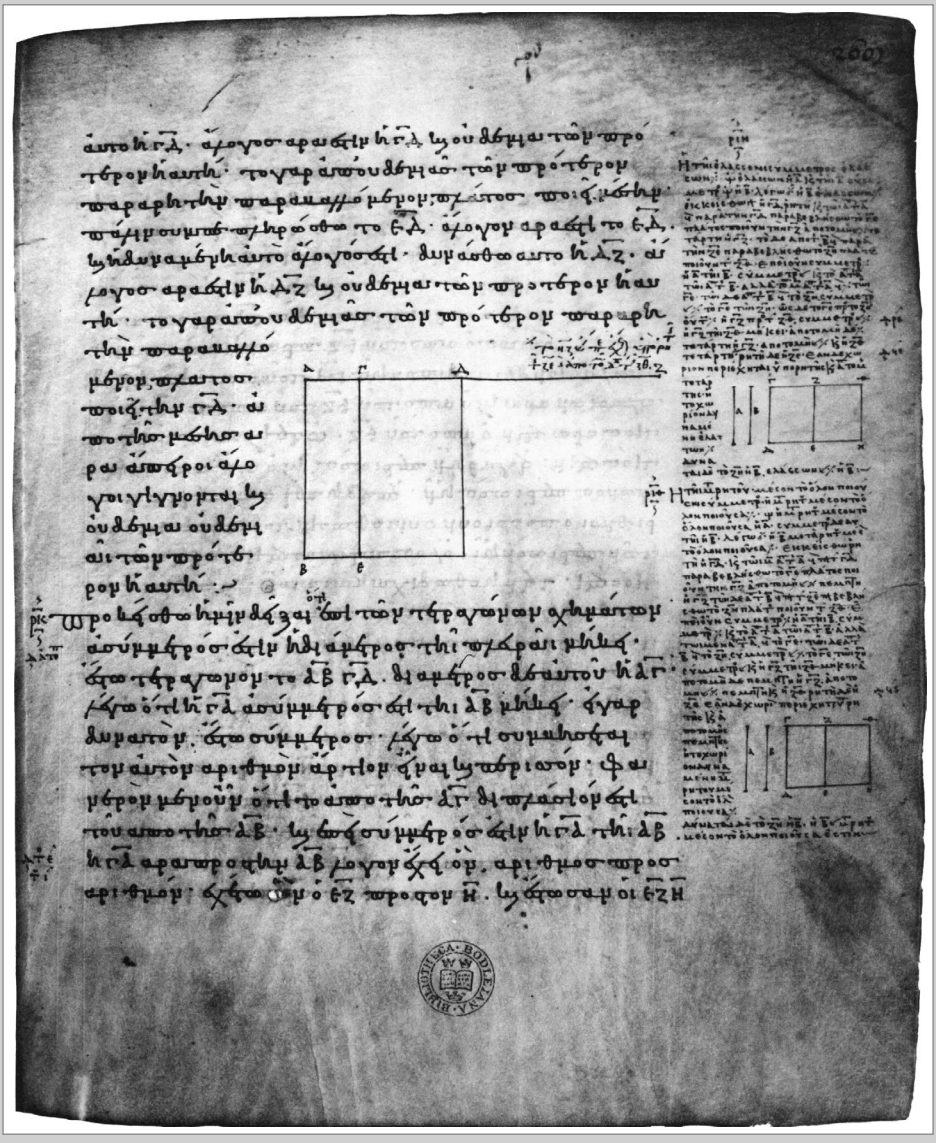


Abbildung 52. Handschrift der Elemente Euklids[38]

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.

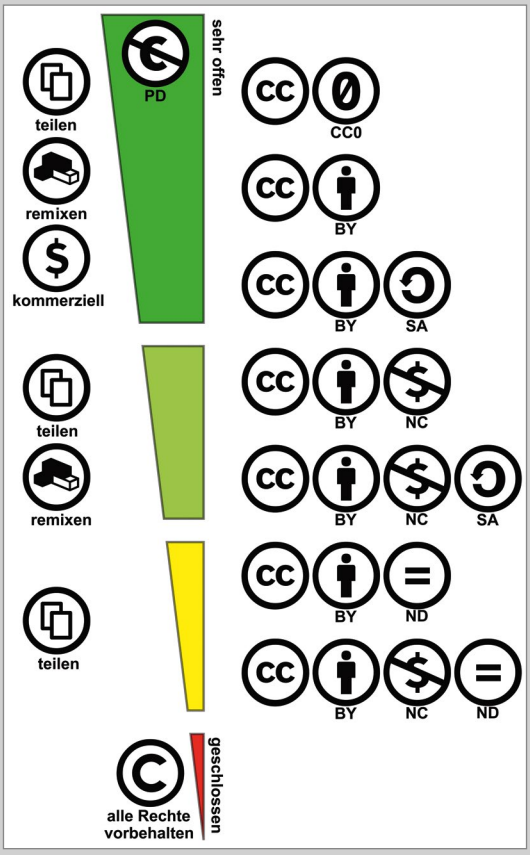


Abbildung 54. CC-Lizenzmodule können miteinander kombiniert werden[40]

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

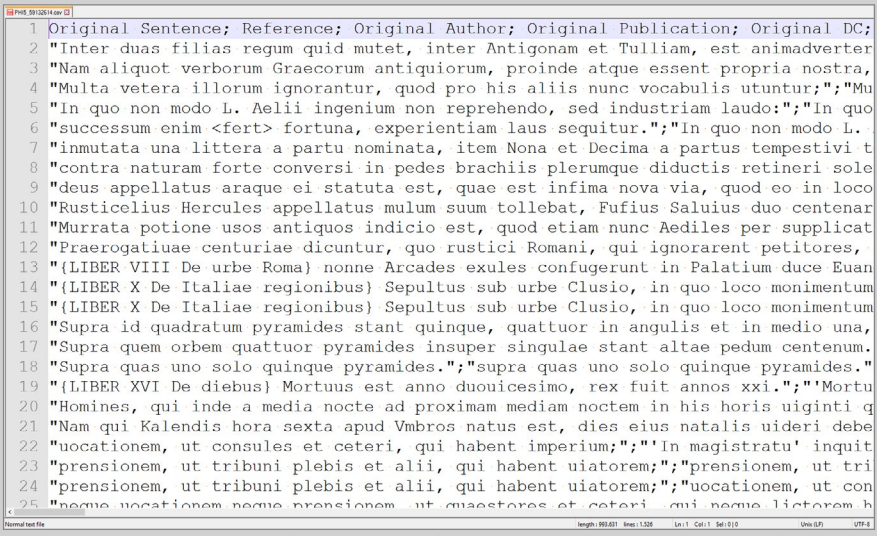
Browser
Ein Computerprogramm, im neuen Sprachgebrauch oftmals auch App (engl. Abkürzung von Application software, deutsch: Anwendungssoftware) bezeichnet, welches speziell zur Darstellung von Webseiten konstruiert wurde, ist, dem englischen Verb browsen (stöbern, umsehen, schmökern) entlehnt, ab 1989 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.

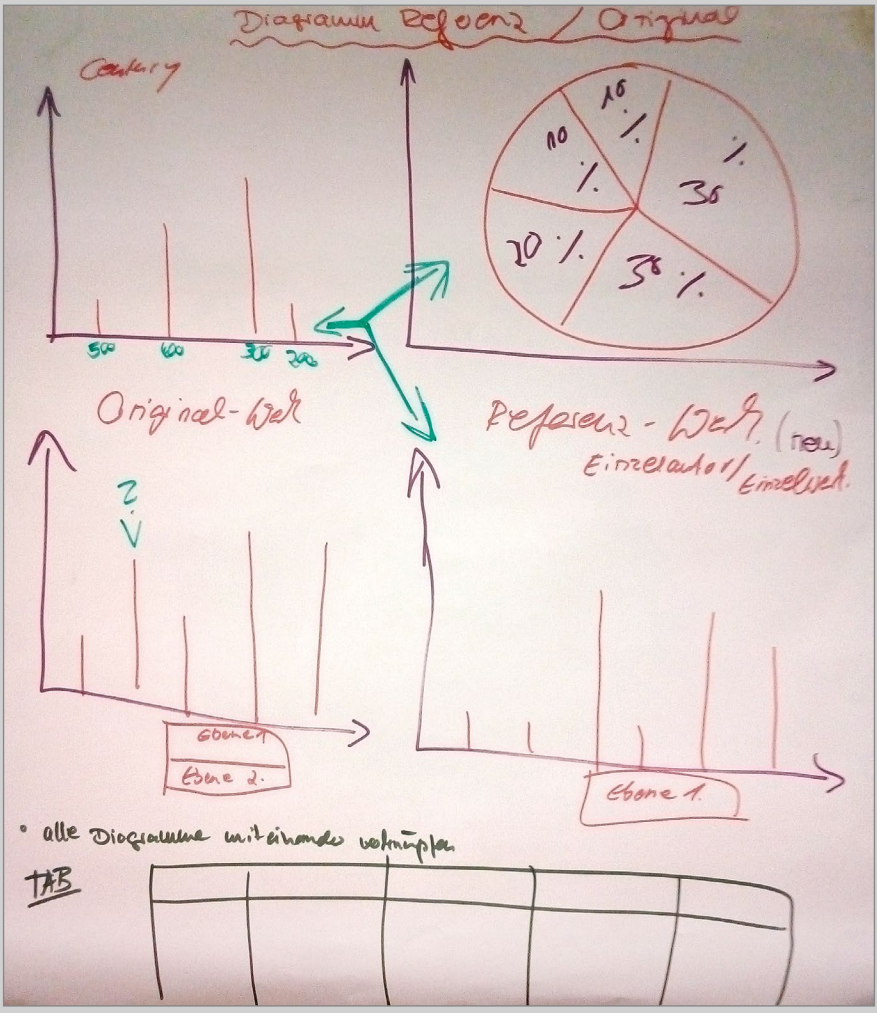


Abbildung 56. eAQUA – Entwurf für eine neue Bedienoberfläche bei der Parallelstellensuche

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**).

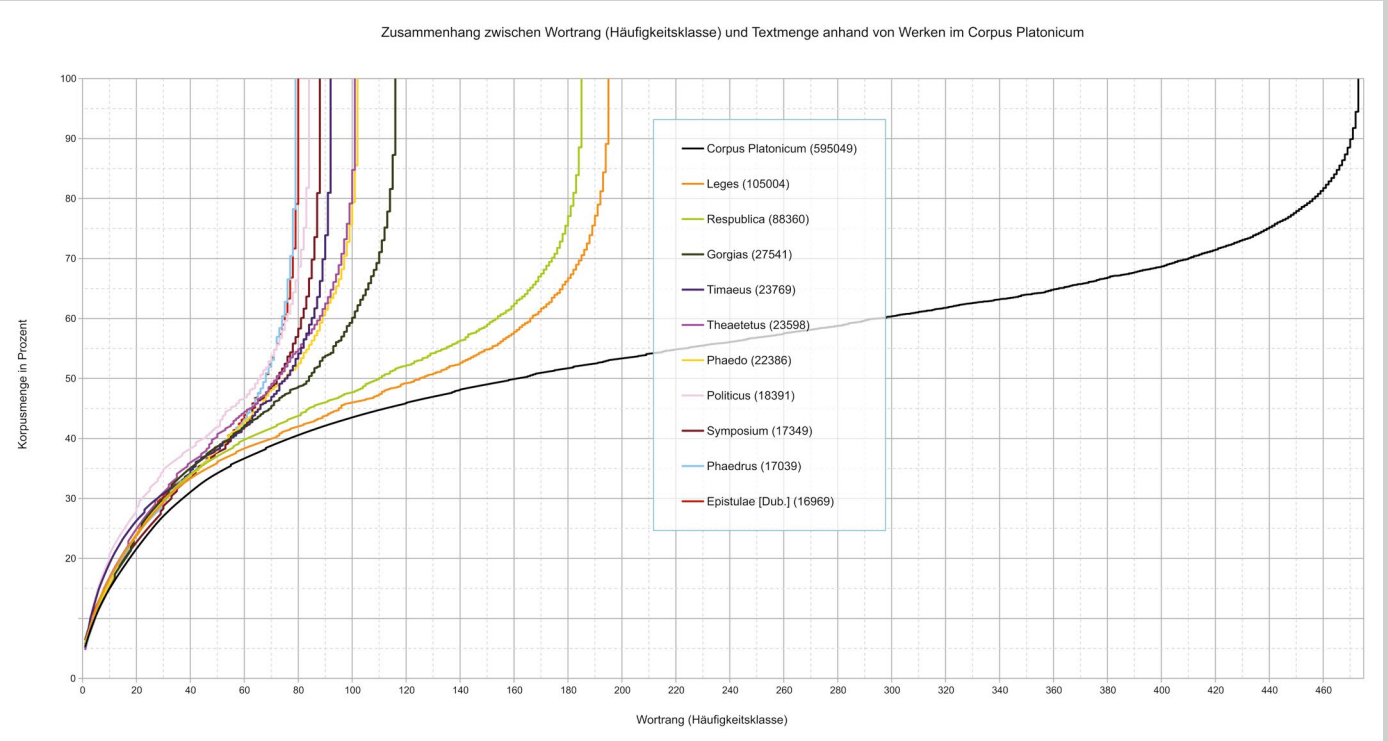


Abbildung 57. Häufigkeitsverteilung von Wörtern in ausgewählten Texten Platons

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).

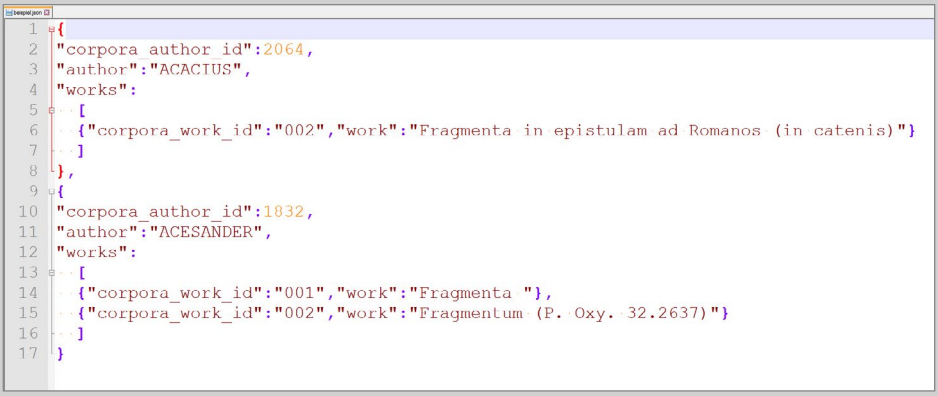
Häufigkeitsklasse

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.

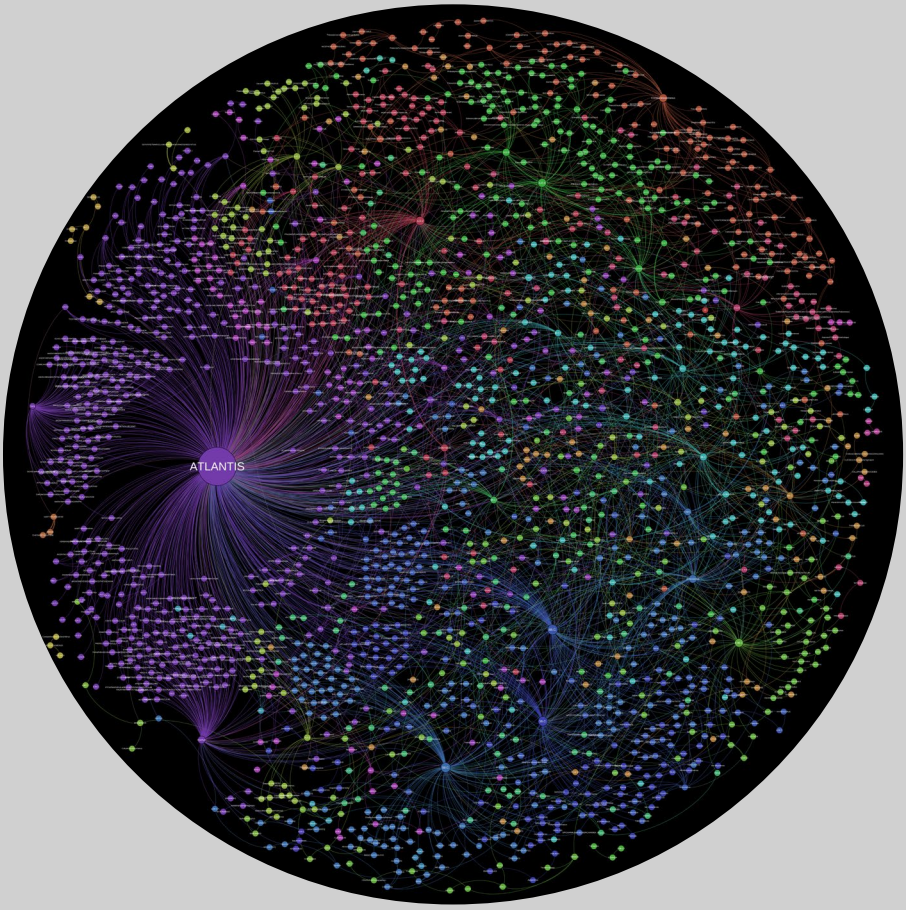


Abbildung 60. Mit Gephi erstellte Visualisierung auf der Basis des Metadatensatzes (Autorennamen, Orte, Epochen) des TLG-E.[45]

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.

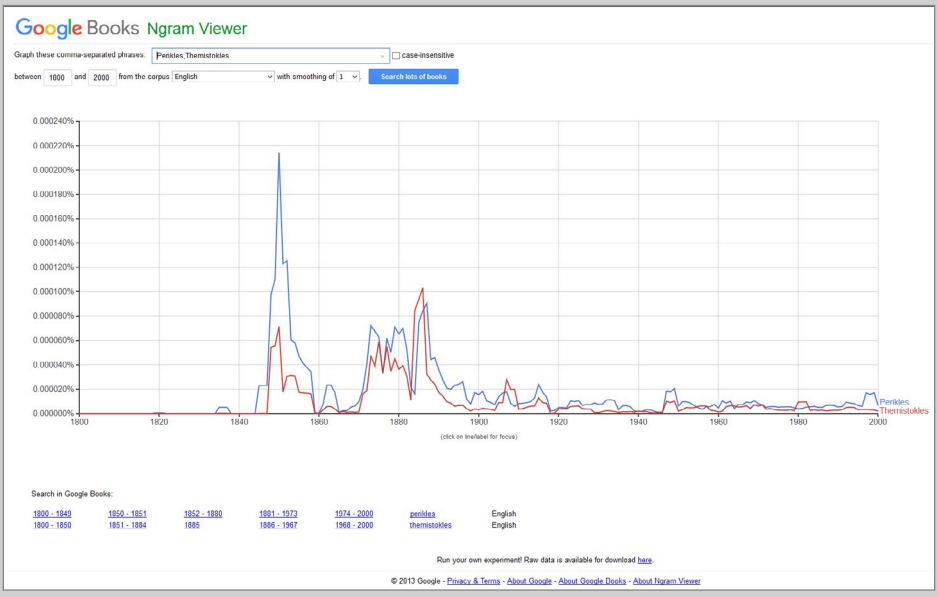


Abbildung 61. Google Books Ngram Viewer[46]

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.

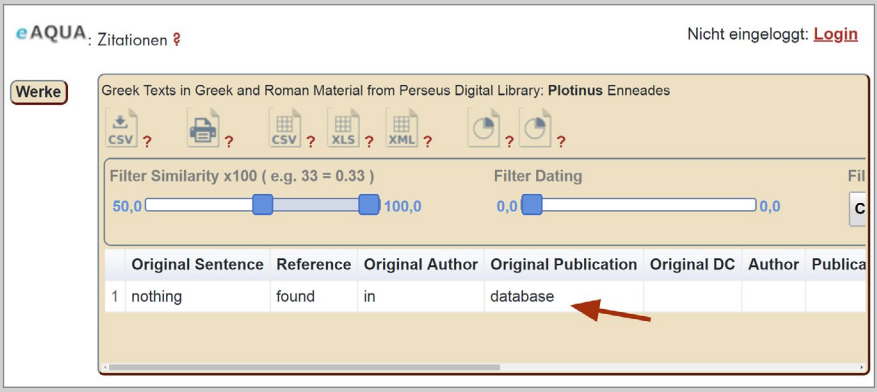


Abbildung 62. Zitationstabelle ohne Treffer

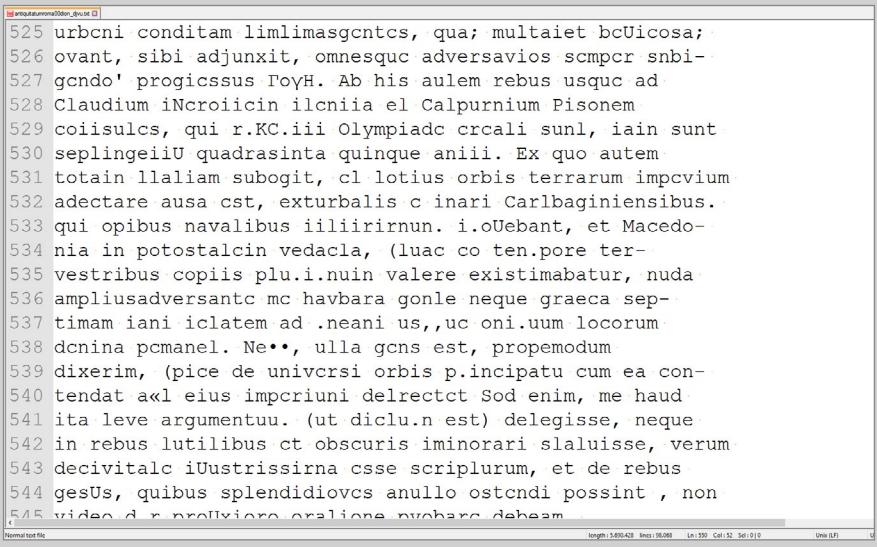


Abbildung 63. Auszug eines per ABBYY FineReader 8.0 erzeugten lateinischen Textes[47]

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.

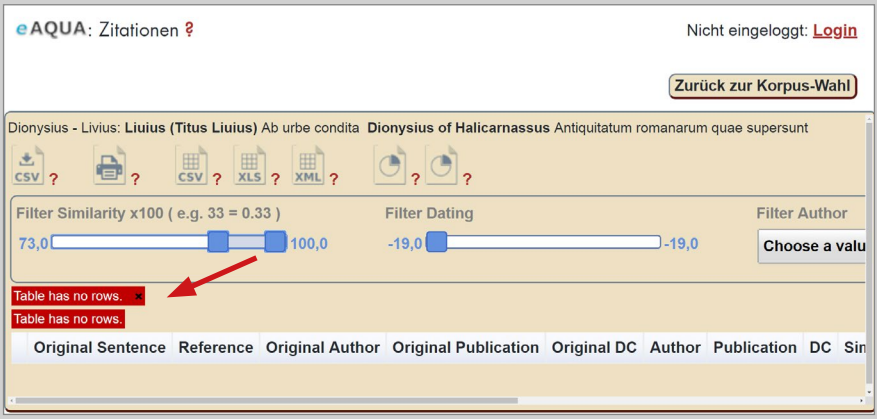


Abbildung 67. Zitationstabelle mit einer Fehlermeldung bei unzutreffenden Filterkriterien, obwohl Treffer vorhanden sind

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**).

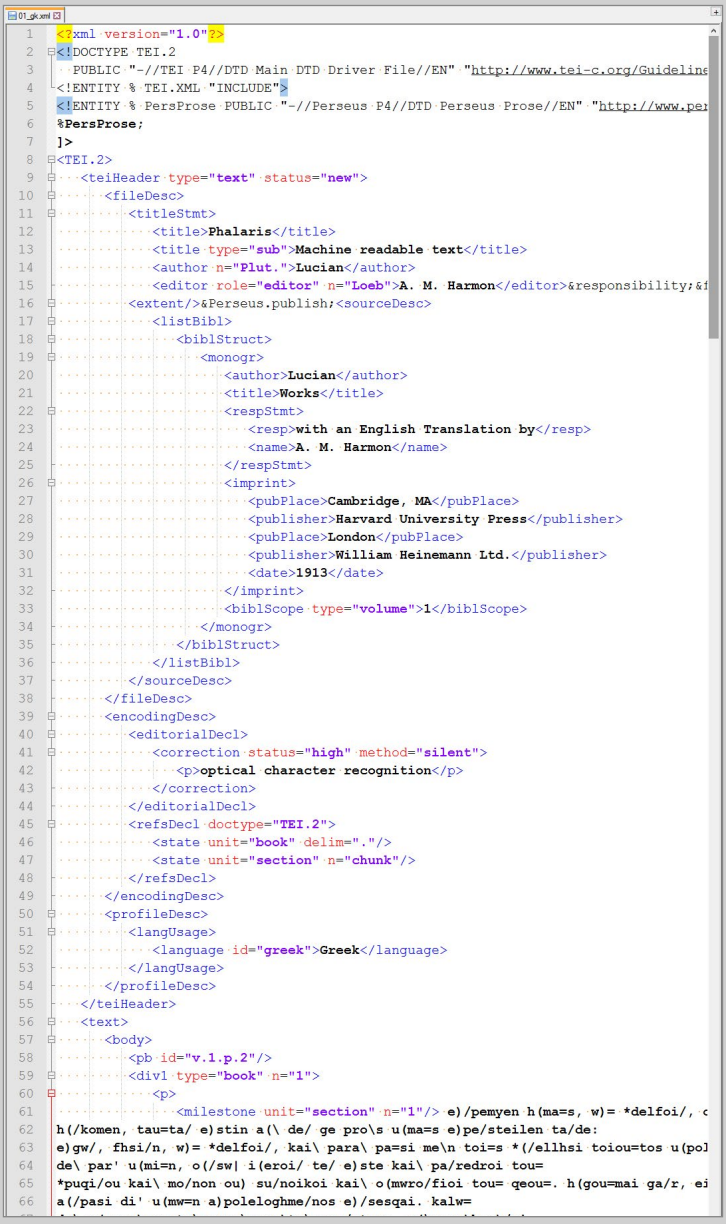
Tag

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.

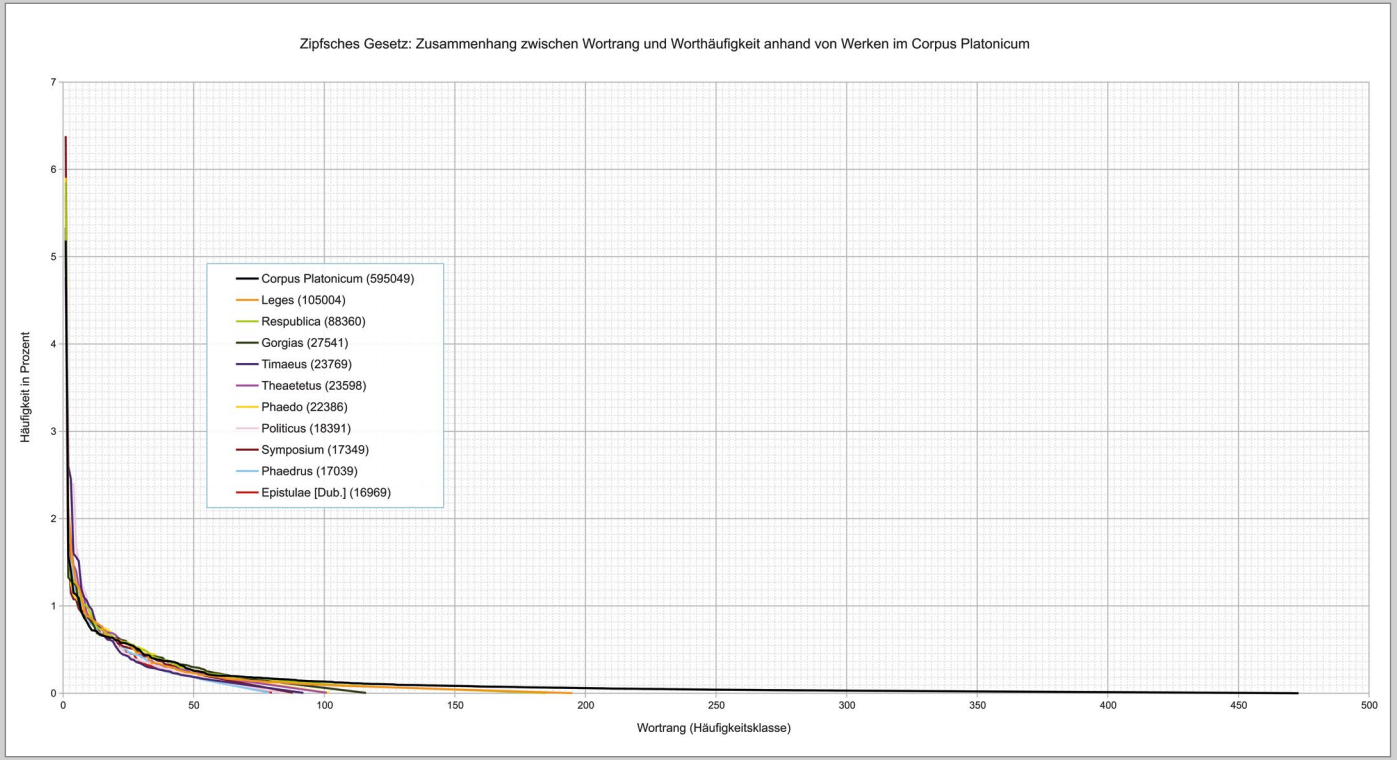


Abbildung 71. Zipfsches Gesetz im Corpus Platonicum

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**).