflektiert auf jeden Fall bereits vorher akkumuliertes linguistisches Wissen.

Inwiefern kann ein annotationsbasiertes Retrieval dann überhaupt selbst zum linguistischen Erkenntnisgewinn beitragen? Eine ganz wesentliche Funktion liegt in der Verifikation und Falsifikation von Hypothesen, die das Verfahren erlaubt. Dabei geht es insbesondere um eine statistische Untermauerung, d.h. die Frage nach der Stringenz linguistischer Annahmen im Hinblick auf in Corpora enthaltene Daten. Warum ein Retrieval auf der Basis von Annotationen dabei einer einfachen Suche nach objektsprachlichen Zeichenketten überlegen ist, sei wiederum an einem Beispiel illustriert.

23 S. http://www.ikp.uni-bonn.de/dt/forsch/fnhd/.

24 Die Ausgabe von XML-Strukturen hängt von dem jeweiligen Verarbeitungsprogramm ab; hier dargestellt sind die frei verfügbaren XML-Editoren von firstobject (http://www.firstobject.com/dn_editor.htm) und XMLFox (http://www.xmlfox.com/download.htm). Auf der Seite http://titus.uni-frankfurt.de/ld/links.htm sind online verfügbare XML-Werkzeuge zusammengestellt.

2.2.3. Echofragen im Kaukasus

Wer die georgische Sprache erlernt, wird sich über kurz oder lang wundern, daß Echofragen in dieser Sprache in spezifischer Weise ausgeprägt sind, insofern sie mit den auf sie folgenden Antworten durch die Konjunktion *da* „und" verbunden werden; man vgl. das folgende Beispiel:[25]

Frage:	*ras aḳetebt axla tkven?* „Was tut ihr gerade?"
Echofrage+Antwort:	*čven ras vaḳetebt da meored motibul tivas vparcxavt.* „Was wir tun? ‚Und' – wir rechen die zweite Mahd zusammen."

Wollte man derartige Konstellationen nun in einem größeren Corpus des (gesprochenen) Georgischen abfragen, so würde man mit einer rein zeichenkettenbasierten Suche schwerlich in vertretbarer Zeit zum Erfolg kommen, da eine Suche nach selbständigem *da* sowohl sämtliche Belege für die Konjunktion (und damit das häufigste Wort überhaupt) als auch solche für das homonyme *da* „Schwester" (Nom.Sg.) erbringen würde. Noch schwieriger würde sich die entsprechende Suche im Svanischen, einer der Schwestersprachen des Georgischen, gestalten, das einen genau entsprechenden Gebrauch der Konjunktion „und" kennt. Diese lautet hier normalerweise *i* wie in der wörtlichen Entsprechung des obigen Beispiels:

Frage:	*im xašdbad atxe sgäy?*
Echofrage+Antwort:	*nä im xwašdbad i mēris xwipcxid.*

In der Position nach vokalischem Auslaut wird das *i* nun aber als [j] realisiert und, sofern svanische Texte niedergeschrieben werden, als solches an das vorhergehende Wort angehängt, so daß eine Suche nach (selbständigem) *i* diese Fälle nicht erbringen würde:

Frage:	*si zurāl, im xašdba?* „Du, Frau, was tust du (gerade)?"
Echofrage+Antwort:	*im xwašdbay ulyāks xwäpšwde.* „Was ich tue? 'Und' – ich schere ein Schaf."

25 Die folgenden Beispiele sind dem im Rahmen des Projekts „Endangered Caucasian Languages in Georgia" (s. `http://titus.fkidg1.uni-frankfurt.de/ecling/ecling.htm`) gesammelten originalsprachlichen Material entnommen. Die Beispiele wurden nicht gezielt eruiert, sondern sind jeweils Bestandteil ungesteuerter Konversation.

Noch komplexer ist die Sachlage in einer zweiten, ihrerseits nicht verwandten Nachbarsprache des Georgischen, dem Tsova-Tuschischen oder Batsischen. Auch hier können, möglicherweise unter georgischem Einfluß, Echofragen mit den folgenden Antworten durch „und" verbunden sein; die Konjunktion liegt dabei jedoch nicht als eigenständiges Wort (und somit als eindeutig abgrenzbare Zeichenkette) vor, sondern manifestiert sich jeweils unterschiedlich in der Längung des auslautenden Vokals des vorangehenden Wortes, der seinerseits in anderen Konstellationen getilgt sein kann; man vgl. das folgende Beispiel:

Frage:	*men yar o pstuyn?* „Wer war diese Frau?"
Echofrage+Antwort:	*men yarē xaxabuyren...* „Wer sie war? ('Und') – sie war aus Xaxabo..."

Es kommen jedoch auch Echofragen ohne eine derartige Längung vor:

Frage:	*qēn nānas vux dīen?* „Was tat (seine) Mutter dann?"
Echofrage+Antwort:	*nānas vux dīen? –nān korlacyīen...* „Was (seine) Mutter tat? (Seine) Mutter wurde verhaftet..."

Die sich hieraus unmittelbar ergebende Frage nach der Konsistenz des Gebrauchs kann durch eine Zeichenkettensuche in keiner Weise beantwortet werden, da eben kein zu suchendes Zeichen gegeben ist. Man wird also, um das Phänomen der Echofragen im Kaukasus im Hinblick auf seine Regelhaftigkeit weiter zu untersuchen, nicht umhin können, eine Annotation vorzusehen, die nicht nur die morphologischen Elemente, sondern sogar syntaktische Einheiten (eben Echofragen als solche) markiert.

2.3. Digitale Morphologie und diachrone Untersuchungen

Einen weiteren Einsatzbereich digitalisierter Morphologie innerhalb der Linguistik stellen diachrone Untersuchungen dar. Dabei gilt wieder das Grundprinzip, daß der Einsatz digitaler Verfahren v.a. zur Verifizierung bzw. Falsifizierung von an Einzelfällen erarbeiteten Hypothesen dienen kann. Auch dies sei durch ein Anwendungsbeispiel aus dem kaukasischen Raum illustriert.

Das Svanische fällt unter den kartvelischen oder südkaukasischen Sprachen insbesondere dadurch auf, daß es in mehrere Dialekte mit höchst unterschiedlicher phonologischer Struktur zerfällt. Die historischen Veränderungspro-

zesse, die diese Diversität herbeigeführt haben, umfassen höchst komplexe Umlautungen, Assimilationen sowie Syn- und Apokopeerscheinungen. So würde man etwa die Form *äžhīdx* „sie sind über euch gekommen" des oberbalischen Dialekts auf eine historisch zugrundeliegende Form **adžihīdex* zurückführen, die sich über die folgenden vier chronologisch aufeinanderfolgenden, an das Altirische erinnernden Schritte verändert haben dürfte:

Umlautung:	**ädžihīdex*
Apokope (Tilgung des Vokals der Auslautssilbe):	**ädžihīdx*
Synkope (Tilgung kurzer Vokale in geraden Silben):	**ädžhīdx*
Clustervereinfachung:	*äžhīdx*

Dabei ist zu bemerken, daß ein anderer Dialekt, der lenṭechische, die Synkope noch nicht vollzogen hat und in älteren svanischen Volksliedtexten sogar noch Formen ohne Apokope begegnen.

Die historisch anzusetzende „ursvanische" Grundform ist nun zugleich auch diejenige, die man bei einer morphologischen Analyse von *äžhīdx* zugrundelegen muß. Tatsächlich zerfällt diese Form in insgesamt sechs morphologische Elemente, nämlich das Präverb {ad} „her", das Zeichen eines Objekts der 2. Person, verbunden mit dem Zeichen der „objektiven Version" {ž+i} „in Bezug auf dich / euch", die Verbalwurzel {hīd} „kommen", das Suffix des Aorists, {e}, sowie das Zeichen für ein Subjekt der 3. Person Plural, {x}, von denen aber eben nicht mehr alle in der heutigen oberbal. Form als solche „sichtbar" sind. Schematisch:

I		II		III		IV		V		VI
{ad}	+	{ž}	+	{i}	+	{hīd}	+	{e}	+	{x}
{Präv.}	+	{2.Ps.Obj.}	+	{Obj.V.}	+	{Wz.}	+	{Aor.Suff.}	+	{3.Pl.Subj.}
äžhīdx										

Wie in diesem Fall ist zu erwarten, daß eine konsistente morphologische Analyse regelmäßig Formen liefern wird, die den mutmaßlichen historischen (ursvanischen) Ausgangsformen zumindest nahekommen. Sie können damit als Inputformen für eine (automatische) Verifizierung der postulierten lautgesetzlichen Entwicklungen vom Ursvanischen zu den einzelnen heutigen Dialekten wie auch für die postulierten interdialektalen Lautentsprechungen dienen.[26]

26 Vgl. bereits Gippert, *Divergences* 2000 mit einigen Fallbeispielen.

3. Resümee

Die Perspektiven, die die Anwendung digitaler Morphologie in der Sprachwissenschaft eröffnet, lassen sich auf der Grundlage der oben behandelten Fallbeispiele in zwei Thesen zusammenfassen:

These I: Je mehr sprachliche Daten mit morphologischer (und, darüber hinaus, syntaktischer, semantischer, pragmatischer, metrischer...) Analyse digital aufbereitet sind, auf desto sichererem Boden steht die Theoriebildung in synchron-deskriptiver und diachron-vergleichender Linguistik.

These II: Durch die Digitalisierung großer Datenmengen wird eine (statistische) Verifizierung bzw. Falsifizierung sprachwissenschaftlicher Hypothesen nicht nur möglich, sondern geradezu als methodisches Postulat erzwungen.

Durch eine entsprechende Aufbereitung sprachlicher Daten die Voraussetzungen hierfür zu schaffen, sehe ich als eine der vorrangigen Aufgaben der heutigen Linguistik an.

Literaturverzeichnis

Abulaʒe, Ilia. (1973). *ʒveli kartuli enis leksiḳoni. (Masalebi).* Tbilisi: Mecniereba.

Corkhill, Alan. (1991). "Zum Sprachdenken Goethes in beziehungsgeschichtlicher Hinsicht." in: „Neophilologus“ 1991 (Band 75), S. 239-259.

Gippert, Jost. (1994). *Index Galenicus. Vollständiger alphabetischer Wortformenindex zu den Schriften Galens.* Dettelbach: Röll.

Gippert, Jost. (2000). "Towards an Electronic Analysis of Svan Dialectal Divergences". in: *Kartvelian Heritage*, 2000 (Band 4), Kutaisi: Kutaisi State University, S. 134-149

Goethe, Johann Wolfgang von. (1889–1896). *Werke* (Weimarer Ausgabe). Weimar: Böhlau, / online unter `http://glc.chadwyck.com/framesets/goethe_frameset.htm`.

Hoffmann, Karl. (1958). "Altiranisch". in: *Handbuch der Orientalistik, I. Abtlg.: Der Nahe und Mittlere Osten*, Bd. IV: *Iranistik*, 1. *Abschnitt: Linguistik*. Leiden/Köln: Brill, S. 1-19.

Hoffmann, Karl. (1975). „Aufsätze zur Indoiranistik“. Hrsg. von Johanna Narten. Bd. I., Wiesbaden: Reichert.

Kroeger, Paul. (2004). *Analyzing Syntax: A Lexical-Functional Approach*. Cambridge: Cambridge University Press.

Lubotsky, Alexander. (1997) *A Ṛgvedic Word Concordance*. Part I–II. New Haven / Conn.: American Oriental Society.

Mylius, Klaus. (1975) *Wörterbuch Sanskrit-Deutsch*. Leipzig: VEB-Verlag.

Sardschweladse, Surab - Fähnrich, Heinz. (1999). *Altgeorgisch-deutsches Wörterbuch*. Hamburg: Buske.

Sarǯvelaǯe, Zurab. (1995). *ʒveli kartuli enis leksiḳoni. Masalebi*". Tbilisi: Tbilisis Universiṭeṭis Gamomcemloba.

Schiller, Friedrich von. (1943-). *Werke (Nationalausgabe)*. Weimar: Böhlau online unter `http://glc.chadwyck.com/framesets/schiller_frameset.htm`.

Schleicher, August. (1871). *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*". Weimar: Böhlau.

Tschenkeli, Kita. (1965-1974). *Georgisch-deutsches Wörterbuch*. Zürich: Amirani-Verlag.

Abbildungen

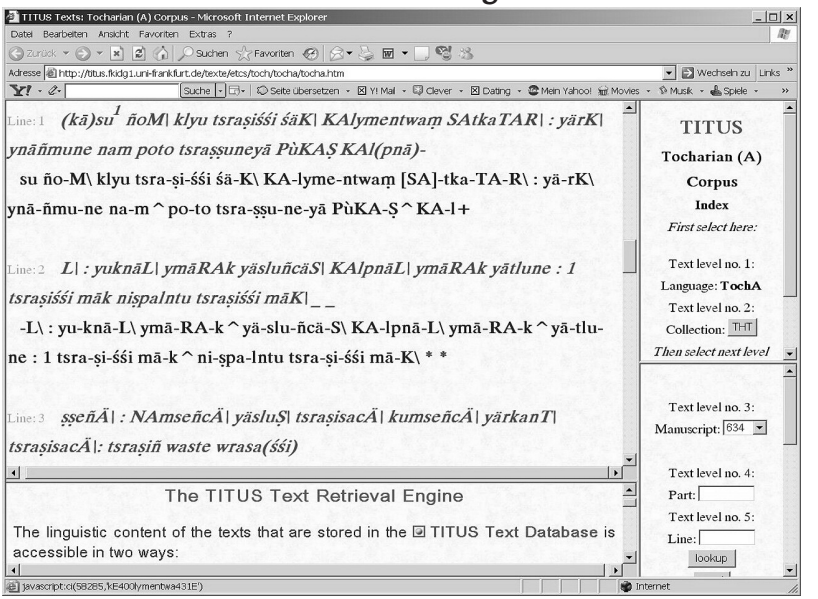

Abb. 1: A-tocharischer Textausschnitt (THT 634r, 1-3)

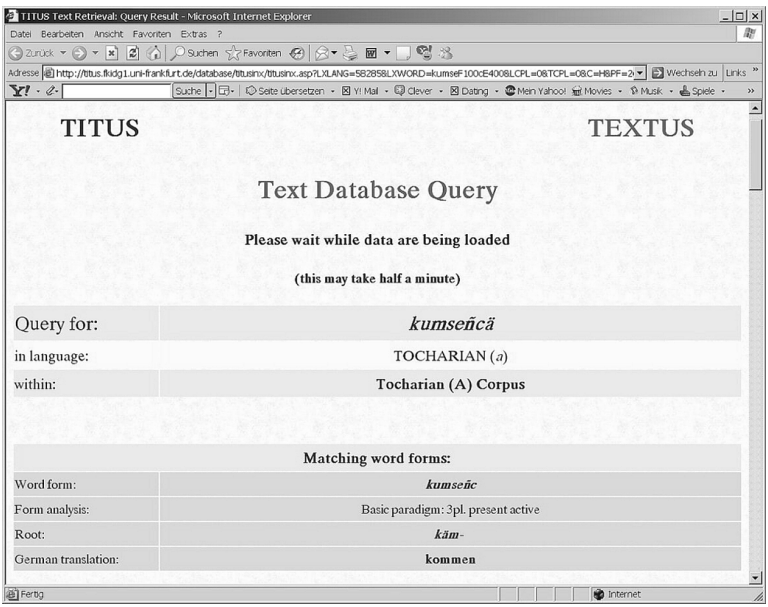

Abb. 2: Aufruf der Datenbank

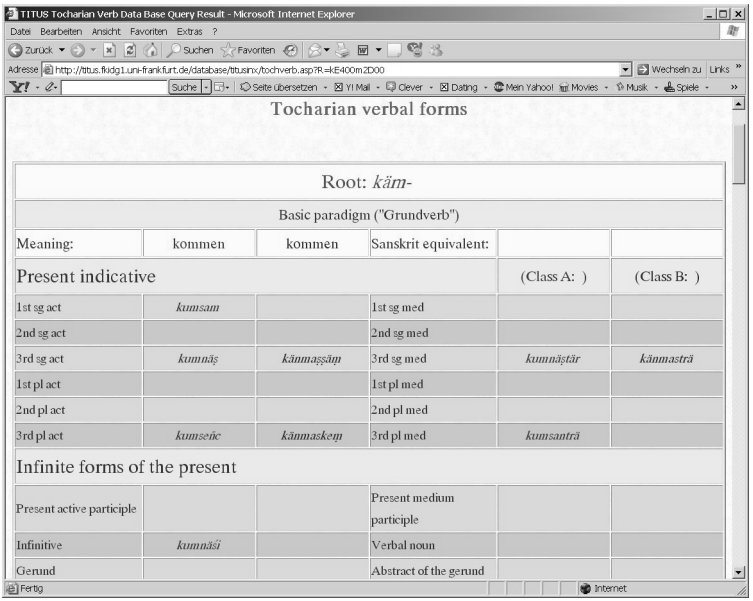

Abb. 3: Paradigma-Ausgabe

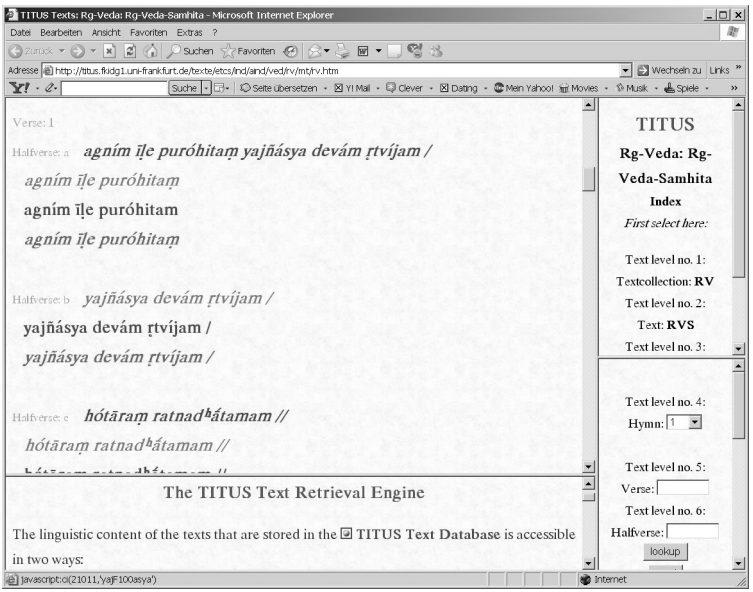

Abb. 4: Vedischer Textausschnitt (RV 1,1,1a)

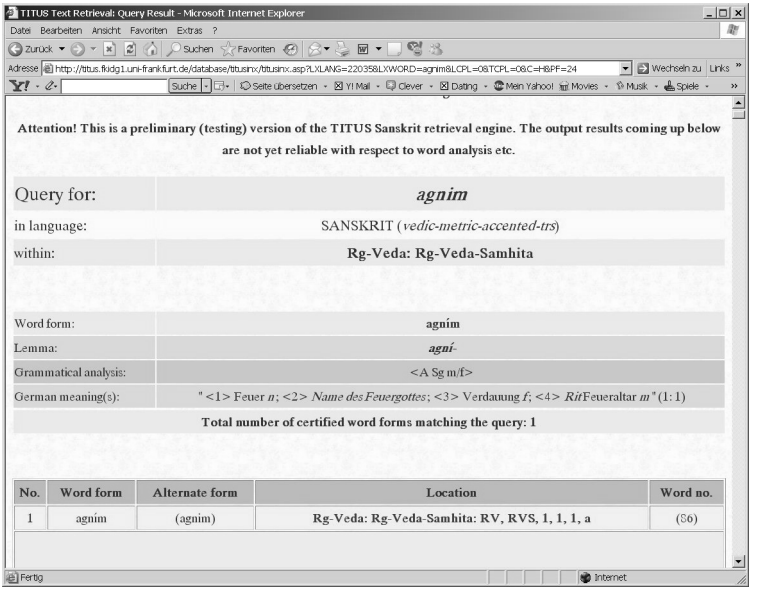

Abb. 5: Bestimmung der Form *agnim*

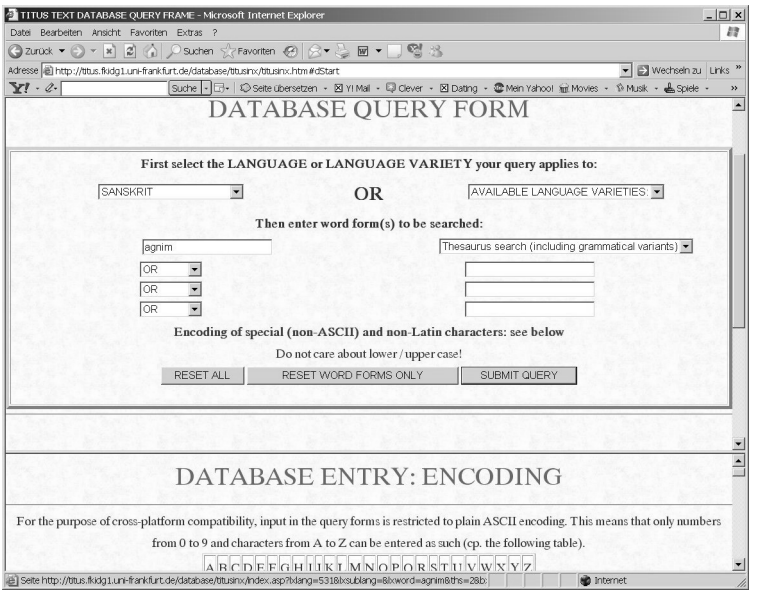

Abb. 6: Aufruf der Thesaurus-Suchmaschine

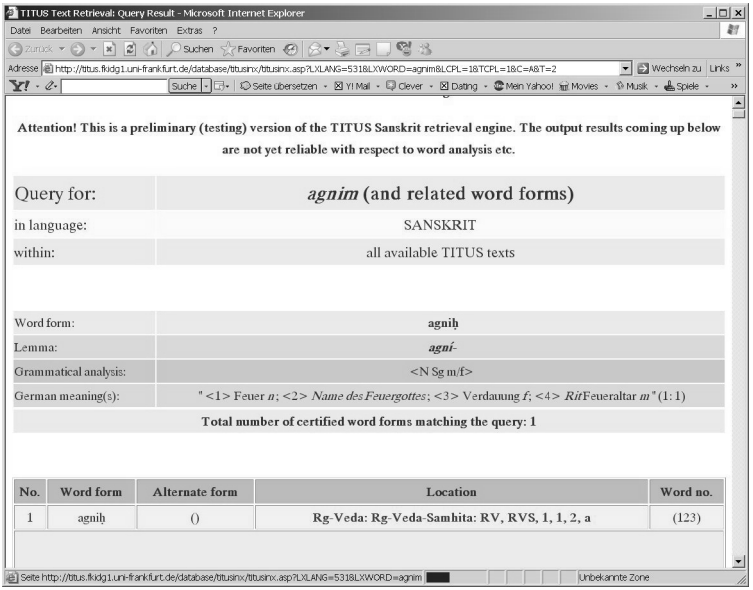

Abb. 7: Ausgabe der Thesaurus-Suche

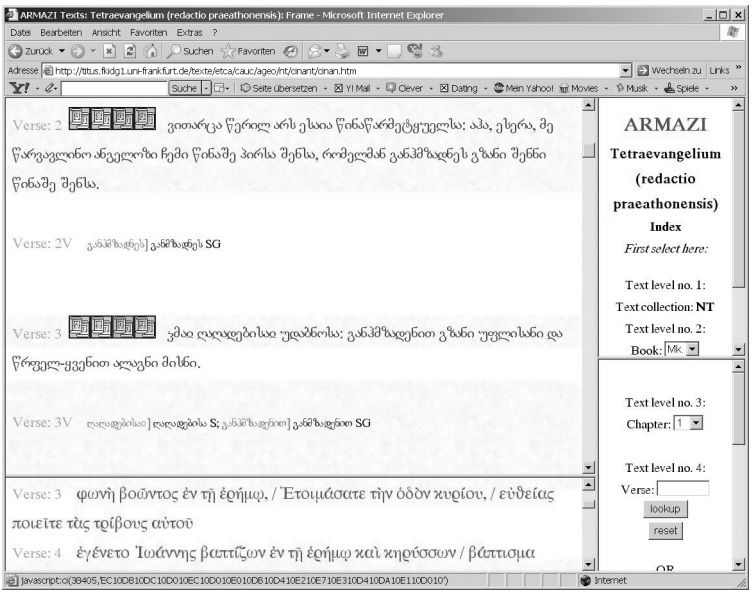

Abb. 8: Mk. 1,2–3 in der altgeorgischen „Protovulgata“

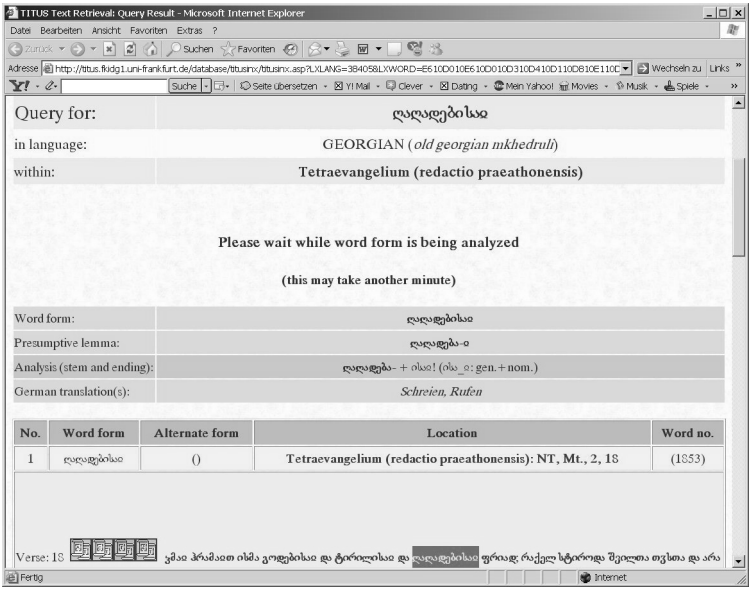

Abb. 9: Bestimmung der altgeorg. Wortform *ġaġadebisay*.

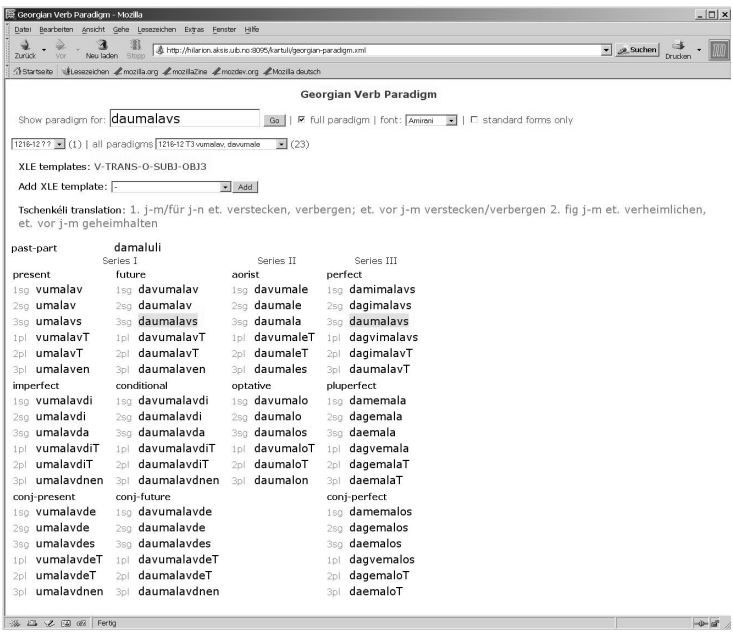

Abb. 10: Verbformengenerator für das Neugeorgische (Paul Meurer)

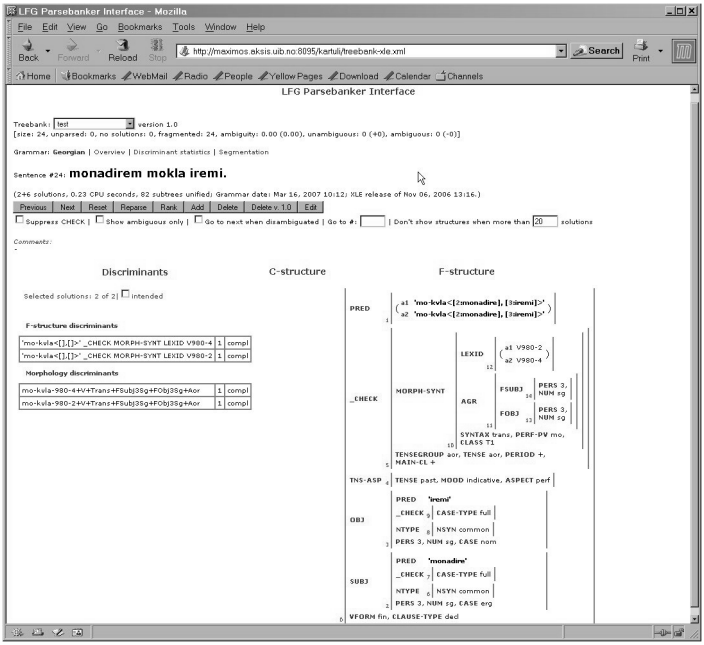

Abb. 11: Georgische Syntaxanalyse (Paul Meurer)

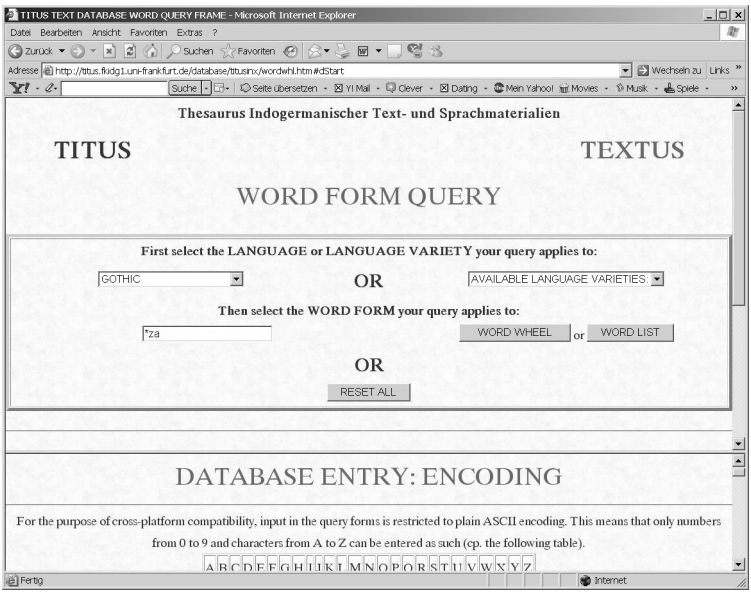

Abb. 12: Abfrage der gotischen Endung *-za*

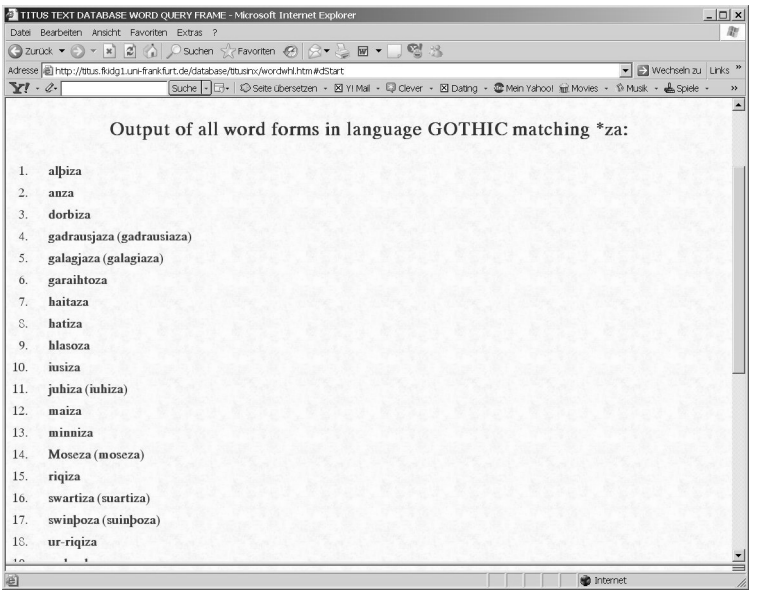

Abb. 13: Ausgabeliste von Wortformen auf *-za*

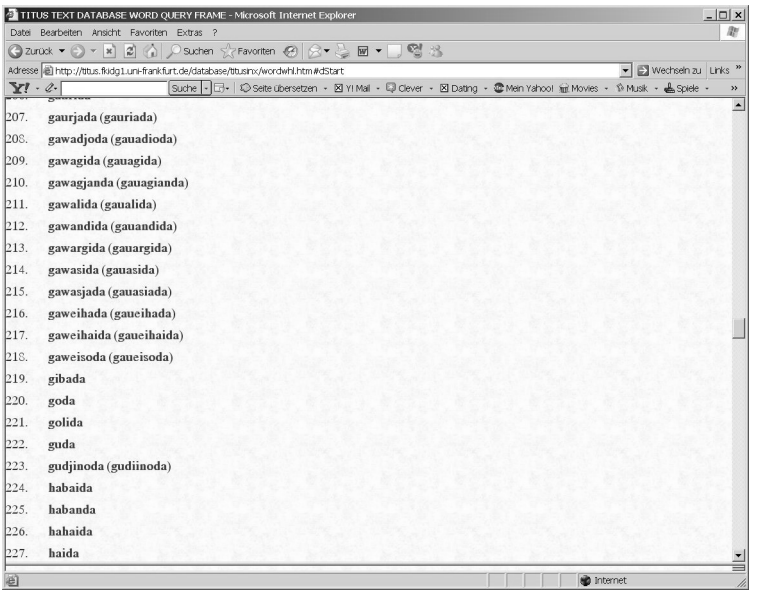

Abb. 14: Ausgabeliste von Wortformen auf *-da*

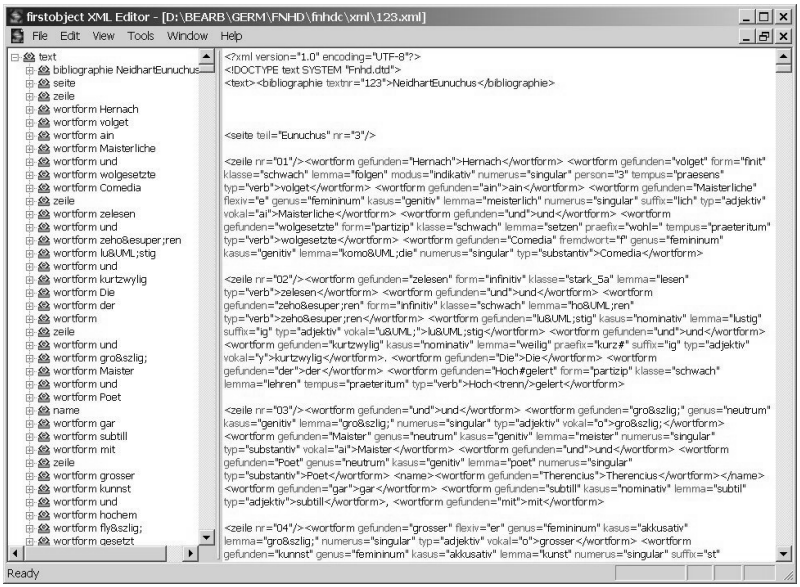

Abb. 15: Annotation (Tagging) im Bonner Frühneuhochdeutschcorpus

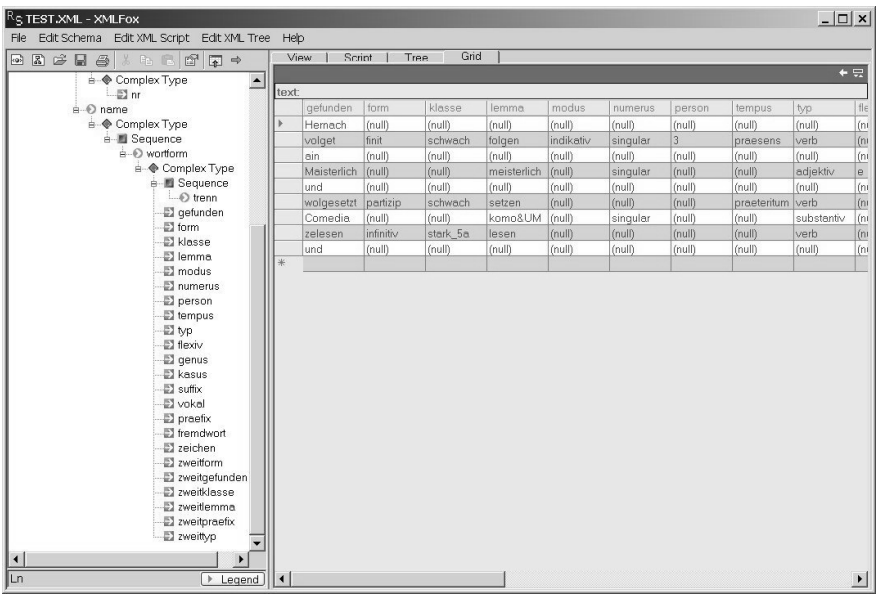

Abb. 16: Dasselbe in systematischer Darstellung

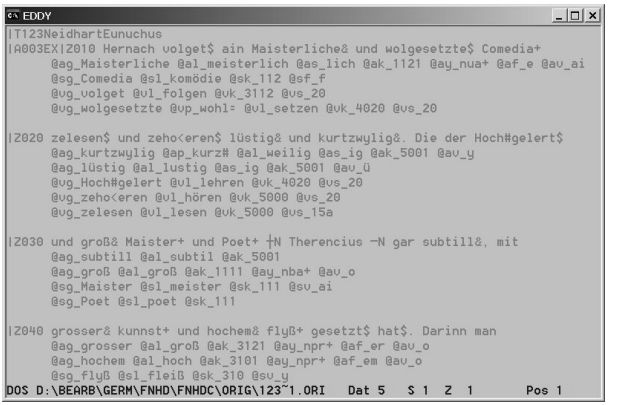

Abb. 17: Dasselbe, in nicht explizitem Tagging

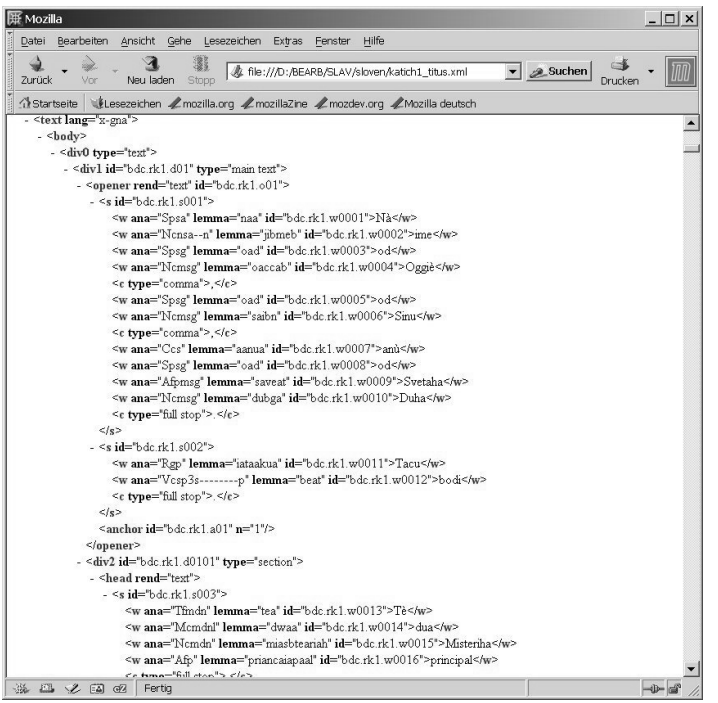

Abb. 18: Implizites Tagging: Slovenischer Katechismus