TEXT MINING UND PAPYRI

Reinhold Scholl

Am interdisziplinären Forschungsprojekt eAQUA (Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft) sind verschiedene altertumswissenschaftliche Disziplinen an verschiedenen Standorten in Deutschland und das Institut für automatische Sprachverarbeitung in Leipzig beteiligt[1]. Das Projekt besteht aus mehreren Teilprojekten : Untersuchung zu den Atthidographen[2] ; Platonrezeption in der Antike[3] ; Plautinische Metrik ; das Wissensnetz der Frühen Neuzeit[4] ; delphische Freilassungsinschriften ; Mental Maps[5] ; Papyrologie und Text Mining.

Ziel des Projektes ist es, für die Altertumswissenschaften aus antiken Quellen spezifisches Wissen zu generieren und über ein Web-Portal der praktischen Forschung nachhaltig zur Verfügung zu stellen. Dafür wird in enger Kooperation zwischen Altertumswissenschaftlern und Informatikern die verfügbare Text Mining Technologie den Bedürfnissen und Anforderungen der Altertumswissenschaften angepaßt.

Mit Text Mining werden computerunterstützte Verfahren für die semantische Analyse von Texten bezeichnet, die die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Texten, unterstützen[6]. Mit Hilfe verschiedener Verfahren der automatischen Textverarbeitung versuchen die einzelnen Teilprojekte in eAQUA ihre Forschungsziele zu erreichen. In einer späteren Phase sollen die so gewonnenen und evaluierten Verfahren kombinierbar sein.

Die verschiedenen Verfahren sind :

1. Nicht speziell papyrologisch

Citationsgraph : Ausgehend von einem selbst gewählten antiken Autor werden diejenigen Sätze angezeigt, die fünf gleich lautende nacheinander folgende Wörter gemeinsam haben, und zwar sowohl bei diesem Autor selbst als auch bei allen anderen Autoren des *Thesaurus Linguae Graecae*, so daß auf diese Weise u.a. potentielle Zitate ermittelt werden können.

Bei der Differenzanalyse können Autoren und deren Werke mit einem anderen Autor oder Werk verglichen werden, um Wörter, die gemeinsam sind, aber auch diejenigen Wörter, die nur bei einem der beiden Autoren vorkommen, anzuzeigen. Dies kann sehr hilfreich bei der Zuweisung einzelner Werke zu bestimmten Autoren sein.

Bei der Plautinischen Metrik soll Text Mining helfen, die komplizierte Metrik automatisch zu analysieren[7].

Ausführlicher werden im Folgenden die Suchmaske und die Textvervollständigung vorgestellt und die Klassifikation kurz gestreift[8]. Bei der Suchmaske hat man die Möglichkeit, über verschiedene altertumswissenschaftliche Textcorpora (TLG-E, PHI5, PHI7, PHI7_INS, PHI7_DDP und Epiduke) und auch modernsprachliche Corpora nach bestimmten Wörtern zu suchen. Als Ergebnis einer solchen Suche werden Wörter mit ähnlichem Kontext angezeigt. Darunter erscheint ein Graph mit Kanten und Knoten, der die

1 <http://www.eaqua.net>. Vgl. Schubert / Heyer (2010). In diesem Band sind aus den Teilprojekten Möglichkeiten der konkreten Arbeit mit den einzelnen Methoden aufgeführt.
2 Vgl. Bünte (2010) 10–25.
3 Vgl. Geßner (2010) 26–41.
4 Vgl. Gruhl (2010) 56–70.
5 Vgl. Kath (2010) 71–90.
6 Vgl. Heyer / Quasthoff / Wittig (2008).
7 Vgl. Blumenstein / Deufert / Gaertner (2010) 101–107.
8 Zur Suchmaske, vgl. Schubert (2010) 42–55 ; zur Textvervollständigung, vgl. Rücker (2010) 91–100, basierend auf einer älteren Version.

Actes du 26e Congrès international de papyrologie (Genève 2010) 695–699

syntaktisch-semantischen Verbindungen der einzelnen Wörter in einem Satz graphisch anzeigt. Auf diese Weise kann man entsprechende Abhängigkeiten erkennen. Anschließend folgen die wichtigsten und häufigsten Kookkurrenten, und zwar sortiert, und zusätzlich getrennt nach linken und rechten Kookkurrenten des gesuchten Wortes. Unter Kookkurrenz ist das gemeinsame Auftreten zweier Wortformen in einem lokalen Kontext zu verstehen. Kookkurrenten heißen zwei Wortformen, die in einem lokalen Kontext gemeinsam auftreten. Wortformen, die in syntagmatischer Relation stehen, sind also stets Kookkurrenten.

Beispiel : κακουχεῖν – Suche in der Epiduke-Datenbank

Als Ergebnis werden zunächst Wörter mit ähnlichem Kontext angezeigt.

Abb. 1 : Ergebnisanzeige der Suche Teil 1

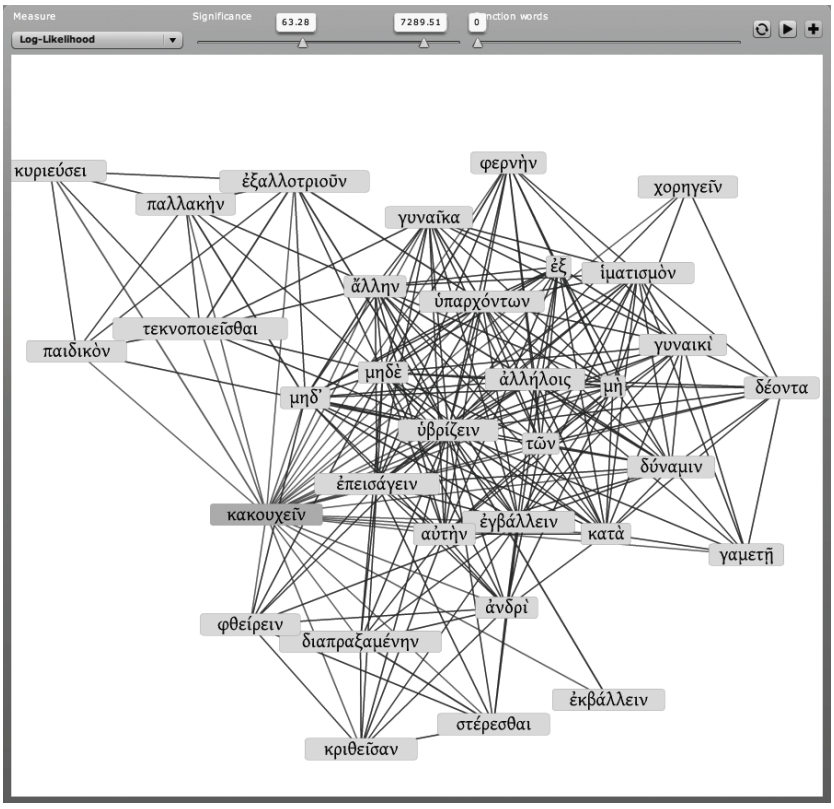
Word *κακουχεῖν (243356)*
Number of occurences *11*
Class of frequency *18*
Words with same normalised form: *κακουχεῖν (11);*

Words with same base form: *κακουχούμενοι (67); κακουχούμενος (14); κακουχουμένων (14); κακουχεῖν (11); κακουχεῖσθαι (7); κακουχουμένους (6); κακουχηθέντες (5); κακουχουμένοις (4); κακουχούμενον (3); κακουχεῖ (3); ἐκακούχησεν (3); κακουχοῦντες (3); κακουχεῖται (3); ἐκακουχήθης (2); ἐκακουχήθη (2); ἐκακούχει (2); κακούχει (2); κακουχηθῆναι (2); κακουχοῦσα (1); κακουχοῦντα (1); κακουχείσθωσαν (1); κακουχήσθαι (1); ἐκακουχήθησαν (1); ἐκακουχήθην (1); ἐκακουχεῖτο (1); Κακούχει (1); κακουχουμένης (1); κακουχηθὲν (1); κακουχοῦνται (1); κακουχῶν (1); κακουχούμεθα (1);*

Words with similar context: *κυριεύσασα (0.5075); καταθεμένη (0.4086); τιμωροῦσιν (0.2857); Εὐρυδίκην (0.2576); ἐκλελοιπέναι (0.2041); νομίσασα (0.1769); παραλαβοῦσα (0.1361); ὑπογραφέων (0.125); ἑαυτῆς (0.1224); ἔλαβε (0.1224); παραδοὺς (0.1224); ἰδίαν (0.1224); ἀρχὴν (0.1224); Οὐετουρία (0.122); ψιλώσασα (0.1212); ὑφηγεῖτο (0.1207); ἐπιγεννηθέντος (0.1176); κατηρτισμένα (0.1163); Κελτοῦ (0.1159); Ζητεῖτε (0.1143); διαλογισμοῦ (0.1127); κατέστελλε (0.1127); μετεμψύχωσιν (0.1088); εὐδαμονίαν (0.1088); καταλιπὼν (0.1088); δείξῃ (0.1088); ἡδονὰς (0.1088); αὐτῆς (0.1088); ἔπεμψε (0.1088); ἤνεγκε (0.1088); εὐτυχίαν (0.1088); ἀνέφερεν (0.1088); ἀσφάλειαν (0.1088); τελευτὴν (0.1088); ἀποδοῦναι (0.1088); φύσεως (0.1088); τό (0.1088); βίου (0.1088); ἀρετὴν (0.1088); Μετὰ (0.1088); ἀρχήν (0.1088); ἡλικίαν (0.1088); τύχην (0.1088); προληφθέντας (0.1053); ἰδοῦσιν (0.1053); Μολοττοὺς (0.1026); κατεγγυωμένη (0.1017); πεντήρη (0.101); ἐναπέθανεν (0.1008);*

Darunter erscheint der Graph zu diesem Wort.

Abb. 2 : Ergebnisanzeige der Suche Teil 2 : Graph zu κακουχεῖν in der Epiduke-Datenbank



Durch Verschieben der Regler am oberen Rand des Graphen können sogenannte *function words* (dazu gehören beispielsweise Artikel, Partikel und sonstige häufige Füllwörter) ausgeblendet werden. Wird der Regler bis zum Anschlag geschoben, werden die 1000 häufigsten Wörter des gesamten Corpus ausgeblendet.

Klickt man κακουχεῖν « schlecht behandeln » im Graphen an, leuchten alle wichtigen Wörter, die syntaktisch und semantisch mit κακουχεῖν in einem Satz verbunden sind, farbig auf. Und in der Tat finden sich in Eheverträgen die Verbote für den Mann: neben μὴ κακουχεῖν auch μὴ ὑβρίζειν und μὴ ἐκβάλλειν, die auch hier im Graphen vertreten sind. Auf diese Weise lassen sich semantische Abhängigkeiten von Wörtern erkennen und visualisieren sowie im Zusammenhang einer möglichen Wortergänzung nutzen.

Es folgt die Anzeige der Kookkurrenten :

Abb. 3 : Ergebnisanzeige der Suche Teil 3 : Anzeige der Kookkurrenten zu κακουχεῖν in der Epiduke-Datenbank

Significant cooccurrences of *κακουχεῖν*

κυριεύσασα (2); τιμωροῦσιν (2); καταθεμένη (2); Εὐρυδίκην (2); νομίσασα (2); ἐκλελοιπέναι (2); συνεχωρήθη (2); ἀσεβέσιν (2); παραλαβοῦσα (2); Ὀλυμπιάς (2); ἀναπλήρωσιν (2); ἀμέλειαν (2); μετανοεῖν (2); νεανίσκων (2); τάνδρὸς (2); εὐτυχίαν (2); βαρέως (2); ἀνθρωπίνως (2); θέλουσιν (2); ἠνεγκεν (2); φέρουσα (2); βασιλικῶν (2); δικαίους (2); αἰχμαλώτους (2); κολάσεως (2); Φίλιππον (2); φυλακὴν (2); τελευτὴν (2); ἐξισχύσει (1); Ὅτι (3); υἱοὺς (2); μεταμεληθέντας (1); χώριζε (1); ἰδίας (2); Μαδιηναῖοι (1); ἐπεκτεῖναι (1); φιλοσόφει (1); μήτηρ (2); ἐκλέ (1); ἀποκτέννει (1); ἐποίησε (2); ζῆν (2); ἄγχειν (1); βασιλείαν (2); σωμάτων (2); τεκμαίρονται (1); φόρει (1); ἀφεθῇ (1); τήν (2); περικειμένου (1); ἐνυπνίου (1); ἀνδρα (2); δακτύλιος (1); φοροῦσιν (1); ἐπιθυμοῦσι (1); τιμωρεῖν (1); χωρὶς (2); ἐκδιδόναι (1); δι' (3); ἐπιθυμοῦσιν (1); περίκειται (1); ἀσκήσει (1); προσφυῶς (1); ἁμαρτήμασιν (1); τὴν (6); ἁρμόζειν (1); ἐπεχείρησεν (1); ἐπεχείρησε (1); εἰληφὼς (1); ἀρξηται (1); νομίμως (1); δεσμὸς (1); δακτύλιον (1); δεσμοῦ (1); τρό (1); τοιοῦτός (1); ἐνύπνιον (1); γεται (1); πραχθέντα (1); προτέροις (1); σῶμα (2); δικαίοις (1); πορνείας (1); κακὸς (1); πον (1); πρῶτον (2); κινδύνου (1); τὸ (6); εἰς (4); κινδύνων (1); δαίμονες (1); μετανοίας (1); τοιαύτῃ (1); καὶ (8); κακά (1); τιμωρίας (1); ἀλλά (1); εἶδον (1); ὢν (1); εἰδὼς (1);

Significant left cooccurrences of *κακουχεῖν*

κυριεύσασα (2); τιμωροῦσιν (2); κακουχεῖν (2); καταθεμένη (2); Εὐρυδίκην (2); νομίσασα (2); ἐκλελοιπέναι (2); ἀσεβέσιν (2); παραλαβοῦσα (2); Ὀλυμπιάς (2); ἀναπλήρωσιν (2); ἀμέλειαν (2); μετανοεῖν (2); νεανίσκων (2); τάνδρὸς (2); εὐτυχίαν (2); βαρέως (2); ἀνθρωπίνως (2); θέλουσιν (2); ἠνεγκεν (2); φέρουσα (2); βασιλικῶν (2); δικαίους (2); κολάσεως (2); Φίλιππον (2); φυλακὴν (2); τελευτὴν (2); Ὅτι (3); υἱοὺς (2); μεταμεληθέντας (1); ἰδίας (2); Μαδιηναῖοι (1); μήτηρ (2); ἐκλέ (1); ἐποίησε (2); ζῆν (2); ἄγχειν (1); βασιλείαν (2); σωμάτων (2); ἀφεθῇ (1); τήν (2); ἀνδρα (2); δακτύλιος (1); φοροῦσιν (1); ἐπιθυμοῦσι (1); τιμωρεῖν (1); χωρὶς (2); ἐκδιδόναι (1); ἐπιθυμοῦσιν (1); περίκειται (1); ἀσκήσει (1); ἁμαρτήμασιν (1); ἀρξηται (1); νομίμως (1); δεσμοῦ (1); τρό (1); ἐνύπνιον (1); γεται (1); πραχθέντα (1); προτέροις (1); δικαίοις (1); κακὸς (1); πον (1); πρῶτον (2); κινδύνου (1); εἰς (4); κινδύνων (1); μετανοίας (1); κακά (1); τιμωρίας (1); εἶδον (1); δι' (2); ὢν (1); εἰδὼς (1); δικαίως (1); τρόπῳ (1); κακόν (1); τοιαύτης (1); ἡνίκα (1); τοὺς (3); νοῦς (1); τοῖς (3); τὸ (5); ἴδιον (1); αὐτὸν (2); καίτοι (1); δηλοῖ (1); καὶ (7); τ' (1); τὴν (4); ὅθεν (1); ἑαυτὸν (1); ποιεῖ (1); ἀπ' (1); τῶν (4); ἐχων (1); ἀλλὰ (2); μηδὲ (1); Τὸ (1); ὅπερ (1);

Significant right cooccurrences of *κακουχεῖν*

κακουχεῖν (2); συνεχωρήθη (2); ἀναπλήρωσιν (2); μετανοεῖν (2); θέλουσιν (2); δικαίους (2); αἰχμαλώτους (2); κολάσεως (2); ἐξισχύσει (1); χώριζε (1); ἰδίας (2); ἐπεκτεῖναι (1); φιλοσόφει (1); ἀποκτέννει (1); τεκμαίρονται (1); φόρει (1); περικειμένου (1); ἐνυπνίου (1); ἐπιθυμοῦσι (1); ἐπιθυμοῦσιν (1); ἀσκήσει (1); προσφυῶς (1); ἁρμόζειν (1); ἐπεχείρησεν (1); ἐπεχείρησε (1); εἰληφὼς (1); δεσμὸς (1); δακτύλιον (1); δεσμοῦ (1); τοιοῦτός (1); σῶμα (2); πορνείας (1); δαίμονες (1); τοιαύτῃ (1); ἀλλά (1); ἐξουσίαν (1); ψυχήν (1); ψυχῇ (1); ἀληθῶς (1); νοῦς (1); ὑπάρχει (1); πως (1); ἑαυτὸν (1); φησὶν (1); ἐχειν (1); ἡμᾶς (1); τῇ (2); τοὺς (2); μὴ (2); ὥσπερ (1); δι' (1); εἰς (2); τις (1); εἰ (1); οὖν (1); τῆς (2); οἱ (1); οὐ (1); ὅτι (1); τὴν (2); ὡς (1); τὸ (2);

Significant left neighbours of *κακουχεῖν*

καταθεμένη (2); ἀσεβέσιν (2); δικαίους (2); υἱοὺς (2); νομίμως (1); μηδὲ (1); καὶ (1);

Significant right neighbours of *κακουχεῖν*

Die Anzeige der Kookkurrenzen für bestimmte Wörter in einem Satz und auch die Differenzierung zwischen linken und rechten Kookkurrenten können ebenfalls zur Überprüfung einer Ergänzung und zum Erkennen von Formeln, Floskeln und Phrasen genutzt werden.

Außer den Kookkurrenten werden auch die linken und rechten Nachbarn angezeigt. Nachbarn sind Wörter, die direkt links und rechts von dem gesuchten Wort stehen. Ganz am Ende folgen noch die Belegstellen mit einem Textauszug. Die Visualisierung ist eines der wichtigen Vorteile dieses tools.

2. Klassifikation

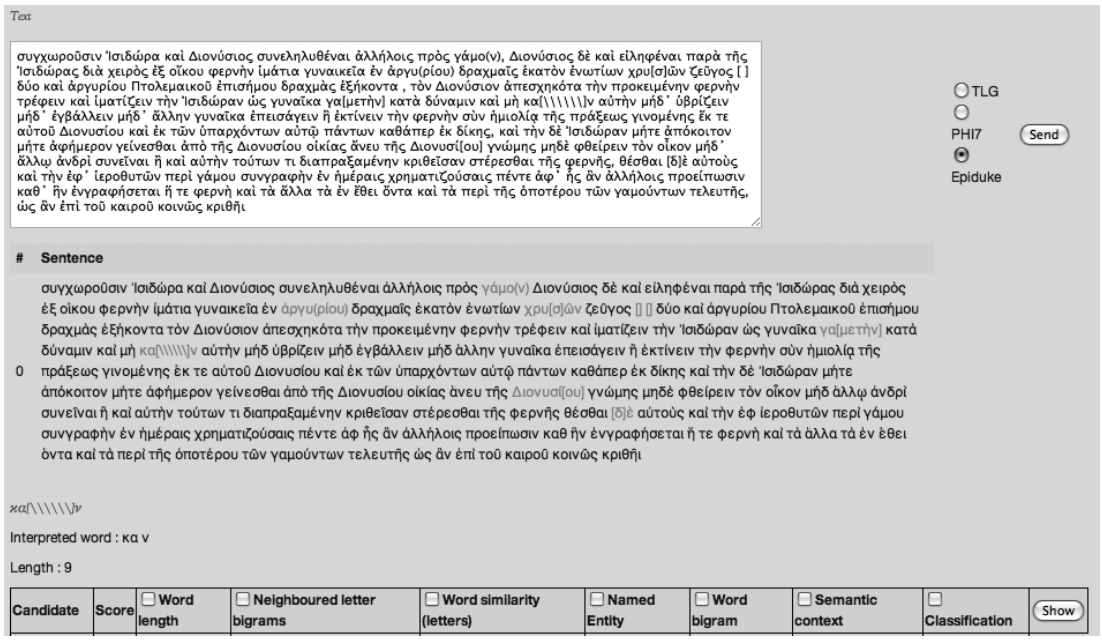
In diesem tool ist das Programm daraufhin trainiert worden, die Klassifikation der griechischen dokumentarischen Papyri automatisch vorzunehmen, und zwar nach der Klassifi-

kation des Sammelbuches, die alle Teilnehmer des Papyrusportals nutzen, die mit MyCoRe arbeiten[9].

3. Textergänzung

Für die Textergänzung, die zur Zeit nur für ein Wort mit bekannter Größe funktioniert, wird als Beispiel ein Papyrus mit κακουχεῖν gewählt, und zwar der erste Beleg aus der Trefferliste der Suchmaske : BGU IV 1050.

Abb. 4 : Maske der Textergänzung



Man fügt den zu bearbeitenden Text in das Textfeld der Maske ein. Für Demonstrationszwecke wurde das Wort κακουχεῖν fragmentiert, in eckige Klammern gesetzt, so daß nur *kappa* und *alpha* am Anfang sowie *ny* am Schluß stehen. Dann setzt man sechs Mal ein Backslash-Zeichen für die fehlenden Buchstaben als Platzhalter ein.

Als Analysecorpus ist die Epiduke-Datenbank gewählt. Nach dem Drücken des Buttons « Send » markiert der Rechner alle mit Klammern versehenen Wörter. Anschließend klickt man auf das zu suchende Wort κα\\\\\\ν.

Der Rechner zeigt das gesuchte Wort und die Zahl der Buchstaben an, aus denen das gesuchte Wort besteht. Dann kann man verschiedene Verfahren der automatischen Sprachverarbeitung wählen und mit « show » die Ergebnisse anzeigen lassen.

– *Wordlength* : Angezeigt werden Wörter, die dieselbe Wortlänge haben wie das gesuchte Wort.
– *Neighboured letter bigrams* : Angezeigt werden Grapheme, die aus so vielen Zeichenketten bestehen, wie das gesuchte Wort hat. Dabei werden die schon bekannten Zeichen vor und nach der Lücke berücksichtigt, um die dazwischen befindlichen möglichen Zeichenketten zu errechnen. Das Ergebnis sind rein mathematisch errechnete Grapheme, die keinerlei lexikalische Bedeutung haben müssen bzw. keine griechischen Wörter sein müssen. Die diakritischen Zeichen werden nicht als eigenständig gezählt. Dieses Verfahren ist somit nur in Verbindung mit anderen Verfahren sinnvoll.

9 Papyrusportal : < http://www.papyrusportal.net>.

- *Word Similarity (letters)* : Angezeigt werden Wörter, bei denen eine Ähnlichkeit über Buchstabenvergleiche errechnet wurde (Levenshtein-Distanz).
- *Named Entity* : Angezeigt werden Wörter, die derselben NE Kategorie (Personennamen, Ortsnamen) angehören.
- *Wordbigram* : Angezeigt werden die signifikanten rechten oder linken Nachbarwörter des gesuchten Wortes nach einer Gewichtung.
- *Semantic context* : Angezeigt werden Wörter, die normalerweise in Sätzen mit dem gesuchten Wort vorkommen. Dabei werden die 200 häufigsten Wörter (*function words*) und Wörter, die weniger als drei Mal vorkommen, nicht berücksichtigt.
- *Classification* : Die automatisch vorgenommene Klassifikation nach dem Sammelbuch wird berücksichtigt.

Die besten Ergebnisse erzielt man in der Kombination der einzelnen Verfahren. Alle genannten tools und Verfahren der Teilprojekte stehen frei unter <http:// www.eaqua.net> zur Verfügung, so daß sie jeder für seine Fragestellungen als Hilfsmittel nutzen kann.

Literaturverzeichnis

Blumenstein, J. / Deufert, M. / Gaertner, J.F. (2010), « Elektronische Analyse der plautinischen Sprechverse : Ein Werkstattbericht », *in* Schubert / Heyer (2010) 101–107.

Bünte, A. (2010), « Text Mining with the Atthidographers », *in* Schubert / Heyer (2010) 10–25.

Geßner, A. (2010), « Das automatische Auffinden der indirekten Überlieferung des Platonischen *Timaios* und die Bedeutung des Tools “CitationGraph” für die Forschung », *in* Schubert / Heyer (2010) 26–41.

Gruhl, R. (2010), « Das Wissensnetz der Frühen Neuzeit. Von der virtuellen Bibliothek zur virtuellen Enzyklopädie », *in* Schubert / Heyer (2010) 56–70.

Heyer, G. / Quasthoff, U. / Wittig, Th. (2008), *Text Mining, Konzepte, Algorithmen, Ergebnisse* (1. Korr. Nachdruck, Bochum).

Kath, R. (2010), « Das Konzept des “einfachen Lebens” in der Antike : Ein Beispiel für die Anwendung von Textmining-Verfahren in der Geschichtswissenschaft », *in* Schubert / Heyer (2010) 71–90.

Rücker, M. (2010), « Die Möglichkeiten der automatischen Textergänzung auf Papyri », *in* Schubert / Heyer (2010) 91–100.

Schubert, Ch. (2010), « Zitationsprofile, Suchstrategien und Forschungsrichtungen », *in* Schubert / Heyer (2010) 42–55.

Schubert, Ch. / Heyer, G. (2010) (Hrsg.), *Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung* I (*Working Papers Contested Order* 1, Leipzig).