BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae 4

Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae

4

BERLIN 2013

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Perspektiven einer corpusbasierten historischen Linguistik und Philologie

Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN

2013

Dieser Band wurde durch die gemeinsame Wissenschaftskonferenz im Akademienprogramm mit Mitteln des Bundes (Bundesministerium für Bildung und Forschung) und des Landes Berlin (Senatsverwaltung für Wirtschaft, Technologie und Forschung) gefördert

URN: urn:nbn:de:kobv:b4-opus-24310

INHALTSVERZEICHNIS

VORWORT 7

GREGORY CRANE & ALISON BABEU
Global Editions and the Dialogue among Civilizations 11

HISTORISCHE CORPUS-PROJEKTE – SYNCHRON UND DIACHRON

STÉPHANE POLIS & JEAN WINAND
The Ramses project. Methodology and practices in the annotation of Late Egyptian Texts 81

SERGE ROSMORDUC
The Ramses project in perspective. Managing evolving linguistic data 109

DIETER KURTH
Das Edfu-Projekt. Ziel, Methode und Verarbeitung der lexikographischen Ergebnisse 121

INGELORE HAFEMANN & PETER DILS
Der Thesaurus Linguae Aegyptiae – Konzepte und Perspektiven 127

GÜNTER VITTMANN
Zur Arbeit an der Demotischen Textdatenbank: Textauswahl 145

GERNOT WILHELM
Das Hethitologie Portal Mainz 155

JOST GIPPERT
The TITUS Project. 25 years of corpus building in ancient languages 169

KURT GÄRTNER & RALF PLATE
Die Doppelfunktion des digitalen Textarchivs als Wörterbuchbasis und als Komponente der Online-Publikation. Am Beispiel des Mittelhochdeutschen Wörterbuchs 193

HANS-CHRISTIAN SCHMITZ, BERNHARD SCHRÖDER & KLAUS-PETER WEGERA
Das Bonner Frühneuhochdeutsch-Korpus und das Referenzkorpus ‚Frühneuhochdeutsch' 205

ALEXANDER GEYKEN
Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv 221

BRYAN JURISH
Canonicalizing the Deutsches Textarchiv 235

WORTGESCHICHTE - TEXTGESCHICHTE - SPRACHGESCHICHTE: TRADITION UND INNOVATION BEI DER TEXTPRODUKTION

FRANK FEDER & SIMON D. SCHWEITZER
Auf dem Weg zu einem integrierten Lexikon des Ägyptisch-Koptischen 245

FRIEDHELM HOFFMANN
Die Demotische Wortliste – virtuell erweitert 263

GÜNTER VITTMANN
Kursivhieratische Texte aus sprachlicher und onomastischer Sicht 269

MATHEW ALMOND, JOOST HAGEN, KATRIN JOHN, TONIO SEBASTIAN RICHTER & VINCENT WALTER
Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen: Lehnwort-Lexikographie im Projekt Database and Dictionary of Greek Loanwords in Coptic (DDGLC) 283

THOMAS GLONING
Historischer Wortgebrauch und Themengeschichte. Grundfragen, Corpora, Dokumentationsformen 317

LOUISE GESTERMANN
Die altägyptischen Sargtexte in diachroner Überlieferung 371

THOMAS STÄDTLER
Überlegungen zu Textsorte und Diskurstradition bei der Beschreibung von Textcorpora und ihr Bezug zur lexikographischen Forschung 385

VORWORT

Die internationale Tagung „Perspektiven einer corpusbasierten historischen Linguistik und Philologie“ vom 12. – 13. Dezember 2011 am Akademienvorhaben „Altägyptisches Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) war dem Thema des Aufbaus und der Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher in den historischen Sprachen gewidmet. Die Teilnehmer, Vertreter der Ägyptologie, der Hethitologie, Indogermanistik sowie Referenten aus der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen diskutierten vor allem über die Veränderungen, die mit dem Einsatz elektronischer Erfassungs- und Verarbeitungsprozeduren einhergehen. Vertreter der Computerlinguistik vom „Zentrum Sprache“ der BBAW wurden in die Diskussionen einbezogen. Dort beschäftigt man sich seit Jahren mit dem Aufbau großer elektronischer Textcorpora (DWDS), darunter auch solcher, die historische Texte (DTA) für die elektronische Nutzung ermöglichen.

Die größte Herausforderung dieser neuen elektronischen Corpora und Wörterbücher ist es, sowohl den Methoden und damit den wissenschaftlichen Ansprüchen der traditionellen Philologie und Lexikographie unbedingt verpflichtet zu bleiben als auch neue Gebiete wie die Corpus- und Computerlinguistik für die historischen Sprachen zu öffnen. Die Teilnehmer haben gemeinsam und disziplinenübergreifend die Möglichkeiten und Grenzen der Datenerfassung, ihrer Präsentation und den Nutzen neuer Auswertungsprozeduren diskutiert.

Unter dem ersten Thema „Historische Corpusprojekte – synchron und diachron“ wurden elektronische Corpora vorgestellt und ein intensiver Austausch darüber geführt, welche Datenstrukturen die linguistischen Inhalte in adäquater Weise abbilden. Wichtig war die Frage, auf welche Resonanz diese elektronischen Corpora bei den Nutzern gestoßen sind und welche Erwartungen und Anforderungen aus den verschiedenen Fachdisziplinen an die Projekte herangetragen werden. Der Austausch über Nutzungsperspektiven elektronischer Corpora schloss auch die Diskussion über die Erarbeitung projektübergreifend einsetzbarer Standards der Codierung und Strukturierung historischer Textdaten mit ein. Hinsichtlich einer mittel- und langfristigen Nutzbarkeit sowie einer langfristigen Datensicherheit stehen solche Fragen zunehmend im Focus und einige aktuelle Initiativen dazu wurden vorgestellt. Spezielle technische Aspekte

elektronischer Datenerfassung und automatischer Analyse- und Speicherungsverfahren elektronischer Textdaten konnten am letzten Tag als ein Themenschwerpunkt mit den Programmierern diskutiert werden.

Ein zweiter Schwerpunkt waren konkrete Fragstellungen aus der historischen Lexikographie und diachronen Textanalyse. Für das Ägyptische ist der diachrone Ansatz auf Grund der über viertausendjährigen Textüberlieferung von großer Relevanz. Themen wie historischer und/oder textgattungsspezifischer Wortgebrauch, die Erarbeitung diachroner Wortlisten und Aspekte des kontaktindizierten Sprachwandels konnten disziplinübergreifend zwischen den Ägyptologen und den Kollegen der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen behandelt werden.

Mit dem Abendreferenten Gregory Crane, dem Begründer der „Perseus Digital Library“, wurde ein breites Publikum angesprochen. In seinem Vortrag hat er noch einmal die hohe Relevanz und die neuen Möglichkeiten der Einbeziehung zahlreicher Wissenschaftler und einer interessierten Öffentlichkeit in die Projektarbeit demonstriert, die das Internet auf völlig neue Weise eröffnet hat. Die Herausgeberin ist sehr froh, seinen programmatischen Beitrag zu diesem Thema, dessen schriftliche Form er gemeinsam mit Alison Babeu erarbeitet hat, ebenfalls in diesem Band präsentieren zu können.

Wir danken der Berlin-Brandenburgischen Akademie der Wissenschaften für die umfassende Unterstützung unserer Projektarbeit und ganz speziell der Vorbereitung dieser Konferenz sowie der Möglichkeit, die Akten auf dem E-Doc-Server der Akademie veröffentlichen zu können.

Der Hermann und Elise geborene Heckmann Wentzel-Stiftung sei hiermit ausdrücklich für die unbürokratische und großzügige finanzielle Unterstützung dieser erfolgreichen Tagung gedankt.

Das Akademienvorhaben „Altägyptisches Wörterbuch“ konnte sich als aktives Mitglied des Weiteren auf das „Zentrum Grundlagenforschung Alte Welt“ stützen, dem alle altertumswissenschaftlichen Vorhaben der BBAW angehören. Dem Zentrum ist es zu danken, dass der Abendvortrag von Gregory Crane einem breiteren Publikum dargeboten werden konnte.

Allen Autoren dankt die Herausgeberin für ihre anregenden Diskussionen und die qualitätvollen Beiträge in diesem Band.

Auf eine Gesamtbibliographie wurde verzichtet und die Abkürzungen der in den ägyptologischen Beiträgen erwähnten Zeitschriften und Reihen folgen dem Lexikon der Ägyptologie, herausgegeben von Wolfgang Helck und Wolfhart Westendorf, Band VII: Nachträge, Korrekturen, Indices, Wiesbaden 1992, XIV-XIX.

Ganz besonders sei schließlich Frau Angela Böhme für die gewissenhafte redaktionelle Bearbeitung der Manuskripte gedankt sowie Dr. Simon Schweitzer für seine Hilfe beim Erstellen des Layouts.

Berlin, Mai 2013 Ingelore Hafemann

GLOBAL EDITIONS AND THE DIALOGUE AMONG CIVILIZATIONS

GREGORY CRANE & ALISON BABEU

> "If we want to identify one idea which through the whole of history is visible in ever broader effect, if any [idea] proves the often contested, but even more often misunderstood perfection of all mankind, it is the idea of Humanity, the struggle to remove the hostile boundaries which prejudices and biased perspectives have placed between human beings and to treat all of humanity without regard to religion, nationality, or color, as one great, closely related family, as a single whole for the achievement of a single goal, the free development of individual power. This is the final, external goal of sociability at the same time the inborn inclination of human beings to the unconstrained expansion of their destiny." – "On the duties of the historian," Wilhelm von Humboldt (1821)[1]

> "By selecting these two specimens of German scholarship we should indeed adduce the most favourable instances which could be found, but should not exemplify the general character of the German philologer. For, in their activity of mind and body, Hermann and Lachmann came nearer to Englishmen than 99 out of 100 Germans." – John William Donaldson (1856)[2]

This paper is about the reinvention of editing source texts from the human record. Editing may be largely a technical, frequently a tedious, and almost always an underappreciated task, but editing can have profound effects upon the world. We have an opportunity, one could argue an urgent necessity, to establish a dialogue among civilizations. When information flows back and forth across the world in real time, the alternative to dialogue is conflict. The quotations above illustrate two fundamental forces that strain against

---

1 VON HUMBOLDT, W., 1821: *Über die Aufgabe des Geschichtsschreibers*, Berlin: „Wenn wir eine Idee bezeichnen wollen, die durch die ganze Geschichte hindurch in immer mehr erweiterter Geltung sichtbar ist; wenn irgendeine die vielfach bestrittene, aber noch vielfacher missverstandene Vervollkommnung des ganzen Geschlechtes beweist: so ist es die Idee der Menschheit, das Bestreben, die Grenzen, welche Vorurteile und einseitige Ansichten aller Art feindselig zwischen die Menschen gestellt, aufzuheben; und die gesamte Menschheit ohne Rücksicht auf Religion, Nation und Farbe als einen großen, nahe verbrüderten Stamm, als ein zur Erreichung eines Zweckes, der freien Entwicklung innerer Kraft, bestehendes Ganzes zu behandeln. Es ist dies das letzte, äußere Ziel der Geselligkeit und zugleich die durch seine Natur selbst in ihn gelegte Richtung des Menschen auf unbestimmte Erweiterung seines Daseins."

2 DONALDSON, J. W., 1856: *Classical scholarship and classical learning considered with especial reference to competitive tests and University teaching*, Cambridge, 157, http://books.google.com/books?id=riACAAAAQAAJ.

one another whenever anyone reflects upon the past. Wilhelm von Humboldt, a Prussian aristocrat and product of the Berlin Enlightenment, sees in the study of history an opportunity to lower the barriers that separate humanity. John Donaldson reduces the study of Greek and Latin to a proxy for the superiority not only of European culture within the world but also of the British upper classes within Europe.

If we follow a path such as Humboldt described, our goal is to increase understanding across humanity. The goal is not to eradicate difference but to promote a dialogue among civilizations – a dialogue that European and North American voices do not impose upon the rest of the world. In 1998, the then Iranian President Mohammed Khatami called for a dialogue among civilizations as an alternative to the "Clash of Civilizations" which thinkers such as a Samuel Huntington had seen as a successor the Cold War.[3] President Khatami's call did not fall upon deaf ears and the United Nations (UN) declared a year of Dialogue among Civilizations. "I see," Secretary General Kofi Annan asserted, "dialogue as a chance for people of different cultures and traditions to get to know each other better, whether they live on opposite sides of the world or on the same street."[4] The official UN English website introduced the topic: "What does a dialogue among civilizations mean? One could argue that in the world there are two groups of civilizations – one that perceives diversity as a threat and the other which sees it as an opportunity and an integral component for growth. The Year of Dialogue Among Civilizations was established to redefine diversity and to improve dialogue between these two groups. Hence, the goal of the Year of Dialogue Among Civilizations is to nurture a dialogue which is both preventive of conflicts – when possible – and inclusive in nature."[5]

It would not be difficult to find similarly contrasting statements in every major language – narrow exclusivity is inherent in our Hobbesian, primate natures, but the cosmopolitan aspirations that we find in Humboldt appear – and will always reappear. Every nation with the opportunity to do so has fallen far short of Humboldt's ideas in the two centuries since they were composed but these failures only emphasize the need to reassert a shared humanity

---

3 HUNTINGTON, S., 1996: *The Clash of Civilizations*, New York.

4 http://www.un.org/dialogue/.

5 http://www.un.org/dialogue/background.html.

and to view in the complexity and diversity of human cultures an opportunity for each of us to learn and to grow. Nor does such a dialogue of civilizations reflect a European or North American attempt to reduce cultures to their own categories. The then president of Iran, Mohammed Khatami, called for such a dialogue and the United Nations responded by declaring a year for the Dialogue among Civilizations. That year was 2001 and the events of 9/11 set in motion a new chain of violence that smothered dialogue but the need for that dialogue remains and is only the greater. When the bombs fall or the door is kicked in before dawn, dialogue may seem a futile, even a laughable instrument. But dialogue, born not only of solemn respect but also of curiosity and delight, provides an essential instrument against violence and for civilization, if that word is to have any meaning.

Greek, Latin, and the Dialogue among Civilizations

As a practical initial goal, we should build a space whereby those who can work with any one of several modern languages can work directly with a range of historical languages.

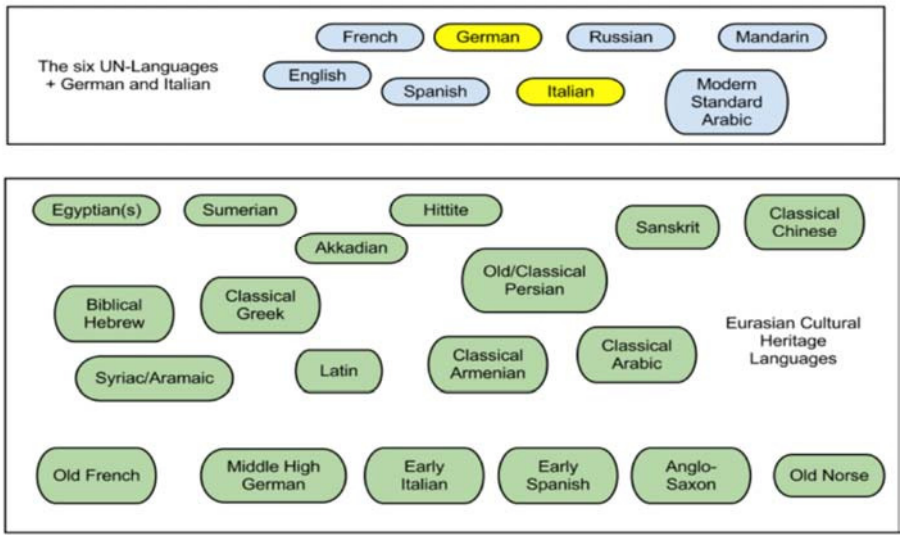


Figure 1: A Euro-centric view of major languages (the six UN languages including German and Italian because of their historical importance in the study of Greek and Latin).

The figure above lists eight modern languages; those in blue boxes are the six official languages of the United Nations. A European contribution to the dialogue among civilizations would probably need to consider including support as well for German and Italian because a great deal of information about the Greco-Roman world is available in these languages. A speaker of Chinese or Russian should, for example, be able to work with information about the Greco-Roman world that is available in French or German. Here the task is to optimize very large systems already emerging to help individuals work across multiple modern languages.[6] Students of historical languages should shrewdly track, exploit and, where appropriate, contribute to new multilingual services such as improving machine translation, information extraction, and cross-language information retrieval.[7] Different communities could extend the coverage to meet their own needs – the European Union might, for example, well want to provide coverage for more European languages, while India might consider support for Hindi, Bengali, Telugu and other major languages.

The lower part of the above figure illustrates a selective and Eurocentric subset of nineteen historical language types. Some of the languages, such as Persian and Egyptian, refer generally to languages that have evolved over thousands of years, from records in cuneiform and hieroglyphics through classical sources in Arabic script. Some of these languages (e.g., Latin, Classical Chinese) remained languages of publication for thousands of years. If we are to support a substantive dialogue among civilizations, we might begin by developing an environment to enable anyone who can understand one of the modern languages above to work directly with materials in any of the other supported modern languages and with any from a subset of historical languages such as those listed below. Thus, a Chinese speaker interested in Alexander the Great should be able to work directly with the lives of Alexander that survive by Plutarch and Quintus Curtius Rufus in Greek and Latin respectively, as well as any

---

[6] For a detailed overview of the use of multilingual technologies to provide cross language access to digital libraries, see DIEKEMA (2012) and for the growing need for such tools in technology enhanced language learning, see ANTONIADIS *et al.* (2009).

[7] A useful overview of the potential of these and other natural language processing technologies for cultural heritage texts and historical languages has been provided by PIOTROWSKI (2012) and SPORLEDER (2010) while a particular focus on the use of these tools for manuscripts has been presented by VERTAN (2010).

supporting scholarship in English, French, German, Italian, Spanish and Russian.

Europe and the Americas can contribute to, but could not, even if they wished, control, a dialogue among civilizations. The people of Europe and the Americas must depend upon their fellows elsewhere to support languages such as classical forms of Chinese, Sanskrit, Persian, Arabic other historical languages. Analysis of the most recent statistics from the Modern Language Association (MLA) indicates that, in the United States at least, the early modern big three of Classical Greek, Latin, and Biblical Hebrew account for more than 95% of all enrollments in historical languages (66,668 of 68,877). Greek and Latin alone accounted for more than three quarters of the total (53,246). Personal experience and conversations with colleagues suggest that the situation in Europe is not much different.

		2006	2009
Latin	Classical and Medieval	31,400	31,369
Greek	Ancient, Koine, Biblical, Old Testament	22,788	21,877
Hebrew	Biblical	14,098	13,422
Aramaic		2,556	562
Sanskrit		607	483
Arabic	Classical	4	285
Chinese	Classical	113	202
Akkadian		129	195
Egyptian[8]		56	110
Slavic	Old Church	133	73
German	Middle High	9	55
Others		223	244
Totals		72,116	68,877
Greek + Latin		54,188	53,246

[8] The MLA statistics do not define what "Egyptian" means in this context. The figure above probably counts those studying the dialect of Arabic currently spoken in Egypt but the figure is included because Egyptian could cover earlier forms of the language (e.g., Coptic, Demotic, Hieroglyphic).

percentage	**75.1%**	**77.3%**
Greek + Latin + Hebrew	**68,286**	**66,668**
percentage	**94.7%**	**96.8%**

Table 1: Enrollments in historical languages based upon figures from the Modern Language Association.[9]

The goal is not to reduce the number of students studying Greek, Latin and Hebrew but to increase the number of those engaged with every historical language – the aggregate 2006 and 2009 enrollments of 72,000 and 68,000 are far too low. Each student of a historical language serves also as a proxy both for broader interest, and access to classes, in a given language. The vanishingly small numbers listed for Classical Sanskrit, Arabic and Chinese reflect the economics of brick and mortar universities and colleges, where each class must draw a minimum number of students to be taught. As distance learning evolves, we will be able to draw upon much larger populations of students and staff courses on more languages – it is easier to find 15 students for a language in a population of 500,000 students (such as represented by the US http://www.cic.net/) than in a liberal arts college of 2000.

In the short run, if we in Europe and the Americas wish to advance a global dialogue among civilizations and to advance a digital infrastructure to support that dialogue, we need to begin by focusing upon Greek and Latin for both diplomatic and practical reasons. First, Greek and Latin are the two major cultural heritage languages to which no region outside of Europe or the Americas can assert a proprietary claim and feel usurped by a Western hegemony. And, second, because there are not enough students of languages other than Greek, Latin, and Hebrew in Europe and the Americas to do the work that is needed – for, as this paper will suggest, our automated systems have now created immense needs and opportunities for intellectual activity of every kind.

Digital Editions

The methods by which we disseminate Greek and Latin are based upon the limitations and possibilities of print technology. They are

[9] http://www.mla.org/2009_enrollmentsurvey; http://www.mla.org/2006_flenrollmentsurvey.

obsolete – indeed, our editions are cultural fossils, retaining archaic forms that now assume and perpetuate a dwindling specialist audience. These forms were, however, originally designed to reach beyond barriers of language, religion and nation. Our task is to re-imagine how to address that ancient goal with the methods available in a digital space.[10]

Non-specialists, interested in the Greco-Roman world, may shake their heads curiously if they happen to pick up the new print editions that specialists still create for one another. The introductions are still, for the most part, exercises in Latin prose composition. The textual notes consist of telegraphic abbreviations that can only partially represent the sources upon which they are based. And the most sophisticated editions still all too often lack an accompanying translation. Editors, of course, have very definite, often distinct, ways of understanding texts in which they have scrutinized every word but the editorial conventions of major editions still assume specialist audiences who can read the Greek or Latin source text on their own. The Greek and Latin editions of the twentieth century were monuments of a closed intellectual culture.

Greek and Latin editions played a different role in early modern culture. When the first editors of printed editions wrote their introductions and notes, even their translations from Greek, in Latin, they were asserting membership in a cosmopolitan European culture that transcended the petty duchies and kingdoms in which they lived. To write in Latin was to advance a transnational republic of letters and to assert a broader identity. The rise of vernaculars – much heralded as a triumph of mass culture – replaced a single language of publication to which no one ethnic group could lay special claim with a handful of culturally dominant dialects. As languages such as French, German, Italian and English emerged as literary media, speakers of these languages could dispense with Latin. Speakers of Croatian and Danish simply had to learn another foreign language – and to accept, in some measure, cultural, if not political domination, of more numerous contemporaries.

The editors of the twenty-first century can now pursue again – and indeed far more effectively – the cosmopolitan goals of their intellectual ancestors. We now have the tools at hand by which to

[10] A series of articles dedicated to this very topic were published in a special issue of *Digital Humanities Quarterly* in 2009, entitled, "Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure," http://www.digitalhumanities.org/dhq/vol/3/1/.

begin developing a new generation of editions, ones designed to serve not merely a European but also a global audience. The grand challenge for editors is not simply to represent a text in a general format but to do so in a format that allows the speaker of Chinese or Arabic to work directly with sources in Greek, Latin, and other European cultural heritage languages.

Adding a translation in a modern language with extensive computational support provides an initial first step: machine translations from English to Mandarin or from French to Arabic may be problematic but they exist and are steadily improving. A great deal more can be done – and the next generation of scholars can congratulate itself on its good fortune in reaching maturity just as our understanding of Greek, Latin, and every cultural heritage language is being reborn. The past is not simply a foreign country but a truly new world, ready to be discovered. Some prototypes exist but we are still in the incunabular stage of invention. No true digital editions exist for any authors.[11] After a generation of experimentation, however, the outlines of new editorial practices are beginning to appear.

The outlines may shift and the subject is in flux – an editor today could put their bets on the wrong services and find their work obsolete even as it is published. We do not know the precise nature of the future – but it hard to believe that the conventions of print will be those of the digital world. Conservative practice is the most promising path to obsolescence and, at best, a sighing sympathy from future readers. The safe bet – producing another edition on the print model – is the safest bet for failure. As students of Greek and Latin, we participate in a conversation that extends centuries and millennia into the past. Our print editions have been mature since Karl Lachmann in the nineteenth century if not before. We have an equal obligation to write, as best we can, for the future and to think in terms of decades and generations to come, rather than the practices that we have inherited.

Digital editions[12] must have the following characteristics:

---

[11] Paolo Monella has also commented on this phenomenon in a recent article, "Why are there no digital scholarly editions of "classical" texts?" http://folk.uib.no/hnooh/filologiadigitale/abstracts/Monella.pdf

[12] The topic of digital editions and how best to design them is a topic of intense discussion within the digital humanities community, and providing support for digital editions is frequently cited as an important task by large humanities cyberinfrastructure research projects, see for example NEDIMAH (Network for

1. Not texts, but multi-texts. Editions must be multi-texts, capable of representing the relationships between any number of versions that the text has assumed.[13] Print conventions present single reconstructions of an original source (a critical edition) or diplomatic representations of particular versions of that text (a diplomatic edition of a manuscript).[14] They represent a finite number of textual differences as manually constructed abbreviated formulas in the notes. These textual notes are often not machine actionable – we cannot dynamically reconstruct from these notes what different versions looked like or see immediately how different versions resembled one another. And different versions should include not only manuscripts and critical editions but also quotations and paraphrases. A digital edition should, as much as possible, trace the entire history of a text.

Within this framework, editors may argue for particular readings or suggest new corrections. They can also create complete networks of suggested readings but these readings constitute – as they have always constituted – a network of annotations that produces one particular version of the text while alluding to many other possible reconstructions. In a truly digital edition, the annotations are immediately separable, whether these constitute the original decisions in an editio princeps or a new anthology of earlier readings.[15]

In some, if not many cases, the earlier states of a text are more important than any new edition, however improved. The works of Galen in Greek, as well as in translations into Arabic and then from Arabic into Latin, served as medical textbooks for more than a thousand years. A new edition of a work by Galen, however much better it captures the original text, should never again inform medical practice. Literary, historical and philosophical works may

---

Digital Methods in the Arts and Humanities) recently announced expert meeting on scholarly editions (http://www.esf.org/index.php?id=8752).

[13] The literature regarding the utility of the digital environment for representing not only different versions of classical or historical texts but also their textual evolution is quite extensive; two recently published books have a number of chapters discussing this topic, see McCarty (2010) and Peursen (2010). For other important work in this area, see also Schmidt & Colomb (2009) and Monella (2008).

[14] For a discussion of "diplomatic editions" in the digital age, see Pierazzo (2011).

[15] For some interesting work in digitally mapping conjectures and variants to textual decisions within editio princeps, see Boschetti (2007) and Cisne *et al.* (2010).

continue to be important in their own right but Machiavelli's text of Livy or the editions behind Gibbon's Decline of the Roman Empire were not those that we use today. If we wish to understand the significance of historical sources in any language, we need editions that help us trace the history of those sources as fully as possible.

2. At least one aligned translation into a modern language. Digital editions must contain at least one major modern language, ideally with a translation that is aligned to the original source text. The modern language translation not only provides basic intellectual access to those who understand that language but also links the original text indirectly to the multi-lingual services available to the modern language (e.g., English) but either not available or not as fully developed for the source language (e.g., Classical Greek). Automatic systems can identify the relationship between most of the words in a Greek or Latin source text and the corresponding words in a modern language translation[16]. Editors can refine these automatic alignments and even optimize their translations to make the alignments more precise. Such optimization can affect the structure and vocabulary. Different translators will, as they always have, pursue different philosophies about how closely the translation should follow the original.

3. Machine actionable annotations as the foundation. Third, digital editions must more fully capture the linguistic interpretations of their editors. Print editions have for centuries added annotations not present in the manuscripts, inscriptions, or other original sources. These include punctuation, capitalization, paragraph breaks, indentation, and indices of people and places. Digital editions should include annotations that represent the editor's understanding[17] and that traditional print markup cannot represent nearly as well if at all.[18] Annotations should include, at a minimum, one or more interpretations of the morphological and syntactic structure of every

---

[16] Work in parallel text alignment is particularly applicable to this task (for a fairly recent overview of the state-of-the-art, see MIHALCEA & SIMARD (2005), and for some interesting work using parallel text alignment and markup projection, see BAMMAN *et al.* (2010)).

[17] O'DONNELL (2009) expands upon this idea of how digital editions can both build upon and improve the traditional practice of print critical editions in representing various textual witness and expert editorial opinions.

[18] For example, the EpiDoc schema (http://epidoc.sourceforge.net/), created for encoding inscriptions can be used to provide for far more sophisticated markup as well as multiple interpretations than is possible with the Leiden conventions, see CAYLESS *et al.* (2009).

word, identifications of every person, place, and similar named entity, metrical analyses, as well as alignments to at least one modern language translation.[19] Since editors traditionally invest a great deal of time pondering the function of every word in a text, the added labor of creating such annotations should be marginal. In practice, annotation should not be a final stage but should constitute a key element of digital editing, with editors using the discipline of linguistic annotation to make sure that they have considered every single word. Digital editions must also contain major alternative annotations.

4. Adequate expository argument to explain the decisions behind the machine-actionable annotations. Digital editions must contain sufficient explanations to justify the choices that their editors make. Even as digital editions exploit machine actionable annotations, expository narrative should justify the substantive decisions that these annotations reflect. There is no reason to have a volume of textual notes separate from the main edition or to create a distinct editio minor without most of the editorial data. The arguments traditionally printed in introductions, commentaries, and accompanying volumes are thus, if anything, more tightly integrated into the edition.

5. Open architectures. Digital editions must have open architectures[20] and can be dynamically constructed from many different elements, each of which has clearly identified provenance. Provenance[21] in turn includes the date at which a conjecture was first published or the number of editors who have endorsed a particular

---

[19] The importance of not only supporting different types of annotations within digital editing and textual scholarship but also the need for shared annotation models to provide interoperability between digital projects is quite vast. For an overview of the nature of digital annotations, see AGOSTI & FERRO (2007), and for recent work combining two of the most prominent annotation models, the Open Annotation Collaboration (http://www.openannotation.org/spec/core/) and the Annotation Ontology, see HUNTER & GERBER (2012).

[20] A number of recent projects have sought to develop open architectures (e.g. shared data models, services, tools and infrastructure) for the creation of digital scholarly editions including Interedition (http://www.interedition.eu), the Virtual Manuscript Room (http://vmr.bham.ac.uk), and TextGrid (http://www.textgrid.de/en/ueber-textgrid.html). For a detailed examination of the importance of developing critical editions as open access texts (including both the marked up text and any code used to generate the edition), see BODARD & GARCÉS (2009). Peter Robinson has also explored the importance of open architectures for the creation of digital editions, see ROBINSON (2010a, 2010b).

[21] For a recent look at designing workflows that support the unique needs of data provenance for philological research, see KÜSTER *et al.* (2011).

variant from one or more manuscripts. Provenance allows readers to reconstruct and to compare particular versions, the contributions that particular sources have made over time and who has endorsed those contributions. The open architecture allows readers to view a new edition in isolation or in conjunction with earlier editions and subsequent reviews. The open architecture also allows readers to link new proposed annotations immediately to the relevant passages in particular texts. The open architecture also allows members of the community to create new translations in a wide range of languages.

6. Dynamic knowledge bases rather than static visualizations. Printed editions – and their PDF imitations – are static visualizations. Digital editions are dynamic entities that evolve over time. Editors may still create comprehensive editions, in which they produce new translations and re-examine many old questions, publishing their own selection of earlier annotations and of their own conjectures. But with digital editions readers can integrate new materials as they appear. Students of the text will add notes on particular passages, studies of particular phenomena, and surveys of the reception of a text.[22] Readers have the freedom to define the texts according to parameters that they choose.

The situation in 2012

According to the criteria listed above, no digital editions yet exist – and no digital editions will soon fully satisfy all six criteria for any textually complex work. But the services, collections and even communities are now in place that can begin to build the textual sources needed to enable broader dialogue and deeper understanding of the human record than has ever before been possible. Computational linguistics, broadly construed, allows us to extract machine actionable text from analogue representations such as images and sound files and then to detect meaningful patterns across vast bodies

[22] There is growing recognition of the need to design digital editions as dynamic sources that lend themselves to both student contributions and collaborative editing between scholars, teachers and students. For example, the Textus Project (http://textusproject.org/), from the Open Knowledge Project, is an "open source platform for working with collections of texts" that "enables students, researchers and teachers to share and collaborate around texts using a simple and intuitive interface." Similarly, the INKE (Implementing New Knowledge Environments) project is examining how best to design tools and interfaces to support an intersection of social media and the creation of "online scholarly editions" (SIEMENS *et al.* 2012). For some other related perspectives, please see BEAULIEU & ALMAS (2012) and GIBBS (2011).

of texts composed in hundreds, if not thousands, of languages.[23] Where computational linguistics focuses largely upon automated processes that can be applied to open ended collections, corpus linguistics develops well-defined, ever more richly annotated corpora to study linguistic phenomena.[24] In the traditional terminology of information retrieval, computational linguists excel at recall (they can detect far more phenomena than human annotators could ever manually examine) while corpus linguists emphasize precision (they focus on annotations of high accuracy in scientifically designed corpora).

As this document is composed in late 2012, many on-going efforts in Europe and the Americas are laying tangible foundations for new digital editions of historical languages such as Greek and Latin. These efforts include at least five different threads, each of which contributes to an emergent fabric of intellectual life; (1) mass digitization, (2) scalable, highly granular collections, (3) customized Optical Character Recognition (OCR), (4) transcription and structural markup, (5) text-reuse detection, (6) machine actionable annotations such as named entity identification and morpho-syntactic analysis, and (7) more decentralized structures for intellectual activity, integrating the contributions of student researchers and citizen scholars.

1. Mass digitization. Gallica[25], Google Books[26], and the Internet Archive[27] are only the most prominent efforts that have made digital images of millions of documents openly accessible to a net public that has, by recent estimates,[28] reached 2.3 billion – one third of humanity. These digital images represent not only books but also manuscripts, papyri, inscriptions and virtually every text-bearing

---

[23] The use of computational linguistics, particularly text mining and data mining, to find patterns across digitized historical corpora, has an ever growing body of literature. One of the best known papers that made us of n-gram detection within Google Books introduced the term "culturomics" to describe this type of work (MICHEL *et al.* 2011). For an overview of the potential of text mining, see UNSWORTH (2011), and for some recent experimental work, see CLEMENT (2012) and ODIJK *et al.* (2012).

[24] For more on the differences as well as the intersection between computational and corpus linguistics, see LÜDELING & ZELDES (2007).

[25] http://gallica.bnf.fr/?lang=EN.

[26] http://books.google.com.

[27] http://www.archive.org.

[28] "The World in 2011: ITC Facts and Figures", International Telecommunications Unions (ITU), Geneva, 2011 (http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf).

object. Documents include every major historical language, from Classical Chinese, Sanskrit, Cuneiform languages of the Near East such as Sumerian, Akkadian, Hittite, and Persian, every form of Egyptian from hieroglyphic through Coptic, Classical Arabic, and every language from Europe for which significant written traces survive.[29]

A great deal needs to be done for the coverage of every language. For Greek and Latin, the raw materials are, however, now available. Virtually every major source surviving from antiquity and an immense body of post-classical Latin is available as a scanned image book from some source. Some editions have been poorly scanned or scanned from damaged originals. And even if we have one version of every major source, the multi-text model assumes that we are able to view the textual history of a work as fully as possible – not just one critical edition but every version, including both critical editions and original sources on manuscript, papyrus or stone.

The mass digitization efforts have provided a foundation upon which library professionals can build. Many libraries can now digitize materials from their own holdings and thus many different institutions can add new content and replace problematic scans. The challenge here is to represent the logical contents of, rather than simply the physical form, of the digitized objects. The objects of interest are no longer simply the physical objects that preserve the textual record of the past.

The focus upon books as physical objects rather than upon their contents emerges quickly if one tries to study change over time using digitized books with the default library metadata. This metadata normally records only the date at which a physical book was published rather than including the date as well when the contents of that book were composed. Thus, we find that the vast majority of books catalogued as being in Latin from the Internet Archive list publication dates in the nineteenth century because most of those books were originally printed in that century. Analysis of a subset of 7,000 books that are in fact in Latin and that contain works that can be reasonably assigned single composition dates reveals the actual distribution, with the classical period providing a major, though, interestingly, not dominant, cluster. Interestingly, the nineteenth century remains the major period at which Latin books were

[29] For an overview of the extensive amount of digitized materials available in these various historical languages, see BABEU (2011).

composed – even in the nineteenth century, a great deal of Latin was being produced.[30]

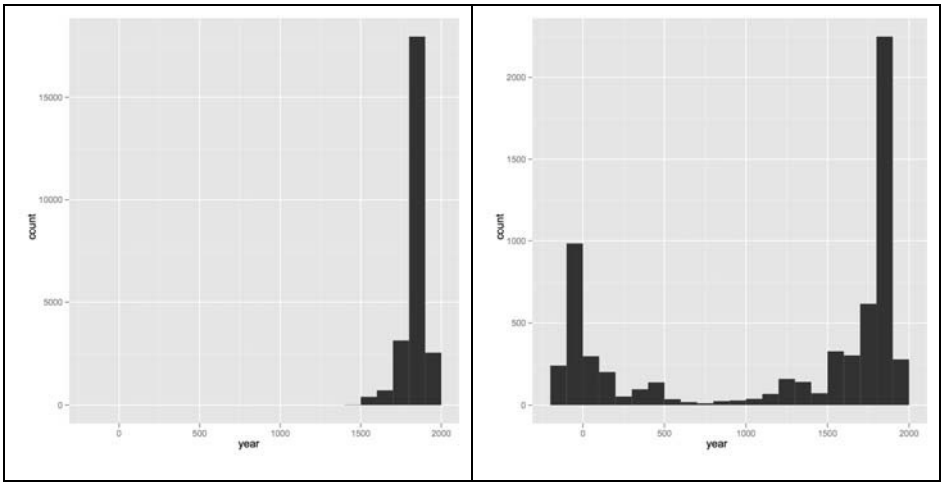


Figure 2: Left, 25,886 books downloaded from the Internet Archive that were catalogued as Latin, charted by publication date; right, analysis of 7,055 Latin books from the Internet Archive, charted by date of composition.

Even the reviewed date figures above provide only preliminary data. The spike of Latin produced in the first century BCE surely reflects not the absolute amount of Latin that survives from that period but the large number of editions for Cicero, Vergil, Horace and other authors from that period. By contrast, the large spike of nineteenth century materials will surely consist much more often of single editions and will thus contain an even larger collection of unique documents than the first century BCE spike. Latin was – and remained through the nineteenth century – a major language of publication within Europe, with many critical scientific, philosophical, and legal as well as literary texts produced in Latin. One could argue that the idea of Europe evolved most purely among those who chose Latin rather than their local language as a means of expression.

In October 2012, the 10,556,524 volumes digitized in the HathiTrust[31] (about ½ the 20 million that Google has already digitized) include 80,069 books identified as being in Latin and

[30] For more on the work that produced this data, see BAMMAN & SMITH (2012).

[31] http://www.hathitrust.org/home.

9,369 as being in Ancient Greek. Such estimates are only rough initial guides – substantial Greek and Latin will appear in books that are not catalogued as being in these languages. Nevertheless, these figures provide a first approximation for a lower bound of Greek and Latin that survive in printed form. An analysis of 9,000 Latin books downloaded from the Internet Archive shows that they include 385 million words. The HathiTrust thus probably contains close to 4 billion words of Greek and Latin. Each of these words is an object of interest that we need to be able to represent and each word can also be the target of an open-ended number of annotations representing an open-ended set of annotation types (e.g., links from a transcribed word to the corresponding section of a page, a link from a name to an encyclopedia entry, a morphological or syntactic analysis of a particular word).[32] By contrast, the *Thesaurus Linguae Graecae* (TLG)[33] contains approximately 100 million words of Ancient and Byzantine Greek. If we focus only upon Classical Greek and Latin (e.g., surviving documents produced through 600 CE – after Justinian and before the Prophet Mohammed), the total is roughly 60 million words of Greek and 40 million words of Latin. The HathiTrust of 2012 already contains about 40 times as many words of Greek and Latin. Of course, many of these books are restricted by copyright law but the counts of Ancient Greek and Latin books in the public domain[34] are 5,587 and 61,659 respectively – about 3 billion words.

catalogued	actual	precision	missed	total	recall
25,886	15,623	60.35%	6,790	22,413	69.71%

Table 2: Book level metadata provides an imperfect tool for locating books in Latin. Out of 1.2 million books downloaded from the Internet Archive, 25,886 were listed as being in Latin. Only 60% of these books were in fact primarily in Latin (many were editions of Greek with Latin introductions) while analysis of the language in 1.2 million book collection revealed 6,790 Latin books that were not catalogued as Latin.[35]

[32] For more on the need to design digital libraries that can deal with analyses at the level of trillions of individual words, see CRANE *et al.* (2012).

[33] http://www.tlg.uci.edu/.

[34] http://www.hathitrust.org/visualizations_languages.

[35] BAMMAN & SMITH (2012).

While it is useful to know that we have 3 billion words of public domain Greek and Latin, such a figure is only a very coarse measurement. Many of these 3 billion words will be represent different versions of the same text – canonical works will have been republished and quoted thousands of times. Each new publication, each excerpt in an anthology, and each quotation represent a decision made at a particular point, with its own context and background. In many instances, readers are not interested in a book but in a logical work such as the Homeric *Iliad* or the *Odes* of Horace. Such logical works often do not correspond to physical books – simple cases such as single volume editions of Dickens' *Oliver Twist* or Shakespeare's *Hamlet* are just one case and even these single volume editions are complex – a text of *Hamlet* will often include not only an introduction but also notes on the bottom of the page below the text.

2. Scalable, highly granular collections. Few researchers actually work with a million, much less ten million, digitized books. Massive collections contain many different potential corpora, each connected to many other corpora but each having its own center of gravity and its own communities. One challenge before us is to create dynamic relationships between smaller, subject-oriented curated collections such as emerged in the first generation of digital scholarship and the massive bodies of data from Gallica, Google and the Internet Archive.

The Perseus Digital Library provides one framework that can be generalized over the 90,000 or so books listed as being in Ancient Greek or Latin and the many citations of Greco-Roman culture scattered throughout millions more books. Perseus serves a number of purposes but its fundamental task is to provide a catalogue of logical documents – it is oriented not around the physical books but around their contents. This approach had evolved already when Perseus began in the 1980s, when CD ROMs had emerged as distribution media and the Internet as it is known today had not yet emerged.[36]

[36] CRANE, G., 2004: Classics and the Computer: An End of the History, in: *Companion to Digital Humanities*, 46-55, Malden Massachusetts. (http://www.digitalhumanities.org/companion).

Figure 3: Visualization of data relevant to chapter 1 of book 1 of Livy's History of Rome in the Perseus Digital Library.

The figure above visualizes results from a query that, in effect, says: "show me everything available about the first chapter of the second book of the *History of Rome* by Livy." The result includes materials of various kinds:

1) Three Latin editions of this particular chapter (with one of these editions the default display for this user). Note that none of the Latin editions contains the whole of Livy's history: two Latin editions come from volumes that contains books 1-10 of Livy, while the third comes from a volume that contains books 1-4. A normal catalogue cannot automatically determine which volumes contain editions of book 2 – or book 32 or 41 – of Livy.

2) Three English Translations of this particular chapter of Livy. Again, each of these translations comes from books that contain varying sections of Livy's work.

3) Three versions of an ancient summary of the first book of Livy's history, two in Latin and one English translation. For most of the works of Livy – and for the works of a number of other authors,

only ancient summaries survive. Summaries are thus an important document type that users need to track.

4) One commentary on this particular chapter. Commentaries are central resources for the study of historical sources. Canonical texts can have not only multiple commentaries composed during centuries of print scholarship but also commentaries preserved in complex formats in earlier manuscripts. These older commentaries are called scholia and a twelfth century CE manuscript can include material produced in Alexandria 1500 years before.[37] Commentaries follow the structure of the work that they explicate, often quoting particular phrases and passages.

5) Livy, like many Greek and Latin authors, has a detailed canonical citation scheme – much as a coordinate system allows people to describe particular regions of the earth, a canonical citation scheme allows scholars to identify particular regions of a text. The existence of these citations allows us to identify passages that mention the first chapter of the second book of Livy's History of Rome. Such references to this chapter of Livy appear (in the figure above) in commentaries on other parts of Livy, in a machine-readable index of Livy, in a reference grammar for Latin, and in an encyclopedia of daily life. Obviously, referenced to Livy will appear in every category of publication.

The structure underlying the figure above is based upon categories that are very old but the visualization depends upon the ability to analyze and manipulate chunks of text dynamically. The volume and page structures of print culture provide a framework out of which the deeper logical structures of logical documents must be extracted and then represented.

Perseus had developed the concept of abstract bibliographic objects (ABO)[38] to represent the distinction between a work, such as Livy's *History of Rome* and the various forms and derivations such as editions, translations, commentaries, and summaries. In the 1990s, the International Federation of Library Associations (IFLA) addressed a similar (though less complex) challenge with its Functional Requirements for Bibliographic Records (FRBR). The FRBR hierarchy provides a framework for organizing dozens--in some cases hundreds

---

[37] For a digital project working with the Scholia of the Homeric Epics, see www.homermultitext.org.

[38] For more on the concept of ABOs, see SMITH *et al.* (2001).

and thousands--of documents associated with canonical works. In the simplest case, FRBR identifies a *work* such as *Hamlet* or *Huckleberry Finn*. Different editions of *Hamlet*, such as those in the Riverside or the Norton Shakespeare, then constitute *expressions* of *Hamlet*. FRBR uses the concept of *manifestations* to distinguish between different physical forms that a particular manifestation can take. The traditional Riverside Shakespeare version of Hamlet, a Braille printing and an audio book constitute three distinct *manifestations* of the same expression. FRBR, in turn, uses the concept of item, to distinguish physical copies of the same manifestation. In traditional libraries, items are central--if the one copy of a book or CD ROM is out on loan or damaged or lost, then no one else can use it. In a digital environment, the *item* still can matter: the FRBR *item* allows us to distinguish the particular copy of a Greek edition of Demosthenes in which John Adams added notes from all other copies of that same edition.

The default FRBR model was originally designed as an entity-relationship model by a study group appointed by IFLA during the period 1991-1997, and was published in 1998.[39] This model was designed to manage print copies of items that frequently had multiple editions. Items become particularly complicated in a digital setting where we can, for example, have multiple scans of the same book, text generated from each scanned page by multiple OCR-engines, then multiple versions of a TEI (Text Encoding Initiative) XML[40] transcription derived from the OCR output (or simply typed in). A more recent effort, FRBRoo,[41] has emerged to provide a metadata standard that mapped the terms of museum documentation and bibliographic description.

For editions of Greek and Latin, Perseus has since 2007 been developing metadata inspired by the FRBR data model.[42] The goal was to develop an extensible bibliography with at least one edition of each Greek and Latin work surviving from antiquity. As an initial focus, the lists of works and editions used by the Lewis and Short Latin-English Lexicon (LS), the Liddell-Scott Jones Greek-English Lexicon (LSJ) and the Oxford Latin Dictionary (OLD) were used to create this initial bibliography. LS dates from the nineteenth century but it covers later Latin, while the more recent OLD focuses upon

---

39 For the full guidelines and model, see IFLA (1998).

40 http://www.tei-c.org.

41 http://www.cidoc-crm.org/frbr_inro.html.

42 For more on this work, see MIMNO *et al.* (2005) and BABEU (2008).

Latin authors through the second century CE. OLD still lists many of the editions that were current when it began work, most of which are now in the public domain. LSJ provides broad coverage for Classical authors, with selective coverage of later sources. Comparison with the TLG Canon – the extensive checklist of editions used by the *Thesaurus Linguae Graecae* – reveals some of the gaps. The largest eleven missing sources are all Christian sources: John Chrysostom (TLG# 2062), Cyril of Alexandria (TLG# 4090), Theodoretus of Cyrrha (TLG# 4089), the series of commentaries on the New Testament known as the Catenae (TLG# 4102), Gregory of Nyssa (TLG# 2017), Didymus the Blind (TLG# 2102), Athanasius (TLG# 2035), Basilius (TLG# 2040), the Ecumenical Councils (TLG# 5000), Epiphanius (TLG# 2021), and Gregory of Nazianzus (TLG# 2022) – a collection that contains more than 13 million words. LSJ documents the great shift of philology away from Christian Greek.

At present, the Perseus FRBR catalogue documents 5,055 Greek and Latin works. Works, at this point, can include not only such well-defined units as Plato's *Republic* or Vergil's *Aeneid*, but also fairly random groups (e.g., the four "epigrams" of Phaedimus that happen to appear in the Byzantine collection known as the *Greek Anthology*) and even phantom works that do not exist in their own right (e.g., the fragmentary quotations and allusions to a lost work or author). The FRBR catalogue represents, however, perhaps the first effort to create a framework by which to track multiple editions of both Greek and Latin authors that may be split among multiple printed volumes or be buried in large, heterogeneous collections such as the *Greek Anthology*.

Out of these 5,055 works, 3,262 have a record describing a particular edition. In 5,935 instances these records include the start and end page of a particular work in a particular printed edition. These records in turn contain 5,195 page level links to image books available in Google Books, the HathiTrust, or the Internet Archive so that users can go directly to a human-readable digitized copy of the books.

	links	works	image books
	0	210	0
	1	962	962
	2	2,037	4,074
	3	53	159
totals		3,262	5,195

Table 3: Image books associated with catalogued works.

This dataset lays the foundation for automatically extracting the sections of books that contain particular works. Ultimately such data will make it possible to feed pages containing particular Greek and Latin works to OCR software and then to use the OCR output to align the new edition with others already online. Rights restrictions still make it impossible often to download the high-resolution versions of the page images needed for best results from OCR software but the underlying data – works, start pages, end pages, and machine actionable links to digital copies – illustrates the necessary architecture for such a system.

Page numbers provide, of course, just a first step towards multitexts. Every word and every character on every surviving object is itself an object of interest. Our metadata must be able to track every word in every surviving version of a work. In addition, students of texts have regularly developed canonical citations schemes as coordinate systems by which to describe very precise chunks of the same text. The surface forms may vary (e.g., Thuc. 4.14 vs. Th. iv, 14) and in cases be ambiguous (e.g., is Th. iv, 14 the fourth *Idyll* of Theocritus or the fourth book of the history of Thucydides) but once properly decoded such citation strings define very precise chunks of text (e.g., chapter 14 of book 4 of Thucydides' *History of the Peloponnesian Wars* or line 14 of the fourth *Idyll* of Theocritus). The contents of these chunks will vary from edition to edition and multitexts need to be able to track those variations, allowing students to recognize, for example, that a particular instance of *fecerit* in one version of a text corresponds to *dixit* in another version. The Canonical Text Services (CTS) protocol[43], which builds upon the FRBR data model, provides a well-defined framework with which to express such relations.

[43] For further explanation of the CTS protocol, see SMITH (2009).

`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1-1.10`

The uniform resource name (URN) above describes a textual object within the Canonical Text Services Name Space. The basic elements above describe the following features:

greekLit: the work belongs to the category Greek literature.

`tlg0012`: This first field describes a **Text Group**, a category for traditional, convenient groupings of texts such as "authors" for literary works, or corpus collections for epigraphic or papyrological texts (e.g. "Homer," "Aristotle", "inscriptions from a given site"). The string `tlg0012` follows the numerical identifier used by the TLG to designate the Homeric epics.

`tlg001`: Within each TextGroup are **Works**, notional entities, each with a unique identifier within a TextGroup. Each work includes one or more titles (such as titles in different languages). The string `tlg001` follows the numeric identifier used by the TLG to designate the *Iliad*.

`perseus-grc1`: Works, in turn, may appear as **Expressions** which are specific versions of a notional work. Each has a unique identifier within the Work. Within the context of Greek and Latin, expressions are commonly **Editions, Translations, Indices, Commentaries**, author-specific **Lexica** (such as a Lexicon of Homer), and **Summaries**. The string `perseus-grc1` designates a particular Greek edition of the Homeric *Iliad*.

`1.1-1.10`: This designates a range within the canonical citation scheme for the particular work, in this case line 1 of book 1 of the *Iliad* through line 10 of book 1 of the *Iliad*. These URNs can provide the basis for precise and sustainable annotations across documents. Thus, for example, we often need to define the relationship between original source texts and modern language translations. If an English translation of the *Odyssey* begins "Tell me, O Muse, of the man of many devices" and we wish to express the assertion that "of many devices" corresponds to the Greek word *polutropon* in the Greek, we can use the following URNs.

```
urn:cts:greekLit:tlg0012.tlg002.perseus-
eng1:1#of[1]-devices[1]
```

The URN above describes a particular translation of the Odyssey (that of A. T. Murray published in Cambridge, MA, in 1919) and does not assume that this translation contains line numbers. It describes instead a string that begins at the first instance of the word "of" and ends with the first instance of "devices" in book 1 of this translation.

```
urn:cts:greekLit:tlg0012.tlg002:1.1#πολύτροπον[1]
```

The URN above defines the first instance of the Greek word πολύτροπον in line 1 of book 1 of the Odyssey. Like many, if not most, references mined from print sources, this URN does not define a particular edition but instead assumes that the text is sufficiently stable that we can resolve this reference across multiple editions. If the URN above exploits the full expressiveness of the CTS URN syntax, it can easily add a string such as `perseus-eng1` (a critical edition in Perseus) or `hmt-msA` (a particular manuscript of the *Iliad*) to resolve any ambiguities:

```
urn:cts:greekLit:tlg0012.tlg002.hmt-
msA:1.1#πολύτροπον[1]
```

The simplified URN above reflects the reality that most canonical citations are not linked to particular editions. The CTS URN syntax allows for graceful degradation for less precisely specified citations.

The examples above do not address every case in a digital space: we will immediately have multiple OCR-generated transcriptions of different scans of the various physical copies of the same page from a print edition, each of which contains errors. In other cases, different editors will transcribe the same word or abbreviation in a manuscript, papyrus or inscription differently and then occasionally change their minds. We thus need additional specificity, including time-stamps.

Ultimately, accessing the URNs above will yield a digital text, an electronic version of an Edition, Translation, or one of their Exemplars, which will contain one **Online** element. This element contains information about the citation scheme as well as information the server could use to translate the abstract reference into terms needed for local retrieval, such as a filename or database

lookup. Nevertheless, the CTS syntax above provides a precise foundation upon which to build.

In a mature digital space, where we need to align multiple versions of the same work, individual TEI XML transcriptions play a different but important role. In the first generation of digital corpora, researchers depended upon having access to a single, reasonable edition of each work represented in a documented format (ideally, TEI XML). In a multitext space, the transcription becomes a framework around which to cluster and to organize many other editions. Thus, if we can associate a line such as

<l>Arma virumque cano, Troiae qui primus ab oris</l>

with a URN such as `cts:latinLit:phi0690.phi003.perseus-lat1:1.1`,[44] we then can find a very large number of other passages that belong to editions of Vergil's *Aeneid*, or that quote all or part of the above line. Where other versions differ from the base text, we can represent those differences in well-established forms for edit operations (e.g., substitute string X with string Y or insert string Y after string X etc). Once we have one edition of a work encoded with a canonical citation scheme, we can align many others, even when other transcriptions consist of noisy OCR-generated text, and allow users to compare different versions. The TEI XML transcription becomes, in a multitext world, an entry point into a network of different versions. A transcription such as that listed above constitutes both data in its own right and metadata (i.e., data to find related data).

Many Greek and Latin sources exist in digital form but do not support digital scholarship because they are in idiosyncratic formats (such as the page layout description language, developed in the 1970s, in which many Greek and Latin texts are stored), have restrictive front-ends that prevent downloading, and include licensing, enforced with threats of legal action, that prevents the re-use, repurposing and redistribution which are central to digital scholarship. At times, sources are restricted because of all of these reasons.[45]

---

[44] PHI stands for Packard Humanities Institute which published a collection of Classical Latin Texts and assigned identification numbers to authors and works. Here phi0690 designates Vergil and phi003 the *Aeneid*: http://latin.packhum.org/.

[45] CAYLESS (2010) has made a strong case for the role of re-use in long-term digital preservation, whereas a panel at the Digital Humanities in 2009 explored the

Approximately 20 million words of Greek and Latin – roughly 20% of the classical corpus – are either already available, or have been entered and are being formatted, in TEI XML with Creative Commons open licenses.[46]

Greek and Latin editions		
versions	TEI XML transcriptions	total
1	970	970
2	22	44
3	3	9
Subtotal	995	1,023
English Translations		
1	539	539
2	96	192
3	2	6
Subtotal	637	737
Total	1,632	1,760

Table 4: TEI XML transcriptions in the Perseus Digital Library representing original language editions and English translations of Greek and Latin sources.

The Perseus Digital Library currently has 995 distinct Greek and Latin sources in TEI XML, along with English translations for 637 of these works. The collections in Perseus provide breadth but the handful of instances where more than one edition and translation are available have provided an opportunity to develop and demonstrate initial methods by which to manage multiple versions of the same work.

We can represent trillions of relationships between billions of words digitally but we cannot transcribe, much less annotate, 4

difficulties of reusing even open-source objects within digital classics (BODARD 2009).

[46] The major sources for on-line TEI XML transcriptions of Greek and Latin are the Perseus Digital Library (http://www.perseus.tufts.edu/hopper/opensource/download) and http://www.papyri.info/. A Mellon-funded Project centered at Harvard has entered, and is now formatting, several million words of Greek scientific and medical texts.

billion words of Greek and Latin. We must depend upon automated methods if we are to organize even such a modest collection as the surviving body of Greek and Latin (which account for less than 1.5% of the digitized books in the HathiTrust). Many of these 4 billion words will be different versions of the same work – but book level metadata alone would not allow us to determine how many versions of book 4 of Vergil's *Aeneid* or of Sophocles' *Oedipus the King* are within this massive collection: one volume may contain three plays of Sophocles, one play, or all seven remaining plays, while many edited documents are quite short and appear as sections in larger publications. And each version of a document is a historical event in its own right – the school anthology may, for example, draw upon a standard edition but the fact that it drew upon a particular edition and the selections that it drew shed light upon intellectual and educational practices of the time. There is no good way to determine how many unique words of Latin from how many works are within this vast space without analyzing the texts themselves.

If we are to manage the vast body of materials already available to us we need a two-fold transformation of scholarship. We obviously need to draw upon automated methods of every kind relevant to the analysis of textual data in many multiple languages. But automated methods are not enough – there is just too much work to be done and too many instances where human input is necessary. Even if all library professionals and advanced researchers shifted their focus away from book-level metadata creation and specialist publications and towards the myriad tasks by which to make these billions of words ever more intellectually accessible to an ever widening set of humanity, the labor available would still not be enough. Professional students of Greek and Latin must welcome student researchers and citizen scholars as collaborators – in the United States, the 3200 or so members of the American Philological Association (APA)[47] must, in other words, turn not only to the 55,000 students of Greek and Latin in postsecondary education but also to the almost 150,000 secondary

[47] This figure is based upon the statement at http://apaclassics.org/index.php/about the APA/director report/executive director report for 2011/ that 800 represents 27% of the individual members of the American Philological Association. This figure, which includes some who are not professional classicists and others who are not from the United States, serves as a rough estimate for the number of professional Classicists in the United States.

school students studying Latin.[48] Such a shift in the relationships between teacher and student and between learning and research would presumably have an effect upon the students who enroll in Greek and Latin and, inevitably, the number of jobs for those teaching them.

The explosion of digital access to Greek and Latin has transformed the relationship between those languages and society. At the least, a global public could view a range of Greek and Latin sources which were previously only available in research libraries. This physical access challenges students of Greek and Latin to provide the intellectual access needed to understand these sources. That challenge in turn provides the most inward looking specialist with a material reason to look outwards and to engage a wider audience. We cannot pursue our research fully without a new collaborative, laboratory culture. Every aspect of digital editing depends upon not only new automated methods but also new, more broadly based forms of collaboration.

3. OCR for historical languages: Human beings can read images of writing – indeed, high resolution, multispectral and 3D scans of text-bearing objects can make some surfaces more readable than the original objects were to the naked eye.[49] But we cannot transcribe billions of words of Greek and Latin. OCR works well for modern printed Latin texts if the OCR system knows that it is analyzing Latin and if it has access to a Latin dictionary/word list so that it does not try to turn Latin into some other language (e.g., Latin t-u-m, "then," can become English t-u-r-n if the OCR system expects English). But commercial OCR performs much less well for earlier printed books in Latin and indeed in any language. Substantial work remains to be done if we are to extract high quality text from these earlier printed sources.[50]

---

[48] The figure of 150,000 is a rough approximation based upon the 148,000 students who registered for the 2012 National Latin Exam: http://www.nle.org/pdf/ExamResults2012.pdf.

[49] For example, using such technologies has provided unprecedented access to the Archimedes Palimpsest (http://www.archimedespalimpsest.org), see SALERNO (2007).

[50] While still a relatively specialized area, the development of OCR tools (both the modification of commercial tools and the adaptation of open source systems) for historical languages has grown dramatically in the last five years. See for example, the results of the recently concluded Improving Access to Text (IMPACT) project (http://www.impact-project.eu) as well as the newly funded Early Modern OCR Project (http://emop.tamu.edu/). For a review of the state-of-the-art in this area, see PIOTROWSKI (2012).

At the same time, while OCR may need to be optimized for recently published Latin, Classicists have never had access to reasonable OCR-generated text for Ancient Greek. For the forty years since the TLG was founded in 1972, they have had to depend upon manual keyboarding – a labor intensive, inherently expensive process. It has not been possible to image working with thousands of books printed in Ancient Greek. That situation changed when Gordon Stewart published the first paper documenting the effective use of OCR for Classical Greek.[51] He demonstrated that in 2007 a modern Greek OCR system (Anagnostis), trained to ignore the accents in Classical Greek, could generate transcriptions of the alphabetic characters in 19th and twentieth century Greek editions. Because this OCR method also included textual variants and because these variants account for between 8 and 15% of the words on a given page, OCR generated text for editions immediately provides better recall than error-free transcriptions that only include the reconstructed text.

In the subsequent five years, Federico Boschetti and Bruce Robertson carried this work further.[52] Commercial OCR systems had serious limitations: they could not be trained to recognize Classical Greek directly or they could not run on large bodies of text or their licensing systems were not designed to support multi-processor systems. Boschetti and Robertson undertook to train open source OCR systems to recognize Classical Greek and to develop the error checking methods needed to correct the output.

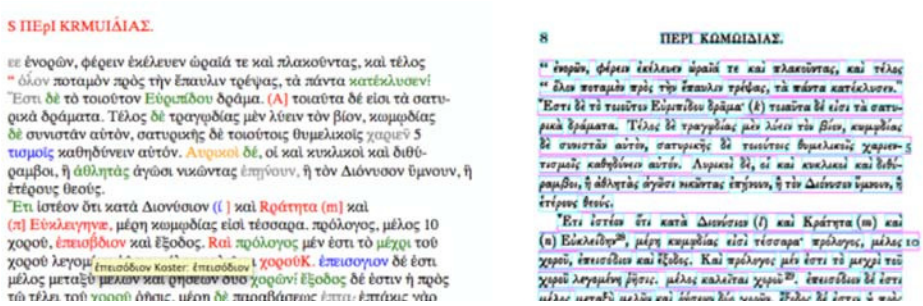


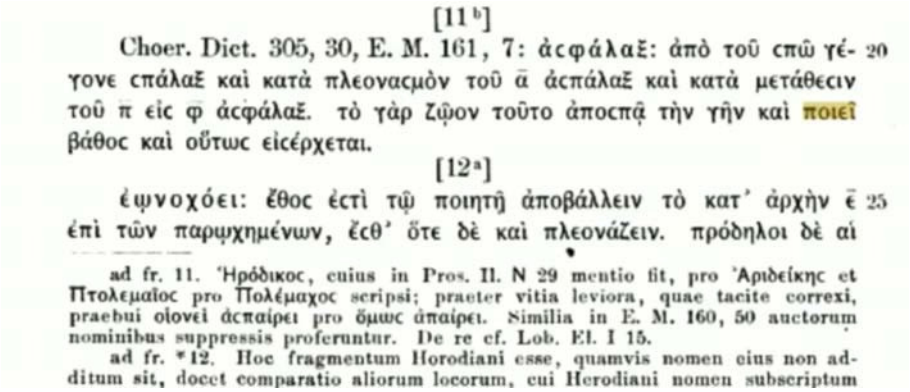
Figure 4: Error identification in Greek OCR developed by Federico Boschetti. Color indicates classes of error. The HOCR format above includes (1) suggestions for corrections based upon standard spell-checking strategies; (2) suggestions based upon words as they appear in another edition on-line (near ground truth).

[51] STEWART *et al.* (2007).

[52] For more on this work, see BOSCHETTI *et al.* (2009) and ALMAS *et al.* (2011).

This multi-text approach to digital editions creates editions that are, in effect, self-correcting as they include OCR-generated text from multiple print editions, even where these individual transcriptions contain substantial error rates. Suppose OCR for two different editions of a text (perhaps one a Teubner and one a Loeb) generates an error in every 10th word. If the errors are randomly distributed, then the probability that one or the other OCR-generated text contains a valid reading rises to 99%. If we add a third edition under the same conditions, the probability that we will have at least one correct transcription rises to 99.9% and so on. Of course, different editions will have different forms up to 5 or 10% of the time but as more editions become available, the probability that the same reading will be correct somewhere will rise. Errors will remain but the nature of the discussion has now shifted from never having variants to doing a better job of capturing a growing body of variants.

Once we align OCR-generated text not only with the page images from which it was derived but also with other editions of the same text, we can create image-front searching long familiar to academics from JSTOR[53]. We search for Greek and Latin and fault tolerant searching locates probable hits and displays the results either as text or as clips from the image of the printed page.


[11 b]

Choer. Dict. 305, 30, E. M. 161, 7: ἀcφάλαξ: ἀπὸ τοῦ cπῶ γέγονε cπάλαξ καὶ κατὰ πλεοναcμὸν τοῦ ᾱ ἀcπάλαξ καὶ κατὰ μετάθεcιν τοῦ π εἰc φ ἀcφάλαξ. τὸ γὰρ ζῷον τοῦτο ἀποcπᾷ τὴν γῆν καὶ ποιεῖ βάθος καὶ οὕτωc εἰcέρχεται.

[12 a]

ἐψνοχόει: ἔθοc ἐcτὶ τῷ ποιητῇ ἀποβάλλειν τὸ κατ' ἀρχὴν ἓ ἐπὶ τῶν παρῳχημένων, ἔcθ' ὅτε δὲ καὶ πλεονάζειν. πρόδηλοι δὲ αἱ

ad fr. 11. Ἡρόδικος, cuius in Pros. Il. N 29 mentio fit, pro Ἀριδείκης et Πτολεμαῖος pro Πολέμαχος scripsi; praeter vitia leviora, quae tacite correxi, praebui οἱονεὶ ἀcπαίρει pro ὅμωc ἁπαίρει. Similia in E. M. 160, 50 auctorum nominibus suppressis proferuntur. De re cf. Lob. El. I 15.

ad fr. *12. Hoc fragmentum Herodiani esse, quamvis nomen eius non additum sit, docet comparatio aliorum locorum, cui Herodiani nomen subscriptum

Figure 5: Image-front, morphologically-aware searching of OCR-generated Greek text. (Bruce Robertson, demo of the Squeegee search prototype, developed as part of a Digging into Data Phase 1 Project).[54]

[53] http://jstor.org.

[54] http://heml.mta.ca/RobertsonGreekOCR/.

The major challenge at this point is to develop the workflow that will feed scanned editions of Greek and Latin to the appropriate OCR software and then allow members of the community to correct the output as they see necessary. This becomes now a question of software development and of the diplomacy needed to make high-resolution scans of public domain Greek and Latin editions available.

4. Transcription and structural markup: More than 25 years ago, the TEI began to develop shared conventions for representing texts in digital form. A major goal of the TEI was to enable semantic markup – rather than labeling a string in italics and then letting the reader determine if the string were in italics because it was the title of a book, because the author wanted to emphasize the text, because the text was in a foreign language, or because of some other reason, the TEI offered conventions to express these deeper purposes. Formatting software could then convert titles and German quotations into italics for printing, while the text preserved these distinctions in a machine-actionable form. The TEI published its fifth edition of Guidelines (TEI P5) in 2007. Off-the-shelf commercial XML editors such as Oxygen[55] exist that support editing TEI XML. Workshops regularly introduce neophytes to the basic (and not, in the end, so terribly challenging) basics of TEI XML.[56] An individual or small working group can now create individual TEI XML transcriptions of texts in Greek, Latin, and many other languages.

The problem now is one of scale. In fall 2012, roughly 35,000 individual users each month work with more than 17 million words of Greek and Latin texts in Perseus. How can we enable any of these users to correct residual data entry errors in, or add additional TEI XML markup, within this corpus as a whole? What happens as the amount of OCR-generated Greek and Latin text ready for editing increases to billions of words and the audience of potential contributors expands beyond the largely English-language users of Perseus?

There are two approaches to this problem. In the simplest case, texts are uploaded to Wikisource[57] and the Wiki community makes corrections as they choose.[58] The Wiki formatting language is not as

---

55 http://oxygenxml.com.

56 See for example the resources offered by the Women Writers Project at Brown University (http://www.wwp.brown.edu/outreach/resources.html).

57 http://wikisource.org/.

58 The potential of collaborative transcription and the creation of TEI-XML documents has been investigated by the Transcribe Bentham project, see CAUSER *et al.* (2012). There are also a number of tools other than WikiSource that have

expressive as TEI XML but it can capture the basic page layout and some fundamental semantic concepts. Texts corrected in a Wikisource space provide an excellent starting point for more elaborate TEI markup. And, with a little work, most corrections to a Wikisource version of a text could, in most cases, be automatically integrated into a parallel TEI XML transcription. In this model, the Wiki infrastructure provides the framework for basic text correction.

Another approach focuses upon the challenge of precisely representing many different changes to a collection, some involving isolated changes to particular documents, others covering thousands of passages. In the Wikipedia model, corrections converge on a single canonical transcription of a master print source. Scholarly editing will, however, produce many different versions of the same text and the editorial workflows quickly diverge as different groups potentially create their own version of the same text. To address this case, papyrologists, funded by the Mellon Foundation, developed a more complex workflow, the *Son of Suda Online*. (SoSOL).[59]

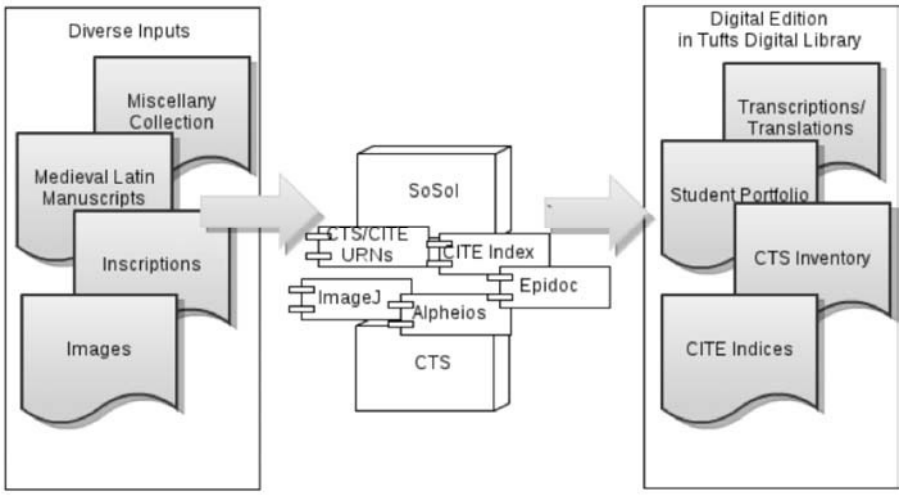


Figure 6: SoSOL as it is being adapted to work with materials in the Perseus Digital Library.

been created to aid in the creation of collaborative manuscript and or text transcriptions, including Scripto (http://scripto.org) and T-Pen (http://t-pen.org/TPEN/).

[59] http://idp.atlantides.org/trac/idp/wiki/. For more details, see SOSIN (2010).

Much work remains, however, to make SoSOL scale up beyond dozens of papyrologists to thousands of contributors working with Greek and Latin in general. Nevertheless, SoSOL can track a large number of very precise editorial events and it constitutes a fundamental step in the direction of scalability.

5. Automatic cataloguing, including language and text reuse detection: Once we have a collection of OCR-generated texts, we can begin to look for instances where one text re-uses another. Book level metadata provides, of course, only a very coarse guide. Books that are primarily in Latin or Ancient Greek can contain distinct documents from different periods (e.g., the Byzantine collection of Greek poetry known as the *Greek Anthology*) and genres (e.g., inscriptions from the same site and covering many genres are customarily published together). Documents also quote each other: Porphyry quotes Plato but Plato also quotes Homer. The self-standing edition and the text that draws upon an earlier text represent two ends of a continuum that we need to track if we are to understand the history of a text.

The Proteus Project,[60] developed with support from the National Science Foundation (NSF)[61] by researchers at the University of Massachusetts, had addressed the problem of identifying duplicate versions of the same work in collections that are large (greater than 1 million books) and that can, in depending upon OCR-generated text contain numerous errors.

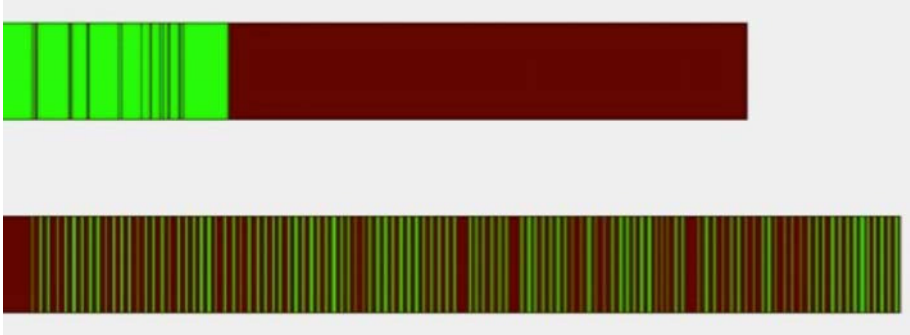


Figure 7: Text alignment is also used for finding groups of texts whose structure corresponds in other ways, such as works published in different languages, or texts and their commentaries. Here, for instance, we see an automatically

[60] http://books.cs.umass.edu/beta-sprint/.

[61] http://www.nsf.gov.

generated alignment between the Latin text of Vergil's Aeneid *and a commentary. The first bar depicts the first eight books of the* Aeneid*. The green in this first bar indicates the aligned portions, from which we can tell that the commentary only deals with the first three books of the* Aeneid*. The second bar depicts the commentary. Its green portions are brief passages from the text of the* Aeneid*, and the intervening red bars are the commentary, which does not align.*[62]

At the other extreme, one text quotes or paraphrases small sections of another (e.g., Plato quoting Homer). In this case, at least three issues complicate the process. First, it is not always clear when one text is directly citing another – we generally need to know the composition dates of various documents so that we can automatically determine which document cites the other. Second, text reuse can include short phrases (e.g., "to be or not to be") and it may not be clear whether the phrase represents an intentional allusion to a particular text (e.g., to *Hamlet*) or has simply become an idiom with no widely recognized single origin. Third, one text may paraphrase, rather than directly quote, another, thus making it hard to detect the textual reuse by searching for repeated strings.

The UMASS Proteus system has also explored methods to detect and to visualize text reuse in large collections. The Proteus visualization of documents that quote *Hamlet* maps one text onto a restricted number of quoting documents. This visualization allows readers to compare a single text with a finite number of documents that quote it.[63]

---

[62] Text drawn from: http://books.cs.umass.edu/beta-sprint/Demonstration/Entries/2011/8/3_Aligning_the_Aeneid_and_commentary.html.

[63] For further discussion of this work, see SMITH *et al.* (2011).

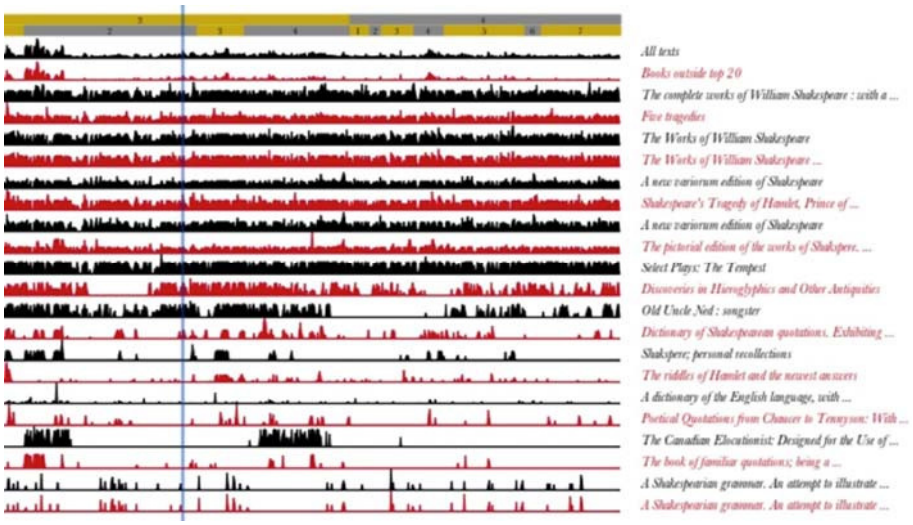


Figure 8: At the top are listed the acts and scenes of the play. Below are histograms showing the amount of textual overlap between each line and various other books. Five Tragedies, for instance, contains the complete text of Hamlet and thus overlaps completely. But we can also see other genres such as a Dictionary of Shakespeare, which uses quotes to illustrate word definitions, or The Canadian Elocutionist, which excerpts speeches for practice by aspiring public speakers, or The riddles of Hamlet *and the newest answers, which is a work of literary criticism*[64]*. The text reuse patterns are represented using the Highbrow visualization tool.*[65]

The eAqua[66] and subsequent eTraces[67] projects, located at the University of Leipzig and funded by the German Federal Ministry of Education[68] also explored the problem of detecting text reuse within a corpus. The first visualization illustrates how subsequent students of Plato used the author's *Timaeus*. The visualization illustrates how this work grew dramatically in importance as Neo-Platonism replaced Middle Platonism. It also shows which passages the Middle and Neo-Platonists most often cited (thus showing a shift in interest within the work). In addition, the visualizations show which authors most often cited this work.

---

64 Text drawn from: http://books.cs.umass.edu/beta-sprint/Demonstration/Entries/2011/8/2_Quotation_detection%3A_Hamlet.html.

65 http://osc.hul.harvard.edu/highbrow/.

66 http://www.eaqua.net/index.php.

67 http://etraces.e-humanities.net/.

68 http://www.bmbf.de/.

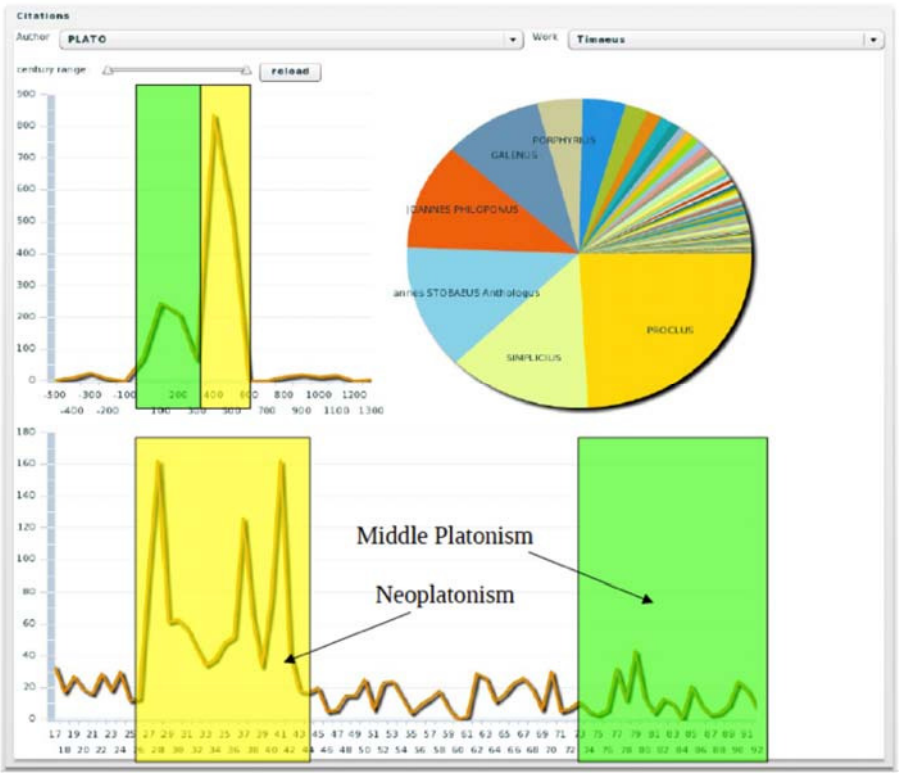


Figure 9: The Leipzig-based eAqua project explored the relationship between different texts. Here we see the frequency with which later authors cite portions of Plato's Timaeus*. In the above left, the green and yellow boxes distinguish quotations by Middle and Neo-platonists, demonstrating the surge of interest in the* Timaeus *among the neo-Platonists. The pie chart on the upper right hand illustrates which authors most frequently cite the* Timaeus*. The graph below shows which sections of the* Timaeus *are most frequently cited. The yellow and green boxes illustrate the sections of greatest interest to the Middle and Neo-Platonists.*

The eAqua and eTraces projects[69] also developed "heat maps" to track which sections of an author's work are most frequently quoted in subsequent Greek literature and thus to see as well which authors are more frequently quoted than others. The heat maps below illustrate the quotation frequency of passages in the surviving works of Xenophon, Plato, Aristotle, and Plutarch. Not surprisingly, Plato

[69] For some related publications regarding the work of both projects, see for eAqua (BÜCHLER *et al.* 2010) and for eTraces (BÜCHLER *et al.* 2012).

and Aristotle are much more heavily quoted than either Xenophon or Plutarch. The heat map for Plato shows a particularly striking pattern of black (i.e., rarely if ever quoted) passages among much more heavily cited passages. The heat map captures a one-to-many relationship (e.g., how often one text is cited in a collection of open ended size). The heat maps above can provide summary views of all subsequent citations while the UMASS visualization shows relationships with specific texts.

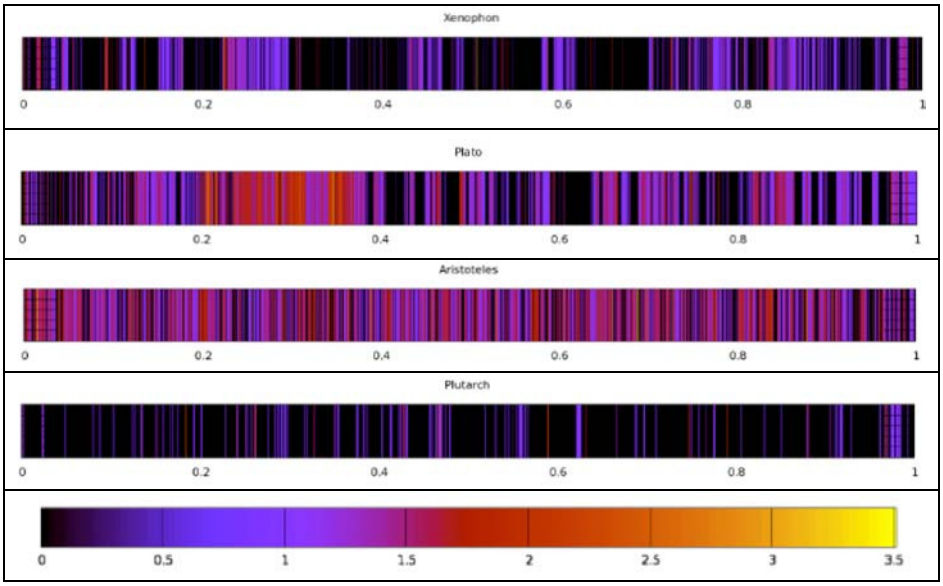


Figure 10: The heat maps above reflect the frequency with which sections of an author's surviving work have been quoted. Sections of the work that have not been quoted appear as black. The more frequently the section has been quoted, the brighter the color, with yellow indicating passages quoted more than three times by other authors.

Translation is a special case of text reuse: a translator takes words in one language and represents them, more or less closely, in another. Automated methods can detect in most cases which words in a source language correspond to their equivalents in a translation – assuming there are enough parallel texts so that the system can learn which words in one language correspond with words in another.[70]

[70] The use of parallel texts for translation alignment has also proven useful as one step in finding translations within massive digitized collections of books (YALNIZ

The Alpheios project[71] has provided tools whereby human editors can refine the results of this machine alignment of source text and translation. The figure below shows a human edited alignment of Greek and English words in the opening of the Homeric *Odyssey*. The textual data is here visualized as a traditional interlinear translation (such as were developed when Greek and Latin were staples of education and many students had to struggle through a few canonical texts).

Text reuse becomes an object of scholarly concern in particular when the quoted source does not itself survive and the quotation is not necessarily verbatim. Thus in the following passage, a speaker in Athenaeus' *Banquet of the Wise Men* quotes an earlier source.

> Ἴστρος δ᾽ ἐν τοῖς Ἀττικοῖς οὐδ᾽ ἐξάγεσθαί φησι τῆς Ἀττικῆς τὰς ἀπ᾽ αὐτῶν γινομένας ἰσχάδας, ἵνα μόνοι ἀπολαύοιεν οἱ κατοικοῦντες: καὶ ἐπεὶ πολλοὶ ἐνεφανίζοντο διακλέπτοντες, οἱ τούτους μηνύοντες τοῖς δικασταῖς ἐκλήθησαν τότε πρῶτον συκοφάνται.
>
> "And Istrus, in his Attics, says that it was forbidden to export out of Attica the figs which grew in that country, in order that the inhabitants might have the exclusive enjoyment of them. And as many people were detected in sending them away surreptitiously, those who laid informations against them before the judges were then first called sycophants." (tr. C. D. Yonge)

Scholars have tried to reconstruct from such fragmentary pieces lost works of Greek and Latin – most of the works of which we know only survive insofar as they are quoted, paraphrased or mentioned.[72] In the passage above, we need to decide what words we believe come from Istros and what words were produced by Athenaeus. We need to mark "says" as the so-called *verbum dicendi* (the word of speaking) so that we can compare it with other similar words (e.g., "asserts", "claims", "reports") and so that we can detect the ways in which one author describes their use of sources. Ultimately we move from automated services that detect textual reuse to close scholarly analysis.

While we may wish to use textual alignment to identify multiple editions and quotations of a work, methods also exist by which to identify translations and then to align many of the words in the

---

& MANMATHA 2012), as well as for markup projection between text in Greek and Latin and modern language translations (BAMMAN *et al.* 2010).

71 http://alpheios.net.

72 For some preliminary work on the encoding of fragmentary works within digital editions and libraries, please see BERTI *et al.* (2009).

original text to their equivalents in the translation. Such parallel texts are fundamental to many, if not most, multilingual services now in use – statistical methods are used to determine automatically which words co-occur. Such parallel texts also enable new lexicographic and semantic tools that grow more and more useful as collections grow larger and purely manual techniques become less feasible.[73]

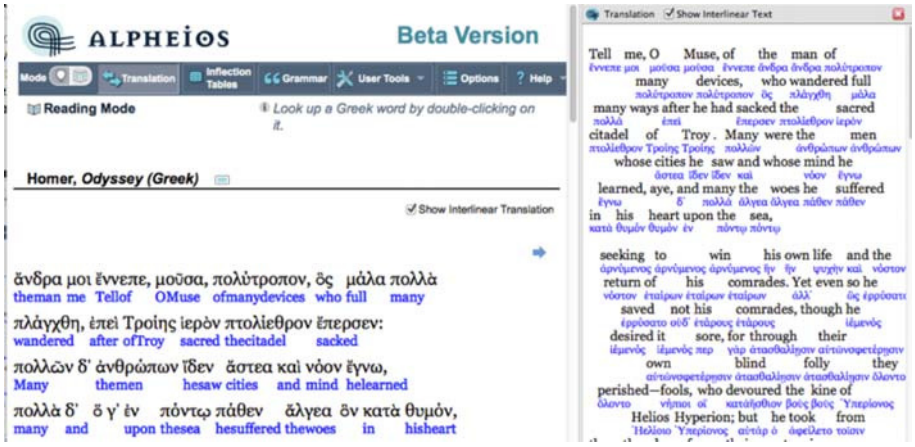


Figure 11: Visualization of Greek and English words aligned to one another in the Alpheios parallel text browser.

The figure above visualizes a Greek text of the Odyssey aligned to an English translation and the corresponding English translation as it is aligned to the Greek. The alignments above were first generated automatically and were then edited.

With the source text/translation alignment, however, we also enter into the world of reading support. The more precisely a source text and its corresponding translation correspond, the more support readers have in picking apart the granular form of a source text in a language that they may have never studied. With aligned source texts and translations we begin to provide a fundamental instrument for global editions that must serve many different linguistic and cultural audiences. The links from Greek to English above connect the Homer text to vast and growing resources being developed to make English (or any other major language) available to a global net audience.

---

[73] For example, see work on the Dynamic Lexicon (BAMMAN & CRANE 2008).

6. Annotation of named entities and morpho-syntactic features: Digital editions should also include machine actionable annotations on various features relevant to their readers. The identification of people, places, ethnic groups and other named entities essentially extends the print practice of adding indices of people and places.[74] Machine actionable annotations for the morpho-syntactic analysis of each word have ancient intellectual roots in pedagogical practice – students have been asked for thousands of years to state which word a given noun or preposition depends upon in a sentence.

Support from the National Endowment for the Humanities (NEH)[75] and the Institute for Museum and Library Services (IMLS)[76] allowed Perseus to develop named entity classification services for Greek and Latin. In the following pasage of Greek text, the names Plato, Menelaos, Homer, Patroklos, and Hector are all classified as being the names of people.

οὐ δεόντως γοῦν <name type="person">Πλάτων</name> τὸν <name type="person">Μενέλεων</name> ἐνόμισεν δειλόν, ὃν ἀρηίφιλον <name type="person">Ὅμηρος</name> λέγει καὶ μόνον ὑπὲρ <name type="person">Πατρόκλου</name> ἀριστεύσαντα καὶ τῷ <name type="person">Ἕκτορι</name> πρὸ πάντων πρόθυμον μονομαχεῖν

Semantic classification by itself is useful, but for many purposes we want to be able to assert that the Plato in a particular passage does indeed describe the famous Greek philosopher rather than the comic playwright of the same name. In some cases, this information can be mined from digitized print indices[77] (although it is not always easy to determine automatically that Alexander-5 in one index is Alexander-3 in the index for another author). In some cases the precise identity of the Antigonus or Alexandria in a given passage is not clear and is the object of scholarly analysis.

---

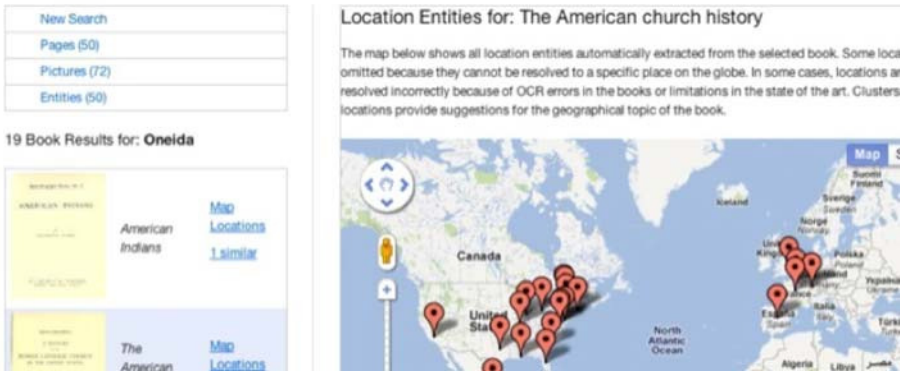
[74] The importance of supporting both the automatic annotation of various named entities within diverse types of historical texts as well as the creation of tools to support users in identifying and annotating such entities has received a great deal of attention in the last few years. For some recent work in these various areas, ZHANG *et al.* (2010), CLOUGH *et al.* (2009), and TOBIN *et al.* (2008).

[75] http://www.neh.gov.

[76] http://www.imls.gov.

[77] For some interesting work on the mining of digitized print indices from historical books for personal and place name identification see PIOTROWSKI (2010) and ROMANELLO *et al.* (2009).

Having the identity of the particular people and places, for example, enables new classes of analysis and visualization. We can, for example, begin to build on machine actionable social network data to trace members of a family or group.[78] A great deal of work has gone into the automatic idenification of places[79] (an inherently easier problem because there are fewer places than people and places do not have children and grand-children nearly so often as do people). The UMASS group has included named entity identification in its architecture. The figure below illustrates frequently mentioned places in a book on church history.



For students of historical languages, richly annotated corpora may be the most important new phenomena from the shift to a digital space. Editors have long included punctuation, capitalization, paragraph breaks and other print annotations based upon their own analysis of the text in order to support contemporary readers. The field of corpus linguistics has developed methods by which to systematically record the linguistic features in a text. An annotated corpus can be queried and its features retrieved and quantified for analysis.

[78] The exact identification of historic individuals is one of the tasks of prosopography and there is growing work in the field of "digital prosopography" with social networks and visualization tools, see for example the project Berkeley Prosopography Services (http://code.google.com/p/berkeley-prosopography-services/) described in SCHMITZ 2009, and also interesting work by GRAHAM & RUFFINI (2007).

[79] Relevant work in place name recognition, particularly in terms of historical language resources and the field of classics, has been reported by the Googling Ancient Places project, see ISAKSEN *et al.* (2012), as well as by the HESTIA project, which has made use of Perseus TEI-XML texts as part of its work, see BARKER *et al.* (2010).

Grammars can then be constructed directly from the full corpus, with explicit statements about the frequency of particular phenomena and links directly back to the

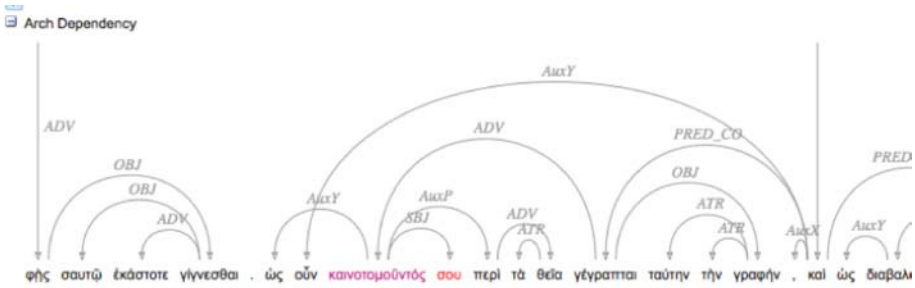
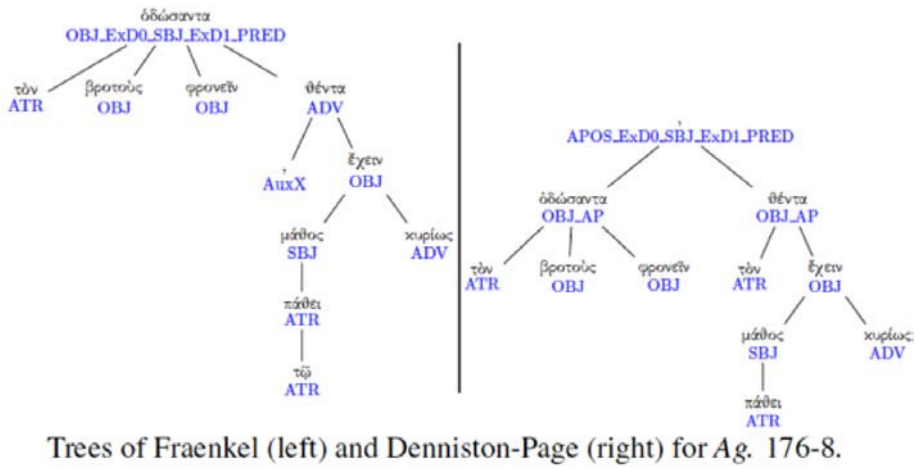


Figure 12: A genitive absolute retrieved from the Euthyprho *of Plato, morpho-syntactically annotated by Giuseppe Celano.*

Richly annotated corpora with systematic morphological and syntactic analyses are often called treebanks because the syntactic structures can be visualized as trees.

Linguists often (in practice) focus upon developing the largest possible corpora because they are looking for typical (and thus repeated) phenomena. More data is, in this case, better data because quantification and statistical significance are fundamental to evidence-driven linguistic research. Philologists focusing intensely on particular texts are often more concerned with exploring multiple ways to construe a particular sentence or phrase. In this case, the goal is not to provide a single plausible interpretation of each sentence but to represent variant interpretations. In the example below, two competing interpretations for one sentence in Aeschylus have been encoded in a dependency grammar. The two hypothetical readings can then be compared to the other sentences in Aeschylus, Greek tragedy or larger corpora as these become available.



Trees of Fraenkel (left) and Denniston-Page (right) for *Ag.* 176-8.

*Figure 13: Interpretations of the same sentence in Aeschylus as proposed by two twentieth-century editors and represented in machine actionable form by Francesco Mambrini (*BAMMAN *et al. 2009).*

Morpho-syntactic analyses are, however, fundamental to global editions because they reveal the underlying structure of a sentence in a general format. Readers with the morpho-syntactic analysis of a sentence and an aligned translation into a language with which they are familiar have the tools with which to pull apart every word in a source of interest to them. The 350,000 morpho-syntactically analyzed Greek and Latin words available in the Perseus Greek and Latin Treebanks provide support for readers regardless of whether their primary language is English, German, Arabic or Chinese.[80] Those who understand English can combine the treebanks with aligned English translations and can begin to work with Greek and Latin directly even before they have begun systematic study of those languages.

Curated treebanks are not only useful for precise study and analysis with methods from corpus linguistics; these curated treebanks also provide data from which automated systems can learn to perform morphological and syntactic analysis. In general, the more morphological and syntactic training data available that is relevant to a given corpus, the more accurate the automatic analyses will be.

[80] http://nlp.perseus.tufts.edu/syntax/treebank/.

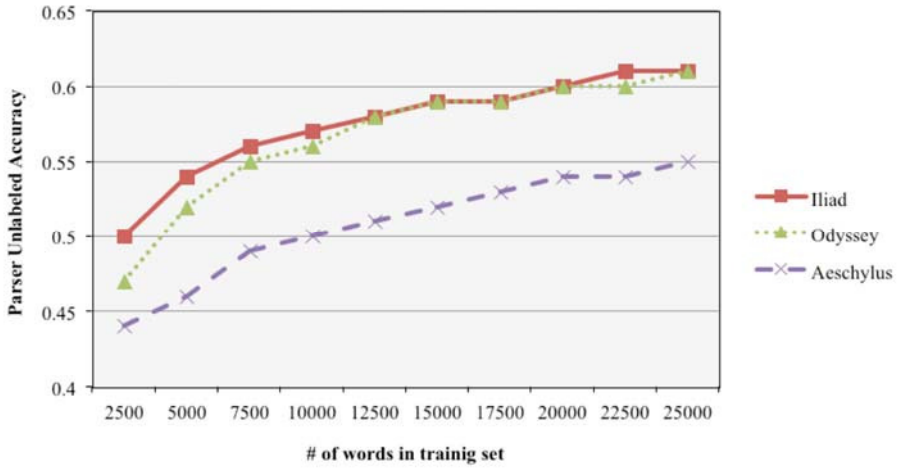


Figure 14: Learning curves for the Iliad*,* Odyssey *and the works of Aeschylus (Saeed Majidi)*

The figure above tracks the growing accuracy of an automatically trained syntactic parser as the training set increases. Saeed Majidi, a PhD candidate in Computer Science at Tufts University computed these figures by using curated syntactic analyses for the Homeric Epics and for Aeschylus, training the parser on part of the curated data and then running the parser against the rest, comparing the parser output with the curated analyses. Two thousand years of students who have worked on Classical Greek would not be surprised to see that Aeschylus is harder than Homer for machines as well as for human beings.

Even noisy syntactic data can be very useful if it is large enough – in effect, errors tend to be random while significant results cluster into significant patterns. In other words, the signal will, in many cases, be visible despite the noise. Relatively modest training sets (10,000-50,000) can generate automatic syntactic analyses that are 50-60% accurate and that provides a great deal of useful data.

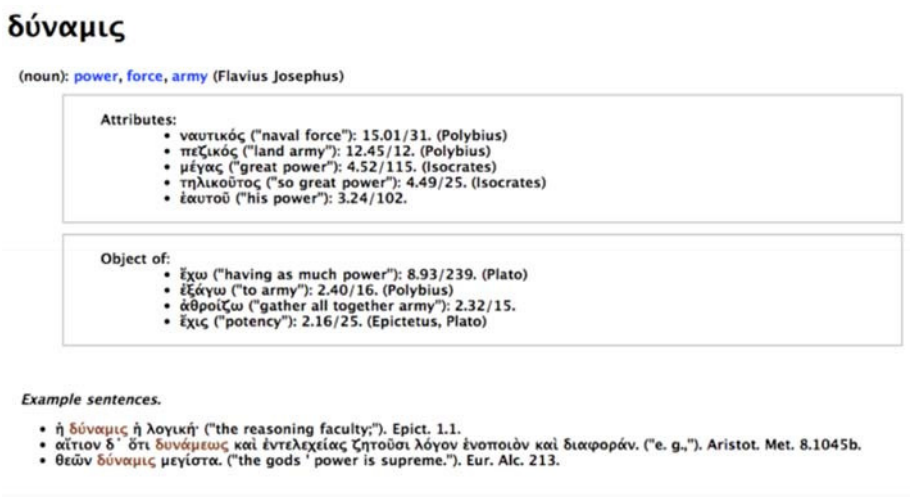


Figure 15: Dynamic Lexicon Entry for the Greek noun δύναμις (David Bamman)

The figure above presents work from the Dynamic Lexicon project,[81] which applied computational methods to extract basic lexical data. The figures above are derived from a corpus of 8 million words of Greek, of which c. 5 million have been aligned with English translations. "While the automatically induced information naturally contains noise (e.g., the misclassification of ἕχις or the mistranslation of the second example sentence), it reveals larger patterns of usage consistent with traditional lexica. In particular, we have automatically induced three categories of information:

- ***Morphology.*** This entry has correctly categorized δύναμις as a noun. Some lexemes have multiple parts of speech – e.g., the very common word καί can be used as a conjunction ("and") and as an adverb ("even") and has different sense and syntactic behavior as a result of this distinction.
- ***Sense.*** By aligning all our Greek source texts with their English translations at the level of individual sentences and then words, we have induced that δύναμις has three predominant senses in all of Greek literature – "power," "force," and "army" – and that "army" itself is an especially dominant sense in the works of Flavius Josephus.
- ***Syntax.*** The availability of syntactically-parsed data allows us to calculate that the most common attributes for δύναμις are ναυτικός

[81] See BAMMAN & CRANE (2008).

("naval") and πεζικός ("on foot") – both especially dominant in the works of Polybius. The alignment of parallel texts lets us present appropriate translations of each (e.g., a naval *force* rather than a naval *army*)

In addition, the availability of Greek/English and Latin/English parallel text that has been aligned at the level of individual sentences also allows us to supplement the lexical entry with several instances of its actual use in text – allowing us to present not only the source text but also its automatically aligned translation."[82]

The Dynamic Lexicon cannot create finished articles on the grammatical usage and meanings of a word but it does provide a starting point – and more importantly it scales to large collections. The *Thesaurus Linguae Latinae* (TLL), begun in 1894, is creating a lexicon for Latin through c. 600CE. Its staff page lists 23 names,[83] including a general editor, four editors, and twelve collaborators. "The work is based on an archive of about 10 million slips which takes account of all surviving texts. In the older texts there is a slip for each occurrence of each word; the later ones are generally covered by a selection of lexicographically relevant examples."[84] As of 2012, published volumes of the TLL had reached the beginning of the letter "r".[85]

There are now billions of words available in Latin. Approaches such as those demonstrated in the Dynamic Lexicon grow more, rather than less, effective as the collection size increases. But the accuracy of those automated processes depends upon the size and quality of the training data. Each digital edition not only serves an immediate circle of human readers but also contributes new data to intelligent services, some already in operation and surely others that we cannot yet predict. The digital edition is distinguished by its ability to support interaction between each individual reader and a growing network of increasingly sophisticated services.

The Greek and Latin Dependency Treebanks available from Perseus represent a basic standard. They encode morphological form and syntactic function but they do not include other features (such as

---

[82] http://nlp.perseus.tufts.edu/lexicon/ -- quoted text and research by D. Bamman.

[83] http://www.thesaurus.badw.de/english/index.htm -- accessed on October 26, 2012.

[84] http://www.thesaurus.badw.de/english/index.htm.

[85] http://www.badw.de/publikationen/kommissionen_publ/thesaurus/index.html Vol. XI 2 Fasc. I: r – rarus. Redaktoren: J. Blundell, S. Clavadetscher, C. G. van Leijenhorst.

co-reference resolution, which specifies who the "he" or "they" are in a given sentence). The Greek and Latin Treebanks represent only a conservative first step, representing only the most obvious annotations that should accompany digital texts. The dominant shape of digital editions will depend upon a social consensus that will evolve over time. The morpho-syntactic analyses reflect a very conservative estimate of what will be expected either a decade or a generation from now.

7. New forms of intellectual production. Wikipedia will almost certainly be remembered as the single most important advance for the humanities from the early twenty-first century. Wikipedia as a particular project may or may not flourish over time but it has nonetheless demonstrated a fundamentally new mode of intellectual production, one that is far more deeply collaborative than any of its immediate print predecessors.[86] Humanists who question the potential of this medium because they find the articles in their area problematic might spend time working with Wikipedia articles on mathematically complex topics (one example of which is shown in the figure below). These cover concepts quite as challenging as any that students of historical languages face. If the articles on Greco-Roman topics are not as impressive as those for various mathematical sciences, then that only means that those of us who advance understanding of the past as a vocation have ourselves not developed the broader community of interest.

[86] The volume of both scholarship and academic commentary on either the importance of or the disaster of Wikipedia as both a collaborative knowledge creation model and as a reference source is far too vast to wade into here, but for some differing perspectives, see the seminal piece on open source history by R. ROSENZWEIG (2006), for an example of using Wikipedia articles as a model to improve student writing (GRAHAM 2012), and for faculty uses of and responses across the disciplines (DOOLEY 2010, WHALLEY 2012).

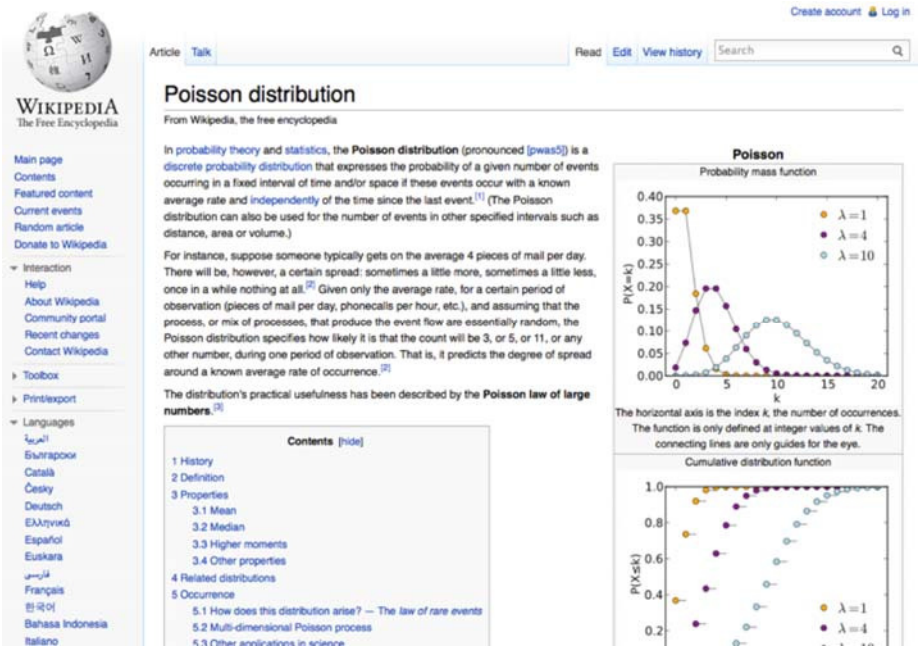


Figure 16: Wikipedia article on the "Poisson distribution" in probability theory (as of October 24, 2012). The decentralized mode of intellectual activity produces an immense amount of clear, accurate exposition on topics quite as complex as those addressed by students of historical languages.

The Homer Multitext Project[87] (HMT) may well be the most important project that has emerged within Classical studies since the beginning of the twenty-first century.[88] Only within that recent time frame have we had the technology to create, store, distribute and license very high-resolution images of manuscripts. The first three changes reflect decreases in the costs of digital cameras, storage and bandwidth. The fourth feature may be less obvious but machine-actionable licenses, such as those available in a growing number of languages, provided by Creative Commons[89] are essential for scalable work with digital sources. In the first generation of digital work, licenses were written in expository prose and could differ in multiple ways. If one wished to create a work with materials from different

[87] http://www.homermultitext.org/.

[88] For more on the history and scholarly future of the HMT, see NAGY (2010), and for an outline of the technical choices, see SMITH (2010).

[89] http://creativecommons.org/.

sources, each source required a separate agreement. Such a procedure does not scale to projects that may draw upon thousands of different sources, especially when projects may dynamically detect and repurpose newly available materials (e.g., a morphological and syntactic analysis engine that generates annotations for Greek and Latin sources as these become available).

The HMT seeks to represent the textual history of the Homeric Iliad and Odyssey in its full complexity. This task is particularly challenging because the Homeric epics emerge from an oral poetic tradition that was formulaic and fluid in nature. Thus the HMT is not attempting to create a single authoritative edition but rather to represent every detectable version of the Homeric epics.[90] To do so requires far more detailed publication of the surviving manuscripts than has ever been feasible before. The general idea behind the HMT is not necessarily new – Milman Parry and Albert Lord articulated models of oral composition for the Homeric epics in the twentieth century. The method behind the HMT represents a sharp departure from recent practices.

Undergraduate researchers play fundamental roles in the HMT.[91] The most knowledgeable experts of particular manuscripts are juniors and seniors who have worked for years on these documents and who publish their findings. The summer of 2012, for example, saw research published by Stephanie Lindeborg on "Catalog of Ships Summary Scholia Part Two: Comparing the Y.1.1 with the Venetus B" and "Catalog of Ships Summary Scholia in the Escorial Y.1.1"[92], Matthew Angiolillo and Christine Roughan on "Scholia to Iliad 14.506 in Two Manuscripts in Venice (Venetus A and Marciana 458)"[93] and Thomas Arralde on "Identifying Aristarchean Commentary in the Venetus A Scholia."[94] The expository form of this research follows the traditions of expository prose that have evolved over millennia.

---

[90] For further discussion of these issues, see DUÉ & EBBOTT (2009).

[91] To read more about the role of undergraduate researchers and the HMT, see BLACKWELL & MARTIN (2009).

[92] http://homermultitext.blogspot.de/2012/08/catalog-of-ships-summary-scholia-part.html; http://homermultitext.blogspot.de/2012/08/catalog-of-ships-summary-scholia-in.html.

[93] http://homermultitext.blogspot.de/2012/07/scholia-to-iliad-14506-in-two.html.

[94] http://homermultitext.blogspot.de/2012/06/identifying-aristarchean-commentary-in.html.

The relationship between the arguments and the data within the manuscript is radically traditional – it departs from the print conventions by more fully realizing the ideals of scholarly argumentation. These publications explicitly document their arguments with high-resolution images of those sections of the manuscripts upon which they base their arguments. At the same time, these particular images contain the coordinate data that allows automatic linking directly into the archival images, available at high resolution and often in multiple spectra of light.[95] Assertion and evidence are far more tightly – and consistently – linked than was ever feasible in print – especially when arguments depended upon extensive visual imagery. The underlying idea is deeply traditional – footnotes have for centuries allowed us to define our sources. But we can realize that traditional idea much more fully.

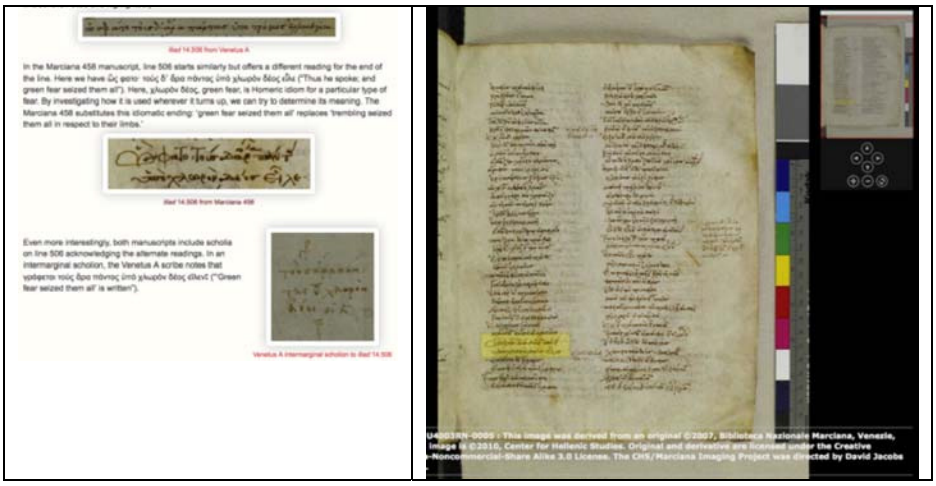


Figure 17: Citations to particular passages in a manuscript include coordinate data that enables dynamic linking into images available at high resolution and in multiple spectra of light.[96]

The HMT demonstrates a new culture of intellectual activity, one in which undergraduates have an opportunity to develop their own

[95] The ability to create tools or programs that can at least semi-automatically link manuscript transcriptions directly to images, particularly at the word level, has been a subject of active research, see for example, FISCHER *et al.* (2011), PORTER *et al.* (2009), and CAYLESS (2008).

[96] Example drawn from http://homermultitext.blogspot.de/2012/07/scholia-to-iliad-14506-in-two.html.

voices and to contribute in substantive ways. The figure below uses different colors to mark different elements and logical relationships within one page of the tenth century Byzantine Venetus A manuscript. There are at least four categories of annotation associated with the text of the Iliad (left of the text, right of the text, interlinear, surrounding the text) and various relationships between the scholia, the text and each other.

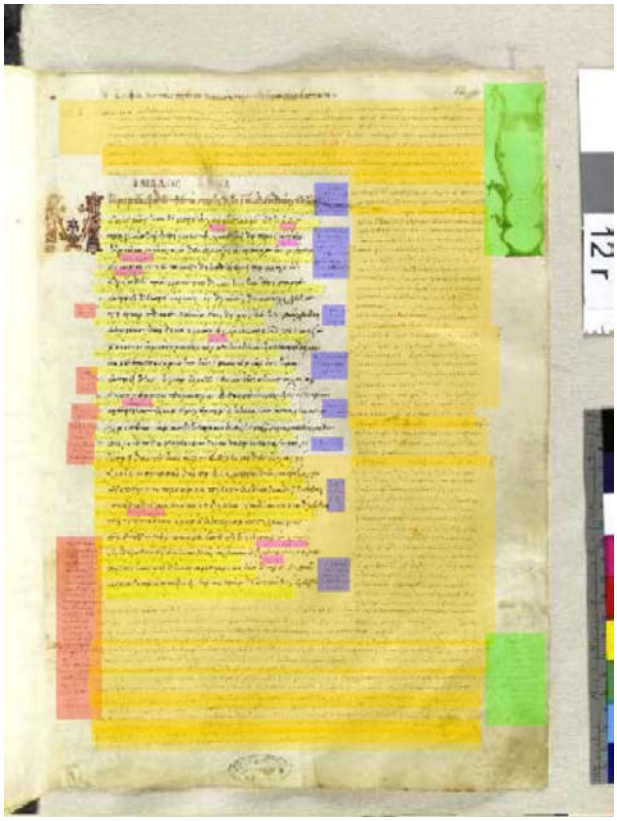


Figure 18: Venetus A, folio 12 recto, with the first 25 lines of the Iliad*; overlays show the location of scholia, color-coded for their class of placement on the folio.*[97] *First year students of Greek were able to create these overlays, providing them with an early opportunity in their careers to use their incipient knowledge of Greek to contribute fundamental data that no machine could provide.*

[97] http://homermultitext.blogspot.de/2012/07/verifying-inventory-of-scholia.html.

No page layout system can identify the regions of the manuscript page above. Nor can existing systems for handwriting analysis determine the first and last lines of the *Iliad* in the central textual section on the page above. These are, however, fundamental tasks for the analysis of the manuscript as a whole. Students of Greek can, however, as early as their first year, begin to contribute such analyses, learning how to interpret the manuscripts as a whole and how to associate the Byzantine script to the characters that they learned in their textbooks and the Greek poetry that they aspire to read.

Ultimately the HMT upon far more detailed transcriptions and representations of the textual data than were ever published in print. In August 2012, the HMT published TEI XML transcriptions of the Iliadic text and scholia from *Iliad* 1-6 in the Venetus A manuscript, and other texts from the first eleven folios of the Venetus A manuscript. Undergraduates at Furman, Holy Cross and the University of Houston produced these transcriptions, working with each other and with their faculty collaborators over several years.

In the twentieth century, the study of manuscripts involved the specialized field of palaeography.[98] Advanced researchers might have an opportunity to take seminars in this subject, working often with facsimiles of the originals produced as large-scale books or as microfilms. Few, if any undergraduates, took such courses – they were expected to focus on learning the standardized Greek and Latin of their critical editions. In the twenty-first century, we find undergraduates energized by access to very high-resolution images of these originals and (like their counterparts in the growing citizen science movement) by the realization that they can contribute to human knowledge. At Holy Cross and Furman, enrollments in Classical Greek have expanded – with 2,898 and 2,951 students each, both schools have more than 25 students in introductory Greek. Undergraduate interest in manuscripts has led to a new open palaeography project.[99] The Holy Cross Manuscripts, Inscriptions and Documents Club – a student organized, volunteer organization – advances "the study of these academic fields: paleography, codicology, epigraphy, as well as the study of languages. We strive for undergraduate

[98] For one perspective on how the study of palaeography is changing with the availability of digital methods, see CIULA (2009).

[99] http://homermultitext.blogspot.de/2012/10/announcing-open-paleography-project.html.

inclusion in work normally reserved for the graduate level."[100] "At the club's first general meeting of the new academic year on Friday, seventeen returning members and three faculty collaborators were joined by twenty newcomers. Six of the club's most active members could not attend Friday's meeting because they are currently studying abroad, but they have already sent back photographs of inscriptions as part of a club project on the epigraphic sources for tribute in fifth-century Athens, just one of an expanded roster of projects the club is hosting this year."[101]

Others have encountered the enthusiasm that students and the general public show when working with original sources.[102] The HMT is important because the Byzantine Greek manuscripts offer great challenges of form (they contain many abbreviations as well as handwriting that is very different from modern Greek fonts) and of content (they contain not only the archaic poetic dialect of the Homeric epics but much later technical prose of commentators writing about grammar, meter, style, and other subjects). The HMT demonstrates the feasibility of a very hard case. If undergraduates working together and with their faculty can produce data about and research on these Homeric manuscripts, they can contribute a wide range of challenging subjects in many languages.

The HMT and the Greek and Latin treebanks each contribute essential components to a mature digital edition. The HMT addresses the challenge of documenting textual witnesses that are inherently complex in form and that cannot be analyzed by methods such as OCR or handwriting recognition. The Greek and Latin treebanks provide the linguistic analyses for the phenomena transcribed from various paper, papyrus or stone sources. Both share a common

---

[100] http://shot.holycross.edu/hcmid/.

[101] http://homermultitext.blogspot.de/2012/09/undergraduate-interest-in-manuscripts.html.

[102] Another example from Classical studies can be found at http://udallasclassics.org/maurer_files/Valla-Intro.htm, which publishes transcriptions of Lorenzo Valla's translation of Thucydides into Latin: "The motive was given by an undergraduate Thucydides course at the University of Dallas, in fall 2008, where at my suggestion, two students chose to transcribe Valla's translation of the Plataean Debate (using Stephanus' text) instead of writing a term paper. I suggested this knowing that it would help both their Latin and their Greek, and give them a glimpse (normally denied to undergraduates) of the rich (in Thucydides' case peculiarly, immensely rich) history of classical scholarship. But when I saw that they did this work with gusto, remarkably carefully and accurately, it occurred to me that it might interest others too; so I added the apparatus, and now put the whole thing online."

philosophy that emphasizes the links between assertions and the data upon which those assertions are based. While the HMT links transcriptions to images, the Greek and Latin treebanks allow us to link assertions about particular linguistic phenomena to the precise places where those phenomena occur.

And like the HMT, the Greek and Latin treebanks depend upon collaboration among students and professional researchers. Two undergraduate or MA-level students independently proposed morphological and syntactic analyses for 230,000 words in the Homeric *Iliad* and *Odyssey*. A professional Homerist, Jack Mitchell, resolved those instances where two different analyses were proposed. The result was a data set in which each sentence has identifiers for the initial annotators and the expert reviewer. Each sentence constitutes a distinct, citable publication that sets out to describe a defensible interpretation.


```
- <sentence id="3044" document_id="Perseus:text:1999.01.0133" subdoc="book=6:card=1" span="pa/ntas0:.4">
    <primary>mpkinn10</primary>
    <primary>millermo</primary>
    <secondary>nicanor</secondary>
    <word id="1" form="pa/ntas" lemma="pa=s1" postag="a-p---ma-" head="3" relation="OBJ"/>
    <word id="2" form="ga\r" lemma="ga/r1" postag="g--------" head="3" relation="AuxY"/>
    <word id="3" form="file/esken" lemma="file/w1" postag="v3siia---" head="0" relation="PRED"/>
    <word id="4" form="o(dw=l" lemma="o(do/s1" postag="n-s---md-" head="5" relation="ADV"/>
    <word id="5" form="e)/pi" lemma="e)pi/1" postag="r--------" head="7" relation="AuxP"/>
    <word id="6" form="oi)ki/a" lemma="oi)ki/on1" postag="n-p---na-" head="7" relation="OBJ"/>
    <word id="7" form="nai/wn" lemma="nai/w2" postag="t-sppamn-" head="3" relation="ADV"/>
    <word id="8" form="." lemma="period1" postag="u--------" head="0" relation="AuxK"/>
  </sentence>
- <sentence id="3045" document_id="Perseus:text:1999.01.0133" subdoc="book=6:card=1" span="a)lla/0::2">
    <primary>mpkinn10</primary>
    <primary>millermo</primary>
    <secondary>nicanor</secondary>
    <word id="1" form="a)lla/" lemma="a)lla/1" postag="d--------" head="14" relation="AuxY"/>
    <word id="2" form="oi(" lemma="e(/1" postag="p-s---md-" head="8" relation="OBJ"/>
    <word id="3" form="ou)/" lemma="ou)1" postag="d--------" head="8" relation="AuxZ"/>
    <word id="4" form="tis" lemma="tis1" postag="p-s---mn-" head="8" relation="SBJ"/>
    <word id="5" form="tw=n" lemma="o(1" postag="l--------" head="4" relation="ATR"/>
    <word id="6" form="ge" lemma="ge1" postag="g--------" head="5" relation="AuxZ"/>
    <word id="7" form="to/t'" lemma="to/te1" postag="d--------" head="8" relation="ADV"/>
    <word id="8" form="h)/rkese" lemma="a)rke/w1" postag="v3saia---" head="14" relation="PRED_CO"/>
    <word id="9" form="lugro\n" lemma="lugro/s1" postag="a-s---ma-" head="10" relation="ATR"/>
    <word id="10" form="o)/leqron" lemma="o)/leqros1" postag="n-s---ma-" head="8" relation="OBJ"/>
```


Figure 19: Morphological and syntactic analyses represented as XML. Each sentence contains a unique identifier for the two annotators (<primary>) and the Homerist (<secondary>) who reviewed their contributions to create the final collaborative entries in the Treebank.

The workflow used to develop the treebank data for Homer was designed to produce data of high accuracy but it was, in its initial form, slow. Months might pass after a first student created an initial annotation before a second annotation was created and the two were compared. There was no mechanism to provide students with significant feedback. The goal was to generate data.

But the treebanking process can be organized to produce data of high accuracy quickly and to give students feedback as they create that data. The figure below illustrates how two different students in a third semester Latin class differently annotated the same sentence. In this scenario, students can independently annotate one or more sentences, then work together to resolve the different interpretations, present the final results (with questions) to the class and instructor and publish, by the end of the class, the results as data for comment. The class can build up their own corpus over a semester, eliciting comments and feedback from the broader community and making such adjustments as they see fit. New interpretations can – and inevitably will – be proposed long after the class. The results can be quite accurate.

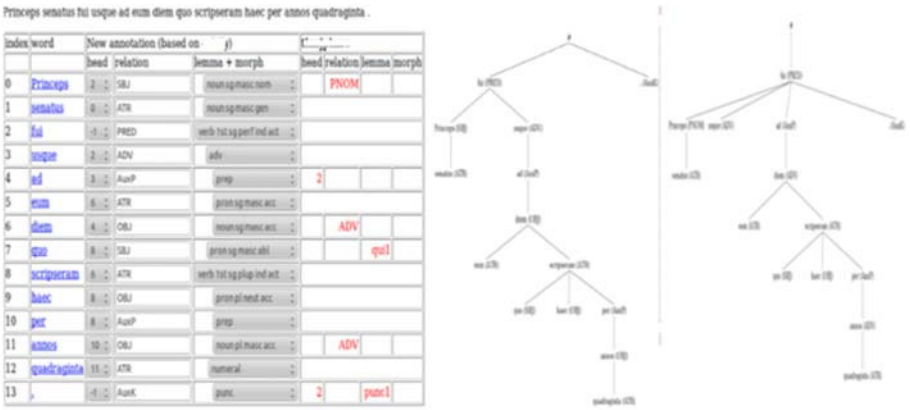


Figure 20: Individual sentences analyzed by third semester Latin students. The left display shows in red where students differed in their analyses. The right display visualizes the interpretations as trees. We will be able to support such dynamic activities, where individuals, whether in the same classroom or in completely different locations, can compare their analyses, revising or defending their choices. In a classroom setting, the instructor can help adjudicate and classroom work can, where a consensus appears, be immediately submitted as a contribution to the Greek or Latin Treebanks, with instructor and students as joint, named contributors.

Students have accounted for the morphological and syntactic functions of words in Greek and Latin since grammatical analysis began in antiquity but this ancient pedagogical practice can now produce much of the linguistic data that we need, both to rebuild our understanding of Greek, Latin and other historical languages on

explicit, evidence-based models and to support a global audience of readers from many different linguistic and cultural backgrounds.

The greatest challenge facing Greek, Latin and other historical languages is social rather than technical. A new intellectual culture has begun to emerge that reflects the strengths and possibilities of a society where ideas circulate primarily in digital, rather than print, forms. The departments that provide doctoral training for new researchers remain, certainly in the field of Greek and Latin, deeply rooted in a traditional print culture that emphasizes single authored, static publications and specialist audiences rather than collaborative research, dynamic knowledge bases (of which a digital edition constitutes a special case) and the relentless effort to use specialized scholarship to advance the general life of society.

A new generation of researchers is increasingly eager to move forward, if only because many realize that fields that do not exploit the strengths of digital culture are at a disadvantage and because students of historical languages have enough disadvantages in the twenty-first century. A NEH-funded three week institute on Working with Texts in a Digital Age[103] attracted almost eighty applications for twenty-five slots. All of the participants – most of them early in their careers and under pressure to complete PhDs or to crank out publications – had agreed to devote a substantial part of their summer time to acquiring new skills and they thus reflected a self-selected group with a stated interest in digital methods. Most expressed profound surprise at how much was, in fact, possible. Even those who were most active already on digital projects had little, if any, exposure to immediately applicable methods from either corpus or computational linguistics.

We are poised for a shift in the intellectual culture of the humanities as a whole and of philology in particular. In the twentieth century, departments of Classics in the United States and elsewhere began, of necessity, to develop curricula for students who studied little or no Greek and Latin. Such a move was necessary because of the decline in the number of students who entered college with background in either of these languages. The APA has even begun serious consideration of changing its name – "the term philology has become so obscure to all but practitioners as to impede

[103] http://sites.tufts.edu/digitalagetext/.

our efforts to gain broader public (even academic) visibility."[104] We have certainly come a long way from 1956, when the mad scientist of the film *Forbidden Planet*[105] was a philologist. But the present obscurity of the term creates an opportunity to reinvent and refashion its meaning and to assert, in fact, a meaning much like that of Friedrich Wolf in eighteenth century Halle and Augustus Boeckh in nineteenth century Berlin, for whom philology aimed at fostering an understanding of antiquity as a whole (*cognitio universae antiquitatis*) and a means to breath life back into the past. As the Greek and Latin sources become accessible to a global audience, the old term for studying these sources directly may reassert itself and become a symbol of a reborn field.

Nevertheless, we see now in the twenty-first century opportunities to re-integrate the language into our curricula, both by making the language more accessible and by making contribution and research feasible for our undergraduates. We have an opportunity in the study of Greek, Latin, and other historical languages to be leaders in fostering a new generation of student researchers and citizen scholars. An opportunity is, however, not inevitability, and no technological determinism will save or overwhelm us. How well we realize the possibilities emerging before us will depend upon decisions that we make as communities and as individuals.

The role of Germany

This paper builds upon a 2011 talk delivered to the Berlin-Brandenburg Academy of Sciences, the twenty-first century successor to the Prussian Academy of Sciences founded by Gottfried Wilhelm Leibniz in 1700 more than 300 years before. In that period, Germany became, for many years, the primary center for scholarship on Greek and Latin. Early in the twenty-first century, Germany has a unique opportunity to build upon this tradition of scholarship and to advance a global dialogue among civilizations.

First, Germany now occupies a unique position within the world. The strongest economic power within the European Union, Germany also lacks the complicating background in global affairs that color perceptions of the geopolitically active Anglo-American nations.

---

[104] Jeff Henderson, APA president: http://apaclassics.org/index.php/apa_blog/apa_blog_entry/request_for_comments_on_possible_name_change_for_association/.

[105] http://www.imdb.com/title/tt0049223/.

Within the diplomatic conditions of the early twenty-first century, no country in Europe or North America is better situated to advance a dialogue among civilizations than is Germany.

Second, in the period between 1700 and the present, more editions of Greek and Latin may have been produced in the area of contemporary Germany than in the rest of the world – Leipzig, in particular, was the greatest center for the publication of Greek and Latin print editions through the Second World War. And German authors produced an immense stream of original Latin in virtually every written genre and on every topic from the medieval period through the twentieth century. This immense body of Greek and Latin represents a major component of German cultural heritage and well deserves digital publication. A library of Greek and Latin produced in the German speaking lands would be of immense value to those interested not only in the texts themselves but also in the intellectual and cultural history of Europe.

Third, German academic traditions do not separate computer science from the humanities – both are instances of *Wissenschaft*, where the English term "science" is used exclusively for the natural and, when qualified, social sciences. The semantic distinction has immense practical consequences in the Anglo-American world. In the United States, for example, the NEH[106] (with a 2010 budget of around US$167 million) and the NSF[107] (with a 2010 budget of around US$6.89 billion) are officially separate organizations that serve different communities. The NSF can support computer scientists working on applications in biology, physics, earth sciences, or any other NSF-supported discipline but the NSF cannot readily support computer science research on subjects that belong to the NEH. With a budget 40 times smaller than the NSF, the NEH simply cannot provide significant support for computer science research, however important that may be for the humanities.[108] Efforts such as

[106] http://www.neh.gov.

[107] http://www.nsf.gov.

[108] The NEH Preservation and Access Research and Development track can provide up to $350,000 (http://www.neh.gov/grants/preservation/preservation-and-access-research-and-development) -- a very large sum for NEH grants but well below the $500,000 cap for small grants awarded for Computer Science research by the NSF: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12765&org=CISE&sel_org=CISE&from=fund.

the Digging into Data Program[109] depend upon ad hoc collaborations to bring NEH and NSF funded research together.

In Germany, computer scientists face no structural barriers if they wish to focus their research upon problems from the humanities. In 2012, the German Ministry of Education[110] announced that it had provided 19.5 million Euros to support research projects that involved computer science and the humanities. In April of 2012, the Humboldt Foundation announced my own election to a Humboldt Chair of Digital Humanities, a chair situated in a Department of Computer Science at Leipzig and bringing with it support of 5,000,000 Euros over five years. Leipzig was already hosting projects with joint humanist and computer scientist teams with aggregate support of c. 1 million Euros a year. Other such collaborations between humanists and computer scientists can be found around Germany. The overall consequence of this for the humanities in a digital world could be profound in the long run. In Germany, emerging researchers in computer science can explicitly build a career on collaboration with humanists. If the 2012 19.5-million euro BMBF investment draws promising computer scientists into long-term research agendas relevant to the humanities, that one program can shape development for decades.

Fourth, Germany passed in 1965 an explicit law to define the rights status of editions. Sophocles and Vergil may be long gone, but German law provides protection to scientific editions for a period of 25 years after publication.[111] Textual notes on the bottom of the page in many editions may be considered a separate original work and qualify for the regular European protection of the life of the author + 70 years. This complicates redistribution of the text as scanned image book because the textual notes on the bottom of the page would have to be excluded. Nevertheless, the reconstructed text can be manually marked before or after the books are scanned, and methods exist to identify the text automatically. The reconstructed texts of editions published through 1987 can be redistributed in 2012, with a moving wall freeing a year's worth of editions with each new calendar year.

The German situation does not reflect the full needs of scholarship. Scholars who handed over their introductions and

[109] http://www.diggingintodata.org/.

[110] www.bmbf.de.

[111] http://de.wikipedia.org/wiki/Schutz_wissenschaftlicher_Ausgaben.

textual notes to publishers can expect that, under current law, their work will not be able to circulate freely for scholarly analysis until all of their immediate colleagues are long dead – a grandchild ten years old at the editor's death would be eighty before the editorial data was available. But, of course, even if the printed editions were released, they do not represent their data in a machine actionable format (e.g., you can't use a digitized apparatus to compare dynamically the contributions of multiple witnesses) and they do not include the full range of data for a true digital edition (e.g., commas, periods, and other annotations from print culture are imposed upon the original text but print editions do not record the morphological, syntactic and other analyses behind punctuation and page layout in any form, machine- or human-readable).

Nevertheless, recently printed books lend themselves to OCR better than do older books. OCR software could be applied to a library of page images from editions whose authors have not been dead 70 years but that were published 25+ years ago. The OCR-generated text can then be aligned to other editions and the scholarly community can then quickly see how individual passages in this edition relate to others that are available online. Because editors worked on Greek and Latin sources from the fifteenth century through the present, one or more complete editions – including introduction and textual notes – is available for digitization for virtually every Greek and Latin source printed from manuscript sources.

Conclusions: what is to be done?

If in creating digital editions we wish to foster a dialogue among civilizations – and not all editors may share this goal – we need to work from the two convergent directions of breadth and depth. First, we need to make very large bodies of linguistic sources accessible with methods that are not only scalable but that become more effective as collections grow larger. Second, we need to build upon methods by which to represent our textual sources and linguistic data more precisely, with dense and growing webs of machine actionable annotations that either perfect print practice (e.g., back-of-the book indices of people and places become links to authority lists) or represent a major step forward (e.g., encoding morpho-syntactic analyses, co-reference resolution etc). In effect, students of historical languages must draw upon the results of computational linguistics to account for phenomena at scale and corpus linguistics for intensive analysis. Our goal must be to serve dozens, if not

hundreds, of historical languages, but Greek and Latin provide a starting point: they are big enough and complicated enough for us to develop methods for working with historical languages embedded in much larger collections of modern language materials.

First, to address breadth, we need to put as much of the human textual record as possible online for computational analysis and for the results of that analysis to be shared freely. A great amount of the underlying scanning has already been done. The Internet Archive offers 3.6 million books for public download, HathiTrust currently has 3.2 million public domain books, and Gallica offers more than 1 million books and manuscripts. The original scans of these books should be made available where researchers can apply OCR software customized for particular languages. Such aggregation requires storage as well as computational power.

Second, one can begin by focusing on subsets such as the 65,000 public domain titles out of c. 90,000 that the HathiTrust lists as being in Ancient Greek or Latin. But the real challenge is to find not only the Ancient Greek and Latin in such obvious places but to also track all the quotations of Greek and Latin scattered throughout the other three million plus books. Such tracking includes recognizing passages written primarily in some other language (e.g., English or German) that have quoted shorter passages in Greek or Latin so that we can run customized OCR on the relevant chunks of those pages. Such tracking also includes the ability to recognize as many instances of text reuse as possible, including quotations of a modern language translation of a Greek or Latin work, paraphrases, citations (e.g., Th. 1.32 refers to Thucydides book 1, chapter 32) and names (distinguishing Aristotle the philosopher from Aristotle Onassis).

The HathiTrust Research Center[112] has provided an initial approach to solving this problem for researchers in the United States. This approach is itself evolving but even if perfected for users in the United States, work needs to be done for researchers in Europe, where copyright laws are different and different materials are in the public domain. Germany has a real opportunity to lead in this case because it can provide funding for computer science and humanities collaborations and because of its special copyright laws for editions, which create a moving wall that brings 25-year old editions into the public domain each year.

[112] http://www.hathitrust.org/htrc.

Third, we need to not only educate philologists about new, more intensive, machine actionable methods of representing textual data (such as providing not only punctuation but the morphological and syntactic analyses that punctuation assumes) but also enable them to make informed decisions about how to fashion their work for a rapidly changing intellectual world.

In this we need to engage not only advanced researchers in editing, and library professionals in documenting, historical sources, but we must also involve a generation of student researchers and citizen scholars upon whom we must rely if we are to make the individual documents within the vast and growing digital collections intellectually accessible. Here the means is also the end – at least, insofar as we believe that the end of our work is to advance the intellectual life of humanity and engage society as broadly and deeply as possible.

Two hundred years ago, Augustus Boeckh saw already that the true aim of philology was to understand the ancient world as fully as possible but he also understood that the study of the past was important because it contributed to the lived experience of society as a whole. And one could find such statements from scholars for hundreds and thousands of years before Boeckh, in every corner of Europe, in Baghdad and Cairo, and, of course, in Alexandria. In the end, our methods may change but our goals do not. We honor in the present those values of the past that we most admire by re-imagining those values to serve the future.

REFERENCES

AGOSTI, M. & N. FERRO, 2007: A Formal Model of Annotations of Digital Content, in: *ACM Transactions on Information Systems* 26(1), Article No. 3, 1-57.

ALMAS, B. *et al.*, 2011: What Did We Do With A Million Books: Rediscovering the Greco-Ancient world and reinventing the Humanities, White Paper Submitted to the NEH, National Endowment for the Humanities. http://hdl.handle.net/10427/75558.

ANTONIADIS, G. *et al.*, 2009: Integrated Digital Language Learning, in: BALACHEFF, N. *et al.* (ed.), *Technology-Enhanced Learning*, Dordrecht, Chapter 6, 89-103.

BABEU, A., 2008: Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. Technical report. http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf.

BABEU, A., 2011: *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*, Council on Library and Information Resources. http://www.clir.org/pubs/abstract/reports/pub150.

BAMMAN, D. & G. CRANE, 2008: Building a Dynamic Lexicon from a Digital Library, in: *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, New York, 11-20. http://hdl.handle.net/10427/42686.

BAMMAN, D. & D. SMITH, 2012: Extracting Two Thousand Years of Latin from a Million Book Library, in: *Journal on Computer and Cultural Heritage* 5(1), Article No. 2. http://dx.doi.org/10.1145/2160165.2160167.

BAMMAN, D. *et al.*, 2009: An Ownership Model of Annotation: The Ancient Greek Dependency Treebank, in: *TLT 2009-Eighth International Workshop on Treebanks and Linguistic Theories*. http://hdl.handle.net/10427/70399.

BAMMAN, D. *et al.*, 2010: Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection, in: *JCD '10: Proceedings of the 10th annual joint conference on Digital libraries*, New York, 11-20. http://hdl.handle.net/10427/70398.

BARKER, E. *et al.*, 2010: Mapping An Ancient Historian In A Digital Age: The Herodotus Encoded Space-Text-Image Archive (HESTIA), in: *Leeds International Classical Studies* 9. http://www.leeds.ac.uk/classics/lics/2010/201001.pdf.

BEAULIEU, M.-C. & B. ALMAS, 2012: Digital Humanities in the Classroom: Introducing a New Editing Platform for Source Documents in Classics, in: *Digital Humanities*. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/digital-humanities-in-the-classroom-introducing-a-new-editing-platform-for-source-documents-in-classics/.

BERTI, M. *et al.*, 2009: Collecting Fragmentary Authors in a Digital Library, in: *JCDL '09: Proceedings of the 2009 joint international conference on Digital libraries*, New York, 259-262. http://hdl.handle.net/10427/70401.

BLACKWELL, C. & T. R. MARTIN, 2009: Technology, Collaboration, and Undergraduate Research, in: *Changing the Center of Gravity*, Vol. 3 No. 1.

BODARD, G., 2009: Digital Classicist: Re-use of Open Source and Open Access. Publications in Ancient Studies, in: *Digital Humanities 2009 Conference Abstracts*, 2. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

BODARD, G. & J. GARCÉS, 2009: Open Source Critical Editions: A Rationale, in: DEEGAN, M. & K. SUTHERLAND (eds.), *Text Editing, Print and the Digital World*, Farnham, Surrey, 83-98.

BOSCHETTI, F., 2007: Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses Onto Reference Text, in: *CL 2007: Proceedings of the Corpus Linguistics Conference*, Birmingham. http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf.

BOSCHETTI, F. *et al.*, 2009: Improving OCR Accuracy for Classical Critical Editions, in: AGOSTI, M. *et al.* (ed.), *Research and Advanced Technology for Digital Libraries*, Lecture notes in computer science 5714, Berlin / Heidelberg, Chapter 17, 156-167. http://hdl.handle.net/10427/70402.

BÜCHLER, M. *et al.*, 2010: Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts, in: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(2). http://letterpress.uchicago.edu/index.php/jdhcs/article/viewArticle/60.

BÜCHLER, M. *et al.*, 2012: Increasing Recall for Text Re-use in Historical Documents to Support Research in the Humanities Theory and Practice of Digital Libraries, in: *Theory and Practice of Digital Libraries (TPDL)*, Lecture Notes in Computer Science 7489, Berlin / Heidelberg, Chapter 11, 95-100.

CAUSER, T. *et al.*, 2012: Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham, in: *Literary and Linguistic Computing* 27, 119-137.

CAYLESS, H. A., 2008: Linking Page Images to Transcriptions with SVG, in: *Balisage: The Markup Conference 2008*, 12-15. http://www.balisage.net/Proceedings/vol1/html/Cayless01/BalisageVol1-Cayless01.html.

CAYLESS, H. A., 2010: Ktêma es aiei: Digital Permanence from an Ancient Perspective, in: BODARD, G. & S. MAHONY (eds.), *Digital Research in the Study of Classical Antiquity*, Burlington, 139-150. http://philomousos.com/papers/Cayless_DRSCA.pdf.

CAYLESS, H. A. *et al.*, 2009: Epigraphy in 2017, in: *Digital Humanities Quarterly* 3. http://www.digitalhumanities.org/dhq/vol/3/1/000030.html.

CISNE, J. L. *et al.*, 2010: Mathematical Philology: Entropy Information in Refining Classical Texts' Reconstruction, and Early Philologists' Anticipation of Information Theory, in: *PloS one* 5: e8661 + . http://dx.doi.org/10.1371/journal.pone.0008661.

CIULA, A., 2009: The Palaeographical Method Under the Light of a Digital Approach, in: *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age*. Norderstedt, 219-237. http://kups.ub.uni-koeln.de/volltexte/2009/2971/.

CLEMENT, T., 2012: Methodologies In The Digital Humanities For Analyzing Aural Patterns In Texts, in: *Proceedings of the 2012 iConference*, New York, 287-293.

CLOUGH, P. D. *et al.*, 2009: Extending Domain-Specific Resources to Enable Semantic Access to Cultural Heritage Data, in: *Journal of Digital Information* 10.

http://journals.tdl.org/jodi/article/view/698.

CRANE, G. *et al.*, 2012: Student Researchers, Citizen Scholars And The Trillion Word Library, in: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, New York, 213-222. http://hdl.handle.net/10427/75559.

DIEKEMA, A. R., 2012: Multilinguality in the Digital Library: A Review, in: *The Electronic Library* 30, 165-181.

DOOLEY, P. L., 2010: Wikipedia and the Two-Faced Professoriate, in: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, New York, 1-2.

DUÉ, C. & M. EBBOTT, 2009: Digital Criticism: Editorial Standards for the Homer Multitext, in: *Digital Humanities Quarterly* 3. http://www.digitalhumanities.org/dhq/vol/3/1/000029.html#.

FISCHER, A. *et al.*, 2011: HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents, in: *International Conference on Document Analysis and Recognition (ICDAR), 2011*, Piscataway, NJ, 53-57.

GIBBS, F. W., 2011: New Textual Traditions from Community Transcription, in: *Digital Medievalist* 7. http://www.digitalmedievalist.org/journal/7/gibbs/.

GRAHAM, S., 2012: The Wikiblitz: A Wikipedia Editing Assignment in a First Year Undergraduate Class, in: DOUGHERTY, J. & K. NAWROTZKI (eds.), *Writing History in the Digital Age*, Forthcoming from the University of Michigan Press, web-book edition. http://WritingHistory.trincoll.edu.

GRAHAM, S. & G. RUFFINI, 2007: Network Analysis and Greco-Roman Prosopography, in: Keats-Rohan, K. S. B. (ed.), *Prosopography Approaches and Applications: A Handbook*, Oxford, 325-336.

HUNTER, J. & A. GERBER, 2012: Towards Annotopia—Enabling the Semantic Interoperability of Web-Based Annotations, in: *Future Internet* 4, 788-806.

IFLA 1998: *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications N.S. 19, München. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

ISAKSEN, L. *et al.*, 2012: GAP: A NeoGeo Approach to Classical Resources, in: *Leonardo* 45, 82-83.

KÜSTER, M. W. *et al.*, 2011: TextGrid Provenance Tools for Digital Humanities Ecosystems, in: *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, IEEE May 31 – June 3 2011, 317-323.

LÜDELING, A. & A. ZELDES, 2009: Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics, in: *Jahrbuch für Computerphilologie* 9, 149-178. http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html.

MCCARTY, W. (ed.), 2010: *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Cambridge. http://www.openbookpublishers.com/reader/64.

MICHEL, J.-B. *et al.*, 2011: Quantitative Analysis of Culture Using Millions of Digitized Books, in: *Science* 331, 176-182.

MIHALCEA, R. & M. SIMARD, 2005: Parallel Texts, in: *Natural Language Engineering* 11, 239-246.

MIMNO, D. *et al.*, 2005: Hierarchical Catalog Records Implementing a FRBR Catalog, in: *D-Lib Magazine*. http://www.dlib.org/dlib/october05/crane/10crane.html.

MONELLA, P., 2008: Towards a Digital Model to Edit the Different Paratextuality Levels within a Textual Tradition, in: *Digital Medievalist*. http://www.digitalmedievalist.org/journal/4/monella/.

NAGY, G., 2010: Homer Multitext project, in: MCGANN, J. *et al.* (ed.), *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia March 26-28, 2010*, Houston, 87-112. http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4087.

O'DONNELL, D. P., 2009: Back to the Future: What Digital Editors Can Learn From Print Editorial Practice, in: *Literary and Linguistic Computing* 24, 113-125.

ODIJK, D. *et al.*, 2012: Semantic Document Selection Theory and Practice of Digital Libraries, in: *Theory and Practice of Digital Libraries (TPDL 2012)*, Lecture Notes in Computer Science 7489, Berlin / Heidelberg 2012, Chapter 24, 215-221.

PEURSEN, W. T. (ed.) *et al.*, 2010: *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text scholarship*, Leiden [u. a.].

PIERAZZO, E., 2011: A Rationale of Digital Documentary Editions, in: *Literary and Linguistic Computing* 26, 463-477.

PIOTROWSKI, M., 2010: Leveraging Back-Of-The-Book Indices To Enable Spatial Browsing Of A Historical Document Collection, in: *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, New York, 1-2.

PIOTROWSKI, M., 2012: Natural Language Processing for Historical Texts, in: *Synthesis Lectures on Human Language Technologies* 5, 1-157. http://dx.doi.org/10.2200/S00436ED1V01Y201207HLT017.

PORTER, D. *et al.*, 2009: Text-Image Linking Environment (TILE), in: *Digital Humanities 2009: Conference Abstracts*, 388-390. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

ROBINSON, P., 2010a: Editing Without Walls, in: *Literature Compass* 7, 57-61.

ROBINSON, P., 2010b: Electronic Editions for Everyone, in: *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, 145-163.

ROMANELLO, M. *et al.*, 2009: When Printed Hypertexts Go Digital: Information Extraction From The Parsing Of Indices, in: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, New York, 357-358.

ROSENZWEIG, R., 2006: Can History be Open Source: Wikipedia and the Future of the Past? in: *Journal of American History* 93, 117-146.

SALERNO, E. *et al.*, 2007: Digital Image Analysis to Enhance Underwritten Text in the Archimedes Palimpsest, in: *International Journal on Document Analysis and Recognition* 9, 79-87.

SCHMIDT, D. & R. COLOMB, 2009: A Data Structure for Representing Multi-Version Texts Online, in: *International Journal of Human-Computer Studies* 67, 497-514.

SCHMITZ, P., 2009: Using Natural Language Processing and Social Network Analysis to Study Ancient Babylonian Society, in: *UC Berkeley iNews*.
http://inews.berkeley.edu/articles/Spring2009/BPS.

SIEMENS, R. *et al.*, 2012: Toward Modeling The Social Edition: An Approach To Understanding The Electronic Scholarly Edition In The Context Of New And Emerging Social Media, in: *Literary and Linguistic Computing* 27, 445-461.

SMITH, D. A. *et al.*, 2001: Management of XML Documents in an Integrated Digital Library.
http://xml.coverpages.org/perseus-hopperExtreme2000.pdf.

SMITH, D. A. *et al.*, 2011: Mining Relational Structure from Millions of Books: Position Paper, in: *Proceedings of the 4th ACM workshop on Online books, Complementary Social Media and Crowdsourcing*, New York, 49-54.

SMITH, D. N., 2009: Citation in Classical Studies, in: *Digital Humanities Quarterly* 3.
http://www.digitalhumanities.org/dhq/vol/003/1/000028.html#

SMITH, D. N., 2010: Digital Infrastructure and the Homer Multitext Project, in: BODARD, G. & S. MAHONY (eds.), *Digital Research in the Study of Classical Antiquity*, Burlington, 121-137.

SOSIN, J., 2010: Digital Papyrology, in: *26th Congress of the International Association of Papyrologists (19 August 2010)*.
http://www.stoa.org/archives/1263.

SPORLEDER, C., 2010: Natural Language Processing for Cultural Heritage Domains, in: *Language and Linguistics Compass* 4, 750-768.

STEWART, G. *et al.*, 2007: A New Generation Of Textual Corpora: Mining Corpora From Very Large Collections, in: *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, New York, 356-365.
http://hdl.handle.net/10427/14853.

TOBIN, R. *et al.*, 2008: Named Entity Recognition for Digitised Historical Texts, in: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC '08)*.
http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf.

UNSWORTH, J., 2011: Computational Work with Very Large Text Collections, in: *Journal of the Text Coding Initiative* 1. http://jtei.revues.org/215.

VERTAN, C., 2010: Towards the Integration of Language Tools Within Historical Digital Libraries, in: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10)*, European Language Resources Association (ELRA). http://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/811_Paper.pdf.

WHALLEY, B., 2012: Wikipedia: Reflections on Use and Acceptance in Academic Environments, in: *Ariadne* 69. http://www.ariadne.ac.uk/issue69/whalley.

YALNIZ, I. Z. & R. MANMATHA, 2012: Finding Translations in Scanned Book Collections, in: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, New York, 465-474.

ZHANG, Z. *et al.*, 2010: A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality, in: CIMIANO, P. & H. PINTO (eds.), *Knowledge Engineering and Management by the Masses*, Lecture Notes in Computer Science 6317, Berlin / Heidelberg, Chapter 21, 301-315.

THE RAMSES PROJECT
METHODOLOGY AND PRACTICES IN THE ANNOTATION OF LATE EGYPTIAN TEXTS

STÉPHANE POLIS & JEAN WINAND

0. Introduction

This paper is an updated presentation of the Ramses project being currently developed at the University of Liège.[1] The first section stresses the main objectives and gives a technical description of the general architecture of Ramses software.[2] The second part describes the encoding procedures and reviews the current state of the annotation. In the third section, some changes brought about by the use of large-scale corpora are discussed from an epistemological viewpoint. The paper ends with the presentation of some new avenues for research that will ensue from the use of a complex multilevel corpus.

1. Goals and Means

1.1 The philosophy behind Ramses

The Ramses project that has been under development in Liège since the end of 2006 is deeply rooted in the fields of expertise of its creators. This explains some critical decisions that have been made

---

[1] Previous reports are POLIS, S., 2006: Le projet Ramsès, in: WINAND, J., Un siècle d'Égyptologie à l'Université de Liège, in: WARMENBOL, E. (ed.), *La caravane du Caire. L'Égypte sur d'autres rives*, Louvain-la-Neuve, 180; ROSMORDUC, S. *et al.*, 2009: Ramses. A new research tool in philology and linguistics, in: STRUDWICK, N. (ed.), *Information Technology and Egyptology in 2008. Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique et Égyptologie), Vienna, 8-11 July 2008*, Bible in Technology 2, New Jersey, 133-142; POLIS, S. *et al.*, 2013: Building an annotated corpus of Late Egyptian. The Ramses Project: Review and perspectives, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 25-44; WINAND, J. *et al.*, Forthcoming: Ramses. An annotated corpus of Late Egyptian, in: KOUSOULIS, P. & N. LAZARIDIS (eds.), *Proceedings of the Tenth International Congress of Egyptologists. University of the Aegean, Rhodes, 22-29 May 2008*, Orientalia Lovaniensia Analecta, Leuven, 10 p.

[2] From a technical point of view, Ramses is a relational database in SQL where the texts are represented and stored in XML; the software interface is written in JAVA.

from its very inception. Ramses is both a philological and a linguistic tool, with perhaps more emphasis on the latter dimension. The database is intended to answer all possible questions that can arise when studying a text language. Such a goal is admittedly very ambitious — and might even sound pretentious —, but given the present technical means, it seems far from being unrealistic.

Indeed, most databases presently available — for ancient text languages and for modern languages alike — are usually very good at retrieving isolated words, with varying degrees of precision when it comes to grammatical inflexions. However, they perform less efficiently when it comes to complex queries concerned simultaneously with several layers of annotation. The situation can even become inextricable if these layers are combined within queries that involve several words, phrases or sentences. It is those kinds of shortcomings that the general architecture of Ramses will hopefully overcome.

Moreover, Ramses has been developed with an evolutionary database design: it has the capability of integrating new layers of annotation (that will eventually be connected to the pre-existing levels of annotation). For instance, it would be possible to add a new layer of analysis for tagging proper names with all relevant socio-professional information and to use it as a filter when analyzing the textual data. This, of course, would be a major improvement for those interested in studying prosopography in relation to written production.

1.2 Software Architecture: The relationship between the modules

As a richly annotated corpus, Ramses required software capable of a fair degree of complexity. The first figure (Fig. 1) gives a schematic overview of the general architecture of Ramses' software.

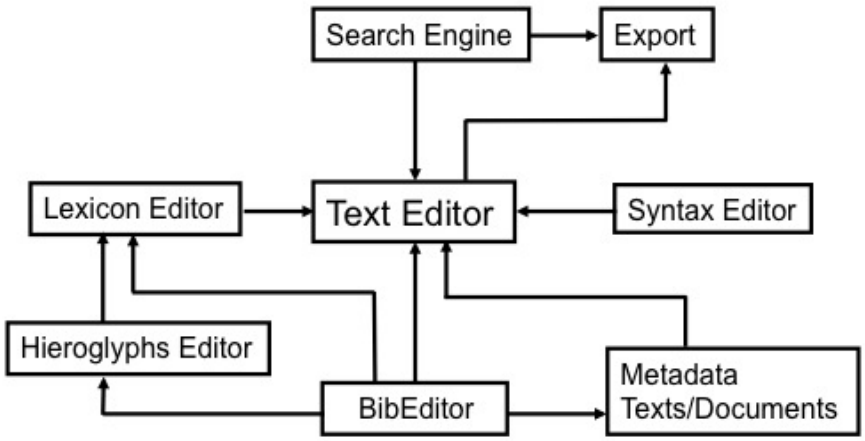


Figure 1. Software architecture of Ramses

1.2.1 The annotation tools: Lexicon, morphology and syntax

The **TextEditor** is the core module. This is the part of the interface that first presents itself on the screen when the database is opened by one of the annotators.[3] The text is segmented in words (Fig. 2); each word is graphically isolated in a box that contains some basic information (Fig. 2, box 1):

- The hieroglyphic spelling;
- The transliteration and the label of the inflexion;
- The standard translation of the lemma.

[3] The interface will obviously be adapted for end-users when we make Ramses available online.

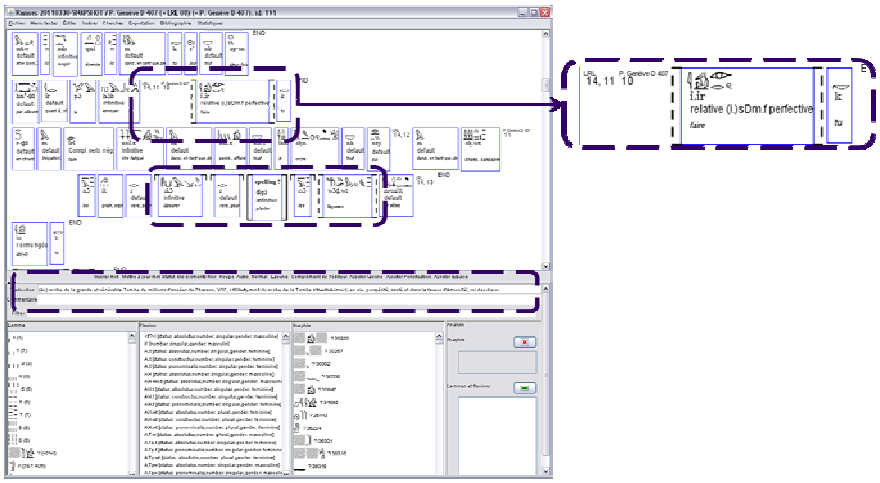


Figure 2. TextEditor

Textual criticism is entirely integrated: lacuna, editor's restoration, erasure, etc. are systematically annotated (Fig. 2, box 2).

For each text, there is a double reference system: first according to the document's materiality (e.g. r° 1,2), second following the modern edition that has been used; a marked preference has been given to well-known collections of texts like *LRL*, *LES*, *KRI*, *LEM*, etc. A translation in French or sometimes in English (depending on the annotator's first language) is provided at the bottom of the main window; it is aligned sentence by sentence (Fig. 2, box 3).

The three lists at the bottom of the screen (see Fig. 2) contain all the lexemes, inflexions and spellings already recorded in the database. Thanks to basic statistical functions, filters help the encoders to find the adequate analysis in context when annotating new occurrences. The result appears in the last box on the right.

Those lists are connected to the data encoded in the **LexiconEditor**. Fig. 3 illustrates what is displayed in this module when the verb *irỉ* "to do" has been selected. Within the central window, all the spellings that have been encoded so far for this verb are displayed: 257 different spellings are stored so far in the database for this very common verb.

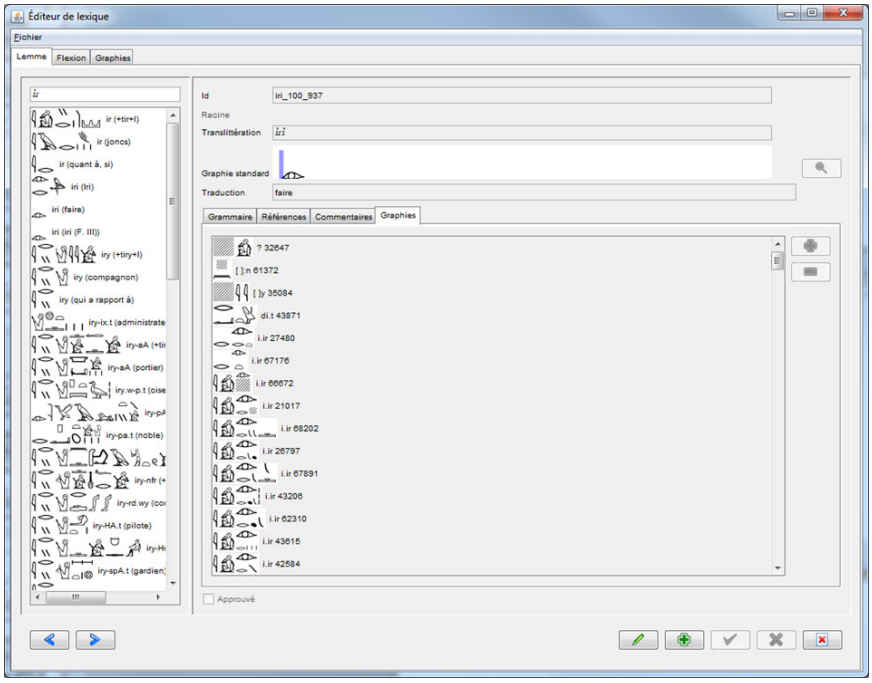


Figure 3. LexiconEditor

Once a lemma has been selected, it is possible to visualize the inflexions that have been linked to it (Fig. 4). If the user picks one of them (e.g. the emphatic form *i.ir=f*), the spellings that have been annotated for this particular inflexion appear in the main window. It is worth noticing that one can also proceed the other way around — which can prove to be extremely useful: starting from a particular spelling, it is possible to visualize all the inflexions that have been linked to it in the corpus.

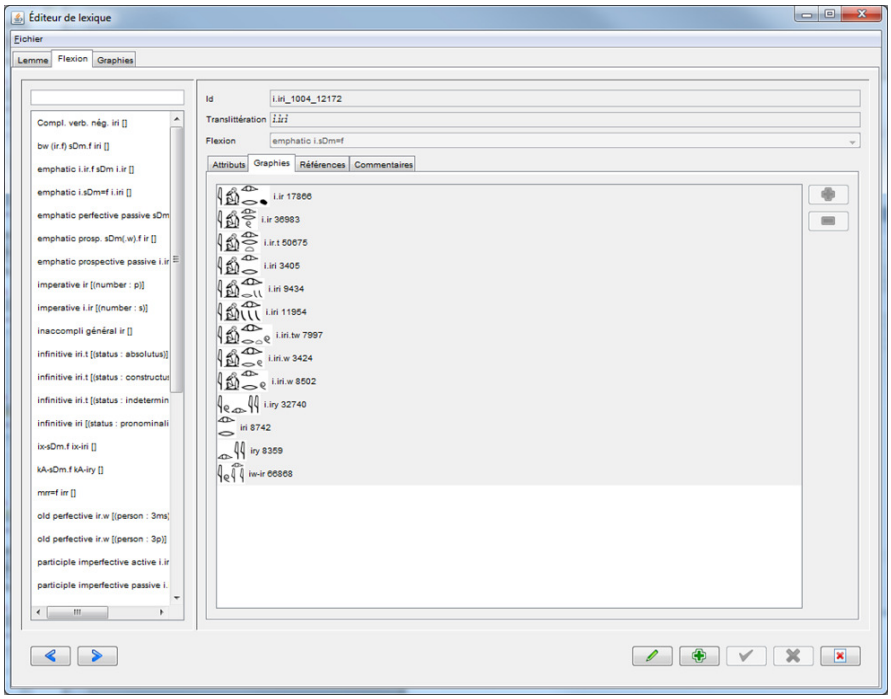


Figure 4. LexiconEditor – Spellings of the "Emphatic (i.)sḏm=f" inflexion

In order to create a new hieroglyphic spelling, a special module has been designed, the **HieroEditor** (Fig. 5), an offspring of Serge Rosmorduc's JSESH hieroglyph editor,[4] that basically works along the principles of the *Manuel de Codage* (with slight modifications and additions).

[4] See http://jsesh.qenherkhopeshef.org.

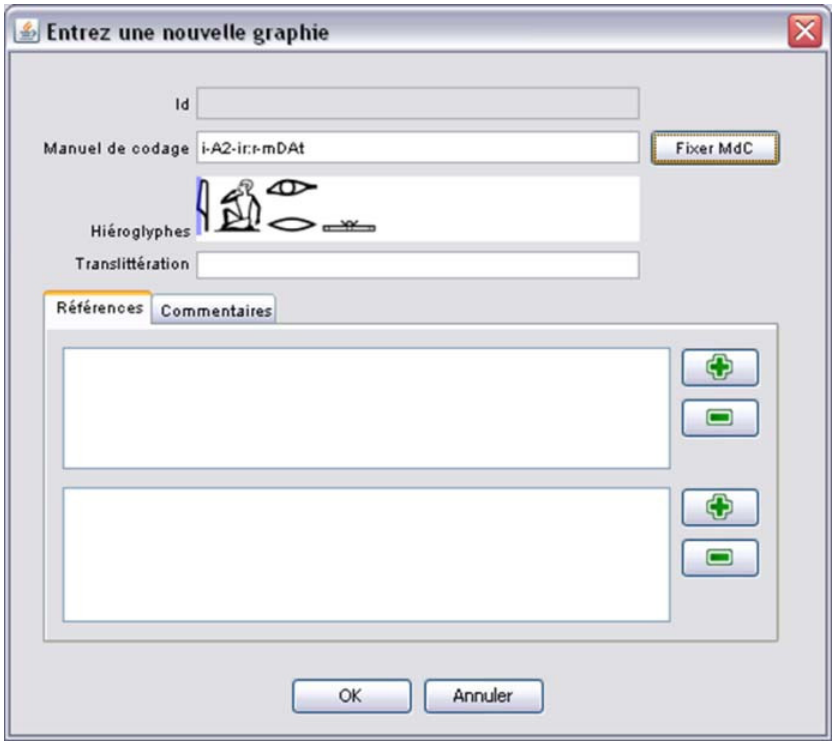


Figure 5. HieroEditor

The **SyntaxEditor** is still under development, but is already in a test-phase.[5] It capitalizes on the data annotated in the TextEditor, and makes them fully available when one performs syntactic annotation.

The functionalities of the SyntaxEditor have been developed in order to allow not only phrasal chunking (supporting discontinuous constituents, as in the simplified Ex. of Fig. 6) and full syntactic analysis of a sentence, but also in order to annotate other dimensions of linguistic analysis like anaphoric relations (field of textual cohesion,

[5] For a complete description of the SyntaxEditor (specifications that are implemented, syntactic formalism, representation format and annotation scheme), see POLIS, S. & S. ROSMORDUC, 2013: Building a construction-based treebank of Late Egyptian: The syntactic layer in Ramsès, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 45-59.

e.g. via the co-indexation of pronouns and noun phrases) and information structure as well as speech acts.

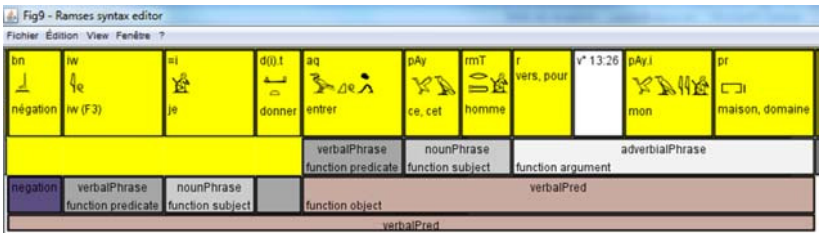


Figure 6. SyntaxEditor

The annotation scheme, which defines the valid types of syntactic annotations as well as the possible set of functions, construction by construction, is a priori neither framed in a constituent structure nor in a dependency-based formalism: we see these representations as two different possible outputs of a single 'construction-based' annotation scheme. This approach — close to the one developed in Potsdam university for the TIGER Treebank[6] — has been developed in order to account for the diversity of linguistic facts found in the Late Egyptian corpus. It is much in agreement with the grammatical tradition in Egyptology, which endorsed a *construction grammar* perspective *avant la lettre* by systematically taking into consideration different grammatical *patterns*.

This perspective takes seriously the assumption that *constructions* are the basic units of any syntactic representation. Accordingly, we consider as a real possibility that the syntactic annotation will lead to generalizations concerning elements across constructions that are not congruent with the pre-existing categorization (e.g. parts-of-speech that are encoded for each lemma in the LexiconEditor). This means that the syntactic annotation will most certainly have feed-back effects on the previous analyses, thereby avoiding the methodologically untenable position of defining a priori categories such as part-of-speech, etc.

From an IT point of view, the TextEditor and the SyntaxEditor will eventually merge into a single JAVA module with visualization facil-

[6] BRANTS, S. *et al.*, 2002: The TIGER Treebank, in: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 24-41.

ities that will enable the annotators to select the level of linguistic analysis they wish to have access to.[7]

1.2.2 Appending metadata to the corpus

The annotation of the linguistic material would be virtually useless without metadata. These are recorded in Ramses with the help of two main modules: the **TextDocumentEditor** and the **BibEditor**.

The annotated texts are identified and described in the **TextDocumentEditor** (Fig. 7). Texts and documents as material objects must be carefully distinguished.[8] In most of the cases, the two categories overlap, but a text is sometimes preserved on many documents (this is of course mostly the case with literary and religious texts; all the parallel versions of a given text are annotated — and will be later aligned — in Ramses), and a single document can also contain more than one text, as is the case with anthologies, for instance.

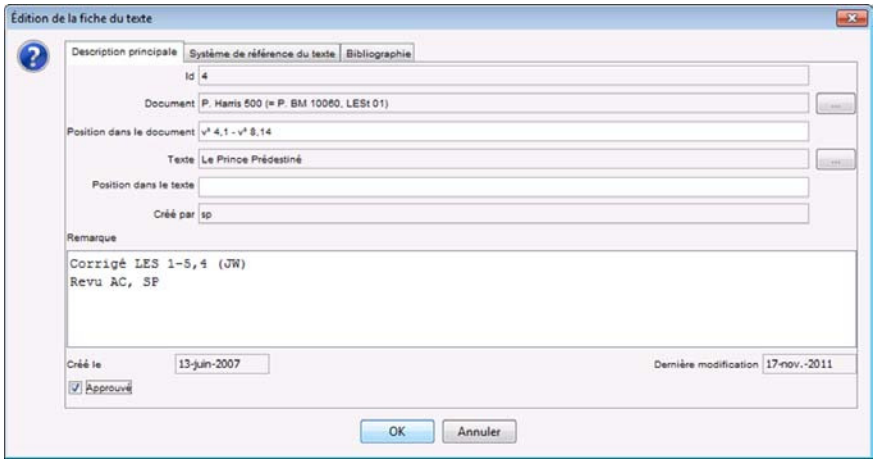


Figure 7. TextDocumentEditor

[7] The SearchEngine for the syntactic layer is not implemented yet. We currently investigate the possibility of using *Annis2* (see http://www.sfb632.uni-potsdam.de/annis/), "an open source, versatile web browser-based search and visualization architecture for complex multilevel corpora with diverse types of annotation."

[8] Cf. the distinction between object and text in the *Thesaurus Linguae Aegyptiae*.

Metadata such as date, provenance, writing system, writing support, language sub-categorization, textual genre, are based on hierarchical thesauri that match recognized standards such as the *Multilingual Egyptological Thesaurus*[9] whenever possible.

Furthermore, modern literature can be appended selectively to different levels of annotation (see Fig. 8) in order to justify the choices and interpretations made by annotators.

Complete references are first encoded in a specialized **BibEditor**. They are then linked, with the appropriate pagination and tags specifying their content, to different objects of the database. The following screen shot (Fig. 8) shows how bibliographical references are instantiated in the LexiconEditor for the lemma *ỉb* "heart" (especially noteworthy are the hyperlinks to other electronic resources such as the *Thesaurus Linguae Aegyptiae*[10], *The Deir el-Medina Database*[11], *Deir el Medine Online*[12] and the *Online Egyptological Bibliography*[13]).

[9] See VAN DER PLAS, D. (ed.), 1996: *Multilingual Egyptological Thesaurus*, Publications Interuniversitaires de Recherches Égyptologiques Informatisées 11, Utrecht/Paris.

[10] See http://aaew.bbaw.de/tla/.

[11] See http://www.leidenuniv.nl/nino/dmd/dmd.html.

[12] See http://dem-online.gwi.uni-muenchen.de/.

[13] See http://oeb.griffith.ox.ac.uk/.

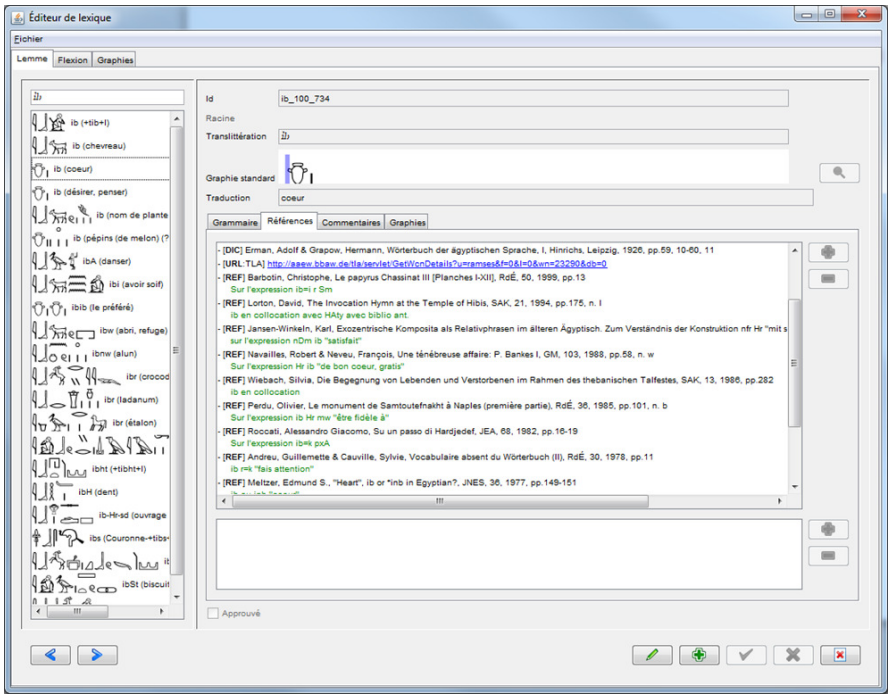


Figure 8. BibEditor

1.2.3 The SearchEngine

A database, however rich and complete, is useless without a powerful system for retrieving the relevant information. As noted above (see §1.1), the SearchEngine has been designed to run ideally any type of queries, without limitation regarding the types of annotations or metadata that can be searched for simultaneously.

Queries can be made on the whole corpus or on sub-corpora by using filters on genres, date, provenance, writing support, writing system, and so on. Fig. 9 shows how one can build a query on a sub-corpus containing all the letters that have been written on ostraca and come from the village of Deir el-Medina.

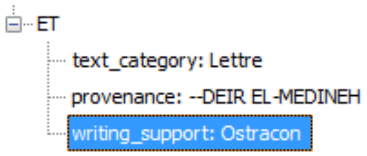


Figure 9. Defining a subcorpus

Any query is built step by step. One specifies successively the layers of annotation that are searched for and the context that will be taken into consideration. In the following example (Fig. 10), the search aims at finding fronted relative clauses that are introduced by *ir* and whose predicate is a verb that has both the infinitive as inflexion and the moving-legs as classifier. The skip operator (*) that appears twice in this example means that unspecified words are allowed between two elements of the query. If needed, the number of these unaccounted blocks can be more or less strictly specified (exactly 3 occurrences or between 1 and 4 occurrences).

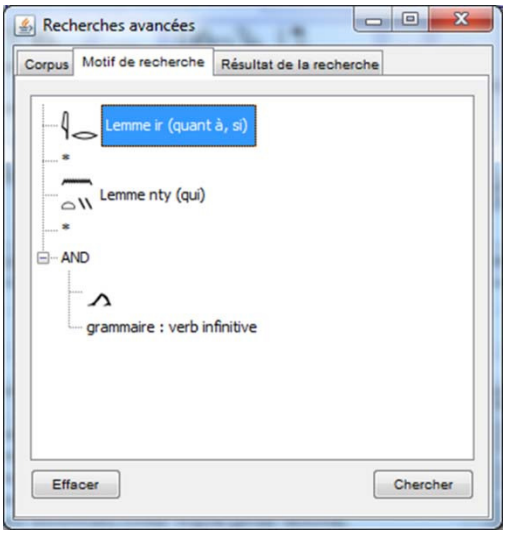


Figure 10. SearchEngine

The results can of course be visualized in a table format. Each line of the results is linked to the TextEditor, so that the end-user can easily access a wider and fuller context with the relevant bibliography, if any.

Finally, the data can be exported in `.pdf`, `.html` or `.gly` file format. All levels of annotation can be exported at once, but it is also possible to select specific data to be exported. In the first example (Fig. 11a), all the data have been exported in `.pdf` format; it should be noted that interlinear grammatical glosses are produced automatically, based on the annotated data. The second example (Fig. 11b) illustrates a lighter option: the hieroglyphic line has been exported in `.gly` format,[14] without the lexical and grammatical tagging.[15]

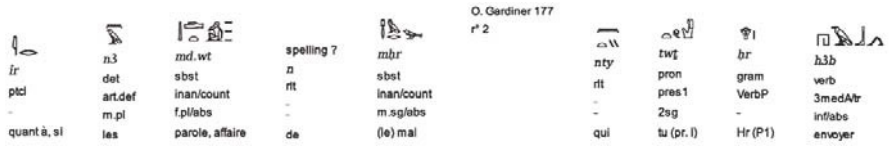


Figure 11a. Export Tool (a sentence in `.pdf` format)

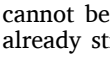
O. Ashmolean Museum 0177 (= O. Gardiner 0177); id: 1404

Figure 11b. Export Tools (same sentence in `.gly` format)

2. Building an annotated corpus: Methodology and current state

In this section, the current state of the annotation process is reviewed with a particular emphasis on the way an annotated corpus like Ramses is actually built. In the first part, we focus on the methodological principles at stake when annotating texts in the corpus and we show how software developments have been used in the fight against time, probably enemy number one in the lengthy task of

[14] The `.gly` export format has recently been implemented by Serge Rosmorduc. It proves to be an especially useful and time-saving tool when data coming from Ramses are used in a later written production.

[15] As the hieroglyphic line is composed automatically by juxtaposing the coding of the individual blocks, the relative position of the signs at the border of two words cannot be accounted for. A sequence like  will appear as . As already stressed, Ramses is primarily a research tool with a clear orientation to questions related to grammar and linguistics; it will never substitute for a sound philological edition nor for a photograph.

building a corpus. In the second part, we comment upon figures summarizing the progress made so far in the encoding and in the annotation of the textual data. Finally, future prospects are outlined in the third part.

2.1 Software ergonomics

As a manually annotated corpus, Ramses had to meet one requirement of paramount importance from the annotator's point of view: the editing software had to be user-friendly so as to meet the criteria of speed (and ideally consistency) of annotation.

In order to meet this requirement, three interrelated JAVA modules (see §1.2.1) have been designed for handling the graphemic, morphological, syntactic and textual levels: a TextEditor, a LexiconEditor and a SyntaxEditor. We will focus here on the relationship between the first two modules when annotating a text.

The goal was to save annotators from reduplicating work by implementing fully the capabilities of relational databases. Therefore, the following principle has been adopted: each occurrence of a word in a text (TextEditor level) is the actuation of a detailed entry in the lexicon (LexiconEditor level).

In Fig. 12, for instance, the verb *gmḥ* in one sentence from the *Doomed Prince*, is an actuation of the lemma *gmḥ* in the LexiconEditor (on the left) and of the inflexion /infinitive_StatusConstructus/ (on the right).

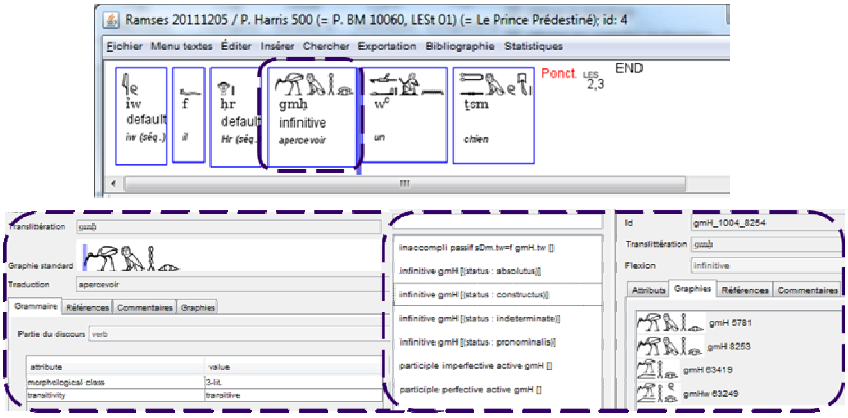


Figure 12. Link between the TextEditor and the LexiconEditor

When encoding a text in the TextEditor, the annotator simply has to select the lemma, the inflexion and the spelling from lists (see Fig. 2) that are fed by the LexiconEditor and sorted according to basic statistics automatically generated about the existing corpus.[16] If any lemma, inflection or spelling is missing, these lists are supplemented by adding new information in the LexiconEditor.

The encoding of texts was obviously quite slow at the beginning of the project (given that every single new occurrence had to be fully encoded in the LexiconEditor), but as the corpus was growing and the data in the LexiconEditor correlatively expanding, the annotator's work became correlatively faster: annotators never have to encode the same data twice. At every single level — from inflexions to spellings, from bibliographical references to documents and texts — data are encoded only once and they are directly available and easily accessible for the all the annotators working on the database.

2.2 Progress in the annotation

Whatever the quality of the tools developed for facilitating the encoding, Ramses remains a manually annotated corpus, which means that the integration of new texts in the database is time consuming.

Besides software developments, an additional strategy has been devised in order to speed up the process of annotation (and hopefully to increase its consistency): since Late Egyptian written registers are highly diverse — in terms of lexicon, phraseology, distribution of inflectional patterns, etc. — the whole Late Egyptian corpus has been split up into sub-corpora according to text genres and chronological periods. Each annotator working in the project is responsible for the annotation of a particular *Textsorte*.

Currently,[17] 1744 texts have been worked out in the database and received multifaceted annotations, which amounts to a little more than 334,000 tokens or words. As shown by Fig. 13a-b, the progress made in the encoding is quite regular. The last two years even testify a slight increase of the number of new words annually annotated in

[16] We are currently developing a context-sensitive semi-automatic tagger that suggests the lemma, inflection and spellings that are the most likely to be accurate for a word (taking into account mark-up data such as the genre, date and support of any new text). This tool should significantly enhance the speed of annotation.

[17] The statistics provided below have been produced on 2012/09/15, which means that the figures for 2012 are not complete yet.

the database, which has resulted from capitalization on the strong base of a well-stocked LexiconEditor.

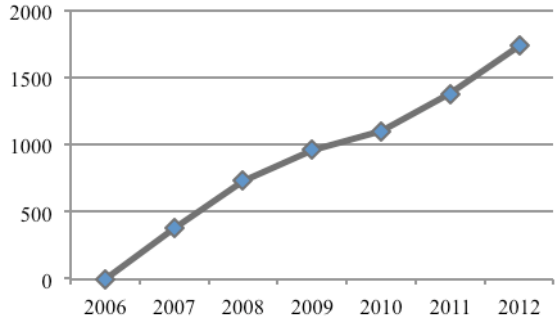


Figure 13a. Number of texts

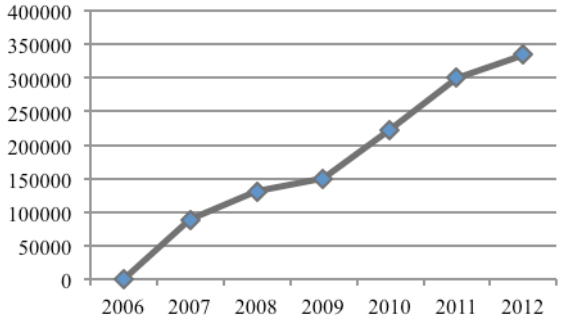


Figure 13b. Number of tokens

Fig. 14 shows the distribution according to genre of the documents written in hieratic script that are encoded and annotated (and the number of documents that await further treatment).[18]

[18] Additionally, more than 400 monumental texts in hieroglyphic script have already been annotated; they represent (a) a selection of 18th dynasty texts whose registers attest evolutionary grammatical features of Late Egyptian; (b) the whole corpus of Ramesside legal decrees; (c) monumental literary texts, like *The Battle of Qadesh*; (d) ideological narratives and rhetorical texts, like the ones of the Medinet Habu inscriptions of Ramses III.

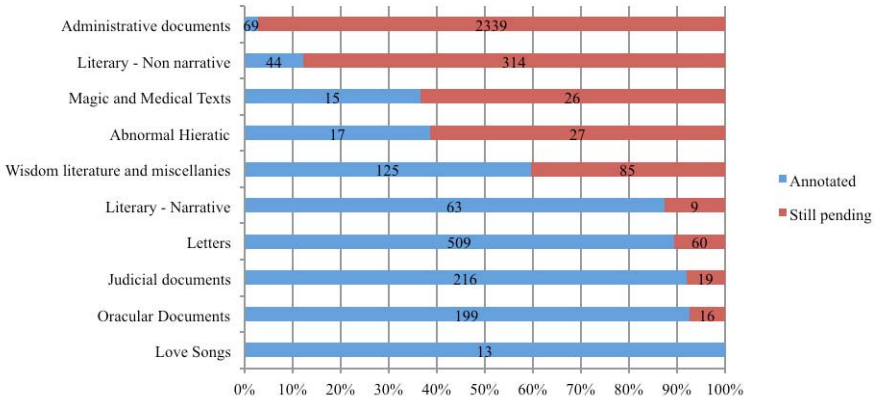


Figure 14. The distribution of hieratic texts according to genres (annotated vs pending)

Given that Ramses is aimed first and foremost at linguistic searches, this figure hardly represents the actual state of the database. Three remarks are warranted here:

(1) Documents deemed more relevant for linguistic analysis have been given high priority. This partially explains the uneven distribution, particularly the small number of administrative documents that have been included in the database up until now.

(2) From the beginning, a deliberate emphasis has been put on the integration of standard editions that contain texts considered to be representative of Late Egyptian: all the texts belonging to the standard collections of texts, such as *LEM*, *LES*, *LRL*, *LRLC*, *RAD*, *TR*, etc. have been completely encoded and annotated.

(3) The length of the documents is highly variable, even within one category. In the category "Wisdom literature and Miscellanies", for examples, among the 85 documents that are still missing, more than 40 are parallel versions on ostraca of the P. Anastasi 1: the longer and/or better preserved documents have been preferred in the first phase of annotation.

Fig. 15a-c show the evolution of the number of lemmata, inflexions, and spellings recorded in the database between 2006 and 2012.

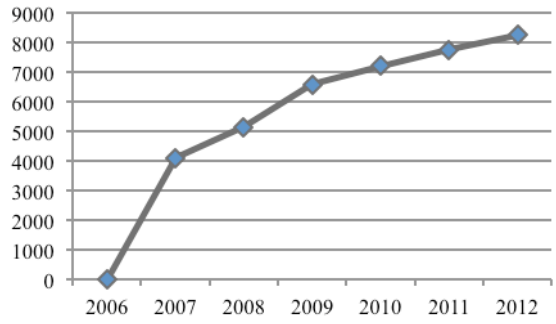


Figure 15a. Number of lemmata

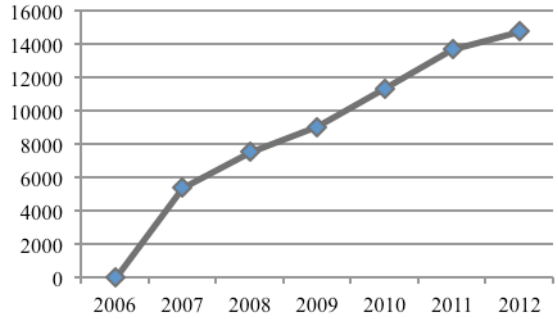


Figure 15b. Number of inflexions

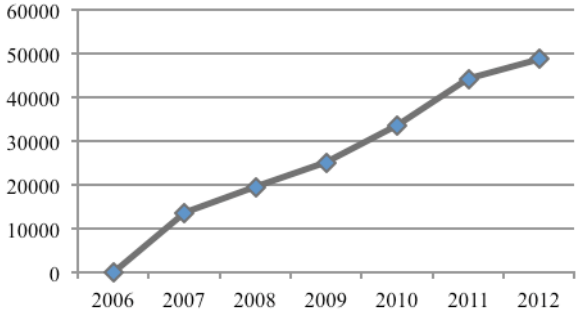


Figure 15c. Number of spellings

As shown in Fig. 15a, the number of lemmata grew quickly during the first year of the project; this results from the fact that the only

dictionary available for Late Egyptian[19] was entirely encoded in the LexiconEditor at the beginning of the project in order to speed up the encoding of the first texts. Otherwise, the progression is regular for the number of inflections and spellings: a bit counter-intuitively, each new text keeps on adding with the same ratio new inflexions and spellings to the database.

2.3 Future perspectives

Before termination of the first phase of the project in October 2013, we will focus on several aspects of Ramses that deserve further attention:

(1) Completion of the encoding and of the annotation of the sub-corpora that we began integrating in Ramses, with a particular focus on the non-narrative literary texts, on the administrative texts and on the texts of the Third Intermediate Period (including the texts written in so-called "abnormal hieratic" or "Kursiv-hieratisch").

(2) New implementations in the TextEditor and SyntaxEditor (ultimately to be merged in a single RamsesEditor). This crucially includes the possibility of defining different levels of access to Ramses (in order to preserve the integrity of the validated data) and a storage of the "history" of successive annotations (when, how and by whom was the annotation carried out? who modified it and when?).

(3) Development of a Web application so as to give the community of Egyptologists and linguists access to the whole range of Ramses data.[20]

Long-term projects include:

(1) The completion of the syntactic annotation of the corpus and the addition of a semantic level of annotation (with word-sense disambiguation).

---

[19] LESKO, L. H., [2]2002-2004: *A Dictionary of Late Egyptian*, 2 vol., 2nd ed., Providence.

[20] We plan to publish the sub-corpora online one after another, immediately after final approval by the team. The end-users will be able to contribute to the enrichment of the corpus thanks to a wiki-like device that will be added in order to allow suggestions regarding the hieroglyphic readings, the addition or emendation of annotations, etc.

(2) The continuation of existing (and development of new) collaborations, e.g. with TXM concerning statistic tools,[21] with the *Thesaurus Linguae Aegyptiae* (see n. 10) in the field of Egyptian lexicography, with the Deir el-Medina Database (see n. 11) regarding the metadata on Late Egyptian texts.

(3) The extension of Ramses' functionalities in order to support earlier and later stages of the Egyptian language, down to Coptic.

3. Changes in methods and practices

The use — the massive use in some cases — of annotated corpora will trigger significant changes in Egyptologists' methods and practices. These changes are, on the whole, indisputably for the better. However, using these new tools without a sharp sense of criticism could potentially lead us in dangerous territories. Here follows a quick review of the main pros and cons.

One of the most obvious advantages of using corpora — even if it is a never ending process — is the exhaustiveness of the data. The textual corpus of Ancient Egyptian (and even a limited subcorpus such as Late Egyptian) is now beyond the reach of a single individual. As one can safely anticipate a regular increase of the data, the benefit of an electronic corpus cannot be overemphasized. Indeed, combined with unlimited numbers of queries on different level of annotation, such corpora should produce falsifiable results in Egyptian linguistics, which is admittedly what is expected from any scientific work.[22]

Electronic corpora, however, could easily give the confidence that they are — even intrinsically, so to speak — objective tools, because they record simple and plain facts. But it would actually be dangerous to assume that databases are neutral from a scientific viewpoint: they are modern ways to organize the rough data. In this respect, Ramses is an annotated corpus, extensively enriched and, as it turns out, choices must be made all the time: in some cases, arbitrary choices that can be explained; in some other cases, choices that are

[21] See http://textometrie.ens-lyon.fr/.

[22] A database like Ramses will make it possible to check hypotheses that unavoidably surface in the course of research projects. This point cannot be overestimated. Scholars are used to the frustrating experience of having failed to take a feature into consideration when reading the corpus. One is then left with two options: neglect it (which can quickly become very problematic from a scientific viewpoint) or start reading the corpus again (which inevitably raises practical problems of time).

the result of the developers' conception of how the grammar of Ancient Egyptian works. In the end, the picture that could emerge from the database is a Late Egyptian grammar *à la liégeoise*, maybe not a bad one in itself, but better to be avoided if one intends to reach a wider audience. To steer clear of such bias, we relied on three strategies aimed at producing a descriptive (i.e. theory-neutral) approach to language structure, with no loss of data because of the resulting method of annotation:

- An analytical approach;
- The possibility of encoding ambiguities;
- The possibility of storing unanalyzable chunks of graphemes.

3.1 An analytical approach to encoding

The choice has been made of coding minimal units rather than larger groups. This is apparent, for instance, in the way lexical composita are handled. In the first place, composita like *mr-mšᶜ* "general" were encoded as one lexical unit. This seemed the most natural way to do it, because it was felt to be very close to every Egyptologist's experience.

This option, however, quickly turned out to be problematic, when less common phrases were to be treated: for instance, is *mr-mšᶜ(-)wr* "general in chief" to be analyzed as a compositum or as two phrases? If *rmṯ-ỉs.t* "crewman" can be safely assumed to be one unit, could this be equally valid for any group with *rmṯ* as its first element? Coptic at first sight seemed to give clear indications,[23] but this turns out to be an illusion. Above all, having large composita would probably have hampered the flexibility of later queries: it would have been impossible, for instance, to look for all the titles containing the qualifier *wr* "great".

Therefore, the decision was finally taken to encode the texts word by word in the TextEditor and to create larger groups with the SyntaxEditor, even in cases where it can safely be assumed that one is dealing with a compositum.

3.2 Encoding ambiguities

Our goal in allowing for the encoding of ambiguities was to lose no piece of information that could be relevant for a query in the corpus.

[23] The Coptic data show that there exist composita built on ⲣⲙ̄- and ⲣⲙ̄ⲙ̄-, which at least suggests different chronological strata.

Accordingly, ambiguities can be encoded at three levels in Ramses (lemma, inflexion, syntactic analysis, and any combination thereof).

Most ambiguities relate to poorly understood contexts, often due to the presence of lacunae. For instance, it is not at all always clear whether a verb is to be analyzed as a perfective or a subjunctive *sḏm=f*. Fig. 16 is an illustration of such a case of morphological ambiguity: in the box of the occurrence, instead of having one analysis, the label <AMBIGUOUS> appears. The two possibilities are recorded in the status line of the word that is displayed at the bottom of the screen (right of Fig. 16). The text can of course been retrieved in any query involving either a perfective or a subjunctive.

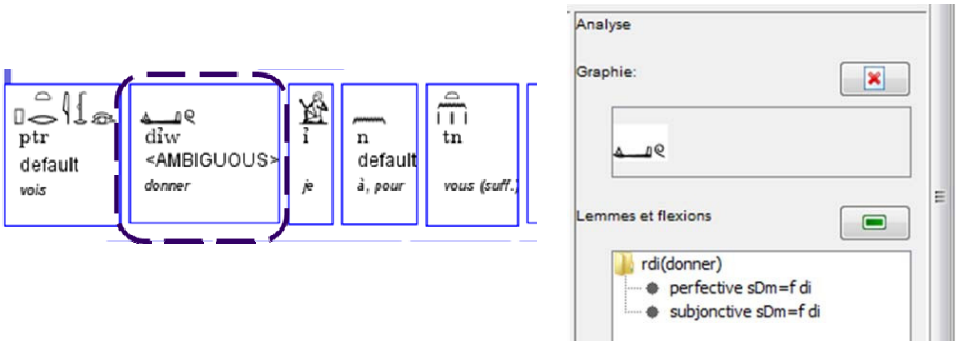


Figure 16. Ambiguities (type 1)

The next example shows another type of ambiguity combining lexical and morphological possibilities (Fig. 17). Due to the fragmentary state of the text, the word *bȝk* can be understood either as a noun "the work" or as verb "to work". According to the option that will be chosen, the morphological analysis has to be adapted.

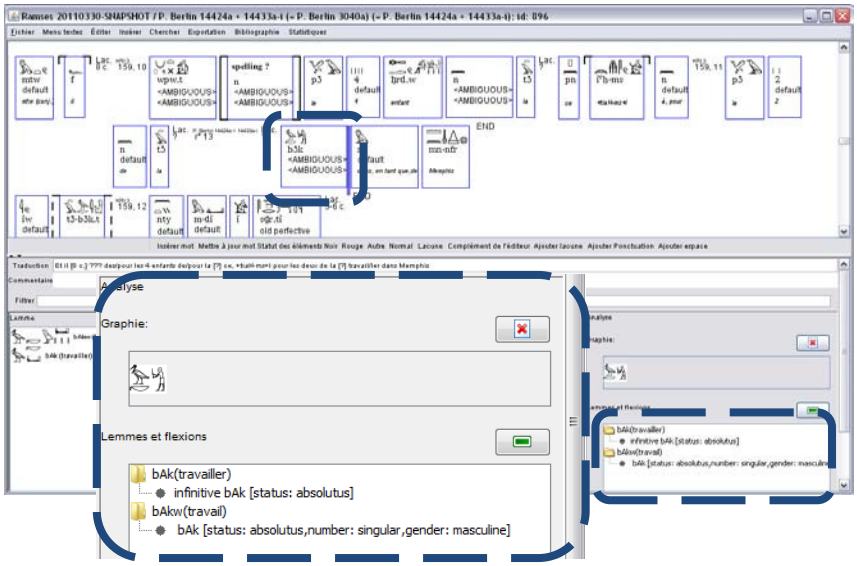


Figure 17. Ambiguities (type 2)

3.3 Encoding unanalyzable sequences of signs

Ramses also makes it possible to encode hieroglyphic signs without linking them to a lemma. This option is of course maximally used in case of lacunae. In doing so, no sign — even if it is completely isolated — is left along the road (Fig. 18).

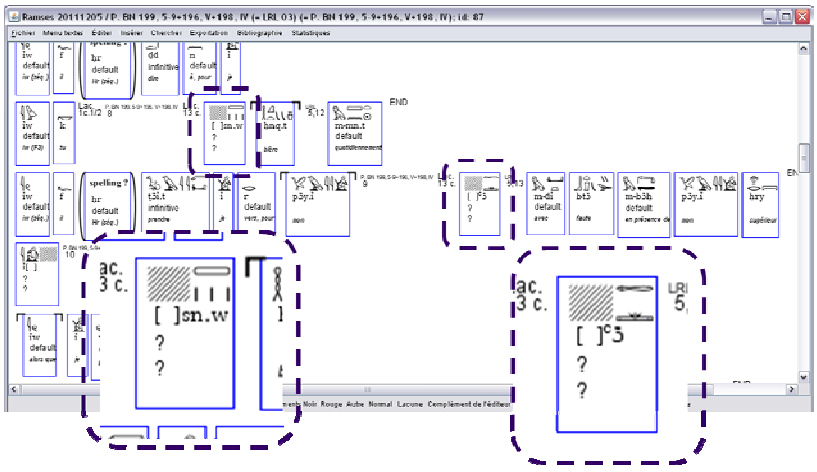


Figure 18. Unanalyzed chunk of graphemes

3.4 Some general reflections

The use of an annotated corpus for data mining seems to offer all possible advantages: it is exhaustive, quick and systematic. But one has to refrain constantly from being naive in the use one makes of an annotated corpus.

First, as remarked above (§3.1), information stored and annotated in the database are never simple facts, directly imported from a supposed objective realm; they always undergo processes of standardization. Second, the extensive — if not sole — use of electronic corpora might entail the risk of discouraging people from developing basic philological skills. There is indisputably some virtue in the old-fashioned habit of reading through whole texts; the exploitation of large corpora solely by means of search engines, even sophisticated ones, usually brings with it a lot of drawbacks, as has become clear for those of us who are accustomed to certain types of typological literature.

Before proceeding to conclusions for this paper, it should be briefly (but plainly) stated what an annotated corpus like Ramses is not, is not yet, and will never be:

(1) Ramses is not a substitute for traditional philological editions. Not only are *facsimile* representations and photographs lacking,[24] but information regarding textual criticism has been kept to a minimum.

(2) Ramses will probably never integrate the vast body of secondary literature that has been written on the texts. In other words, it will never exempt scholars from going back to the secondary literature. As a matter of fact, bibliographical references aim at justifying choices in the annotation (§1.2.2), not at collecting all possible references.

(3) Ramses will never be a substitute — in this case a very bad one to be sure — for a grammar or a dictionary of Late Egyptian. This paper is not the proper place to discuss the all-important issue of lexicographical tools in Egyptology. Some time ago, the *Wörterbuch* team decided that they would not engage a new version of the *Wörterbuch*, but that they would instead provide scholars with an electronic thesaurus, the *Thesaurus Linguae Aegyptiae* (see

[24] From a technical point of view, this issue can be very easily addressed, but problems regarding the copyrights and credits for the pictures are still to be dealt with.

n. 10). As a consequence, it is now up to everyone to write his/her own lexicographical notes based on the data of the *TLA*, which is a complete change of paradigm. On the contrary, in our eyes, Egyptologists need a proper and modern dictionary. A new dictionary of Late Egyptian is thus one of the major achievements that could be produced with the help of Ramses, but the output will clearly be outside the scope of the Ramses project.[25]

4. Conclusions: New anvenues for research

Notwithstanding the foregoing observations, annotated corpora like the *TLA* or *Ramses* will bring significant positive changes in the study of the Ancient Egyptian language(s) and texts.

The SearchEngine under development in the framework of Ramses will indeed not only make queries far easier to execute than ever before, but — above all — it will allow queries that could not have been previously achieved on account of the high degree of complexity and/or the size of the corpus to be examined. By way of conclusion, we will point to some research domains that were, on the one hand, already accessible with traditional tools but that can now be approached faster, more systematically, and more exhaustively and, on the other hand, new avenues for research that were simply impossible to pave without such richly annotated corpora.

In the sphere of traditional philology, a corpus like Ramses could help considerably in taking up the challenge of the identification and grouping of hundreds of pieces of literary texts on ostraca that are scattered in collections and museums all over the world. If one is faced, for example, with the sequence of graphemes  , the identification of the text (even if well-known) is a long and uncertain endeavor. A simple query in Ramses — that is built according to the most probable segmentation for this sequence of graphemes (see Fig. 19) — gives two results, both from copies of the *Teaching of Amennakhte*.

[25] This issue has been discussed by Winand in a conference held in Leipzig in November 2012.

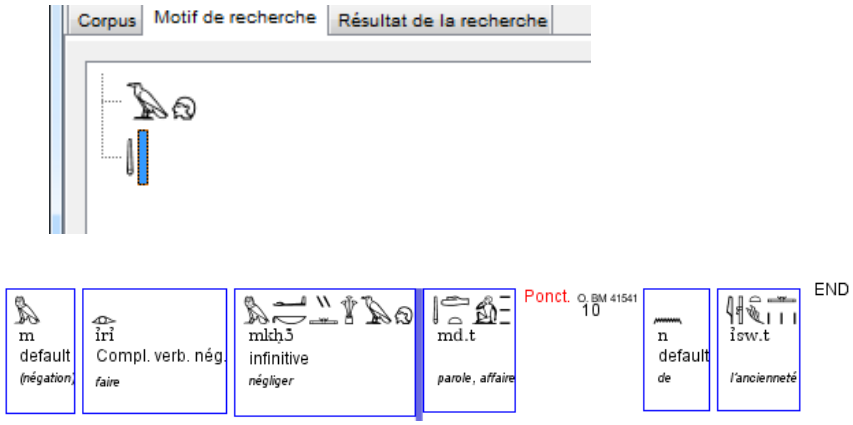


Figure 19. The identification of literary fragments

In the domain of graphemics, researches on the system of classifiers can be broached much more easily. For instance, listing all the lexemes that can have the 𓄿 classifier was a long, fastidious and possibly non exhaustive task without an annotated corpus, while Ramses produces a list of 53 related lexemes in the corpus instantaneously. More problematic would be studies that involve the combination of the graphemic level with other level of analysis. One can think, for example, of the combination of the divine classifier 𓇋 with pronominal elements. A query like the one of Fig. 20 gives directly 1358 matches that can be sorted according to any kind of criterion.

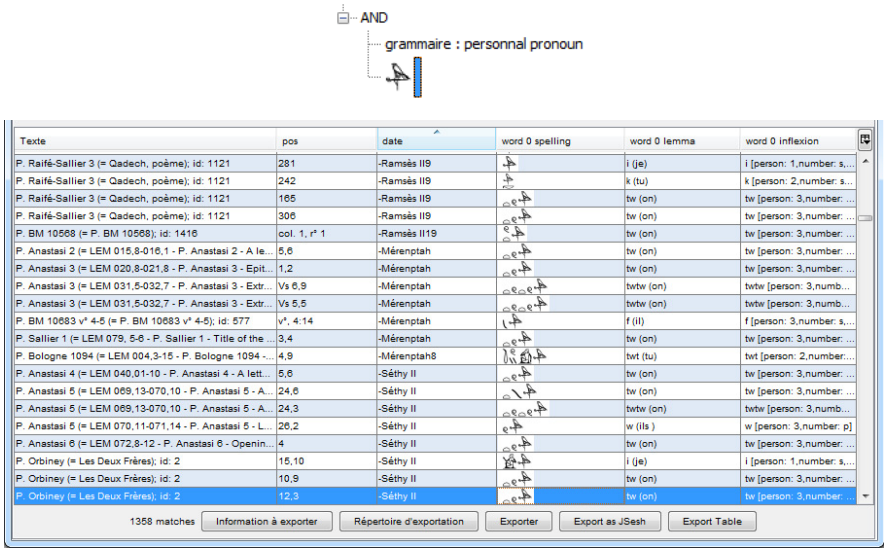


Figure 20. Research on classifiers

If the benefits of an annotated corpus in the field of morphology and syntax are obvious, it should be stressed that the use of semantic information that are stored in the LexiconEditor in combination with morphosyntactic features opens new opportunities for checking hypotheses, e.g. about (the evolution of) the selectional restrictions of constructions. The query of Fig. 21 in the TextEditor, for instance, allows one to find all the occurrences of Future III with inanimate subjects.

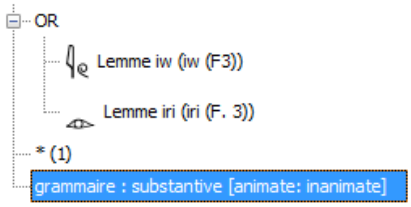


Figure 21. Types of subject with the Future III: Inanimate subjects

Finally, Natural Language Processing — an entirely new field for Egyptology — will be explored in close collaboration with computer scientists. The first applications that come to mind are: the develop-

ment of taggers and parsers, automatically generated indexes and concordances, the application of methods for automatic text categorization (e.g. with decision trees) and information retrieval, as well as advanced statistical tools.

THE RAMSES PROJECT IN PERSPECTIVE
MANAGING EVOLVING LINGUISTIC DATA

SERGE ROSMORDUC

1. Introduction

Back in 2006, when we started the Ramsès Project [WINAND *et al.* (2008), ROSMORDUC *et al.* (2008)], our primary concern was to develop an efficient tool for data input, and to deliver it to the encoding team as soon as possible. We had to work on a tight schedule, and, keeping in mind ease of use, and linguistic coherence as prime requisites, we did not have time to handle fully the problem of the evolution of our data and of our own view of the data.

Needless to say, we have paid the price for this a number of times, and, in more than one case, we had to perform large-scale modifycations of the data, which usually required both automated and manual processing. Most of the time, those modifications involved lemma or parts of speech modifications, for which our system is rather versatile. In a few instances, we did modify the very structure of our lexicon (that is, the way the database tables are organized). Even if, in retrospect, no major problem occurred, the process was quite taxing.

As the initial phase of development of Ramsès is almost done, with a working prototype of a syntactic editor, we have started to think about ways of improving the encoding process, and securing our data consistency. This paper explains the current state of our ideas on the subject.

2. A short typology of changes in Ramsès

We will start by looking how data evolves in the "Ramsès" ecosystem, and try to roughly classify those changes. Doing so will probably amount to beating a dead horse, as most similar databases will have met the same problems.

Anyway, a short description of the organisation of the Ramsès database is necessary to explain the extent of our problems. In Ramsès, the structure for representing the texts is quite complex. First, each entry is documented both in terms of content (a text, which might be known from multiple copies, and which has a number of characteristics, such as a particular blend of language

(from Middle Egyptian to full Late Egyptian) or genre of text. The entries are also documented in terms of original documents, which have they own characteristics (date, writing support or script for instance).

The content of the text is a sequence of lemma analysis, which record word spelling, lemma and inflexions. Actually, the spellings, lemma and inflexions are recorded in a lexicon, and a text content is a list of references to the said lexicon.

In this respect, Ramsès is quite different from non-lemmatised text databases. In a non lemmatised database, the information in a particular entry is relatively stand-alone, and quite impervious to changes in the rest of the base. In Ramsès, the data in a given entry depends heavily on the lexical database.

This being said, we can now categorise changes in Ramsès in two families. The first kind of change is due to modifications of data, which can occur at multiple points. This is the most mundane type of changes, although, in the current state of the database, it can have huge consequences in some cases.

The second and most drastic change is structural. In some cases, one can consider modifying the structure of the database itself.

2.1 Data Modification

Data modification can alter the texts, the lexicon, and, in Ramsès, even the lexicon structure. In normal edition mode, the Ramsès encoders will create and modify text entries routinely. They will also need to enter new words, new spellings and new inflexions, thus modifying the lexicon. Let's consider a few examples, and the problems they incur.

The first example is *simple text modification*. An encoder can modify an existing text, adding, deleting or replacing part of its content. Our current system does not remember the modifications made to a text and, thus, present only the latest stage of editing. However, it is quite desirable to keep a history, for a variety of reasons, especially if text edition is in part collaborative. It can allow to reverse bad manipulations, like accidental erasures of parts of the text.

A second type of modifications is *lexical modification*. In Ramsès, all encoders can add, delete, or modify lemmas, inflexions and spellings in the lexicon. Obviously, a change in a given lemma (for instance, a change in its transliteration) will have large consequences on *all* the texts which use it. It is quite important to be able to review

those changes and to be able to cancel them if needed. Even the creation of a lemma can be a problem, for the said lemma can be already encoded (especially if we consider the variability of Late Egyptian orthography).

Finally, a large number of features of the lexicon are encoded as data, and can be modified by the administrators of the system. For instance, parts of speech, the inflexions associated with them, and their attributes can be modified (which would allow, for instance, to use the database to create a purely Middle Egyptian database). Those modifications are quite critical, for they potentially affect, directly or indirectly, large segments of the database. Let's take a few examples.

Creating a new part of speech, adding information to existing parts of speeches or inflexions is not a simple addition. It can affect encoded texts, in that their encoding might become partly obsolete.

For instance, if the database contains only one form for the infinitive, and if we decide to distinguish *status absolutus*, *status constructus* and *status pronominalis*, and introduce this new distinction in the lexicon, all texts encoded prior to the addition will become obsolete and need reviewing (or, at the very least, will need to be specifically marked).

Other modifications can call for more drastic text revisions. For instance, when we added the notion of "phrase determiners", to indicate the determinative which can occasionally appear after a relative clause or a complex noun phrase, a number of texts had already been encoded, in which those determinatives had been somehow artificially related to the last word of the noun phrase. We had to change the segmentation of the texts in a few other cases, in particular when, after many discussions, we decided that most compound words (and in particular titles) were to be analysed when it made sense and not encoded as a unit (*ḥry.w-š* is one word, but *ḥry pḏ.t* is encoded as two units).

2.2 Structural Modifications

In extreme cases, the core structure of the database itself can change. For instance, the spellings were initially dependent of lemma and inflexions. We had an entry "  " as spelling for the circumstantial *jw*, and another entry with the same glyphs for the unusual spelling of the "r" preposition. At some point, it appeared that we really needed to consider spellings as independent entities which could be shared between lemma and inflexions: sometimes there is no doubt upon a spelling, but the corresponding lemma is unclear, for instance. In this

case, we had to change the structure of the database tables, and to correct all the texts. It was however possible to do it through an automated process.

In the near future, a number of important structural changes will probably take place. In our current system, we have used a number of hierarchical thesauri for texts descriptions: dating, writing system, geographical information use a thesaurus. However, the lexeme descriptions do not. In a number of cases, it might be interesting to allow some kind of inheritance between lexical attributes. For instance, substantives have a number of attributes like “animate”, “proper name”, “god”... and, obviously, it would be better to consider “proper name” as a sub-type of “animate”. Inflexions could benefit from the same system: a meta-category “*sḏm=f*”, could be considered as a super-category for perfective and subjunctive Late Egyptian *sḏm=f*, avoiding the current problem in the prologue of many royal texts, where it is often difficult to choose between the two forms.

A last kind of structural change which will take place at some point is that the lemma themselves need to be hierarchically organised, with links to their roots, meaning classification, etc... as it is now the case in the *thesaurus linguae aegyptiae*. When this is done, it is quite likely that parts of the texts will require a revision.

Fortunately, most of those large, structural changes, can be *partially* automated. Actually, we would probably not think of doing them if it wasn’t the case. However, changes made to such an extent mean drastic modifications, and, hence, we have sometimes considered necessary changes with reluctance.

2.3 Temporary Conclusions

To summarise our current discussion, the core of the problem is that our data is precious. It amounts to thousands of hours of work. We don’t want to loose part of it due to an incorrect manipulation. Another point is that, as a collaborative work, we wish it to be coherent. As the world is not perfect, the system must be able to deal with semi obsolete data (if we delete a lemma, we probably don’t want to delete the texts which use it altogether).

We want to know why the database changes and how it changes. This will allow us to cancel erroneous modifications, and to understand systematic errors too.

3. Managing data history

A text database like Ramsès is a collaborative project. It is rather structured, with a validation process for texts, but still, even if it is not a Web 2.0 application, we do have a significant team of encoders, and the decisions of each one is important and impact the others (especially as far as the lexicon is concerned). In this section, we are going to consider the options for monitoring the data changes.

3.1 State of the Art

The first possible solution would be to centralise most decisions. One or two experts would need to approve some operations (like lemma creation) before they are done. However, this would not solve all the problems, as even experts can change their minds, or mis-use the user interface.

Still on an organisational level, some projects have introduce interesting practices: in the *Syntactic Reference Corpus of Medieval French* [PRÉVOST, S. & A. STEIN (2012:6)], a text is encoded twice independently. For each entry where there is a disagreement, a discussion takes place between the two encoders, and, if needed, a senior encoder takes the final decision choice. However, this improve the initial content of the database, but does not constitute a management tool for existing entries.

The Open Richly Annotated cuneiform corpus (http://oracc.museum.upenn.edu), uses the notion of versions of the database. This is quite interesting, because if someone quotes the database for a given result, the quotation can be made obsolete by modifications of the texts. Versioning allows consistent quotation, and might even allow to reproduce searches at a given point in time.

The IDP project (http://papyri.info), which regroups the Duke Databank, the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), and the Advanced Papyrological Information System (APIS), has used systems like *subversion* and *git* [BODARD *et al.* (2011)] for managing their text database history. *Subversion* and *git* have been developed to manage software projects. In large development teams, it is rather usual that a number of programmers modify the same files, and version control systems like these ones save a precise history of modifications and help to keep a consistent view of the code, reverting changes if needed.

A number of other systems can be considered, such as wiki oriented ones, which stores a precise history of each page.

A last possibility, which is not currently used by textual databases (as far as we know), is to keep a precise history, not of the *text content*, as is done explicitly or implicitly by the previous systems, but of *operations made on the database*. Actually, this is what most softwares use in-memory to implement "undo/redo" facilities.

3.2 History Management and the needs of Ramsès

One peculiarity of Ramsès, as compared with the databases listed above, is that it has a very structured lexicon. A text in Ramsès is basically a list of references to its lexicon. The lexicon itself cannot be considered as a text, and the granularity of actions taken on it needs to be very fine. In other words, it will probably be enough to store a number of versions of each text in the base (we might even consider a policy where only a few versions would be kept indefinitely). But, on the other hand, when we come to the lexicon, we must be much more precise. For instance, some operations on the lexicon are much more sensitive than others. Adding a new spelling is something relatively innocuous. On the other hand, removing a lemma from the lexicon has potentially huge consequences on existing texts. Indeed, the importance of a change in the database could be measured by considering how many modifications it would entail in the rest of the data. In a similar way, adding a new attribute to a lemma entry is not as serious as changing its part of speech.

The solution to this problem is to record the *operations* on the database: creation of an entry, modifications of an attribute, deletion or an entry. In theory, starting with an empty set of data, and having the whole history, we would be able to re-create the data as seen at any date. In practice, the database must contain two different parts: on one hand, a large history, recording for each modification its date, its author, and the data needed to perform (and maybe to undo) the modification. On the other hand, a "view" of the current state of the database content is needed for efficient access.

An interesting property of this model is that it allows to examine and eventually to cancel some operations selectively. For instance, attribute changes are not all equal. With a precise list of operations, it is possible to review changes in lemma parts of speech (and only them), and to cancel them *without cancelling the rest of the work*. That is, if a particular change is thought to be erroneous, we don't need to revert the whole database to a state anterior to the change. We can compute precisely what modifications are dependant on the change, and revert them. It can even be done on a rather fine level. For in-

stance, we could allow links between a lemma and a spelling to be done, while cancelling a part of speech change on the lemma.

In the future, we want a better control on user modifications. Most of this control will be done up-front (that is, some modifycations won't be available to all encoders in the first place). However, the history management will allow us to catch even the problems we haven't foreseen, and provide a safer environment even for expert encoders.

Last, but not least, the reversibility of all operations will be a welcome addition when performing large (and semi-automated) changes, alleviating the burden on the software engineer performing the modifications (who has usually been the present author until now).

3.3 The status of Deletion

Among all modifications, deletion has a special status. If we delete a reference to a lemma, spelling or document, what should happen to the texts which use it? Even when the deletion is "correct", if we perform it blindly, we will lose information. Suppose for instance that we had created a "ghost" lemma, which should be merged with others (consider for instance *snj* vs. **sš* in Middle Egyptian dictionaries). If the "ghost" lemma is actually deleted from the database, all references to it in the database texts will be point nowhere, and be difficult to amend. Thus plain deletion is not a good idea.

The solution is simple: we should not actually delete data, but mark it as "obsolete". Obsolete entries will be kept, but it will be impossible to select an obsolete lemma when typing a new text, for instance. Creating a list of obsolete entries and their uses is quite easy, and the database administrators will then be able to rectify all occurrences of the virtually deleted data.

3.4 Final Words on History Management

The only real problem with this approach is that the size of the database might grow a bit too much. It is not sure that such fears are founded, as they depend on the actual behaviour of encoders. Actual data is needed there. In any case, it is possible to imagine some clean-up phases, where some part of the database history will be forgotten.

Finally, a last interesting feature of the process we want to use in the future is that keeping an exact history of what happens is also a way to improve our understanding of the encoding process itself.

Patterns of operations made by encoders can be found, and used to improve the interface. Common errors can be detected more easily and prevented, etc.

4. Managing Structural Change

Keeping a log of changes works well for data modifications. However, at least parts of our database are liable to structural modifycations. How can we deal with them? We can use an age-old technique of computer scientists for this: transform structure into data. This is more or less the approach used by the Notabene system [MAZZIOTTA (2010)].

4.1 Graph databases

In recent years, for reasons quite alien to ancient languages databases, a family of software globally called "NoSQL" databases has become fashionable. SQL relational databases are very efficient and very secure for well structured, well understood data. But they are not very well suited for changing, distributed and semantic data. Typically, keeping information on web pages, managing the data for very large but loosely structured databases (like those used by *tweeter* or *facebook*), is not a task where classical relational databases perform well.

Huge distributed databases are not really an issue for ancient languages projects, as even an exhaustive database of Egyptian would be very small if we compare it with, for instance, a newspaper archive. But the relative freedom they offer in terms of modelling is interesting.

Databases like Neo4J or OrientDB are built on graph theory. The same theory stands behind the semantic web and the *Resource Description Framework* [MANOLA & MILLER (2004)]

In these systems, there are basically two kinds of elements: "*nodes*", and "*links*" between nodes.

If we consider, form instance, inflexions and lemma in our current (relational) system, we have a "lemma" table and an "inflexion" table. Saying that entry 123332 is a lemma, with an inflexion with id 56562 means in our current system that there is an entry with id 123332 in the lemma table, and that the entry with id 56562 in the inflexions table has a foreign key attribute pointing to 123332.

If we ever decide to change the architecture, we need to modify the database table structure, which is a heavy operation. Keeping a history of the operations done would not be simple, either.

Now, with a graph database, all of 123332, 56562, “lemma” and “inflexion” would be nodes. A “is-a” link would be used to specify the class of each node (Figure 1). That is, both data and structure would use the same formalism.

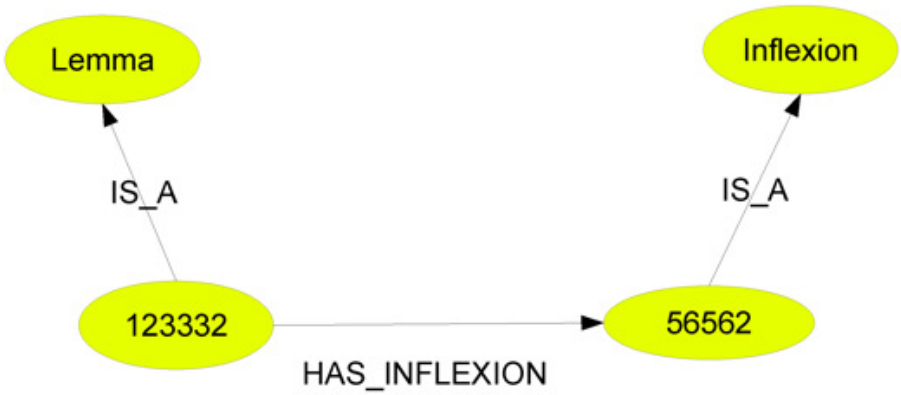


Figure 1: graph database example

The advantage of this representation is that it is very uniform. It is possible to accommodate changing structures with it, and to keep for some time parts of an old model alongside a new one.

Another interesting element is that it can be easily coupled with a history system. There is a small set of possible modifications: node, link, and attribute creation, deletion or modification. By recording them, one gets a history which covers all possibilities on a database.

The downside of such a general system is that it’s more complex to manage and to interrogate than an SQL database. However the current version of Ramsès works with the whole data in-memory, and the database itself is only a convenience to store the information on a permanent basis.

4.2 The Case of Syntax

The Ramsès database is planned to contain ultimately a fully parsed syntactic annotation of all the texts. The problem is that, whereas lexical annotation is relatively stable, our syntactic formalism will certainly evolve a lot. The way we deal with it currently might be seen as a forerunner of the general system to come. We have decided to represent the tagging system for syntax as a “loose grammar”, which states what kinds of phrases are available, and how they can

relate. The grammars are described as simple text, in an *ad-hoc* formalism, which is described in POLIS *et al.* (2013); the idea of keeping an external, user-created and explicit definition of the grammar is taken from N. MOZZIOTA (2010).

The encoder analysing a text needs to choose which grammar he will use, and the system ensures the analysis respects the chosen system.

The system is relatively rich: it's possible to create syntactic categories, to assign them attributes, to state which kinds of groups are fit to fulfil a particular function in a parent group (for instance that the subject is supposed to be a noun phrase). Provision is also made for non hierarchical inter-group links, which can occur for instance between a pronoun and its antecedent, etc.

As an example, the following fragment of grammar describes a simple annotation scheme, which defines four kinds of syntactic groups: *groups*, *noun phrases*, *adverbial phrases*, and *propositions*. A simple inheritance mechanism allows us to factor common attributes. The "group" can bear common attributes, in our present case, comments (which are free text).

The noun phrase has a specific attribute, *defined*, whose possible values are declared in the "TYPE" declaration at the beginning of the file. "Definiteness" can be either "unset", "defined", "undefined", or "doubtful".

The last interesting feature in this file is the definition of "proposition", in which we define two possible children: subject, which must be a noun phrase, and adjunct, which must be an adverbial phrase.

```
ANNOTATION SCHEME "st_1"
TYPE definiteness ENUM unset defined undefined
doubtful ENDTYPE
GROUP group
  ATTR comment TEXT * ENDATTR
ENDGROUP
GROUP nounPhrase EXTENDS group
  ATTR defined definiteness ONE unset ENDATTR
ENDGROUP
GROUP adverbialPhrase EXTENDS group
ENDGROUP
// proposition
GROUP proposition EXTENDS group
```

```
  CHILD subject CHILDTYPE nounPhrase ENDCHILD
  CHILD adjunct CHILDTYPE adverbialPhrase ENDCHILD
ENDGROUP
```

5. Conclusion

The next step for us is to create an experimental implementation of the graph database. It will be a good stepping stone for programming an historical log. Meanwhile, we will try to generalize the database. The software behind Ramsès (excluding libraries such as Jsesh) is already 80000 lines long, and it would be nice if it could benefit other projects in the near future.

6. BIBLIOGRAPHY

BODARD, G. *et al.*, 2011: Lessons from the conversion of the Duke Databank of Documentary Papyri from legacy formats into EpiDoc TEI XML, abstracts of the Digital Humanities 2011 conference, Stanford, in: https://dh2011.stanford.edu/wp-content/uploads/2011/05/DH2011_BookOfAbs.pdf, p. 31.

MANOLA, F. & E. MILLER, 2004: *RDF Primer*, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

MAZZIOTTA, N., 2010: Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples, in: *Recherches qualitatives. Hors-série "Les actes"*, 83-94.

POLIS, S. & S. ROSMORDUC, 2013: Building a construction-based treebank of Late Egyptian. The syntactic layer in Ramsès, in: POLIS, S. & J. WINAND (eds.), *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique & Égyptologie), Liège, 6-8 July 2010*, Ægyptiaca Leodiensia 9, Liège, 45-59.

PRÉVOST, S. & A. STEIN, 2012: *Syntactic Reference Corpus of Medieval French et l'ordre des compléments du verbe en ancien français*, Séminaire "Lectures en linguistique experimentale", Université Paris 7, http://www.uni-stuttgart.de/lingrom/stein/downloads/prevost-stein-objets-afr-handout.pdf.

ROSMORDUC, S. *et al.*, 2008: Ramsès, a new Research Tool in Philology and Linguistics, in: STRUDWICK, N. (ed.), *Information Technology and Egyptology in 2008. Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique et Egyptologie), Vienna, 8-11 July, 2008*, Piscataway, NJ, 155-166.

WINAND, J. *et al.*, 2008: Ramses. An Annotated Corpus of Late Egyptian, in: KOUSOULIS, P. (ed.), *Proceedings of the Xth IAE Congress, Rhodos, 2008*, (in press).

OrientDB: http://www.orientdb.org/.

Neo4J: http://neo4j.org/.

DAS EDFU-PROJEKT
ZIEL, METHODE UND VERARBEITUNG DER LEXIKOGRAPHISCHEN ERGEBNISSE

DIETER KURTH

Eine „Corpusbasierte historische Linguistik und Philologie“ ist natürlich ohne die umfassende philologische Bearbeitung der vorhandenen Textcorpora nicht denkbar. Eines der großen Textcorpora, die Altägypten hinterlassen hat, sind die Inschriften des Tempels von Edfu, der in Oberägypten unter ptolemäischer Herrschaft erbaut wurde.

Umfang und Bedeutung der Edfutexte sind enorm. Sie wurden von dem französischen Kollegen Emile Chassinat in jahrzehntelanger Arbeit auf über 3.000 DIN A4-Seiten in hieroglyphischen Drucktypen publiziert, und seitdem werden diese Texte in den meisten ägyptologischen Publikationen innerhalb der verschiedensten Sachzusammenhänge zitiert.

Die Sachzusammenhänge betreffen vor allem Religion, aber auch in nicht geringem Umfang Architektur, Astronomie, Botanik, Geographie, Geschichte, Landwirtschaft, Pharmakologie, Philosophie, Wirtschaft und – was man von den Inschriften eines Sakralbaus nicht erwartet – sogar Geometrie, Mathematik und vieles andere mehr.

Dennoch waren bis zur Gründung des Edfu-Projekts nur relativ wenige Inschriften des Tempels übersetzt worden, und zwar zu einem Teil in speziellen Untersuchungen, die sich bestimmten Bereichen widmeten, etwa dem Schriftsystem, dem täglichen Kult, dem Inhalt der mythologischen Texte oder einigen der über 2000 Ritualszenen. Der größte Teil der Inschriften war jedoch in Art eines Steinbruchs verwendet worden, indem man einzelne Textpassagen nur *punktuell* herangezogen hatte, um die Ergebnisse anderer Untersuchungen zu bestätigen oder zu widerlegen, wobei natürlich wegen des nicht beachteten Zusammenhangs manche Texte nicht richtig aufgefasst werden konnten. Eine Bearbeitung der Texte *um ihrer selbst willen*, das heißt vor allem ihre zusammenhängende Übersetzung, war ein Desiderat geblieben.

Im Jahre 1986 konnte das von mir gegründete Edfu-Projekt seine Arbeit an der Beseitigung dieses Desiderats beginnen. Das Projekt war als Langzeitprojekt der Deutschen Forschungsgemeinschaft angenommen worden, und zu seiner Ausstattung gehörten zunächst

drei halbe BAT2a-Stellen für Wissenschaftliche Mitarbeiter sowie Sachmittel.

Seit 2002 ist das Projekt im Programm der Akademie der Wissenschaften der Universität zu Göttingen. Diese hat die Finanzmittel kräftig aufgestockt, derart, dass zurzeit ein Mitarbeiter auf einer ganzen, und fünf Mitarbeiter auf halben BAT2a-Stellen beim Projekt beschäftigt sind. Außerdem stellt die Göttinger Akademie dem Projekt einen beträchtlichen Bücheretat zur Verfügung, was nötig wurde, nachdem die Universität Hamburg die international renommierte Hamburger Ägyptologie liquidiert hatte, ohne die Leistungen des Faches zu würdigen und ohne die Bedeutung des Faches im Fächerkanon der Universität zu beachten.

Anlass zur Gründung des Projekts war eine betrübliche Erfahrung bei meiner Arbeit mit den Tempelinschriften der griechisch-römischen Epoche Ägyptens, also mit den Texten der Tempel von Philä, Edfu, Kom Ombo, Esna und Dendera – um nur die größten zu nennen. Betrüblich war nämlich, dass ohne eine Übersetzung und Aufbereitung aller Texte die überaus zahlreichen Informationen und erhellenden Parallelen der Texte in ihrer eigenen Masse *begraben* waren. Nur einen Teil davon zu finden, erforderte sehr viel Zeit, denn jede neue Fragestellung verlangte ein erneutes Durchsuchen Tausender Seiten – und für die Suche standen außer den Szenentiteln keine Indizes zur Verfügung.

So hatte ich denn beschlossen, diesen einer Wissenschaft unwürdigen Zustand wenigstens für einen Tempel zu beenden. Warum gerade der Tempel von Edfu? Nun, von den fünf großen Tempeln der griechisch-römischen Epoche waren bis 1984, dem Beginn der Planungen, nur die Inschriften Edfus vollständig publiziert worden. Und einer der großen Tempel sollte es sein, um über möglichst viele interne Parallelen verfügen zu können.

Die Ziele des Projekts ergaben sich also aus der praktischen philologischen Arbeit. Dabei war das wichtigste Anliegen eine geschlossene Übersetzung aller Inschriften sowie deren Aufbereitung für das Fach Ägyptologie und seine Nachbarwissenschaften. Was die Aufbereitung angeht, ihr Herzstück ist die Anlage sehr ausführlicher analytischer Indizes, mit deren Hilfe ein rascher und gezielter Zugriff auf die reichen Informationen der Texte möglich wird.

Diesen Zielen musste sich die Methode unterordnen, und dabei zeigte sich, dass zwei hauptsächliche Arbeitsphasen nötig sind. In der *ersten*

Phase wird eine vorläufige Übersetzung angefertigt, in deren Verlauf zum einen alle fraglichen Stellen notiert, und zum anderen alle für das Verständnis wichtigen Informationen erfasst und eingeordnet werden. Das System der Einordnung besteht aus folgenden Kategorien:

- Formular (sich wiederholende formelhafte Wendungen)
- Vokabular
- Schreibungen
- Grammatik
- Gottheiten
- Ortsnamen
- Inhalte von besonderem Interesse.
- Hinzu kommen die Ergebnisse der Arbeit vor Ort, also eine Photosammlung sowie die Aufzeichnungen der Kollationierungskampagnen, die beide dazu dienen, eine epigraphisch sichere Textgrundlage zu schaffen; denn die Photos in Chassinats Publikation und eigene Aufnahmen zeigen, dass die vorhandenen Textabschriften zahlreiche Fehler enthalten.
- Außerdem wird die gesamte Sekundärliteratur zu den Edfutexten gesammelt, um den Fortschritt der ägyptologischen Forschung zu nutzen.

Diese Materialien sind zunächst nur für den projektinternen Gebrauch gedacht, um in der *zweiten Phase* – nun von der Höhe der aufbereiteten Materialien aus und im Besitz aller Parallelen – eine weitgehend abgesicherte Übersetzung liefern zu können.

Nach Abschluss der Übersetzungsarbeit sollen die obengenannten und zunächst nur dem projektinternen Gebrauch dienenden Materialien in gesonderten Bänden publiziert werden.

Auf Anraten des Kollegen Wolfgang Schenkel, der sich in der Arbeit mit den Computern des frühen Computerzeitalters gut auskannte und mich vor einer Verdoppelung der Probleme gewarnt hatte, wurde die Arbeit ab 1986 zunächst mit Hilfe der vertrauten Zettelkästen durchgeführt. Als die Computer eine für meine Zwecke taugliche Leistungsfähigkeit erreicht hatten, wurden die Daten der Zettelkästen in elektronische Datenträger eingebracht.

Die nun immer leistungsstärker gewordenen Computer erlaubten die Herstellung neuer Arbeitsmittel, die auf das Projekt zugeschnittenen sind und die philologische Arbeit enorm erleichtern und fördern. Über die beiden überaus hilfreichen Programme „Edfu Explorer“ und

„Vector Office“ kann Herr Graeff, als ihr Erfinder und als Mitarbeiter des Projekts, Genaueres berichten.

Nun möchte ich noch einmal auf den Nutzen des Projekts für die Ägyptologie und ihre Nachbarwissenschaften zurückkommen. Einige Materialien konnten bereits vor Abschluss des Projekts ins Netz gestellt werden, wie beispielsweise ein Teil der Datenbanken sowie ein Teil der Edfu-Begleithefte und der Übersetzungen.

Auch meine zweibändige „Einführung ins Ptolemäische“, eine Grammatik mit Zeichenliste und Übungsstücken, ist inzwischen erschienen. Zwar fußt diese Arbeit nicht, wie anfangs geplant, auf der abschließenden Übersetzung aller Edfutexte, doch ist sie auch ohne dies für das laufende Projekt hilfreich, weil sowohl für die Grammatik als auch für die Zeichenliste alle hieroglyphischen Inschriften ausgewertet wurden, die bis zum Jahre 2007/8 erschienen waren oder die mir Kollegen, wie Jochen Hallof, Peter Dils und Sylvie Cauville, freundlicherweise vorab zur Verfügung gestellt hatten. – Auch mit Blick auf den möglichen Nutzen für das Fach hatte ich es für sinnvoll gehalten, die „Einführung ins Ptolemäische“ bereits vor dem Abschluss der Übersetzung aller Edfutexte zu veröffentlichen; denn es war nach und nach immer deutlicher geworden, dass die Arbeit an der Übersetzung erst nach vielen weiteren Jahren enden würde.

Ein *Wörterbuch* der Edfutexte gehört von Beginn an zum Publikationsplan des Edfu-Projekts. Das Wörterbuch soll nicht nur neue Wörter, sondern auch neue Bedeutungen und Schreibungen der Wörter enthalten; der Maßstab für „neu“ ist das Berliner „Wörterbuch der ägyptischen Sprache“.

Die Schreibungen werden hierarchisch angeordnet, von der ausführlichsten Schreibung bis zur Abkürzung, wobei sich die Allographen des ptolemäischen Schriftsystems und die teils verschiedenen Determinative in ihrer Abfolge an den Sachgruppen der Zeichenlisten orientieren. Mit Hilfe dieser Anordnung lässt sich die Genese der meisten abgekürzten und ungewöhnlichen Schreibungen nachvollziehen, beziehungsweise überhaupt erst verstehen.

Abschließend möchte ich noch einiges zur Frage anmerken, wie sich die Solidität der traditionellen philologischen Arbeit mit den großartigen neuen Möglichkeiten der EDV zum Nutzen eines neuen Wörterbuchs optimal verbinden lassen.

Mein ursprünglicher Vorschlag, den ich 1997 in Berlin während der Tagung „Textcorpus und Wörterbuch“ vorgetragen habe, sah folgende computergestützte Hauptschritte vor:

●● Die Bearbeiter der Textcorpora oder auch anderer Textgruppen oder Einzeltexte, allesamt ausgewiesene Philologen, liefern ihre lexikographischen Ergebnisse zum Wörterbuch; sie haben in diesem Bereich die höchste Kompetenz, weil sie die Kontexte am besten kennen.

●● Die Bearbeiter des Wörterbuchs, ausgewiesene Philologen und Lexikographen, sammeln die Beiträge, prüfen sie und weisen die Bearbeiter der Textcorpora auf Widersprüche und Unstimmigkeiten hin; in diesem Bereich haben die Lexikographen die höchste Kompetenz, weil sie alle angegebenen Wortbedeutungen überblicken.

●● Von diesem hermeneutischen Zirkel, der mehrere Durchgänge haben wird, profitieren sowohl die Einträge des Wörterbuchs als auch die Übersetzungen der Textcorpora.

Bei diesem Verfahren würden, wie es einer historischen Wissenschaft mit stets wachsendem Quellenmaterial angemessen ist, keine Fehler festgeschrieben; es würde vielmehr der lexikographische Fortschritt in die notwendigen Aktualisierungen eingearbeitet und nach und nach zu einer beständigen Verbesserung des Wörterbuchs führen. In diesem „offenen“ Wörterbuch würden meines Erachtens die neuen Möglichkeiten der EDV bestens genutzt.

Nun ist dieser schlichte und aus der philologischen Praxis erwachsene Vorschlag nicht aufgegriffen worden, wobei ich trotz verbliebener Zweifel hoffe, dass die an seiner Ablehnung beteiligten Kollegen, wie Adolf Erman es einmal ausgedrückt hat, „mit höherer Weisheit gesäugt“ waren.

So habe ich denn die Materialien des Edfu-Projekts dem Berliner Wörterbuchunternehmen zur Verfügung gestellt, dergestalt, dass eine Mitarbeiterin die bereits übersetzten Edfutexte nach den Vorgaben des Wörterbuchunternehmens aufbereitet und nach Berlin schickt.

Dabei ist aber ein Problem nicht zu übersehen. Es geht darauf zurück, dass man beim Berliner Unternehmen erst relativ spät bemerkt hatte, wie sehr man doch auf die Zulieferungen der philologischen Projekte angewiesen ist. Nun verfolgen aber diese teils schon seit längerer Zeit laufenden philologischen Projekte eigene Ziele und haben jeweils verschiedene Formalia entwickelt. Für deren Anpassung an die Vorgaben des Berliner Unternehmens und für die Übertragung ihrer teils sehr umfangreichen Materialien können die

philologischen Projekte in der Regel keinen Mitarbeiter eigens abstellen, und es fragt sich, wer die erforderlichen Personalkosten trägt.

Abschließend sei hierzu noch angemerkt, dass mein oben skizziertes dreigliedriges Konzept eines „offenen Wörterbuchs“ notwendigerweise *feste* Stellen braucht, also nicht solche, die in mehr oder weniger kurzlebigen Projekten angesiedelt sind. Dieser Bedarf ließe sich gut begründen, weil der Fortschritt des Faches im wichtigen Bereich der Lexikographie im Interesse aller Fachvertreter ist.

DER THESAURUS LINGUAE AEGYPTIAE – KONZEPTE UND PERSPEKTIVEN

INGELORE HAFEMANN & PETER DILS

Teil I: Einführung, Grundstrukturen und Nutzung (Ingelore Hafemann)

Zu den interessanten Internetangeboten auf dem Gebiet der historischen Textwissenschaft und Lexikographie gehört seit einigen Jahren zweifellos auch der *Thesaurus Linguae Aegyptiae* (kurz TLA). Er umfasst ein digitales Corpus von Abschriften ägyptischer Texte aus der Pharaonenzeit (von ca. 2400 v. Chr. bis zum 5. Jh. n. Chr.). Bei den Texten handelt es sich um Dokumente aus der gesamten Geschichte der Pharaonenherrschaft, die nicht nur in unterschiedlichen Sprachstufen des Ägyptischen verfasst, sondern auch in verschiedenen Schriften niedergeschrieben wurden – in hieroglyphischer und hieratischer sowie in demotischer Schrift. Das Corpus bietet die Texte komplett in ägyptologischer Transkription und mit Übersetzungen und Kommentaren an. Teils werden auch Fotos und Umzeichnungen der Texte gegeben. Die Volltexte stehen für lexikalische und philologische Recherchen im Internet zur Verfügung. Hauptnutzer sind dabei Ägyptologen, aber dank der Übersetzungen können auch Vertreter verschiedenster Nachbardisziplinen sowie Kulturwissenschaftler oder interessierte Laien dieses Corpus mit Gewinn einsehen. Konkret stellt dieser Thesaurus die Publikationsplattform dreier Akademienvorhaben dar.[1] Das sind einmal die beiden Vorhaben mit dem Namen *Altägyptisches Wörterbuch*, eines an der Berlin-Brandenburgischen Akademie der Wissenschaften und eines an der Sächsischen Akademie der Wissenschaften zu Leipzig. Das dritte Partnerprojekt ist die *Demotische Textdatenbank* an der Akademie der Wissenschaften und der Literatur, Mainz. Der Zusammenarbeit dieser drei Akademienprojekte haben sich weitere Projektgruppen angeschlossen, wie das Projekt *Leuven online index of Ptolemaic and Roman Hieroglyphic Texts* der Katholieke Universiteit Leuven und das *Bonner Totenbuchprojekt*, das an der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste verankert ist, sowie das abgeschlossene Corpusprojekt *Digital Heka* der Universität Leipzig, das sich der Erfassung magischer Sprüche und Zaubertexte widmete.

---

[1] URL: http://aaew.bbaw.de/tla/.

Konzepte

Der *Thesaurus Linguae Aegyptiae* ist eine lexikalische Datenbank ägyptischer Texte. Durch die Verknüpfung einer Textdatenbank mit einem lexikalischen Wort-Thesaurus innerhalb eines modernen Navigationsprogramms sind vielfältige Abfragemöglichkeiten geschaffen worden, die weit über gedruckte Formate von Texteditionen und Wörterbüchern hinausgehen, diese aber nicht ersetzen sollen und können. Der TLA ist ein völlig neuartiges Instrument der Recherche und Forschung, das im Folgenden beschrieben wird.

Das Konzept eines solchen digitalen Textcorpus muss im Rahmen dieses Tagungsbandes nicht erläutert werden. Inzwischen ist der TLA auch in dem kleinen Fach der Ägyptologie als ein stabiles Forschungswerkzeug etabliert. Allerdings ist das Gesamtkonzept eines digitalen Corpus mit allen seinen Nutzungsstrategien bisher nur wenigen Ägyptologen bekannt, da die Arbeit mit digitalen Corpora in der Ägyptologie wie in den meisten historischen Sprachwissenschaften noch Neuland ist.

Die Arbeit am elektronischen Corpus – die 1992 gestartet wurde – ist von Beginn an einem Grundprinzip verpflichtet gewesen: dem der corpusbasierten Lexikographie. Im Fokus stand demnach zwar ein Text-Thesaurus – das Corpus – aber in erster Linie ist das Neuprojekt als ein lexikographisches Unternehmen anzusehen. Dieses folgte dem methodischen Grundprinzip des alten Vorgängerprojektes an der damals Berliner Akademie, das im Jahre 1897 von Adolf Erman als das internationale Großprojekt „Wörterbuch der ägyptischen Sprache“ an der Preußischen Akademie initiiert wurde. In diesem Projekt verfolgte Erman konsequent den Ansatz der Darstellung des Wortschatzes auf der Grundlage der kompletten Quelltexte. Sein großes Wörterbuch[2] ist bis heute das Standardwerk zur ägyptischen Lexikographie. Das eigens dafür angelegte Archiv mit Textabschriften, auf über 1,5 Millionen lexikalisch-lexikographisch sortierten Zetteln niedergelegt – das sogenannte „Zettelarchiv des ägyptischen Wörterbuchs“ – ist neben dem Wörterbuch ebenfalls ein Forschungswerkzeug ersten Ranges für Ägyptologen geblieben.

Eben diesem Prinzip der textbasierten Arbeitsweise folgte auch das 1992 an der Berlin-Brandenburgischen Akademie der Wissenschaften neu gegründete Projekt „Altägyptisches Wörterbuch“. Ein elektronisches Textcorpus mit einer integrierten elektronischen Wort-

[2] ERMAN, A. & H. GRAPOW (Hrsg.), *Wörterbuch der aegyptischen Sprache*, 7 Bde., Leipzig / Berlin 1926-1963, Belegstellen, 5 Bde., Leipzig / Berlin 1935-1953.

liste soll ein völlig neuartiges Forschungswerkzeug zur Verfügung stellen, das nun Wörterbuch und Textarchiv miteinander verbindet und dem Nutzer komplexe Recherchen ermöglicht. Erstmals stehen ihm in diesem System nun alle ausgewerteten Quellentexte zur Verfügung.

Strukturen

Die erwähnte elektronische Wortliste (Lemmaliste) ist das lexikalische Rückgrat der Struktur. Diese Lemmaliste hat zweierlei Funktionen: zum einen ist sie ein Nachschlagewerk zum Wortschatz des Ägyptischen für jeden Nutzer, zum anderen dient sie intern als Lemmatisierungswerkzeug innerhalb des Programms der Texterfassung. Sie wurde mit dem bis heute bekannten Wortschatz des Ägyptischen entsprechend dem aktuellen Forschungsstand angefüllt.[3] Hierzu wurden lexikographische Standardwerke ausgewertet sowie mit neuen Wörtern aus den eingegebenen Einzeltexten vervollständigt.

Die lexikographische Beschreibung der einzelnen Lemmata in der Liste ist relativ flach und setzt sich aus den Angaben zur Transkription, hieroglyphischen Schreibweise, Bedeutungsangabe in Form von Übersetzungen (Deutsch und Englisch), der Wortkategorie und den bibliographischen Referenzen einiger lexikographischer Standardwerke zusammen (vgl. Abb. 1). Den Wörtern des Spezialwortschatzes sind Sachgruppenbezeichnungen zugewiesen worden, wie Toponym, Königs- und Göttername, Personenname, Titel und Epitheton sowie Eigenname von Sachen und Institutionen, die in der Rubrik Wortkategorie stehen. Es handelt sich bei diesen Wörtern immer um Substantive und insbesondere oft um Komposita.

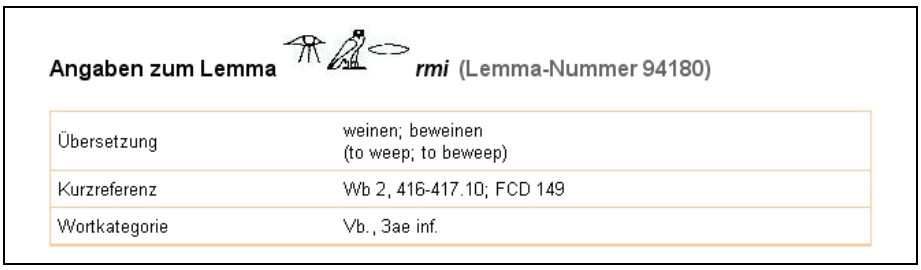


Abb. 1: Detailansicht zum Lemma rmị im TLA

[3] Ende 2012 zählte die ägyptische Lemmaliste 31.400 Lemmata, inklusive der 3.000 Personennamen, und die demotische Lemmaliste 11.400 Lemmata, einschließlich der 1.800 Personennamen.

Die Angaben zum Gebrauch der Wörter und zu ihren semantischen Lesarten sowie zur phraseologischen und idiomatischen Verwendung, die sonst in Wörterbüchern verzeichnet werden, sind nur teilweise als Einträge in die Lemmaliste aufgenommen. Der vielfältige Wortgebrauch aber kann aus dem elektronischen Corpus selbst entnommen werden, denn erst im Sprachgebrauch in den Texten entwickeln sich die lexikalischen und grammatischen Potenzen der Wörter, zeigt sich ihre Polysemie und werden Metonyme sichtbar oder auch die dialektale Färbung von Wortbedeutungen und sogar Ideosynkrasien einzelner Autoren.

Daher lag und liegt die tatsächliche Hauptarbeit des Projektes in der Schaffung eines Corpus der Quellentexte. Die Texteingabe erfolgt ausschließlich manuell. Unmittelbar bei der Eingabe der einzelnen Textwörter wird in einem eigenen Arbeitsschritt Wort für Wort mit der Lemmaliste verlinkt – also lemmatisiert. Diese Lemmazuweisung erfolgt halbautomatisch und muss jeweils manuell bestätigt werden. Damit werden gleichzeitig die lexikographischen Beschreibungslabel des Lemmas mit jedem Textwort verknüpft, wie z.B. die Wortkategorie. Danach werden grammatisch-morphologische Annotationen, wie Flexionsform, für die aktuellen Wortformen jedes Textwortes vorgenommen[4] und es wird eine gebundene Satzübersetzung angefertigt.

Den Texten selbst werden auch Metadaten zugewiesen wie Datierung, Textsorte, Herkunft, Textträger oder Schriftform. Diese Metadaten werden als Filter für Abfragen dienen und dann können bei den Wortrecherchen nur bestimmte Texte einer bestimmten Zeit oder Herkunft einbezogen werden. Hier wird jeder Nutzer frei sein und kann seine Fragen nach eigenen Bedürfnissen formulieren.

Was wir gewinnen, ist ein vollständig corpusbasiertes Wörterverzeichnis – oder aber ein lexikonbasiertes Textcorpus, je nachdem, aus welcher Perspektive man es betrachtet. Im Laufe der Arbeit am Corpus ergab sich tatsächlich auch ein Perspektivwechsel vom Lexikon zum Corpus, und die wissenschaftliche Dimension eines elektronischen Corpus als solchem wurde immer deutlicher. Damit gerieten die Methoden der Corpuslinguistik in den Blick, einer neuen

[4] Es sind nicht alle Texte grammatisch annotiert worden; das gesamte Corpus der Demotischen Textdatenbank ist nicht grammatisch annotiert. Die grammatische Annotation ist bis Ende 2012 noch nicht im TLA abrufbar.

Disziplin, die mit dem Aufbau elektronischer Corpora neue Wege in der Linguistik aufgetan hat.[5]

Leistungen und Nutzungsstrategien des TLA

Ursprünglich stand also die Idee einer reinen Belegstellenabfrage für die Wörter des ägyptischen Lexikons im Mittelpunkt. Da das gesamte Textcorpus Wort für Wort lexikalisch verknüpft ist, werden den Lemmata der elektronischen Wortliste sukzessive mit der laufenden Texterfassung immer weitere Belege zugewiesen. Mit dem Corpus wächst die Belegmenge je Lemma.

So wurde das Konzept eines virtuellen Wörterbuchs entwickelt. Die Philosophie dahinter ist die folgende: Der durch den Lexikographen erarbeitete traditionelle Wörterbucheintrag, der zu den einzelnen Wörtern bestimmte Angaben abschließend in einem Buch oder auch einem digitalen Dokument verzeichnet, wird hier durch ein Werkzeug ergänzt, das dem Forscher erlaubt, die gesamte Materialbasis zur Aufklärung seiner spezifischen Wortrecherche zu nutzen. Bekanntlich ist jeder Eintrag in gedruckten Wörterbüchern eine Notlösung. Es lassen sich meist viele weitere Wortbedeutungen oder Lesarten eines Lemmas bzw. semantische Eigenschaften durch kotextuelle Zusammenhänge im tatsächlichen Sprachgebrauch finden. Die Fragen des Forschers lassen sich mit einem lexikonbasierten Corpus frei und immer wieder neu formulieren. Die angebotenen Funktionen des TLA sind vornehmlich auf ägyptologische Bedürfnisse abgestimmt. Es können Formenbildung von Verben und Substantiven sowie Vorkommen bestimmter Phonemfolgen ermittelt werden. Neben den wortspezifischen Abfragen können auch Recherchen zu phraseologischen und idiomatischen Wendungen oder zum Wortgebrauch im Gesamtcorpus oder in bestimmten Textgruppen angestellt werden. Durch den Zugang zum Gesamtcorpus lassen sich allgemeine Fragen der Textlinguistik, die sich mit satzübergreifenden Strukturen beschäftigen, erstmals am ägyptischen Textmaterial durchführen. Dazu gehören Fragen nach der Lexemvariation eines Textes oder seiner Kohärenz. Auch thematische und funktionale Aspekte der Textstruktur sind auf neuer Datengrundlage untersuchbar. Einige Analysen, die aus dem Bereich der Corpuslinguistik bekannt sind, wie Wortstatistiken, Verteilung von Worthäufigkeiten,

[5] LEMNITZER, L. & H. ZINSMEISTER, *Korpuslinguistik*, 2. Aufl., Tübingen 2012; LÜDELING, A. & M. KYTÖ, *Corpus linguistics: An International Handbook*, Berlin / New York, Vol. 1 2008, Vol. 2 2009.

Schlüsselwortanalysen und Kollokationsanalysen, sind bereits im TLA lauffähig implementiert und können nach Eingabe der gewünschten Parameter per Mausklick gestartet werden (s. detaillierter im Teil II unten). Mit Hilfetexten in Deutsch und Englisch und einem Handbuch (als PDF) soll allen Nutzern der multifunktionale Zugang zum TLA erleichtert werden.

Seit 2009 kann der Nutzer auch nach hieroglyphischen Schreibungen innerhalb der Lemmata der Wortliste recherchieren, wobei jeweils eine Standardschreibung je Lemma geboten wird.

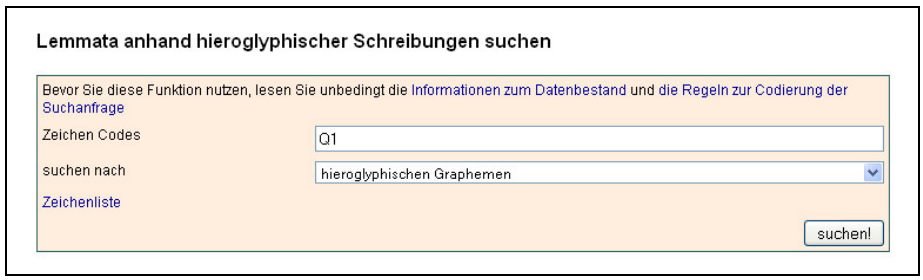


Abb. 2a: Suche nach hieroglyphischen Schreibungen im TLA

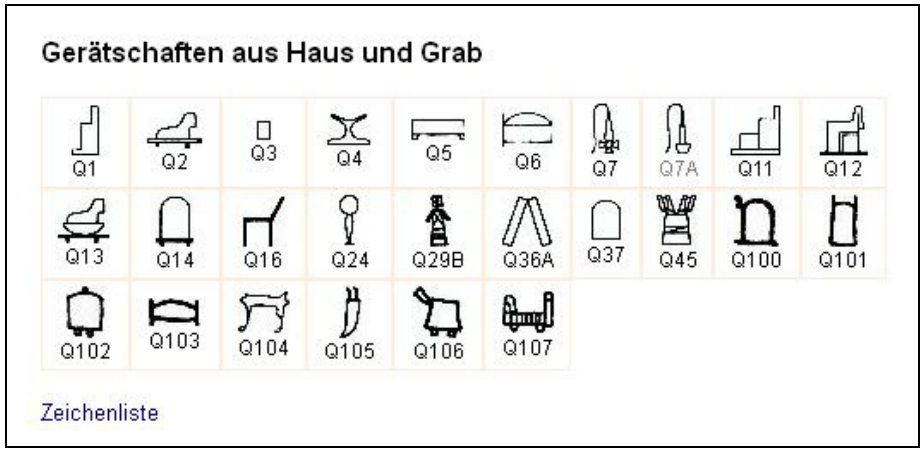


Abb. 2b: Ausschnitt aus der Zeichenliste

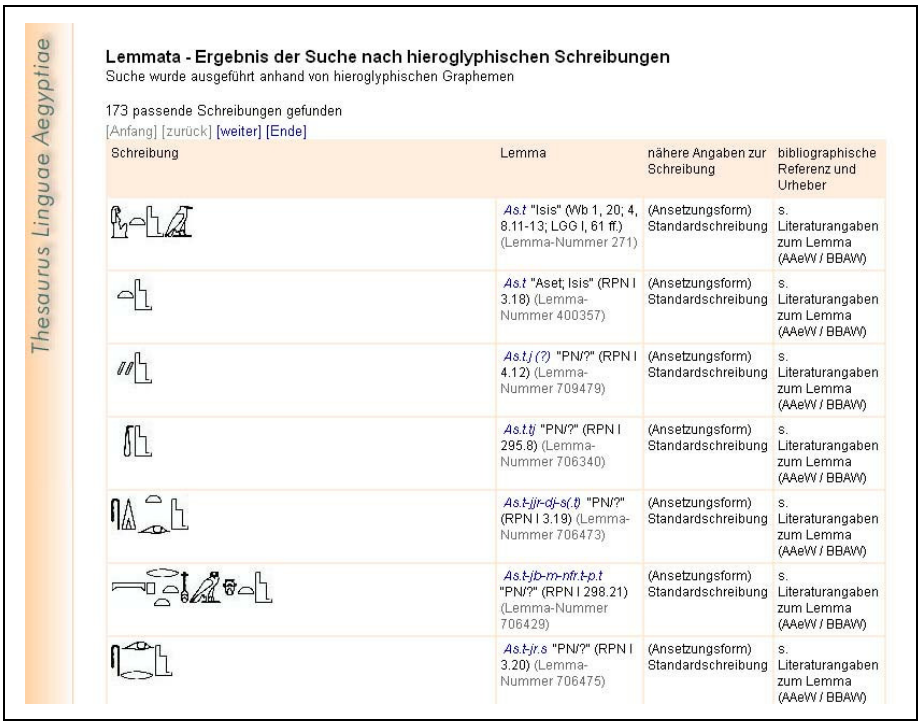


Abb. 2c: Ergebnisliste der Suche nach der Hieroglyphe Q1

Seit 2011 kann man hieroglyphische Schreibungen auch für einige Textwörter des Corpus einsehen, vorerst probeweise in einem ausgewählten Teilcorpus. Dieser Ansatz wird ausgebaut und eröffnet neue Perspektiven für die Erforschung von Verteilungen und Frequenzen hieroglyphischer Zeichen in den Wortschreibungen, die dann nach linguistischen, soziolinguistischen, chronologischen oder regionalen Parametern angeordnet und untersucht werden können.

Teil II: Funktionalitäten und Anwendungen des TLA (Peter Dils)

Im Bereich der Funktionalität bietet der TLA zwei Einstiegsmöglichkeiten in die digitalen Daten. Die erste Möglichkeit läuft über die einzelnen Lemmata der hieroglyphisch-hieratischen oder alternativ die der demotischen Lemmaliste. Die zweite startet über die Texte des Textcorpus.

Die erste Einstiegsmöglichkeit über die Lemmaliste dient vorrangig ausgebildeten Ägyptologen, wobei die Belegstellenabfrage im Vordergrund steht. Die zweite Möglichkeit des Einstiegs in das Text-

corpus oder auch nur in Teilcorpora wird gern von Studierenden genutzt – besonders für das Corpus der literarischen Texte, die häufig Gegenstand von Seminaren zur Textlektüre sind. Dabei werden die Transkriptionen, Übersetzungen und Kommentierungen als Vorbereitungsgrundlage für den Unterricht konsultiert.

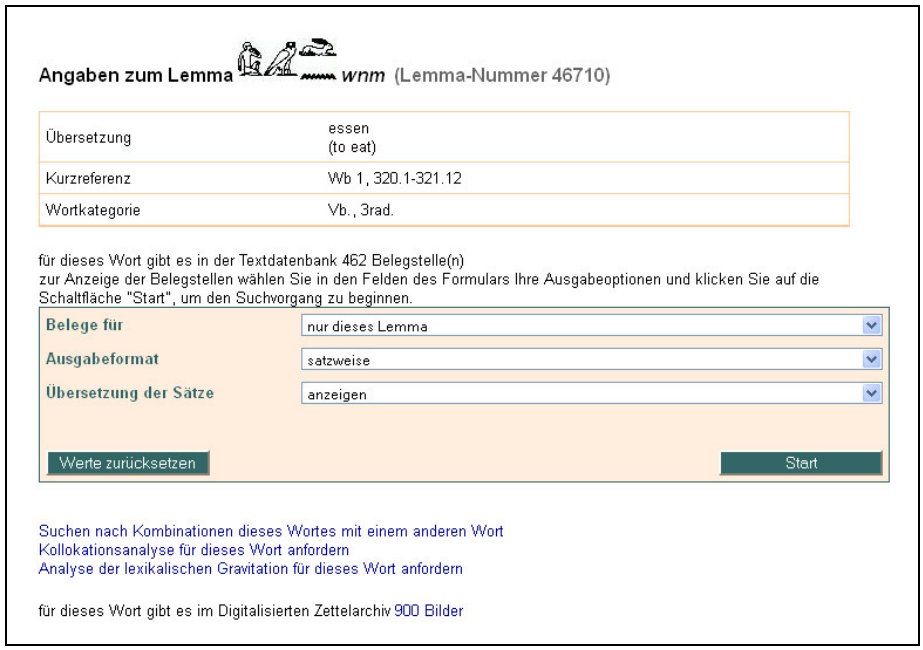


Abb. 3a: Detailansicht zum Lemma wnm im TLA

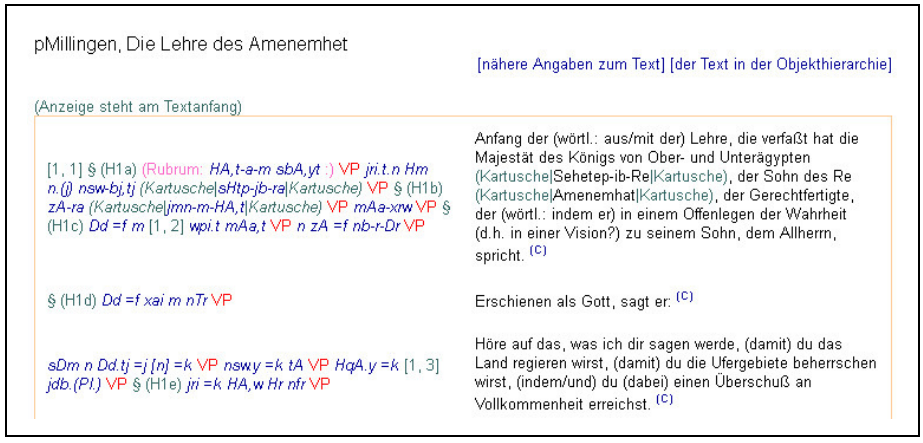


Abb. 3b: Beispieltext aus dem TLA

Die einzelnen transkribierten Wortformen im Text liefern den Zugang zu den passenden Lemmata in der Wortliste, über die dann alle Belegstellen dieses Wortes im Corpus aufgelistet werden, und zwar in ihrem textlichen Zusammenhang. Kotextuelle Übersetzungen pro Satz und Quellenangabe sowie chronologische und geographische Parameter liefern per Mausklick schnell einen Überblick über Wörter und ihren Gebrauch. Parallel und synchron können für jedes Lemma der Wortliste die betreffenden Zettel aus dem digitalisierten Zettelarchiv – im Beispiel 900 Zettel – von den insgesamt über 1,5 Millionen Zetteln des Wörterbucharchivs angezeigt werden, das seinerzeit – wie oben beschrieben – Adolf Erman und Hermann Grapow angelegt hatten.

Spezielle Anzeigefunktionen

Die Ergebnisse können in doppelter Hinsicht konkretisiert werden. Zum einen kann man nach dem gemeinsamen Vorkommen von zwei bestimmten Lemmata suchen, oder der Nutzer sucht nach einem Lemma und dessen Kombination mit einem Wort einer bestimmten Wortart wie zum Beispiel dem Vorkommen eines bestimmten Verbs der Bewegung mit bestimmten Präpositionen. Zum anderen kann eine Kollokationsanalyse für ein Lemma angefordert werden, so dass eine Liste von – statistisch gesehen – regelmäßigen gemeinsam auftretenden Lemmata ausgegeben wird oder auch von solchen, die unerwartet zusammen auftreten. Bei der Kollokationsanalyse wird also versucht, Nuancen des Wortgebrauchs durch Sammlung und Ordnung typischer Wortkombinationen zu erfassen, was auf der Basis von Prüfstatistiken errechnet und in Tabellenform dargestellt werden kann. Im in Abb. 4 gezeigten Beispiel wird gefragt, welche Lemmata typischerweise eine Stelle rechts vom ägyptischen Verb *wnm*: „essen“ vorkommen.[6] Im TLA sind hierfür zwei statistische Analyseverfahren – *T-score* und *MI-score* genannt[7] – implementiert. Nach den Regeln der ägyptischen Grammatik ist an der Stelle rechts vom Verb am ehesten entweder das Subjekt oder das Objekt zu erwarten und das wird durch die Analyse mittels *T-score* bestätigt: „ich esse, er isst, du isst, sie essen“ einerseits sowie andererseits „Brot essen, Nahrungsration/Brot essen, Kleinvieh essen, Kot essen, ihn essen, mich essen“.

---

[6] Für die Analyse sind in der Eingabemaske die Parameter für die gewünschten Wortabstände einstellbar; s. Handbuch.

[7] *T-scores* zeigen häufige Kollokationen, *MI-scores* weisen eher seltene, aber enge Zusammenhänge terminologischen Charakters aus.

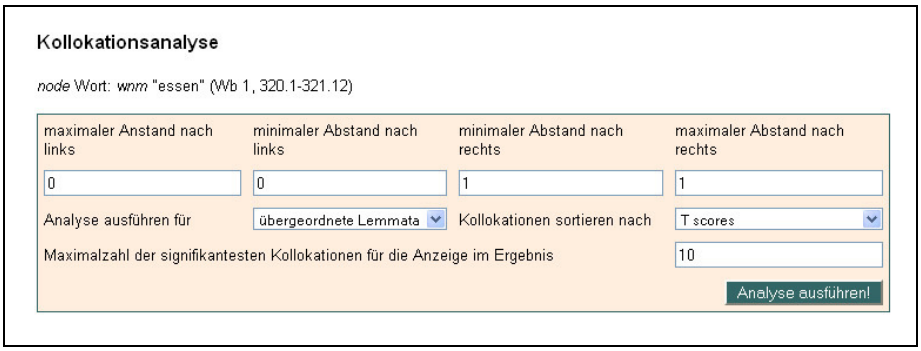


Abb. 4a: Kollokationsanalyse des Wortes wnm, sortiert nach T-scores

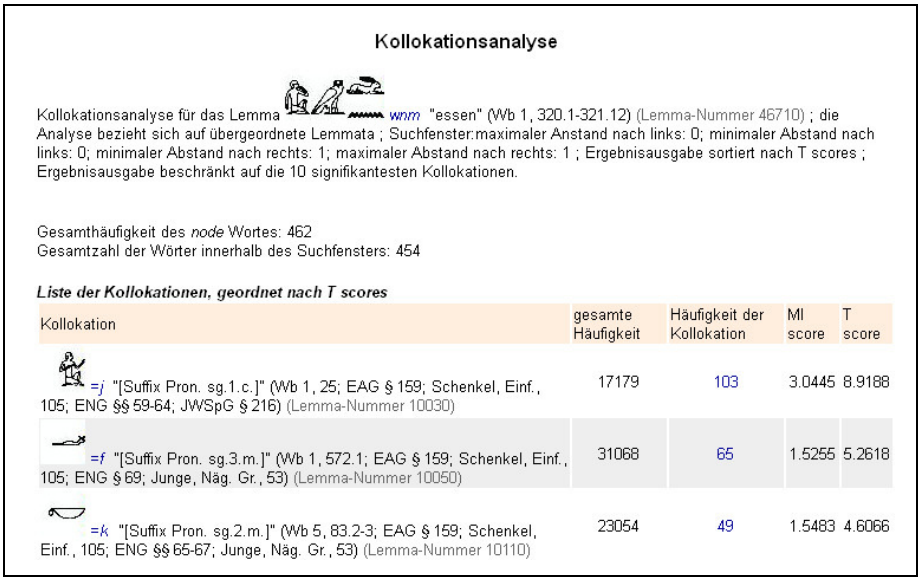


Abb. 4b: Ergebnis der Kollokationsanalyse

Bei der Analyse mittels *MI-score* tauchen keine Verbsubjekte auf, sondern nur gegessene Lebensmittel, allerdings oft ziemlich skurrile, die mit Bedrohungen im Jenseits zusammenhängen.

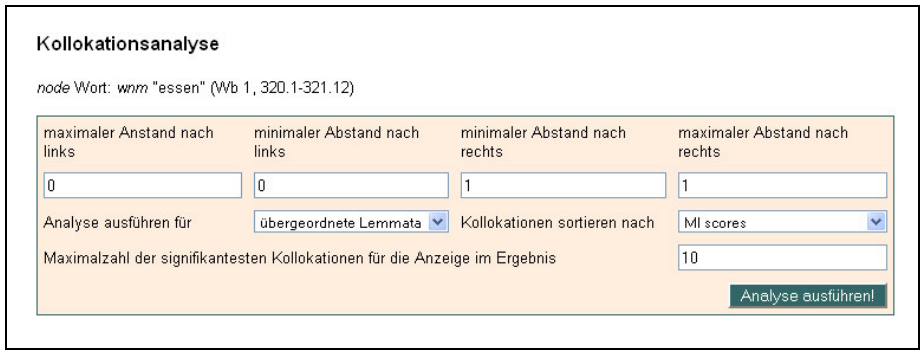


Abb. 4c: Kollokationsanalyse des Wortes wnm, sortiert nach MI scores

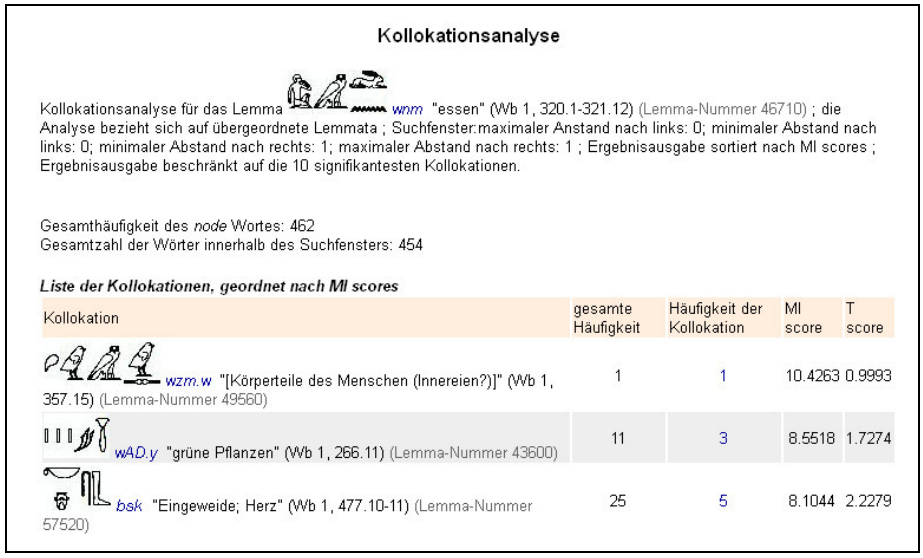


Abb. 4d: Ergebnis der Kollokationsanalyse

Einstieg über die Texte

Bei der zweiten Einstiegsmöglichkeit in den TLA stehen nicht die einzelnen Lemmata im Vordergrund, sondern der Gesamttext oder eine Gruppe von Texten. Der TLA ermöglicht es, automatisch Indizes zu Texten oder Textgruppen zu erstellen. Der TLA unterscheidet dabei zwischen den Nomina propria und den Nomina Apellativa (Eigennamen und Titel): Bei den Nomina propria kann der Index rein alphabetisch oder nach Wortarten sortiert ausgegeben werden.

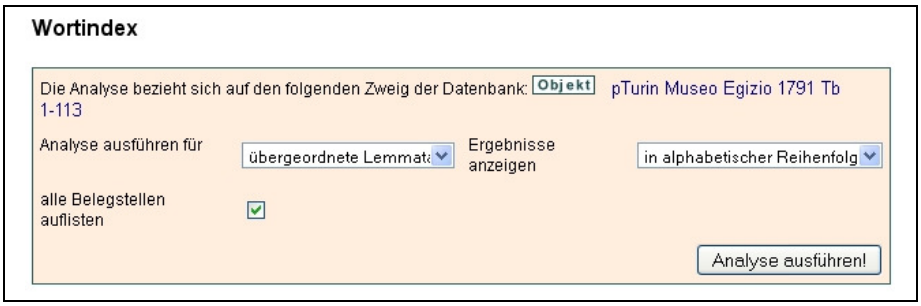


Abb. 5a: Erstellen eines Wortindexes des pTurin Museo Egizio 1791

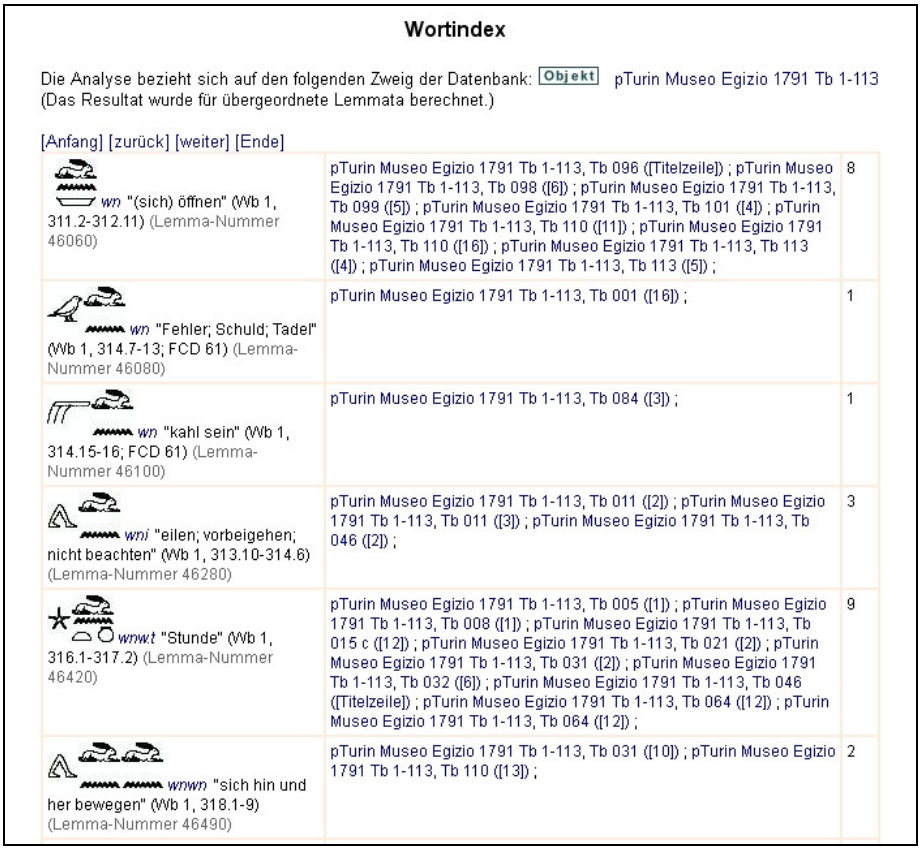


Abb. 5b: Ergebnisliste des Wortindexes des pTurin Museo Egizio 1791

Ein solcher Index kann ein Ziel für sich sein, z.B. der „Wortindex zum späten Totenbuch (pTurin 1791)“ von Burkhard Backes[8], der auf der Grundlage der Digitalisierung dieses Totenbuchs im TLA erstellt wurde. Solche Indizes können aber auch die Ausgangslage für weitere Forschungen sein. Zum Beispiel wurden die im TLA erstellten Indizes von literarischen Texten schon für Clusteranalysen eingesetzt. Simon Schweitzer ist der Frage nachgegangen, ob man auf der Grundlage seines Wortschatzes die Urschrift der „Lehre für König Merikare“, von der nur Abschriften aus dem Neuen Reich überliefert sind, dem literarischen Schaffen des Mittleren oder doch erst des Neuen Reiches zuweisen kann. Nach der hier angewandten Analysemethode sind die nächstliegenden Texte dieser Lehre zwei Texte – die „Lehre des Ptahhotep“ und der „Beredte Bauer“ –, die unbestritten aus dem Mittleren Reich stammen.[9]

Eine weitere Anwendung, die der TLA bereitstellt, ist die Schlüsselwortanalyse. Diese kann dem Philologen helfen, Thematik und Stil eines Textes genauer zu bestimmen.[10]

Eine Schlüsselwortanalyse für die Substantive, die in der Geschichte des „Beredten Bauern“ vorkommen, zeigt folgendes Ergebnis: Es werden Lemmata ausgeworfen, die jeder Ägyptologe bei diesem Text als thematisch wichtig einstufen würde: Der „Bauer“ erleidet „Böses“, weil einer seiner Esel „oberägyptische Gerste“ am Wegrand frisst. Der Getreidebesitzer konfisziert daraufhin unberechtigterweise den Besitz des Bauern. Der Bauer fühlt sich als „Bedrängter“ und versucht in neun Reden beim Distriktverwalter Recht zu bekommen. Diesen vergleicht er u.a. mit einer „Waage“, die im „Gleichgewicht“ sein und nicht absichtlich zu einer Seite abweichen sollte.

---

[8] BACKES, B., *Wortindex zum späten Totenbuch (pTurin 1791)*, Studien zum ägyptischen Totenbuch 9, Wiesbaden 2005.

[9] SCHWEITZER, S., Dating Egyptian literary texts: lexical approaches, in: MOERS, G. *et al.* (eds.), *Dating Egyptian Literary Texts*, Lingua Aegyptiae. Studia Monografica 11, Hamburg 2012.

[10] Schlüsselwörter sind Wörter, die den Text besonders charakterisieren. Sie kommen in einem Text häufiger vor als in anderen Texten und geben so Aufschluss über das Thema oder stilistische Besonderheiten des Textes. Kulturwissenschaftlich gesehen zielt der Begriff auf Wörter, die in kulturellen Diskursen eine zentrale Rolle spielen.

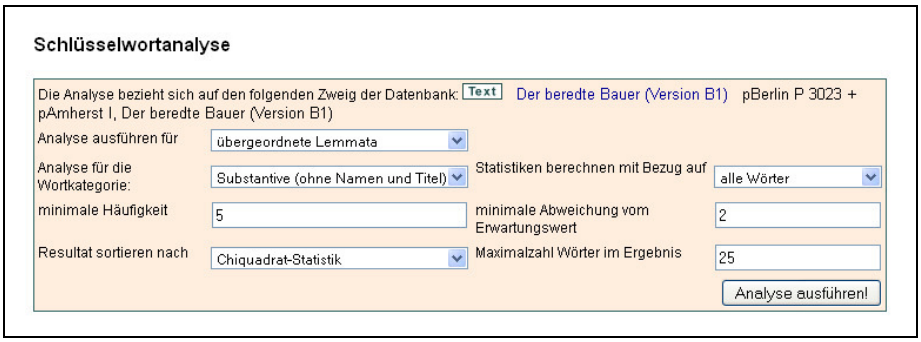


Abb. 6a: Schlüsselwortanalyse für Substantive aus dem beredten Bauern

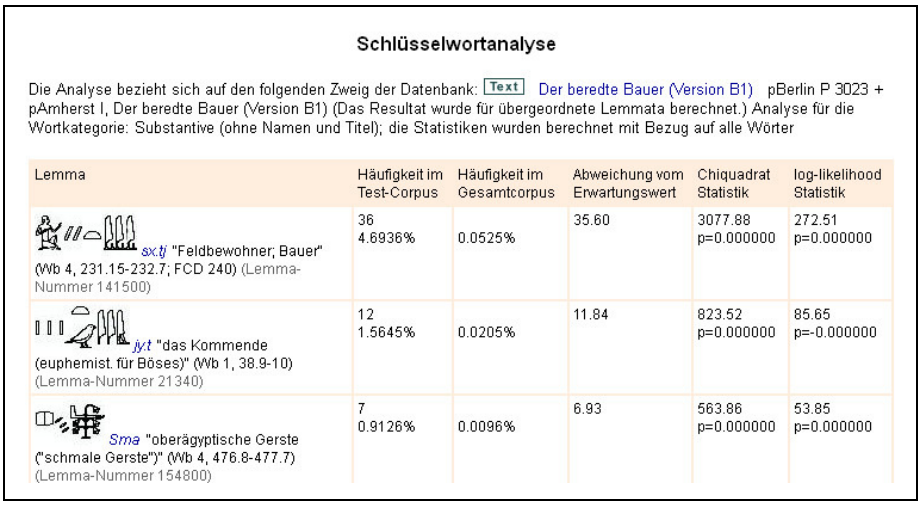


Abb. 6b: Ergebnisliste der Schlüsselwortanalyse für Substantive aus dem beredten Bauern

Eine entsprechende Analyse für die Verben bestätigt diesen Eindruck: Verben des Redens sowie Verben zum Thema „schief sein, betrügen, beseitigen, schädigen“ sind prominent vertreten.

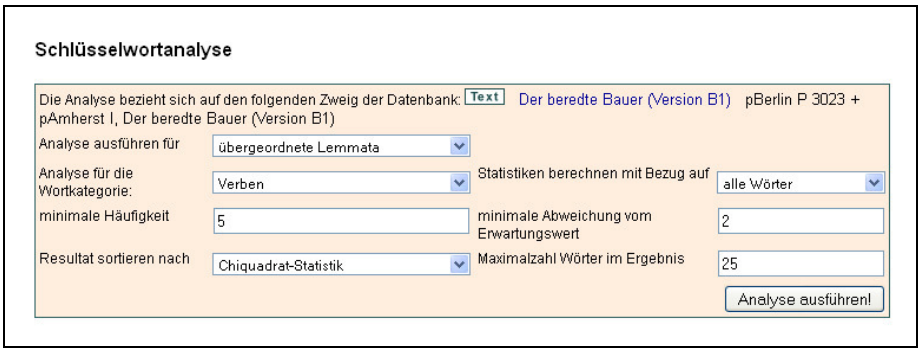


Abb. 7a: Schlüsselwortanalyse für Verben aus dem beredten Bauern

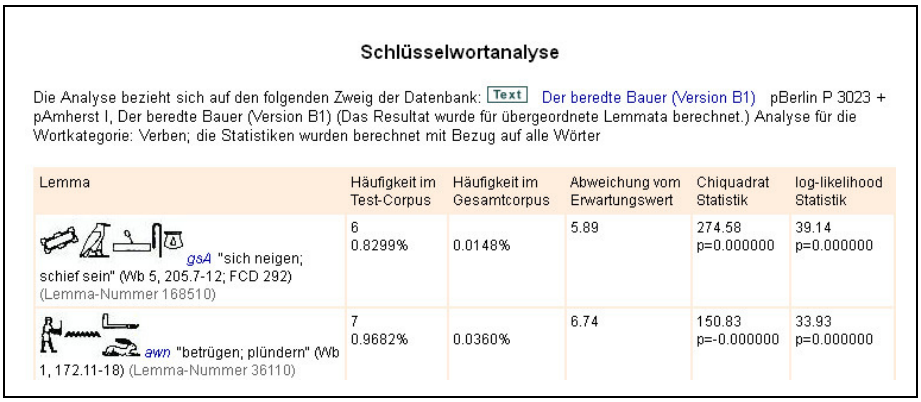


Abb. 7b: Ergebnisliste der Schlüsselwortanalyse für Verben aus dem beredten Bauern

Führt man eine solche Schlüsselwortanalyse für alle Lemmata durch, tauchen auch Funktionswörter auf. Diese beleuchten nicht die Thematik, sondern den Stil. Auffällig beim Beredten Bauern ist, dass keine Personalpronomina auftauchen, dafür aber Hervorhebungspartikel sowie der negative Imperativ „tue nicht!“. Im Gegensatz dazu steht z.B. die autobiographische Erzählung des Sinuhe, in dem die Pronomina „ich“ und „mich“ wichtig sind, sowie indikativische Negativsätze.

Der TLA liefert auch Frequenzdaten von Wörtern und Wortarten in einem Text, wie z.B. Angaben zu den häufigsten Wörtern und

Wortarten oder das *type/token*-Verhältnis.[11] Man kann dort die Parameter so hoch stellen, dass nicht nur die häufigsten Wörter, sondern alle Lemmata berücksichtigt werden. Auf diese Weise erhält man ausreichend Daten, um sie für Recherchen im Bereich der Lexikostatistik zu nutzen. Der TLA bietet zum ersten Mal in der Ägyptologie die Möglichkeit, solche Recherchen ohne gigantischen manuellen Zählaufwand durchzuführen.

Die digitalen Rohdaten der altägyptischen Textdatenbank können natürlich auch außerhalb der bisherigen Analysemöglichkeiten des TLA eingesetzt werden. Tatsächlich sind die Möglichkeiten, die in der öffentlich zugänglichen Version der Datenbank geboten werden, für corpuslinguistische Fragestellungen im Augenblick noch relativ beschränkt, ganz im Gegensatz zu dem Potential an Forschungsmöglichkeiten, die die im XML-Format abgelegten Informationen bieten. So hat Katharina Stegbauer (Universität Leipzig) die eingegebenen magischen Texte des Projektes *Digital Heka* mit computerlinguistischen Methoden untersucht, um stilistische Analysen durchzuführen.[12]

Des Weiteren sollen in einem Kooperationsprojekt mit der Ruhr-Universität Bochum und dem Kompetenzzentrum für die Erschließungs- und Publikationsverfahren der Universität Trier ebenfalls die XML-Daten des TLA innerhalb eines Projekts zur Religionssoziologie zur Verfügung gestellt werden. Eine semantische Netzwerkanalyse der in Texten zum Ausdruck gebrachten Relationen zwischen Inhalten und Akteuren soll anhand der digitalen Sprachdaten durchgeführt werden. Im Ergebnis werden so die altägyptischen Relationen mit denen anderer Kulturen auf neuer Basis vergleichbar.[13]

Für die vergleichenden statistischen Analysen eines Textes ist ein Referenzcorpus erforderlich. Im TLA dient die ganze Textdatenbank als Referenzcorpus. Hierbei ist zu bemerken, dass dieses Referenzcorpus bisher kein ausbalanciertes Corpus im Sinne der Corpuslinguistik darstellt.

---

[11] Die *type-token*-Ratio berechnet das Verhältnis zwischen der Zahl der Wörter im Text insgesamt und der Anzahl der einzelnen Lexeme. So sieht man, ob der Text im Verhältnis zur Textlänge einen verhältnismäßig reichen oder armen Wortschatz aufweist. Vgl. SCOTT, M., *Oxford WordSmith Tools*, Oxford 2004.

[12] http://www.uni-leipzig.de/~digiheka/.

[13] http://aktuell.ruhr-uni-bochum.de/pm2012/pm00248.html.de; http://www.ceres.ruhr-unibochum.de/de/project/forschungsprojekte/senereko/?print=1.

Die hier beschriebenen Operationen kann man sowohl für das hieroglyphisch-hieratische als auch für das demotische Corpus durchführen. In zukünftigen Projekten ist ein gleichzeitiger Zugriff auf beide Teilcorpora geplant, die dann in einem Gesamtcorpus integriert zur Nutzung aufbereitet werden sollen. In diesem neuen diachronen Ansatz der Sprachforschung soll die wissenschaftshistorisch bedingte Trennung der Sprachstufen des Ägyptischen überwunden werden.[14]

[14] Ein entsprechender Antrag für ein Projekt zur diachronen Aufbereitung und Erforschung des Ägyptischen ist bei der Union der Akademien eingereicht und befürwortet worden.

ZUR ARBEIT AN DER DEMOTISCHEN TEXTDATENBANK: TEXTAUSWAHL[1]

GÜNTER VITTMANN

Einer Anregung von Frau Hafemann folgend, werde ich speziell über die Textauswahl im Mainzer Akademieprojekt „Datenbank demotischer Texte“ und gewisse damit verbundene Probleme sprechen. Gleich vorweg: Das theoretische Ideal einer vollständigen Aufnahme sämtlicher oder auch nur der meisten in demotischer Schrift geschriebenen Texte aus gut einem Jahrtausend ist im Rahmen eines zeitlich wie personell begrenzten Unternehmens – das stand von Anfang an völlig fest – nicht zu realisieren. Auf Grund verschiedener Umstände ist die Textaufnahme oft sehr zeitaufwendig: Erhaltungszustand, schriftbedingte Probleme, Unklarheiten im sprachlichen und sachlichen Verständnis, Publikationsstand, Länge des jeweiligen Textes etc. spielen eine Rolle.

Die durch diese Umstände nötige Beschränkung muss aber keineswegs ein Nachteil sein: Wenn man z.B. wissen will, in welchen sprachlichen und inhaltlichen Zusammenhängen bestimmte Begriffe, Wendungen oder Kollokationen wo und wann vorkommen, ist es nicht zwingend erforderlich, von einer allgegenwärtigen Urkundenformel wie etwa „Du hast mein Herz zufriedengestellt“[2] *sämtliche* Belege verfügbar zu haben. Natürlich wäre auch dies grundsätzlich sinnvoll und letztlich wünschenswert, aber hierzu bedürfte es umfassender Teilcorpora nach Textgattungen, Archiven etc., was im Rahmen des Projekts „Datenbank demotischer Texte“ nicht geleistet werden kann, da es schlichtweg über dessen Zielsetzung hinaus und vielleicht sogar daran sogar vorbeigehen würde. Eines der angestrebten Ziele war und ist auch, beim interessierten nichtdemotistischen Benutzer – und d.h. nicht zum geringsten bei Kollegen mehr „klassisch“-ägyptologischer Observanz – gewisse traditionelle

---

[1] Um die Anmerkungen erweiterter und nur gelegentlich geringfügig modifizierter Text meines im Rahmen der Tagung gehaltenen Vortrags. Zur Benutzung der Datenbank sei auf VITTMANN, G., Ein neues demotistisches Hilfsmittel. Die „Datenbank demotischer Texte“, in: *Enchoria* 31, 2008/2009, 144-152 hingewiesen.

[2] Die Formel hat sogar in den Titel einer Papyruspublikation Eingang gefunden: SCHENTULEIT, M. & G. VITTMANN, *„Du hast mein Herz zufriedengestellt…“. Ptolemäerzeitliche demotische Urkunden aus Soknopaiu Nesos*, Corpus Papyrorum Raineri 29, Berlin & New York 2009.

Berührungsängste mit dem Demotischen abzubauen und dem ja bisweilen immer noch anzutreffenden Vorurteil, demotisches Schrifttum sei langweilig und für die Ägyptologie wenig bedeutsam, da fast nur aus stereotypen Rechtsurkunden und Abrechnungen aus einer Zeit von Fremdherrschaft und Verfall bestehend, entgegenzuwirken.

Aus diesem Bestreben heraus, aber gleichzeitig motiviert durch die Absicht, eine möglichst große Anzahl lexikalisch, phraseologisch, grammatisch und inhaltlich ergiebiger Texte der verschiedensten Gattungen aufzubereiten, werden – ohne eher formelhafte Dokumente mit notorisch repetitiven Elementen zu vernachlässigen – bevorzugt Texte ausgewählt, die in irgendeiner Weise etwas Besonderes zu bieten haben. Während man bei der großen Masse von Rechtsurkunden und Verwaltungsdokumenten, wie schon angedeutet, mit einer repräsentativen Auswahl, bei der der Benutzer zwangsläufig dieses oder jenes vermissen wird, vorliebnehmen muss, sind bestimmte Textgattungen des demotischen Schrifttums dazu prädestiniert, soweit der aktuelle Publikationsstand dies erlaubt, in besonders hohem Umfang aufgenommen zu werden:

- *Literarische Texte* („schöne Literatur“), von denen der größte Teil des verfügbaren Materials, angefangen von den demotischen „Klassikern“ wie den beiden Setne-Erzählungen oder den großen Inaros-Petubastis-Erzählungen, bis hin zu mythologischen (vor allem Mythos vom Sonnenauge) und „prophetischen“ Texten (Demotische Chronik; „Lamm des Bokchoris“) sowie auch erst kürzlich publizierten Texten wie der Geschichte von Padipep, der dem Pharao eine Geschichte erzählt,[3] aufgenommen wurde.[4]
- (Im weitesten Sinne) *religiöse und magische Handschriften*. Hier ist ebenfalls recht viel aufgenommen worden; genannt seien hier vor allem die Handschriften Berlin 8351, BM 10507, der Papyrus

[3] TAIT, J., Pa-di-pep tells Pharaoh the Story of the Condemnation of Djed-her: Fragments of Demotic Narrative in the British Museum, in: *Enchoria* 31, 2008/2009, 113-143.

[4] Friedhelm Hoffmann und Joachim Quack ließen mir bereits vor Erscheinen ihrer Anthologie (HOFFMANN, F. & J. F. QUACK, *Anthologie der demotischen Literatur*, Berlin 2007) dankenswerterweise Berichtigungen zu früheren Versionen der Demotischen Textdatenbank zukommen. Nach Erscheinen dieser Anthologie habe ich die alten Eingaben sukzessive erneut revidiert, wobei sicher immer noch Fehler stehen geblieben sind. Auf die kürzlich erschienene Anthologie von AGUT-LABORDÈRE, D. & M. CHAUVEAU, *Héros, magiciens et sages oubliés de l'Égypte ancienne. Une anthologie de la littérature en égyptien démotique*, Paris 2011, habe ich bei den entsprechenden Texten verwiesen; kritische Vergleiche der Übersetzungen in dieser neuen Anthologie mit denen in der Datenbank werden nach und nach vorgenommen werden.

Harkness, die beiden Totenpapyri Rhind und das demotische Totenbuch.[5] Bei den im engeren Sinne magischen Texten musste ich mich allerdings im wesentlichen auf den längsten und wichtigsten, den bekannten magischen Papyrus London-Leiden sowie die besser erhaltenen Partien des erst vor wenigen Jahren veröffentlichten divinatorischen Papyrus Wien D 12006[6] beschränken; die anderen längeren magischen Texte werde ich wohl nicht mehr schaffen, jedenfalls nicht während der offiziellen Laufzeit des Projekts.

- *Weisheitstexte*: die großen demotischen Klassiker (Anchscheschonki, Papyrus Insinger) sind natürlich eingearbeitet worden, außerdem eine Reihe kleinerer Weisheitstexte.[7]
- *Briefe*, einschließlich der sog. Briefe an Götter: Da die Briefe alles in allem mit Ausnahme der einleitenden und abschließenden Formeln sprachlich (auch hinsichtlich Onomastik) wie inhaltlich durchaus abwechslungsreich sind – und in der Regel auch von ihrem Umfang her zu bewältigen sind – nehme ich möglichst jeden mir unterkommenden demotischen Brief auf:[8] Bisher sind ca. 150 in der Datenbank, das ist schon ein großer Teil der publizierten Quellen dieser Textgattung, und es werden weitere hinzukommen.
- Auch die zahlreichen *Stelen* sowie *Graffiti* bzw. gegebenenfalls *Dipinti* auf Felsen und in Gebäuden erschöpfen sich keineswegs immer nur in Formeln: nicht selten gibt es, teilweise sogar sehr ausführlich wie in den berühmten Synodaldekreten (Rosette) oder in vielen, teilweise recht langen, späten Graffiti im Tempel von Philae,[9] historisch, biographisch und kulturgeschichtlich wertvolle

---

[5] Eine neue Übersetzung aller größeren hieratischen und demotischen religiösen Texte bietet SMITH, M., *Traversing Eternity. Texts for the Afterlife from Ptolemaic and Roman Egypt*, Oxford 2009.

[6] Vgl. zuletzt QUACK, J. F., in: *Omina, Orakel, Rituale und Beschwörungen*, TUAT NF 4, Gütersloh 2008, 362-367.

[7] Nicht aufgenommen, da zu fragmentarisch und noch zu lückenhaft verständlich, wird der Wiener Weisheitstext D 6212 (mit zugehörigen Fragmenten in verschiedenen Sammlungen), an dessen Publikation ich arbeite.

[8] Eine große Hilfe war und ist hier natürlich das Standardwerk von DEPAUW, M., *The Demotic Letter*, Dem. Stud. 14, Sommerhausen 2006.

[9] Besonders genannt sei Graffito Philae 416, das längste erhaltene demotische Graffito überhaupt. Die Neuedition durch POPE, J., The Demotic Proskynema of a Meroïte Envoy to Roman Egypt (Philae 416), in: *Enchoria* 31, 2008/2009, 68-103, bot eine günstige Gelegenheit zu einer neuerlichen Revision der bereits vor Erscheinen erfolgten Aufnahme in die Datenbank (Entsprechendes gilt mutatis mutandis natürlich auch für viele andere Texte, von denen seit ihrer Einarbeitung

Informationen. Was die Stelen betrifft, ist – von den Stelen der Apismütter abgesehen – das meiste in der Datenbank schon berücksichtigt,[10] wobei ich auf ein spezielles Problem, den Umgang mit Bilinguen in einer Datenbank, noch zu sprechen kommen werde.

Die Aufnahme von Kurztexten wie *Mumienschildern* (Abb. 1) mit kurzen religiösen Formeln sowie Personendaten ist, anders als etwa bei umfangreichen Rechtsurkunden, für das Einzelexemplar wenig zeitaufwendig, solange man nicht gezwungen ist, den Gesamtbestand aufzunehmen. Ein besonderer Reiz der Arbeit an der Datenbank demotischer Texte besteht für mich eben auch darin, sowohl Notwendigkeit als auch Freiheit der Auswahl zu haben – in der Einsicht, dass das Material, das Aufnahme verdient, ganz bestimmt nie ausgehen wird. Was z.B. die genannten Mumienschilder betrifft, von denen derzeit ca. 170 in der Datenbank verfügbar sind (ein Drittel davon die Stücke der Berliner Sammlung), weist die Leuvener Datenbank Trismegistos („DAHT“)[11] die stattliche Anzahl von 1396 Nummern auf.[12] Selbst wenn man die gut 500 dort registrierten unpublizierten Objekte im Louvre abzieht und einkalkuliert, dass auch von den übrigen 900 ein beträchtlicher Teil nicht in Publikationen und Abbildungen vorliegt, mit denen sich einigermaßen zuverlässig arbeiten ließe, bleibt immer noch genug übrig, was schon ausreichend Material für ein Einzelprojekt abgäbe. Man muss sich unter den gegebenen Umständen also bescheiden: Ein seltener oder ungewöhnlicher Personenname, ein Titel oder eine Datierung, ein Orts-

in die Datenbank Neueditionen oder auch „nur“ neue Übersetzungen erschienen sind).

[10] Für die Aufnahme von Stelen und Weihinschriften bildete VLEEMING, S. P., *Some Coins of Artaxerxes and other Short Texts in the Demotic Script* (…), Studia Demotica 5, Leuven [u.a.] 2001, die willkommene Grundlage. Bei den Weihinschriften ergab sich neuerdings ein für die Demotische Textdatenbank gewinnbringender Synergieeffekt durch die Gelegenheit, eine Auswahl derartiger Texte für TUAT zu bearbeiten (VITTMANN, G., Demotische Weihinschriften, in: *Grab-, Sarg-, Bau- und Votivinschriften*, TUAT NF 6, Gütersloh 2011, 129-144). Ähnliche Synergien gab und gibt es während der Arbeit an der Datenbank natürlich auch sonst des öfteren.

[11] http://www.trismegistos.org/daht/search.php.

[12] Letzter Zugriff 27. Juni 2012. – Im Spätsommer 2012 erschien VLEEMING, S. P., *Demotic and Greek-Demotic Mummy Labels and Other Short Texts Gathered from Many Publications*, Studia Demotica 5, Leuven [u.a.] 2011 (sic), worin ca. 750 Mumienschilder (mit Absicht unter weitgehender Auslassung des Materials im Britischen Museum und im Louvre) behandelt sind.

name, eine Altersangabe kann schon den Ausschlag dafür geben, den Text aufzunehmen (was ähnlich übrigens auch für andere Texte gilt): Es kommt durchaus vor, dass ich ein Dokument in die Datenbank um einiger weniger Besonderheiten willen – z.B. eines seltenen oder überhaupt erstmals belegten Wortes aufnehme[13] –, auch wenn das Dokument als Ganzes vielleicht nicht so spannend sein mag. Es ist aber grundsätzlich wichtig, die Belege nicht isoliert wie in einem herkömmlichen Wörterbuch, sondern im Textzusammenhang betrachten zu können.

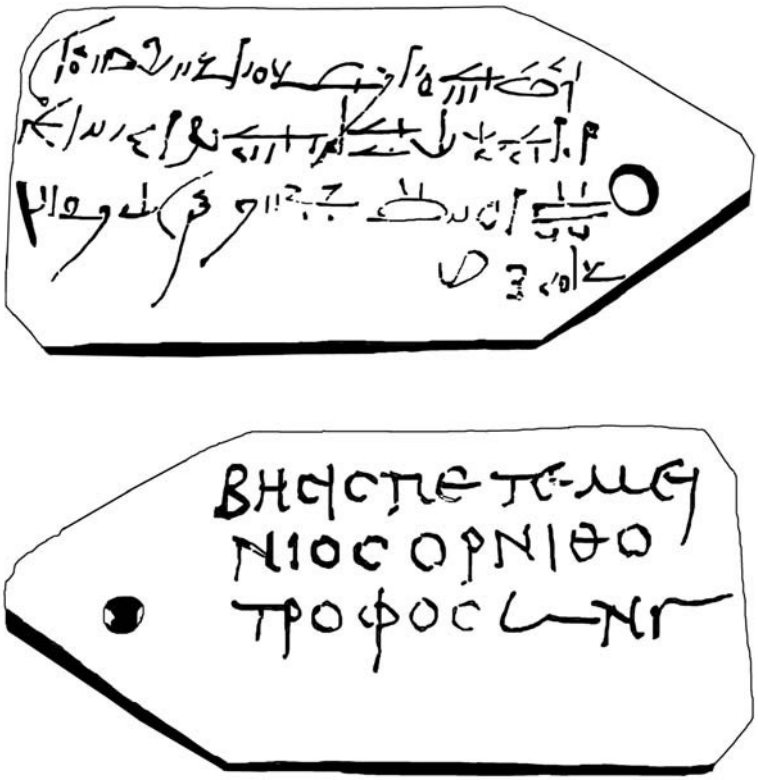


Abb. 1: Mumienschild BM 24533 (nach Photo aus dem Archiv des „Demotischen Namenbuchs“).

[13] So verdankt beispielsweise das Mumienschild BM 24533 (Abb. 1; Nr. 64 in der in der folgenden Anmerkung genannten Publikation; vgl. jetzt auch VLEEMING, *Mummy Labels*, Nr. 851) seine Aufnahme vor allem dem hier – und m.W. bisher nur hier – belegten Titel *s-n-ppj* „Vogelzüchter, Vogelhändler“, der im griechischen Teil mit ὀρνιθοτρόφος wiedergegeben wird.

Die Mumienschilder eignen sich übrigens auch gut als Beispiel für ein weiteres Phänomen, mit dem ich oft konfrontiert bin und das sich auch auf die Textauswahl auswirkt: Da bei einer per definitionem philologischen Disziplin wie der Demotistik immer wieder neue Textpublikationen erscheinen, und zwar nicht nur Neueditionen an sich bekannter Texte – die natürlich ebenfalls verglichen und ausgewertet werden müssen – sondern durchaus editiones principes wie die kürzlich im Druck erschienene Würzburger Dissertation über die Mumienschilder des Britischen Museum,[14] ist eine gewisse Flexibilität in der Aufnahme von Texten einem starren Festhalten an einem im Projektantrag konzipierten Zeitplan vorzuziehen, und ich bin da für die Freiheit und Selbständigkeit, die mir Herr Thissen als gleichermaßen großzügiger wie einsichtiger Projektleiter gelassen hat, sehr dankbar. Es versteht sich also, möchte ich meinen, von selbst, dass ich nach dem lange erwarteten Erscheinen einer Würzburger Dissertation mit der Publikation der Mumienschilder des Britischen Museum im Jahr 2011 eine Anzahl der wichtigeren (ca. 20 Stück) umgehend in die Textdatenbank aufgenommen habe. Ähnliches gilt auch für andere neue Publikationen, die jetzt natürlich nicht aufgezählt zu werden brauchen.[15]

Während es bei Ersteditionen von nicht zu langen und nicht allzu schlecht erhaltenen wichtigen (und manchmal vielleicht auch weniger wichtigen) Texten, die man sich als Demotist ja ohnehin näher ansieht, naheliegt, sie für die Datenbank aufzubereiten, besteht ein Problem, wenn die Bedingungen weniger günstig sind, d.h. ein Text zwar durchaus wichtig ist, aber man mit beträchtlichen Verständnisschwierigkeiten kämpft und der Erhaltungszustand teilweise auch nicht immer optimal ist. Das gilt vor allem für das in mehreren Handschriften überlieferte sog. *Thotbuch*, eine erst vor wenigen Jahren veröffentlichte, relativ umfangreiche und inhaltlich höchst bedeutsame Textkomposition, deren Publikation durch Richard Jasnow und Karl-Theodor Zauzich[16] Joachim Quack[17] immerhin als

[14] ARLT, C., *Deine Seele möge leben für immer und ewig. Die demotischen Mumienschilder im British Museum*, Studia Demotica 10, Leuven [u.a.] 2011.

[15] Genannt sei nur die kürzlich erschienene Publikation der Stelen der Apismütter (SMITH, H. S. *et al.*, *The Sacred Animal Necropolis at North Saqqara. The Mother of Apis Inscriptions*, London 2011, von denen ich – zusätzlich motiviert durch die aktuelle Arbeit an einer Rezension für *BiOr* (69, 2012, 460-463) – eine Anzahl für die Textdatenbank aufbereitet habe.

[16] JASNOW, R. & K.-TH. ZAUZICH, *The Ancient Egyptian Book of Thoth*, Wiesbaden 2005; hierzu wichtig J. F. QUACK, Die Initiation zum Schreiberberuf im Alten Ägypten, in: *SAK* 36, 2007, 249-295.

„das bedeutendste ägyptologische Buch des Jahres, wenn nicht des Jahrzehnts“, bezeichnet wurde. Ich bin bisher vor der Eingabe dieses Texts wegen der damit verbundenen nicht unerheblichen Schwierigkeiten und des zu erwartenden Zeitaufwands zurückgeschreckt und konnte (und kann) mich bis zu einem gewissen Grad damit beruhigen, dass es ja immer noch genügend andere aufnahmewürdige Texte gibt. Zumindest die besser erhaltenen Abschnitte des Thotbuchs müssten aber auf jeden Fall noch in die Datenbank eingespeist werden, zumal es demotisch sonst nichts Vergleichbares gibt. Ihn ganz herauszuhalten mit der Begründung, dass der Text ja zum Zeitpunkt der groben Arbeitsplanung für das Datenbankprojekt ja noch lange nicht verfügbar war, erscheint schwer zu akzeptieren.

Obwohl bisher ein sehr großer Teil des im Augenblick verfügbaren demotischen Schrifttums in der Datenbank, wie ich meine, ausreichend repräsentiert ist, gibt es immer noch Bereiche, die auf Weiterführung bzw. Inangriffnahme warten. Das sind – abgesehen von dem eben genannten Thotbuch sowie dem Rechtsbuch des sog. Codex Hermopolis, von dem erst die ersten vier Kolumnen eingearbeitet sind, was nicht ausreicht[18] – vor allem die sog. „wissenschaftlichen“ Texte, von denen bisher noch nicht viel in der Datenbank zu finden ist. Dies war vom Arbeitsplan her auch so vorgesehen, und es hat zugegebenermaßen nicht zuletzt einen subjektiven Grund: Es handelt sich dabei um *medizinische*, *mathematische* und *astronomische* Texte, die auch von ihrer Inhaltsseite her spezielle Anforderungen an den Bearbeiter stellen, und da kommt man als Philologe doch leicht an die Grenzen seiner Kompetenz.[19] Trotzdem ist es selbstverständlich unumgänglich, soll die Datenbank nicht allzusehr als Ruine dastehen, auch solche Texte in einem ausreichend repräsentativen Umfang zu berücksichtigen. Während man sich bei mathematischen und astronomischen Texten noch teilweise damit herausreden kann, dass da manches tatsächlich nur aus Zahlentabellen und Ähnlichem besteht, was in der Tat weder lexikalisch noch grammatisch sonderlich ertragreich ist, sind Texte medizini-

---

[17] QUACK, J. F., Rezension zu Jasnow & Zauzich, *Book of Thoth*, in: *OLZ* 101, 2006, 610-615, hier 610.

[18] 2012 wurden zahlreiche weitere Abschnitte aufgenommen, so dass der größte Teil dieses Dokuments nun eingearbeitet ist.

[19] Zur Problematik der Übersetzungen antiker wissenschaftlicher Texte vgl. IMHAUSEN, A. & T. POMMERENING (eds.), *Writings of Early Scholars in the Ancient Near East, Egypt, Rome, and Greece. Translating Ancient Scientific Texts*, Beiträge zur Altertumskunde 286, Berlin / New York 2010.

schen Inhalts, von ihrem enormen kulturgeschichtlichen Interesse ganz zu schweigen, auch lexikalisch, wie sich denken lässt, wertvoll. Gerade hier ist aber tut sich dank der in Vorbereitung befindlichen Arbeiten von Friedhelm Hoffmann[20] einiges, so dass in Zukunft noch neues Material aufgenommen werden kann, freilich aber aller Wahrscheinlichkeit nach erst nach Ende der offiziellen Laufzeit des Unternehmens „Datenbank demotischer Texte“.

Zum Schluss möchte ich noch einiges zum praktischen Umgang mit demotischen Texten sagen, die in irgendeiner Form mit anderen Sprachen und Schriften vergesellschaftet sind. Am wenigsten problematisch ist es noch, wenn derselbe Text griechisch und demotisch, also als klassische Bilingue, überliefert ist, wie bei den Synodaldekreten (Rosette, Kanopus), aber häufig auch bei Mumienschildern (vgl. Abb. 1). Ich füge dann die griechische Version – leider aus technischen Gründen nur in Umschrift und ohne Akzente – einfach in das Kommentarfeld ein.

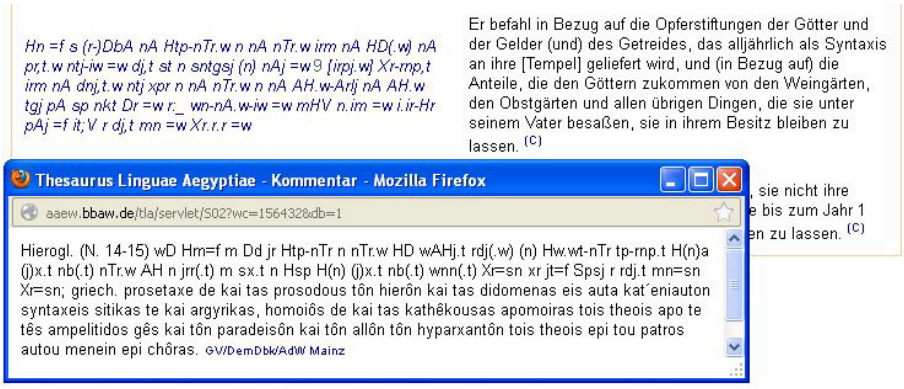


Abb. 2: Passage aus Z. 9 der Rosettana im TLA mit eingeblendetem Kommentarfeld zum untersten Absatz mit Angabe der hieroglyphischen und griechischen Parallelversionen.

Ist der Text, wie bei den Synodaldekreten, auch noch in hieroglyphischer Version überliefert, wird diese in Umschrift ebenfalls passagenweise im Kommentarfeld eingegeben (Abb. 2). (Für Nichtspezialisten sei bemerkt, dass sich die Versionen nicht nur durch die Schrift, sondern auch sprachlich, lexikalisch und stilistisch unter-

[20] Vgl. HOFFMANN, F., Zur Neuedition des hieratisch-demotischen Papyrus Wien D 6257 aus römischer Zeit, in: IMHAUSEN & POMMERENING, *Writings of Early Scholars*, 201-218.

scheiden.) An sich müssten natürlich diese hieroglyphischen Versionen eigenständig in den Thesaurus aufgenommen, übersetzt und lemmatisiert werden, was aber nicht im Rahmen der Demotischen Textdatenbank geschehen kann, da in diese, wie es von Anfang an geregelt war, begreiflicherweise nur demotisch geschriebene Texte Eingang finden können. Ähnliches gilt für die hieratisch-demotischen Totenpapyri Rhind.

Es kommt aber auch einmal vor, dass in einen langen demotischen Text ein hieratischer Abschnitt eingeschoben ist, der kein Paralleltext zu ersterem ist und darum nicht einfach sozusagen durch die Hintertür im Kommentarfeld präsentiert werden kann: Ich meine die gegen Ende des P. Rylands 9 eingeschobenen Abschriften zweier älterer Stelen.[21] Hier blieb aus Gründen der Gesamtkonzeption nichts anderes übrig, als die ganze Passage (immerhin zwei Kolumnen, wenigstens ist der Text aber in sich abgeschlossen) einfach wegzulassen. Objektiv gesehen ist das natürlich misslich, aber im derzeitigen Stadium der einzelnen Teilprojekte war das nicht anders zu machen. Es ist jedoch zu hoffen, dass die nichtdemotischen Abschnitte außerhalb der Demotischen Textdatenbank, in die sie technisch nicht hineingehören, in Zukunft in ein größeres Ganzes integriert werden können.

In manchen spätdemotischen Texten sind einzelne Elemente hieratisch geschrieben, manchmal sogar innerhalb eines Wortes (wie z.B. in einem Ostrakon aus Medinet Madi im Fayum, auf dem die Präposition *r* hieratisch *ỉw* () = [e] und der Relativkonverter *ntj* gemischt hieratisch-demotisch  *ỉw.ntj* = [et] geschrieben ist[22]). Solche „allographen“ Elemente müssen natürlich innerhalb der laufenden Transkription berücksichtigt (und lemmatisiert) werden, wobei vermerkt wird, dass es sich um eine hieratische Schreibung handelt. Kopfschmerzen, weil nicht in das Datenbankschema passend, bereiten mir solche Texte, in denen abwechselnd längere demotische und hieratische Passagen einander folgen, die aber insgesamt ein Ganzes bilden. Es sind das zum Glück nicht viele Texte, aber der mit Abstand umfangreichste und wichtigste ist das Wiener

---

[21] Col. XXI, 12 – XXII, 7 (Stele vom Jahr 14 Psammetichs I.); XXII 9 – XXIII 9 (Stele vom Jahr 34 desselben Herrschers); vgl. GRIFFITH, F. LL., *Catalogue of the Demotic Papyri in the John Rylands Library Manchester*, Manchester / London 1909, I, pl. XLIII-XLV (Photos); II, Pl. 38-40; letzte Übersetzungen HOFFMANN & QUACK, *Anthologie*, 49-51; AGUT-LABORDÈRE & CHAUVEAU, *Héros, magiciens et sages*, 193-195.

[22] GALLO, P., *Ostraca demotici e ieratici dall'archivio bilingue di Narmouthis* II *(nn. 34-99)*, Pisa 1997, 30-31, Nr. 49.

Balsamierungsritual für den heiligen Apis (Abb. 3).[23] Hier müsste noch nach einer Möglichkeit gesucht werden, diesen gemischt hieratisch-demotischen Text im Thesaurus zu präsentieren. Was für eine Edition kein Problem ist, ist sehr wohl eines für eine Datenbank, in der bisher schon aus technischen Gründen unvermeidlicherweise nach Demotisch einerseits und Hieroglyphenägyptisch sowie Hieratisch andererseits geschieden werden muss.

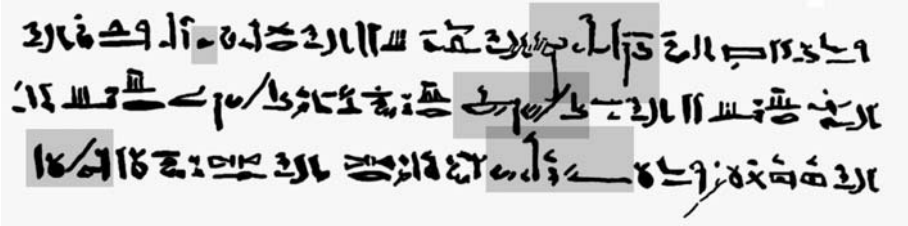


Abb. 3a: Aus pWien 3873 (Apisritual), Verso III 15-17; demotisch geschriebene Passagen grau hinterlegt (nach unnumerierter Tafel bei VOS, Embalming Ritual).

(15) *sbn r-ḥrj ỉw nꜣ* **nm.w r-ḫ pꜣ ntj sḫ** *ỉw=w ˁq p(ꜣ) nṯr ỉw=w wrḥ* **pꜣ** *nṯr n sqn (= sgn) ỉw=w* (… … …)
(16) *r-bnr ḥr p(ꜣ) nṯr ỉw=w dj.t* **wrs** ˁꜣ *ḥr tꜣ šnb.t* **wrs** *ḥm ḥr p(ꜣ) ḥr* (… … …)
(17) *ỉw=w tštš=f (n) sbn* **ṯꜣj tꜣ ḥꜣ.t r pꜣ** *ḥꜣ ṯs-pḥr ỉw=w ṯs nꜣ ḫbs.w* **skr.w** (… … …)

Übersetzung:
[…] (15) die *seben*-Binde darauf, indem die **Tragstangen** (so) sind, **wie es (vor)geschrieben ist.** Man soll den Gott eintreten lassen, man soll **den** Gott mit Salbe salben, man (… … …), (16) außen unter dem Gott. Man soll (ihm) eine **große Kopfstütze** unter die Brust und eine kleine **Kopfstütze** unter das Gesicht geben (… … …) (17) Man soll es mit *seben*-Binden befestigen **von vorne bis** hinten und umgekehrt und die ***seker*-**Bandagen knoten (… … …)

Abb. 3b: Aus dem Wiener Apisritual (pWien 3873, Verso III 15-17); demotisch geschriebene Passagen in Fettdruck.

[23] VOS, R. L., *The Apis Embalming Ritual P. Vindob. 3873*, OLA 50, Leuven 1993.

DAS HETHITOLOGIE PORTAL MAINZ

GERNOT WILHELM

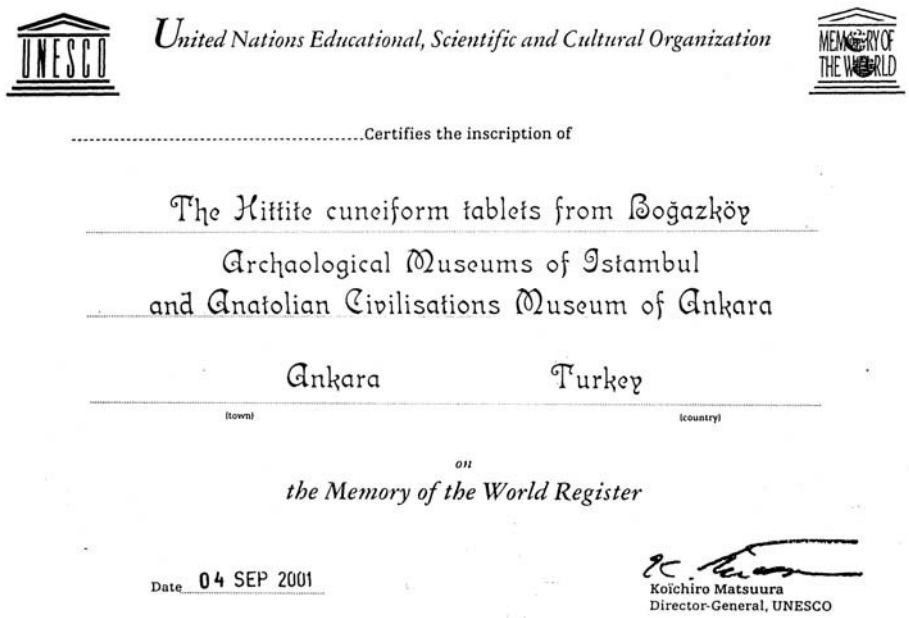
1. Texte der Hethiter und ihre Edition

In Ḫattuša, der beim heutigen Landstädtchen Boğazkale, früher Boğazköy – „Schluchtdorf“ – 160 km östlich von Ankara gelegenen Hauptstadt des hethitischen Reiches, wurden bei den 1905 begonnenen und noch heute andauernden Ausgrabungen ca. 30.000 Fragmente von Keilschrifttafeln gefunden. Die Mehrzahl der Texte ist in hethitischer Sprache abgefasst, doch fanden sich auch solche in anderen Sprachen (Sumerisch, Akkadisch, Hattisch, Hurritisch, Luwisch, Palaisch).

Die Texte umfassen:

- die wichtigsten und vielfältigsten Quellen für die politische Geschichte des mittel- und spätbronzezeitlichen Alten Orients sowie die Sprachen, Literaturen, Kulte, Mythen, Orakelpraktiken, Gesetze etc. Anatoliens im 2. Jahrtausend v. Chr.;
- die in ihrem Umfang und ihrer Detailliertheit einzigartigen Quellen für religiöse und magische Rituale des 2. Jahrtausends v. Chr.;
- die beinahe einzigen Quellen für den ältestbezeugten Zweig der indogermanischen Sprachenfamilie;
- die einzigen bzw. die umfangreichsten Quellen für weitere frühe, schon im 2. Jahrtausend v. Chr. ausgestorbene Sprachen wie Hattisch, Hurritisch und Palaisch.

Die Bedeutung der Tontafelfunde aus der Hethiterhauptstadt wird durch ihre 2001 erfolgte Aufnahme in das „Memory of the World Register“ der UNESCO bestätigt.


United Nations Educational, Scientific and Cultural Organization

MEMORY OF THE WORLD

.....Certifies the inscription of

The Hittite cuneiform tablets from Boğazköy

Archaological Museums of Istambul
and Anatolian Civilisations Museum of Ankara

Ankara (town) Turkey (country)

on
the Memory of the World Register

Date 04 SEP 2001

Koïchiro Matsuura
Director-General, UNESCO

Abb. 1: UNESCO-Urkunde

Ca. 11.000 Tafeln und Tafelfragmente stammen aus den 1905-1912 von Hugo Winckler und Theodor Makridi durchgeführten Grabungen. Bei den nach längerer Unterbrechung 1931 unter der Leitung von Kurt Bittel wieder aufgenommenen und unter mehreren Grabungsleitern (Peter Neve, Jürgen Seeher, Andreas Schachner) bis heute fortgeführten Grabungen wurden bisher fast 18.000 Fragmente von Keilschrifttafeln gefunden. Dazu kommen zahlreiche Fragmente, die nicht in den regulären Ausgrabungen entdeckt wurden und sich in verschiedenen Museen und Privatsammlungen befinden.

Die traditionelle Editionsform von Keilschrifttexten ist die „Handkopie“ („Autographie“), eine Abzeichnung, bei der die dreidimensionalen Eindrücke der Keilschrift zweidimensional reduziert werden, womit bereits ein erster, wichtiger Teil der Texterschließung geleistet ist. Bisher liegen 130 Bände mit Autographien von Texten aus Ḫattuša vor, die größtenteils in den beiden Reihen *Keilschrifttexte aus Boghazköi* (KBo, 1916ff.) und *Keilschrifturkunden aus Boghazköi* (KUB, 1921ff.) erschienen sind.

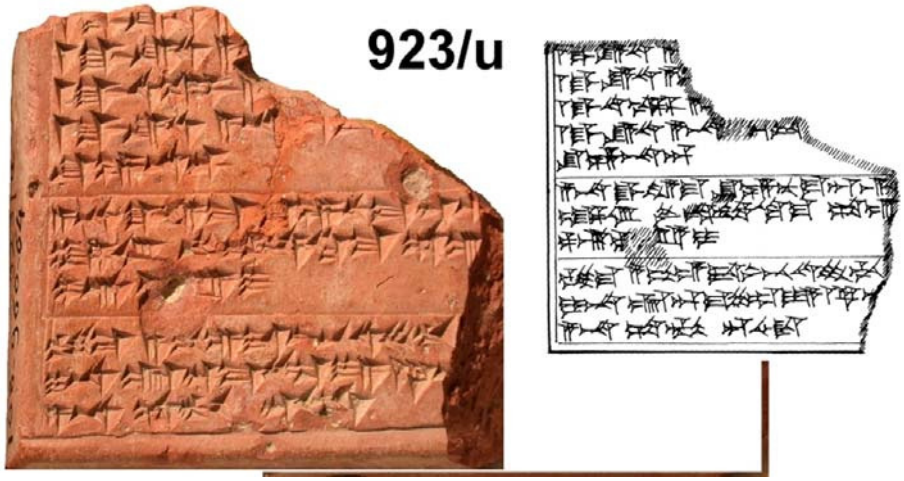


Abb. 2: Photo und Abzeichnung eines Fragments

2. Das Akademie-Projekt „Hethitische Forschungen“

1936 wurde der Hethitologe Heinrich Otten (1913-2012; s. Abb. 3) als „Grabungsphilologe“ Mitglied der Boğazköy-Expedition mit der Zuständigkeit für die Publikation aller Texte einschließlich der in den jährlichen Grabungskampagnen zu Tage kommenden Neufunde. Hierfür baute er eine lexikalische Zettelsammlung auf, die er – zunächst in Berlin, dann in Marburg – kontinuierlich erweiterte (heute mehr als 1 Mio. Zettel) und um eine gleichfalls stetig wachsende Photosammlung ergänzte. Bald nach seiner 1959 erfolgten Aufnahme in die Akademie der Wissenschaften und der Literatur, Mainz, wurde sein Arbeitsgebiet unter der Bezeichnung „Hethitische Forschungen“ als Mainzer Akademievorhaben mit Sitz in Marburg etabliert und zunächst mit DFG-Mitteln gefördert. 1979 wurde es als Langzeitprojekt in das Akademienprogramm übernommen und wenig später nach Mainz, an den Sitzort der Akademie, verlagert.

Abb. 3: Die lexikalische Zettelsammlung des Akademie-Projekts

Ungeachtet seiner allgemeiner formulierten Bezeichnung war die zentrale Aufgabe des Projekts „Hethitische Forschungen“ die Edition der seit 1931 bei den deutschen Ausgrabungen in Ḫattuša/Boğazköy, Türkei, gefundenen Texte. Dieser Teil des Projekts wird voraussichtlich entsprechend der Planung mit dem Laufzeitende 2015 zum Abschluss gekommen sein. Ein zweiter Aufgabenbereich des Akademie-Projekts, nämlich die Erarbeitung von Texteditionen in Umschrift und Übersetzung, ist mit der Überführung des Projekts ins digitale Zeitalter in ein grundlegend neues Stadium getreten.

3. Gründung und Ziele des „Hethitologie Portal Mainz“

Nach der Übernahme der Projektleitung 2001 begründete Verf. ein von der DFG gefördertes Projekt mit der Bezeichnung „Informationsinfrastruktur für digitale Publikation keilschriftlicher Staatsverträge der Hethiter und für darauf bezogene netzbasierte Forschungskooperation“. Der Hauptmitarbeiter des DFG-Projekts war Gerfrid Müller, der seine Tätigkeit anschließend in der Akademie fortsetzte und dem die Programmierung und laufende Optimierung aller Angebote des Portals zu verdanken ist.

Durch die Zusammenführung des 2007 abgeschlossenen DFG-Projekts mit dem Akademie-Projekt wurde dieses in erheblichem Maße

ergänzt und modernisiert. Das damit geschaffene und – auch in Zusammenarbeit mit Kooperationspartnern im In- und Ausland – stetig ausgebaute *Hethitologie Portal Mainz* (http://www.hethiter.net) ist aus der hethitologischen Forschung und Lehre nicht mehr fortzudenken.

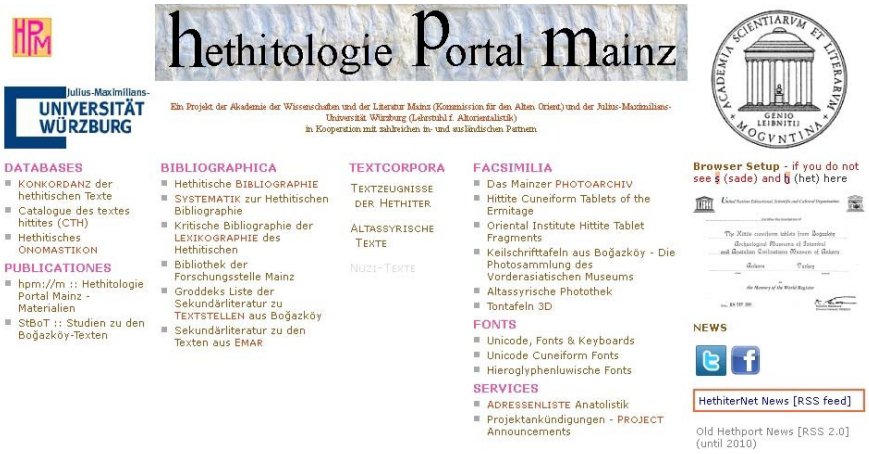


Abb. 4: Startseite des Hethitologie Portals Mainz (2012)

4. Die Konkordanz hethitischer Texte

1985 begann Silvin Košak im Rahmen des Akademieprojekts „Hethitische Forschungen“ mit dem außerordentlich arbeitsaufwändigen Projekt einer auf 15 Bände veranschlagten Konkordanz der Keilschrifttexte der Hethiter. Sie sollte alle Tafeln und Fragmente mit Texten der Hethiter (unabhängig von der Sprache) nach Inventarnummern aufführen und Textzusammenschlüsse, Editionen, Datierungen nach dem Schriftduktus, Gattungs- und Inhaltsbestimmungen nach E. LAROCHE, *Catalogue des textes hittites* (CTH), Fundortangaben nach Edition oder Grabungsdokumentation sowie Behandlungen in neuerer Forschungsliteratur angeben. Bis 2000 erschienen vier Bände im Druck. Es zeigte sich bald, dass die Konkordanz steter Verbesserung und Ergänzung bedurfte. Der zuerst erschienene nur acht Jahre alte Band war im Jahr 2000 schon stark ergänzungsbedürftig. Auch war die gedruckte Form für die schnelle Ausschöpfung des Erkenntnispotentials der Konkordanz unzulänglich.

Da die Daten in eine vom Trierer Kompetenzzentrum konfigurierte Datenbank eingegeben worden waren, wurde im Rahmen des DFG-Projekts von Gerfrid Müller eine Ausgabemaske programmiert, die

multiple Recherchemöglichkeiten bietet. 2001 ist die Konkordanz ins Netz gestellt und seitdem ständig ergänzt und verbessert worden.

Abb. 5: Abfragemaske der Konkordanz

Für die hethitologische Forschung hat sie sich mittlerweile zum Zentralbereich des Portals entwickelt, von dem aus zahlreiche andere Angebote angesteuert werden, die zunehmend durch neue Arten von arbeitserleichternden Links verknüpft sind. In Zusammenarbeit mit den Nutzern konnte die Suchmaske ständig optimiert und erweitert werden. So ist es nun z.B. möglich, sehr schnell zu ermitteln, welche Fragmente aus der Tafelsammlung x zur Textgattung y gehören und in mittelhethitischem Duktus geschrieben sind. Die Antwort auf diese Frage wäre zuvor allenfalls nach langwieriger Arbeit zu beantworten gewesen.

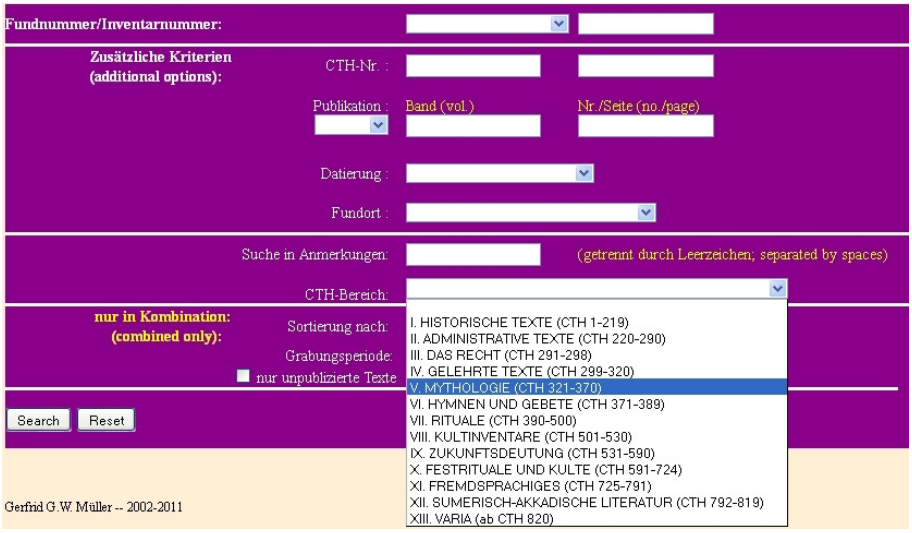


Abb. 6: Abfragemaske der Konkordanz mit geöffneter Liste der Textgattungen

Die Keilschrifttafeln aus Boğazköy sind meist stark fragmentiert, und ein inhaltlich verwertbarer Text kommt oft erst durch die mühevolle Gewinnung von Fragmentzusammenschlüssen („Joins“) zustande. Da ständig neue Joins entdeckt werden, bieten die gedruckt vorliegenden Editionsbände sehr oft nur die Einzelfragmente. Die Joinangaben der Konkordanz repräsentieren zwar stets den aktuellen Kenntnisstand, jedoch ist es bei einer Angabe wie z.B. „350/z + 895/z + Bo 992 + Bo 1187 + ...“ schwierig und mühevoll, die direkten Anschlüsse dieser Fragmente untereinander festzustellen, wenn nur Autographien der Einzelstücke vorliegen. Hier sind die von Silvin Košak in den letzten Jahren hergestellten „Joinskizzen“ eine große Hilfe.

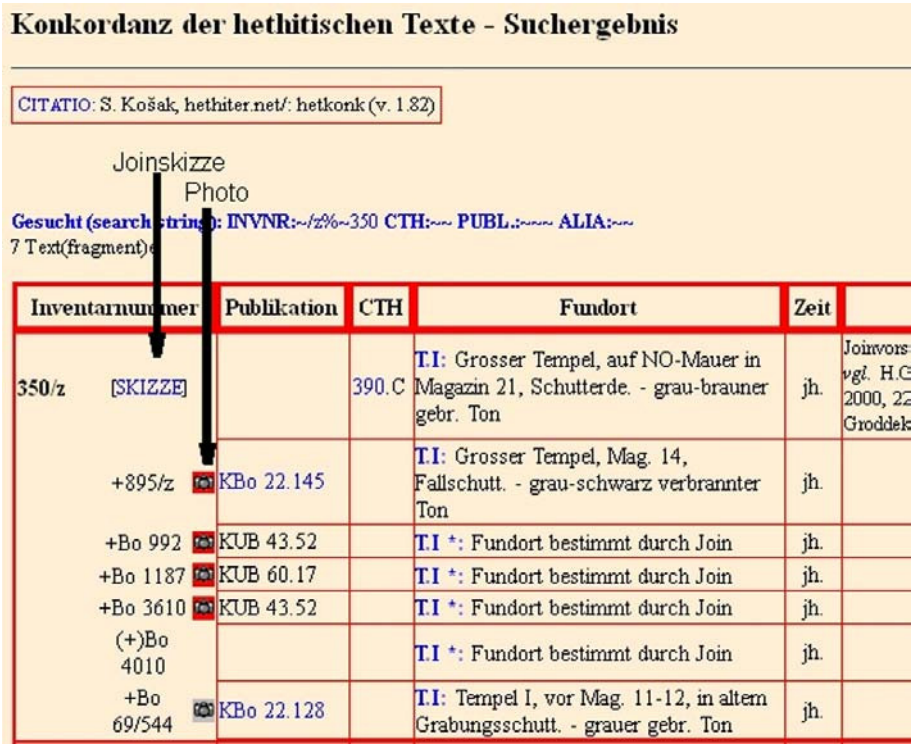


Abb. 7: Abfragemaske der Konkordanz mit Hinweis auf Joinskizzen

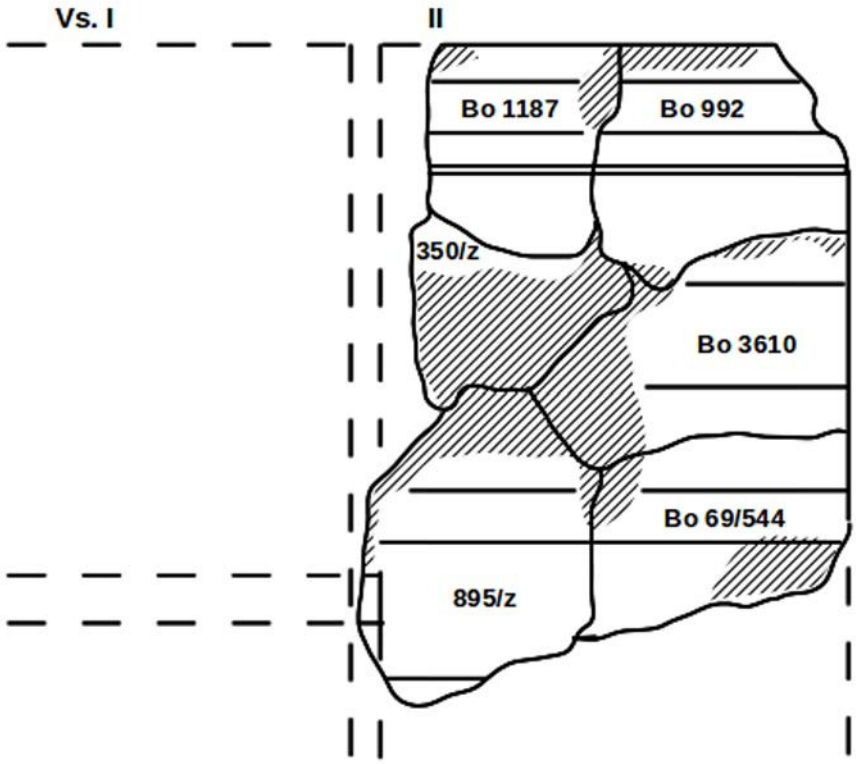


Abb. 8: Beispiel einer Joinskizze

Da es sich als praktisch erwiesen hat, die Materialien, die im „Portal“ bereitgestellt werden, von Zeit zu Zeit in gedruckter Form zu fixieren, soweit sich dies anbietet, wurde 2005 eine neue Reihe „Hethitologie Portal Mainz – Materialien“ (HPMM) begründet, in der die Konkordanz erschienen sind. Ein Jahr darauf wurde diese Druckfassung auch als PDF über das Portal kostenfrei bereitgestellt.

Seit der Buchpublikation der Konkordanz (Version 1.0) wurden zahlreiche Updates vorgenommen (derzeit Version 1.84).

5. Das digitale Photoarchiv

Die Sammlung von Photographien hethitischer Keilschrifttafeln in der Arbeitsstelle des Akademie-Projekts wurde seit 2001 fortschreitend digitalisiert und im Umfang von ca. 35.000 Bilddatensätzen über das Hethitologie Portal Mainz online verfügbar gemacht. Im Jahre 2009 schloss die Akademie mit der Stiftung Preußischer

Kulturbesitz eine Vereinbarung ab, der zufolge die im Vorderasiatischen Museum aufbewahrte, in den 30er Jahren entstandene Photosammlung von ca. 15.000 Photos ebenfalls digitalisiert und katalogisiert werden konnte. Die zahlreichen Aufnahmen mit mehreren Fragmenten wurden in Einzelaufnahmen der Fragmente zerteilt und jeweils mit Maßstab und Signaturen versehen. Dadurch ergaben sich ca. 37.000 Bilddatensätze, die inzwischen gleichfalls mit der Konkordanz verlinkt und über das Portal einsehbar gemacht wurden.

Die Photos sind insbesondere für die Datierung von Texten auf Grund paläographischer Analysen und die Verifikation von Textzusammenschlüssen („Joins“) wichtig.

Die Publikation der Photos hat auch einen erheblichen Ausbildungswert für den wissenschaftlichen Nachwuchs der Hethitologie, der nun die Möglichkeit hat, seine Kompetenz im Umgang mit Originaltafeln frühzeitig zu entwickeln, was mit der ausschließlichen Verwendung von Autographien, die ja in gewissem Maße bereits eine Interpretation des Befundes darstellen, nicht erreicht wird. Auch ist die Quellengrundlage für paläographische Forschungen nun allgemein zugänglich.

6. Digitale Edition von Keilschrifttexten der Hethiter

Die Präsentation von Quellentexten aus altorientalischen Keilschriftsprachen bietet besondere Probleme, die sich in dieser Form in den meisten anderen Philologien nicht stellen. Sie ergeben sich aus der Besonderheit des Schriftträgers (Tontafel), der Schriftart (Keilschrift), dem fragmentarischen Zustand der Überlieferung, dem steten Quellenzuwachs und dem raschen Forschungsfortschritt in Grammatik und Lexikon.

Die Edition in (Lateinschrift-)Transkription hat sich traditionell an Editionen lateinischer, griechischer und hebräischer Texte orientiert. Dabei wird zumeist ein gut erhaltenes Textexemplar als Haupttext definiert, während die Varianten der anderen Exemplare im Apparat aufgeführt werden. Da die Keilschrifttafeln der Hethiter nur in sehr seltenen Fällen vollständig erhalten sind, müssen bei dieser Editionsform Textabschnitte aus unterschiedlichen Exemplaren als „Haupttext“ herangezogen werden. Hieraus ergibt sich eine große Unübersichtlichkeit, wenn es um die Bestimmung der Charakteristika der einzelnen Exemplare geht.

Besonders unzulänglich ist diese Methode für graphematische und grammatikalische Untersuchungen, da aus dem Variantenapparat nicht hervorgeht, ob ein dort berücksichtigtes Exemplar in einem

bestimmten zu untersuchenden Phänomen mit dem Haupttext übereinstimmt oder ob die entsprechende Textstelle in dem betreffenden Exemplar abgebrochen ist.

Aus diesem Grund wird für die Edition von Texten, die in mehreren Exemplaren erhalten sind, seit einigen Jahrzehnten zunehmend eine zeilensynoptische Editionsform („Partitur“) bevorzugt, in der die einzelnen Exemplare voll ausgeschrieben und in übersichtlicher Weise zeilenweise untereinander gestellt werden. Meist wird zusätzlich als graphisch hervorgehobene Textzeile ein aus allen Exemplaren gewonnener Text in zusammenhängender (nicht wie bei den Einzelexemplaren syllabischer) Umschrift (*master text*) geboten.

Dieses Verfahren bietet gegenüber der herkömmlichen Editionsform Vorteile, ist aber teurer. Es bietet nicht die Möglichkeit, den *master text* oder ein einzelnes Exemplar bequem in größeren Stücken zu überblicken, falls sie nicht noch einmal gesondert geboten werden.

Hier eröffnet die digitale Edition neue Möglichkeiten, die durch das Hethitologie Portals Mainz erstmals in der Altorientalistik genutzt werden. Hier ist ein in mehreren Exemplaren bezeugter Text sowohl zeilensynoptisch als auch im Textzusammenhang des *master text* und der einzelnen Exemplare einsehbar. Als Beispiel sei eine Stelle aus dem Staatsvertrag Muršilis II. mit Manapa-Tarḫunta vom Šeḫaflussland (CTH 69) gewählt. Abb. 9 zeigt einen Ausschnitt aus der zeilensynoptischen Darstellung. Da in den einzelnen Exemplaren hethitischer Texte die Zeilen oft unterschiedlich umbrochen werden, trennt die Textwiedergabe in der „Partitur“ den Text nach Kola, die fortlaufend nummeriert werden (im angegebenen Beispiel Kola 140-141). Um den Text an den Keilschriftabzeichnungen kontrollierbar zu halten, werden die zusammengesetzten Fragmente in ihrer Abfolge von links nach rechts bezeichnet (hier z.B. B1 + 4 + 5 + 2) und die Zeilennummerierung der keilschriftlichen Edition des jeweiligen Fragments ebenso angegeben.

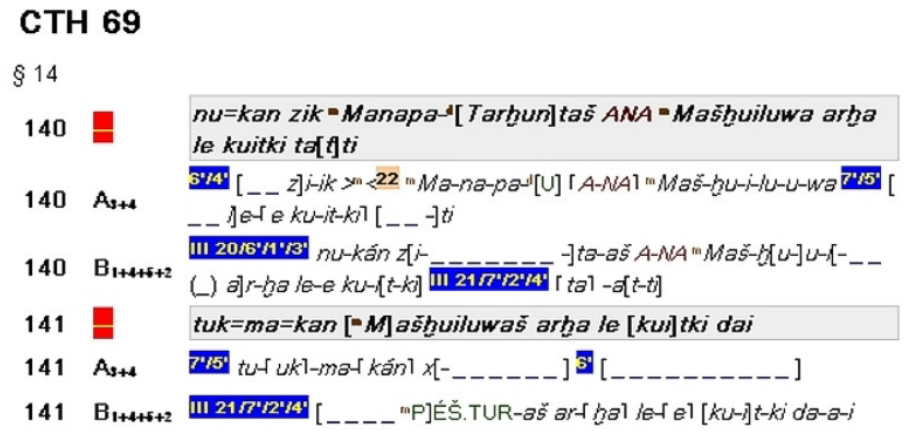


Abb. 9: Zeilensynoptische Wiedergabe („Partitur“) eines in zwei Exemplaren erhaltenen Textes.

Die Wiedergabe eines einzelnen Exemplars wird durch Anklicken des entsprechenden Sigles (hier: A oder B) erreicht. Hier wird die Zeile wie auf der Tafel umbrochen. Die Kolon-Nummer wird im Text in hochgestellten weißen Ziffern auf schwarzem Grund angegeben.

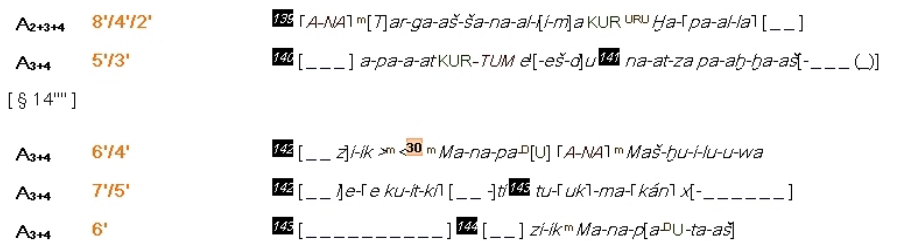


Abb. 10: Wiedergabe eines Exemplars

Für manche Zwecke (z.B. schnelle inhaltliche Übersicht) ist es nützlich, nur den Mastertext aufzurufen.

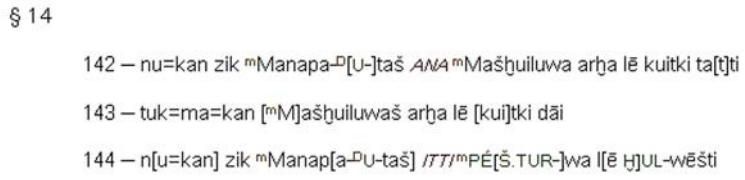


Abb. 11: Wiedergabe nur des master text

Im Rahmen des Akademie-Projekts werden die Staatsverträge der Hethiter und andere historische Texte in dieser Weise ediert. Die Edition anderer Textgruppen haben Kooperationspartner übernommen, die nach denselben Prinzipien und mit denselben Kodierungen arbeiten, wie sie im Akademie-Projekt entwickelt wurden. So werden an der Universität Mainz unter der Leitung von Doris Prechel hethitische Beschwörungsrituale ediert und an der Universität Marburg unter der Leitung von Elisabeth Rieken die hethitischen mythologischen Texte (abgeschlossen) und die hethitischen Gebete.

7. Weitere wissenschaftliche Serviceleistungen des Portals

Das Portal bietet über die genannten Zentralbereiche (Konkordanz, Joinskizzen, Photosammlung, digitale Textedition) zahlreiche Hilfsmittel, die teilweise innerhalb des DFG-Projekts erarbeitet, überwiegend aber von Kooperationspartnern zur Verfügung gestellt wurden.

Hier ist zunächst eine umfassende digitale Bibliographie zu nennen, die mit Mitteln der DFG in Zusammenarbeit mit Jana Siegelová-Součková (Prag) und Massimiliano Marazzi (Neapel) erstellt und 2004 über das Portal benutzbar gemacht wurde. Seitdem sorgt M. Marazzi mit seinen Mitarbeitern für die laufende Ergänzung. Dazugetreten ist eine von J. Siegelová-Součková (Prag) und Gerfrid Müller (AWL Mainz) erarbeitete systematische Bibliographie. Für die Erschließung der Forschungsliteratur ist von hohem Wert auch ein von Detlev Groddek erstellter ca. 250.000 Einträge umfassender Index zu in der Forschungsliteratur zitierten Stellen aus Texten der Hethiter. Dasselbe gilt für einen lexikographischen Index zur neueren Forschungsliteratur, den M. Marazzi aufgebaut hat und laufend ergänzt.

S. Vanséveren, Leuven, und G. Anders, Reinach/Schweiz, haben hethitische Keilschrift- bzw. Hieroglyphenfonts entwickelt, die über das Portal angeboten werden. Im Rahmen des Projekts sind schon frühzeitig UNICODE-basierte Fonts erstellt worden, die die in der Altorientalistik üblichen Sonderzeichen umfassen und frei verfügbar sind (SemiramisUnicode).

8. Akzeptanz

Mit dem Hethitologie Portal Mainz ist ein nunmehr kaum wegzudenkendes Forschungshilfsmittel entstanden, das die internationale Bereitschaft zu Kooperation und Kontribution auf dem Gebiet der

Hethitologie in eindrucksvoller Weise angeregt hat und für weitere Textgruppen angewandt werden wird. „Some years ago on the internet the so-called Hethitologie Portal Mainz was introduced ..., which made available for everyone old and new information about published and unpublished Hittite texts and fragments. The project of digitalization of photographs of tablets from Boğazköy-Ḫattuša is extremely attractive and useful. Other projects are added regularly and studying Hittite is unthinkable without the Portal.“ (J. DE ROOS, in: *Bibliotheca Orientalis* 64, 2007, 187).

THE TITUS PROJECT
25 YEARS OF CORPUS BUILDING IN ANCIENT LANGUAGES

JOST GIPPERT

The article summarizes the contents and the structurtal premises of the "Thesaurus Indogermanischer Text- und Sprachmaterialien" (TITUS), focussing on search functions and facilities and questions of the encoding of ancient languages written in various scripts. Examples are taken from Tocharian, Greek, Vedic Sanskrit, and other ancient Indo-European languages covered by TITUS.

In September 1987, a group of Indo-Europeanists decided to join efforts in the digitization of primary sources that are essential for their research, by creating a common pool of the electronic texts to be prepared. Eversince,[1] the text pool has developed into a comprehensive retrieval system covering a large amount of relevant materials. The scope, the contents and the structural premises of the "Thesaurus Indogermanischer Text- und Sprachmaterialien" (TITUS) are summarized in the following pages.[2]

1. Since its foundation, the primary goal of the TITUS project consisted in the compilation of a comprehensive text database of ancient Indo-European languages that were not covered by concurrent projects such as the *Thesaurus Linguae Graecae*.[3] To reach this aim, a practical way of cooperation was decided upon: everybody who was able to contribute to the database was granted, as a member of the TITUS team, access to the complete database. In the 1980ies, this still presupposed data exchange via floppy or, later, compact disks, as internet facilities were not yet available in a sufficient way. Nevertheless, as early as 1988 the complete text of the Old Indic Rigveda Samhita, which had been electronically prepared as a text file of ca. 1.5 MB by H.S. Ananthanarayana under the supervision of W.P. Lehmann in Austin/Texas, was successfully transferred via a data line from the USA to the Berlin Free University, which hosted

---

[1] The project was announced under the title "Thesaurus altindogermanischer Textcorpora auf Datenträgern" in: *Die Sprache* 32/2, 1987 [1988], 151t.

[2] For previous accounts of the TITUS project cf. GIPPERT (1995a; 1995b; 1996; 2001; 2010).

[3] Project of the University of California at Irvine; cf. http://stephanus.tlg.uci.edu/.

the data pool then. By 1994, when the facilities of the internet emerged, the exchange of data was put on an online basis by establishing an FTP server at the University of Frankfurt, and soon after, the first web pages of the project were launched under the new name of "TITUS" which had meanwhile been agreed upon by the members.[4] Since 1996, the TITUS project has been promoting the use of Unicode to ensure a reliable encoding of its data, and the independent web server of the project established then[5] was one of the first sites world-wide to make a considerable amount of textual data available in this way of encoding. Thanks to a generous grant of the WordCruncher Company, the project was able in 1997 to install, along with its web site, a special "WordCruncher" server for the search and retrieval of data from the database. This service has been maintained until recently but has now been given up as most of the facilities it provides have meanwhile been implemented in an SQL-based online retrieval engine that has been publicly accessible since 2000.[6] Today, the TITUS database comprises not only corpora of ancient Indo-European languages such as Avestan, Vedic Sanskrit, Phrygian, or Umbrian, many of them covering the complete textual heritage of the languages involved, but also materials in more recent Indo-European as well as neighbouring languages, among them the largest corpus of Old and Middle Georgian available world-wide.[7] Many of the TITUS corpora have been the basis for more specialized corpus projects such as, e.g., the Referenzkorpus Altdeutsch project,[8] which aims at a full annotation of all textual materials in Old High German and Old Saxon; the Sanskrit Library project at Brown University, which aims at providing grammatical and other information pertinent to Sanskrit texts;[9] or the National Corpus of the Georgian Language, an international project aiming to cover the complete

[4] The clumsy URL was http://www.rz.uni-frankfurt.de/home/ftp/pub/titus/public_html/.

[5] URL: http://titus.uni-frankfurt.de/.

[6] URL: http://titus.fkidg1.uni-frankfurt.de/search/query.htm.

[7] Cf. http://titus.uni-frankfurt.de/texte/texte.htm for a full account of available corpora and texts.

[8] A common project of the universities of Berlin (Humboldt), Frankfurt and Jena, financed by the Deutsche Forschungsgemeinschaft since 2009 and part of the initiative "Deutsch-diachron-digital"; cf. http://www.deutschdiachrondigital.de.

[9] Cf. http://sanskritlibrary.org.

written history of Georgian from the 5th century A.D. up to the present day.[10]

2. With the establishment of the WordCruncher server in 1997, the TITUS project has moved far away from its original concept of being a mere exchange base of text files. Instead, the focus has shifted towards providing sophisticated search facilities within and across the text corpora, thus supporting online research into the languages and literatures in question. A few examples may suffice to illustrate the facilities that have been developed meanwhile.

2.1 One of the Indo-European languages for which TITUS may claim to cover the complete textual heritage in its corpus, is Tocharian, a language that was spoken in two different varieties in East Turkestan in the first millennium of our era. The textual remnants of the two Tocharian varieties (East- or A- and West- or B-Tocharian) are contained in ca. 5,000 manuscripts written in a "Northern" type of Brahmi script that were found in a region extending from Kucha to Turfan and Dunhuang along the Silk Road.[11] The largest part of these manuscripts is preserved in the Turfan Collection of the Berlin-Brandenburg Academy of Sciences (BBAW) today (ca. 4,000 manuscripts);[12] other major collections are hosted in London, Paris, and St. Petersburg. Within the TITUS project, work on the Tocharian manuscripts started in 1996 with the digitization of the printed editions of A- and B-Tocharian texts of the Berlin collection (in Romanized transcription), which formed the foundation of the emerging corpus. In the same year still, TITUS and the BBAW agreed upon preparing a complete set of digital images of the Tocharian manuscripts from Berlin to provide them online along the transcribed texts;[13] this endeavour, which was kindly supported by T. Tamai, resulted, in 2000, in one of the first frame-based online editions providing images, transcribed texts, and metadata as to each manuscript side by side. Today, this online-edition comprises the complete set of Berlin manuscripts including the ca. 3,000 hitherto unpublished

[10] Cf. http://georgiannationalcorpus.ac.ge.

[11] Cf. http://titus.uni-frankfurt.de/didact/karten/turkstan/turkst.htm for a map showing the locations.

[12] Cf. http://www.bbaw.de/en/research/turfanforschung as to the Turfan Studies project of the BBAW.

[13] Cf. GIPPERT (1997 and 1998) as to the technical foundations of the digitization project.

fragments, all manually transliterated by T. Tamai (cf. Fig. 1 showing a screen-shot of the site).[14]

2.2 In parallel to the online edition of the Berlin collection, which provides access to the corpus only via the catalogue number of a given manuscript,[15] the Tocharian data of all major collections have been prepared for a word-form based retrieval via the TITUS search engine.[16] This is built upon a more fine-grained referencing system where every single line of a manuscript (page) can be addressed directly (cf. Fig. 2 showing line 3 of the recto of the A-Tocharian fragment THT 634, which is part no. 1a in the edition by SIEG & SIEGLING 1921). To facilitate investigations into the paleography of the Brahmi script used for Tocharian, each line is further provided with an *akṣara*-based transliteration alongside the "normal" word-based transcription (as visible in Fig. 2). On the basis of a preindexation of the complete corpus, this allows for searching for both word-forms and individual *akṣara*s, either by clicking upon an item as displayed in the text or by using a query form. E.g., clicking upon the word-form *kumseñc* 'they come' in the given text line invokes (via a javascript underlying the word) a query for all (eight) occurrences of the same word-form throughout the A-Tocharian corpus,[17] which is output as a list of keyword-in-context entries with full referentiation of the text passages in question (cf. Figs. 3 and 4). Each entry is linked to the corresponding text passage so that the wider context can be accessed at will (cf. Fig. 5 showing the context of THT 935 = 302a, line 5). Note that in the transliteration, "Ä\" stands for the combination of the diaeresis-like vowel mark (which usually stands

[14] Cf. http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/thtframe.htm; most elements of the edition were prepared in cooperation by K. Kupfer and T. Tamai.

[15] In the TITUS edition, the Berlin manuscripts are referenced according to their catalogue number in the Turfan Archive ("THT"). Of the 4074 manuscripts listed there, nos. 1–633 are B-Tocharian, and nos. 634–1099 are A-Tocharian (numbered 1–467 in the printed edition by SIEG & SIEGLING 1921). Several manuscripts have been missing since the Second World War; in some cases, digitized images could be provided from existing photographs.

[16] Cf. http://titus.uni-frankfurt.de/texte/texte2.htm#toch; for the time being, access to the B-Tocharian corpus, which is still under construction, is restricted to TITUS members and other registered users (cf. http://titus.uni-frankfurt.de/titusstd.htm for a form to apply for registration).

[17] The javascript causes an SQL-query to be sent to the following ASP script: http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.asp?LXLANG=58285&LXWORD=kumseF100c&LCPL=0&TCPL=0&C=H&PF=26.

for the shewa vowel rendered as *ä* in transcriptions of Tocharian) with a *virāma* in word-final position (i.e., *kumseñc*$^{\ddot{a}}_{\backslash}$ in the transcription system applied in the editions by Sieg and Siegling); in the word-based search, it is ignored. In a similar way, the so-called "Fremdzeichen" are represented by capital letters, with "A" standing for their inherent vowel (i.e., KA etc. stands for *ka̲* etc. in Sieg/Siegling's transcription); in the word-based search, *A* is treated as an equivalent of *ä* and the difference between "Fremdzeichen" and "Indian" akṣaras is ignored. The unsyllabic *u* vowel indicated by subscript *u* with a bent line above in the editions is represented by *ù* in the corpus; in the word-search, this is treated as being equivalent with plain *u*. In the *akṣara*-based transliteration, * stands for a (missing or unreadable) complete *akṣara* and +, for a (vocalic or consonantal) element of an *akṣara*; ˆ stands for a word boundary within an *akṣara* (ignored in the search). All this is warranted by a specific structure of the underlying relational database, which contains "normalised" variants of the word-forms wherever applicable (cf. Fig. 6).[18]

2.3 A more flexible and powerful query method than the hyperlink-based retrieval is provided via special input forms. In general, the TITUS "Search Engine" comprises two different methods of input-based access to its data, one yielding a list of word-forms matching the query input, and one, the keyword-in-context output of occurrences as shown above (cf. Fig. 7).[19] In both input forms, the language of the search must be determined first, either specifically (e.g., "Tocharian A" as in Fig. 8) or more generally (e.g., "Tocharian" as in Fig. 9). The word-form to be searched can then be entered either in toto or partially, in exact Unicode encoding or in a substitutional plain-ASCII representation (or in a mixed representation), and with two types of "wild cards" replacing explicit characters: the question mark, "?", stands for one single character, and the asterisk, "*", for any sequence of characters (including zero). E.g., the word-form *kumseñc* can be entered as such or in the form *kumsen~c* (with the diacritic adscripted to match ASCII-based keyboards, cf. Fig. 10), and *kumseñc* will also be found if the query string is reduced to

[18] The database used at present is IBM DB-2 Express version 9.1, a powerful and yet free SQL-based system with full Unicode support.

[19] Both query types are accessible via http://titus.fkidg1.uni-frankfurt.de/search/query.htm. – A third query type, styled "unspecified", consists of a mere link to a Google search over the TITUS site.

*k?ms*c*, i.e., with one character between *k* and *m* and any sequence of characters between *s* and *c* (cf. Fig. 11), or kums*, i.e., with any sequence of characters following *kums*, as the resulting word-list shows (cf. Fig. 12). Similarly, the word-form *NAmseñc* 'they revere', which occurs in the same line of THT 634 as *kumseñc*, can be retrieved by entering *NAmseñc*, *nämseñc*, *n*ms??c*, *näm**, etc. (cf. Fig. 13 showing the variant spellings in parentheses). In the word-list output, every word-form is provided with a hyperlink to the relevant context query; this means that by clicking upon *NAmseñc* or *nämseñc* in the list, all 3 occurrences of the word-form (the number of occurrences is indicated in square brackets for each list entry) will be listed in an extra window (cf. Fig. 14).[20] Of course, the same result can also be achieved with the "context output" form, entering, e.g., *n?mse?c* (cf. Fig. 15). If the query string has more than one match as, e.g., in the case of *kums** (cf. above), the occurrences of the different matching forms will be output in alphabetic order (cf. Fig. 16 showing first an occurrence of *kumsanträ*, 3rd pers.pl.pres.ind.med., then one of *kumse*, 3rd pers.pl.pres.ind.act., shortened form). In addition, in the case of verb forms, the header of the output list indicates the underlying root (if determinable), again provided with a hyperlink (cf. Fig. 17); this leads to a special table which illustrates for every Tocharian verb which of its paradigm forms are attested in the two dialects (cf. Fig. 18).

2.4 As was stated above, the input of search strings in the query forms can be exactly "as is", i.e., in Unicode encoding, or in a substitutional plain-ASCII format with adscript diacritics. This is true not only for the input of Latin-based scripts (or transcriptions) but also for other scripts. Thus, e.g., to search for the attestations of Greek ἄνδρα (acc.sg. of ἀνήρ 'man'), both the Greek spelling and Latin *andra* can be entered (cf. Figs. 19 to 21). Note that the entry of Greek diacritics is not necessary as unaccented variants are stored in the database for all word-forms; this means that all occurrences of ἄνδρα will also be found in a search for (less specified) ανδρα. This is even true for the same word spelt with an initial capital (Ἄνδρα) or with an acute on the word-final vowel (ἄνδρά, to be expected in the

[20] This is again invoked by a javascript which sends the SQL-query to the following URL: http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.asp?lxlang=941&lxword=N4100mseF100c&LCPL=1&TCPL=1&C=H.

position preceding a clitic), which are matched by ανδρα (and *andra*) but not by ἄνδρα (cf. Figs. 21 and 22).

2.5 A special feature of the context-related search is the "combined search" function. Up to four query patterns (word-forms, stems, etc.) can be entered in parallel for a search of co-occurrences in a given context; cf., e.g., Figs. 24 and 25 showing a combined search for *thaz* 'the' and *uuort* 'word' in the Old High German corpus. The amount of context envisaged here can be determined by the user. If the "distance" is set to "0" (the default setting), the context in question is the lowest reference level of a given text (usually a sentence, a verse or a line); in the given example, this yields 111 co-occurrences of *thaz* and *uuort*, irrespective of the order of the two words (and including spelling variants such as *tház* and *uuórt*). Setting the distance to "1 - exact" (cf. Fig. 26) yields but 33 co-occurrences, with *uuort* immediately following *thaz*. (cf. Fig. 27).

2.6 A feature that has only been implemented for Old Indic (Sanskrit) and Avestan so far is the "thesaurus search" function. Different from the word-form based queries illustrated above, this function admits of searching for complete paradigms of words irrespective of a common ("matching") string structure of the individual word-forms. Starting, e.g., from *ṛ̥tvíjo* as the genitive singular case form of the Vedic noun *ṛ̥tvíj-* 'priest' (cf. Fig. 28), the output displays all occurrences of all case forms of this word as met with in the corpus[21] beginning with the nominative singular *ṛ̥tvík*, provided with a grammatical analysis of each form[22] and a German translation of the respective lemmata[23] (cf. Fig. 29).

3. It is obvious that the latter type of retrieval presupposes a thorough modelling of the morphology of the language concerned. To implement similar facilities for all languages covered by TITUS

---

[21] Including spelling variants (caused by sandhi and accentuation), the list comprises the following forms: *ṛ̥tvik*, *ṛtvík*, *ṛ̥tvig*, *ṛ̥tvíg*, *ṛtvijam*, *ṛtvíjam*, *ṛtvijaṃ*, *ṛ̥tvíjaṃ*, *ṛtvijā*, *ṛtvíjā*, *ṛtvijaḥ*, *ṛtvíjaḥ*, *ṛtvijas*, *ṛtvíjas*, *ṛtvijaś*, *ṛtvíjaś*, *ṛ̥tvijo*, *ṛtvíjo*, *ṛtvija*, *ṛ̥tvíja*, (*ṛtvijaḥ*, *ṛ̥tvijaḥ*), *ṛtvijām*, *ṛtvíjām*, *ṛtvijāṃ*, *ṛtvíjāṃ*.

[22] The analysis was worked out by R. Gehrke in the course of the AUREA project ("Avesta und Rigveda: Elektronische Analyse") financed by the Deutsche Forschungsgemeinschaft in 1997-1999; cf. http://titus.uni-frankfurt.de/curric/aurea/aurea.htm.

[23] Based upon the dictionary by K. MYLIUS (1992).

therefore means an immense task for the project that still has to be undertaken. Another task for the future that can be envisaged today concerns improvements in the rendering of non-Latin scripts as in the case of the cuneiform inscriptions of Old Persian for which a Unicode-based encoding in the original script has recently been provided by A. Sarhadi and M. Esnaashari (cf. Fig. 30).[24] In some cases, this is still hampered by the fact that the corresponding code-points of the Unicode standard are not yet available; e.g., it would be possible now to encode the Avestan texts in the original script[25] but for Middle Persian (Pahlavi) passages that are often met with in Avestan contexts, a Unicode rendering is not yet available. As in former cases, the members of the TITUS project are ready to support the standardisation process with their expertise.

[24] Cf. http://titus.uni-frankfurt.de/texte/etcs/iran/airan/apers/apers.htm.

[25] Cf. GIPPERT (forthcoming) as to details.

BIBLIOGRAPHY

GIPPERT, J., 1995a: TITUS. Das Projekt eines indogermanistischen Thesaurus, in: *LDV-Forum* 12/1, 35-47.

GIPPERT, J., 1995b: TITUS. Von der Keilschrifttafel zur Textdatenbank, in: *Forschung Frankfurt* 4/1995, 46-56.

GIPPERT, J., 1996: TITUS – Alte und neue Perspektiven eines indogermanistischen Thesaurus, in: *Studia Iranica, Mesopotamica et Anatolica* 2, 1996 [1997], 49-76.

GIPPERT, J., 1997: Digitization of Tocharian Manuscripts. Short notice about a new project, in: *Tocharian and Indo-European Studies* 7, 265-266.

GIPPERT, J., 1998: Digitization of Tocharian Manuscripts from the Berlin Turfan Collection, in: *Manuscripta Orientalia. International Journal for Oriental Manuscript Research* 4/1, 49-57.

GIPPERT, J., 2001: Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems, in: WILLÉE, G. *et al.* (eds.), *Computerlinguistik. Was geht, was kommt? / Computational Linguistics. Achievements and Perspectives. Festschrift für Wilhelm Lenders*, Bonn 2002, 81-85.

GIPPERT, J., 2010: Manuscript Related Data in the TITUS Project, in: *Comparative Oriental Manuscript Studies Newsletter* 1, 2011, 7-8.

GIPPERT, J., forthcoming: The Encoding of Avestan: Problems and Solutions, to appear in: *Journal for Language Technology and Computational Linguistics*, 2012.

MYLIUS, K., 1992: *Wörterbuch Sanskrit-Deutsch*, 4. Auflage, Leipzig [u.a.].

SIEG, E. & W. SIEGLING, 1921: *Tocharische Sprachreste*, Band I: *Die Texte*, A: *Transcription*, B: *Tafeln*, Berlin [u.a.].

FIGURES

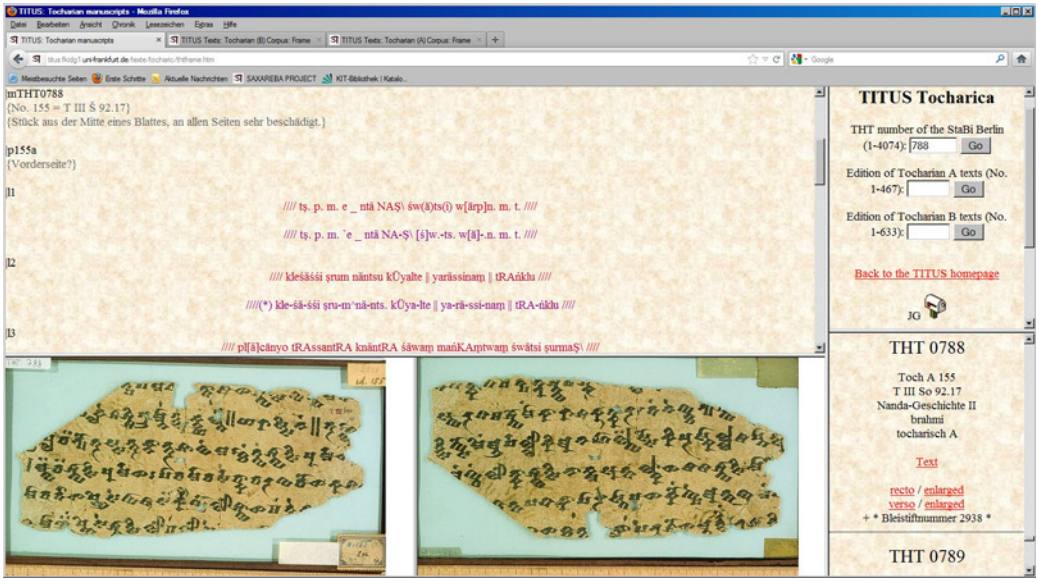


Figure 1

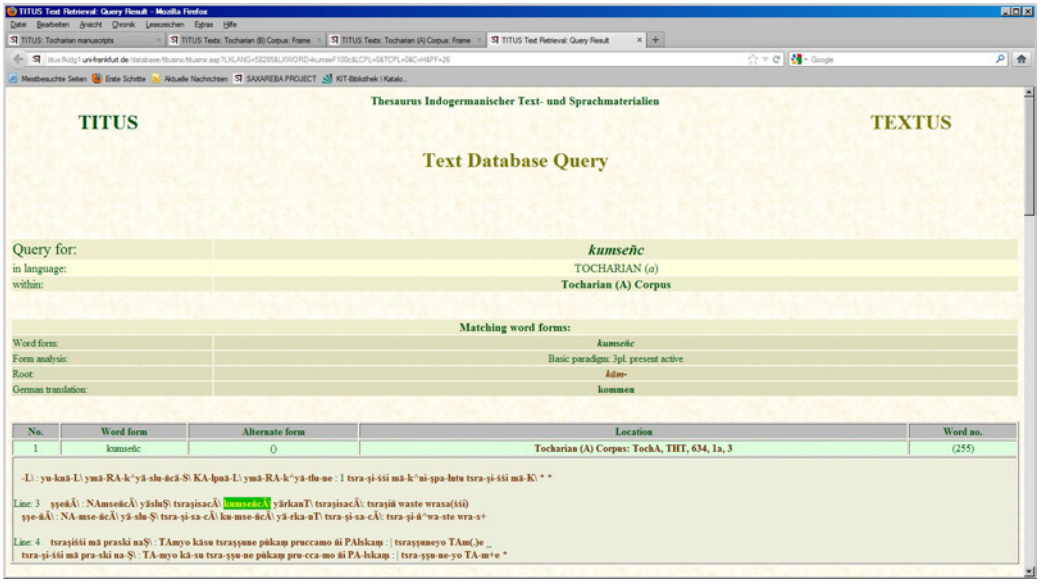


Figure 2

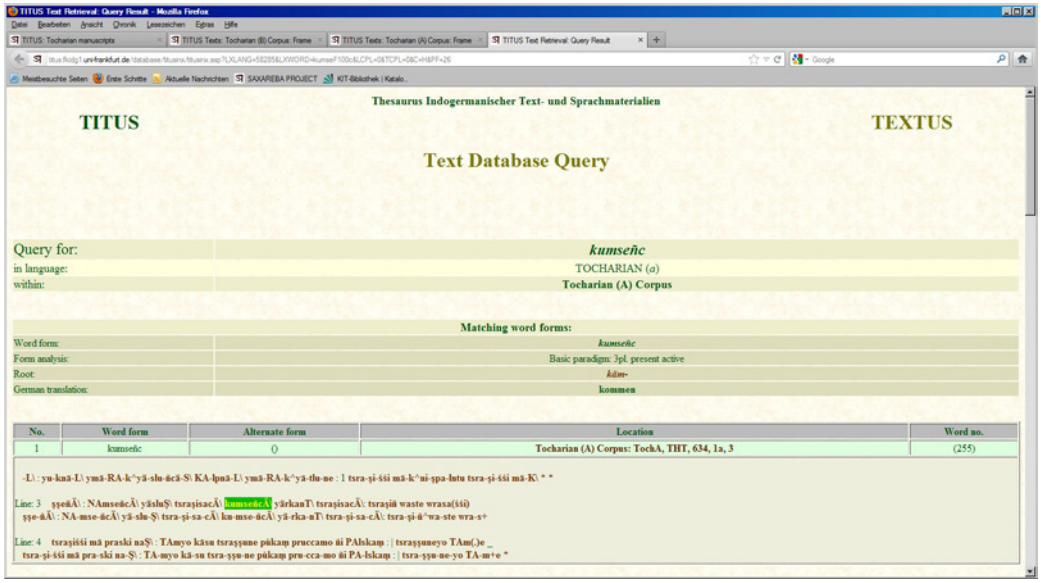


Figure 3

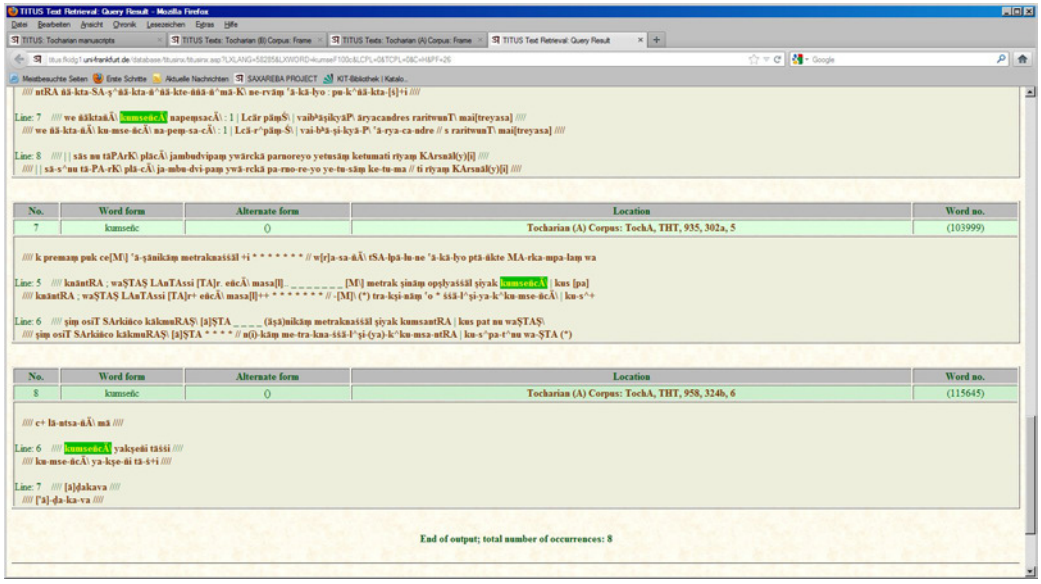


Figure 4

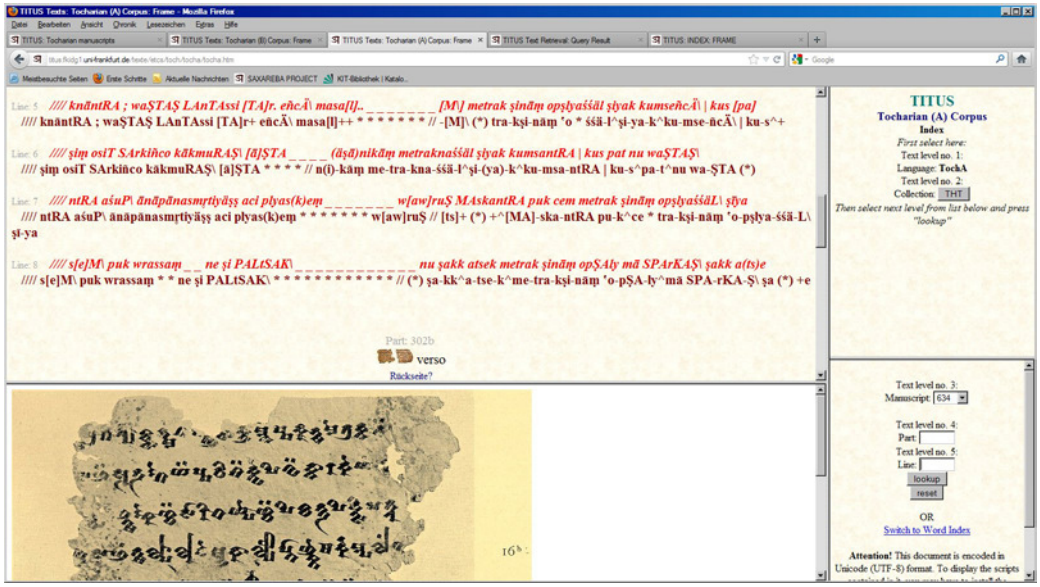


Figure 5

941	57	tlu		26	6	232
941	57	ne		26	6	233
941	57	tsra		26	6	236
941	57	ṣi		26	6	237
941	57	śśi		26	6	238
941	57	mā		26	6	239
941	57	kni		26	6	240
941	57	spa		26	6	241
941	57	lntu		26	6	242
941	57	tsra		26	6	243
941	57	ṣi		26	6	244
941	57	śśi		26	6	245
941	57	mā		26	6	246
941	57	K\	k\	26	6	247
941	57	*		26	6	248
941	57	*		26	6	249
941	56	ṣṣeñ		26	7	250
941	56	NAmseñc	nämseñc	26	7	252
941	56	yäsluṢ	yäsluṣ	26	7	253
941	56	tsraṣisac		26	7	254
941	56	kumseñc		26	7	255
941	56	yärkanT	yärkant	26	7	256
941	56	tsraṣisac		26	7	257
941	56	tsraṣiñ		26	7	259
941	56	waste		26	7	260
941	56	wrasaśśi		26	7	261
941	57	ṣṣe		26	7	262

Figure 6

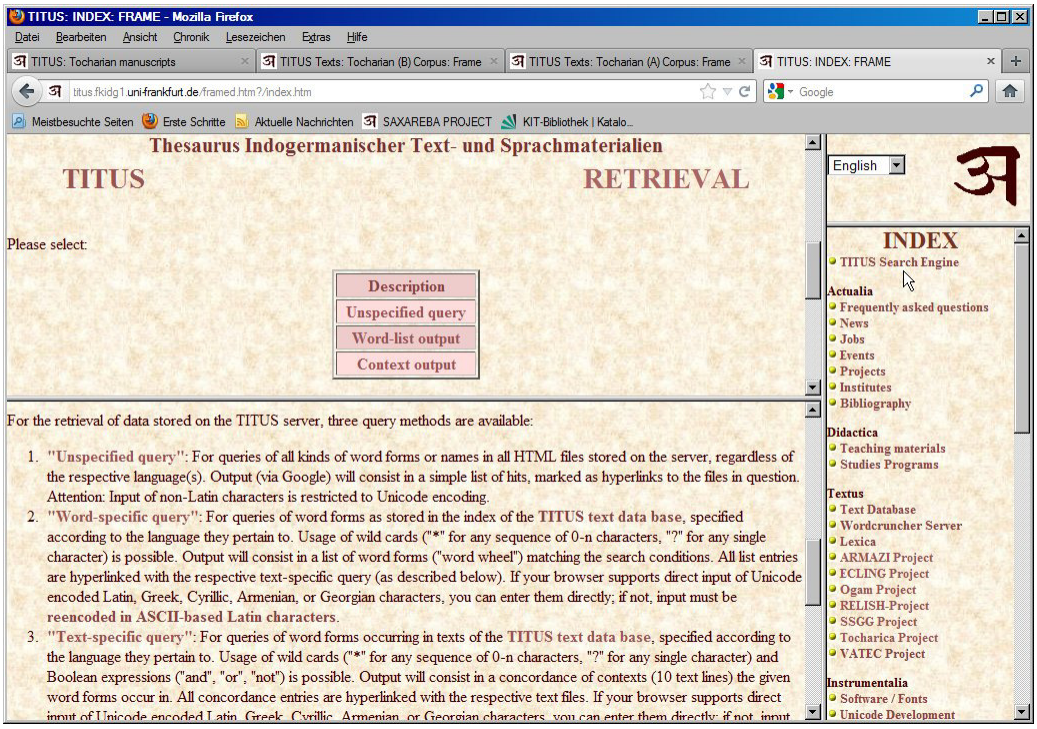


Figure 7

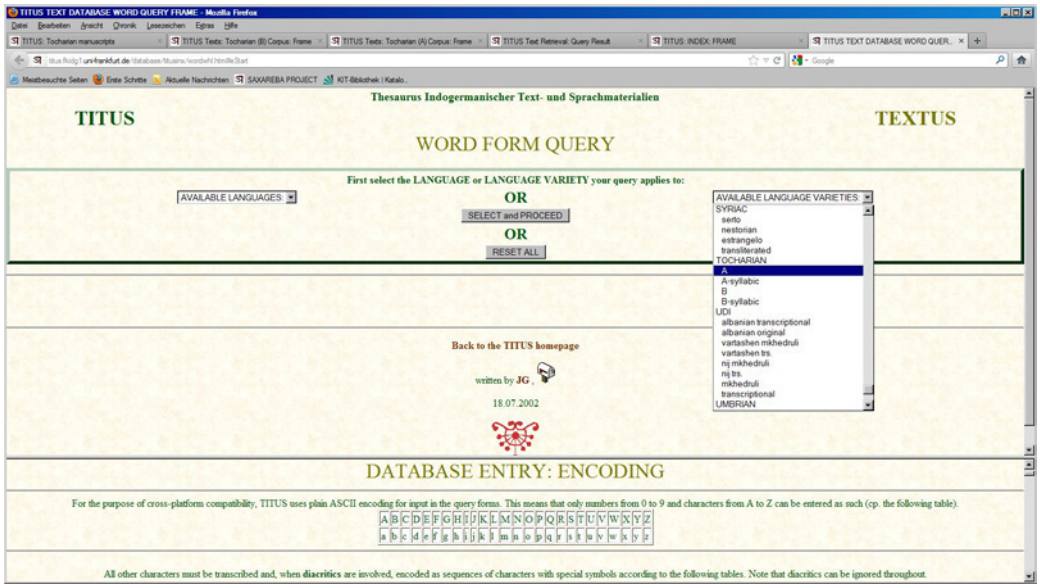


Figure 8

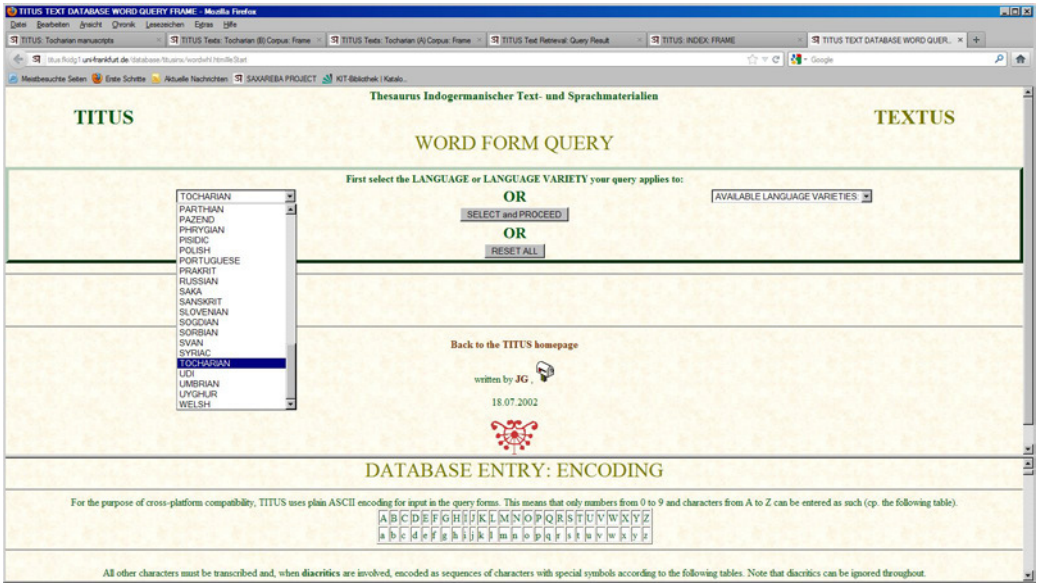


Figure 9

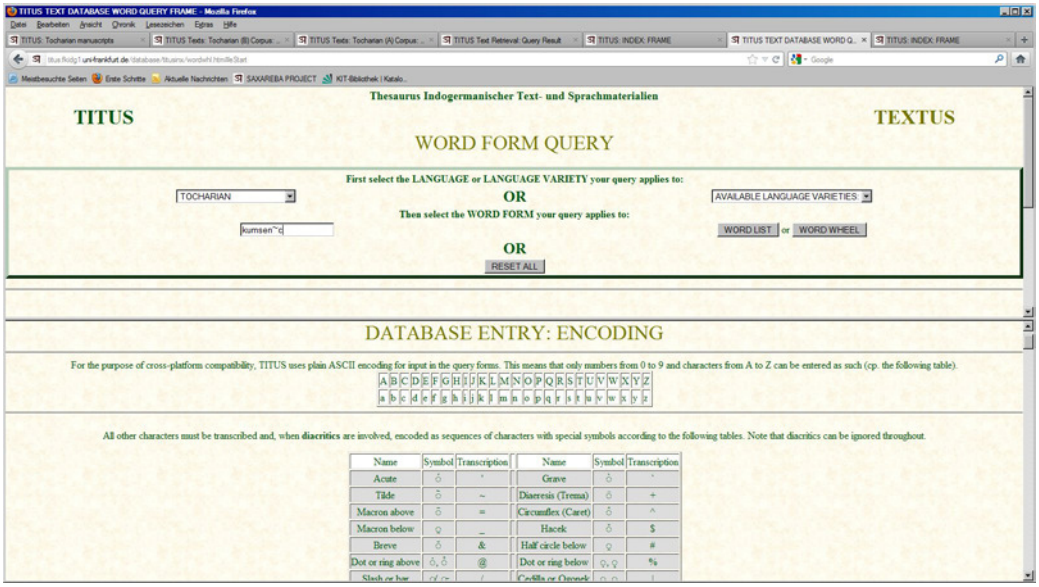


Figure 10

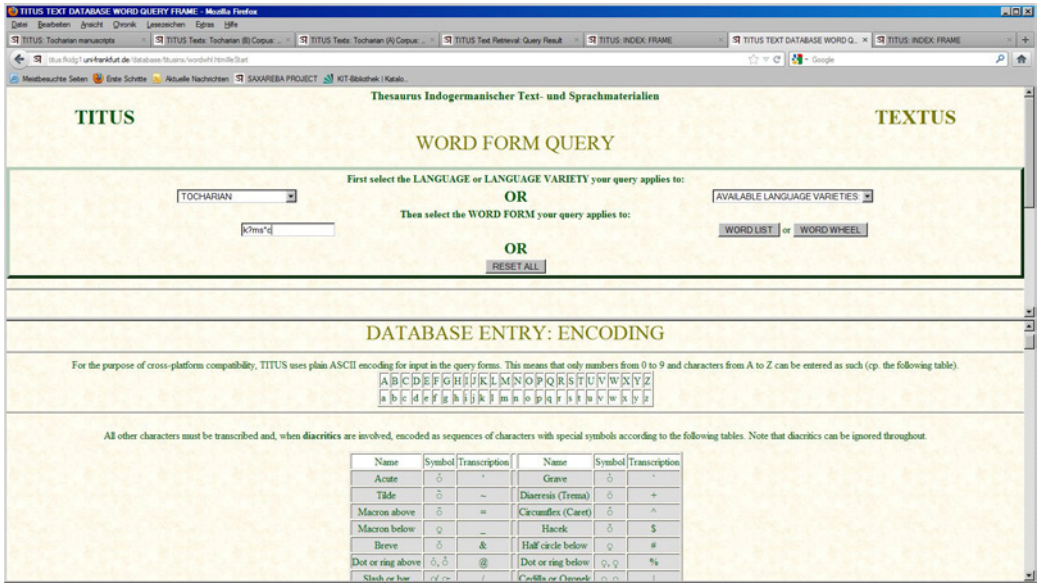


Figure 11

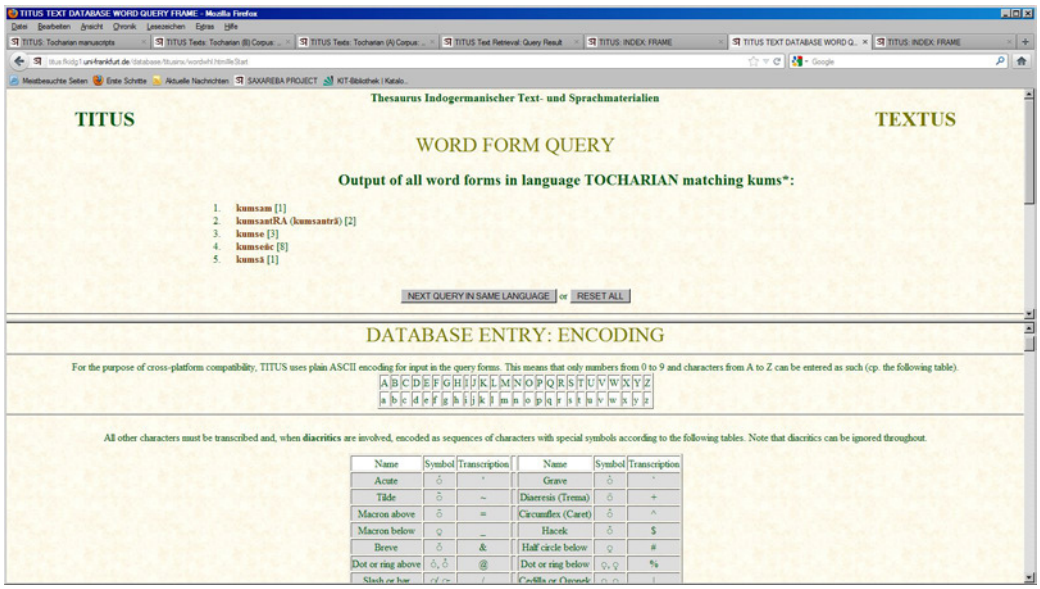


Figure 12

Figure 13

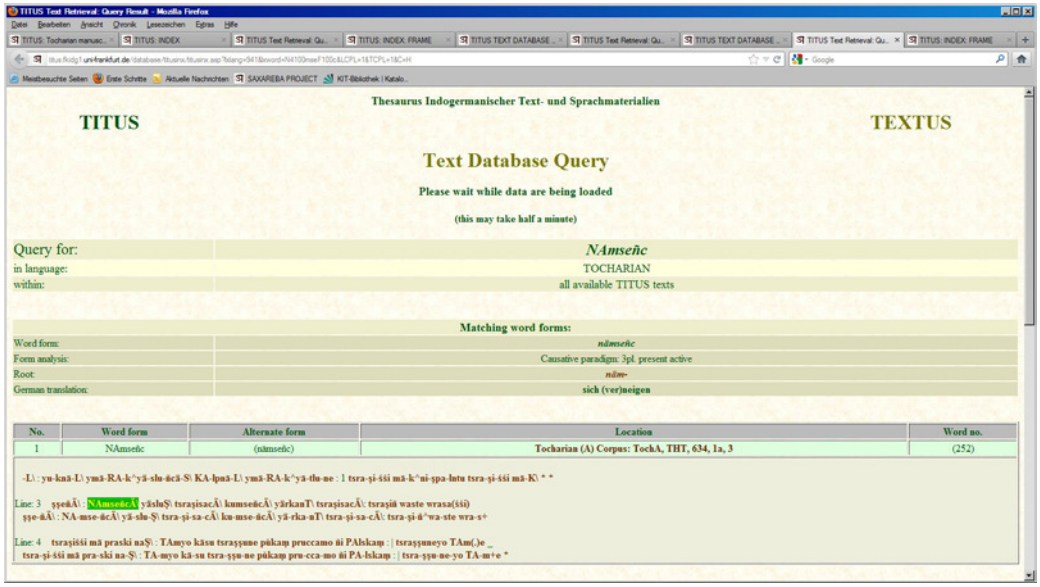


Figure 14

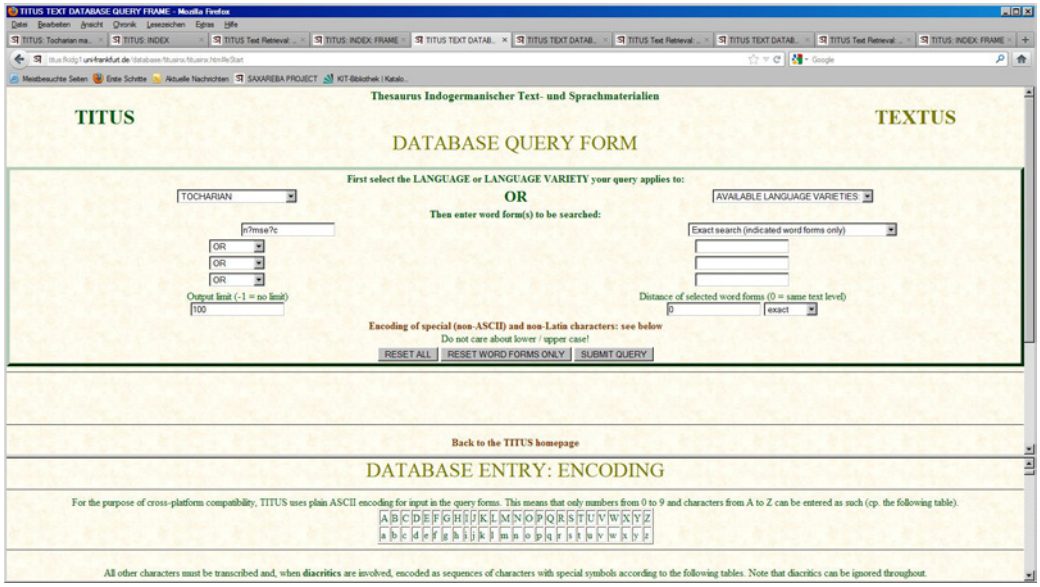


Figure 15

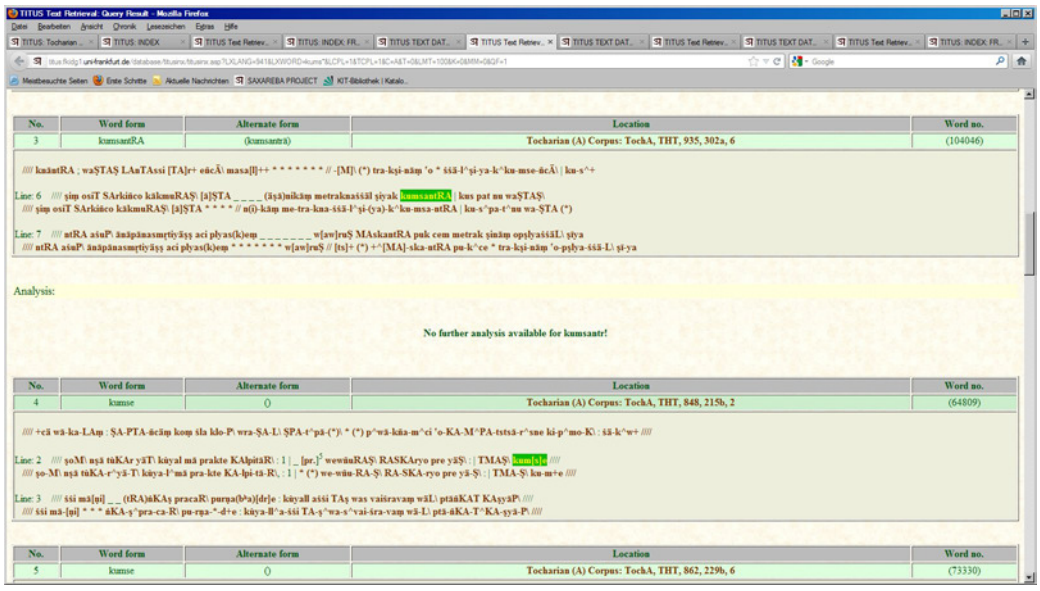
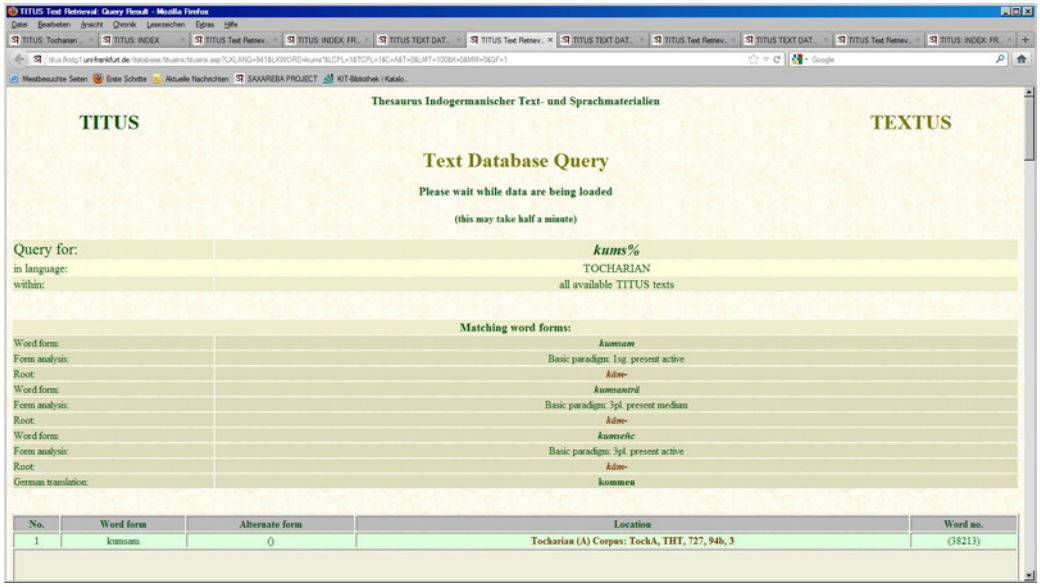


Figure 16



Thesaurus Indogermanischer Text- und Sprachmaterialien

TITUS TEXTUS

Text Database Query

Please wait while data are being loaded

(this may take half a minute)

Query for:	*kums%*
in language:	TOCHARIAN
within:	all available TITUS texts

Matching word forms:

Word form:	*kumsam*
Form analysis:	Basic paradigm: 1sg. present active
Root:	*käm-*
Word form:	*kumsanträ*
Form analysis:	Basic paradigm: 3pl. present medium
Root:	*käm-*
Word form:	*kumsehc*
Form analysis:	Basic paradigm: 3pl. present active
Root:	*käm-*
German translation:	**kommen**

No.	Word form	Alternate form	Location	Word no.
1	kumsam	()	Tocharian (A) Corpus: TochA, THT, 727, 94b, 3	(38213)

Fugure 17

Thesaurus Indogermanischer Text- und Sprachmaterialien

TITUS TEXTUS

Tocharian verbal forms

Root: *käm-*					
Basic paradigm ("Grundverb")					
Meaning:	kommen	kommen	Sanskrit equivalent:		
Present indicative				(Class A:)	(Class B:)
1st sg act	*kumsam*		1st sg med		
2nd sg act			2nd sg med		
3rd sg act	*kumnäṃ*	*kämmaṣṣäṃ*	3rd sg med	*kumnäṣtär*	*kämmaṣträ*
1st pl act			1st pl med		
2nd pl act			2nd pl med		
3rd pl act	*kumsenlc*	*kämmaskeṃ*	3rd pl med	*kumsanträ*	
Infinite forms of the present					
Present active participle			Present medium participle		
Infinitive	*kumnäši*		Verbal noun		
Gerund			Abstract of the gerund		
Verbal adjective			Agent noun		
Imperfect					
1st sg act			1st sg med		
2nd sg act			2nd sg med		
3rd sg act	*kumpäk/kumpä*		3rd sg med		
1st pl act			1st pl med		
2nd pl act			2nd pl med		

Figure 18

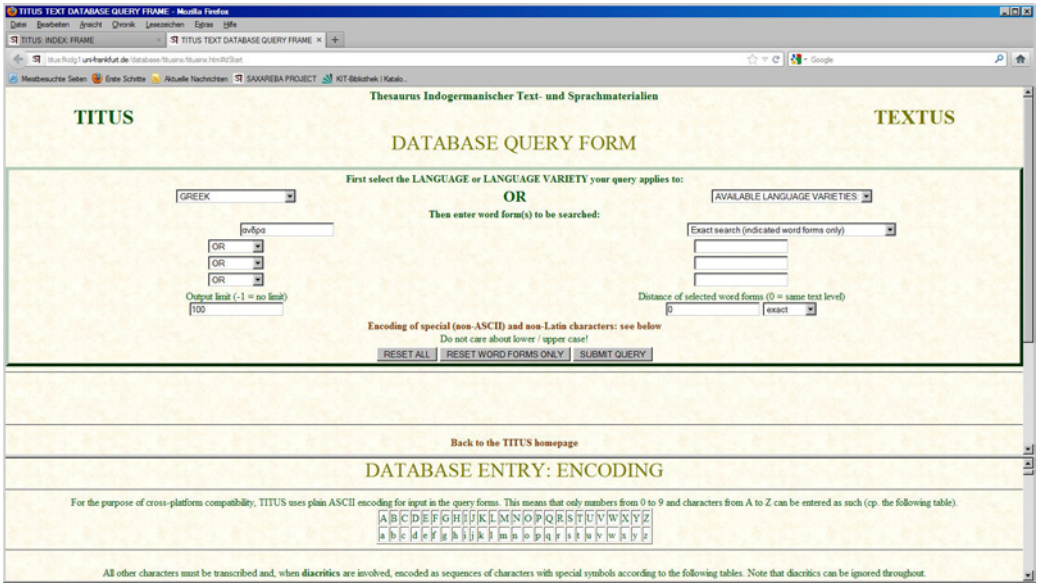


Figure 19

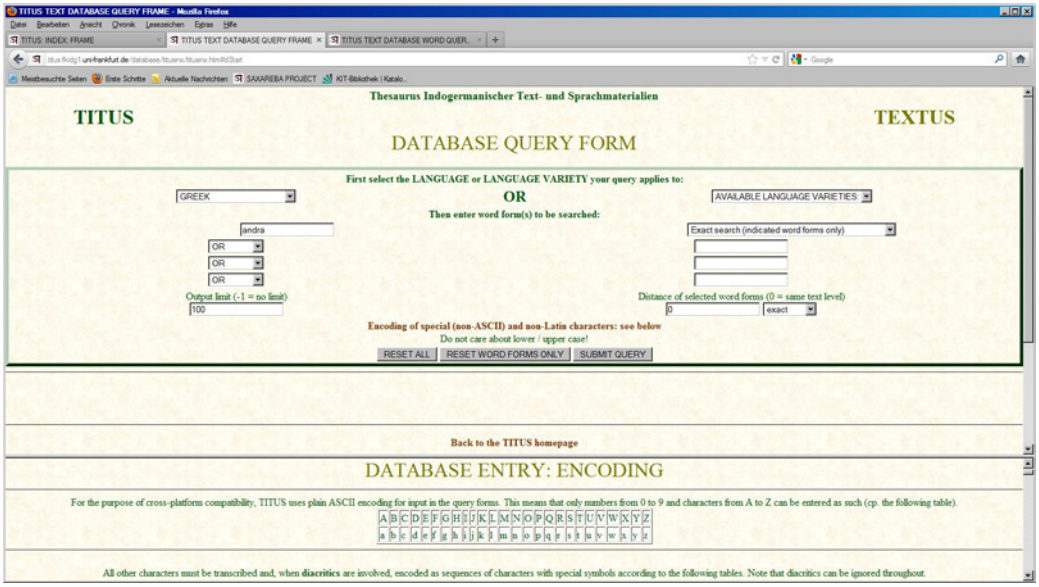


Figure 20

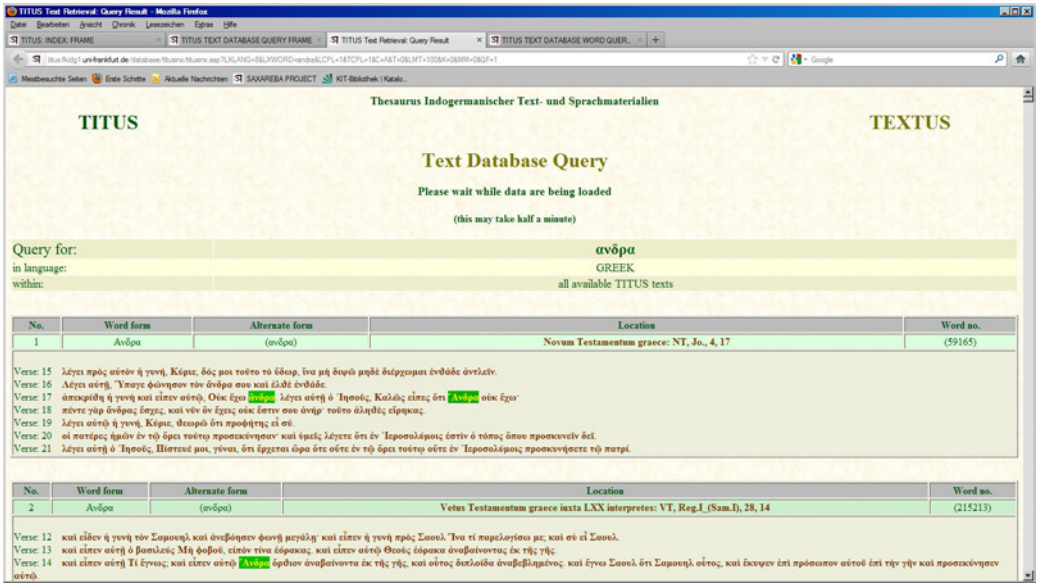


Figure 21

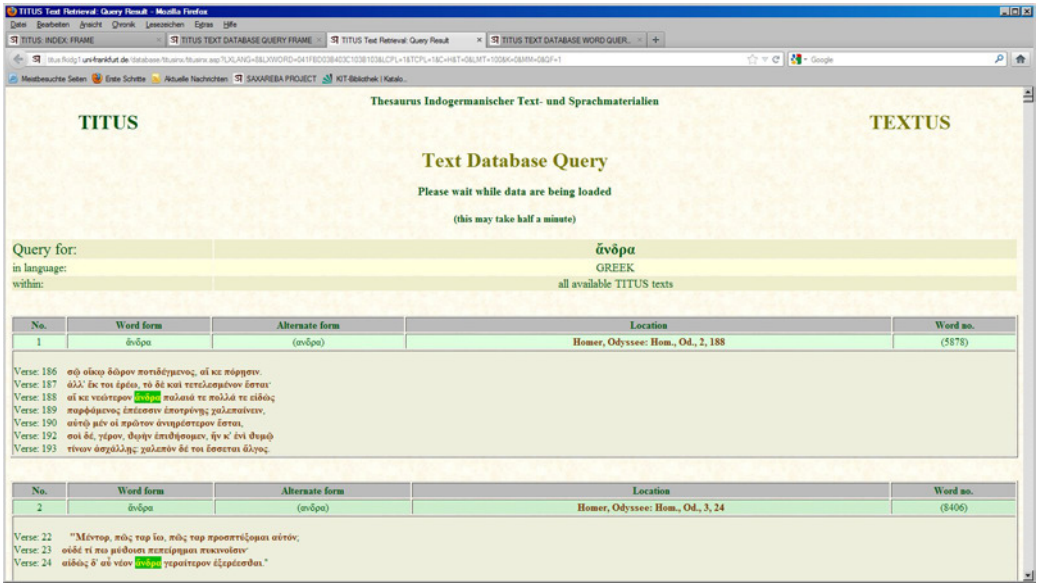


Figure 22

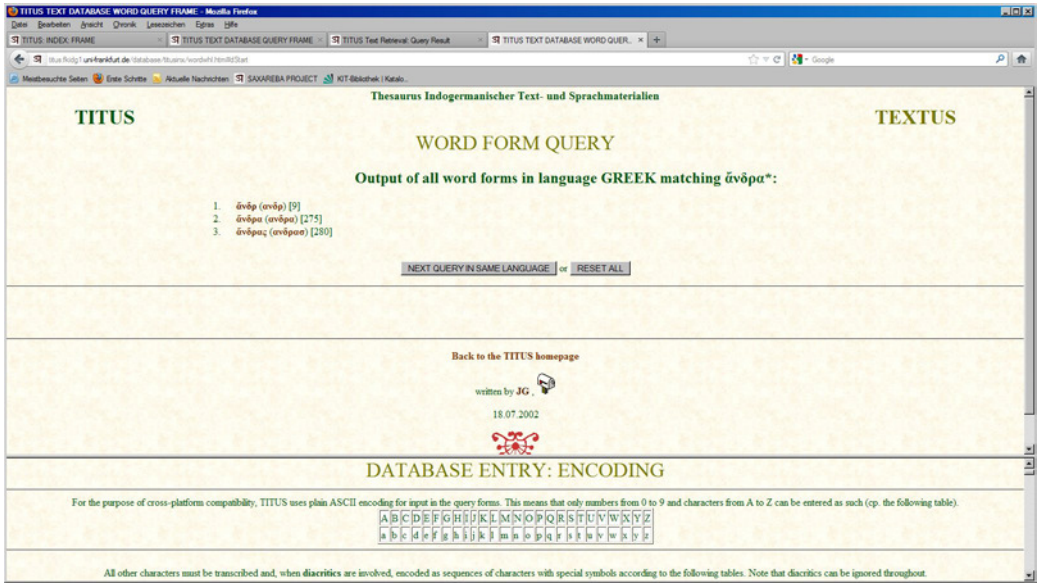


Figure 23

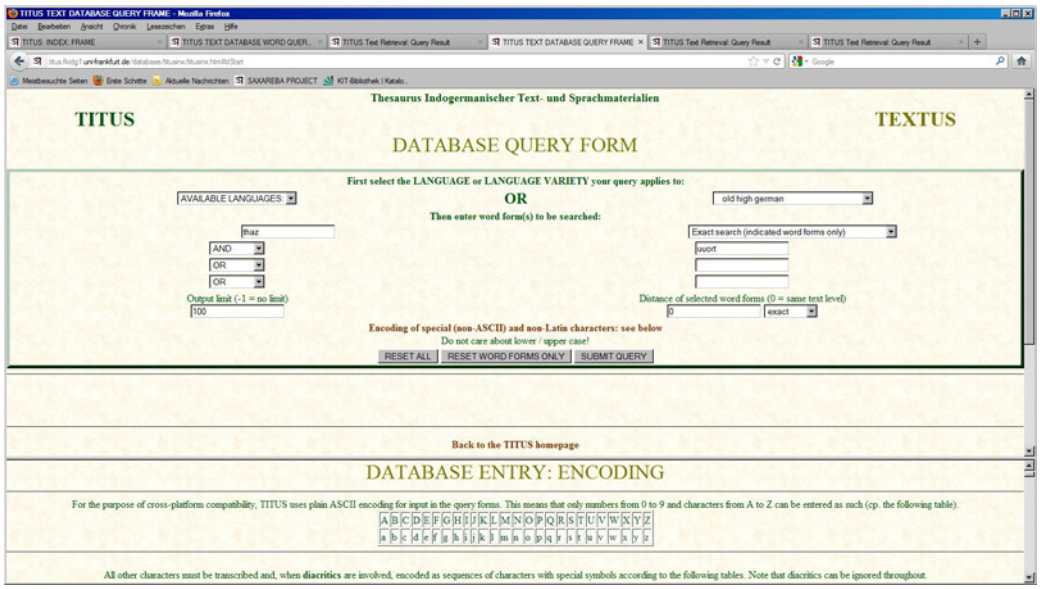


Figure 24

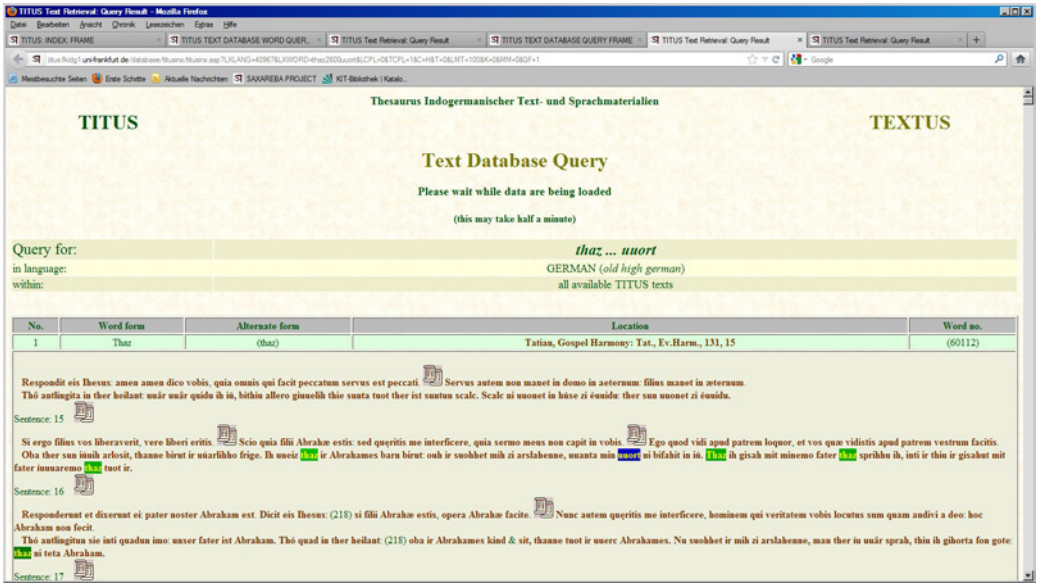


Figure 25

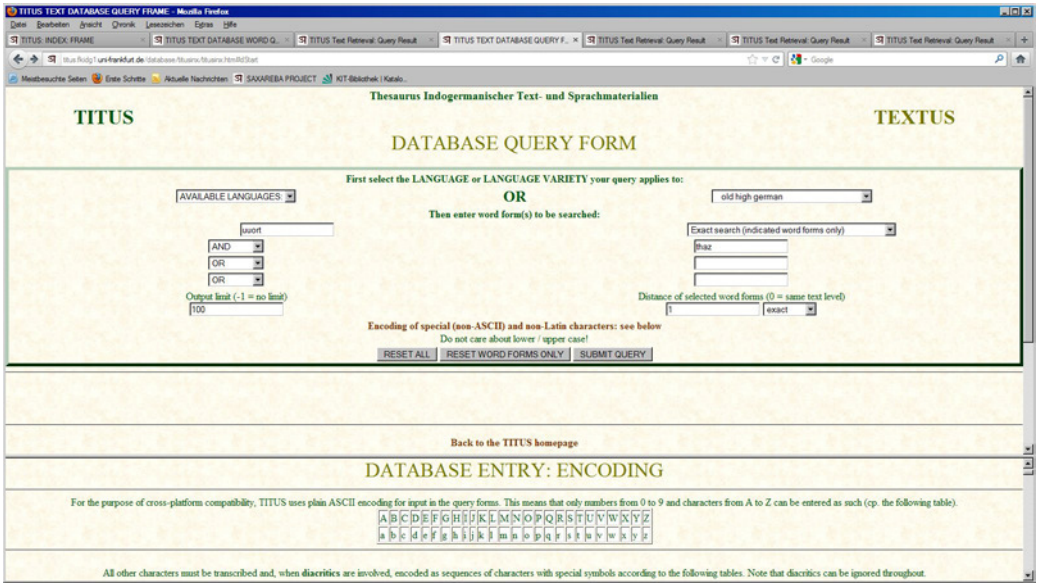


Figure 26

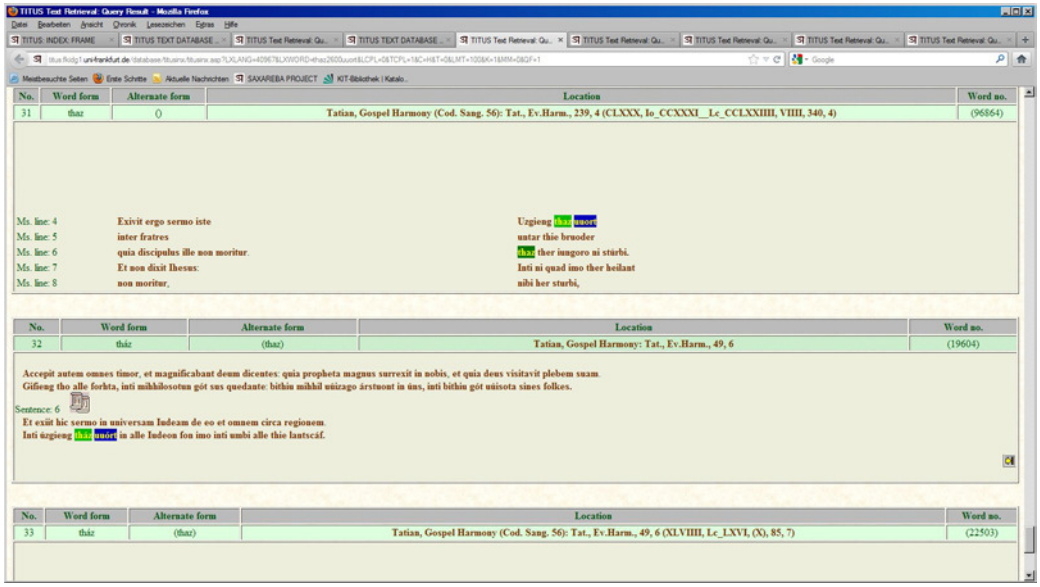


Figure 27

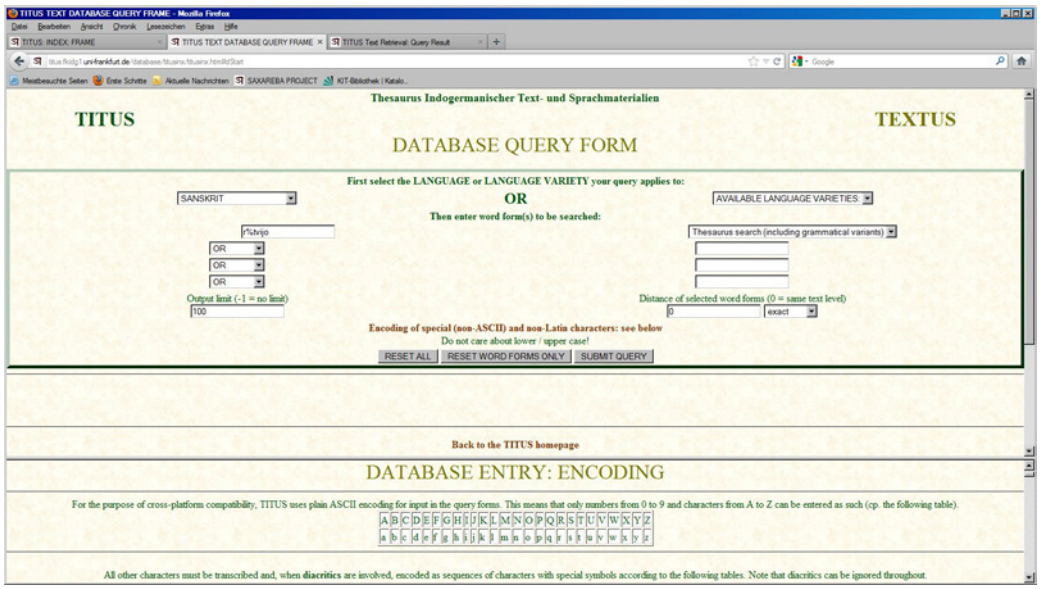


Figure 28

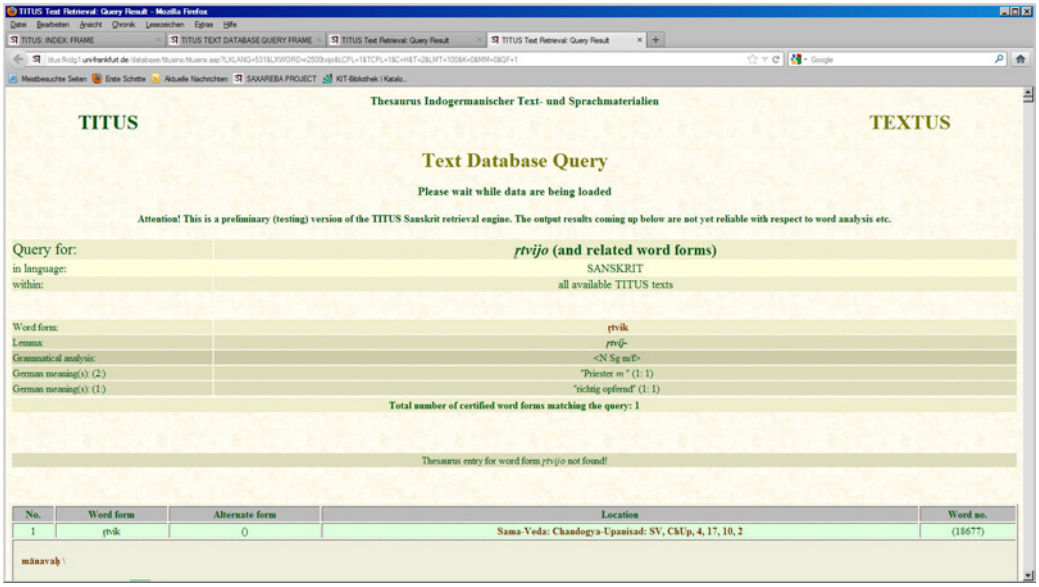


Figure 29

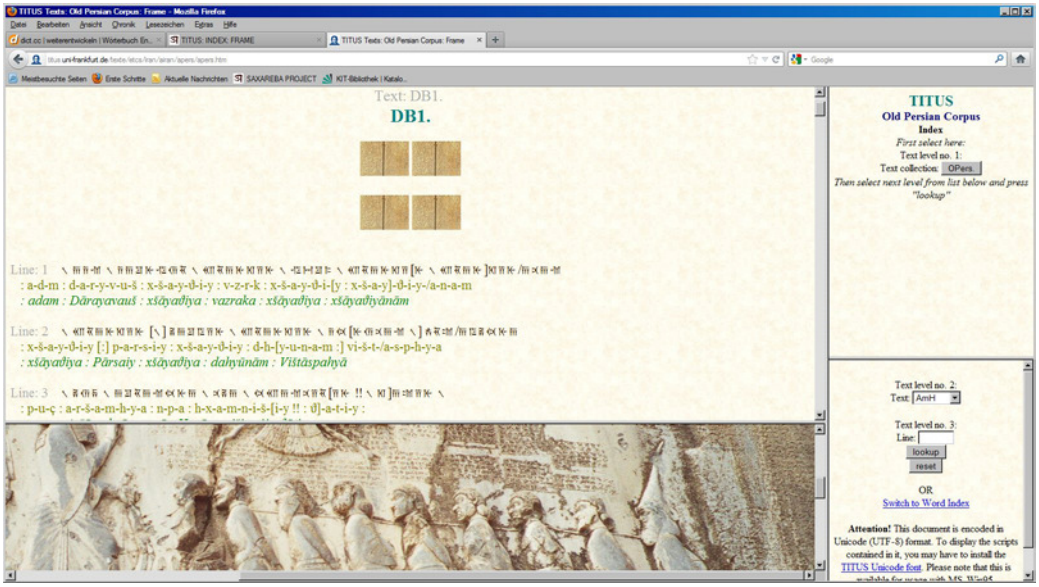


Figure 30

DIE DOPPELFUNKTION DES DIGITALEN TEXTARCHIVS ALS WÖRTERBUCHBASIS UND ALS KOMPONENTE DER ONLINE-PUBLIKATION

AM BEISPIEL DES MITTELHOCHDEUTSCHEN WÖRTERBUCHS

KURT GÄRTNER & RALF PLATE

Abstract
Auf dem Hintergrund des umfassenden EDV-Einsatzes, der die vorbereitende Materialbereitstellung seit 1986 (*Findebuch zum mittelhochdeutschen Wortschatz*) bzw. 1994 (für das neue *Mittelhochdeutsche Wörterbuch* selbst) ebenso wie die Ausarbeitung und Publikation des neuen *Mittelhochdeutschen Wörterbuchs* seit 2006 kennzeichnet, beleuchtet der Beitrag den Gewinn von digitalen Textcorpora für Macher und Nutzer von historischen Belegwörterbüchern: Den Lexikographen verschafft ein umfangreiches digitales Textarchiv und daraus durch halbautomatische Lemmatisierung gewonnenes Belegarchiv, das in einem Redaktionssystem für die Artikelarbeit bereitgestellt wird, größere und leichtere Übersicht über den historischen Sprachgebrauch und entlastet sie von zeitraubenden Exzerptions- und Korrekturarbeiten; den Wörterbuchbenutzern wird es durch die Verknüpfung der Belegzitate bzw. Belegstellenangaben mit den digitalisierten und im Online-Angebot zur Verfügung gestellten Volltexten der Wörterbuchquellen ermöglicht, den Quellenbezug der lexikographischen Befunde zu rekontextualisieren. Aus der Erfahrung eines Vierteljahrhunderts der Vorbereitung und Ausarbeitung eines großen digitalen Belegwörterbuchs werben die Autoren für die Bereitstellung von umfassenden digitalisierten Textcorpora (einschließlich der Retrodigitaliserung der älteren lexikographischen Hilfsmittel) für die historische Sprachforschung selbst wie für die Nutzer ihrer Forschungsergebnisse. Eine umfassende Textdigitalisierung von historischen Sprachquellen des Deutschen steckt aber leider (im Gegensatz zur Bilddigitalisierung von Handschriften und gedruckten Büchern) trotz großer Anstrengungen und überzeugender Ergebnisse in einzelnen Vorhaben immer noch in den Anfängen oder wird zum Teil unter zu engen Gesichtspunkten betrieben.

1

Das *Mittelhochdeutsche Wörterbuch* (MWB) soll den mittelhochdeutschen Wortschatz in seinen zeitlichen Grenzen von ca. 1050 bis ca. 1350 beschreiben. Die Beschreibung des deutschen Wortschatzes in den Zeiträumen davor und danach erfolgt durch das von Theodor Frings und Elisabeth Karg-Gasterstädt begründete *Althochdeutsche Wörterbuch* (AWB) und das von Oskar Reichmann begründete *Frühneuhochdeutsche Wörterbuch* (FWB). Wie AWB und FWB ist das MWB

ein Belegwörterbuch;[1] sein Ziel ist eine repräsentative Darstellung des mittelhochdeutschen Wortschatzes in seiner sprachsystematischen Gliederung und seiner zeitlichen, räumlichen und textsortenspezifischen Erstreckung. Die Vorarbeiten zum MWB bis zum Einsetzen der Artikelarbeit wurden von 1994 bis 1999 durch die DFG gefördert, zum 1.1.2000 erfolgte die Weiterförderung im Rahmen des Akademienprogramms.

Bei der Planung des MWB war von vornherein vorgesehen, dass auf allen Stufen der Wörterbucharbeit die EDV eingesetzt werden sollte.[2] Zum Wörterbuchplan seien zunächst einleitend einige kurze Erläuterungen gegeben. Dieser sah als zentrale Einheit eine digitalisierte Quellensammlung bzw. ein digitales Textarchiv mit einem auf Zuwachs angelegten Corpus von retro-digitalisierten Volltexten vor, das die Basis für ein digitales Belegarchiv bildete, das aus dem Textarchiv durch Verfahren der halbautomatischen Lemmatisierung gewonnen wurde. Unter den weiteren digital nutzbaren Hilfsmitteln sind als wichtigste Ergänzung des Quellencorpus und des Belegarchivs die Vorgängerwörterbücher zu nennen. Jedes neue Wörterbuch hat lexikographische Vorfahren, auf denen es aufbaut und die daher einen wesentlichen Teil der Wörterbuchbasis bilden. Das gilt für Wörterbücher vergangener Sprachepochen in viel höherem Maße als für Wörterbücher zur Gegenwartssprache. Mit DFG-Unterstützung haben wir daher in Trier von 1997 an die alten mittelhochdeutschen Wörterbücher zusammen mit ihren Quellenverzeichnissen maschinenlesbar gemacht und miteinander verknüpft. Es entstand so ein digitaler Wörterbuchverbund, der außer den Quellenverzeichnissen vier lexikographische Hilfsmittel vereinigt:

1) Das erste und bis heute nicht ersetzte Wörterbuch zum Mittelhochdeutschen wurde von 1854 bis 1866 mit Benutzung des Nachlasses von Georg Friedrich Benecke durch Wilhelm Müller und Friedrich Zarncke ausgearbeitet. Der Benecke-Müller-Zarncke (= BMZ) ist ein exzellentes Wörterbuch, aber nicht leicht zu benutzen,

---

[1] Vgl. dazu ausführlich PLATE, R., Das Mittelhochdeutsche Wörterbuch: Beleglexikographische Konzeption, EDV, Vernetzungspotentiale, in: *Lexicographica* 23 (2007), 77-95.

[2] Zur Erprobung des EDV-Konzepts vgl. ausführlich PLATE, R. & U. RECKER, Elektronische Materialgrundlage und computergestützte Ausarbeitung eines historischen Belegwörterbuchs. Erfahrungen und Perspektiven am Beispiel des neuen Mittelhochdeutschen Wörterbuchs, in: LEMBERG, I. *et. al.* (Hrsg.), *Chancen und Perspektiven computergestützter Lexikographie*, Lexicographica; Series Maior 107, Tübingen 2001, 155-177.

denn er ist nach Wortstämmen angeordnet, unter denen alle Glieder einer Wortfamilie erscheinen. 2) Um den BMZ besser benutzbar zu machen, arbeitete Matthias Lexer von 1872 bis 1878 einen alphabetischen Index aus, der zugleich als Supplement und Handwörterbuch auf der Basis des BMZ fungieren sollte. 3) Da schon während der kurzen Ausarbeitungszeit des Mittelhochdeutschen Handwörterbuchs zahlreiche neue Quellen zum Mittelhochdeutschen erschlossen wurden, ließ Lexer 1878 zusammen mit dem dritten Band seines Wörterbuchs umfangreiche Nachträge erscheinen, die das dritte Wörterbuch des digitalen Verbunds bilden. 4) Das vierte Werk ist das sog. *Findebuch zum mittelhochdeutschen Wortschatz*, das an der Universität Trier zwischen 1986 und 1992 ausgearbeitet wurde. Es vereinigt 106 Glossare zu den wichtigsten nach Lexer, d.h. nach 1878 erschienen Textausgaben des Mittelhochdeutschen. Die vier Komponenten wurden miteinander und mit dem Quellenverzeichnis von Eberhard Nellmann zu BMZ und Lexer sowie dem Quellenverzeichnis zum *Findebuch* verknüpft und vielfältig recherchierbar gemacht. Man kann nun also ein Stichwort in allen vier Komponenten, wie Abb. 1 am Beispiel von *abbet* zeigt, nebeneinander „aufschlagen“ und vergleichen.

Der Ausbau des digitalen mittelhochdeutschen Textarchivs, auf dessen Bedeutung für die Ausarbeitung des MWB wie für seine Nutzung im Internet im Folgenden genauer eingegangen wird, wurde 2001 bis 2003 in einem deutsch-amerikanischen Kooperationsprojekt mit DFG/NSF-Förderung vorangetrieben; erfasst wurden vor allem die über 100 *Findebuch*-Quellen, die den größten Zuwachs an neuem Wortschatz (rund 8000 Lexeme) lieferten. Wieder mit DFG-Förderung wurde 2002 bis 2005 ein Artikelredaktionssystem entwickelt, das mit Beginn der Artikelarbeit zur Verfügung stand.

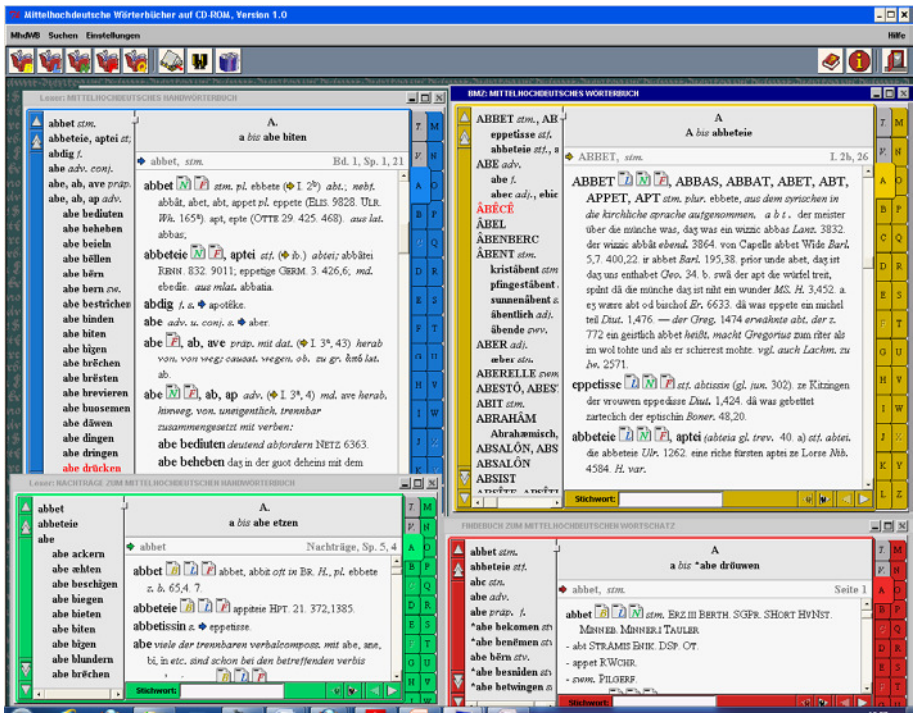


Abb. 1: Mittelhochdeutscher Wörterbuchverbund (Offline-Version): Lemma abbet

2

Die Planung des neuen Mittelhochdeutschen Wörterbuchs sah also von vornherein, bereits zu Beginn der 1990er Jahre, vor, dass das Wörterbuch eine möglichst breite Basis elektronischer Quellen nutzen und dass es mit den Mitteln der elektronischen Datenverarbeitung ausgearbeitet werden sollte. Die technischen und konzeptionellen Überlegungen dafür waren lange vor der WWW-Zeit, nämlich schon seit Ende der 1960er Jahre entwickelt und erprobt worden, auf dem Gebiet der historischen Lexikographie des Deutschen maßgeblich von Paul Sappler, auf dessen Expertise sich die Planungen für das MWB stützen konnten.[3] Als dann das WWW und die neuen elektronischen Publikationsmöglichkeiten aufkamen, war sofort klar, dass das neue Wörterbuch zusammen mit seiner Basis

[3] Zusammenfassend: SAPPLER, P., Prinzipien des EDV-Konzepts, in: GÄRTNER, K. & K. GRUBMÜLLER (Hrsg.), *Ein neues Mittelhochdeutsches Wörterbuch. Prinzipien, Probeartikel, Diskussion*, Nachrichten der Akademie der Wissenschaften in Göttingen, Phil.-hist. Klasse, Jg. 2000, Nr. 8, Göttingen 2000, 43-52.

auch für elektronische Nutzung bereitgestellt werden sollte; die Publikation des eben vorgestellten Verbunds der digitalisierten Vorgängerwörterbücher auf CD und im Internet war der erste Schritt dazu, die Bereitstellung der wachsenden digitalen Belegsammlung für das neue Wörterbuch in einem Online-Angebot seit dem Jahr 2000 der zweite.

Seit 2006 wird nun das neue Wörterbuch selbst lieferungsweise im Druck und nach einer kurzen Schutzfrist auch elektronisch publiziert. Bisher liegen vier Doppellieferungen vor, die 2006, 2007, 2009 und Ende 2011 erschienen sind und auf 2224 Spalten die Wortstrecke *a* bis *êvrouwe* enthalten. In einer um Druckfehler bereinigten Fassung werden sie zusammen mit einem zusammengefassten und überarbeiteten Quellenverzeichnis (das das Verzeichnis der ersten Lieferung mit den drei Nachträgen der Lieferungen 2-4 vereinigt) und mit einem Vorwort als Band 1 voraussichtlich Ende 2012 erscheinen.

Aus den Erfahrungen bei der Vorbereitung, der Ausarbeitung und der Publikation des neuen Wörterbuchs soll heute zu der Leitfrage dieser Tagung Stellung genommen werden, also zur Frage nach den Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher, und zwar gerade in ihrem wechselseitigen Verhältnis: Welche Rolle spielt für uns das elektronische Textcorpus bei der Ausarbeitung des Wörterbuchs, und welche Rolle spielt es dann anschließend für die Nutzer der elektronischen Publikation des Wörterbuchs?

Zur Frage der Rolle des elektronischen Textcorpus für die Ausarbeitung des Wörterbuchs seien zunächst einige Zahlen zum Stand der elektronischen Materialbasis vorgelegt, die nach Erscheinen der dritten Doppellieferung erhoben wurden.[4] Die Stichwortliste für das Wörterbuch umfasst insgesamt rd. 84.000 Artikelkandidaten, die wir zu Projektbeginn aus den Vorgängerwörterbüchern des 19. Jahrhunderts und dem *Findebuch* kompiliert haben; diese Liste bildete dann später auch die Grundlage für die Vernetzung der Vorgängerwörterbücher untereinander, die oben unter 1 beschrieben worden ist, und ebenso auch für die Anbindung dieses Wörterbuchverbunds an das neue Wörterbuch und seine Materialien im Online-Angebot des MWB, das unten vorgestellt wird. Tatsächlich artikelwertig dürften von diesen 84.000 Kandidaten für das neue Wörterbuch nach den bisherigen Erfahrungen nur etwa zwei Drittel sein, das MWB

[4] PLATE, R., Das Mittelhochdeutsche Wörterbuch. Bearbeitungsstand 2009, Erfahrungen und Perspektiven, in: SCHMID, H. U. (Hrsg.), *Perspektiven der germanistischen Sprachgeschichtsforschung*, Jahrbuch für germanistische Sprachgeschichte 1, Berlin / New York 2010, 254-268, hier 256-259.

wird demnach also rd. 60.000 Wortartikel enthalten. Die Gründe für die gegenüber den Vorgängerwörterbüchern geringere Artikelzahl sind ein engerer Quellenzeitraum und abweichende Lemmatisierungsprinzipien.

Die zweite Hauptkomponente der elektronischen Materialbasis ist die digitalisierte Quellensammlung. Sie besteht zurzeit aus rd. 210 Texten mit zusammen rd. 7 Mio. laufenden Wortformen. Aus diesen Texten ist vor allem durch Verfahren halbautomatischer Lemmatisierung, ergänzend auch durch Lemmatisierung von Einzelstellen, das elektronische Belegarchiv erhoben worden. Es umfasst (nach Ausarbeitung der dritten Doppellieferung) rd. 1.425.000 Textbelege, die sich auf rund 27.000 Artikelkandidaten verteilen. Diese Zahlen besagen, dass trotz umfassender Digitalisierungs- und Lemmatisierungsanstrengungen durch das Projekt selbst immer noch nur rund die Hälfte – nämlich 27.000 von voraussichtlich knapp 60.000 artikelwertigen Wortschatzeinheiten – im digitalen Belegarchiv vertreten sind. Entsprechendes gilt dann wiederum für die Ebene der einzelnen Wortschatzeinheit und dem Belegmaterial, das für die Beschreibung ihres Gebrauchs – Bedeutungen, Wendungen, syntaktisches Verhalten – nötig ist. Die dazu für die ersten drei Doppellieferungen (wie Anm. 2, S. 257) erhobenen Zahlen ergeben, dass sich vom Gesamt der in einer gedruckten Doppellieferung zitierten oder mit einer Stellenangabe vertretenen Belege nur rd. die Hälfte im elektronischen Belegarchiv der Wörterbuchdatenbank fanden, während die anderen nachexzerpiert werden mussten. Zwar ist der Anteil des elektronischen Belegarchivs kontinuierlich gesteigert worden von 40 % in der ersten Doppellieferung bis auf 54 % in der dritten Doppellieferung, und in der gerade im Druck erschienenen vierten Doppellieferung ist er abermals leicht erhöht worden, aber mindestens ein Drittel der Belege stammt auch in ihr immer noch aus Nachexzerption, die zu einem großen Teil auf den Nachweisen in den Vorgängerwörterbüchern beruht.

Die[5] nachexzerpierten Belege haben gegenüber den Datenbankbelegen auch ganz abgesehen von dem Aufwand, der für ihre Erhebung bei laufender Artikelarbeit getrieben werden muss, verschiedene Nachteile, von denen hier nur der gravierendste für Lexikographen wie für Benutzer des Wörterbuchs hervorgehoben werden soll: Die Exzerpte haben keinen Volltext hinter sich, den man als Artikelbearbeiter im Redaktionssystem oder als Benutzer der Online-Fassung

[5] Dieser Absatz ist unverändert wiederholt aus PLATE (wie Anm. 4), 258f.

des Wörterbuchs aufrufen könnte, um den Textzusammenhang über den exzerpierten Ausschnitt hinaus zu prüfen. – Hierzu sei eine Seitenbemerkung gestattet, die von allgemeinerem methodischen Interesse für die historische Linguistik ist. Als wir in den 1990er Jahren damit begannen, das elektronische Text- und Belegarchiv für das neue Wörterbuch einzurichten, wurde die Diskussion über den nötigen Umfang stark von der Befürchtung einiger Beteiligter bestimmt, dass schnell Belegmengen entstehen könnten, die nicht mehr bearbeitbar wären und geeignet, das eigentliche Vorhaben, die Ausarbeitung des Wörterbuchs, lahmzulegen. Dies betrifft tatsächlich jedoch nur eine kleine Gruppe von höchstfrequenten Lexemen; für sie müssen bei der halbautomatischen Lemmatisierung geeignete Filter oder Sperren eingebaut werden. Der größte Teil des Wortschatzes dagegen bedarf einer sehr großen Textmenge, um angemessen vertreten zu sein in der Belegsammlung. Der Aufbau einer umfassenden elektronischen Quellensammlung für die deutsche Sprachgeschichte ist aber leider immer noch ein dringendes Desiderat; es sollte heute nicht mehr Aufgabe einzelner Projekte sein, sich die elektronischen Corpora für ihre Untersuchungszwecke selbst zu erarbeiten.

Die Leitfrage nach den Nutzungsperspektiven elektronischer Corpora ist also soeben für die Wörterbuchmacher wie folgt beantwortet worden: Der Nutzen ist enorm, man sollte keine Kosten und Mühen scheuen, solche Corpora für die historische Linguistik bereitzustellen, und man sollte sich dabei vor allem nicht von zu engen Gesichtspunkten leiten lassen. Bereits gestreift worden ist eben auch schon die Perspektive der Wörterbuchnutzer, also von uns allen. Zu dieser Perspektive jetzt abschließend einige Anmerkungen am Beispiel unseres Internetangebots MWB Online.

Auf MWB Online wird nicht nur das Wörterbuch selbst für elektronische Nutzung bereitgestellt, sondern zusätzlich auch die elektronischen Materialien, die für die Ausarbeitung genutzt worden sind bzw. für die weitere Ausarbeitung bereitstehen. Zentraler Einstiegsort in das Online-Angebot ist die Komponente Lemmaliste / Belegarchiv (vgl. Abb. 2). Sie enthält die gesamte Stichwortliste des Vorhabens, also auch die Stichwörter, die nur in den alten Wörterbüchern mit einem Artikel bzw. mit einem eigenen Artikel vertreten sind. In der Lemmaliste wird auf ökonomische Weise angezeigt, zu welchen Stichwörtern es Belegmaterial gibt und wie viele Belege, außerdem, ob es dazu Artikel im neuen und / oder in den alten mittelhochdeutschen Wörterbüchern gibt. Am Beispiel des Stichworts *abbet* ‚Abt, Vorsteher eines Klosters': Dazu findet sich in Klammern eine

Häufigkeitsangabe (26), die die Zahl der Stellen im elektronischen Belegarchiv anzeigt, anschließend folgen die Siglen MWB und MWV für das neue Wörterbuch und den Verbund der alten Wörterbücher im Internet. Die Häufigkeitsangabe und die beiden Siglen sind als Hyperlinks gestaltet, über die man in die entsprechenden Komponenten gelangt.

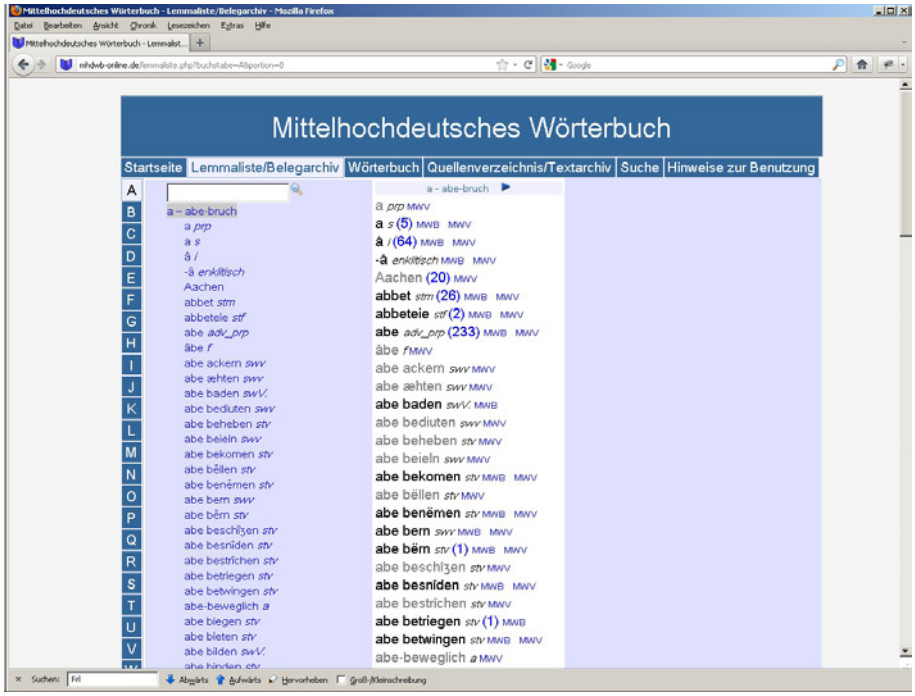


Abb. 2: MWB Online: Komponente Lemmaliste/Belegarchiv

Abb. 3 zeigt die Ausgabe des elektronischen Belegarchivs zum Stichwort *abbet*, die bereitgestellt wird, wenn man auf die eingeklammerte Frequenzangabe klickt. Wenn man dagegen dem Link in den Artikel selbst des neuen Wörterbuchs folgt, erhält man die Ausgabe, die in Abb. 4b abgebildet ist; dem elektronischen Artikel gegenübergestellt ist dort in Abb. 4a die Satzausgabe des Artikels im gedruckten Wörterbuch.

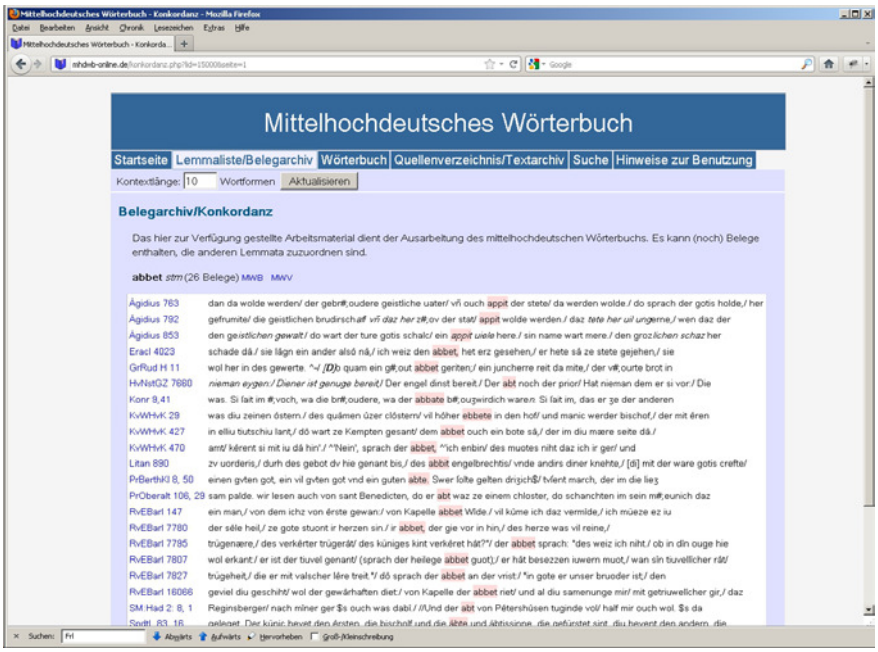


Abb. 3: MWB Online: Komponente Belegarchiv, Konkordanz der Belege für abbet

abbet *stM. Pl.* eb(be)te; *auch* abbât [:slât] UVZLANZ 3864, abbate KONR 9,41, abte [*Akk.*] PRBERTHKL 8,50; *verschoben* -p(p)- *(zu den Schreibformen s. WMU 1,13); lat.* abbas, -atis, *ahd.* abbat *(vgl. Etymol.Wb.d.Ahd. 1,19f.). 'Abt, Vorsteher einer Mönchsgemeinschaft'* der gebr#dere geistliche uater / vnd ouch appit der stete ÄGIDIUS 763; wir lesen auch von sant Benedicten, do er abt waz ze einem chloster PROBERALT 106,29; der meister über die münche was, / daz was ein wizzic abbas UVZLANZ 3832; ich han getan gehorsam ir, / reht als ein munch eim apte tut MINNEB 1887; *sprichwörtl.* swâ der apt die würfel treit, / spilnt dâ die münche daz ist niht ein wunder MARNER S. 160 *(vgl. TPMA 1,17f.). – als hoher Würden- und Herrschaftsträger:* die äbte und äbtissinne, die gefürstet [*in den Fürstenstand erhoben*] sint SPDTL 83,16; eines tages keiser Ott reit / mit den fürsten […] kurzwîlen an daz velt, / […] der abt von Vult neben im reit ENIKWCHR 27811; des quâmen ûzer clôstern / vil hôher ebbete in den hof / und manic werder bischof KVWHVK 29; nur die man infel tragen sach, / bischolf, ebt und cardinal OTTOK 13525. – *Gebrauch als Titel in Verbindung mit Namens- und/oder Ortsangabe:* von Kapelle abbet Wîde RVEBARL 147; gedenke […] zv vorderis […] des abbit Engelbrechtis LITAN 890; der appet sante Brandan RVEWCHR 3060; *in der Anrede mit* her(re): der herzog Albreht […] sprach: 'nû verjeht, / her abt, waz iu werre.' OTTOK 36064

Abb. 4a: Artikel abbet *(Druckfassung)*

Etymol.Wb.d.Ahd. 1,1f.; umfangr. mhd. Belegsammlung von I.V. Zingerle, Germania 7 (1862), S. 257-267 (älteste Belege dort „gegen das Ende des 12. Jh.s", S. 266)

abbet *stM. Pl.* eb(be)te; *auch* abbât [.slât] UVZLANZ 3864, abbate KONR 9,41, abte [*Akk.*] PRBERTHKL 8,50; *verschoben* -p(p)- *(zu den Schreibformen s. WMU 1,13); lat.* abbas, -atis, *ahd.* abbat *(vgl. Etymol.Wb.d.Ahd. 1,19f.). 'Abt, Vorsteher einer Mönchsgemeinschaft'* der gebr#;oudere geistliche uater / vnd ouch appit der stete ÄGIDIUS 763; wir lesen auch von sant Benedicten, do er abt waz ze einem chloster PROBERALT 106,29; der meister über die münche was, / daz was ein wizzic abbas UVZLANZ 3832; ich han getan gehorsam ir, / reht als ein munch eim apte tut MINNEB 1887; *sprichwörtl.* swâ der apt die würfel treit, / spilnt dâ die münche daz ist niht ein wunder MARNER S. 160 *(vgl. TPMA 1,17f.). – als hoher Würden- und Herrschaftsträger:* die äbte und äbtissinne, die gefürstet [*in den Fürstenstand erhoben*] sint SPDTL 83,16; eines tages keiser Ott reit / mit den fürsten [...] kurzwîlen an daz velt, / [...] der abt von Vult neben im reit ENIKWCHR 27811; des quâmen ûzer clôstern / vil hôher ebbete in den hof / und manic werder bischof KVWHVK 29; nur die man infel tragen sach, / bischolf, ebt und cardinal OTTOK 13525. – *Gebrauch als Titel in Verbindung mit Namens- und / oder Ortsangabe:* von Kapelle abbet Wîde RVEBARL 147; gedenke [...] zv vorderis [...] des abbit Engelbrechtis LITAN 890; der appet sante Brandan RVEWCHR 3060; *in der Anrede mit* her(re): der herzog Albreht [...] sprach: 'nû verjeht, / her abt, waz iu werre.' OTTOK 36064

abbeteie, abbetîe, abdîe, abtei *stF. auch* abbateie, abbatîe, abbacie, appiteie, aptie; *vereinzelt sw., mit Uml.* eptigen *(Dat. Sg.)* PILGERF 9922. *Neben Formen mit dem Suffix* -eie, *die zuerst in Hss. des 12./13. Jh.s im* SUMMHEINR *belegt sind, stehen solche mit dem*

Abb. 4b: MWB Online: Artikel abbet *(Online-Version)*

Hier soll weder inhaltlich auf den Artikel eingegangen werden, noch sollen die Vorzüge und Nachteile der beiden Darstellungen diskutiert werden. Worauf es ankommt, sind die weiterführenden Benutzungsmöglichkeiten des Bildschirmartikels. Sie sind angedeutet in den farblich hervorgehobenen Links, die sich bei Quellensiglen und bei Stellenangaben finden. Die Quellensiglen sind durchgehend mit Links versehen, die zu den bibliographischen Angaben im Quellenverzeichnis führen. Die Stellenangaben sind nur dann als Hyperlinks gestaltet, wenn es sich um Belege aus den Quellen des elektronischen Textarchivs handelt. Durch Anklicken der Stellenangabe kann dann eine elektronische Version der Textausgabe an der betreffenden Textstelle aufgeschlagen werden.

Dies ist in Abb. 5 für den Beleg aus Konrads von Würzburg Vers-Erzählung ‚Heinrich von Kempten' vorgeführt, der sich im gedruckten Artikel in Zeile 20-22 findet. Es handelt sich um eine Volltextanzeige, in der man beliebig weit nach vorne oder hinten zurückblättern bzw. -rollen kann, um so die Interpretation eines Textbelegs, wie sie der Wörterbuchartikel ausdrücklich vornimmt oder durch den Artikelzusammenhang zu verstehen gibt, im ursprünglichen Textzusammenhang zu prüfen oder vertiefend nachzuvollziehen.

Zusammenfassend lässt sich zur Leitfrage der Tagung festhalten: Der Nutzen elektronischer Textcorpora für die Ausarbeitung von historisch-philologischen Wörterbüchern ist evident, auch wenn diese Einsicht – wenigstens für die historische Lexikographie des Deutschen – noch nicht hinreichend in die Tat umgesetzt worden ist. Mindestens ebensosehr müsste die breite elektronische Fundierung von Wörterbüchern, aber auch von anderen Hilfsmitteln der historischen Linguistik, eine Forderung der Nutzer dieser Hilfsmittel sein, wie der Kurzdurchgang durch MWB Online zeigen sollte.

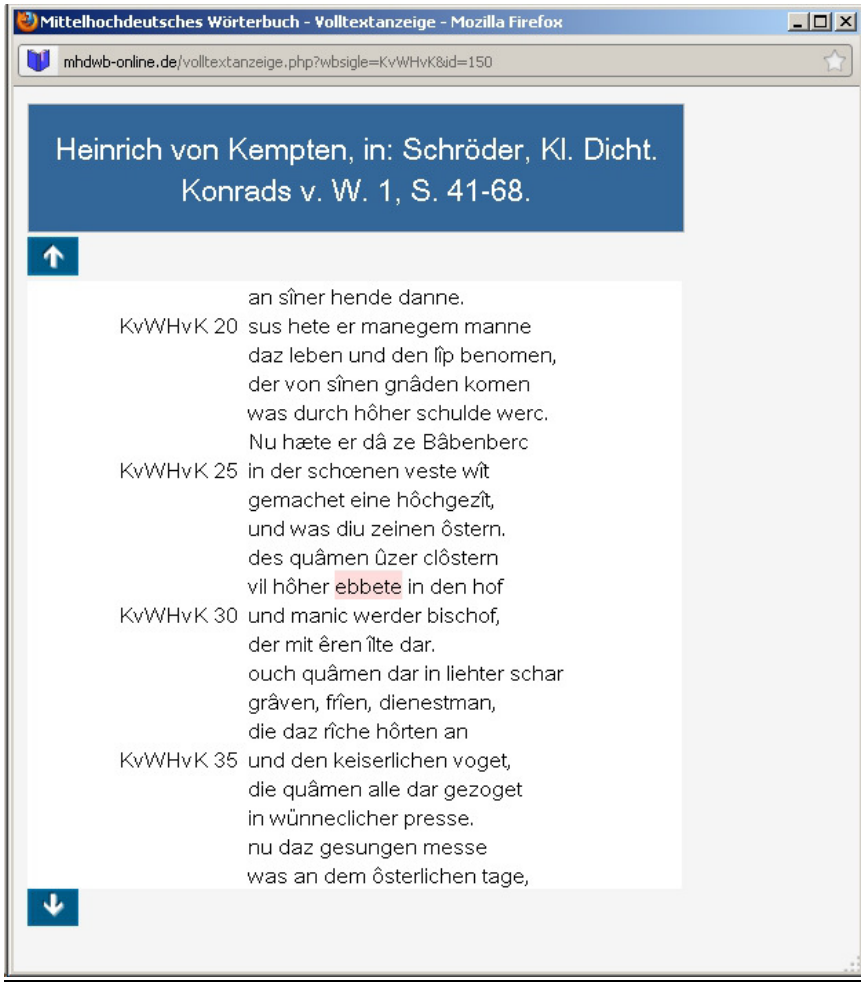


Abb. 5: MWB Online: Volltextanzeige eines Wörterbuchbelegs (Konrad von Würzburg ‚Heinrich von Kempten' v. 29)

Netzadressen der erwähnten Hilfsmittel (dort auch die bibliographischen Angaben zu den Druckwerken) und von Seiten mit weiterführenden Informationen:

http://www.MWV.uni-trier.de/ (Die Vorgängerwörterbücher im Mhd. Wörterbuchverbund [MWV] online).

http://www.mhdwb-online.de/ (Online-Angebot des MWB).

http://www.mwb.uni-trier.de/ (Homepage der Trierer Arbeitsstelle des MWB).

http://www.zfda.de/beitrag.php?id=782&mode=maphilinet (Überblicksartikel zur Online-Lexikographie des mittelalterlichen Deutsch, aus ZfdA 138,1 [2009]).

http://www.uni-trier.de/fileadmin/forschung/maw/MWB/Plate/Digitale_Editionen.pdf (Referat über die E-Texte des MWB; Abstract: http://www.telota.de/telota/nachrichten/abstracts_edworkshop/plate).

DAS BONNER FRÜHNEUHOCHDEUTSCH-KORPUS UND DAS REFERENZKORPUS ‚FRÜHNEUHOCHDEUTSCH'

HANS-CHRISTIAN SCHMITZ, BERNHARD SCHRÖDER & KLAUS-PETER WEGERA

1. Einleitung

Das Bonner Frühneuhochdeutsch-Korpus

Es gibt zwei Bonner Korpora des Frühneuhochdeutschen, nämlich erstens ein ‚großes' Korpus aus 1500 Texten, welches wir hier das Gesamtkorpus nennen. Das Gesamtkorpus ist nicht in maschinenlesbarer Form vorhanden. Zweitens gibt es ein elektronisches Korpus, das aus 40 Texten des Gesamtkorpus besteht. Wir nennen dieses Korpus das elektronische Teilkorpus oder kürzer das Teilkorpus.

Das Gesamtkorpus wurde zwischen 1972 und 1974 an der Forschungsstelle Frühneuhochdeutsch der Universität Bonn zusammengestellt, um „in großer Breite die Textüberlieferung des 14.-17. Jahrhunderts für sprachgeschichtliche Untersuchungen verschiedenster Art bereit[zu]stellen" (HOFFMANN & WETTER, 1985, XIV). Die Texte des Gesamtkorpus entstammen 22 Sprachlandschaften und sieben Zeitschnitten à 50 Jahre (1350-1700).

Das Auswahlkorpus entstand zwischen 1972 und 1985 im Rahmen des Projekts ‚Flexionsmorphologie des Frühneuhochdeutschen' unter Leitung von Werner Besch, Winfried Lenders (ab 1976), Hugo Moser und Hugo Stopp (bis 1981). Es diente als Materialgrundlage zur Analyse der Flexionsmorphologie des Frühneuhochdeutschen und zur Erarbeitung von mehreren Bänden der Grammatik des Frühneuhochdeutschen, nämlich zur Flexion der Substantive (WEGERA, 1987), zur Flexion der starken und schwachen Verben (DAMMERS *et al.*, 1988) und zur Flexion der Adjektive (SOLMS & WEGERA, 1991). Mit Fertigstellung dieser Bände hatte das Korpus seinen primären Zweck erfüllt. Der Datensatz wurde daraufhin nicht mehr genutzt, bis er im Jahr 2002 nach XML und HTML transferiert und über das WWW öffentlich bereitgestellt wurde. Das Korpus ist seitdem in verschiedenen Formen und über verschiedene Schnittstellen allgemein verfügbar. Es wird in Forschung und Lehre genutzt.

Das Referenzkorpus ‚Frühneuhochdeutsch'

Im Rahmen eines im Oktober 2011 begonnenen Projekts wird an der Ruhr-Universität Bochum in Zusammenarbeit mit den Universitäten

Halle und Potsdam ein Referenzkorpus des Frühneuhochdeutschen erstellt. Dieses Korpus wird mit über 300 Texten vom Umfang her kleiner sein als das Bonner Gesamtkorpus, aber weitaus größer als das elektronische Teilkorpus. Das Bonner Teilkorpus spielt für die Konstruktion des Referenzkorpus eine zentrale Rolle, da ein Großteil seiner Texte übernommen, korrigiert und an die Standards für Referenzkorpora angepasst wird. Das Referenzkorpus wird das Bonner Korpus mittelfristig in Forschung und Lehre ablösen.

Gliederung des Artikels

Der vorliegende Artikel ist wie folgt gegliedert: In Abschnitt 2 referieren wir kurz den Umfang und die Konzeption des Bonner elektronischen Teilkorpus. Daraufhin erläutern wir in Abschnitt 3 die XML-Kodierung und die Zugriffsmöglichkeiten auf das Korpus, insbesondere über eine HTML-Darstellung für die explorative Erschließung (FnhdC/HTML) und über eine Maske für die gezielte Durchsuchung des Korpus (FnhdC/S). In Abschnitt 4 nennen wir beispielhaft einige Anwendungen in Forschung und Lehre, bevor wir in Kapitel 5 den Blick auf das Referenzkorpus ‚Frühneuhochdeutsch' und damit die Zukunft des Bonner Korpus richten. Wir übernehmen in diesem Artikel Passagen aus dem Arbeitsbericht zur XML-Kodierung des Bonner Frühneuhochdeutsch-Korpus von DIEL *et al.* (2002).

2. Konzeption und Umfang des Bonner Teilkorpus

Die Anfänge des Bonner Frühneuhochdeutsch-Korpus liegen im Beginn der 1970er Jahre. Die Grundlagen, die Zusammenstellung und die Probleme dieser frühen Phase der computerunterstützten Grammatikographie sind ausführlich dokumentiert in SOLMS & WEGERA (1998).

Das elektronische Teilkorpus besteht aus 40 Texten, respektive Textausschnitten. Die Texte sind nicht immer vollständig in das Korpus aufgenommen, sondern jeweils zu einem Ausschnitt von ca. 30 Normalseiten mit etwa 400 Wortformen. Der relativ kleine Umfang ist durch den damals großen Aufwand der Digitalisierung begründet. Der Umfang reicht allerdings aus, um die flexionsmorphologische Exemplarität des Korpus zu gewährleisten und dadurch seinen Zweck als Materialgrundlage für die Erarbeitung von drei der vier flexionsmorphologischen Bände der Grammatik des

Frühneuhochdeutschen zu erfüllen.[1] Die Texte sind durchgängig mit Wortklassen- und Formenbestimmungen versehen (s.u., Abschnitt 3).

Die Korpustexte entstammen zehn verschiedenen Sprachlandschaften, nämlich dem Mittelbairischen (Wien), Schwäbischen, Ostfränkischen (Nürnberg), Obersächsischen, Ripuarischen (Köln), Osthochalemannischen, Oberschwäbischen (Augsburg), Elsässischen (Straßburg), Hessischen und Thüringischen. Die Texte entstammen außerdem vier Zeitschnitten à 50 Jahre, nämlich 1350-1400, 1450-1500, 1550-1600 und 1650-1700. Jede Sprachlandschaft ist mit insgesamt vier Texten vertreten, wobei diese Texte aus den verschiedenen Zeitschnitten stammen; entsprechend ist jeder Zeitschnitt mit zehn Texten aus den zehn verschiedenen Landschaften vertreten.

Das Korpus wurde nachträglich ergänzt durch insgesamt 20 Texte für die Zeiträume 1500-1550 und 1600-1650 (mit jeweils einem Text pro Sprachlandschaft); darüber hinaus wurden 22 weitere Texte für die zusätzlichen Sprachlandschaften (Nord-)Hessisch, Schlesisch, Westhochalemannisch, Mittelbairisch 2 (München) und den norddeutschen Raum zusammengestellt. Diese Ergänzungen wurden allerdings nicht digitalisiert und sind ergo im elektronischen Teilkorpus nicht vorhanden. Auch nachträgliche Korrekturen wurden nicht immer in die elektronische Fassung eingearbeitet.

Das Korpus stellt eine authentische Mischung verschiedener Textsorten dar: Insgesamt sind fünf kirchlich-theologische Fachtexte, zwölf chronikalische und Berichtstexte, zwölf erbauliche Texte, ein Bibeltext, sechs unterhaltsame Texte (‚schöne Literatur') und vier Realientexte (Fachprosa) vorhanden. Es fehlen metasprachliche Texte, Texte mit gebundener Sprache, Privattexte, Rechts- und Geschäftstexte. Die Authentizität der Textsortenmischung gilt für das gesamte Teilkorpus. Innerhalb der Sprachlandschaften und der Zeitschnitte liegt allerdings eine weitgehend zufällige Textsortenmischung vor. Wenn also beispielsweise die Texte eines Zeitschnitts verglichen werden, dann können festgestellte Unterschiede nicht ohne weiteres als Unterschiede zwischen den verschiedenen Sprachlandschaften gedeutet werden. Es ist möglich, dass die Unterschiede vielmehr textsortenbedingt sind. Wollte man nun den Einfluss der Textsorte überprüfen, müssten für jede Sprachlandschaft und jeden Zeitschnitt systematisch alle verschiedenen Textsorten vorliegen. Das ist nicht der Fall.

---

[1] Band V der Grammatik des Frühneuhochdeutschen – WALCH (1988) – basiert nicht auf dem Korpus.

Die Auswahl der Texte wurde an strenge Qualitätskriterien gebunden, die allerdings nicht immer eingehalten werden konnten. Grundsätzlich galt, dass ein Text nur dann aufzunehmen sei, wenn erstens seine Datierung und Lokalisierung belegt oder zu ermitteln war und zweitens der Verfasser bzw. Hersteller aus der Entstehungslandschaft stammte oder wenigstens zehn Jahre lang in ihr tätig war. Schließlich sollten zwischen dem Verfassen des Originaltexts und der Herstellung der gegebenen Textfassung nicht mehr als 50 Jahre liegen.

Für die Digitalisierung der Texte – hierin liegt zweifellos ein Manko des Korpus – wurden teilweise Editionen verwendet, ohne dass diese mit den originalen Handschriften abgeglichen wurden.

Die Zusammenstellung des Korpus hat sich für die Flexionsmorphologie insbesondere der Substantive bewährt. Die Belegdichte für Adjektive und Verben ist hingegen geringer und manchmal sogar unbefriedigend. SOLMS & WEGERA (1998: 27) stellen fest, dass das Korpus „bezüglich der Belegverhältnisse in seiner Gesamtheit zumeist hinreichend [ist]; unzureichend wird es bei grundsätzlich oder textsortenbezogen wenig frequenten Phänomenen. Die Erfahrung zeigt jedoch auch, daß solche Probleme durch entsprechende Zusatzexzerption gelöst werden können.“ Während die flexionsmorphologischen Phänomene relativ unabhängig von Textsorten vorkommen, sind Wortbildungsphänomene stark textsortenabhängig. Die fehlende Textsortenorientierung wird dabei bei der Untersuchung von Wortbildungsphänomenen spürbar. Das Korpus kann hier nur bedingt als Materialgrundlage dienen.

Fassen wir die Vor- und Nachteile des Korpus zusammen: Der entscheidende Vorteil des Bonner Teilkorpus ist, dass es existiert. Es ist frei verfügbar, kann gelesen und durchsucht werden und ist insofern schon mangels Alternativen eine wichtige Ressource der Sprachwissenschaft. Es handelt sich bei ihm ferner um einen gut durchdachten, wohl strukturierten Bestand, der sowohl einen hohen Qualitätsstandard hinsichtlich der Quellenauswahl als auch der Exemplarität bezüglich flexionsmorphologischer Phänomene beanspruchen kann. Das Korpus aus den 1970er Jahren wird allerdings nicht den Standards moderner Referenzkorpora gerecht: Seine Texte basieren z.T. auf Editionen, und sie sind nicht vollständig, sondern nur in Ausschnitten im Korpus vorhanden. Das Korpus hat mit 40 Teiltexten einen relativ geringen Umfang, so dass es zur systematischen Untersuchung von Phänomenen jenseits der Flexionsmorphologie nur bedingt geeignet ist. Hier fällt insbesondere die umfangs-

bedingt geringe Repräsentanz einzelner Textsorten ins Gewicht, aufgrund derer es nicht die Materialgrundlage für die erschöpfende Untersuchung von textsortenabhängigen Phänomenen wie solchen der Wortbildung bieten kann. Schließlich wurden Korrekturen und Ergänzungen des Korpus nur teilweise eingearbeitet, so dass die vollständige Richtigkeit des Korpus nicht immer garantiert ist.

3. Kodierung und Zugriffsmöglichkeiten

Kommen wir zur Kodierung der digitalisierten Texte und den Möglichkeiten des Zugriffs auf die Daten: Die erste Version des elektronischen Teilkorpus wurde beginnend 1972 auf Lochkarten gespeichert, dann in den Zeichensatz der MS-DOS Codepage 437 übertragen und auf Disketten gespeichert. Die durch diese Übertragung entstandene zweite Version nennen wir die Original-kodierte Version.[2] Die Lochkarten sind unseres Wissens nicht erhalten. 2002 wurde die Original-kodierte Version vollständig nach XML transformiert. Maßgabe der Transformation war, die Kodierung in den neuen Standard zu überführen und dabei inhaltlich so wenig wie möglich zu verändern. Offensichtliche Fehler wurden jedoch korrigiert und Kodierungen in den Originaltexten, deren Bedeutung trotz Konsultation von LENDERS & WEGERA (1982) und Befragung ehemaliger Projektmitglieder nicht festgestellt werden konnte, wurden gelöscht.[3] Auf Grundlage der XML-kodierten Version wurde im gleichen Jahr eine ‚lesbare', an Information reduzierte HTML-Version erstellt. Im Jahre 2007 wurde die XML-Version einer Korrektur unterzogen. Darüber hinaus wurde eine neue HTML-Version erzeugt (FnhdC/HTML, s.u.) und eine Suchmaschine für das Korpus implementiert (FnhdC/S, s.u.). Alle vorhandenen Versionen des Korpus – die Original-kodierte Version, die XML-Version und die beiden HTML-Versionen – wurden als komprimierte Archive über die Webseite korpora.org verfügbar gemacht.[4]

Die XML-Kodierung ist in einer kommentierten Document Type Definition (DTD) und in DIEL *et al.* (2002) beschrieben:[5] In den

---

[2] Sie stimmt nicht in allen Punkten mit der von BERG (1982) beschriebenen ersten Version überein.

[3] Den ehemaligen Projektmitgliedern war klar, dass nicht alle Korrekturen in das elektronische Teilkorpus eingearbeitet wurden und dass es ergo nicht vollständig richtig ist. Daher rührte ihre Skepsis bezüglich der Veröffentlichung des Korpus.

[4] http://www.korpora.org/Fnhd

[5] Die DTD befindet sich zusammen mit den XML-Dateien im komprimierten Archiv. Ferner steht sie im Anhang von DIEL *et al.* (2002).

Texten sind Seiten, Blätter, Kapitel und Zeilen markiert und nummeriert.[6] Darüber hinaus sind Eingriffe (Bearbeitungen), Hervorhebungen, Überschriften, Zitate und Namen als solche annotiert. Endlich sind Substantive, Adjektive und Verben mit morphosyntaktischen Angaben versehen. Dazu gehören erstens die Wortklassen. Zweitens werden für Adjektive und Substantive, deren Formen bestimmt werden konnten, Kasus, Numerus und Genus angegeben. Entsprechend werden von Verben, deren Formen bestimmt werden konnten, Tempus, Modus, Person und Form (Infinitiv, Partizip, finite Form) genannt. Drittens sind Flexive, Lemmata, Prä- oder Suffixe in normalisierter Form und die Vokale der Stammsilbe gesondert aufgeführt.

Zu mehreren Wortformen sind im Annotationsteil des Originalkodierten Korpus zwei Wortklassen angegeben. Dabei handelt es sich zumeist um substantivierte Infinitive, die sowohl als Substantive als auch als Infinitive annotiert sind. Beide Angaben, für den Erst- und den Zweittyp, wurden in die XML-Version überführt.

Die Zeichen der XML-Version sind gemäß dem Unicode™-Standard in UTF-8 kodiert. Auf Sonderzeichen wird mit Referenzen verwiesen. Alle Zeichen mit Diakritika sind aufgespalten. Sie bestehen aus Grundzeichen und darauf folgendem Diakritikum. Ein ,,ä“ ist somit nicht als einzelnes Zeichen kodiert (wie in ISO-8859-1), sondern als ,,a“ gefolgt vom Diakritikum „ ¨ “; ein „å“ ist kodiert als ,,a“ gefolgt vom Diakritikum „ ̊ “.

Alle Texte des Korpus sind über ein Quellenverzeichnis erschlossen, das wie die Texte selbst in XML- und HTML-kodierten Fassungen vorliegt. Die XML-kodierten Daten sind aufgeteilt in Basis- und Zusatzdaten. Die Basisdaten umfassen Kurztitel, Titel, Autoren, Quellen und Herausgeber, Erscheinungsorte, -jahre und andere Editionsspezifika, Sprachlandschaften, Entstehungszeiten sowie die Anzahl der aufgenommenen Seiten. Die Zusatzdaten umfassen editorische Anmerkungen, eventuelle Vorlagen, Biographien der Verfasser, Übersetzer, Schreiber und Drucker, die Textarten sowie eventuelle editorische Eingriffe (z.B. Kürzungen).

Aus der 2002er HTML-Version des Korpus wurden alle linguistischen Annotationen entfernt, um die Texte leicht lesbar zu machen. Diese Lösung war nicht optimal, weil die für die Rezeption wertvolle linguistische Zusatzinformation verloren ging. Deshalb wurde 2007

[6] Genauer: Zur Vermeidung konkurrierender Strukturen sind Seiten-, Blatt-, Kapitel- und Zeilen*wechsel* markiert.

eine neue Version erstellt (FnhdC/HTML, vgl. FISSENI *et al.*, 2007), die sowohl lesbar ist als auch die linguistischen Annotationen enthält. In ihr werden die linguistischen Annotationen für eine Wortform angezeigt, wenn man den Mauszeiger auf die entsprechende Wortform legt. Abbildung 1 zeigt einen Ausschnitt von Wilhelm Durandus' Rationale (Text 111[7] des Korpus) in dieser Darstellung. Der Mauszeiger steht hier auf dem Wort „chůnig". Neben dem Text werden die zugehörigen morphosyntaktischen Informationen wie der Stammvokal, die Wortart und wortartspezifische Angaben, hier Kasus, Numerus und Genus, angezeigt. Außerdem zeigt die Lemmaangabe das Wort in der neuhochdeutschen Schreibung.

Abbildung 1: Darstellung eines Ausschnitts von Text 111 (Wilhelm Durandus: Rationale, Wien 1384) in FNHD/HTML.

[7] Zur Nummerierung der Korpustexte: Das Korpus ist in zwei Hälften à 20 Texte geteilt. Der erste Teil trägt die Nummer 1, der zweite die Nummer 2. In jedem Teil sind Texte aus fünf Sprachlandschaften vertreten, jeweils nummeriert von 1-5. Schließlich wurden die Zeitschnitte mit den Zahlencodes 1, 3, 5, und 7 versehen. Die Nummer eines Textes ist eine dreistellige Ziffernfolge, bestehend aus dem Code für den Teil des Korpus, die Sprachlandschaft und den Zeitschnitt. Durandus' Rationale trägt die Nummer 111, weil der Text dem ersten Teil, darin der ersten Sprachlandschaft (Mittelbairisch, Wien) und dem ersten Zeitschnitt (1350-1400) entstammt.

Zusätzlich wurde eine Suchmaschine implementiert, mittels derer das Korpus und seine einzelnen Texte durchsucht werden können (FnhdC/S, vgl. FISSENI *et al.*, 2007). Die Herausforderung bestand darin, einerseits komplexe Suchanfragen auch unter Einbezug der Annotationen zuzulassen und andererseits die effektive Nutzbarkeit auch für technisch wenig affine Linguisten zu gewährleisten, die Bedienung also einfach zu machen. Diese Aufgabe wurde mithilfe einer Suchmaske gelöst, die die Eingabe von (Mustern von) Wortformen oder Lemmata und die Auswahl linguistischer Eigenschaften (Kasus, Numerus, etc.) ermöglicht. Abbildung 2 zeigt einen Ausschnitt der Suchmaske mit Eingaben für die Suche nach Substantiven im Dativ Singular mit dem Lemma „erde“. Abbildung 3 zeigt einen Ausschnitt des entsprechenden Suchergebnisses.

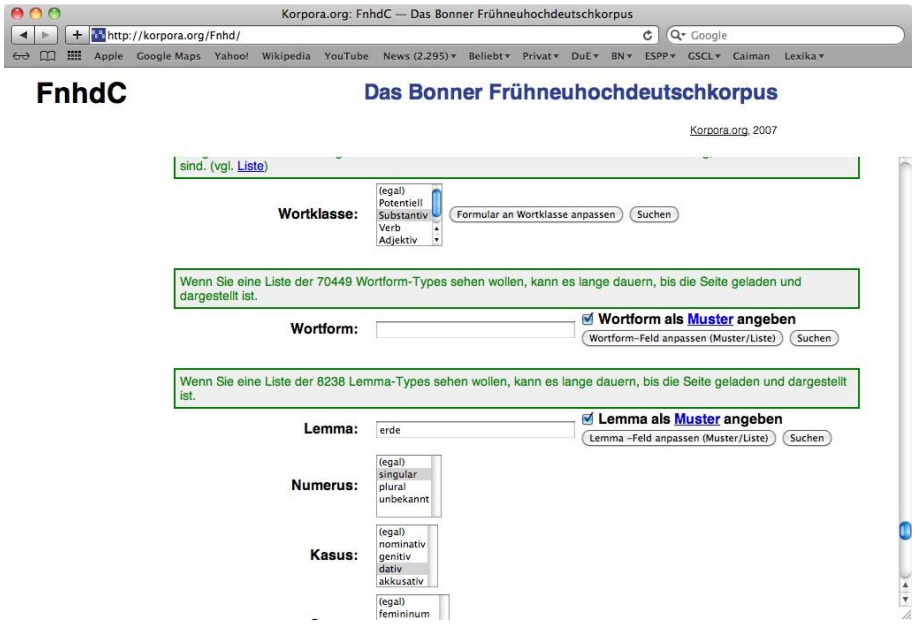


Abbildung 2: Suchanfrage nach den Flexionsformen von Erde im Dativ Singular in FnhdC/S.

Abbildung 3: Ergebnisdarstellung der Suchanfrage von Abbildung 2.

4. Nutzung in Forschung und Lehre

An einigen Nutzungsbeispielen der XML-Fassung des Bonner Frühneuhochdeutsch-Korpus in Forschung und Lehre lässt sich illustrieren, welcher Mehrwert sich bereits aus der wortbezogenen morphosyntaktischen Etikettierung des Korpus und seinen diachronen und diatopischen Auswahlprinzipien ergibt.

Das Bonner Frühneuhochdeutsch-Korpus stellt die Materialgrundlage für flexionsmorphologische Untersuchungen des Frühneuhochdeutschen dar, wie sie in oben genannten Bänden der von MOSER *et al.* (1970 ff.) herausgegebenen Grammatik des Frühneuhochdeutschen ihren Niederschlag fand. Die linguistische Etikettierung der Texte war genau auf diese Untersuchungszwecke abgestellt. Die Korpusannotation erweist sich allerdings auch in anderen Bereichen der Morphologie als hilfreich. So arbeiteten PRELL (1991), PRELL & SCHEBBEN-SCHMIDT (1996) und DOERFERT (1994) mit diesem Korpus im Kontext von Untersuchungen zur frühneuhochdeutschen Wortbildungsmorphologie.

Die XML-Annotation ist gegenüber über die Wortgrenze hinausgehenden syntaktischen oder semantischen Etikettierungen offen. Helmut Weiß und Anna Voldina (Goethe-Universität Frankfurt, IDS

Mannheim) untersuchen die diachrone Entwicklung von Null-Argumenten im Deutschen anhand mittelhochdeutscher und frühneuhochdeutscher Texte. So zeigt sich, dass frühere Sprachstufen des Deutschen offenbar ausgeprägtere Möglichkeiten der Pronominalellipse kannten als das Neuhochdeutsche. VOLODINA & ONEA (2012) zitieren das Beispiel aus dem Korpus-Text 113 *Die Denkwürdigkeiten der Helene Kottannerin*, das verdeutlicht, dass hier im mit *ob* eingeleiteten Nebensatz das Personalpronomen der ersten Person in Subjektposition nach einem ebensolchen Subjekt im Hauptsatz im Gegensatz zum Neuhochdeutschen nicht realisiert zu werden braucht:

(1) ich solt auf das haws vnd solt versu̍ehen, ob ø i̍r kran vnd ander ir klainat mocht hinab zu i̍r bringen [...][8]

Zum Zweck dieser Untersuchung planen Weiß und Volodina, ein Korpus mittel- und frühneuhochdeutscher Texte, um die entsprechenden Markierungen von Pro-Formen und Nullrealisierungen von Argumenten zu ergänzen, darunter auch Texte des Frühneuhochdeutsch-Korpus. Eine Pilotstudie verwendet die bereits erwähnten *Denkwürdigkeiten der Helene Kottannerin* aus dem Bonner Teilkorpus.

Das elektronische Korpus in der gegenwärtigen Präsentationsform erweist sich auch in der universitären Lehre als überaus nutzbringend. Die Anzeige der Lemmata zumeist in der neuhochdeutschen Schreibung hilft Studierenden bei der Erschließung der Texte, wie sich in verschiedenen Seminaren zu Themen des Sprachwandels und der Sprachgeschichte an der Universität Duisburg-Essen gezeigt hat. Hier erweist sich insbesondere die Darstellungsform einer ‚Glossierung on-demand' als günstig. Im Gegensatz zum Angebot einer Übersetzung bleibt das Augenmerk auf den Originaltext gelenkt. Dass die Glossierung mithilfe des Mauszeigers für Einzelwörter angefordert werden muss und nicht ständig für alle Wortformen eingeblendet ist, stellt einen Anreiz dar, sich auf die historische Schreibung einzulassen (vgl. Abbildung 1 im vorherigen Abschnitt 3).[9]

---

[8] KOTTANNERIN, Bonner Frühneuhochdeutschkorpus, Text 113, S. 13, Z. 13f., Markierung des Nullsubjekts mit „ø" nach VOLODINA & ONEA (2012).

[9] Die vereinheitlichte Lemmaschreibung wird auch im Forschungskontext genutzt: M. FISCHER (2003) beispielsweise greift in seinen Untersuchungen zum Gebrauch von „zweifeln" im Frühneuhochdeutschen ausdrücklich darauf zurück.

Die wortformenbezogene Suche nach morphosyntaktischen Merkmalen wurde in verschiedenen Seminaren zum Sprachwandel an der Universität Duisburg-Essen illustrativ und explorativ eingesetzt. Dies bietet sich beispielsweise bei Kurseinheiten zum Flexionswandel an. So lässt sich die diachrone und diatopische Verteilung der Endungen -(*e*) und -(*en*) auf die Singular-Kasus von *Erde* leicht durch FnhdC/S-Abfragen für die einzelnen Kasus, wie in den in Abschnitt 3 gezeigten Abbildungen 2 und 3 für den Dativ-Singular gezeigt, erkunden. Das unterschiedlich schnelle Fortschreiten der Kasusnivellierung bei den einzelnen Kasus wird unmittelbar veranschaulicht.

Auch für die didaktisch motivierte Erkundung von Phänomenen des Syntaxwandels hat sich FnhdC/S schon als nützlich erwiesen, obgleich weder nach Phrasen noch nach syntaktischen Funktionen gesucht werden kann. Für die tendenziell fortschreitende Rechtsverschiebung der rechten Satzklammer im vom Korpus abgedeckten Zeitraum findet man allein durch die Suche nach Infinitiven und Partizipien leicht eine Fülle an illustrierendem Material. Die folgende Abbildung 4 zeigt das Ergebnis der Suchabfrage nach Infinitiven und Partizipien.

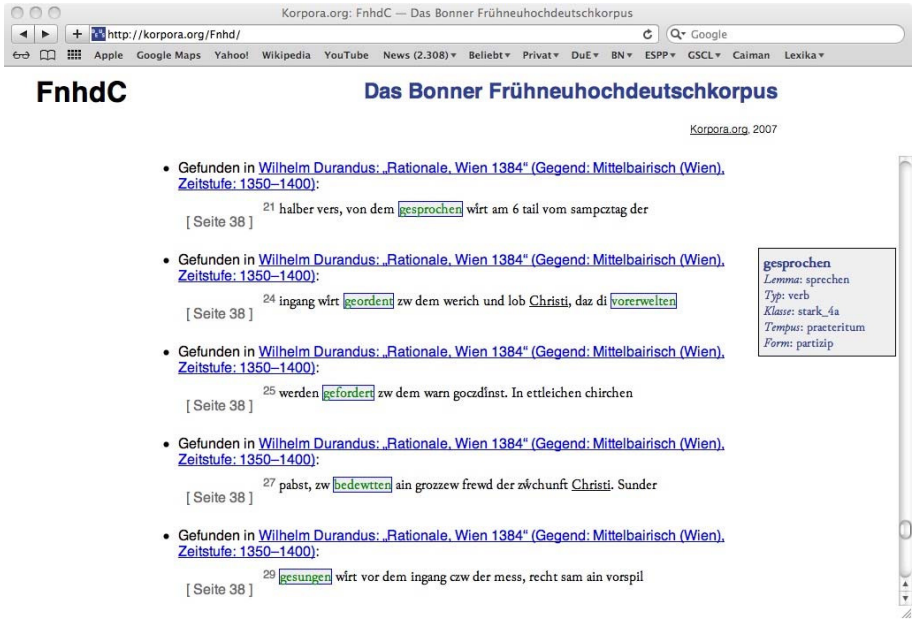


Abbildung 4: Ergebnis der Suchabfrage nach Infinitiven und Partizipien in Text 111 (Wilhelm Durandus: Rationale, Wien 1384).

5. Ausblick: Das Referenzkorpus ‚Frühneuhochdeutsch'

Im Rahmen des umfassenden Historischen Referenzkorpus des Deutschen (früher DDD = Deutsch Diachron Digital) wurde das Bonner Frühneuhochdeutsch-Korpus als vorhandenes, teillemmatisiertes und teilannotiertes Textarchiv erneut interessant. Die Planung des Historischen Referenzkorpus – eigentlich ein Archiv, das als Korpus genutzt werden kann[10] – sieht ein Textarchiv vor, das von den Anfängen der deutschsprachigen Überlieferung bis ca. 1800 reicht. Dabei werden die überlieferten (und erhaltenen) Handschriften bis 1200 komplett erfasst,[11] für die Zeiträume danach wird jeweils eine strukturierte Textauswahl für das Archiv aufbereitet.

Für die Erarbeitung des Frühneuhochdeutsch-Referenzkorpus spielt das Bonner Frühneuhochdeutsch-Korpus eine zentrale Rolle. 31 der 40 Texte[12] der zweiten Hälfte des 14. Jahrhunderts bis 1600 – erweitert um die (nicht-digitalisierte) Auswahl 1600-1650 – werden aus dem Bestand übernommen, korrigiert und an die Standards der Referenzkorpora angepasst. Die Textausschnitte werden von 12.000 Wortformen auf max. 20.000 Wortformen erweitert. Die 14 Texte, die bisher nur nach der jeweiligen Edition eingelesen sind, sollen zusätzlich nach der Handschrift beziehungsweise dem Druck bearbeitet werden.

Die Annotation wird umfangreicher sein als dies bisher der Fall ist; es werden *alle* Wortformen annotiert. Die Annotation wird neben

---

[10] Vgl. WEGERA (2013) zur Unterscheidung von Archiv und Korpus.

[11] Dies ist mit den derzeitigen technischen Mitteln innerhalb eines finanzierten Zeitrahmens möglich. Die Überlieferung des Althochdeutschen (und Altsächsischen) wird in Zusammenarbeit der Humboldt-Universität Berlin (Karin Donhauser, Anke Lüdeling), der Goethe-Universität Frankfurt (Jost Gippert) und der Friedrich-Schiller-Universität Jena (Rosemarie Lühr) erarbeitet; das Mittelhochdeutsche an der Ruhr-Universität Bochum (Klaus-Peter Wegera, Stefanie Dipper) und der Friedrich-Wilhelms-Universität Bonn (Thomas Klein, Claudia Wich-Reif), das Frühneuhochdeutsche ebenfalls an der Ruhr-Universität Bochum (Klaus-Peter Wegera, Stefanie Dipper), an der Martin-Luther-Universität Halle (Hans-Joachim Solms) und der Universität Potsdam (Ulrike Demske). Geplant sind das Mittelniederdeutsche an der Westfälischen Wilhelms-Universität Münster (Robert Peters) und der Universität Hamburg (Ingrid Schröder) und das Frühe Moderne Deutsch (1650-1800) in Zusammenarbeit deutscher Universitäten (Bochum, Halle, Potsdam) mit der University of Manchester (Martin Durrell) und der Universitetet i Oslo (Kirsten Bech). Zu diesem Zweck wird das Manchester-Korpus (GerManC; vgl. DURRELL *et al.*, 2007) an die Standards der übrigen Referenzkorpora angepasst und erweitert.

[12] Die Texte aus dem Zeitraum 1650-1700 werden evtl. im Rahmen der Überarbeitung des GerManC Teil des Referenzkorpus Frühes Modernes Deutsch (1650-1800) berücksichtigt.

den Wortarten (POS) im Falle der flektierten Wortarten jeweils eine Angabe der Flexionsklasse und Angaben zu dem flexionsmorphologischen Merkmalen Kasus, Numerus und Genus (Substantive, Adjektive und Pronomina), Person, Numerus, Modus und Tempus (Verben) beinhalten. Eine syntaktische Annotation erfolgt für eine kleinere, noch nicht festgelegte, Auswahl von Texten des Korpus.[13] Für das Referenzkorpus wurde das STTS für Erfordernisse der Annotation historischer Texte modifiziert (STTS(H)). Mit diesem Tagset annotierte Texte werden in das XML-Standoff-Format PAULA (Potsdamer Austauschformat für linguistische Annotationen)[14] überführt und über die linguistische Datenbank ANNIS (Annotation von Informationsstruktur)[15] verfügbar gemacht.

Es stellt sich die Frage, ob und in welcher Form das Bonner Frühneuhochdeutsch-Korpus neben dem Referenzkorpus weiter gepflegt werden soll. Da es sich um einen gut durchdachten strukturierten Bestand handelt, ist eine weiterhin separate Nutzung sicherlich möglich. Als Form wäre allerdings die erweiterte Fassung wünschenswert. Diese wäre dann aber auch über ANNIS möglich, da auf strukturierte Teilkorpora als solche ein separater Zugriff möglich sein wird.

Es lässt sich vorhersehen, dass das Bonner Korpus für die Forschung und Lehre weitgehend vom Referenzkorpus ‚Frühneuhochdeutsch' abgelöst wird. Es bleibt bestehen als Beispiel für ein exemplarisches und hinsichtlich der Textauswahl hochwertiges Korpus, und möglicherweise besteht auch ein anhaltender Bedarf an der spezialisierten Oberflächen des Bonner Korpus (FnhdC/HTML und FnhdC/S). Bis zur Ablösung durch das Referenzkorpus fungiert das Bonner Korpus als Labor für die Untersuchung verschiedener Zugriffsmöglichkeiten und Analysen, zur Erhebung von Anforderungen für die eHumanities, als Prototyp zur Erprobung spezialisierter Suchmöglichkeiten und für den Einsatz in der Lehre.

---

[13] Zur Methode vgl. DEMSKE (1996).

[14] http://www.sfb632.uni-potsdam.de/en/paula-en.html

[15] http://www.sfb632.uni-potsdam.de/annis/; vgl. DIPPER (2005), CHIARCOS *et al.* (2008).

LITERATUR

BERG, E., 1982: Entwicklung eines Kodierungssystems am Beispiel frühneuhochdeutscher Texte, in: LENDERS, W. & K.-P. WEGERA (Hrsg.), *Maschinelle Auswertung sprachhistorischer Quellen*, Sprache und Information 3, Tübingen, 19-50.

CHIARCOS, C. *et al.*, 2008: A flexible framework for integrating annotations from different tools and tagsets, in: *Traitement Automatique des Langues* 49 (2), 217-246.

DAMMERS, U. *et al.*, 1988: *Flexion der starken und schwachen Verben*, Band IV der Grammatik des Frühneuhochdeutschen, hrsg. von MOSER, H. *et al.*, Heidelberg.

DEMSKE, U., 1996: Bestandsaufnahme zum Untersuchungsbereich ‚Syntax', in: FRITZ, G. & E. STRASSNER (Hrsg.), *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*, Tübingen, 70-125.

DIEL, M. *et al.*, 2002: *XML-Kodierung des Bonner Frühneuhochdeutschkorpus*, IKP-Arbeitsbericht NF 02, Bonn.

DIPPER, S., 2005: XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation, in: *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, Berlin, 39-50.

DOERFERT, R., 1994: *Die Substantivableitung mit -heit, -keit, -ida, -î im Frühneuhochdeutschen*, Berlin.

DURRELL, M. *et al.*, 2007: 'GerManC: A historical corpus of German 1650-1800', in: *Sprache und Datenverarbeitung* 31, 71-80.

FISCHER, M., 2003: Ein Zweifelsfall: *zweifeln* im Deutschen, in: ECKARDT, R. (Hrsg.), *Questions and Focus*, ZAS Papers in Linguistics 30, Berlin.

FISSENI, B. *et al.*, 2007: FnhdC/HTML und FnhdC/S, in: *Sprache und Datenverarbeitung* 31, 67-69.

GRASER, H. & K.-P. WEGERA, 1978: Zur Erforschung der frühneuhochdeutschen Flexionsmorphologie, in: *Zeitschrift für Deutsche Philologie* 97, 74-91.

HOFFMANN, W. & F. WETTER, 1985: *Bibliographie frühneuhochdeutscher Quellen. Ein kommentiertes Verzeichnis von Texten des 14.-17. Jahrhunderts (Bonner Korpus)*, Frankfurt am Main [u.a.].

LENDERS, W. & K.-P. WEGERA (Hrsg.), 1982: *Maschinelle Auswertung sprachhistorischer Quellen. Ein Bericht zur computerunterstützten Analyse der Flexionsmorphologie des Frühneuhochdeutschen*, Tübingen.

MOSER, H. *et al.* (Hrsg.), 1970 ff.: *Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre*, Heidelberg.

PRELL, H.-P., 1991: *Die Ableitung von Verben aus Substantiven in biblischen und nichtbiblischen Texten des Frühneuhochdeutschen*, Frankfurt am Main.

PRELL, H.-P. & M. SCHEBBEN-SCHMIDT, 1996: *Die Verbableitung im Frühneuhochdeutschen*, Studia linguistica Germanica 41, Berlin.

SOLMS, H.-J. & K.-P. WEGERA, 1991: *Flexion der Adjektive*, Band VI der Grammatik des Frühneuhochdeutschen, hrsg. von MOSER, H. *et al.*, Heidelberg.

SOLMS, H.-J. & K.-P. WEGERA, 1998: Das Bonner Frühneuhochdeutsch-Korpus. Rückblick und Perspektiven, in: BERGMANN, R. (Hrsg.), *Probleme der Textauswahl für ein elektronisches Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung 1. und 2. November 1996*, Stuttgart, Leipzig.

VOLODINA, A. & E. ONEA, 2012: Am Anfang war die Lücke, in: BÄR, J. & M. MÜLLER (Hrsg.), *Geschichte der Sprache – Sprache der Geschichte. Probleme und Perspektiven der historischen Sprachwissenschaft des Deutschen*, Berlin, 207-237.

WALCH, M., 1988: *Flexion der Pronomina und Numeralia*, Band V der Grammatik des Frühneuhochdeutschen, hrsg. von MOSER, H. *et al.*, Heidelberg.

WEGERA, K.-P., 1987: *Flexion der Substantive*, Band III der Grammatik des Frühneuhochdeutschen, hrsg. von MOSER, H. *et al.*, Heidelberg.

WEGERA, K.-P., 2013, im Druck: Language Data Exploitation. Design and Analysis of Historical Language Corpora, in: BENNETT, P. *et al.* (Hrsg.), *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and Interdisciplinary Perspectives on Language 3, Tübingen.

WEGE ZU EINEM HISTORISCHEN REFERENZKORPUS DES DEUTSCHEN: DAS PROJEKT DEUTSCHES TEXTARCHIV

ALEXANDER GEYKEN

1. Einleitung

Der Nutzen umfassender Referenzkorpora für die Sprachwissenschaft, digitalen Textsammlungen also, die ausgewogen nach Textsorten und hinreichend groß für den Gegenstand der Untersuchungen sind, ist unbestritten: Referenzkorpora dienen unter anderem als Basis für Forschungen zum Wortschatz, zur Wortgeschichte, zur Grammatik der Textorganisation oder zu kontrastiven Studien zum Vergleich Fachwortschatz – normaler Wortschatz (vgl. SINCLAIR 2005, LEMNITZER 2010).

Für die deutsche Gegenwartssprache existieren mit DeReKo (KUPIETZ 2010) und dem DWDS-Kernkorpus (GEYKEN 2007) zwei für die o.g. Fragestellungen geeignete Korpora. Für die älteren Stadien des Neuhochdeutschen (ca. 1650–1900) fällt die Situation derzeit weitaus weniger befriedigend aus[1]. Hierfür gibt es keine hinreichend großen, nach einheitlichen Standards aufbereiteten und übergreifend abfragbaren Korpora, die als Referenzkorpora verwendbar wären und somit eine Grundlage für die o.g. Untersuchungen bilden könnten. Eine Reihe von Ursachen ist hierfür zu nennen, von denen im Folgenden die wichtigsten aufgeführt werden.

(1) Es fehlen bislang einheitliche verwendete Qualitätsstandards für die Erfassungsgenauigkeit auf Zeichenebene und die vorlagengetreue Wiedergabe der Textbasis. Dies gilt sowohl für die zahlreichen Einzelsammlungen, nicht zuletzt aber auch für die großen Textsammlungen Google Bücher, Wikisource, Zeno.org oder Gutenberg.org[2] und Gutenberg-DE. Diese unterscheiden sich von Referenzkorpora nicht nur durch die opportunistische Vorgehensweise bei der Textaufnahme, sondern auch auf der Ebene der

[1] Für das Althochdeutsche (8. Jh. bis ca. 1050) und das Mittelhochdeutsche (ca. 1050-1350) gibt es u.a. die Initiativen von Titus, der Trierer Arbeitsgruppe und die Initiative DeutschDiachronDigital (DDD). Für das Frühneuhochdeutsche wird das seit Herbst 2011 angelaufene Projekt zu einem 300 Texte umfassenden Frühneuhochdeutsch-Corpus im Rahmen von DeutschDiachronDigital (DDD) die Lage verbessern.

[2] Wir beschränken die Diskussion hier auf die deutschsprachigen Anteile dieser Sammlungen.

Meta- und Objektdaten sowie hinsichtlich der erzielten Erfassungsgenauigkeit und der Vorlagentreue von der für ein Referenzkorpus wünschenswerten Genauigkeit. Bei den Metadaten bleibt man beispielsweise bei Google Bücher oder Gutenberg-DE oft im Unklaren, welche Ausgabe dem Volltext zugrunde liegt; bei Zeno.org gibt es zu den Volltexten keine zugehörigen Bild-Digitalisate, was die Nachprüfbarkeit von Transkriptionen ohne einen Gang in die Bibliothek nahezu unmöglich macht. Gutenberg.org oder Wikisource enthalten oftmals modernisierte Transkriptionen und sind somit für manche sprachhistorische Untersuchungen nur von eingeschränktem Wert. Die Transkriptionsgenauigkeit von Google Bücher ist, da sie ausschließlich per OCR entstanden sind, bei historischen Werken ohne gründliche Nachkorrektur als Bestandteil eines Referenzkorpus nicht verwendbar.

(2) Textannotationsstandards, insbesondere die der Text Encoding Initiative (TEI; BURNARD & BAUMAN 2012) sind im wissenschaftlichen Kontext mittlerweile zunehmend verbreitet und stellen für historische Textsammlungen nahezu einen De-Facto-Standard dar[3]. Man sollte daher annehmen, dass dadurch die Austauschbarkeit der Daten als auch die Interoperabilität, also die unmittelbare Einsetzbarkeit TEI-kodierter Daten auf Korpusplattformen, gewährleistet ist. In der Praxis steht dem entgegen, dass für die Textkodierung verschiedene TEI-‚Dialekte' verwendet werden, deren Unterschiede im Allgemeinen so groß sind, dass die Interoperabilität nicht per se gegeben ist. In jüngerer Zeit wurden zwar einige Basisformate geschaffen (s. hierzu Abschnitt 3.2), um die Nutzbarkeit der Texte in verschiedenen Kontexten zu ermöglichen. Diese Basisformate haben bislang jedoch nur eine geringe Verbreitung gefunden.

(3) Bis vor kurzem wurden Korpusprojekte im Wesentlichen als Einzelprojekte durchgeführt, in denen die Textsammlungen projektspezifisch für einen bestimmten Anwendungszweck transkribiert und annotiert wurden. Erst in den letzten Jahren wurde, nicht zuletzt durch die strategischen Schritte der Forschungspolitik in

[3] Dies gilt nicht für die o.g. nicht in wissenschaftlichen Kontexten entstandenen Textsammlungen. Diese liegen entweder in html oder anderen proprietären nicht xml-Formaten vor und müssen erst in ein konsistentes TEI-Format konvertiert werden. Für einen Teil, vorrangig der literarischen Texte aus Zeno.org, ist dies schon geschehen (allerdings mit Informationsverlust), für Wikisource, Gutenberg.org und Gutenberg-DE werden Teile vom DTA konvertiert (www.deutschestextarchiv.de/dtae).

Richtung Open Access, die Weitergabe von Forschungsdaten thematisiert und rechtliche (offene Lizenzen[4]) und technische Rahmenbedingungen (interoperable Annotationsformate) geschaffen. Auf Forschungsebene wurden diese neuen technischen Möglichkeiten aber bislang noch nicht umfassend umgesetzt.

Mit dem Deutschen Textarchiv (DTA) und dessen Verfügbarkeit im Rahmen des großen Infrastrukturverbundes CLARIN-D[5] wird die technische Basis für ein dynamisch erweiterbares historisches Referenzkorpus geschaffen. Im Folgenden soll zunächst das DTA-Korpus als Grundstock für ein Referenzkorpus beschrieben werden (Abschnitt 2). In Abschnitt 3 werden die Anforderungen beschrieben, die sich für den Aufbau einer solchen Korpus-Infrastruktur ergeben. Abschnitt 4 fasst die Ergebnisse zusammen und gibt einen Ausblick auf weitere Arbeiten.

2. Der Grundstock: Das Deutsche Textarchiv

Ziel des von der Deutschen Forschungsgemeinschaft geförderten und an der BBAW beheimateten Projekts Deutsches Textarchiv (DTA) ist es, einen disziplinenübergreifenden Bestand deutschsprachiger Texte aus der Mitte des 17. bis zum Ende des 19. Jahrhunderts nach den Erstausgaben zu digitalisieren und als linguistisch annotiertes Volltextkorpus im Internet bereitzustellen. Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die ersten selbstständigen Ausgaben der jeweiligen Werke zugrunde gelegt. Die Volltexterfassung erfolgt möglichst vorlagengetreu und unter Verzicht auf textkritische Eingriffe und Kommentierungen. Hierzu werden die Texte in einem standardisierten Prozess größtenteils manuell (im *Double Keying*-Verfahren) erfasst. Dies ist aufgrund der Textvorlagen, die überwiegend in Fraktur vorliegen, bedeutend zuverlässiger als eine Texterfassung durch OCR (mit anschließender manueller Nachkontrolle).

Hinsichtlich der Entstehungszeit der für das DTA erfassten Texte sowie in Bezug auf die dabei berücksichtigten Textsorten wird eine größtmögliche Ausgewogenheit angestrebt. Derzeit[6] stehen im DTA Werke im Umfang von etwa 247.000 Seiten aus dem Zeitraum von

---

[4] Z.B. mit der Empfehlung, Forschungsdaten unter einer Creative Commons Lizenz zu veröffentlichen.

[5] CLARIN-D: eine web- und zentrenbasierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften, www.clarin-d.de.

[6] Stand Dezember 2011.

1780 bis 1900 als elektronische Volltexte und digitale Faksimiles zur Verfügung. In der zweiten Projektphase (Dezember 2010–2013) soll das Textkorpus auf die Zeit bis ca. 1650 ausgeweitet werden. Im Durchschnitt wird täglich ein weiteres Werk digitalisiert und über das Internet bereitgestellt. Mit einem geplanten Umfang von ca. 1.300 Texten des 17.–19. Jahrhunderts (ca. 100 Millionen Textwörtern bzw. 1 Milliarde Zeichen) entsteht mit dem Deutschen Textarchiv ein großes historisches TEI-kodiertes Kernkorpus deutschsprachiger Texte.

Dieses Korpus dient als Basis und Ausgangspunkt für das Referenzkorpus, welches sich aus dem DTA-Korpus und weiteren Texten aus externen Quellen speisen soll.

3. Anforderungen an eine Infrastruktur zum Aufbau eines integrierten Referenzkorpus

In diesem Abschnitt werden die notwendigen Anforderungen formuliert werden, die an eine technische Infrastruktur zum Aufbau eines integrierten historischen Referenzkorpus für das ältere Frühneuhochdeutsche zu stellen sind.

3.1 Aufbau eines Textsorteninventars

Der Aufbau eines Textsorteninventars sowie die Gewichtung der Textsorten untereinander stellt eine wichtige Voraussetzung für die Erstellung eines Referenzkorpus dar. Dies beinhaltet nicht, dass die dem Referenzkorpus zugrunde liegende Textsammlung alle Textsorten entsprechend ihrer Gewichtung enthalten muss. Es bedeutet jedoch, dass alle Texte über ausreichende Metadaten verfügen müssen, insbesondere Datierung, Textsorte, diatopische Merkmale sowie Textumfang. Aus diesen Metadaten lassen sich dann kriteriengestützt Referenzkorpora extrahieren, die den vorher festgelegten Textsorten und ihrer Gewichtung untereinander möglichst optimal Rechnung tragen. Ein entsprechendes Optimierungsverfahren wurde beim Auswahlprozess des DWDS-Kernkorpus aus einer etwa drei Mal so großen Kernkorpusbasis angewendet (GEYKEN 2007). Als Ergebnis entstand ein optimal nach Textsorten ausgewogenes Textkorpus mit einer Größe von 100 Millionen laufenden Textwörtern.

Neben der Auswahl der Textsorten und deren Gewichtung zueinander müssen auch für die eigentliche Textauswahl Kriterien bestimmt werden. Für das DTA, welches den Grundstock des Referenzkorpus bilden soll, fand die Textauswahl unter sprachwissenschaftlich-lexikographischen Gesichtspunkten statt. In das Korpus wurden

Werke aufgenommen, die in der Geschichte der (deutschsprachigen) Literatur oder für die Entwicklung wissenschaftlicher Disziplinen einflussreich waren und die intensiv rezipiert wurden. Neben solchen, als kanonisch geltenden Werken wurden auch einige weniger bekannte Texte berücksichtigt, um die Ausgewogenheit des Korpus zu erhöhen. Die Textauswahl bietet ein großes Spektrum an Genres der literarischen und wissenschaftlichen Produktion.

Die Verteilung der Texte soll hinsichtlich der unterschiedlichen Disziplinen und Textsorten möglichst ausgewogen sein. Die nachstehenden Diagramme 1 und 2 zeigen die diesbezügliche Verteilung der DTA-Texte auf Basis der jeweiligen Anzahl der Titel.

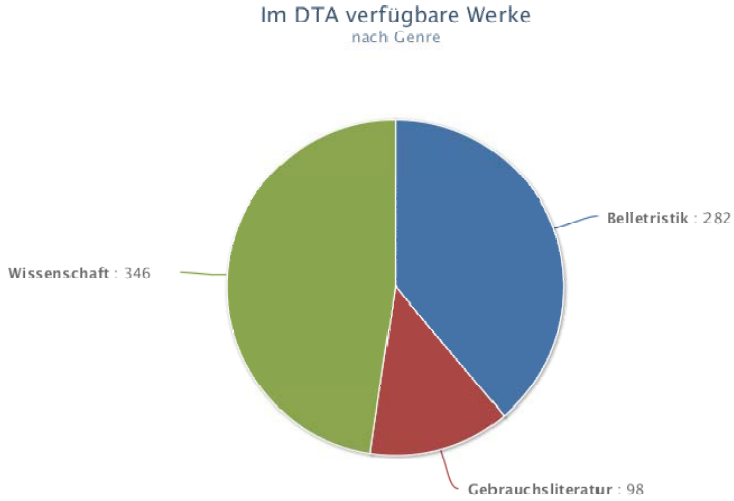


Abb. 1: Textsorteninventar und Verteilung für des DTA (Zeitraum 1780-1900)

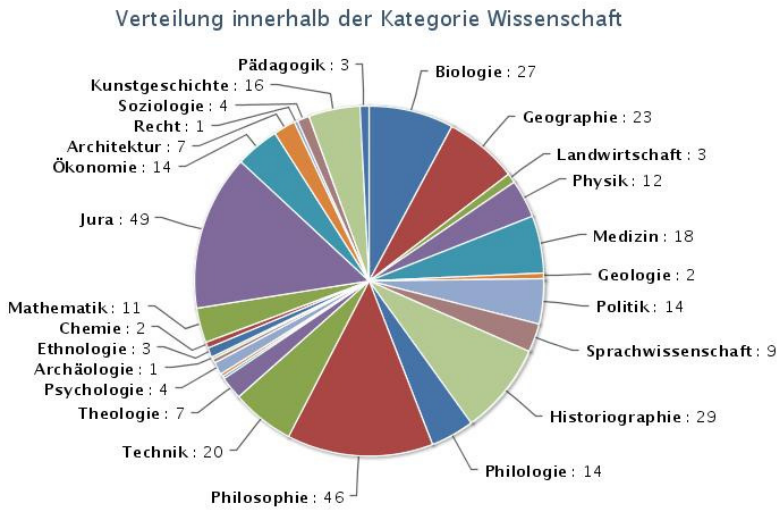


Abb. 2: Verteilung innerhalb der Kategorie Wissenschaft (Zeitraum 1780-1900)

3.2 Einheitlichkeit der Formate

Wie in Abschnitt 1 bereits erwähnt, ist die TEI auf einem guten Wege, ein De-Facto-Standard für die Annotation historischer Texte zu werden. Aufgrund ihrer großen Flexibilität ist die TEI jedoch nicht per se interoperabel (z.B. UNSWORTH 2011). Interoperabilität ist jedoch ein entscheidender Eckpfeiler für eine Korpusinfrastruktur, die auf verteilten Ressourcen basiert. Auf Metadatenebene ermöglicht Interoperabilität, dass Texte einheitlich verarbeitet werden und über große Datenbanken verfügbar gemacht werden können. Auf der Ebene der Text- bzw. Objektdaten erleichtert die einheitliche Annotation die Auswertung der Texte für weitere Analysen, wie z.B. Text Mining oder die tiefere Annotation des Texts mit Diskursinformationen. Darüber hinaus lassen sich einheitlich formatierte Texte auch indizieren, ohne dass dafür Konvertierungen oder manuelle Arbeitsschritte notwendig wären. Für die Abfrage hat dies den Vorteil, dass auch semantische Suchabfragen, sofern sie vorher annotiert wurden, konsistent abfragbar sind. Beispielsweise könnten aus großen Textsammlungen nur dann alle Abschnitte vom Typ „Brief“ mit einer Suchabfrage extrahiert werden, wenn diese vorher auch einheitlich ausgezeichnet wurden. Schließlich können einheitlich kodierte Meta- und Objektdaten in großen Repositorien gespeichert werden und damit nachhaltig vorgehalten werden.

Wie lässt sich die Interoperabilität von Korpusdaten gewährleisten? Die TEI selbst ist sich dessen bewusst, dass die TEI als solche zu unspezifisch für einen interoperablen Einsatz ist und empfiehlt daher die Erstellung von geeigneten Untermengen von TEI-P5 (BURNARD & BAUMANN 2012). Die TEI selbst liefert auch Vorschläge wie TEI Tite (TROLARD 2011), TEI Lite (BURNARD & SPERBERG-MCQUEEN 2006) oder Best Practices for TEI in Libraries (TEI SIG on Libraries 2011). Darüber hinaus gibt es projektspezifische Anpassungen wie beispielsweise TEI Analytics (UNSWORTH 2011; PYTLIK ZILLIG *et al.* 2009), IDS-XCES (Institut für deutsche Sprache, Mannheim) oder Textgrid's Baseline Encoding for Text Data in TEI P5 (Textgrid 2007–2009). Eine Gemeinsamkeit dieser Formate ist, dass sie nur eine gegenüber der gesamten TEI-P5 reduzierte Zahl von Elementen verwenden, um eine bessere Kontrolle über Annotation der Dokumente zu erhalten. Solche Formate, technisch als XML-Schemata beschrieben, sollten genügend ausdrucksstark sein, um eine Basisstrukturierung von Texten zu ermöglichen, die wiederum als Ausgangspunkt für tiefere projektspezifische Annotationen dienen kann.

Das Basisformat des Projekts DTA[7] (fortan DTA-BF) ist ebenso wie die anderen oben genannten Formate eine Untermenge des TEI-P5. Ziel des DTA-BF ist es, für die heterogenen Anforderungen des DTA-Projekts – da Texte aus unterschiedlichen Zeiträumen, Orten und Textsorten aufgenommen werden – eine weitestmöglich eindeutige Kodierungszuordnung für die unstrittigen strukturellen Aspekte bereitzustellen. Insbesondere soll durch das DTA-BF sichergestellt werden, dass semantisch gleiche Phänomene strukturell gleich und damit eindeutig annotiert werden. Ziel des DTA-BF ist somit, die Interoperabilität von Texten auf der Ebene der Metadaten und der Textstrukturierungsebene zu gewährleisten und damit eine übergreifende Einheitlichkeit der Abfragemöglichkeiten sicherzustellen. Ein wichtiger Unterschied gegenüber den oben genannten Formaten besteht vor allem darin, dass das DTA-BF nicht versucht, eine möglichst große Anzahl verschiedener Textkollektionen durch die Bildung der Vereinigungsmenge über all diese Formate zu gewährleisten, wie dies beispielsweise bei TEI-Analytics geschieht. Vielmehr sollen bei der Integration neuer Texte im Vorfeld Redundanzen vermieden werden. Zugleich werden möglichst viele Informationen aus den externen Texten bewahrt, statt einen Teil dieser nach dem Prinzip des ‚kleinsten gemeinsamen Nenners' zu ignorieren. Dies

[7] www.deutschestextarchiv.de/doku/basisformat.

geschieht dadurch, dass neue Elemente auf bereits bestehende Elemente oder Attribut-Wert-Paare abgebildet werden. Somit soll die Anzahl der verschiedenen Elemente auf eine möglichst klein bleiben. Ähnliches gilt für die Attribut-Wert-Paare des DTA-BF. Diese werden durch eine vorgegebene Liste von Werten beschrieben. Bei der Auswertung ist somit gewährleistet, dass es zu keinem „unkontrollierten Wachstum" verschiedener Bezeichnungen für gleiche semantische Inhalte kommt.

Bei dem DTA-BF handelt es sich um ein flexibles Format: das bedeutet, dass Annotationen eines Texts im DTA-BF in unterschiedlicher Tiefe vorgenommen werden können, die durch sogenannte Levels voneinander unterschieden werden. Dies stimmt auch mit den Empfehlungen der TEI überein, die für die Kodierung von Korpora vier Ebenen der Annotation vorschlägt: obligatorische, empfohlene, optionale und verbotene Elemente[8].

3.3 Qualitätssicherung

Um die einheitliche Transkriptions- und Annotationsqualität zu gewährleisten, ist es notwendig, dass die Infrastruktur eine Qualitätssicherungsumgebung enthält, in der die importierten Textquellen evaluiert und gegebenenfalls korrigiert werden können. Die Qualitätskontrolle findet sowohl formativ (im Zuge der Auswahl des geeigneten Exemplars und durch Vorannotation) wie auch summativ statt. Summative Plattformen für das verteilte Korrigieren existieren für die großen Textsammlungen wie Wikisource, Gutenberg-DE[9] oder Gutenberg.org, sie sind jedoch bisher nicht für die Korrektur von TEI-Dokumenten implementiert worden. Für das DTA wurde daher eine solche Umgebung entwickelt: DTAQ[10]. DTAQ ist web-basiert und ermöglicht die verteilte Prüfung (und bislang in eingeschränktem Maße auch die Korrektur) von TEI-Dokumenten. In DTAQ können Strukturierungs- und Erfassungsfehler, aus der Vorlage übernommene Druckfehler sowie im Erfassungsprozess unterlaufene Transkriptionsfehler gemeldet, ggf. kommentiert und nachvollziehbar behoben werden. Für die Textkontrolle haben Nutzer die Möglichkeit, Texte seitenweise in der Gegenüberstellung von Text und Bild anzusehen und Fehler (Transkriptionsfehler, Druckfehler, Strukturierungs- und

[8] www.teic.org/release/doc/tei-p5-doc/en/html/CC.html.

[9] Z.B. www.gaga.net.

[10] www.deutschestextarchiv.de/dtaq/about.

Darstellungsfehler) zu melden. Verschiedene Textansichten sind verfügbar:

- die originale XML/TEI-Fassung,
- eine HTML-Darstellung,
- eine reine Textansicht,
- die Transkription nach automatischer linguistischer Analyse.

DTAQ ist sein Juni 2011 im Einsatz. Seither wurden etwa 17.000 Textseiten vollständig Korrektur gelesen und insgesamt mehr als 30.000 Fehler gemeldet. Derzeit sind knapp 100 Nutzer in DTAQ angemeldet.[11] Spezialanwendungen erlauben die Fokussierung auf bestimmte Phänomene (typischerweise fehlerhafte Zeichenketten, Textmaterial nicht-lateinischer Alphabete etc.) sowie auf wünschenswerte Erweiterungen des Korpus (z.B. die Transkription von Formeln, deren Vorhandensein bisher nur durch ein leeres <formula/>-Element angedeutet wurde, mithilfe eines speziellen Formel-Editors).

3.4 Übergreifende Abfragbarkeit

Für gegenwartssprachliche Texte ist eine übergreifende wortformen- bzw. lemmabasierte Volltextsuche mittlerweile eine Standardanwendung, da für die heutige Orthographie umfassende morphologische Analyseprogramme vorliegen. Damit kann beispielsweise die Suche nach der Wortform „Kleid“ auch Treffer für alle flektierten Formen des morphologischen Paradigmas „Kleids“, „Kleider“, „Kleidern“ etc. zurückliefern.

Da historische Texte, wenn sie gemäß ihrer Originalausgaben transkribiert sind, in keiner standardisierten Rechtschreibung vorliegen (die erste Normierung der deutschen Rechtschreibung fand erst 1902 statt), ist die übergreifende Abfragbarkeit nicht durch Standardprogramme zu leisten. Für das Deutsche gibt es derzeit zwei Verfahren, die beide darauf basieren, historische Formen, die von der heutigen Orthographie abweichen, auf die zugehörige gegenwartssprachliche Form abzubilden (GOTSCHAREK *et al.* 2009, JURISH 2010, JURISH 2011). Mit diesen Verfahren ist es möglich, die orthographisch uneinheitlichen Texte zu durchsuchen, indem die orthographisch gültige Form als Eingabewort verwendet wird. Beispielsweise würde eine Suche nach der heutigen Form „Kleid“ auch die dazugehörigen historischen graphematischen Varianten für „Kleidt“, „Kleidts“ „Kleydt“, „Cleyd“, „Cleit“ etc. finden. Durch diese Analyse-

[11] Stand August 2012.

schritte ist das Korpus somit schreibweisentolerant und orthographieübergreifend durchsuchbar.

3.5 Nachnutzbarkeit der Texte

Die Nachnutzbarkeit von Texten hat eine rechtliche und eine technische Ebene. Auf der rechtlichen Ebene müssen Texte, wenn sie durch Dritte für Forschungszwecke nutzbar sein sollen, mit einer offenen Lizenz versehen sein, damit Analysen oder Zusatzannotationen zu einem Text wieder zusammen mit den Basistexten veröffentlicht werden können. Mittels einer Creative Commons Lizenz ist dies möglich. Texte können damit nicht nur zum Download angeboten werden, sondern auch in andere Repositorien eingebettet und dort für weitere Nutzungsformen zur Verfügung stehen. Im Projekt DTA beispielsweise stehen alle vom DTA produzierten Texte unter einer CC-BY-NC Lizenz. Sie sind somit unter Namensnennung des DTA für nichtkommerzielle Zwecke in beliebigen Forschungskontexten nachnutzbar.

Auf technischer Ebene müssen verschiedene Maßnahmen bezüglich Zugriff, Objektpersistenz und Formatstandardisierung getroffen werden. Zunächst einmal sollten die Dokumente autonom in Repositorien vorgehalten werden und nicht etwa in Datenbanken gekapselt. Des Weiteren sollten Texte besteht mit persistenten Identifizierern (PID) ausgestattet sein, damit sie von Dritten verlässlich referenziert werden können. Die Referenzierungsgenauigkeit sollte dabei mindestens auf Dokumentebene, besser aber auf Seiten- bzw. Zeilenebene erfolgen. Drittens sollten Meta- und Objektdaten gemäß transparenter und weit verbreiteter Standards beschrieben sein. Verwendet man beispielsweise Dublin Core oder klar definierte TEI-Header, können die Metadaten über eine OAI-PMH-Schnittstelle verfügbar gemacht werden und stehen somit für viele Formen der Nachnutzung, insbesondere zur Verzeichnung und Kontextualisierung in den großen Metadatenkatalogen, zur Verfügung. In Abschnitt 3.2 wurde bereits darauf eingegangen, dass Objektdaten, die in einem TEI-Basisformat (Beispielsweise dem DTA-BF) vorliegen, auch interoperabel einsetzbar und somit auch leichter nachnutzbar in anderen Kontexten sind.

3.6 Infrastruktur für die Anreicherung der Textbasis

Die Korpusinfrastruktur muss offen für externe Beiträger gestaltet werden, die im Rahmen ihrer Forschungsarbeiten Texte des späten 16. bis frühen 20. Jahrhunderts bearbeiten und/oder digitalisieren. Dies geschieht zum wechselseitigen Nutzen: Die Beiträger erreichen

durch die Verbreitung ihrer Daten im DTA eine höhere Sichtbarkeit, als wenn sie diese nur privat über die eigene Website oder den Universitätsserver zur Verfügung stellen. Insbesondere bietet das DTA externen Beiträgern folgende Möglichkeiten:

- eine Text-/Bild-Ansicht und eine Leseansicht, jeweils mit einer ausführlichen Präsentation der Metadaten, im DTA sowie – durch Einbettung eines Inlineframes (<iframe>) – auf ihrer eigenen Internetseite,
- einen eigenen Suchindex für die Textkollektion,
- die Korrekturmöglichkeit vor der Veröffentlichung durch die Korrekturplattform DTAQ (s. Abschnitt 3.3), bei der die Texte von einer wissenschaftlichen Nutzergemeinschaft evaluiert werden,
- standardisierte Schnittstellen (OAI-PMH) zum Austausch von Metadaten.

Das DTA kann auf solche Weise das DTA-Kernkorpus fortlaufend durch Primärtexte ergänzt werden, wodurch die Vielfalt und Umfang weiter erhöht wird. Darüber hinaus werden so vergleichende linguistische Untersuchungen zwischen den angelagerten Spezialkorpora und dem DTA-Kernkorpus ermöglicht.

Kandidaten für Ergänzungen zum DTA zeichnen sich durch ihre hohe Wirkungsmächtigkeit aus oder stellen Schlüsseltexte innerhalb wichtiger Diskurse dar. Für die Integration der Texte in das DTA ist die strukturelle Aufbereitung entsprechend dem DTA-Basisformat vonnöten. Das DTA oXygen-Framework unterstützt die Erarbeitung von Texten entsprechend dem DTA-Basisformat. Ziel ist ein hinsichtlich der Transkription und der Annotation homogenes Korpus, dessen Texte uneingeschränkt miteinander interoperabel sind.

Derzeit unterhält das DTA in Kooperation mit 10 Projekten, deren Texte sukzessive über das DTA verfügbar gemacht werden[12] und bemüht sich zudem um die Integration einzelner, qualitativ hochwertiger Textzeugen aus den o.g. kleinen und großen Textsammlungen.

3.7 Aktives Archiv, d.h. lebende elektronische Texte

Ein grundsätzlicher Unterschied zwischen gedruckten Publikationen und elektronische Texten besteht in dem unterschiedlichen Lebenszyklus: gedruckte Publikationen sind statisch, Korrekturen werden allenfalls über eine zweite Auflage kenntlich gemacht. Im Unterschied dazu sind digitale Publikationen einer dynamischen Verände-

[12] www.deutschestextarchiv.de/dtae.

rung unterworfen. Transkriptions- oder Druckfehler können jederzeit vermerkt werden und führen durch die Vergabe einer neuen PID zu einer neuen Version des Dokuments. Wichtiger ist die Tatsache, dass auch die Annotationen des Texts stets verfeinert werden können: beispielsweise können im Laufe der Lektüre oder der Arbeit mit dem elektronischen Text Eigennamen oder Zitate in der Annotationsumgebung von DTAQ markiert und wieder in das Archiv zurückgespielt werden. Notwendig für den Aufbau eines im skizzierten Sinne aktiven Archivs ist somit die Schaffung einer Annotationsumgebung, in der die Annotationen zusammen mit dem Basistext (und dessen Versionen) verwaltet werden können.

4. Zusammenfassung und Ausblick

In diesem Beitrag wurde das Korpus des Deutschen Textarchivs als Basis für ein dynamisch erweiterbares historisches Referenzkorpus vorgestellt. Es wurden sieben Anforderungen für eine Korpus-Infrastruktur benannt, die dazu dienen sollen, in systematischer Weise weitere Texte für die historische Korpusforschung nutzbar zu machen. Dabei wurden rechtliche (OpenAccess) und technische (Standardisierung der Formate) Eckpfeiler benannt.

Notwendige Voraussetzung für den Erfolg ist jedoch die Bereitschaft der Beiträger, sich aktiv an einem solchen Vorhaben zu beteiligen – sei es durch das Bereitstellen eigener Daten oder über die kollaborative Arbeit mit und an bereitgestellten Texten. Dabei darf die vorgeschlagene Infrastruktur nicht als Einbahnstraße erscheinen, sondern muss den Beiträgern glaubhaft vermitteln, dass die in die Infrastruktur hineingegebenen Texte genauso wie alle anderen Texte der Infrastruktur wieder nachgenutzt und in andere Analyse- und Präsentationsumgebungen eingespeist werden können. Mit seiner offenen Lizenzpolitik (CC-BY-NC), die jedem Beiträger genau diese Nachnutzungen ermöglicht, leistet das DTA hierfür einen Beitrag. Ein Projekt alleine ist zu klein, um die verschiedenen potentiellen Beiträger tatsächlich beraten zu können. Die Schaffung einer örtlichen und über verschiedene Institutionen verteilten Arbeitsgruppe ist daher ein nächstes Ziel. Diese soll dazu beitragen, die Sogwirkung zu entfalten, die dafür notwendig ist, dass mehr Beiträger für das Referenzkorpus gewonnen werden können. Eine solche Arbeitsgruppe, bestehend aus dem DTA, der HAB Wolfenbüttel, dem IDS Mannheim und der Universität Gießen wurde im Rahmen des CLARIN-D Projekts im Juni 2012 von der Facharbeitsgruppe Deutsche Philologie gegründet (CLARIN-D Kurationsprojekt).

5. BIBLIOGRAPHIE

BURNARD, L. & S. BAUMAN, 2012: *P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.1.0, June 17th, 2012*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html.

BURNARD, L. & SPERBERG-MCQUEEN, C. M., 2006: *TEI Lite: Encoding for Interchange: an introduction to the TEI – Revised for TEI P5 release*. http://www.tei-c.org/Vault/P5/2.1.0/doc/tei-p5-exemplars/html/teilite.doc.html.

CLARIN-D Kurationsprojekt, 2012: »Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur«, Vorhabensbeschreibung für ein Kurationsprojekt der F-AG 1 »Deutsche Philologie«. http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.

GEYKEN, A., 2007: The DWDS corpus: A reference corpus for the German language of the 20th century, in: FELLBAUM, CHR. (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, London, 23-41.

GEYKEN, A. *et al.*, 2012: The DTA 'base format': A TEI-subset for the compilation of Interoperable Corpora. To appear in: *Proceedings of Konvens 2012*.

GEYKEN, A. *et al.*, 2012: TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv, in: *Jahrbuch für Computerphilologie*, online-Version vom 05.08.2012. http://www.computerphilologie.de/jg09/geykenetal.html.

GEYKEN, A. *et al.* (Panel): Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints. Panel DH2012, Hamburg (Abstract und Video Lecture). http://www.dh2012.uni-hamburg.de.

GOTSCHAREK, A. *et al.*, 2009: Enabling information retrieval on historical document collections: the role of matching procedures and special lexica, in: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, New York, 69–76.

HAAF, S. *et al.*, 2012: Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-

annotated historical text, in: *Journal of the Text Encoding Initiative* 4 (Forthcoming Paper).

JURISH, B., 2010: More than Words: Using Token Context to Improve Canonicalization of Historical German, in: *Journal for Language Technology and Computational Linguistics (JLCL)*, vol. 25/1, 23–39.

JURISH, B., 2011: *Finite-state canonicalization techniques for historical German* (URN: urn:nbn:de:kobv:517-opus-55789) (URL: http://opus.kobv.de/ubp/volltexte/2012/5578/).

KUPIETZ, M. *et al.*, 2010: The German Reference Corpus DeReKo: A primordial sample for linguistic research, in: CALZOLARI, N. *et al.* (eds.), *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*, 1848-1854.

LEMNITZER, L. & H. ZINSMEISTER, 2010: *Korpuslinguistik. Eine Einführung*, 2. Aufl., Tübingen.

PYTLIK ZILLIG, B. L., 2009: TEI Analytics: converting documents into a TEI format for cross-collection text analysis, in: *Literary and Linguistic Computing*, 24 (2), 187-192.

SINCLAIR, J., 2005: Corpus and Text – Basic Principles, in: WYNNE, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, 1-16.

TROLARD, P., 2011: *TEI Tite – A recommendation for off-site text encoding. Version 1.1, September 2011.* www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

UNSWORTH, J., 2011: Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI, in: *Journal of the Text Encoding Initiative* 1. http://jtei.revues.org/215 (accessed June 24th, 2012).

CANONICALIZING THE DEUTSCHES TEXTARCHIV

BRYAN JURISH

1. Introduction

Virtually all conventional text-based natural language processing techniques – from traditional information retrieval systems to full-fledged parsers – require reference to a fixed lexicon accessed by surface form, typically trained from or constructed for synchronic input text adhering strictly to contemporary orthographic conventions. Unconventional input such as historical text which violates these conventions therefore presents difficulties for any such system due to lexical variants present in the input but missing from the application lexicon.

Traditional approaches to the problems arising from an attempt to incorporate historical text into such a system rely on the use of additional specialized (often application-specific) lexical resources to explicitly encode known historical variants. Such specialized lexica are not only costly and time-consuming to create, but also – in their simplest form of static finite word lists – necessarily incomplete in the case of a morphologically productive language like German, since a simple finite lexicon cannot account for highly productive morphological processes such as nominal composition [KEMPKEN *et al.* (2006)].

To facilitate the extension of synchronically-oriented natural language processing techniques to historical text while minimizing the need for specialized lexical resources, one may first attempt an automatic canonicalization of the input text. Canonicalization approaches treat orthographic variation phenomena in historical text as instances of an error-correction problem, seeking to map each (unknown) word of the input text to one or more extant *canonical cognates*: synchronically active types which preserve both the root and morphosyntactic features of the associated historical form(s). To the extent that the canonicalization was successful, application-specific processing can then proceed normally using the returned canonical forms as input, without any need for additional modifications to the application lexicon.

This paper provides an informal overview of the various canonicalization techniques currently employed by the *Deutsches*

Textarchiv[1] (DTA; [GEYKEN & KLEIN (2010)]) project at the Berlin-Brandenburg Academy of Sciences and Humanities to prepare a corpus of historical German text for part-of-speech tagging, lemmatization, and integration into a robust online information retrieval system. For more details on the methods employed, the interested reader is referred to [JURISH (2012)].

2. Canonicalization Techniques

It is useful to distinguish between *type-wise* and *token-wise* canonicalization techniques. Type-wise canonicalization techniques are those which process each input word in isolation, independently of its surrounding context, and are fully specified by a binary *conflation relation*[2] over surface strings. Token-wise canonicalization techniques on the other hand make use of the context in which a given instance of a word occurs when determining the optimal canonical cognate, and can thus better account for ambiguities in the mapping from historical to contemporary forms, insofar as these can be resolved by reference to the immediate context. In the sequel, $w \sim_r v$ indicates that the words (types or tokens) w and v are related by the conflator r, and $[w]_r$ denotes the set of all words conflated with w by the conflator r. If r returns a unique (canonical) value v for each input word w, the standard notation for functions $r(w) = v$ will be used.

2.1 String Identity

String identity (id) is the easiest conflator to implement (no additional programming effort or resources are required) and provides a high degree of precision, "false friends" being limited to historical homographs such as the historical form *wider* when it occurs as a variant of the contemporary form *wieder* ("again") rather than the lexically distinct contemporary homograph *wider* ("against"). Since its coverage is restricted to valid contemporary forms, string identity cannot account for any spelling variation at all, resulting in very poor recall – many relevant types will not be retrieved in response to a query in current orthography.

As an example, consider the historical form *Abſtände*, a variant of the contemporary cognate *Abstände* ("distances"). The conflation set $[\textit{Abſtände}]_{\text{id}} = \{\textit{Abſtände}\}$ does not contain the desired contemporary cognate, so no instances of the historical variant *Abſtände* will be

[1] "German Text Archive", http://www.deutschestextarchiv.de.

[2] Prototypically, every conflation relation will be a true equivalence relation.

retrieved via string identity for a query of the contemporary form *Abstände*. In the DTA canonicalization architecture, string identity is used only as a fallback conflator. Each input word is treated as its own canonical form if all other canonicalizations methods have failed, or if it passes some simple heuristic tests for detecting "uncanonicalizable" strings such as punctuation, abbreviations, mathematical formulæ, or foreign language material in non-latin script.

2.2 Transliteration

A slightly less naïve family of conflation methods are those which employ a simple deterministic transliteration function to replace input characters which do not occur in contemporary orthography with extant equivalents. A transliteration conflator is defined in terms of a character transliteration function xlit which maps each possible input character to a (possibly empty) output string over the contemporary alphabet, the concatenation of which yields the candidate canonical form for the input word.

In the case of historical German, deterministic transliteration is especially useful for its ability to account for typographical phenomena, e.g. by mapping 'ſ' (long 's', as commonly appeared in texts typeset in fraktur) to a conventional round 's', and mapping superscript 'e' to the conventional *Umlaut* diacritic '¨', as in the transliteration xlit(*Abſtände*) = *Abstände* ("distances"). Given this transliteration, a query for the contemporary form Abstände will successfully retrieve all instances of the historical form *Abſtände*.

The DTA canonicalization cascade uses a fast conservative transliteration function based on the `Text::Unidecode` Perl module.[3] Despite its efficiency, and although it outdoes even string identity in terms of its precision, deterministic transliteration suffers from its inability to account for spelling variation phenomena involving extant characters such as the *th/t* and *ey/ei* allographs common in historical German. As an example, consider an instance of the historical form *Theyl* corresponding to the contemporary cognate *Teil* ("part"). Both historical and contemporary forms will be transliterated to themselves, since both strings contain only extant characters, but the historical form will not be retrieved by a query for the contemporary form, since their transliterations are distinct: xlit(*Teil*) = *Teil* ≠ *Theyl* = xlit(*Theyl*).

[3] http://search.cpan.org/~sburke/Text-Unidecode-0.04/.

2.3 Phonetization

A more powerful family of conflation methods is based on the dual intuitions that graphemic forms in historical text were constructed to reflect phonetic forms[4] and that the phonetic system of the target language is diachronically more stable than its graphematic system. A phonetic conflator maps each input word w to a unique phonetic form pho(*w*) by means of a computable function pho, conflating those strings which share a common phonetic form. The phonetic conversion module used in the DTA was adapted from the phonetization rule-set distributed with the IMS German Festival package [MÖHLER *et al.* (2001)], a German language module for the Festival text-to-speech system [BLACK & TAYLOR (1997)], and compiled as a finite-state transducer.[5]

Phonetic conflation offers a substantial improvement in recall over conservative methods such as transliteration or string identity: variation phenomena such as the *th/t* and *ey/ei* allographs mentioned above are correctly captured by the phonetization transducer: pho(*Theyl*) = [taɪl] = pho(*Teil*), which implies that all instances of the historical form *Theyl* will be retrieved in response to a query of the contemporary form *Teil*. Unfortunately, these improvements come at the expense of precision: in particular, many high-frequency types are misconflated by the simplified phonetization rule-set, including **in* ∼ *ihn* ("in" ∼ "him") and **wider* ∼ *wieder* ("against" ∼ "again"). While such high-frequency cases can easily be dealt with by a small exception lexicon (cf. section 2.6), the underlying tendency of strict phonetic conflation either to over- or to under-generalize – depending on the granularity of the phonetization function – is likely to remain, expressing itself in information retrieval tasks as reduced precision or reduced recall, respectively.

2.4 Rewrite Transduction

Despite its comparatively high recall, the phonetic conflator fails to relate unknown historical forms with any extant equivalent whenever the graphemic variation leads to non-identity of the respective phonetic forms (e.g. pho(*umb*) = [ʔʊmp] ≠ [ʔʊm] = pho(*um*) for the historical variant *umb* of the preposition *um* ("around")),

[4] [KELLER (1978)] codifies this intuition as the imperative "write as you speak" governing historical spelling conventions.

[5] In the absence of a language-specific phonetization function, a general-purpose phonetic digest algorithm such as SOUNDEX [RUSSELL (1918)] may be employed instead [ROBERTSON & WILLETT (1993)].

suggesting that recall might be further improved by relaxing the strict identity criterion implicit in the definition of the phonetic conflator. A conflation technique which fulfills both of the above desiderata is *rewrite transduction*,[6] which can be understood as a generalization of the well-known *string edit distance* [DAMERAU (1964), LEVENSHTEIN (1966)].

A rewrite conflator (rw) is defined in terms of a *target lexicon* of contemporary forms and a weighted *error model* [KERNIGHAN *et al.* (1990), BRILL & MOORE (2000)] which associates each known pattern of diachronic variation with a non-negative weight or "distance". The conflation set $[w]_{\mathrm{rw},k}$ is computed as the set of k nearest neighbors of the input word w which are themselves members of the target lexicon. Importantly, such a rewrite conflation set can be computed even in the presence of an infinite target lexicon,[7] provided that both lexicon and error model can be represented as (weighted) finite-state transducers [MOHRI (2002)].

The DTA canonicalization architecture uses a finite-state rewrite cascade whose error model was compiled from a set of manually constructed rules and whose target lexicon was extracted from the the high-coverage TAGH morphology system for contemporary German [GEYKEN & HANNEFORTH (2006)] to compute rewrite conflation sets containing at most only a single "best" contemporary form ($k=1$). Although this rewrite cascade does indeed improve both precision and recall with respect to the phonetic conflator, these improvements are of comparatively small magnitude, precision in particular remaining well below the level of conservative conflators such as naïve string identity or transliteration, due largely to interference from "false friends" such as the valid contemporary compound *Rockermehl* ("rocker-flour") for the historical variant *Rockermel* of the contemporary form *Rockärmel* ("coat-sleeve").

2.5 Hidden Markov Model Disambiguation

Systematic evaluations of the type-wise techniques described above revealed a typical precision-recall trade-off pattern: the ultra-conservative string identity conflator – despite its near-perfect precision – shows quite poor recall, while the more ambitious high-recall

---

[6] Related approaches to historical variant detection include [RAYSON *et al.* (2005), ERNST-GERLACH & FUHR (2006), GOTSCHAREK *et al.* (2009)].

[7] E.g. as arising from morphologically productive phenomena such as German nominal composition.

conflators such as phonetic identity or rewrite transduction tend to be disappointingly imprecise. In order to recover some of the precision offered by conservative conflation techniques such as transliteration while still benefiting from the flexibility and improved recall provided by more ambitious techniques such as phonetization or rewrite transduction, the DTA canonicalization architecture makes use of a Hidden Markov Model (HMM) disambiguator which operates on the token level, using sentential context to determine a unique "best" canonical form for each input token, in a manner similar to the spelling correction technique described by [MAYS *et al.* (1991)].

Treating the conflation sets returned by all active type-wise conflators as candidate canonicalization hypotheses, the HMM disambiguator chooses an optimal sequence of token-wise unique canonical forms for each input sentence by application of the well-known Viterbi algorithm [VITERBI (1967)]. Lexical probabilities are dynamically computed as a Maxwell-Boltzmann distribution over the candidate conflations for each input word, and a static trigram model of contemporary German is used to model local syntactic and semantic context constraints. The disambiguator is thus able to resolve ambiguous conflation sets such as {*in, ihn*} or {*wider, wieder*} in a context-dependent manner.

Using a simple smoothing mechanism, the disambiguator is also able to override the decisions of the type-wise conflators by selecting a canonical form not explicitly enumerated in the target lexicon.[8] This behavior is particularly useful for proper names, which are not exhaustively represented in the TAGH morphology system, and which were excluded entirely from the rewrite target lexicon because their presence lead to too many spurious conflations with valid historical forms, e.g. the TAGH lexical entry for the surname *Aehnlich* caused the rewrite conflator to treat all instances of the type as their own canonical forms rather than mapping them to the correct contemporary form *ähnlich* ("similar").

2.6 Exception Lexicon

The HMM disambiguator performs very well at the token level, but its reliance on a static *n*-gram model over contemporary forms is

[8] Technically, the possibility of selecting the input word itself as its own canonical form is implemented by allowing the identity conflator id to provide a candidate conflation hypothesis. In practice, the DTA canonicalization architecture uses the transliteration conflator xlit whenever it returns a non-empty string, and only resorts to a pure identity hypothesis when the transliterator fails.

problematic for input words whose canonical cognate was not present in the training data: in such cases, the model effectively reverts to a type-level canonicalization, choosing the most likely conflation candidate based only on criteria of word length and source conflator. Due to the conflator-dependent distance functions employed, short input words in particular are likely to be subjected to such treatment, which was designed primarily to handle low-frequency unlexicalized types such as proper names, and thus often results in a fallback identity canonicalization. Many common historical variants of high-frequency words fall into this category, usually due to (irregular) patterns of variation not captured by the type-wise conflators such as exhibited by the (strongly inflected) historical variant *frug* of the (weakly inflected) contemporary form *fragte* ("asked"). On the other hand, "false friends" in the training data can cause also spurious canonicalizations: even a single occurrence of the given name *André* in the training data is sufficient to cause the historical variant *andre* of the contemporary form *andere* ("other") to be miscanonicalized.

To handle problematic cases such as these, the DTA canonicalization architecture incorporates a semi-automatically generated exception lexicon [JURISH *et al.* (forthcoming)] which operates on incoming word types before they are passed to the disambiguator. If the exception lexicon contains an entry for an input type, only that entry is considered by the disambiguator as a candidate canonicalization for the input word. This technique ensures that the exception lexicon entry will in fact be the canonical form chosen on the one hand, and allows the disambiguator to make use of the provided entry for context-dependent resolution of nearby items on the other.

3. Summary

Historical text presents unique challenges for typical synchronically oriented natural language processing tasks. In particular, violations of contemporary orthographic conventions are problematic for any task requiring reference to a fixed lexicon keyed by surface word type. Part-of-speech tagging, lemmatization, and information retrieval (corpus indexing & query) are all affected. Canonicalization approaches address this problem by attempting to map unknown historical variants to extant contemporary forms and deferring synchronically oriented analysis to the returned (canonical) forms.

The canonicalization techniques currently used to preprocess the *Deutsches Textarchiv* corpus of historical German were briefly described. String identity on its own does not provide an adequate solution, since it cannot account for any orthographic variation at all, but it can be useful in conjunction with additional heuristics for detecting non-lexical material. Transliteration provides an efficient and very precise canonicalization method for dealing with extinct characters such as the long 's' common in historical German, but cannot account for any variation involving extant characters. More ambitious techniques such as conflation by phonetic identity or rule-based rewrite transduction are able to account for a much wider range of variation, but these improvements come at the cost of precision. Use of a Hidden Markov Model to disambiguate canonicalization hypotheses at the token level using sentential context effectively recovers much of this lost precision while still benefitting from the improved recall. Remaining systematic canonicalization errors are accounted for by a type-wise exception lexicon. The fully canonicalized corpus was subsequently tagged and lemmatized before being indexed by a robust information retrieval system which uses the canonical-lemma token-level conflation relation to implement an intuitive linguistically motivated search term expansion operator for non-expert user queries.

REFERENCES

BLACK, A. W. & P. TAYLOR, 1997: *Festival speech synthesis system*, Technical Report HCRC/TR-83, University of Edinburgh, Centre for Speech Technology Research.
URL: http://www.cstr.ed.ac.uk/projects/festival.

BRILL, E. & R. C. MOORE, 2000: An improved error model for noisy channel spelling correction, in: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 286-293.

DAMERAU, F. J., 1964: A technique for computer detection and correction of spelling errors, in: *Communications of the ACM* 7, 171–176. DOI: 10.1145/363958.363994.

ERNST-GERLACH, A. & N. FUHR, 2006: Generating search term variants for text collections with historic spellings, in: LALMAS, M. *et al.* (eds.), *Advances in Information Retrieval*, Lecture Notes in Computer Science 3936, Berlin, 49–60.
DOI: 10.1007/11735106_6.

GEYKEN, A. & T. HANNEFORTH, 2006: TAGH: A complete morphology for German based on weighted finite state automata, in: *Finite State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Revised Papers*, Lecture Notes in Computer Science 4002, Berlin, 55–66.
DOI: 10.1007/11780885_7.

GEYKEN, A. & W. KLEIN, 2010: Deutsches Textarchiv, in: *Jahrbuch 2009*, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 320–323. URL:
http://edoc.bbaw.de/volltexte/2010/1515/pdf/BBAW_Jahrbuch_2009.pdf.

GOTSCHAREK, A. *et al.*, 2009: Enabling information retrieval on historical document collections: the role of matching procedures and special lexica, in: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, AND '09*, New York, 69–76. DOI: 1568296.1568309.

JURISH, B., 2012: Finite-State Canonicalization Techniques for Historical German, PhD thesis, Universität Potsdam.
URL: http://opus.kobv.de/ubp/volltexte/2012/5578/.

JURISH, B. *et al.*, forthcoming: Constructing a canonicalized corpus of historical German by text alignment, in: BENNETT, P. *et al.* (eds.),

New Methods in Historical Corpus Linguistics, vol. 3 of Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP), Tübingen.

KELLER, R. E., 1978: *The German Language*, London.

KEMPKEN, S. *et al.*, 2006: Comparison of distance measures for historical spelling variants, in: BRAMER, M. (ed.), *Artificial Intelligence in Theory and Practice*, Boston, 295–304.
DOI: 10.1007/978-0-387-34747-9_31.

KERNIGHAN, M. D. *et al.*, 1990: A spelling correction program based on a noisy channel model, in: *Proceedings COLING-1990*, vol. 2, 205–210.

LEVENSHTEIN, V. I., 1966: Binary codes capable of correcting deletions, insertions, and reversals, in: *Soviet Physics Doklady* 10, 707–710.

MAYS, E. *et al.*, 1991: Context based spelling correction, in: *Information Processing & Management* 27, 517–522.
DOI: 10.1016/0306-4573(91)90066-U.

MÖHLER, G. *et al.*, 2001: *IMS German Festival manual, version 1.2*. Institute for Natural Language Processing, University of Stuttgart.
URL: http://www.ims.uni-stuttgart.de/phonetik/synthesis.

MOHRI, M., 2002: Semiring frameworks and algorithms for shortest-distance problems, in: *Journal of Automata, Languages and Combinatorics* 7, 321–350.

RAYSON, P. *et al.*, 2005: VARD versus Word: A comparsion of the UCREL variant detector and modern spell checkers on English historical corpora, in: *Proceedings of the Corpus Linguistics 2005 conference*, Birmingham, UK, July 14-17 2005.

ROBERTSON, A. M. & P. WILLETT, 1993: A comparison of spelling-correction methods for the identification of word forms in historical text databases, in: *Literary and Linguistic Computing* 8, 143–152.

RUSSELL, R. C., 1918: Soundex coding system, in: *United States Patent* 1,261,167.

VITERBI, A. J., 1967: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, in: *IEEE Transactions on Information Theory* 13, 260–269.

AUF DEM WEG ZU EINEM INTEGRIERTEN LEXIKON DES ÄGYPTISCH-KOPTISCHEN

FRANK FEDER & SIMON D. SCHWEITZER

Einleitung

Die ägyptisch-koptische Sprache ist seit dem späten 4. Jahrtausend v. Chr. bezeugt und gilt damit als die am längsten schriftlich belegte Sprache der Welt. Sie gehört zur afroasiatischen Sprachfamilie und wird üblicherweise in die fünf Sprachstufen Altägyptisch, Mittelägyptisch, Neuägyptisch, Demotisch und Koptisch gegliedert.[1] Jedoch erscheinen diese Sprachstufen im überlieferten Material nicht in einer direkten Folge. So finden sich beispielsweise in Tempeltexten der griechisch-römischen Zeit Textpassagen, die eher der mittelägyptischen Sprachstufe zuzurechnen sind.[2] Die ägyptisch-koptische Sprache wird in drei verschiedenen Schriften überliefert. Zunächst sind die Hieroglyphen mit ihrer kursiven Variante, dem Hieratischen, zu nennen. Als Weiterentwicklung der kursiven Schrift im ersten Jahrtausend ist das Demotische anzusehen. Schließlich fand in christlicher Zeit das so genannte koptische Alphabet Verwendung, das auf dem griechischen mit einigen Zusatzbuchstaben basiert. Das Verhältnis zwischen Sprachstufe und verwendeter Schrift ist kein triviales. In Hieroglyphen geschriebene Texte können der altägyptischen, der mittelägyptischen, der neuägyptischen und der demotischen Sprachstufe zugewiesen werden.[3] In demotischer Schrift finden sich neben dem überwiegenden Teil der Texte demotischer Sprachstufe auch Texte, die zu älteren Sprachstufen zu zählen sind.[4] Vereinzelt bieten auch Texte in koptischer Schrift älteres Sprachgut.[5]

---

[1] Neuerdings wird die Sprache der Frühdynastischen Periode vom Altägyptischen getrennt und als Frühägyptisch bezeichnet, vgl. KAHL, J., *Frühägyptisches Wörterbuch*, Wiesbaden 2002-.

[2] Vgl. QUACK, J. F., Von der Vielfalt der ägyptischen Sprache in der griechisch-römischen Zeit, in: *ZÄS* 140 (im Druck). Ob nun diese Textpassagen auf ältere Vorlagen zurückgehen oder als Neuschöpfungen aus der Zeit ihrer Niederschrift anzusetzen sind, spielt in diesem Rahmen keine Rolle.

[3] Zu hieroglyphischen Texten in demotischer Sprache vgl. QUACK, J. F., Monumental-Demotisch, in: GESTERMANN, L. & H. STERNBERG-EL HOTABI (Hrsg.), *Per aspera ad astra. Wolfgang Schenkel zum neunundfünfzigsten Geburtstag*, Kassel 1995, 107-121.

[4] QUACK, J. F., Inhomogenität von ägyptischer Sprache und Schrift in Texten aus dem späten Ägypten, in: LEMBKE, K. *et al.* (Hrsg.), *Tradition and Transformation:*

Diese vor allem im 1. Jahrtausend auftretende vielschichtige Situation wird zumeist dem Phänomen einer „literarischen Diglossie“[6] zugerechnet, deren Beginn Pascal Vernus in der Ramessidenzeit zeitlich verortet hat.[7] Nach diesem Modell ist ein erheblicher Teil der erhaltenen Texte des ersten Jahrtausends und bis in die römische Kaiserzeit (zumeist in Hieroglyphen und Hieratisch geschrieben) in einer „Gelehrtensprache“[8] verfasst, deren Kenntnis auf die höher gebildeten unter den Schriftkundigen, am Ende gar auf wenige Spezialisten unter der gelehrten Tempelpriesterschaft beschränkt war. Ganz Prestige und Privileg der Oberschicht Ägyptens war diese auch einem schriftkundigen Ägypter, der nur die gesprochene Sprache seiner Zeit und ihre jeweilige Schriftform beherrschte, kaum mehr und im Laufe der Zeit gewiss gar nicht mehr verständlich. Natürlich war diese ‚Sprache‘ für die Ägypter keine Gelehrtensprache, sondern die klassische und heilige Sprache der als Vorbild empfundenen klassischen Epochen der ‚alten Zeit‘, des Alten und Mittleren Reiches. Sie war den nicht-alltäglichen Textformen der religiösen Literatur für das Diesseits und das Jenseits (Tempel und Grab) und der Prestige und Gedächtnis stiftenden Selbstdarstellung der Herrscher und Elite des Landes in Tempel und Grab sowie der (schriftlichen) Heiligung politisch-ideologischer Institutionen (z.B. Dekrete) der jeweiligen Regentschaft vorbehalten. Wir haben es also in der textlichen Überlieferung aus Ägypten im ersten Jahrtausend v. Chr. mit einer parallelen Textkultur zu tun, die uns Texte in der jeweils aktuellen Sprachstufe der jüngeren Phase der Entwicklung des Ägyptischen (Neuägyptisch, Demotisch und Koptisch) und Texte in einer nicht gesprochenen und artifiziell den Texten der älteren Phase des Ägyptischen (Alt- und Mittelägyptisch) nachempfundenen Sprache erhalten hat. Diese ‚Kunstsprache‘ entspricht nun weder wirklich der alten Sprache noch ist sie frei von Einsprengseln der jüngeren Sprachstufen und stellt daher ein vielschichtiges, oft zeitlich und örtlich beeinflusstes Phänomen dar.[9] Aufgrund dieser Viel-

Egypt under Roman Rule. Proceedings of the International Conference, Hildesheim, Roemer- and Pelizaeus-Museum, 3–6 July 2008, Leiden & Boston 2010, 313-341.

5 OSING, J., *Der spätägyptische Papyrus BM 10808*, ÄA 33, Wiesbaden 1976.

6 JANSEN-WINKELN, K., Diglossie und Zweisprachigkeit im Alten Ägypten, in: *WZKM* 85, 1995, 85-115.

7 VERNUS, P., Langue Litteraire et Diglossie, in: LOPRIENO, A. (ed.), *Ancient Egyptian Literature: History and Forms*, PÄ 10, Leiden [u.a.] 1996, 555-564, hier 563.

8 Etwa der Rolle des Lateinischen im Mittelalter und in der Neuzeit vergleichbar.

9 Vgl. z.B. QUACK, Inhomogenität.

schichtigkeit hat man sich auch bisher schwer getan, einen adäquaten und allgemein akzeptieren Terminus dafür zu finden. Pascal Vernus prägte den Begriff „Égyptien de tradition“.[10] Der Begriff „Spätmittelägyptisch“[11] oder „Late Middle Egyptian“[12] hat sich ebenso eingebürgert, fand aber auch berechtigte Kritik.[13] Friedrich Junge versuchte zugleich, einen adäquaten Begriff zu finden und ihn in ein Modell der Sprachstufen und Sprachgeschichte einzubinden.[14] In diesem Modell steht der Begriff „Spätmittelägyptisch“ für die letzte Entwicklungsstufe des Mittelägyptischen vom Mittleren Reich bis in die 19. Dynastie (Spätmittelägyptisch I-II), also im direkten Anschluss an die mittelägyptische Sprachstufe. Für das „Égyptien de tradition“ prägte er den Begriff „Neo-Mittelägyptisch“. In Anlehnung an Vernus' „Égyptien de tradition“ wird auch im Deutschen „traditionelles Mittelägyptisch“ gebraucht. Joachim Quack brachte vor kurzem mit gewisser Zurückhaltung auch „klassisches Ägyptisch“[15] ins Spiel.

In dieser Situation einer vielschichtigen Textüberlieferung mit ineinander greifenden Sprachstufen hat sich kein allgemeiner Zugang zum Wortschatz des Ägyptisch-Koptischen etabliert. Vielmehr entstanden in der lexikographischen Tradition des Ägyptisch-Koptischen für die Teildisziplinen eigenständige Wörterbücher. Diese Tradition orientiert sich aber vorrangig nicht an den Sprachstufen, sondern an der verwendeten Schrift. So behandelt das Berliner Wörterbuch[16] in seinen Einträgen den außerhalb der demotischen und koptischen Schrift überlieferten Wortschatz. Auch für die übrigen Schriften haben sich eigene Wörterbücher und damit auch eigene lexiko-

---

[10] VERNUS, P., Thème I: Diachronie et Synchronie dans la Langue Égyptienne, in: *L'Égyptologie en 1979 – Axes prioritaire de recherches*, vol. I, Paris 1982, 17-18 und 81-84.

[11] JANSEN-WINKELN, K., *Spätmittelägyptische Grammatik der Texte der 3. Zwischenzeit*, ÄUAT 34, Wiesbaden 1996.

[12] LOPRIENO, A., *Ancient Egyptian. A linguistic introduction*, Cambridge 1995, 6.

[13] WINAND, J., Rezension zu: JANSEN-WINKELN, K., *Text und Sprache der 3. Zwischenzeit. Vorarbeiten zu einer Spätmittelägyptischen Grammatik*, ÄUAT 26, Wiesbaden 1994, in: *LingAeg* 5, 1999, 224-225.

[14] JUNGE, F., Sprachstufen und Sprachgeschichte, in: *ZDMG Supplement VI*, 1985, 17-34; vgl. auch JUNGE, F., Sprache, in: *LÄ* V, 1176-1211.

[15] QUACK, Inhomogenität, 313-314.

[16] ERMAN, A. & H. GRAPOW, *Wörterbuch der ägyptischen Sprache*, 7 Bde., Leipzig & Berlin 1926-1963.

graphische Traditionen in den jeweiligen Teildisziplinen gebildet.[17] Diese Differenzierung des Ägyptisch-Koptischen anhand der verwendeten Schrift ist auch im Thesaurus Linguae Aegyptiae (TLA) zu bemerken, da das Ägyptische und das Demotische eigene Textdatenbanken aufweisen, die nicht gemeinsam durchsuchbar sind bzw. die es nicht erlauben, dass eine Ergebnisliste auf die beiden Textdatenbanken gleichzeitig zurückgreift. Wörterbücher, die ein Teilsegment des Ägyptisch-Koptischen behandeln, führen natürlich zu einer Stärkung der jeweiligen Teildisziplin, die zu einer gewissen Eigenständigkeit führen kann, die aber auch die Ausblendung gerade dieser Teildisziplin in den anderen Teildisziplinen nach sich ziehen kann. Dies betrifft nicht nur die Lexikographie, sondern auch andere ägyptologische Bereiche. So schreibt Quack zur demotischen Literatur: „Demotische Literatur ist in der Ägyptologie insgesamt unzureichend bekannt."[18] Vittmann äußert sich folgendermaßen zur Bedeutung der Demotistik innerhalb der Ägyptologie: „Daß vieles nicht schon längst veröffentlicht wurde, liegt wohl nicht zuletzt daran, daß die Zahl der Demotisten früher viel geringer war als heutzutage und zudem die demotischen Studien ohnehin schon immer eher ein Schattendasein am Rande der „klassischen" Ägyptologie fristeten."[19] Diese Segmentierung der lexikographischen Forschung zum Ägyptisch-Koptischen erschwert somit den Versuch, den ägyptisch-koptischen Wortschatz in Gänze zu betrachten. Neben diesem globalen Problem führt die Trennung des Wortschatzes zu handfesten Schwierigkeiten. So führt die Orientierung anhand der verwendeten Schrift zu einer Trennung eines Dokumentes in mehrere Wörterbücher, wenn mehrere Schriften verwendet werden. So müsste man das von Osing herausgegebene onomasiologische Wörterbuch aus Tebtunis[20] teilweise im ägyptischen, im demotischen und im koptischen Lexikon wiederfinden, da über dem hieratischen Text demotische und altkoptische Glossen stehen. Wenn man bei diesem Dokument bleibt und sich über den Wortschatz des Ägyptisch-Koptischen dieser Zeit, nämlich des

---

[17] ERICHSEN, W., *Demotisches Glossar*, Kopenhagen 1954; CRUM, W. E., *A Coptic Dictionary*, Oxford 1939.

[18] QUACK, J. F., *Einführung in die altägyptische Literaturgeschichte III. Die demotische und gräko-ägyptische Literatur*, Einführungen und Quellentexte zur Ägyptologie 3, Münster [2]2009, IX.

[19] VITTMANN, G., Tradition und Neuerung in der demotischen Literatur, in: *ZÄS* 125, 1998, 63.

[20] OSING, J., *The Carlsberg Papyri 2. Hieratische Papyri aus Tebtunis I*, CNI Publications 17, Kopenhagen 1998.

2. nachchristlichen Jahrhunderts informieren möchte, wäre man gezwungen in drei Wörterbuchtraditionen nachzuschlagen, da zu dieser Zeit hieratische, demotische und (alt-)koptische Texte nebeneinander vorkommen. Durch die Trennung in eigenständige Textdatenbanken kann man ebenso wenig in der digitalen Welt einen synchronen Schnitt für diese Zeit herstellen. Ferner trennt die Aufteilung des Wortschatzes in verschiedene Wörterbücher die innerägyptischen Texttraditionen. So finden sich einzelne Sprüche der ursprünglich altägyptisch verfassten Pyramidentexte auch in demotisch geschriebenen Texten.[21] Die dort anzutreffende Lexik müsste in einem demotischen Wörterbuch stehen, womit aber die Texttradition unterbrochen wird. Anders gesagt: die Textüberlieferung und ihre lexikographischen Besonderheiten können nur so lange betrachtet werden, wie sie innerhalb der hieroglyphisch-hieratischen Schrift erfolgt.

Diese Nachteile verdeutlichen, dass ein einheitlicher Zugang zum ägyptisch-koptischen Wortschatz unabhängig von der Schriftform geschaffen werden muss, der die wissenschaftshistorisch bedingte Trennung der Teildisziplinen aufhebt. Als weitere Vorteile sind zu nennen: Man kann den Wortschatz unabhängig von der Schriftform synchron betrachten oder Teilcorpora nach eigenen Vorstellungen, beispielsweise anhand der Geographie, zusammenstellen. Die Laufzeiten von Wörtern und auch die Entwicklung von Kollokationen werden sichtbar.

Wie erreicht man nun diese Wunschvorstellung? Welche Schwierigkeiten ergeben sich, wenn man die verschiedenen einzelnen Zugänge zum ägyptisch-koptischen Wortschatz miteinander verknüpfen möchte?

Sprachwandel

Wenn man das gesamte ägyptisch-koptische Lexikon in den Blick nimmt, hat man als Ergebnis keinen synchronen Schnitt, sondern ist in hohem Maße der Diachronie verpflichtet. Im Laufe der Sprachgeschichte vollzogen sich viele Lautwandel, sodass die Relation[22]

---

[21] Vgl. Bodl. MS. Egypt. a. 3(P), SMITH, M., *Traversing Eternity. Texts for the Afterlife from Ptolemaic and Roman Egypt*, Oxford 2009, 661-662.

[22] Nicht jede Aufhebung der Opposition zweier Grapheme muss notwendigerweise mit einem Lautzusammenfall einhergehen. So hat das Koptische für die Dentale nur noch ein Graphem. Dieses Graphem ⲁ [X1] repräsentiert aber mehrere Phoneme, die vor dem Neuen Reich auf der Graphemebene noch auseinandergehalten wurden, vgl. SCHENKEL, W., Glottalisierte Verschlusslaute, glottaler Verschlusslaut und ein pharyngaler Reibelaut im Koptischen, in: *LingAeg* 10, 2002, 1-57.

zwischen Phonem und Graphem nicht als eine Konstante anzusprechen ist, sondern vielmehr Veränderungen unterworfen ist. So bewirkte beispielsweise der Lautzusammenfall der Phoneme, die durch  [S29] und  [O34] repräsentiert werden, dass die Grapheme ab dem Mittleren Reich austauschbar verwendet werden können. Die Veränderungen der Phonem-Graphem-Relationen spiegeln sich auch in den Transkriptionsalphabeten: So besitzt die Umschrift des Demotischen kein *d*. Wörter, die in den Transkriptionsalphabeten der vordemotischen Sprachstufen ein *d* aufweisen, werden im Demotischen mit einem *t* transkribiert. Neben der Reduktion des Zeicheninventars gibt es aber auch Zeichen, die vor dem Demotischen nicht verwendet werden, dies ist bei *l* der Fall. Somit sind die verwendeten Zeichen der Transkriptionsalphabete nicht eineindeutig aufeinander abbildbar. Für das Koptische hat sich im Gegensatz zu den vorkoptischen Sprachstufen kein eigenes Transkriptionssystem durchgesetzt. Vielmehr wird für die Wiedergabe der koptischen Wörter das koptische Schriftbild verwendet, das auf dem griechischen Alphabet mit einigen Zusatzbuchstaben beruht. Der Unterschied besteht aber nicht nur im Zeicheninventar, auch die Alphabetreihenfolge ist unterschiedlich. So wird im ägyptischen und im demotischen Transkriptionsalphabet das *ḥ* vor das *š* sortiert. Das koptische Alphabet setzt jedoch das ϣ, das Nachfolger des *š* ist, vor das ϩ, das Nachfolger des *ḥ* ist. Würde man nun für einen einheitlichen Zugang zum ägyptisch-koptischen Wortschatz eine einheitliche Transkription eines Wortes, welche sich an der Umschrifttradition einer bestimmten Sprachstufe orientiert, wählen, entstünden zunächst Umschreibungen ägyptisch-koptischer Wörter, die außerhalb der bevorzugten Sprachstufe in keiner Tradition stehen und somit zunächst schwer lesbar sind. Zur Verdeutlichung könnte man zum Beispiel annehmen, man wollte für den einheitlichen Zugang zum Wortschatz ein am Mittelägyptischen orientiertes System verwenden. Dann entstünde folgende Umschrift eines koptischen Satzes: *r-ḏbꜣ wˁ (tw=)k mdw.t jr=f r-ḏd* … Schwerlich erkennt man darin die Vorlage: ⲉⲧⲃⲉ ⲟⲩ ⲕⲙⲟⲩⲧⲉ ⲉⲣⲟⲕ ϫⲉ … Man mag einwenden, für die Darstellung des Wortschatzes seien die Probleme der Lesung eines Fließtextes und die Wiedergabe grammatisch-syntaktischer Eigenheiten wie hier das Personalpronomen im Präsens I nicht von Belang. Aber Koptisch lässt sich aus vier Gründen kaum in am Mittelägyptischen orientierten Transkriptionszeichen schreiben. Zunächst verhindern die oben angesprochenen Lautwandel die richtige Darstellung. Ein koptisches ⲧ kann auf mittelägyptisches *t*, *ṯ*, *d* oder *ḏ* zurückgehen. Wenn der Vorläufer eines

koptischen Wortes aus vorkoptischer Zeit nicht bekannt ist, kann man nicht sicher beurteilen, welches Umschriftzeichen man wählen muss. Wie soll z.B. ⲧⲣⲟⲡ in mittelägyptischer Umschrift aussehen? Darüber hinaus ist manches Mal nicht sicher, wie der Vorgänger eines koptischen Wortes aussieht. Das Fragepronomen ⲟⲩ ist nach dem Vorschlag von Sethe[23] als *wꜥ*, „eins“[24] aufgefasst. Nach Westendorf[25] geht es aber vielleicht auf *ꜥ.w*, „Person; Individuum“[26] zurück. Die eindeutige Ansprechbarkeit, die durch die Wiedergabe der koptischen Vorlage gegeben ist, kann also durch die fehlende Sicherheit in den Etymologien verloren gehen. Ferner besitzt das Koptische eine sehr hohe Zahl griechischer Lehnwörter. Man hätte in diesem Fall einen ahistorischen Standard zu schaffen, wie man diese Wörter mit mittelägyptischen Umschriftzeichen wiedergibt. Schließlich notiert ein Umschriftsystem des Mittelägyptischen keine Vokale, welche aber in der koptischen Schrift erscheinen. Somit geht eine wichtige Information zur lexikalischen Disambiguierung verloren. Leider birgt das entgegen gesetzte Vorgehen auch Probleme: Wenn alle mittelägyptischen Wörter mit koptischen Buchstaben geschrieben werden sollen, wird man auf die koptischen Nachfolger der mittelägyptischen Wörter zurückgreifen. Hierbei gibt es nun Wörter, die in der Hieroglyphenschrift unterschiedbar sind, aber im Koptischen nicht. So lauten die Nachfolger von *mḏ*, „zehn“[27] und *mtj*, „richtig“[28] jeweils ⲙⲛⲧ. Wenn es solche Nachfolger aber nicht gibt, kann nicht gefolgert werden, wie ein mittelägyptisches Wort notwendigerweise im Koptischen gelautet hätte, da die Lautwandelprozesse nicht in allen Fällen durchgeführt worden sind. Ein mittelägyptisches *ṯ* kann im Koptischen als ⲧ oder auch als ϫ erscheinen. Somit ist die Darstellung eines Wortes in einem einheitlichen, über alle Sprachstufen und Zeiten hinweg gültigen Transkriptionsäquivalent schwerlich durchführbar.

Im Laufe der Sprachgeschichte sind die ägyptischen Wörter nicht nur Lautwandelprozessen unterworfen, die Bedeutungen der Wörter können sich auch verändert haben. So entwickelt sich das im Neuen Reich belegte Wort *ḫr*, „Syrer“ zu einer allgemeinen Bezeichnung für

---

[23] SETHE, K., Untersuchungen über die ägyptischen Zahlwörter, in: *ZÄS* 47, 1910, 4.

[24] WCN:44150.

[25] WESTENDORF, W., *Koptisches Handwörterbuch*, Heidelberg 1964/1977, 264.

[26] WCN:34510.

[27] WCN:78340.

[28] WCN:77420.

„Diener“, die im Koptischen als ϩⲁⲗ erscheint. Verknüpft man direkt das Wort *ḫr* des Neuen Reiches mit dem ϩⲁⲗ der koptischen Zeit, erzeugt man u.U. in Abfragen eine geringe *Precision*. Wenn man nämlich Belege für ägyptische Dienerbezeichnungen sucht, erhält man dann auch die nicht gewollten Belege mit der Bedeutung „Syrer“. Verknüpft man aber nicht, lässt sich nicht herausfinden, ob dieses Wort trotz des Bedeutungswandels gewisse Konstanzen vielleicht im Bereich der verwendeten Konstruktionen oder der beteiligten Mitspieler ausweist.

Abbildbarkeit der Listen

Ingelore Hafemann konnte für die Berliner Wortliste und für die Tübinger Liste nachweisen, dass zwischen den Lemmata der einzelnen Listen die Relationen 1:1, 1:n, n:1 und n:m bestehen.[29] Die beiden Listen können somit nicht direkt maschinell aufeinander abgebildet werden. Die für die Verknüpfung so hinderliche Relation n:m tritt vor allem bei Verben auf, da es dort keine eindeutige Maßgaben gibt, wann man noch unterschiedliche Lesarten eines Verbums oder doch schon ein neues Lemma anzusetzen hat. Die n:m-Relation tritt auch im Vergleich zwischen der Berliner Wortliste und der Demotischen Wortliste auf, sodass auch hier keine direkte Verknüpfung hergestellt werden kann. Für eine Verknüpfung mit der noch zu schaffenden Liste des Koptischen ist Ähnliches zu erwarten. Für eine Abbildbarkeit der Listen gibt es zwei unterschiedliche Strategien: 1) Man weist zunächst innerhalb der Listen die Lemmata ihren Wurzeln zu. Danach verknüpft man die Wurzeln. Dieser arbeitstechnisch relativ geringe Aufwand hat den entscheidenden Nachteil, dass damit die *Precision* stark leidet: Man kann damit nicht mehr die Entwicklung eines Verbums im Laufe der Sprachgeschichte betrachten. Die Ergebnismenge ist durch die Belege der deverbalen Ableitungen durchsetzt. 2) Man überarbeitet die Ansetzungen in den Listen; d.h. man gleicht in den zu verknüpfenden Listen die Einträge derart an, dass keine n:m-Relationen bestehen bleiben. Bei dieser Umarbeitung der Listen ist darauf zu achten, dass die den Lemmata zugeordneten Belegstellen auch überprüft und gegebenenfalls neu zugewiesen werden. Dadurch entsteht zwar eine gewünschte Abbildbarkeit der

[29] HAFEMANN, I., Die Verknüpfung der Tübinger Lemmaliste mit der Berliner Wortliste, in: HAFEMANN, I. (Hrsg.), *Wege zu einem digitalen Corpus ägyptischer Texte. Akten der Tagung „Datenbanken im Verbund“ (Berlin, 30. September – 2. Oktober 1999)*, Thesaurus Linguae Aegyptiae 2, Berlin 2003, 84-85.

Listen, jedoch ist ein immens hoher Arbeitsaufwand zu veranschlagen.

Struktur des Ergebnisses

Schließlich muss man sich fragen, wie das Ergebnis einer Verknüpfung verschiedener Listen aussehen soll. Wenn man eine einheitliche Wortliste für das gesamte Ägyptische generiert, so weisen die Einträge ein unterschiedliches Zeicheninventar auf und sind nicht sortierbar. Stattdessen könnte man eher an ein Netzwerk denken, in dem die verschiedenen synchron definierten Listen des Wortschatzes verknüpft sind. Bei einer Netzwerklösung ist aber zu bedenken, dass zum einen die Pflege des Netzwerkes aufwändiger ist als bei einer einheitlichen Liste, da man bei jeder Änderung in einer Liste auch die damit verbundenen Einflüsse auf das Netzwerk zu berücksichtigen hat. Zum anderen ist auch die Gestaltung der Abfrageprozeduren an ein Netzwerk komplexer als die an eine Liste. Insofern ist doch eher an eine gemeinsame Liste zu denken. Das Problem der Sortierbarkeit beispielsweise von Trefferlisten ist gesondert zu lösen.

Auf dem Weg zu einer integrierten Wortliste / einem integrierten Lexikon

Welche Probleme und Hindernisse ergeben sich nun, wenn man sich praktisch an die Umsetzung der Forderung nach einem einheitlichen Lexikon des Ägyptisch-Koptischen, ausgehend von den getrennten Wortlisten und Textcorpora des Ägyptischen und Demotischen, macht? Betrachten wir dazu als Beispiel einen Lemmaeintrag für ein Wort, das über den gesamten Zeitraum der ägyptisch-koptischen Sprachgeschichte belegt ist, mit seinem Eintrag im TLA.

wꜣḥ legen, setzen; opfern, weihen; dauern; lassen AR – Gr/Röm

Abb. 1: Lemmaeintrag w³ḥ in der ägyptischen Wortliste

Ein Blick in das gedruckte *Wörterbuch der Ägyptischen Sprache* (Wb) macht sofort den Unterschied zu dem hier abgebildeten Eintrag im TLA deutlich. Die semantische Differenzierung von *w³ḥ* im Wb I 253.1-257.6 ist subtiler strukturiert und geht über fast fünf Seiten. Der TLA kann diesen Nachteil vor allem durch den direkten Zugriff auf das Textcorpus und die Ausgabe der Wortverwendung im Kontext kompensieren. Natürlich wäre auch für das elektronische Lexikon des TLA eine stärkere semantische Strukturierung wünschenswert, um sich gezielt Belegstellen zu einer bestimmten Bedeutung zusammenstellen zu können. So sollten natürlich die Bedeutungen *legen/setzen, dauern, opfern, lassen* hierarchisch strukturiert und einzeln abfragbar sein.

In der demotischen Lemmaliste stellt sich der Lemmaeintrag für *w³ḥ* folgendermaßen dar:

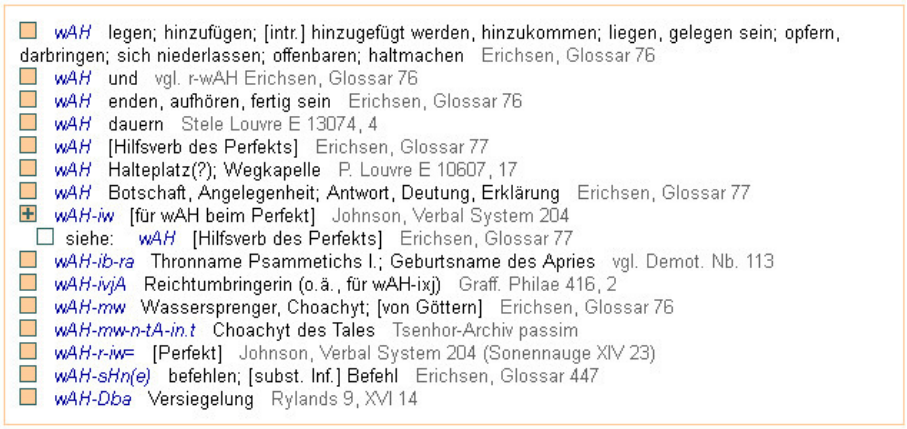


Abb. 2: Lemmaeintrag w³ḥ in der demotischen Wortliste

Hier ist auffällig, dass die semantische Differenzierung der schon im älteren Ägyptisch belegten Bedeutungen noch weniger strukturiert ist. Eine tiefer gehende Strukturierung erscheint daher unbedingt wünschenswert. Allerdings kommen nun andere und neue Bedeutungen und Grammatikalisierungen hinzu, die sich offensichtlich im Laufe der Sprachgeschichte ergeben haben. Schauen wir uns einige Details an und versuchen sogleich, wenn möglich, eine „Netzwerkstruktur“ vom älteren Ägyptisch bis zum Koptischen darzustellen.

w³ḥ (Wb I 253.1-257.6)

→ *w³ḥ* (demotisch) *legen; hinzufügen; liegen, gelegen sein; opfern, darbringen; sich niederlassen; offenbaren; haltmachen* (Übersetzungsvarianten der demotischen Lemmaliste im TLA) Erichsen, Demotisches Glossar (EDG), 76

→ ⲟⲩⲱϩ Westendorf, Koptisches Handwörterbuch (KHB), 284-285 *legen, setzen, werfen, stürzen, stoßen, hinzufügen; sich niederlassen, sich aufhalten, bleiben, wohnen*

Aufgrund der verloren gegangenen Differenzierung in der Schreibung ist im Demotischen eine Kontamination der Lemmata (Wurzeln) *w³ḥ* und *wḥ*ʿ zu *w³ḥ/wḥ* zu beobachten. Die hier folgende Bedeutungsvarianz wird in den demotischen Lexika unter *w³ḥ* aufgeführt, ist jedoch von *wḥ*ʿ entlehnt. Ein „Durchgriff“ von der demotischen zur ägyptischen Lemmaliste müsste das berücksichtigen:

*wḥ*ʿ (Wb I 349.5-7)
→ wꜣḥ (demotisch) *enden, aufhören, fertig sein* EDG 76
→ ογω *aufhören, beenden; zu Ende sein, verweilen, bleiben* KHB, 266; CED[30], 210; DELC[31] 230

Obwohl EDG (Verweis auf: Wb I 253) und CED (Verweis auf: Wb I 255, top) auf das Wb und das ältere Ägyptisch verweisen, findet sich die spezifische Bedeutung *enden, aufhören, fertig sein* dort nicht. DELC verweist gleich auf den ganzen Lemmaeintrag von *wꜣḥ* (Wb I 253.1-257.5, notamment 256,1-5). Im TLA ist es jedoch möglich, dass ein Beleg für diese Bedeutung im TLA unter dem wenig differenzierten Lemmaeintrag *legen/setzen, dauern, opfern, lassen* subsumiert (lemmatisiert) wurde. Dieser wäre anhand der Belegstellenabfrage zwar mühselig zu suchen, aber durchaus herauszufinden.

wꜣḥ (demotisch) in der Bedeutung *dauern* ist in der demotischen Lemmaliste jetzt separiert und hat auch nur eine Belegstelle im Textcorpus (Stele Louvre E 13074, 4). EDG hat diese Bedeutung nicht aufgeführt aber das *Chicago Demotic Dictionary* (CDD)[32], allerdings scheinen die Belege auch hier selten und zum Teil problematisch zu sein. Auch in der *Demotischen Wortliste online* ist die Bedeutung *dauern* nur mit einem Fragezeichen vermerkt.[33] Es hat den Anschein als konnte *wꜣḥ* im Demotischen die Bedeutung *dauern* nicht mehr, oder nur noch relikthaft tragen.

Erst ab der demotischen Sprachstufe sind folgende Bedeutungsvarianten als Nomina belegt:
wꜣḥ/wḥ (demotisch) *Botschaft, Angelegenheit; Antwort, Deutung, Erklärung*;
jr wꜣḥ antworten, *ḏd/tꜣy wꜣḥ antworten, ein Orakel geben, verkünden, ankündigen*: EDG 77; CDD w (9:01) Pages 13-15
→ ογωϱ *Botschaft, Nachricht* KHB 285
→ ογω *Botschaft, Nachricht, Antwort, Verkündigung*
ⲣ̄ ογω, ϫⲓ (ⲛ)ογω *antworten, berichten; verkünden, ankündigen, melden* KHB 266; CED 210

[30] ČERNÝ, J., *Coptic Etymological Dictionary*, Cambridge 1976.
[31] VYCICHL, W., *Dictionnaire Étymologique de la Langue Copte*, Leuven 1983.
[32] Buchstabe *w* (7 August 2009): 9.01 = CDD w (9:01) Page 8; http://oi.uchicago.edu/research/pubs/catalog/cdd/.
[33] http://www.dwl.aegyptologie.lmu.de/suche.php?nummer=01193 (aufgerufen am 25.05.2012).

Der Verbstamm transportiert diese Bedeutung aber ebenso vom älteren Ägyptisch bis zum Koptischen:
wḥˁ (Wb I 348.3-348.15) *lösen, erklären, deuten*
→ *wꜣḥ* (demotisch) *lösen, erklären* EDG 77; CDD w (9:01) Pages 13-14
→ ογω *antworten, verkünden, aufklären* KHB 266; DELC 230
→ ογωϩ *erklären, deuten, lösen* KHB 285; CED 222; DELC 241-242

Die allgemeine Tendenz des jüngeren Ägyptisch, Verben durch Funktionsverben (vor allem *jrỉ* und *rd̲ỉ*) + Substantiv zu ersetzen, spielt hier natürlich eine Rolle. Doch die aus dieser Darstellung offensichtlich erscheinende Ableitung des Bedeutungsfeldes *lösen, erklären*, mit der Erweiterung zu *antworten, verkünden, ankündigen* (+ Substantive) *wḥˁ* → *wꜣḥ/wḥ* → ογω wurde durch die Separierung der Lexikographie für Ägyptisch, Demotisch und Koptisch bisher nicht transparent, das gilt leider auch für den TLA. In der demotischen Lemmaliste kann man die Kollokationen *jr wꜣḥ* und *d̲d/tꜣy wꜣḥ* nur über die Abfrage zum Substantiv in der Textdatenbank zusammensuchen (aktuell 81 Belege!). Das vereinzelte Auftreten von ογωϩ in dieser Bedeutung bezeugt auch noch für das Koptische die lautliche Kontamination der einst getrennten Wurzeln *wꜣḥ* und *wḥˁ*.

wꜣḥ (demotisch) *Halteplatz(?); Wegkapelle* P. Louvre E 10607, 17 → CDD w (09:1) Page 12-13
→ ογωϩ KHB 284.
Die genannte Bedeutung als Substantiv scheint im Demotischen noch recht selten zu sein, entwickelt aber im Koptischen ein breites Bedeutungsspektrum: *Aufenthaltsort, Wohnort, Platz* usw., dazu korrespondierend das Verb mit der Bedeutung *sich niederlassen, sich aufhalten, bleiben, wohnen*. Etymologisch liegt wohl auch hier eine Kontamination von *wꜣḥ* und *wḥˁ* (vgl. Wb I 349.8-13) vor.

Grammatikalisierte Bildungen des Stammes *wꜣḥ*, die erst in der demotisch-koptischen Sprachstufe auftreten, sind besonders wegen der im Koptischen in den Dialekten verfolgbaren Unterschiede in Bildung und Verwendung interessant, z.B. die Konjunktion:

r-wꜣḥ/e-wꜣḥ (demotisch) *und* geht auf den neuägyptischen Imperativ *j:w(ꜣ)ḥ* zurück[34] *füge hinzu* o.ä. EDG 76; CED 14; KHB 14; DELC 18-19, 241. Besonders relevant ist hierbei, dass die in den koptischen Dialekten[35] auftretenden verschiedenen Formen der Konjunktion unterschiedlichen grammatischen Bildungen entsprechen:[36]
r-wꜣḥ (Imperativ) → ⲁⲩⲱ (*S, L, F, M*), ⲁⲟⲩ (*A*), ⲁⲩⲟⲩ, ⲁⲩ (*S, F*)
w(ꜣ)ḥ (Infinitiv) → ⲟⲩⲟⲉ (*B*), ⲟⲩⲁⲉ, ⲟⲩⲱⲉ (Altkoptisch), ⲟⲩⲁⲉⲁ (*L*)

Die erst im Demotischen auftretende Verwendung von *wꜣḥ* als ‚Hilfsverb des Perfekts', also als grammatikalisiertes Morphem oder Konjugationsbasis, erscheint nur in einigen koptischen Dialekten, noch dazu in differenzierter Form und Verwendung.
wꜣḥ (demotisch) EDG 77; CDD w (9:01) Page 13
→ ⲁⲉ-, ⲁⲉⲁ-, ⲁⲉ= (*L*) KHB 16-17; CED 17; DELC 22
findet sich nur im Lykopolitanischen (*L*), in thebanischen Urkunden und altkoptischen Texten.[37] Es handelt sich folglich wohl um eine oberägyptische Varietät.[38] Allerdings konkurriert dieses auch in *L* mit der weiter verbreiteten sahidischen Bildung des Perfekts ⲁ-, ⲁ=.[39] Im Achmimischen (*A*) wird nur -ⲁⲉ- mit dem Relativkonverter ⲉⲧⲁⲉ- ((ⲉ)ⲛⲧⲁⲉ-) verwendet, wenn das Subjekt des Relativsatzes identisch mit dem Antecedens ist.[40] Diese Formen gehen auf das perfektische

[34] Vgl. WINAND, J., *Études de néo-égyptien 1, La morphologie verbale*, AegLeo 2, Liège 1992, 164; *wꜣḥ* ist schon hier offensichtlich zu einem zweiradikaligen Stamm *wḥ* reduziert worden.

[35] Zur modernen Einteilung der koptischen Dialekte und ihren aktuellen Sigla, die z.B. nicht mehr den in den einschlägigen Lexika (KHB, CED) entsprechen, vgl. KASSER, R., Dialects; Dialects, Grouping and Major Groups of, in: *The Coptic Encyclopedia* 8, 87-101.

[36] Wahrscheinlich ist die besondere koptische Form ⲱ (*S, F*; KHB, 14) aus einem im Demotischen in dieser Funktion belegten *jw-ꜥ* entstanden (EDG 76; CDD w (9:01) Page 8).

[37] HAARDT, R., Koptologische Miszellen, in: *WZKM* 57, 1961, 96-97.

[38] RICHTER, T. S., Das demotische Konjugationspräfix *wꜣḥ* und die ⲉ-haltigen Konjugationsformen des Koptischen, in: *Enchoria* 24, 1997/98, 76-77, vgl. 68-71; FUNK, W.-P., Die Morphologie der Perfektkonjugation im NH-subachmimischen Dialekt, in: *ZÄS* 111, 1984, 110-130, vgl. 111 und 117.

[39] Vgl. NAGEL, P., Lycopolitan (or Lyco-Diospolitan or Subachmimic), in: *The Coptic Encyclopedia* 8, 157.

[40] Sporadisch auch in den Nag-Hammadi Texten in *L* und im Proto-Sahidischen P.Bodmer VI (*P*), vgl. FUNK, Morphologie der Perfektkonjugation, 111-112.

Partizip bzw. die perfektische Relativform des Neuägyptischen/Demotischen *j:wꜣḥ** zurück.[41]

Dagegen ist → ⲉⲁ-, ⲉⲁ= (*L, F, M*) KHB 348; CED 269; DELC 284 aus der *sḏm=f*-Form *wꜣḥ=f* entstanden, ebenso wie das weiter verbreitete sahidische ⲁⲩ- aus *jrị=f*. Es erscheint in *L* selten und überwiegend in einer bestimmten Textgruppe[42], war aber in Mittelägypten und im Fajjum weiter verbreitet, ebenso – oftmals im gleichen Text – wiederum in Konkurrenz mit ⲁ-, ⲁ=.[43] Hier wäre von besonderem Interesse, die geographische Verteilung des Auftretens von *wꜣḥ* in der genannten Funktion im Demotischen zu verfolgen.

Es wurden bereits, ausgehend von der ägyptischen und der demotischen Lemmliste, eine ganze Reihe von spezifischen Problemen dargelegt, die sich bei der Verbindung der ägyptischen Lemmata über alle Sprachstufen hinweg zur Strukturierung eines integrierten ägyptisch-koptischen Lexikons ergeben können. Einige Kernbereiche lexikographischer Erschließung sind aber bisher noch gar nicht einbezogen worden, da sie im TLA nur sekundär über eine Kollokationsabfrage erreichbar sind. Dazu gehören in erster Linie die Kollokationen von Verben mit Präpositionen (und Substantiven). Sie machen einen wichtigen Teil der semantischen Strukturierung der Lemmaeinträge des Wb aus und sind auch im Standardlexikon des Koptischen, dem *Coptic Dictionary* von W. E. Crum (CCD)[44], in einer Ausführlichkeit geboten, die die bisher genannten Lexika des Koptischen nicht aufweisen. Allerdings gibt CCD keine oder nur sporadische Hinweise auf die etymologischen Beziehungen zum vorkoptischen Ägyptisch, dafür aber die griechischen Entsprechungen der koptischen Lemmata aus der Übersetzungsliteratur, die die Bibel und die christliche Literatur aus dem Griechischen ins Koptische übertragen hatte. Diese bieten einen weiteren Aspekt zur semantischen Erschließung des Ägyptischen der koptischen Sprachstufe.

---

[41] WINAND, *Études de néo-égyptien*, 344-353 und 376-384; JOHNSON, J., *The Demotic Verbal System*, SAOC 38, Chicago 1976, 182.

[42] NAGEL, Lycopolitan, 157; RICHTER, Konjugationspräfix, 69-70; FUNK, Morphologie der Perfektkonjugation, 111 und 115.

[43] BOSSON, N., *Wörterverzeichnis zu Gawdat Gabras Ausgabe des Psalters im Mesokemischen (Oxyrhynchitischen/Mittelägyptischen) Dialekt des Koptischen (Mudil-Psalter)*, CSCO 568, Subsidia 96, Leuven 1997, 313; BOSSON, N., Remarques sur la « structure (ⲉ)ⲁ- . . . (ⲉ)ⲁⲩ- », in: *LingAeg* 14, 2006, 281-300.

[44] CRUM, *Coptic Dictionary*.

Hier sollen einige Beispiele für semantische Veränderungen bzw. Konstanten von Verb + Präposition Kollokationen von *wꜣḥ*, vom Ägyptischen über das Demotische zum Koptischen, folgen, wie sie aus dem Wb, EDG, CDD und CCD erschließen lassen.

Es zeigen sich u.a. semantische Konstanten, die im Laufe der Sprachentwicklung mit veränderten „Mitspielern" realisiert werden. Auffällig ist auf der Ebene der demotisch-koptischen Sprachstufe die Zunahme der Varianz der Präpositionen bzw. der Kombinationsmöglichkeiten der Präpositionen, die mit *wꜣḥ* in Verbindung treten können. Leider lässt sich das in vielen Fällen für das Demotische (noch) nicht zeigen, da die Lexika des Demotischen (EDG, CDD sowie die demotische Wortliste des TLA) solche Kollokationen nur in sehr begrenzter Auswahl aufführen. Dagegen bieten Wb und CCD schon eine sehr weitgehende Strukturierung der Lemmaeinträge, die im Bereich solcher Kollokationen wiederum als Modell zur Strukturierung und Verlinkung der integrierten Wortliste aus drei Teillisten (Ägyptisch, Demotisch, Koptisch) dienen könnte. Für die demotische Lemmaliste (aber großflächig auch für die ägyptische Lemmaliste) im TLA ist eine solche Strukturierung anhand des Textcorpus noch durchzuführen.

■ *etw. ablegen, niederlegen (zur Erde)*
wꜣḥ r tꜣ Wb 1, 253.7-9
→ CDD w (09:1) Page 8
→ ⲟⲩⲱϩ ⲉⲡⲉⲥⲏⲧ CCD 507b

■ *etw. ablegen, niederlegen, an einen Ort legen (bringen), etw. vor jmd. niederlegen* usw.
wꜣḥ / m / m-bꜣḥ Wb 1, 253.10-15
→ ⲟⲩⲱϩ ⲛ-, ⲛⲁϩⲣⲏⲛ-, ϩⲁ-, ϩⲓⲣⲉⲛ-, ϩⲁ(ϩ)ⲧⲏ̄-, ϩⲓⲝⲏ̄-, ⲉⲃⲟⲗ, ⲉϩⲟⲩⲛ CCD 506b-507b

■ *die Hand legen an/auf jmd./etw., (auf) etw. zeigen*
wꜣḥ ḏr.t ḥr / r Wb 1, 253.16-18
→ ⲟⲩⲱϩ ⲉⲧⲛ-, ⲉⲧⲟⲟⲧ= CCD 506a

■ *liegen, sich befinden, wohnen (an einem Ort)*
wꜣḥ ḥr / r / m Wb 1, 253.19-23
→ ⲟⲩⲱϩ ⲉ-, ⲉⲝⲏ̄-, ⲙ̄ⲛ, ⲛ̄-, ϩⲁ-, ϩⲓ-, ϩⲛ̄-, ϩⲓⲝⲛ̄- CCD 508b-509a

■ *hinzufügen (etw. zu etw.)*
wꜣḥ r / ḥr Wb 1, 254.7-12
→ Erichsen, Glossar, 76; CDD w (09:1) Page 11[45]
→ ⲟⲩⲱϩ ⲉ-, ⲉⲣⲛ̄-, ⲉⲧⲛ̄-, ⲉⲧⲟⲟⲧ=, ⲉϫⲛ̄-, ⲛ̄- CCD 505b-506b

■ *jmd. folgen, hinter jmd. her sein*
wꜣḥ m-sꜣ (Wb 4, 10.4-13 bei *m-sꜣ*!)
→ Erichsen, Glossar, 76; CDD W (09:1) Page 10
→ ⲟⲩⲱϩ ⲛ̄ⲥⲁ-, ϩⲓⲡⲁϩⲟⲩ ⲛ̄- CCD 506b + 507a; KHB 285

Aus CCD könnte man noch die griechischen Entsprechungen zu den koptischen Kollokationen gewinnen und erhielte so einen tieferen Einblick in die Semantik des Koptischen. Die Gräzistik und die Übersetzungswissenschaft wiederum gewönnen einen Einblick in Übersetzungsäquivalente strukturell so verschiedener Sprachen wie Ägyptisch und Griechisch. Die ägyptisch-koptische Kollokation *wꜣḥ m-sꜣ* → ⲟⲩⲱϩ ⲛ̄ⲥⲁ- konnte nach CCD 506b in Griechisch folgendermaßen wiedergegeben werden:

ⲟⲩⲱϩ ⲛ̄ⲥⲁ- → (ἐπ-), (παρ-), (συν-), ἀκολουτεῖν; πορεύεσθαι ὀπίσω; καταδιώκειν ὀπίσω; προσκεῖσθαι πρός; ἐπιτιθέναι ἐπί; περιάγειν

Abschließend soll noch ein Blick auf Kollokationen mit den Funktionsverben *jrị* und *rḏị* geworfen werden, die im Ägyptisch-Koptischen besonders häufig sind und seit dem älteren Ägyptisch breite Verwendung finden. Im jüngeren Ägyptisch werden Grammatikalisierungen aus diesen Verben zu Hilfsverben und schließlich zu Konjugationsbasen als Präfixe. Sie sind somit die im Textcorpus am besten und zahlreichsten belegten Lemmata. Einige Beispiele für Bedeutungskonstanzen und Veränderungen zum Koptischen hin:

jrị hrw den Tag verbringen Wb 1, 109.24-25 (Lemma: *jrị*)
→ ⲣ̄ (ⲡⲉ)ϩⲟⲟⲩ den Tag verbringen (KHB, 403; Lemma: ϩⲟⲟⲩ)

[45] dazu: *m wꜣḥ ḥr / r* Wb 1, 254.9-12 → ERICHSEN, *Glossar*, 76; CDD w (09:1) Page 12 (vgl. oben).

jrị ḥ3.w Besitz ergreifen (von); (jmdn.) verhaften (jurist.) Wb 2, 478.17-18 (Lemma: *jrị*)
→ ⲣ̄ ϩⲏⲩ *nützen* (KHB, 402; Lemma: ϩⲏⲩ)

rḏị j3.w (jmdn.) preisen Wb 1, 28.3 (Lemma: *j3.w*)
→ ϯ ⲉⲟⲟⲩ *rühmen* (KHB, 402; Lemma: ⲉⲟⲟⲩ)

rḏị ḥ3.tj sich sorgen um (u.Ä.) Wb 3, 27.17 (Lemma: *rḏị*)
→ ϯ (ⲛ̄)ϩⲧⲏ= *aufmerksam sein, beachten* (KHB, 219; Lemma: ϯ)

rḏị ḥr Weisung erteilen; Aufmerksamkeit schenken Wb 3, 127.9 (Lemma: *ḥr*) [Wb ibidem: vgl. Kopt. ϯ ϩⲟ?]
→ ϯ ϩⲟ *bitten, trösten* (KHB, 219; Lemma: ϯ)

rḏị ḥꜥ aufstellen; bereitstellen Wb 1, 219.15 (Lemma: *ḥꜥ*)
→ ⲧⲁϩⲟ *aufstellen, zufrieden stellen, festsetzen, erreichen* etc.; mit ⲉⲣⲁⲧ= *auf die Füße stellen, errichten* etc. (KHB, 257)

Es ist heute gar keine Frage mehr, dass die Schaffung eines integrierten digitalen Lexikons des Ägyptisch-Koptischen in Verbindung mit einem möglichst umfangreichen Referenzcorpus einerseits – wissenschaftsgeschichtlich – eine notwendige und überfällige, andererseits – perspektivisch – eine vielversprechende und zeitgemäße Aufgabe für die Ägyptologie ist. Das hier Dargelegte sollte einen Einblick in die Aufgaben und Probleme, aber auch in die Chancen und Vorteile gewähren, die sich auf dem Weg zu einem integrierten Lexikon des Ägyptisch-Koptischen ergeben werden.

DIE DEMOTISCHE WORTLISTE – VIRTUELL ERWEITERT

FRIEDHELM HOFFMANN

Die Demotische Wortliste (DWL), seit 2005 online,[1] ist, wie der Name sagt, kein Wörterbuch, sondern eine Liste von Wörtern, in der man Wortschreibungen suchen kann. Verschiedene orthographische Varianten eines und desselben Wortes stehen als gleichberechtigte Einträge in der Datenbank nebeneinander und sind nicht hierarchisch einem Lemma untergeordnet. Die DWL erfasst also in erster Linie nicht den demotischen Wortschatz, sondern demotische Schreibungen, denn sie soll ein Werkzeug für die Entzifferungsarbeit bereitstellen. Deshalb ist die DWL so konzipiert, dass man in ihr auch nach etwas suchen kann, von dem man noch gar nicht weiß, was es ist.

Ich möchte kurz den Aufbau und die Möglichkeiten der Demotischen Wortliste erklären.[2] Der Eingangsbildschirm präsentiert zugleich auch schon das Menü, das die Punkte „Start“, „Information“, „Benutzungshinweise“, „Downloads“, „Kontakt“ und „Suche“ umfasst.

Der wichtigste Menüpunkt ist die „Suche“. Voreinstellung ist die Komfortsuche, auf die ich mich im Folgenden beschränke. Die Suchmaske bietet zwei grundsätzlich verschiedene und unabhängig voneinander operierende Suchmöglichkeiten an: erstens über die Wortnummer, zweitens über das Wort, also das demotische Wort in Umschrift, die Determinierung und die Übersetzung. Die Wortnummer kann dazu herangezogen werden, aus der Lemmaliste der Mainzer bzw. Berliner Demotischen Textdatenbank[3] auf meine Daten zu verweisen.

Wichtiger ist die zweite Art der Datenbankabfrage, nämlich die über das Wort in Umschrift, die Determinierung und ggf. auch die Übersetzung. Die drei Felder „Wort“, „Determinativ“ und „Über-

---

[1] http://www.dwl.aegyptologie.lmu.de.

[2] Ausführlicher hierzu HOFFMANN, F., Ein demotistisches EDV-Werkzeug: die Demotische Wortliste (DWL), in: WIDMER, G. & D. DEVAUCHELLE (eds.), *Actes du IXe congrès international des études démotiques. Paris, 31 août - 3 septembre 2005*, BdE 147, Kairo 2009, 145-155. Das erste Drittel meines hier vorgelegten Beitrages stellt eine Zusammenfassung, streckenweise aber auch eine wörtliche Übernahme von Passagen dieser früheren Beschreibung dar.

[3] http://www.adwmainz.de/index.php?id=44&L=0; Zugang zur Datenbank über http://aaew.bbaw.de/tla/.

setzung“ sind kombinierbar durch „enthält“ oder „enthält nicht“ bzw. „UND“/„ODER“-Verknüpfungen usw.

Möglich sind selbstverständlich ganz simple Abfragen nach Wörtern, z.B. „sntr“. Man kann dabei spezifizieren, dass „sntr“ z.B. nicht ein beliebiger Teil des gesuchten Wortes sein soll, sondern der Wortanfang. Über andere Schalter lässt sich festlegen, ob die Suchzeichenfolge exakt das Wort repräsentieren oder vom Wortende stammen soll. Vorgesehen ist auch die Suche nach beliebigen Wortbruchstücken.

Nach Start und Abschluss der Suche wird das Ergebnis in Form einer Liste präsentiert. Sie enthält die Felder „Nummer“ (nämlich des Wortes), „Wort“ (in Umschrift), „Determinativ“, „Übersetzung“ und „Belegstellen“.

Jedes Determinativ ist durch die Folge eines Groß- und eines Kleinbuchstabens codiert, die möglichst sprechend gewählt sind. Für das Hausdeterminativ beispielsweise habe ich „Hs“ genommen. Die einzelnen Codierungen findet man in den Listen unter dem Button „Hinweise“ beim Determinativfeld der Suchmaske. Verschiedene Determinierungen ein und desselben Wortes sind in den Datenbankeinträgen berücksichtigt.

Einen Kernbestandteil der Komfortsuche bilden die sog. Erweiterten Suchoptionen. Über sie sind komplexe Abfragen, bei denen Alternativen berücksichtigt werden, sowohl nach der Umschrift als auch nach den Determinativen sehr leicht zu realisieren. Die Demotische Wortliste soll ja, wie gesagt, in erster Linie ein Hilfsmittel für die Entzifferungsarbeit sein. Nun kann in vielen Situationen aber noch gar nicht richtig klar sein, wonach man eigentlich suchen muss. Ist das fragliche Zeichen z.B. ein *b* oder ein *ḏ*, ein *ꜥ.wy*, *ꜥn*, *bn* oder *tn*? Handelt es sich um das Determinativ des sitzenden Kindes oder das Metalldeterminativ? Alle solche *graphisch* gleichen oder ähnlichen Zeichen muss man ja vielleicht zunächst als Möglichkeiten in seine Abfrage einbeziehen. In den Gleichbehandlungslisten der Erweiterten Suchoptionen ist dies durch Anklicken des entsprechenden Menüs und Aktivierung der benötigten Gleichsetzung leicht machbar. Dann braucht man in der Suchanfrage beispielsweise nur „*b*“ zu schreiben, und Wörter mit *ḏ* an derselben Stelle werden mitgefunden. Der Menüpunkt zu graphisch gleichen Determinativen funktioniert im Prinzip genauso.

Ich habe auch die Möglichkeit berücksichtigt, dass in einem neuen Text ein eigentlich altbekanntes Wort in einer neuen Schreibweise vorkommen könnte. Vielleicht ist die fajumische Form mit *l* ja noch

nicht in der Datenbank, sondern nur die mit *r*; vielleicht stammt ein Text aus römischer Zeit, und die Unterscheidung von ꜣ und ꜥ ist längst hinfällig. Dann sollen verschiedene Zeichen *phonetisch* gleichbehandelt werden. Das ist über die entsprechende Auswahlliste unter „Optionen für ‚Wort'" ebenfalls in den Erweiterten Suchoptionen leicht möglich.

Eventuell möchte der Benutzer ja auch *funktionsverwandte* Determinative bei seiner Abfrage automatisch mit berücksichtigen lassen, z.B. Baum- und Pflanzendeterminativ. Auch das ist über die Erweiterten Suchoptionen zu bewerkstelligen.

Die bisher geschilderte Funktionalität der DWL berechtigt mich eigentlich noch nicht dazu, meinen Beitrag in den Zusammenhang der corpusbasierten Philologie zu stellen. Denn meiner Datensammlung liegt zwar ein Corpus zugrunde, nämlich die demotischen Texte. Aber es ist kein digitalisiertes Textcorpus – und um solche geht es hier doch. Wenn ich aber aus der Perspektive des Nutzers konsequent weiterdenke, stoße ich schnell an eine ganz banale Grenze der DWL und anderer mir bekannter digitaler Corpora: Sie enthalten nur Wörter, die auch belegt sind. Das ist irgendwie logisch und selbstverständlich. Wenn man aber bedenkt, dass bisher erst ca. 19.000 verschiedene demotische Wörter oder genauer Wortschreibungen aus etwa 1000 Jahren bekannt sind (ohne Personennamen), während in hieroglyphischer und hieratischer Schrift deutlich mehr ägyptische Wörter, nämlich ca. 25.000, belegt sind (allerdings aus insgesamt ca. 3000 Jahren), wird sofort einsichtig, dass in noch unveröffentlichten demotischen Texten mit einer Vielzahl neuer Wörter zu rechnen ist. Und wer sich jemals an die Erstpublikation demotischer Texte gemacht hat, kann die gerade angestellte theoretische Erwartung aus der Praxis voll und ganz bestätigen: Pro Zeile ein bisher unbekanntes Wort ist für neue Textsorten eine durchaus realistische Zahl.

Um einen solchen Text trotz unserer mangelhaften Kenntnis des demotischen Wortschatzes zu verstehen, schaut man natürlich, ob das demotische Wort vielleicht einen hieroglyphischen oder hieratischen Vorläufer oder einen koptischen Nachfahren hat. Und oft gibt es tatsächlich wenigstens einen von beiden.

Die Suche kann mühsam sein und muss viele Alternativen berücksichtigen. Eine demotische Schreibung *wt* beispielsweise könnte auf älteres *wt, wṯ, wd, wḏ, wꜣt, wꜣṯ, wꜣd, wꜣḏ, wtỉ, wṯỉ, wdỉ, wḏỉ* und noch viel, viel mehr zurückgehen. Es stellt sich die Frage, ob man diese Sucharbeit nicht wenigstens ein wenig automatisieren und systema-

tisieren könnte. Man müsste alle diese Wörter unter einer quasi demotischen Form *wt* auffindbar machen, indem aus den älteren Wörtern unter Berücksichtigung bekannter Lautveränderungen die zu erwartenden demotischen Formen erzeugt würden. Diese vielleicht nicht wirklich demotisch belegten Schreibungen müsste man also in die DWL aufnehmen. Ausgangspunkt für eine derartige virtuelle Erweiterung der DWL kann eine Liste der hieroglyphisch und hieratisch belegten Wörter sein, die man ihrerseits gewissermaßen als ein elektronisches Corpus versteht, aus dem die demotischen Formen generiert werden. Das Ganze muss natürlich automatisch erfolgen; nicht von Hand, sonst wäre man lange beschäftigt.

Ich selbst habe auf der Grundlage von *Wörterbuch*,[4] MEEKS: *Année lexicographique*,[5] HANNIG[6] und vielen Texteditionen eine inzwischen 25.000 Einträge umfassende Wortliste zusammengetragen. Natürlich ist sie wie jede ägyptische Wortliste nie vollständig und fertig. Doch sie bietet einen praktikablen Ausgangspunkt.

Die Einträge sind denkbar einfach strukturiert. Umschrift, Übersetzung und Belegstelle (meist *Wörterbuch* oder eine andere Sekundärliteratur) stehen, von entsprechenden eineindeutigen Kennungen eingeschlossen, in einer Textdatei. Ich habe nun ein kleines TUSTEP-Programm[7] geschrieben, das die Umschrift nach bekannten ägyptischen Lautveränderungen hin zum Demotischen modifiziert und einen neuen, veränderten Eintrag anlegt. Z.B. fallen, wie wir gesehen haben, ganz häufig die Dentale zu *t* zusammen, Aleph im Wortinnern schwindet, in der Römerzeit werden *ȝ* und *ʿ* nicht mehr unterschieden, ebenso können *h* und *ḥ* füreinander eintreten, im Fajumischen erscheint *l* für *r* usw. Oder die Femininendung wird im Demotischen oft nicht geschrieben, dafür erscheint bei vielen weiblichen Wörtern ein *y* am Wortende. Ich mache nun nicht anderes, als dass ich rein mechanisch beispielsweise alle Wörter, die auf *.t* enden wie etwa *p.t* „Himmel“ auch unter der Form *p* (also ohne Femininendung) und unter der Form mit *y* anstelle der Femininendung abspeichere, also in diesem Fall *py*, und auf die Ausgangsform *p.t* verweise:

---

[4] ERMAN, A. & H. GRAPOW (Hrsg.), *Wörterbuch der aegyptischen Sprache*, 7 Bde., Berlin [4]1982.

[5] MEEKS, D., *Année lexicographique*, 3 Bde., Paris 1980-1982.

[6] HANNIG, R., *Die Sprache der Pharaonen. Großes Handwörterbuch Ägyptisch – Deutsch (2800–950 v. Chr.)*, Mainz 1995.

[7] Zum Tübinger System von Textverarbeitungs-Programmen (TUSTEP) siehe http://www.tustep.uni-tuebingen.de/.

```
,.woa p ,.woe --> py --> p.t ,.üba Himmel
,.woa p.t ,.üba Himmel
,.woa pA ,.woe --> py --> p.t ,.üba Himmel
,.woa py ,.woe --> p.t ,.üba Himmel
```

(`,.woa` und `,.woe` markieren das Lemma, auf das die Suchabfragen zugreifen; `,.üba` trennt das Übersetzungsfeld vom Vorangehenden ab.)

Fände man nun in einem demotischen Text die Schreibungen *p* oder *py*, würde man beim Nachschauen in meiner Liste fündig und auf die Ausgangsform *p.t* „Himmel“ verwiesen. Übrigens: Die drei Schreibungen *p.t*, *p* und *py* gibt es tatsächlich in demotischen Texten.[8]

Mein Programm berücksichtigt eine ganze Menge derartiger Lautveränderungen und ist jederzeit leicht erweiterbar. Ich habe auch daran gedacht, dass es in einem einzigen Wort zu mehreren Veränderungen kommen kann, z.B. zum Ersatz der Femininendung durch *y*, das seinerseits mit *ỉ* wechseln kann. Darum wird zu *p.t* auch die Form *pỉ* generiert. Ich möchte bemerken, dass ich diese Schreibung noch in keinem demotischen Text gesehen habe, dass aber das Himmelszeichen im hieroglyphischen Schriftsystem der griechisch-römischen Zeit durchaus auch als Schreibung für den Artikel *pỉ* dienen kann,[9] der freilich seinerseits da schon zu *p* geworden ist.

Programmiertechnisch gehe ich einfach iterativ vor: Die als Resultat einer Lautveränderung generierten Formen werden zu den Ausgangsdaten hinzukopiert und können alle weiteren Lautveränderungen mit durchlaufen.

Natürlich entstehen bei einem solchen rein mechanischen Vorgehen auch Dubletten und Formen, die man vermutlich niemals real belegt finden wird. Nicht jede Lautveränderung findet ja immer statt. Manchmal wird sie z.B. von der lautlichen Umgebung verhindert. Aber überschüssige automatisch erzeugte Formen sind aus meiner Sicht kein Schaden. Denn wir suchen ja in der Liste nur, wenn wir einer wirklichen Form in einem neuen Text begegnen. Wenn wir

---

[8] ERICHSEN, W., *Demotisches Glossar*, Kopenhagen 1954, 127.

[9] Z.B. Esna 356, 11, 22 und 24; 367, 20 und 24 (SAUNERON, S., *Le temple d'Esna*, Esna 3, Kairo 1968, 310, 311, 330 und 331) und öfter. Zu diesem sprachlich demotischen Text beachte QUACK, J. F., Das Monumental-Demotische, in: GESTERMANN, L. & H. STERNBERG-EL HOTABI (Hrsg.), *Per aspera ad astra. Wolfgang Schenkel zum neunundfünfzigsten Geburtstag*, Kassel 1995, 107-121, bes. 110 und 119.

diese bisher unbelegte Schreibung in der virtuell erweiterten DWL finden, hat diese ihren Zweck erfüllt.

Insgesamt werden aus dem 25.000 hieroglyphischen und hieratischen Einträgen an die 350.000 Formen generiert. Auch wenn davon nur 10 Prozent, d.h. 35.000, jemals in einem demotischen Text auftauchen sollten, hätte sich die Erweiterung der DWL um virtuelle Einträge auf jeden Fall gelohnt. Denn bisher sind dort ja erst ca. 19.000 tatsächlich belegte Formen versammelt. Auch wenn 10 Prozent noch zu optimistisch angesetzt sein sollten, ist mit der bequem nachschlagbaren Bereitstellung von vielen Hunderten bisher unbelegter demotischer Schreibungen und Wörter zu rechnen, da manche Teile des Wortschatzes im Demotischen noch deutlich unterrepräsentiert sind, etwa im Bereich der Religion.

Abschließend möchte ich betonen, dass selbstverständlich auch die sprachgeschichtlich umgekehrte Richtung möglich, sinnvoll und technisch ohne weiteres realisierbar wäre: Aus einer koptischen Wortliste könnte man ebenfalls virtuelle demotische Wörter ableiten.

KURSIVHIERATISCHE TEXTE AUS SPRACHLICHER UND ONOMASTISCHER SICHT[1]

GÜNTER VITTMANN

Im Unterschied zu dem Terminus „Demotisch“, der zwar primär eine Schriftart bezeichnet, darüber hinaus aber auch eine Sprachstufe, die überwiegend eben in demotischer Schrift geschrieben wurde, zielt der etwas pleonastische anmutende Ausdruck „Kursivhieratisch“ bzw. alternativ, aber auch nicht besser, „abnormhieratisch“ (abnormal hieratic, hiératique anomal)[2] ausschließlich auf eine bestimmte Entwicklungsform der hieratischen Schrift ab. Eine bestimmte Sprachform bzw. Sprachstufe ist damit also ebensowenig wie bei „Hieroglyphisch“ oder „Hieratisch“ impliziert. Im engeren Sinne wird mit „Kursivhieratisch“ die hauptsächlich, aber inzwischen keineswegs ausschließlich, in Theben bezeugte Geschäftsschrift der zwei Jahrhunderte etwa zwischen 750 und 550 v. Chr., als es durch das Demotische verdrängt wurde,[3] verstanden. Im weiteren Sinne kann man auch schon gewisse administrative Dokumente aus dem ersten Viertel des 1. Jahrtausends v. Chr. dazurechnen, da deren Schriftduktus oft schon chronologisch wie auch äußerlich in der Mitte zwischen der spätramessidischen Kursive und dem klassischen Kursivhieratisch der 25. und 26. Dynastie steht,[4] doch sind die von mir hier zitierten und ausgewerteten Quellen nahezu sämtlich in die zwei genannten Jahrhunderte datierbar. Anders als beim Demotischen ist die Zahl der publizierten(!) Dokumente relativ begrenzt – die Datenbank „Trismegistos“ enthält derzeit nicht mehr als 68 Einträge[5] – und somit ganz gut überschaubar, allerdings erhöht sich diese Zahl durch eine beträchtliche Reihe unveröffentlichter

---

[1] Um einige Beispiele erweiterte Fassung meines in Berlin gehaltenen Kurzreferats.

[2] Zur Forschungsgeschichte vgl. MALININE, M., L’hiératique anormal, in: *Textes et langages de l’Égypte pharaonique. Hommage à Jean-François Champollion* I, BdÉ 64/1, Le Caire 1972, 31-35.

[3] Vgl. DONKER VAN HEEL, K., The lost battle of Peteamonip son of Petehorresne, in: *Egitto e Vicino Oriente* 27, 1994, 115-124; MARTIN, C. J., The Saite ‘Demoticisation’ of Southern Egypt, in: LOMAS, K., *et al.* (ed.), *Literacy and the State in the Ancient Mediterranean*, London 2007, 25-38.

[4] Vgl. MALININE, L’hiératique anormal; VLEEMING, S. P., *Papyrus Reinhardt. An Egyptian Landlist from the Tenth Century B.C.*, Berlin 1993, und hier 78-80 („Appendix II. Survey of Related Texts from the Third Intermediate Period“).

[5] Letzter Zugriff 8. Juni 2012.

Dokumente, die teils unbeachtet in verschiedenen Sammlungen schlummern, teils überhaupt erst in den letzten Jahren ans Licht gekommen sind. Auch manche der schon publizierten Quellen bedürften dringend einer verbesserten Neubearbeitung.

Die Benennung dieses Kurzvortrags „Kursivhieratische Texte aus sprachlicher und onomastischer Sicht" soll andeuten, dass im Rahmen dieser Veranstaltung der Schwerpunkt nicht, wie es an sich beim Kursivhieratischen naheläge, auf paläographischen Aspekten liegt – hier bleibt für künftige Arbeit noch viel zu tun[6] –, sondern eben auf sprachlichen (und dass auch Personennamen eine wichtige Quelle für die Lexikographie sein können, braucht nicht eigens betont zu werden). Bedauerlicherweise hat das Kursivhieratische schon immer nur äußerst wenige Adepten gefunden, die dann nicht zufällig zumeist auch Demotisten waren bzw. sind, was vermutlich nicht zuletzt damit zusammenhängt, dass die Inhalte von Verwaltungsdokumenten für die meisten Ägyptologen nicht „interessant" genug sind, um sich mit den Tücken der Schrift herumzuschlagen. So ist es vielleicht nicht verwunderlich, dass kursivhieratische Quellen nicht allzu oft, und noch seltener kritisch, ausgewertet werden. Ebensowenig überrascht andererseits, dass gerade ein verhältnismäßig umfangreicher literarischer Text, der unerwarteterweise tatsächlich in klassischem Kursivhieratisch niedergeschrieben wurde, von einem Nichtdemotisten, aber Spezialisten in altägyptischer Literatur und ausgezeichneten Kenner des Hieratischen, nämlich Hans-Werner Fischer-Elfert, ediert werden wird.[7]

An dem eben genannten Papyrus, dessen Veröffentlichung in absehbarer Zeit abgeschlossen werden soll, zeigt sich besonders schön, dass eine verstärkte Berücksichtigung kursivhieratischen Materials auch für Nichthieratisten im Fach sinnvoll und wünschenswert ist, z.B. eben – von der Inhaltsseite ganz zu schweigen – unter lexikographischen und sprachgeschichtlichen Gesichtspunkten. Ich will der Publikation hier nicht allzu weit vorgreifen, abgesehen davon wäre

---

[6] Verf. beabsichtigt, entsprechende Arbeiten in den nächsten Jahren stärker voranzutreiben. Ein *Abnormal Hieratic Reading Book* von K. DONKER VAN HEEL mit Paläographie von JOOST GOLVERDINGEN ist in Vorbereitung und soll demnächst im Netz veröffentlicht werden.

[7] pQueen's College (vgl. BAINES, J. *et al.*, Abnormal Hieratic in Oxford: Two New Papyri, in: *JEA* 84 (1998), 234-236). Die weniger erbaulichen, dafür nicht eben leichter zu lesenden Abrechnungen auf der Rückseite, die vom Referenten bearbeitet werden, können hier außer Acht bleiben. Hans-Werner Fischer-Elfert arbeitet übrigens auch an der Publikation der kursivhieratischen Papyri aus Qasr Ibrim.

das Anführen zu vieler Details in diesem Rahmen zu speziell, darum nur Weniges: Grammatisch und syntaktisch steht der Text, wie zu erwarten, der „normalhieratisch" überlieferten Erzählung des im TLA von Lutz Popko aufgenommenen pVandier[8] und damit dem Demotischen als Sprachstufe sehr nahe, die Handschrift stammt aber aus der 25. Dynastie, also einer Zeit, da sich die demotische Schrift noch nicht herausgebildet hat. Die nachfolgende, auch den pQueen's College berücksichtigende Auswahl mit lexikalischen, grammatischen und (einigen wenigen ausgewählten) onomastischen Besonderheiten kursivhieratischer Texte soll demonstrieren, dass die sprachliche Auswertung solcher Quellen umso wichtiger ist, als aus dieser Zeit – also grob zwischen 750 und 550 – nur wenig Material in derselben Sprachstufe, aber anderen Schriftformen (hieroglyphisch, „normalhieratisch" und demotisch), erhalten ist.[9]

1. Lexikalisches

ỉb „Herz" ist kursivhieratisch häufig in der Urkundenformel *m ỉb hr(=j)/n* „zu meiner/unserer Zufriedenheit" u.ä. belegt.[10] Demotische Verkaufsurkunden sowie die spätesten, sprachlich schon stark

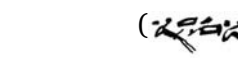
---

[8] Letzte Übersetzung AGUT-LABORDÈRE, D. & M. CHAUVEAU, *Héros, magiciens et sages oubliés de l'Égypte ancienne. Une anthologie de la littérature en égyptien démotique*, Paris 2011, 3-11 und 323-325, mit Literatur.

[9] Von herausragender Wichtigkeit sind hier die Königsrede auf der Kleinen Sandsteinstele des Pianchi vom Gebel Barkal (JANSEN-WINKELN, K., *Die Inschriften der Spätzeit*, II: *Die 22.-24. Dynastie*, Wiesbaden 2007, 350-351) der hieratische Brooklyner Weisheitspapyrus (vgl. zuletzt AGUT-LABORDÈRE & CHAUVEAU, *Héros, magiciens et sages*, 213-221 und 343-344, mit Literatur), die von VERNUS, P., Inscriptions de la Troisième Période Intermédiaire (I), in: *BIFAO* 75, 1975, 1-66, hier 26-66 edierte und bearbeitete Inschrift des Taharka aus Karnak (Text jetzt auch bei JANSEN-WINKELN, K., *Inschriften der Spätzeit*, III: *Die 25. Dynastie*, Wiesbaden 2009, 84-87 [Text 48.33]) sowie die „protodemotische" Übersetzung des in Urk. VI veröffentlichten Vernichtungsrituals, vgl. VERNUS, P., Entre néo-égyptien et démotique: La langue utilisée dans la traduction du Rituel de Repousser l'Agressif (Étude sur la diglossie I), in: *RdÉ* 41, 1990, 153-208; ALTMANN, V., *Die Kultfrevel des Seth. Die Gefährdung der göttlichen Ordnung in zwei Vernichtungsritualen der ägyptischen Spätzeit (Urk. VI)*, Studien zur spätägyptischen Religion 1, Wiesbaden 2010.

[10] *m ỉb hr* bzw. *m ỉb hr(=j)*: pBM 10800, 4 (EDWARDS, I. E. S., Bill of Sale for a Set of Ushebtis, in: *JEA* 57, 1971, 120-124); pLouvre E 3228e, 4 (MALININE, M., *Choix de textes juridiques en hiératique « anormal » et en démotique (XXV^e – XXVII^e dynasties)*, I, Paris 1953, 36-37; II, RAPH 18, Le Caire 1983, 14 und pl. V); pWien D 12002, I 6 (VITTMANN, G., Nochmals der kursivhieratische Papyrus Wien D 12002, in: *GM* 154, 1996, 103-112); pTurin 2118, 65 (MALININE, *Choix* I, 64-65 [irrig *m-ỉb-hr=n*]; II, 29 [korrigiert]); pTurin 2120, 9 (MALININE, *Choix* I, 72-73; II, 34); *m ỉb hr=n* pTurin 2118, 12. 13. 29 etc. (MALININE, *Choix* I, 58-67; II, 23-32).

vom Demotischen beeinflussten kursivhieratischen Urkunden formulieren *dj=k mtr ḥ3tj=j/n* „du hast mein/unser Herz zufrieden gestellt“[11]; das normale Wort für „Herz“ ist im Demotischen *ḥ3tj* (ϩнт), während sich *ỉb* nur in einigen wenigen Verbindungen erhalten hat.[12]

ỉprṯ dšr.t „roter Typ“ (o.ä., von einer Kuh) pWien D 12002, I 4[13] (); *jprṯ* ist ein Hapax von unklarer Etymologie.

ʿbšnn „Frosch“ pQueen’s College, x+II 8 (), zweifellos mit dem im pEbers sowie in späten Tempelinschriften belegten *ʿbḫn* (Var. *ʿbnḫ*) Wb I, 178, 15-17 zu identifizieren.

bnbn(?), ein technischer Ausdruck, der häufig in den kursivhieratischen Ostraka aus Mut (Dachla) in der Verbindung *p3 bnbn*(?) *rm k3m* NN „das ..?.. mit dem Winzer NN“ erscheint. Ein typisches Beispiel zeigt Abb. 4[14] (oMut 38/7, hierin Z. 1). Eine Lesung *bnbn* für und andere gleichartig geschriebenen Belege wie oMut 38/24, 1 und 38/130, 1 ist sehr wahrscheinlich, doch lässt sich – ebensowenig wie für eine an sich denkbare Lesung *ḏnḏn* – ein Anschluss finden.

m3ʿ als Terminus technicus für Opferwein ist überaus häufig in den eben genannten Ostraka belegt, und zwar immer in abgekürzter Schreibung mit der Feder () und gelegentlich mit dem Krug-determinativ versehen wie in oMut 38/7, 2 (, Abb. 4). Ansonsten ist mir diese Verwendung von *m3ʿ* lediglich aus Krugetiketten aus Amarna bekannt.[15]

---

[11] Vgl. hierzu DONKER VAN HEEL, K., *Abnormal Hieratic and Early Demotic Texts Collected by the Theban Choachytes in the Reign of Amasis*, Diss. Leiden 1995, 79 (V).

[12] Z.B. *ỉb-n-Rʿ* „Herz des Re“ als Bezeichnung des Thot, vgl. LGG I 208-209; *ỉb-ls* „Herz-Zunge“, QUACK, J. F., Korrekturvorschläge zu einigen demotischen literarischen Texten, in: *Enchoria* 21, 1994, 63-72, hier 70 (24); *ḥr-ỉb* „Mitte“; „inmitten von“ ERICHSEN, *Glossar* 321; *swḏ3 ỉb n ỉt(=f)* „der das Herz (seines) Vaters erfreut“ Titel der Priester von Teudjoi, vgl. GRIFFITH, F. LL., *Catalogue of the Demotic Papyri in the John Rylands Library Manchester*, III, Manchester / London 1909, 429.

[13] VITTMANN, Wien D 12002, bes. 108 ad loc.

[14] Diese Ostraka wurden von der australischen Mission unter Colin Hope im Januar/Februar 2011 im Bezirk des Seth-Tempels von Mut entdeckt. Im Februar 2012 – also erst nach der Berliner Tagung – hatte ich Gelegenheit, diese neuen Funde im Magazin von Ismant kennenzulernen. Ich habe darüber inzwischen bei der 7. Internationalen Konferenz des Dakhleh Oasis Project (Leiden, 20.-24. Juni 2012) berichtet.

[15] Vgl. WAHLBERG, E.-L., *The Wine Jars Speak: A text study*, Uppsala 2012, 41. 115 (Amarna 143; 146)

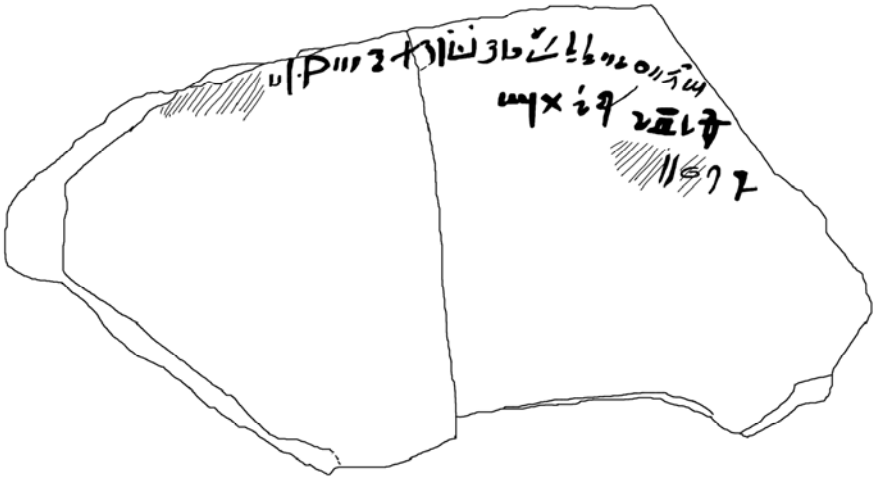


Abb. 1: Ostrakon aus dem Areal des Seth-Tempels von Mut / Dachla (Nr. 38/7; nach eigenem Photo).

Von den beiden Titeln *mnjw-ʿnḫ* „Ziegenhirt“ pBM 10907, 2. 6-7 und *wrš-ʿnḫ* pBM 10906, 2. 3; 10907, 3. Vso 1[16] „Ziegenwächter“ ist ersterer auch neuägyptisch belegt[17], letzterer jedoch neu.

Der Ausdruck *mtr-sḫ* „Zeugenschreiber“ bezeichnet in kursivhieratischen Urkunden den Schreiber des betreffenden Dokuments. Er findet sich zuerst in dem paläographisch noch nicht den klassischen Typ repräsentierenden pBM 10800, 10[18], aber auch in der etwas jüngeren „steinhieratischen“ Kleinen Dachlastele (Pianchi, Jahr 24), 16.[19] Die jüngsten kursivhieratischen Texte, die bereits

(http://uu.diva-portal.org/smash/record.jsf?pid=diva2:528049).

[16] DONKER VAN HEEL, K., A day in the life of the ancient Egyptian goatherd Ityaa: abnormal hieratic P. Michaelides 1 and 2 (P. BM EA 10907 and 10906), in: *JEA* 90, 2004, 153-166.

[17] LESKO, L., *A Dictionary of Late Egyptian*, I, Providence 1982, 218.

[18] EDWARDS, Bill of Sale. Auf welchen König sich die Datierung „Jahr 14“ bezieht, lässt sich nicht sagen.

[19] JANSSEN. J. J., The Smaller Dâkhla Stela (Ashmolean Museum no. 1894. 107 b), in: *JEA* 54, 1968, 165-172.

stark unter dem Einfluss des demotischen Formular stehen, kennen diesen Begriff nicht mehr.

rḫ „Narr“ pQueen’s College, x+IV 12 (), bisher nur demotisch seit der Ptolemäerzeit, vor allem aus der Weisheitsliteratur, bekannt (*lḫ*[20]).

ḥm „Diener, Sklave“ ist kursivhieratisch nicht nur in Titeln und Personennamen (*Ḥm-ḫnsw*[21], *Ḥm-nꜣ-nf*[22]) belegt, sondern auch selbständig in den Ausdrücken *ḥmw ḥmy* „Sklaven und Sklavinnen“ pBM 10800, 3 ( , von Uschebtis);[23] *ḥm.w šr.w* „sclaves-enfants“ pTurin 2121, 7 ();[24] *ḥm ḥm.t* „Sklave, Sklavin“ pBM 10113, 6 ( , 570 v. Chr.)[25] in einer Aufzählung von Garantien (hinter *pr* „Haus“ und vor *šr šr.t* „Sohn, Tochter“. Im Demotischen ist *ḥm* (abgesehen von bestimmten Titeln, vor allem natürlich *ḥm-nṯr* > ϩⲟⲛⲧ) zugunsten von *bꜣk* völlig außer Gebrauch gekommen.

sḫm.t Jmn „Amunsfrau“ pKairo CG 30886, 2 (),[26] ein recht merkwürdiger Titel, der mir sonst nur von einer Schenkungsstele der 22. Dynastie bekannt ist, wo er sich auf eine Frau namens *Tꜣ-nt-mrḳwrs* bezieht.[27]

sḏd Kairo JE 94478 (): In dieser Schülertafel aus einem Kuschitengrab aus dem Asasif steht auf der einen Seite in

---

[20] ERICHSEN, *Glossar* 263; Chicago Demotic Dictionary *l*, 15 s.v. *lḫ* (online-Version 29.06.2001).

[21] pTurin 2121, 2. 3, s. MALININE, *Choix* I, 118-119; II, 53.

[22] pQueen’s College, x+III 14; x+IV 16. In diesem literarischen Text heißt so der Fürst von Athribis in bewusster Abänderung des historisch belegten *Bꜣk-n-nf*; vgl. VITTMANN, G., Zur Familie der Fürsten von Athribis in der Spätzeit, in: *SAK* 10, 1983, 333-339.

[23] EDWARDS, Bill of a Sale.

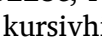
[24] MALININE, *Choix* I, 118-119; II, 54. Facsimile nach Photo des Museums.

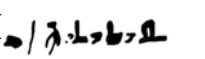
[25] MALININE, *Choix* I, 16-17; II, 6; DONKER VAN HEEL, *Abnormal Hieratic and Early Demotic Texts* 231 [Text 23], mit pl. 30/30A.

[26] SPIEGELBERG, W., *Die demotischen Denkmäler*, II: *Die demotischen Papyrus*, Straßburg 1906-1908, Taf. 67 (im Textband S. 194 kurz beschrieben, aber nicht bearbeitet); kollationiert mit Photo und Original. In Z. 4 erscheint der Titel nochmals; es folgt dort *Wsir* (ebenfalls mit Gottesdeterminativ). Auch wenn mir die Fortsetzung in Z. 2 und 4 unklar ist, zeigt die Verwendung des Gottesdeterminativs bei *ꜥImn*, dass es sich – entsprechend einer für das Kursivhieratische geltenden Regel – um keinen Teil eines PN handelt.

[27] JANSEN-WINKELN, K., *Inschriften der Spätzeit* II: *Die 22.-24. Dynastie*, Wiesbaden 2007, 257 (Text 26.7); vgl. dazu KOCH, C., *„Die den Amun mit ihrer Stimme zufriedenstellen“. Gottesgemahlinnen und Musikerinnen im thebanischen Amunstaat von der 22. bis zur 26. Dynastie*, Studien zu den Ritualszenen altägyptischer Tempel 27, Dettelbach 2012, 82.

normalem Hieratisch der Anfang der Lehre des Cheti, auf der anderen in Kursivhieratisch *nꜣ sḏd Ḏḥwtj-ỉw=f-ʿnḫ* etc. „die Worte / Aussprüche / Erzählungen des Djedthotiufanch“.[28] *sḏd* ist quasi als demotischer Gattungsbegriff für „Erzählung“ bestimmt worden;[29] ob genau dies auch auf die Schülertafel zutrifft, lässt sich wegen der Kürze des Texts nicht sagen, ist aber natürlich gut möglich.

qnb.t hat seine eigentliche Bedeutung „Gericht, Gerichtshof“ im Ausdruck *tꜣ qnb.t ꜥ.t n Nw.t* „der große Gerichtshof von Theben“ pLouvre E 3228c, I 5. 10 ( [30] ).[31] Im selben Dokument und in anderen kursivhieratischen Texten ist dagegen – in Verbindung mit *ỉr* bzw. *ḏd* – nur die spezielle Bedeutung „prozessieren“ üblich;[32] ebenso im Demotischen, wo *qnb* auch die „Gerichtsurkunde“ bezeichnet, aber keine Institution mehr.

qḏwḏ „Sklave aus Gaza“ > „Sklave /Diener; Junge, Bursch“ pBM 10906, 6[33]; pLouvre E 3228e, 3 (). 12 (). 22 ()[34].

---

[28] VITTMANN, G., Eine spätzeitliche Schülertafel, in: *Ägypten und Levante* 16, 2007, 187-193; zum Fundzusammenhang s. BUDKA, J., *Bestattungsbrauchtum und Friedhofsstruktur im Asasif. Eine Untersuchung der spätzeitlichen Befunde anhand der Ergebnisse der österreichischen Ausgrabungen in den Jahren 1969-1977*, Untersuchungen der Zweigstelle Kairo des Österreichischen Archäologischen Instituts 34, Wien 2010, 595-596.

[29] Vgl. Diskussion bei VITTMANN, Schülertafel, 190-191 (b).

[30] Facsimile nach pQueen's College, x+IV 6; an den anderen Stellen aber sehr ähnlich geschrieben.

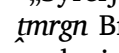
[31] MALININE, M., Un jugement rendu à Thèbes sous la XXVe dynastie (pap. Louvre E. 3228c), in: *RdÉ* 6, 1951, 157-178; neuere Übersetzung KAPLONY-HECKEL, U., in: *TUAT* I, 227-230.

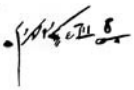
[32] Mit *ỉr* pLouvre E 3228c, I 5 (MALININE, Jugement); P. Wien D 12003, I 6 (MALININE, M., Une affaire concernant un partage (pap. Vienne D 12003 et D 12004), in: *RdÉ* 25, 1973, 192-208); mit *ḏd* pQueen's College, x+IV 6; pBM 10113, 7 (MALININE, *Choix* I, 16-17; II, 6; DONKER VAN HEEL, *Abnormal Hieratic and Early Demotic Texts* 231, mit pl. XXX/XXXA); pLouvre E 7845 B, 10 DONKER VAN HEEL, *Abnormal Hieratic and Early Demotic Texts* 117, mit pl. VII/VIIA) (beide Male in der im Demotischen geläufigen Klausel *ỉwtj ḏd qnb.t nb*).

[33] DONKER VAN HEEL, K., Goatherd.

[34] MALININE, *Choix* I, 36-39; II, 14-16 und pl. V.

Als Personenname ist *Gḏḏ* u.ä. seit dem Neuen Reich[35] belegt als normales Substantiv abgesehen von den kursivhieratischen Stellen, jedoch m.W. nur in Demot. Chronik VI 17, wo *rš̈j=tn n gḏwḏ ỉw=tn (r) gm r wnm* von Quack[36] „Freut euch, ihr Burschen, ihr sollt zu essen finden“ übersetzt wird. Tatsächlich wird das Substantiv in der unmittelbar folgenden Erklärung durch *ḫm-ḫl* „Junge, Diener“ (< „Syrerjunge“) wiedergegeben.

ṯmrgn Brief pKairo *CG* 30865, 6 ();[37] derselbe etymologisch unklare Ausdruck im Moskauer literarischen Brief, V 5 zur Bezeichnung einer Art Krieger. David Klotz[38] stellt dies zu *tmrhtnt* in der bekannten Stele des *M-ḥb* (Emhab) und zu Temüjin / Temürjin, dem etymologisch „Schmied“ bedeutenden Geburtsnamen von Dschingis Khan.[39] Bei dem enormen zeitlichen und geographischen Abstand ist natürlich große Vorsicht geboten, auch wenn es sich grundsätzlich um eine Art Wanderwort handeln könnte.

ṯk{g}r pr-ꜥ „Eunuch/Kastrat des Pharaos“ Brief pKairo CG 30865, 1 ();[40] vgl. *ṯkr pr-ꜥ* Stele Wien 165 ()[41]; demot. *ṯkr* pRylands 9, XVI 17 (mit „Phallus“-Determinativ).

---

[35] PN I 429, 21; II 404; ein Beleg auf einer Statue der 25. Dynastie bei JANSEN-WINKELN, K., *Inschriften der Spätzeit*, III, 293 (Text 51.50); *Demot. Nb.* 1014; akkadisch als [1]*Ga-ṣu-ṣu* (VITTMANN, G., Zu einigen keilschriftlichen Umschreibungen ägyptischer Personennamen, in: *GM* 70, 1984, 65-66); kopt. ⲕⲁϧⲱϧ; griech. Καθύτης.

[36] HOFFMANN & QUACK, *Anthologie*, 191.

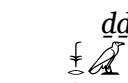
[37] VITTMANN, G., Ein kursivhieratisches Brieffragment (P. Kairo CG 30865), in: *Enchoria* 27, 2001, 155-163; zum Titel 159 (q).

[38] KLOTZ, D., Emhab versus the *tmrhtn*: Monomachy and the Expulsion of the Hyksos, in: *SAK* 39, 2010, 211-241; zu *tmrgn* 225-227.

[39] Türk. *demirci*; mongol. *temürci(n)* „Schmied“; zu alttürk. *tämir* „Eisen“ mit Wortbildungssuffix *–či*. Vgl. RYBATZKI, V., *Die Personennamen und Titel der mittelmongolischen Dokumente. Eine lexikalische Untersuchung*, Diss. Helsinki 2006 (http://ethesis.helsinki.fi/julkaisut/hum/aasia/vk/rybatzki/), 390.

[40] VITTMANN, Brieffragment; zum Titel 157-158 (d); VITTMANN, G., *Der demotische Papyrus Rylands 9*, ÄAT 38, Wiesbaden 1998, 527-531.

[41] Die Publikation durch H. SATZINGER in einem Band über die Spätzeitstelen des Wiener Kunsthistorischen Museums ist im Druck.

ḏḏw „Gerichtskollegium“ pQueen’s College, x+IV 8. 15 ( [42]), das offensichtlich auf *ḏ3ḏ3.t* zurückgeht und einen Genuswandel durchgemacht hat, im Demotischen und erst recht im Koptischen aber nicht mehr gebräuchlich ist. In einem der 2011 in Mut / Dachla entdeckten kursivhieratischen Ostraka (38/144, 3) konnte ich kürzlich ein weiteres Beispiel für einen Genuswandel[43] identifizieren: *ḥq.t* „Bier“ wird dort wie sein koptisches Derivat ϩⲛⲕⲉ als Maskulinum behandelt, was bisher erst seit der Ptolemäerzeit belegt war[44].

2. Bemerkenswertes aus Grammatik und Syntax

Zunächst ist noch einmal darauf aufmerksam zu machen, dass hier nur einige spezifische Eigenheiten, wie sie in kursivhieratischen Texten zu finden sind, angeführt und besprochen werden. Es ist also keineswegs meine Absicht, eine umfassende Darstellung von Grammatik und Syntax in Vergleich mit dem Neuägyptischen zu geben, es geht mir auch in diesem Abschnitt in erster Linie darum, exemplarisch die Bedeutung der kursivhieratischen Texte für die Erforschung der ägytischen Sprache und Sprachgeschichte aufzuzeigen.

Ein neuägyptisches Relikt, das sich bis in die 3. Zwischenzeit erhalten hat,[45] im Demotischen aber nicht mehr eindeutig nachzuweisen und bis dahin wohl aus der Sprache verschwunden ist,[46] ist das narrative *ỉw=f (ḥr) sḏm* („non-initial main sentence /NIMS“) im pQueen’s College. Die Analyse der hierfür in Betracht kommenden Stellen muss der Publikation vorbehalten bleiben; explizit zitiert sei hier lediglich x+IV 13, wo auf eine Kette von drei durch *ỉm dj=f* eingeleiteten Imperativen die Erzählung durch *ỉw=w dj.t t3 700(?)*

---

[42] Zum Determinativ vgl. die Beispiele bei LESKO, L., *A Dictionary of Late Egyptian*, IV, Providence 1989, 152.

[43] Zum Thema vgl. BRUNSCH, W., Zum vermeintlichen Genuswandel im Koptischen durch den Einfluß des Griechischen, in: *ZÄS* 110, 1983, 122-126.

[44] pLoeb 5, 44 (= Vso 18) *p3j ḥq.t* (das von SPIEGELBERG, W., *Die demotischen Papyri Loeb*, München 1931, 15, hinzugesetzte Fragezeichen ist entbehrlich). Im Übrigen haben die derzeit in der Demotischen Datenbank enthaltenen Belege für *ḥq.t* entweder den Pluralartikel (so auch pVandier, I 4) oder überhaupt keinen Artikel, so dass das Genus des Wortes nicht ersichtlich ist.

[45] Vgl. VERNUS, Entre néo-égyptien et démotique, 182-183.

[46] Positiv äußerte sich SHISHA-HALEVY, A., Papyrus Vandier *Recto*: An Early Demotic Literary Text?, in: *JAOS* 109, 1989, 421-435, hier 424 (b); VERNUS, Entre néo-égyptien et démotique, 183 unten.

dbn ḥḏ n ʾIḫj „und sie gaben dem Ihi 700(?) Silberdeben“ weitergeführt wird.

Die Hervorhebungs- und Konditionalpartikel *ỉr* ist im Kursivhieratischen geläufig, z.B. in der Urkundenklausel *ỉr pȝ ntj ỉw=f (r) md.t n.ỉm=w* „Was den betrifft, der dagegen Einspruch erheben wird“ pTurin 2118, 33[47] (ähnlich pWien D 12002, I 12); man findet sie sogar noch in der Endphase des Kursivhieratischen in pLouvre E 7846, 4[48] *ỉr ỉw(=ỉ) ḫȝʿ sḫm.t X* „Wenn ich Frau X entlasse“ (549 v. Chr.). Das Demotische kennt *ỉr* als solches nicht mehr: Ein Satzteil wie der zu Beginn dieses Abschnitts zitierte müsste demotisch gleichlautend, aber eben ohne *ỉr*, formuliert werden, während anstelle von *ỉr ỉw(=ỉ) ḫȝʿ* demotisch – lautlich vermutlich aus der neuägyptischen Form mit Abfall des auslautenden *r* hervorgegangen – der Konditionalis *ỉw=j sḏm* geworden ist.[49] Kursivhieratischem *ỉr ỉw=j tm dj.t s n=k* „Wenn ich es dir nicht gebe“ pBM 10113, 3-4[50] (570 v. Chr.) steht folglich – um nur ein einziges von vielen analog konstruierten Beispielen zu wählen – demotisch *ỉw(=j) tm dj.t st n=f* „Wenn ich sie (Pl.) ihm nicht gebe“ pBerlin P 3110, 7[51] gegenüber.

ptr sw „siehe“ ist in der ersten Hälfte des Jahrtausends als „auxiliaire d’énoncé“ kursivhieratisch und im pVandier wiederholt bezeugt.[52] Demotisch ist *ptr sw* nicht mehr gebräuchlich; es ist aber angenommen worden, dass sich die funktional ähnliche demotische Partikel tws auch lautlich daraus entwickelt hat.[53]

Die neuägyptische disjunktive Partikel *m-rȝ-pw* ( < mäg. *rȝ-pw*) ist dem Demotischen verlorengegangen; hier wird in dieser Funktion – sofern überhaupt eine Bezeichnung nötig erscheint – *gr* (ERICHSEN, *Glossar* 582-583) gebraucht. Das Kursivhieratische kennt

---

[47] MALININE, *Choix* I, 60-61; II, 26.

[48] MALININE, M., Transcriptions hiéroglyphiques de quatre textes du Musée du Louvre écrits en hiératique anormal, in: *RdÉ* 34, 1982/83, 93-100, hier 99 und pl. 7.

[49] Zum (positiven) Konditionalsatz im Demotischen und seiner sprachgeschichtlichen Stellung vgl. JOHNSON, J. H., *The Demotic Verbal System*, SAOC 38, Chicago 1976, 233-260.

[50] MALININE, *Choix* I, 16-17; II, 5; DONKER VAN HEEL, *Abnormal Hieratic and Early Demotic Texts*, 230, mit pl. XXX/XXXA.

[51] MALININE, *Choix* I, 32-33; II, 13.

[52] Den von VERNUS, Entre néo-égyptien et démotique, 199-200 § 24 genannten Beispielen sind pQueen’s College, x+II 4 und öfter; pDuke Library 648, 2. 12 (unpubl.); oMut (Dachla) 38/140, 7 (unpubl.) hinzuzufügen.

[53] Vgl. VITTMANN, *Papyrus Rylands 9*, 274-279 mit Literatur. Parallelen sind pKairo *CG* 30907, 6 und pLouvre E 7849, 6; vgl. LÜDDECKENS, E., *Ägyptische Eheverträge*, ÄA 1, Wiesbaden 1960, Urk. 3-4.

jedoch *m-r3-pw* noch in einer Klausel der Eheurkunden, wo es sich bis zum Ende dieser Schrift (und der Auflösung der mit Kursivhieratischen verknüpften Rechtstraditionen) erhalten hat: In pLouvre E 7846, 5 aus dem Jahr 549 v. Chr. heißt es in Fortführung des Konditionalsatzes *ir iw(=j) ḫ3ʿ sḫm.t* NN etc. „Wenn ich Frau NN entlasse“ etc., *m-r3-pw mr(=j) k.t-ḫ.t sḫm.t i.r=s* „oder (wenn) ich eine andere Frau als sie will“.[54]

Die spätneuägyptische Relativkonstruktion *ntj iw=f m-b3ḥ N* (Brief,) „der bestimmt ist für N“,[55] ist kursivhieratisch noch gut belegt,[56] im Demotischen aber nicht mehr üblich.

tw (geschrieben ∫, was man am besten nach demotistischem Usus als ß umschreibt) als Personalpronomen („man“) und zur Passivbildung erscheint in den Formeln *bn sḏm.t̂ (= sḏm.tw) r3=f m ḫ3 nb n sḫ* „seine Aussage soll nicht in irgendeinem Archiv gehört werden“ pLouvre E 3228c, I 24[57]; pLouvre E 3228e, 9[58]; Var. *bn sḏm.t̂ r3=w m s.t nb n sḫ* pTurin 2118, 33[59], und *iw=tw ir mj-qd tp n dj.t st* „Man wird handeln entsprechend der (festgesetzten) Art, sie zu geben (d.h. dementsprechend zahlen)“ pLouvre E 3228e, 4[60]; pTurin 2118, 16-17.[61] Statt Verwendung des obsoleten *tw/t̂* ist auch einmal die aktive Umformung *bn sn sḏm r3=f n ḫ3 nb n sḫ* pLeiden 1942/5.15, 8[62] belegt, während die singuläre Formulierung *bn sḏm r3=f m s.t nb n sḫ* pWien D 12003, I 12[63] entweder als Fehler bzw. Ungenauigkeit oder

---

[54] DONKER VAN HEEL, *Abnormal Hieratic and Early Demotic Texts*, 127 und pl. IX/IXA.

[55] Vgl. WINAND, J., *Études de néo-égyptien*, I. *La morphologie verbale*, Aegyptiaca Leodiensia 2, Liège 1992, 433-434, § 672.

[56] pBerlin P 3048 Verso, Texte 6, 3; 10, 1; 16 (unpubl.); pDuke Library 648, 1 (unpubl.); pKairo *CG* 30865, 1 (VITTMANN, Brieffragment); Holztafel Leiden I 431, 1 (ČERNÝ, J., The Abnormal-Hieratic Tablet Leiden I, 431, in: *Studies Presented to F. Ll. Griffith*, London 1932, 46-56). Unsicher ist pBrooklyn 37.1799 E, 1 (JASNOW, R. & G. VITTMANN, An Abnormal Hieratic Letter to the Dead (P. Brooklyn 37.1799 E), in: *Enchoria* 19/20, 1992/93, 23-43), da das angebliche *ntj* in *iw=s m-b3ḥ* dem Aussehen nach eher ein ꜥ und dann Teil des vorangehenden Namens sein dürfte (*P3-dj-p3-nb-p3-ꜥ*; Hinweis Koenraad Donker van Heel).

[57] MALININE, Jugement, 160 und pl. II.

[58] MALININE, *Choix* I, 36-37; II, 15 und pl. V.

[59] MALININE, *Choix* I, 60-61 (mit irriger Umschrift *ḫ3* statt *sḫ*); II, 26 (korrigiert).

[60] MALININE, *Choix* I, 36-37; II, 14 und pl. V; speziell zu *tp n* + Infinitiv 39-40 (11); vgl. auch die Hinweise bei MALININE, Partage, 206 (o).

[61] MALININE, *Choix* I, 58-59; II, 24.

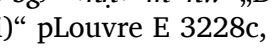
[62] VLEEMING, S. P., The Sale of a Slave in the Time of Pharaoh *Py*, in: *OMRO* 61, 1980, 1-17.

[63] MALININE, Partage.

aber, wie dies noch im pVandier belegt ist, tatsächlich als Relikt des neuägyptischen endungslosen Passivs.[64] Im Demotischen ist *tw/ṯ* mit seltenen relikthaften Ausnahmen – vor allem *ḏd.ṯ* zur Nennung des Beinamens – weitgehend verschwunden.[65]

3. Bemerkenswertes aus der Onomastik

Nur ganz wenige spezielle Fälle sollen hier vorgestellt werden:

P3-bg3-<nḥ>m-ʿnw „Der ... der (Göttin) Nehem-an (= Nehmet-awai)“ pLouvre E 3228c, I 18 ()[66]; sonst nicht belegt.

P(3)-n-smn-ỉmn Louvre E 3228d, 2 (). 16. 22; in Z. 11 fehlerhaft *Smn-ỉmn*.[67] Die Schreibung von *smn* wie „befestigen“ legt eine Analyse als *P3-ỉ.smn-ỉmn* „Der, den Amun dauern ließ“ nahe, stünde dann ähnlich wie demot. *p3j*[68] für *p3* mit folgendem Augment der Relativform.

P3-ḥrj-sḏm „Der hörende (göttliche) Herr“ oder „Der Herr hört“, pTurin 2118, 10 ()[69] und in den folgenden unpublizierten Quellen: pAshmolean 1998.3, Fr. 1+2, 3; Kruginschrift aus Gurna, II 21; oMut (Dachla) 151, Konvexseite, 5. Hieroglyphisch und demotisch nicht belegt.

---

[64] QUACK, J. F., Notes en marge du papyrus Vandier, in: *RdÉ* 46, 1995, 163-170, hier 168 (zu 5,16).

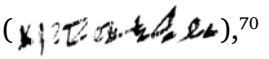
[65] Vgl. QUACK, J. F., [Rez. zu VOS, *Embalming Ritual*,] in: *Enchoria* 21, 1994, 186-191, hier 190 (zu vs. I,5); VITTMANN, *Papyrus Rylands 9*, 496-497 mit weiterer Literatur. In religiösen Texten in demotischer Schrift, aber nicht ausschließlich demotischer Sprache kommt das *tw*-Passiv auch sonst noch gelegentlich vor. Alle bis dato eingearbeiteten Belege sind in der Demotischen Textdatenbank einfach über die Wortsuche unter dem Eintrag *.ṯ* (als Passivendung) abrufbar.

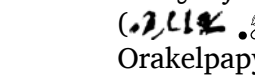
[66] MALININE, Jugement, 160 und pl. II; 168 Anm. 40.

[67] MALININE, *Choix* I, 44-47; II, 17-19 und pl. VI.

[68] PARKER, R. A., The Orthography of Article plus Prothetic *r* in Demotic, in: *JNES* 33, 1974, 371-376.

[69] Nach Photo des Museums; MALININE, *Choix*, 58-59 (mit überflüssigem Fragezeichen); II, 23.

P3-ḥ… Holztafel Leiden I 431, Vso 17 (),[70] in einer Reihe zum guten Teil unklarer Personennamen. Lesung und Identifizierung des Namens sind bisher nicht gelungen.

Nḫt-t3j=f-mw.t, Var. *Nḫt-t3-mw.t*, in pKairo CG 30884 + 30864, 10 ().12 ()[71]; Brooklyner Saitischer Orakelpapyrus, C 12; hieroglyphisch *Nḫt.t=f-mw.t, Nḫt=f-mw.t* PN 212, 17; II 372 „Sein Schutz ist Mut“; kursivhieratisch der Schreibung nach offenbar als „Stark ist seine/die Mutter“ verstanden. Demotisch ist der Name nicht mehr gebräuchlich.

[70] ČERNÝ, Abnormal-Hieratic Tablet, bes. 55 Anm. 59 und pls. 3 und 7 (mit Umschrift ).

[71] VITTMANN, G., Zwei kursivhieratische Urkunden in Kairo, in: *Enchoria* 26, 2000, 125-150, hier 136 und dazu 142 (x) und Taf. 17.

KONTAKTINDUZIERTER SPRACHWANDEL DES ÄGYPTISCH-KOPTISCHEN: LEHNWORT-LEXIKOGRAPHIE IM PROJEKT DATABASE AND DICTIONARY OF GREEK LOANWORDS IN COPTIC (DDGLC)

MATHEW ALMOND, JOOST HAGEN, KATRIN JOHN,
TONIO SEBASTIAN RICHTER & VINCENT WALTER

Der griechisch-ägyptische Sprachkontakt, wie er sich in der Anreicherung des ägyptischen Lexikons im 1. Jahrtausend n. Chr. mit mehr als 4500 griechischen Wörtern der meisten Wortarten und semantischen Felder darstellt, ist einer der am breitesten und dichtesten bezeugten Fälle von intensiver lexikalischer Entlehnung in der Antike. Doch die Erfassung und elementare Aufbereitung der relevanten Sprachdaten, d.h. die Lexikographie griechischer Lehnwörter im Koptischen, scheiterte mehrmals während des 20. Jahrhunderts und ist zu einem kardinalen Desiderat der ägyptischen Wortforschung geworden. Vom 1. April 2010 bis zum 31. März 2012 arbeitete am *Ägyptologischen Institut der Universität Leipzig* das in der Ausschreibung „Geistes- und Sozialwissenschaftliche Forschung“ des *Sächsischen Staatsministeriums für Wissenschaft und Kunst* und der *Sächsischen Akademie der Wissenschaften zu Leipzig* bewilligte Projekt *Database and Dictionary of Greek Loanwords in Coptic (DDGLC)*. In dieser zweijährigen Pilotphase wurde die Möglichkeit getestet, das umfangreiche und in sich reich untergliederte Gesamtcorpus koptischer Texte lehnwortlexikographisch aufzuarbeiten. Dazu wurden in einer Arbeitsdatenbank konzeptuelle und technische Voraussetzungen geschaffen und eine lexikographische Praxis entwickelt und erprobt, die hier in gebotener Kürze vorgestellt werden sollen.

1. Problemstellung

1.1 Koptische Texte

> „Das Koptische stammt vom Altägyptischen, und ist, so weit man bis jetzt beurteilen kann, davon etwa so verschieden, wie das Italiänische vom Lateinischen, oder das Neuhochdeutsche vom Althochdeutschen. Der Name K o p t en wird am wahrscheinlichsten von: Ae – g y p t – er hergeleitet. Er kam bald nach der Arabischen Eroberung in Gebrauch und bezeichnete zuerst die Jakobitischen Christen, dann aber alle Eingeborene von Aegypten im Gegensatze zu den Arabern. – Die Kopten bedienen sich des Hellenischen Alphabetes mit Hinzufügung einiger Zeichen für eigenthümliche Aegyptische Laute.“ VATER, J. S., *Litteratur der Grammatiken. Lexika und Wörtersammlungen aller Sprachen der Erde*, 2., völlig umgearbeitete Ausgabe von B. JÜLG, Berlin 1847, 207.

Johann Severin Vaters frühe Charakterisierung der koptischen Sprache ist im Wesentlichen gültig. Seine lapidare Aussage: „Die

Kopten bedienen sich des Hellenischen Alphabetes mit Hinzufügung einiger Zeichen für eigenthümliche Aegyptische Laute“ bringt ein Phänomen zur Sprache, das, historisch betrachtet, einerseits wohlbekannt ist, andererseits recht erstaunlich erscheint: Ist die Schriftentlehnung vom Griechischen in der Kulturgeschichte des antiken und spätantiken Mittelmeerraumes zu einem typischen Szenario der Alphabetisierung von Sprachen geworden, so unterscheidet sich die Alphabetisierung des Ägyptischen als eine der seltenen *Neuverschriftungen* von jener etwa des Gotischen, Etruskischen oder Altkirchenslavischen, um nur drei der wesentlich häufigeren Beispiele für *Erstverschriftungen* zu nennen (Fig. 1).

Donor System (Basisalphabet), z.B.	Griechisch						
Recipient System 1°, z.B.	Gotisch	Etruskisch	Koptisch	Altkirchenslavisch (Kyrillisch)			
Recipient System 2°, z.B.		Lateinisch	Altnubisch	Russisch	Ukrainisch	Rumänisch	Serbisch

Fig. 1. Schriftentlehnung vom Griechischen

Die in den ersten Jahrhunderten n. Chr. erfolgte Aufgabe des nach 3000 Jahren wahrlich altbewährten, überdies an den phonologischen und strukturellen Eigenarten der ägyptischen Sprache entwickelten hieroglyphischen Schriftsystems ist als *sprach- und schriftgeschichtlicher Prozess* ungefähr nachvollziehbar (vgl. GESSMAN 1976, 1986, QUAEGEBEUR 1982, RICHTER 2009a, SATZINGER 1984, 1985, 1990, 1991, 2003). In dem zugrundeliegenden soziolinguistischen Prozess ist als ein Grundmoment der seit der Eroberung Ägyptens durch Alexander den Großen im 4. Jh. v. Chr. gesellschaftlich etablierte Sprach- und Kulturkontakt zwischen Ägyptern und Griechen in Ägypten (vgl. HOFFMANN 2000, VITTMANN 2003, VIERROS 2012) auszumachen.

Die koptische Schriftsprache war über einen Zeitraum von ca. 1000 Jahren in produktivem Gebrauch. Die frühesten literarischen Texte, typischerweise Übersetzungen aus dem Griechischen, sind um 300 n. Chr. datierbar. Die spätesten neu verfassten Texte – neben Inschriften auch vereinzelt noch umfangreichere Textkompositionen, wie das bohairische Martyrium des Neo-Märtyrers Johannes von Panajôt (ZABOROWSKI 2005) oder die sahidische Versdichtung *Triadon* (VON LEMM 1903, NAGEL 1983) – stammen aus der ersten Hälfte des

14. Jahrhunderts. Eine reproduzierende Manuskripttradition lässt sich noch weiter verfolgen, wie im sahidischen Manuskript P.Bodl.Hunt. 393 aus dem Jahr 1393 (HEBBELYNCK 1900/1, BANDT 2007) und vor allem in den liturgischen Manuskripten des unterägyptischen Hochdialekts Bohairisch, die noch bis ins 18. Jh. kopiert wurden. Dagegen erlischt die Produktion von dokumentarischen Texten des Alltags, wie Rechtsurkunden oder Briefen, bereits im 11. Jahrhundert (RICHTER 2009a, DELATTRE *et al.* 2012) – ein wichtiger Indikator für die Chronologie des Sprachwechsels der ägyptischen Christen zum Arabischen (BJÖRNESJÖ 1996, PAPACONSTANTINOU 2007 & 2012, RICHTER 2009a, ZABOROWSKI 2008). Während der gesamten Dauer seiner Anwendung als schriftsprachliches Medium war das Koptische durch andere, dominante Schriftsprachen sozial und funktionell beschränkt auf bestimmte Milieus und Textdomänen. Seit der Neuverschriftung des Ägyptischen um 300 n. Chr. bis ins 8. Jh. war es das Griechische, das als Prestige-Sprache mit weitaus größerem schriftsprachlichen Anwendungsbereich neben dem Koptischen in Gebrauch stand; von der Mitte des 8. Jh.s an wuchs das Arabische in diese Rolle hinein (RICHTER 2009a, 2010, DELATTRE *et al.* 2012).

Wirkte sich die soziolinguistische Konstellation der Mehrsprachigkeit Ägyptens im ersten Jahrtausend somit restringierend auf das Repertoire der in Koptisch geschriebenen Texte und Textsorten aus, so war die gleichsam inoffizielle Stellung des Koptischen in dieser Konstellation sicherlich ein Grund dafür, dass sich kein einheitlicher Standard der koptischen Schriftsprache durchsetzen konnte. Stets waren mehrere lokale Varietäten, zeitweilig bis zu etwa einem Dutzend, in Gebrauch (HINTZE 1984, FUNK 1988, 1991, KASSER 1991a-c), abgesehen von prä- und destandardisierten Normen, wie sie im frühesten und spätesten Koptisch sowie in bestimmten nicht- und semiliterarischen Textsorten anzutreffen sind (KAHLE 1954, GROSSMAN 2007, RICHTER 2008a).

Zu den topolektalen und chronolektalen varietätenlinguistischen Parametern, denen das Koptische unterworfen ist, tritt als diversifizierender Faktor die textlinguistische Spezifik unterschiedlicher Texttypen und Textsorten hinzu. Koptische *Übersetzungsliteratur* aus dem Griechischen, wie z.B. *LXX*, *NT*, ATliche und NTliche Apokryphen, patristische Literatur, manichäische, gnostische und hermetische Texte und koptische *Originalliteratur*, wie z.B. monastische, homiletische und hagiographische Literatur, liturgische Texte (vgl. COQUIN 1993, EMMEL 2004, KRAUSE 1980, ORLANDI 1995, 1998, 2004, 2005), ferner koptische *Wissensliteratur*, wie medizinische und

magische Ablagetexte, mathematische Exempel, alchemistische Rezeptsammlungen (MEYER & SMITH 1994, RICHTER 2009b, TILL 1951a) und darüber hinaus Tausende von koptischen *dokumentarischen* Texten (RICHTER 2008a, 2008b), wie Briefe, Rechtsurkunden, Listentexte, geschäftliche Kurztexte, medizinische und magische Anwendungstexte und Inschriften (TUDOR 2011), bieten sprachliches Material aus ganz unterschiedlichen sprachlichen Registern des späten Ägyptisch mit ihren jeweiligen phraseologischen und stilistischen Eigenarten, ihren Sonder- und Fachwortschätzen. Diese Diversität des koptischen Textcorpus, abgesehen von seinem schieren Umfang, macht die sprachliche und so auch die lexikographische Beschreibung und Analyse des Koptischen sehr aufwendig, aber auch besonders lohnend, da im Ergebnis reich und komplex.

1.2 Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen

Während ihrer durch Textüberlieferung mehr als 4000 Jahre lang bezeugten Geschichte dürfte die ägyptisch-koptische Sprache (LOPRIENO 1995, LOPRIENO 2001, LOPRIENO & MÜLLER 2012) permanent in Kontakt mit afrikanischen, semitischen oder indoeuropäischen Sprachen gestanden haben. Nur in bestimmten Perioden und unter bestimmten Bedingungen haben diese Sprachkontakte jedoch Spuren im ägyptischen Textcorpus hinterlassen, lassen sich auf einer oder mehreren der Strukturebenen der ägyptischen Schriftsprache Phänomene *kontaktinduzierten Sprachwandels* erkennen. Tatsächlich scheint von den Sprachstufen *Altägyptisch* (ca. 2700 – 2200 v. Chr.) bis *Demotisch* (ca. 650 v. Chr. – 300 n. Chr.) der Sprachwandel des Ägyptischen kaum oder wenig durch Sprachkontakt motiviert, affiziert oder gelenkt worden zu sein (semitische Lehnwörter im Ägyptischen: BURCHARDT 1909/10, HOCH 1994, QUACK 2005, VITTMANN 1996, WINAND [in Vorbereitung]; nichtsemitische Lehnwörter im Ägyptischen: KNIGGE 2004, SCHNEIDER 2004).

Dagegen stellt sich der Unterschied zwischen dem im *Demotischen* fixierten Sprachzustand und dem der jüngsten ägyptischen Sprachstufe, des *Koptischen*, nicht zum geringsten Teil als eine Bilanz des intensiven griechisch-ägyptischen Kultur- und Sprachkontakts seit der Eroberung Ägyptens durch Alexander den Großen dar (CLARYSSE 1993, FEDER 2004, FEWSTER 2002, KAPSOMENOS 1953, MCBRIDE 1989, PEREMANS 1964 & 1983, QUAEGEBEUR 1974, SATZINGER 1984, SIDARUS 2008, TORALLAS TOVAR 2004, 2005, VERGOTE 1984, VIERROS 2012). Schießlich tritt in spätkoptischen Texten seit dem 9., verstärkt dann im 10. und 11. Jahrhundert n. Chr., die linguistische Interferenz der

ägyptischen Sprache mit dem Arabischen zutage (RICHTER 2001, 2006).

Neben der Neuverschriftung der ägyptischen Sprache auf Basis des griechischen Alphabetes ist das auffälligste Indiz des jahrhundertelangen griechisch-ägyptischen Sprachkontakts in Ägypten die hohe Anzahl und Frequenz griechischer Lehnwörter beinahe aller Wortarten und semantischen Felder in koptischen Texten. Wortentlehnung aus dem Griechischen dürfte quantitativ und qualitativ für die Gesamtstruktur des koptischen Lexikons wie für die Architektur vieler seiner semantischen Domänen von großer Bedeutung sein. Andererseits wird der *via* Koptisch überlieferte griechische Wortschatz als wichtige Nebenüberlieferungen für unsere Kenntnis des griechischen Wortschatzes der post-hellenistischen Zeit beurteilt. Generell dürfte gelten, dass kaum ein sprachlicher Lehnvorgang der Antike breiter und dichter bezeugt ist als dieser. Dennoch ist der griechische Lehnwortschatz im Koptischen bisher kaum erforscht, ja, auch nur einigermaßen überblickt.

1.3 Lexikographie der Griechischen Lehnwörter im Ägyptisch-Koptischen: Der Forschungsstand vor dem DDGLC-Projekt

> „Es würde eine umfangreiche Arbeit sein, die die Wechselbeziehungen der beiden Volks- und Sprachgeister vielfach beleuchten müsste, ein Lexikon dieser koptisch-griechischen Worte anzufertigen, die Fälle ihres Vorkommens zu zählen, und ihr Verhältniss zu ihren rein-koptischen Synonymen numerisch und semasiologisch zu erörtern.“ ABEL, C., *Koptische Untersuchungen*, Berlin 1876, 549-550.

Als der vielschreibende und vielgeschmähte Carl Abel 1876 diesen ‚Projektvorschlag‘ zu einer den Zielen des *DDGLC*-Projekts gar nicht fernen Lehnwort-Lexikographie des Koptischen vortrug, herrschte weit und breit die Meinung, dass griechische Wörter im Koptischen nichts als griechische Wörter in verballhornter Schreibung seien – ein wohlvertrautes Element im weniger vertrauten Umfeld, zu dessen Verständnis ein griechisches Wörterbuch genügt – und dass diese Wörter dem koptischen Lexikon nichts Wesentliches hinzufügten; vgl. z.B. SCHWARTZE & STEINTHAL 1850, 4: „Ist nun durch die Aufnahme dieser fremden Wörter der Umfang der Koptischen Sprache in materieller Hinsicht verringert worden? Diese Frage ist unbedingt zu verneinen, weil, mit Absehung von ganz speciellen Benennungen …, äußerst wenig Griechische und Lateinische Wörter gefunden werden möchten, für welche sich nicht auch der entsprechende Koptische Ausdruck nachweisen liesse“. Wiewohl in der Folgezeit revidiert, hat

diese Meinung maßgeblich die Praxis der koptischen Lexikographie bestimmt: Alle existierenden koptischen Wörterbücher schließen die griechischen Lehnwörter systematisch aus. Sie basieren auf einer etymologischen Selektion, vergleichbar einem deutschen Wörterbuch, in dem allein Wörter mit germanischer Etymologie berücksichtigt sind und Wörter wie Mauer, Fenster, Wein, Tabak, Tomate, Zucker, Reis, Rakete, Turban, Sekunde, Minute, spazieren, Tresor etc. unauffindbar wären. Damit ist freilich weder der quantitativen und qualitativen Bedeutung des griechischen Lehnwortschatzes für den Wortbestand und die semantische Architektur des ägyptischen Lexikons im 1. Jahrtausend n. Chr., noch dem Einfluss, den die koptische Sprachumgebung und der koptische Gebrauch auf das griechische Wortmaterial ausübten, Rechnung getragen.

Im 20. Jahrhundert wurde nach und nach der griechische Lehnwortschatz im Koptischen als philologisches Thema entdeckt und als lexikologische Aufgabe erkannt (ALLBERRY 1937, BLOK 1927, GASELEE 1929/30, HOPFNER 1918, JERNSTEDT 1929, LEFORT 1934, RAHLFS 1900, 1912). In den frühen fünfziger Jahren begann Alexander Böhlig in Halle mit der Arbeit an einem Wörterbuch der griechischen Lehnwörter im Koptischen (BÖHLIG 1953b, 1954a-c, 1956, 1960, 1962; vgl. NAGEL 2013). Die ‚Republikflucht' Böhligs im Jahr 1963, bei der er *nolens-volens* die bis dato angelegten lexikographischen Zettelkästen mit ca. 65.000 Einträgen (TUBACH 1999a, 414) in Halle zurückließ, bedeutete eine Zäsur und im Endeffekt das Scheitern seines Projekts. Auf der Basis der in der DDR verbliebenen Zettelkästen versuchte zunächst noch einer von Böhligs Schülern, der Neutestamentler Hans-Friedrich Weiß, das Projekt zu vollenden (WEISS 1966, 1968, 1969, 1972). Dreißig Jahre später nahm Jürgen Tubach am Institut für Christlichen Orient in Halle einen neuen Anlauf, um das Projekt wieder in Gang zu setzen (TUBACH 1999a-b, DEMARIA 2005). Auf der anderen Seite der Mauer initiierte Alexander Böhlig, der das Institut für Christlichen Orient an der Universität Tübingen begründet hatte, einen Neubeginn der lexikographischen Verzettelung koptischer Texte (BAUER 1975, SIEGERT 1982). In diesen Kontext gehören auch die Zettelkästen von Gertrud Bauer mit ca. 14.000 Einträgen zu griechischen Partikeln, Konjunktionen und Präpositionen sowie ein abgeschlossenes Buchmanuskript darüber, die durch Peter Nagel im Sommer 2010 dem *DDGLC*-Projekt übergeben wurden und in den vergangenen Jahren für die weitere Nutzung aufbereitet worden sind: <http://www.uni-leipzig.de/~ddglc/docs/GertrudBauerCardindex.pdf> (s.u., 2.5).

In die Forschungsgeschichte der Lehnwortlexikographie des Koptischen gehören schließlich die Arbeiten zu einem Gesamtwörterbuch des Koptischen, die in den 1960er Jahren von dem Genfer Koptologen Rodolphe Kasser in Zusammenarbeit mit Werner Vycichl angestellt wurden. Doch dem ersten Faszikel dieses *Dictionnaire auxiliaire, étymologique et complet de la langue Copte*, das von *A(lpha)* bis *baukalion* reicht (KASSER & VYCICHL 1967), ist kein zweites gefolgt. Die obsolet gewordenen Vorarbeiten sind teilweise aufgegangen in VYCICHL 1983.

In der Zwischenzeit dienten Lehnwortindizes und Konkordanzen als pragmatische Ersatzlösung. Auf diese Weise wurden etwa das sahidische Neue Testament (DRAGUET 1960, LEFORT 1950a), Teile des griechischen Wortschatzes der koptischen Manichaica (BÖHLIG 1954a, 1958a, CLACKSON *et al.* 1998), des Neuen Testaments im bohairischen Dialekt (BAUER 1975, BÖHLIG 1954b-c, 1958b-c) und der Werke des Klosterabts Schenute (BEHLMER 1997/8; FUNK 2007), das Corpus der koptischen dokumentarischen Texte (FÖRSTER 2002) und die Manuskripte der Bibliothek von Nag Hammadi (CHARRON 1992, 1995, CHERIX 1993, 1995, 2000, FUNK 1997, 2000, FUNK & POIRIER 2006), wenngleich in einem reduzierten Modus, lehnwortlexikographisch dokumentiert.

Dem defizitären Stand der lexikographischen Aufarbeitung geschuldet, konnten am griechischen Lehnwortschatz des Koptischen bisher nur sporadisch und punktuell semantische und semasiologische Wortforschung (z.B. DRESCHER 1969/76, FUNK 1982, GASELEE 1914, GODRON 1983, LEFORT 1948, 1950b, SATZINGER 1970, SCHILLER 1950, TILL 1951b) oder andere lexikologische Arbeiten geleistet werden (*zur Phonologie und Morphologie* der griechischen Lehnwörter im Koptischen z.B. BÖHLIG 1953a, 1955, 1995, FÖRSTER 2002, xiv-xxix, GASELEE 1916, GIRGIS 1963-2001, RAHLFS 1900, HOPFNER 1918, TILL 1951a; *zur Syntax* der griechischen Lehnwörter im Koptischen und kontaktlinguistischen Fragestellungen z.B. ALMOND 2010, 2011, BRUNSCH 1983, FUNK 1984, GROSSMAN 2009, GROSSMANN & RICHTER [in Vorbereitung], HASZNOS 2012, KASSER 1966, 1991d, NAGEL 1971, ORÉAL 1999, POLOTSKY 1950, REINTGES 2001, 2004, RICHTER 2008a, SHISHA-HALEVY 2009).

2. Das DDGLC-Projekt und seine Pilotphase

2.1 Das Ziel des DDGLC-Projekts und seiner Pilotphase

Das Projekt *Database and Dictionary of Greek Loanwords in Coptic* (*DDGLC*) hat die lehnwortlexikographische Aufarbeitung des Gesamtcorpus' der koptischen literarischen und nichtliterarischen Texte zum Ziel. Dazu ist die Erfassung aller Formen (*types*) und Belege (*tokens*) griechischer Wörter im Koptischen mit ihren syntaktischen und semantischen Eigenschaften und Funktionen zunächst in einer Datenbank, dann auch in einem Wörterbuch geplant. In gleicher Weise sollen später auch die griechischen Lehnwörter im vorkoptischen Ägyptisch und die arabischen Lehnwörter im späteren Koptisch dokumentiert werden. Damit können der historisch-linguistischen, kontaktlinguistischen und lehnwort-typologischen Forschung Daten zum kontaktinduzierten Sprachwandel des ägyptischen Lexikons über einen Zeitraum von 1.500 Jahren bereitgestellt werden.

Während der zweijährigen Pilotphase, die vom 01.04.2010 bis zum 31.03.2012 am Ägyptologischen Institut der Universität Leipzig lief, wurden in einer Arbeitsdatenbank konzeptuelle und technische Voraussetzungen geschaffen und eine lexikographische Praxis entwickelt, die dazu geeignet sind, dieses hochgesteckte Projektziel langfristig zu erreichen. Seit November 2012 arbeitet das *DDGLC*-Projekt als *DFG*-Langzeitprojekt mit einer geplanten Laufzeit von zwölf Jahren.

2.2 Technische Aspekte des DDGLC-Projekts: Modellierung und Realisierung der Datenbank

Entsprechend den Voraussetzungen und Zielen des Projektes wurden in den ersten drei Monaten der Pilotphase in Besprechungen und Workshops zwischen dem Projektleiter, den Lexikographen und der IT-Verantwortlichen die Anforderungen für die zu erarbeitende Datenbank präzisiert; Probeartikel und Testdatensätze wurden ausgewertet. Im Zentrum dieses Prozesses stand die Erstellung eines Fachkonzeptes zur lexikographischen Aufnahme und Bearbeitung griechischer Lehnwörter im Koptischen. Dieser Entwurf wurde als Datenverarbeitungs-Konzept mit der Beschreibung der relevanten Daten, ihrer Strukturierung und Verarbeitung weitergeführt; seine Umsetzung erfolgte im relationalen Datenbank-Verwaltungssystem FileMaker (Version 11).

Die zweijährige Pilotphase verfolgte auch das Ziel, das erarbeitete Konzept aus lexikographischer und technischer Sicht ausreifen zu

lassen. So wurden permanent Korrekturen und Weiterentwicklungen eingearbeitet. Es waren diese bereits erwarteten Änderungen, die es empfehlenswert erscheinen ließen, ein Datenbank-Management-system wie FileMaker zu verwenden, bei dem in effektiver Weise eine benutzerfreundliche Datenbankversion mit graphischer Oberfläche zur Verfügung gestellt werden kann, während Neuerungen umgehend und buchstäblich bei laufendem Betrieb durchgeführt werden können. In gewissem Umfang waren hierbei auch Änderungen im Datenbankschema möglich.

Die FileMaker-Realisierung war nur für die Verwendung in der Pilotphase vorgesehen. Im Hauptprojekt soll, aufbauend auf den Strukturen und Erfahrungen der Formierungsphase, eine Evaluation des Datenverarbeitungs-Konzeptes stattfinden. Anschließend soll die Realisierung im Datenbank-Managementsystem MySQL erfolgen. Nicht mehr benötigte Funktionen werden entfernt, neue treten über die gesamte geplante Projektzeit nach und nach hinzu, wie z.B. solche Funktionen, die bei der späteren Kompilation des Wörterbuches Möglichkeiten eines Artikelredaktionssystems bieten werden.

Nach der Erstellung dieser Datenbank soll eine Online-Version für die Öffentlichkeit verfügbar gemacht werden.

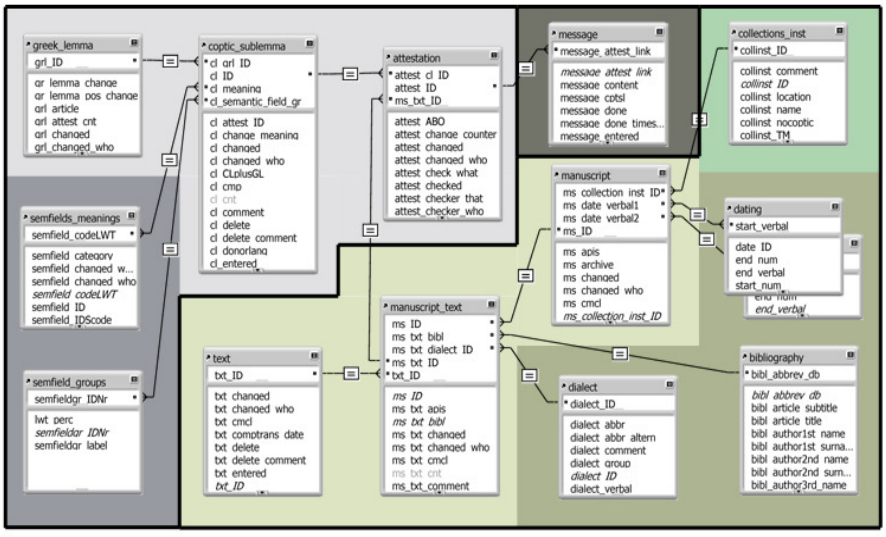


Fig. 2. Datenbankschema der DDGLC-Datenbank

Die erste *DDGLC*-Datenbankversion wurde am 10. August 2010 für die Bearbeiter in Dienst gestellt und konnte seitdem durch zahlreiche Änderungen und Erweiterungen verbessert werden. Dies geschah auf den Ebenen des zugrundeliegenden Schemas, der Benutzeroberfläche und der zur Verfügung gestellten Funktionen.

An den Basis-Pfeilern dieses Schemas waren keine wesentlichen Änderungen nötig. Die Daten sind in 14 Tabellen mit insgesamt 218 Feldern strukturiert (Fig. 2).

Diese Anzahl von Feldern muss den Bearbeiter nicht stören, da dieser im Wesentlichen lediglich mit den Hauptinformationen in 48 Feldern in Berührung kommt. Andere Felder, größtenteils ausgeblendet, dienen organisatorischen und statistischen Zwecken und werden automatisch ausgefüllt. Komplexere Wertelisten sind als eigene Relation integriert, was die Anzahl der Tabellen und Felder ein wenig erhöht. Für die tägliche Arbeit, die Aufnahme der Belegstellen, sind übersichtliche acht Felder auszufüllen. Umfangreiche Möglichkeiten der Kommentierung (34 Felder) und des Nachrichtenaustauschs innerhalb der Datenbank stehen den Bearbeitern zur Verfügung. Abgesehen von Feldern mit automatischer Erzeugung von Feldinhalten, bzw. im Zusammenhang damit, erleichtern augenblicklich 48 Skripte die Arbeit mit und an der Datenbank.

Die Relationen, deren Zusammenschluss das lexikographische Modul bildet, sind (1) die *Belegstellen* (*Attestations*), (2) die koptischen *Sublemmata* und (3) die Tabelle griechischer *Lemmata*. In einem zweiten Modul der Datenbank werden Informationen zu den Texten und Manuskripten mit (thesaurusgestützten) Angaben zu Datierung, Herkunft, ggf. Archivzugehörigkeit, Dialekt, Aufenthaltsort, Edition und Texttypologie verbunden.

Die Unterscheidung von Text-Daten, Manuskript-Daten und Manuskript-Text-Daten ermöglicht es, die Zugehörigkeit mehrerer Texte zu einem Manuskript (z.B. die in einem Nag-Hammadi-Kodex enthaltenen Einzeltexte) sowie Texte, die in verschiedenen Manuskripten vorkommen (z.B. ein neutestamentliches Buch, das in mehreren Dialekten und etlichen Einzelhandschriften bezeugt sein kann), angemessen abzubilden.

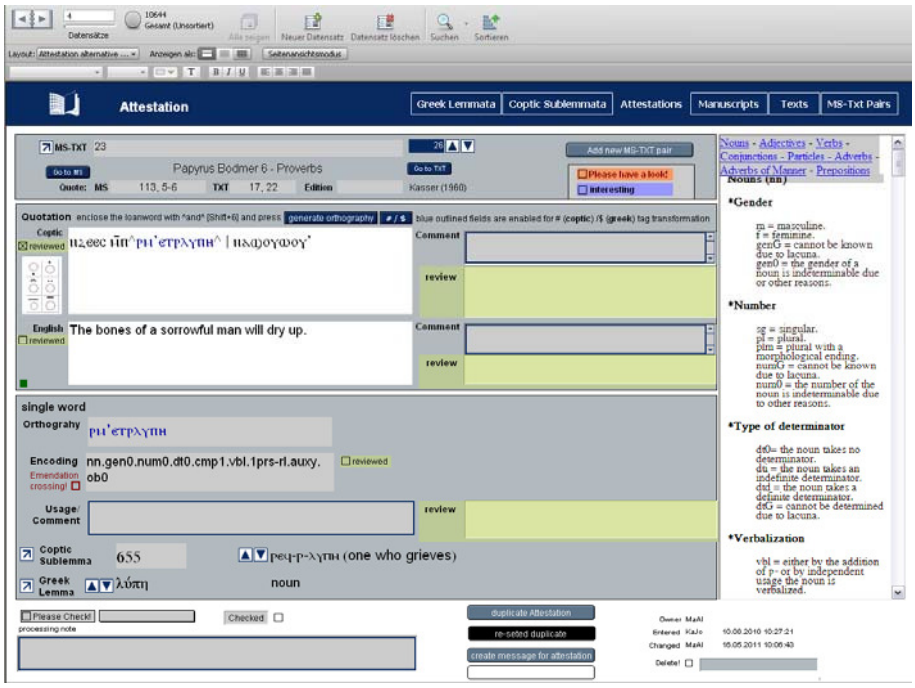


Fig. 3. Hauptansicht der DDGLC-Datenbank: Belegstellen-Eingabe

Für die Dateneingabe und -ansicht stehen momentan zehn Layouts zur Verfügung, von denen hauptsächlich drei (ausnahmslos aus dem lexikographischen Modul) genutzt werden, besonders die Hauptansicht zur Belegstellen-Eingabe (Fig. 3). Bei der entstandenen Datenbank handelt es sich um ein *Werkzeug* für die tägliche Arbeit der Lexikographen, sie ist ausschließlich für die interne Nutzung gedacht. Eine öffentliche Online-Version für ein breites Publikum war im Pilotprojekt selbst noch nicht vorgesehen.

2.3 Die „Demonstration Videos of the DDGLC Database“

Um bereits diese vorläufige Datenbank präsentieren und der Kritik der Fachwelt stellen zu können, wurden neun Demonstrationsvideos erstellt und auf der Projekt-Homepage des *DDGLC*-Projekts verfügbar gemacht. Diese neun Kurzfilme thematisieren jeweils einzelne Ebenen oder wichtige Einzelaspekte, wie etwa die grammatische Codierung, mit aussagekräftigen Beispielen und leiten den Betrachter audio-visuell durch die *DDGLC*-Datenbank.

Video 1
<http://www.uni-leipzig.de/~ddglc/docs/videos/v1/v1.html>
„Introduction“ bietet Informationen zur Affiliation des *DDGLC*-Projekts und zu den generellen Zielen seiner Arbeit.

Video 2
<http://www.uni-leipzig.de/~ddglc/docs/videos/v2/DDGLC%20Video%202.html>
„Historical-Archaeological framework“ geht auf jenen Teil der Datenbank ein, in dem Metadaten über die lexikographisch aufgearbeiteten Handschriften und Texte aufbewahrt werden. Standardisierte Informationen zu den historischen Umständen, unter denen ein Text produziert oder kopiert wurde, wie regionale Herkunft, Dialekt, Datierung, Textsorte, sind mit der Ebene der Einzelbelege verknüpft, so dass der Einfluss dieser Parameter auf Entlehnung und Entlehnbarkeit griechischer Wörter systematisch getestet werden kann.

Video 3
<http://www.uni-leipzig.de/~ddglc/docs/videos/v3/v3.html>
„Greek Lemma Level“ führt die griechische Lemmaliste, die oberste Struktureinheit der Datenbank, vor. In dieser Liste sind alle im Koptischen belegten individuellen griechischen Wörter (*types*) als Ausgangsformen (*input forms*) in griechischer Schrift und Orthographie nach den Konventionen der griechischen Lexikographie aufgelistet; nach derzeitigem Stand (Mai 2013) mehr als 5.200 Wörter.

Video 4
<http://www.uni-leipzig.de/~ddglc/docs/videos/v4/v4.html>
„Coptic Sublemma Level“ erklärt die nächsttiefere Struktureinheit, in der die von den Einheiten der abstrakten Gesamt-Lemmaliste derivierten koptischen Wörter in einer standardisierten Form mit ihren konkreten Bedeutungen im koptischen Gebrauch aufgelistet sind. Diese koptischen Sublemmata sind einerseits mit den griechischen Lemmata, andererseits mit den Einträgen der Einzelbelege (*attestations*) verknüpft.

Video 5
<http://www.uni-leipzig.de/~ddglc/docs/videos/v5/v5.html>
„Attestation Level“ erklärt die Strukturebene der Einzelbelege (*tokens*), deren Einträge (*attestations*) die individuelle Schreibweise eines Wortes (*single word orthography*) der standardisierten Sublemma-Form zuweist, eine Übersetzung des Wortes im spezifischen Kontext bietet sowie zu jedem Wortbeleg morphologische und grammatische Informationen nach einem differenzierten Kodierungssystem (*encoding*) gibt. Dieses Kodierungssystem wird in den Videos 6-9 wortklassenweise erläutert:

Video 6
<http://www.uni-leipzig.de/~ddglc/docs/videos/v6/v6.html>
„Attestation Level – Encoding: Nouns“,

Video 7
<http://www.uni-leipzig.de/~ddglc/docs/videos/v7/v7.html>
„Attestation Level – Encoding: Verbs“,

Video 8
<http://www.uni-leipzig.de/~ddglc/docs/videos/v8/v8.html>
„Attestation Level – Encoding: Adjectives“,

Video 9
<http://www.uni-leipzig.de/~ddglc/docs/videos/v9/v9.html>
„Attestation Level – Encoding: Functional Language“.

2.4 Die lexikographische Arbeit des DDGLC-Projekts in seiner Pilotphase

Um der Zielsetzung des *DDGLC*-Projekts zu entsprechen, neben der technischen auch eine lexikographische Grundlage für die Erforschung des griechisch-koptischen Sprachkontakts zu schaffen, lag ein Hauptaugenmerk der zweijährigen Pilotphase auf der Eingabe von Lehnwort-Daten aus verschiedenen Teilcorpora des Koptischen. Die Auswahl dieser Teilcorpora war unter der Maßgabe erfolgt, einerseits in sich geschlossene Datensets zu erfassen, die an sich von wissenschaftlichem Interesse sind, andererseits durch Heterogenität in Bezug auf dialektale, diachrone und funktionale Varietäten möglichst viele der künftig anfallenden Problemtypen zu antizipieren.

Eine grundsätzliche Schwierigkeit von Grundlagenforschung, wie sie im *DDGLC*-Projekt stattfindet, besteht darin, künftige Forschungsinteressen im Voraus zu kalkulieren. So wurden im Laufe der Konzeption mögliche Problemfelder, angefangen von scheinbaren Trivialitäten wie die Repräsentation von Informationen zu Manuskripten und Texten (sowie deren in der Praxis nicht so triviale Verknüpfung und Unterscheidung) bis hin zu zentralen linguistischen Problemen, im Modus des *educated guess* antizipiert und nach Möglichkeit berücksichtigt.

Exemplarisch ist etwa die Frage, wie mit verschieden tradierten Versionen ein und desselben Textes umzugehen ist – eine in Anbetracht der für die koptische Literatur so wichtigen Bibeltexte virulente Problematik. Eine Verlinkung der Belegstellen allein auf der Ebene des Textes würde der Existenz verschiedener Versionen nicht Rechnung tragen, während eine Verlinkung auf der Ebene der

Manuskripte wiederum die Tatsache ignorieren würde, dass es sich grundsätzlich um Varianten des selben Textes handelt. Oder: Wie sind Manuskripte zu benennen, deren Fragmente auf verschiedene Sammlungen verteilt und inzwischen als zusammengehörig identifiziert sind, die aber noch keine eigene Bezeichnung in der Wissenschaft erhalten haben?

Auch aus linguistischer Perspektive stellten sich Fragen, die für die potentiellen Ergebnisse des Projekts und damit für seinen künftigen Erfolg zentral sind: Welche grammatikalischen Informationen sind für die spätere Auswertung relevant und müssen daher in der Datenbank festgehalten werden? Wie viel Kontext ist zur syntaktischen und semantischen Analyse notwendig? Nach welchem Prinzip sollen die entlehnten griechischen Wörter lemmatisiert werden?

Bezüglich der Lemmatisierung fiel die Entscheidung auf ein hierarchisches Modell, an dessen Spitze ein griechisches *Lemma* steht. Mit diesem griechischen Lemma sind die sogenannten koptischen *Sublemmata* verknüpft, die durch eine koptische Standard-Morphologie und eine bestimmte entlehnte Bedeutung konstituiert werden. Mit diesen Sublemmata sind schließlich die einzelnen *Belegstellen* verknüpft, die neben dem koptischen Text auch eine englische Übersetzung sowie Informationen zur Morphologie bzw. Orthographie individueller Lehnwortbelege und zur Syntax der individuellen Lehnwort-Verwendung bieten (s.o., 2.3).

Bei Abschluss der Pilotphase Ende März 2012 enthielt die Datenbank 10.644 Belegstellen, die mit 951 koptischen Sublemmata verknüpft waren. Folgende Teilcorpora wurden bis dahin erfasst:

1) Frühe koptische literarische Varietäten: P.Bodmer III (*B4*), P.Bodmer VI (*P*), P.Hamb.Bil. 1 (*F7*), Ascensio Iesaiae ed. Lefort (*i7*);

2) frühe koptische nichtliterarische Varietäten: P.Lond. VI, P.Nepheros, P.Nag Hammadi, die Apa-Johannes-Korrespondenz aus P.Ryl.Copt.;

3) aus dem mittelägyptischen Corpus: die Matthäus-Evangelien der Codices Scheide und Schøyen (Dialekt *M*), aus kleineren mittelägyptischen Dialekten (*V*, *W*) das Johannes-Evangelium P.Mich. 3521, der Ecclesiastes-Text und die katholischen Briefe des P.Mich. 3520 und des P.Mich. 6868a;

4) aus dem Corpus des achmimischen Dialekts (*A*): das Wiener Ms. der Kleinen Propheten, der Proverbientext des Ms. Berl. inv.

orient. oct. 987 und die Exodus-, Sirach- und Makkabäer-Fragmente.

Die Verteilung der Belegstellen nach Wortarten stellt sich wie folgt dar (wobei Belegstellen und Sublemmata, die noch nicht abschließend kategorisiert werden konnten, ausgespart sind):

Wortart	Belege	Prozentualer Anteil
Nomina	582	61,2 %
Verben	197	20,7 %
Adjektive	74	7,8 %
Konjunktionen und Partikeln	77	8,1 %
Präpositionen	11	1,2 %

Fig. 4. Types (Sublemmata) in der DDGLC-Datenbank, März 2012

Wortart	Belege	Prozentualer Anteil
Nomina	4792	45,0 %
Verben	1059	9,9 %
Adjektive	775	7,3 %
Konjunktionen und Partikeln	3599	33,8 %
Präpositionen	179	1,7 %

Fig. 5. Tokens in der DDGLC-Datenbank, März 2012

Obwohl es in Anbetracht der geringen Anzahl der bisher erfassten Texte und des eingeschränkten Repertoires der bisher vertretenen Textsorten für Generalisierungen zu früh ist, zeigen doch diese vorläufigen Zahlen gewisse Tendenzen.

Die Daten zur Beleghäufigkeit (*token frequency*) haben sich über die Laufzeit der Pilotphase signifikant gewandelt.

Bei den *types* sind die Zahlenverhältnisse zwischen den Wortarten stabiler geblieben, wobei zu erwarten steht, dass sich auch dort mit fortschreitender Dateneingabe das Verhältnis langsam verschieben wird, da im Bereich der *content words* (*Autosemantika*) ein ungleich größeres Repertoire an entlehnbaren Wörtern verfügbar ist als im Bereich der *functional language* (*Synsemantika*).

Fallbeispiel: Die Matthäus-Evangelien der Codices Scheide und Schøyen

Als Beispiel für die Informationen, welche die *DDGLC*-Datenbank der koptischen Lehnwortforschung bieten kann, soll hier die Differenz im Lehnwortgebrauch der beiden Versionen des Matthäus-Evangeliums im mittelägyptischen Dialekt des Koptischen – Codex Scheide (SCHENKE 1981) und Codex Schøyen (SCHENKE 2001) – vorgestellt werden. Auch wenn die absoluten Belegzahlen (*token frequency*) im Codex Schøyen aufgrund des geringeren erhaltenen Textbestandes niedriger sind, lassen sich doch bereits in der Verteilung der Lehnwortbelege auf Wortklassen erstaunliche Unterschiede feststellen.

Wortart	Belege	Prozentualer Anteil
Nomina	989	38,8 %
Verben	364	14,3 %
Adjektive	74	2,9 %
Konjunktionen und Partikeln	1059	41,5 %
Präpositionen	26	1,0 %

Fig. 6. Tokens in Codex Scheide (EvMt, Dialekt M)

Wortart	Belege	Prozentualer Anteil
Nomina	647	51,6 %
Verben	175	14,0 %
Adjektive	56	4,5 %
Konjunktionen und Partikeln	319	25,4 %
Präpositionen	22	1,8 %

Fig. 7. Tokens in Codex Schøyen (EvMt, Dialekt M)

Ein auffälliger Unterschied zeigt sich in der Häufigkeit von Synsemantika. Recherchiert man die Gründe dafür im Detail, so stellt sich heraus, dass dabei die Häufigkeit der postpositiven Partikel δέ „und; aber“ maßgeblich ist (Fig. 8):

	Belege	„und“	„aber“	unsicher
Scheide	496 (19,4 % aller Lehnwörter)	331 (66,7 %)	160 (32,2 %)	5 (1 %)
Schøyen	95 (7,6 % aller Lehnwörter)	30 (31,6 %)	65 (68,4 %)	

Fig. 8. Die postpositive Partikel δέ in den Codices Scheide und Schøyen

Im Codex Scheide tritt sie auffallend häufig auf (496 Belege = 19,4 % aller Lehnwörter). Diese Häufigkeit nun korrespondiert mit ihrem regelmäßigen Gebrauch in semantisch reduzierter Funktion, nämlich zum einfachen Anschluss des folgenden Satzes oder Gedankens. Die Belege für δέ „und“ verhalten sich in Codex Scheide zu den Belegen für δέ „aber“ wie zwei zu eins.

Im Codex Schøyen hingegen ist der Gebrauch von δέ deutlich seltener (95 Belege = 7,6 % aller Lehnwörter), doch tritt hier die disjunktive Valeur dieser Partikel viel deutlicher zutage. Unter diesen insgesamt 95 Belegstellen konnte die Bedeutung „und“ nur 30 Belegen, die Bedeutung „aber“ dagegen 65 Belegen zugeordnet werden.

Auch was die *types* der verwendeten Lehnwörter angeht, sind die Unterschiede zwischen den beiden mittelägyptischen Matthäus-Texten beträchtlich: Von den 360 im Codex Scheide und 285 im Codex Schøyen belegten griechischen Wörtern treten 88 ausschließlich im Codex Scheide auf; 31 sind ausschließlich im Codex Schøyen zu finden. Dabei sind verschiedene Phänomene zu beobachten, welche an dieser Stelle anhand einiger Beispiele nur vorgeführt, nicht erklärt werden sollen:

- Griechischer Ausdruck vs. koptisches Synonym: Im Codex Scheide wird konsequent der griechische Fachterminus γραμματεύς „Schreiber, Schriftgelehrter“ verwendet, wo Codex Schøyen den allgemeineren koptischen Begriff ⲥⲉϩ „Schreiber“ nutzt.
- Griechischer Ausdruck vs. griechisches Synonym: Während Codex Scheide den gewöhnlichen Ausdruck σκάνδαλον „Ärgernis“ bietet, verwendet Codex Schøyen die in originär griechischen Texten seltene Form σκάνδαλος; Codex Schøyen bietet die normale griechische Verbform ὑμνεῖν „besingen, preisen“, Codex Scheide dagegen die in den gängigen griechischen Wörterbüchern nicht verzeichnete Form ὑμνεύειν.

Derartige Unterschiede im Lehnwortschatz der beiden Texte sind nicht auf *content words* beschränkt. Man findet sie auch bei den

Elementen der *functional language*: So sind εἴτε … εἴτε „sei es, dass ... sei es, dass“ und πῶς „wie“ auf Codex Scheide beschränkt, ὅπως „damit“, χωρίς „ohne“ und ἀπό „von“ dagegen auf Codex Schøyen.

2.5 Die Zettelkästen Dr. Gertrud Bauers

Mit Unterstützung der *Gertrud-und-Alexander-Böhlig-Stiftung* konnten während der Pilotphase des Projekts *Database and Dictionary of Greek Loanwords in Coptic* die von Gertrud Bauer für Alexander Böhlig erarbeiteten Zettelkästen, die Peter Nagel im Sommer 2010 dem *DDGLC*-Projekt übergeben hatte, aufgearbeitet werden. Bei der Sichtung und Ordnung der ca. 14.000 lexikographischen Zettel zeigte sich, dass die im Büro des *DDGLC*-Projekts aufbewahrten Zettelkästen eine umfassende und repräsentative, semantisch strukturierte Belegsammlung zu griechischen Konjunktionen, Partikeln und Präpositionen im Koptischen darstellen. Um diese wertvolle lexikographische Quelle möglichst rasch einer weiteren Öffentlichkeit bekanntzumachen, wurde eine Liste erarbeitet, in der die Lemma-Struktur der Zettelkästen expliziert ist:
<http://www.uni-leipzig.de/~ddglc/docs/GertrudBauerCardindex.pdf>

Die weitere Planung geht dahin, den gesamten Zettelbestand nach Vorbild des Zettelarchivs des Berliner Altägyptischen Wörterbuchs online zugänglich zu machen. Dazu wurden die Zettel eingescannt und die Scans mit Kennzeichnung ihrer Stellung innerhalb der Lemma-Struktur in eine Datenbank importiert (Fig. 9). Bis auf einige Nacharbeiten ist die elektronische Zettel-Datei fertig ausgearbeitet und bereit dazu, online gestellt zu werden. Ein abgeschlossenes Manuskript Getrud Bauers wurde in elektronischen Text umgewandelt und soll ebenfalls der Öffentlichkeit zugänglich gemacht werden.

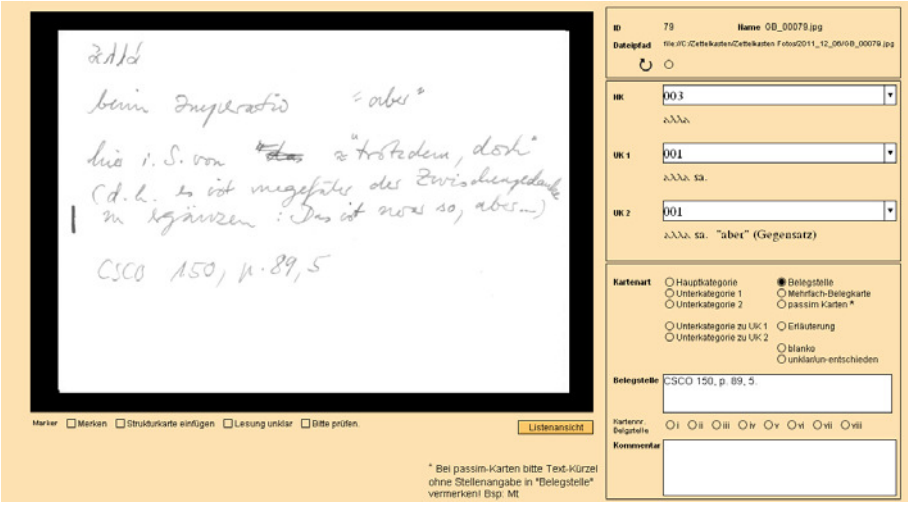


Fig. 9. Elektronische Zetteldatei der Bauer-Zettelkästen zu griechischen Konjunktionen, Partikeln und Präpositionen im Koptischen

3. BIBLIOGRAPHIE

ALLBERRY, C. R. C., 1937: Greek and Latin Words in the Coptic Manichaean Papyri, in: *Proceedings of the 5th International Congress of Papyrology*, Oxford, 20.

ALMOND, M., 2010: Language Change in Greek Loaned Verbs, in: *Lingua Aegyptia* 18, 19-31.

ALMOND, M., (in Vorbereitung): Greek Adjectives borrowed into Coptic, in: DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

ALMOND, M., 2011: *A Comparative Study of Loanword Integration in Fourth-Century Coptic Literature*, PhD Macquarie University Sydney.

BANDT, C., 2007: *Der Traktat „Vom Mysterium der Buchstaben“*, TU 162, Berlin [u.a.].

BAUER, G., 1975: *Konkordanz der nichtflektierten griechischen Wörter im bohairischen Neuen Testament*, GOF VI/6, Wiesbaden.

BEHLMER, H., 1997/8: Index der Lehnwörter und Namen in Amélineau, Œuvres de Shenoudi, in: *Enchoria* 24, 1-33.

BJÖRNESJÖ, S., 1996: L’arabisation de l’Égypte: le témoignage papyrologique, in: *Égypte – Monde Arabe* 27/28, 93-106.

BLOK, H. P., 1927: Die griechischen Lehnwörter im Koptischen, in: *ZÄS* 62, 49-60.

BÖHLIG, A., 1953a: *Ein Lexikon der griechischen Wörter im Koptischen. Die griechisch-lateinischen Lehnwörter in den koptischen manichäischen Texten*, Studien zur Erforschung des christlichen Ägyptens Heft 1, 1. Aufl., München.

BÖHLIG, A., 1953b: Griechische Deponentien im Koptischen, in: *Aegyptus* 33, 91-96.

BÖHLIG, A., 1954a: *Ein Lexikon der griechischen Wörter im Koptischen. Die griechisch-lateinischen Lehnwörter in den koptischen manichäischen Texten*, Studien zur Erforschung des christlichen Ägyptens Heft 1, 2. Aufl., München.

BÖHLIG, A., 1954b: *Die griechischen Lehnwörter im sahidischen und bohairischen Neuen Testament*, Studien zur Erforschung des christlichen Ägyptens Heft 2, 1. Aufl., München.

BÖHLIG, A., 1954c: *Die griechischen Lehnwörter im sahidischen und bohairischen Neuen Testament. Register und Vergleichstabellen zu Heft 2*, Studien zur Erforschung des christlichen Ägyptens Heft 2a, 1. Aufl., München.

BÖHLIG, A., 1955: Beiträge zur Form griechischer Wörter im Koptischen, in: *ZÄS* 80, 90-97.

BÖHLIG, A., 1956: Die Fortführung der Arbeit am Lexikon der griechischen Wörter im Koptischen, in: *Wissenschaftliche Zeitschrift der Martin-Luther-Universität Halle-Wittenberg* 5/4, 655-657.

BÖHLIG, A., 1958a: *Ein Lexikon der griechischen Wörter im Koptischen. Die griechisch-lateinischen Lehnwörter in den koptischen manichäischen Texten*, Studien zur Erforschung des christlichen Ägyptens Heft 1, 3. Aufl., München.

BÖHLIG, A., 1958b: *Die griechischen Lehnwörter im sahidischen und bohairischen Neuen Testament*, Studien zur Erforschung des christlichen Ägyptens Heft 2, 2. Aufl., München.

BÖHLIG, A., 1958c: *Die griechischen Lehnwörter im sahidischen und bohairischen Neuen Testament. Register und Vergleichstabellen zu Heft 2*, Studien zur Erforschung des christlichen Ägyptens Heft 2a, 2. Aufl., München.

BÖHLIG, A., 1960: Griechische Elemente im Koptischen als Zeugnis für die Geschichte der griechischen Sprache, in: DÖLGER, F. & H.-G. BECK (Hrsg.), *Akten des XI. Internationalen Byzantinistenkongresses, München 1958*, München, 62-67.

BÖHLIG, A., 1962: Griechische Wörter im Koptischen, in: *Forschungsinformation Halle* 1, Abt. D, Bl. 24.

BÖHLIG, A., 1995: Die Form der griechischen Verben in den Texten von Nag Hammadi, in: FLUCK, C. *et al.* (Hrsg.), *Divitiae Ægypti. Koptologische und verwandte Studien zu Ehren von Martin Krause*, Wiesbaden, 19-28.

BRESCIANI, E. & R. PINTAUDI, 1987: Textes démotico-grecs et grecodémotiques des ostraca de Medinet Madi: un problème de bilinguisme, in: VLEEMING, S. P. (ed.), *Aspects of Demotic Lexicography. Acts of the Second International Conference for Demotic*

Studies, Leiden, 19-21 September 1984, Studia demotica 1, Leuven [u.a.], 123-126.

BRUNSCH, W., 1983: Zum vermeintlichen Genuswechsel im Koptischen durch den Einfluß des Griechischen, in: *ZÄS* 110, 122-126.

BURCHARDT, M., 1909/10: *Die altkanaanäischen Fremdworte und Eigennamen im Aegyptischen*, Leipzig.

CHARRON, R., 1992: *Concordances des textes de Nag Hammadi: Le Codex VII*, Bibliothèque Copte de Nag Hammadi, section concordances 1, Québec [u.a.].

CHARRON, R., 1995: *Concordances des textes de Nag Hammadi: Le Codex III*, Bibliothèque Copte de Nag Hammadi, section concordances 3, Québec [u.a.].

CHERIX, P., 1993: *Concordances des textes de Nag Hammadi: Le Codex VI*, Bibliothèque Copte de Nag Hammadi, section concordances 2, Québec [u.a.].

CHERIX, P., 1995: *Concordances des textes de Nag Hammadi: Le Codex I*, Bibliothèque Copte de Nag Hammadi, section concordances 4, Québec [u.a.].

CHERIX, P., 2000: *Lexique analytique du parchemin pBodmer VI version copte du Livre des Proverbes*, Instruments pour l'étude des langues de l'orient ancien 2, Lausanne.

CLACKSON, S. et al., 1998: *Dictionary of Manichaean texts*. Vol. I: *Texts from the Roman Empire (Texts in Syriac, Greek, Coptic and Latin)*, Corpus Fontium Manichaeorum, Subsidia II, Turnhout.

CLARYSSE, W., 1987: Greek loan-words in Demotic, in: VLEEMING, S. P. (ed.), *Aspects of Demotic lexicography. Acts of the Second International Conference for Demotic Studies, Leiden, 19-21 September 1984*, Studia demotica 1, Leuven [u.a.], 9-33.

CLARYSSE, W., 1993: Egyptian Scribes Writing Greek, in: *CdE* 68, 186-201.

COQUIN, R.-G., 1993: Langue et littérature copte, in: *Christianismes orientaux. Introduction à l'étude des langues et des littératures*, Paris, 169-217.

DELATTRE, A. *et al.*, 2012: Écrire en arabe et en copte. Le cas de deux lettres bilingues, in: *CdE* 87, 170-188.

DEMARIA, S., 2005: Die griechischen Entlehnungen in den koptischen manichäischen Texten, in: VAN TONGERLOO, A. & L. CIRILLO (eds.), *Il Manicheismo. Nuove Prospettive della Richerca (Quinto Congresso Internazionale di Studi sul Manicheismo, Napoli, 2-8 Settembre 2001)*, Manichaean Studies 5, Turnhout, 96-114.

DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

DRAGUET, R., 1960: *Index copte et grec-copte de la concordance du Nouveau Testament sahidique*, CSCO 196, Subsidia 16, Louvain.

DRESCHER, J., 1969/76: Graeco-Coptica, parts I-III, in: *Le Muséon* 82, 85-100; 83, 139-155; 89, 307-321.

EMMEL, ST., 2004: *Shenoute's Literary Corpus*, CSCO 599-600, Subsidia 111-112, Louvain.

FEDER, F., 2004: Der Einfluß des Griechischen auf das Ägyptische in ptolemäisch-römischer Zeit, in: SCHNEIDER, TH. (Hrsg.), *Das Ägyptische und die Sprachen Vorderasiens, Nordafrikas und der Ägäis. Akten des Basler Kolloquiums zum ägyptisch-nichtsemitischen Sprachkontakt, Basel 9.-11. Juli 2003*, AOAT 310, Münster, 509-521.

FEWSTER, P., 2002: Bilingualism in Roman Egypt, in: ADAMS, J. N. *et al.*, *Bilingualism in Ancient Society. Language Contact and the Written Text*, Oxford, 220-245.

FÖRSTER, H., 2002: *Wörterbuch der griechischen Wörter in den koptischen dokumentarischen Texten*, Texte und Untersuchungen zur altchristlichen Literatur 148, Berlin [u.a.].

FUNK, W.-P., 1982: Polis, Polites und Politeia im Koptischen. Zu einigen Fragen des einschlägigen koptischen Lehnwortschatzes, in: WELSKOPF, E. CH. (Hrsg.), *Das Fortleben altgriechischer sozialer Typenbegriffe in den Sprachen der Welt*, 2. Teil, Berlin, 283-320.

FUNK, W.-P., 1984: Bemerkungen zum Sprachvergleich Griechisch-Koptisch, in: NAGEL, P. (Hrsg.), *Graeco-Coptica. Griechen und Kopten im byzantinischen Ägypten*, Wissenschaftliche Beiträge der Martin-Luther-Universität Halle-Wittenberg 48 (I 29), Halle/Saale, 147-180.

FUNK, W.-P., 1988: Dialects wanting homes: A numerical approach to the early varieties of Coptic, in: FISIAK, J. (ed.), *Historical Dialectology: Regional and Social*, Trends in Linguistics, Studies and Monographs 37, Berlin, 149-192.

FUNK, W.-P., 1991: Dialects, Morphology of Coptic, in: *The Coptic Encyclopedia* 8, 101-108.

FUNK, W.-P., 1997: *Concordances des textes de Nag Hammadi: Le Codex VIII et IX*, Bibliothèque Copte de Nag Hammadi, section concordances 5, Québec [u.a.].

FUNK, W.-P., 2000: *Concordances des textes de Nag Hammadi: Le Codex X et XIa*, Bibliothèque Copte de Nag Hammadi, section concordances 6, Québec [u.a.].

FUNK, W.-P., 2007: *Concordance of Shenoute's Canons*, Privatdruck Battlefield Québec City.

FUNK, W.-P. & P.-H. POIRIER, 2006: *Concordances des textes de Nag Hammadi: Le Codex XIb, XII et XIII*, Bibliothèque Copte de Nag Hammadi, section concordances 7, Québec [u.a.].

GASELEE, ST., 1914: ΕΛΛΗΝ in Coptic, in: *JEA* 1, 207-208.

GASELEE, ST., 1916: The Pronunciation of Greek Words in Christian Egypt, in: *Classical Review* 30, 6-7.

GASELEE, ST., 1929/30: Greek words in Coptic, in: *Byzantinische Zeitschrift* 30, 224-228.

GESSMAN, A. M., 1976: The Birthdate of the Coptic Script, in: *The University of South Florida Language Quarterly* 14/2-3, 2-4.

GESSMAN, A. M., 1986: The Birthdate of the Coptic Script, in: *Coptologia* 7, 57-66.

GIRGIS, W. A. [ANBA GEORGIOS], 1963-2001: Greek Loan Words in Coptic, parts I-VII, in: *BSAC* 17 (1963/4), 63-73; 18 (1965/6), 71-96; 19 (1967/8), 57-87; 20 (1969/70), 53-67; 21 (1971/73), 33-53; 23 (1976/78), 199-220; 30 (1991), 77-92; 40 (2001), 61-88.

GIRGIS, W. A. [ANBA GEORGIOS], 2010: *Greek Words in Coptic Usage*, Cairo.

GODRON, G., 1983: *limên* 'portrait', 'image', in: *BSAC* 25, 1-52.

GROSSMAN, E., 2007: Worknotes on the Syntax of Nitrian Bohairic: A Hitherto Unnoticed Circumstantial Conversion, in: BOSSON, N. & A. BOUD'HORS (eds.), *Actes du Huitième Congrès International*

d'Études Coptes, Paris, 28 juin – 3 juillet 2004, OLA 163, Leuven [u.a.], 711-726.

GROSSMAN, E., 2009: The syntax of argument clauses in Sahidic Coptic, in: *ZÄS* 136, 19-32.

GROSSMANN, E. & T. S. RICHTER, (in Vorbereitung): Lexical borrowing into Coptic. A case study in loanword typology, in DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

GROSSMAN, E. *et al.* (eds.), (in Vorbereitung): *Egyptian-Coptic Linguistics in Typological Perspective*, Trends in Linguistics. Studies and Monographs, Berlin – New York.

HASITZKA, M. R. M. & H. SATZINGER, 2004/5: Index der gräkokoptischen Wörter in nichtlitarischen Texten oder: Was ist ein Wörterbuch?, in: *Enchoria* 29, 19-31.

HASZNOS, A., 2012: *Graeco-Coptica. Greek and Coptic Clause Patterns*, GOF IV/52, Wiesbaden.

HEBBELYNCK, A., 1900/1901: Les Mystères des Lettres Grecques. Texte Copte, Traduction, Notes, in: *Le Muséon* 19 (1900), 5-36, 105-136, 269-300; 20 (1901), 5-33. 369-414.

HINTZE, F., 1984: Eine Klassifizierung der koptischen Dialekte, in: *Studien zu Sprache und Religion Ägyptens. Zu Ehren von Wolfhart Westendorf überreicht von seinen Freunden und Schülern*, Göttingen, Bd. 1, 411-432.

HOCH, J., 1994: *Semitic Words in Egyptian Texts of the New Kingdom and Third Intermediate Period*, Princeton.

HOFFMANN, F., 2000: *Ägypten. Kultur und Lebenswelt in römischer Zeit. Eine Darstellung nach den demotischen Quellen*, Berlin.

HOPFNER, TH., 1918: *Über Form und Gebrauch der griechischen Lehnwörter in der koptisch-sa'idischen Apophthegmenversion*, DAAW 62/2, Wien.

JERNSTEDT, P. V., 1929: Graeco-Coptica, in: *ZÄS* 64, 122-135.

KAHLE, P. E., 1954: *Bala'izah. Coptic Texts from Deir el-Bala'izah in Upper Egypt*, Oxford – London.

KAPSOMENOS, S. G., 1953: Das Griechische in Ägypten, in: *Museum Helveticum* 10, 248-263.

KASSER, R., 1966: La penetration des mots grecs dans la langue copte, in: *Wissenschaftliche Zeitschrift der Martin-Luther-Universität Halle-Wittenberg* 15, 419-425.

KASSER, R., 1991a: Dialects, in: *The Coptic Encyclopedia* 8, 87-97.

KASSER, R., 1991b: Dialects, Grouping and Major Groups of, in: *The Coptic Encyclopedia* 8, 97-101.

KASSER, R., 1991c: Geography, Dialectal, in: *The Coptic Encyclopedia* 8, 133-141.

KASSER, R., 1991d: Vocabulary, Copto-Greek, in: *The Coptic Encyclopedia* 8, 215-222.

KASSER, R. & W. VYCICHL, 1967: *Dictionnaire auxiliaire, étymologique et complet de la langue copte*, Fascicule I: ⲁ – ⲃⲁⲩⲕⲁⲗⲓⲟⲛ, Genève.

KNIGGE, C., 2004: Sprachkontakte und lexikalische Interferenz im ersten vorchristlichen Jahrtausend, in: SCHNEIDER, TH. (Hrsg.), *Das Ägyptische und die Sprachen Vorderasiens, Nordafrikas und der Ägäis. Akten des Basler Kolloquiums zum ägyptisch-nichtsemitischen Sprachkontakt, Basel 9.-11. Juli 2003*, AOAT 310, Münster, 33-88.

KRAUSE, M., 1980: Koptische Literatur, in: *LÄ* Bd. III, 694-728.

LEFORT, L.-TH., 1934: Le copte: source auxiliaire du grec, in: *Mélanges Bidez*, AIP 2, Brussels, 569-578.

LEFORT, L.-TH., 1948: EIMHTI dans le NT sahidique, in: *Le Muséon* 61, 153-170.

LEFORT, L.-TH., 1950a: *Concordance du Nouveau Testament sahidique,* I: *Les mots d'origine grecque*, CSCO 124, Subsidia 1, Louvain.

LEFORT, L.-TH., 1950b: Gréco-Copte, in: MALININE, M. (ed.), *Coptic studies in honor of Walter Ewing Crum*, The Bulletin of the Byzantine Institute 2, Boston, 65-71.

LEMM, O. V., 1903: *Das Triadon. Ein sahidisches Lehrgedicht mit arabischer Übersetzung*, St. Pétersbourg.

LOPRIENO, A., 1995: *Ancient Egyptian. A linguistic Introduction*, Cambridge.

LOPRIENO, A., 2001: From Ancient Egyptian to Coptic, in: HASPELMATH, M. *et al.* (eds.), *Language Typology and Language*

Universals / Sprachtypologie und sprachliche Universalien / La Typologie des langues et les universaux linguistiques: An International Handbook / Ein internationales Handbuch / Manuel international, Berlin – New York, 1742-1761.

LOPRIENO, A. & M. MÜLLER, 2012: Ancient Egyptian and Coptic, in: FRAIZYNGIER, Z. & E. SHAY (eds.), *The Afroasiatic Languages*, Cambridge Language Surveys, Cambridge, 102-144.

MCBRIDE, D., 1989: The Development of Coptic: Late-pagan Language of Synthesis in Egypt, in: *Journal of the Society for the Study of Egyptian Antiquities* 19, 89-111.

MEYER, M. & R. SMITH, 1994: *Ancient Christian Magic. Coptic Texts of Ritual Power*, Princeton/New Jersey.

NAGEL, P., 1971: Die Einwirkung des Griechischen auf die Entstehung der koptischen Literatursprache, in: ALTHEIM, F. & R. STIEHL (Hrsg.), *Christentum am Roten Meer*, Berlin – New York, 327-355.

NAGEL, P., 1983: *Das Triadon. Ein sahidisches Lehrgedicht des 14. Jahrhunderts*, Halle (Saale).

NAGEL, P., 2013: Einleitung: Alexander Böhlig – ein Leben für die Wissenschaft vom christlichen Orient, in: BÖHLIG, A. (†), *Die Bibel bei den Manichäern und verwandte Studien*, hrsg. von NAGEL, P. & S. G. RICHTER, Nag Hammadi and Manichaean Studies 80, Leiden – Boston, 1-17.

ORÉAL, E., 1999: Contact linguistique. Le cas du rapport entre le grec et le copte, in: *Lalies* 19, 289-306.

ORLANDI, T., 1995: La documentation patristique copte. Bilan et prospectives, in: FREDOUILLE, J.-CL. & R.-M. ROBERGE (eds.), *La documentation patristique. Bilan et prospectives*, Québec – Paris, 127-147.

ORLANDI, T., 1997: Letteratura copta e cristianesimo nazionale egiziano, in: CAMPLANI, A. (ed.), *L'Egitto cristiano. Aspetti e problemi in età tardo-antica*, Roma, 39-120.

ORLANDI, T., 1998: Koptische Literatur, in: KRAUSE, M. (Hrsg.), *Ägypten in spätantik-christlicher Zeit. Einführung in die koptische Kultur*, Wiesbaden, 117-147.

ORLANDI, T., 2004: Coptic Monastic Literature. The Forgotten Names, in: BIELAWSKI, M. & D. HOMBERGEN (eds.), *Il Monachesimo tra eredità e aperture. Atti del simposio „Testi e temi nella tradizione del*

monachesimo cristiano" per il 50° anniversario dell'Istituto Monastico di Sant'Anselmo, Rome 28 maggio - 1 giugno 2002, Roma, 175-195.

ORLANDI, T., 2005: La letteratura copta e la storia dell'Egitto cristiano, in: SINISCALCO, P. (ed.), *Le antiche Chiese orientali. Storia e letteratura*, Roma, 85-117.

PAPACONSTANTINOU, A., 2007: «They shall speak the Arabic language and take pride in it»: Reconsidering the fate of Coptic after the Arab conquest, in: *Le Muséon* 120, 273-299.

PAPACONSTANTINOU, A., 2012: Why did Coptic fail where Aramaic succeeded? Linguistic developments in Egypt and the Near East after the Arab conquest, in: MULLEN, A. & P. JAMES (eds.), *Multilingualism in the Graeco-Roman Worlds*, Cambridge, 58-76.

PEREMANS, W., 1964: Über die Zweisprachigkeit im ptolemäischen Ägypten, in: *Studien zur Papyrologie und antiken Wirtschaftsgeschichte. Friedrich Oertel zum achtzigsten Geburtstag gewidmet*, Bonn, 49-60.

PEREMANS, W., 1983: Le bilinguisme dans les relations gréco-égyptiennes sous les Lagides, in: VAN'T DACK, E. *et al.* (eds.), *Egypt and the Hellenistic World*, Stud. Hellen. 27, Leuven, 253-280.

POLOTSKY, H. J., 1950: Modes grecs en copte?, in: MALININE, M. (ed.), *Coptic studies in honor of Walter Ewing Crum*, The Bulletin of the Byzantine Institue 2, Boston, 73-90.

QUACK, J. F., 2005: Zu den vorarabischen semitischen Lehnwörtern des Koptischen, in: BURTEA, B. *et al.* (eds.), *Studia semitica et Semitohamitica. Festschrift für Rainer Voigt anläßlich seines 60. Geburtstages am 17. Januar 2004*, AOAT 317, Münster, 307-338.

QUAEGEBEUR, J., 1974: The Study of Egyptian Proper Names in Greek Transcription. Problems and Perspectives, in: *Onoma* 18, 403-420.

QUAEGEBEUR, J., 1982: De la préhistoire de l'écriture Copte, in: *OLP* 13, 125-136.

RAHLFS, A., 1900: ⲑⲁⲗⲁⲥⲥⲁ im Koptischen, in: *ZÄS* 38, 152-153.

RAHLFS, A., 1912. Griechische Wörter im Koptischen, in: *SPAW*, 1036-1046.

REINTGES, CH., 2001: Code-mixing strategies in Coptic Egyptian, in: *Lingua Aegyptia* 9, 193-237.

REINTGES, CH., 2004: Coptic Egyptian as a Bilingual Language Variety, in: BÁDENAS DE LA PEÑA, P. *et al.* (eds.), *Lenguas en contacto: el testimonio escrito*, Madrid, 69-86.

REMONDON, R., 1964: Problèmes du bilinguisme dans l'Égypte lagide (UPZ I 148), in: *CdE* 39, 126-146.

RICHTER, T. S., 2001: Arabische Lehnworte und Formeln in koptischen Rechtsurkunden, in: *The Journal of Juristic Papyrology* 31, 75-89.

RICHTER, T. S., 2006: Coptic[, Arabic loanwords in], in: *Encyclopedia of Arabic Language and Linguistics*, vol. 1, Leiden, 595-601.

RICHTER, T. S., 2008a: *Rechtssemantik und forensische Rhetorik. Untersuchungen zu Wortschatz, Stil und Grammatik der Sprache koptischer Rechtsurkunden*, 2. überarbeitete Aufl., Philippika 20, Wiesbaden.

RICHTER, T. S., 2008b: Coptic letters, in: GROB, E. M. & A. KAPLONY (eds.), *Documentary letters from the Middle East. The evidence in Greek, Coptic, South Arabian, Pehlevi, and Arabic (1st-15th c CE)*, Asiatische Studien 62/3, special issue, Bern, 739-770.

RICHTER, T. S., 2009a: Greek, Coptic, and the 'Language of the Hijra'. Rise and Decline of the Coptic Language in Late Antique and Medieval Egypt, in: COTTON, H. M. *et al.* (eds.), *From Hellenism to Islam: Cultural and Linguistic Change in the Roman Near East*, Cambridge, 398-443.

RICHTER, T. S., 2009b: What Kind of Alchemy is Attested by Tenth-Century Coptic Manuscripts?, in: *Ambix. Journal of the Society for the History of Alchemy and Chemistry* 56/1 (March), 23-35.

RICHTER, T. S., 2010: Language choice in the Qurra papyri, in: PAPACONSTANTINOU, A. (ed.), *The multilingual experience in Egypt, from the Ptolemies to the Abbasides*, Farnham, 189-219.

RICHTER, T. S., (in Vorbereitung): The other Story: Lexical borrowing into Coptic from Arabic, in: DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

ROCHETTE, B., 1996: Sur le bilingualisme dans l'Égypte gréco-romaine, in: *CdE* 71, 153-168.

RUTHERFORD, I., 2010: Bilingualism in Roman Egypt? Exploring the Archive of Phatres of Narmuthis, in: EVANS, T. & D. OBBINK (eds.), *The Language of the Papyri*, Oxford, 198-207.

SATZINGER, H., 1970: ⲕⲁⲑⲁⲣⲱⲥ ⲕⲁⲓ ⲁⲡⲟⲕⲣⲟⲧⲱⲥ in koptischen Urkunden, in: *CdE* 45, 417-420.

SATZINGER, H., 1975: The Old Coptic Schmidt Papyrus, in: *JARCE* 12, 37-50.

SATZINGER, H., 1984: Die altkoptischen Texte als Zeugnisse der Beziehungen zwischen Ägyptern und Griechen, in: NAGEL, P. (Hrsg.), *Graeco-Coptica. Griechen und Kopten im byzantinischen Ägypten*, Wissenschaftliche Beiträge der Martin-Luther-Universität Halle-Wittenberg 48 (I 29), 137-146.

SATZINGER, H., 1985: On the Origin of the Sahidic Dialect, in: ORLANDI, T. & F. WISSE (eds.), *Acts of the Second International Congress of Coptic Studies, Roma 22-26 September 1980*, Roma, 307-312.

SATZINGER, H., 1990: On the Prehistory of the Coptic Dialects, in: GODLEWSKI, W. (ed.), *Coptic Studies. Acts of the Third International Congress of Coptic Studies, Warsaw, 20-25 August 1984*, Warszawa, 413-416.

SATZINGER, H., 1991: Old-Coptic, in: *The Coptic Encyclopedia* 8, 169-175.

SATZINGER, H., 2003: Das Griechisch, aus dem die koptischen Alphabete stammen, in: BELTZ, W. *et al.* (Hrsg.), *Sprache und Geist. Peter Nagel zum 65. Geburtstag*, Hallesche Beiträge zur Orientwissenschaft 35, 201-213.

SCHENKE, H.-M., 1981: *Das Matthäus-Evangelium im mittelägyptischen Dialekt des Koptischen (Codex Scheide)*, TU 127, Berlin.

SCHENKE, H.-M., 2001: *Das Matthäus-Evangelium im mittelägyptischen Dialekt des Koptischen (Codex Schøyen)*, Coptic Papyri in the Schøyen Collection 1, Oslo.

SCHILLER, A. A., 1950: ⲕⲁⲛⲱⲛ and ⲕⲁⲛⲱⲛⲓⲍⲉ in the Coptic Texts, in: MALININE, M. (ed.), *Coptic Studies in Honor of Walter Ewing Crum*, The Bulletin of the Byzantine Institute 2, Boston, 175-184.

SCHNEIDER, TH., 2004: Nichtsemitische Lehnwörter im Ägyptischen. Umriss eines Forschungsgebietes, in: SCHNEIDER, TH. (Hrsg.), *Das Ägyptische und die Sprachen Vorderasiens, Nordafrikas und der Ägäis.*

Akten des Basler Kolloquiums zum ägyptisch-nichtsemitischen Sprachkontakt, Basel 9.-11. Juli 2003, AOAT 310, Münster, 11-31.

SCHWARTZE, G. M. & H. STEINTHAL, 1850: *Koptische Grammatik*, Berlin.

SHISHA-HALEVY, A. 2009: Work-Notes on Shenoute's Rhetorical Syntax: ϵϣϫϵ and ⲁⲣⲁ. Suspension of Disagreement, Irony and reductio ad absurdum, in: GIEWEKEMEYER, A. *et al.* (eds.), *Liber Amicorum: Jürgen Horn zum Dank*, GM Beihefte 5, Göttingen, 113-129.

SIDARUS, A., 2008: Plurilinguisme en Égypte sous la domination gréco-romaine, in: *Journal of Coptic Studies* 10, 183-202.

SIEGERT, F., 1982: *Nag-Hammadi-Register. Wörterbuch zur Erfassung der Begriffe in den koptisch-gnostischen Schriften von Nag-Hammadi*, Wissenschaftliche Untersuchungen zum Neuen Testament 26, Tübingen.

TILL, W. C., 1951a: *Die Arzneikunde der Kopten*, Berlin.

TILL, W. C., 1951b: ϵⲗϵⲩⲑϵⲣⲟⲥ = unbescholten, in: *Le Muséon* 64, 251-259.

TORALLAS TOVAR, S., 2004: Egyptian Lexical Interference in the Greek of Byzantine and Early Islamic Egypt, in: SIJPESTEIJN, P. M. & L. SUNDELIN (eds.), *Papyrology and the History of Early Islamic Egypt*, Islamic History and Civilization 55, Leiden, 163-198.

TORALLAS TOVAR, S., 2005: *Identidad lingüística e identidad religiosa en el Egipto grecorromano*, Barcelona.

TORALLAS TOVAR, S., (in Vorbereitung): Egyptian borrowing into Greek: the problem of the corpus, in: DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

TUBACH, J., 1999a: Bemerkungen zur geplanten Wiederaufnahme des Wörterbuchprojekts „Griechische Lehnwörter im Koptischen" in Halle, in: EMMEL, ST. *et al.* (Hrsg.), *Ägypten und Nubien in spätantiker und christlicher Zeit. Akten des 6. Internationalen Koptologenkongresses Münster, 20.-26. Juli 1996*, Bd. 2: *Schrifttum, Sprache, Gedankenwelt*, Sprachen und Kulturen des Christlichen Orients 6/2, Münster, 405-419.

TUBACH, J., 1999b: Griechische Lehnwörter in den koptischen Manichaica. Zur Problematik eines Lehnwortschatzes in einer Übersetzung aus einem anderen Kulturbereich, in: GRUNERT, ST. & I. HAFEMANN (Hrsg.), *Textcorpus und Wörterbuch. Aspekte zur ägyptischen Lexikographie*, PÄ 14, Leiden [u.a.], 329-343.

TUDOR, B., 2011: *Christian Funerary Stelae of the Byzantine and Arab Periods from Egypt*, Marburg.

VERGOTE, J., 1984: Bilinguisme et calques (translation loan-words) en Égypte, in: *Atti del XVII congresso internazionale di papirologia, Napoli*, vol. 3, 1385-1389.

VIERROS, M., 2012: *Bilingual Notaries in Hellenistic Egypt. A Study of Greek as a Second Language*, Collectanea Hellenistica 5, Brussel.

VITTMANN, G., 1996: Semitisches Sprachgut im Demotischen, in: *WZKM* 86, 435-447.

VITTMANN, G., 2003: *Ägypten und die Fremden im ersten vorchristlichen Jahrtausend*, Kulturgeschichte der antiken Welt 97, Mainz.

VYCICHL, W., 1983: *Dictionnaire étymologique de la langue copte*, Leuven.

WEISS, H.-F., 1966: Zum Problem der griechischen Fremd- und Lehnwörter in den Sprachen des christlichen Orients, in: *Helikon* 6, 183-209.

WEISS, H.-F., 1968: Beobachtungen zur Frage der griechischen Komponente in der Sprache des Schenute, in: NAGEL, P. (Hrsg.), *Probleme der koptischen Literatur*, Halle, 173-185.

WEISS, H.-F., 1969: Ein Lexikon der griechischen Wörter im Koptischen, in: *ZÄS* 96, 79-80.

WEISS, H.-F., 1972: Zum Problem der Einwirkung des Griechischen auf die Sprachen des byzantinischen Orients, in: NAGEL, P. (Hrsg.), *Von Nag Hammadi bis Zypern*, Halle, 28-34.

WINAND, J., (in Vorbereitung): Grammatical integration of loan-verbs in Late Egyptian, in: DILS, P. *et al.* (eds.), (in Vorbereitung): *Language Contact and Bilingualism in Antiquity: What Linguistic Borrowing Into Coptic Can Tell Us About It. Papers Read on the DDGLC Inaugural Conference, Leipzig, Saxonian Academy of Sciences, April 2010*, Lingua Aegyptia, Studia Monographica.

ZABOROWSKI, J. R., 2005: *The Coptic Matyrdom of John of Phanijôit. Assimilation and Conversion to Islam in Thirteenth-Century Egypt*, Leiden.

ZABOROWSKI, J. R., 2008: From Coptic to Arabic in Medieval Egypt, in: *Medieval Encounters* 14, 15-40.

HISTORISCHER WORTGEBRAUCH UND THEMENGESCHICHTE. GRUNDFRAGEN, CORPORA, DOKUMENTATIONSFORMEN

THOMAS GLONING

1. Ausgangspunkte, Gegenstand, Fragestellungen

Bei der Benutzung der historischen **Wörterbücher** des Deutschen macht man immer wieder zwei Erfahrungen: Zum einen sind verschiedene **thematische Felder** des Sprachgebrauchs in sehr unterschiedlicher Dichte, mit unterschiedlich zuverlässiger Erwartbarkeit und in schwankender Differenzierung bearbeitet, sowohl im Hinblick auf die Chronologie als auch im Hinblick auf die Verteilung über das Varietätenspektrum. Wörter wir *Kläranlage*, *Atomstrom*, *Frauenwahlrecht*, *Großraumbüro*, *Urning* oder *Naturschutz* sind einige Beispiele aus meiner eigenen Nachschlagebiographie. Der Hintergrund dafür ist die Tatsache, dass bei der Planung von lexikographischen Beleggrundlagen bzw. von Textcorpora thematische Gesichtspunkte bislang offenbar nicht systematisch berücksichtigt werden konnten.

Eine zweite Erfahrung besteht darin, dass es derzeit in der Regel nicht möglich ist, den spezifischen Wortschatz einzelner **thematischer Felder** bzw. themenspezifische Verwendungsweisen in den Wörterbüchern systematisch ›**abzufragen**‹. Wenn man sich zum Beispiel fragt, wie sich der Wortgebrauch im Bereich der Musik, des Sports, der Sexualität, des Kriegs usw. in den letzten 200 Jahren entwickelt hat oder welche lexikalischen Mittel zum Thema Inflation in den 1920er Jahren gebraucht wurden, dann ist man auf monographische Untersuchungen angewiesen, soweit es sie denn gibt. Die Wörterbücher helfen bei solchen Interessen nur dann, wenn man die ›Suchadressen‹, also die einschlägigen Stichwörter bereits kennt und dann wiederum nur mit den oben genannten Einschränkungen im Hinblick auf Fragen der thematischen Abdeckung.

Ich möchte diese beiden Erfahrungen nicht als Kritik an den vorhandenen und äußerst verdienstvollen Wörterbüchern, sei es in gedruckter oder digitaler Form, formuliert und verstanden wissen, sondern als ein Ausgangspunkt für die **Frage**, wie man **themengeschichtliche Gesichtspunkte**, die aufs Engste auch mit kultur-, sozial- und ideengeschichtlichen Entwicklungen verbunden sind, stärker in **zukünftigen Formen der lexikographisch-lexikologischen Dokumentation integrieren** kann.

Themen sind zentrale Aspekte der Kommunikations- und der Ideengeschichte, sie sind – wie gerade schon betont – eng verwoben mit Fragen der kulturellen, der sozialen, der technischen und im Hinblick auf Themenkarrieren auch der medialen Entwicklung. Mit der Erweiterung des sprachwissenschaftlichen Gebietes um eine im weitesten Sinne pragmatische Perspektive, also um Fragen des sprachlichen Handelns mit Texten und in Gesprächen, wurde auch der Themen-Begriff als ein pragmatischer Grundbegriff aufgenommen.[1] Thematisches Wissen und kommunikative Fähigkeiten im Bereich des Themenmanagements sind wichtige Elemente der Alltagskommunikation und der darauf beruhenden sozialen Ordnung, dies ein Gedanke, der schon früh in der ethnomethodologischen Gesprächsforschung betont wurde (ADATO 1971). Thematisches Wissen hängt weiterhin eng zusammen mit dem jeweils zeitbedingten Wissen bzw. den zeitbedingten Annahmen und Auffassungen über die Welt und die soziale Ordnung. Schließlich sind Themen auch ein zentraler Parameter für die Strukturierung von lexikalischen Aspekten der privaten, der öffentlichen und der fachlichen Kommunikation, auch in einer sprachhistorischen Perspektive.

Der **Themenhaushalt** in seiner historischer Entwicklung, seiner dynamischen Entfaltung und in seinen Zusammenhängen mit lexikalischen Mitteln und sprachlichen Verfahren wird damit ein wichtiger Gegenstand auch für **Corpus-Planungen**, für sprachgeschichtliche Analysen und für die lexikographische Dokumentation.

Im vorliegenden Beitrag stehen drei **Fragestellungen** im Mittelpunkt:

(i) Welche Rolle spielt die kommunikationsgeschichtliche Entwicklung von Themen für die Entwicklung des Wortgebrauchs und für thematisch geprägte Wortschatzsektoren zu einem bestimmten historischen Zeitpunkt, im historischen Längsschnitt und im Hinblick auf das Varietätengefüge einer Sprache?

(ii) Wie kann man Kriterien der Themengeschichte für die historische Lexikographie und Lexikologie fruchtbar machen (insbesondere beim Aufbau von Textcorpora für die historische Lexikographie und im weiteren Sinne auch für andere, zum Beispiel monographische Formen der Erschließung historischer Wortschatzentwicklungen)?

---

[1] Für eine grundlegende Darstellung zum Themenbegriff und zur Rolle von Themen für die Organisation von Texten, Gesprächen und zusammenhängenden Kommunikationsverläufen siehe nun FRITZ (2013, Kap. 4).

(iii) Welche Arbeitsformen und Darstellungsmittel sind geeignet, um thematische Gesichtspunkte in den unterschiedlichen Forschungsbereichen zum historischen Wortgebrauch und beim Aufbau historischer Textcorpora zu verankern?

Im folgenden Abschnitt erläutere ich zunächst zentrale Aspekte des Zusammenhangs von Themen, Themengeschichte und Wortgebrauch. In den darauf folgenden Abschnitten wird es dann um mögliche Folgerungen und Anwendungen in der historischen Lexikographie und Lexikologie sowie bei der thematischen Erschließung historischer Textcorpora gehen. In diese konzeptionell orientierten Darlegungen baue ich an geeigneten Stellen jeweils Beispiele aus unterschiedlichen thematischen Bereichen bzw. diskursiven Strängen der deutschen Sprach- und Kommunikationsgeschichte ein, im hinteren Teil folgen dann systematischer ausgebaute Beispiele.

2. Grundlagen: Themen, Themengeschichte und Wortgebrauch

In der alltäglichen Verständigung sind **Themen**, Teilthemen, thematische Zusammenhänge und auch Verfahren des Themenmanagements grundlegende Aspekte der Organisation sowohl von Texten als auch von Gesprächen.[2] Mit Fragen wie zum Beispiel »Wie war es im Urlaub« eröffnen wir bestimmte thematische Stränge im Gespräch, ModeratorInnen von Talkshows geben zu Beginn thematische Übersichten über Teilfragen und arbeiten die einzelnen thematischen ›Punkte‹ dann systematisch ab, zu den grundlegenden Fertigkeiten bei der Produktion von Texten gehört es unter anderem, die thematische Struktur durch geeignete Verfahren durchsichtig zu machen.

Auch das **Wortgebrauchsprofil** von Texten und Gesprächen ist von thematischen Erfordernissen geprägt: Ob man *über* Energiepolitik spricht oder ein Lehrbuch *über* Biologie schreibt, hat Folgen für den dafür erforderlichen, thematisch geprägten Wortschatz.[3]

Auf thematische Aspekte des Sprachgebrauchs beziehen sich vielfältige **Forschungstraditionen** mit je eigenen Ergebnissen und Methoden, die in unterschiedlichen disziplinären Kontexten verankert sind, so zum Beispiel die Lehre vom ›Stoff‹ und seiner Disposition in der Rhetorik, die Verfahren der qualitativen und quantitativen Inhaltsanalyse in der Soziologie sowie vielfältige Ansätze im

---

[2] Vgl. dazu und zu den folgenden Ausführungen FRITZ 2013, Kap. 4; FRITZ 1982, Kap. 7; FRITZ 1994, Kap. 2.5.

[3] Vgl. GLONING 2003, Kap. 2.2.2; vgl. auch GLONING 2004; 2011; 2012.

Rahmen der Textlinguistik, der gesprächsanalytischen Forschung, der Diskursanalyse und der thematisch kontextualisierten Begriffsgeschichte. Im Rahmen der Fachsprachenforschung spielt der Zusammenhang zwischen Sprachgebrauch, fachlichen Kommunikationsbereichen und damit auch fachlich geprägten Themenbereichen ebenfalls eine zentrale Rolle.

Themen sind von unterschiedlicher **kommunikativer Reichweite** in einer Sprachgemeinschaft. Die Skala reicht von hochgradig privaten Themen über gruppenspezifische Themen bis hin zu ›öffentlichen‹ Themen, deren Reichweite mit bestimmt wird von den Kommunikationskreisen und ggf. auch den Medien, in denen sie behandelt werden. Für das **Verstehen** von Themen, thematischen Verläufen und auch thematischen Zusammenhängen ist spezifisches **Wissen** bei den Kommunikationspartnern von grundlegender Bedeutung. Wenn es zum Beispiel zum gemeinsamen Wissen bzw. zu den geteilten Annahmen von Kommunikationspartnern gehört, dass bestimmte Fortbewegungsmittel die Umwelt in unterschiedlichem Ausmaß belasten, dann kann man *über* persönliche Einstellungen eines Menschen zur Umwelt sprechen, indem man eine Festellung über seine Fortbewegungsart macht (»Mein Chef hat auch so einen dicken Geländewagen«).

Mit dem Begriff des **thematischen Wortschatzes** kennzeichnen wir diejenigen lexikalischen Mittel, die in einer mehr oder weniger spezifischen Weise dazu beitragen, ein bestimmtes Thema kommunikativ zu bewältigen. Die grundlegenden Einheiten sind zum Teil Wörter mit einer einzigen Verwendungsweise (*Arbeitsamt*), vielfach aber auch nur bestimmte Verwendungsweisen von Wörtern (z.B. *stempeln* im Sinne von ›arbeitslos sein‹). Hinzu kommen Mehrwortverbindungen und ihre Verwendungsweisen, die ebenfalls einen besonderen thematischen Bezug aufweisen können (z.B. *scharfer Löffel* zur Bezeichnung eines seitlich angeschliffenen löffelförmigen Werkzeugs, das in der Medizin zur Entfernung von Gewebe verwendet wird). Den Themenbezug kann man ermitteln mit Fragen wie z.B.: »Welche lexikalischen Mittel dienen im Deutschen dazu, über das Thema Arbeitslosigkeit zu sprechen?« Oder: »Welche lexikalischen Mittel in einem Text sind in spezifischer Weise durch das Thema des Textes bedingt?« Auch wenn man auf diese Weise einen Ausdruck wie *Arbeitsamt* einem thematischen Wortschatz zugeschrieben hat, weisen doch nicht alle seine Verwendungen notwendig diesen Themenbezug auf, denn man kann z.B. über das Arbeitsamt einer Stadt auch als Gebäude mit seinen baulichen Eigenschaften z.B. in

einem architekturkritischen Text schreiben (»Die Fassade des Arbeitsamts hingegen ist von ausgesuchter Scheußlichkeit«).

Fachwortschätze kann man betrachten als thematische Wortschätze, deren Beherrschung und Nutzung in der Regel auf fachliche Gemeinschaften eingeschränkt ist. Auch Sachgebiete und Wissensgebiete sind eng verwandt mit dem Begriff des Themengebiets: Sachgebiete und Wissensgebiete *sind* auch zusammenhängende und organisierte Themengebiete, insofern sie nämlich Gegenstände kommunikativer Behandlung in Gesprächen und mehr noch in spezifischen Texten sein können. Auch wenn sich der Grad der Themenspezifik und die Art des Beitrags von Ausdrücken bzw. Verwendungsweisen zu Bewältigung thematischer Aufgaben in manchen Fällen nicht streng bestimmen lässt, so behält doch die Idee einer thematischen Prägung des Wortgebrauchs sein Anregungspotential als eine Klammer, die Wortschatzstrukturen und die kommunikative Rolle von Ausdrücken bei der Behandlung von Themen verbindet.

Thematisches Wissen organisiert auch unsere Sichtweisen größerer thematischer »Felder« wie z.B. Sport, Militär oder Kultur und den Stellenwert einzelner Elemente wie z.B. Arten von Ereignissen oder Arten von Personen. Wissensbestände dieser Art bilden komplexe Schema-Zusammenhänge ab, die nicht nur für das Verstehen von Kommunikationen erheblich ist, sondern auch für die Bewältigung alltäglicher Lebensvollzüge. Thematische Bereiche weisen also in der Regel eine interne, frame-artige Strukturierung auf, so dass man von thematischen Systemstellen sprechen kann, die sich auch für die Wortschatzstrukturierung nutzen lassen.

Ich will diesen Gesichtspunkt der systematischen thematischen Strukturierung des Wortgebrauchs mit einem kurzen, aber dichten **Beispiel**-Textausschnitt aus einem Werk zur **Rassenhygiene** veranschaulichen, aus *Hygienische Erziehung im Volksgesundheitsdienst* (1940) von Gottfried FREY, gekennzeichnet als *Fünfte erweiterte Auflage*. Die früheren Auflagen erschienen teilweise als Handbuchartikel. Der einleitende Abschnitt mit der Überschrift *Zweck der volkshygienischen Erziehung* lautet wie folgt, die für das Denksystem der Rassenhygiene wesentlichen Ausdrücke sind durch meine Unterstreichung markiert:

> »Fußend auf dem Urrecht der Selbsterhaltung und auf dem Willen, im Wettstreit der Nationen nach Stärke und Wesensausdruck eine geachtete Stellung einzunehmen, muß der Staat eine zielbewußte Bevölkerungspolitik treiben. Diese hat für alle Zukunft den Volksbestand zu sichern, durch geeignete Maßnahmen zur Hebung der Geburtenzahl und zur Verminderung der Sterblich-

> keit auf dessen Wachsen hinzuarbeiten, durch Ausmerzen des erblich Minderwertigen und Ausscheidung des Fremdrassigen das Gesamterbgut zu verbessern und die erblich Hochwertigen und Rassetüchtigen zu bevorrechten. Indem wir die Reinheit der Rasse und das stete Rauschen des gesunden Erbstromes als ethisch bedingte Grundlage jeder gemeinnützigen gesundheitlichen Bestrebung ansehen, setzen wir sie an die Spitze auch der hygienischen Erziehung im Volksgesundheitsdienst.«

In diesem kurzen Textstück werden mehrere **Grundbegriffe** (z.B. *Rasse, Volksbestand, Gesamterbgut*) und zentrale Ausgangspunkte der Rassenhygiene und der damit verbundenen **Sichtweisen** eingeführt. Sie dienen als Grundlage für die Formulierung von Forderungen und öffentlichen Aufgaben (*den Volksbestand sichern; das Gesamterbgut verbessern*). Dabei sind unter anderem Kontrastierungen (*erblich Minderwertiges, erblich Hochwertiges*), Hinweise auf Berechtigungsgrundlagen (*Urrecht*) und die Berufung auf Reinheits-Auffassungen (*Reinheit der Rasse*) erkennbar.

Der **Wortgebrauch** der Rassenhygiene-Ideologie weist eine interne, quasi-terminologische Strukturierung auf, die sich mit Hilfe von **thematischen Systemstellen** rekonstruieren lässt. In dem kurzen Beispieltext lassen sich bereits zwei wichtige Systemstellen ausmachen.

Eine erste Systemstelle umfasst Wörter bzw. Verwendungsweisen von Wörtern, die zentrale Anschauungen des Volks- und Rassegedankens zum Ausdruck bringen. Hierzu gehören: *Gesamterbgut, Rassetüchtige, Erbstrom; gesund, rein, Volksgesundheit, Fremdrassiges*.

Eine zweite Gruppe von Ausdrücken bezieht sich auf Maßnahmen im Bereich der Rassenhygiene, der Ausdruck *geeignete Maßnahmen* wird im Text auch als Oberbegriff für diese Systemstelle gebraucht. Hierzu gehören z.B.: *Volksgesundheitsdienst, Rassenpflege, ausmerzen, ausscheiden* oder *bevorrechten*. Die drei verbalen Elemente zeigen darüber hinaus die Unterscheidung zwischen negativen (*ausmerzen, ausscheiden*) und ›positiven‹ (*bevorrechten*) Arten von Maßnahmen, die in der Ideologie der Rassenhygiene eine lange Tradition hat.

Der kurze Text zeigt weiterhin, von welchen Ressourcen beim Aufbau dieses quasi-terminologischen Systems Gebrauch gemacht wird: Es sind dies zum einen die Mittel der Wortbildung, u.a. mit *erb-, volks(s)-, rasse-, art-*, zum anderen Formen der Metaphorik (z.B. der Gebrauch von *Strom* in *Erbstrom*), sodann auch die Übernahme von Schlagwörtern und ideologischen Kernwörtern aus verwandten Bereichen (*Wettstreit der Nationen*) und die Ausbildung bestimmter Verwendungsweisen, die sich im Grunde nur im Zusammenhang

einer Beschreibung des ideologischen Systems selbst rekonstruieren lassen (z.B. der spezifische Gebrauch von *Rasse*, *Gesamterbgut* oder *Volksgesundheit*). – Nun zurück zur Lehre von den Themen.

Themen und **thematische Wortschätze** haben auch eine **evolutionäre Dimension**, sie sind historisch veränderlich und weisen eine eigene Dynamik auf. Zu den zentralen Aspekten der Themengeschichte gehören unter anderem die Fragen, wann Themen aufkommen (z.B. Arbeitslosigkeit oder die Frage einer *Hebung der Rasse*), in welchen kommunikativen Kreisen, in welchen Medien und mit welcher Intensität sie behandelt werden, wie Themen von bestimmten Kommunikationskreisen in andere übernommen werden, mit welchen anderen Themen ein Thema verknüpft ist (das Thema Arbeitslosigkeit stand 1890 in anderen Zusammenhängen als heute), wann Themen wieder verschwinden bzw. an Intensität der Behandlung oder kommunikative Reichweite verlieren.

Im Hinblick auf die **historische Entwicklung thematischer Wortschätze** stellt sich nicht nur die Frage, mit welchen lexikalischen Mitteln bestimmte Themen zu unterschiedlichen Zeitpunkten bewältigt wurden, sodann auch die Frage, wie sich der Gebrauch spezifischer lexikalischer Mittel in unterschiedlichen Kommunikationskreisen und Texttypen ausgebildet und entwickelt hat. So wurden z.B. die Themen Computertechnik und AIDS zunächst innerfachlich behandelt, mit ihrer Entwicklung zu allgemeinen Themen gewannen auch entsprechende thematische Wortschätze weitere Verbreitung in der Sprachgemeinschaft.[4] Ein weiterer Aspekt der Entwicklungsgeschichte ist darüber hinaus auch die Frage, wie sich Alternativen und Konkurrenzverhältnisse in thematischen Wortschatzsektoren entwickelten und wie sich die Grade der Beherrschung in unterschiedlichen Personengruppen darstellen bzw. entwickelt haben.

Nun kann man fragen, wie sich die soeben vorgetragenen Aspekte einer thematischen Prägung des Wortgebrauchs bei der Entwicklung und Erschließung historischer Textcorpora, in der traditionellen **Wörterbuchschreibung** und bei der Organisation von neuartigen lexikographisch-lexikologischen Informationssystemen fruchtbar machen lassen.

---

[4] Vgl. hierzu u.a. WICHTER 1991, BUSCH 2004 (Computerwortschatz); TÖNNESEN 1995, EITZ 2005 (Aids).

3. Themengeschichte, thematische Corpora und Wortgebrauch in historischer Perspektive

Corpora dienen dazu, die unüberschaubare Vielzahl der Sprachereignisse in kontrollierter Weise zu reduzieren auf eine forschungspraktisch handhabbare Menge dessen, was in Texten, Tonbandaufnahmen, Videoaufnahmen usw. davon erhalten geblieben ist. Für die auswählende Reduktion spielt einerseits die zu dokumentierende Größe – z.B. die Gesamtsprache, eine bestimmte Varietät, ein Texttyp, ein Sprachstadium, ein Kommunikationsbereich, ein thematischer Sektor – eine wesentliche Rolle, zum anderen aber auch die Art der Fragestellung.

Will man ein historisches Textcorpus nutzen, um den Wortgebrauch im Hinblick auf die Kommunikationsgeschichte von Themen und Themenfeldern zu dokumentieren – sei es in lexikographischer Form, sei es in Form von lexikologischen Untersuchungen –, dann muss der **thematische Gesichtspunkt beim Aufbau eines Corpus** bzw. bei der Zusammenstellung spezifischer Teilcorpora planvoll berücksichtigt werden. Hierbei spielen unterschiedliche Teilfragen und Gesichtspunkte eine Rolle.

Zunächst geht es darum, **historische Themenstränge** zu identifizieren, zu **konturieren** und im Themenhaushalt der Zeit zu verorten. Hierfür kann man einerseits zeitgenössische[5] Indikatoren und Ressourcen nutzen, andererseits können moderne Sekundärquellen dazu beitragen, historische Themengebiete und ihre netzwerkartigen Zusammenhänge zu erschließen.

Zu den **zeitgenössischen Ressourcen** gehören zunächst **Themen-Bezeichnungen**, die in einem historischen Zeitraum als ›Adressen‹ für thematische Stränge, für Sach- und Fachgebiete unterschiedlicher Größe sowie für öffentliche Streitfragen aller Art gebraucht wurden. Als Beispiele wären etwa die zahlreichen Komposita auf *-frage* um 1900 zu erwähnen, mit denen öffentliche Themen benannt wurden (*Frauenfrage*, *Wahlrechtsfrage*, *Alkoholfrage*; vgl. auch Wendungen wie *sociale Frage*, *sexuelle Frage*). Diese Bezeichungen tauchen nicht selten im Titel von Beiträgen oder an prominenter Stelle in Texten mit einer entsprechenden thematischen Prägung auf. Die Bestimmung, in welchem Zeitraum eine Bezeichnung wie *Alkoholfrage* diese Adressierungsfunktion erfüllte und in welchem Sinn sie jeweils gebraucht

[5] Ich verwende *zeitgenössisch* mit Bezug auf eine historische Betrachtzeit, also zum Beispiel die Zeit um 1900 oder um 1600.

wurde, ist bereits ein wichtiger Teil nicht nur der thematischen Analyse, sondern auch der lexikalischen Charakterisierung dieser Ausdrücke. Insofern kann man sich auch die Frage stellen, wie man ein Inventar von ›thematischen Adress-Ausdrücken‹ anlegen könnte, ein Inventar derjenigen Ausdrücke also, die in einer jeweils zeitgenössischen Perspektive als Anlagerungspunkte für die damals jeweils aktuellen Themen gedient haben.

Eine zweite Art von zeitgenössischer Quelle sind **Lehrbücher**, **Handbücher** und thematisch brauchbare **Übersichtsdarstellungen** aller Art. Ihr besonderer Wert liegt darin, dass sie den Anspruch erheben, ein Themengebiet und ggf. auch die Teilgebiete in einer mehr oder weniger umfassenden Weise zu strukturieren, zu dokumentieren und auch sprachlich zu organisieren. Als Beispiel nenne ich hier das fast 700 Seiten starke Werk *Die deutschen Leibesübungen. Großes Handbuch für Turnen, Spiel und Sport* (NEUENDORFF 1927), in dem nicht nur die drei großen Traditionen der Turn-, der Sport- und der Spielbewegung, sondern auch die einzelnen Bewegungsformen in ihrer Ausprägung in den 1920er Jahren beschrieben werden. Der Wert solcher Darstellungen beruht für die lexikalische Analyse darin, dass dabei auch die sprachlichen Mittel des thematischen Bereichs der Bewegungskulturen und seiner Teilbereiche in einiger Breite gebraucht werden. So bieten Werke dieser Art nicht nur einen fachlich-thematischen, sondern auch einen sprachlich-lexikalischen Querschnitt für einen Zeitpunkt. Neben solchen großen und umfassenden Werken haben aber auch kleinere Texte ihren Wert, zum Beispiel zeitgenössische Zeitungsberichte, die aktuelle Entwicklungen eines bestimmten Bereichs in abgerundeter Form darstellen sollen. Ein Beispiel für einen solchen Kurztext stelle ich im Abschnitt 4 vor.

Eine dritte Aufschlussquelle können **moderne Sekundärdarstellungen** (z.B. Monographien, Handbuchartikel) sein, in denen Themen in ihrem historischen Kontext strukturiert werden und in denen auch die dafür zentralen Quellentexte genannt, bibliographisch verzeichnet und im Hinblick auf ihre Relevanz bewertet werden. Als Beispiel für diese Kategorie nenne ich das *Handbuch der deutschen Reformbewegungen 1880-1933* (KERBS & REULECKE 1998), das zu zahlreichen Themensträngen der Zeit um 1900 kurze Überblicksartikel bietet, in denen in der Regel Hauptthemen, Spielarten der Thematisierung, diskursive Zusammenhänge mit anderen Teilthemen, zentrale Schlagwörter, wichtige Quellentexte und ausgewählte weiterführende Sekundärliteratur genannt und kommentiert sind. In diesem Zusammenhang kann man weiterhin auch thema-

tische Informationen nutzen, die – für Bücher und teilweise auch für Aufsätze – im Rahmen der bibliothekarischen Verschlagwortung zur Verfügung gestellt werden. Sie geben in der Regel immerhin Auskunft zu Sachgebieten und ggf. zu Teilgebieten, die in einem Text thematisiert werden. Man kann weiterhin thematische bzw. diskursorientierte Bibliographien nutzen, exemplarisch sei hierfür die Bibliographie von SVEISTRUP & VON ZAHN-HARNACK (1934) zu den Teilthemen und diskursiven Verzweigungen der ersten Frauenbewegung genannt.

Neben diesen stärker systematischen Erschließungsformen gibt es darüber hinaus auch die mehr oder weniger ›zufälligen‹ **Lektürefunde**, die Hinweise, die sich beim Verfolgen von Fußnoten oder bei der Recherche einzelner Wortgeschichten ergeben. Auch diese Hinweise haben ihr Recht, sofern ihr Stellenwert für die Geschichte eines Teilthemas kritisch beurteilt wird.

Einen **historischen Themenstrang** vorläufig **konturieren** heißt also zunächst einmal, einen zentralen Gegenstand bzw. eine zentrale Frage zu bestimmen und im Anschluss daran die bereits bekannten Teilthemen und Systemstellen, bereits bekannte Denkfiguren und Topoi, die zeitliche Erstreckung und Aspekte der zeitlichen Dynamik, die zentralen TeilnehmerInnen, die kommunikative Reichweite (in unterschiedlichen Texttypen, in verschiedenen Formen der Mündlichkeit, in Medien), erste Beobachtungen zu Thematisierungspraktiken und auch zu prominenten sprachlichen Mitteln darzulegen.

Eine weitere zentrale Aufgabe, die zeitlich in vielen Fällen parallel zu den Arbeitsschritten der Konturierung von thematischen Feldern und Strängen läuft, ist die Ermittlung von **thematisch einschlägigen Texten** bzw. Sprachdaten, die Anreicherung der bibliographischen Angaben mit **themenbezogenen Metadaten** und ggf. die **Auswahl** geeigneter Texte für ein eingeschränktes Teilcorpus. Die traditionellen variationslinguistischen Angaben zu Datierung, Genre/Textsortenbereich, ggf. Lokalisierung etc. von Corpustexten müssen also ergänzt werden mit Angaben dazu, welchen Themenfeldern, Teilthemen und/oder Diskursthemen bestimmte Texte zuzuordnen sind. Man muss davon ausgehen, dass sich die Sichtweise historischer Themenstränge bei laufender Arbeit und mit der Bearbeitung neuer Texte verändert, differenziert und modifiziert. Insofern muss ein Dokumentationssystem für die Verwaltung historischer Themen, ihrer Systemstellen, ihrer Entwicklungen sowie für die darauf

bezogenen Corpustexte offen, dynamisch und revidierbar angelegt sein.[6]

Das **thematische Kriterium** spielt auch eine wichtige Rolle für die Chronologie und die Dynamik der **Wortschatzentwicklung** in unterschiedlichen **Texttypen und Kommunikationsbereichen**. Eine Leitfrage, die sich auf diesen Zusammenhang bezieht, kann man so formulieren: Wie wirkt sich die Kommunikationsgeschichte eines Themas aus auf die Verwendungscharakteristik der lexikalischen Mittel, die dabei gebraucht werden.

Ich gebe ein **Beispiel**: Die Idee des **Naturschutz**es entstand in Deutschland gegen Ende des 19. Jahrhunderts, mit ihr begann auch eine entsprechende Thematisierungsgeschichte, deren Ursprung zunächst im Bereich der Forstwirtschaftskunde und ihrer Zeitschriftentexte lag. Im ersten Viertel des 20. Jahrhunderts wurden Aspekte des Themas auch außerhalb des Ursprungsbereichs und in weiteren Kommunikationsmedien (z.B. der Alpenvereinszeitschrift) behandelt. Auch in den Tageszeitungen wird gelegentlich über Naturschutzthemen berichtet, etwa in Form von Veranstaltungsberichten. Zu Beginn des 20. Jahrhunderts wird Naturschutz auch staatlich institutionalisiert, zu den späteren Entwicklungen gehört die Naturfreundebewegung und die eng mit der Naturschutzthematik verwandte Umweltschutzthematik, die in der zweiten Hälfte des 20. Jahrhunderts mehrere Teilthemen (z.B. das Waldsterben) mit erheblicher, aber auch schwankender medialer Breitenwirkung hervorbrachte. Das Auf und Ab der Themenkarrieren wird zum Teil in den Medien selbst reflektiert und kommentiert:

> »Zuviel verlangt? Es ist nicht lange her, da waren Umwelt- und Naturschutz des Deutschen größte Sorge. Die sechs Nationalparke in den neuen Bundesländern galten sogar als das „Tafelsilber der deutschen Einheit“ (Klaus Töpfer). Heute taucht das Thema bei aktuellen Meinungsumfragen nicht mal

[6] Diese Forderung ist nicht trivial. Nehmen wir an, eine ForscherInnengruppe hat 40 lange Texte mit einem thematischen Profil X lexikographisch erschlossen. Nehmen wir weiterhin an, dass beim Text Nr. 41 deutlich wird, dass die thematische Binnenstrukturierung verändert werden muss. Die Frage, wie man solche Probleme im Schnittpunkt zwischen lexikologisch-lexikographischer Arbeit, Corpus-Aufbereitung und thematischer Markierung ganz praktisch löst, gehört mit zu den Zukunftsaufgaben bei der Weiterentwicklung der technischen und der konzeptuellen Grundlagen von digitalen Systemen. – Eine weitere Frage ist auch, wie man bereits vorhandene Corpora thematisch erschließen kann, vgl. hierzu u.a. WEISS 2005 und laufende Arbeiten an den IDS-Corpora. – Vgl. zur Rolle thematischer Corpora für diskurslinguistische Untersuchungen weiterhin FELDER *et al.* 2012.

mehr auf den hinteren Plätzen auf.« (Die Zeit, 19.12.1997; http://www.dwds.de; 06.01.2013)

Der zentrale Ausdruck aus der Frühgeschichte des Themas ist *Naturschutz*. Im *DWb* ist dem Wort kein Eintrag gewidmet, was nicht verwundert, denn der entsprechende Band mit der N-Strecke erschien 1883, also noch vor der Themenkarriere des Naturschutzes. Dass aber auch die Volltextsuche im gesamten Bestand des digitalen *DWb* keinen Treffer ergibt, verwundert dann doch. Das PAULsche *Wörterbuch* und WEIGAND & HIRT haben keinen Eintrag dazu. Im *DWDS* findet sich eine kurze Bedeutungsparaphrase (allerdings ohne Hinweis auf die Tradition des Naturschutzgedankens), im DWDS-Corpus zum 20. Jahrhundert finden sich Belege, die bis ins Jahr 1914 reichen. Wesentliche Teile der Gebrauchsgeschichte lassen sich damit gut verfolgen mit Ausnahme der frühen Entstehungs- und Verbreitungsgeschichte.

Ich ziehe ein erstes **Zwischenfazit**:

(a) Viele Wortgebräuche sind *auch* geprägt durch ihre Rolle in der Geschichte von Themen, thematischen Feldern und Thematisierungspraktiken.

(b) Dem steht gegenüber, dass diese Prägung in der lexikographischen Tradition bislang in der Regel nicht systematisch berücksichtigt wird. Dies betrifft mindestens drei Dimensionen der lexikographischen Arbeit:

(i) Die alphabetische *Zugriffsstruktur* erlaubt keine themenorientierte Recherche;

(ii) Bei der Planung und Zusammenstellung von lexikographischen *Corpora* spielten Gesichtspunkte der Themengeschichte bislang keine systematisch nachvollziehbare Rolle;

(iii) Bei der Bedeutungsbeschreibung bzw. bei der *Charakterisierung von Gebrauchsprofilen* und ihrer Entwicklung wird die thematische Prägung von Ausdrücken vielfach nicht systematisch mit dargestellt.

(c) Umgekehrt werden in den Spezialuntersuchungen zu Diskursen und zu einzelnen Themen in der Regel die verwendeten Corpustexte nicht öffentlich dokumentiert und die einschlägigen Aus-

drücke auch nicht einzelwortbezogen beschrieben und dokumentiert.[7]

Es ist eine naheliegende Überlegung, die jeweiligen **Quellen**, die z.T. an entlegener Stelle erschienen sind, auch digital zu **dokumentieren** und die Belegwortregister zu **thematischen Glossaren** bzw. thematischen Wörterbüchern auszubauen. Im Hinblick auf die Quellendokumentation sind dabei leider dornige rechtliche Probleme zu erwarten, sofern es sich um Quellen handelt, die noch dem Urheberrecht oder sonstigen Leistungsschutzrechten unterliegen. Unabhängig davon können aber die themen- und diskursgeschichtlichen Untersuchungen auch die anstehenden Überlegungen befruchten, wie man themen- und diskursgeschichtliche Aspekte auch stärker in der historischen Lexikographie des Deutschen verankern könnte.

4. Lexikalische Profile thematisch einschlägiger Quellentexte: Schlüsseltexte aus der Frühgeschichte des Sports

Ein wertvolles heuristisches Verfahren zur Ermittlung und Charakterisierung historischer thematischer Wortschätze, ihrer Struktur und ihrer Entwicklungsdynamik nutzt die Idee, die Darstellungsform des einzeltextbezogenen Glossars für die Erschließung historischer thematischer Wortschätze zu nutzen. Die Heuristik besteht darin, zunächst **thematisch einschlägige Quellentexte** zu bestimmen, sodann die einzelnen Texte im Hinblick auf ihren Anteil an thematisch spezifischem Wortschatz zu analysieren und die einzelnen Wortschatzelemente dann je nach den eigenen Zielsetzungen lexikographisch zu bearbeiten und im Hinblick auf thematische Systemstellen zu markieren.

Nehmen wir an, unser Fernziel sei dabei ein **historisches Wörterbuch** bzw. ein Digitales Lexikalisches System[8] zu Spielarten der Bewegungskultur auf der Grundlage von deutschsprachigen Quellen.

---

[7] Um ein Beispiele zu nennen: In den ganz hervorragenden Beiträgen zum Sammelband *Kontroverse Begriffe* (STÖTZEL & WENGELER 1995) und in den darauf bezogenen Dissertationen zu einzelnen Themen (z.B. JUNG 1994: *Öffentlichkeit und Sprachwandel. Zur Geschichte des Diskurses über die Atomenergie*) werden die in beeindruckender Breite ermittelten Quellengrundlagen jeweils nur bibliographisch dokumentiert, die einzelnen Ausdrücke und ihre Verwendungsweisen werden aber nur im Rahmen monographischer Darstellungen und mit jeweils unterschiedlicher Fokussierung (z.B. Erzeugung von Sichtweisen, Rolle von Wortbildungen oder Entlehnungen, Formen der Metaphorik) kommentiert und dann auch in Form von Belegwortregistern erfasst.

[8] Vgl. hierzu KLEIN 2004; KLEIN & GEYKEN 2010.

Teilgegenstände seien jeweils die themenspezifischen Wörter bzw. Verwendungsweisen (z.B. zu einzelnen Sportarten), ihr Gebrauchsprofil im Rahmen von thematischen Strängen sowie Markierungen, die einen gezielten Zugriff auf thematische Systemstellen erlauben. Ich möchte das Verfahren nun anhand eines **Beispiel**s aus der Frühgeschichte des **Fußballspiel**s veranschaulichen, eine voll ausgebaute Darstellung wird an anderer Stelle erfolgen.

In einem ersten Schritt stellen wir fest, dass es im letzten Viertel des 19. Jahrhunderts, also in der **Etablierungsphase des Fußballs** im deutschen Sprachgebiet unterschiedliche **Texttypen** gab, die auf die neue Praxis bezogen waren und die sich vorläufig in vier Gruppen einteilen lassen. (i) Zur ersten Gruppe gehören Formulierungen von Vereinsstatuten und Spielregeln, zum Beispiel die 1875 von KONRAD KOCH veröffentlichten *Regeln des Fußball-Vereins der mittleren Classen des Martino-Catharineums zu Braunschweig*. (ii) Eine zweite Gruppe bilden frühe Formen der Spielberichterstattung (vgl. NAIL 1983). (iii) Ein dritter Typ sind handbuchartige Überblicksdarstellungen (HEINEKEN 1898). (iv) Zahlreiche Texte umfasst schließlich ein breiter Diskurs über den erzieherischen Wert und über Fragen der schulisch-pädagogischen Umsetzung des Fußballspiels in Zeitschriften wie dem *Jahrbuch für Volks- und Jugendspiele* oder dem *Pädagogischen Archiv*, an dem neben Konrad Koch auch andere Schulmänner beteiligt waren. Es war erklärtes Ziel dieser pädagogisch orientierten Beiträge, Fußball als ein Bewegungsspiel vor allem für die Schuljugend einzuführen.

Betrachten wir nun als **Beispiel** einen frühen **Beitrag von KONRAD KOCH** in der Zeitschrift *Pädagogisches Archiv* (**1877**) mit dem Titel *Fußball, das englische Winterspiel*. In einem ersten Schritt soll es exemplarisch darum gehen zu bestimmen, welche **Teilthemen** der Text aufweist und welche **Wortschatzsektoren** bzw. lexikalischen Mittel jeweils darauf bezogen sind.

Ein erstes wesentliches Teilthema des Beitrags ist die Beschreibung unterschiedlicher Charakteristika des Fußballspiels, auch in Abgrenzung von anderen Bewegungsformen und den dort vorgesehenen Bestimmungen. Der Ausdruck *Fußball* wird sowohl für den Gegenstand als auch für das Spiel gebraucht. Die Spielbezeichnung wird im 19. Jahrhundert allerdings noch weiter verwendet als heute, die historischen Spielvarianten, die wir heute als *Rugby* bezeichnen, fielen damals ebenfalls noch unter die Bezeichnung.

Im Rahmen der Charakterisierung des Spiels werden unter anderem relevante **Gegenstände** und **Aspekte der Spielpraxis** eingeführt

und benannt. Dies sind vor allem Personen, Spielgegenstände und die räumlichen Gegebenheiten, aber auch die zentrale Spielidee und Handlungsformen bzw. Ereignisse. Auf diese thematischen Systemstellen sind jeweils bestimmte Ausdrücke und **Wortgebrauchssektoren** bezogen:

Zu den Bezeichnungen für **Personen** gehören zunächst allgemeine Bezeichnungen wie *Spieler* (163) oder *Fußballspieler* (162, 163) sowie *Gespielschaft* im Sinne von ›Mannschaft als eine strukturierte Gemeinschaft von Personen; Partei in einem Spiel‹ (162). Sodann gibt es bereits früh eine Gruppe von Bezeichnungen für die unterschiedlichen Arten von Spielern in einer Mannschaft, die sich später weiter ausdifferenzieren. Im Text von 1877 finden wir unter anderem folgende Bezeichnungen mit ihren englischen Gegenstücken: *Stürmer* (*forward*; *player up*; 168), *Markmann* (*half-back*; 168), *Malmann* (*back*; 168). Über die Malmänner heißt es, dass sie »als eine Art Nachhut das Mal gegen solche Ueberfälle [Angriffe auf das eigene Mal bzw. Tor; TG] zu schützen haben« (168). Diese Bezeichnungen beziehen sich also auf funktional differenzierte Spielpositionen, die durch einen spezifischen räumlichen Aufgabenbereich und eine je eigene Aufgabe charakterisiert sind. Das ist von einem modernen Standpunkt aus zwar selbstverständlich, es ist aber nicht trivial, denn es gab historische Formen und gibt auch heute noch Spielweisen, bei denen alle Spieler gleichzeitig dem Ball nachjagen, z.B. unter Kindern. Die Idee der funktionalen Differenzierung und der systematischen Anlage des Zusammenspiels in einer Mannschaft ist also ein Element, das in einer historischen Perspektive nicht selbstverständlich ist. Die Idee steht auch im thematischen System an einem ganz anderen Ort: Sie wird u.a. im Rahmen der Diskussion gemeinschaftsförderlicher pädagogischer Werte eigens herausgestrichen, sie bezieht sich also nicht oder nicht nur auf die Idee der Leistungssteigerung und der Erfolgsorientierung durch funktionale Differenzierung, wie es für eine moderne Sichtweise charakteristisch ist. Mit der Verbindung *Kaiser der Gespielschaft* wird der Mannschaftsführer, der Mannschaftskapitän bezeichnet und dem englischen Ausdruck *captain* in Klammern zugeordnet. (Später wird dafür auch die Wortbildung *Gespielschaftskaiser* üblich.)

Aufgrund der relativen Einfachheit des Fußballspiels finden wir nur eine überschaubare Anzahl von Bezeichnungen für wesentliche **Spielgegenstände**. Hierzu gehören unter anderem *Ball* (161, 162); *Fußball* (als Bezeichnung für den Spielgegenstand, nicht für das Spiel als solches, 162) sowie *Stange* (166) und *Mal* ›Tor‹ (166). Zwar

werden in einigen Passagen auch Fragen der in England üblichen Kleidung behandelt, dies gehört aber nicht zu den wesentlichen Gesichtspunkten des Spiels.

Ein anderer wichtiger Aspekt der Charakterisierung des Spiels ist die Bestimmung **räumlicher Verhältnisse**. Hierfür dienen unter anderem eine Reihe von Bezeichnungen für räumliche Ausstattungsaspekte: *Spielplatz* (166), *Spielraum* (166), *Ecke* (166), *Maß* (166), *Länge* (166), *Breite* (166), *Schritt* als Maßeinheit (166), *Ausdehnung* (166), *Längenseite* (166), *Marklinie* (166), *Breitseite* (166), *Mallinie* (166). Die folgende Textpassage zeigt, dass diese thematisch geprägten Ausdrücke durchaus unterschiedlichen Status haben. Ein Ausdruck wie *Längenseite* wird in einer Einführungssituation verwendet, um den spezifischen Ausdruck *Marklinie* zu erklären, dabei werden in Klammern auch die englischen Gegenstücke mit angeführt.

> Die Längenseiten heißen Marklinien (*touch lines*), die Breitseiten Mallinien (*goal lines*). In der Mitte der Letzteren stehen die Male (*goals*). (166; die Kursive ersetzt den originalen Antiquadruck in einer Frakturumgebung)

Räumliche Aspekte spielen dann auch eine wichtige Rolle bei der Charakterisierung von Handlungsweisen bzw. **Spielereignissen**. So sind räumliche Aspekte impliziert bei: *fernhalten* (162), *gegenüber stehen* (162), *heranbringen* (162), *über X weg* (162) und anderen.

Bei der Kennzeichnung des übergeordneten **Spielziel**s dienen sowohl *Aufgabe* (162) als auch *Hauptaufgabe* (162) als Signalausdrücke, Ausdrücke wie *Spielregeln* (162), *Gesetzsammlung* (›gesammelte Kodifizierung von Spielregeln‹; 170) oder *verboten* (165) beziehen sich auf die normativen Grundlagen des Spiels. Bei der Kodifizierung bzw. Charakterisierung selbst dienen Ausdrücke wie *Laufen* (163), *Mal* (162), *stark* (zur Bezeichnung der Anzahl; 162) oder *Zeit* ›Spielsaison‹ (163) zur Darstellung der unterschiedlichen Aspekte des Spiels.

Ich breche die Hinweise zur Behandlung des ersten Teilthemas, die Charakterisierung von Aspekten des frühen Fußballspiels im Spiegel seiner Wortschatzsektoren hier ab und wende mich einem zweiten wichtigen Teilthema zu, der Diskussion des **pädagogischen Werts des Fußballspiels** im letzten Viertel des 19. Jahrhunderts. In diesem Zusammenhang werden eine ganz Reihe von **Sinndimensionen** angesprochen, die teils körperlicher, teils charakterlicher, teils sittlicher Natur sind und die mit anderen Diskursen der Zeit eng verbunden sind. Auch bestimmte Denkfiguren, z.B. der Gegensatz

von Stadt/Natur bzw. Stube/frischer Luft spielen hier eine wesentliche Rolle. Einige Beispiel-Ausdrücke aus diesem Sektor sind: *Abspannung für den Geist* (163); *Anstrengung* (als positiver Wert; 163); *Bildung des Charakters* (163); *Erholung* (163); *erschöpft* (als positiver Zustand; 163); *frische Luft* (164); *Gesundheit* (164); *Körperkraft* (162); *Lungen* (164); *faules Nichtstun* (163; als Gegenbegriff); *rüstig* (163); *Stubenhockertum* (163; als Gegenbegriff); *Uebung* (162); *Vergnügen* (162), *Wert* (163). Einzelne Ausdrücke und Wendungen werden dabei auch kontrastiv verwendet, z.B. in Dichotomien wie *Körper* vs. *Geist* (163) oder *dumpfe Stubenluft* vs. *im Freien* (164). Auch die Gegenüberstellung des Einzelnen im Gegensatz zum pädagogisch wertvollen Gemeinschaftsaspekt des Fußballspiels schlägt sich lexikalisch nieder: *Einzelner* (161) vs. *Zusammenspiel* (161); *geschlossen* (161 ›in einer geordneten Ganzheit organisiert‹).

Es ist klar, dass man Wörtern wie *Zusammenspiel* oder *Stubenluft* ihre besondere Rolle in diesem **diskursiven Zusammenhang** nicht ansieht. Wenn man die Rolle, die einzelne Ausdrücke in spezifischen thematischen Zusammenhängen und teilweise auch nur zeitlich begrenzt gespielt haben, in sprachgeschichtlichen Darstellungen mit erfassen will, dann muss man auch über neue Formate nachdenken, wie diese Gebrauchsaspekte dokumentiert werden können, sei es in Form einer Ausweitung von lexikographischen Darstellungsformen, sei es als Kombination von traditionellen lexikographischen Formaten mit stärker diskurs- und themenorientierten monographischen Darstellungsformen.

Ein drittes Teilthema bezieht sich auf den aus deutscher Sicht **fremden Ursprung des Spiels**, der unter den Bedingungen des 19. Jahrhunderts explizit thematisiert bzw. entschärft werden musste. Konrad Koch spielt diesen Punkt geschickt herunter, indem er die Entlehnung auf einen harmlosen Ball reduziert und anfügt, dass die deutsche Jugend davon schon eigenen Gebrauch machen werde. Zu den Wortschatzelementen, die in diesem Zusammenhang gebraucht werden, gehören zunächst die Kontrastierung *englisch/deutsch* (162), sodann Bezeichnungen für nationale Eigenheiten wie z.B. *Art* (162), *Eigentümlichkeit* (162) oder *Sitten* (162), ihre explizite Kennzeichnung (*fremdartig* 162) sowie Ausdrücke wie *Selbstbewußtsein* (162), *Selbstständigkeit* (162) oder *verpflanzen* (162) für Aspekte einer Dynamik kultureller Kontakte.

Man sieht an den bisher genannten Beispielen aus den drei thematischen Teilbereichen, dass **thematisch relevante Ausdrücke** bzw. ihre Verwendungsweisen einen ganz **unterschiedlichen Status**

haben. Manche sind nicht sportspezifisch (z.B. *stark* zur Angabe einer Anzahl kann auch in anderen Zusammenhängen, etwa beim Militär genutzt werden), sie leisten aber natürlich dennoch einen wichtigen Beitrag zu ›Bewältigung‹ thematischer Aspekte, der jeweils mit erfasst und mit beschrieben werden muss. Andere Verwendungsweisen (nicht Wörter) dagegen sind spezifisch wie z.B. der Gebrauch von *stoßen* für das, was wir heute als das Schießen eines Balles bezeichnen.

Bemerkenswert ist darüber hinaus auch die **Rolle der Wortbildungen** für die Bewältigung thematischer Aufgaben. So finden wir im Text von Koch z.B. eine ganze Reihe unterschiedlich komplexer Wortbildungen zum Stamm *Spiel*. Hierzu gehören: *Ballspiel* (162, 163); *Fußball-Wettspiel* (161); *Fußballspieler* (162); *Jugendspiel* (161); *Laufspiel* (163); *Spielausdruck* (162); *Spieler* (163); *Spielplatz* (162, 163); *Spielweise* (162); *Winterspiel* (161).

Ein weiterer Gesichtspunkt, der quer zur sportsystematischen Perspektive liegt, ist die Beobachtung, dass zahlreiche **Ausdrücke aus dem Bereich Kampf/Krieg** für die Thematisierung sportlicher Ereignisse genutzt werden. Die Nutzung dieser Art von Vergleich bzw. Metaphorik wird im Text auch explizit diskutiert und mit Beispielen aus der englischen Literatur veranschaulicht. Im deutschen Text werden u.a. folgende sprachlichen Mittel aus diesem Sektor für die Charakterisierung sportlicher Ereignisse beim Fußball gebraucht: *Abwehr* (als Ereignis 161); *Fußballkampf* (161); *Gegner* (161); *Kampf* (161); *Nahgefecht* (161); *Niederlage* (161); *Sieg* (161); *Sturm* (161); *Übermacht* (161); *Feind* (172), *feindlich* (162; 166 u.ö.). Die Konzeptualisierung von Ereignissen und Gegenständen in der Begrifflichkeit von Kampf und Krieg ist in zweierlei Hinsicht besonders aufschlussreich: (i) Sie bietet Materialien für Fragestellungen einer historischen Metaphorik und damit auch zur Frage, wie sich kognitive Aspekte der Metaphorik zu Gesichtspunkten der historischen Tradition bzw. der kulturellen Prägung verhalten. (ii) Sie bietet darüber hinaus auch weitere Materialien zur Ideengeschichte des Militarismus und seiner sprachlichen Realisierung im 19. und frühen 20. Jahrhundert.[9]

Man kann also zum einen festhalten, dass **Schlüsseltexte** wie dieser aufschlussreiche Quellen sind für bestimmte Themen und für zentrale Thematisierungspraktiken zu einem historischen Zeitpunkt.

---

[9] Auch die Sprach- und Kommunikationsgeschichte des frühen Alpinismus zeigt vergleichbare militaristische Elemente, etwa beim Gebrauch von Formen der Kampf- und Eroberungsmetaphorik.

Aber der Text gibt auch Anlass zu **methodischer Vorsicht** und zu Einschränkungen: Denn eine zweite Lehre, die sich aus dem Studium dieses Textes ergibt, besagt, dass manche Passagen sich im damaligen Verständnis auf eine ganz andere *Spielweise* (169) des **Fußball**spiels beziehen, die wir heute als ein eigenes Spiel (**Rugby**) betrachten und die auch an späterer Stelle im Beitrag von Koch nicht nur als andere *Spielweise*, sondern eben als eigenes *Spiel* bezeichnet wird. Auch im Verständnis der Zeitgenossen haben sich die Spielweisen also schnell differenziert. So ist dann auch das Buch von HEINEKEN (*Das Fußballspiel. Association und Rugby*; 1898), in dem immerhin noch beide Spielweisen behandelt werden, in zwei klar getrennte Teile gegliedert, die auch die sportgeschichtliche Entwicklung widerspiegeln, gleichwohl erscheint das Werk noch unter dem übergeordneten Titel *Das Fußballspiel*. Konrad Koch geht in seinem Beitrag von 1877 offenkundig von der engeren, der Association-Sichtweise des Fußballspiels aus und behandelt die Rugby-Spielweise als eine konservativere Variante, die er präzise auf bestimmte Elemente der Spielregeln zurückführt (»Die entscheidende Regel, an welcher die Vermittlung zwischen den beiden Spielweisen gescheitert ist, ...«; 165.25ff.).

Für uns Wortschatz-HistorikerInnen hat das zur Konsequenz, dass wir diejenigen Wortschatz-Elemente, die zur **Thematisierungsgeschichte** des Rugby-Spiels gehören, aus der **Wortschatzgeschichte** des Fußballs in einem engeren, modernen Sinne ausschließen müssen, falls wir die entsprechende thematische Eingrenzung vornehmen wollen. So wird zum Beispiel auf Seite 169.10ff. des Beitrags ein Typ von sportlicher Aktivität, ein Situationstyp beschrieben, der als *Mengen* bzw. als *Gemenge* bezeichnet wird und der im heutigen Rugby mit dem Ausdruck *Gedränge* bezeichnet wird. Das bei Koch erwähnte englische Gegenstück ist *scrummage*. Sowohl *Mengen* als auch *Gemenge* gehören also nicht eigentlich in die Thematisierungsgeschichte des Fußballspiels im engeren, modernen Sinne, obwohl sie im Beitrag von Koch mit der Überschrift *Fußball* vorkommen.

Das Beispiel lehrt, dass man **Quellentexte** nicht unbesehen aufgrund von Titelstichwörtern oder von im Text verwendeten Schlüsselwörtern bestimmten **Themenbereiche**n zuordnen kann. So hat sich gerade gezeigt, dass der thematisch verwendete Signalausdruck *Fußball* als Spielbezeichnung im Titel des Beitrags das im zeitgenössischen Verständnis als Variante betrachtete Rugby-Spiel bzw. seine historischen Vorstufen noch mit umfasste. Es hat sich weiterhin gezeigt, dass man bestimmte Textpassagen und lexikalische Mittel ausschließen muss, wenn man einen engeren, modernen

Begriff von Fußball zugrundelegt, auch wenn Schlüsselwörter wie *Stürmer* bereits fußballspezifisch in einem engeren Sinn im Gesamttext verwendet sind. Umgekehrt sind die hier aus der Perspektive des modernen Fußballspiels auszuschließenden Passagen zur damaligen Rugby-Variante natürlich sehr wertvoll für Untersuchungen zur Frühgeschichte des Rugby-Spiels und seiner Thematisierung im deutschen Sprachraum.

Neben solchen mittelgroßen und thematisch schon recht breit angelegten Werken haben aber auch **kleinere Texte** ihren Wert, zum Beispiel zeitgenössische Zeitungsberichte, die aktuelle Entwicklungen eines bestimmten Bereichs in abgerundeter Form darstellen sollen. Auch hierfür ein **Beispiel**, das ich kurz auswerten möchte: Die Wochenzeitschrift *Die Woche* veröffentlichte Anfang April 1902 einen dreiseitigen Beitrag mit dem Titel *Mädchensport*, der mit drei »Momentaufnahmen«, Fotografien beim Rudertraining, beim Golfspiel und beim Volleyball (*Handballspiel*) bebildert war. Dieser Beitrag bietet in seinem Textanteil, der neben den Bildern nur wenig mehr als eine Zeitschriftendruckseite umfasst, mehrere **thematisch relevante Wortschatzsektoren** in recht breiter Belegung. Hierzu gehören in verkürzter Aufzählung (der Text ist im Anhang dokumentiert):

- Überbegriffe für Arten von Bewegungen: *Sport*, *Sportspiel*, *Ballspiel*, *Sportart*, *Spiel*, *Mädchensport*;
- Bezeichnungen für einzelne Bewegungsformen: Lawntennis, Tennis, Golf / Golfspiel, Hockey, Wiesenballspiel; Handball, Handballspiel ‚Volleyball', Korbballspiel, Basketballspiel, Fußball, Bowling, Badminton, Rudern, Radeln;
- Bezeichnungen für einzelne Arten von Ereignissen: Rudertraining, Meisterschaftskampf, Turnier, Match; Schlag; gewinnen, hinüberschlagen, herüberschlagen; Sieg, Niederlage; Zählen der Points; servieren ‚Aufschlag machen; das Aufschlagrecht innehaben';
- Bezeichnungen, die sich auf Sinndimensionen der Bewegungskultur beziehen: *Gesundheit*, *Jugendfrische*, *Schönheit*, *Gewandtheit*; *stählen*; *widerstandsfähig*; *im Freien*, *Natur*;
- mehr oder weniger spezifische Bezeichnungen für Gegenstände: Rakett, Kolben, club ‚Schläger, Schlagholz', Schlagholz; Ball, Guttaperchaball, Korkball, Kugel, Holzkugel, Jack; Lawn, Platz, Rasenfläche, Bahn, Court; Goal, Mal, Netz, Loch; Stahlroß, Boot;
- Bezeichnungen, die sich auf unterschiedliche Aspekte der sozialen Organisation beziehen: *Partei*, *Team*, *Gegenpartei*; *Spieler*, *Spielerin*,

Teilnehmer, *Sportsdame*; *Nationalität*; *Spielregeln*, *Grundregeln*; *Sieger*;
- auch die thematische Systemstelle der zeitlichen bzw. der saisonalen Organisation der Spielpraxis ist angelegt mit dem Ausdruck *Winterpause*.

Die Auswertung dieses Textes und die Gruppierung der Beispiele soll verdeutlichen, dass auch **schon kurze Texte wertvolle Aufschlüsse** geben können über den **Bestand** thematischer lexikalischer Einheiten eines Bereichs und auch über die Strukturierung **thematischer Systemstellen** zu einem bestimmten Zeitpunkt. Wir erfahren auf diese Weise, welche Themen zu einem bestimmten historischen Zeitpunkt behandelt wurden und welche sprachlichen Mittel in welchen Verwendungsweisen dafür gebraucht wurden. Anschlussfragen können sich unter anderem darauf beziehen, welche Ressourcen genutzt werden, um die entsprechenden sprachlichen Werkzeuge bereitzustellen. Das Wort *Team* ist einen Entlehnung, die wir heute noch gebrauchen, nicht nur im Sport. Die Wortbildung *Wiesenballspiel* ist offenbar als Eindeutschung für den englischen Ausdruck bzw. die Entlehnung *Lawntennis* gebraucht worden, beide Ausdrücke sind über einen noch zu bestimmenden Zeitraum hinweg Konkurrenzausdrücke oder – wenn man eher den kommunikativen Nutzen dieser Konstellation hervorhebt – Mittel der Variation im Ausdruck.

5. Wie kann man themenorientierte Aspekte in lexikographisch-lexikologischen Darstellungsformaten verankern?

Im Hinblick darauf, ob und ggf. wie die besprochenen thematischen Aspekte des Wortgebrauchs in der lexikographischen Praxis aufgegriffen werden können, stellen sich eine Reihe von wichtigen **Anschlussfragen**. Hierzu gehören u.a. die Fragen,
- ob und ggf. wie auch thematisch geprägte **ad-hoc-Mittel** dokumentiert werden sollen oder können,
- sodann die Frage, ob und ggf. wie Aspekte des Themenbezugs auch in **Wortartikeln** beschrieben bzw. markiert werden können,
- und schließlich auch die Frage, wie sich ggf. Formen der **Verbindung** mit den Ergebnissen aus **monographischen Darstellungsformen** nutzen lassen, um Aspekte der thematischen Prägung in lexikographisch-lexikalischen Dokumentationssystemen abzudecken.

5.1 Thematische Prägung und Grade der Etablierung von Wortgebräuchen

Zunächst stellt sich die Frage der Konventionalisierung und der z.T. sehr eingeschränkten Stabilität von thematisch geprägten lexikalischen Mitteln: Was soll, was kann man tun mit **ad-hoc-Mitteln**, mit lexikalischen Eintagsfliegen oder auch mit sprachlichen Ausdrücken bzw. Verwendungsweisen von Ausdrücken, deren Gebrauchsgeschichte jeweils in einer zeitlich, räumlich oder auf einen Kommunikationsbereich begrenzten Themenkarriere verankert ist?

Die traditionelle Antwort auf diese Frage lautet: Einen lexikographischen Eintrag gibt es erst ab einer bestimmten **Gebrauchsfrequenz** relativ zu einem ausgewogenen, allgemeinsprachlichen Materialcorpus. Diese Einschränkung wird u.a. mit dem begrenzten Raum der lexikographischen Darstellung, mit Aspekten der Arbeitskapazität und auch mit einer Orientierung am Sprachsystem begründet.

Gegen diese traditionelle Auffassung kann man zunächst einwenden, dass der begrenzte Raum in der lexikographischen Darstellung im digitalen Medium kein relevanter Gesichtspunkt mehr ist. Sodann kann man sich im Hinblick auf die Arbeitskapazität gestufte Dokumentationstiefen denken, mit denen lexikalische Eintagsfliegen bzw. sprachliche Mittel mit einer eingeschränkten Gebrauchsgeschichte immerhin in ihrer Existenz dokumentiert werden, auch wenn sie nicht lexikographisch bearbeitet werden. Die Tatsache, dass ein Ausdruck wie *Scud-Rakete* in den Zeiten des Golf-Kriegs eine beschränkte Gebrauchsgeschichte hatte, legt die Frage nahe, wie sich solche zeitweise genutzte sprachliche Mittel zum Bestand der langfristig etablierten Mittel verhalten.

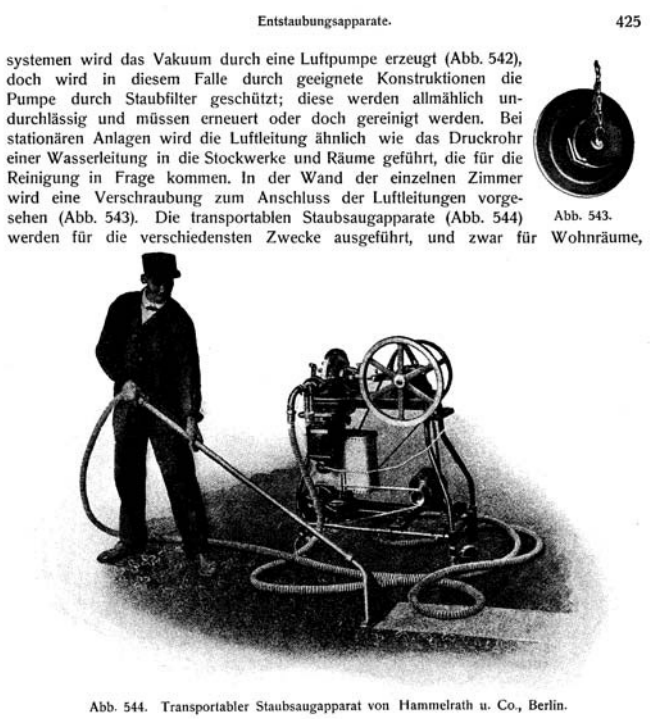
Damit ist man aber bei einer **Perspektive**, die nicht mehr nur die »Langue« im Blick hat, sondern die **kommunikative**n **Ressourcen**, die in bestimmten sprachhistorischen Zeiträumen, und seien dies ein paar Monate mit einem prominenten Themenstrang, genutzt wurden. Dazu gehören demnach auch die sprachlichen Mittel, die nur mehr oder weniger kurzlebigen Themensträngen zugeordnet waren, und auch die evolutionär weniger erfolgreichen lexikalischen Mittel, die eine Zeit lang zwar im Pool der Varianten verfügbar waren, die dann aber nicht langfristig weiter gebraucht wurden.

Wenn man als Lexikograph nur begrenzte Zeit zur Bearbeitung von Material hat oder wenn man nur begrenzten Platz zur Darstellung hat, dann sind Ausschlusskriterien wie die fehlende Konventionalisierung gut nachvollziehbar und begründbar. Die Folge ist aber, dass das, was aus einer kommunikativen Perspektive besonders

interessant ist, durch die Maschen fällt. Wenn man dagegen an den Grundlagen der **Verständigung** (»Welche Rolle spielen nicht-konventionalisierte sprachliche Mittel für die Verständigung?«), an den Prinzipien der **sprachlichen Entwicklung** (»Welche Rolle spielen die zu einem bestimmten Zeitpunkt nicht-konventionalisierten Mittel für die sprachliche Evolution?«) und auch an einer genauen Beschreibung der **sprachlichen Situation zu einem bestimmten Zeitpunkt** (»Wie kann man die komplexe Architektur des Wortgebrauchs um 1900 differenziert beschreiben?«) interessiert ist, dann sind auch genau die Mittel, die sich nicht oder noch nicht »durchgesetzt« haben, sehr wertvolle Elemente in einem umfassenderen Bild der sprachlichen Dynamik. Zur sprachlichen Dynamik gehören auch die sprachlichen Entwicklungen, die sich aus der Themengeschichte, den Thematisierungspraktiken usw. in einem bestimmten Zeitraum ergeben, auch in den jeweils eigenen thematischen Vernetzungszusammenhängen (z.B. Sport und Wehrkraft; Sicherheit und Terrorismus; Ernährung und Gesundheit; Gesundheit und Sport).

5.2 Zum Beispiel: »Staubsauger« und sein thematisches Umfeld

Ich gebe nun ein Beispiel, die Geschichte des Wortes ***Staubsauger*** und seiner Konkurrenten bzw. seiner nahen Verwandten. In der Medizin um 1900, speziell in der damals neu konstituierten Teildisziplin der **Hygiene**, wurde auch der Zusammenhang von Gesundheitsfürsorge und Sauberkeit thematisiert, in diesem Zusammenhang wurden auch neue Geräte besprochen, mit denen sich Staub auf technischem Wege, durch Absaugung, beseitigen ließ. Im *Atlas und Lehrbuch der Hygiene* (PRAUSNITZ 1909, 422-426) zum Beispiel werden solche Geräte mit ganz unterschiedlichen Wörtern und Wendungen bezeichnet: *Entstaubungsapparat*, *Sauganlage*, *Entstaubungssystem*, *Staubsaugapparat*, *Entstaubungsanlage*, *Vakuumapparat*, *Wohnungsentstaubungsapparat*, *Apparate zur Staubabsaugung*.



Entstaubungsapparate. 425

systemen wird das Vakuum durch eine Luftpumpe erzeugt (Abb. 542), doch wird in diesem Falle durch geeignete Konstruktionen die Pumpe durch Staubfilter geschützt; diese werden allmählich undurchlässig und müssen erneuert oder doch gereinigt werden. Bei stationären Anlagen wird die Luftleitung ähnlich wie das Druckrohr einer Wasserleitung in die Stockwerke und Räume geführt, die für die Reinigung in Frage kommen. In der Wand der einzelnen Zimmer wird eine Verschraubung zum Anschluss der Luftleitungen vorgesehen (Abb. 543). Die transportablen Staubsaugapparate (Abb. 544) werden für die verschiedensten Zwecke ausgeführt, und zwar für Wohnräume,

Abb. 543.

Abb. 544. Transportabler Staubsaugapparat von Hammelrath u. Co., Berlin.

Quelle:
PRAUSNITZ 1909, S. 425

Der Ausdruck *Staubsauger* kommt in diesem Text nicht vor, er wird aber in anderen Texten seit dem späten 19. Jahrhundert immer wieder verwendet (siehe hierzu die Belegdokumentation im Anhang). Die Gebrauchs- und **Verbreitungsgeschichte** des Wortes *Staubsauger* und seiner Konkurrenten bzw. nahen Verwandten lässt sich drei **thematischen Bereichen** zuordnen, in denen über Staubsauger bzw. Staubsaugvorrichtungen geschrieben wird: dem Bereich der Technik, dem Bereich der Hygiene (insbesondere der Gewerbehygiene) und – zeitlich etwas später – dem Bereich der alltäglichen Haushaltsführung. Im Feld der Konkurrenzausdrücke wurde mit *Staubsauger* vor allem eine trag- bzw. fahrbare, also bewegliche Variante von Geräten zur Staubabsaugung bezeichnet, während die großen, ortsfest installierten Geräte z.B. mit Ausdrücken wie *Entstaubungsanlage* bezeichnet wurden, wobei dann in einem Fall mit *Staubsauger* auch ein Teil einer solchen großen Entstaubungsanlage gemeint war. Im *Maschinentechnischen Lexikon* (Hg. Felix KAGERER, Wien 1912) finden wir zum Beispiel nur einen Verweiseintrag »Staubsauger s. Entstaubungsanlagen« (S. 890), im Artikel *Entstaubungsanlagen* wird der

Ausdruck *Staubsauger* selbst aber nicht verwendet, obwohl er in der Zeit durchaus gebräuchlich war. Hier zeigt sich offenbar die Unterscheidung großer, industriell genutzter Anlagen von den kleineren, beweglichen Geräten, wie sie im Haushalt verwendet wurden.

Die **Verbreitungsgeschichte** des Ausdrucks *Staubsauger*, soweit sie sich im Moment schon überblicken lässt, folgt der technischen **Nutzungsgeschichte** in den entsprechenden **Literaturbereiche**n: So finden wir Belege zum einen in der Literatur zu technischen Neuerungen, in den Spezialblättern einzelner Gewerbe, in denen Staubsauger eingesetzt wurden (z.B. Buchdruckerei, Eisenbahntechnik, Bergbau), sodann in Texten zur Gewerbehygiene und auch zur allgemeinen Hygiene, über die allgemeine Hygiene dann offenbar auch in Texte, in denen die alltägliche Haushaltsführung thematisiert wird, zum Teil auch in einer kulturgeschichtlichen oder kulturvergleichenden Perspektive, hier ist insbesondere der amerikanische Einfluss auf die Technikentwicklung und Techniknutzung erkennbar. In den Belegen sind auch Hinweise interessant, die den Gegenstand Staubsauger als neue Entwicklung kennzeichnen, z.B. der Gebrauch des Ausdrucks in Verbindung mit *sogenannt* (»mittels sogenannter Staubsauger«), oder auch Bemerkungen, die den Etablierungsgrad der neuen Technik thematisieren:

> »Hier ist es daher angebracht, möglichst oft den Staub mittels eines Staubsaugers zu entfernen; neuerdings verwendet man hierfür elektrische Apparate. Erwähnt sei hierbei, daß früher die Setzerlehrlinge die Aufgabe hatten, mit einem Blasebalg die Kästen von Staub zu befreien, wobei die jungen Leute naturgemäß den schlimmsten Gefahren ausgesetzt waren« (1913; Hervorhebung T. G).

Wie stellen unsere sprachhistorischen **Wörterbücher** die Wortgeschichte von *Staubsauger* dar? Im **PAUL**schen Wörterbuch (2002, Spalte 957a) finden wir im Artikel *Staub* die Angabe: »dazu *Staubsauger* (1910 *DWb*)«. Von den lexikalischen Mitbewerbern wie *Entstaubungsapparat* ist erwartungsgemäß in einem einbändigen Werk dieser Art nicht die Rede. Im *Deutschen Wörterbuch* (***DWb*** 17, 1074) lesen wir im Artikel *Staub* unter der Position II.1.h:

> »h) in den räumen eines hauses, schon nach kurzer zeit auf alles gerät sich lagernd; gegen ihn kämpft die hausfrau mit dem *staubbesen, staubtuch, staubwedel* u. s. w., jetzt mit der *staubmaschine*, dem *staubsauger* (s. unten).«

Mit der Formulierung »jetzt mit ... dem *staubsauger*« wird signalisiert, dass es sich beim Gebrauch von Staubsaugern im Haushalt um eine neuere Entwicklung handelt, ohne dass aber die damit zusammen-

hängende wortgeschichtliche Entwicklung explizit erläutert würde. Auch der erwähnte eigene Wortartikel zu *Staubsauger* (*DWb* 17, 1122) enthält keine Belege und lautet folgendermaßen:

> STAUBSAUGER, m. gerät, welches gegenstände durch absaugen von dem darauf lagernden staub reinigt. vgl. oben staubsammler, zu dem es sich als eine ergänzende wortbildung stellt, ohne aber in mehr als der idee mit ihm übereinzustimmen.

Immerhin haben wir mit diesen beiden Einträgen eine relativ frühe lexikographische Buchung; die Lieferung mit den Artikeln *Staub* und *Staubsauger* erschien im Jahr 1910.

Prüfen wir nun noch einige weitere wortgeschichtliche Auskunftsmittel: Im **TRÜBNER**schen Wörterbuch (VI, 544) wird das Wort *Staubsauger* im Artikel *Staub* wohl erwähnt, aber ohne jegliche semantische oder wortgeschichtliche Kommentierung und ohne Belege. Im 1910 erschienenen zweiten Band des Wörterbuchs von **WEIGAND** & **HIRT** ist das Kompositum im Artikel *Staub* nicht aufgeführt. Auch in der zweiten Auflage des dritten Bandes von Moriz **HEYNE**s *Deutschem Wörterbuch* (1906; III 762) findet sich zwischen *Staubregen* und *Staubtuch* kein Eintrag *Staubsauger*. Das *Etymologische Wörterbuch* von **PFEIFER** enthält ebenfalls keinen Eintrag dazu. Im **DWDS-Corpus** sind 190 Belege verzeichnet, 160 davon sind nach Anmeldung einsehbar, sie reichen bis ins Jahr 1924 hinab und sind überwiegend dem thematischen Bereich der alltäglichen Haushaltsführung zuzuordnen. Die frühe Verbreitungsgeschichte und die technischen und die medizinisch-hygienischen Thematisierungsanteile sind in dieser Dokumentation also nicht erkennbar.

Ich ziehe ein **Zwischenfazit** aus der exemplarischen Betrachtung zur Wortgeschichte von *Staubsauger* und ihrer Dokumentation:

(i) Die **Wortgeschichte** der Bezeichnung *Staubsauger*, die exemplarisch ist für ein Stück Technikgeschichte und auch für die Veränderungen des lebensweltlichen Alltags, ist in den wortgeschichtlichen Auskunftsmitteln unzureichend dargestellt.

(ii) Die Wortgeschichte von *Staubsauger* enthält bislang keinerlei Hinweise auf ein zeitgenössisches Feld von Konkurrenzausdrücken bzw. von funktional nahe verwandten Ausdrücken, die in den Texten zum Teil für Zwecke der Variation im Ausdruck verwendet werden; ihre Kenntnis ist auch deshalb wichtig, weil es keineswegs ausgemacht ist, dass *Staubsauger* der kommunikationsgeschichtlich aussichtsreichste Kandidat war: Warum verwenden wir heute nicht *Vakuumreiniger*? Hinzu kommt, dass sich

offenbar funktionale Differenzierungen stabilisiert haben zwischen Bezeichnungen für die großen, ortsfest installierten Anlagen und die kleineren, beweglichen Geräte.

(iii) Die kommunikative Verankerung der Wortgeschichte in den Themenfeldern Technik, (Gewerbe-)Hygiene und alltägliche Lebensführung, wie ich sie oben in ihrer Entwicklungsdynamik skizziert habe, ist bislang auch in Ansätzen nicht erkennbar.

(iv) Vom Feld der Konkurrenzausdrücke und vom lexikalischen Variantenpool aus betrachtet, kann man sich spekulativ auch die Frage stellen, welche Ausdrucksweisen in der Zeit um 1900 *nicht* genutzt wurden, obwohl sie prinzipiell möglich gewesen wären: Wenn man z.B. sieht, dass wir heute den Ausdruck *Folie* auch für die Resultate von Beamer-Projektionen verwenden, für die technischen Nachfolger von Folien also, die mit ›richtigen‹ Folien nichts mehr zu tun haben, dann hätte man sich auch vorstellen können, die frühen Staubsauger um 1900 z.B. mit dem Wort *Vakuum-Staubtuch* zu bezeichnen. Aber so ist es nicht gewesen. Soweit ich sehe, haben die Zeitgenossen um 1900 diese Möglichkeit nicht genutzt.

5.3 Aspekte der thematischen Prägung des Wortgebrauchs in Wörterbuchartikeln

Nun stellt sich die Frage, ob und ggf. wie sich Aspekte der thematischen Prägung des Wortgebrauchs auch in lexikographischen Beschreibungen einzelner Wörter bzw. einzelner Verwendungsweisen erfassen, veranschaulichen und dokumentieren lassen? Zu diesen Aspekten gehören u.a. die Verankerung in bestimmten thematischen Bereichen (wie z.B. Technik, Gewerbehygiene), die Enfaltung der Thematisierungsgeschichte im zeitlichen Längsschnitt, die Verteilung in unterschiedlichen Kommunikationsbereichen und Texttypen, die Stellung innerhalb einer thematischen Systematik, die Frage der Offenheit und der Verfestigung des Wortgebrauchs in bestimmten Phasen usw.

5.3.1 Wortgeschichtliche Überblicksdarstellungen; erneut »Staubsauger«

Ein erster Vorschlag besteht darin, Aspekte der **thematischen Prägung** mit Hilfe von **narrativen Darstellungsmustern** in die Organisation wortgeschichtlicher Entwicklungen einzubauen. Hier ein Versuch, der die oben skizzierte Entwicklung des Gebrauchs von

Staubsauger in kondensierter Form organisiert und dabei durchaus auch Aspekte der lexikalischen Variation integriert:

STAUBSAUGER Das Wort *Staubsauger* ist seit dem Ende des 19. Jahrhunderts in deutschsprachigen Texten belegt, zunächst in Texten, in denen Staubsauger (bzw. unterschiedliche Geräte zur Beseitigung von Staub mittels Absaugtechnologien) als technische Innovationen beschrieben und diskutiert werden, sodann in Texten, in denen der medizinisch-hygienische Wert von Staubsaugern thematisiert wird (vor allem in Bezug auf Felder der Gewerbehygiene wie z.B. Buchdruck, Eisenbahnbetrieb, Gebäudereinigung), schließlich in Bezug auf die Rolle von Staubsaugern im Haushalt als Arbeitserleichterung und als gesundheitsförderliches Gerät. Seit den 1920er Jahren steht der Gebrauch mit Bezug auf lebensweltliche Haushaltsgeräte im Vordergrund (siehe DWDS-Corpus). Um 1900 brachte die neue Technologie der Staubabsaugung einen breiten Variantenpool von Bezeichnungen hervor, u.a. *Entstaubungsapparat*, *Entstaubungsanlage*, *Vakuumreiniger* (zu engl. *vacuum cleaner*) und zahlreiche andere. Der Ausdruck *Staubsauger* bezog sich dabei auf kleinere, bewegliche und nicht ortsfest installierte Geräte, während Ausdrücke wie *Entstaubungsanlage* sich auf große, ortsfest installierte Einrichtungen bezog. Vereinzelt wird *Staubsauger* in Bezug auf solche größeren Anlagen auch für den Teil genutzt, mit dem die eigentliche Reinigungsarbeit vollzogen wird. Spätestens um 1910 erscheint aber der Gebrauch zur Bezeichnung eines Haushaltsgeräts stabil etabliert. – Lit.: *DWb* 17, 1074 und 1122 (ohne Belege; Buchung: 1910); TRÜBNER VI 544 (ohne wortgeschichtlichen Kommentar und ohne Belege);

(1889) Zeitschrift für Nahrungsmittel-Untersuchung und Hygiene (= Oesterreichische Chemiker-Zeitung), Jg. 3, S. 188

(1894) Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft, Bd. 18/1-2, S. 254f.

(1897) Theodor WEYL, *Handbuch der Hygiene*, Bd. 8, S. 727f.

(1903) Die Gartenlaube. Illustriertes Familienblatt, S. 613

(1904) Bericht über die I. Versammlung der Tuberkulose-Ärzte, S. 38

(1905) Gordian. Zeitschrift für die Kakao-, Schokoladen- und Zuckerwaren-Industrie und für alle verwandten Erwerbszweige, Bd. 11, S. 61

(1906) Museumskunde. Zeitschrift für Verwaltung und Technik öffentlicher Sammlungen, Bd. 2, S. 27

(1907) Vierteljahrsschrift für gerichtliche Medizin und öffentliches Sanitätswesen, S. 168

(1907) Journal suisse de médecine, S. 182f.

(1907) Charité-Annalen, S. 29
(1907) Zeitschrift für Tuberkulose, Bd. 10, S. 407
(1907) *Dinglers Polytechnisches Journal*, Jg. 88, Bd. 322, Heft 2 (»Einige bemerkenswerte Neuerungen auf der Ausstellung zu Mailand 1906«), S. 19
(1908) Handbuch des Eisenbahnmaschinenwesens, Bd. 3 (Werkstätten), S. 371
(1908/09) STRINDBERG, A., *Fröhliche Weihnacht* (Gesammelte Werke I/10, 1920), S. 105
(1912) SCHLEKER, K., *Die Frau und der Haushalt*, S. 194
(1912) Zeitschrift der Deutschen Gesellschaft für Mechanik und Optik, S. 234
(1913) FISCHER, A., Grundriss der sozialen Hygiene, S. 262
(1914) First German Reader, S. 217
(1914) *Weyl's Handbuch der Hygiene*, 2. Aufl., Bd. 4, S. 266
(1918) Technik und Industrie. Jahrbuch der Technik, Jg. 4 (1917/18), S. 141
(1918) Technik und Industrie. Jahrbuch der Technik, Jg. 4 (1917/18), S. 141
(1918) Technik und Industrie. Jahrbuch der Technik, Jg. 4 (1917/18), S. 143
(1918) Technik und Industrie. Jahrbuch der Technik, Jg. 4 (1917/18), S. 144
(1918) Technik und Industrie. Jahrbuch der Technik, Jg. 4 (1917/18), S. 145
(1921) HEDIN, A., *Arbeitsfreude*, S. 164

5.3.2 Thematische Markierungen zu Verwendungsweisen

Narrative Formen haben den Vorteil der Flexibilität und der Freiheit der Darstellung, sie sind allerdings schwer ›abfragbar‹ und kaum systematisch erschließbar. Diesen Nachteil kann man durch Formen der **Markierung** von Aspekten der **thematischen Prägung** auffangen.

Zum einen kann man hierfür die einzelnen **Quellentexte**, die in die Belegdokumentation zu einzelnen Verwendungsweisen eingehen, mit thematischen Deskriptoren versehen, sofern das thematische Profil des jeweiligen Textes entsprechend einheitlich ist und dies zulässt (siehe hierzu oben den einschränkenden Hinweis zu Fußball vs. Rugby).

Zum anderen kann man aber auch die einzelnen **Verwendungsweisen** mit kontrollierten thematischen Markierungen wie z.B. »Technik«, »Medizin/Gesundheit« oder »Haushaltsführung« anreichern. Neben alltagssprachlichen Themen-Schlagwörtern (z.B. »Technik«) sind auch formale Markierungen im Rahmen einer historischen Themen-Ontologie denkbar, die historische Denkfiguren und historische Thematisierungsgesichtspunkte in einen handhabbaren Markierungszusammenhang bringen müsste. Eine hypothetische Markierung wie z.B. »134.12.03« wäre etwa ein denkbares Gegenstück für 134 = Medizinische Säftelehre, 134.12 = Idee der Primär-

qualitäten im Rahmen der medizinischen Säftelehre, 134.12.03 = Medizinische Säftelehre, Idee der Primärqualitäten, Primärqualität ›feucht‹.

Die Erstellung einer **thematischen »Ontologie«** für eine derartige Verschlagwortung ist nicht trivial, da sich thematische Strukturen und ihre Zusammenhänge im zeitlichen Wandel laufend verschieben können. Dennoch erscheinen solche Formen der thematischen Verschlagwortung aufschlussreich, in einer nächsten heuristischen Phase wäre es sinnvoll, einzelne Themenfelder (Fußball, Rassenhygiene, Sexualität, Technik, Kleidung/Mode, ...) über zeitlich gestaffelte Schlüsseltexte zu erschließen, in strukturierter Form zu markieren und damit auch abfragbar zu machen mit Anfragen wie z.B.: »Welche Bezeichnungen für Spielerpositionen im Fußball / für Kleidungsstücke / für sexuelle Praktiken / für Haushaltsgeräte / für Ideologeme der Rassenhygiene usw. gab es im Deutschen in den 1920er Jahren?« Wenn darüber hinaus auch noch weitere lexikologische Gesichtspunkte wie z.B. Aspekte der Entlehnung, der Metaphorik, der Wortbildungsstruktur, der Regionalität usw. kodiert sind, erlaubt dies auch Quer-Abfragen, zum Beispiel solche zur Rolle von Wortbildungen, zur Funktion der Metaphorik usw. in bestimmten thematischen Feldern.

5.3.3 Integration lexikologischer Untersuchungsbefunde zu Aspekten der thematischen Prägung: Literaturverweise und Zusammenfassungen

Ein weiterer Fragenkomplex bezieht sich auf die Möglichkeit der Verknüpfung von **lexikographische**n Darstellungsformen mit den Erträgen **monographische**r **Darstellungen**: Wie lassen sich ggf. die Befunde aus lexikalischen Analysen, die thematische Gesichtspunkte in den Vordergrund stellen, mit traditionellen Wörterbuchartikeln bzw. mit den Einträgen in neuartigen digitalen lexikographischen Systemen verbinden? Hier kann man die auch jetzt schon in Wörterbüchern geübte Praxis nutzen, Untersuchungsbefunde zusammenzufassen und mit weiterführenden Literaturhinweisen zu versehen.

Hierzu ein **Beispiel**: In der Neubearbeitung des *Deutschen Wörterbuchs* sind die Artikel zu ***Atomkraft*** und *Atomenergie* nur äußerst knapp formuliert und dokumentiert, ein Eintrag *Atomstrom* fehlt ganz.[10] Hier böte sich die Möglichkeit an, auf die von Matthias JUNG

[10] Man vergleiche aber die sehr viel ausführlichere, differenzierte und auch die unterschiedlichen Kommunikationsbereiche dokumentierende Darstellung in der Neubearbeitung des *Deutschen Fremdwörterbuchs* (*DFWbN* 2, 460ff.), auf die in der Neubearbeitung des *Deutschen Wörterbuchs* beim Grundwort *Atom* auch im

veröffentlichte Arbeit *Öffentlichkeit und Sprachwandel. Zur Geschichte des Diskurses über die Atomenergie* (1994) sowie auf seine kleineren Beiträge, z.B. den entsprechenden Beitrag im Sammelband *Kontroverse Begriffe* (JUNG 1995; vgl. auch 1994b) als Teil einer wortgeschichtlichen Dokumentation zu verweisen. Eine wortgeschichtliche Dokumentation, in der die thematische Prägung mit beschrieben wird, müsste neben dem Hinweis auf den fachsprachlichen Ursprung im Themenbereich der Kernphysik auch auf die Verbreitung in die Themenbereiche der Militärtechnik und der zivilen Energieversorgung dokumentieren, sie müsste darüber hinaus auch zeigen, dass die beiden letztgenannten Themenfelder auch Gegenstände lang andauernder öffentlicher Diskurse waren, in denen Ausdrücke wie *Atomkraft*, *Atomenergie* und weitere Komposita, aber auch Konkurrenzausdrücke auf *Kern-* eine zum Teil zentrale Rolle gespielt haben. Es geht bei diesem Vorschlag nicht darum, einen DWbN-Artikel, der ja bestimmten Bearbeitungsvorgaben folgt, zu kritisieren, es geht um die Frage, wie der wortgeschichtlich relevante Aspekt der thematischen Prägung in lexikographisch-lexikologische Beschreibungen zukünftig da besser integriert werden kann, wo es sinnvoll und nötig erscheint. Die Darstellung im *DFWbN* kommt diesen Vorschlägen schon sehr nahe, auch wenn für eine digitalisierte Fassung einer solchen Darstellungsform Ergänzungen sinnvoll sind: zum Beispiel auch hier der Verweis auf die genannten monographischen, diskursorientierten Untersuchungen, weiterhin die Markierung und Suchbarkeit der artikelinternen Stichwörter von Wortbildungen, die in komplexeren Artikeln unter den einzelnen Bedeutungspositionen gesucht werden müssen. Will man zum Beispiel prüfen, ob *Atomstrom* im *DFWbN* behandelt ist, muss man zunächst nach der entsprechenden Bedeutungsposition von *Atom-* suchen, unter der das Kompositum verzeichnet sein müsste. Im Beispiel kommen hierfür 2c ‚durch Kernenergie erzeugt, angetrieben, Kern-' (*DFWbN* 2, 462) infrage, sodann aber auch die Bedeutungsposition 2e ‚im Hinblick auf negative Folgen der Kernenergie gesehen' (463).

Sinne einer vernünftigen Arbeitsteilung verwiesen wird. Der zweite Band des *DFWbN* mit der Wortstrecke *Atom-* ist 1996 erschienen, die entsprechende frühere Lieferung konnte von den diskursanalytischen Arbeiten zur Atomenergiediskussion (JUNG 1994; JUNG 1995) vermutlich nicht mehr profitieren, die entsprechenden Diskussionen sind aber genannt, zeitlich verortet und mit Belegen dokumentiert.

6. Themen – Diskurse – Wortgebrauch

Diskurse sind in der deutschen Sprachwissenschaft seit den 1980er Jahren als ›neue‹ Gegenstände in den Blick gekommen. Auch für die Diskursforschung sind Themen von zentraler Bedeutung. Unter **Diskurse**n versteht man in dieser Perspektive Ensembles von Texten, die in erster Linie durch den gemeinsamen Bezug auf ein bestimmtes **Thema** (ein ›Problem‹, eine ›Fragestellung‹, einen öffentlichen Streitpunkt etc.) und zum anderen durch Verweisungen bzw. intertextuelle Bezüge gekennzeichnet sind. Da es forschungspraktisch in vielen Fällen nicht möglich ist, *alle* Texte zu untersuchen, die sich auf ein Thema beziehen, werden ›virtuelle Corpora‹ herangezogen, also Zusammenstellungen von Texten, die wesentliche Eigenschaften des entsprechenden Diskurses abbilden sollen und die unter den Bedingungen des wissenschaftlichen Alltags mit seinen Beschränkungen realistischerweise auch bearbeitbar sind. Eine frühe Kernstelle aus dem Beitrag von BUSSE & TEUBERT (1994) lautet:

> »Unter Diskursen verstehen wir im forschungspraktischen Sinn virtuelle Textkorpora, deren Zusammensetzung durch im weitesten Sinne inhaltliche (bzw. semantische) Kriterien bestimmt sind. Zu einem Diskurs gehören alle Texte, die
> – sich mit einem als Forschungsgegenstand gewählten Gegenstand, Thema, Wissenskomplex oder Konzept befassen, untereinander semantische Beziehungen aufweisen und/oder in einem gemeinsamen Aussage-, Kommunikations-, Funktions- oder Zweckzusammenhang stehen,
> – den als Forschungsprogramm vorgegebenen Eingrenzungen in Hinblick auf Zeitraum/Zeitschnitte, Areal, Gesellschaftsausschnitt, Kommunikationsbereich, Texttypik und andere Parameter genügen,
> – und durch explizite oder implizite (...) Verweisungen aufeinander Bezug nehmen (...)« (BUSSE & TEUBERT 1994, 14).

Themen, Teilthemen und Thematisierungspraktiken – z.B. Verfahren des Argumentierens oder der Etablierung von Sichtweisen auf umstrittene Gegenstände der öffentlichen Diskussion – hängen häufig mit bestimmten Formen des **Wortgebrauch**s zusammen. Ob jemand ein Ereignis X mit dem Ausdruck *Annexion* oder mit *Grenzberichtigung* bezeichnet, gibt Aufschluss darüber, wie er/sie das Ereignis X beurteilt und welche Sichtweise er/sie mit der sprachlichen Thematisierung verbindet.[11] Die Analyse des Wortgebrauchs kann in diesen Fällen Aufschluss geben über kommunikative Ziele und die Sicht-

[11] Beispiel aus STÖTZEL & WENGELER 1995, 32.

weisen, die einzelne SchreiberInnen oder Verbünde von SchreiberInnen im Rahmen von Diskursen vertreten. Untersucht man diese sprachlichen Praktiken in einem größeren Zusammenhang, dann kann die Analyse des Wortgebrauchs einen wesentlichen Beitrag leisten zur Frage, wie zeitspezifische Auffassungen sprachlich konstitutiert werden, seien sie nun politischer, wissenschaftlicher, fachlicher oder alltagsweltlicher Natur. Inzwischen gibt es vielfältige sprachwissenschaftliche Befunde zur Rolle des Wortgebrauchs in Diskursen, z.B. in dem methodisch und substantiell grundlegenden Sammelband »Kontroverse Begriffe« (STÖTZEL & WENGELER 1995) und in weiteren Arbeiten aus dem Umkreis der Düsseldorfer Schule um Georg STÖTZEL, aber auch in mehreren Arbeiten von Heidrun KÄMPER.[12] Als eine wesentliche Konsequenz für die **lexikographische Beschreibung diskursgeprägter Wörter und Verwendungsweisen** ergibt sich, dass die Verankerung und die spezifische Rolle von Wörtern und Verwendungsweisen in Diskursen auch ein Teil der semantisch-lexikologischen Beschreibung in traditionellen Wörterbüchern und Informationssystemen werden sollte, sei es als eigenes Beschreibungselement, sei es immerhin als Verweisung auf einschlägige Studien.

Im Hinblick auf Darstellungsformen und **Beschreibungsformate** lassen sich zwei Traditionen ausmachen, die Darstellung der thematisch-diskursiven Prägung von Wortgebräuchen in quasi-monographischer Form (Bücher, Artikel) sowie diskurslexikographische Ansätze.

In quasi-**monographisch**en Darstellungen werden die Wortgebräuche in Diskursen narrativ charakterisiert und in der Regel durch exemplarische Textzitate und ggf. durch weitere Stellenangaben dokumentiert, die zugrundeliegenden Corpus-Texte selbst sind häufig nicht dokumentiert und möglicherweise auch nur schwer dokumentierbar, den Zugriff auf einzelne Wörter ermöglicht ein alphabetisches Register, das dann zu den Stellen der jeweiligen Untersuchung führt, an denen einzelne Wörter, ihre Verwendungsweisen und ihre Rolle in bestimmten Diskursen kommentiert wird. Es ist eine naheliegende Idee, die einzelwortbezogenen lexikographisch-lexikologischen Dokumentationen in digitalen Wortgeschichtssystemen auch zu verbinden mit den Resultaten der monographischen Untersuchungen zu einzelnen Diskursthemen und auch mit den verfügbaren thematisch zentralen Volltexten eines bestimmten

[12] Vgl. u.a. JUNG 1994; BÖKE *et al.* 1996; KÄMPER 2006, 2007, 2012.

Diskurses. Wenn man die *Atom*-Wortstrecke im *DFWbN* abgleicht mit den Wortregistern in den Arbeiten von Matthias JUNG, dann ergibt sich manche Ergänzung und auch die Möglichkeit einer vertieften und stärker kontextualisierten Einbettung in diskursive Sprachgebrauchszusammenhänge. Wo im Wörterbuch Komprimierung angesagt ist, erlaubt die Monographie breitere Entfaltung kommunikationsgeschichtlicher Zusammenhänge, die man durch Verweise und ggf. durch Kurzzusammenfassungen auch nutzen kann.

Neben den Untersuchungen gibt es inzwischen auch gut ausgearbeitete Vorschläge und Ansätze für lexikologisch-**lexikograpische Umsetzungen** von diskursorientierten Fragestellungen sowie Ansätze und Beispiele einer historischen Diskurslexikographie. Hier sind exemplarisch zu nennen der Darstellungenverbund zum Schulddiskurs der Nachkriegszeit von Heidrun KÄMPER, der eine groß angelegte Untersuchung (2005), ein darauf bezogenes Wörterbuch (2007) und eine methodische Reflexion über diskurslexikographische Verfahren (2006) umfasst. Das *Wörterbuch zum Schulddiskurs* verbindet mit dem *Diskurshistorischen Wörterbuch zur Einwanderung seit 1945* (JUNG *et al.* 2000) und der früheren zeitgeschichtlich orientierten Dokumentation *Brisante Wörter* (STRAUSS *et al.* 1989), dass hier jeweils Aspekte der Diskursstruktur (Beteiligungsrollen; zentrale Themen) als Grundlage für die Organisation des Materials gewählt wurden. In vergleichbarer Weise verbindet Katja FAULSTICH in ihrer Untersuchung zum Sprachnormierungsdiskurs im 18. Jahrhundert (2008) die Darstellung von Diskursbereichen mit einer jeweils anschließenden Dokumentation einschlägiger »Diskurslexik«, auch hier wird im zweiten Kapitel über die Theorie und Methode dieser Verfahrensweise reflektiert. Jochen BÄR (1998; 1999) hat in seinen Arbeiten zum frühromantischen Diskurs insbesondere die Verbindung von diskurs- und textlexikographischen Verfahren betont.

Als Fazit dieser Teildiskussion können wir festhalten. (i) Auch diejenigen Themen, die in öffentlichen Diskursen verankert sind, sind mit ihren Schlüsselwörtern, Leitvokabeln und thematisch geprägten Verwendungsweisen wichtige Gegenstände der lexikographischen Dokumentation, auch wenn die sprachlichen Mittel nur eine begrenzte zeitliche, räumliche oder kommunikative Verbreitung gehabt haben sollten. (ii) Die Diskurslexikographie stellt inzwischen wertvolle empirsche Untersuchungsbefunde, aber auch wichtige konzeptionelle Vorschläge für lexikographische Darstellungs- und Dokumentationsformen bereit, die sich auch für digitale Systeme fruchtbar nutzen lassen. Das gilt auch für themen- und diskursorientierte

Beschreibungsansätze und -elemente, die in traditionellen Wörterbüchern bereits integriert sind, zum Beispiel in der Neubearbeitung des Deutschen Fremdwörterbuchs.

7. Zusammenfassung und Rückblick

Der zentrale Gegenstand dieses Beitrags ist die Frage, wie Themen als grundlegende Aspekte der sprachlichen Verständigung mit dem Wortgebrauch zusammenhängen und wie diese Zusammenhänge – die thematische Prägung des Wortgebrauchs – auch für die lexikographisch-lexikologische Dokumentation des Wortgebrauchs fruchtbar gemacht werden kann.

Für eine elementare Klärung des Themenbegriffs kann man auf Ergebnisse der Gesprächsforschung, der Textlinguistik und auch der Diskursforschung zurückgreifen. Themen sind nicht nur ein wesentliches Strukturierungsprinzip aktueller Kommunikation, sie sind auch ein zentraler Aspekt der sprach- und kommunikationsgeschichtlichen Entwicklung. Themen entwickeln sich historisch, Themen und Teilthemen weisen historisch veränderliche Zusammenhänge untereinander auf, auch die lexikalischen Mittel für die Bewältigung von Themen im historischen Längsschnitt weisen eine eigene Dynamik auf. In einem eigenen Abschnitt werden auch Zusammenhänge zwischen dem thematischen Gesichtspunkt und den Grundannahmen der (historischen) Diskursforschung beleuchtet, die Themen untersucht, insofern sie Gegenstände öffentlicher Diskussionen sind und insofern die entsprechenden Texte bzw. Kommunikationsereignisse untereinander Verweisungen zeigen.

In der historischen Lexikographie des Deutschen sind Themen als zu kontrollierende Variationsparameter noch nicht sehr deutlich ausgeprägt: nicht bei der Kontrolle der Textauswahl und auch nicht in den Überlegungen zur Beschreibungs- und Zugriffsmethodik. Anhand von Beispielen (u.a. aus den Themenbereichen Rassenhygiene, Naturschutz, Sport/Fußball, Hygiene, Technik und Haushalt) mache ich Vorschläge, wie man den Themenbezug historischer Kommunikation und die thematische Prägung lexikalischer Mittel auch in lexikographischen Darstellungen stärker verankern kann. Die Vorschläge beziehen sich u.a. auf folgende Aspekte:

- Fragen der Corpusplanung und der Kontrolle thematischer Anteile bei der Zusammenstellung historischer Textcorpora;
- Möglichkeiten der narrativen Darstellung von Wortgeschichten in thematischen Bezügen;

– Möglichkeiten der Markierung thematischer Aspekte des Wortgebrauchs in digitalen lexikalischen Systemen.

Anhand von zwei Textbeispielen wird auch erläutert, wie thematische Schlüsseltexte lexikographisch-lexikologisch genutzt werden können, um historische Systemstellen von Themen und Teilthemen mit den entsprechenden Wortschatzsektoren zu füllen.

8. LITERATUR

ADATO, A., 1971: *On the sociology of topics in ordinary conversation. An investigation into the tacit concerns of members for assuring the proper conduct of everyday activities*, Diss. University of California, Los Angeles.

BÄR, J. A., 1998: Vorschläge zu einer lexikographischen Beschreibung des frühromantischen Diskurses, in: WIEGAND, H. E. (Hrsg.), *Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium*, Tübingen, 155-211.

BÄR, J. A., 1999: *Sprachreflexion der deutschen Frühromantik. Konzepte zwischen Universalpoesie und grammatischen Kosmopolitismus. Mit lexikographischem Anhang,* Berlin.

BÖKE, K., 1994: »Gleichberechtigung« oder »natürliche Ordnung«. Die Diskussion um die rechtliche Gleichstellung der Frau in den 50er Jahren, in: BUSSE, D. *et al.* (Hrsg.), *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik*, Opladen, 84-106.

BÖKE, K., 1995a: »Männer und Frauen sind gleichbereichtigt«. Schlüsselwörter in der frauenpolitischen Diskussion seit der Nachkriegszeit, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 447-516.

BÖKE, K., 1995b: »Lebensrecht« oder »Selbstbestimmungsrecht«? Die Debatte um den Par. 218, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 563-592.

BÖKE, K. *et al.* (Hrsg.), 1996: *Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Georg Stötzel zum 60. Geburtstag gewidmet*, Opladen.

BUSCH, A., 2004: *Diskurslexikologie und Sprachgeschichte der Computertechnologie*, Tübingen.

BUSSE, D. *et al.* (Hrsg.), 1994: *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik*, Opladen.

BUSSE, D. & W. TEUBERT, 1994: Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik, in: BUSSE, D. *et al.* (Hrsg.), *Begriffsgeschichte und Diskursgeschichte.*

Methodenfragen und Forschungsergebnisse der historischen Semantik, Opladen, 10-28.

Deutsches Fremdwörterbuch (*DFWbN*) (1995-2010), begonnen von H. SCHULZ, fortgeführt von O. BASLER, 2. Aufl., völlig neu erarbeitet im Institut für deutsche Sprache, Bd. 1-7, Berlin / New York.

Deutsches Wörterbuch (DWb) von Jacob und Wilhelm Grimm, 16 Bände (32 Teile) und ein Quellenverzeichnis, Leipzig 1854-1971, Nachdruck München 1984. [Online: http://www.DWB.uni-trier.de; CD-Fassung bei Zweitausendeins].

DWDS (*Digitales Wörterbuch der deutschen Sprache*): http://www.dwds.de.

EITZ, TH., 2005: Aids. Krankheitsgeschichte und Sprachgeschichte, in: WENGELER, M. (Hrsg.), *Sprachgeschichte als Zeitgeschichte*, Germanistische Linguistik 180-181, Hildesheim [u.a.], 371-398.

FAULSTICH, K., 2008: *Konzepte des Hochdeutschen. Der Sprachnormierungsdiskurs im 18. Jahrhundert*, Berlin / New York.

FELDER, E. *et al.* (Hrsg.), 2012: *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*, Berlin / Boston.

FREY, G., 1940: *Hygienische Erziehung im Volksgesundheitsdienst*, 5. erweiterte Aufl. von ‚Hygienische Volksbelehrung, ihre Wege und Hilfsmittel', Handbücherei für den öffentlichen Gesundheitsdienst 12 A, Berlin.

FRITZ, G., 1982: Thema und thematischer Zusammenhang, in: FRITZ, G., *Kohärenz. Grundfragen der linguistischen Kommunikationsanalyse*, Tübingen, 205-223 (Kap. 7).

FRITZ, G., 1994: Grundlagen der Dialogorganisation, in: FRITZ, G. & F. HUNDSNURSCHER (Hrsg.), *Handbuch der Dialoganalyse*, Tübingen, 177-201.

FRITZ, G., 2013: *Dynamische Texttheorie*, Gießen, Gießener Elektronische Bibliothek.
<http://geb.uni-giessen.de/geb/volltexte/2013/9243/>

GLONING, TH., 2003: *Organisation und Entwicklung historischer Wortschätze. Lexikologische Konzeption und exemplarische Untersuchungen zum deutschen Wortschatz um 1600*, Tübingen.

GLONING, TH., 2004: Ernst Jüngers Aufzeichnungen und ihr Wortschatz-Profil, in: HAGESTEDT, L. (Hrsg.), *Ernst Jünger. Politik – Mythos – Kunst*, Berlin / New York, 145-165.

GLONING, TH., 2011: Humoraler Wortgebrauch in der Prosavorrede zum deutschen »Macer« (13. Jh.), in: PLATE, R. & M. SCHUBERT (Hrsg.), *Mittelhochdeutsch. Festschrift für Kurt Gärtner zum 75. Geburtstag*, Berlin / Boston, 375-386.

GLONING, TH., 2012: Wortgebrauch älterer Kochbücher und textbezogene Glossare, in: BERGMANN, H. & R. M. UNTERGUGGENBERGER (Hrsg.), *Linguistica culinaria. Festgabe für Heinz-Dieter Pohl zum 70. Geburtstag*, Wien, 205-237.

HEINEKEN, PH., 1898/1993: *Das Fußballspiel. Association und Rugby*, Reprint nach der Originalausgabe Stuttgart 1898 mit zusätzlichen Abbildungen aus zeitgenössischen Werken, Klassiker der Sportliteratur 2, Hannover.

HEYNE, M., 1905-06: *Deutsches Wörterbuch*, 3 Bde., 2. Aufl., Leipzig.

JUNG, M., 1994a: Zählen oder deuten? Das Methodenproblem der Diskursgeschichte am Beispiel der Atomenergiedebatte, in: Busse, D. *et al.* (Hrsg.), *Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik*, Opladen, 60-81.

JUNG, M., 1994b: *Öffentlichkeit und Sprachwandel. Zur Geschichte des Diskurses über die Atomenergie*, Opladen.

JUNG, M., 1995: Umweltstörfälle. Fachsprache und Expertentum in der öffentlichen Diskussion, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 619-678.

JUNG, M. *et al.*, 2000: *Ausländer und Migranten im Spiegel der Presse. Ein diskurshistorisches Wörterbuch zur Einwanderung seit 1945*, Göttingen.

JUNG, M. & M. WENGELER, 1995: »Nation Europa« und »Europa der Nationen«. Sprachliche Kontroversen in der Europapolitik, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 93-128.

KÄMPER, H., 2005: *Der Schulddiskurs in der frühen Nachkriegszeit. Ein Beitrag zur Geschichte des sprachlichen Umbruchs nach 1945*, Berlin / New York.

KÄMPER, H., 2006: Diskurs und Diskurslexikographie. Zur Konzeption eines Wörterbuchs des Nachkriegsdiskurses, in: *Deutsche Sprache* 34, 334-353.

KÄMPER, H., 2007: *Opfer – Täter – Nichttäter. Ein Wörterbuch zum Schulddiskurs 1945-1955*, Berlin / New York.

KÄMPER, H., 2012: *Aspekte des Demokratiediskurses der späten 1960er Jahre. Konstellationen – Kontexte – Konzepte*, Berlin / Boston.

KERBS, D. & J. REULECKE (Hrsg.), 1998: *Handbuch der deutschen Reformbewegungen 1880-1933*, Wuppertal.

KLEIN, W., 2004: Vom Wörterbuch zum Digitalen Lexikalischen System, in: *Zeitschrift für Literaturwissenschaft und Linguistik* 136, 10-55.

KLEIN, W. & A. GEYKEN, 2010: Das Digitale Wörterbuch der Deutschen Sprache (DWDS), in: *Lexikographica* 26, 79-93.

KOCH, K., 1875: *Fußball. Regeln des Fußball-Vereins der mittleren Classen des Martino-Catharineums zu Braunschweig*, Braunschweig 1875, Nachdruck des von Konrad Koch handschriftlich annotierten Exemplars aus dem Besitz von Kurt Hoffmeister, Braunschweig o.J.

KOCH, K., 1877: Fußball, das englische Winterspiel, in: *Pädagogisches Archiv. Centralorgan für Erziehung und Unterricht in Gymnasien, Realschulen und höheren Bürgerschulen* 19/3, 161-176.

Mädchensport (1902), in: *Die Woche*, 4. Jg., Nr. 14., 5. April, 614-616. (Verfasserkürzel: M. O.).

NAIL, N., 1983: Die Lokalzeitung als Hilfsmittel der Sprachgeschichtsforschung. Beobachtungen am Beispiel der »Oberhessischen Zeitung« (Marburg/Lahn) in den Jahren 1866-1966, in: *Sprache und Literatur in Wissenschaft und Unterricht* 14, Heft 52, 30-42.

NEUENDORFF, E., 1927: *Die deutschen Leibesübungen. Großes Handbuch für Turnen, Spiel und Sport*, Berlin / Leipzig.

PAUL, H., 2002: *Deutsches Wörterbuch. Bedeutungsgeschichte und Aufbau unseres Wortschatzes*, 10., überarbeitete und erweiterte Aufl. von Helmut Henne *et al.*, Tübingen.

PFEIFER, W. (Hrsg.), 1989: *Etymologisches Wörterbuch des Deutschen*, erarbeitet von einem Autorenkollektiv des Zentralinstituts für Sprachwissenschaft unter der Leitung von W. PFEIFER, 3 Bde., Berlin.

PRAUSNITZ, W. (Hrsg.), 1909: *Atlas und Lehrbuch der Hygiene mit besonderer Berücksichtigung der Städte-Hygiene*, München.

STÖTZEL, G., 1988: Konkurrierender Sprachgebrauch in der deutschen Presse, in: HERINGER, H. J. (Hrsg.), *Holzfeuer im hölzernen Ofen*, 2. Aufl., Tübingen, 277-289.

STÖTZEL, G., 1993: Sprachgeschichte als Problemgeschichte der Gegenwart. Vorstellung eines Konzepts, in: HERINGER, H. J. & G. STÖTZEL (Hrsg.), *Sprachgeschichte und Sprachkritik*, Berlin / New York, 111-128.

STÖTZEL, G. & TH. EITZ (Hrsg.), 2002: *Zeitgeschichtliches Wörterbuch der deutschen Gegenwartssprache*, Darmstadt.

STÖTZEL, G. & M. WENGELER, 1995: *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York.

STRAUSS, G. *et al.*, 1989: *Brisante Wörter von Agitation bis Zeitgeist. Ein Lexikon zum öffentlichen Sprachgebrauch*, Berlin / New York.

SVEISTRUP, H. & A. VON ZAHN-HARNACK (Hrsg.), 1934: *Die Frauenfrage in Deutschland. Strömungen und Gegenströmungen 1790-1930. Sachlich geordnete und erläuterte Quellenkunde*, Burg.

TÖNNESEN, C., 1995: Die Terminologie der Sexual- und Partnerschaftsethik im Wandel, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 593-618.

Trübners Deutsches Wörterbuch, 1939-57, im Auftrag der Arbeitsgemeinschaft für deutsche Wortforschung hrsg. von A. GÖTZE, fortgeführt von E. BRODFÜHRER *et al.*, 8 Bde., Berlin.

WEIGAND, F. L. K. & H. HIRT, 1909-10: *Deutsches Wörterbuch*, 5. Aufl., nach des Verfassers Tode vollständig neu bearbeitet von K. VON BAHDER *et al.*, hrsg. von H. HIRT, 2 Bde., Gießen.
<http://digisam.ub.uni-giessen.de/diglit/weigand-bd1>
<http://digisam.ub.uni-giessen.de/diglit/weigand-bd2>

WEISS, CH., 2005: Die thematische Erschließung von Sprachkorpora. Mannheim: IDS (= OPAL 1/2005).
<http://pub.ids-mannheim.de/laufend/opal/pdf/opal2005-1.pdf>

WENGELER, M., 1995: »Multikulturelle Gesellschaft« oder »Ausländer raus«? Der sprachliche Umgang mit der Einwanderung seit 1945, in: STÖTZEL, G. & M. WENGELER (Hrsg.), *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*, Berlin / New York, 711-749.

WICHTER, S., 1991: *Zur Computerwortschatz-Ausbreitung in die Gemeinsprache. Elemente der vertikalen Sprachgeschichte einer Sache*, Frankfurt a.M. [u.a.].

ANHANG 1: Zeitschriftenartikel »Mädchensport« (*Die Woche*, 1902)

[Spalte **614a**]
[Legende zu Abb. 1:] Rudertraining.

Mädchensport.

Hierzu 3 Momentaufnahmen.
Sobald die ersten milden Lüfte wehen und es sich in der Natur überall zu regen beginnt, treibt es die lebenslustige Jugend hinaus, um im Freien bei Sport und Spiel die Glieder zu stählen. Kaum ist die letzte Feuchtigkeit von den Sonnenstrahlen aufgesogen, so wird alles zur Ausübung der Sportspiele hergerichtet. Bei uns hat sich das Lawntennis so eingebürgert, daß man es in sehr vielen Familien schon lange als eine Forderung des guten Tones erachtet, den heranwachsenden Töchtern Gelegenheit zu geben, sich in dem bereits über die ganze Erd verbreiteten englischen »Wiesenballspiel« zu vervollkommnen. Bekannt sind die Turniere in Homburg, Wiesbaden, an der Riviera u.s.w. In den letzten großen Meisterschaftskämpfen, die im Februar und März in Nizza und Monte Carlo zwischen Teilnehmern der verschiedensten Nationalitäten ausgefochten wurden; zeichnete sich eine bekannte deutsche Gräfin in hervorragender Weise aus.

Seit kurzem sind auch Golf und Hockey in Deutschland eingeführt, doch wird es wohl noch lange dauern, ehe unsere Damen sich dem einen oder andern dieser beiden Spiele mit solcher Leidenschaft hingeben werden, wie es jenseits des Kanals, im klassischen Land des Sports, und in Nordamerika geschieht. Besonders die Amerikanerinnen möchte man in dieser Hinsicht immer wieder der heutigen weiblichen Generation Deutschlands als Muster hinstellen. Jene Frauen und Mädchen haben eben längst erkannt, daß sie sich durch Ausübung der mannigfaltigsten Sportarten und Spiele im Freien widerstandsfähige Gesundheit sichern. Und wo Gesundheit ist, da ist auch Jugendfrische und Schönheit.

Mit welcher Begeisterung die jungen und »jüngeren« Damen jenseits des Weltmeers sich vornehmlich den

Sportspielen widmen, erkennt man an dem Eifer, mit dem sie in jedem Jahr von neuem Raketts, Kolben, Schlagholz und Bälle hervorholen. Auf den wohlgepflegten »Lawns«, die zu den Villen und Palästen der exklusiven Vierhundert gehören, wie auf den aus-
[**614b**] gedehnten Rasenflächen im Weichbild der Stadt herrscht um die Jahreszeit stets lustiges Leben und Treiben. Zu den beliebtesten Spielen der freien Töchter Kolumbias gehört unstreitig das Hockey. Der zu diesem Ballspiel erforderliche Platz muß etwa 90 Meter lang und halb so breit sein. An jedem Ende des Terrains befindet sich ein sogenanntes »Goal«, das von zwei mit einer Querstange und einem Netz verbundenen Markpfählen gebildet wird. Die Spielenden teilen sich in zwei Parteien, von denen jede bemüht ist, einen kleinen hellen Guttaperchaball durch das Mal der gegnerischen Partei unter dem nicht ganz bis zum Erboden reichenden Netz hindurchzutreiben. Sobald dies einer Spielerin gelingt, hat sie für ihre Partei das Spiel gewonnen. Zum Treiben des Balls, der nie mit der Hand berührt werden darf, bedient man sich eines Schlagholzes, das mit einem schlichten Krückstock große Aehnlichkeit hat.

Ein neueres, in Amerika in allen Damenuniversitäten und Pensionaten eingeführtes Sportspiel nennt sich »Handball«. Die Spielregeln weichen wenig von denen des Lawntennis ab, nur daß der Ball statt mit dem Rakett mit den Händen zwischen den aus je zwei, drei oder vier Personen bestehenden Parteien hin- und herüber geschlagen wird. Unser Bild auf Seite 615 zeigt zwei Teams im kritischen Moment einer Partie. Der nächste Augenblick muss über Sieg und Niederlage entscheiden. Von dem moderneren Handball ist das vor einigen Jahren auftauchende interessante Korbballspiel noch nicht ganz in den Hintergrund gedrängt worden. Die transatlantischen Schönen sind einmal sehr für Abwechslung. Und sie thun gut daran, aus ihrer reichen Auswahl von Sportspielen bald dieses, bald jenes auf die Tagesordnung zu setzen. Das »Basketballspiel« ist dem Fußball ähnlich. Von den beiden Parteien sucht jede den ziemlich großen Ball in einen

in beträchtlicher Höhe vom Erdboden an einer Mauer angebrachten Korb hineinzuwerfen. Die Mitglieder

[**615**] [Legende zu Abbildung 2:] Mädchensport: Beim Handballspiel.
[**616**] [Legende zu Abbildung 3:] Mädchensport: Beim Golfspiel.

[**616a**] des Team, denen dies gelingt, trotz energischer Bemühungen der Gegenpartei, den Ball für sich zu erobern und in das korbartige Netz zu werfen, sind natürlich Sieger.

Ein allgemein beliebter Sport ist auch das Bowling, ein Spiel, das Gewandtheit und ein sehr gutes Auge erfordert. Auf ebener Rasenfläche von mindestens dreißig Meter Breite und Länge läßt jeder Beteiligte zwei größere dunkle Holzkugeln, die jedoch an einzelnen Stellen etwas platt sind, derart über die abgesteckte Bahn rollen, daß sie eine kleine weiße Kugel, den »Jack«, berühren oder ihr doch so nahe wie möglich zu liegen kommen. Der Jack muß bei Beginn der Partie ungefähr 21 Meter weit von einer kleinen Fußmatte entfernt sein, auf der jeder Spieler Posto faßt, sobald er seine Kugeln entsenden will. Die Partei, deren Kugeln in die nächste Nähe des Jack gelangen, gewinnt. Die Bewegung, die man sich bei Ausübung des Spiels macht, ist nicht übermäßig, und aus diesem Grunde eignet sich Bowling besonders für das zarte Geschlecht. Beim Golf, dem »alten Königsspiel«, besteht der Zweck der zwei Spielenden darin, einen leichten Guttaperchaball aus dem einen »Loch« in das nächstfolgende zu treiben. Wem dies mit Hilfe seiner Kolben (clubs), deren jeder Spieler ein ganzes Sortiment bei [**616b**] sich hat, mit den wenigsten Schlägen gelingt, der geht als Sieger aus dem Match hervor. Die Löcher, in der Regel 18 an der Zahl, bilden einen Kreis und sind je nach dem 100 bis 400 Meter voneinander entfernt.

Bei uns kaum dem Namen nach bekannt ist Badminton, das neuerdings auch in Homburg eingeführt ist. Wie man behauptet, soll Lawntennis davon hergeleitet sein. Jedenfalls spielten in Madras und Kalkutta lebende Engländer Badminton schon Jahrzehnte, bevor Tennis populär wurde, dessen Grundregeln mit denen des älteren Spiels fast übereinstimmen. Nur ist das Netz bedeutend

höher, der Court dagegen viel kleiner als beim Wiesenballspiel. Es wird mit einem durch Blei beschwerten, federngekrönten Korkball, dem wie ein Babyspielzeug aussehenden »shuttlecock« und einem ganz leichten indischen Rakett gespielt. Die Art des Zählens der Points ist etwas anderes als bei Lawntennis, da nur die »servierende« Seite zählt.

Ueber allen diesen Sportspielen wird die echte Sportsdame aber niemals das Radeln und Rudern vernachlässigen. Eine wahre Lust ist es, zum erstenmal wieder nach monatelanger Winterpause auf dem Stahlroß in die Frühlingsluft hinauszueilen oder im schlanken Boot auf schimmernder Wasserfläche dahinzugleiten. m. o.

[Texterfassung: Clara Gloning; Korrekturen: Clara Gloning, Thomas Gloning]

ANHANG 2: Belege zur Wortgeschichte von *Staubsauger*

Grundlage dieser Beleg-Erhebung war das digitale Angebot von books.google.com. Die Hervorhebungen durch Fettdruck stammen von mir (T. G.).

(1889) *Zeitschrift für Nahrungsmittel-Untersuchung und Hygiene* (= *Oesterreichische Chemiker-Zeitung*), Jg. 3, S. 188

> »An Apparaten waren vorhanden die Ventilatoren von Arioni (Amsterdam) ohne Lichtverlust und Zug; Luft und **Staubsauger** von Wing in verticaler und horizontaler Anwendung, derselbe mit Maschine in Verbindung für solche Fabriken und Anlagen, wo Triebkraft entweder nicht vorhanden ist oder Treibachsen und Riemen nicht gelegt werden können oder wo die Kraftmaschine nicht dauernd in Thätigkeit ist.«

(1894) *Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft*, Bd. 18/1-2, S. 254f.

> »Die Gewinnung des Rohmaterials geschieht heute im Tagebau in den Brüchen: Stollenbruch, [...] Der Betrieb vereinigt nun: 6 Gatter, 65 verschiedene Drehbänke, 5 Zirkelsägen, 4 Hobelwerke, 2 Fraismaschinen, 4 Bohrmaschinen, 3 Schleif- und Polierwerke, 1 Sandgebläse, **Staubsauger** und Ventilatoren. [...] Wasserschmierung und -Rieselung, **Staubsauger** und Ventilatoren sorgen bei dem Betrieb noch für die Verwirklichung der hygienischen Forderungen an einen solchen von heutzutage, und machen so denselben auch nach dieser Hinsicht zu einem modernen«

(1897) WEYL, Th. (Hrsg.), *Handbuch der Hygiene*, Bd. 8, S. 727f.

> »Zu den gefährlichsten Operationen gehört das Verpacken des Bleiweißes in Fässer. Man hat daher durch geeignete **Vorrichtungen die Staubentwickelung zu vermindern** gesucht. Besonders verdienen auch hier die Einrichtungen der Firma Leyendecker & Co. in Cöln hervorgehoben zu werden. Wir geben in nachfolgendem eine Beschreibung verschiedener Vorrichtungen zum Abfangen des Staubes beim Packen (s. Fig. 20). [...] Das Faß wird auf einem Schütteltisch unter die vorher beschriebenen **Staubaufsaugeapparate** gestellt. Der beim Einfüllen entstehende Staub wird durch die Exhaustoren abgesaugt. Ist das Faß gefüllt, so werden zwei Tücher darüber gelegt und mit einem Riemen festgeschnallt. Alsdann wird eine mechanische Schüttelvorrichtung in Bewegung gesetzt um das Bleiweiß zusammenzuschütteln. Die Schüttelvorrichtung besteht aus einem mit dem Schütteltisch verbundenen Hebel, welcher den in Lagern ruhenden Tisch nebst dem Fasse in die Höhe hebt. Durch seine eigene Schwere fällt der Tisch wieder auf einen darunter befindlichen Ambos, durch die Erschütterung sinkt das Bleiweiß in dem Fasse zusammen. Das letztere wird 80mal in der Minute gehoben. Nach einem

Schütteln von einigen Minuten wird die Schüttelvorrichtung zur Ruhe gebracht, der **Staubsauger** niedergelassen, die Tücher, welche über das Faß gebunden waren, losgeschnallt und der Staub abgesaugt. Die Operation des Nachfüllens und Schüttelns wird solange wiederholt, bis das Faß sein gehöriges Gewicht hat. Um die ganze Maschine läuft ein Eisengitter, damit niemand durch den Tisch, während derselbe in Bewegung ist, verletzt werden kann. Bei richtiger Ausführung wird die Staubentwickelung so viel wie möglich vermieden.«

(1903) *Die Gartenlaube. Illustriertes Familienblatt*, S. 613

»Unter den Ausstellungsgegenständen, die der Gewerbehygiene dienen, sind am interessantesten die **Staubsauger** und **Exhaustoren**. **Staub und Gas** sind ja die schlimmsten Feinde der Gesundheit des Arbeiters. Sie rufen eine große Zahl von Berufskrankheiten, besonders die stark verbreitete Lungenschwindsucht hervor. Eine Reihe anatomischer Modelle zeigt die Wirkungen des Kohlen-, Eisenoxyd-, Ultramarinstaubes auf die menschliche Lunge. Man kann wohl sagen, die Reinhaltung der Atmungsluft sei eine Hauptaufgabe der Gewerbehygiene.«

(1904) *Bericht über die I. Versammlung der Tuberkulose-Ärzte*, S. 38

»Die ausserordentliche Zimmerreinigung wollen wir jetzt auf andere Art als bisher vornehmen. Es ist vor einiger Zeit ein **mechanischer Staubabsauger mittels Vakuums** erfunden. Diese Apparate sind noch teuer, aber so ausserordentlich gut, dass die Charité damit umgeht, einen anzuschaffen. Der Apparat verhindert jede Staubaufwirbelung, wird elektrisch angetrieben und kann entweder als stationäre Anlage benutzt oder fahrbar gemacht werden. Die Eisenbahndirektion z.B. hat auf der Station Grunewald einen solchen Apparat aufgestellt, in kurzer Zeit kann sie damit einen D-Zug von beträchtlicher Länge vollständig reinigen. Von dem Apparat gehen Schläuche aus, an deren Enden Mundstücke angebracht sind, die den Ecken, Winkeln und Flächen angepasst werden können. Wie intensiv die Wirkung, ist, kann man daran beobachten, dass man beispielsweise Mehl unter einen Teppich streut und durch den Teppich hindurch aufsaugen lässt. Der Apparat kostet mit zwei Schlauchleitungen bei einer fahrbaren Anlage und elektrischem Antrieb etwa. 7000 Mark. Wir haben in der Charité fast überall Elektrizität und können den Apparat überall in Gang setzen. Wir glauben, dass der Apparat sich trotz der hohen Anschaffungskosten rentieren wird. Wir werden nicht mehr genötigt sein, die Ölanstriche der Wände zu waschen, sondern können sie mit dem **Staubsauger** reinigen, ohne den Ölanstrich zu verletzen. Dadurch wird viel Arbeitskraft und Material gespart. Die Reinigung der Betten kann jetzt geschehen, ohne dass die Bettstellen herausgenommen oder die daneben liegenden Kranken belästigt werden.«

(1905) *Gordian. Zeitschrift für die Kakao-, Schokoladen- und Zuckerwaren-Industrie und für alle verwandten Erwerbszweige*, Bd. 11, S. 61

»Allerlei. Wilde Ausstellungen: Es wurde uns ein Ausschnitt aus dem Berliner Tageblatt vom 1. Juni d.J. eingeschickt, folgenden Inhalts: „Auch eine Ausstellung. Eine ‚Ausstellung gewerblicher Erzeugnisse' wurde gestern in Südende im Beisein der schwarzbefrackten Aussteller und unter gänzlicher Abwesenheit des für eine Ausstellung doch so wichtigen Publikums eröffnet. Die auf einem Bauplatz etablierte ‚Ausstellung' umfasst ca. 70 Fabrikate heterogenster Art; neben Gartenmöbeln zeigt ein Zahnkünstler aus Leipzig seine dauerhaften Zähne, **Staubsauger** wechseln ab mit Grammophonen, Wacholderbranntwein, Seifen, Uhrketten, Leitern, Pianinos, Schornsteinaufsätzen und Zuckerbretzeln. Von einer wirklichen, ernst zu nehmenden Gewerbeschau kann gar keine Rede sein."«

(1906) *Museumskunde. Zeitschrift für Verwaltung und Technik öffentlicher Sammlungen*, Bd. 2, S. 27

»Da man aber bis jetzt solche kleinen **Vakuumreiniger** nicht im Handel bekommen kann und doch, gerade hier bei uns in Darmstadt, eine große Menge sehr zarter Objekte, vor allem ausgestopfte Vögel, zu entstauben sind, habe ich zunächst für unsere Bedürfnisse einen Apparat konstruiert, welcher vielleicht auch anderwärts Verwendung finden könnte und der deshalb hier kurz beschrieben werden soll: [...] Nachtrag – Während der Niederschrift des Vorstehenden erhielt ich Kunde von dem Vorhandensein eines **Staubsaugers** mit Handbetrieb ‚Atom' genannt, ohne aber vor Abschluß derselben einen solchen untersuchen zu können. Erst in den letzten Wochen habe ich einen ‚Atom'-Apparat erhalten, nach den verschiedensten Seiten hin ausprobiert und beeile mich nun das Wichtigste darüber nachzutragen. [...] Der ganze Apparat ist sehr leicht gebaut und daher ohne Mühe überallhin zu transportieren, dabei aber für alle Arbeiten, welche eine nicht gar zu große Saugkraft« erfordern, ausreichend. Die Saugkraft kann durch verschieden schnelles Drehen und durch die Anbringung eines Windkessels [...] sehr variiert und dadurch den verschiedenen Objekten leicht angepaßt werden. Da der **Staubsauger** Atom auch mit Elektromotor zum Antrieb geliefert wird [...], so dürfte er die von mir als Desiderat bezeichneten kleineren Vakuumapparate vollständig ersetzen.«

(1907) *Vierteljahrsschrift für gerichtliche Medizin und öffentliches Sanitätswesen*, S. 168

»Ist durch eine gute Wetterführung dafür gesorgt, die Grubenluft möglichst frei von Schlagwettern zu halten, so muss sie in zweiter Linie vor der Kohlenstaubexplosionsgefahr geschützt werden. Den Staub mechanisch zu beseitigen, ist sehr mühevoll. Durch die Arbeit selbst, z.B. beim Stürzen der Kohlen entstandenen Staub kann man mittels **sogenannter Staubsauger**

entfernen. Als Sauger kommt ein kleiner Ventilator in Anwendung, welcher den Staub durch eine Blechlutte hindurch in einen bis zur Hälfte mit Grubenwasser gefüllten Behälter befördert.«

(1907) *Journal suisse de médecine*, S. 182f.

»Sehr zweckmässig ist der **sogen. Vacuumcleaner**, der mit Saugluft arbeitet. Er besteht aus einer Pumpe mit ein oder zwei Cylindern, die meist elektrisch angetrieben wird, und einem Filter. Die Apparate sind entweder stabil und werden dann im Kellerraume untergebracht, oder fahrbar. In letzterem Fall ist Motor, Pumpe und Filter auf einem Wagen angebracht. Der Elektromotor kann, wo eine solche vorhanden, an die Lichtleitung des betreffenden Hauses angeschlossen werden. [...] Stabile Anlagen sind bereits auch auf verschiedenen Bahnhöfen erstellt worden, wo sie namentlich zum Reinigen der Sitze in den Coupes angewendet werden. [...] Ein Apparat, der nicht selten in Haushaltungen Anwendung findet, ist der **Staubsauger** ‚Atom', bei dem das Vacuum durch zwei Blasebälge bewerkstelligt wird, die mittelst Kurbel und Schwungrad in Bewegung gesetzt werden. Auch diese Apparate vermögen, wenigstens für den Kleinbetrieb, Gutes zu leisten. Bei Versuchen, welche der Vortragende angestellt hatte, wurde auch aus Polstermöbeln und Teppichen, die vorher in gewöhnlicher Weise durch Klopfen und Bürsten gereinigt worden, durch den Apparat noch eine erhebliche Menge Staub entfernt. Es sollte jedoch dieser letztgenannte **Staubsauger** nur mit vorgeschaltetem Wassergefäss verwendet werden, das dem Apparat auf Wunsch beigegeben wird.«

(1907) *Charité-Annalen*, S. 29

»Der vor 3 Jahren angeschaffte und durch elektrischen (Licht-)Strom angetriebene **Vakuum-Staubsauger** hat neben der grossen Sauberkeit, die er hervorbringt, dadurch besondere Vorteile, dass er die alljährliche Räumung der Krankenräume zur Reinigung überflüssig macht. Da der Apparat fast geräuschlos arbeitet und keinerlei Staub oder andere Belästigungen verursacht, kann er die Krankensäle Stück für Stück und Fläche nach Fläche reinigen, ohne dass die im Zimmer liegenden Kranken gestört werden. Dadurch ist die völlige Räumung der Krankensäle, die stets einen erheblichen Einnahmeverlust bewirkte, nur noch alle paar Jahre nötig, wenn der Oelanstrich der Wände erneuert werden muss.«

(1907) *Zeitschrift für Tuberkulose*, Bd. 10, S. 407

»Ich möchte hier noch einer Erwägung Platz geben, die sich mir gelegentlich praktischer Besichtigungen aufgedrängt hat. Selbst ein noch so vorsichtig arbeitendes Personal vermag derartige Reinigungen nicht zu bewerkstelligen ohne Aufwirbelung des trockenen, oft infektiösen Staubes. Meist wird die Arbeit von mehreren verrichtet. Reden ist dabei unvermeidlich, ein Herabgleiten des Nasen- und Mundschutzes häufig die Folge. Ich spreche aus

Erfahrung, wenn ich, im Interesse einer möglichst vollkommenen und sauberen Diensterledigung nicht weniger als der Gesundheit der arbeitenden Leute, die Anregung zu Versuchen geben möchte, jeder größeren Desinfektionsanstalt künftig einen, oder mehrere wirksame **Staubsauger, etwa Vakuumreiniger**, beizugeben.«

(1907) *Dinglers Polytechnisches Journal*, Jg. 88, Bd. 322, Heft 2 (»Einige bemerkenswerte Neuerungen auf der Ausstellung zu Mailand 1906«), S. 19

»Zum Schluß sei einer von A. Borsig, Berlin-Tegel, vorgeführten Entstäubungseinrichtung gedacht, die auf einem völlig neuen Prinzip beruht und anscheinend gute Resultate gibt.« Dazu Abbildung: »Fig. 19: **Staubsauger** von Borsig«.

(1908) *Handbuch des Eisenbahnmaschinenwesens*, Bd. 3 (*Werkstätten*), S. 371

»Über dem Windzylinder ist ein Zwischenkühler mit Wasserumlauf für den Kompressor angeordnet. Ferner ist eine Vorrichtung zur selbsttätigen Regelung des Ganges der Pumpe vorhanden, ein großes und sechs kleine Filter, ein stehend angeordneter schweißeiserner Windkessel von 900 mm lichter Weite und etwa 3350 mm Gesamthöhe, acht Stück Staubsauger für die Polstersitze und ein **Teppichstaubsauger** nebst den erforderlichen Schläuchen. Die Rohrleitung besteht aus mehreren Teilen von verschiedener Lichtweite und einer Gesamtlänge von 1090 m und besitzt dreißig Anschlußstellen zur Entnahme von Druckluft für die Sauger.« [Mit „Staubsauger“ ist hier offenbar der Teil der gesamten Entstaubungsanlage gemeint, mit dem man tatsächlich an den Polstern arbeitet.]

(1908/09) STRINDBERG, A., *Fröhliche Weihnacht* (Gesammelte Werke I/10, 1920), S. 105

»Ellen wird beschuldigt, einen Ring genommen zu haben! – Das hat sie nicht! Ellen nimmt keinen Ring! Ebba hätte es tun können! Ich kenne alle hier im Hause! Alle Herrschaften und alle Mädchen. Ellen weint! Ich werde den Ring suchen, vom Keller bis zum Boden, im Aufzuge, im Badezimmer, im **Staubsauger**: alle Löcher und Winkel kenne ich…«.

(1912) SCHLEKER, K., *Die Frau und der Haushalt*, S. 194

»Von allen durchlässigen Geweben: Jute, Kokos, Haargarn, Holländern und ähnlichem, ist abzuraten, falls man nicht einen **Staubsauger** zum Reinigen hat, der auch den Staub entfernt, welcher sich in unglaublicher Menge sonst durch das harte, lockere Gewebe stiehlt und darunter ablagert, bis ihn der nächste Tritt wieder in tanzender Bewegung durch die Lüfte schwirren läßt. Wer auch nur einmal zugesehen hat, welch eine langwierige und mühsame Arbeit das Fortnehmen und Wiederhinlegen von Treppenläufern ist, wird

wohl finden, daß man sie tunlichst verringern soll, und das Mittel ist ja heute in den Staubsaugern gegeben. Das ist sehr erfreulich; denn so hygienisch gut auch Linoleum ist, ist es doch weder für Fuß noch Auge auf Treppen angenehm [...] Legt man dagegen einen geeigneten Teppich auf die Treppe, so kann man ruhig die etwa freibleibenden Stufenecken bohnern, die naturgemäß niemand betritt. Man hat dann eine ebenso hübsche wie sichere und leicht zu reinigende Treppe, immer den **Staubsauger** als Bestandteil des häuslichen Arsenals vorausgesetzt!«

(1912) *Zeitschrift der Deutschen Gesellschaft für Mechanik und Optik*, S. 234

»Die Kolbenpumpen nach der Art der Guerickeschen Pumpe sind noch heute viel im Gebrauch, wobei sie in ihrer Konstruktionsform den jeweiligen Bedürfnissen angepaßt werden. Erinnert sei beispielsweise an die modernen von Hand betriebenen **Staubsauger**, die in manchem Haushalt zu einem unentbehrlichen Hilfsmittel geworden sind.«

(1913) FISCHER, A., *Grundriss der sozialen Hygiene*, S. 262

»Hier ist es daher angebracht, möglichst oft den Staub mittels eines **Staubsaugers** zu entfernen; neuerdings verwendet man hierfür elektrische Apparate. Erwähnt sei hierbei, daß früher die Setzerlehrlinge die Aufgabe hatten, mit einem Blasebalg die Kästen von Staub zu befreien, wobei die jungen Leute naturgemäß den schlimmsten Gefahren ausgesetzt waren.« – Dazu eine Abbildung: »Fig. 44. Elektrischer **Staubsauger** für Druckereien (nach Serényi)«.

(1914) *First German Reader*, S. 217

»der **Staubsauger** (-) [*literally* dust-sucker] vacuum cleaner« [Hervorhebung im Original; T. G].

(1914) *Weyl's Handbuch der Hygiene*, 2. Aufl., Bd. 4, S. 266

»Über die Reinigung der Möbel, Teppiche, Gardinen usw. durch den **Staubsauger (vacuum cleaner)** vgl. den Abschnitt über das Wohnhaus in diesem Band.«

(1918) *Technik und Industrie. Jahrbuch der Technik*, Jg. 4 (1917/18), S. 141

»Einige dieser amerikanischen **Staubsauger** werden in Mengen bis zu 30000 Stück jährlich in einer einzigen Fabrik hergestellt und – verkauft. Nichts kennzeichnet die Wichtigkeit der maschinellen Entstaubung besser, als diese Zahl, die noch vor einem Jahrzehnt ins Reich der unerfüllbaren Phantastik verwiesen worden wäre.«

(1918) *Technik und Industrie. Jahrbuch der Technik*, Jg. 4 (1917/18), S. 141

»Selbst kleinere Haushaltungen dürfen heutzutage an die Anschaffung eines elektrisch betriebenen **Entstaubungsapparats** denken. Wir werden von Tag zu Tag praktischer. Wir wissen, daß die bisherige Art und Weise der Zimmerentstaubung unverhältnismäßig viel Arbeit und Mühe erfordert, und daß auch der dauerhafteste Teppich unter den gutgemeinten Schlägen eines Ausklopfers sehr leidet. Aber gerade der Kauf eines solchen kleinen **Staubsaugers** ist heute noch durchaus Vertrauenssache und bedarf immer des fachmännischen Rates.«

(1918) *Technik und Industrie. Jahrbuch der Technik*, Jg. 4 (1917/18), S. 143

»Es ist lehrreich, zu sehen, wie der Wasserverbrauch dieser **Staubsauger** mit der Höhe des verfügbaren Wasserdrucks sich verschiebt.«

(1918) *Technik und Industrie. Jahrbuch der Technik*, Jg. 4 (1917/18), S. 144

»Während bei den ortsfesten Anlagen vor allem die hydraulischen **Entstaubungsvorrichtungen** bei Erfüllung gewisser Vorbedingungen als vollkommen zweckentsprechend zu bezeichnen sind, hätte man bei den beweglichen **Staubsaugern** bis in die neueste Zeit hinein vergeblich nach einem Apparat gesucht, der der Vollkommenheit wenigstens einigermaßen nahekommt. Betriebskosten, Saugkraft, Dauerhaftigkeit und Hygiene spielten sich gegenseitig einen Schabernack.«

(1918) *Technik und Industrie. Jahrbuch der Technik*, Jg. 4 (1917/18), S. 145

»Unsere Lebensführung ist so anspruchsvoll geworden, daß uns das **Staubtuch** bereits als Schreckmittel aus alten Tagen erscheint. In wenigen Jahrzehnten werden die staubspeienden Balkone an der Rückseite der großstädtischen Mietshäuser völlig der Vergangenheit angehören. Ein modernes Krankenhaus, ein Sanatorium, ein großes Gasthaus oder eine Fabrik unserer Tage ist ohne **Entstaubungsanlage** nicht mehr möglich. Wenn man sich nur die eine Tatsache vor Augen hält, daß aus einem mittelgroßen Teppich, der von Dienstboten mit Bürste und Klopfer ‚gründlich' gereinigt worden war, mit einem **Staubsaugapparat** noch gut 1 kg Staub entfernt werden konnte, so versteht man ohne viele Worte, daß die alte Art der Staubbeseitigung unserer hygienischen Zeit nicht mehr genügt. Die Aktenregistratur unserer Verwaltungsbetriebe verlangt ebensosehr nach dem **Staubsauger** (Abb. 6), wie die Arbeitsräume der Webereien und Tabakfabriken. Die Setzkästen der modernen Druckereien haben sich schon lange an den immer wiederkehrenden Saugrüssel, der den so gefährlichen Bleistaub wegnimmt (Abb. 7),

gewöhnt, und auch der Gegenpol des Setzkastens, das früher stets dick verstaubte, spinnwebüberzogene Büchergestell der Bibliotheken, wird heute saugend entstaubt. Die staubsammelnden Vielfachumschalter unserer Fernsprechämter muß sich die neue Reinigungsart ebenso gefallen lassen, wie die staubfreudigen Polstersitze der Eisenbahnabteile (Abb. 8). Auf der Bühne finden wir den **Staubsauger** beim Reinigen der Ausstattungsstücke, und im Haushalt sind seiner Pflichten so viel, daß sie sich gar nicht auszählen lassen (vgl. Abb. 5 und 9 bis 14). Überall faßt die moderne Technik der Staubbeseitigung festen Fuß: wirtschaftlich im Aufwand menschlicher Arbeit stellt sie einen wertvollen Fortschritt in unserer gesundheitlichen Lebensführung dar.«

(1921) HEDIN, A., *Arbeitsfreude*, S. 164

»Amerika ist uns unendlich weit voraus in allem, was praktische Wohnungseinrichtung anlangt. Drei, vier, fünf, auch sechs Zimmer sieht man als Bedarf für Arbeiterfamilien an, je nach der Zahl der Familienmitglieder. Eine Wohnung ohne eignes Bad ist nicht denkbar. Alle möglichen, Arbeit sparenden praktischen Erfindungen stehen zur Verfügung. Elektrische Waschmaschinen und **Staubsauger** sind an Stelle der veralteten Systeme des Handwaschens und Ausklopfens in Gebrauch. Abgesehen davon, daß Zeit und Arbeit gespart werden, bleibt man auf diese Weise auch damit verschont, alle seine Nachbarn Teppiche, Möbel und Betten ausklopfen zu hören: es ist das in einer dicht bevölkerten Gemeinde eine Qual, die ganz besonders an den geplagten Nerven der Menschen zehrt.«

DIE ALTÄGYPTISCHEN SARGTEXTE IN DIACHRONER ÜBERLIEFERUNG

LOUISE GESTERMANN

I. Die Textgrundlage

Ausgangspunkt der folgenden Betrachtungen und Überlegungen ist ein Textcorpus, das mit Hinweis auf die altägyptische Nutzung umrissen werden kann, für dessen Abgrenzung aber auch die ägyptologische Erschließung mit einzubeziehen und verantwortlich ist.

Es handelt sich bei der Textsammlung, die erstmals zum Ende der Ersten Zwischenzeit und zu Beginn des Mittleren Reiches belegt ist (ca. 2000 v.u.Z.), um funeräres Spruchmaterial, das vornehmlich auf Särgen von Privatleuten, aber auch auf anderen Gegenständen ihrer Grabausstattung (Masken und Betten z.B.) zu finden ist oder auf den Wänden ihrer Grabanlagen niedergeschrieben wurde.[1] Die Texte spiegeln Jenseitsvorstellungen wider, geben Auskunft über den Weg, den der Verstorbene in das Jenseits nimmt, wie auch über sein dortiges ewiges Leben.[2] Inhaltlich und formal sind die Sargtexte als Nachfolgecorpus der Pyramidentexte zu verstehen. Mit ihnen wurden seit Unas, der etwa 300 Jahre früher lebte (5. Dynastie, um 2330 v.u.Z.), die Wände der unterirdischen Räume in (königlichen) Pyramiden wie auch die in Pyramiden königlicher Frauen ausgestattet. Durch Umarbeitungen von Sprüchen aus diesem Textcorpus entstand die späterhin von Privatleuten genutzte Sammlung der Sargtexte, die darüber hinaus aber noch eine Vielzahl neuer, bislang nicht belegter Sprüche beinhaltet. Bei einer diachronen Betrachtung der Sargtexte wäre demzufolge (in vielen Fällen) diese frühere Überlieferung mit einzubeziehen, ebenso das Totenbuch als das Corpus, in dem die Sargtexte aufgingen.[3]

---

[1] Die aktuellste Zusammenstellung der Quellen findet sich bei WILLEMS, H., *Chests of Life. A Study of the Typology and Conceptual Development of Middle Kingdom Standard Class Coffins*, MVEOL XXV, Leiden 1988, 19-34, zu den spätzeitlichen Bezeugungen GESTERMANN, L., *Die Überlieferung ausgewählter Texte altägyptischer Totenliteratur („Sargtexte") in spätzeitlichen Grabanlagen*, ÄA 68, Wiesbaden 2005, op.cit., 349-351 zu denen des Neuen Reiches und der Dritten Zwischenzeit.

[2] Zusammenfassend GESTERMANN, L., *Sargtexte*, in: www.wibilex.de, mit weiteren Hinweisen.

[3] Dazu HORNUNG, E., *Das Totenbuch der Ägypter*, Zürich / München 1979.

Das Textcorpus der Sargtexte ist (bis zu einem bestimmten Punkt jedenfalls) auch ägyptologisch definiert und bleibt unter ägyptischen Gesichtspunkten unvollständig. Auf Särgen oder anderen Schriftträgern, die mit Sargtexten versehen sind, finden sich in der Regel auch „andere" Texte, mitunter auch bildliche Elemente (z.B. Gerätefries oder Opferformeln), zudem sind durch neue Quellen inzwischen auch neue Texte bekanntgeworden (s. auch noch im Folgenden). Grundlage dafür, was ägyptologisch unter einem Sargtext zu verstehen ist, bildet die Textausgabe von Adriaan de Buck.[4] Das solchermaßen definierte Corpus der Sargtexte umfasst knapp 1.200 Sprüche (1.185 bzw. 1.186[5]), die zumeist mehrfach belegt sind und in Sequenzen oder Spruchfolgen eingebunden sein können. Zusammengekommen ist eine wenig homogene Textsammlung: Die Sprüche sind von unterschiedlicher Länge, zeigen hinsichtlich ihres formalen Aufbaus keine grundlegende Übereinstimmung, benutzen verschiedene Stilmittel, sind mit den in ihnen zum Ausdruck gebrachten Gedankenwelten vielfältig ebenso wie sie Hinweise auf unterschiedliche Herkunft und Verwendungszweck („Sitz im Leben") beinhalten.[6] Als zusammengehörig sind sie vor allem durch den Anbringungsort gekennzeichnet, der ihnen gemeinsam ist.

Und es handelt sich (nicht zuletzt) um ein Textcorpus, das sich innerhalb kurzer Zeit (etwa einer Generation) geradezu explosionsartig über Ägypten ausbreitete und intensiv genutzt wurde, und zwar von einer Oberschicht, die in den Provinzen des Landes als dezentrale Schaltstellen und lokale Machthaber eingesetzt waren.[7] Mit Wegfall dieser Gruppe erfuhr auch die Nutzung der Sargtexte Niedergang und Verschiebungen (hin zur Residenz). Die Überlieferung bricht allerdings nicht gänzlich ab, auch die nachfolgenden Zeiten,

---

[4] DE BUCK, A., *The Egyptian Coffin Texts* I-VII, OIP XXXIV, XLIX, LXIV, LXVII, LXXIII, LXXXI, LXXXVII, Chicago 1935-1961.

[5] Wie schon angedeutet, existieren mehr als die von DE BUCK, *The Egyptian Coffin Texts*, erschlossenen „Sargtexte". Ein erster neuer Text wurde von WILLEMS, H., *The Coffin of Heqata (Cairo JdE 36418), A Case Study of Egyptian Funerary Culture of the Early Middle Kingdom*, OLA 70, Leuven 1996, 138-139 und 411-412 benannt („CT 1186").

[6] Vgl. etwa die von ASSMANN, J., Der literarische Begriff im Alten Ägypten. Versuch einer Begriffsbestimmung, in: *OLZ* LXIX, 1974, 117-126, angestoßene Diskussion um den Begriff der Gattungen, die in Sammlungen funerärer Texte vertreten sein können und die es nicht erlauben, diese Spruchsammlungen als ein in sich geschlossenes Corpus zu sehen.

[7] Hierzu WILLEMS, H., *Les textes des sarcophages et la démocratie. Élements d'une histoire culturelle du Moyen Empire égyptien. Quatre conférences présentées à l'École Pratique des Hautes Études. Section des Sciences religieuses. Mai 2006*, Paris 2008.

und insbesondere die 18. Dynastie unter Hatschepsut und Thutmosis III. sowie die Spätzeit bzw. die 25.-27. Dynastie, liefern noch Belege für Sargtexte (vgl. Anm. 1).

Es gab und gibt diverse Unternehmungen, die sich zum Ziel gesetzt haben, die Informationen dieses Textcorpus systematisch zu erschließen. Darauf wird noch zurückzukommen sein. Zunächst soll jedoch der Begriff der diachronen Überlieferung näher beleuchtet und dargelegt werden, welche Ansatzpunkte m.E. im Fall der Sargtexte damit in Verbindung zu bringen sind und Berücksichtigung finden sollten, um die Überlieferung der Sargtexte über die Zeiten hinweg oder einzelne Sprüche daraus adäquat bewerten zu können.

II. Überlieferung und Textkritik

Ein erster Gesichtspunkt beschäftigt sich mit der Besonderheit der Sargtexte (s. zuvor), dass für einen Spruch (in der Regel) mehrere Bezeugungen vorliegen.[8] Dies kann sich einerseits z.B. bei lückenhaften, weil zerstörten oder schwer verständlichen Textstellen als vorteilhaft und überaus hilfreich erweisen, weil auf diese Weise ein Zugang zu einem Text überhaupt erst möglich wird. Andererseits konfrontiert diese Situation mit der Notwendigkeit, Ordnung in diese Mehrfachbelegungen zu bringen, um darauf basierend z.B. Hinweise auf sprachliche Veränderungen oder Entwicklungen zu bekommen.

Auf der Grundlage bekannter Kriterien (Fundumstände, Beifunde, Paläographie z.B.) lässt sich die Zeitstellung der Quelle, auf der ein Text niedergeschrieben wurde, zumeist mit mehr oder weniger großer Genauigkeit festlegen.[9] Dies ist natürlich nicht gemeint, versteht man Überlieferung als einen Prozess der Tradierung eines Textes und der Veränderungen, die er möglicherweise durchgemacht hat. Es geht vielmehr darum, die verschiedenen Textbezeugungen

---

[8] SCHENKEL, W., Eine Konkordanz zu den Sargtexten und die Graphien der 1. Person Singular, in: WILLEMS, H., *The World of the Coffin Texts. Proceedings of the Symposium Held on the Occasion of the 100th Birthday of Adriaan de Buck. Leiden, December 17-19, 1992*, Egyptologische Uitgaven IX, Leiden 1996, 115-127 (118) gibt einen Durchschnitt von 3,8 Bezeugungen pro Spruch an.

[9] Gerade unter den Quellen mit Sargtexten befinden sich allerdings einige prominente Beispiele mit weit auseinanderklaffenden Datierungsvorschlägen. D1C und KH1KH etwa gehören dazu, ganz abgesehen davon, dass bei zahlreichen Textträgern keine wirklich genaue zeitliche Einordnung möglich ist. Bei anderen Quellen ist ihre „späte“ Datierung innerhalb des Mittleren Reiches zwar unumstritten, doch schwankt die exaktere Festlegung von der ausgehenden 12. über die 13. bis in die 17. Dynastie, s. L1/2Li oder S8X, WILLEMS, *Chests of Life*, 19-34, mit Angaben zu den genannten Quellen.

eines Spruches hinsichtlich ihrer entwicklungs- oder überlieferungsgeschichtlichen Beziehung(en) einander zuzuordnen und festzulegen, in welcher Abhängigkeit sie (bzw. die ihnen zugrundeliegenden Vorlagen oder Abschriften) zueinander stehen.

Um genau diese Zusammenhänge zu rekonstruieren, besitzen wir das (methodische) Instrumentarium der Textkritik. Zum textkritischen Verfahren selbst sollen im Folgenden keine weiteren Ausführungen gemacht werden, da es Standard sein sollte, dass diese Methode entweder beherrscht wird oder wenigsten in ihren Grundlagen bekannt ist.[10] Für die hiesigen Zwecke geht es zudem eher darum, mit welcher Absicht textkritisches Arbeiten angewandt wird oder angewandt werden kann, was sich durchaus unterschiedlich darstellt. An die Anmerkung zuvor knüpft ein allgemeines Ziel der textkritischen Analyse an, das die Weitergabe eines Textes durch sein Abschreiben oder Kopieren und über Zwischentextträger zu rekonstruieren sucht, visualisiert in einem Stammbaum oder Stemma. Es zeichnet den Weg nach, den ein Spruch im Laufe seiner Überlieferung genommen hat.[11] Ursprüngliche Intention war es indes, mittels der Textkritik den Urtext oder Archetyp eines Textes zu gewinnen. Mit diesem Anliegen wurde sie in der Altphilologie und den Bibelwissenschaften entwickelt und genutzt, und auch innerhalb der Ägyptologie zielt ihre Anwendung häufig darauf, eine gesicherte Textgrundlage zu gewinnen, die in der ursprünglich abgefassten Spruchversion gesehen werden kann.[12] Für die Sargtexte, aber auch für andere Textcorpora, ist auf Grund der Zufälligkeit der Belege besser von der frühesten greifbaren oder belegbaren Textversion zu sprechen und von der Textfassung, die sich nach den bekannten

[10] Für die Anwendung dieser Methode, ihre Möglichkeiten und Grenzen vgl. die Darstellung von BACKES, B., Zur Anwendung der Textkritik in der Ägyptologie. Ziele, Grenzen und Akzeptanz, in: VERBOVSEK, A. *et al.* (Hrsg.), *Methodik und Didaktik in der Ägyptologie. Herausforderungen eines kulturwissenschaftlichen Paradigmenwechsels in den Altertumswissenschaften*, Ägyptologie und Kulturwissenschaft IV, München 2011, 451-479. Nahezu zeitgleich ist eine Arbeit von mir selbst entstanden, deren Druck sich in Vorbereitung befindet, s. GESTERMANN, L., Möglichkeiten und Grenzen textkritischen Arbeitens, in: BICKEL, S. (Hrsg.), *Ancient Egyptian Funerary Literature. Tackling the Complexity of Texts, Basel December 9-11, 2010*.

[11] In der Ägyptologie sind inzwischen eine Reihe von Arbeiten entstanden, die sich dieser Fragestellung widmen, vgl. hierzu und zum Folgenden BACKES, Zur Anwendung der Textkritik, 452-453.

[12] So z.B. ausdrücklich JÜRGENS, P., *Grundlinien einer Überlieferungsgeschichte der altägyptischen Sargtexte. Stemmata und Archetypen der Spruchgruppen 30-32 + 33-37, 75(-83), 162 + 164, 225 + 226 und 343 + 345*, GOF IV/31, Wiesbaden 1995, 4.

Bezeugungen als die früheste herausschälen lässt. Sie wird unterschiedlich weit von einer möglicherweise vorausgegangenen einheitlichen und einmaligen Konzeption des Textes entfernt sein.

So wie sich die früheste Textfassung rekonstruieren lässt, ist auch die Textfassung jeder Vorlage, die das Stemma enthält, mehr oder weniger sicher zu bestimmen, d.h. es ist an jeder beliebigen Stelle der Überlieferung der Textbestand eines Spruches festzuschreiben. Die textkritische Analyse bietet diese Möglichkeit zwar, für die Ägyptologie ist allerdings zu konstatieren, dass die Perspektiven, die sich daraus entwickeln ließen, (bislang) nicht genutzt werden (dazu noch im Folgenden).

Ist solchermaßen die Entwicklung eines Textes nachgezeichnet, sind auf der Grundlage einer textkritischen Aufbereitung eines Textes weitergehend auch Textveränderungen zu greifen und ist die Entwicklung von Sprache oder ausgewählter sprachlicher Erscheinungen zu verfolgen, zudem die Textentwicklung in größerem oder umfassendem Kontext.

Die Anwendung der Textkritik geht mit einer idealen Vorstellung dessen einher, was mit ihr oder durch sie erreicht werden kann – anders ließe sich mit ihr auch nicht arbeiten. Doch gibt es durchaus Einschränkungen oder Grenzen, mit denen sie behaftet ist, sei es nun, dass sie der Textkritik immanent oder im Fall der Sargtexte auf bestimmte Gegebenheiten zurückzuführen sind. So setzt die erfolgreiche Anwendung der Textkritik bestimmte Bedingungen voraus.[13] Eine ausreichende Anzahl von Textbezeugungen gehört z.B. dazu, um auf diese Weise einen möglichst breiten Überblick über die Tradierung eines Textes zu bekommen und möglichst viele der überlieferten Textbezeugungen einzubinden, des weiteren auch eine ausreichende Länge des Textes, so dass Veränderungen überhaupt entstehen können. Es sind inzwischen eine gewisse Anzahl von Sprüchen aus dem Corpus der Sargtexte textkritisch untersucht worden, letztlich aber nur ein geringer Teil, der sich überschlagsweise auf etwa 15 % belaufen dürfte.[14] Für diese Sprüche sind die Erfordernisse zur Anwendung der Textkritik weitestgehend gegeben, unklar bleibt

---

[13] Z.B. KAHL, J., *Steh auf, gib Horus deine Hand. Die Überlieferungsgeschichte von Altenmüllers Pyramidentext-Spruchfolge D*, GOF IV/32, Göttingen 1996, 4, vgl. BACKES, Zur Anwendung der Textkritik, 457-458.

[14] Zu nennen sind die Arbeiten von JÜRGENS, *Grundlinien einer Überlieferungsgeschichte*, GESTERMANN, *Überlieferung ausgewählter Texte*, BACKES, B., *Das altägyptische Zweiwegebuch. Studien zu den Sargtext-Sprüchen 1029-1130*, ÄA 69, Wiesbaden 2005.

aber, ob auch für die übrigen Sprüche eine Anwendung lohnend wäre und zu Ergebnissen führen würde.

Abgesehen davon, dass bislang von den knapp 1.200 Sprüchen, die unter dem Terminus Sargtexte subsumiert werden, ein nur sehr kleiner Teil textkritisch untersucht worden ist, geben die erarbeiteten Stemmata allenfalls eine vage Vorstellung davon, wie sich die Sargtexte als Gesamtcorpus entwickelt haben. Generelle Traditionslinien abzuleiten, ist in einem gewissen Umfang zwar möglich,[15] aber mit nicht zu kalkulierenden Unsicherheiten verbunden. Jeder neue textkritisch ausgewertete Spruch könnte das Gesamtbild verändern.

Eine weitere Überlegung wiegt schwerer, da sie die Wertigkeit der Textzeugen betrifft. Es ist schnell festgestellt, dass eine Textstelle zwei oder (sogar) mehr Textversionen überliefert. Dabei kann es sich um offensichtliche Fehler handeln, durch die ein Text entstellt wird. Auslassungen gehören dazu, fehlerhafte pronominale Bezüge oder die Verwechslung von Schriftzeichen, um nur einige Beispiele zu nennen.[16] Schwieriger zu beurteilen ist indes, ob es sich tatsächlich um einen Fehler handelt, wenn zwei (oder mehr) Versionen einer Textstelle jeweils für sich genommen verständlich bleiben. Auch in einem solchen Fall wird sich mitunter die Richtung der Textveränderung ermitteln lassen (das Alter einer Niederschrift kann für eine solche Beurteilung eine Rolle spielen oder die soziale Stellung des Nutznießers – mit aller Vorsicht). Es drängt sich aber dennoch die Frage auf, ob eventuell auf Grund bestimmter Gegebenheiten oder wegen inner- oder außertextlicher Bezüge bewusste Veränderungen am Text vorgenommen wurden, inhaltliche Verschiebungen einkalkuliert. Darüber kann, muss aber nicht ein Blick über den eigentlichen Text hinaus Aufklärung geben. Einfluss kann z.B. die Tatsache nehmen, dass ein Sarg für eine Frau angefertigt wurde,[17]

---

[15] In diese Richtung weisen die Arbeiten von JÜRGENS, *Grundlinien einer Überlieferungsgeschichte*, und GESTERMANN, *Überlieferung ausgewählter Texte*.

[16] Hierzu und zum Folgenden BACKES, Zur Anwendung der Textkritik, 454-457, ZEIDLER, J., *Pfortenbuchstudien. Teil I: Textkritik und Textgeschichte des Pfortenbuches*, GOF IV/36, Wiesbaden 1999, 11-84.

[17] Dabei handelt es sich allerdings um eine erst unzureichend untersuchte Überlegung, vgl. die Hinweise von MEYER-DIETRICH, E., *Nechet und Nil. Ein ägyptischer Frauensarg des Mittleren Reiches aus religionsökologischer Sicht*, Acta Universitatis Upsaliensis, Historia Religionum 18, Uppsala 2001, 138 und 140. Die Textveränderung von Seth zu Sachmet in CT 353 (op.cit., 141) ist allerdings abweichend zu erklären, s. GESTERMANN, *Überlieferung ausgewählter Texte*, 277-278, s.a. MEYER-DIETRICH, E., *Senebi und Selbst. Personenkonstituenten zur rituellen Wiedergeburt in einem Frauensarg des Mittleren Reiches*, OBO 216, Fribourg / Göttingen 2006, 280ff.

oder es können Veränderungen darauf zurückgehen, dass ein Text in einem speziellen Umfeld mit bestimmten anderen Texten zusammen auftritt, dazu noch im Folgenden. Die Anwendung der Textkritik und das Ergebnis, die überlieferungsgeschichtliche Gewichtung der einzelnen Bezeugungen eines Spruches (z.B. der Sargtexte), ignorieren diese Veränderungen, mit denen ein Text „optimiert" wird, bzw. orientieren sich hinsichtlich einer Bewertung (in der Regel) an der Nähe oder Ferne einer Textniederschrift zum Urtext. Dies soll in einem eigenen Abschnitt noch einmal aufgriffen werden (III).

Ein spezielles Problem stellen in diesem Zusammenhang die Varianten dar, die zu einem Spruch existieren können – Varianten verstanden als zwei (unterschiedliche) Textversionen, die sich nicht auf eine gemeinsame Urfassung zurückführen lassen.[18] Das Phänomen zweier Varianten eines Textes ist eine alte Erscheinung und begegnet bereits in den Pyramidentexten und späterhin eben auch in den Sargtexten. Sie könnte anzeigen, dass der Verschriftlichung der Sprüche eine lebhafte und durchaus variantenreiche mündliche Überlieferung vorausging. Sie könnte zudem belegen, dass eine solche Varianz und auch die Vielschichtigkeit und Vielfältigkeit der Annäherung an einen bestimmten Inhalt durchaus gewollt war.[19] Eine Anwendung der Textkritik kann in solchen Fällen nicht ergebnisführend oder erfolgversprechend sein. Diese besondere Beleglage rüttelt auf den ersten Blick auch an der Prämisse, mit der die Textkritik arbeitet, dass nämlich die verschiedenen Versionen eines Spruches, die im Laufe der Zeit entstanden sind, auf eine einzige, demzufolge bewusst erstellte Textfassung zurückgehen. Die Textkritik negiert bis zu einem gewissen Grad die Existenz solcher Varianten, allerdings ist genau dieser Unterschied zu machen: Die Annahme eines einzigen (Ur-)Textes als Ausgangspunkt der Überlieferung muss nicht auf jeden einzelnen Text zutreffen, die erfolgreiche Anwendung der textkritischen Methode bestätigt diese Annahme aber dann, wenn ein Stemma in sich widerspruchsfrei konstruiert werden kann.

---

[18] Vgl. das Stemma zu Spruch 227 der Sargtexte mit den beiden abweichenden Versionen des Textes auf Pap.Gardiner II, dazu GESTERMANN, *Überlieferung ausgewählter Texte*, 232-234. Die fehlende Einbindung einer Textversion in das Stemma könnte mit der Beleglage begründet werden, doch muss diese Annahme nicht zwingend zutreffen. Allgemein zu diesem Problem der offenen Überlieferung BACKES, Zur Anwendung der Textkritik, 463-467.

[19] S.a. BAUER, TH., *Die Kultur der Ambiguität. Eine andere Geschichte des Islams*, Berlin 2011, insbes. 109-114.

III. Überlieferung und Textgeschichte

Wie bereits dargelegt wurde, schafft die textkritische Auswertung ein Wertesystem, das die Qualität einzelner Handschriften bzw. Textbezeugungen danach definiert, wie nah oder fern diese der jeweils rekonstruierten frühesten Textversion stehen. Dies allerdings stellt durchaus ein Problem dar oder kann ein Problem darstellen, denn es muss ein hoher Grad der Abweichung eines Textzeugen gegenüber der frühesten bekannten Textfassung nicht zwangsläufig als fehlerhaft verstanden werden. Sie kann vielmehr Berechtigung dadurch besitzen, dass er mit einer speziellen Nutzung des überlieferungsgeschichtlich späten Textbeleges einhergeht – Eine andere, an dieser Stelle nicht weiter verfolgte Möglichkeit besteht darin, dass ein Text bewusst verändert wurde, weil der Schreiber die Fehlerhaftigkeit einer Textstelle erkannte und korrigierte, wenn mit einiger Wahrscheinlichkeit auch nicht zurück in die ursprüngliche Fassung. In diesem Fall hätte ein Text keine offene Überlieferung („produktive" Textüberlieferung) erfahren, wie sie zuvor dargelegt wurde, sondern es würde sich um gängige Textarbeit handeln, die sich beim Niederschreiben eines Textes ergibt oder ergeben kann. In der Konsequenz ist von der Wertigkeit, die eine textkritische Analyse suggerieren könnte, etwas, wenn auch nicht pauschal abzurücken und sind die einzelnen belegten Niederschriften (auch) als „Endgestalt" eines Textes zu verstehen, die jede für sich eine eigene, wertzuschätzende Version darstellt.

Einen weiteren Gesichtspunkt gilt es deshalb nicht aus den Augen zu verlieren: Sprüche aus dem Corpus der Sargtexte treten (bis auf einzelne Ausnahmen) nicht isoliert auf, sondern im Verbund mit anderen Sprüchen, mitunter auch im Verbund mit bildlichen Komponenten. Die textkritische Analyse löst diesen Verbund auf und lässt ihn weitgehend beiseite, wenngleich die Zusammenführung eines Spruches mit anderen Texten, wie sie die Quellen belegen, durchaus textkritisch ausgewertet werden kann und sollte („textexterne Daten"). Die Umgebung, die ein Sargtext haben kann, ist zudem keineswegs festgelegt, sondern kann variieren. Und sie besitzt für die Geschichte eines Textes nicht zu unterschätzende Relevanz, da ein Text auf diese Weise einen bestimmten, traditionsgeschichtlich bedeutsamen Kontext erhält. Abgesehen davon, dass ein Spruch der Sargtexte mit verschiedenen anderen Sprüchen zusammen auftreten und um bildliche Elemente erweitert sein kann, lässt sich die Fragestellung auch auf den architektonischen Kontext (im weitesten Sinn) ausweiten und auf den Anbringungsort, der z.B. auf eine bestimmte

rituelle oder kommunikative Praxis hindeuten kann. Diese offene Überlieferung als bewusste Handhabung und als Anpassung an eine bestimmte Umgebung oder an bestimmte Anforderungen findet sich z.B. bei Bezeugungen von Sprüchen aus dem Corpus der Sargtexte im Neuen Reich, bei denen sich komplett neue, ungewöhnliche Text- und Bildarrangements ergeben können. Hingewiesen sei auf das Auftreten von Sargtexten im Grab des Rechmire (TT 100). Dort sind ausgewählte Sprüche der Sargtexte und solche aus dem Corpus der Pyramidentexte neben Abbildungen von Göttern gestellt und begleiten in ihrer Gesamtkomposition ein rituelles Geschehen, das an der Opferstätte des Grabes für den Verstorbenen vollzogen wurde.[20] In der Grabanlage des Monthemhet aus der ersten Hälfte des 7. Jhd.s v.u.Z. sind Sprüche der Sargtexte offensichtlich dazu benutzt worden, die Thematik eines Raumes auszugestalten. Bindeglied zu den benachbarten Texten ist demzufolge nicht die Herkunft aus einem gemeinsamen Textcorpus, dessen Definition davon abgesehen auch Probleme bereitet (ein aus diesem Grund sowieso problematischer Ansatz), sondern der Inhalt der Sprüche.[21] Eine solche Kontextualisierung kann, muss aber nicht, zu Überarbeitungen und Umformulierungen eines Textes führen.

Schlussfolgerung des Gesagten kann nur sein, den einzelnen Textbezeugungen unabhängig von ihrer jeweiligen überlieferungsgeschichtlichen Stellung, die sich aus einer textkritischen Analyse ergibt, eine eigene Bedeutung zuzumessen. Oder anders ausgedrückt: Die Textgeschichte eines Spruches umfasst auch die möglicherweise wechselnde Kontextualisierung eines Textes und hat zu berücksichtigen, dass trotz nachweislich langer Überlieferung ein Text entstehen kann, der für sich eine korrekte Textversion wiedergibt.

---

[20] Hierzu ASSMANN, J., *Altägyptische Totenliturgien Bd. 2: Totenliturgien und Totensprüche in Grabinschriften des Neuen Reiches*, Supplemente zu den Schriften der Heidelberger Akademie der Wissenschaften, Philosophisch-historische Klasse 17, Heidelberg 2005, 59-146, mit einer im Detail, nicht aber im grundlegenden Verständnis abweichenden Einschätzung GESTERMANN, L., Rezension zu Assmann, Totenliturgien 2, in: *OLZ* 104, 2009, 278-289.

[21] Etwas früher datiert die Anbringung von Pyramidentexten in der Kapelle der Amenirdis in Madīnat Hābū, die ca. 740 v.u.Z. als Gottesgemahlin installiert wurde. Die Versionen dieser Textniederschriften unterscheiden sich von jeweils früheren und einer möglichen ursprünglichen Textfassung teilweise erheblich, sie schaffen aber ein eigenes und in sich schlüssiges Szenario, vgl. hierzu GESTERMANN, L., Pyramidentexte und Sargtexte, in: JANOWSKI, B. *et al.* (Hrsg.), *Grab-, Sarg-, Bau- und Votivinschriften*, TUAT NF 6, Gütersloh 2011, 221-235.

IV. Fazit: Die Anwendung

Im Ergebnis stellt sich natürlich die Frage, welche Konsequenzen aus dem bislang Gesagten für den Aufbau und die Nutzungsperspektiven eines elektronischen Textcorpus wie den Sargtexten zu ziehen sind. Dazu ist zunächst allerdings zu beachten, dass im Fall der Sargtexte einige Erschließungsprozesse bereits stattgefunden haben. Dazu gehört die synoptische Ausgabe der Sargtexte von Adriaan de Buck (erschienen 1935-1961),[22] die nach wie vor Grundlage aller weiteren Untersuchungen und Annäherungen unterschiedlicher Art ist. Dazu gehört auch das von Wolfgang Schenkel initiierte und von ihm durchgeführte, computergestützte Projekt („Sargtexteprojekt") das die Sargtexte in lexikalischer, morphologischer und graphematischer Hinsicht erschließt[23] und bereits Grundlage für zahlreiche Artikel von Schenkel zur Morphologie in den Sargtexten war.[24] Des Weiteren sind ein Index zu den Sargtexten von Dirk van der Plas und J. F. Borghouts zu nennen, der ebenfalls computerbasiert entstanden ist,[25] sowie Wörterbuch und Konkordanz zu den Sargtexten, die von Rami van der Molen noch in Handarbeit erstellt wurden.[26] Hinsichtlich einer Erschließung der Sargtexte ist auf die bisherigen Untersuchungen textkritischer Art zu einzelnen Sprüchen oder Spruchgruppen zu verweisen (s. zuvor) wie auch auf die Veröffentlichung einzelner Quellen in Buchform.[27]

Es liegt demzufolge eine Gemengelage vor, die sehr unterschiedlich ist und sehr unterschiedlichen Ansprüchen genügt. Daran lässt sich nun nichts mehr ändern. Abseits von dem Material, das Adriaan de Buck seinerzeit zusammengestellt hat, existiert inzwischen aber eine Gruppe brachliegender Quellen und Texte, die für eine neuerliche Erschließung zur Verfügung ständen. Und mit ihnen stellt sich

---

[22] DE BUCK, *The Egyptian Coffin Texts*.

[23] SCHENKEL, W., *Konkordanz zu altägyptischen Sargtexten auf der Grundlage von A. de Buck, The Egyptian Coffin Texts 1: Lexikographisch-morphologischer Index* (angekündigt).

[24] Abrufbar über www.aigyptos.uni-muenchen.de.

[25] PLAS, D. VAN DER *et al.*, *Coffin Texts Word Index*, PIREI VI, Utrecht / Paris 1998.

[26] VAN DER MOLEN, R., *A Hieroglyphic Dictionary of Egyptian Coffin Texts*, PÄ 15, Leiden / Boston 2000, und VAN DER MOLEN, R., *An Analytical Concordance of the Verb, the Negation and the Syntax in Egyptian Coffin Texts*, 2 Bde., HdO Sect. 1, Vol. 77, Leiden / Boston 2005.

[27] Vgl. etwa WILLEMS, *Coffin of Heqata*. Eine Übersetzung der Sargtexte ist von DORIS TOPMANN für den *Thesaurus Linguae Aegyptiae* angefertigt worden und abgeschlossen, aber noch nicht ins Netz gestellt.

die Frage danach neu, wie ein elektronisches Textcorpus aufgebaut und in welcher Weise es genutzt werden könnte oder sollte. Um eine ungefähre Vorstellung von der Größenordnung zu geben: de Buck hatte 159 einzelne Quellen ausgewertet und insgesamt 1.185 Sprüche isoliert und (als Sargtext) definiert. Die neuen und von de Buck nicht mit einbezogenen Quellen belaufen sich, soweit ich es überblicke, auf knapp 80 Objekte. Näherungsweise sind mir nicht ganz 500 Nachträge zu etwa 260 Sprüchen bekannt.[28] Diese Bezeugungen sind mit bestimmten Unsicherheiten behaftet, die für diese Zwecke aber unberücksichtigt bleiben können (Zerstörungen, einige vermutlich neue Texte, Problem der Variante). Demzufolge könnten sich bei den genannten Zahlen noch Änderungen ergeben. Gleichwohl steht damit ein Textmaterial zur Verfügung, das zum einen für eine ganzheitliche (wissenschaftliche) Betrachtung der Sargtexte und ihre formale wie inhaltlich Erschließung an sich nicht beiseite gelassen werden kann, und für das sich zum anderen eine elektronische, computerbasierte Erschließung lohnen könnte. Diese hat sich damit auseinanderzusetzen, wie sie aussehen müsste, vor allem aber damit, welchem Zweck sie dienen soll. Vier Bereiche sind m.E. zu berücksichtigen und abzugrenzen.

1 In einem ersten Schritt müsste es um die Erarbeitung und Bereitstellung des Materials gehen. Dazu wäre ein Index der Quellen anzufertigen, in dem die Texte einer Quelle benannt sind bzw. allgemeiner das Text- und eventuell Bildprogramm.[29] Des Weiteren wären die Texte in eine Synopse umzusetzen, dies in Anlehnung – soweit es sich um Paralleltexte zu dem Bestand in der Edition von Adriaan de Buck handelt – nach dessen Ausgabe und Textaufteilung.[30] Eine Synopse, wie sie de Buck noch per Hand angefertigt hat, computerbasiert zu konzeptualisieren und zu erstellen, bereitet inzwischen keine oder nur noch im Detail Schwierigkeiten.[31] Sie bewahrt den Zustand eines unfertigen Produktes, da sie beliebig erweitert werden kann oder bei der Anordnung der Textzeugen Umstellungen vorgenommen werden können. Sie entzieht sich damit auch bestimmten

---

[28] 219 Nachträge zu 108 Sprüchen sind inzwischen aufgenommen, ca. 250 Nachträge zu etwa 150 Sprüchen sind noch zu bearbeiten. Hinzu kommen Belege von über 20 Quellen aus der Zeit nach dem Mittleren Reich.

[29] Vgl. LESKO, L. H., *Index of the Spells on Egyptian Middle Kingdom Coffins and Related Documents*, Berkeley 1979.

[30] DE BUCK, *The Egyptian Coffin Texts*.

[31] S. ALLEN, J. P., *The Egyptian Coffin Texts*, Vol. 8: *Middle Kingdom Copies of Pyramid Texts*, OIP 132, Chicago, Ill. 2006.

Fragen, die an sich vor der Erstellung einer Synopse zu beantworten wären. Solchermaßen gesehen ist die Textsynopse Ausgangs- und Endpunkt gleichermaßen.

Dafür, dass überhaupt eine Synopse angefertigt wird, spricht die Bereitstellung bislang nicht erschlossener Sprüche nach dem etablierten Ordnungssystem dieses Textcorpus (CT) und die damit verbundene bessere Handhabung der Sprüche. Des Weiteren bewahrt die synoptische Zusammenstellung den hieroglyphischen (eventuell hieratischen oder kursivhieroglyphischen) Charakter der Texte und damit die Möglichkeit, die verschiedenen Bezeugungen eines Spruches (auch) hinsichtlich ihrer orthographischen Unterschiede, Übereinstimmungen oder Besonderheiten zu untersuchen.

2 Schon diese ersten Ergebnisse, Quellenindex und Synopse, sind mit diverser Textarbeit verbunden, zu der sich weitere gesellt. Dazu sind die Übersetzung der Texte zu zählen, und zwar jedes Textzeugen, selbst wenn es immer wieder Abschnitte geben wird, die übereinstimmend formuliert sind. Diese Übersetzungen sollten über Index und Synopse abrufbar sein. Es wird zudem erforderlich sein, einige Sprüche textkritisch zu analysieren, um so die Struktur des Textes (besser) zu verstehen.

3 Ein weiterer Schritt umfasst die Erschließung des Materials an Hand der Präsentation von Indizes, und zwar solcher lexikalischer, morphologischer und graphematischer Art, jeweils nach Belegstellen aufgeschlüsselt. Sie können auf der Grundlage der Synopse erfolgen, indem schon bei der Eingabe der Texte Lexeme bzw. Morpheme bestimmt und für eine weitere Bearbeitung markiert und kodiert werden. Was eine Erschließung der Texte in syntaktischer Hinsicht angeht, d.h. mit Blick auf den Textzusammenhang, so mag es der Umfang des Materials zulassen, diesen auf Satzebene zu präsentieren.

Grundsätzlich sind diese Indizes als Erweiterung des Referenzcorpus der Sargtexte zu betrachten, mit den daraus resultierenden Erwartungen und Hoffnungen. Gerade mit Blick auf bislang nicht erschlossene Texte und solche, die sich als Varianten zu bereits bekannten Sprüchen einordnen lassen (mit den damit einhergehenden Abweichungen), dürften neue Einblicke zu erwarten sein.

4 Mit der beschriebenen Erschließung des Textmaterials, teilweise sicher parallel zu den beschriebenen Arbeitsschritten, ist eine wesentliche Voraussetzung dafür geschaffen, weitergehende Forschungsansätze zu formulieren. So könnten sprachliche Phänomene im Corpus der Sargtexte verfolgt werden – weiter verfolgt, muss man sagen, denn es liegen bereits Arbeiten vor, an denen sich

entsprechende Fragestellungen orientieren können. Weitergehend könnte auch die diachrone Überlieferung der Sargtexte in ihrem engeren, eigentlichen Sinn als Forschungsgegenstand befördert werden. Es gibt durchaus Hinweise darauf, dass in den Sargtexten Umformulierungen vorgenommen wurden, mit denen ein moderner, zeitgemäßer oder zeitgerechter Sprachgebrauch in das Spruchcorpus eingeführt worden ist,[32] ganz abgesehen davon, dass sich in den Sargtexten eine Entwicklung weg vom Sprachgebrauch des Alten Reiches beobachten lässt.[33] Allerdings ist in Rechnung zu stellen, dass es sich bei den Sargtexten (wie auch bei Pyramidentexten, Totenbuch, Unterweltbüchern etc.) um religiös-funeräre Texte handelt und diese tendenziell konservativ und (sprachlich gesehen) bewahrend formuliert sind.[34] Bisherige Untersuchungen legen demzufolge nahe, dass es Überarbeitungen und Anpassungen an den aktuellen sprachlichen Gebrauch gegeben hat, entsprechende Untersuchungen könnten aber noch gezielter in den Focus wissenschaftlicher Diskussion wandern, vielleicht auch nicht nur mit Blick auf die sprachliche Entwicklung innerhalb der Sargtexte. Erfolgversprechend ist es möglicherweise auch, das Auseinanderdriften zwischen religiösen Texten einerseits und solchen aus anderen Kontexten (literarische Texte, Gebrauchsliteratur wie Briefe z.B.) zu beobachten bzw. eine solche Entwicklung zu postulieren und einen entsprechenden Forschungsansatz zu formulieren.[35]

Eine weitere Fragestellung, die auf den dargelegten Arbeitsschritten aufbauen kann, zielt auf die ganzheitliche Textbetrachtung – erarbeitet an ausgewählten Sprüchen der Sargtexte (und unter Einbeziehung des Materials von Adriaan de Buck). Entsprechende Beleglage vorausgesetzt, wären für eine solche Fragestellung vornehmlich

---

[32] Ein bekanntes Beispiel betrifft die zunächst fehlerhafte Umsetzung einer negierten Aussage *n msj.y=i is ms.yt* „Daß ich geboren wurde, ist nicht durch Gebären“ in *n msj.t(w)=i ms.yt* „Ich wurde nicht geboren, auch nicht durch Gebären“, bevor in die klassisch-ägyptische („mittelägyptische“) Entsprechung der ursprünglichen Version korrigiert wurde, nämlich *n msj.n.t(w)=i is ms.yt*, s. JÜRGENS, *Grundlinien einer Überlieferungsgeschichte*, 130-131.

[33] S. die Vorgaben von ALLEN, J. P., *The Inflection of the Verb in the Pyramid Texts*, BAe 2, Malibu 1984, dazu die Rezension von SCHENKEL, W., in: *BiOr* 42, 1985, 481-491, EDEL, E., *Altägyptische Grammatik* 1/2, AnOr 34/39, Roma 1955/1964, sowie die Arbeiten von W. SCHENKEL zur Morphologie in den Sargtexten.

[34] ASSMANN, J., *Re und Amun. Die Krise des polytheistischen Weltbilds im Ägypten der 18.-20. Dynastie*, OBO 51, Fribourg / Göttingen 1983, S. 8.

[35] S.a. ALLEN, J. P., Old and New in Middle Kingdom, in: SILVERMAN, D. P. *et al.*, *Archaism and Innovation: Studies in the Culture of Middle Kingdom Egypt*, New Haven / Philadelphia 2009, 263-275.

aber „neue“ Texte (im weitesten Sinn) interessant. Für diese Texte wäre die Überlieferung auf der Grundlage einer textkritischen Analyse zu rekonstruieren und der Weg der Tradierung nachzuzeichnen.

Es könnte sich die Betrachtung dessen anschließen, was als Textgeschichte bezeichnet werden kann, nämlich die jeweilige Kontextualisierung eines Spruches, womit im Wesentlichen seine Zusammenführung mit weiteren textlichen und bildlichen Elementen auf einem gemeinsamen Dokument gemeint ist. Ziel dieses Vorgehens (einer zeitlichen und räumlichen Zuordnung) ist es, sofern dies möglich ist, zu einem besseren Textverständnis zu gelangen, Textspezifisches herauszufiltern, auf Grund von Textveränderungen sich wandelnde Intentionen des Textes nachzuvollziehen und (eventuell) sich wandelnde geistige Welten. Es kann des Weiteren auch darum gehen, einen Text in einem außertextlichen Rahmen zu verankern, z.B. hinsichtlich einer speziellen rituellen oder kommunikativen Einbindung. Wesentliche Gesichtspunkte für solchermaßen ausgerichtete Fragestellungen bieten sich in großer Zahl an (u.a. Literar-, Form- oder Redaktionskritik oder die Analyse des Textinhaltes bzw. weitere Ansatzpunkte der New Philology). In diesem Zusammenhang kann auch eine Variantenkritik angesiedelt sein, d.h. die Diskussion solcher Texte, die zwar deutliche formale und inhaltliche Affinität zueinander zeigen und sicher einen gemeinsamen Ursprung haben, die aber nicht auf eine ihnen gemeinsame (frühe) Textversion zurückgeführt werden können (s. zuvor).

V. Schlussbemerkung

Wie dargelegt wurde, ist die Bedeutung eines Textes, auch eines Spruches der Sargtexte, m.E. in einer Art Fadenkreuz zu sehen. Jeder Spruch besitzt überlieferungsgeschichtlich Wurzeln, die unterschiedlich weit, mitunter aber sehr weit zurückreichen können. Ein Text ist immer aber auch als Produkt der Zeit zu verstehen und somit als ein Einzelmanuskript zu behandeln, das in jeweils neuem Umfeld auftreten kann. Dies angemessen zu berücksichtigen, ist m.E. die große Herausforderung bei der Erschließung und Bereitstellung eines Textcorpus mit Mehrfachbezeugungen, wie es die Sargtexte darstellen.

ÜBERLEGUNGEN ZU TEXTSORTE UND DISKURSTRADITION BEI DER BESCHREIBUNG VON TEXTCORPORA UND IHR BEZUG ZUR LEXIKOGRAPHISCHEN FORSCHUNG

THOMAS STÄDTLER

Für die diachronische Analyse einer Sprache und ihrer Wörter, also für unsere alltägliche lexikographische Praxis, ist es von zentraler Bedeutung, die linguistisch relevanten Charakteristika von verschiedenen Texten ein und derselben Gattung – und natürlich auch ein und derselben Epoche – miteinander zu vergleichen um erkennen zu können, was die Texte verbindet und was sie unterscheidet. Was im gelungenen Fall das Ergebnis eines solchen Vergleiches sein könnte, ist das systematisierte Wissen darum, was eine Fachsprache oder eine gattungsspezifische Sprache zu dem macht, was sie ist – vorausgesetzt freilich, dem Text eignen eine Sprache und entsprechende Wörter, die typische Merkmale aufweisen. Das Ergebnis wäre des Weiteren eine systematisierte Kenntnis der Beziehungen zwischen den sprachlichen Merkmalen dieser Texte und denen anderer Gattungen, und das Ergebnis wäre schließlich eine Kenntnis der Ähnlichkeiten und Unterschiede, die zwischen den einzelnen Fachsprachen und gattungsspezifischen Sprachen beziehungsweise besonderen sprachlichen Ausprägungen existieren. Diese Kenntnisse könnten uns dann in die Lage versetzen, sowohl die linguistischen Charakteristika innerhalb einer speziellen Sprache als auch zwischen den unterschiedlichen Diasystemen einer Sprache – ich werde darauf zurückkommen – zu beschreiben und solcherart das jeweilige Funktionieren der Sprache verschiedener Texte zu begreifen. Das wäre der Idealfall, aber wir wissen wohl alle aus eigener Erfahrung, dass wir, was die Untersuchung des Wortschatzes eines Textes anbelangt, oft weit von diesem Ideal entfernt sind und die Realität uns einen deutlich bescheideneren Arbeitsspielraum lässt.

Um meinen Ausführungen hier eine gewisse Eindeutigkeit zu geben, erscheint es mir angesagt, zunächst terminologische Klarheit zu schaffen und auf die Begriffe Textsorte oder Textgattung und Diskurstradition einzugehen, die oft gleichsam als Synonyme verwendet werden bzw. in den einschlägigen Diskussionen problemlos ausgetauscht werden. Doch das scheint mir so unproblematisch nicht zu sein.

Textsorte oder Textgattung sind ursprünglich Begriffe aus der Literaturgeschichte, mit deren Hilfe wir unsere literarischen und nichtliterarischen Texte grosso modo beschreiben, qualifizieren und ordnen. Die bestehende Klassifizierung der französischen Literatur nach Textgattungen ist das Ergebnis von Überlegungen, die sich über mehrere Jahrzehnte erstreckten. Nach dem Grundriß von Gröber[1] gab es den *Manuel bibliographique* von Bossuat[2], wir haben den *Grundriß der romanischen Sprachen des Mittelalters*[3], wir haben den *Inventaire systématique des premiers documents des langues romanes*[4], wir haben, für die anglo-normannische Literatur, das Handbuch von Ruth Dean[5], wobei all diesen Werken gemeinsam ist, dass sie in ihren Ordnungen jeweils von Textgattungen ausgehen. Und wir haben nicht zuletzt die Bibliographie des *Dictionnaire étymologique de l'ancien français* (DEAF), die für den Bereich des Altfranzösischen Vollständigkeit anstrebt und deren Notizen in aller Regel auch Angaben zur Gattung enthalten[6]. Die Ergebnisse, die so im Laufe vieler Jahre in all diesen Nachschlagewerken zusammengetragen wurden, sind beachtlich, aber es sei die Frage erlaubt, wozu sie eigentlich nützlich sind. Ist die Kenntnis der Gattung eines Textes mehr als hilfreich bei literarischen Studien, bei Untersuchungen zu syntaktischen oder stilistischen Fragen, so ist doch ihr Wert für die lexikologische Arbeit alles andere als unumstritten.

Ich möchte Ihnen das am Beispiel zweier altfranzösischer Texte verdeutlichen, zum einen am *Pèlerinage de Charlemagne*, der Pilgerreise Karl des Großen (PelCharlK[7]), und zum anderen an der *Chanson*

---

[1] GRÖBER, G. (Hrsg.), *Grundriss der romanischen Philologie*, 4 Bde., Straßburg 1888-1902.

[2] BOSSUAT, R., *Manuel bibliographique de la littérature française du moyen âge*, Bibliothèque elzévirienne, Nouvelle série, Etudes et documents, Melun 1951; Supplément (1949-1953) avec le concours de J. MONFRIN, Paris 1955; Second supplément (1954-1960), Paris 1961 (réimpr. 1986 en 1 vol.); Troisième Supplément (1960-1980), t. 1, *Les origines, les légendes épiques, le roman courtois*, t. 2, *L'ancien français (ch. IV à IX), le moyen français*, p. p. F. VIELLIARD & J. MONFRIN, Paris 1986-1991.

[3] KÖHLER, E., (Hrsg.) (zusammen mit H. R. JAUSS & H. U. GUMBRECHT), *Grundriß der romanischen Literaturen des Mittelalters (GRLMA)*, 11 Bde., Heidelberg 1972-1993.

[4] FRANCK, B. *et. al.* (Hrsg.), *Inventaire systématique des premiers documents des langues romanes*, 5 Bde., Tübingen 1997.

[5] DEAN, R. J., *Anglo-Norman Literature. A Guide to Texts and Manuscripts*, London 1999.

[6] Einzusehen im Internet unter www.deaf-page.de.

[7] Die hier verwendeten Sigel sind die des DEAF, s. unter der in Anm. 6 genannten Adresse.

d'Audigier (AudigierJ). Das erste Werk erzählt die Reise Karls und seiner Pairs nach Jerusalem und Konstantinopel, das zweite erzählt die vorehelichen Erlebnisse des jungen Protagonisten Audigier bis zu seiner Hochzeit. In der romanistischen Forschung gilt es seit geraumer Zeit als gesicherte Erkenntnis, dass es sich bei beiden Werken um Parodien der chanson de geste handelt, also um Parodien des altfranzösischen Heldenepos[8]. Verglichen mit wahrhaftigen Heldenepen sind die beiden Texte mit 870 beziehungsweise 517 Versen relativ kurz und man könnte ihnen den Status einer kleinen Sondergattung zuschreiben, nämlich den der Untergattung der epischen Parodie oder besser der Parodie auf das Epos. Man möchte nun vordergründig annehmen, dass zwei Texte, die sich auf gemeinsame literarische Modelle beziehen, eine ganz ähnliche Sprache oder zumindest ein ähnliches Vokabular verwenden.

Schauen wir uns zur Überprüfung dieser Annahme die Glossare in den Ausgaben der beiden Texte an – und man tut das ja in der Hoffnung, dass die Glossare über die interessanten und wichtigen Wörter eines Textes Auskunft geben –, schauen wir also in die Glossare der Ausgabe Koschwitz für den *Pèlerinage* und der Ausgabe Jodogne für den *Audigier*. Unter den hunderten von Einträgen – und besonders das Glossar von Koschwitz ist ziemlich vollständig – finden sich lediglich vier Wörter, die für beide Texte aufgenommen sind[9]. Das Ergebnis dieser bewusst rein mechanischen Abfrage ist also äußerst mägerlich, und man kann daraus den Schluss ziehen, dass Texte, die derselben Gattung angehören, nicht zwangsläufig ein Vokabular enthalten müssen, das auf den ersten Blick sehr ähnlich erscheint, zumindest nach dem zu schließen, was den Glossaren zu entnehmen ist. Im vorliegenden Fall ist dieser Sachverhalt vergleichsweise einfach zu erklären. Der Text der Karlsreise behält durchgängig den Stil der *chanson de geste* bei, selbst in den Passagen, die die *gaberies* erzählen, also die großmäuligen und aufschneiderischen Anekdoten der Pairs, die diese über angeblich von ihnen begangene Heldentaten verbreiten. Das alles wird erzählt, als handle es sich in der Tat um ein echtes Epos, und die Themen, die angesprochen werden,

---

[8] Vgl. etwa GRIGSBY, J. L., Le voyage de Charlemagne, Pélerinage ou parodie? in: *Senefiance* 20, 1987, 567-584.

[9] Es handelt sich um das Substantiv *chevel* (P 181; A 296), *envers*, Adjektiv im Pèlerinage (789) und Adverb im Audigier (413), das Verbum *paistre*, im Pèlerinage auf Ochsen bezogen (318), auf Menschen im Audigier (474), das Verbum *veer* im Sinne von „untersagen" (P 845; A 311) (auch wenn das Glossar zum Audigier fälschlich „éviter" definiert).

entsprechen durchweg solchen, wie sie auch im Epos vorkommen. Darin liegt auch der Grund, weswegen manch einer in diesem alten Text keine Parodie sehen wollte. Aber es sind die Inhalte und Verläufe der Handlungen, die so gar nicht zu dem passen, was man sich von einem Heldenepos verspricht, und die dadurch aus dem Text eine Parodie machen. In der *Chanson d'Audigier* ist es ebenfalls und ausschließlich der formale Rahmen, der an eine *chanson de geste* erinnert. Die Handlungen und etliche Wörter hingegen sind von einer skatologischen Komik, die in der altfranzösischen Literatur ihresgleichen sucht. Eine lexikologische Untersuchung dieses Textes würde problemlos seine Zuordnung zur Gattung der *fabliaux*, also der Schwankerzählungen, erlauben. Ich schlussfolgere daraus, dass die Kenntnis der Gattung eines Textes nicht zwangsläufig Rückschlüsse auf das Vokabular zulässt, welches in dem Text vorkommt, und umgekehrt dieses Vokabular nicht unbedingt Hinweise auf die Textsorte geben muss.

An dieser Stelle nun kommt die Diskurstradition mit ins Spiel, ein Konzept, welches aus der Textlinguistik hervorgegangen ist und im Vergleich zur Gattungsforschung eine noch junge Spielwiese für Linguisten, Literaturwissenschaftler und Philologen darstellt. Man kennt die verschiedenen diasystematischen Varietäten, mit deren Hilfe eine Diskurstradition oder Diskursform beschrieben wird und die ihr ihren Platz auf der Nähe-Distanz-Achse zuweisen, auf der Achse, die auch das Verhältnis zu Mündlichkeit oder Schriftlichkeit ausweist. Die Literatur einer zurückliegenden Epoche ist in aller Regel und fast notwendigerweise sehr weit von der Mündlichkeit entfernt, es sei denn, es handelt sich um eine fiktive Mündlichkeit, etwa in einer direkten Rede, die in eine Erzählung einfließt oder in einem Theaterstück. Eine Sonderrolle spielen möglicherweise Predigtsammlungen, deren Redestil gegebenenfalls der Mündlichkeit näher kommen kann. Wie dem auch sei, es gibt auch in den älteren und alten Literaturen Texte, die Züge aufweisen, die im Allgemeinen eher als zur Mündlichkeit gehörig betrachtet werden:

- Texte, die in einer bestimmten Skripta, in einem bestimmten Dialekt geschrieben sind und in denen die diatopische Varietät stark markiert ist,
- Texte, bei denen in der diastratischen Varietät ein niedriges Stilniveau zu beobachten ist, wie etwa in einigen der vorhin erwähnten Schwankerzählungen,
- Texte schließlich mit einer deutlichen Markierung der diaphasischen Varietät, wie zum Beispiel noch einmal die Schwank-

erzählungen oder verschiedene Dichtungsformen, wie etwa die Schäferdichtung (Rondeau oder Pastourelle), die ein bestimmtes Vokabular enthalten können, das man nicht so ohne weiteres andernorts finden wird.

Die Einteilung nach Diskurstraditionen erlaubt meiner Ansicht nach eine feinere Untergliederung auch der älteren Literaturen als dies mit der Einteilungen in literarische Gattungen möglich ist. Ich komme noch einmal auf die *Chanson d'Audigier* zurück, und zwar auf die Verse 410-416:

Grinberge a descouvert et *cul* et *con*
Et sor le vis li ert *a estupon*;
Au *cul* li chiet la *merde* a grant foison.
Quant Audigier se siet sor un fumier envers
Et Grinberge sor lui qui lui froie les ners,
Deus foiz li fist *baisier son cul* ainz qu'il fust *ters*
Et Audigier i ert par ses *lievres aers*.

[Grinberge hat ihren Arsch und ihr Geschlecht entblößt, und über seinem (Audigiers) Gesicht hat sie den Hintern entblößt; aus dem Arsch fällt ihr massenhaft die Scheiße. Als Audigier sich rücklings auf einen Misthaufen setzt, und Grinberge über ihn, die ihm die Muskeln zerquetscht, lässt ihn zweimal ihren Arsch küssen, bis er abgewischt ist, und Audigier klebt mit seinen Lippen daran.] Hier finden wir innerhalb weniger Verse die Wörter *cul* „Arsch", *con* „weibliches Geschlecht", *a estupons* „seinen Hintern zeigend"[10], noch einmal *cul*, *merde* „Scheiße", *fumier* „Misthaufen", *baisier son cul* „seinen Arsch küssen", *lievres aers* „mit klebenden Lippen", und schließlich den *cul ters* „den abgewischten Arsch", und Sie werden mir vielleicht zustimmen, dass wir es hier mit ziemlich deutlich markierten diastratischen und diaphasischen Varietäten zu tun haben, die man einer Diskurstradition mit der Bezeichnung ‚skatologischer Scherz' zuordnen könnte[11]. Mit Hilfe der Analyse des Wortschatzes ist es also möglich, hier eine Art Spezialsprache auszumachen, die einer bestimmten Diskurstradition zuzurechnen ist. Es stellt sich die Frage, ob es gleichfalls möglich ist, einen Text aufgrund seines Vokabulars einer Gattung zuzuordnen. Wir sehen uns dazu die folgende Passage an:

---

[10] Ein Syntagma, welches, wie M. ROQUES schrieb, « semble en tout cas ne pas s'être répandu beaucoup en dehors de la langue familière (et même grossière) » in: *Romania* 41, 1912, 611.

[11] Ein anderes Beispiel hierfür wäre das Fabliau Jouglet.

Car la montoit li *aloëte*
En fuiant la terre peu nette,
Et la li *lousignous* cantoit
Qui cuers a amer enortoit.
La faisoit son ni li *agaiche*
A deux pertruis que fuïr saiche.
Li *cucus* la se despoulloit
Ou tronc et en printamps cantoit.
La voloit li *chinnes* canter,
Quant sa vie devoit finer,
Et li *calandre* y regardoit
Le malade qui respassoit.
Et li *tourtereule* simplete
S'aseoit sur seke branquete,
Et li *fenix* la s'embrasoit
De cui cendres autres venoit;
Et li *quaille* y faisoit un cant
D'os qui se vont entreferant.
Li *chawe* y assanloit deniers
Qu'elle embloit en autrui greniers,
Et li *cos* par nuit y cantoit
Et par jours ses *glines* paissoit.
Li *escouffles* par grant cembel
Cachoit a prendre un pouchinel.
Le *queville,* comme soutiex,
Estoit de sen ni li espiex.
La ravivoit li *pellicans*
Du sanc de son ceur ses enfans.
Li *hurepel* la remplumoient
Pere et mere quant nu estoient,
Et li *espriviers* deportoit
L'oisiel qui escauffé l'avoit.
Et li *aigles* aloit brisier
Sen bek au perron ou gravier,
Et li *cherïos* au baler
Estoit prins et au hault treper;
Et li *aronde* ralumoit
Ses enfans qu'avulles trouvoit.
Li *huraus* a droit ou a tort
Y cantoit des aucuns la mort.
Li *faucons* l'*anete* y abat

Qui a la terre se debat,
Et li *grue* par nuit la veille
Sur la pierre ou son piet traveille.

In diesem Textabschnitt finden wir in 44 Versen (AnticlLudR 129-172) nicht weniger als 25 Vogelnamen. Es sind geläufige Namen dabei wie *aloete* „Lerche“, *coq* „Hahn“, *geline* „Henne“, *cygne* „Schwan“ usw., aber auch eher ausgefallene Namen wie *agaiche* „Elster“, *calandre* „Kalanderlerche“ oder *hurepel* „Wiedehopf“. Der *huraus* des Verses 167 ist *hurans* zu lesen und bezeichnet wahrscheinlich den Steinkauz. Nicht identifizieren konnte ich den *cherïos* des Verses 163, aber offensichtlich handelt es sich um einen Vogel, der gerne tanzt und hüpft. Man möchte auf den ersten Blick annehmen, man habe es hier mit dem ornithologischen Kapitel eines enzyklopädischen Textes zu tun oder mit einem Ausschnitt aus einem Bestiarium – doch weit gefehlt. Der Text ist in einem der eingangs erwähnten bibliographischen Handbücher unter die religiöse Literatur geräumt, genauer gesagt unter die allegorischen Dichtungen[12]. Es handelt sich um die altfranzösische Übersetzung aus dem ersten Drittel des 14. Jahrhunderts des *Ludus super Anticlaudianum* von Adam de la Bassée, einer Bearbeitung und Umformung des *Anticlaudianus* von Alain de Lille. Der *Anticlaudianus* ist ein ziemlich langes allegorisches Gedicht, welches eine Zusammenschau der christlichen Doktrin bietet. Unsere Textstelle findet sich in der Beschreibung eines von Natur getragenen Mantels auf einem allegorischen Gemälde. Alle erwähnten Vögel sind auf diesem Mantel abgebildet. Und da die Natur ja Platz nicht nur für die Vögel hat, begegnen wir etwas weiter im Text einem ganzem zoologischen Garten sowie einer langen Aufzählung von Blumen und wertvollen Steinen, die alle auf dem Bild Platz finden. Wir haben hier mit Sicherheit die Widerspiegelung bestimmter Diskurstraditionen oder aber sogar Passagen, die aus einer anderen Gattung entlehnt und in eine ganz und gar andere, wenn nicht sogar fremde Umgebung eingefügt sind. Für diese Passagen schlage ich den Begriff Gattungsinsel vor, auf französisch *îlots genresques*, eine Bezeichnung, die mir geeignet erscheint um dieses Phänomen zu beschreiben. Diese Gattungsinseln finden sich ein bisschen überall in der Literatur und oft in größeren Kontexten, in denen man sie nicht erwarten würde. Wer etwa hätte geglaubt, dass sich in der Vita des heiligen Edmund, der *Vie seint*

[12] BOSSUAT, Second supplément, n° 7764.

Edmond le rei (EdmK), einem Text mit über 4000 Versen, der zweifelsfrei der hagiographischen Dichtung zuzuordnen ist, eine erstaunliche Anhäufung von Wörtern aus der Schifffahrtsterminologie findet, wie man sie eher in einem einschlägigen Traktat erwarten würde, und die den modernen Herausgeber dazu veranlasste, im Autor des Textes einen erfahrenen Seemann zu erkennen[13]. Aus Gründen, die uns nicht zugänglich sind, die für unsere Überlegungen aber auch keine Rolle spielen, hat der Autor seinem Text den Anstrich wenn nicht einer ganzen Gattung, so doch einer bestimmten Texttradition gegeben. Vielleicht sind es just solche Gattungsinseln, auf denen die Interferenzen zwischen Textgattung und Diskurstradition stattfinden, wobei der Akzent je nach Art der Insel mehr auf der einen oder der anderen liegen kann. Worauf es ankommt, ist, diese Inseln zu erkennen um in der Lage zu sein, das sich auf ihnen befindliche Vokabular richtig interpretieren zu können. Um sie bei der Beschreibung der Werke eines Textcorpus benennen zu können, bleibt freilich keine andere Möglichkeit als eine profunde Kenntnis dieser Werke. Und je umfangreicher ein solches Corpus geplant ist, umso schwieriger bis schier unmöglich wird eine solche Aufgabe. Wenn ich Ihnen sage, dass wir am DEAF von mehreren tausend Texten ausgehen, die zu bearbeiten sind, werden Sie verstehen, was ich meine.

Neben den Gattungsinseln oder, anders ausgedrückt, neben dem Auftauchen von Spuren einer unerwarteten Diskurstradition in einem Text, findet sich natürlich auch noch das Phänomen, dass ein Text einige Zeilen oder Verse eines anderen Textes enthalten kann, ganze Passagen eines anderen Textes oder gar einen anderen Text in seiner Gänze enthalten kann. Hier geraten wir nun in den Bereich der Intertextualität, bei der es im von mir gebrauchten Sinn darum geht, die Beziehungen zwischen konkreten Texten oder Textteilen zu klären und zu systematisieren. In der altfranzösischen Literatur haben wir einen schönen Fall etwa in der Erzählung von Philomena und in der Geschichte von Pyramus und Thisbe, die beide in den umfangreichen *Ovid moralisé* eingebettet sind. Solche Texteinschlüsse sind in aller Regel leichter zu erkennen als dies bei den von mir so genannten Inseln der Fall ist. Ein solcher Einschluss kann als explizites oder implizites Zitat in einem Text auftauchen, wobei es in letzterem Fall dann doch schwierig sein kann, dieses Zitat zu erkennen und in die

[13] « Il faut donc croire qu'il était un homme expérimenté dans les choses de la mer » (EdmK S. cxvi).

Werkbeschreibung aufzunehmen. Grundsätzlich aber ist es machbar. Aber auch wenn man den Eindruck gewinnt, eine Beschreibung sei bereits ganz anständig gemacht, kann sie doch oft noch ergänzt und verbessert werden. Ein Beispiel: in der Beschreibung des Sigels für den Text *Le Menagier de Paris* (MenagB) lesen wir in der Bibliographie des DEAF, dass es sich um ein Buch zum Hauswesen handelt, genauer gesagt um ein Traktat moralischer und haushaltstechnischer Unterweisungen, welche ein Pariser Bürger, ein nicht weiter bekannter Jurist, seiner offensichtlich deutlich jüngeren Ehefrau erteilt. Man erfährt des weiteren, dass es in diesem aus ganz unterschiedlichen Kapiteln zusammengesetzten Buch, in dem sich zum Beispiel auch eine ganze Reihe von Kochrezepten findet, mehrere eingebaute Texte gibt, nämlich die Geschichte von Griseldis von Philippe de Mézières, das Buch von Melibee und Prudence von Renaud de Louhans sowie den Weg der Armut, *leChemin de povreté*, eine Art Nachahmung oder Fortsetzung des Rosenromans. Das ist schon gar nicht so schlecht, aber man könnte etwa noch hinzufügen, dass dieser *Menagier* auch noch die Episode der *Tentamina* aus dem *Roman des sept sages de Rome*, dem Roman der sieben Weisen aus Rom, enthält sowie die drei *exempla* Papire, Raymonde und Lucrèce, die den *Moralitez sur le jeu des echecs*, moralisierenden Betrachtungen über das Schachspiel, entnommen sind.

Ich gehe davon aus, dass die von mir so genannten Gattungsinseln und die intertextuellen Bezüge durchaus keine Eigentümlichkeit der altfranzösischen Literatur sind, sondern dass sie sich, vielleicht mehr oder weniger stark ausgeprägt, in allen Literaturen finden. Und ich bin daher davon überzeugt, dass die Identifizierung und Benennung dieser Gattungsinseln und intertextuellen Bezüge die Beschreibung von Textcorpora – und ich komme hiermit wieder auf meinen Ausgangspunkt zurück – deutlich verfeinern und verbessern kann. Was bleibt, ist die Frage der Umsetzungsmöglichkeiten. Diese sind sicher von Projekt zu Projekt ganz unterschiedlich, je nach Umfänglichkeit des zu erstellenden oder bereits erstellten Corpus. Dass eine verfeinerte Beschreibung in elektronischer Form ganz neue Abfragemöglichkeiten erlaubt und somit ganz neue Perspektiven eröffnet, um etwa Verbindungen zwischen einzelnen Texten erkennen zu können, die man nicht immer alle vor seinem geistigen Auge haben kann, sei hier lediglich als anregender und abschließender Ausblick vermerkt.