Ingelore Hafemann

Ein corpusbasiertes Belegwörterbuch des Altägyptischen und seine Nutzungsperspektiven

1. Die Anfänge

Das bedeutendste lexikographische Projekt innerhalb der Ägyptologie ist eng mit der Berliner Akademie verbunden. Es begann mit einem echten Großprojekt - dem „Wörterbuch der Ägyptischen Sprache“, das von 1897 bis 1925 vorbereitet und schließlich in fünf Hauptbänden von 1926-1931 und fünf späteren Belegstellenbänden publiziert wurde[1]. Dieses Wörterbuch gründete sich auf 1.2 Millionen Belegzettel, die in über 30 Jahren von zahlreichen Ägyptologen erarbeitet wurden. In der Regel wurde der Volltext der Textquelle oder lange Passagen derselben erfasst, in dem diese auf mehrere Zettel zu je 30-40 laufenden Textwörtern kopiert wurden. Somit stand im Zettelarchiv nicht nur jeweils ein exzerpierter Belegsatz aus einem Text zur Konsultation zur Verfügung, sondern der Gesamttext, auf mehrere Zettel verteilt. Die Zettelsammlung im Ganzen stellt noch heute die umfangreichste Sammlung mit kompletten Textkopien aus allen Sprachstufen der ägyptischen Sprachgeschichte dar. Der einzelne Zettel bietet den sprachlichen Kontext eines ausgezeichneten Lemmas (rot unterstrichen) mit ca. 30-40 laufenden Textwörtern (Abb. 1).

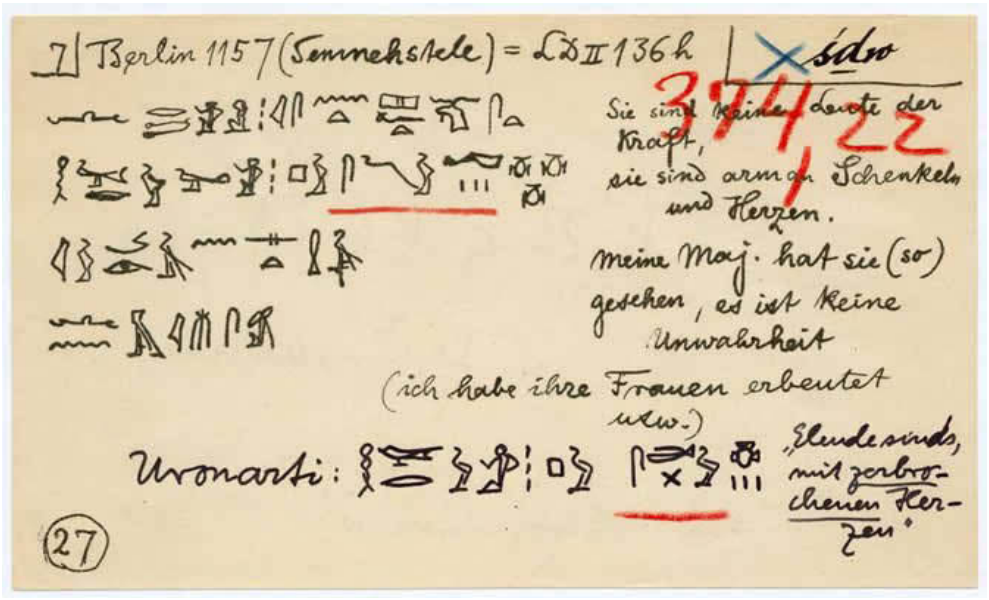

Abb.1

1 Adolf Erman, Herrmann Grapow (Hg.), Wörterbuch der aegyptischen Sprache, 7 Bde., Berlin und Leipzig 1926-1963; Belegstellen, 5 Bde., Berlin und Leipzig 1935-1953.

Dem Prinzip nach entspricht das einem modernen *Key-Word-In-Context* Index (KWIC).

Die Zettel stehen im Archiv hinter entsprechenden Lemmakarten und eine lexikalische Feinsortierung gliedert die Zettel hinter dem Lemmakarten weiter nach Gebrauchsweisen, d.h. nach sachlichen oder nach phraseologischen und syntaktisch-semantischen Aspekten. Man hat also ein lexikalisch-lexikographisch sortiertes Wortarchiv mit direktem Durchgriff auf das Textarchiv mit allen Belegen.

Dieses methodische Grundprinzip, die lexikographische Arbeit auf ein umfassendes und erschlossenes Corpus ägyptischer Texte zu gründen, wurde von Adolf Erman (1854-1937) im Jahre 1897 initiiert. Es hat sich als sehr fruchtbar auch für das neue lexikographische Projekt an der BBAW erwiesen. Das wertvolle und materialreiche Zettelarchiv, bis heute eine Fundgrube für jeden Ägyptologen, wurde daher auch in die neue Projektarbeit einbezogen. Es wurde komplett digitalisiert und durch eine lexikalische Indizierung erschlossen. Seit 2002 steht es vollständig im Internet als indiziertes *Digitalisiertes Zettelarchiv* zur Verfügung und wird seitdem von den ägyptologischen Fachkollegen weltweit genutzt (Abb. 2).

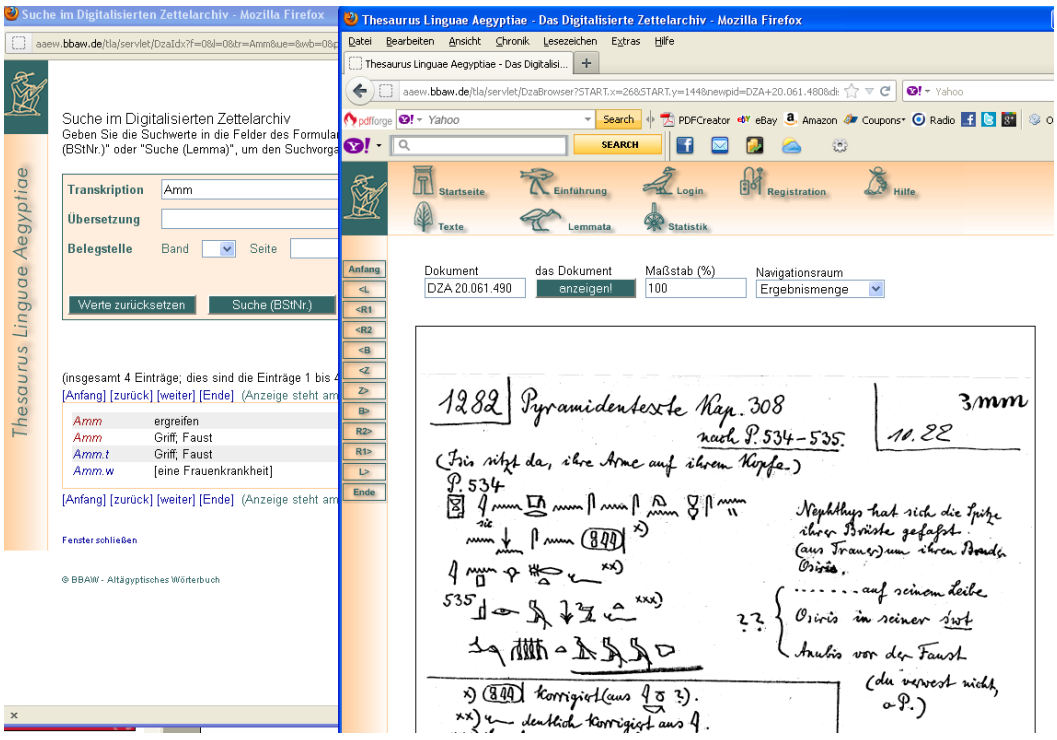

Abb. 2

2. Elektronisches Corpus und Lexikon. Die Strukturen des *Thesaurus Linguae Aegyptiae*

Das Prinzip corpusbasierter Lexikographie ist für die Ägyptologie also nicht neu. Daher war es folgerichtig, ein neues Projekt mit der Anlage eines neuen – nunmehr elektronischen – Corpus ägyptischer Texte zu starten. Das Projekt folgt weiterhin dem Prinzip der Darstellung des Wortschatzes auf der Grundlage der kompletten Quelltexte und unter ständigem direktem Be-

zug auf diese. Das ist für tote Sprachen unabdingbar, bei denen die Texte die ersten und einzigen Quellen bilden. Aber die Schaffung großer Corpora wird auch zunehmend für die Textlinguistik und Lexikographie moderner Sprachen interessant. In den vergangenen dreißig Jahren wurden daher umfangreiche elektronische Textcorpora in vielen verschiedenen Sprachen aufgebaut.

Strukturell haben wir ein lexikonbasiertes Textcorpus oder aber - je nach Perspektive – ein textbasiertes Lexikon in den Focus unserer Arbeiten gestellt. Corpus und Lexikon bilden also die zwei Grundsäulen des neuen elektronischen lexikalischen Systems. Grundsätzlich sind dazu zwei Typen von Daten – und zwar sowohl für das Corpus als auch für das Lexikon – zu erfassen. Das sind

a. die extralinguistischen Beschreibungsdaten - die so genannten Passportdaten zu Texten und Wörtern und

b. die Sprachdaten selbst, die als fortlaufende Textwörter sowie als Lemmata des Wörterbuchs aufgeführt werden.

Die Passportdaten der Texte geben Angaben zur Herkunft, Datierung, Textsorte und Bibliographie eines Textes. Dabei müssen die Angaben für den Textträger nicht identisch mit denen des Textes selbst sein. Wir vermerken daher Textträger/Objekt-Daten und Textdaten getrennt, denn auf einem Objekt können mehrere Texte sein, z.B. auch solche aus verschiedenen Zeiten.

Die Beschreibungsdaten der Lemmata sind die lexikographischen Angaben zu den einzelnen Einträgen eines Wörterbuchs. Unser elektronisches Wörterbuch umfasst den gesamten Wortschatz der ägyptischen Sprache. Es gibt Übersetzungsäquivalente in Deutsch und wahlweise in Englisch an sowie Angaben zur Wortart, Flexionsklasse, zum Spezialwortschatz und bibliographische Referenzen.

Diese so genannten Passportdaten dienen v.a. für Filterfunktionen bei Recherchen im Corpus und im Lexikon. Texte des Corpus können mit ihrer Hilfe nach Herkunft oder Datierung ausgefiltert werden. Im Lexikon der Wörter können z.B. nur solche einer bestimmten Wortart oder ggf. eines Spezialwortschatzes in die Recherche einbezogen werden. Wir führen z. B. Toponyme, Personen-, Götter- und Königsnamen, sowie Titel und Epitheta von Personen und sonstige Eigennamen als extra gekennzeichnete Wörter des Spezialwortschatzes auf.

Der zweite Typ der Datenerfassung sind die Sprachdaten selbst: Das sind die Wörter in ihrer originalen Form und einer ägyptologischen Transkription. Die Wörter werden in zweierlei Form zur Nutzung aufbereitet.

a. Als Wortformen, die sich in Sequenzen zu Texten formieren. Die Text-Wörter werden in Form einer ägyptologischen Transkription Wort für Wort erfasst und zu Satzeinheiten zusammengefasst.

b. Als separierte Einzelwörter in einer „Normalform“, die in ägyptologischer Transkription und hieroglyphischer Schreibung innerhalb einer strukturierten Wortliste (Lexikon) erfasst werden.

3. Texterfassung – Corpusaufbau

In der Texterfassung wird der ägyptische Text, der in Hieroglyphen oder einer Kursive, dem so genannten Hieratischen, niedergelegt ist, in einer ägyptologischen Umschrift notiert. Alle lexikalischen Recherchen laufen praktischerweise über diese Transkription und nicht über die hieroglyphischen Zeichenformen. Die einzelnen Wörter werden tokenisiert, die Wortfolge wird in Sätze oder satzähnliche Strukturen segmentiert. Jedem Satz oder jeder Sinneinheit wird in einem separaten Feld eine gebundene Übersetzung in Deutsch, Englisch oder Französisch zugefügt. Das Ganze sieht wie eine Edition aus und folgt auch partiell deren Prinzipien. So notieren wir z.B. die Textkritik, um die Verlässlichkeit einer Quelle zu zeigen. Außerdem erfolgt eine grammatische Annotation jeder einzelnen Wortform. Die Eingabe der Textwörter selbst erfolgt manuell, denn eine automatische Erkennung der verschiedenen hieroglyphischen Formen oder Transkriptionen in den Quellpublikationen und deren Zuweisung zu einer Normform gelingen noch nicht. In einem halbautomatischen Lemmatisationsvorgang wird jedes Textwort mit einer elektronischen Lemmaliste verknüpft. Damit wird auch gleichzeitig jedem Eintrag der Lemmaliste eine neue Belegstelle zugewiesen.

Was wir gewinnen ist ein vollständig corpusbasiertes Wörterverzeichnis oder ein lexikonbasiertes Corpus - je nachdem, welchen Focus man legt. Im Laufe der Arbeit ist uns erst die wissenschaftliche Dimension eines elektronischen Corpus als solchem immer klarer geworden und die Methoden der Corpuslinguistik gerieten zunehmend in unseren Focus. Diese erlauben es im Corpus zu recherchieren.

Unser Corpus ist relativ klein, denn mit 960.000 Textwörtern bleiben wir bis Ende 2012 knapp unter der magischen Millionengrenze. Aber es ist ein nach einheitlichen Prinzipien strukturiertes und lexikographisch und grammatisch annotiertes Corpus. Als solches ist es ausbaufähig in vielerlei Hinsicht.

4. Die Wortliste – als elektronisches Lexikon

Das elektronische Lexikon ist das lexikalische Rückgrad der Struktur. Es ist eine Liste mit Wörtern, die den bis heute bekannten Wortschatz des Ägyptischen in seinem aktuellen Forschungsstand widerspiegelt. Genau genommen sind es zwei Listen – eine mit Wörtern, die hieroglyphisch-hieratisch niedergeschrieben und eine mit Wörtern, die in demotischer Schrift niedergelegt wurden, einer späten Schriftnorm des Ägyptischen (7. Jhd. v. Chr. bis 5. Jhd. n. Chr.). Insgesamt umfasst die erste Liste ca. 28.000 Lemmata, ohne Personennamen und die demotische Liste ca. 9200 Lemmata, ebenfalls ohne Personennamen. Diese Wortlisten haben auch eine hierarchische Binnenstruktur. Die Hierarchie bildet semantische Lesarten oder die erstgenannte sogar teilweise phraseologische Wortverbindungen ab (Abb. 3).

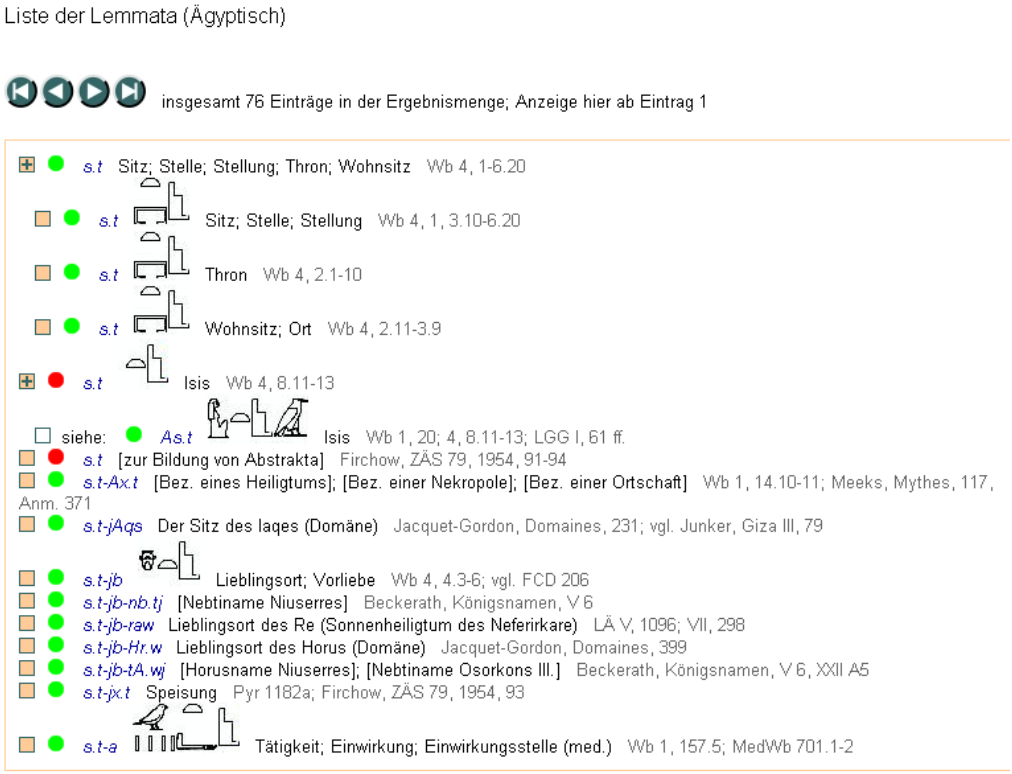

Abb. 3

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

Diese Lemmalisten dienen als Werkzeuge für die oben beschriebene Lemmatisation bei der Texteingabe. Gleichzeitig stehen sie aber auch zum Nachschlagen wie normale Wörterbücher im Internet zur Verfügung.

4. Nutzungsstrategien

Ursprünglich stand die Idee einer bloßen Belegstellenabfrage für die Wörter des ägyptischen Lexikons im Mittelpunkt unserer Projektarbeit. Da das gesamte Textcorpus Wort für Wort lexikalisch erschlossen - sprich lemmatisiert wird, haben wir das Konzept eines virtuellen Wörterbuchs entwickelt. Die Philosophie dahinter ist die folgende: Der durch den Lexikographen erarbeitete feste Wörterbucheintrag, der zu den einzelnen Wörtern bestimmte Angaben quasi abschließend im Lexikon verzeichnet, wird durch ein Werkzeug ergänzt, das es dem Forscher erlaubt, die gesamte Materialbasis zur Aufklärung seiner sehr speziellen Wortrecherche zu nutzen.

Ist nicht jeder Eintrag in gedruckten Wörterbüchern auch irgendwie eine Notlösung? Es las-

sen sich meist etliche weitere Wortbedeutungen oder Lesarten eines Lemmas bzw. semantische Eigenschaften durch kotextuelle Zusammenhänge im tatsächlichen Sprachgebrauch finden. Die Fragen des Forschers lassen sich in einem Corpus frei und immer wieder neu formulieren. Die Werkzeuge, die wir zu solchen Abfragen zur Verfügung stellen, werden im Folgenden noch etwas näher beschrieben. Sie sind z. T. auf ägyptologische Bedürfnisse abgestimmt, stellen aber gleichzeitig Fragestellungen für Analysen aller natürlichen Sprachen dar, wie sie auch die Corpuslinguistik benutzt.

Der lexikonbasierte Ansatz setzt ein gutes Lexikon voraus. Damit steht und fällt die Corpusanalyse. Viele Wörter sind polysem. Hier unterscheiden Lexika oft mehrere Lesarten eines Lexems. Die Lesartendifferenzierung stellt immer wieder ein Problem dar. Wie überall bieten auch in der Ägyptologie verschiedene gedruckte Wörterbücher sehr unterschiedliche Ansätze für dasselbe Lexem. Bei der Ansammlung von Belegen auf die ägyptischen Lemmata während der Texterfassung nutzen wir in vielen Fällen bereits eine lexikalische Differenzierung, die in Form eines Subeintrages zum Hauptlemma zur Verfügung gestellt wird (Abb. 4).

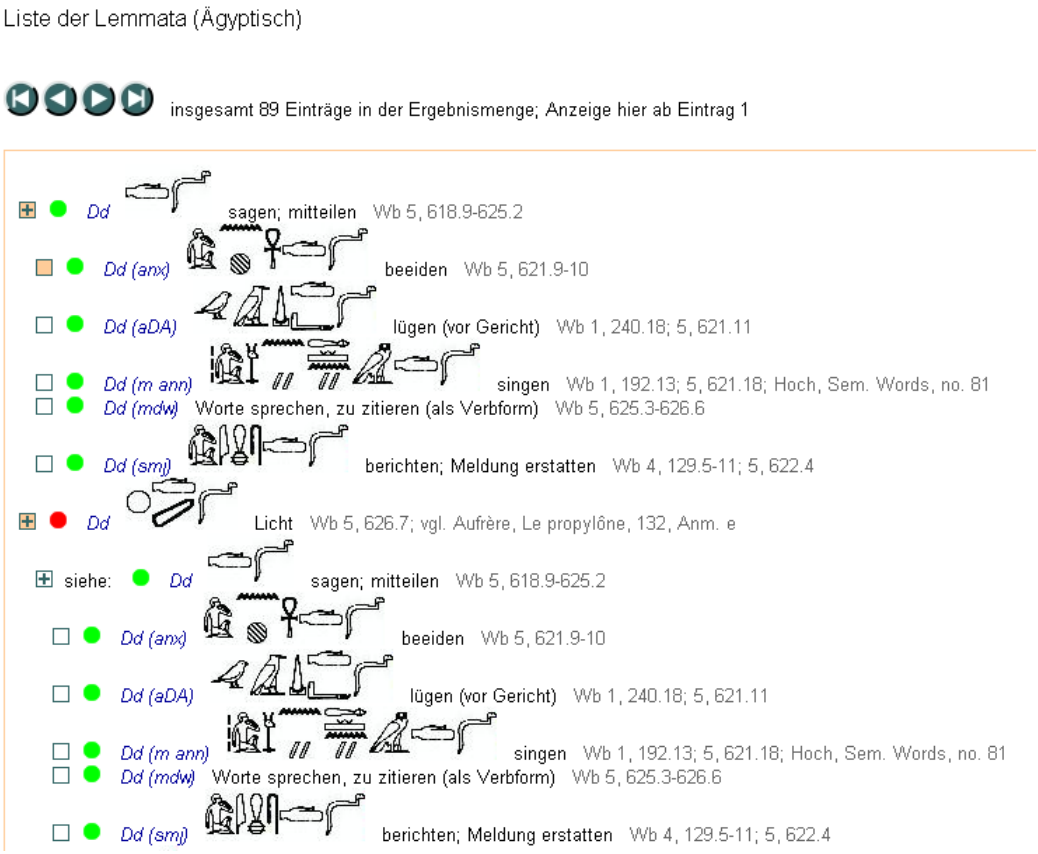

Abb. 4

Beim Ansetzen solcher Lesarten oder Sublemmata spielen gewöhnlich Frequenzdaten eine Rolle. Oft können wir uns dabei auf Vorarbeiten der Bearbeiter des alten Wörterbuches beziehen, indem wir die jeweiligen Informationen darüber aus dem alten Zettelarchiv mit den konkreten Zettelzahlen ziehen. Viele solcher Untereinträge ergeben sich aber auch erst im Laufe der Arbeit am Corpus durch wachsende Belege. Jeder neu eingegebene Text liefert neue Wortbelege.

Solche Lesartendifferenzierungen sind aber seinerzeit längst nicht für alle Lemmata vorgenommen worden und ihre Ansetzung ist nicht trivial. Übersetzungsäquivalente der Zielsprachen sind wenig hilfreich und hier müssen die Probleme der kontrastiven Lexikographie reflektiert werden. Bei Verben können die Argumentstrukturen dabei helfen, verschiedene Lesarten zu bestimmen. Allgemein lassen sich nun Kookurenten von Wörtern in einem elektronischen Corpus besonders gut recherchieren. Das kann gerade bei der Lesartenunterscheidung von Lemmata helfen. Phraseologische und idiomatische Wendungen, der Wortgebrauch in bestimmten Texten oder Textgruppen und das unter Umständen auch nur zu bestimmten Zeitepochen sowie andere, jeweils aktuelle Fragestellungen eines Forschers können mit einem elektronischen Corpus auf völlig neuer Art ermittelt werden.

So kann der Zugang zum Wortschatz nicht mehr nur über rein formal alphabetische Sortierung von Lexika gewonnen werden, sondern zusätzlich über Aspekte der natürlichen Struktur des Wortschatzes.

In den vergangenen fünf Jahren konnten wir auch einige Analysewerkzeuge aus der Corpuslinguistik in unsere Publikationsplattform implementieren, die wir für die Nutzer im Internet zur Verfügung stellen. Es handelt sich neben Funktionen zur Erstellung eines Wortindexes für Einzeltexte oder ganze Textgruppen, oder für Titel- bzw. Personennamenindizes vor allem um statistische Funktionen. So kann man wahlweise für einen Text, eine Textgruppe oder das Gesamtcorpus folgende statistische Analysen durchführen:

- Suchen nach den häufigsten Wörtern,
- Analyse der Verteilung der Worthäufigkeiten,
- Schlüsselwortanalyse,
- Häufigkeit der Wortarten und type / token Statistik.

Für einzelne Wörter der Wortliste können neben der Belegstellenabfrage zu einem Lemma und deren Anzeige in satzweiser Anordnung oder als KWIC-Datei auch Corpusanalysen durchgeführt werden, wie

- die Kollokationsanalyse oder
- die Suche nach dem kombinierten Auftreten zweier Wörter.

Solche Analyseverfahren der Textlinguistik sind jetzt erstmals am ägyptischen Textmaterial durchführbar. Wir hoffen, mit diesem dynamischen Corpus und seiner lexikalischen Erschließung neue Wege für die ägyptologische und auch für die allgemeine Sprachforschung zu öffnen. Das Ägyptische bietet hier als die am längsten überlieferte Schriftsprache der Welt ein besonders interessantes Sprachgut an. Daher wird der Focus der zukünftigen Arbeit auch auf die diachrone Sprachbetrachtung gelegt. Phänomene wie Diglossie, Sprachwandel, Sprachkontakt sowie Texttraditionen und Innovationen werden mit einem wachsenden elektronischen Corpus ägyptischer Texte völlig neu analysierbar sein. Unser Angebot ist als *Thesaurus Linguae Aegyptiae* unter der Adresse **http://aaew.bbaw.de/tla/** erreichbar.