Der *Thesaurus Linguae Aegyptiae* – die Verknüpfung eines elektronischen Corpus ägyptischer Texte und dem Wortschatz der ägyptischen Sprache

Dr. Ingelore Hafemann

Akademienvorhaben Altägyptisches Wörterbuch
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22/23
10117 Berlin
hafemann@bbaw.de

Abstract: The *Thesaurus Linguae Aegyptiae* (TLA) is an electronic corpus of texts written in Ancient Egyptian language (from about 2500 B.C. to 300 A.D). The TLA is today with actually about 800.000 text words the largest corpus of Egyptian texts available online. It is a lexicon based corpus - every word of the corpus is linked to a corresponding lemma in the electronic lexicon that comprises meanwhile 26.450 lemma entries with lexicographical description. The original intention of the project was mainly to create a considerably large corpus. The more the corpus grew larger the more it became obvious that it would be necessary to improve the analytical tools both for the corpus and the lexicon.

Das lexikographische Projekt

Die Projektarbeit des Akademienvorhabens Altägyptisches Wörterbuch knüpft an das größte lexikographische Unternehmen in der Ägyptologie an - dem Wörterbuch der Ägyptischen Sprache, das von 1897 bis 1925 vorbereitet und schließlich in fünf Hauptbänden von 1926-31 publiziert wurde. Dieses Wörterbuch gründete sich auf 1.2 Millionen Belegzettel, die in der Regel den Volltext der Textquelle erfassen. Diese Zettelsammlung liefert sämtliche Belege eines Wortes in seinen Kontexten (Abbildung 1). Dem Prinzip nach ist das ein KWIC-Index. Seit 2002 steht diese Zettelsammlung komplett im Internet als indiziertes *Digitalisiertes Zettelarchiv* und wird von der internationalen Ägyptologie weltweit genutzt [Se2000].

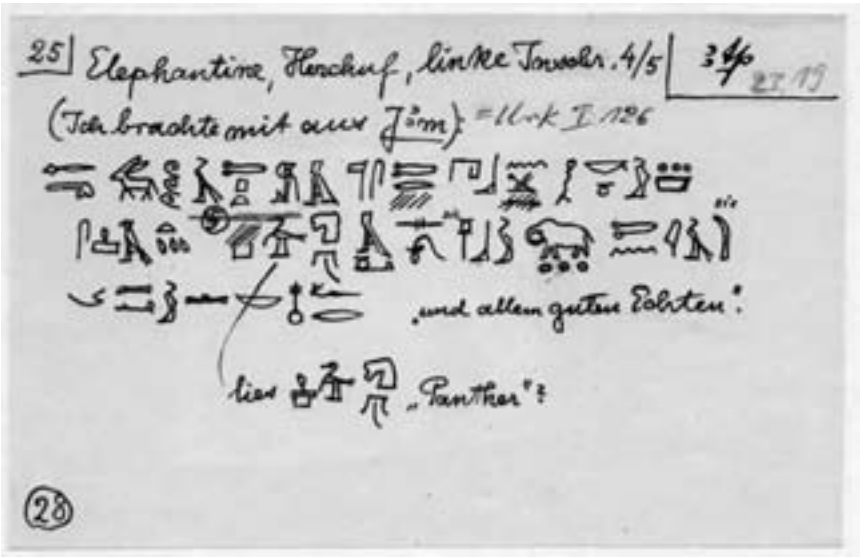

Abbildung 1: Textzettel mit *node word* in seinem Kontext - rot unterstrichen und in der rechten oberen Ecke für die lexikalische Einsortierung noch einmal in Transkription notiert

Jeder Zettel steht hinter einer Lemmakarte und eine lexikalische Feinsortierung gliedert die Zettel nach Gebrauchsweisen, d.h. nach phraseologischen und syntaktisch-semantischen Aspekten. Dieses methodische Grundprinzip - die Darstellung des Wortschatzes ganz und gar aus den Quellen zu erarbeiten - und die Entscheidung, die lexikographische Arbeit auf ein umfassendes und erschlossenes Corpus zu gründen, hat sich als fruchtbar, auch für das neue Projekt erwiesen.

Der *Thesaurus Linguae Aegyptiae*

Das Prinzip corpusbasierter Lexikographie ist für die Ägyptologie also nicht neu. In diesem Ansatz liegt das Spezifikum der Berliner Wörterbucharbeit bis heute. Das neue Projekt, das 1990 mit Vorüberlegungen und 1992 mit der realen Arbeit an der Berlin-Brandenburgischen Akademie der Wissenschaften startete, entschied sich für die Anlage eines Corpus ägyptischer Texte in einer relationalen Datenbank. Die Schwierigkeiten der Organisation großer Materialmengen sind jetzt beherrschbarer. Allerdings erfordert der Einsatz moderner Datenverarbeitung genaues Vorausdenken - Strukturierung und Kontrolle der Dateneingabe anhand von standardisierten Beschreibungsbegriffen waren erforderlich. Grundsätzlich sind zwei Datengruppen zu erfassen - die extralinguistischen Beschreibungsdaten und die Sprachdaten selbst (Abbildung 2):

1. Beschreibungsdaten der ägyptischen Texte und Wörter (Begriffsthesauri)
2.a. Texte als fortlaufende Wortsequenzen mit Satzsegmentierung
2.b. Lexikalischer Thesaurus als Lemmaliste, mit hierarchischer Binnenstruktur.

Eine Bilddatenbank soll Wörterbuch (Bildwörterbuch) und Textcorpus (Faksimiles, Photos der originalen Denkmäler) ergänzen.

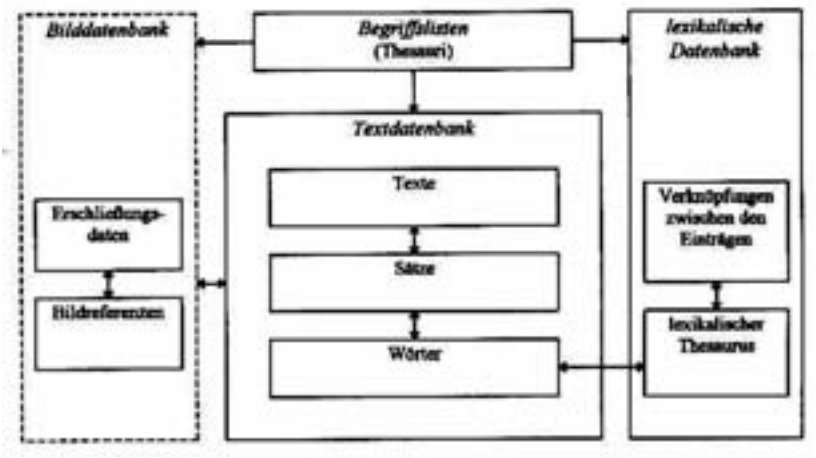

Abbildung 2: Strukturskizze der Datenbank des Altägyptischen Wörterbuches

Die Texterfassung war und ist das Kernstück der Arbeit. Mit Anwachsen der Datenbank wurden zunehmend Filter und Analysefunktionen notwendig, die parallel entwickelt und implementiert wurden. Im Jahr 2004 startete der Internetauftritt mit dem *Thesaurus Linguae Aegyptiae* als der neuen Publikationsplattform des Projektes [Se2004] und ist seitdem unter der Adresse http://aaew.bbaw.de/tla erreichbar.

Die Struktur des *Thesaurus Linguae Aegyptiae*

Die Texte

In der Texterfassung wird der ägyptische Text, der in Hieroglyphen oder einer Kursive, dem Hieratischen, niedergelegt ist, in einer Umschrift notiert. Dabei werden die Wörter des Textes in Folge erfasst und durch *blanks* voneinander getrennt. Die Wortsequenzen werden segmentiert in Sätze oder satzähnliche Strukturen, die semantisch-syntaktische Sinneinheiten widerspiegeln. Jedem Satz / Sinneinheit wird in einem separaten Feld eine Übersetzung in Deutsch, Englisch oder Französisch zugefügt. Außerdem erfolgt eine grammatische Annotation der Wortform (Flexionsform).

In einem halbautomatischen Lemmatisationsvorgang wird jedes Textwort mit einer elektronischen Lemmaliste verknüpft.

Das Lexikon

Das Wörterbuch steht in Gestalt einer elektronischen Lemmaliste zur Verfügung. In ihr ist der gesamte bekannte Wortschatz des Ägyptischen erfasst. Die Lemmaliste bietet je Eintrag die ägyptologische Transkription, die hieroglyphische Standartschreibung, eine oder mehrere Übersetzungen in Deutsch und Englisch, bibliographische Angaben (Wörterbücher, Glossare, Einzeltextzitate) und ist mit einigen lexikographischen *labels* versehen wie einer Wortartangabe und teilweise einer Klassifizierung von Eigennamen als Personennamen, Götternamen, Titel und Epitheta, Ortsnamen und Namen von Institutionen. Diese Wortliste hat auch eine hierarchische Binnenstruktur. Die Hierarchie bildet semantische Lesarten oder sogar teilweise phraseologische Wortverbindungen ab (Abbildung 3)

Abbildung 3: Lemmaliste des Altägyptischen Wörterbuches mit Binnengliederung

Die vielen syntagmatischen und paradigmatischen Wortverbindungen sind nie in einem Wörterbuch abzubilden und jeder Lexikograph muss hier selektieren. Daher ist die Verknüpfung mit dem gesamten Textcorpus von unschätzbarem Wert. So lässt sich der vielfältige Gebrauch von Wörtern, verteilt über Epoche oder Textsorten abfragen.

Die Recherche-*tools* des *Thesaurus Linguae Aegyptiae*

Da mit nunmehr fast 800.000 laufenden Textwörtern das größte elektronische Corpus ägyptischer Texte geschaffen wurde, das weiter anwächst, sind erstmals auch Frequenzdaten für Wortvorkommen verfügbar. Das Corpus ist zur Zeit für statistische Analysen im Ganzen noch nicht geeignet, da es kein ausgewogenes Corpus darstellt. Für viele Texte und Textgruppen aber sind eine Reihe von Abfragen und Analysen möglich, die über die lexikalische und thematische Struktur dieser Texte informieren.

Folgenden Analysen sind aufgrund der relationalen Datenbankstruktur möglich:

- Jedes Wort des Wortschatzes kann in der Lemmaliste nachgeschlagen werden. Die Suche erfolgt über die Transkriptionsmorpheme, auch unter Verwendung regulärer Ausdrücke sowie alternativ über die hieroglyphischen Schreibungen. Die Selektion nach Wortarten oder Spezialwörtern (Eigennamen) ist möglich.
- Für jedes Lemma der Wortliste können die Belege des Corpus aufgelistet werden. Die Ausgabe kann satzweise oder als KWIC-Datei, sortiert nach rechtem oder linkem Kotextwort, ausgegeben werden.
- Für jeden Textbeleg ist ein direkter Zugang zur gesamten Wortfolge des jeweiligen Textes möglich. Jedes Textwort dieses Textes ist erneut durch seine Verknüpfung mit der Lemmaliste analysier- und recherchierbar.
- Es können Wortindizes für alle Texte oder Corpussegmente erstellt werden und der Spezialwortschatz kann aufgelistet werden.
- Es kann nach dem gemeinsamen Auftreten zweier konkreter Wörter gesucht werden. Die Suche nach dem kombinierten Auftreten von Wörtern kann auch auf Wortarten bezogen werden.

Darüber hinaus sind eine Reihe statistischer Analyse-*tools* im TLA implementiert:

- Kollokationsanalysen zu einem Lemma
- Analyse der *type/token*-Ratio und der Zusammensetzung nach Wortarten
- Analyse der Verteilung der Häufigkeiten der Wörter, einschließlich ihrer Verteilung auf Wortarten
- Schlüsselwortanalyse (*key-words*) für einzelne Texte oder Corpussegmente
- Analyse der lexikalischen Gravitation eines Lemmas

Der *Thesaurus Linguae Aegyptiae* als virtuelles Wörterbuch

Der lexikonbasierte Ansatz setzt ein gutes Lexikon voraus. Viele Wörter sind polysem, wie Verben, die - abhängig von beteiligten Präpositionen und Substantiven - den propositionalen Gehalt des Satzes bestimmen und verändern können. Hier unterscheiden Lexika oft mehrere Lesarten eines Lexems. Die Lesartendifferenzierung stellt immer wieder ein Problem dar. Lexika bieten hier oft sehr verschiedene Ansätze für dasselbe Lexem. Bei der Ansammlung von Belegen auf die ägyptischen Lemmata während der Texterfassung nutzen wir in vielen Fällen eine lexikalische Differenzierung, die in Form eines Untereintrages zum Hauptlemma zur Verfügung gestellt wird. Beim Ansetzen solcher Lesarten spielen gewöhnlich Frequenzdaten eine Rolle. Oft können wir uns auf Vorarbeiten der Bearbeiter des alten Wörterbuches (Zettelarchiv mit Zettelzahlen) beziehen. Viele solcher Untereinträge ergeben sich aber auch erst im Laufe der Arbeit am Corpus durch wachsende Belege. Um Bedeutungswandel zu erfassen, sind solche Fragen von großer Bedeutung. Da sowohl das alte Wörterbuch der Ägyptischen Sprache (1926-1931) als auch der *Thesaurus Linguae Aegyptiae* das Ägyptische in allen seinen Sprachstufen abbilden ist auch die elektronische Lemmaliste diachron angelegt. Bedeutungs- und Sprachwandel zu erforschen ist eine zentrale Aufgabe. Wir hoffen, mit einem dynamischen Corpusansatz und der modernen Texterschließung neue Wege für die ägyptologische und die allgemeine Sprachforschung zu öffnen. Das Ägyptische bietet sich hier als die am längsten überlieferte Schriftsprache der Welt in besonderer Weise an.

Literaturverzeichnis

[Se00] Seidlmayer, S. u.a.: Das Zettelarchiv des Wörterbuches der Ägyptischen Sprache. Aufbau, Digitalisierung, Erschliessung und Konsultation im Internet, Thesaurus Linguae Aegyptiae 1, Achet Verlag, Berlin 2000.

[Se04] Seidlmayer , S.: Der Thesaurus Linguae Aeyptiae im Internet. In Göttinger Miszellen 203, Göttingen 2004, S. 99-104.