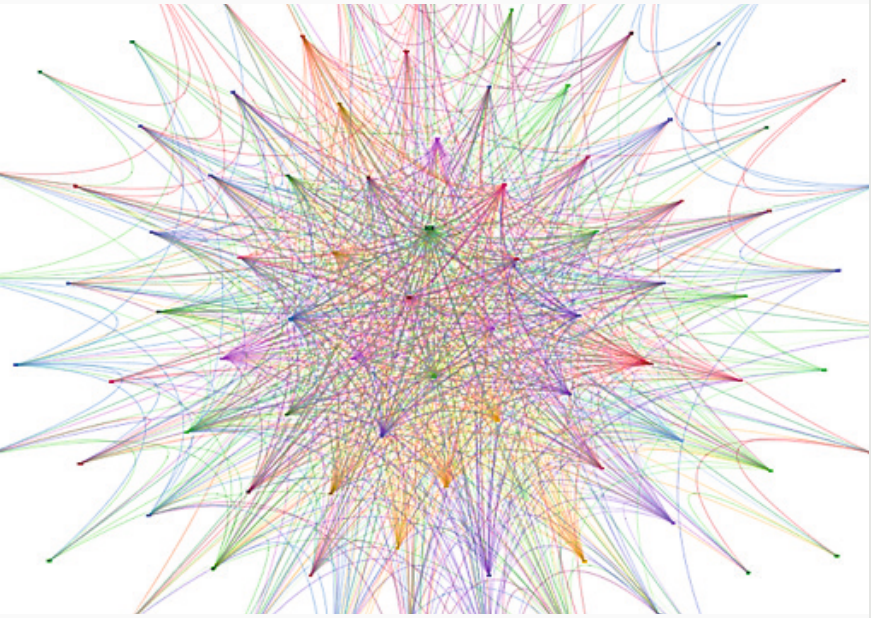
Jens Wittig / Corina Willkommen

Digital Classics Books 2

DIGITAL CLASSICS IN DER PRAXIS



Arbeiten mit eAQUA:
Eine Einführung mit Beispielen

Propylaeum
FACHINFORMATIONSDIENST ALTERTUMSWISSENSCHAFTEN

Arbeiten mit eAQUA:
Eine Einführung mit Beispielen

Digital Classics in der Praxis

DIGITAL CLASSICS BOOKS – 2

Arbeiten mit eAQUA:
Eine Einführung mit Beispielen

DIGITAL CLASSICS IN DER PRAXIS

Jens Wittig und Corina Willkommen

Propylaeum
FACHINFORMATIONSDIENST
ALTERTUMSWISSENSCHAFTEN

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://dnb.dnb.de abrufbar.

Propylaeum
FACHINFORMATIONSDIENST ALTERTUMSWISSENSCHAFTEN

Publiziert bei Propylaeum,
Universitätsbibliothek Heidelberg 2019.

Diese Publikation ist auf https://www.propylaeum.de dauerhaft frei verfügbar (open access).
urn: urn:nbn:de:bsz:16-propylaeum-ebook-431-4
doi: https://doi.org/10.11588/propylaeum.431

Umschlagabbildung: Beispiel einer Kookkurrenzvisualisierung mit eAQUA.

eISSN: 2627-5988
ISSN: 2566-7890

ISBN: 978-3-947450-27-5 (PDF)
ISBN: 978-3-947450-28-2 (Softcover)
ISBN: 978-3-947450-31-2 (Hardcover)

Inhaltsverzeichnis

Vorwort der Herausgeber

„Arbeiten mit eAQUA: Eine Einführung mit Beispielen“ soll, wie der Untertitel DIGITAL CLASSICS IN DER PRAXIS besagt, in die praktische Arbeit mit einigen aus dem Textmining stammenden Tools des Webportals eAQUA einführen. Das Buch kann im Rahmen von althistorisch-philologischen Übungen als Lehrbuch für den Einsatz digitaler Tools zur Textanalyse verwendet werden. Aber auch unabhängig von der Basis in den Altertumswissenschaften können die Erklärungen und Beispiele als Begleitmaterial für Lehrveranstaltungen eingesetzt werden, in denen die Grundlagen des Textmining vermittelt und angewendet werden.

Die Einführungen, Erklärungen und Beispiele, die die Autoren hier in verständlicher Sprache und mit Illustrationen anschaulich präsentieren, sind in verschiedenen Lehrveranstaltungen erprobt und evaluiert worden. Sie haben einen Stand der Konsolidierung erreicht, der heute Replikation und Transfer in andere Bereiche ermöglicht. Der vorliegende Band antwortet damit auf ein nach unserem Eindruck steigendes Bedürfnis nach solchen praktischen Anleitungen, die mit Rücksicht auf die sehr unterschiedlichen Ausgangsvoraussetzungen insbesondere diejenigen ansprechen soll, die an dem praktischen Einsatz von Tools aus dem Textmining für die jeweils eigenen, fachspezifisch geprägten Fragestellungen interessiert sind, jedoch nicht über informationswissenschaftliches Basiswissen verfügen.

An dieser Stelle sei auch den vielen Teilnehmerinnen und Teilnehmern der Lehr- und Demonstrationsveranstaltungen gedankt, deren Fragen und Anregungen bei der Konzeption und Ausgestaltung dieser Anleitung eine wichtige Rolle spielten.

Die Publikation nicht nur als gedrucktes Buch, sondern vor allem auch in kostenfreier, digitaler Fassung (als PDF- und HTML-Datei) soll einen niedrigschwelligen und anwendungsorientierten Zugang für die Nutzer und Nutzerinnen ermöglichen.

Roxana Kath – Michaela Rücker – Reinhold Scholl – Charlotte Schubert

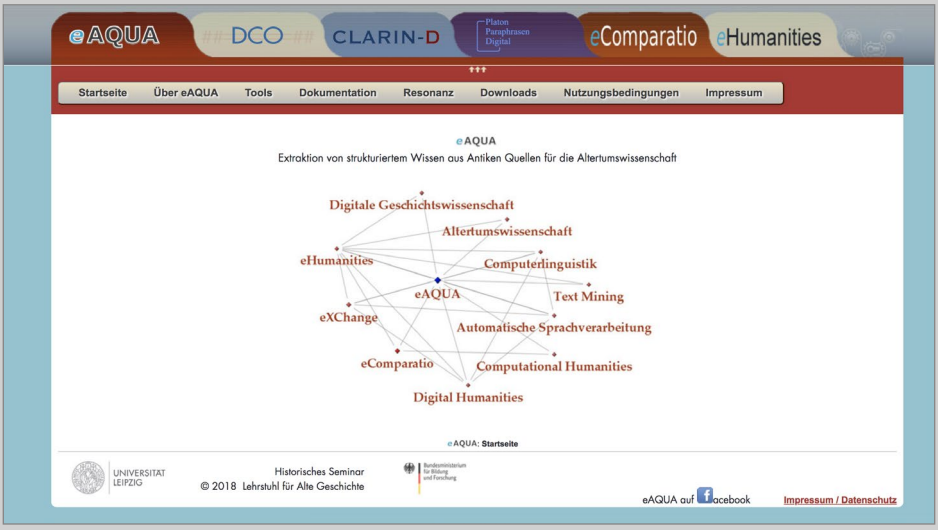


Abbildung 1. Startseite www.eaqua.net

eAQUA – Was ist das?

Zur Geschichte des Projektes

Das heutige Portal eAQUA (Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft) ist aus einer Projektförderung des Bundesministeriums für Bildung und Forschung (BMBF) hervorgegangen (■ **Abbildung 1**). Die Förderung umfasste eine erste Phase von 2008 bis 2011, in der Altertumswissenschaftler[1], Frühneuzeitler und Informatiker zusammenarbeiteten, um die Anwendung fortgeschrittener Werkzeuge aus dem Bereich des Text Mining für die beteiligten Fachdisziplinen erstmals experimentell zu erproben. Das Projekt wurde damals im Rahmen des vom BMBF im Förderschwerpunkt „Geistes- und Sozialwissenschaften" aufgelegten Programms „Wechselwirkungen zwischen Geistes- und Naturwissenschaften" (2008–2011) gefördert und umfasste in diesem Zeitrahmen acht Teilprojekte, die Verfahren aus dem Bereich des Text Mining für Anwendungsszenarien unterschiedlicher Genres und Quellengattungen entwickelten:

- Projekt Atthidographen (Leitung: Ch. Schubert, Alte Geschichte, Universität Leipzig)
- Projekt Platon (Leitung: K. Sier, Gräzistik, Universität Leipzig)
- Projekt Metrik (Leitung: M. Deufert, Latinistik, Universität Leipzig)
- Projekt Camena (Leitung: W. Kühlmann, Germanistik, Universität Heidelberg)
- Projekt Inschriften (Leitung: B. Meißner, Alte Geschichte, Universität der Bundeswehr Hamburg)
- Projekt Papyri (Leitung: R. Scholl, Alte Geschichte, Universität Leipzig)
- Projekt Fehlererkennung (Leitung: G. Heyer, Informatik, Universität Leipzig)
- Projekt Mental Maps (Leitung: Ch. Schubert, Alte Geschichte, Universität Leipzig).

1 Das Autorenteam hat sich beim Formulieren der Texte um eine genderneutrale Sprache bemüht. In den wenigen Fällen, in denen dies nicht möglich war, haben wir uns zugunsten der Lesbarkeit und des flüssigen Textlaufs für das generische Maskulinum entschieden.

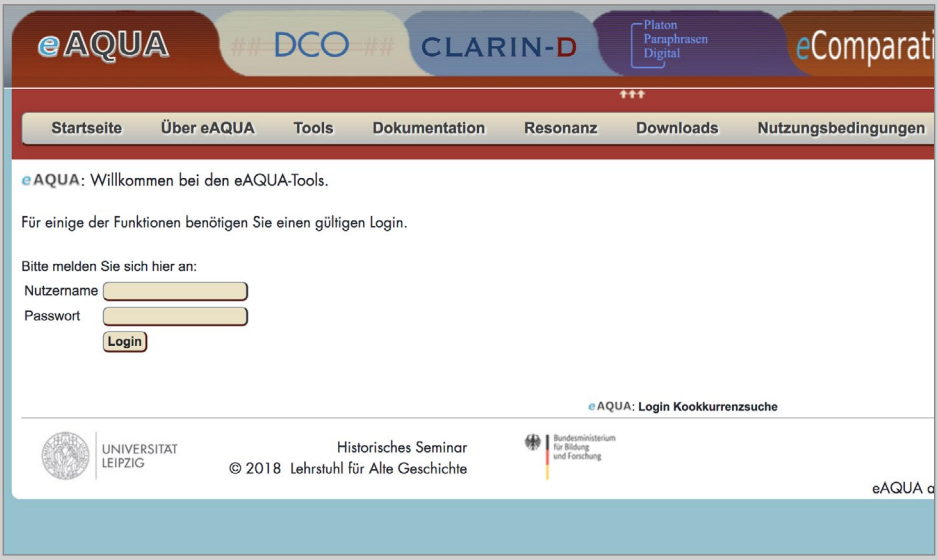


Abbildung 2. Login geschützter Zugang

Ziel der interdisziplinären Zusammenarbeit zwischen geisteswissenschaftlichen Fächern und der Informatik war es, neues und strukturiertes Wissen aus antiken oder frühneuzeitlichen Quellen zu gewinnen und dabei Werkzeuge und Verfahren aus dem Segment des Text Mining weiterzuentwickeln.

In der zweiten Projektphase (2011–2013) wurde das Projekt in der alleinigen Verantwortung des Lehrstuhls für Alte Geschichte vom BMBF weitergefördert, die über den reinen Projektstatus hinausführen sollte, um eine nachhaltige und breite Nutzung zu ermöglichen.

Die in diesem Buch vorrangig behandelten Suchvarianten Kookkurrenzsuche und Parallelstellensuche,[2] die in dieser zweiten Projektphase maßgeblich auf Anwendungsstabilität hin weiterentwickelt wurden, gehen über die üblichen Möglichkeiten digitaler Bibliotheken hinaus und ermöglichen die Erschließung von Abhängigkeiten, Einflüssen und Transferwegen des Wissens in größerem Maßstab.

Die Kookkurrenzsuche, vorrangig im eAQUA-Teilprojekt Atthidographen 2008–2011 entwickelt, führt heute zur Erschließung von semantischen Zusammenhängen, die Zitationssuche, vorrangig im eAQUA-Teilprojekt Platon 2008–2011 entwickelt, ermöglicht heute die Auflistung von Textpassagen, die Parallelen zwischen einem Werk und dem gesamten Referenzkorpus darstellen. Die hier vorgestellten Tools sind in der zweiten Förderphase des Projektes 2011–2013, sowie auch seither umfangreich weiterentwickelt und verbessert worden. Sie werden auf der Internetseite von eAQUA[3] zur Anwendung sowohl für frei zugängliche Textkorpora als auch für Korpora, die einer Lizenzpflicht unterliegen, für alle Interessenten angeboten. Bei Benutzung lizenzpflichtiger Korpora ist nach gegenwärtigem Stand des Urheberrechts[4] ein individuell zu vergebender Zugang in den geschützten Bereich der Webseite notwendig (■ **Abbildung 2**). Der geschützte Zugang trägt dabei dem Umstand Rechnung, dass es laut Urheberrechtsgesetz zwar erlaubt ist, 75 Prozent eines Werkes für die eigene Forschung weiterzuverarbeiten, aber nur bis zu 15 Prozent eines veröffentlichten Werkes vervielfältigt, verbreitet, öffentlich zugänglich gemacht oder wiedergegeben werden dürfen.[5]

---

2 Parallelstellen- und Zitationssuche werden synonym verwendet, auf eine genaue Unterscheidung wird hier verzichtet.

3 URL: http://www.eaqua.net.

4 Urheberrechtsreform 2018 mit dem Namen „Urheberrechts-Wissensgesellschafts-Gesetz (UrhWissG)".

5 Gesetz über Urheberrecht und verwandte Schutzrechte – UrhG: Teil 1, Abschnitt 6, Unterabschnitt 4: Gesetzlich erlaubte Nutzungen für Unterricht, Wissenschaft und Institutionen; insbesondere §60c und §60d. URL: https://www.gesetze-im-internet.de/urhg/.

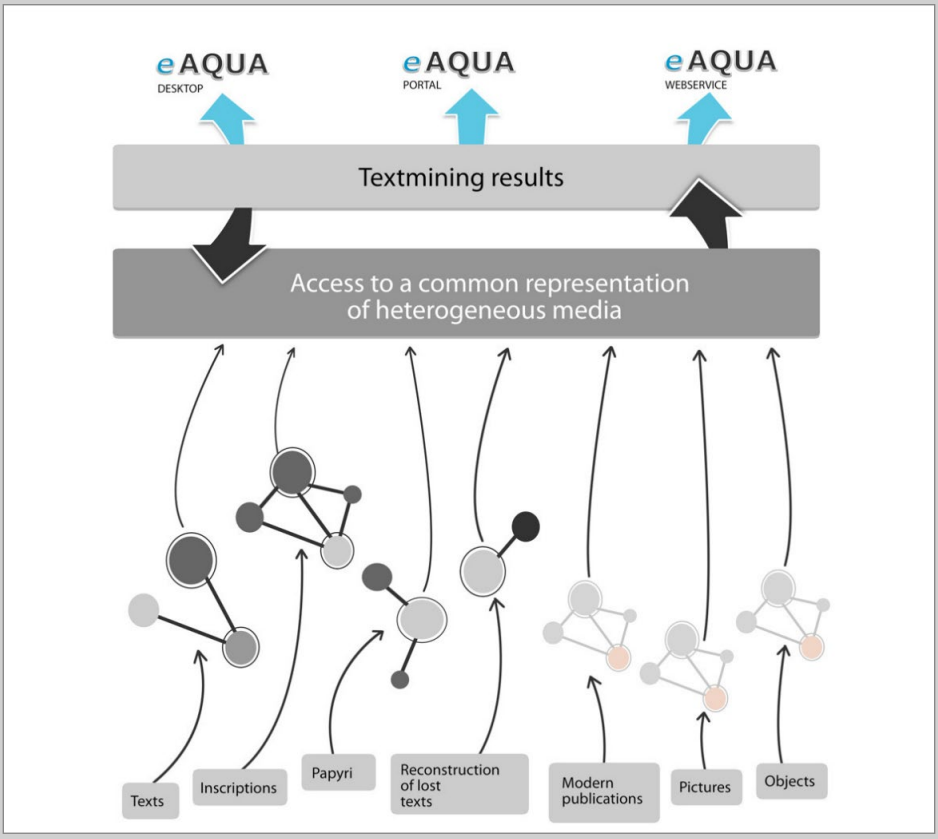


Abbildung 3. Das Portal eAQUA[6]

6 Charlotte Schubert, Gerhard Heyer: Working Papers Contested Order: Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I – Eine Einführung in das Portal eAQUA; Abb. 1, Seite 7; DOI: https://doi.org/10.11588/ea.2010.0.

Technische Komponenten von eAQUA

In technischer Sicht bestehen die unter dem Begriff eAQUA subsumierten Software-Entwicklungen aus zwei Kernbereichen (■ **Abbildung 3**). Auf der einen Seite aus einer Reihe von Tools, die sogenanntes Text Mining betreiben, die also digitalisierte Korpora verarbeiten, berechnen und Ergebnisse ermitteln. Diese Tools laufen in separierten Serverumgebungen mit entsprechender Rechenkapazität, sind jedoch zumeist so konzipiert, dass sie auch auf Desktop-Rechnern zum Einsatz kommen könnten, wenn die benötigten Laufzeitumgebungen installiert sind. Bei der Konzeption wurde auf eine aufwendig zu programmierende GUI verzichtet, die Programme werden mittels Konsolenbefehlen in der Shell gestartet.

Die so bezeichnete eAQUA-Toolchain besteht aus einer Ansammlung von JAVA-Programmen und Shell-Skripten, die mittels Apache Ant erstellt wurden und demzufolge eine Java-Laufzeitumgebung (JRE) benötigen. Darüber hinaus sollten die für die Erzeugung der Datenbanken nötigen Datenbankmanagementsysteme (DBMS), MySQL oder MariaDB vorhanden und installiert sein, da die Berechnungsergebnisse in diese eingespielt werden.

Auf der anderen Seite gibt es webbasierte Anwendungen zur Präsentation der Ergebnisse und zur Interaktion von Nutzern in dem aufbereiteten Datenmaterial. Im Wesentlichen setzen die webbasierten Programme auf einen LAMP Stack (Linux, Apache, MySQL, PHP) auf, sind also so konzipiert, dass sie auf einem üblichen Web-Server einsatzfähig sind. Konkret werden alle dynamischen Inhalte mittels PHP aus MySQL- (bzw. MariaDB-) Datenbanken ausgelesen und zu statisch auslieferbaren Inhalten verarbeitet. Dabei werden nicht nur HTML-Seiten erzeugt, sondern die Daten auch in JSON- oder CSV-Dateien umgeschrieben, um sie später über JavaScript-Bibliotheken darzustellen bzw. interaktiv herunterladen zu können.

Zur Visualisierung innerhalb der Internetseiten kommen freie JavaScript-Bibliotheken zum Einsatz, die entweder unter Creative Commons, Apache 2.0 oder MIT lizensiert sind. Dabei handelt es sich um Google Visualization API, jQuery, jquery-svg-pan-zoom und vis.js.

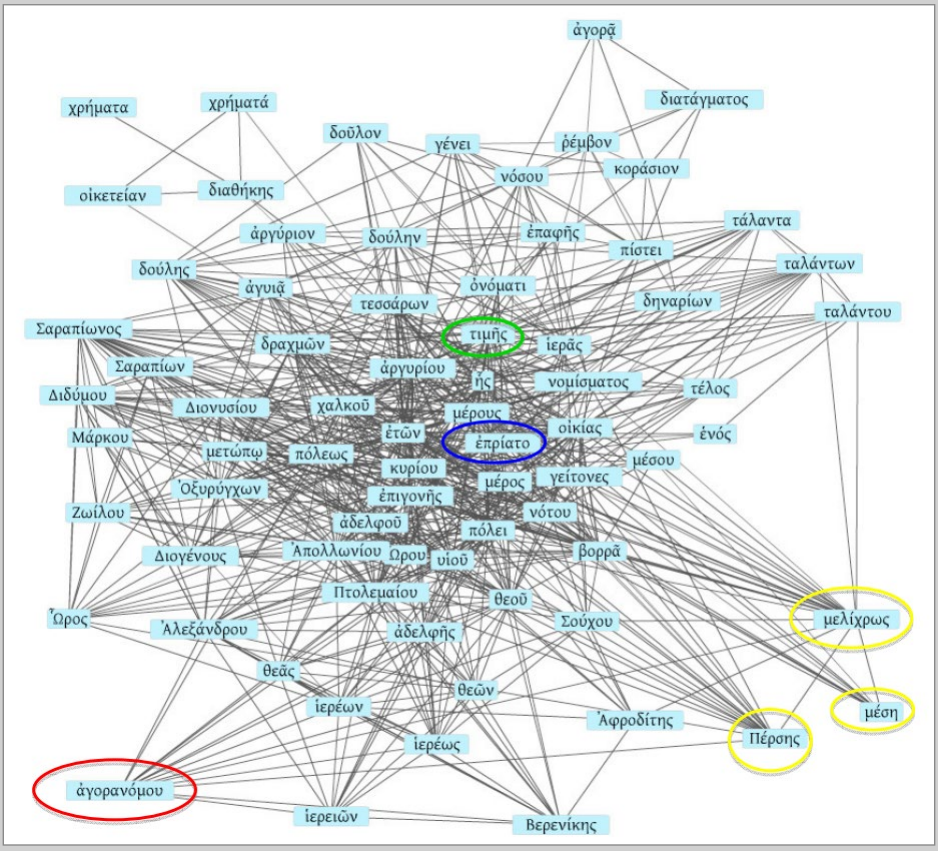


Abbildung 4. Der ursprüngliche Kookkurrenzgraph in Flash[7]

7 Charlotte Schubert, Gerhard Heyer: Working Papers Contested Order: Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I – Eine Einführung in das Portal eAQUA; Abb. 7, Seite 7; DOI: https://doi.org/10.11588/ea.2010.0.

Im älteren Begleitmaterial zum Portal eAQUA[8] ist von einer Cocoon-basierten Architektur die Rede, welche bei der internen und externen Darstellung vollständig auf XML setzte. Aus Performance- und Lesbarkeitsgründen lassen sich die Ergebnisse, hier insbesondere die der Kookkurrenz- und Parallelstellenanalyse nur unzureichend in XML darstellen, so dass im Zuge der Neugestaltung der webbasierten Anwendungen nicht nur auf die zuvor eingesetzte Rich-Client-Technologie Flash,[9] sondern auch auf die Java-Servlet-Technologie zur Darstellung von Webinhalten verzichtet wurde (■ **Abbildung 4**). Im Bereich des Preprocessing wird weiterhin auf die ANT-basierte Toolchain zurückgegriffen, die für die Aufbereitung und Extraktion der Texte, die Segmentierung der Sätze, Tokenisierung von Wörtern bis hin zur Erstellung der Datenbank eingesetzt wird.

8 Beispielsweise in den eAQUA Working Papers, einer Reihe der Working Papers Contested Order des Profilbildenden Bereichs Contested Order der Universität Leipzig; URL: https://journals.ub.uni-heidelberg.de/index.php/eaqua-wp.

9 Insbesondere als Visualisierungskomponente.

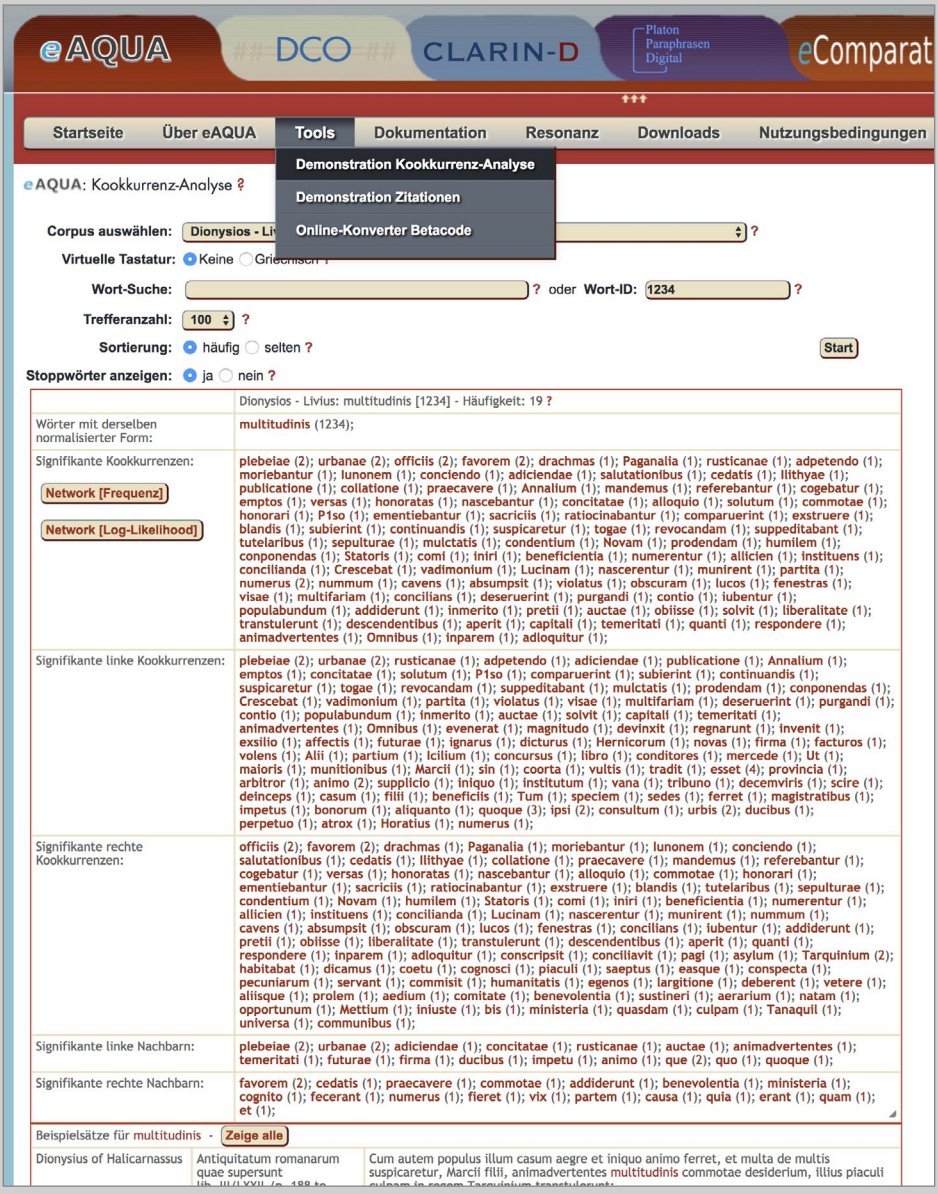


Abbildung 5. Demonstration Kookkurrenz-Analyse

Die Online-Tools von eAQUA

Demonstration Kookkurrenz-Analyse

Die Demonstration der Kookkurrenz-Analyse ist über den Menü-Punkt Tools erreichbar (■ **Abbildung 5**). Sie beinhaltet einige grundsätzliche Analyseergebnisse aus dem Projekt-Portal und wurde im Laufe der Dissemination um weitere Korpora ergänzt. Für den Login in den geschützten Bereich, in dem Ergebnisse von Korpora abgerufen werden können, die an Benutzungslizenzen gebunden sind, ist am rechten Bildschirmrand ein Link zum Login vorgesehen.

Nach Auswahl des Korpus im oberen Bereich gibt es mehrere Möglichkeiten für eine Suche in der Datenbank. Im Feld Wort-Suche kann nach einem Wort oder Wort-Teil in der Datenbank gesucht werden. Das Feld ist mit einer Autovervollständigung versehen, die ab drei Buchstaben wirksam wird. Diese listet alle gefundenen Wörter (mit ihren Häufigkeiten in Klammern) auf. Bei Klick auf eines der Listenergebnisse wird dieses Wort übernommen. Es ist daran zu erkennen, dass rechts im Feld Wort-ID eine Zahl steht. Bei Klick auf die Schaltfläche Start würde dieses Wort analysiert.

Falls keine Vorschläge unterbreitet werden, kann in der Datenbank nach Treffern gesucht werden, indem man nur die Anfangsbuchstaben in das Feld eingibt und dann auf Start klickt. Es erscheint eine Liste aller Terme, die mit diesen Buchstaben beginnen. Die Begriffe sind mit einem Link versehen, der die Kookkurrenz-Analyse für den Suchbegriff startet.

Falls durch eine vorhergehende Suche die ID des Wortes bekannt ist, kann diese auch in das Feld Wort-ID eingetragen und so die Analyse gestartet werden.

Es gibt bei der Kookkurrenzanalyse eine Beschränkungsmöglichkeit hinsichtlich der maximal anzuzeigenden Treffer, die über den Dropdown-Button Trefferanzahl einstellbar sind. Dies ist wegen der dynamisch geladenen Visualisierung notwendig, damit die Seite nicht allzu lang zum Laden benötigt. Um auch unsignifikante Kookkurrenzen finden zu können, ist ein zusätzlicher Sortierungsfilter eingebaut. So lassen sich zu einem Wort zum Beispiel die 100 unsignifikantesten Kookkurrenzen anzeigen, indem die Trefferanzahl auf 100 und die Sortierung auf selten gesetzt wird.

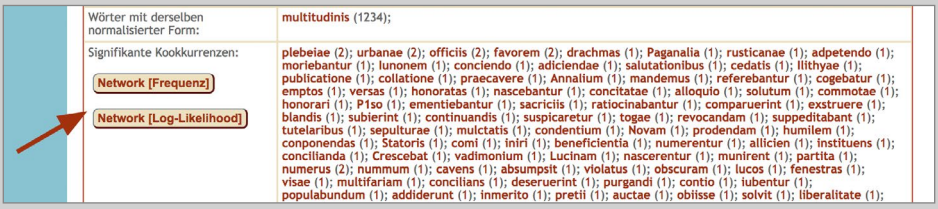


Abbildung 6. Schaltflächen zum Aufruf der Netzwerk-Visualisierung

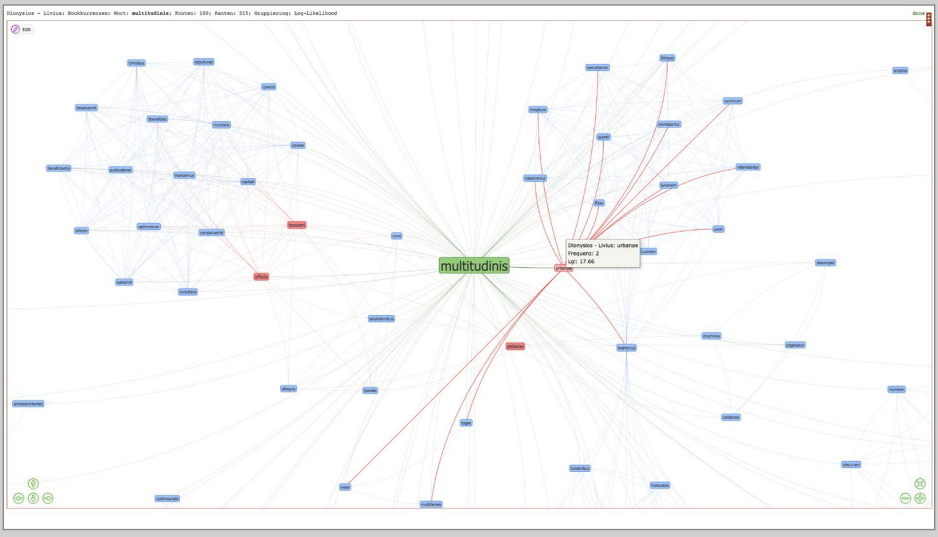


Abbildung 7. Netzwerk-Visualisierung von Kookkurrenzen

Das Ergebnis präsentiert sich als Liste unterteilt in Wörter mit ähnlichem Zusammenhang, den signifikantesten (bzw. hier auch unsignifikantesten) Satzkookkurrenzen (gesamt / links / rechts) sowie den (un-)signifikantesten rechten und linken Nachbarn. Außerdem werden die (un-)signifikantesten Kookkurrenzen in einer Visualisierung dargestellt. Zu deren Aufruf werden in der linken Spalte unterhalb von „Signifikante Kookkurrenzen" zwei Schaltflächen eingeblendet (■ **Abbildung 6**).

Ganz am Ende finden sich Belegstellen, in denen das Wort Verwendung findet.

Als Satzkookkurrenz wird das gemeinsame Auftreten zweier Wörter innerhalb eines Satzes verstanden. Stehen die beiden Wörter direkt nebeneinander, wird auch von Nachbarschaftskookkurrenz gesprochen. Steht ein Kookkurrent links / rechts vom Ausgangswort, wird er unter linke / rechte Kookkurrenzen einsortiert, steht er direkt, ohne ein Wort dazwischen, daneben, wird er zusätzlich unter linke / rechte Nachbarn gelistet. Die Listen der linken / rechten Nachbarn sind demzufolge jeweils eine Teilmenge der Listen der linken / rechten Kookkurrenzen. Wörter mit ähnlichem Zusammenhang ergeben sich dann, wenn die Kookkurrenzprofile vergleichbar sind.

Die Listen der Kookkurrenzen sind jeweils mit Querverweisen hinterlegt. Bei Klick auf ein Wort wird die Kookkurrenzsuche genau für diesen Term gestartet, bei Klick auf die Zahl in Klammern, die die Häufigkeit der Vorkommen angibt, werden Belegstellen mit beiden Ausdrücken angezeigt.

Visualisierung der Kookkurrenzen

In der Liste der Kookkurrenzen zu einem Wort werden in der linken Spalte unterhalb des Eintrags „Signifikante Kookkurrenzen" dynamisch zwei Schaltflächen eingeblendet. Ein Klick darauf öffnet ein neues Browserfenster mit einer in JavaScript geschriebenen Netzwerk-Visualisierung vis.js[10] (■ **Abbildung 7**). Es besteht die Möglichkeit, die Knoten einmal nach der Frequenz (Häufigkeit) der Kookkurrenzen, zum anderen nach dem Signifikanzmaß Log-Likelihood zu gruppieren. Zu den Signifikanzmaßen und deren Bedeutung finden sich die Erklärungen weiter unten.

Gruppierungen werden optisch durch unterschiedliche Farben dargestellt. Ähnliche Werte erhalten die gleiche Farbe. Für die Stabilisierung des sich darstellenden Netzwerkes können verschiedene Algorithmen benutzt werden, die in der rechten Konfigurationsspalte auswählbar sind. Voreingestellt ist hierbei der Barnes-Hut-Algorithmus,[11] der die Anzahl der zu berechnenden Wechselwirkungen durch geschicktes Zusammenfassen reduziert.

10 URL: http://visjs.org/; Apache 2.0 / MIT-Lizenz.

11 Josh Barnes, Piet Hut (1986): A hierarchical O(N log N) force-calculation algorithm, in Nature 324, 446–449; https://doi.org/10.1038/324446a0.

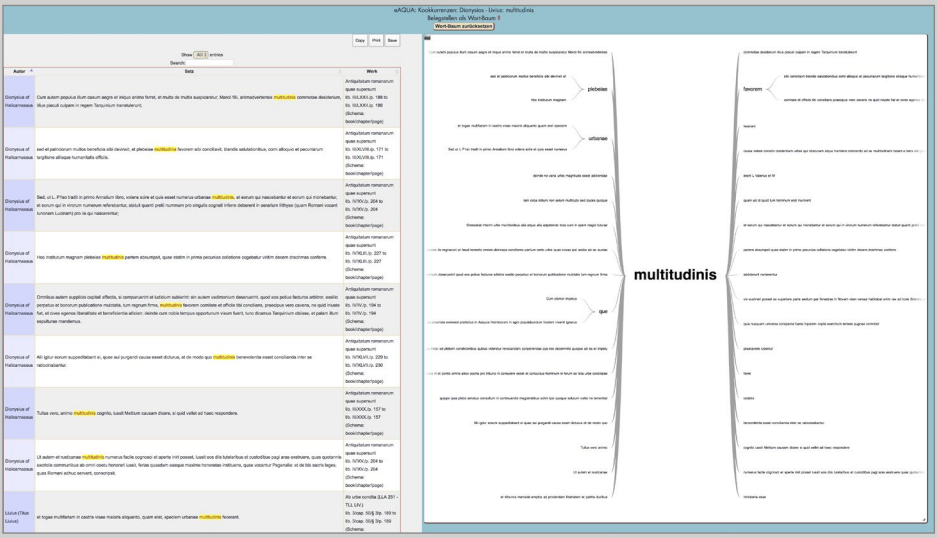


Abbildung 8. Belegstellenanzeige mit Wortbaum

Um die Darstellungszeiten zu beschleunigen, werden die ersten Berechnungsrunden[12] im Hintergrund durchgeführt, danach werden lediglich die Knoten und erst ganz zum Schluss die Kanten (Verbindungslinien) dargestellt.

Auch noch während der Stabilisierung können die Kanten an- oder abgestellt werden und zwar rechts in den Einstellungs-Optionen edges: hidden.

Belegstellen und Wortbaum anzeigen

Der Wortbaum stellt mehrere parallele Wortfolgen dar, um zu zeigen, welche Wörter einem Zielwort am häufigsten folgen oder vorangehen (■ **Abbildung 8**). Als Basis dienen die kompletten Tokens (Sätze) aller Satz-Kookkurrenzen zu einem Wort, die hier als Belegstellen bzw. Beispielsätze bezeichnet sind.

Wurde zuvor ein bestimmtes Wort als Kookkurrent zum Zielwort ausgewählt, reduziert sich die angezeigte Belegstellenliste auf diejenigen, die beide Begriffe enthalten und der Wortbaum rückt an die Stelle vor. Der gleiche Effekt lässt sich erreichen, indem innerhalb des Wortbaumes auf einen Begriff geklickt wird.

Es gibt mehrere Möglichkeiten, das Fenster mit den Fundstellen und dem Wortbaum aufzurufen:

- Nach der Suche in den Listen zu den signifikanten Kookkurrenzen, indem auf die Zahl in den runden Klammern geklickt wird.
- Unterhalb der signifikanten Kookkurrenzen, indem auf die Schaltfläche „Zeige alle“ geklickt wird.
- In der Netzwerk-Visualisierung, indem auf einen Knoten geklickt wird.

12 Genaugenommen ist es die achtfache Anzahl der Knoten.

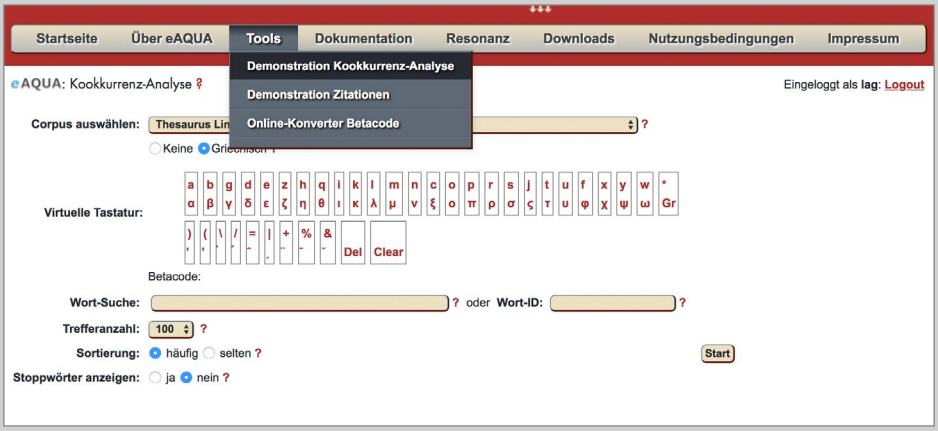


Abbildung 9. Kookkurrenzanalyse Auswahl TLG-E

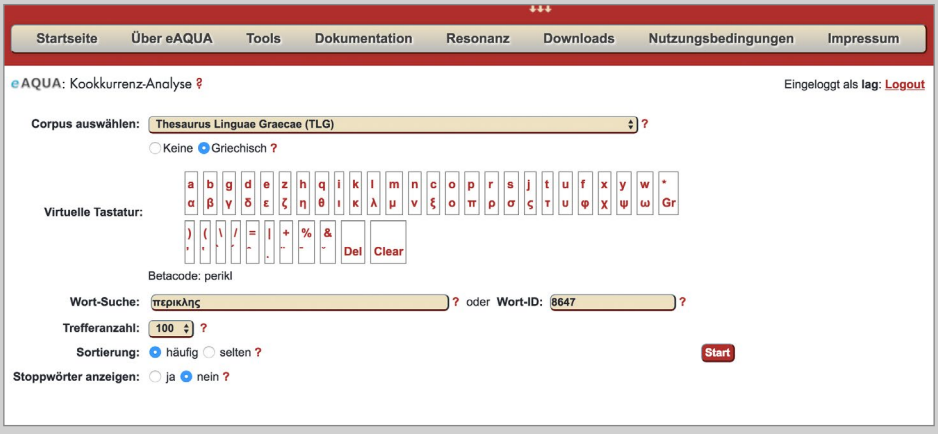


Abbildung 10. Kookkurrenzanalyse Suchmaske zu Perikles

Beispielbenutzung Schritt für Schritt

Korpus auswählen

Unter dem Reiter Tools öffnet sich das Tool „Kookkurrenz-Analyse“, anschließend eine Suchmaske. Der erste Arbeitsschritt ist die Auswahl des Textkorpus. Über die Vorauswahl des Korpus ist eine Wahl zwischen verschiedenen freien und lizenzpflichtigen Korpora bzw. Subkorpora möglich. In diesem Beispiel dient der „Thesaurus Linguae Graecae“ (TLG-E) als Ausgangsbasis. Hierfür sind derzeit Zugangsdaten und Berechtigungsnachweise notwendig. Die Zugangsdaten können eingegeben werden, indem rechts auf Login geklickt wird (■ **Abbildung 9**).

Suchmaske: Wort-Suche

Das griechische (bzw. in der Auswahl eines lateinischen Korpus das lateinische) Wort wird nach Installation einer Altgriechisch-Tastatur mit Akzenten und ohne Trunkierungen in das Eingabefeld eingegeben bzw. direkt über den Kopierbefehl eingefügt. Für den Fall, dass das Programm das gesuchte Wort nicht erkennt, werden vergleichbare Wörter vorgeschlagen. Bei wiederholter Suche des gleichen Wortes empfiehlt es sich, die Word ID zu notieren und in das dafür vorgesehene Feld einzutragen. Mit dem Befehl „Start“ wird das gesamte ausgewählte Korpus, basierend auf den satzweise enthaltenen Texten, nach der eingegebenen Wortform hin analysiert.

Suchmaske: Word-ID

Die Word-ID legt für jede Wortform im zugrundeliegenden Korpus eine bestimmte Nummer fest. In diesem Beispiel ist die ID für das Suchwort Περικλῆς die Zahl 8647 (■ **Abbildung 10**).

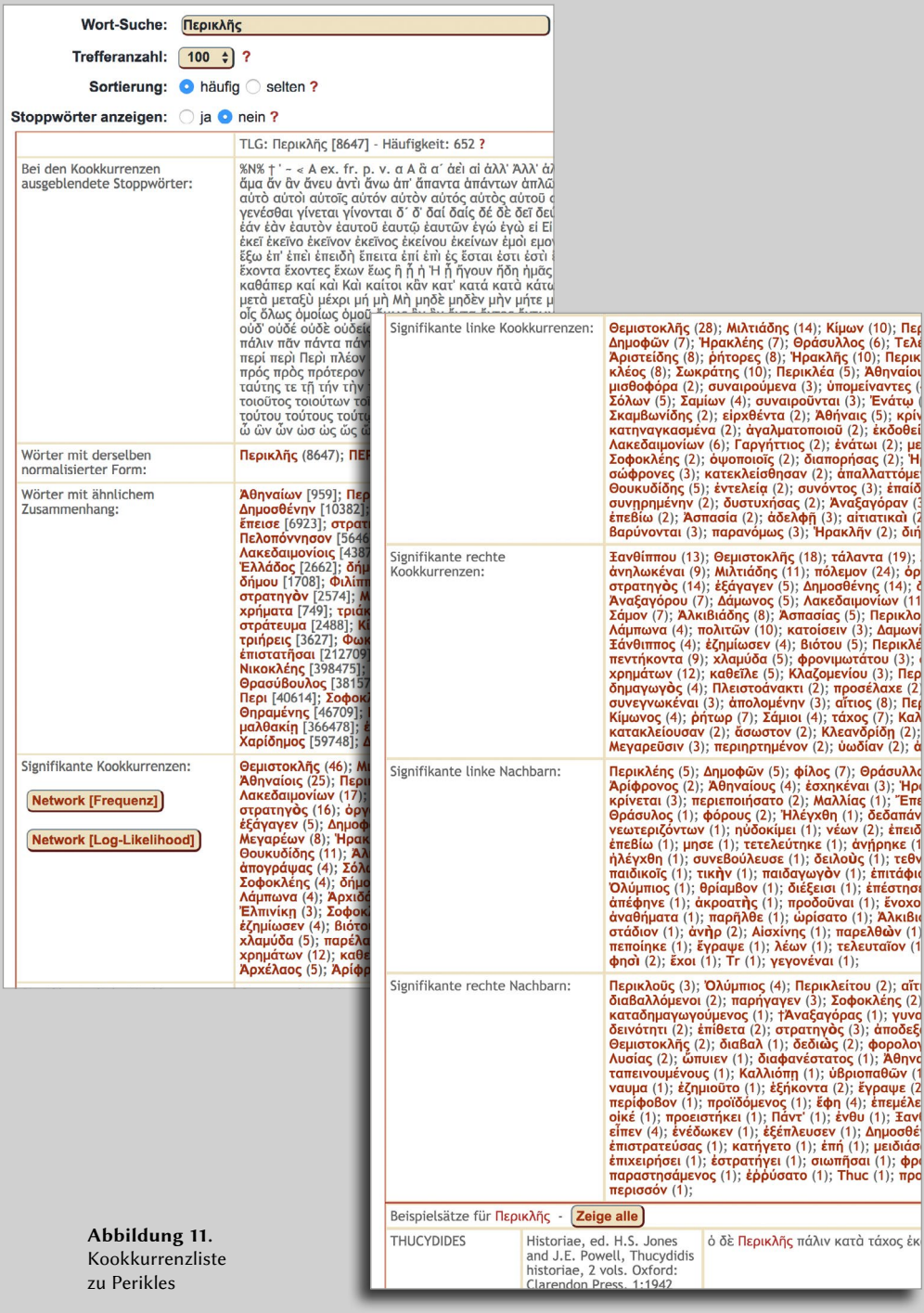


Abbildung 11.
Kookkurrenzliste zu Perikles

Trefferanzahl und Sortierung

Nachfolgend kann das Ergebnis nach der Trefferanzahl eingegrenzt werden oder aber man lässt sich alle Treffer anzeigen. Für das genannte Suchwort ergäbe das eine Trefferanzahl von 652, wie in der Ergebnisanzeige zu sehen ist (■ **Abbildung 11**). Weiterhin kann die Ergebnistabelle nach besonders häufigen oder selten vorkommenden Kookkurrenzen sortiert werden.

Stoppwörter

Eine Liste mit sogenannten Stoppwörtern ermöglicht es, diese aus dem Gesamtergebnis auszusortieren. Dabei handelt es sich um vorab definierte Wörter wie Artikel, Pronomen etc., die keine inhaltliche Aussagekraft haben, aber so häufig sowohl im Graphen wie in der Ergebnistabelle vorkommen, dass die Aussagekraft der Ergebnisse davon nachhaltig beeinflusst wird. Es empfiehlt sich, zuerst die Stoppwörter auszublenden, d. h. mit eingeschalteter Stoppwortliste zu suchen und erst – bei Bedarf – in einem zweiten Schritt mit eingeblendeten Stoppwörtern (d. h. mit ausgeschalteter Stoppwortliste).

Ergebnisanzeige und Kategorien

Nachdem auf diese Weise die Darstellung des Ergebnisses eingestellt ist, erhält man eine tabellarische Auflistung der analysierten Kookkurrenz mit nachfolgender Untergliederung.

Im unteren Bildabschnitt zeigt sich eine Auswahl der Quellenbelege für das Suchwort Perikles. Hier hat man die Möglichkeit sich alle Quellenbelege anzusehen und herunterzuladen. Dafür ist die Schaltfläche „Zeige alle“ zu wählen. Daraufhin öffnet sich ein neues Fenster mit der gesamten Quellentabelle, einer weiteren Suchmaske und den Möglichkeiten, die Tabelle zu exportieren.

Die Ergebnisanzeige wird zunächst als tabellarische Aufschlüsselung des Ergebnisses in mehreren Unterkategorien mit einer Vielzahl an Wörtern und Kookkurrenzen präsentiert. Im unteren Bildabschnitt schließt sich eine Tabelle mit allen Quellen- und Belegstellen der Textpassagen an, in denen das Suchwort Perikles vorkommt. Doch zunächst zum tabellarischen Ergebnis: Im Kopf der Tabelle steht ein Hinweis zu dem Suchwort mit Word-ID und Trefferanzahl. Unterhalb des Tabellenkopfes liegt nun das Ergebnis nach folgenden Kategorien aufgeschlüsselt vor:

- Ausgeblendete Stoppwörter: Keine Verlinkung.
- Wörter mit derselben normalisierten Form: Verlinkung zu neuer Suche (Wort).
- Wörter mit ähnlichem Zusammenhang: Verlinkung zu neuer Suche (Wort).

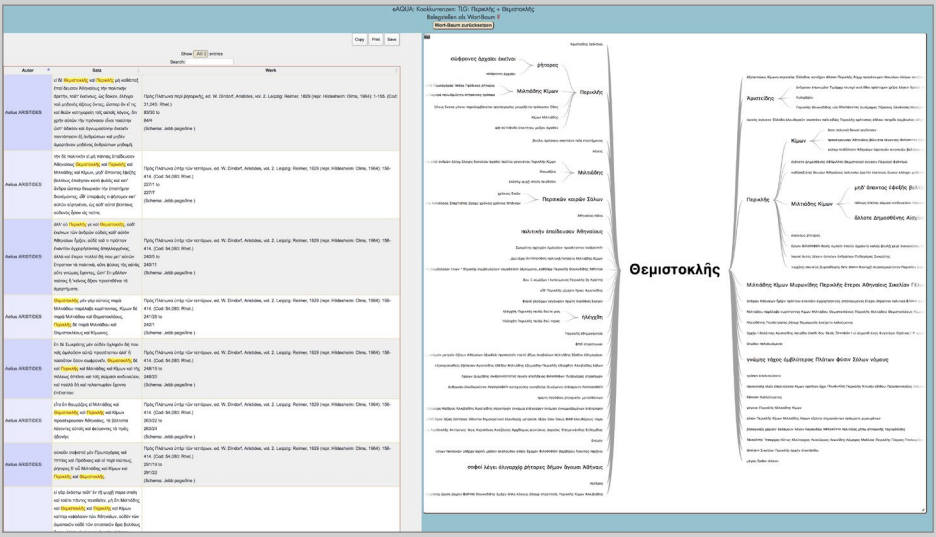


Abbildung 12. Belegstellen und Wortbaum der Kookkurrenz Perikles und Themistokles

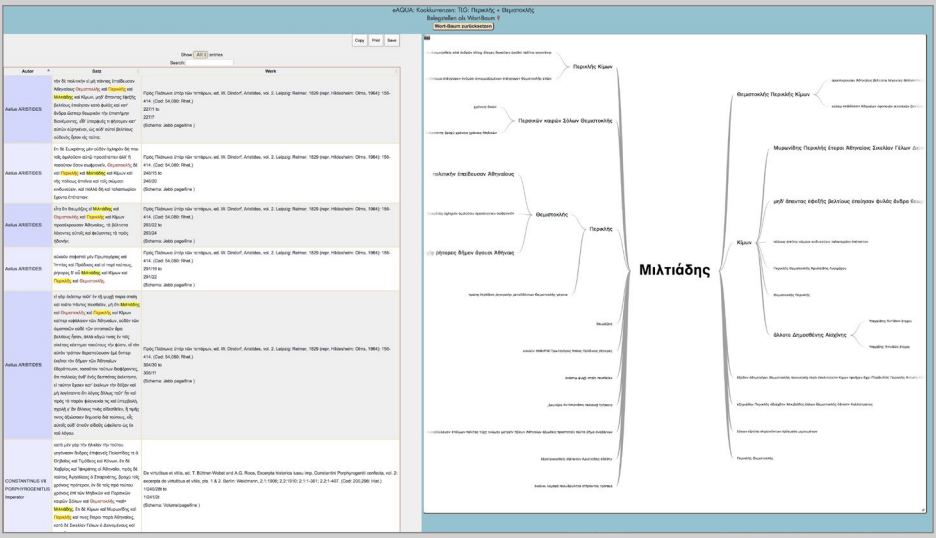


Abbildung 13. Belegstellen und Wortbaum der Kookkurrenz Perikles und Miltiades

- Signifikante Kookkurrenzen: Verlinkung zu neuer Suche (Wort), Verlinkung Belegstellen mit Wortbaum (Zahl), Verlinkung zur Netzwerkvisualisierung (Schaltflächen).
- Signifikante linke Kookkurrenzen: Verlinkung zu neuer Suche (Wort), Verlinkung Belegstellen mit Wortbaum (Zahl).
- Signifikante rechte Kookkurrenzen: Verlinkung zu neuer Suche (Wort), Verlinkung Belegstellen mit Wortbaum (Zahl).
- Signifikante linke Nachbarn: Verlinkung zu neuer Suche (Wort), Verlinkung Belegstellen mit Wortbaum (Zahl).
- Signifikante rechte Nachbarn: Verlinkung zu neuer Suche (Wort), Verlinkung Belegstellen mit Wortbaum (Zahl).
- Auswahl an Belegstellen: Sämtliche Quellenstellen, in denen das Suchwort vorkommt.

Belegstellen und Wortbaum

Sobald man die neben der Kookkurrenz in Klammern stehende Zahl anklickt, öffnet sich ein neues Fenster mit einer tabellarischen Aufstellung aller Quellenbelege, in denen das Suchwort Perikles mit der Kookkurrenz gemeinsam auftritt. Im Beispiel wurde auf die Zahl 46 neben Themistokles geklickt (■ **Abbildung 12**). Beide Suchwörter sind eingefärbt, um diese im Text schneller auszumachen. Neben der Tabelle befindet sich im rechten Bildabschnitt ein „Wortbaum", der die Position der Kookkurrenz als Satzbestandteil innerhalb der Quellenstellen darstellt und gleichzeitig einen Überblick über diese Quellenstellen liefert.

Dabei ist zu beachten, dass jedes Wort innerhalb des Wortbaumes wiederum eine verlinkte Kookkurrenz darstellt, die mit dem Suchwort Perikles verbunden ist. Mit einem Klick auf die entsprechende Kookkurrenz richten sich sowohl der Wortbaum als auch die Quellentabelle nach dieser Neuwahl aus.

Alle Kookkurrenzen sind mit einer neuen Suchfunktion verknüpft. Das heißt, wenn eine Kookkurrenz angeklickt wird, öffnet dies eine neue Suchmaske für das angeklickte Wort, das aber nicht im Zusammenhang mit der bereits durchgeführten Suche steht, sondern eine neue Suche darstellt. Die Beziehung zu dem ersten Suchwort lässt sich darstellen, indem man auf die Trefferanzahl klickt. Für die Kookkurrenz Themistokles bedeutet dies also 46 Belegstellen, die in direktem Zusammenhang zu dem Suchwort Perikles stehen. Innerhalb der Quellentabelle ist es möglich, nach bestimmten Worten zu suchen. Dies erfolgt über die Suchmaske im Tabellenkopf.

Mit der Auswahl einer neuen Kookkurrenz aus dem Wortbaum richten sich Quellentabelle und Wortbaum nach dieser neuen Kookkurrenz aus und ergeben eine neue Quellenausgabe nach der Kombination Perikles und Miltiades, sowie einen neuen Wortbaum mit Miltiades im Zentrum (■ **Abbildung 13**).

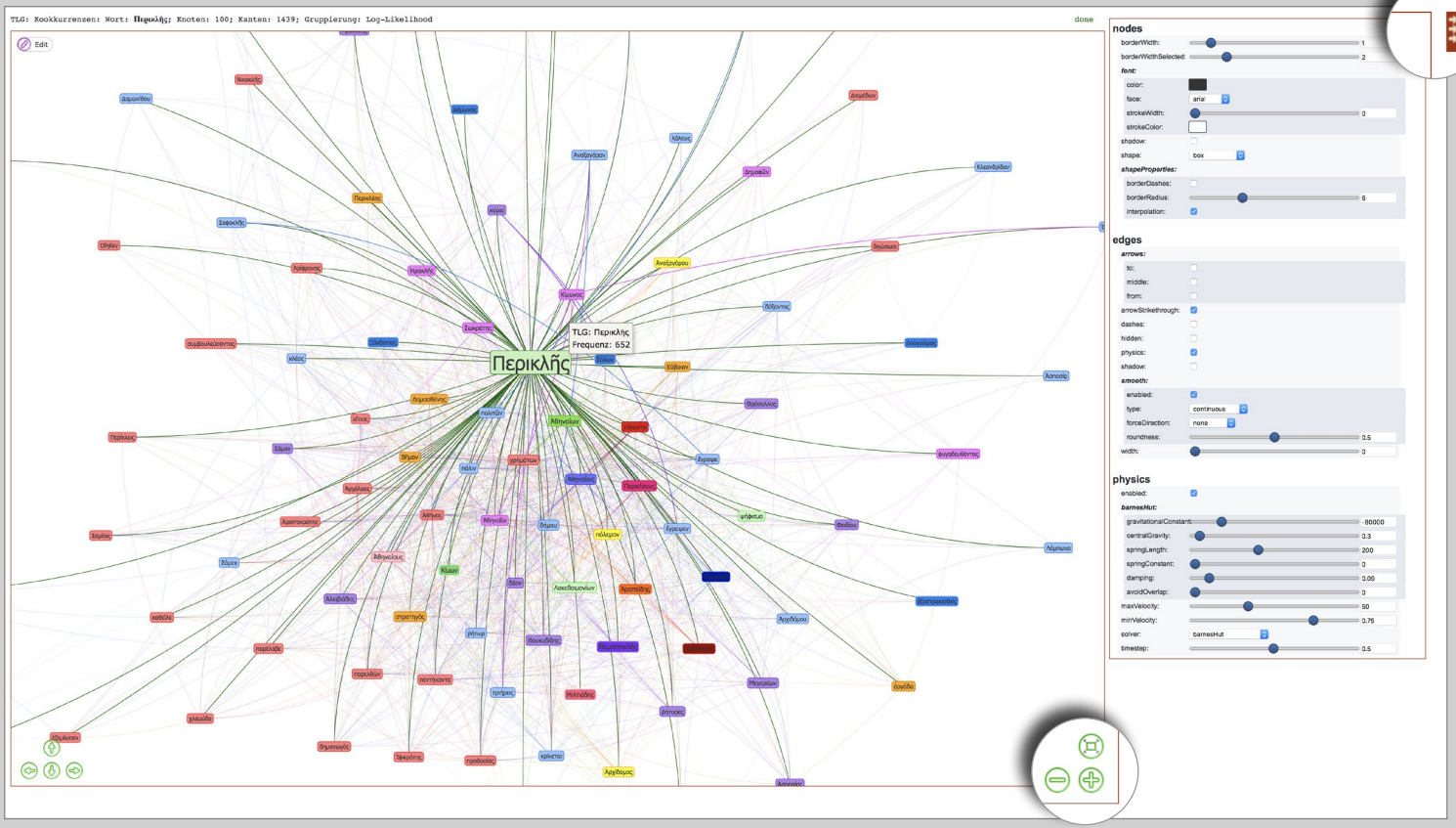


Abbildung 14. Kookkurrenzgraph zu Perikles

Kookkurrenzgraph

Über die Schaltflächen „Network [Frequenz]“ bzw. „Network [Log-Likelihood]“ gelangt man von der rein tabellarischen Darstellung in die graphische Visualisierung. Mit der Auswahl einer der beiden Graphenvarianten, die sich lediglich in der Art der visuellen Gruppierung der Knoten unterscheiden, öffnet sich ein neues Fenster mit einem sich bewegenden Graphen (■ **Abbildung 14**).

Im rechten Bildabschnitt öffnet sich zudem über die Pfeilschaltfläche eine Bildbearbeitungsleiste mit verschiedenen Werkzeugen, um die Darstellung des Graphen zu beeinflussen. Die Bearbeitungsmöglichkeiten sind dabei vielfältig und einmal in die darstellende sowie in die berechnende Methode zu unterscheiden. Darstellerische Veränderungen beziehen sich hauptsächlich auf die Farbgebung von Kanten und Knoten, Schriftgröße, Feldfenster etc. Durch die Berechnungsmethode besteht die Möglichkeit, den Graphen neu zu visualisieren. So können sich mit Variationen von

- force Direction
- Type
- Solver
- Physics

mehrere Graphen mit der gleichen Datenausgangsbasis ergeben.

Sollte der Graph einmal zu groß werden und über das Fenster hinauswachsen, kann er mittels der Zoomfunktion in der rechten, unteren Ecke wieder justiert werden.

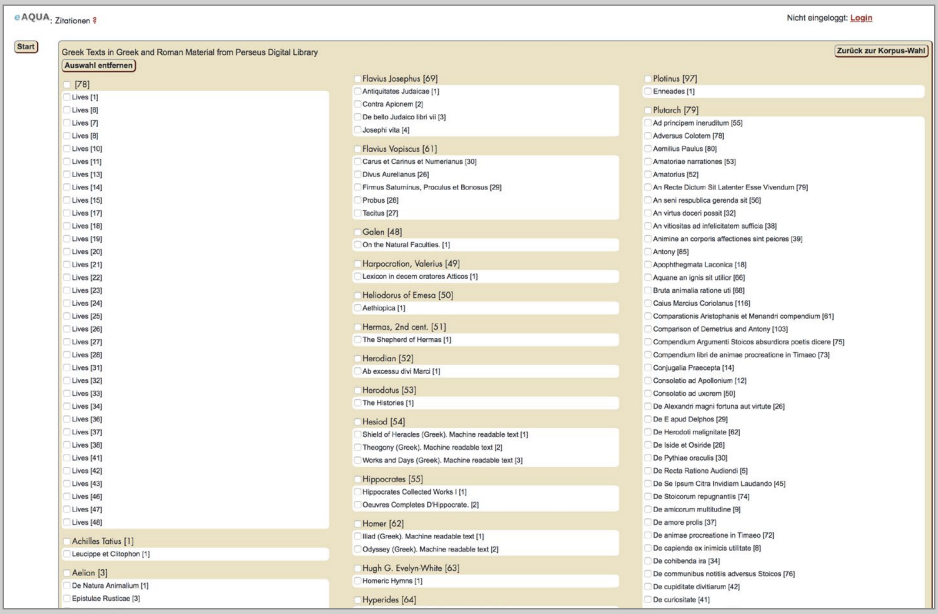


Abbildung 15. Werkauswahl bei der Parallelstellensuche

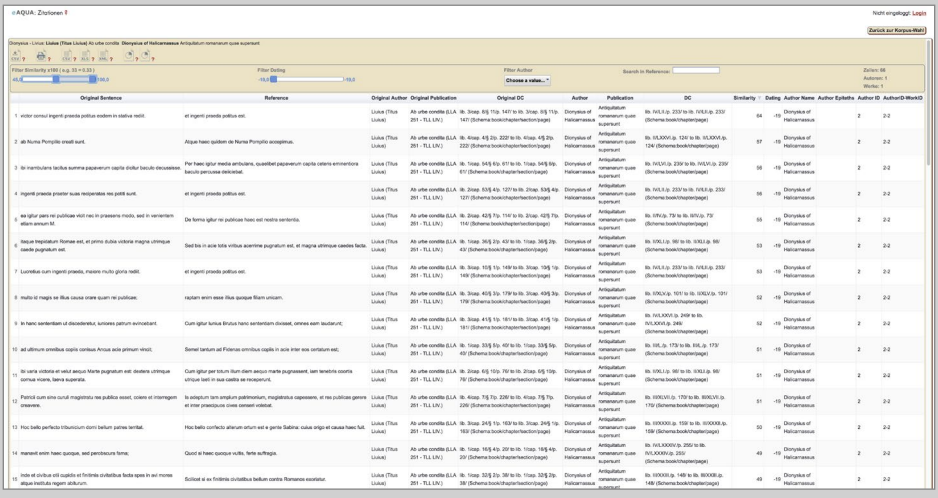


Abbildung 16. Parallelstellenanzeige in Tabellenform

Demonstration Zitation

Die Demonstration Zitationserkennung (Parallelstellensuche) ist über den Menü-Punkt Tools erreichbar. Sie beinhaltet neben den Analyse-Ergebnissen des Projekts einige zusätzliche Korpora, die während oder nach der zweiten Förderphase des eAQUA-Projektes auch zu Lehrzwecken angelegt wurden.

Für den Login in den geschützten Bereich, in dem Ergebnisse von Korpora abgerufen werden können, die an Benutzungslizenzen gebunden sind, ist am rechten Bildschirmrand ein Link Login vorgesehen.

Der Auswahlprozess erfolgt in drei Schritten: Korpus auswählen – Autor(en) und Werk(e) anhaken – auf die Schaltfläche Start ganz links im Fenster klicken (■ **Abbildung 15**). Bei den als Subkorpora bezeichneten Auswahlmöglichkeiten sind alle Werke eines Autors zusammengefasst, so dass hier die beiden letzten Schritte entfallen und sogleich mit dem Laden der Daten begonnen wird. Insbesondere bei diesen Subkorpora ist zu beachten, dass die Ladezeiten der Tabellenansicht sehr lange dauern können. Zum Beispiel hat das Subkorpus ARISTOTELES et CORPUS ARISTOTELICUM eine Ergebnismenge in der Größe von ca. 180 MB. Diese müssen sowohl vom Server an den Browser übertragen werden, was abhängig von der Anbindung des Nutzers ist, und darüber hinaus vom Browser weiterverarbeitet und dargestellt werden, was gelegentlich auch zum Absturz des Browsers führen kann, wie es besonders bei älteren Systemen oder exotischen Konstellationen beobachtet wurde. Aber auch schon bei geringeren Ergebnismengen können Wartezeiten von mehreren Minuten entstehen. Hierbei ist etwas Geduld gefragt. Nachdem in der Tabellenansicht die Ergebnisse der Parallelstellensuche geladen sind, ergeben sich verschiedene Filter- und Sortiermöglichkeiten (■ **Abbildung 16**).

Wie die gefilterten Daten exportiert werden können, ist weiter unten beschrieben. Bei größeren Ergebnismengen ist nur der Export in das Format CSV zu empfehlen, weil hier die Daten direkt vom Browser an den Client geschickt werden können. Sowohl bei XML als auch XLS werden die Listen vom Browser nochmals an den Server zur Umwandlung transportiert und von dort wieder geladen, wodurch die Ausführungszeit erheblich verlängert wird.

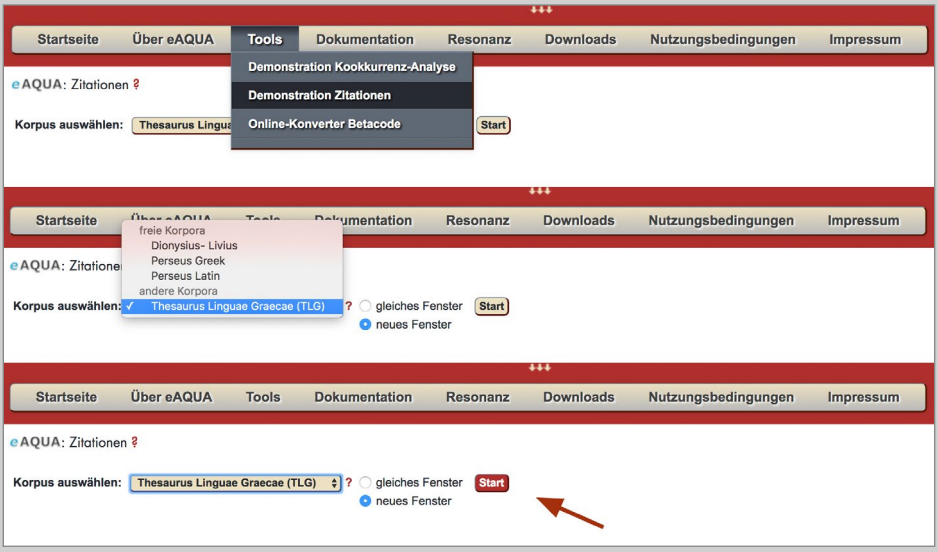


Abbildung 17. Zitation – Auswahl TLG-E

Beispielbenutzung Schritt für Schritt

Korpusauswahl

Nach dem Einloggen – für einige Korpora sind Zugangsdaten notwendig – wird unter dem Reiter „eAQUA“ das Tool „Demonstration Zitationen“ und – für die folgenden Beispiele des Arbeitsweges – die Texte der Textdatenbank Thesaurus Linguae Graecae (TLG-E) als Ausgangsbasis für die folgende Analyse griechischer Autoren und Werke der Antike ausgewählt. Dieser erste Schritt ist für die Tools „Zitationen“ und „Kookkurrenz – Analyse“ gleich (■ **Abbildung 17**).

Nach der Auswahl des Korpus öffnet sich ein neues Fenster mit einer Liste aller Autoren und der zugehörigen Werke. Außerdem sind zu jedem Autor und jedem Werk die entsprechenden IDs aus der Datenbank verzeichnet. Über die Schaltfläche „Zurück zur Korpus-Wahl“ gelangt man wieder zurück zum ersten Schritt, in dem zwischen den verschiedenen Datenbanken gewählt werden kann. Durch Anwählen der Auswahlkästchen besteht die Möglichkeit, ein spezifisches Analysekorpus zusammen zu stellen.

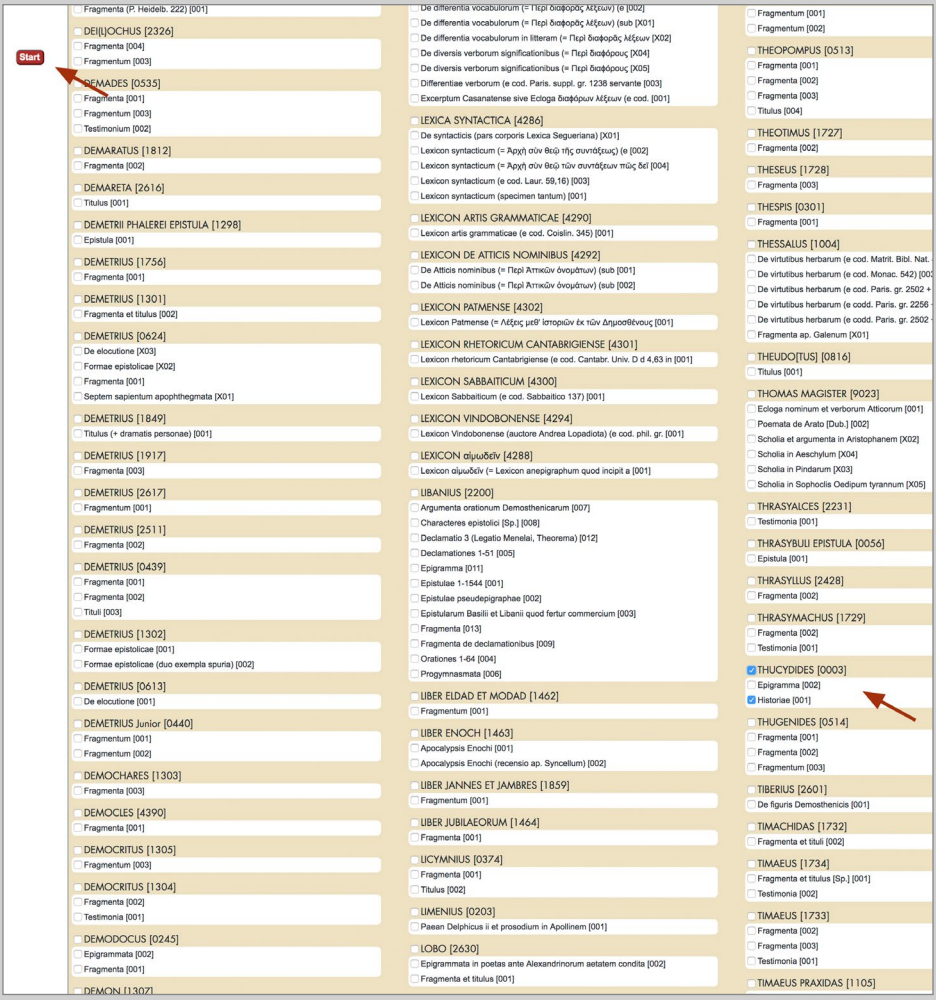


Abbildung 18. Zitation – Werkauswahl Thukydides Historien

Werkauswahl

Es können Autoren und Werke frei zusammengestellt werden. Dabei ist zu beachten, dass die Zusammenstellung von der Ausgangsfrage abhängig ist. Die Praxis hat jedoch gezeigt, dass es sinnvoller ist, sich zuerst auf einen Autor zu beschränken und dafür mehrere Einzelanalysen durchzuführen, und, sofern man mehrere Autoren untersuchen will, dafür mehrere Schritte vorzusehen. Auch bei der Analyse des Gesamtwerkes eines Autors empfiehlt es sich, mehrere Einzelwerke nacheinander zu betrachten, da die Datenmenge in manchen Fällen sehr umfangreich sein kann.

Mit der Bestätigung der Werkauswahl über die Start-Schaltfläche beginnt das Programm mit der Analyse, wobei das zu untersuchende Korpus im Verhältnis zum gesamten Textbestand des gewählten Korpus hin untersucht wird (■ **Abbildung 18**).

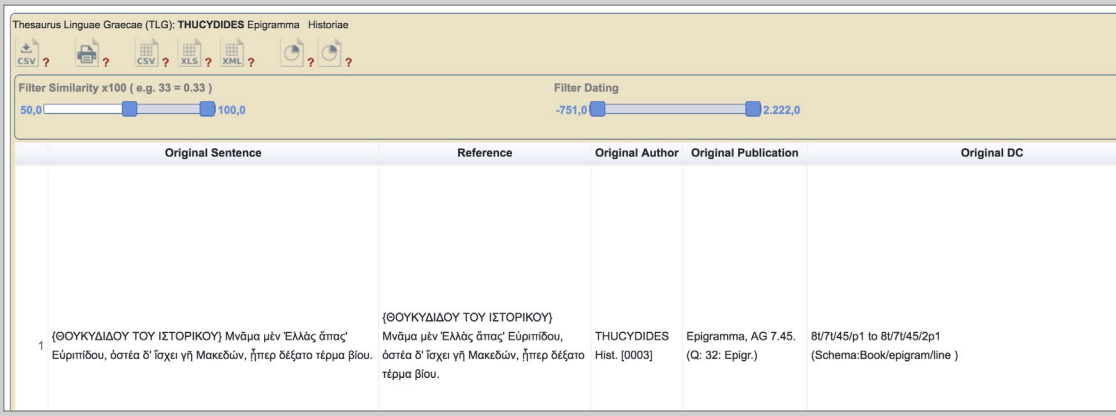


Abbildung 19. Zitation Ergebnistabelle Thukydides

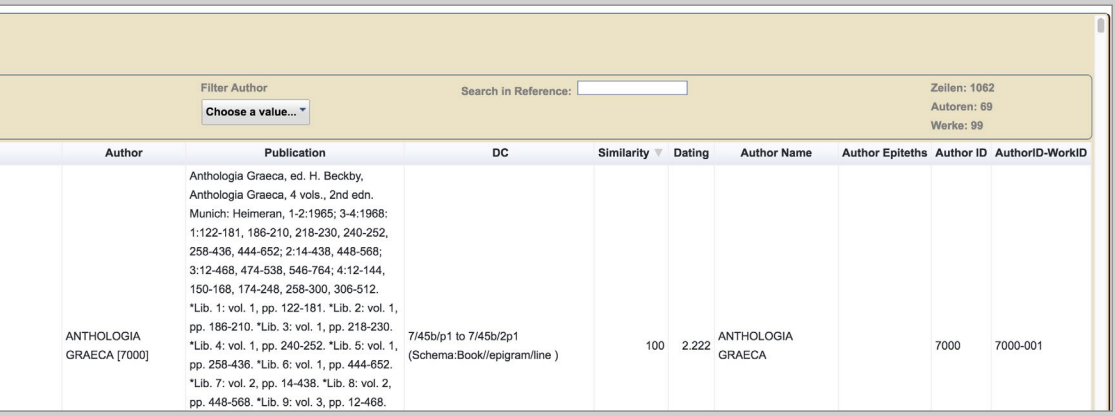
Ergebnistabelle

Nachdem die Auswahl bestätigt ist, beginnt das Tool mit der Analyse und liefert alle gefundenen Ergebnisse als durchnummerierte Ergebnistabelle (■ **Abbildung 19**). Die Ergebnisse sind in der Standardeinstellung nach den errechneten Similaritätswerten absteigend sortiert, beginnend mit dem Wert 100 (100-prozentige Übereinstimmung).

Im Folgenden kann die Tabelle im Hinblick auf die jeweilige Fragestellung sortiert und nach bestimmten Einstellungen ausgerichtet werden. Dazu besteht die Möglichkeit, über die entsprechende Auswahl innerhalb der Tabelle das Ergebnis nach den jeweiligen Spalten alphabetisch zu sortieren oder nach den Einstellungsmöglichkeiten im Tabellenkopf zu ordnen. In der Standardausgabe sind diese Bereiche bereits vorgegeben mit einer Similarity von 100 bis 60, ohne zeitliche Einschränkung und ohne Auswahl an Referenzautoren. Im folgenden Schritt werden die weiteren Einstellungen erklärt.

Original Sentence

In der Spalte Original Sentence werden alle vom Programm berechneten Originaltextstellen des gewählten Subkorpus in fortlaufender Nummerierung aufgeführt. Diesen Textstellen werden Reference Textstellen zugeordnet, die sich aus dem Vergleich des Originals mit dem gesamten Textbestand des Korpus ergeben. Sofern sich bei der Analyse eine Übereinstimmung von fünf aufeinander folgenden Wörter ergibt, wird die Textstelle als Ergebnis angegeben. Dieser Kategorie sind die Spalten: „Original Author", „Original Publication" und „Original DC" zugeordnet, wobei erstere den Autor des ausgewählten Korpus nennt, gefolgt von der dem Textbestand zugrundeliegenden Edition. Letzte gibt den entsprechenden Quellenverweis innerhalb der Edition wieder.



Filter Author: Choose a value...
Search in Reference:
Zeilen: 1062
Autoren: 69
Werke: 99

Author	Publication	DC	Similarity	Dating	Author Name	Author Epiteths	Author ID	AuthorID-WorkID
ANTHOLOGIA GRAECA [7000]	Anthologia Graeca, ed. H. Beckby, Anthologia Graeca, 4 vols., 2nd edn. Munich: Heimeran, 1-2:1965; 3-4:1968: 1:122-181, 186-210, 218-230, 240-252, 258-436, 444-652; 2:14-438, 448-568; 3:12-468, 474-538, 546-764; 4:12-144, 150-168, 174-248, 258-300, 306-512. *Lib. 1: vol. 1, pp. 122-181. *Lib. 2: vol. 1, pp. 186-210. *Lib. 3: vol. 1, pp. 218-230. *Lib. 4: vol. 1, pp. 240-252. *Lib. 5: vol. 1, pp. 258-436. *Lib. 6: vol. 1, pp. 444-652. *Lib. 7: vol. 2, pp. 14-438. *Lib. 8: vol. 2, pp. 448-568. *Lib. 9: vol. 3, pp. 12-468.	7/45b/p1 to 7/45b/2p1 (Schema:Book//epigram/line)	100	2.222	ANTHOLOGIA GRAECA		7000	7000-001

References

Neben dem reinen Textbestand werden die Namen der Original- und Referenzautoren mit entsprechender TLG-Nummer angegeben.

Dieser Kategorie sind die Spalten: „Author“, „Publication“, „DC“, „Dating“, „Author Name“, „Author Epithes“, „Author ID“ und „Author Work-ID“ zugeordnet. Dem Referenzautor werden editorische Hinweise zugefügt sowie die dem Textbestand zugrundeliegende Edition und der entsprechende Quellenverweis innerhalb der Edition. Die Datierung bezieht sich auf die Lebensdaten des Referenzautors und kann demnach schwanken.[13] Ist eine Datierung nicht möglich, wird ein Fantasiewert von 2222 eingesetzt, da technisch bedingt eine Zahl in dem entsprechenden Feld eingetragen sein muss und die Null dafür ja nicht infrage kommt.

Search in Reference

Über die Wortsuche kann die gesamte Ergebnistabelle durchsucht werden.

Similarity

Bei der Similarity handelt es sich um die prozentuale Übereinstimmung von Original- und Referenztext, angegeben in ganzen Zahlen von 1 bis 100. Man kann die Similarity über einen Regelschalter beliebig einstellen. Aus der Praxis hat sich der Bereich von 100 bis 60 als effizient erwiesen, daher ist dieser Bereich voreingestellt.

Das Tool ist darauf hin konzipiert, Übereinstimmungen von zwei Texten zu finden, die in fünf Wörtern (Token) übereinstimmen und beschreibt damit die

13 Die Datierung ist ein Schätzwert, der sich aus den Lebensdaten der Autoren ergibt. Falls die Lebensdaten bekannt sind, wird die ungefähre Lebensmitte als Zeitpunkt errechnet, da der genaue Entstehungszeitpunkt zumeist nicht überliefert ist.

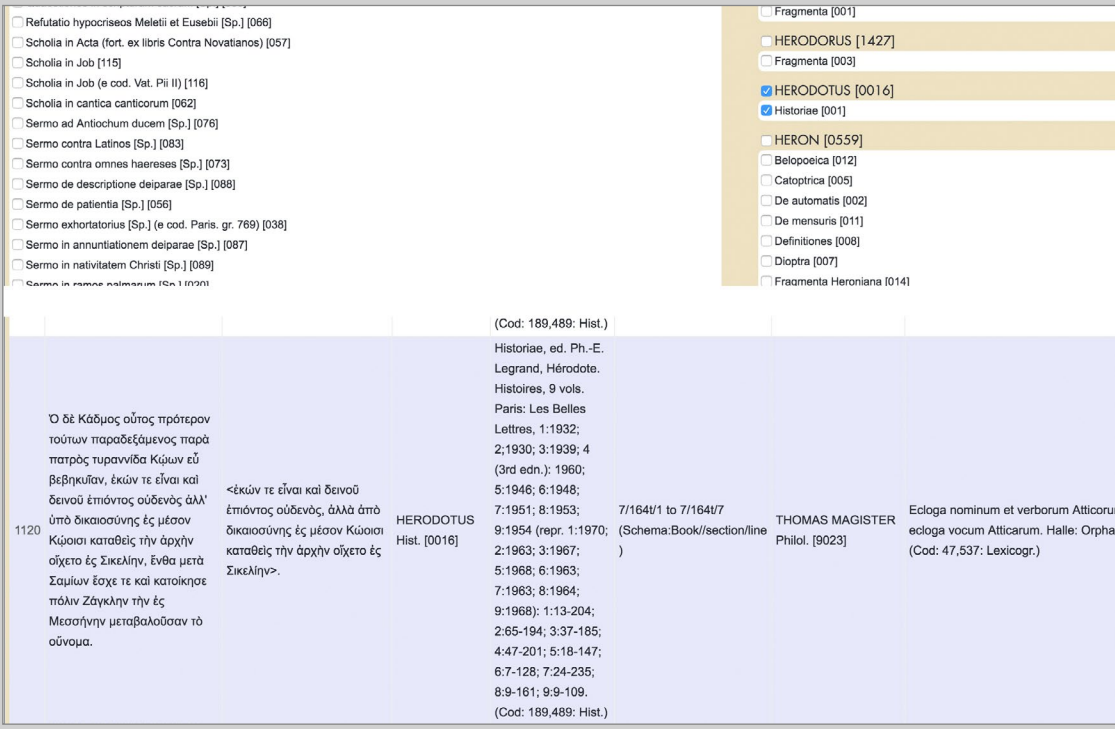
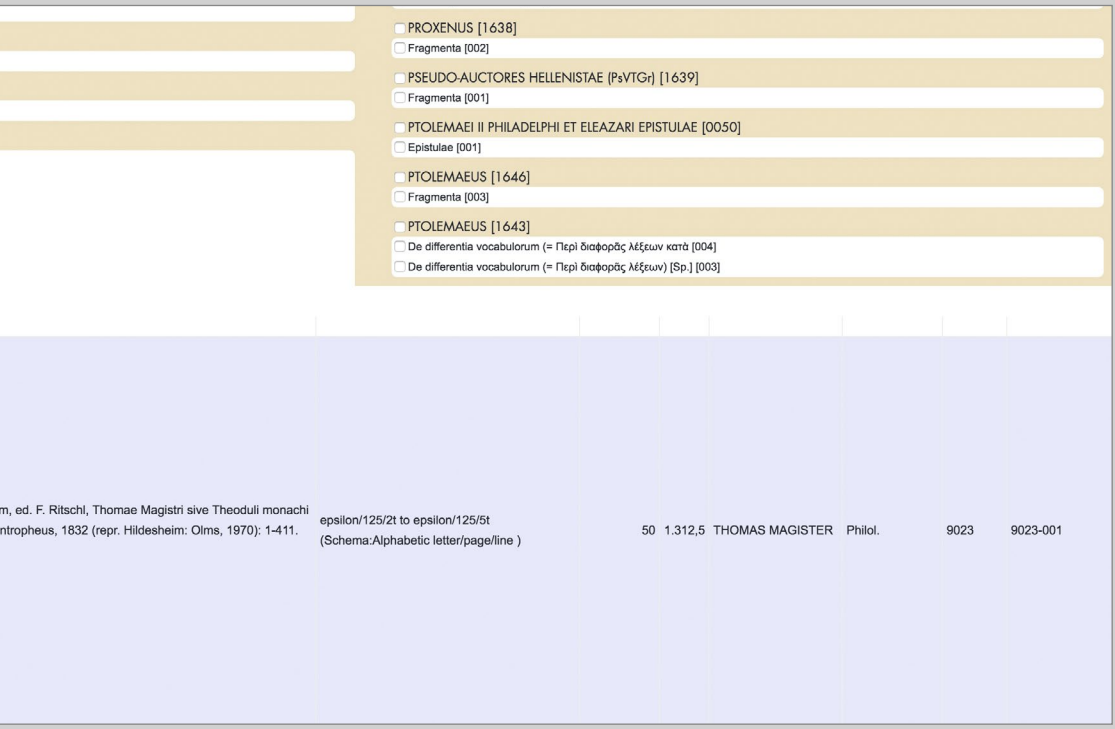


Abbildung 20. Zitation Herodot Historien: Ergebnis Nummer 1120

Übereinstimmung von „Original Sentence“ zu „Reference“. Im Falle, dass zwei ungleich lange Textstellen miteinander in Beziehung stehen, kann die Similarity zu uneindeutiger Aussagekraft führen. Daher ist es unumgänglich, auch bei geringen Similaritywerten die Textstellen genau zu analysieren. Ein solcher Fall soll anhand eines Suchkorpus „Herodot“ verdeutlicht werden.

Beispiel Herodot Historien 7.164.1-7

Das Ergebnis Nummer 1120 im Korpus Herodot (TLG – Herodotus [0016] – Historiae [001]) beschreibt das Verhältnis zwischen der Herodotstelle 7.164.1–7 und dem Referenzwerk „Ecloga nominum et verborum Atticorum“ des Theodoulos Monarchos, genannt Thomas Magister (■ **Abbildung 20**). Errechnet wurde eine Similarity von 50. Die Similarity bietet also keine eindeutige Aussage über den Wert der Stelle als Parallelstelle, die Übereinstimmung der verglichenen Textpassagen liegt bei 50 Prozent. Betrachtet man nun die Ausgangstextpassage, ergibt sich jedoch ein anderes Bild: der gesamte Text der Referenzstelle ist in dieser als Nebensatz zu finden, was wiederum den Schluss nahelegt, dass der Referenzautor nicht den



gesamten Referenztext als Einheit übernommen, sondern ihn als Einzelsatz neu strukturiert hat.

Ὁ δὲ Κάδμος οὗτος πρότερον τούτων παραδεξάμενος παρὰ πατρὸς τυραννίδα Κῴων εὖ βεβηκυῖαν, ἑκών τε εἶναι καὶ δεινοῦ ἐπιόντος οὐδενὸς ἀλλ' ὑπὸ δικαιοσύνης ἐς μέσον Κῴοισι καταθεὶς τὴν ἀρχὴν οἴχετο ἐς Σικελίην, ἔνθα μετὰ Σαμίων ἔσχε τε καὶ κατοίκησε πόλιν Ζάγκλην τὴν ἐς Μεσσήνην μεταβαλοῦσαν τὸ οὔνομα.	<ἑκών τε εἶναι καὶ δεινοῦ ἐπιόντος οὐδενὸς, ἀλλὰ ἀπὸ δικαιοσύνης ἐς μέσον Κώοισι καταθεὶς τὴν ἀρχὴν οἴχετο ἐς Σικελίην>.

Um den Wert der Übereinstimmung zu überprüfen ist ein Blick in den Referenztext nötig. In diesem leitet der Autor die Referenzstelle mit den Worten Ἡρόδοτος ἐν Πολυμνίᾳ ein und gibt damit klar zu erkennen, dass es sich um ein direktes Zitat aus Herodots siebtem Buch (benannt nach der Muse Polymnia) handelt, welches aus dem Text extrahiert wurde.

Herodot. Historien 4.133		Athenaeus. Deipnosophistae 4.23.	
133.1-2	Ἡμέρην δὲ ἁπασέων μάλιστα ἐκείνην τιμᾶν νομίζουσι τῇἕκαστος ἐγένετο.	[Ἡρόδοτος δὲ συγκρίνων τὰ τῶν Ἑλλήνων συμπόσια πρὸς τὰ παρὰ Πέρσαις φησίν·]	
133.2-3	Ἐν ταύτῃ δὲ πλέω δαῖτα τῶν ἀλλέων δικαιοῦσι προτίθεσθαι	‘ἡμέρην δὲ Πέρσαι ἁπασέων μάλιστα ἐκείνην τιμᾶν νομίζουσι τῇ ἕκαστος ἐγένετο.	4.23.2-4
133.3-6	[ἐν τῇ οἱ εὐδαίμονες αὐτῶν βοῦν καὶ ἵππον καὶ κάμηλον καὶ ὄνον προτιθέαται ὅλους ὀπτοὺς ἐν καμίνοισι,] οἱ δὲ πένητες αὐτῶν τὰ λεπτὰ τῶν προβάτων προτιθέαται.	ἐν ταύτῃ δὲ πλέω δαῖτα τῶν ἄλλων δικαιεῦσι προτίθεσθαι· [ἐν τῇ οἱ εὐδαίμονες αὐτῶν βοῦν καὶ ὄνον καὶ ἵππον καὶ κάμηλον προτιθέαται ὅλους ὀπτοὺς ἐν καμίνοις·]	4.23.4-5
1.133.6-7	Σίτοισι δὲ ὀλίγοισι χρέωνται, ἐπιφορήμασι δὲ πολλοῖσι καὶ οὐκ ἁλέσι	οἱ δὲ πένητες αὐτῶν τὰ λεπτὰ τῶν προβάτων προτίθενται.	4.23.7-8
1.133.7-10	καὶ διὰ τοῦτό φασι Πέρσαι τοὺς Ἕλληνας σιτεομένους πεινῶντας παύεσθαι, ὅτι σφι ἀπὸ δείπνου παραφορέεται οὐδὲν λόγου ἄξιον, → Zitationsbruch durch Satzzeichen. εἰ δέ τι παραφέροιτο, ἐσθίοντας ἂν οὐ παύεσθαι.	σίτοισί τε ὀλίγοισι χρέονται, ἐπιφορήμασι δὲ πολλοῖσι καὶ οὐκ ἁλέσι.	4.23.8-9
		καὶ διὰ τοῦτό φασι Πέρσαι τοὺς Ἕλληνας σιτεομένους πεινῶντας παύεσθαι, ὅτι σφίσιν ἀπὸ δείπνου παραφορέεται οὐδὲν λόγου ἄξιον· → Zitationsbruch durch Satzzeichen.	4.23.9-12
1.133.12-13	[Οἴνῳ δὲ κάρτα προσκέαται. Καί σφι οὐκ ἐμέσαι ἔξεστι, οὐκὶ οὐρῆσαι ἀντίον ἄλλου. Ταῦτα μέν νυν οὕτω φυλάσσεται. Μεθυσκόμενοι δὲ ἐώθασι βουλεύεσθαι τὰ σπουδαιέστατα τῶν πρηγμάτων.]	εἰ δέ τι παραφέροιτο, ἐσθίοντας ἂν οὐ παύεσθαι.	4.23.12-13
		[Οἴνῳ δὲ κάρτα προσκέαται· καί σφιν οὐκ ἐμέσαι ἔξεστιν, οὐκ οὐρῆσαι ἀντίον ἄλλου. ταῦτα μέν νυν οὕτω φυλάσσεται. μεθυσκόμενοι δὲ εἰώθασι βουλεύεσθαι τὰ σπουδαιότατα τῶν πρηγμάτων·]	4.23.13-16
1.133.14-16	Τὸ δ’ ἂν ἄδῃ σφι βουλευομένοισι, τοῦτο τῇ ὑστεραίῃ νήφουσι προτιθεῖ ὁ στέγαρχος, ἐν τοῦ ἂν ἐόντες βουλεύωνται·	τὸ δ’ ἂν ἄδῃ σφίσι βουλευομένοισι, τοῦτο τῇ ὑστεραίῃ νήφουσι προτιθεῖ ὁ στεγέαρχος ἐν τοῦ ἂν ἐόντες βουλεύωνται.	4.23.17-19
1.133.16-17	καὶ ἢν μὲν ἄδῃ καὶ νήφουσι, χρέωνται αὐτῷ, ἢν δὲ μὴ ἄδῃ, μετιεῖσι	καὶ ἢν μὲν ἄδῃ καὶ νήφουσι, χρέονται αὐτῷ· εἰ δὲ μή, μετιεῖσιν.	4.23.19-20
1.133.17-18	τὰ δ’ ἂν νήφοντες προβουλεύσωνται, μεθυσκόμενοι ἐπιδιαγινώσκουσι.	τὰ δ’ ἂν νήφοντες προβουλεύσωνται, μεθυσκόμενοι ἐπιδιαγινώσκουσι.	4.23.20-21

Tabelle 1. Original Sentence Herodot I. 133.1-25 und Referenzwerk Athenaeus Deipnosophistae

Ebenso kann es vorkommen, dass einzelne Ergebnisse ganze zusammenhängende Textstellen oder Passagen sind, die jedoch aufgrund der Berechnungsgrundlage als einzelne Ergebnisse mit unterschiedlichen Similaritywerten angeführt werden. Da das Tool innerhalb von Satzeinheiten rechnet, können solche Passagen nur eindeutig gefunden werden, wenn man die Tabelle nach „Original DC" vorsortiert und somit die Ergebnisse nach dem Aufbau des Werkes gelistet werden. Ein Beispiel soll einen solchen Fall verdeutlichen.

Einschätzung der Fundstelle

In Kapitel 133 des ersten Buches der Historien beschreibt Herodot die Tischsitten von Griechen und Persern am Beispiel, wie die Perser den Geburtstag feiern und stellt diese vergleichend gegenüber. Athenaeus greift diesen Passus in Kapitel 23 des vierten Buches des Gelehrtenmales auf und übernimmt es gewissermaßen wortwörtlich.

Anhand dieser Textpassage (■ **Tabelle 1**) lassen sich mehrere Aussagen über die Arbeitsweise der Suche nach Parallelstellen festmachen. Zum einen wird nicht der ganze Passus als Parallele und mögliches Zitat angezeigt, da das Tool von Satzzeichen zu Satzzeichen analysiert und daher das Kapitel auch in Sätze zerlegt wird. Diese Sätze wiederum werden mit den Referenzsätzen in Beziehung gesetzt und mit einem Similaritywert bestimmt. Aufgrund minimaler Unterschiede liegt die Similarity im Durchschnitt bei 85, auch wenn dem Betrachter natürlich auffällt, dass die Texte quasi identisch sind. Allein die Groß- und Kleinschreibung macht aber einen Unterschied, wie in Hdt. I. 133.6-7 „Σίτοισι" und Ath. IV. 23.8-9 „σίτοισί" erkennbar wird. Auch die Übernahme einzelner Satzbestandteile ist entscheidend. So übernimmt Athenaeus zwar den gesamten Satz von Hdt. I. 133.7-10, teilt ihn aber in zwei Sätze auf; es entsteht ein „Bruch" und dies führt zu zwei Ergebnisanzeigen.

Ein weiteres Problem wird deutlich, wenn es sich bei den Originaltextstellen um solche mit weniger als fünf Worten handelt, da auf fünf aufeinanderfolgende Terme hin analysiert wird. So wird die Passage Hdt. I. 133.12-13 nicht erkannt. Es obliegt also in jedem Fall der philologisch-historischen Überprüfung und Auswertung, um ein Ergebnis zu erhalten, das in eine wissenschaftliche Arbeit integriert werden kann.

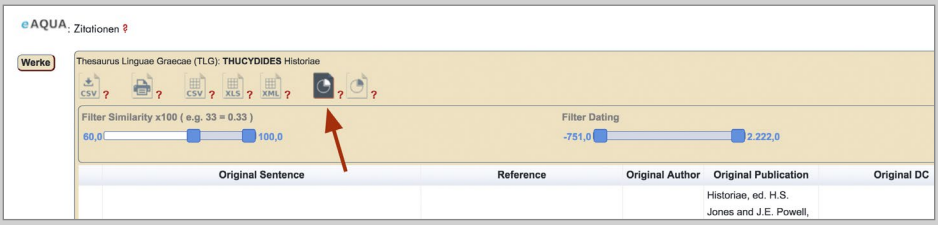


Abbildung 21. Chartview 1 aufrufen

Checkliste für Ergebnisanalyse

Da sich aufgrund der technischen Voraussetzungen und den anderen Zitier- und Verweisgepflogenheiten der antiken Autoren dem modernen Historiker eine Vielzahl an Herausforderungen stellt, sollte man folgende Checkliste bei der Analyse der Ergebnisse beherzigen:

Handelt es sich bei den Referenzstellen um:

1. Wörtliche Zitate?
2. Wörtliche Zitate innerhalb eines größeren Kontextes?
3. Inhaltliche Übernahmen und damit Bezugnahmen zwischen Autoren?
4. Abhängigkeiten zwischen Autoren und Werken?
5. Allgemeinen Sprachgebrauch (Sprichworte, etc.)?
6. Direkte Zitate eines Autors oder haben mehrere Autoren auf den gleichen Ausgangstext zurückgegriffen und beziehen sich darauf?
7. Einzelne Sätze, Satzbestandteile oder ganze Textpassagen, die übernommen werden?
8. Zwischen den Zeilen lesen: Sofern die Ergebnisse darauf hindeuten, dass ein Referenzautor mehrere zusammenhängende Textstellen übernommen hat, kann es sein, dass auch weiterer Text übernommen wurde, wenngleich nicht wörtlich und damit nicht als Zitat.
9. Gibt der Referenzautor möglicherweise einen Quellenbezug in einer vorangegangenen Textstelle an?

Chartview 1: Über zeitliche Auswahl und Autorenschaft zum Ergebnis

Von der ersten Ergebnisanzeige mittels Ergebnistabelle gelangt man zu zwei weiteren Lösungs- bzw. Arbeitswegen, um die Analyse weiter detailliert zu verfolgen. Im oberen Bereich der Tabelle stehen neben den Symbolen der Exportfunktion zwei Symbole mit Kreisdiagrammen, die in den Bereich Chartview 1 und Chartview 2 führen.

Über die Chartview 1 betritt man einen spezifischen Lösungsweg für die Analyse, und zwar ausgehend von der chronologischen Aufteilung des Ergebnisses über die Referenzautoren zu den Referenzstellen (■ **Abbildung 21**). Mit der Auswahl des ersten Symbols öffnet sich ein neues Fenster mit der Darstellung eines Balken- und eines Kreisdiagramms (■ **Abbildung 22,** *siehe Seite 44/45*).

Die Analyse beginnt mit der Entscheidung, ob man das Ergebnis auf einen bestimmten Zeitpunkt oder einen bestimmten Autor einschränken will. In der Chartview 1 kann man allerdings nur einen Autor und einen Zeitabschnitt auswählen. Die Auswahl auf mehrere Autoren ist über die Chartview 2 möglich, doch dazu später. Im folgenden Anwendungsbeispiel dient der Peloponnesische Krieg des Thukydides als Ausgangsbasis. Dazu wählt man in der Korpusauswahl „Thucydides“ „Historien“ aus.

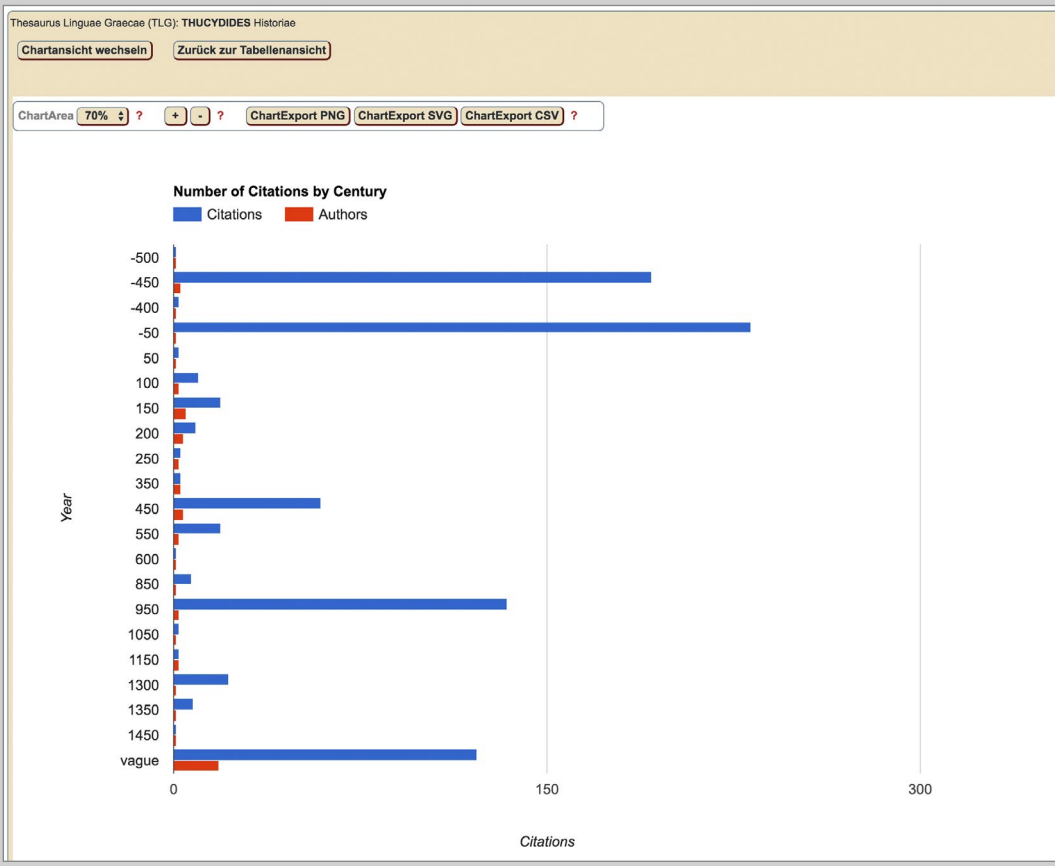
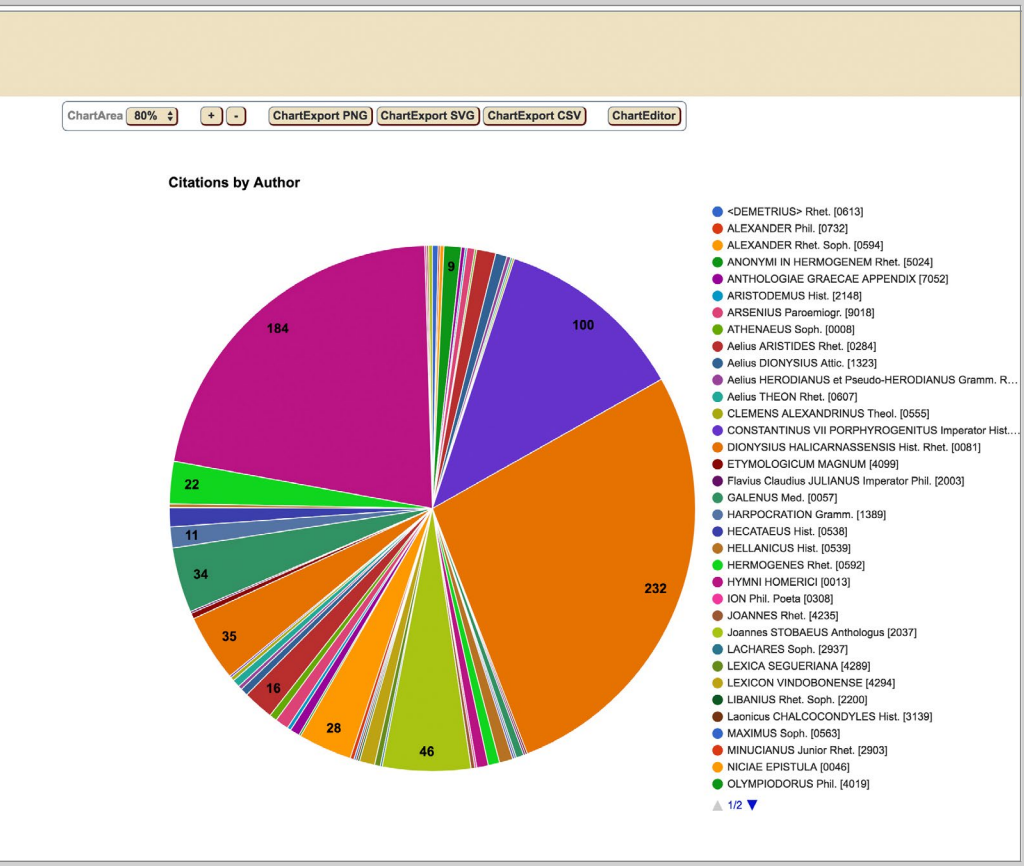


Abbildung 22. Chartview 1: Thukydides Historien in chronologischer Ordnung

Die Chartview 1 zeigt das gleiche Ergebnis wie die Tabellenansicht. Sollten also bereits im ersten Schritt die Einstellungen verändert worden sein, zum Beispiel die Datierung eingeschränkt, der Similaritywert erweitert oder eine Auswahl an Referenzautoren getroffen worden sein, so spiegelt sich dieses Ergebnis auch in der jeweiligen Chartview wieder. Während des Arbeitsprozesses kann man jedoch jederzeit wieder in die Tabelle oder die Chartview 2 zurückkehren und zwar über die Schaltflächen „Chartansicht wechseln“ und „Zurück zur Tabellenansicht“.

Chartview 1: Century Range

Im linken Bildabschnitt erscheint ein Balkendiagramm, das als Zeitleiste fungiert und in 50er Jahresabschnitte aufgeteilt ist. Die blauen Balken stellen die Anzahl



der gefundenen Referenzstellen dar, die roten beschreiben die Anzahl der Autoren. Die Referenzautoren können auch vor dem Schaffenszeitraum des Originalautors eingeordnet sein; diese sind als mögliche Quellen für den Autor in Betracht zu ziehen. Die Balken sind einzeln auswählbar und führen im nächsten Schritt zu einer weiteren Eingrenzung des Ergebnisses. Sofern man einen blauen oder roten Balken anklickt, wählt man einen Zeitraum mit möglicherweise mehreren Referenzautoren und Werken.

Chartview 1: Autorendiagramm

In der rechten Bildhälfte öffnet sich ein Kreisdiagramm, das die prozentuale Verteilung der Referenzautoren anzeigt. Eine fortlaufend alphabetisch sortierte Liste der Referenzautoren befindet sich rechts davon, mit der Möglichkeit, über die Pfeil-

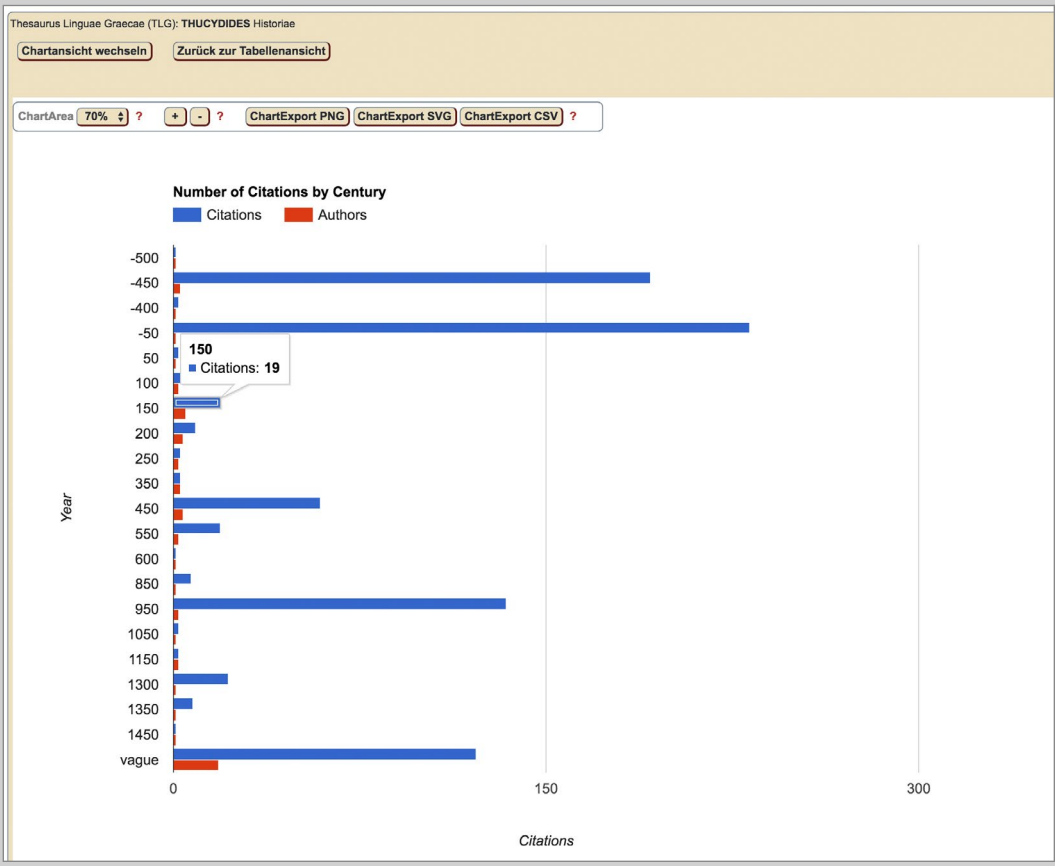
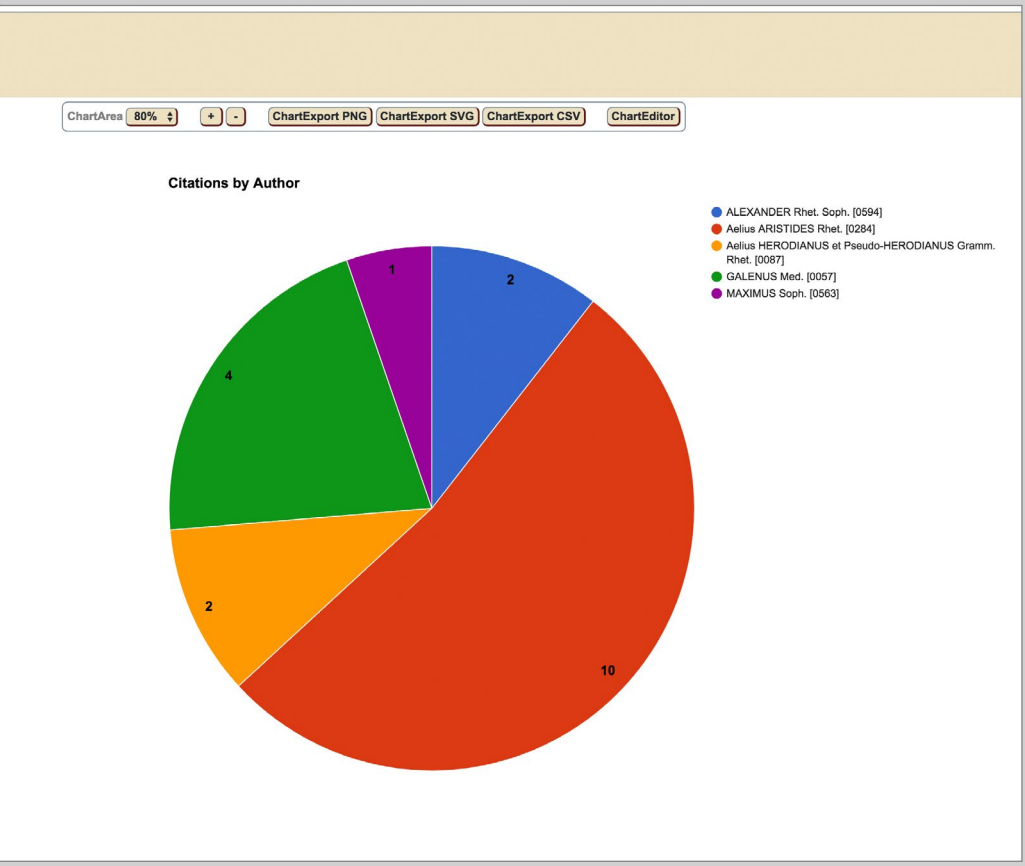


Abbildung 23. Chartview 1: Thukydides Historien eingegrenzt auf 150 n. Chr.

schaltflächen durch die Liste zu klicken. Durch die Auswahl eines Autors im Diagramm öffnet sich ein weiteres Diagramm, das die Möglichkeit gibt, zwischen verschiedenen Werken dieses Referenzautors auszuwählen.

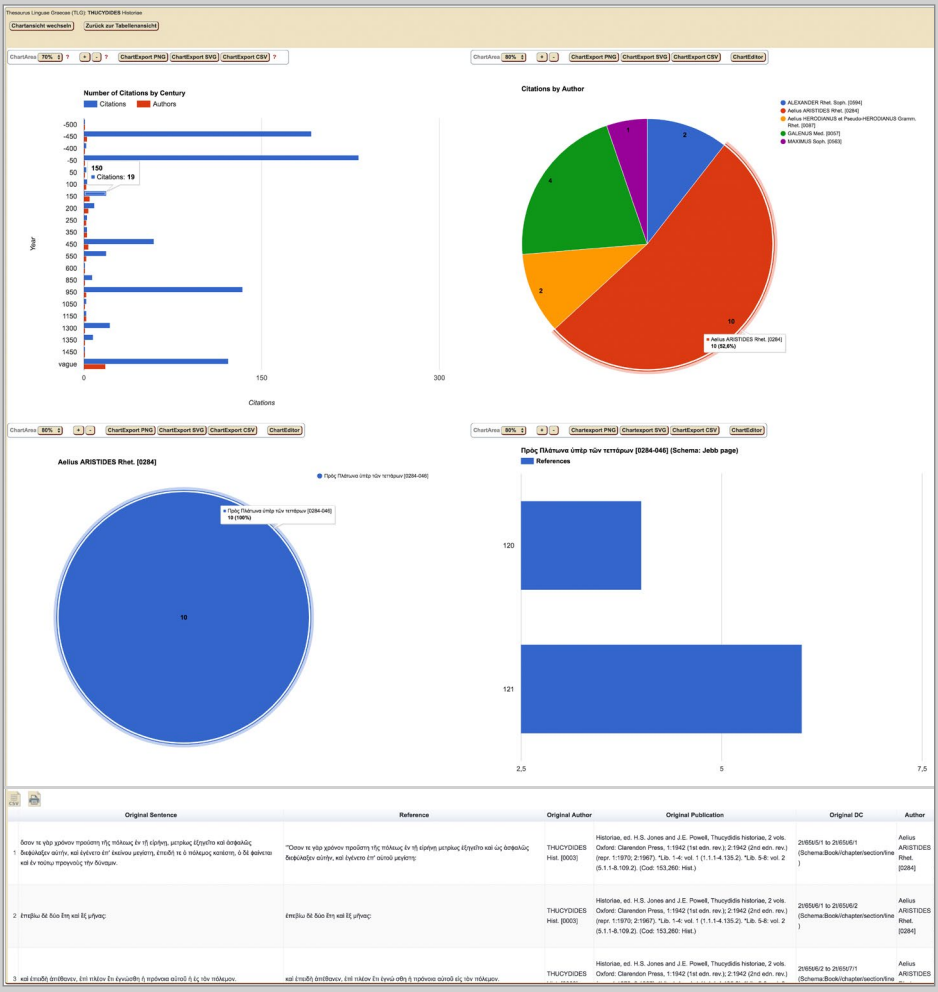
Für das folgende Beispiel wird zuerst eine zeitliche Auswahl im Balkendiagramm getroffen, und zwar der Zeitraum 150 nach Christus. Die Similarity ist auf den Bereich 100 bis 60 voreingestellt. Sobald der Cursor über den entsprechenden Balken fährt, erscheint ein Pop-up Fenster mit dem Zeitraum und der Anzahl der Referenzstellen (■ **Abbildung 23**).

Nachdem der zeitliche Horizont in der Century Range auf den Bereich auf 150 n. Chr. eingestellt wurde, ändert sich gleichzeitig das Autorendiagramm und zeigt die prozentuale Verteilung der Referenzstellen; in diesem Beispiel sind es 19 Referenzstellen aus 5 Autoren. Im nächsten Schritt erfolgt die Auswahl eines Re-



ferenzautors; in diesem Beispiel „Aelius Aristides“. Durch einen Klick auf den roten Bereich im Diagramm wird die Auswahl bestätigt.

Mit der Auswahl des Referenzautors „Aelius Aristides“ ist der letzte Schritt des Arbeitsweges erreicht. Da der Autor nur mit einem Werk vertreten ist, muss an dieser Stelle keine Auswahl mehr zwischen einzelnen Werken getroffen werden. Durch einen Klick in das Kreisdiagramm wird die endgültige Auswahl bestätigt, worauf sich ein weiteres Balkendiagramm öffnet, dass die Verteilung der Referenzstellen im Werk numerisch wiedergibt. Im unteren Bildfeld hat sich derweil eine Ergebnistabelle geöffnet. Diese Ergebnistabelle hat das gleiche Layout wie die eingangs besprochene Ergebnistabelle und folgt der gleichen Handhabung. Sie enthält nun das endgültige Ergebnis der Analyse zwischen dem Originalautor Thukydides,



	Original Sentence	Reference	Original Author	Original Publication	Original DC	Author
1	ὅσον τε χἠρ χρόνον προύστη τῆς πόλεως ἐν τῇ εἰρήνῃ, μετρίως εἴργαστο καὶ ἀσφαλῶς διεφύλαξεν αὐτήν, καὶ ἐγένετο ἐπ' ἐκείνου μεγίστη, ἐπειδὴ τε ὁ πόλεμος κατέστη, ὁ δὲ φαίνεται καὶ ἐν ποσῳ προγνοὺς τὴν δύναμιν.	"Ὅσον τε χἠρ χρόνον προύστη τῆς πόλεως ἐν τῇ εἰρήνῃ μετρίως εἴργαστο καὶ ὡς ἀσφαλῶς διεφύλαξεν αὐτήν, καὶ ἐγένετο ἐπ' αὐτοῦ μεγίστη:	THUCYDIDES Hist. [0003]	Historiae, ed. H.S. Jones and J.E. Powell, Thucydidis historiae, 2 vols. Oxford: Clarendon Press, 1:1942 (1st edn. rev.) 2:1942 (2nd edn. rev.) (repr. 1:1970; 2:1967). "Lib. 1-4: vol. 1 (1.1.1-4.135.2). "Lib. 5-8: vol. 2 (5.1.1-8.109.2). (Cod: 153,260: Hist.)	2\1\65\5\1 to 2\1\65\6\1 (Schema:Book\chapter\section\line)	Aelius ARISTIDES Rhet. [0284]
2	ἐπεῖδε δὲ δύο ἔτη καὶ ἑξ μῆνας:	ἐπεῖδε δὲ δύο ἔτη καὶ ἑξ μῆνας:	THUCYDIDES Hist. [0003]	Historiae, ed. H.S. Jones and J.E. Powell, Thucydidis historiae, 2 vols. Oxford: Clarendon Press, 1:1942 (1st edn. rev.) 2:1942 (2nd edn. rev.) (repr. 1:1970; 2:1967). "Lib. 1-4: vol. 1 (1.1.1-4.135.2). "Lib. 5-8: vol. 2 (5.1.1-8.109.2). (Cod: 153,260: Hist.)	2\1\65\6\1 to 2\1\65\6\2 (Schema:Book\chapter\section\line)	Aelius ARISTIDES Rhet. [0284]
3	καὶ ἐπειδὴ ἀπέθανον, ἐπὶ πλέον ἐν ἐγνώσθη ἢ πρόνοια αὐτοῦ ἢ ἐς τὸν πόλεμον.	καὶ ἐπειδὴ ἀπέθανον, ἐπὶ πλέον ἐν ἐγνώ σθη ἢ πρόνοια αὐτοῦ εἰς τὸν πόλεμον.	THUCYDIDES	Historiae, ed. H.S. Jones and J.E. Powell, Thucydidis historiae, 2 vols. Oxford: Clarendon Press, 1:1942 (1st edn. rev.) 2:1942 (2nd edn. rev.)	2\1\65\6\2 to 2\1\65\7\1 (Schema:Book\chapter\section\line	Aelius ARISTIDES

Abbildung 24. Chartview 1: Ergebnis Thukydides – Aristides

dessen Werk als Ausgangstext gewählt wurde, und dem Referenzautor Aelius Aristides (■ **Abbildung 24**).

Alternativ können Referenzautoren auch mit mehreren Werken vertreten sein, wie beispielsweise der Referenzautor Galen, der mit vier Referenzstellen aus drei Werken vertreten ist (grüner Kreissektor in ■ **Abbildung 24**). Auch in diesem Fall gilt die gleiche Handhabung wie bereits beschrieben. Durch die Auswahl eines Werkes öffnen sich ein neues Balkendiagramm mit der Verteilung der Referenzstellen im Werk und eine Ergebnistabelle im unteren Bildabschnitt.

Dieses Ergebnis wäre auch dann zustande gekommen, wenn man bereits in der ersten Tableview in der Kategorie „Filter Author" den Referenzautor ausgewählt hätte. Mit dieser Auswahl gehen aber bestimmte Vorüberlegungen einher, die z. B. den Pfad bedingen, dass man nicht vom Ergebnis der Analyse hin zu einem – unbekannten – Ergebnis steuert, sondern sich bereits mit dem Vorwissen aus einer ganz bestimmten Fragestellung der Auswertung nähert. Die Chartview 1 bietet daher die Möglichkeit, mit einer bestimmten Fragestellung ein Werk auf unbekannte Parameter hin zu untersuchen. In dieser Ansicht sind also folgende Aspekte besonders effizient und schnell zu überblicken:

1. In welchem Zeitraum beziehen sich wie viele Autoren auf das Ausgangswerk?
2. Gibt es Zeitspannen, in denen die Rückbesinnung und Einbeziehung des Autors und seiner Werke ganz besonders auffällig ist?
3. Welche Autoren bzw. Werke hat der Ausgangsautor als Quellen für sein Werk verwendet?
4. Verwendet der Ausgangsautor bevorzugt bestimmte Autoren und Werke und wie sind diese Textbezüge im Werk des Ausgangsautors verarbeitet?
5. Gibt es Parallelen, Zitate oder Textbezüge, die in der bisherigen Forschung noch nicht oder nicht genügend Beachtung gefunden haben?
6. Wie stellt sich das Verhältnis zweier Autoren im direkten Vergleich zueinander dar?
7. Handelt es sich überhaupt um echte Zitate oder beziehen sich mehrere Autoren unabhängig voneinander auf denselben Ausgangsautor? Oder handelt es sich bei den festgestellten Parallelen um allgemein gebräuchliche Redewendungen?

Da der Vergleich des Ausgangswerks über den gesamten Textbestand des Korpus läuft, wird es notwendigerweise auch mit sich selbst verglichen, wodurch z. B. folgendes Ergebnis zustande kommt (■ **Abbildung 25,** *siehe Seite 50*).

Das Ergebnis erweckt den Eindruck, dass Thukydides sich selbst 184 Mal „zitiert" hat. Da der Ausgangstext als Bestandteil des gesamten Korpus auch mit

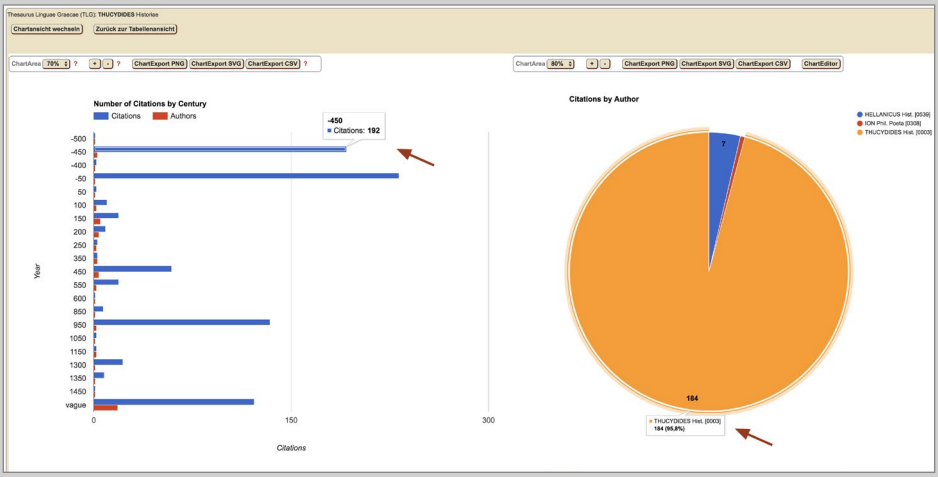


Abbildung 25. Chartview 1: Thukydides – Thukydides

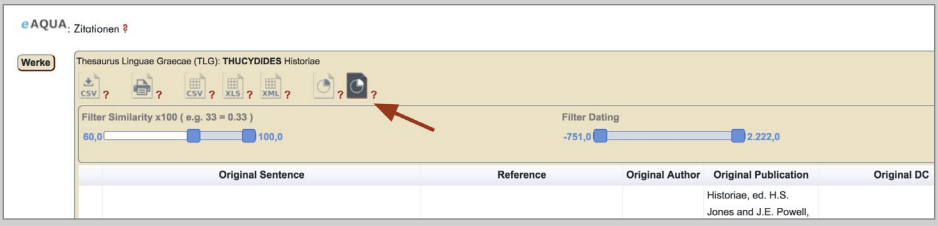


Abbildung 26. Chartview 2 aufrufen

sich selbst verglichen wird, kommt jede Textstelle notwendigerweise zweimal vor, einmal als Ausgangstext und einmal als Referenzstelle. Es handelt sich also nur um 92 mögliche Übereinstimmungen bei einer Similarity von 100 bis 60. Dennoch könnte davon auszugehen sein, dass es sich bei einer Similarity von 96 bei zwei Stellen des gleichen Autors um ein „Selbstzitat" bzw. eine Wiederholung von gleichlautenden Textbausteinen handeln könnte. Im folgenden Beispiel – dem Vergleich von Thuk. IV. 51.1.6 und Thuk. VI. 93.4 – wird dies veranschaulicht:

4t/51t/1/5 to 4t/51t/1/6	**6t/93t/4/4 to 6t/93t/4/6**
καὶ ὁ χειμὼν ἐτελεύτα, καὶ ἕβδομον ἔτος τ�ῳ πολέμῳ ἐτελεύτα τῶδε ὃν Θουκυδίδης ξυνέγραψεν.	καὶ ὁ χειμὼν ἐτελεύτα, καὶ ἕβδομον καὶ δέκατον ἔτος τῳ πολέμῳ ἐτελεύτα τῶδε ὃν Θουκυδίδης ξυνέγραψεν.

Dieses Beispiel eines vermeintlichen Selbstzitates zeigt die ganze Komplexität des Falles. Tatsächlich sind die beiden Textstellen identisch, bis auf den Zusatz καὶ δέκατον, welcher angibt, dass es sich nicht um das siebte Jahr des Peloponnesischen Krieges handelt, sondern um das siebzehnte. Der Text Thuk. IV. 51.6 lautet demnach: „So endete der Winter, und das siebte Jahr des Krieges endete, den Thukydides beschrieben hat." Bzw. Thuk. VI. 93.4: „So endete der Winter, und das siebzehnte Jahr des Krieges endete, den Thukydides beschrieben hat." Dieses Beispiel zeigt, dass vermeintliche Selbstzitate auf Textparallelen hinweisen, die einen ganz unterschiedlichen Charakter haben können. In diesem Beispiel zeigt die Sichtung des Textes sofort, dass die textuelle Gleichheit hier keine Übereinstimmung im Sinne eines Zitats oder einer Parallelstelle ist.

Chartview 1: Über den Referenzautor zum Ergebnis

Der zweite Lösungsweg innerhalb der Chartview 1 entspricht weitestgehend dem ersten und führt auch zu einer Ergebnistabelle mit Referenzstellen eines Referenzautors. Allerdings umgeht dieser Weg die „Century Range" und startet direkt über das Autorendiagramm in die Analyse.

Chartview 2: Von der Verteilung der Parallelstellen im Originalwerk zum Ergebnis

Über die Chartview 2 erfolgt der Zugang zu einem weiteren spezifischen Lösungsweg für die Analyse, und zwar ausgehend von der Verteilung der Referenzstellen auf die Sektionen des Originalwerkes in zwei Ebenen (■ **Abbildung 26**). Abhängig vom Aufbau des Originalwerkes können dies Bücher, Kapitel, Briefe oder sonstige Kategorien sein. Die Aufteilung des zu analysierenden Werkes in einzelne Sektio-

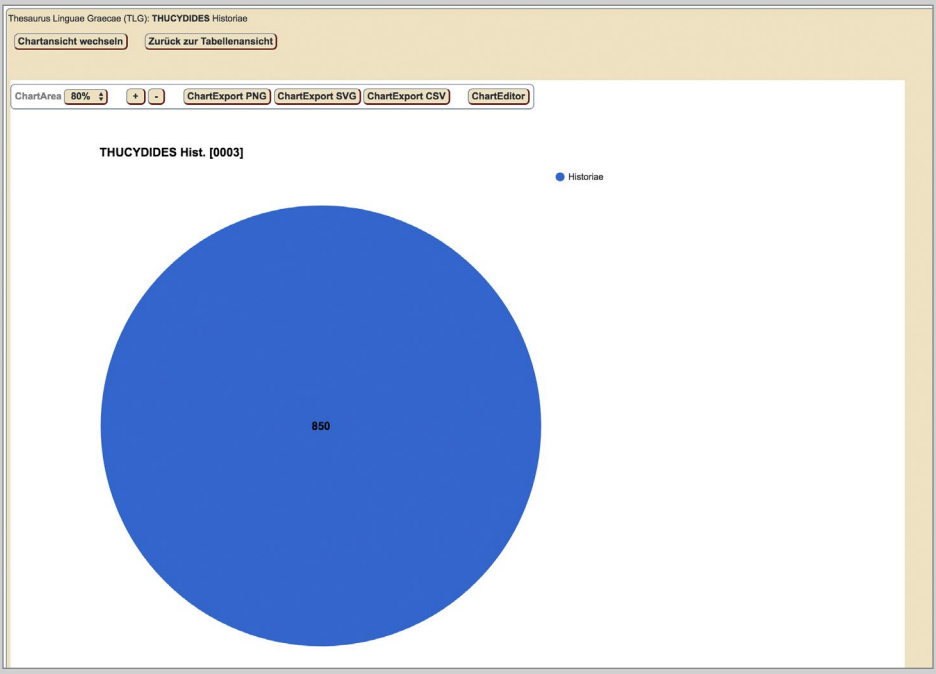


Abbildung 27. Chartview 2: Thukydides Historien

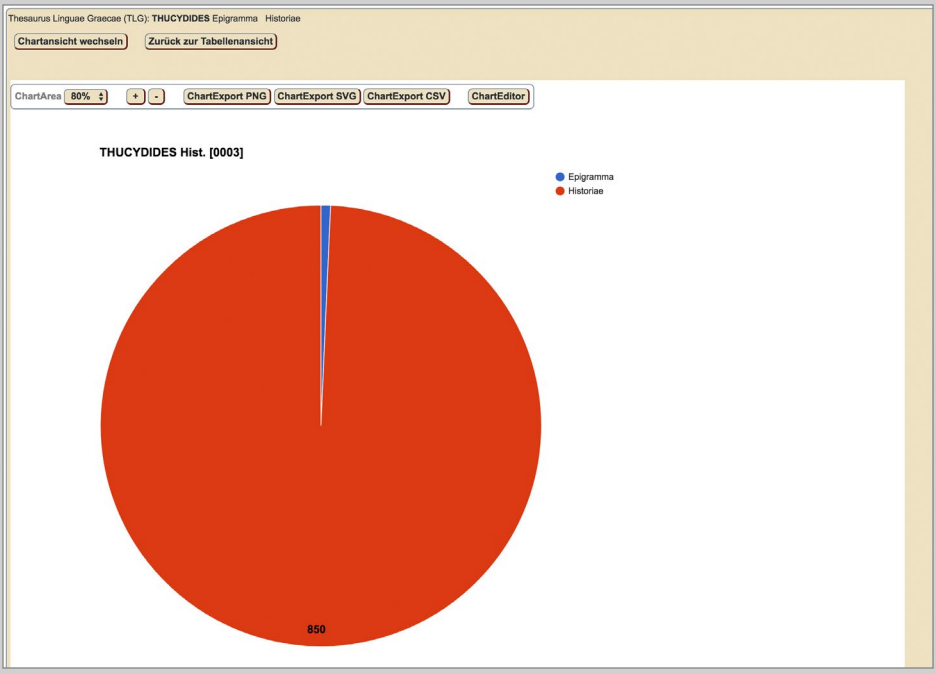


Abbildung 28. Chartview 2: Thukydides Historien und Epigramme

nen dient der unmittelbaren visuellen Erfassung des ansonsten numerischen bzw. tabellarischen Ergebnisses.

Dazu kommt im Bereich unterhalb des Diagramms wie schon in der Chartview 1 eine tabellarische Darstellung in Form des Ergebnisses. Auch sonst entsprechen Darstellung und Arbeitsverlauf der Chartview 1. Auch die Funktionen zum Exportieren sowie die graphischen Darstellungsmechanismen sind gleich.

Das Ergebnis entspricht der vorher getroffenen Auswahl an Regularien in der Tableview (Similarity, Date, etc.). Mit der Auswahl der Chartview 2 öffnet sich ein neues Fenster mit einem Kreisdiagramm; dieses entspricht dem Kreisdiagramm aus Chartview 1. Wiederum bilden die Historien des Thukydides die Ausgangsbasis für die folgenden Ausführungen.

Mit der Auswahl der Historien des Thukydides beschränkt sich die Auswahl von vornherein auf ein Werk, daher ist das erste Kreisdiagramm wenig spektakulär, weil es die prozentuale Verteilung der Parallelstellen auf die Ausgangswerke des gewählten Korpus anzeigt (■ **Abbildung 27**). Hätte man in der Auswahl noch die Epigramme des Thukydides hinzugewählt, sähe das Ergebnis folgendermaßen aus (■ **Abbildung 28**).

Bei mehreren Werken muss an dieser Stelle entschieden werden, welches Werk analysiert werden soll. Vor allem bei Autoren mit einem umfassenden Werkbestand ist eine vorherige Eingrenzung der zu analysierenden Werke sinnvoll, wie beispielsweise bei Plutarch, dessen Gesamtwerk mit 30.000 Referenzstellen auch eine Herausforderung im Hinblick auf die Rechenleistung und -zeit darstellt. In diesem Fall fällt die Entscheidung zugunsten des thukydidischen Werkes mit 850 Referenzstellen.

An dieser Stelle sind bereits einige Vorüberlegungen zu bedenken:

1. Gibt es bereits eine Textpassage des Ausgangstextes, die analysiert werden soll?
2. Interessieren die am häufigsten oder auch die am seltensten zitierten Textpassagen des Ausgangstextes?
3. Wie verhält es sich mit Leerstellen; warum werden bestimmte Texte zitiert und andere überhaupt nicht?

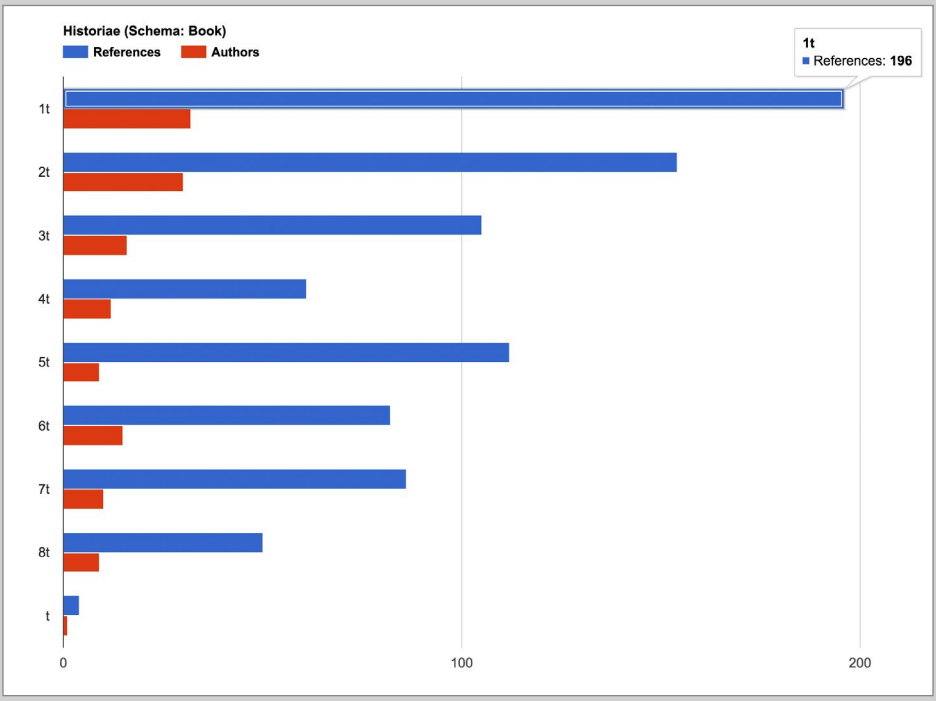


Abbildung 29. Chartview 2: Sektionsebene 1

Book	References	Authors
1t	196	32
2t	154	30
3t	105	16
4t	61	12
5t	112	9
6t	82	15
7t	86	10
8t	50	9
t	4	1

Tabelle 2. Chartview 2: Export Sektionsebene 1

Chartview 2: Erste Sektionsebene

Entsprechend der Korpuszusammensetzung gibt das Kreisdiagramm die prozentuale Verteilung von Parallelstellen im Vergleich zum gesamten Korpus als Kreisdiagramm wieder. Durch Anklicken des entsprechenden Diagrammbereiches bestätigt man die Auswahl und es öffnet sich ein Balkendiagramm, das im Folgenden als Sektionsdiagramm geführt wird. Innerhalb dieses Diagramms werden die Referenzstellen den Sektionen numerisch zugeordnet.

In diesem, hier auf Büchern basierenden Sektionsdiagramm, wird ein positives Ergebnis präsentiert, das heißt, es werden nur solche Kapitel angezeigt, zu denen Parallelen ermittelt werden konnten. Bücher bzw. Sektionen ohne Ergebnisse werden nicht angezeigt. Das hat zum einen praktische Gründe, aber auch darstellerische, da vor allem bei Werken mit einer beträchtlichen Anzahl an Sektionen und Untersektionen der visuelle Effekt verloren ginge. Die Anzahl der Parallelstellen und Autoren erscheinen in einem Pop-up Fenster, sobald der Cursor über den Balken fährt. Mit Anklicken eines Balkens wird der Bereich ausgewählt. Hier fällt die Wahl auf das erste Buch mit 196 Parallelstellen (■ **Abbildung 29**).

Anders als in der Chartview 1 befindet sich bereits ohne eine bestimmte Auswahl an Sektionen die Ergebnistabelle im unteren Bildabschnitt und entspricht exakt der Tabelle der Tableview. Durch die weitere Auswahl reduziert sich das Ergebnis. Man kann sich zum besseren Überblick auch die gesamte Sektionsleiste als CSV Datei herunterladen und numerisch anzeigen lassen (■ **Tabelle 2**).

Chapters	References	Authors						
1t	6	1	63t	1	1	110t	5	3
2t	8	5	64t	1	1	118t	5	2
6t	6	5	69t	1	1	120t	6	3
17t	1	1	70t	17	5	122t	2	1
18t	1	1	71t	3	2	124t	1	1
21t	2	1	73t	3	3	126t	1	1
22t	11	2	75t	2	2	128t	3	1
23t	13	1	76t	1	1	129t	2	1
24t	9	6	77t	2	2	130t	4	2
25t	6	1	78t	1	1	132t	3	3
28t	4	1	88t	1	1	135t	2	1
31t	1	1	89t	1	1	136t	6	1
33t	2	2	91t	1	1	137t	7	2
34t	1	1	92t	1	1	138t	10	2
35t	3	3	96t	1	1	140t	2	2
37t	3	2	97t	9	2	141t	2	2
49t	1	1	99t	5	1	143t	2	2
53t	1	1	100t	1	1			
58t	1	1	102t	1	1			

Tabelle 3. Chartview 2: Export Sektionsebene 2

Chartview 2: Zweite Sektionsebene

Nach der Auswahl des ersten Buches mit 196 Parallelstellen öffnet sich eine neue Sektionsleiste im unteren Bildabschnitt, welche die Aufteilung des gewählten Buches in Sektionen wiedergibt und damit zur Sektionsebene 2 übergeht. Wie bereits erwähnt, wird nur ein positiver Befund dargestellt. Mit einem Blick auf die Sektionsleiste wird deutlich, dass die praktikable Darstellung bei 143 Kapiteln an ihre Grenzen käme.

Auch diese Sektionsleiste lässt sich als CSV Datei herunterladen und numerisch anzeigen (■ **Tabelle 3**).

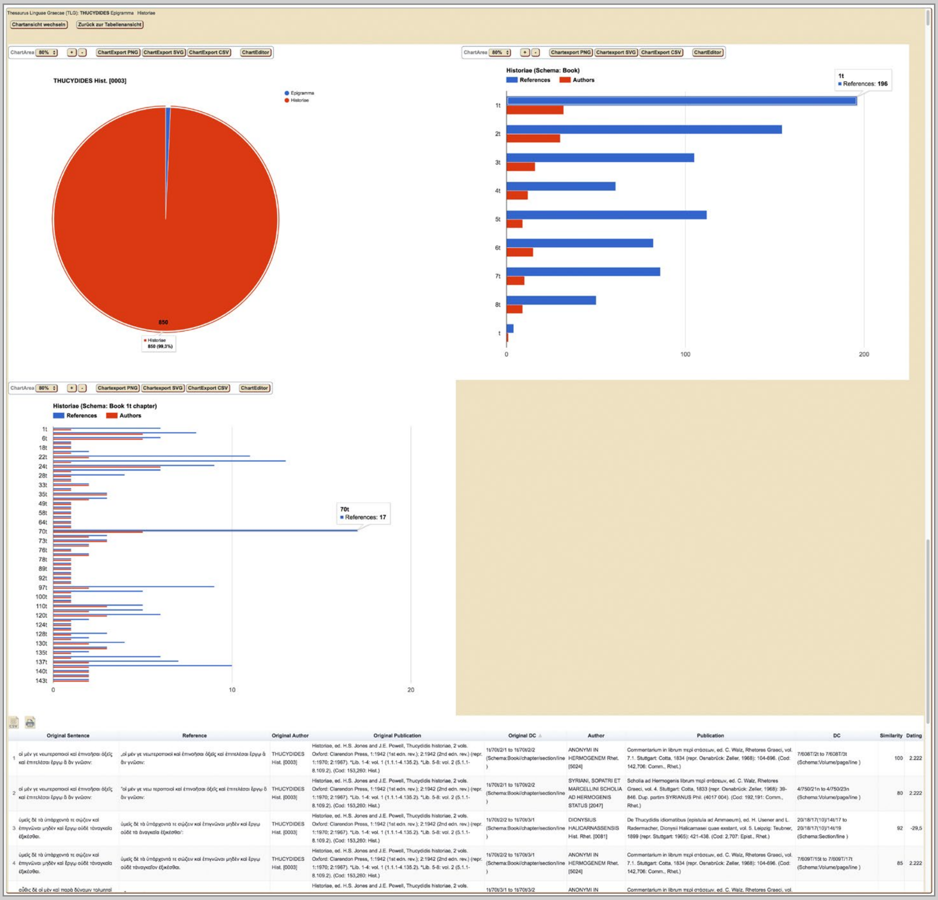


Abbildung 30. Chartview 2: Sektionsebene 2 mit Ergebnistabelle

Ergebnistabelle

Mit der Auswahl eines Unterkapitels wird das Ergebnis weiter eingegrenzt, hier zum Beispiel mit der Auswahl des Kapitels 1. 70, in dem sich 17 Parallelstellen befinden. Mit dieser letzten Auswahl endet der Arbeitsweg. Die Ergebnistabelle beinhaltet nun alle 17 Parallelstellen zu Thuk. I. 70 und kann, wie alle Ergebnistabellen davor nun gedruckt oder als CSV-Datei exportiert werden (■ **Abbildung 30**).

Abbildung 31. Online-Konverter für altgriechischen Beta Code

Online-Konverter Beta Code

Texte in Beta Code zu transkribieren ermöglicht die unkomplizierte Eingabe von Altgriechisch mittels handelsüblicher Tastatur und ohne Verwendung von virtuellen Tastaturen oder Umstellungen des Tastaturlayouts.

Griechischer Beta Code ist die 7-Bit-sichere Kodierung mittels des US-ASCII-Zeichensatzes. Jedes diakritische Zeichen wird durch ein eigenes Zeichen dargestellt, welches dem Buchstaben folgt (Ausnahme: bei Großbuchstaben vor dem Buchstaben). Beta Code unterscheidet nicht zwischen Klein-/Großschreibung, Großbuchstaben werden durch Voranstellung von * Asteriskos (griech. ἀστερίσκος) gekennzeichnet. Einige Projekte benutzen nur Großbuchstaben (z. B. TLG), andere nur Kleinbuchstaben (z. B. das Perseus Project). Beide Varianten werden vom Konverter berücksichtigt. Darüber hinaus gibt es traditionell unterschiedliche Schreibweisen des Endsigmas, welche ebenfalls einstellbar ist.

Der Online-Konverter wandelt den Beta Code in Unicode UTF-8 bzw. Unicode in Beta Code (■ **Abbildung 31**). Dafür sind aus der UTF-8-Codetabelle aus den Bereichen Greek and Coptic (U+0370 – U+03FF) und Greek Extended (U+1F00 – U+1FFF) die Zeichenkombinationen mit ihren Beta Code Varianten versehen. Ein fallback-Mechanismus versucht dabei, die Codepositions aus dem Bereich Combining Diacritical Marks (U+0300 – U+036F) in ihr Äquivalent aus den anderen Bereichen umzuwandeln.

Alle zusätzlichen Formatierungen, wie beispielsweise Zeilenumbrüche oder HTML-Formatierungen werden vom Konverter entfernt. Über den De-bug-mode, einzustellen am rechten unteren Rand, können die Umwandlungspaare schrittweise angezeigt werden.

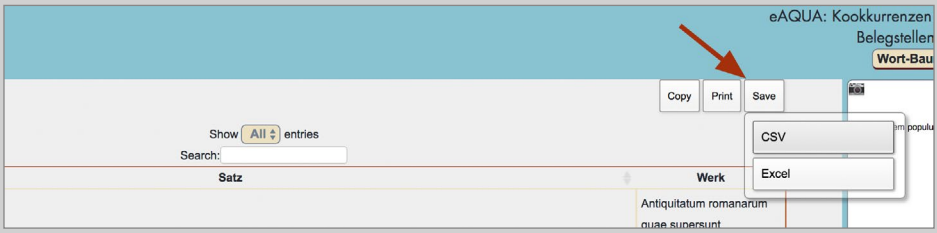


Abbildung 32. Export von Belegstellen der Kookkurrenzsuche

Export von Suchergebnissen aus den Online-Tools

Kookkurrenzsuche

Belegstellen aus der Tabelle

Der Export aus der Tabelle mit den Belegstellen wird über Funktionen des Adobe Flash Players realisiert (■ **Abbildung 32**). Deshalb ist hier ein entsprechendes Plugin im Browser notwendig. Alternativ besteht bei HTML-Tabellen immer die Möglichkeit, die komplette Tabelle im Browser zu markieren, zu kopieren und sie dann in ein Schreibprogramm einzufügen.

Copy
- Content-Type: text/csv
- Separator: Tabulator
- Zeichensatz: UTF-8

Der Text wird in die Zwischenablage kopiert und kann von dort aus weiterverwendet werden.

Save CSV
- Content-Type: text/csv
- Separator: Komma
- Texttrenner: Doppelte Anführungszeichen
- Zeichensatz: UTF-8

Es wird ein Speichern-unter-Dialogfenster zur Auswahl des Speicherorts angeboten.

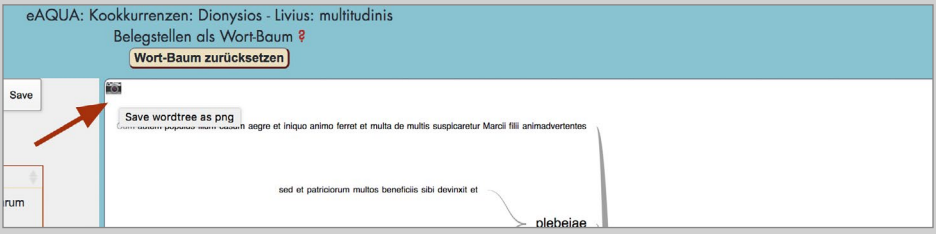


Abbildung 33. Export der Wortbaumansicht

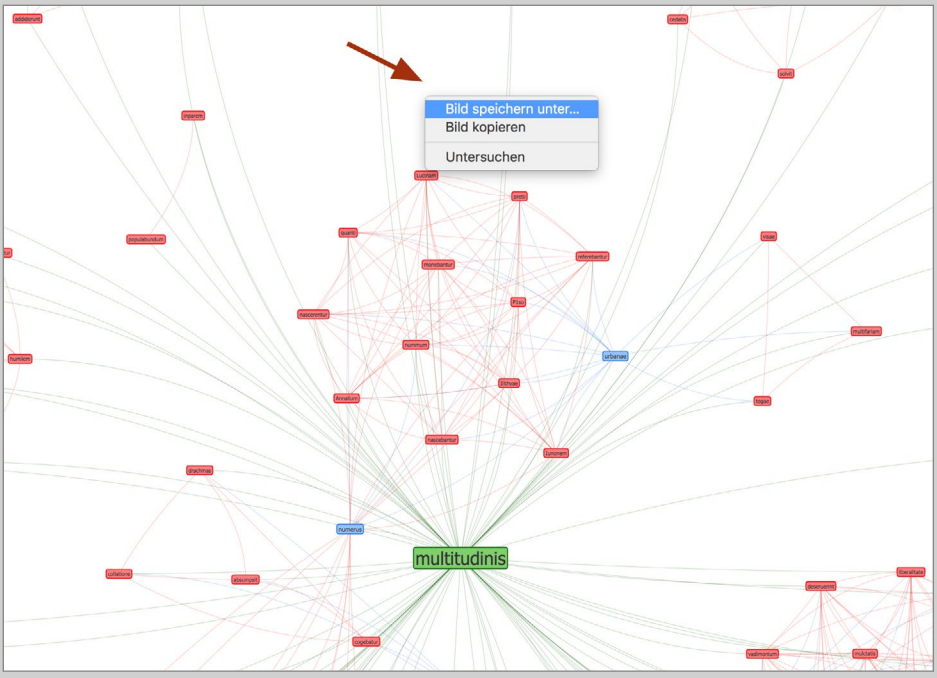


Abbildung 34. Export der Netzwerk-Visualisierung

Save Excel
- Content-Type: text/csv
- Separator: Tabulator
- Zeichensatz: UTF-8

Es wird ein Speichern-unter-Dialogfenster zur Auswahl des Speicherorts angeboten.

Wortbaum
- Content-Type: image/png

Links oben im Wortbaum findet sich ein kleines Icon. Bei Klick darauf öffnet sich ein neues Fenster mit der Grafik. Wenn mit der rechten Maustaste innerhalb der Grafik geklickt wird, erscheint ein Kontextmenü des Browsers (Chrome, FF, Opera). Dort auf Grafik bzw. Bild speichern klicken (■ **Abbildung 33**).

Netzwerk-Visualisierung
- Content-Type: image/png

Wenn mit der rechten Maustaste innerhalb der Visualisierung geklickt wird, erscheint ein Kontextmenü des Browsers (Chrome, FF, Opera). Dort auf Grafik bzw. Bild speichern klicken (■ **Abbildung 34**).

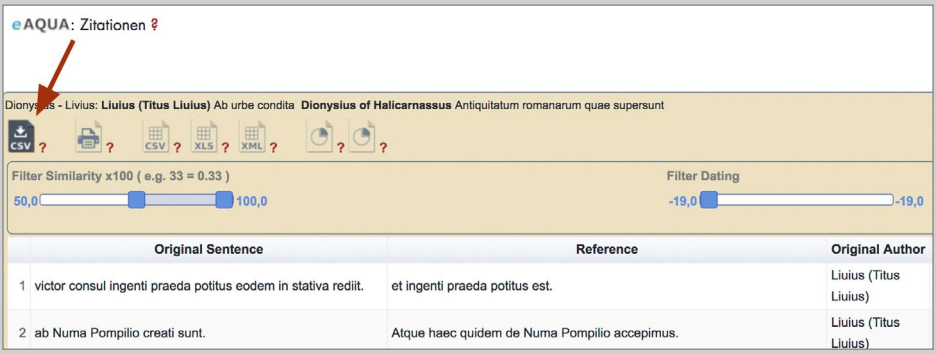


Abbildung 35. Direktdownload der gesamten Tabelle (ohne eingestellten Filter)

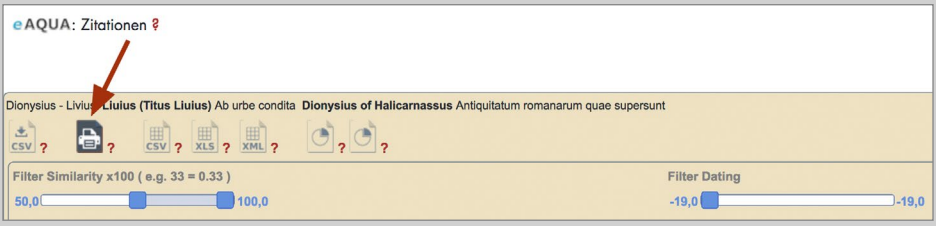


Abbildung 36. Drucken der Tabelle

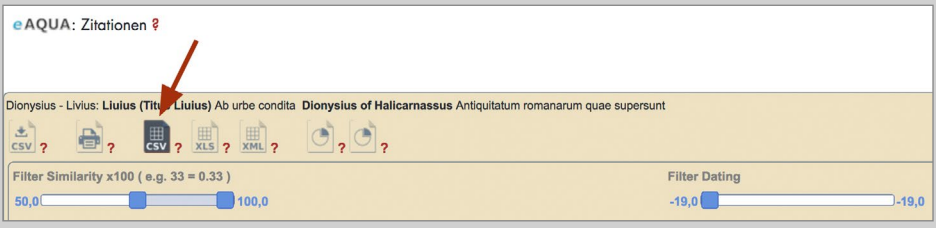


Abbildung 37. Empfohlener Tabellenexport nach CSV – direkt aus dem Browser

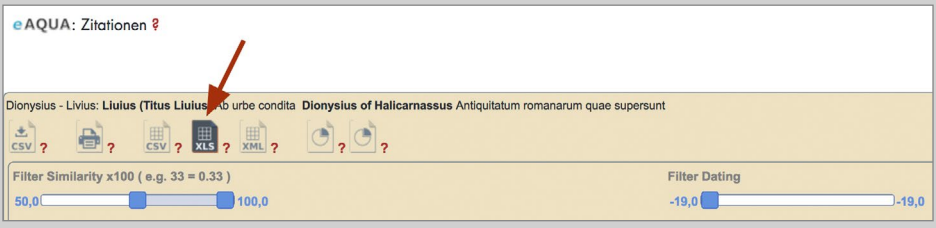


Abbildung 38. Tabellenexport nach XLS

Zitation

Daten aus der Visualisierungs-Tabelle

Direktdownload CSV

- Content-Type: text/csv
- Separator: Semikolon
- Zeichensatz: UTF-8

Beim Direktdownload werden die kompletten Daten nochmals vom Server geladen. Eingestellte Filter in der Visualisierungs-Tabelle, wie z. B. die Eingrenzung des Wertes „Similarity“ finden dabei keine Berücksichtigung (■ **Abbildung 35**).

Tabelle Drucken

Die Tabelle wird in einem neuen Fenster zum Ausdrucken geöffnet (■ **Abbildung 36**).

Tabellenexport CSV

- Content-Type: text/csv
- Separator: Semikolon
- Zeichensatz: UTF-8

Beim Tabellenexport werden die gefilterten Daten zum Download aufbereitet. Bei der Variante CSV geschieht dies, im Gegensatz zu XLS und XML, direkt im Browser ohne Umwege zum Server (■ **Abbildung 37**).

Tabellenexport XLS

- Content-Type: application/vnd.ms-excel

Beim Tabellenexport werden die gefilterten Daten zum Download aufbereitet. Bei der Variante XLS geschieht dies, indem die relevanten Datensätze an den Server zurückgeschickt werden, um sie von dort in einem anderen Format wieder zu laden. Bei großen Datenmengen, wie beispielsweise bei einigen Subkorpora, kann dies sehr lange dauern oder unter Umständen zum Abbruch führen. Der Export nach CSV ist in diesem Falle zu präferieren (■ **Abbildung 38**).

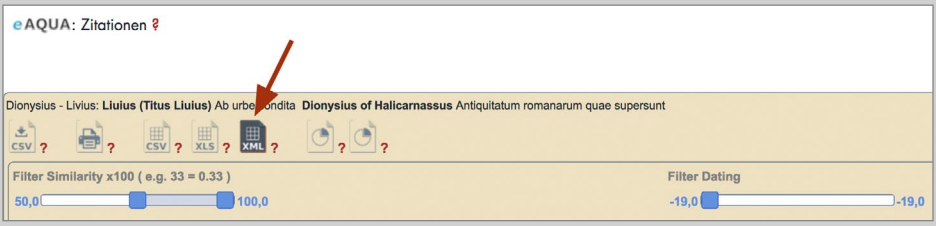


Abbildung 39. Tabellenexport nach XML

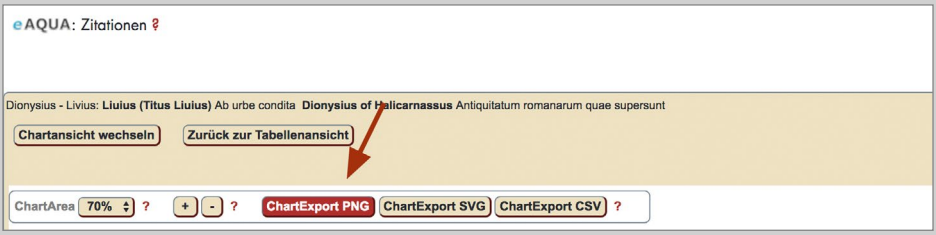


Abbildung 40. Chartexport nach PNG

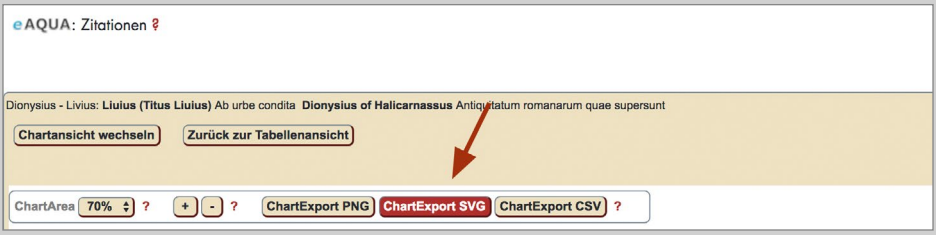


Abbildung 41. Chartexport nach SVG

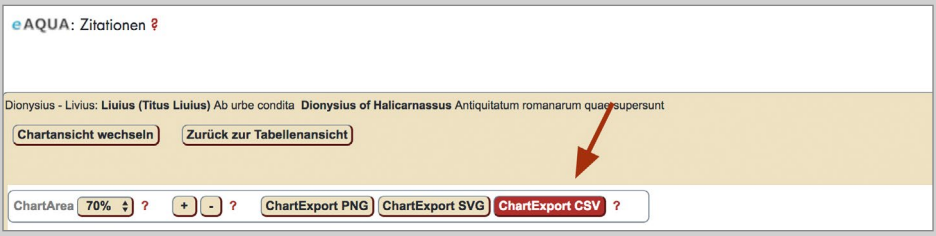


Abbildung 42. Chartexport nach CSV

Tabellenexport XML
- Content-Type: application/tei+xml
- Zeichensatz: UTF-8

Beim Tabellenexport werden die gefilterten Daten zum Download aufbereitet. Bei der Variante XML geschieht dies, indem die relevanten Datensätze an den Server zurückgeschickt werden, um sie von dort in einem anderen Format wieder zu laden. Bei großen Datenmengen, wie beispielsweise bei einigen Subkorpora, kann dies sehr lange dauern oder unter Umständen zum Abbruch führen. Der Export nach CSV ist in diesem Falle zu präferieren (■ **Abbildung 39**).

Charts aus der Visualisierungs-Tabelle

Oberhalb der Charts finden sich jeweils drei Schaltflächen, über die der Export gestartet werden kann. Die ersten beiden sind für den Export in ein Grafikformat, die letzte bedient ein Textformat.

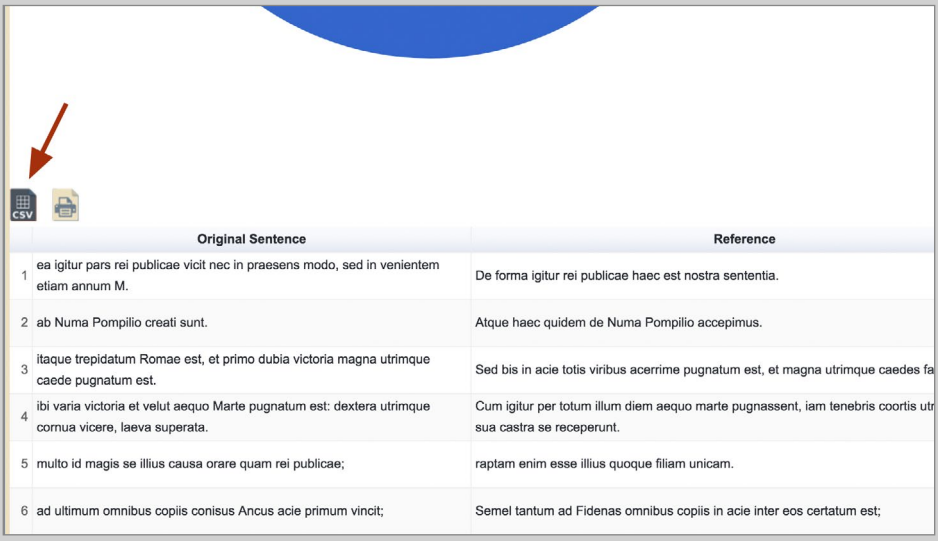
Der Download-Dialog wird nur von einigen Browsern (Chrome, FF, Opera) automatisch gestartet. Andere, wie beispielsweise Safari, öffnen ein neues Fenster mit der Grafik. Dort mit der rechten Maustaste auf die Grafik klicken und sichern unter wählen, um die Grafikdatei auf dem Rechner zu speichern.

Chartexport PNG (■ Abbildung 40)
- Content-Type: image/png

Chartexport SVG (■ Abbildung 41)
- Content-Type: image/svg+xml
- Zeichensatz: UTF-8

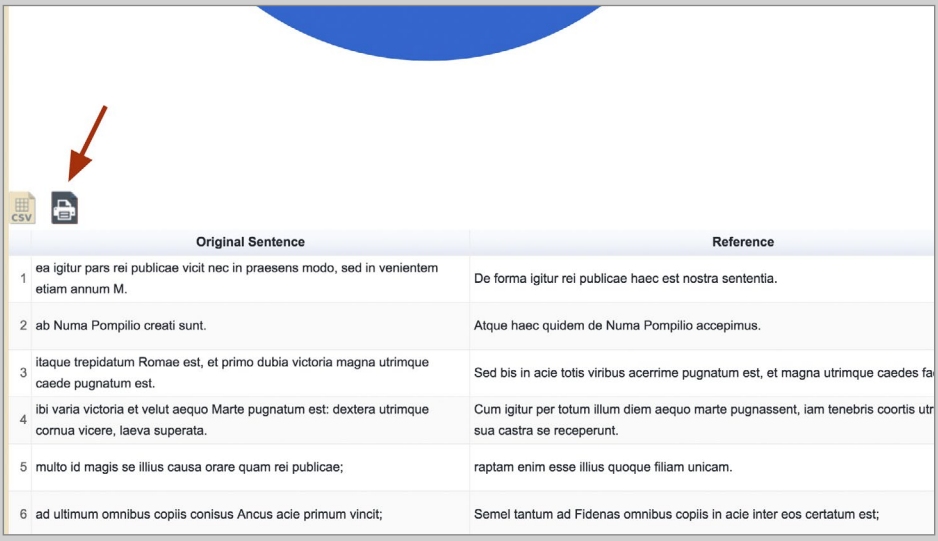
Chartexport CSV
- Content-Type: text/csv
- Separator: Semikolon
- Zeichensatz: UTF-8

Beim Chartexport nach CSV werden die der Darstellung zu Grunde liegenden Daten exportiert (■ **Abbildung 42**).



	Original Sentence	Reference
1	ea igitur pars rei publicae vicit nec in praesens modo, sed in venientem etiam annum M.	De forma igitur rei publicae haec est nostra sententia.
2	ab Numa Pompilio creati sunt.	Atque haec quidem de Numa Pompilio accepimus.
3	itaque trepidatum Romae est, et primo dubia victoria magna utrimque caede pugnatum est.	Sed bis in acie totis viribus acerrime pugnatum est, et magna utrimque caedes fa
4	ibi varia victoria et velut aequo Marte pugnatum est: dextera utrimque cornua vicere, laeva superata.	Cum igitur per totum illum diem aequo marte pugnassent, iam tenebris coortis utr sua castra se receperunt.
5	multo id magis se illius causa orare quam rei publicae;	raptam enim esse illius quoque filiam unicam.
6	ad ultimum omnibus copiis conisus Ancus acie primum vincit;	Semel tantum ad Fidenas omnibus copiis in acie inter eos certatum est;

Abbildung 43. Tabelle aus der Chartview exportieren nach CSV



	Original Sentence	Reference
1	ea igitur pars rei publicae vicit nec in praesens modo, sed in venientem etiam annum M.	De forma igitur rei publicae haec est nostra sententia.
2	ab Numa Pompilio creati sunt.	Atque haec quidem de Numa Pompilio accepimus.
3	itaque trepidatum Romae est, et primo dubia victoria magna utrimque caede pugnatum est.	Sed bis in acie totis viribus acerrime pugnatum est, et magna utrimque caedes fa
4	ibi varia victoria et velut aequo Marte pugnatum est: dextera utrimque cornua vicere, laeva superata.	Cum igitur per totum illum diem aequo marte pugnassent, iam tenebris coortis utr sua castra se receperunt.
5	multo id magis se illius causa orare quam rei publicae;	raptam enim esse illius quoque filiam unicam.
6	ad ultimum omnibus copiis conisus Ancus acie primum vincit;	Semel tantum ad Fidenas omnibus copiis in acie inter eos certatum est;

Abbildung 44. Tabelle aus der Chartview Drucken

Tabellenexport CSV

- Content-Type: text/csv
- Separator: Semikolon
- Zeichensatz: UTF-8

Beim Tabellenexport nach CSV werden die angezeigten Belegstellen direkt aus dem Browser zum Download aufbereitet angeboten (■ **Abbildung 43**).

Tabelle Drucken

Die Tabelle wird in einem neuen Fenster zum Ausdrucken geöffnet (■ **Abbildung 44**).

Wort	Anzahl	Wort	Anzahl
et	373596	sunt	44550
in	250758	per	42081
est	171009	se	40733
que	150721	enim	40511
ut	134350	ab	37589
non	131925	nec	36390
cum	108737	etiam	35838
ad	96350	autem	31760
quod	89124	id	31421
qui	71400	atque	30777
si	71196	ac	30579
quae	68120	ne	30039
sed	65265	quid	29619
ex	57516	haec	28993
de	56801	te	27494
a	56216	quo	27143
quam	51254	vel	27108
aut	50196	me	27015
esse	49777	nam	26992
hoc	48359	sit	26983

Tabelle 4. Frequenzsortierte Wortliste BTL als Basis einer Stoppwortliste

Korpusanalyse

Computergestützte Verarbeitung von Sprache

Für die Gewinnung strukturierter Informationen aus Texten kommen, je nach Anwendungsfall, verschiedene Sprachtechnologie-Komponenten zum Einsatz. Bei der Verarbeitung antiker Texte ergeben sich, beispielsweise durch das Fehlen von sogenannten Metadaten, einige Besonderheiten, so dass nicht alle Komponenten berücksichtigt werden. Nachfolgend soll eine grobe Aufstellung der zum Einsatz kommenden Sprachtechnologie gegeben werden.

Grundsätzlich wird innerhalb von Data-Mining bei der Verarbeitung von Sprache von drei Bereichen gesprochen:

- domänenspezifische Verarbeitung
- dokumentspezifische Verarbeitung
- sprachspezifische Verarbeitung

Hierbei handelt es sich um eine rein thematische, nicht chronologische Aufzählung.

UTF-8 (hex.)	Unicode Codepos.		Name	Beta-Code
e1bc80	U+1F00	ἀ	GREEK SMALL LETTER ALPHA WITH PSILI	a)
e1bc81	U+1F01	ἁ	GREEK SMALL LETTER ALPHA WITH DASIA	a(
e1bc82	U+1F02	ἂ	GREEK SMALL LETTER ALPHA WITH PSILI AND VARIA	a)\
e1bc83	U+1F03	ἃ	GREEK SMALL LETTER ALPHA WITH DASIA AND VARIA	a(\
e1bc84	U+1F04	ἄ	GREEK SMALL LETTER ALPHA WITH PSILI AND OXIA	a)/
e1bc85	U+1F05	ἅ	GREEK SMALL LETTER ALPHA WITH DASIA AND OXIA	a(/
e1bc86	U+1F06	ἆ	GREEK SMALL LETTER ALPHA WITH PSILI AND PERISPOMENI	a)=
e1bc87	U+1F07	ἇ	GREEK SMALL LETTER ALPHA WITH DASIA AND PERISPOMENI	a(=
e1bc88	U+1F08	Ἀ	GREEK CAPITAL LETTER ALPHA WITH PSILI	*)a
e1bc89	U+1F09	Ἁ	GREEK CAPITAL LETTER ALPHA WITH DASIA	*(a
e1bc8a	U+1F0A	Ἂ	GREEK CAPITAL LETTER ALPHA WITH PSILI AND VARIA	*)\a
e1bc8b	U+1F0B	Ἃ	GREEK CAPITAL LETTER ALPHA WITH DASIA AND VARIA	*(\a
e1bc8c	U+1F0C	Ἄ	GREEK CAPITAL LETTER ALPHA WITH PSILI AND OXIA	*)/a
e1bc8d	U+1F0D	Ἅ	GREEK CAPITAL LETTER ALPHA WITH DASIA AND OXIA	*(/a
e1bc8e	U+1F0E	Ἆ	GREEK CAPITAL LETTER ALPHA WITH PSILI AND PERISPOMENI	*)=a
e1bc8f	U+1F0F	Ἇ	GREEK CAPITAL LETTER ALPHA WITH DASIA AND PERISPOMENI	*(=a
e1bc90	U+1F10	ἐ	GREEK SMALL LETTER EPSILON WITH PSILI	e)
e1bc91	U+1F11	ἑ	GREEK SMALL LETTER EPSILON WITH DASIA	e(
e1bc92	U+1F12	ἒ	GREEK SMALL LETTER EPSILON WITH PSILI AND VARIA	e)\
e1bc93	U+1F13	ἓ	GREEK SMALL LETTER EPSILON WITH DASIA AND VARIA	e(\
e1bc94	U+1F14	ἔ	GREEK SMALL LETTER EPSILON WITH PSILI AND OXIA	e)/
e1bc95	U+1F15	ἕ	GREEK SMALL LETTER EPSILON WITH DASIA AND OXIA	e(/
e1bc98	U+1F18	Ἐ	GREEK CAPITAL LETTER EPSILON WITH PSILI	*)e
e1bc99	U+1F19	Ἑ	GREEK CAPITAL LETTER EPSILON WITH DASIA	*(e
e1bc9a	U+1F1A	Ἒ	GREEK CAPITAL LETTER EPSILON WITH PSILI AND VARIA	*)\e
e1bc9b	U+1F1B	Ἓ	GREEK CAPITAL LETTER EPSILON WITH DASIA AND VARIA	*(\e
e1bc9c	U+1F1C	Ἔ	GREEK CAPITAL LETTER EPSILON WITH PSILI AND OXIA	*)/e
e1bc9d	U+1F1D	Ἕ	GREEK CAPITAL LETTER EPSILON WITH DASIA AND OXIA	*(/e
e1bca0	U+1F20	ἠ	GREEK SMALL LETTER ETA WITH PSILI	h)
e1bca1	U+1F21	ἡ	GREEK SMALL LETTER ETA WITH DASIA	h(
e1bca2	U+1F22	ἢ	GREEK SMALL LETTER ETA WITH PSILI AND VARIA	h)\
e1bca3	U+1F23	ἣ	GREEK SMALL LETTER ETA WITH DASIA AND VARIA	h(\
e1bca4	U+1F24	ἤ	GREEK SMALL LETTER ETA WITH PSILI AND OXIA	h)/
e1bca5	U+1F25	ἥ	GREEK SMALL LETTER ETA WITH DASIA AND OXIA	h(/
e1bca6	U+1F26	ἦ	GREEK SMALL LETTER ETA WITH PSILI AND PERISPOMENI	h)=
e1bca7	U+1F27	ἧ	GREEK SMALL LETTER ETA WITH DASIA AND PERISPOMENI	h(=
e1bca8	U+1F28	Ἠ	GREEK CAPITAL LETTER ETA WITH PSILI	*)h
e1bca9	U+1F29	Ἡ	GREEK CAPITAL LETTER ETA WITH DASIA	*(h
e1bcaa	U+1F2A	Ἢ	GREEK CAPITAL LETTER ETA WITH PSILI AND VARIA	*)\h
e1bcab	U+1F2B	Ἣ	GREEK CAPITAL LETTER ETA WITH DASIA AND VARIA	*(\h

Tabelle 5. Auszug Beta Code Altgriechisch und die UTF-8-Entsprechung

Domänenspezifische Verarbeitung

Teilaufgabe	Erläuterung
Eigennamenextraktion	Erkennung von spezifischen Entitäten; meist auf der Basis manuell annotierter Datensätze. Hierbei sind nur die für die Domäne (das Korpus) typischen gemeint.[14]
Stoppwortliste erstellen	Eine Stoppwortliste ist eine Liste mit Begriffen, die bei der späteren Verarbeitung ausgenommen werden sollen (■ **Tabelle 4,** *siehe Seite 72*).[15]
Topic-Modellierung	Automatische Zuordnung von Begriffen zu Themen auf Basis von Worteigenschaften und Kontextinformationen.
Faktenextraktion	Vorher definierte Arten von Informationen werden durch die Verarbeitung modelliert. Viele Verfahren nutzen dafür die Abfolge unterschiedlicher Wörter in einem Satz.[16]
Relationsextraktion	Erkennung von Beziehungen zwischen Entitäten in einem Text.

Dokumentspezifische Verarbeitung

Teilaufgabe	Erläuterung
Metadaten erfassen	Metadaten, im Falle der Korpusanalyse z. B. Entstehungsort, Entstehungszeit, Autorenschaft, Editor, Editionszeit usw., sind bei der Textanalyse wertvolle Informationsquellen, um beispielsweise die Auswahl der zu verarbeitenden Daten einzugrenzen.
Bereinigung und Normalisierung	Abhängig davon, wie die Daten erfasst wurden, müssen sie vor der Analyse von allen irrelevanten Informationen, wie z. B. die für Auszeichnungssprachen üblichen Markup Tags, bereinigt werden. Eventuell abweichende Zeichenkodierungen, wie z. B. transkribierter altgriechischer Beta Code, müssen vor der Verarbeitung in eine einheitliche Zeichenkodierung konvertiert werden (■ **Tabelle 5**).

14 Zum Beispiel die im Bühnenstück von Shakespeare „KING HENRY the Fourth" abgekürzten „Speaker"-Segmente „North." und „West." sind Personenbezeichner, keine Himmelsrichtungen.

15 Solche Listen können sowohl domänenübergreifend, beispielsweise typisch für eine Sprache, als auch domänenspezifisch, beispielsweise typisch für eine Autorenschaft, sein. In eAQUA werden diese Listen auf Basis von Wortzählungen des Gesamtkorpus erstellt.

16 In eAQUA ist dies beispielsweise mit der Kookkurrenzanalyse vollzogen worden.

```
<pb n="62"/>
<p>VII— IX. ΜΑΘΗΜΑΤΙΚΑ.</p>
<p>11 l. [ VII 1] ΠΕΡΙ ΔΙΑΦΟΡΗΣ ΓΝΩΜΗΣ ἢ ΠΕΡΙ ψΑΥΣΙΟΣ ΚΥΚΛΟΥ ΚΑΙ
   <note type="marginal">390</note>
   ΣΦΑΙΡΗΣ.</p>
<p>11 m. [ VII 2] ΠΕΡΙ ΓΕΩΜΕΤΡΙΗΣ. Vgl. B 155.</p>
<lb n="5"/> <p>11 n. [ VII 3] ΓΕΩΜΕΤΡΙΚΩΝ &#x003C;A&#x772;B&#x772;?&#x003E;</p>
<p> o. [ VII 4] ΑΡΙΘΜΟΙ.</p>
<p>11p. [VIII 1] ΠΕΡΙ ΑΛΟΓΩΝ ΓΡΑΜΜΩΝ ΚΑ| ΝΑΣΤΩΝ</p>
<p>11 q. [ VIII 2] ΕΚΠΕΤΑΣΜΑΤΑ.</p>
<p>11r. [ VIII 3] ΜΕΓΑΣ ΕΝΙΑΥΤΟΣ ἡ ΑΣΤΡΟΝΟΜΙΗ. ΠΑΡΑΠΗΓΜΑ. Vgl.</p>
<lb n="10"/> <p>B 14, 5; 15 a. Diog. v 43 Theophrasts Schrift Περὶ τῆς Δημοκρίτου
   ἀστρολογίας ᾱ.</p>
<p>12. Censor. 18, 8 est et Philolai annus [32 A. 22] . . . et Demooriti
   ex annis LXXXII cum intercalariis [nämlich mensibus] perinde
   Kallippos] viginti octo.</p>
<lb n="15"/> <p>13. Apollon. de pronom. p. 65, 15 Schneid. καὶ Φερεκύδης ἐν τῇι
   Θεολογίαι καὶ ἔτι Δ ἐν τοῖς Περὶ ἀστρονομίας καὶ ἐν τοῖς
   ὑπολειπομένοις συντάγμασι συνεχέστερον χρῶνται τῇι ἐμεῦ
   καὶ ἔτι τῇι ἐμέο. Vgl. B 29 a.</p>
<p>VII— IX. MATHEMATISCHE</p>
<p>11 r. [ VIII 3] WELTJAHR oder ASTRONOMIE SAMT STECKKALENDER.</p>
<p>12. Das Weltjahr Demokrits besteht aus 82 gewöhnlichen Jahren
   28 Schaltmonaten.</p>
<p>13. Meiner [ kontrahierte und unkontrahierte Form].</p>
<note type="footnote">2 ΓΝΩΜΗΣ] ΓΝΩΜΟΝΟΣ Cobet. Allmann Hennathena IV 206 meint,
   durch die Differenz des Gnomon sei er auf die Anfänge der Infinitesimalmethode
   worden, da die Atomistik der 'schen Monadenlehre verwandt sei:
   ΓωΝιΗο verm. Gromperz; die Überlieferung haltend übersetzt Über Verschiedenheit
   der Auffassung oder über Kreis- und Kugelberührung H. Vgl. Bibl.
   III. F., X (1910) 146. Er sieht darin eine Polemik gegen ' Angriff auf
   die Geometer [74 B 7] 7 Über verhältnislose (nicht irrationelle) Linien
   Atome erklärt H. Vgl a. O. 147; Hultsch verm. κλαστῶν Jahrb. f kl. Phil.
   579 8 vgl. Ptol. geogr. vii 7 ὑπογραφὴ τοῦ ἐκπετάσματος. ὑπογραφὴ δ' ἔσται
   καὶ τῆς τοιαύτης ἐκπετάσεως ἁρμόζουσά τε καὶ κεφαλαιώδης. ἡ τοιαύτη τῆς κρικωτῆς
   σφαίρας ἐπιπέδωι καταγραφή κτλ. Also Projektion der Armillarsphäre
   die Ebene 9 ΠΑΡΑΠΗΓΜΑ] »Steckkalender«, ein ehernes oder marmornes
   Verzeichnis der Tage des Sonnenjahres nach dem Zodiakus nebst den üblichen
   Episemasien ( Wettexeichen). Neben den Tagen befanden sich Löcher, in die
   Tage des bürgerlichen Monats eingesteckt werden konnten, S. *
   aus Milet, Berl. Sitz. B. 1904, 92 ; 266; Heron de Villefosse Comt. Rend.
   de l' Ac. des Inscript. 1898 p, 267. Das Parapegma des Meton und Euktemon
   (27. Juni 432) zeigt bereits genau die Einrichtung des Demokritischen</note>

<pb n="63"/>
<p>14. ÜBERRESTE DES PARAPEGMA DER ΑΣΤΡΟΝΟΜΙΗ.</p>
```

Abbildung 45. Spracherkennung bei Mehrsprachigkeit: Der griechische Text ist mit deutschen Kommentaren und Überschriften versehen. Einzelne Passagen oder Quellenangaben können auch lateinisch sein. XML-Auszug aus: Die Fragmente der Vorsokratiker, Berlin 1912. S. 62f. Open Greek and Latin Project.
URL: https://github.com/OpenGreekAndLatin/fragmentary-dev/blob/master/fragmenteVorsokratiker_2.xml

Sprachspezifische Verarbeitung

Teilaufgabe	Erläuterung
Spracherkennung	Die verwendeten Sprachen werden ermittelt (■ **Abbildung 45**).[17]
Segmentierung	Strukturiert den Text in einzelne Teile, die separat untersucht werden können. Üblich ist die Segmentierung in Sätze anhand der Satzzeichen.
Tokenisierung	Segmentiert auf der Basis der Wortebene in einzelne Teile (Token), indem beispielsweise das Leerzeichen als Wortgrenze aufgefasst wird.
Wortstammreduktion	Die Wörter werden auf ihren Wortstamm zurückgeführt, um bei einer späteren Suche auch Flexionen zu finden.
Lemmatisierung	Die Grundform eines Wortes (Lemma) wird gebildet.
Part-of-Speech Tagging	Zuordnung von Wörtern und Satzzeichen in Wortarten.
Parsing	Der Text wird in eine neue syntaktische Struktur überführt. Dabei ist für den Parser ein Token die atomare Eingabeeinheit.
Koreferenz (Referenzidentität) auflösen	Eine Koreferenz liegt vor, wenn sich innerhalb einer Äußerung zwei sprachliche Ausdrücke auf das gleiche linguistische Objekt beziehen, beispielsweise mittels Verwendung von Pronomen.
Eigennamenextraktion	Bei der Eigennamenerkennung, auch Named Entity Recognition (NER), werden die Begriffe eines Textes bestimmten Typen (z. B. Ort oder Person) zugeordnet.

17 Wenn diese in den Metadaten nicht annotiert sind, ist dies, gerade bei multilingualen Texten, ein nichttriviales Problem, welches häufig durch sprachspezifische (Stich-)Wortlisten gelöst wird.

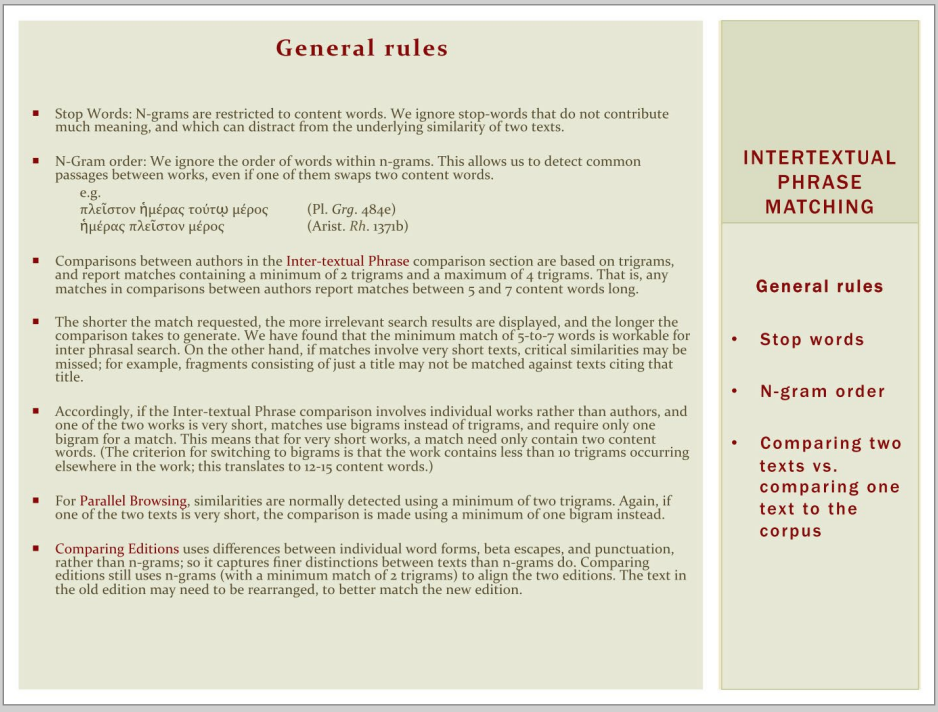


Abbildung 46. Regeln des Inter-textual Phrase-Matching beim TLG-Online[18]

18 URL: http://stephanus.tlg.uci.edu/help.php bzw. http://stephanus.tlg.uci.edu/helppdf/ngrams.pdf.

Parallelstelle, Zitat, Paraphrase, Kookkurrenz

Eine der verständlichsten, wenn auch häufig kontrovers diskutierten Methoden innerhalb der als Digital Humanities bezeichneten Teildisziplin klassischer Wissenschaftszweige ist die Parallelstellen- oder auch Zitationsanalyse. Genaugenommen ist noch nicht einmal die Frage geklärt, was ein Zitat denn letztlich ausmacht: Ist es der genaue Wortlaut oder reicht bereits eine synonyme Umschreibung des Inhalts? Diese eher begriffliche Betrachtungsweise soll nicht Gegenstand der Überlegungen sein. Wir wollen lediglich den Blick auf technische Teilaufgaben bei der Verarbeitung von Sprache richten.

Die Zitationsanalyse beschäftigt sich als Teilgebiet der Bibliometrie mit der qualitativen Untersuchung von zitierten und zitierenden Arbeiten. Im Ergebnis werden Regelmäßigkeiten und Strukturen eines Autors oder einer Autorengruppe aufgezeigt, im ungünstigen Fall führt es zu Plagiatsvorwürfen oder gar zur Aberkennung eines akademischen Grades, wenn in der Abschlussarbeit entsprechende Zitate als solche nicht kenntlich gemacht werden.

Suche über direkte Nachbarn – Nachbarschaftskookkurrenzen

Bei Zitationsanalysen, genaugenommen bei Ähnlichkeitsbestimmungen von Teiltexten, kommen sogenannte String-Matching-Algorithmen häufig zum Einsatz. Dies sind Verfahren, die unter Definition bestimmter Toleranzkriterien nach exakten Übereinstimmungen innerhalb von Texten suchen. Die Art und Weise, wie innerhalb der Zeichenketten (Strings) nach Treffern (Matches) gesucht wird, ist mitunter verschieden, im Ergebnis wird meist aufgrund von Häufigkeiten und Bewertungsmaßstäben (Signifikanzkriterien) geurteilt (■ **Abbildung 46**).

Eine relativ simple Möglichkeit besteht darin, die zu untersuchenden Texte auf wesentliche Terme zu reduzieren, indem beispielsweise häufig benutzte Worte, im Jargon auch Stoppworte (engl. Stopwords) genannt, zusammen mit den Satzzeichen herausgerechnet werden und das so reduzierte Gesamtkorpus auf Nachbarschaftskookkurrenzen, also das gemeinsame Auftreten von Wortgruppen, hin untersucht wird. Wenn anfänglich von einer simplen Möglichkeit gesprochen wurde, dann deshalb, weil hier gewisse linguistische Feinheiten außer Acht gelassen werden und nur der genaue Wortlaut ausschlaggebend ist. Beispielsweise können Synonyme oder Reflexivpronomen dabei nicht berücksichtigt werden. Ausschlaggebend ist die genaue Abfolge der Terme, wobei eines der schwierigsten Probleme dabei ist, innerhalb des Gesamtkontextes sowohl den Anfang als auch das Ende einer Parallelstelle zu finden. Es wird in dem Zusammenhang auch von „gierigen" Suchausdrücken gesprochen, die im Ergebnis eine größere Fundstelle liefern, als tatsächlich vorhanden.

$$sim = \frac{n_{ab} \times 2}{n_a + n_b}$$

Formel 1. Similar-Text.

Hier geben n_a und n_b jeweils die Länge der jeweiligen Zeichenketten und n_{ab} die Anzahl identischer Zeichen, also die Differenz zur Levenshtein-Distanz,[19] an.

$$sim = \frac{\big(max(n_a, n_b) - lev(a,b)\big) \times 2}{n_a + n_b}$$

Formel 2. Similar-Text mit Angabe der Levenshtein-Distanz

Beispiel: Similar-Text			$sim = \frac{(max(n_a, n_b) - lev(a,b)) \times 2}{n_a + n_b}$		
n_a	n_b	$lev(a,b)$	$max(n_a, n_b) - lev(a,b)$	sim	Similar-Text
a = Beispieltext 1 b = Beispiel Text 2					
14	15	3	12	$\frac{12 \times 2}{14 + 15} = \frac{24}{29}$	0,83
Zeichenkette a = The quick brown fox jumps over the lazy dog Zeichenkette b = The fox jumps over the lazy dog					
43	31	12	31	$\frac{31 \times 2}{43 + 31} = \frac{62}{74}$	0,84

Abbildung 47. Beispielberechnung Similar-Text

19 Eine von dem russischen Mathematiker Vladimir I. Levenshtein 1965 eingeführte Methode, zwei Zeichenketten zu vergleichen, indem die minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen gezählt wird, um die eine in die andere umzuwandeln.

Bewertung der Übereinstimmung von Parallelstellen

Wenn die Grenzen der (möglichen) Parallelstelle gefunden sind, muss der Grad der Übereinstimmung bewertet werden. Eine unkonventionelle Berechnungsmethode basiert auf der Editierdistanz oder auch Levenshtein-Distanz.[20] Sie besagt, wie viele Einfüge-, Lösch- und Ersetzoperationen notwendig sind, um eine Zeichenkette in eine andere zu verwandeln.

Der Vorteil eines solchen Signifikanzmaßes ist, dass es sich um einen absoluten Wert zwischen 0 und 1 (bzw. wahlweise eines Prozentwertes 0–100) handelt. Nachteilig bemerkbar macht sich hier die unterschiedliche Länge der zu untersuchenden Zeichenketten. Je größer der Längenunterschied, umso weniger signifikant ist die Parallelstelle, selbst bei exakter Übereinstimmung.

Die Parallelstellen werden in eAQUA schlussendlich unter Verwendung der Editierdistanz mit einem Similaritätswert belegt, der zwischen 0 = nicht identisch und 1 = vollständig identisch liegt. Berechnet wird nach einem Algorithmus Similar-Text, der bei Ian Oliver[21] mittels eines Pseudo-Codes beschrieben ist (■ **Formel 1**, ■ **Formel 2**).

Die berechneten Similaritätswerte beziehen sich immer auf die komplett tokenisierten Segmente, nicht allein nur auf die Suchmaske. Dies führt dazu, dass auch komplett identische Passagen mit einem von 1 abweichenden Wert belegt werden können, wenn sie innerhalb eines größeren Segments benutzt werden. Im zweiten Beispiel Similar-Text ergeben sich die Abweichungen lediglich durch den Einschub quick brown (■ **Abbildung 47**).

Similar-Text-Berechnungen sind nur bei kurzen Segmenten, wie der Satz-Tokenisierung in eAQUA, sinnvoll, da die Werte mit der Länge der untersuchten Segmente tendenziell abnehmen.

20 Vladimir I. Levenshtein: Binary codes capable of correcting deletions, insertions, and reversals. In: Doklady Akademii Nauk SSSR. Band 163, Nr. 4, 1965, S. 845–848. Englische Übersetzung: Soviet Physics Doklady, 10(8), 1966, S. 707–710.

21 Ian Oliver: Programming Classics: Implementing the World's Best Algorithms. Prentice Hall PTR New York, 1993.

Abbildung 48. N-Gramm basierte Suche im TLG-Online[22]

22 URL: http://stephanus.tlg.uci.edu.

n-Gramm basierte Suche

Eine Variante, die Suche nach Parallelstellen über einfache Vergleiche von Elementen der Texte zu realisieren, ist, die Wörter in eine Aneinanderreihung von Buchstabenkombinationen zu zerlegen – es wird hier von der Bildung von n-Grammen gesprochen. Das n steht für eine zu definierende Zahl. Gebräuchlich bei der Textverarbeitung sind Unigramm, Bigramm und Trigramm (■ **Abbildung 48**).

Das Verfahren für eine Suche würde exemplarisch wie folgt aussehen:

- Häufig benutzte Wörter (Stoppwörter) werden aus den Texten herausgerechnet.
- Beide Texte werden in einzelne Sequenzen zerlegt. Üblich können zum Beispiel ganze Sätze sein. Es ließen sich aber auch kleinere Einheiten bilden, indem Segmente mit fünf aufeinander folgenden Wörtern gebildet werden, unabhängig von Satzzeichen, die zuvor entfernt wurden.
- Die Sequenzen werden in einzelne Terme (Token) gesplittet.
- Für jeden Term werden Trigramme gebildet. Ist dieser zu kurz, bleibt er so, wie er ist, stehen.
- Die n-Gramme mit maximaler Größe von drei des einen Segments werden auf das Vorhandensein im zweiten Segment hin untersucht.
- Ist innerhalb eines Terms ein n-Gramm gefunden, wird dieser Term als Fundstelle markiert und der nächste Term genommen.
- Werden innerhalb des Segmentes insgesamt vier Treffer erzielt, gilt das Segment als Parallelstelle.
- Die Reihenfolge einzelner Fundstellen innerhalb der zu überprüfenden Segmente spielt keine Rolle.
- Anschließend wird die Ähnlichkeit beider Segmente bewertet, indem mittels Editierdistanz nach Levenshtein ein Prozentwert ermittelt wird.

Der Vorteil dieser Variante der Annäherung an die Parallelstellen ist, dass automatisch einige, mitunter aufwendige Verfahren der Normalisierung des Textes wegfallen. So kann, zumindest der Theorie nach, auf Stammwortreduktion oder Lemmatisierung verzichtet werden, da durch die Bildung von Trigrammen automatisch der Wortstamm erfasst würde. Denn vielfach werden Wörter nach einem einfachen Schema [Präfix] + Wortstamm + [Suffix] gebildet. Wobei Prä- oder Suffix optional sind. Deswegen sind sie in eckige Klammern gesetzt (in der IT insbesondere beim Programmieren werden optionale Parameter so gekennzeichnet). Ein Nachteil dieses Vorgehens wird aber auch sofort klar. Trigramme können auch Prä- oder Suffixe sein, so würde in der deutschen Sprache allein die Wortendung „ung“ zu einem Treffer führen, so wie beispielsweise in lateinischen Texten das Suffix *ion*, wodurch eine

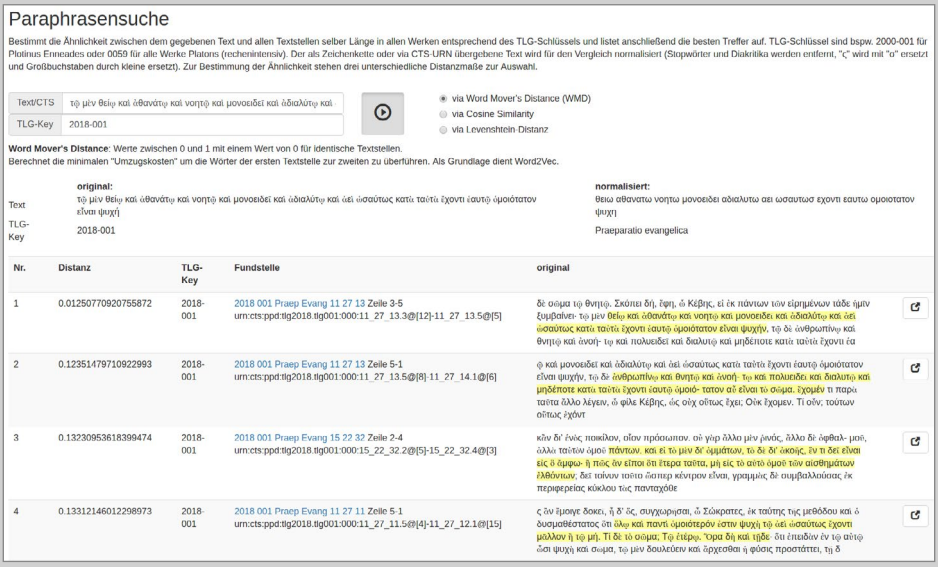


Abbildung 49. Paraphrasensuche mit der Word Mover's Distance

erheblich längere Liste von möglichen Parallelstellen zustande kommt, die über eine Similaritätsberechnung bewertet und abschließend reduziert werden muss.

Der Ansatz einer n-Gramm basierten Suche ist als ein erster halbwegs praktischer Versuch anzusehen, ohne Beachtung und Kenntnis der benutzten Sprache, rein über linguistische Statistik und ohne semantische Interpretation, zu Ergebnissen zu gelangen.

Vektorenbasierte Vergleiche

Einen weiteren Ansatz, auf computerlinguistische, und damit sprachspezifische, Vorverarbeitung weitgehend zu verzichten, bilden Verfahren auf Basis der distributionellen Hypothese.[23] Dabei wird von der Prämisse ausgegangen, dass Wörter, die in einem ähnlichen semantischen Kontext benutzt werden, auch eine ähnliche Bedeutung besitzen müssen. Die distributionelle Semantik repräsentiert Wortbedeutung mittels sogenannter Kontextvektoren, die eine statistische Verteilung eines Wortes über relevante sprachliche Kontexte erfassen. Mittels Verfahren aus der linearen Algebra können aus den Kontextvektoren semantische Ähnlichkeiten einzelner Wörter oder gar die Bedeutung komplexer Phrasen berechnet werden.

In dem durch die Volkswagen Stiftung geförderten Verbundprojekt „Digital Plato“,[24] welches sich der Untersuchung der Rezeption und Nachwirkung des platonischen Werkes in der griechischen Literatur bis zur Spätantike widmet und in dem einer der Projektpartner die Alte Geschichte der Universität Leipzig ist, werden beispielsweise mit Hilfe der Word Mover's Distance alle Textstellen eines ausgewählten Werkes mit der zu suchenden Passage verglichen und die ähnlichsten Treffer ausgegeben[25] (■ **Abbildung 49**).

23 Zellig S. Harris (1954): Distributional Structure, WORD, 10:2-3, 146–162, DOI: https://doi.org/10.1080/00437956.1954.11659520.

24 URL: https://digital-plato.org/.

25 Marcus Pöckelmann, Jörg Ritter, Eva Wöckener-Gade, Charlotte Schubert: Paraphrasensuche mittels word2vec und der Word Mover's Distance im Altgriechischen. In: Digital Classics Online, DCO 3,3 (2017), S. 24–36. DOI: https://doi.org/10.11588/dco.2017.0.40185.

Korpus	Anzahl Kookkurrenzen	Kookkurrenzen freq = 1	in Prozent
BTL[26]	137.486.214	110,876,836	80,65
MPL[27]	580.247.568	398.935.822	68,75
Perseus Shakespeare[28]	6.746.602	5.027.170	74,51
TLG[29]	355.021.014	258.961.566	72,94

Tabelle 6. Gesamtmenge von Kookkurrenzen diverser Korpora im Verhältnis zur Menge mit der Häufigkeit 1

26 Bibliotheca Teubneriana Latina, Online-Version, Stand Februar 2014.
27 Patrologia Latina Database, CD-ROM Version, November 1995c.
28 William Shakespeare in Perseus Digital Library, Renaissance Materials, Stand Mai 2013.
29 TLG-E, CD-ROM Version aus dem Jahre 1999.

Signifikanzmaße bei der Beurteilung von Kookkurrenzen

In der Statistik wird unter Signifikanz eine Kennzahl verstanden, welche die Wahrscheinlichkeit eines systematischen Zusammenhangs zwischen Variablen, im Falle von Textanalysen also zwischen Teiltexten (z. B. Wörtern), bezeichnet. Die Signifikanz drückt aus, ob ein scheinbarer Zusammenhang rein zufälliger Natur sein könnte oder mit hoher Wahrscheinlichkeit tatsächlich vorliegt.

Zur Berechnung werden abhängig vom Untersuchungsgegenstand unterschiedliche Formeln herangezogen, welche in erster Linie aus der Computerlinguistik stammen. Die Signifikanzmaße sollen dabei helfen, wichtige von unwichtigen Kookkurrenzen zu trennen. Dabei werden statistische Kenngrößen wie Korpusgröße, Häufigkeit der einzelnen Wörter oder Frequenz des gemeinsamen Auftretens ins Verhältnis gesetzt.

Eines der einfachsten Signifikanzmaße ist eine frequenzsortierte Kookkurrenzliste, die die Häufigkeit des gemeinsamen Auftretens zweier Worte im Gesamtkorpus angibt. Ein Nachteil frequenzsortierter Listen ist, dass nach dem Zipfschen Gesetz, dem Beginn der quantitativen Linguistik, sehr viele Wörter sehr selten auftreten. Demzufolge lassen sich mit einem Schwellenwert größer 1, also dem mehrmaligen gemeinsamen Auftreten eines Wortpaares, etwa zwei Drittel der Kookkurrenzen herausfiltern. Berechnet von den eAQUA-Tools sieht dies für ausgewählte Korpora wie in nebenstehender Tabelle aus (■ **Tabelle 6**).

Wie aus der kleinen Übersicht zu erkennen ist, sind ein Großteil der gefundenen Kookkurrenzen eher als niedrigfrequent zu bezeichnen. Um daraus die wichtigen zu filtern, sind Berechnungsmethoden erforderlich, von denen hier einige vorgestellt werden.

$$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$$

Formel 3. Dice

Beispiel: Dice	$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$
Bigramm	Trigramm
a = Tür b = Tor	
$a = \{\S T\ T\text{ü}\ \text{ür}\ r\S\}$ $b = \{\S T\ To\ or\ r\S\}$ $d_{Tür,Tor} = \frac{2 \times 2}{4 \times 4} = \frac{4}{8} = 0{,}5$	$a = \{\S\S T\ \S T\text{ü}\ Tür\ \text{ür}\S\ r\S\S\}$ $b = \{\S\S T\ \S To\ Tor\ or\S\ r\S\S\}$ $d_{Tür,Tor} = \frac{2 \times 2}{5 + 5} = \frac{4}{10} = 0{,}4$
a = Spiegel b = Spargel	
$a = \{\S S\ Sp\ pi\ ie\ eg\ ge\ el\ l\S\}$ $b = \{\S S\ Sp\ pa\ ar\ rg\ ge\ el\ l\S\}$ $d_{Spiegel,Spargel} = \frac{2 \times 5}{8 + 8} = \frac{10}{16} = 0{,}625$	$a = \{\S\S S\ \S Sp\ Spi\ pie\ ieg\ ege\ gel\ el\S\ l\S\S\}$ $b = \{\S\S S\ \S Sp\ Spa\ par\ \arg rge\ gel\ el\S\ l\S\S\}$ $d_{Spiegel,Spargel} = \frac{2 \times 5}{9 + 9} = \frac{10}{18} \approx 0{,}556$

Abbildung 50. Beispielberechnung Dice

Dice

Beim Dice-Koeffizienten[30] wird die Ähnlichkeit zweier Terme mittels einer Zahl zwischen 0 und 1 angegeben. Berechnungsgrundlage sind sogenannte n-Gramme. Ermittelt wird die Anzahl der n-Gramme, die in beiden Termen vorhanden sind, um diese ins Verhältnis zur Gesamtzahl der n-Gramme zu setzen.

Berechnet wird nach der nebenstehenden Formel (■ **Formel 3**), wobei n_{ab} die Schnittmenge beider Terme und n_a bzw. n_b die Anzahl der gebildeten n-Gramme pro Term angibt (■ **Abbildung 50**).

Bei der Bewertung von Kookkurrenzen kann der Dice-Koeffizient genutzt werden, indem die Häufigkeiten (Frequenzen) der Wörter ins Verhältnis gesetzt werden. n_a und n_b sind dabei die Frequenzen der Terme, n_{ab} die Anzahl des gemeinsamen Auftretens.

Aus der angeführten Formel ergeben sich relativ einfache Bewertungsmaßstäbe. Je frequenter die beiden Begriffe gemeinsam benutzt werden, umso mehr nähert sich der Wert 1. Treten beide Begriffe nur gemeinsam auf, wird die höchste Signifikanz mit 1 erreicht. Wie oft diese Kookkurrenz im Korpus zu finden ist, spielt dabei keine Rolle. Daraus ergibt sich eine wichtige Eigenschaft des Dice-Koeffizienten: Kookkurrenzen, die selten zusammen auftreten, bei denen ein Wort hoch- und das andere niedrigfrequent sind, werden als unsignifikant bewertet.

30 Auch als Sørensen-Dice-Koeffizient bezeichnet, benannt nach den Botanikern Thorvald Sørensen und Lee Raymond Dice.

$$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$$

Formel 4. Berechnung Jaccard-Koeffizient

Beispiel: Jaccard	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
Bigramm	Trigramm
a = Tür b = Tor	
$a = \{\S T\ T\text{ü}\ \text{ür}\ r\S\}$ $b = \{\S T\ To\ or\ r\S\}$ $j_{Tür,Tor} = \frac{2}{4+4-2} = \frac{2}{6} \approx 0{,}334$	$a = \{\S\S T\ \S T\text{ü}\ Tür\ \text{ür}\S\ r\S\S\}$ $b = \{\S\S T\ \S To\ Tor\ or\S\ r\S\S\}$ $j_{Tür,Tor} = \frac{2}{5+5-2} = \frac{2}{8} = 0{,}25$
a = Spiegel b = Spargel	
$a = \{\S S\ Sp\ pi\ ie\ eg\ ge\ el\ l\S\}$ $b = \{\S S\ Sp\ pa\ ar\ rg\ ge\ el\ l\S\}$ $j_{Spiegel,Spargel} = \frac{5}{8+8-5} = \frac{5}{11} \approx 0{,}455$	$a = \{\S\S S\ \S Sp\ Spi\ pie\ ieg\ ege\ gel\ el\S\ l\S\S\}$ $b = \{\S\S S\ \S Sp\ Spa\ par\ \text{arg}\ rge\ gel\ el\S\ l\S\S\}$ $j_{Spiegel,Spargel} = \frac{5}{9+9-5} = \frac{5}{13} \approx 0{,}385$

Abbildung 51. Beispielberechnung für den Jaccard-Koeffizienten

n_a	n_b	n_{ab}	Dice	Jaccard
100	100	1	0,01	0,005
100	100	10	0,1	0,05
100	100	50	0,5	0,33
100	100	90	0,9	0,82
100	100	100	1	1

Tabelle 7. Vergleich Dice- und Jaccard-Koeffizient bei 100 n-Grammen und verschiedenen Schnittmengen

Jaccard

Beim Jaccard-Koeffizienten (nach dem Botaniker Paul Jaccard) wird die Ähnlichkeit zweier Terme mittels einer Zahl zwischen 0 und 1 angegeben. Berechnungsgrundlage bei Text Mining-Verfahren sind sogenannte n-Gramme. Bei n-Grammen wird ein Term bzw. ein Text in gleich große Teile zerlegt. Diese Fragmente können Buchstaben, Phoneme, ganze Wörter oder ähnliches sein.

Ermittelt wird die Anzahl der n-Gramme, die in beiden Termen vorhanden sind, um diese ins Verhältnis zur Gesamtzahl der n-Gramme zu setzen (■ **Formel 4**). Berechnet wird nach der nebenstehenden Formel, wobei n_{ab} die Schnittmenge beider Terme und n_a bzw. n_b die Anzahl der gebildeten n-Gramme pro Term angibt (■ **Abbildung 51**).

Für die Bewertung von Kookkurrenzen gilt beim Jaccard-Koeffizienten ähnliches wie beim Dice-Koeffizienten. Beide berechnen den Signifikanzwert ähnlich, die relative Ordnung der Kookkurrenzen bleibt gleich, nur der absolute Signifikanzwert unterscheidet sich marginal.

Eine Modell-Berechnung mit mittlerer Frequenz von 100 ist in nebenstehender Tabelle vorgenommen (■ **Tabelle 7**). Bei einer Schnittmenge von 50 übereinstimmenden n-Grammen bei gleicher Länge der Ausdrücke von 100 n-Grammen wird der Dice-Koeffizient mit ½, der Jaccard-Koeffizient dagegen mit ⅓ errechnet.

$$poisson_{n,k} = \frac{1}{k!}\gamma^k \times e^{-\gamma}$$

Formel 5. Poisson-Verteilung

$$poisson\left(n_a, n_b, k, n\right) = \frac{k \times \left(\log k - \log \gamma - 1\right)}{\log n}$$

Formel 6. Poisson-Maß

$$\gamma = \frac{n_a \times n_b}{n}$$

Formel 7. Grundannahme vor der Umstellung

$$poisson = \frac{n_{ab} \times \log \frac{n_{ab} \times n}{n_a \times n_b} - n_{ab}}{\log n}$$

Formel 8. Berechnung Poisson-Maß

Poisson

Ein Ansatz zur Berechnung von signifikanten Kookkurrenzen basiert auf der Poisson-Verteilung,[31] einer diskreten Wahrscheinlichkeitsverteilung (■ **Formel 5**).

Auf der Basis der Poisson-Verteilung geben Quasthoff / Wolff[32] das Poisson-Maß mit der nebenstehenden Formel an, welche beispielsweise für Berechnungen von Korpora im Wortschatz-Portal[33] genutzt wurde, und in der die zwei Faktoren n (Anzahl der Sätze im Korpus) und k (Häufigkeit des gemeinsamen Auftretens, auch n_{ab} bezeichnet) maßgeblich sind (■ **Formel 6**).

Nach einer Umstellung und der Grundannahme ergibt sich schlussendlich die Berechnung (■ **Formel 7**, ■ **Formel 8**).

Somit ließe sich das Poisson-Maß auf die Differenz zwischen Local Mutual Information und Frequenz reduzieren.

31 Benannt nach dem Mathematiker Siméon Denis Poisson.

32 Uwe Quasthoff, Christian Wolff. The Poisson Collocation Measure and its Applications. In Second International Workshop on Computational Approaches to Collocations, 2002. URN: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:355-epub-68241. Vgl. ebenso: Gerhard Heyer, Uwe Quasthoff, Thomas Wittig. Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. Herdecke; Bochum: W3L-Verl. 2006. S. 338 ff.

33 URL: http://wortschatz.uni-leipzig.de/de.

$$p(K=k) = p^k (1-p)^{n-k} \binom{n}{k}$$

Formel 9. Binomialverteilung[34]

$$\begin{aligned} -2\log\lambda = &\left[\log L(p_1,k_1,n_1) + \log L(p_2,k_2,n_2)\right. \\ &\left. -\log L(p_1,k_1,n_1) - \log L(p_2,k_2,n_2)\right] \end{aligned}$$

Formel 10. Log likelihood[35]

$$\log L(p,n,k) = k\log p + (n-k)\log(1-p)$$

Formel 11. Log likelihood Voraussetzung[36]

34 Dunning, S. 64.
35 Dunning, S. 67.
36 Ebd.

Log-Likelihood

Eines der populärsten Signifikanzmaße bei der Analyse großer Textkorpora ist nach Dunning[37] das Log-Likelihood-Maß, welches auf der Binomialverteilung, einer der wichtigsten diskreten Wahrscheinlichkeitsverteilungen, basiert (■ **Formel 9**).

Dunning kommt schließlich bei der Berechnung von log likelihood zu der Formel unter der Log-Likelihood Voraussetzung (■ **Formel 10,** ■ **Formel 11**).

Charakteristisch für das Log-Likelihood-Maß ist, im Gegensatz beispielsweise zum Poisson-Maß, die Gleichbehandlung von signifikant häufigen und signifikant seltenen Ereignissen. So finden sich in den Digitalisaten vom TLG in der Version TLG-E bei rund 73,8 Millionen Wörtern etwa 1,3 Millionen Kookkurrenzen, die nur einmal auftreten und trotzdem mit einem lgl-Wert von 30 und ein wenig mehr belegt sind. Einen ähnlich großen Wert von 34,553 haben zum Beispiel καί und τό, die zusammen 14311 Mal gezählt wurden.

37 Ted Dunning: „Accurate Methods for the Statistics of Surprise and Coincidenc". In: Computational Linguistics 19, 1 (1993), 61–74. URL: http://aclweb.org/anthology/J/J93/J93-1003.pdf.

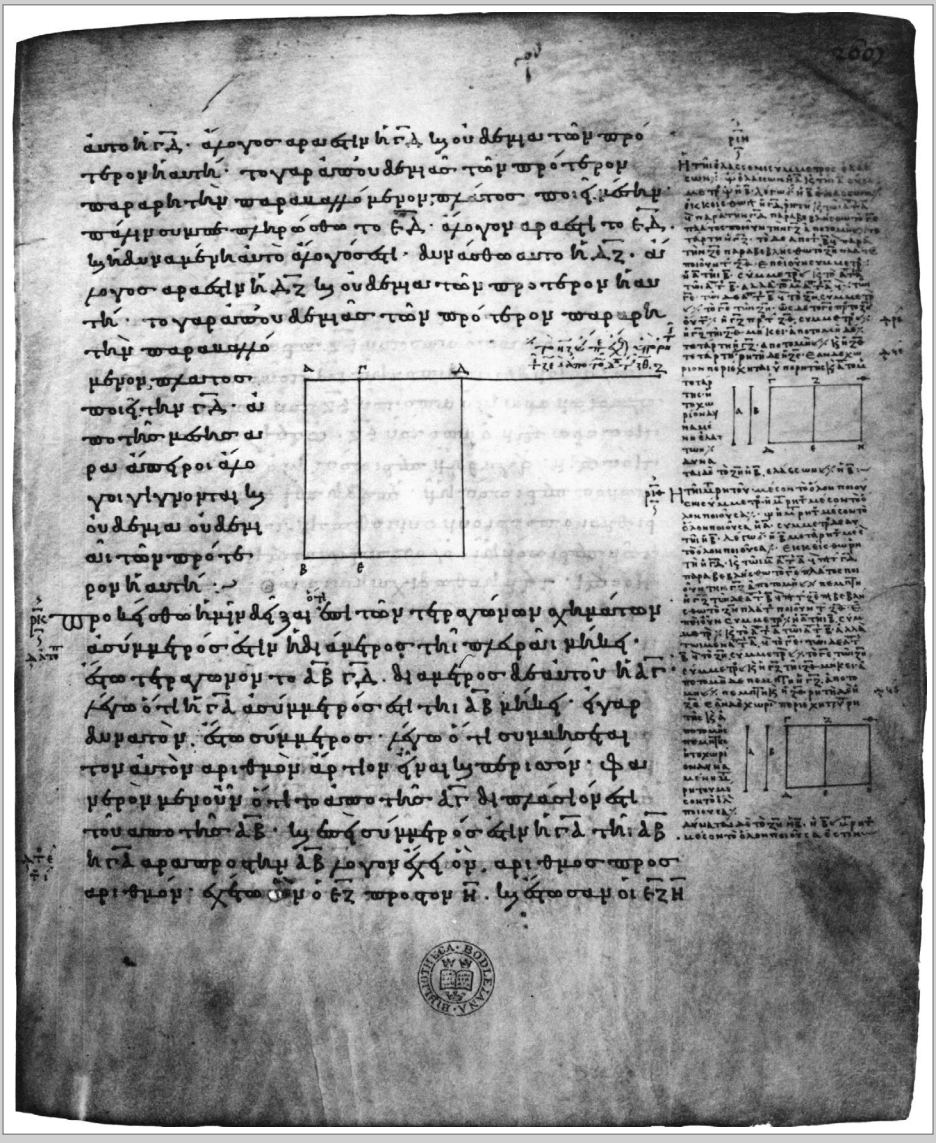


Abbildung 52. Handschrift der Elemente Euklids[38]

38 Diese älteste, erhaltene griechische Handschrift der Elemente wurde im September 888 vom Schreiber Stephanos Clericus fertiggestellt und von Arethas von Caesarea gekauft. Sie wird heute in der Bodleian Library (Oxford) aufbewahrt.
Euklid, Elemente 10, Appendix in der Handschrift Oxford, Bodleian Library, MS. D'Orville 301, fol. 268r.
Lizenz: Public Domain. Quelle: Wikimedia.

Glossar

Algorithmus

Algorithmen sind wesentliche Bestandteile der Informatik und der Mathematik. Sie beschreiben den Lösungsweg eines Problems oder einer Klasse von Problemen, indem eine endliche Anzahl von Anweisungen oder Prozeduren zur Durchführung bestimmter Aufgaben aneinandergereiht werden.

Der Algorithmusbegriff ist etymologisch arabischen Ursprungs, wurde aber historisch im Rahmen von Mathematik, Logik und Philosophie bereits im antiken Griechenland geprägt. Aus dem Altertum ist beispielsweise der mathematische Algorithmusbegriff des Euklid von Alexandria bekannt. Der griechische Mathematiker, der wahrscheinlich im 3. Jahrhundert v. Chr. in Alexandria wirkte, beschreibt in seinem Werk „Die Elemente“ (im Original Στοιχεῖα Stoicheia) ein Verfahren, mit dem sich der größte gemeinsame Teiler zweier natürlicher Zahlen berechnen lässt (■ **Abbildung 52**). Aus dieser Zeit ist etwa auch das Sieb des Eratosthenes, ein Verfahren zur Ermittlung einer Liste von Primzahlen innerhalb eines vorgegebenen Wertebereichs, überliefert.

Apache-Lizenz

Die Apache-Lizenz ist eine Freie-Software-Lizenz der Apache Software Foundation. Sie besitzt keinen Copyleft-Vermerk, dies bedeutet, sie verzichtet auf die Pflicht des Lizenznehmers, Bearbeitungen des Werkes unter dieselbe Lizenz wie die des ursprünglichen Werkes zu stellen. Die Apache-Lizenz ist von der Free Software Foundation, einer gemeinnützigen Organisation zur Förderung freier Software, als Lizenz für freie Software anerkannt.

RFC 20 ASCII format for Network Interchange October 1969

2. Standard Code

B I T S	b7	0	0	0	0	1	1	1	1
	b6	0	0	1	1	0	0	1	1
	b5	0	1	0	1	0	1	0	1
	COLUMN	0	1	2	3	4	5	6	7
b4	b3	b2	b1	ROW					
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

Abbildung 53. RFC 20, page 1[39]

39 URL: https://tools.ietf.org/html/rfc20.

ASCII

Der American Standard Code for Information Interchange ist eine 7-Bit-Zeichenkodierung, die 128 Zeichen definiert, bestehend aus 33 nicht druckbaren Steuerzeichen sowie 95 druckbaren Zeichen. Die ersten 32 Codes sind für Steuerzeichen (control character) reserviert. Sie sind historisch begründet und dienen beispielsweise dem Wagenrücklauf (Drucker) oder Zeilenumbruch. Das letzte Zeichen mit dem Code 127 (DEL für Delete) diente ursprünglich dazu, falsch gestanzte Zeichen auf Lochstreifen zu entfernen, indem alle sieben Bits ausgestanzt wurden.

Die druckbaren Zeichen umfassen das lateinische Alphabet in Groß- und Kleinschreibung, die zehn arabischen Ziffern sowie einige Interpunktions- und andere Sonderzeichen, also im Wesentlich den Zeichenvorrat, der auf einer englischen Tastatur zu finden ist.

Aus diesem Grund musste zu Beginn des Computerzeitalters altgriechischer Text mit Hilfe des sogenannten Beta Code transkribiert werden. Die meisten nachfolgend entwickelten Zeichenkodierungen sind so konzipiert, dass sie für Zeichen zwischen 0 und 127 den gleichen Code verwenden wie ASCII und den Bereich über 127 für weitere Zeichen (■ **Abbildung 53**).

Beta-Code Altgriechisch

Griechischer Beta Code ist die 7-Bit-sichere Kodierung mittels des US-ASCII-Zeichensatzes. Jedes diakritische Zeichen wird durch ein eigenes Zeichen dargestellt, welches dem Buchstaben folgt (Ausnahme: bei Großbuchstaben vor dem Buchstaben). Beta Code unterscheidet nicht zwischen Klein-/Großschreibung, Großbuchstaben werden durch Voranstellung eines * Asterisks (griech. ἀστερίσκος) gekennzeichnet. Einige Projekte benutzen nur Großbuchstaben (z. B. TLG), andere nur Kleinbuchstaben (z. B. das Perseus Project).

ἀστερίσκος in Beta Code Altgriechisch: a)steri/skos

Big Data

Analyse großer Datenmengen aus verschiedenen Quellen mit dem Ziel, wirtschaftlichen Nutzen daraus zu erzeugen.

Bigramm

Zwei aufeinander folgende Wortformen oder Buchstaben werden als spezieller Typ von n-Grammen mit Bigramm bezeichnet.

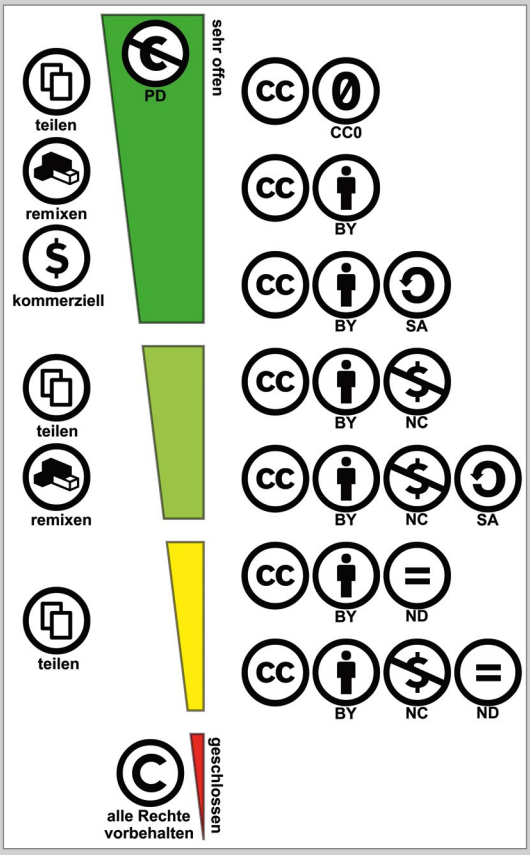


Abbildung 54. CC-Lizenzmodule können miteinander kombiniert werden[40]

40 Von JoeranDE – Creative commons license spectrum.svg by Shaddim, Gemeinfrei, https://commons.wikimedia.org/w/index.php?curid=60988847.

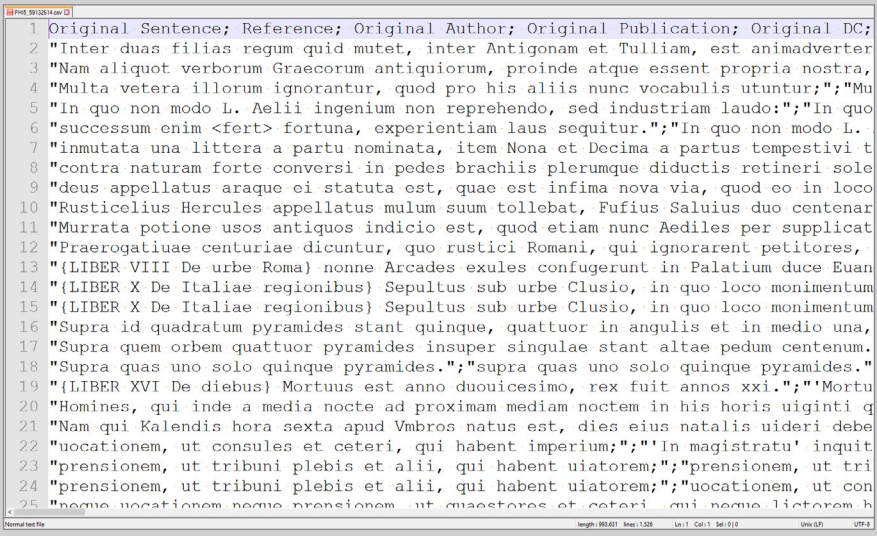
Browser
Ein Computerprogramm, im neuen Sprachgebrauch oftmals auch App (engl. Abkürzung von Application software, deutsch: Anwendungssoftware) bezeichnet, welches speziell zur Darstellung von Webseiten konstruiert wurde, ist, dem englischen Verb browsen (stöbern, umsehen, schmökern) entlehnt, ab 1989 als Webbrowser bezeichnet worden. Zu Beginn zeigten Browser lediglich Text zum Lesen an, später kamen dann Funktionen zum Anzeigen von Bildern, dem Navigieren zwischen Webseiten (Hyperlinks) und der Ausgabe von audiovisuellen Medien hinzu. Bekannte aktuelle Browser sind beispielsweise Google Chrome, Internet Explorer (Microsoft), Mozilla Firefox, Microsoft Edge, Apple Safari, Opera oder Vivaldi.

BTL
Bei der Bibliotheca Teubneriana Latina handelt es sich um eine Sammlung lateinischer Literatur von der Römischen Republik bis zur Kaiserzeit und Spätantike. Es ist die elektronische Version der lateinischen Texte aus der Bibliotheca scriptorum Graecorum et Romanorum Teubneriana, eine 1849 begründete Schriftenreihe, die im Leipziger Verlag B.G. Teubner erschien.

CC
Unter dem Begriff Creative Commons (CC) wird eine Sammlung von Lizenzen verstanden, mit denen ein Autor Nutzungsrechte für sein Werk einräumen kann. Durch die Kombination der Rechtemodule

- by (Attribution) Namensnennung
- nc (Non-Commercial) Nicht kommerziell
- nd (No Derivatives) Keine Bearbeitung
- sa (Share Alike) Weitergabe unter gleichen Bedingungen

kann die Freigabe nach den Wünschen des Urhebers abgestuft werden (■ **Abbildung 54**).



```
Original Sentence; Reference; Original Author; Original Publication; Original DC;
"Inter duas filias regum quid mutet, inter Antigonam et Tulliam, est animadverter
"Nam aliquot verborum Graecorum antiquiorum, proinde atque essent propria nostra,
"Multa vetera illorum ignorantur, quod pro his aliis nunc vocabulis utuntur;";"Mu
"In quo non modo L. Aelii ingenium non reprehendo, sed industriam laudo:";"In quo
"successum enim <fert> fortuna, experientiam laus sequitur.";"In quo non modo L.
"inmutata una littera a partu nominata, item Nona et Decima a partus tempestivi t
"contra naturam forte conversi in pedes brachiis plerumque diductis retineri sole
"deus appellatus araque ei statuta est, quae est infima nova via, quod eo in loco
"Rusticelius Hercules appellatus mulum suum tollebat, Fufius Saluius duo centenar
"Murrata potione usos antiquos indicio est, quod etiam nunc Aediles per supplicat
"Praerogatiuae centuriae dicuntur, quo rustici Romani, qui ignorarent petitores,
"{LIBER VIII De urbe Roma} nonne Arcades exules confugerunt in Palatium duce Euan
"{LIBER X De Italiae regionibus} Sepultus sub urbe Clusio, in quo loco monimentum
"{LIBER X De Italiae regionibus} Sepultus sub urbe Clusio, in quo loco monimentum
"Supra id quadratum pyramides stant quinque, quattuor in angulis et in medio una,
"Supra quem orbem quattuor pyramides insuper singulae stant altae pedum centenum.
"Supra quas uno solo quinque pyramides.";"supra quas uno solo quinque pyramides."
"{LIBER XVI De diebus} Mortuus est anno duouicesimo, rex fuit annos xxi.";"'Mortu
"Homines, qui inde a media nocte ad proximam mediam noctem in his horis uiginti q
"Nam qui Kalendis hora sexta apud Vmbros natus est, dies eius natalis uideri debe
"uocationem, ut consules et ceteri, qui habent imperium;";"'In magistratu' inquit
"prensionem, ut tribuni plebis et alii, qui habent uiatorem;";"prensionem, ut tri
"prensionem, ut tribuni plebis et alii, qui habent uiatorem;";"uocationem, ut con
"neque uocationem neque prensionem, ut quaestores et ceteri, qui neque lictorem h
```

Abbildung 55. Daten in Tabellenform in einer CSV-Datei. Spalten werden durch Semikolon getrennt. Anführungszeichen begrenzen Textfelder.

Cookie

Als Cookies werden in der Informatik kleine Datenpakete bezeichnet, die zwischen Computerprogrammen ausgetauscht werden. Eine frühe Verwendung des Begriffs Magisches Cookie ist in einer Routine der C-Standardbibliothek fseek mit dem Jahr 1979 datiert. Im aktuellen Sprachgebrauch wird der Begriff synonym für HTTP-Cookie verwendet. Diese speichern Informationen in kleinen Textdateien auf dem Rechner eines Anwenders, um sie bei Bedarf wieder an den Server zu übermitteln. Damit lassen sich Webseiten individualisierbar gestalten und Authentifizierungen realisieren, weil das zugrundeliegende HTTP als zustandsloses Protokoll solche Möglichkeiten nicht vorsieht.

Copyleft

Als Copyleft wird eine Klausel in Nutzungslizenzen bezeichnet, die festlegt, dass alle Änderungen an einem Werk nur dann statthaft sind, wenn sie im Wesentlichen unter den gleichen Lizenzbedingungen verbreitet werden.

CSV

Das textbasierte Dateiformat CSV (Comma-separated values) ist eine Form von DSV (Delimiter-separated values). Die Daten sind in Tabellenform, also zweidimensional, gespeichert. Jede Zeile ist ein Datensatz. Felder werden mittels Komma oder Semikolon separiert (■ **Abbildung 55**).

CTS

Das Notationssystem CTS (Canoncial Text Services, entwickelt von Christopher Blackwell und Neel Smith[41], weiterentwickelt von Hannes Kahl[42]) als Teil der CITE Architektur bietet einen netzbasierten Service zur Identifikation klassischer Texte basierend auf URN. CTS URNs sind in fünf Teile untergliedert, die von Doppelpunkten voneinander getrennt sind:

urn:ctn:ctnNameSpace:WorkIdentifier:PassageIdentifier.

41 URL: http://www.homermultitext.org.
42 URL: https://github.com/ecomp-shONgit/ecomparatio.

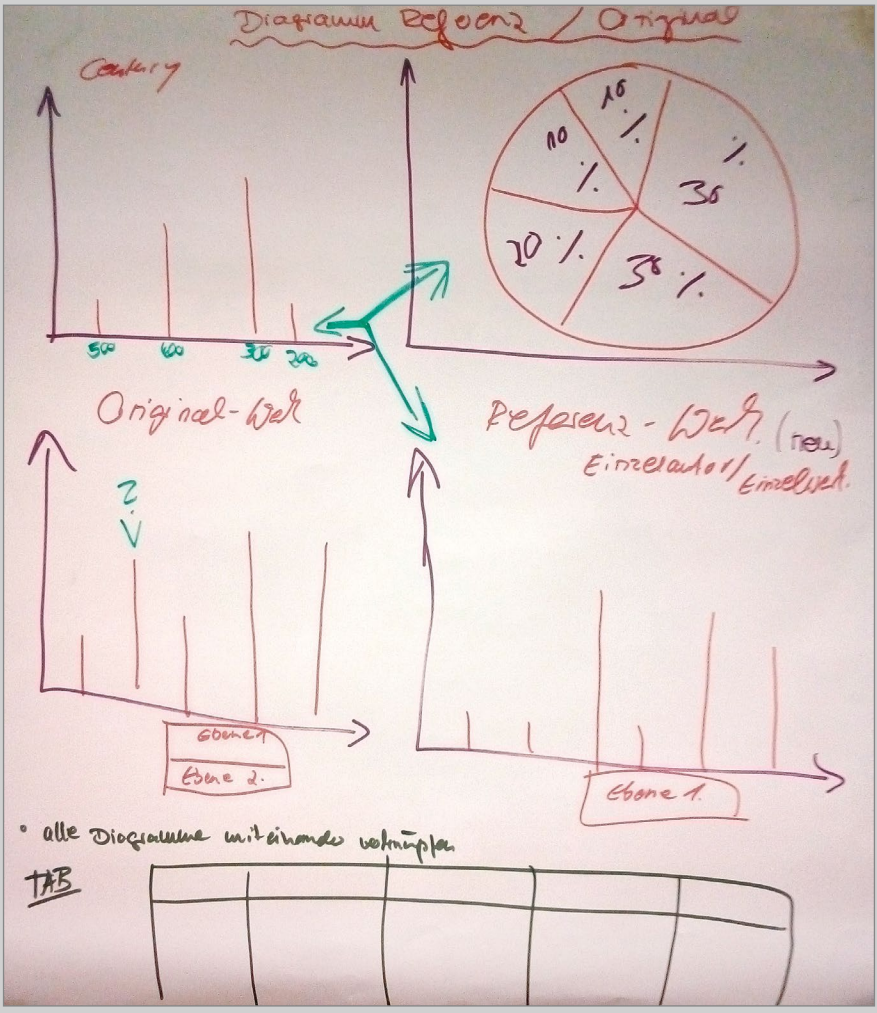


Abbildung 56. eAQUA – Entwurf für eine neue Bedienoberfläche bei der Parallelstellensuche

Digitalisierung

Mit der Digitalisierung von Texten werden allgemein zwei Verfahren bezeichnet, die unabhängig voneinander funktionieren können. Zum einen wird damit die Praxis bezeichnet, ein originalgetreues Abbild eines Dokumentes mittels Scanner oder Fotografie anzufertigen. Elektronische Abbilder von Dokumenten, die in Dokumentenmanagement-, Archiv- oder Enterprise-Content-Management-Systemen eingepflegt werden, sind oftmals auch als Faksimile bezeichnet.

Weiterhin ist damit die Arbeitsweise umrissen, ursprünglich in analoger Form vorliegende Texte, beispielsweise Bücher, Handschriften, Papyri, in einen elektronischen Zeichensatz zu übertragen, der nur den sprachlichen Inhalt erfasst und ihn damit reproduzierbar, übertragbar und analysierbar macht. Dazu werden Texterkennungsprogramme und die sogenannte OCR-Technik benutzt.

DOI

Digital Object Identifier (DOI) werden seit 1998 durch die International DOI Foundation (IDF) koordiniert. Mit DOI können sowohl physische, digitale als auch abstrakte Objekte dauerhaft eindeutig identifiziert und lokalisiert werden. Dem Schema, welches immer mit 10 beginnt, wird zur Identifikation die Bezeichnung doi vorangestellt: doi:10.ORGANISATION/ID. Bei Internetadressen wird der DOI-Resolver („https://doi.org/“) in Form einer URL vorangestellt.

Ein Beispiel:
Ch. Schubert (Hg.): Working Papers Contested Order (NO. 10): Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung V

DOI: https://doi.org/10.11588/ea.2013.2

eAQUA

Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft war ein vom Bundesministerium für Bildung und Forschung im Zeitraum 2008–2013 im Rahmen der eHumanties-Initiativen gefördertes Projekt der Digital Humanities an der Universität Leipzig. Fachspezifische Digitalisate in den historischen Sprachen Griechisch und Latein, wie sie beispielsweise in den Editionen des Thesaurus Linguae Graecae (TLG), des Packard Humanities Institute (PHI), der Bibliotheca Teubneriana Latina (BTL) oder Digitalisierungsprojekten wie der Perseus Digital Library vorkommen, wurden hinsichtlich semantischer Zusammenhänge, lokaler oder chronologischer Abhängigkeiten und Einflüsse systematisch algorithmusgesteuert untersucht (■ **Abbildung 56**).

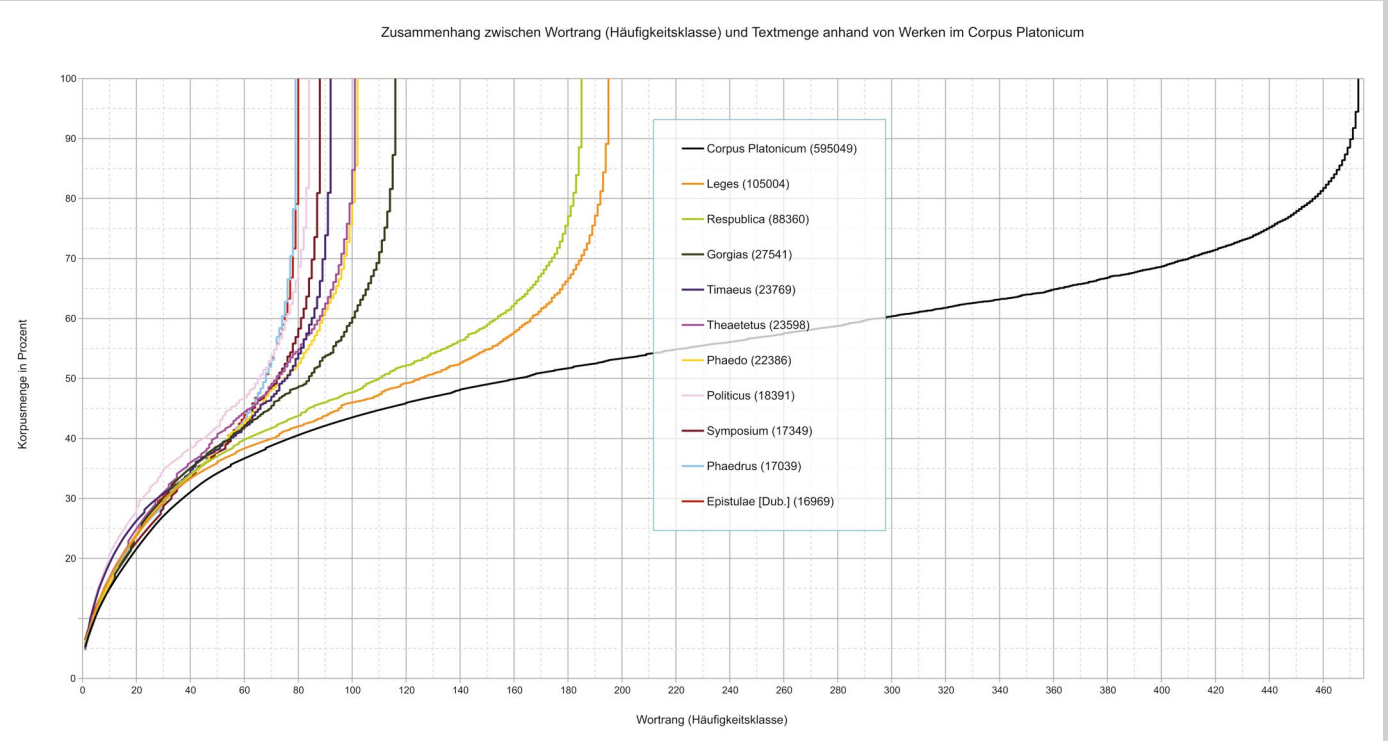


Abbildung 57. Häufigkeitsverteilung von Wörtern in ausgewählten Texten Platons

Editierdistanz

siehe ▶ Levenshtein-Distanz

Entropie

Entropie in der Informationstheorie gibt an, wieviel Bits im Durchschnitt benötigt werden, um einen Wert einer Zufallsvariablen als ein Ereignis (als Teil einer Nachricht) zu codieren. Je mehr Bits benötigt werden, desto höher ist die Entropie und umso schwieriger die Vorhersagen eines Ereignisses.

GPL

Die GNU General Public License (auch GPL oder GNU GPL) ist eine Lizenz, die es erlaubt, eine Software kostenlos zu nutzen, zu verbreiten, zu studieren oder auch zu verändern. Alle von der Software abgeleiteten Programme müssen ebenfalls zu den Bedingungen der GPL lizenziert werden (Copyleft).

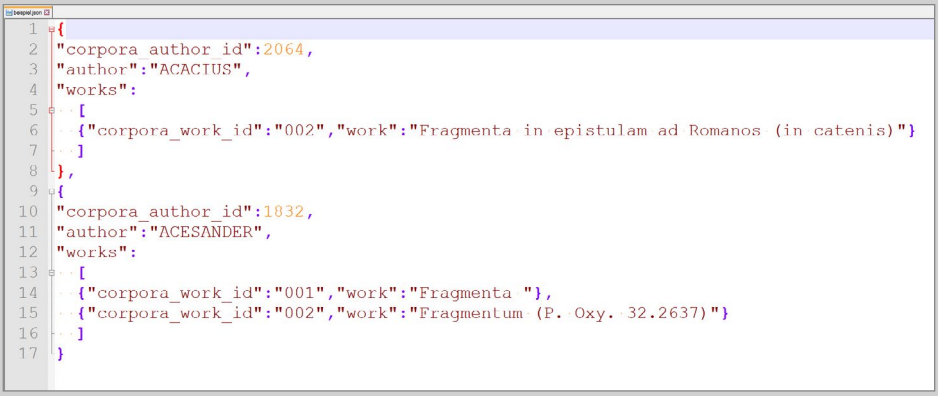
Häufigkeitsklasse

Eine Häufigkeitsklasse ist die Einteilung von Wortformen in Gruppen nach ihrer Frequenz (Häufigkeit) im Korpus.

Häufigkeitsverteilung

Die Häufigkeitsverteilung ist in der deskriptiven (beschreibenden) Statistik eine Funktion, die zu jedem möglichen Wert angibt, wie oft dieser vorgekommen ist. So lassen sich beispielsweise die benutzen Wörter innerhalb von Texten zählen und deren Häufigkeit in Bezug zur Gesamtmenge ermitteln[43] (■ **Abbildung 57**). Wörter, die gleich oft benutzt wurden, können dann in einzelne Klassen (Wortrang) eingeteilt werden. Solche Verteilungen lassen sich als Tabelle, als Grafik oder modellhaft als Funktionsgleichung darstellen. Die Häufigkeitsverteilung hat in der Wahrscheinlichkeitstheorie eine Entsprechung in der Wahrscheinlichkeitsverteilung.

43 Die grafische Darstellung der Häufigkeitsverteilung der benutzten Wörter in Bezug zur Gesamtmenge innerhalb ausgewählter Texte Platons. Es werden aufsteigende Häufigkeitsklassen gebildet. Das häufigste Wort erhält die 1, das nächste die 2 usw. Bei gleicher Frequenz teilen sich mehrere Wörter den Rang, wodurch der gerade Verlauf am Ende Kurve (Frequenz von 1) erklärt wird.



```
{
"corpora_author_id":2064,
"author":"ACACIUS",
"works":
  [
  {"corpora_work_id":"002","work":"Fragmenta in epistulam ad Romanos (in catenis)"}
  ]
},
{
"corpora_author_id":1832,
"author":"ACESANDER",
"works":
  [
  {"corpora_work_id":"001","work":"Fragmenta "},
  {"corpora_work_id":"002","work":"Fragmentum (P. Oxy. 32.2637)"}
  ]
}
```

Abbildung 58. Auszug von TLG-Metadaten in JSON-Notierung

	Deutsches Textarchiv - Grimms Märchen: König [146] - Häufigkeit: 519 ?
Wörter mit ähnlichem Zusammenhang:	**seine** [162]; **der** [103]; **sey** [324]; **er** [105]; **Tochter** [250]; , [12]; **dem** [118]; **daß** [125]; **Prinzessin** [185]; **hatte** [142]; **ihm** [124]; **aber** [112]; **nun** [147]; **ward** [164]; . [14]; **und** [101]; **als** [134]; **sollte** [217]; **wollte** [153]; **ließ** [197]; **wäre** [230]; **ihr** [130]; **Frau** [179]; **das** [106]; **zu** [115]; **mit** [122]; **erzählte** [513]; **von** [145]; **nach** [156]; **auch** [141]; **ihn** [131]; **die** [102]; **da** [113]; **sein** [172]; **Vater** [188]; **wie** [128]; **vor** [154]; **sich** [116]; **es** [109]; **gab** [225]; **keine** [266]; **ein** [111]; **aus** [155]; **Der** [144]; **Reich** [565]; **nichts** [182]; **ihre** [222]; **sah** [165]; **war** [114]; **nicht** [117]; **auf** [120]; **sie** [104]; **sagte** [132]; **wieder** [139]; **Gemahlin** [527]; **seinen** [206]; **kam** [140]; **noch** [149]; **in** [108]; **eine** [133]; **Es** [261]; **Königin** [239]; **im** [161]; **alles** [176]; **haben** [195]; **des** [209]; **Braut** [317]; **Schneider** [368]; **einen** [135]; **wär** [316]; **so** [110]; **doch** [171]; **waren** [167]; **sprach** [136]; **ging** [137]; **Prinz** [257]; **Da** [123]; **den** [107]; **machen** [328]; **zur** [232]; : [16]; **an** [127]; **drei** [175]; **bekannt** [1248]; **wenn** [151];
Signifikante Kookkurrenzen: Network [Frequenz] Network [Log-Likelihood]	**der** (344); **dem** (159); **Tochter** (47); **seine** (71); **und** (384); **Der** (98); **er** (223); **Königin** (41); **Droßelbart** (10); **Prinzessin** (51); **Gemahlin** (20); **befahl** (16); **alte** (35); **ließ** (43); **hatte** (81); **Hof** (19); **Reich** (17); **schickte** (11); **seiner** (27); **Grafen** (8); **Braut** (24); **sollte** (32); **sagte** (87); **daß** (95); **vor** (53); **ward** (48); **zu** (127); **zum** (42); **brachte** (18); **werden** (29); **ihm** (89); **die** (219); **Kater** (12); **verlangte** (10); **heirathen** (10); **solle** (15); **Küchenjungen** (4); **vermählt** (5); **genommen** (8); **wem** (8); ' (28); **sein** (41); **zur** (27); **Land** (14); **Jäger** (13); **glich** (4); **denselben** (4); **vermählte** (4); **verirrte** (4); **Brodsuppe** (4); **befohlen** (5); **Töchter** (7); **Deck** (3); **bestäubt** (3); **hättst** (3); **Denkt** (3); **anstellen** (3); **näßt** (3); **regierte** (3); **Julian** (3); **Falken** (3); **vollbracht** (3); **diese** (9); **Liebe** (6); **gehört** (12); **alten** (14); **lieb** (11); **Schwiegermutter** (4); **Urtheil** (4); **offenbarte** (4); **Centner** (4); **habe** (23); **Schloß** (24); **sang** (5); **geholt** (5); **einzige** (5); **glaubte** (8); **gehalten** (8); **versprochen** (9); **Gold** (16); **Eselein** (6); **haben** (29); **Diener** (10); **könnte** (11); **als** (66); **gefangen** (5); **wäre** (24); **Hauptmann** (4); **stumm** (4); **schönsten** (4); **zart** (4); **nach** (43); **eine** (67); **Kriegsmann** (3); **rußiger** (3); **Schwesterlein** (3); **Wänden** (3); **Waldschloß** (3); **Spinnräder** (3); **Rebhühner** (3);

Abbildung 59. Signifikante Kookkurrenzen zum Wort König bei den Märchen der Gebrüder Grimm[44]

44 Grimm, Jacob; Grimm, Wilhelm: Kinder- und Haus-Märchen. Bd. 1. Berlin, 1812. URN: urn:nbn:de:kobv:b4-200905191950.
Grimm, Jacob; Grimm, Wilhelm: Kinder- und Haus-Märchen. Bd. 2. Berlin, 1815. URN: urn:nbn:de:kobv:b4-200905191965.

HTML
Hypertext Markup Language ist eine textbasierte Auszeichnungssprache zur strukturierten Darstellung von Inhalten in elektronischen Dokumenten.

JPEG
Verschiedene Methoden der Bildkompression, die vom Gremium Joint Photographic Experts Group 1992 in Form einer Norm vorgestellt wurden, werden unter dem Begriff JPEG zusammengefasst.

JSON
JavaScript Object Notation ist ein kompaktes Datenformat, welches zur Übertragung von Daten zwischen Client und Server konzipiert wurde (■ **Abbildung 58**).

Kollokation
Eine genaue Definition des Wortes ist selbst unter Linguisten umstritten. Oftmals ist von einer charakteristischen, häufig auftretenden Wortverbindung die Rede. Das gemeinsame Auftreten der Wörter beruhe auf der Regelhaftigkeit gegenseitiger Erwartbarkeit, sei also semantisch begründet. Oft wird das Wort auch synonym zu Kookkurrenz benutzt, obgleich nicht jede Kookkurrenz automatisch eine Kollokation ist. Wir verzichten hier auf die Verwendung des Begriffs und benutzen nur das eher statistisch (nicht semantisch) geprägte Wort Kookkurrenz.

Kookkurrenz
Für den Begriff gibt es sowohl einen eng als auch einen weit gefassten Sinn. Im weiteren Sinne wird das gemeinsame Auftreten zweier lexikalischer Einheiten, z. B. Wörter, innerhalb eines übergeordneten Segmentes, z. B. Satz, in der Allgemeinen Linguistik als Kookkurrenz bezeichnet (■ **Abbildung 59**). Im engeren Sinn in der Korpuslinguistik ist dafür noch ein statistisches Merkmal notwendig: die Einheiten sollten signifikant häufiger zusammen auftreten, als es ihre kombinierte individuelle Auftretenshäufigkeit erwarten ließe.

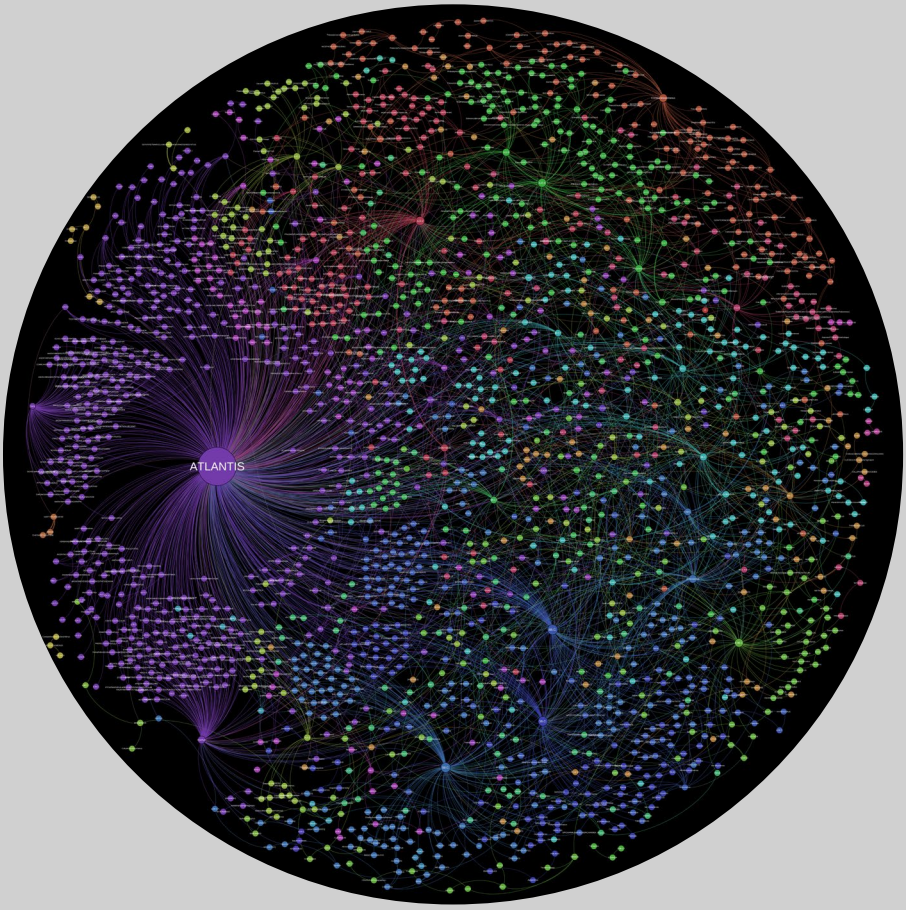


Abbildung 60. Mit Gephi erstellte Visualisierung auf der Basis des Metadatensatzes (Autorennamen, Orte, Epochen) des TLG-E.[45]

45 In: Ch. Schubert, Digital Humanities: Laboratorium der Geisteswissenschaften oder Weg nach Atlantis? Aus: Musikgeschichte zwischen Ost und West: von der ›musica sacra‹ bis zur Kunstreligion. Festschrift für Helmut Loos zum 65. Geburtstag, hrsg. v. Stefan Keym und Stephan Wünsche. Leipziger Universitätsverlag, Leipzig 2015, S. 747–758, ISBN 978-3-86583-958-9. URN: urn:nbn:de:bsz:16-propylaeumdok-25032.

Korpus
Korpus ist die Kurzform von Textkorpus und bezeichnet eine Sammlung von Texten.

Konkordanz
Unter Konkordanzen werden traditionell alphabetisch geordnete Listen von Wörtern oder Phrasen verstanden, die in einem Werk zur Verwendung kamen. Ursprünglich wurden solche Listen per Hand erstellt, waren dementsprechend zeitaufwendig, und wurden deshalb nur für vermeintlich wichtige Werke, wie religiöse Texte oder Werke angesehener Schriftsteller, erzeugt. Synonym zu Konkordanz werden auch die Ausdrücke Register, Index oder Key Word in Context (KWIC) benutzt.

Lemmatisierung
Reduktion auf die Grundform eines Wortes, also diejenige Form, unter der der Begriff in einem Nachschlagewerk zu finden ist.

Levenshtein-Distanz
Anzahl von Einfüge-, Lösch- und Ersetz-Operationen, um eine Zeichenkette in eine andere zu verwandeln.

Markup
Eine Markup Language (ML) oder Auszeichnungssprache beschreibt den Inhalt eines Dokumentes oder das Verfahren, welches zur Verarbeitung der Daten notwendig ist. HTML, XML oder LaTeX sind Auszeichnungssprachen.

Metadaten
Metadaten oder auch Metainformationen sind allgemein Daten, die Informationen über Merkmale beinhalten, die nicht Bestandteil der Daten selbst sind (■ **Abbildung 60**). Bei einer Korpusanalyse werden z. B. alle bibliographischen Informationen als Metadaten behandelt.

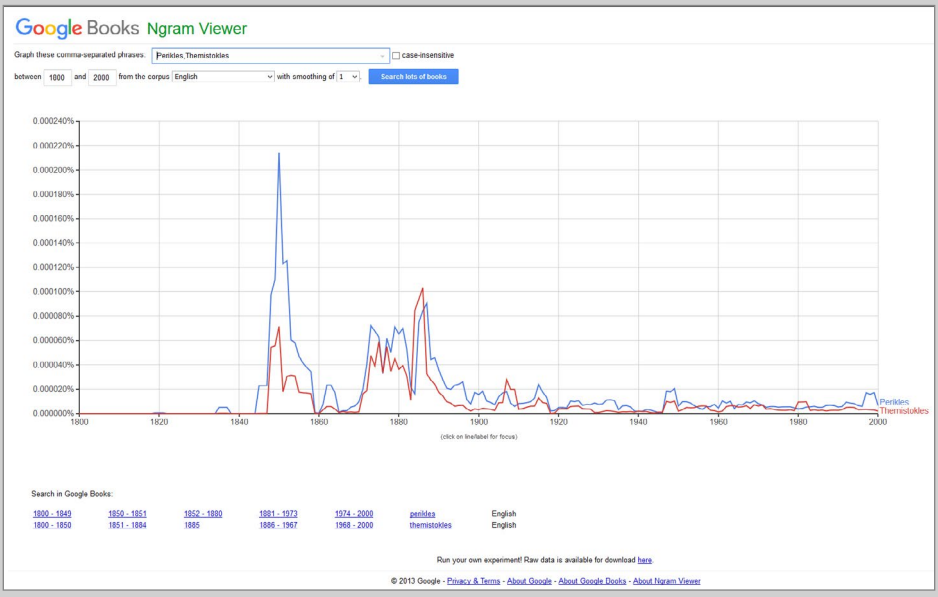


Abbildung 61. Google Books Ngram Viewer[46]

46 Suche nach Perikles und Themistokles in Google Books im Korpus English. URL: https://books.google.com/ngrams/graph?content=Perikles%2CThemistokles&year_start=1800&year_end=2000&corpus=15&smoothing=1.

MIT-Lizenz

Die MIT-Lizenz (auch X-Lizenz oder X11-Lizenz) ist eine aus dem Massachusetts Institute of Technology stammende Lizenz für die Software-Benutzung, die erlaubt, die Software zu verwenden, zu kopieren, zu ändern, zu fusionieren, zu verlegen, zu verbreiten, unterlizenzieren und/oder zu verkaufen, sofern ein Urheberrechtsvermerk und der Erlaubnisvermerk den Kopien beigelegt sind.

N3

Notation 3 ist eine formale Sprache, die beispielsweis als Syntax für RDF-Daten genutzt werden kann:

```
<#Tim Berners-Lee> <#entwickelte> <#N3> .
```

n-Gramm

Zerlegung eines Textes in einzelne Fragmente der Anzahl n. Die Fragmente können Buchstaben, Phoneme oder auch Wörter sein. In der Computerlinguistik finden sich oft Bi- oder Trigramme aus Zeichen (Buchstaben und/oder Satzzeichen) (■ **Abbildung 61**).

NER

Named Entity Recognition – Eigennamenerkennung. Begriffe eines Textes werden bestimmten Klassen zugeordnet, z. B. Orte oder Personen.

Normalisierung

Allgemein wird darunter die Vereinheitlichung von Text verstanden.

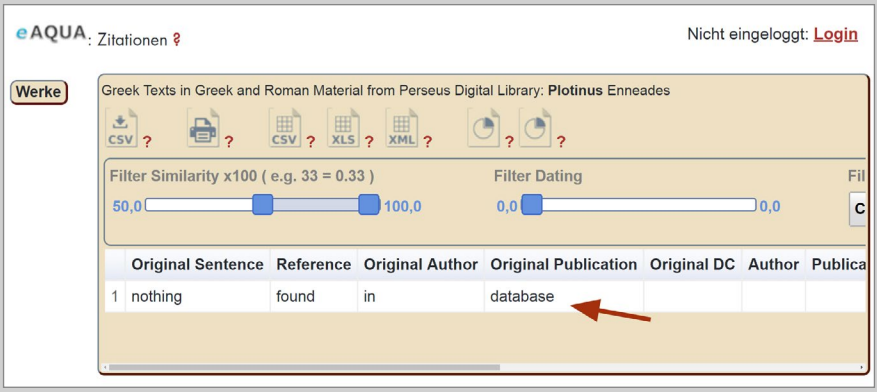


Abbildung 62. Zitationstabelle ohne Treffer

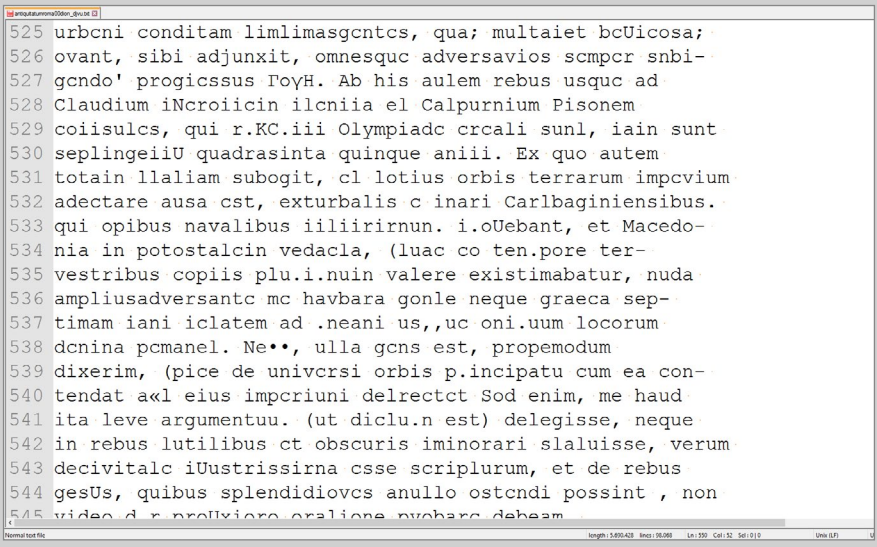


Abbildung 63. Auszug eines per ABBYY FineReader 8.0 erzeugten lateinischen Textes[47]

47 URL: https://archive.org/details/antiquitatumroma00dion.

Nothing found in database

Bei der Zitationsabfrage sind keine Ergebnisse in der Datenbank verzeichnet (■ **Abbildung 62**).

OCR

OCR ist die englische Abkürzung für optical character recognition und bezeichnet die automatisierte Texterkennung innerhalb von Bildern, die per Scanner, Digitalfotografie oder Videokamera erzeugt wurden. Die Texterkennung versucht aus den in Zeilen und Spalten angeordneten Punkten unterschiedlicher Färbung (Pixel) Buchstaben zu identifizieren und ihnen einen Zahlenwert zuzuordnen, der ihnen nach üblicher Textcodierung zukommt (ASCII, Unicode) (■ **Abbildung 63**).

Parser

Ein Parser ist ein Programm, welches eine Eingabe zerlegt und in ein für die Weiterverarbeitung brauchbares Format umwandelt.

Persistent Identifier

Ein künstlich zugewiesenes Merkmal zur eindeutigen, dauerhaften Identifizierung eines Subjektes / Objektes wird als persistent Identifier (persistent ID oder PID) bezeichnet.

PHI

Das Packard Humanities Institute[48] ist eine 1987 gegründete Stiftung zur Unterstützung von Langzeitprojekten auf den Gebieten der Archäologie, Musik, Filmkonservierung, Aufbewahrung historischer Dokumente und der Früherziehung. Die Stiftung veröffentlicht unter anderem antike Textsammlungen, wie beispielsweise alle lateinischen literarischen Texte, die vor 200 n. Chr. geschrieben wurden (PHI 5.3) oder griechische Inschriften und Papyri (PHI 7).

48 URL: https://packhum.org/.

Abbildung 64. eAQUA-Logo als PNG mit transparentem Hintergrund

```
172
173 _:node1cq0hov24x2219693 gndo:personalName "Plato" ;
174     gndo:nameAddition "Atheniensis" .
175
176 <http://d-nb.info/gnd/118594893> gndo:variantNameForThePerson "Plato, Athenensis" ;
177     gndo:variantNameEntityForThePerson _:node1cq0hov24x2219694 .
178
179  :node1cq0hov24x2219694 gndo:personalName "Plato" ;
180     gndo:nameAddition "Athenensis" .
181
182 <http://d-nb.info/gnd/118594893> gndo:variantNameForThePerson "Plato, Philosophus" ;
183     gndo:variantNameEntityForThePerson _:node1cq0hov24x2219695 .
184
185 _:node1cq0hov24x2219695 gndo:personalName "Plato" ;
186     gndo:nameAddition "Philosophus" .
187
188 <http://d-nb.info/gnd/118594893> gndo:variantNameForThePerson "Platon, Philosoph" ;
189     gndo:variantNameEntityForThePerson  :node1cq0hov24x2219696 .
190
191 _:node1cq0hov24x2219696 gndo:personalName "Platon" ;
192     gndo:nameAddition "Philosoph" .
193
194 <http://d-nb.info/gnd/118594893> gndo:variantNameForThePerson "Platon, Sohn des Ariston" ;
195     gndo:variantNameEntityForThePerson _:node1cq0hov24x2219697 .
196
197 _:node1cq0hov24x2219697 gndo:personalName "Platon" ;
198     gndo:nameAddition "Sohn des Ariston" .
199
```

Abbildung 65. Auszug der RDF-Repräsentation des GND-Datensatzes zu Platon bei der DNB[49]

49 URL: http://d-nb.info/gnd/118594893.

PNG
Portable Network Graphics ist ein Grafikformat, welches verlustfrei komprimieren kann. Es wurde als freier Ersatz für Graphics Interchange Format (GIF) entwickelt und unterstützt die Transparenz per Alphakanal (■ **Abbildung 64**).

PoS
Part-of-Speech Tagging ordnet die Wörter eines Textes Wortarten zu.

PURL
Ein Persistent Uniform Resource Locator verweist in Form einer URL nicht direkt auf eine Ressource, sondern auf einen Resolver, der die aktuelle Internet-URL liefert. DOI oder URN existieren alternativ dazu.

Resolver
Als Resolver wird in der Informatik allgemein eine Software zur Namensauflösung bezeichnet. Ein Linkresolver löst Metadaten z. B. in Form einer URN in lokale Bestandsdaten auf und liefert den dazu passenden Hyperlink.

RDA
Resource Description and Access bezeichnet einen neuen Standard für die Erschließung von Ressourcen in Bibliotheken, Archiven und Museen als Nachfolger der Anglo-American Cataloguing Rules (AACR2).

RDF
Das Resource Description Framework wurde vom World Wide Web Consortium (W3C) zur Beschreibung von Metadaten entwickelt. Es gilt mittlerweile als wesentlicher Bestandteil des sogenannten semantischen Webs. Aussagen im RDF-Modell werden als Tripel von Subjekt, Prädikat und Objekt gebildet, zumeist in Form von XML oder N3 (■ **Abbildung 65**).

%N%	αὐτὴ	Διὰ	ἔνθα	ἦν	ν
†	αὐτῇ	διὸ	ἑνὸς	ἣν	νοῦν
‘	αὕτη	διότι	ἐνταῦθα	ης	νῦν
~	αὐτὴν	δύναται	ἐντεῦθεν	ἧς	ὁ
⩽	αὐτῆς	δύο	ἐξ	ἦσαν	Ὁ
A	αὐτὸ	ε	ἔξω	ἤτοι	ὅ
ex.	αὐτοὶ	ἐάν	ἐπ’	ι	ὃ
fr.	αὐτοῖς	ἐὰν	ἐπεὶ	ἴδιον	ὅδε
p.	αὐτόν	ἑαυτὸν	ἐπειδὴ	ἵνα	ὅθεν
v.	αὐτὸν	ἑαυτοῦ	ἔπειτα	καθ’	οἱ
α	αὐτός	ἑαυτῷ	ἐπί	καθάπερ	Οἱ
A	αὐτὸς	ἑαυτῶν	ἐπὶ	καί	οἳ
ἃ	αὐτὸς	ἐγώ	ἐς	καὶ	οἶμαι
α′	αὐτοῦ	ἐγὼ	ἔσται	Καὶ	οἷον
ἀεὶ	αὐτούς	εἰ	ἐστι	καίτοι	οἷς
αἱ	αὐτοὺς	Εἰ	ἐστὶ	κἂν	ὅλως
ἀλλ’	αὐτῷ	εἴ	ἔστι	κατ’	ὁμοίως
Ἀλλ’	αὐτῶν	εἴη	ἐστίν	κατά	ὁμοῦ
ἀλλὰ	ἀφ’	εἰμί	ἐστὶν	κατὰ	ὅμως
Ἀλλὰ	B	εἶναι	ἔστιν	κάτω	ὃν
ἄλλα	β′	εἴπερ	ἕτερον	λοιπὸν	ὃν
ἀλλήλων	Γ	εἰς	ἔτη	μάλιστα	ὄντα
ἄλλο	γ′	εἷς	ἔτι	μᾶλλον	ὄντος
ἄλλοι	γάρ	εἰσι	εὖ	με	ὄντων
ἄλλοις	γὰρ	εἰσὶ	εὐθὺς	μέγα	ὅπερ
ἄλλος	γε	εἰσιν	ἐφ’	μεθ’	ὅπως
ἄλλων	γέγονεν	εἶτα	ἔχει	μέν	ὅς
ἄλλως	γενέσθαι	εἴτε	ἔχειν	μὲν	ὃς
ἅμα	γίνεται	ἐκ	ἔχον	μέντοι	ὅσα
ἄν	γίνονται	ἕκαστον	ἔχοντα	μετ’	ὅσον
ἄν	δ′	ἐκεῖ	ἔχοντες	μετά	ὅστις
ἂν	δ’	ἐκεῖνο	ἔχων	μετὰ	ὅταν
ἄνευ	δαί	ἐκεῖνον	ἕως	μεταξὺ	ὅτε
ἀντὶ	δαίς	ἐκεῖνος	ἢ	μέχρι	ὅτι
ἄνω	δέ	ἐκείνου	ᾗ	μή	Ὅτι
ἀπ’	δὲ	ἐκείνων	ἡ	μὴ	οὐ
ἅπαντα	δεῖ	ἐμοὶ	Ἡ	Μὴ	Οὐ
ἁπάντων	δεύτερον	εμον	ᾗ	μηδὲ	οὗ
ἁπλῶς	δή	ἐμός	ἤγουν	μηδὲν	οὐδ’
ἀπό	δὴ	εμου	ἤδη	μὴν	οὐδέ
ἀπὸ	δηλοῖ	ἐμοῦ	ἡμᾶς	μήτε	οὐδὲ
ἄρα	δῆλον	ἐν	ἡμεῖς	μίαν	οὐδείς
αὖ	δι’	Ἐν	ἡμῖν	μοι	οὐδεὶς
αὖθις	διά	ἓν	ἡμῶν	μόνον	οὐδὲν
αὐτὰ	διὰ	ἕνα	ην	μου	οὐκ

Abbildung 66. Beginn einer Stoppwortliste für Altgriechisch

Satzkookkurrenz

Das statistisch auffällige gemeinsame Auftreten von zwei Wortformen in einem Satz wird Satzkookkurrenz bezeichnet.

Signifikanz

In der Statistik wird unter Signifikanz eine Kennzahl verstanden, welche die Wahrscheinlichkeit eines systematischen Zusammenhangs zwischen Variablen bezeichnet.

Similar-Text

Ein Algorithmus, der die Ähnlichkeit zweier Texte auf Zeichenbasis und mit Hilfe der Editierdistanz berechnet.

SQL

Datenbanksprache in relationalen Datenbanken. SQL (Allgemeiner Sprachgebrauch: Structured Query Language) unterscheidet drei Befehlskategorien:

- Data Manipulation Language (DML) – Befehle zur Datenmanipulation
- Data Definition Language (DDL) – Befehle zur Definition des Datenbankschemas
- Data Control Language (DCL) – Befehle für die Rechteverwaltung und Transaktionskontrolle.

Stoppwort

Eine Liste von Wörtern, die bei der Verarbeitung eines Textes nicht berücksichtigt werden sollen, wird Stoppwortliste genannt. Werden die häufigsten Wörter einer Sprache zur Bildung der Liste herangezogen, wird von einer festen Stoppwortliste gesprochen. Werden die häufigsten Wörter innerhalb eines bestimmten Korpus genutzt, so ist von einer berechneten Stoppwortliste auszugehen. Im Einzelfall kann es durchaus hilfreich sein, einzelne Wörter aus der berechneten Liste wieder zu entfernen. Stoppwörter stammen zumeist aus geschlossenen Wortklassen[50]. Sie sind kaum Veränderungen ausgesetzt und ihre grammatische Bedeutung steht im Vordergrund. Sie werden auch Funktionswörter genannt (■ **Abbildung 66**).

50 Zu den geschlossenen Wortklassen zählen die Präpositionen, Partikel, Konjunktionen und Artikel, in manchen Sprachen auch die Adjektive.

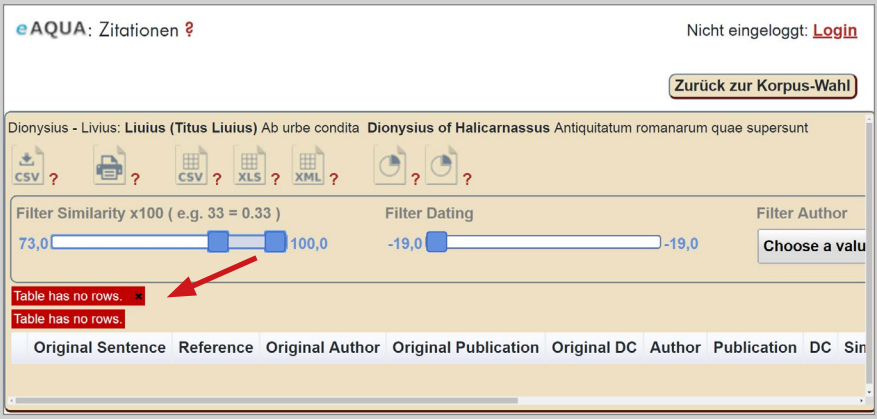


Abbildung 67. Zitationstabelle mit einer Fehlermeldung bei unzutreffenden Filterkriterien, obwohl Treffer vorhanden sind

SVG

Scalable Vector Graphics basiert auf XML und beschreibt zweidimensionale Vektorgrafiken.

Table has no rows

Bei der Datentabelle Zitation kann es vorkommen, dass die eingestellten Filterkriterien eine Anzeige von Datensätzen verhindern, obgleich Daten verfügbar sind. In diesem Fall zeigt die Visualisierung mit dem Hinweis „Table has no rows“ an, dass keine Datensätze den Filterkriterien entsprechen (■ **Abbildung 67**).

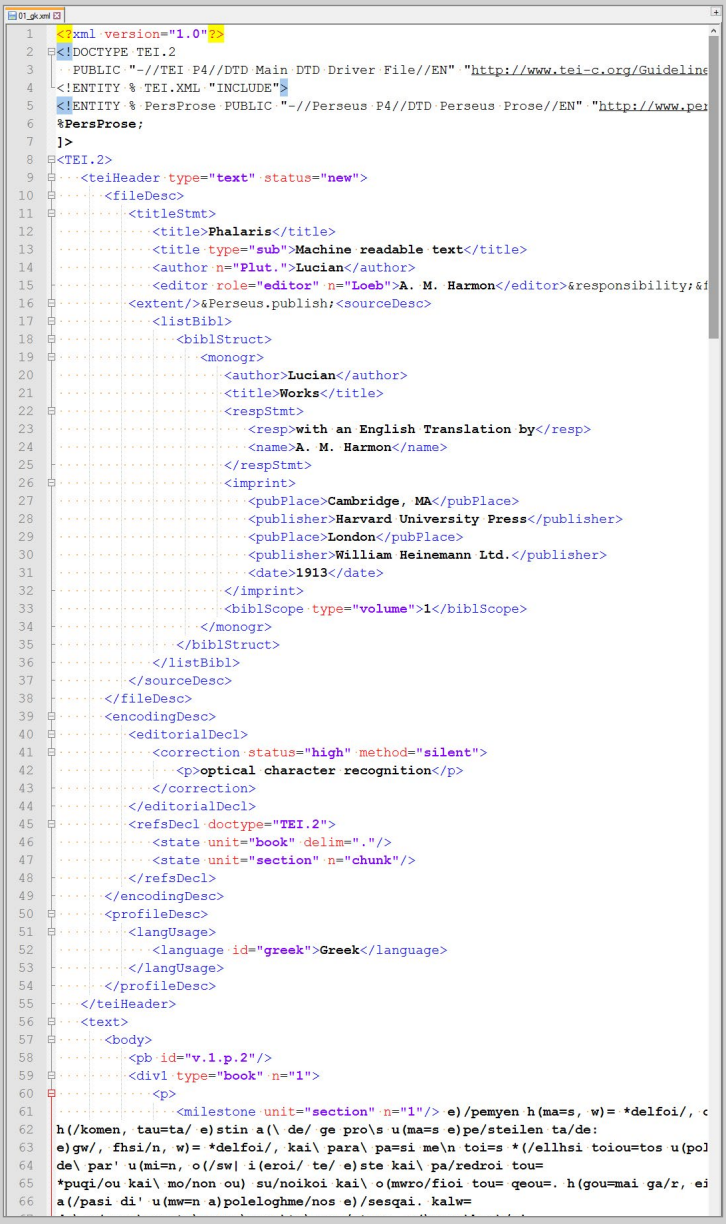
Tag

Ist dem Englischen entlehnt und zeichnet einen Datenbestand mit zusätzlichen Informationen aus. Grundsätzlich gibt es bei Auszeichnungssprachen drei unterschiedliche Markierungsarten:

- <starttag>: Markiert den Beginn einer Auszeichnung
- </endtag>: Markiert das Ende einer Auszeichnung
- <emptyelementtag/>: Ein Element, welches nur aus Attributen besteht und Anfang und Ende gleichzeitig markiert.

Text Mining

Unter Text Mining werden allgemein Verfahren bezeichnet, die mit statistischen und linguistischen Mitteln weitgehend automatisiert Informationen aus Texten erschließen und strukturieren.



```
<?xml version="1.0"?>
<!DOCTYPE TEI.2
 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN" "http://www.tei-c.org/Guidelins
<!ENTITY % TEI.XML "INCLUDE">
<!ENTITY % PersProse PUBLIC "-//Perseus P4//DTD Perseus Prose//EN" "http://www.per
%PersProse;
]>
<TEI.2>
  <teiHeader type="text" status="new">
    <fileDesc>
      <titleStmt>
        <title>Phalaris</title>
        <title type="sub">Machine readable text</title>
        <author n="Plut.">Lucian</author>
        <editor role="editor" n="Loeb">A. M. Harmon</editor>&responsibility;&f
      <extent/>&Perseus.publish;<sourceDesc>
        <listBibl>
          <biblStruct>
            <monogr>
              <author>Lucian</author>
              <title>Works</title>
              <respStmt>
                <resp>with an English Translation by</resp>
                <name>A. M. Harmon</name>
              </respStmt>
              <imprint>
                <pubPlace>Cambridge, MA</pubPlace>
                <publisher>Harvard University Press</publisher>
                <pubPlace>London</pubPlace>
                <publisher>William Heinemann Ltd.</publisher>
                <date>1913</date>
              </imprint>
              <biblScope type="volume">1</biblScope>
            </monogr>
          </biblStruct>
        </listBibl>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <editorialDecl>
        <correction status="high" method="silent">
          <p>optical character recognition</p>
        </correction>
      </editorialDecl>
      <refsDecl doctype="TEI.2">
        <state unit="book" delim="."/>
        <state unit="section" n="chunk"/>
      </refsDecl>
    </encodingDesc>
    <profileDesc>
      <langUsage>
        <language id="greek">Greek</language>
      </langUsage>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
      <pb id="v.1.p.2"/>
      <div1 type="book" n="1">
        <p>
          <milestone unit="section" n="1"/> e)/pemyen h(ma=s, w)= *delfoi/, c
h(/komen, tau=ta/ e)stin a(\ de/ ge pro\s u(ma=s e)pe/steilen ta/de:
e)gw/, fhsi/n, w)= *delfoi/, kai\ para\ pa=si me\n toi=s *(/ellhsi toiou=tos u(pol
de\ par' u(mi=n, o(/sw| i(eroi/ te/ e)ste kai\ pa/redroi tou=
*puqi/ou kai\ mo/non ou) su/noikoi kai\ o(mwro/fioi tou= qeou=. h(gou=mai ga/r, ei
a(/pasi di' u(mw=n a)poleloghme/nos e)/sesqai. kalw=
```

Abbildung 68. TEI- XML-Auszug aus einem Dokument der Perseus Digital Library[51] mit altgriechischem Beta Code

51 http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:2008.01.0420.

TEI

Das von der Text Encoding Initiative[52] entwickelte, gleichnamige Dokumentenformat basiert in der aktuellen Version P5 auf XML und hat sich zum De-facto-Standard zur Kodierung gedruckter Werke in den Geisteswissenschaften entwickelt (■ **Abbildung 68**).

TIFF

Tagged Image File Format ist ein Bilddateiformat, welches insbesondere für hochaufgelöste Bilder in druckfähiger, verlustfreier Qualität benutzt wird.

TLG

Der Thesaurus Linguae Graecae ist eine heute kommerziell arbeitende Institution der University of California, Irvine. Seit der Gründung 1972 hat das Projekt die meisten griechischen Texte von Homer (8. Jh. v. Chr.) bis zum Fall von Byzanz im Jahre 1453 gesammelt und digitalisiert. TLG-Texte wurden der wissenschaftlichen Gemeinschaft zunächst auf Magnetbändern (Mitte der 70er Jahre) und später im CD-ROM-Format zur Verfügung gestellt. Die CD-ROMs A (1985), C (1988) und D (1992) wurden mit technischer Unterstützung des Packard Humanities Institute (PHI) produziert. TLG-E (2000) wurde vom TLG-Team nach der Migration des Corpus vom Ibycus-System in die Unix-Umgebung komplett selbst produziert. Seit 2001 wurde das Projekt als Webanwendung konzipiert und sowohl technisch wie inhaltlich neu aufgesetzt. Diese neue Datenbank ist seither für Abonenten online erhältlich.

Tokenisierung

In der Computerlinguistik wird damit die Zerlegung in Segmente auf Wortebene bezeichnet.

Trigramm

Ein spezieller Typ von n-Grammen, der aus drei aufeinander folgenden Buchstaben oder Wortformen besteht, wird Trigramm bezeichnet.

52 URL: http://www.tei-c.org.

Berners-Lee [Page 2]

RFC 1630 URIs in WWW June 1994

URIs

This document defines a way to encapsulate a name in any registered name space, and label it with the the name space, producing a member of the universal set. Such an encoded and labelled member of this set is known as a Universal Resource Identifier, or URI.

The universal syntax allows access of objects available using existing protocols, and may be extended with technology.

The specification of the URI syntax does not imply anything about the properties of names and addresses in the various name spaces which are mapped onto the set of URI strings. The properties follow from the specifications of the protocols and the associated usage conventions for each scheme.

URLs

For existing Internet access protocols, it is necessary in most cases to define the encoding of the access algorithm into something concise enough to be termed address. URIs which refer to objects accessed with existing protocols are known as "Uniform Resource Locators" (URLs) and are listed here as used in WWW, but to be formally defined in a separate document.

URNs

There is currently a drive to define a space of more persistent names than any URLs. These "Uniform Resource Names" are the subject of an IETF working group's discussions. (See Sollins and Masinter, Functional Specifications for URNs, circulated informally.)

The URI syntax and URL forms have been in widespread use by World-Wide Web software since 1990.

Abbildung 69. RFC 1630, S. 2.

TSV

Das textbasierte Dateiformat TSV (Tab-separated values) ist eine Form von DSV (Delimiter-separated values). Die Daten sind in Tabellenform, also zweidimensional, gespeichert. Jede Zeile ist ein Datensatz. Felder werden mittels Tab-Stop separiert.

Unigramm

Ein Unigramm (oder auch Monogramm) ist ein spezieller Typ von n-Grammen, welches aus einem Buchstaben oder einer Wortform besteht.

URI

Laut RFC 1630 von T. Berners-Lee aus dem Jahr 1994[53] ist URI ein Akronym für Universal Resource Identifiers (■ **Abbildung 69**), inzwischen wird es als Uniform Resource Identifier verstanden. Ein URI dient zur Identifizierung einer abstrakten oder physischen Ressource und kann aus fünf Teilen bestehen, von denen aber nur scheme und path zwingend vorhanden sein müssen:
scheme://[authority]/path?[query]#[fragment].

URL

Uniform Resource Locator identifiziert eine Ressource anhand der zu verwendenden Zugriffsmethode. Der eAQUA-Internetauftritt wird z. B. über http://www.eaqua.net erreichbar gemacht, eine E-Mail-Adresse mit dem Schema mailto:max.mustermann@example.org erkannt.

URN

Publikationen können im Netz dauerhaft und zuverlässig zitiert werden, indem eindeutige, standortunabhängige Identifikatoren URNs (Uniform Resource Name) anstelle von URLs verwendet werden. URNs sind URIs mit dem Schema urn:namensraum:namensraum-spezifischerTeil, also z. B. urn:nbn:de:101-2012121200 für das Werk „Policy für die Vergabe von URNs im Namensraum urn:nbn:de (Version 1.0, Stand: 29. November 2012)" der Deutschen Nationalbibliothek.

53 URL: https://tools.ietf.org/html/rfc1630.

rang	word	count	frequenz	sum frequenz	zipf(rang*count)
1	καὶ	4125322	5.5615	5.5615	4125322
2	δὲ	1505608	2.0298	7.5913	3011216
3	τὸ	1417237	1.9106	9.502	4251711
4	τοῦ	1148784	1.5487	11.0507	4595136
5	τῶν	1055097	1.4224	12.4731	5275485
6	τὴν	993288	1.3391	13.8122	5959728
7	τῆς	851238	1.1476	14.9598	5958666
8	ὁ	828861	1.1174	16.0772	6630888
9	ἐν	796323	1.0736	17.1508	7166907
10	γὰρ	693988	0.9356	18.0864	6939880
11	τὸν	680758	0.9178	19.0041	7488338
12	τὰ	627478	0.8459	19.8501	7529736
13	μὲν	591571	0.7975	20.6476	7690423
14	ἡ	529144	0.7134	21.361	7408016
15	τῷ	517482	0.6976	22.0586	7762230
16	ὡς	455688	0.6143	22.6729	7291008
17	εἰς	433158	0.584	23.2569	7363686
18	πρὸς	392607	0.5293	23.7862	7066926
19	τοῖς	379993	0.5123	24.2985	7219867
20	ἢ	369947	0.4987	24.7972	7398940
21	τε	361643	0.4875	25.2848	7594503
22	ἐπὶ	346314	0.4669	25.7516	7618908
23	ὅτι	345938	0.4664	26.218	7956574
24	διὰ	335376	0.4521	26.6702	8049024
25	κατὰ	329999	0.4449	27.115	8249975
26	τοὺς	326014	0.4395	27.5546	8476364
27	μὴ	323599	0.4363	27.9908	8737173
28	οἱ	322708	0.4351	28.4259	9035824
29	οὐ	314577	0.4241	28.85	9122733
30	τῇ	308985	0.4166	29.2665	9269550

Abbildung 70. Liste der häufigsten Wörter im TLG-E mit berechneten Werten nach George Kingsley Zipf

UTF

Unicode Transformation Format. Zeichen werden zum Zwecke der elektronischen Verarbeitung auf eine Folge von Bytes abgebildet. Übliche Kodierungsverfahren sind

- UTF-8 – Zwischen 1 und 4 Byte. Die Codepoints 0 bis 127, die dem ASCII-Zeichensatz entsprechen, werden mit Hilfe von sieben Bits kodiert. Das achte leitet ein längeres Unicode-Zeichen ein, welches die nachfolgenden 1–3 Bytes belegt. UTF-8 speichert lateinische Zeichen am effizientesten.
- UTF-16 – Ein oder zwei 16-Bit-Einheiten (2 oder 4 Bytes) werden zur Kodierung eines Zeichens verwendet.
- UTF-32 – Kodiert immer 32 Bit (4 Byte). Durch die feste Länge am einfachsten zu handhaben, benötigt dafür mehr Speicher.

Wahrscheinlichkeitsverteilung

Die Wahrscheinlichkeitsverteilung ist das theoretische Pendant zur empirisch ermittelbaren Häufigkeitsverteilung. Sie beschreibt, mit welchen Wahrscheinlichkeiten eine Zufallsvariable ihre möglichen Werte annimmt.

Wortstammreduktion

Auch Stemming, Stammformreduktion oder Normalformenreduktion genannt. Verschiedene morphologische Varianten eines Wortes werden auf ihren gemeinsamen Wortstamm zurückgeführt.

W3C

Das World Wide Web Consortium standardisiert die Techniken im World Wide Web. Es wurde 1994 am MIT gegründet.

XLS

Binäres Dateiformat von Microsoft Excel, welches bis 2007 ausschließlich gebräuchlich war.

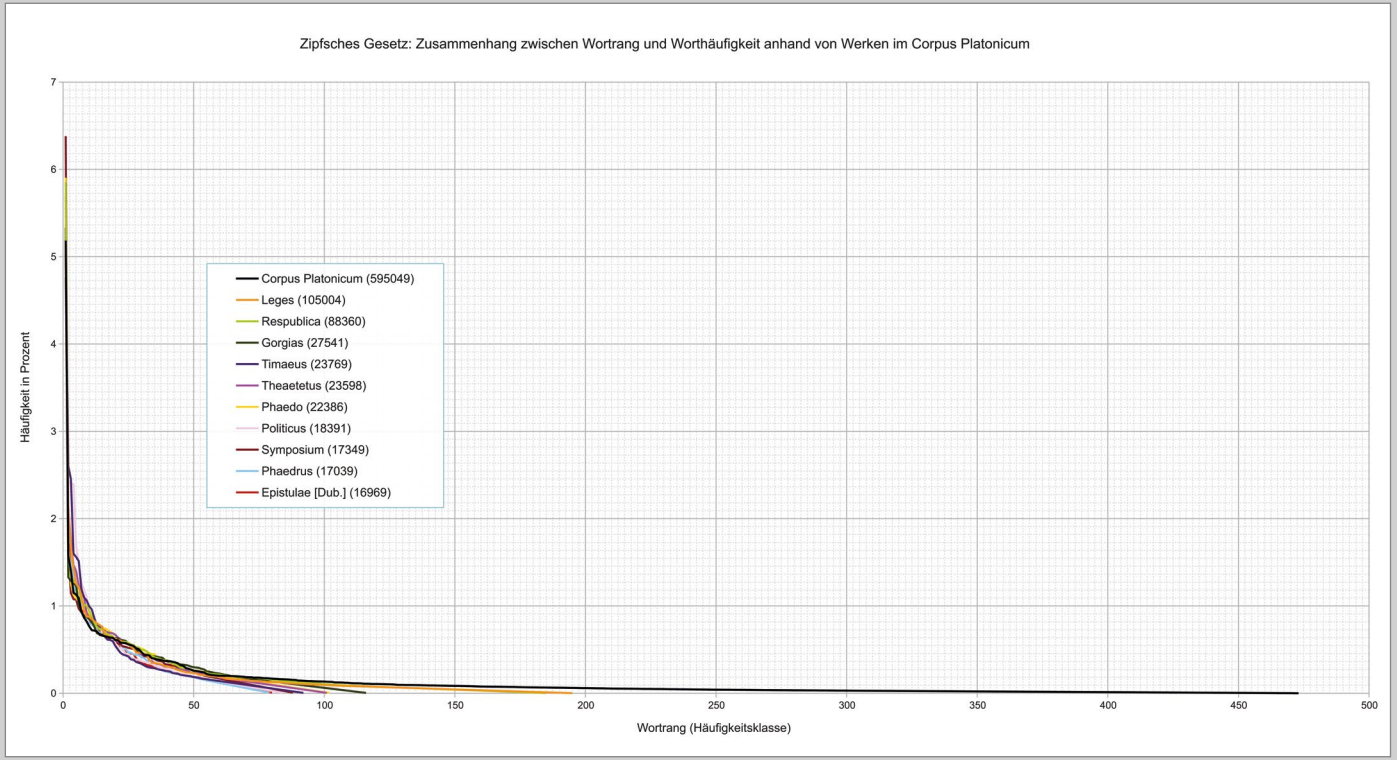


Abbildung 71. Zipfsches Gesetz im Corpus Platonicum

XML

Extensible Markup Language ist eine Auszeichnungssprache zur Darstellung strukturierter Daten in Textform. Sie wird vor allem als Austauschformat zwischen verschiedenen Computersystemen genutzt.

Zipfsches Gesetz

Das Gesetz besagt, wenn man die Elemente einer Menge, zum Beispiel die Wörter eines Textes, ihrer Häufigkeit f nach ordnet und ihnen dabei jeweils einen Rang r zuweist, dann ergibt das Produkt von f und r jeweils einen konstanten Wert k. Es hat seinen Ursprung in der Linguistik und impliziert, dass bestimmte Wörter häufiger auftreten als andere und die Verteilung einer Hyperbel 1/n ähnelt (■ **Abbildung 70,** *siehe Seite 126,* ■ **Abbildung 71**).

Abbildungsverzeichnis

Formelverzeichnis

Tabellenverzeichnis

„Digital Classics in der Praxis: Arbeiten mit eAQUA. Eine Einführung mit Beispielen“ führt in die praktische Arbeit mit den aus dem Textmining stammenden Suchmöglichkeiten der Kookkurrenzsuche und Parallelstellensuche ein. Das Buch ist als niedrigschwellige Einführung konzipiert, die keine Vorkenntnisse erfordert. Ein Einsatz als Lehrbuch zur Textanalyse in althistorisch-philologischen Übungen ist daher problemlos möglich. In einem allgemeinen Einführungsteil werden die Tools anhand ihrer im Portal eAQUA angebotenen Funktionalitäten erklärt. Die Kookkurrenzsuche zeigt semantische Zusammenhänge an, die Suche nach Parallelstellen listet Übereinstimmungen zwischen einem Werk und einem gesamten Referenzkorpus der lateinischen oder griechischen Literatur auf. Abhängigkeiten, Einflüsse und Transferwege des antiken Wissens lassen sich mit den Tools rekonstruieren und bringen Schülern und Studierenden die antiken Texte auf neue Weise nahe.
Weiterhin werden Begriffe aus der Korpusanalyse sowie die Bedeutung und der Einsatz von Signifikanzmaßen erklärt. Ein ausführliches Glossar erläutert die heute häufigsten Begriffe aus den Digital Humanities.

ISBN 978-3-947450-31-2