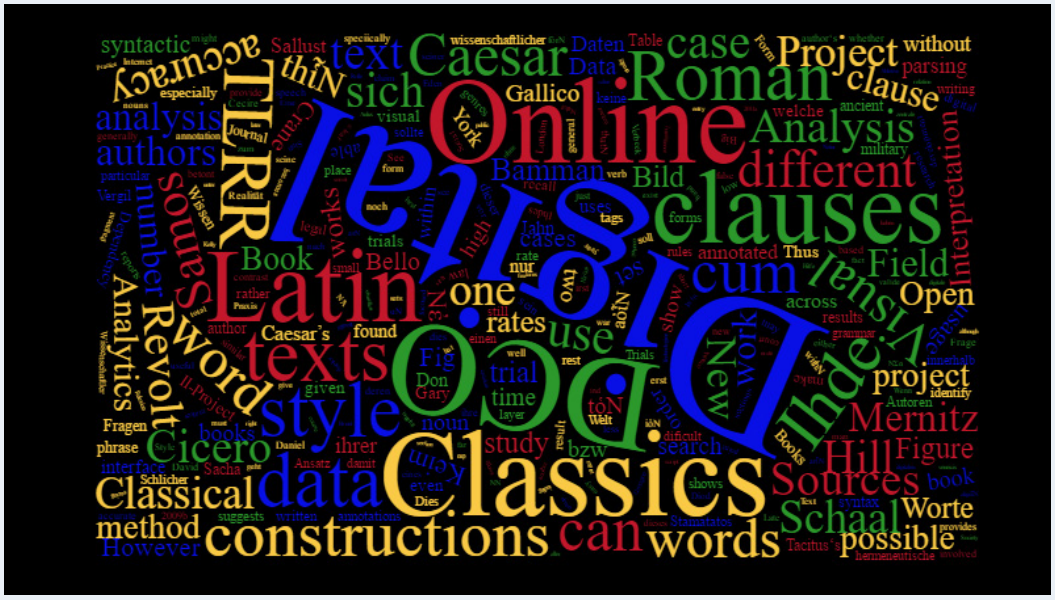
DIGITAL CLASSICS ONLINE

Bd. 2,3 (2016)

Aus dem Inhalt



ISSN: 2364-7957

DIGITAL CLASSICS ONLINE

Editorial

Charlotte Schubert

In eigener Sache: Open Access

Derzeit brandet die Diskussion um Open Access in Deutschland in nie gekannter Schärfe hoch: Da DCO sich dezidiert als wissenschaftliches Open Access Journal versteht, soll das Editorial diesmal einer kurzen Positionsbestimmung mit einem Kommentar zur derzeitigen Diskussion in Deutschland gewidmet sein, die sich speziell auf die Interessen der Wissenschaftlerinnen und Wissenschaftler bezieht.

Die Open Access–Strategie des BMBF (https://www.bmbf.de/de/open-access-das-urheberrecht-muss-der-wissenschaft-dienen-846.html), untermauert durch die Studie von Justus Haucap, Professor für Volkswirtschaftslehre an der Universität Düsseldorf, et al. (Studie „Ökonomische Auswirkungen einer Bildungs- und Wissenschaftsschranke im Urheberrecht“),[1] betont die Vorteile für Wissenschaft und Bibliotheken. Diese liegen in dem komfortableren Umgang mit wissenschaftlicher Literatur, in der Erleichterung des wissenschaftlichen Arbeitens und der verstärkten Anregung zur Produktion wissenschaftlicher Werke. Insbesondere die Bibliotheken werden in diesem Prozeß weitere – nicht nur finanzielle – Handlungs- und Gestaltungsfreiheit gewinnen. Doch in überregionalen Printmedien wird gegen diesen Transformationsprozeß getrommelt und auch innerhalb der Universitäten regt sich mancherorts Widerstand gegen universitäre Open Access-Richtlinien.

Worum geht es?

„Open Access ist eine Idee aus der Wissenschaft für die Wissenschaft.“[2] Das ist nun ausnahmsweise keine euphemistische Camouflage, denn die Deutsche Forschungsgemeinschaft, die Helmholtz-Gemeinschaft, die Max-Planck-Gesellschaft, die Leibniz-Gemeinschaft und die Fraunhofer-Gesellschaft, der Wissenschaftsrat, die Hochschulrektorenkonferenz und der Deutsche Bibliotheksverband haben sich bereits 2003 in der Berliner Erklärung gemeinsam zu diesem Weg bekannt: Hiernach zeigen der sog. Goldene Weg und der sog. Grüne Weg zwei Wege auf, die eigentlich für alle Interessen das geeignete Open Access Modell bieten:[3] „Grüner Weg“ bedeutet, daß eine bereits in einem Verlag publizierte Veröffentlichung zusätzlich im Internet eingestellt wird, etwa auf einer Webseite, einem Repositorium oder auf einem Dokumentenserver der Hochschulen oder Forschungseinrichtungen. „Goldener Weg“ bedeutet, daß die Veröffentlichung sofort im Internet eingestellt wird, d.h. in einem digitalen Medium wie etwa in der hier online erscheinenden Open Access Zeitschrift Digital Classics Online.

Es geht allerdings bei Open Access auch um handfeste wirtschaftliche Konflikte, die Rolle der Verlage, die Rolle der Printmedien und verbunden damit auch um Einflußmöglichkeiten, die ganz grundsätzlich mit Fragen von Kontrolle und Transparenz verbunden werden. Vor

1 http://www.dice.hhu.de/fileadmin/redaktion/Fakultaeten/Wirtschaftswissenschaftliche_Fakultaet/DICE/Ordnungspolitische_Perspektiven/86_OP_Haucap_Loebert_Spindler_Thorwarth.pdf (abgerufen am 25.11.2016).

2 https://www.bmbf.de/de/open-access-das-urheberrecht-muss-der-wissenschaft-dienen-846.html (abgerufen am 25.11.2016).

3 https://open-access.net/informationen-zu-open-access/was-bedeutet-open-access/ (abgerufen am 25.11.2016).

allem steht der Vorwurf im Raum, durch Open Access komme es zu einer Art von „digitale [r] Wissenschaftskontrolle“ (Uwe Jochum, FAZ, 23.11.2016, Seite N4, nur kostenpflichtig zugänglich, jedoch über Social Media wie Twitter breit angekündigt: https://twitter.com/hobohm?lang=de).

So wird Open Access von den Gegnern des Modells folgendermaßen beschrieben: „... soll die digitale Publikation auf den universitären Volltextservern zu „Open Access“-Konditionen erfolgen, das heißt eine beliebige und für die interessierten Leser kostenfreie Nachnutzung der Veröffentlichung erlauben. Das, so glaubt man, sei die gelungene Synthese aus einer digital sich selbst organisierenden und dank Ausschaltung der Verlage ökonomiefreien und daher billigeren Wissenschaft, die übers Internet mit der interessierten Öffentlichkeit direkt in Kontakt kommen und in diesem Direktkontakt die Demokratisierung der Gesellschaft voranbringen könne.“ (Uwe Jochum a.a.O.). Als Gegenargument zu Open Access wird vor allem angeführt, daß „Nachweisinstrumente“ wie Web of Science und Scifinder über Open Access eine „Kontrolle der Nutzer“ ermöglichen würden, die monetären Belohnungsflüssen sowie wissenschaftlicher und industrieller Spionage offenstehe: „Es gibt keinen Grund, von Open Access etwas anderes zu erwarten.“ Die Hauptzielrichtung dieses mit Verdächtigungen und Denunziationen arbeitenden Angriffs wird ebenfalls offengelegt: „Am Ende hat der Staat seine Wissenschaft verschenkt, aber es ist in Wahrheit kein Geschenk an seine Bürger, sondern ein Geschenk an Google und Konsorten. Viele glauben, das sei kein Problem, weil dadurch eine Win-win-Situation entstehe: Die Bürger bekommen im Netz eine kostenlose Wissenschaft, und andere verdienen damit auch noch Geld. Dass die Verlage bei diesem Spiel nichts mehr zu melden haben, sei ein hinnehmbarer Kollateralschaden. Sie übersehen aber, dass der eigentliche Schaden in der Wissenschaft angerichtet wird. Denn um wissenschaftliche Publikationen zu verschenken, muss man über die Geschenke auch verfügen können. Und hier liegt die eigentliche Crux: Die wissenschaftlichen Veröffentlichungen sind das Eigentum der Autoren.“
Es ist schon erstaunlich, wie hier Akteurskonstellationen verdreht werden: Wenn wissenschaftliche Veröffentlichungen das Eigentum der Autoren sind, dann ist es selbstverständlich, daß diese Autoren auch frei entscheiden können, wo und im Rahmen von welchem Modell sie publizieren. Niemand, keine Universitätsverwaltung, kein Ministerium, kein Drittmittelgeber o.ä. kann sich darüber hinwegsetzen, wenn ein Autor oder eine Autorin von einem verbürgten Grundrecht wie der Wissenschaftsfreiheit Gebrauch machen will.
Aber auch Wissenschaftlerinnen und Wissenschaftler bewegen sich nicht im Niemandsland, sondern sind meistens Angehörige einer öffentlichen Einrichtung, bewegen sich also innerhalb eines gewissen institutionellen Rahmens. Und die von mir schon mehrfach erwähnte Dreifachfinanzierung wissenschaftlicher Publikationen durch die öffentliche Hand (Gehälter der Wissenschaftlerinnen und Wissenschaftler, Druckkostenzuschüsse für deren Publikationen, Finanzierung der Bibliotheken, die diese Publikationen dann kaufen,[4] s. meinen Kommentar zu der harschen Kritik an Open Access – von Roland Reuss in der FAZ -: http://redaktionsblog.hypotheses.org/3041) sollte dabei aber nicht vergessen werden. Die öffentliche Finanzierung der Wissenschaft impliziert m.E. nicht nur ein Recht auf Wissenschaftsfreiheit, sondern auch eine Verpflichtung der Wissenschaftlerinnen und Wissenschaftler der Öffentlichkeit gegenüber: Nicht nur hat die Öffentlichkeit ein Recht auf Transparenz, d.h. daß ihr die Ergebnisse dessen, was durch ihre Mittel (vulgo: Steuern) finanziert wird, ohne weitere Hemmschwelle

4 „Im Jahr 2014 gaben die öffentlichen Bildungs- und Wissenschaftseinrichtungen etwa eine Milliarde Euro für den Erwerb von urheberrechtlich geschütztem Material aus. Davon entfiel etwa die Hälfte auf die wissenschaftlichen Bibliotheken von Hochschulen und Forschungseinrichtungen. Sie verwenden rund 40 Prozent ihres Etats auf die Lizenzen von E-Journals und E-Books.“ (https://www.bmbf.de/de/urheberrecht-im-dienst-der-wissenschaft-3229.html, abgerufen am 25.11.2016).

oder andere Hindernisse (vulgo: Kaufpreis) zur Verfügung stehen müssen, sondern dies muß auch das Recht umfassen, die weitere Verwertung zu bestimmen. Das kann so geschehen, daß über ein Gesetz festgelegt wird, daß Wissenschaftlerinnen und Wissenschaftler selbst entscheiden können, wie die Verwertung ihrer Ergebnisse bzw. Publikationen aussehen soll, oder es kann auch so aussehen, daß Universitäten, Bundes- oder Landesregierungen festlegen, daß die Veröffentlichung von Ergebnissen, die im Rahmen der Beschäftigung an einer öffentlichen Institution wie der Universität entstanden ist, auch innerhalb der öffentlichen Institutionen (d.h. z.B. durch die Bibliotheken) zu erfolgen hat. Natürlich entstehen für diese Bereitstellung im Open Access auch Kosten (z.B. für das Betreiben der Publikationssysteme), diese gehören jedoch in den Bereich der Finanzierungsmodalitäten aus öffentlichen Mitteln und sollten nicht mit privatwirtschaftlichen Bereichen vermischt werden. Ein gelungenes Beispiel für die Übernahme der Bereitstellung durch die Bibliotheken ist PROPYLAEUM, der Fachinformationsdienst Altertumswissenschaften (http://www.propylaeum.de/home/ und die Diskussion bei: http://www.hsozkult.de/debate/id/diskussionen-3880).

Wie absurd die derzeitige Diskussion mit den Vorwürfen der Open Access-Gegner abläuft, läßt sich anhand einiger Beispiele demonstrieren:
Man stelle sich dies übertragen auf die Inanspruchnahme von öffentlichen Einrichtungen vor: Deren Dienste (Polizei, Schulen, Bibliotheken) können von uns ohne Kosten in Anspruch genommen werden. Deren Bereitstellung kostet natürlich auch, allerdings werden diese Kosten der Bereitstellung aus Steuern finanziert. Übertragen auf die Rolle der privatwirtschaftlichen Verlage im Hinblick auf die Publikation wissenschaftlicher Ergebnisse müßte man sich hier vorstellen, daß eine private Einrichtung Gebühren – vergleichbar mit dem Kaufpreis für ein wissenschaftliches Buch – für die Zugänglichkeitsmachung der Dienste der öffentlichen Institutionen erheben würde.
Ein anderes Beispiel: Wenn die Transparenz im Hinblick auf wissenschaftliche Publikationen im Open Access zu vermehrter Kontrolle führt, so sollte man sich vor Augen halten, wie die Situation in früheren Zeiten ohne diese Transparenz gewesen ist und was überhaupt Nachweissysteme bedeuten. In den Zeiten, in denen wir noch keine Online-Bibliographien und keine offen und für jedermann und jede Frau frei zugänglichen Texte bzw. Bücher im Internet hatten und in denen es generell nur möglich gewesen war, über eine Fernleihe oder einen Kauf des begehrten Objektes zu überprüfen, ob eine wissenschaftliche Behauptung richtig war, waren wir in vielerlei Hinsicht blind (das zeigen die jetzt vermehrt sichtbar werdenden Plagiatsfälle in aller Deutlichkeit). Und um dies auf eine grundsätzliche Ebene zu heben: Nachweissysteme sind nicht grundsätzlich schlecht und Kontrolle ist ebenfalls nicht generell von Übel. Wir haben uns an diverse öffentliche Nachweissysteme gewöhnt, deren Sinn heute weder infrage gestellt wird, noch das zugrunde liegende Recht der staatlichen Institutionen bestritten wird, diese Nachweissysteme zu betreiben (z.B. die Ausweispflicht anhand der Personalausweise oder Steuererklärungen). Mißbrauch ist natürlich nie auszuschließen, aber dieser ist in der Regel strafbar und kann daher zwar nicht verhindert, aber durchaus begrenzt werden.
Um den Weg nun wieder zurück zu den Open Access Publikationen zu führen, soll hier – obwohl dies an vielen Orten und mit guten Argumenten schon geschehen ist – nur auf die Budapester Open Access Initiative hingewiesen werden (http://www.budapestopenaccessinitiative.org/translations/german-translation):
„Open Access meint, dass diese Literatur kostenfrei und öffentlich im Internet zugänglich sein sollte, so dass Interessierte die Volltexte lesen, herunterladen, kopieren, verteilen, drucken, in ihnen suchen, auf sie verweisen und sie auch sonst auf jede denkbare legale Weise benutzen können, ohne finanzielle, gesetzliche oder technische Barrieren jenseits von denen, die mit dem Internet-Zugang selbst verbunden sind. In allen Fragen des Wiederabdrucks und der Verteilung und in allen Fragen des Copyright überhaupt sollte die einzige Einschränkung darin bestehen, den jeweiligen Autorinnen und Autoren Kontrolle über ihre Arbeit zu belassen und

deren Recht zu sichern, dass ihre Arbeit angemessen anerkannt und zitiert wird."
Schließlich und endlich: Es geht hier nicht nur um ein Recht von Wissenschaftlerinnen und Wissenschaftlern im Rahmen ihrer Verpflichtungen, sondern es geht auch und gerade um wissenschaftliche Arbeitsmethoden. Durch die Digitalisierung unserer Publikationsmedien und die an die zunehmende Digitalität von Texten und Objekten anschließenden Methoden (z.B. Text- und Datamining) eröffnen sich neue Arbeitsfelder und neue Erkenntnisbereiche. Gerade Open Access ist (aus den o.g. Gründen und Beispielen gut ableitbar) eine wesentliche Voraussetzung dafür, daß unsere Wissenschaften diese Möglichkeiten wahrnehmen und erschließen: Die in viel höherem Maße durch Diskursivität und Fluidität geprägte Welt des Digitalen erfordert ein Mehr an kritischem Bewußtsein, ein Mehr an Bemühen um wissenschaftlich abgesicherte Leitplanken, ein Mehr auch an Verantwortungsbewußtsein. Solchen Herausforderungen ist nicht durch das Hochziehen von Mauern und Restriktionen zu begegnen. Vielmehr ist es Aufgabe der Wissenschaft, dies anzunehmen und in den jeweiligen Disziplinen fachspezifisch angemessene Antworten zu finden.

Autorenkontakt[5]

Prof. Dr. Charlotte Schubert
Universität Leipzig
Historisches Seminar
Lehrstuhl für Alte Geschichte

Email: schubert@uni-leipzig.de

URL: https://www.gko.uni-leipzig.de/historisches-seminar/seminar/alte-geschichte/professur.html

5 Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei der Autorin.

DIGITAL CLASSICS ONLINE

Ein Bild sagt mehr als 1000 Worte? Visualisierungen in den Digital Humanities.

Gary S. Schaal & Kelly Lancaster

Abstract: In Digital Humanities, computer-generated visualizations are viewed as highly significant in obtaining scientific insights. However, only through a reflection on their theoretical foundations can we exhaust the epistemological potential of visualizations abiding by the principles of validity and reliability. Digital Humanities is still lacking both an epistemological basis and a best practice for an (hermeneutic) interpretation of visualizations generated by algorithms. This paper will address precisely this research gap in raising the question whether, and to what extent, approaches to the hermeneutic interpretation of computer-generated visualizations in the natural sciences can be applied to analyses in the Digital Humanities. It will provide an answer to this issue with recourse to Don Ihde's theory of Postphenomenology. Though Postphenomenology supplies an epistemology and a visual hermeneutics for visualizations, both originate from and target solely the natural sciences. Whether the theory is applicable in the Digital Humanities is subject of further research.

1. Einleitung und Erkenntnisinteresse

Visualisierungen sind ein Instrument der Wissensgenerierung in allen Wissenschaften.[1] Es existieren hinsichtlich der Nutzung und Bedeutung algorithmisch generierter Visualisierungen zeitliche Differenzen zwischen den Naturwissenschaften und den Geisteswissenschaften. Während algorithmische Visualisierungen in den Naturwissenschaften seit mehreren Jahrzehnten konstitutiv für die Wissensgenerierung sind,[2] musste DiSalvo noch 1998 feststellen: „However, evidence of the use of such extensive computational visualization techniques in the humanities is lacking."[3] Erst in den letzten Jahren avancierten Visualisierungen in den digitalen geistes- und sozialwissenschaftlichen Disziplinen – den Digital Humanities (DH) und den Computational Social Sciences (CSS) – zu einer zentralen Quelle der Wissensgenerierung.[4]
Auf einer deskriptiven Ebene muss konstatiert werden, dass die Naturwissenschaften einen Wissens- und Erfahrungsvorsprung in der Nutzung von algorithmisch erzeugten Visualisierungen für die Generierung von Wissen besitzen. Zugleich existiert noch keine Epistemologie und Methodologie der (hermeneutischen) Interpretation von algorithmisch produzierten Visualisierungen in den Geisteswissenschaften.[5]
Dies wirft die erkenntnisleitende Frage auf, ob in den Naturwissenschaften vorhandene oder auf die Naturwissenschaften zielende Theorien und Ansätze für die hermeneutische

1 Vgl. Keim et al. 2010 und Ihde 1998.

2 Vgl. Rosenberger/Verbeek 2015 und Rosenberger 2009.

3 DiSalvo 1998, 83.

4 Vgl. Huang 2014; Seifert et al. 2014 und Marchese/Banissi 2013.

5 Vgl. Kitchin 2014 und Cecire 2011a, 2011b.

Interpretation von Visualisierungen in den Digital Humanities adaptiert werden können. Dieses Erkenntnisinteresse fokussieren wir aufgrund seiner herausragenden Bedeutung innerhalb des Feldes spezifischer auf das Werk des Technikphilosophen Don Ihde.[6] Er identifiziert computergenerierte Visualisierungen als die primäre Form der Erkenntnisgenerierung in der Gegenwart und hat eine korrespondierende Phänomenologie und Hermeneutik von Visualisierungen – *postphenomenology and visual hermeneutics* – entwickelt, die auf eine Epistemologie der Bildgebungen in den Naturwissenschaften zielt.[7]
Die grundlegende Frage der Anwendbarkeit von Ihdes Postphänomenologie in den Digital Humanities soll im Rahmen des Aufsatzes in drei Teilfragestellungen konkreter aufgearbeitet werden:

> *Welche grundlegenden Unterschiede existieren hinsichtlich der visualisierten Daten in den Natur- und den Geisteswissenschaften und welche Konsequenzen besitzen diese auf der hermeneutischen Ebene?*
> *Welche Ansätze existieren, die die hermeneutische Interpretation von Visualisierungen rückbinden an individuelle wie kulturelle Bedingungen, Voraussetzungen und Kontexte von visueller Wahrnehmung?*
> *Können die Überlegungen zu diesen Fragen so mit der Rekonstruktion von Ihde ‚geschnitten' werden, dass eine erste Skizze für eine Hermeneutik der Interpretation von Visualisierungen in den Geisteswissenschaften vorgelegt werden kann?*

Vor dem Hintergrund dieses Erkenntnisinteresses gliedert sich der Aufsatz in vier Argumentationsschritte. Im ersten Schritt wird der Stand der Forschung zur grundlagentheoretischen Fundierung der Digital Humanities unter besonderer Berücksichtigung von Visualisierungen rekonstruiert. Im zweiten Schritt diskutieren wir die für valide Erkenntnisgewinnung notwendige Einbettung von Visualisierungen in eine digitale Forschungsinfrastruktur. Im dritten Schritt rekonstruieren wir den aktuellen Diskurs innerhalb der Postphänomenologie-Studien unter besonderer Berücksichtigung des Ansatzes von Don Ihde. Abschließend werden wir die Erkenntnisse transponieren und auf ihre Anwendbarkeit in den Geisteswissenschaften hin überprüfen. Dies kulminiert darin, die erste Skizze eines eigenen Ansatzes – die *New Visual Hermeneutics* – zu präsentieren und zu verdeutlichen, wo er über den bisherigen Stand der Forschung hinausgeht. Wir schließen mit einem Ausblick auf notwendige Anschlussforschung.

2. Diskussion des Forschungsstandes

2.1 Diskurse über die grundlagentheoretische Fundierung der Digital Humanities

Die postulierte Notwendigkeit einer Hermeneutik für die Interpretation algorithmisch generierter Visualisierungen in den Digital Humanities ist für uns eingebettet in die vorgängige Notwendigkeit der Explikation einer grundlagentheoretischen Fundierung der Digital Humanities. Deshalb wenden wir uns dem Stand der Forschung in diesem Bereich zu.
Die Bedeutung von epistemologischen Fragen für die weitere Entwicklung der Digital Humanities wurde bereits 2004 adressiert,[8] später jedoch weder systematisch aufgegriffen noch in

6 Vgl. Ihde 1998; 2009b.

7 Vgl. auch Verbeek 2007.

8 Vgl. Schreibman et al. 2004.

der Forschungspraxis berücksichtigt.[9] Trotz Literatur zur Bedeutung der theoretischen Fundierung der Digital Humanities[10] stand bis vor Kurzem im Kern des Selbstverständnisses der Digital-Humanities-Community – zumindest in den USA – die Position, dass sie sich maßgeblich in der praktischen Arbeit bewähren und nicht in der theoretischen Reflexion.[11] Entsprechend kritisiert Cecire, „Digital Humanities is undertheorized“,[12] während Takahashi betont, dass „the emphasis on creating and the process and creation itself being the theory seems fitting for the Digital Humanities“.[13]
In der aktuellen Literatur findet sich jedoch vermehrt die Argumentation, dass die Digital Humanities einer grundlagentheoretischen und methodologischen Reflexion bedürfen.[14]

2.2 Diskurse über Visualisierungen in den Digital Humanities

Es sind – dem Selbstverständnis der Community folgend – maßgeblich Visualisierungen, mit deren Hilfe in den Digital Humanities (latente) Informationen in (manifestes) Wissen überführt werden. Die Bedeutung von Visualisierungen für die Digital Humanities ist daher nicht zu überschätzen: „Visualization is an effective enabler for exploratory analysis, making it a powerful tool for gaining insight into unexplored data sets.“[15] Wenn Visualisierungen zentral für Erkenntnisgewinne in den Digital Humanities sind, ist die Erwartung naheliegend, dass eine ausgearbeitete Methodologie oder Hermeneutik für die Interpretation von Visualisierungen vorliegt. Dies ist jedoch nicht der Fall.[16] Deshalb diskutieren wir im Folgenden als Vorstufe einer Hermeneutik von Visualisierungen zentrale Herausforderungen von Visualisierungen in den Digital Humanities.
Die erste zentrale Herausforderung für die Anwendbarkeit von Visualisierungen zur Generierung von relevanten Einsichten in den Digital Humanities wird von Rensink adressiert: „Is there a best way to visually display a given dataset for a given task, and if so, can we find it?“[17] In den Digital Humanities kann grundlegend zwischen deduktiven/hypothesentestenden und induktiven/explorativen Erkenntnisinteressen unterschieden werden. Die Frage, welche konkrete Form der Visualisierung am besten ist „for a given task“, verlangt bei deduktiven oder explorativen/abduktiven Erkenntnisinteressen nach unterschiedlichen Antworten. Während Evaluationskriterien für deduktive/hypothesentestende Visualisierungen auf einer abstrakten Ebene unkritisch spezifiziert werden können (kann mit Hilfe einer Visualisierung eine konkrete Hypothese falsifiziert werden?), gestaltet sich eine Antwort für explorative Erkenntnisinteressen deutlich schwieriger, da a priori keine Kenntnis über die Informationen vorhanden ist, die eine Visualisierung zur Darstellung bringt. Dies liegt in der Natur einer explorativen Analyse. Wenn Visualisierungen in explorativer Perspektive genutzt werden, sollte eine Visualisierung *insights* ermöglichen – wie relevanter und innovativer Erkenntnisgewinn im Diskurs bezeichnet wird. Liu et al. argumentieren, „what is important is the ability to disco-

9 Vgl. Berry 2011, 4 und Ramsay/Rockwell 2012.

10 Vgl. Cecire 2011a; 2011b und Berry 2011.

11 Vgl. kritisch Cecire 2011a, 2011b und affirmativ Takahashi 2012.

12 Cecire 2011, 45.

13 Takahashi 2012, o. S.

14 Vgl. Rosenberger 2011; Berry 2011, 4; Rieder/Röhle 2012; Ramsay/Rockwell 2012; Kitchin 2014; Manovich 2016; Lemke/Wiedemann 2016; Scheuermann 2016, 61.

15 Vgl. Seifert et al. 2014, 190; vgl. auch Keim et al. 2010 und Sacha et al. 2016.

16 Vgl. Kitchin 2014 und Schaal/Lancaster/Dumm 2016.

17 Rensink 2014, 148.

ver high-value insights that somehow are ‚hidden' in the unimaginably vast amount of low-value digital data that is scattered and unstructured".[18] Rensink paraphrasiert diese Erwartung: „Just find something interesting. [...] But a precise specification should be attempted whenever possible."[19] Eine Visualisierung sollte somit in der Lage sein, das latente Wissen, das in den Daten vorhanden ist, durch eine angemessene Form ihrer graphischen Repräsentation für die Wissenschaftlerin erkennbar werden zu lassen. Daraus folgt ein Qualitätskriterium von Visualisierungen: die Ermöglichung von „high-quality insights".

Damit wird die zweite zentrale Herausforderung für Visualisierungen als Erkenntnisgeneratoren in den Digital Humanities adressiert, d. h. die Frage, wie *empirisch gemessen* werden kann, *was* eine *gute* Visualisierung auszeichnet: „What is the best way to measure how a given visualization works? How could we find the perceptual and cognitive factors that limit its performance? Could we determine if its design is optimal?"[20] Die größte epistemologische Herausforderung dieser Aufgabe besteht in der Konzeptdefinition von *insights* und ihrer empirischen Operationalisierung. Auch hier gilt es wieder grundlegend zwischen deduktiven und explorativen Erkenntnisinteressen zu differenzieren. Existiert ein Referenzpunkt – d. h., ist potenziell zu bestimmen, welche Informationen eine Visualisierung zur Darstellung bringt –, können Evaluationsstandards als Abweichungen zwischen latentem (in der Visualisierung „vorhandenem") und manifestem (von der Forscherin identifiziertem) Wissen definiert werden. Konstruieren Visualisierungen Realität(en) und führen erst hermeneutische Interpretationen zu Erkenntnisgewinnen, ist diese Form der Operationalisierung und Messung unmöglich.

Als weiteres abstraktes Qualitätskriterium definiert Rensink: „One of these is variability, the extent to which the extended system gives the same answers when given the same data."[21] Dieses Qualitätskriterium wird auch von anderen Autorinnen vertreten. Es folgt dem Ideal der Reproduzierbarkeit wissenschaftlicher Erkenntnis und ihrer daraus resultierenden Intersubjektivität. Wir stellen die normative Auszeichnung der Eineindeutigkeit von algorithmisch generierten Visualisierungen infrage, und dies nicht nur mit Blick auf explorative Erkenntnisinteressen. Vielmehr sollte in den Digital Humanities die Eineindeutigkeit – erkenntnisabhängig – ergänzt werden um die Wertschätzung der *Alterität*, der *systematischen* (nicht kontingenten!) Varianz von Visualisierungen. Denn Visualisierungen erzeugen Pfadabhängigkeiten der wissenschaftlichen Erkenntnis, die – wie wir im Rekurs auf Don Ihde zeigen werden – der Wissenschaftlerin nicht transparent sein müssen. *Alterität* kann Irritationen erzeugen, um latente Pfadabhängigkeiten der Erkenntnisgenese manifest werden zu lassen. Diese Zusammenhänge sind in der Literatur – jenseits der Digital Humanities – intensiver diskutiert worden.

Eindeutig ist, dass die Interpretation jener Aspekte einer Visualisierung, die *objektiv bedeutungstragend* sind, von individuellen Faktoren, dem Erfahrungshorizont der Forschenden und kulturellen Bias (verschiedene Lebenswelten) abhängig ist. Erkennen, Verständnis und Interpretation visueller Darstellungen hängen zusammen mit jeweiligen kulturellen bzw. gemeinsamen *visuellen Praktiken*, die das „how to read" oder „learning-to-see" (Ihde) einer Visualisierung ‚prädeterminieren' (visuelles Training, Vorannahmen des ‚richtigen' Sehens, Deutungsmuster). Foster unterscheidet zwischen „vision" und „visuality".[22] Er schlägt vor, „vision" auf anatomische, physische und geometrische/dimensionale Aspekte des Sehens zu verwenden, während „visuality" angereichert ist mit einem „variegated bundle of social factors involved in the process of seeing".

Damit betont Foster die *Historizität des Sehens*, d. h., Sehen ist eine soziale, subjektive, mit

18 Liu et al. 2013, 543.

19 Rensink 2014, 154.

20 Rensink 2014, 148.

21 Rensink 2014, 157.

22 Foster 1998, zit. n. Hentschel 2014.

Bedeutung aufgeladene Handlung und keine rein ‚objektive' körperliche Wahrnehmung. Folglich konkurrieren – kulturell und in den unterschiedlichen disziplinären „visual cultures" – verschiedene Weisen des Erkennens, der Fokussierung auf bestimmte Elemente oder ihres Ignorierens sowie wechselseitigen Vermittlung.[23] Umgekehrt existiert eine *Historizität der Bildproduktion*, die über epochenspezifische Symbolisierungsprozesse (vgl. u. a. Cassirers Philosophie der symbolischen Formen) und epochenspezifische soziokulturelle Kriterien und organisatorische Prinzipien stattfindet, quasi eine Art Syntax oder ‚Choreografie', und bei ihrer Interpretation mit dementsprechenden Appräsentations- und Apperzeptionsprozessen einhergeht.[24]

Anknüpfungspunkte existieren zur Gestaltpsychologie, z. B. zu Arnheims Differenzierung von „seeing of" und „seeing as".[25] Im ersten Fall ist sinnliche Wahrnehmung eine passive Rezeption, im zweiten eine aktive Interpretation, die auf dem Erkennen von sog. Gestalten, der Mustersuche und Mustererkennung basiert. Eine weitere Anknüpfung besteht zu Gestaltwechseln bei Figure-Ground-Prozessen. Beide Aspekte finden sich auch bei Ihdes Postphänomenologie und seiner Konzeptionalisierung der „multistability" bzw. „polymorphy". Zwischen den jeweiligen interpretativen Zuständen kann die Betrachterin *bewusst* und *intentional* entscheiden – und dieses Moment der Entscheidung transformiert die ‚passive' Beobachterin zu einer Konstrukteurin dessen, was die von ihr interpretierte visuelle ‚Realität' darstellt.

Aus diesen Überlegungen kann erstens gefolgert werden, dass eine Hermeneutik für Visualisierungen notwendig ist. Zweitens ist die Angemessenheit einer Visualisierung nicht nur abhängig vom Erkenntnisinteresse, sondern auch von der Beobachterin (in diesem Fall der Wissenschaftlerin und ihrer Zugehörigkeit zu einer *visual culture*). Drittens kann ein und dieselbe Visualisierung mehrere *valide* Deutungen besitzen – aber auch objektiv falsch sein.

Kommen wir ein letztes Mal auf Rensink zurück. Er argumentiert: „Given the difficulties faced in searching through all the alternatives possible for a design, and the fact that much is still unknown about the perceptual and cognitive mechanisms involved, the search for optimal – or even good – designs must be supplemented by empirical assessment."[26] Es steht außer Frage, dass eine Hermeneutik der Interpretation von algorithmisch generierten Visualisierungen auch empirisch fundiert, d. h. getestet, sein muss. Zugleich darf sie die erkenntnistheoretischen Fragestellungen, die sich bei Visualisierungen ergeben, nicht ausblenden. Hiermit kommen wir zur dritten Problemdimension.

Visualisierungen sind – aufgrund ihrer Alterität und Kontingenz – *eine Interpretation* der Daten; in ihnen vereinigen sich das phänomenologische Moment der algorithmischen Generierung einer Wirklichkeit mit dem hermeneutischen Moment ihrer Interpretation. Bei vielen Nutzerinnen von Visualisierungen in den Digital Humanities bestehen einerseits Unklarheiten über die Intensität der konstruktivistischen Dimension einer Visualisierung – häufig wird sie daher *at face value* interpretiert. Andererseits existiert häufig Unklarheit bzw. Desinteresse darüber, welche Algorithmen – und innerhalb der Anwendung von Algorithmen: welche Spezifikation ihrer Parameter – valide, d. h. bedeutungsvolle Visualisierungen generieren.[27] Chen et al. hierzu: „Das Problem beim Einsatz dieser tools ist, dass bei den AnwenderInnen häufig kaum Wissen über die verwendeten Algorithmen und damit über den Bedeutungsgehalt einzelner Dimensionen der Visualisierung vorhanden sind."[28]

Nicht ohne Grund hat sich in vielen Projekten der Digital Humanities die Praxis einer

23 Hentschel 2014, 28.

24 Vgl. Husserl 1973; Breckner 2012, 146.

25 Vgl. Arnheim 2004.

26 Rensink 2014, 165.

27 Vgl. Keim et al. 2010, 102.

28 Chen et al. 2008, 4.

Methodentriangulation mit *street-level-epistemology* etabliert, welche die Angemessenheit von Visualisierungen anhand von der Forscherin bekannten Interpretationen bzw. in der Literatur vorhandenem Wissen validiert. Schmidt greift dieses Problem auf und weist darauf hin, dass Visualisierungen von Textkorpora, die die Forscherin nicht gut kennt, selten einem Plausibilitätscheck unterzogen werden und „interpretive leaps are extraordinarily easy to make with texts".[29] Ähnlich argumentiert Brett: „The only way to know if your results are useful or wildly off the mark is to have a general idea of what you should be seeing."[30] Zusammenfassend muss daher Rieder/Röhle zugestimmt werden: „This does not mean that questions of visual arrangement are epistemologically innocent, quite the contrary."[31] Gerade deshalb benötigen die Digital Humanities eine grundlagentheoretische Fundierung, aus der eine Hermeneutik für die Interpretation von Visualisierungen bruchlos hervorgeht.

2.3 Visualisierungen und ihre Integration in eine Forschungsinfrastruktur

Die oben adressierten Fragen nach einem besseren Verständnis davon, unter welchen Bedingungen Visualisierungen zu bedeutsamen Einsichten beitragen können, werden weniger in den Digital Humanities, sondern im Bereich der Information Visualization[32] und der Visual Analytics diskutiert. Die aktuellen und elaborierten Ansätze in der Visual Analytics betten hierfür das erkenntnisgenerierende Potenzial von Visualisierungen *methodisch* in eine Forschungsinfrastruktur ein.[33] In diesem Feld sind in den letzten Jahren mehrere Sammelbände veröffentlicht worden, die Teilaspekte der oben angesprochenen Fragen behandeln.[34]
Zu den Protagonisten im Feld der avancierten Visual Analytics gehört die Gruppe des Konstanzer Visualisierers Daniel A. Keim. Seit ihrem grundlegenden Werk[35] arbeiten Keim und seine Arbeitsgruppe an der systematischen, methodischen Entfaltung einer Forschungsinfrastruktur, die auf Visualisierungen als zentralem Tool der Wissensgenerierung basiert. In neueren Publikationen wird der Fokus auf den Faktor „Unsicherheit" im Rahmen der Forschungsinfrastruktur gerichtet. So konzeptualisieren Sacha et al. die Forschungsinfrastruktur als einen Prozess, der zwei ineinandergreifende komplexe, dynamische und iterative Blöcke umfasst und dazu dient, Unsicherheiten („uncertainties") auf der Ebene des Umgangs mit den Daten aufzulösen und zu einer Zuverlässigkeit („trust") und damit der Validität von Visualisierungen und Wissenskonstruktion in den Visual Analytics zu führen.[36] Effizienz und Effektivität werden einerseits von der Rechnerleistung (Block 1), andererseits vom menschlichen Faktor, d. h. dem Bewusstsein („awareness") der Forschenden um diese Unsicherheiten bzw. Fehler beeinflusst, das mit dem statistischen Wissen um Zufälligkeiten, Samplegröße, Regressionen, Korrelationen etc. als auch „domain knowlegde", Vorwissen, „perceptual competence and visualisation literacy"[37] einhergeht (Block 2). Die Forschungsinfrastruktur ist für Keim und seine Gruppe

29 Schmidt 2012.

30 Brett 2012, 14.

31 Rieder/Röhle 2012, 69.

32 Vgl. Keim 2002 und James 2004.

33 Vgl. Keim et al. 2010; Kang/Stasko 2012; Endert et al. 2014; Nguyen et al. 2013; Santucci 2013 und Dietrich 2015.

34 Vgl. Huang 2014; Agosti et al. 2012 und Marchese/Banissi 2013.

35 Vgl. Keim et al. 2010.

36 Vgl. Sacha et al. 2016 und Sacha et al. 2014.

37 Sacha et al. 2016, 243.

eine interaktive Mensch-Computersystem-Relation („the human is the loop“ nach Endert et al.).
Der erste Block („system“) referiert auf Prozesse und die ihnen inhärenten Fragen innerhalb des rechnerbasierten bzw. informationstechnischen Systems. Dies beginnt mit der Datenquelle, der Datenaufbereitung, der Wahl eines adäquaten Modells, dem Setzen der Parameter bis hin zur Visualisierung. Im Laufe dieser Prozesse nehmen (kumulativ) Unsicherheiten und Fehleranfälligkeiten zu, die bereits mit der Datenqualität beginnen. Es gilt, die Verbreitung von Unsicherheiten wiederum zu minimieren, was durch konstante Feedbackloops über den zweiten Block („human“) verläuft. Innerhalb des zweiten Blocks – der menschlichen Expertise – interpretieren die Forschenden die jeweilige Visualisierung und generieren auf deren Basis Hypothesen, überprüfen die Visualisierung bzw. das ihr vorausgehende Modell, indem sie Analysestrategien anwenden, Verfeinerungen, Anpassungen und Neukalibrierungen an ihnen vornehmen. Berücksichtigt und bereinigt werden müssen Wahrnehmungs- und kognitive Bias der Visualisierung. Die Ergebnisse werden dem System zurückgespiegelt, sodass ein iterativer Analyseprozess in Gang gesetzt wird.

3. Hermeneutik und Postphänomenologie

Die exemplarisch skizzierten Probleme und Herausforderungen bei der Interpretation von Visualisierungen von Daten in den Geisteswissenschaften stehen nicht nur einer breiteren Akzeptanz der Digital Humanities im Wege, sondern auch der Generierung *valider* Erkenntnisse auf der Basis von Visualisierungen. Die Betonung der Relevanz einer Forschungsinfrastruktur für auf Visualisierungen basierender Wissensgenerierung, die bei Keim und seiner Gruppe methodisch-konzeptionell verankerten Rekursivitäten und Interaktionsmöglichkeiten sind ein bedeutsamer Schritt für die Entwicklung einer *Praxis* der validen Interpretation von algorithmischen Visualisierungen. Die Ausarbeitung einer Forschungsinfrastruktur in der Visual Analytics ist jedoch nur eine notwendige, aber keine hinreichende Bedingung für valide Wissensgenerierung. Hierfür muss aus der Perspektive des Autors und der Autorin die Epistemologie und Methodologie der Interpretation in den Fokus genommen und auf die Philosophie, insbesondere die Erkenntnistheorie, rekurriert werden. Im Folgenden werden zwei Ziele verfolgt: Auf der Basis einer Rekonstruktion der zentralen Elemente von Don Ihdes Postphänomenologie soll erstens plausibilisiert werden, dass Ihde ein gut geeigneter *Ausgangspunkt* ist für die Ausarbeitung einer Hermeneutik der Interpretation von Visualisierungen in den Geisteswissenschaften. Darüber hinaus soll zweitens verdeutlicht werden, dass Grenzen der Übertragbarkeit in die Geisteswissenschaften existieren, die einer präzisen Explikation bedürfen, um produktive weitere Entwicklungspfade in den Geisteswissenschaften aufzeigen zu können.
Vor dem Hintergrund der Science and Technology Studies (STS) entwickelt Don Ihde seinen empirischen Ansatz einer nichtfundamentalistischen und nichtessenzialistischen (experimentellen) Phänomenologie und visuellen Hermeneutik als eine Erkenntnistheorie für die Natur- und Technikwissenschaften, die er als „Postphenomenology“ bezeichnet.[38]
Ihde möchte die auf Dilthey zurückzuführende erkenntnistheoretische Dichotomie zwischen „Erklären“ und „Verstehen“ aufheben. Dilthey weist die Hermeneutik als die Methode des ‚Verstehens‘ den Geisteswissenschaften zu, während er das ‚Erklären‘ und Beschreiben von ‚Tatsachen‘ den Naturwissenschaften vorbehält. Ihde argumentiert, dass die Hermeneutik sich nicht auf die Geisteswissenschaften beschränkt, sondern die naturwissenschaftliche Praxis zur

38 Vgl. Ihde 1998; 2008; 2009a; 2009b; 2012.

Wissensgenerierung auch interpretierende, hermeneutische Anteile besitzt.[39] Zu diesem Zweck erweitert er die Hermeneutik, die sich ursprünglich auf die *Textexegese* bezog, zu einer visuellen Hermeneutik, indem er sie in Zusammenhang mit der sinnlichen Wahrnehmung, über die der Zugang zur ‚Realität' erst möglich wird, bringt. Dabei nimmt hier insbesondere die Visualität eine zentrale Funktion bei der hermeneutischen Deutung der ‚Realität' ein. Auf die Naturwissenschaften übertragen bedeutet dies, dass das zuvor nicht Sichtbare erst durch Instrumente und Technologien – z. B. einen Algorithmus – durch Visualisierung wahrnehmbar, d. h. ein phänomenologischer Zugang und die anschließende Sinngebung ermöglicht wird. Ihde geht von einer *relationalen Ontologie* aus, bei der Sinn und Bedeutung erst durch die *Beziehung(en)* zwischen den Dingen entstehen – und ihnen nicht als Substanz inhärent sind – und daher durch eine Veränderung der Beziehung(en) neuer Sinn generiert wird.[40]
Ihde konstatiert für das 20. und 21. Jahrhundert bahnbrechende Transformationen und Fortschritte in den (Natur-)Wissenschaften, die maßgeblich auf Technologien wie bildgebenden Verfahren und Visualisierungen gründen. Aufgrund dieser Prozesse plädiert er für eine gemeinsame Epistemologie für die Wissenschafts- und die Technikphilosophie und betont, dass Wissenschaft und Technologie nicht getrennt voneinander reflektiert werden sollten und im Konzept der „Technoscience" vereint werden können. Um dieser Situation gerecht zu werden, entwickelt Ihde seine Postphänomenologie, die mit Elementen der klassischen Phänomenologie und des Pragmatismus John Deweys angereichert ist und den „empirical turn", d. h. die Betonung von *case studies*, in der Philosophie der Technik als konstitutiv vorsieht.[41] Seine Überlegungen veranschaulicht er mithilfe von Gedankenexperimenten mit Bildern und wissenschaftlichen Visualisierungen bzw. Imaging-Technologien.
Von der klassischen Phänomenologie übernimmt Ihde für seine *postphenomenology* erstens die „variational theory" (Husserl), die aufzeigt, dass Visualisierungen mehrere Wahrnehmungsoptionen und damit multiple valide Interpretationen besitzen können, welche er als „multistability" oder „polymorphy"[42] bezeichnet. Ihdes Anwendung der *variational theory* ist im Gegensatz zu Husserl nichtessenzialistisch. Damit distanziert sich Ihde von Husserl, der mithilfe der *variational theory* qua eidetischer Reduktion die Essenz von Phänomenen ableiten wollte.[43] Zweitens integriert er das Prinzip des „embodiment" (Merleau-Ponty), welches besagt, dass der Zugang zur Welt über das körperliche Erleben (Wahrnehmung) und die körperliche Interaktion mit der Welt inklusive aller ‚prothetischen' (Hilfs-)Objekte, Technologien oder Interfaces ermöglicht wird (z. B. eine Brille). *Embodiment* und Interaktion mit der Welt inkludieren auch die Einnahme einer bestimmten Perspektive, sodass ein Perspektivwechsel zu einer veränderten Wahrnehmung führt. Drittens nimmt die Postphänomenologie Bezug zur intersubjektiv geteilten „Lebenswelt", d. h., die hermeneutische Analyse von Visualisierungen muss unterschiedliche soziokulturelle und epochenspezifische Lebenswelten als das jeweils ‚So-Gegebene' berücksichtigen, die demselben Phänomen variierende Bedeutungen unabhängig von der Wahrnehmung verleihen[44] und die *kulturelle Situiertheit* der Wahrnehmung im Auge behalten.
Sehen ist somit kein passiver Zustand, sondern bei der Konstruktion von Wissen und Wirklichkeit aktiv beteiligt. Da die Phänomenologie das körperliche Erleben bzw. die Erfahrung in der Welt untersucht, ist eine der zentralen Fragen, die Ihde stellt, wie Technologien ihre Vermittlungsfunktion zwischen unseren Körpern und der Welt erfüllen und auf welche Weise sie damit

39 Ihde 1998.

40 Vgl. Ihde 2008.

41 Vgl. Ihde 2008, 2009a.

42 Ihde 2008; 2009b; 524.

43 Ihde 2008.

44 Ihde 2009a, 11–19.

unsere Wahrnehmungsfähigkeiten in sogenannten „human-technology-relations“ verändern.[45] Das postphänomenologische Modell komplettiert Ihde, indem er vom Pragmatismus John Deweys den Aspekt der „experience“ übernimmt, der experimentelle Erfahrungen und Analysen der Lebenswelt einschließt.[46] Dewey vertritt die Auffassung, dass der Mensch in konkreter Interaktion mit seiner (soziokulturellen) Umgebung Erkenntnisse gewinnt – womit er Empirie und Praxis betont – und diese Erkenntnisse immer provisorisch sind.

Diese postphänomenologischen Vorannahmen bieten für den Zusammenhang von technologischem Fortschritt, technischen Instrumenten als „visual hermeneutic devices“[47] und der Generierung wissenschaftlicher Erkenntnis eine epistemologische Grundlage. Die Verflechtung von Wissenschaft und Technik liefert einen Kontext der Wissenskonstruktion, der sowohl sozial als auch technisch ist: „[P]ostmodern technologies used by science are active in the sense that they are more and more *constructive* rather than passive“.[48] Ähnlich argumentieren die Wissenschaftshistorikerinnen Lorraine Daston und Peter Galison. Für sie fungieren algorithmisch generierte Visualisierungen als *Werkzeuge* und sind „Teil des Herstellungsprozesses“, wobei das „Machen und Sehen“[49] zugleich auftreten. Auch sie verorten die Forscherin nach dem *digital* bzw. *computational turn*[50] in der Rolle eines „konstruierende[n] Selbst“.[51]

Die epistemologische Reflexion dieser empirischen naturwissenschaftlichen Praxis beinhaltet zusammengefasst den soziokulturellen und wissenschaftlichen Erfahrungshorizont der Forschenden, deren Datenauswahl, die Auswahl der Instrumente bzw. Technologie (hier der jeweilige Algorithmus und wie damit ein bestimmter Teilaspekt der ‚Realität‘ erzeugt wird), die Visualisierung und ihre Interpretation(en). Ihde betont vor allem die große Bedeutung des Messinstruments bzw. der Technologie und wie damit Wissen gewonnen wird. Je nach der verwendeten Technologie wird eine ganz bestimmte Beziehung zwischen den Forschenden und der ‚Realität‘ vermittelt, sie fungiert als ein „Interface“ zwischen Forschenden und der ‚Wirklichkeit‘. Dieser Zugang ist nicht neutral, da zum einen durch die Wahl des Algorithmus/Codes nur bestimmte Informationen in der Visualisierung zum Ausdruck gebracht werden (Pfadabhängigkeit), zum anderen die Visualisierung eine „multistability“ mit mehreren *validen* Interpretationen sein kann.

Ihde plädiert folglich für eine ontologische Priorisierung von Technologie gegenüber Wissenschaft und Theorie, da erst Technologie zu einem empirischen Zugang, einem *erweiterten* Erleben der Welt – einer „technologically mediated lifeworld“[52] – und dem Experimentieren in ihr ermächtige. Die über die Instrumente bzw. Algorithmen/den Code ‚nahegebrachten Phänomene‘ der Visualisierungen erlauben erst eine visuelle Hermeneutik. Phänomenologie und Hermeneutik durchdringen sich und das Unsichtbare wird sichtbar gemacht durch Technologie: „[A]ll this instrumentation designed to turn all phenomena into visualizable form for a ‚reading‘ illustrates [...] ‚hermeneutic practices‘ [since] imaging technologies [...] make nonvisual sources into visual ones“.[53] Dabei unterscheidet Ihde zwischen Visualisierungstech-

45 Vgl. Ihde 1998; Rosenberger 2011. Die zwei „Blöcke“, die Keim bei der Explikation seiner Forschungsinfrastruktur identifiziert, stellen eine „human-technology-relation“ im Sinne Ihdes dar. Hieran wird deutlich, dass die Postphänomenologie von Ihde eine Epistemologie bereithält, die avancierte Forschungsinfrastrukturen in den Visual Analytics grundlagentheoretisch fundieren kann, um so bessere Praxen valider Wissensgenerierung zu ermöglichen.

46 Ihde 2009a, 11.

47 Ihde 2009b, 518.

48 Ihde 2009a, 62, Hervorhebung d. A.

49 Daston/Galison 2007, 409.

50 Vgl. Berry 2011.

51 Daston/Galison 2007, 413.

52 Vgl. Tripathi 2004.

53 Ihde 2009b, 518.

nologien, die „isomorphically visual forms“ erzeugen – u. a. Kernspintomographie, Positronenemissionstomographie, Computertomographie etc. –, und semiotischen Visualisierungen wie Graphen, Diagramme, Spektrographen als „translation technologies“.[54]
Greift man auf die skizzierte Phänomenologie von Ihde zurück, *konstruieren* Visualisierungen für die Forscherin erst eine Welt interpretierbarer Daten. Im Kern handelt es sich um eine Konstruktion zweiter Ordnung, da die Visualisierung auf den Ergebnissen algorithmischer Analysen basiert, die ihrerseits eine Konstruktion erster Ordnung darstellt. Für z. B. qua Größe oder Komplexität nicht mehr direkt interpretierbarer Daten stellen *Visualisierungen hermeneutische Instrumente der Sinngenerierung* dar.
An das Werk von Ihde schließt ein postphänomenologischer Diskurs an, der zentral aus Fallstudien aus dem naturwissenschaftlichen Bereich besteht. Gleichwohl charakterisieren Rosenberger/Verbeek „postphenomenology“ als eine „developing school of thought“[55], an die unterschiedliche Disziplinen anknüpfen können, da die Postphänomenologie sowohl theoretischer Rahmen als auch Methodologie für empirische Studien sein kann – was zahlreiche *case studies* auch beweisen. Die postphänomenologische Idee findet in anderen Feldern innerhalb der Philosophie wie der Ethik oder Philosophie des Selbst, aber auch in anderen Disziplinen wie der Anthropologie, Soziologie, Cultural Studies oder Medienwissenschaft Anwendung. Rosenberger/Verbeek verstehen Postphänomenologie als flexiblen Rahmen, wobei Fallstudien als „laboratories within which postphenomenological ideas are interrogated and refined“[56] fungieren. Innerhalb der Postphenomenology Studies liegt ein inhaltlicher Fokus auf vergleichenden bildgebenden Verfahren.[57] Hier existieren direkte Verbindungen zu unserer oben ausgeführten Argumentation zugunsten von Alterität und systematischer Varianz von Visualisierungen als grundlegende Strategie zur Generierung validen Wissens durch Visualisierungen in den Digital Humanities.
Trotz des Framings der Postphänomenologie als eines sich entwickelnden epistemologischen Rahmens für unterschiedliche Disziplinen ist die überwältigende Mehrheit aller (empirischen) Studien in diesem Feld in den Naturwissenschaften verortet. Damit stellen sich alle Fragen der Übertragbarkeit und Fruchtbarmachung für die Geisteswissenschaften, die sich bei Ihde gestellt haben, auch für den daran anschließenden Postphänomenologiediskurs.

4. Eine postphänomenologische Forschungsinfrastruktur – der Ansatz der New Visual Hermeneutics

Die obigen Ausführungen konnten verdeutlichen, dass der Ansatz von Don Ihde für die grundlagentheoretische Fundierung der Digital Humanities vielversprechend ist. Sein Ansatz der *nonfoundationalist* Postphänomenologie ist in der empirischen und experimentellen Dimension der Naturwissenschaften verortet. Die Basis seiner epistemologischen Reflexion über den Zusammenhang von technischem Fortschritt, Imaging Technologies/Instrumenten und wissenschaftlicher Erkenntnis ist das *messbare Ereignis*, ein empirisches Datum, das im Zuge technologischen Fortschrittes zunehmend der Sichtbarmachung, der Visualisierung, bedarf, d. h.: Daten werden algorithmisch prozessiert und mit Hilfe von Algorithmen visualisiert.

54 Ihde 2009b, 518.

55 Rosenberger/Verbeek 2015, 1.

56 Rosenberger/Verbeek 2015, 32.

57 Vgl. Hasse 2008; Verbeek 2008; Rosenberger 2011; Carusi/Hoel 2014; Hoel/Carusi 2015 und Friis 2015.

Visualisierungen besitzen *in der Form ihrer Darstellung* keine Entsprechung in der Realität; trotzdem kann eine Forscherin Expertise in der Interpretation von Visualisierungen erwerben und darüber den Erkenntnisfortschritt fördern.
Ob sich die *Postphenomenology* mit ihrer *visual hermeneutics* aus ihrem naturwissenschaftlichen Kontext auf die digitalen Geisteswissenschaften transponieren lässt, bedarf einer eingehenderen Analyse, als wir sie hier vorlegen können. Fest steht, dass in den Digital Humanities Hardware, Algorithmen und die auf ihrer Basis generierten Visualisierungen im Sinne Ihdes eine Form von Technologie darstellen, ein Medium der Wissensgenerierung sind und Realität konstruieren. Soweit ist ihm auch in den Geisteswissenschaften zu folgen. In den Digital Humanities werden *soziokulturelle* Phänomene, *sinnbehaftete* kulturelle Artefakte, visualisiert und wären – aus Ihdes Perspektive – somit prinzipiell einer *materiellen* Hermeneutik zugänglich. Doch markiert ‚Sinn' die Scheidelinie zwischen Natur- und Geisteswissenschaften. Daher muss ein kategorialer Unterschied zwischen der materiellen Hermeneutik in den Natur- und den Geisteswissenschaften existieren, obwohl beide in der Postphänomenologie wurzeln.
Unabhängig von der zentralen Frage, welche Gestalt dieser kategoriale Unterschied annimmt und welche Konsequenzen für eine materielle Hermeneutik der Visualisierung in den Digital Humanities daraus resultieren, folgt aus der Rekonstruktion von Ihde die Einsicht: Wenn Visualisierungen hermeneutische Instrumente sind, müssen sie als Teil des Erkenntnisprozesses modelliert werden. Sie markieren nicht den Abschluss einer Forschungspipeline, sondern müssen als ein konstitutiver Teil eines iterativ aufgesetzten Prozesses der Erkenntnisgewinnung, in der Visualisierungen alterieren, verstanden werden. Das Moment der Alterität betrifft dabei sowohl die Ausgestaltung einer konkreten Visualisierungsform, die Form der Visualisierung als auch die algorithmische Grundlage der Visualisierung. Da Visualisierungen immer eine Interpretation von algorithmisch prozessierten Daten und somit Konstruktionen zweiter Ordnung sind, wird für deren Verständnis eine Form der Hermeneutik benötigt, eine Hermeneutik zweiter Ordnung, die den *Prozess der Generierung von Visualisierungen konstitutiv* berücksichtigt. Einen solchen Ansatz entwickelt eine Gruppe um die Autoren dieses Aufsatzes unter dem Titel *New Visual Hermeneutics*.[58] Die *New Visual Hermeneutics* ist ein methodischer Ansatz für die Generierung von Wissen mittels algorithmischer Analysen aus unstrukturierten Textdaten mit Hilfe von *Information Visualization*.[59] Unser Ansatz ist im Forschungsfeld der Visual Analytics[60] verortet und kann auch als Forschungsinfrastruktur spezifiziert werden.[61]

58 Vgl. Kath/Schaal/Dumm 2015; Lemke/Niekler/Schaal/Wiedemann 2015; Schaal/Kath/Dumm 2016 und Schaal/Lancaster/Dumm 2016.

59 „One core difference between Information Visualization and Visual Analytics lies in the support of analytical work flows and the generation and validation of hypothesis" (Seifert et al. 2014, 197).

60 Keim et al. 2010; Endert et al. 2014; Sacha et al. 2014 und Sacha et al. 2016.

61 „Visual Analytics is an interdisciplinary field based on information visualization, knowledge discovery and cognitive and perceptual sciences, which deals with designing and applying interactive visual user interfaces to facilitate analytical reasoning" (Seifert et al. 2014, 190).

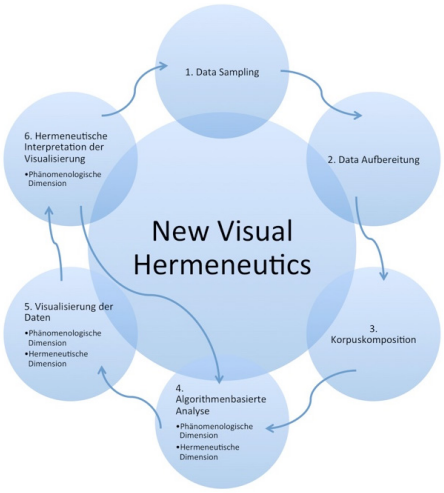


Abb. 1: Die Forschungspipeline aus der Perspektive der New Visual Hermeneutics

Die *New Visual Hermeneutics* fokussiert auf die Forschungsinfrastruktur und betont die grundlagentheoretische, epistemologische und methodologische Dimension der *gesamten* Forschungsinfrastruktur. Dies impliziert, dass die Qualität der Arbeit in den Digital Humantities im Rahmen einer Forschungsinfrastruktur zentral davon abhängt, dass die Forscherinnen die methodischen, theoretischen und epistemischen Implikationen *aller* Phasen des Forschungsprozesses kennen und um die Herausforderungen beim Übergang von einer Phase zur nächsten wissen.

5. Ausblick

Vor dem Hintergrund der ausgeführten Überlegungen muss der nächste Schritt in die empirische Praxis führen. Angeleitet durch grundlagentheoretisch fundierte Forschungsinfrastrukturen – wie z. B. die New Visual Hermeneutics – gilt es zukünftig die Frage zu beantworten, welche *praktischen* Unterschiede für die Hermeneutik von Visualisierungen aus unterschiedlichen Typen von Korpora resultieren. Als Leitdifferenz des vorliegenden Aufsatzes diente ‚naturwissenschaftliche vs. geisteswissenschaftliche' Daten. Hinter dieser Leitdifferenz steht die Materialität der Daten in Verbindung mit ihrer Sinndimension. *Innerhalb* der geisteswissenschaftlichen Seite gilt es zukünftig nach unterschiedlichen *Typen* von Daten und Datenkorpora zu differenzieren. So sollte zwischen *unimodalen* (z. B. Textkorpora) und *multimodalen* Korpora (z. B. Textkorpora und Geodaten) in den Geisteswissenschaften differenziert werden, da die epistemische und hermeneutische Komplexität der Visualisierung von multimodalen Daten weitaus höher ist als jene von unimodalen Daten.

Obwohl in verschiedenen Digital Humanities-Projekten bereits eine Vielzahl von Tools und Verfahren entwickelt wurden, mit denen computergestützt geisteswissenschaftliche Fragestellungen beantwortet werden können, fehlt es doch an einer Best Practice, an der sich Forscherinnen in den Digital Humanities orientieren können.[62] Auf einer methodischen und methodologischen Ebene sind wir noch deutlich von einem Konsens entfernt, wie geisteswissenschaftlich ‚sauber' aufgesetzte und durchgeführte Analysen in den Digital Humanities aussehen. Durch die Orientierung an einer Forschungsinfrastruktur wie den New Visual Hermeneutics in der Forschungspraxis kann dieses Defizit überwunden und in forschungspragmatischer Perspektive eine Best Practice generiert werden. Darüber hinaus würde eine Best Practice – durch die erhöhte Vergleichbarkeit von Studien – auch einen zentralen Beitrag zu schnellerem *kumulativen Wissensgewinn* in den Digital Humanities leisten.

62 Vgl. v. a. http,//www.clarin-d.de/de/.

Literaturverzeichnis

Agosti, Maristella / Ferro, Nicola / Forner, Pamela / Müller, Henning / Santucci, Giuseppe (Hg.) (2012): Information Retrieval Meets Information Visualization, New York / Heidelberg / London.

Anderson, Chris (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, in: Wired [www.wired.com/2008/06/pb-theory] 10.04.2016.

Arnheim, Rudolf (2004): Visual Thinking, Berkeley.

Berry, David M. (2011): The Computational Turn: Thinking About the Digital Humanities. In: Culture Machine 12, 1–22.

Breckner, Roswitha (2012): Bildwahrnehmung – Bildinterpretation. Segmentanalyse als methodischer Zugang zur Erschließung bildlichen Sinns. In: Österreichische Zeitschrift für Soziologie 37, 143–164.

Brett, Megan N. (2012): Topic Modeling: A Basic Introduction. In: Journal of Digital Humanitites 2, 12–16; http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/, 01.12.2016.

Burkhard, Remo (2006): Knowledge Visualization: Die nächste Herausforderung für Semantic Web Forschende? In: Tassilo Pellegrini/Andreas Blumauer (Hg.), Semantic Web, Berlin/Heidelberg, 201–212.

Carusi, Annamaria / Hoel, Aud Sissel (2014): Toward a New Ontology of Scientific Vision. In: Catelijne Coopmans / Janet Vertesi, / Michael E. Lynch / Steve Woolgar (Hg.), Representation in Scientific Practices Revisited, Cambridge (MA) / London, 201–221.

Cassirer, Ernst (2010): Philosophie der symbolischen Formen 1: Sprache, Hamburg.

Cecire, Natalia (2011a): Introduction: Theory and the Virtues of Digital Humanities. In: Journal of Digital Humanities 1, 45–53; http://journalofdigitalhumanities.org/1-1/introduction-theory-and-the-virtues-of-digital-humanities-by-natalia-cecire/, 01.12.2016.

Cecire, Natalia (2011b): When Digital Humanities was in Vogue. In: Journal of Digital Humanities 1, 54–59; http://journalofdigitalhumanities.org/1-1/when-digital-humanities-was-in-vogue-by-natalia-cecire/, 01.12.2016.

Daston, Lorraine / Galison, Peter (2007): Objektivität, Frankfurt a. M.

Dietrich, David. 2015. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Indianapolis; http://public.eblib.com/choice/publicfullrecord.aspx?p=1908952, 14.05.2016.

Endert, Alex / Hossain, M. Shahriar / Ramakrishnan, Naren / North, Chris /Fiaux, Patrick / Andrews, Christopher (2014): The human is the loop: new directions for visual analytics. In: Journal of Intelligent Informations Systems 43, 411–435.

Friis, Jan-Kyrre Berg Olsen (2015): Gestalt descriptions: embodiments and medical image interpretation. In: AI & Society 30, 1–9.

Hasse, Cathrin (2008): Postphenomenology: Learning Perception in Science. In: Human Studies 31, 43–61.

Hentschel, Klaus (2014): Visual Cultures in Science and Technology, Oxford.

Heyer, Gerhard / Schaal, Gary S. / Dumm, Sebastian / Lemke, Matthias / Niekler, Andreas/ Wiedemann, Gregor (2016): Postdemokratie und Neoliberalismus. Textmining, VS, i. E.

Hoel, Aud Sissel / Carusi, Annamaria (2015): Thinking Phenomenology with Merleau-Ponty. In: Robert Rosenberger / Peter-Paul Verbeek (Hg.), Postphenomenological Investigations – Essays on Human-Technology Relations, Lanham u. a., 73–84.

Huang, Weidong (Hg.) (2014): Handbook of Human Centric Visualization, New York / Heidelberg / Dordrecht / London.

Husserl, Edmund (1973): Cartesianische Meditationen und Pariser Vorträge, hg. v. S. Strasser, Husserliana Band 1, Den Haag / York / Albany.

Ihde, Don (1998): Expanding Hermeneutics: Visualism in Science, Evanston (IL).

Ihde, Don (2008): Introduction: Phenomenological Research. In: Human Studies 31, 1–9.

Ihde, Don (2009a): Postphenomenology and Technoscience, New York / Albany.

Ihde, Don (2009b): Scientific Visualism. In: David M. Kaplan (Hg.), Readings in the Philosophy of Technology, 2. Auflage, Lanham, 517–533.

Ihde, Don (2012): Experimental Phenomenology: Multistabilities, Albany.

James, Kathryn (2004): Mapping Scientific Frontiers: The Quest for Knowledge Visualization. In: Isis 95, 325.

Kang, Youn-ah / Stasko, John (2012): Examining the Use of a Visual Analytics System for Sensemaking Tasks: Case Studies with Domain Experts. In: IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology) 18, 2869–2878.

Kath, Roxana / Schaal, Gary S. / Dumm, Sebastian (2015): New Visual Hermeneutics. In: Sonderheft der Zeitschrift für Germanistische Linguistik „Automatisierte Textanalyse“ 43, 27–51.

Keim, Daniel A. (2002): Information Visualization and Visual Data Mining. In: IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology) 8, 1–8.

Keim, Daniel A. / Kohlhammer, Jörn / Ellis, Geoffrey / Mansmann, Florian (Hg.) (2010): Solving Problems with Visual Analytics, Goslar.

Kitchin, Rob (2014): Big Data, New Epistemologies and Paradigm Shifts. In: Big Data & Society 1, 1–12.

Kurt, Ronald (2008): Vom Sinn des Sehens: Phänomenologie und Hermeneutik als Methoden visueller Erkenntnis. In: Jochen Dreher / Michaela Padenhauer (Hg.), Phänomenologie und Soziologie, Wiesbaden, 369–378.

Lemke, Matthias / Niekler, Andreas / Schaal, Gary S. /Wiedemann, Gregor (2015): Content Analysis between Quality and Quantity. Fulfilling Blended-Reading Requirements for the Social Sciences with a Scalable Text Mining Infrastructure. In: Datenbank-Spektrum. Zeitschrift für Datenbanktechnologien und Information-Retrieval, Januar 2015, 7–14.

Lemke, Matthias / Schaal, Gary S. 2013: Paradigmenpluralität in der Politikwissenschaft. Eine Bestandsaufnahme des Faches in Deutschland. In: Gerhard Schurz / Stephan Kornmesser (Hg.), Die multiparadigmatische Struktur der Wissenschaften. Koexistenz, Komplementarität und (In)Kommensurabilität, Wiesbaden, 63–101.

Liu, Qing / Vorvoreanu, Mihaela / Madhavan, Krishna P. C. / McKenna, Anne F. (2013): Designing Discovery Experience for Big Data: A Case of Web-Based Knowledge Mining and Interactive Visualization Platform. In: Aaron Marcus (Hg.), Design, User Experience, and Usability. Web, Mobile, and Production Design, Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21–26, 2013, Proceedings, Part IV, Berlin / Heidelberg, 543–552.

Manovich, Lev (2016): The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. In: Mirko T. Schäfer / Karin van Es (Hg.), The Datafied Society: Social Research in the Age of Big Data, Amsterdam, i. E.

Marchese, Francis T. / Banissi, Ebad (Hg.) (2013): Knowledge Visualization Currents, New York / Heidelberg / London.

Nguyen, Quang Vinh / Qian, Yu / Huang, MaoLin / Zhang, JiaWan (2013): TabuVis: A Tool for Visual Analytics Multidimensional Datasets. In: Science China Information Sciences 56, 1–12.

Ramsay, Stephen / Rockwell, Geoffrey 2012, Developing Things: Notes Toward an Epistemology of Building in the Digital Humanities. In: Matthew K. Gold (Hg.), Debates in the Digital Humanities, Minneapolis, 75–84.

Raschke, Michael / Blascheck, Tanja / Burch, Michael (2014): Visual Analysis of Eye Tracking Data. In: Weidong Huang (Hg.), Handbook of Human Centric Visualization, New York / Heidelberg / London, 391–409.

Rensink, Ronald A. (2014): On the Prospects for a Science of Visualization. In: Weidong Huang (Hg.), Handbook of Human Centric Visualization, New York, 147–178.

Rieder, Bernhard / Röhle, Theo (2012): Digital Methods: Five Challenges. In: David M. Berry (Hg.), Understanding Digital Humanities, Basingstoke, 67–84.

Rosenberger, Robert (2009): Quick-freezing philosophy: An analysis of imaging technologies in neurobiology. In: Jan-Kyrre Berg Olsen / Evan Selinger / Søren Riis (Hg.), New waves in philosophy of technology, New York, 65–82.

Rosenberger, Robert (2011): A Case Study in the Applied Philosophy of Imaging: The Synaptic Vesicle Debate. In: Science, Technology & Human Values 36, 6–32.

Rosenberger, Robert / Verbeek, Peter-Paul (2015): A Field Guide to Postphenomenology. In: Dies. (Hgg.), Postphenomenological Investigations – Essays on Human-Technology Relations, Lanham u. a., 9–41.

Sacha, Dominik / Senaratne, Hansi / Kwon, Bum C. / Ellis, Geoffrey / Keim, Daniel A. (2016): The Role of Uncertainty, Trust, and Awareness, in Visual Analytics. In: IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology) 22, 240–249.

Sacha, Dominik / Senaratne, Hansi / Kwon, Bum C. / Keim, Daniel (2014): Uncertainty Resolution and Trust in Visual Analytics. Workshop-Poster IEEE VIS – Provenance for Sensemaking Workshop 2014.

Sacha, Dominik / Stoffel, Andreas / Stoffel, Florian / Kwon, Bum B. C. / Ellis, Geoffrey / Keim, Daniel A. (2014): Knowledge Generation Model for Visual Analytics. In: IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology) 20, 1604–1613.

Santucci, Giuseppe (2013): Visual Analytics and Information Retrieval. In: Agosti, Maristella / Ferro, Nicola / Forner, Pamela / Müller, Henning / Santucci, Giuseppe (Hg.), Information Retrieval Meets Information Visualization, Berlin / Heidelberg, 116–131.

Schaal, Gary S. / Ewert, Björn / Lancaster, Kelly / Stulpe, Alexander (2016): Die Herausforderungen der Digitalität für demokratische Staatlichkeit. In: Stefanie Hammer et al. (Hg.), Staat, Internet und digitale Gouvernementalität, Wiebaden, i. E.

Schaal, Gary S. / Kath, Roxana / Dumm, Sebastian (2016): New Visual Hermeneutics. In: Cybernetics & Human Knowing 23, 51–76.

Schaal, Gary S. / Kath, Roxana (2014): Zeit für einen Paradigmenwechsel in der Politischen Theorie? Der Ansatz der neuen visuellen Hermeneutik. In: André Brodocz / Daniel Schulz / Julia Schulze-Wessel (Hg.), Die Verfassung des Politischen, Wiesbaden, 331–350.

Schaal, Gary S. / Lancaster, Kelly / Dumm, Sebastian (2016): Politikwissenschaft und Big Data. Eine epistemologische Reflexion über Herausforderungen, Chancen und Risiken, In: Joachim Behnke / Andreas Blätte / Kai-Uwe Schnapp / Claudius Wagemann (Hg.), Big Data: Große Möglichkeiten oder große Probleme? Wiesbaden, i. E.

Scheinfeldt, Tom (2012): Sunset for Ideology, Sunrise for Methodology? In: Matthew K. Gold (Hg.), Debates in the Digital Humanities, Minneapolis, 124–126.

Scheuermann, Leif (2016): Die Abgrenzung der digitalen Geisteswissenschaften, in: Digital Classics Online 2, 58–67; http://journals.ub.uni-heidelberg.de/index.php/dco/article/view/22746, 01.12.2016.

Schmidt, Benjamin (2012): When You Have a Mallet, Everything Looks like a Nail. In: http://sappingattention.blogspot.de/2012/11/when-you-have-mallet-everything-looks.html, 14.05.2016.

Schreibman, Susan / Siemen, Ray / Unsworth, John (2004): The Digital Humanities and Humanities Computing: An Introduction. In: Dies. (Hg.), A Companion to Digital Humanities. Oxford, http://www.digitalhumanities.org/companion/, 14.05.2016.

Seifert, Christin / Vedran, Sabol / Kienreich, Wolfgang / Lex, Elisabeth / Granitzer, Michael (2014): Visual Analysis and Knowledge Discovey for Text. In: Aris Gkoulalas / Abderrahim Labbi (Hg.), Large-Scale Data Analytics, New York u. a., 189–218.

Takahashi, Jade (2012): Is Theory the Doing?, http://dh201.humanities.ucla.edu/2013/?p=398, 01.12.2016.

Tripathi, Arun K. (2004): Technologically Mediated Lifeworld. In: Ubiquity [http://ubiquity.acm.org/article.cfm?id=1670827] 12.04.2016.

Verbeek, Peter-Paul (2007): Beyond the human eye: Technological mediation and posthuman visions. In: Petran Kockelkoren (Hg.), Mediated vision, Rotterdam, 43–53.

Verbeek, Peter-Paul (2008): Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis. In: Human Studies 31, 11–26.

Autorenkontakt[63]

Prof. Dr. Gary S. Schaal
Helmut-Schmidt-Universität
Holstenhofweg 85
22043 Hamburg

Email: gschaal@hsu-hh.de

Kelly Lancaster M.A.
Helmut-Schmidt-Universität
Holstenhofweg 85
22043 Hamburg

Email: klancaster@hsu-hh.de

63 Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.

DIGITAL CLASSICS ONLINE

The TLRR II-Project – Providing a Digital Infrastructure to Research Roman Republican Trials

Kirsten Jahn

Abstract: The project *Trials in the Late Roman Republic II* (TLRR II)[1] aims at collecting, organizing, and analyzing information about Roman legal cases in an XML database. M. Alexander published the book "Trials in the Late Roman Republic, 149 BC to 50 BC" (TLRR I) in 1990, and initiated the current project that will make Roman republican trials easily accessible with modern technology. For each case a short description is provided, a clear distinction between assumptions and facts is made, and an updated bibliography can be found at the end of each entry. The open access database can serve both as a reference work and as a starting point for further research in Roman Republican history. It could be a connecting link within the developing digital infrastructure for that era.

1. Relevance of a Digital Companion for Roman Republican Trials

In his speech *Pro Milone* Cicero not only reminds the judges of the misdeeds of Clodius but also mentions several members of the court; among them the presiding judge, L. Domitius. But was Pompey, sole consul and author of the law establishing the court, really seeking justice, dignity, humanity and honesty by letting this man preside?[2] To answer this question one has to research who L. Domitius was. This is an easy task if a good historical commentary is at hand. Otherwise, the task will amount to finding the right L. Domitius in Pauly-Wissowa, among the 88 male members of the gens Domitia in an article stretching over 196 columns.[3] Of course, one could use a less comprehensive reference work but a lesser-known person will perhaps not be mentioned there, or the information available will focus on something other than his political or judicial accomplishments. In order to find a person in modern encyclopedias, usually the *nomen gentile* is needed, but neither Cicero nor other ancient sources constantly use this part of a Roman name. Hence, the identification of the person referred to will require special knowledge and time, while promising only meager results.

---

1 The current homepage is http://tlrr.blackmesatech.com/ with information about the project, a simple search interface and a provisional editing interface. I would like to thank C. M. Sperberg-McQueen for working on the TLRR II, answering my questions and improving the draft.

2 Cic. Mil. 22: *Quod vero te, L. Domiti, huic quaestioni praeesse maxime voluit, nihil quaesivit aliud nisi iustitiam, gravitatem, humanitatem, fidem*.

3 The Pauly-Wissowa is not completely accessible online, some volumes are available as pdf on https://archive.org/ and many entries are can be found through http://de.wikisource.org/wiki/RE.

For prosopographical questions, especially in the context of Roman legal history, we are still dependent on printed books and indices. This comes as a huge disadvantage because the participation in trials in one way or another was and is used as an important fact(or) establishing who belongs to which political faction, and thus represents a key tool for clarifying the connection between politics and the courts during the Republic.[4]
In speeches, historiographies, and biographies hundreds of people are mentioned. Only a few of them are important enough to make following their traces through ancient sources appear a worthwhile task on its own, but if all the small pieces of evidence for defendants, prosecutors, witnesses, judges are compiled, our picture gradually becomes complete and new questions can be answered quickly.
This is especially true for research interests involving more than just the biography of one person. Even a simple question like how many tribunes of the plebs were involved in cases *de repetundis* requires a lot of work. The most efficient approach today would be to use the index of M. Alexander's *Trials in the Late Roman Republic* in order to determine all cases *de repetundis* and to check the entry for each of those 63 cases to see whether a *tribunus plebis* is mentioned.[5] With the TLRR II database a search for 'de repetundis' and 'tribunus plebis' would result in a list of all the cases which fit those requirements within seconds. Thus, the scholar has more time to explore, e.g. in which roles and under what circumstances tribunes of the plebs acted in court.

2. Tradition and Innovation – TLRR I and II in Comparison

During the 19th century different collections of Roman trials were published; most of them were barely more than a re-narration of ancient sources.[6] A new era in the study of Roman criminal law and Roman republican history began in 1899 with Th. Mommsen's *Römisches Strafrecht*. He offered a systematic approach to the development and functionality of Roman Republican criminal law, and the influence of this book cannot be overstated.[7] On the other hand, the *Strafrecht* is not very helpful for determining which sources contain information about a specific trial. Of course, this was not Mommsen's aim, but the dominance of his work made all the previous books that might have listed all known references for a judicial proceeding hard to find.[8]
Only in 1968 with Gruen's book about *Roman Politics and the Criminal Courts, 149–78 B.C.* was a new work widely available where one could use the extensive footnotes to gather the sources relating to a trial. But the real turning point was 1990 when Alexander's *Trials of the Late Roman Republic, 149 BC to 50 BC* (TLRR I) was published. Inspired by the model of Broughton's *Magistrates of the Roman Republic*, Alexander developed a standardized format to record all the available information for each Roman trial in the covered period.

4 Cf. the classical work Gruen (1968). Brunt (1988) argued convincingly against the concept of *clientela* which was an essential basis for the prosopographical approach. Hence modern scholarship like Powell / Paterson (2007), 41 is more skeptical about the value of court appearances as evidence for political alliances.

5 Alexander (1990). If they exist, a thematic treatment of the desired topic can be consulted, in this case Thommen (1989), could be used alternatively.

6 Geib (1842); Rein (1844); Zumpt (1871), 468–558.

7 Mommsen (1899).

8 With reference to his age – he was 82 at the time as the *Römisches Strafrecht* came out – Mommsen declined to cite the scholarly literature of his time and thus making it even harder to track it down now (Mommsen, Th. (1899), XXIV).

The year 149 BC was chosen to begin with because according to Cicero the first permanent jury court was then established by L. Calpurnius Piso.[9] Since Caesar crossed the Rubicon in January 49 BC and started the civil war, the last orderly trials in the Roman Republic would have taken place in the year before.

Alexander decided to record eleven attributes for each trial. These are (1) date; (2) charge / claim; (3) defendant; (4) advocates / speakers of the defense; (5) prosecutors / plaintiffs; (6) presiding magistrate; (7) jurors; (8) witnesses; (9) parties to a civil suit; (10) verdict and (11) "miscellaneous information", i.e. facts concerning legal and formal aspects of the trial.[10] Following those categories, there is a list of all the ancient sources for the case, sometimes a few bibliographic references and very often footnotes explaining controversial details are added. As much information regarding the legal procedures is far from certain, doubtful entries are marked with a question mark while only in case of scholarly debate an endnote lists the most important opinions and works on the controversy.

Since June 1988 when the last revisions on TLRR I were made much has changed: important books and articles have been published, some omissions and mistakes have become known and, not least, the development of information technology and the internet has offered new opportunities for the analysis of the material.[11] Thus Alexander issued a call for a team to update the book and convert it into a state-of-the-art database.[12] As Alexander has agreed to the re-use of the existing data and since the existing division of information has proven very reasonable the basic structure of TLRR I will be kept. However TLRR II will not be only a slightly revised edition with an updated bibliography. It will offer (1) more internal differentiation, (2) more information about the cases and (3) it will make use of the modern technology.

First, while TLRR I listed 'charge / claim', i.e. usually the statute under which the trial was held, as one item, TLRR II provides the type of the court (ranging from *apud centumviros* to *quaestio extraordinaria*) and the procedure (ranging from *actio ad exhibendum* to *vis*) to the specific law (ranging from *lex Acilia de repetundis* to *lex Varia*). Such accuracy of discrimination provides the opportunity to show what is known for sure about the legal procedure of a case and what is just conjecture. For example, the ancient sources (including the legal ones) do not mention a *lex venefici* or a *lex de veneficis* and the earliest texts attesting a *lex Cornelia de sicariis et veneficis* belong to the 3rd century AD.[13] Nevertheless, TLRR I, in accordance with most of the scholarly literature, assigns 13 cases to such a law, among them three cases of attempted poisoning. The TLRR II database, in contrast, discloses whether we know if it was trial by jury, if the crime was attempted poisoning and if it is only assumed that the trial took place under a *lex Cornelia de sicariis et veneficis*. Thus TLRR II will give future scholars an opportunity to check which views the ancient authors held and which views were developed within the tradition of Roman legal scholarship.

---

9 Cic. Brut. 106.

10 Alexander (1990), IX–X . On http://indigo.uic.edu/handle/10027/99 a legal download of the book as PDF is possible.

11 E.g. David (1992); Brennan (2000).

12 After the project had begun, health issues compelled Alexander to step down as editor-chief of TLRR II, although he has remained available for consultation as needed. Dr. T. Deline (Grant MacEwan University, Edmonton, AB) has taken over the position as editor-in-chief. Dr. V. Arena (University College London); Dr. A. Borgna (Università di Torino); Dr. A. Raggi (Università di Pisa); Dr. F. Russo (Universität Konstanz/ München) and the author of this article make up the TLRR II team working on structure and content while Dr. C. M. Sperberg-McQueen (Black Mesa Technologies LLC) develops the technical infrastructure.

13 One doubtful source is Liv. per. 8: *Lex de veneficio tunc primum constituta est,* because the same story is told in Liv. 8,18 without any mention of a law (*neque de veneficiis ante eam diem Romae quaesitum est*). For a *lex Cornelia de sicariis et veneficis* cf. D. 48,8,1pr., 3, 5 (Marcian) and Coll. 1,2,1 (Paulus).

Second, TLRR II provides additional information for every trial. The most important change is the following one: in order to get a quick and general idea about the facts of a case there will be a short description of the trial. In TLRR I, the information for case #161 is:

date:	between 74 and 70
charge:	lex Cornelia de ambitu
defendant:	Ti. Gutta (1) sen.
prosecutors:	people condemned for electoral bribery (ambitus condemnati)
outcome:	C(ondemned)
sources:	Cic. Clu. 98, 103, 127; Quint. Inst. 5.10.108; [Asc.] 216St

In TLRR II, this description of the case will follow:
"Gutta was accused of attempting to bribe the jury in the case of Oppianicus (#149); his fellow conspirators were C. Aelius Paetus Staienus and M. Atilius Bulbus."
Furthermore additional information will be given about unique details of the case, e.g. name of the province involved for cases *de repetundis* or reported omen. The bibliography will be more closely connected to each single case in form of bibliographic entry accompanying each case record.[14]
Third, the advantages of a database for searching and counting sets of data are obvious. The same goes for the possibility to constantly change and update the data entered. But the TLRR database II will also offer some smart ways to move within the datasets, e.g. there will be links to follow related (or most likely related) cases through the database. For trial #161 there will be a link to trial #149, in which the bribery allegedly had taken place, and to the cases against the other corrupt judges. Another feature will show the names of individuals that belong to one side of the trial (accusation, defense or staff) next to each other and in similar colours.

14 TLRR I closes with an extensive bibliography and only some works cited and often quoted to illustrate the controversies appear in the footnotes for a specific case.

3. Technical Choices and Current Status

The manuscript of TLRR I used Waterloos Script. While the first task was to decide which modern format to use, the second one was to transfer the data. The technical advisor to the project, Dr. C. M. Sperberg-McQueen (co-editor of the Extensible Markup Language [XML] and chair of the W3C XML Schema working group), believes that XML has better capabilities to manage uncertain, semi-structured, and fragmentary data, like the data for the Roman trial, than a relational databank management system like MySQL.[15] Using XSLT he transferred the data into XML.[16] Now each individual, law, and trial is represented by a different XML document.

```
- <trial id="ZMM" tlrrl="309" sortdate="">
  - <date xml:space="preserve">
      52, Milo charged on March 26, trial on April 4-7/[8])
    - <en>
      - <p xml:space="preserve">
          On the chronology of this trial and related trials, see Ruebel (1979) 245-47.
        </p>
      </en>
    </date>
  - <ccGrp>
    - <charge>
        <procedure pid="c-lex_Pompeia_de_vi">lex Pompeia de vi</procedure>
      - <p xml:space="preserve">
          (murder of Clodius)
        - <en>
          - <p xml:space="preserve">
              On the meeting of Clodius and Milo, see Davies (1969);
              <i>contra</i>
              Wellesley (1971).
            </p>
          </en>
        </p>
      </charge>
    </ccGrp>
  - <defGrp>
    - <defendant>
      - <namelist>
        - <person-entry xml:space="preserve">
            <person pid="pAnnius67T.Milo" ix="2" form="Annius (+67), T. Milo">T. Annius Milo (67)</person>
            pr. 55
          </person-entry>
        </namelist>
      </defendant>
    - <advocate label="advocates">
      - <namelist>
        - <person-entry xml:space="preserve">
            <person pid="pClaudius229M.Marcellus" ix="3" form="Claudius (229), M. Marcellus">M. Claudius Marcellus (229)</person>
            cos. 51
          </person-entry>
        - <person-entry xml:space="preserve">
            <person pid="pTullius29M.Cicero" ix="3" form="Tullius (+29), M. Cicero">M. Tullius Cicero (29)</person>
```

Fig. 1: TLRR 2: XML for the trial Pro Milone in 52 (detail)

A first query interface, named ‘Balbus’, relying on the data of TLRR I is available since January 2016 on the homepage, and any user wanting to know more about L. Domitius would get this result:

15 Sperberg-McQueen (2016).

16 All programs are available at http://tlrr.blackmesatech.com/lib/

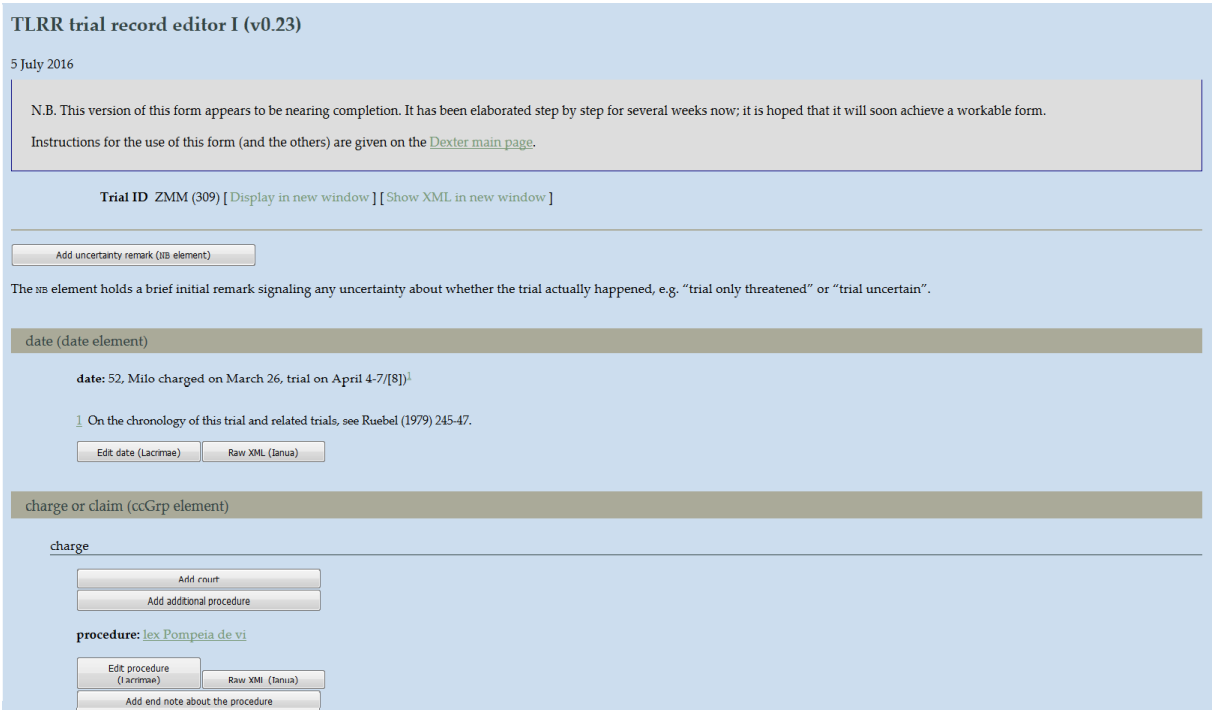


Fig. 2: TLRR 2 search interface 'Balbus' (detail)

Instead of having to look through 88 Domitii in the Pauly-Wissowa a result of 13 trials is given. Because 'Balbus' is designed to prefer high recall, any trial involving a Domitius is given as a result. But only two of them are named Lucius: one is a witness in the trial against Verres and one is the presiding judge named in *Pro Milone*. The user now has his full name: L. Domitius Ahenobarbus; his number within the Domitii in the Pauly-Wissowa: 27; the knowledge that Domitius was consul in 54 BC and that this case is his only legal involvement we know of (or one might speculate that this Domitius and the witness in the Verres case are identical). For people familiar with XPath a further query tool is available on the project website.

An additional query interface is in development and will support Boolean operators making it possible to search for all cases in which Cicero AND Milo were involved or all cases in which Cicero appeared AND NOT acted as an advocate. The different search interfaces will be available on the website all the time catering for different needs of different users.

Another challenge of the Roman trials data is the fact that, frequently, we do not have a precise date of a trial. Sometimes the precise year is known, sometimes we only know that one trial took place prior to another, or we have to deal with a larger or smaller time span. In some cases trials are dated differently by different scholars and a few trials have no reliable dating at all. Such material makes a simple question like "How many ambitus trials took place between 80 and 50 BC?" hard to answer. In order to solve this problem the TLRR II search assesses the quality of the match, i.e. 75 BC is within the desired period and hence a 100% match, a case dated in the period between 90 and 70 BC is not a 100% match. A scale to allocate equal probability to each year has to take into account the years 509 to 27 BC. This is necessary because for a few trials it can only be said that they took place within Roman Republican times (i.e., between 509 and 27 BC). If it is highly probable that a case can be assigned to the desired timespan it will show up equally high in the search results.

Still in testing phase is the editing interface. Because the project members are geographically dispersed, it is of great importance that the interface fosters a consistent entry of data. In the current state the record editor data can be changed or added via two different interfaces: one, 'Lacrimae', that provided a more structured display of the XML which makes it easier to use and one called 'Ianua', giving the XML.

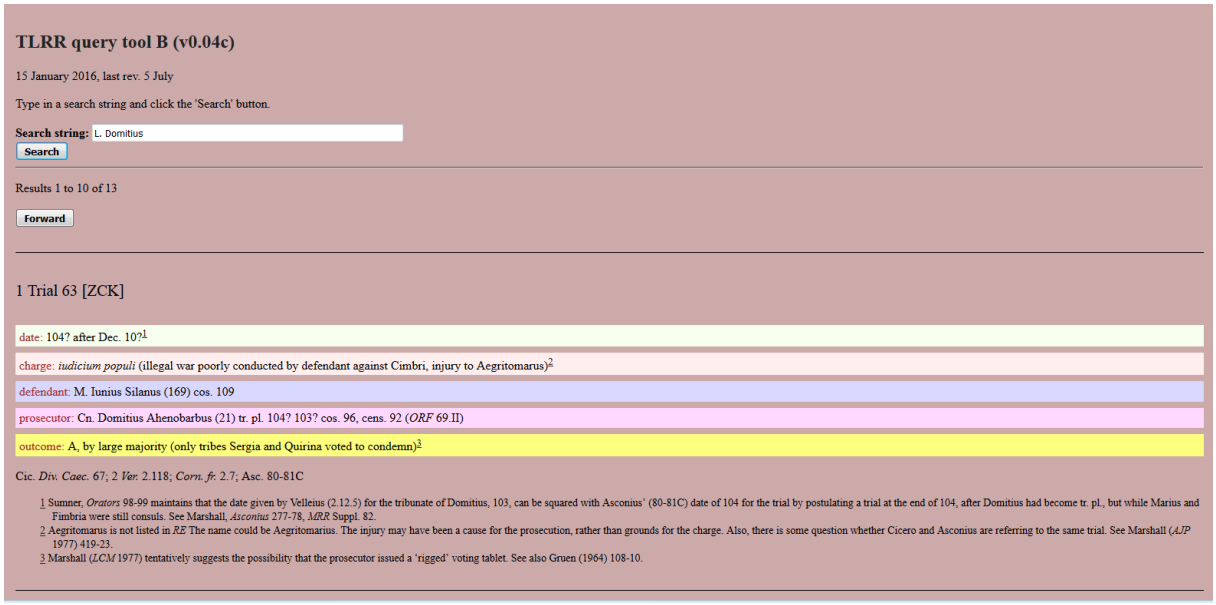


Fig. 3: TLRR 2: record editor 'Dexter'

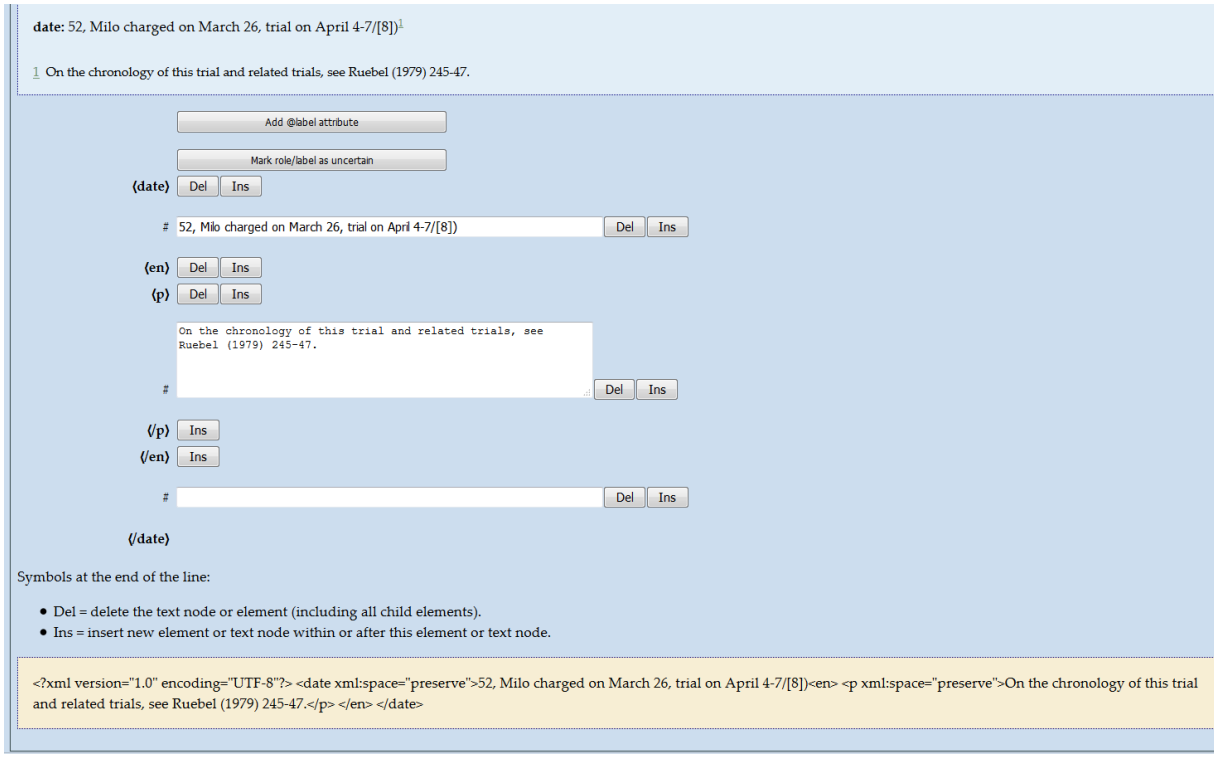


Fig. 4: TLRR 2: editing interface 'Lacrimae'

4. Projects and Possibilities – The TLRR II within the New Digital Landscape on Roman Republican History

In recent years several web-based projects have started to make the period of the Roman Republic more accessible. The ANHIMA Research Center worked on *Leges Populi Romani* (LEPOR) till fall 2014.[17] The database details the known information for every comitial law of the Roman people, including a commentary and bibliographic references.

At King's College London a team headed by H. Mouritsen is creating an open-access searchable digital database for all known members of the republican elite, updating and expanding the data of Broughton's *Magistrates of the Roman Republic*.[18] The *Digitizing the Prosopography of the Roman Republic* (DPRR) project works on the same set of individuals as TLRR II and it would be very useful to interlink both databases. As the DPRR plans to give URIs for each individual this should be a manageable task.

The project *Fragments of the Republican Roman Orators* (FRRO) with C. Steel as principal investigator at the University of Glasgow is less concerned with prosopography and all the more with texts.[19] The project aims to collect, edit, annotate and translate all evidence for public speech during the Roman Republic by all men other than Cicero. The project is using Malcovati's *Oratorum Romanorum fragmenta* and adding information about public speeches which are not transmitted as direct quotation. While some of these texts are merely public speeches by aspiring politicians and magistrates trying to foster their careers, many were delivered in the courts of law. For the moment TLRR 2 follows TLRR I in providing the ORF number of Malcovati and this may be one way to connect both projects in the future.

The ancient sources are essential for the TLRR 2, and with the Canonical Text Services (CTS) it will be possible to use concise citations in URN form, i.e. refer directly to *Pro Milone* paragraph 22 where L. Domitius is mentioned.[20] Ideally, all the sources mentioned will be connected via CTS with a standard edition of the ancient text. Thus, the reader will only need to simply click on the citation to see what, e.g. Valerius Maximus wrote and in which context he referred to the case. At the moment, the TLRR II team is trying to figure out whether this very useful feature can be added to our database without delaying the completion of TLRR 2 too far into the future.

17 Website: http://www.cn-telma.fr/lepor/introduction/

18 A public website still does not exist but http://gtr.rcuk.ac.uk/projects?ref=AH/K007211/1 and Robb (2012) give insights about the DPRR. Broughton (1951/52); Broughton (1986).

19 Website: http://www.frro.gla.ac.uk/

20 Blackwell / Schubert in DCO 2,1 (2016) and Schubert's CTS project „Erstellung und Implementierung eines Referenzierungstools zum Zitieren antiker Quellen aus Online-Datenbanken", funded by the German Research Foundation.

5. Bibliography

Alexander (1990): M. Alexander, Trials in the Late Roman Republic, 149 BC to 50 BC (Phoenix Suppl. 26), Toronto 1990.

Blackwell / Schubert (2016): C. B. Blackwell / C. Schubert, Annotating and Editing with Canonical Text Services (CTS) Project funded by the Andrew W. Mellon Foundation: 2016–2017, in: DCO 2,1 (2016), 94–99; DOI: http://dx.doi.org/10.11588/dco.2016.1.28180.

Brennan (2000): T. C. Brennan, The Praetorship in the Roman Republic, 2 vol., Oxford 2000.
Broughton (1951/52): T.R.S. Broughton, The Magistrates of the Roman Republic, 2 vol, New York 1951/52.

Broughton (1986): T.R.S. Broughton, The Magistrates of the Roman Republic, Bd. 3 suppl., New York 1986.

Brunt (1988): P.A. Brunt, Clientela, in: P.A. Brunt: The Fall of the Roman Republic and Related Essays, Oxford 382–442.

David (1992) : J.-M. David, Le patronat judiciaire au dernier siècle de la République romaine (Bibliothèque des Écoles françaises d'Athènes et de Rome 267), Rom 1992.

Geib (1842): K. G. Geib, Geschichte des römischen Kriminalprozesses bis zum Tode Justinians, Leipzig 1842.

Gruen (1968): E. S. Gruen, Roman Politics and the Criminal Courts, 149–78 B.C., Cambridge 1968.

Leges Populi Romani (LEPOR): http://www.cn-telma.fr/lepor/introduction/ Powell / Paterson (2007): J. Powell, / J. Paterson, Introduction, in: J. Powell, / J. Paterson (ed.), Cicero the Advocate, Oxford 2007, 1–57.

Mommsen (1899): Th. Mommsen, Römisches Strafrecht, Leipzig.

Rein (1844): W. Rein, Das Kriminalrecht der Römer von Romulus bis auf Justinian, Leipzig 1844.

Robb (2012): M. Robb, Digitising the Prosopography of the Roman Republic, presented at Institute of Classical Studies Digital Seminar 2012: http://www.digitalclassicist.org/wip/wip2012-07mr.html.

Sperberg-McQueen (2016): C. M. Sperberg-McQueen, Trials of the Late Roman Republic: Providing XML infrastructure on a shoe-string for a distributed academic project, in: Balisage Series on Markup Technologies 17 (2016); doi:10.4242/BalisageVol17.Sperberg-McQueen01.

Thommen (1989): L. Thommen, Das Volkstribunat der späten römischen Republik, Stuttgart 1989.

Zumpt (1871): A. W. Zumpt, Der Criminalprozeß der römischen Republik, Leipzig 1871.

Autorenkontakt[21]

Kirsten Jahn
Otto-von-Guericke-Universität Magdeburg
Fakultät für Humanwissenschaften - Institut II: Bereich Geschichte
Zschokkestraße 32
39104 Magdeburg

Email: kirsten.jahn@ovgu.de

21 Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.

DIGITAL CLASSICS ONLINE

The Digital Hill Project
Sources on the Revolt of Samos

Marcel Mernitz

Abstract: This article covers the work done for the Digital Hill project of the Alexander von Humboldt Chair of Digital Humanities at Leipzig University. After a short introduction about the book on which the project is based and the arrangement of the chosen chapter in this book the goals of the project are presented, which are the creation of EpiDoc TEI compatible XML files for the sources, the production of treebank annotations and text alignments and the provision of the results on a web page. The paragraphs concerning treebank annotations and text alignments present working with the interfaces of Arethusa and Alpheios in the Perseids platform. Users can interact with the web page. For that reason jQuery scripts have been written, whose functionality is explained in the visualization paragraph. Some issues on the creation of the EpiDoc files are presented there as well as the applied solutions.

1. Introduction

"Digital Hill" is a project of the Alexander von Humboldt Chair of Digital Humanities at Leipzig University engaged in the production of a digital edition of the *Sources for Greek History between the Persian and Peloponnesian Wars* edited by G. F. Hill in 1897.[1] This volume is a collection of sources encompassing the fifty years of Greek history (Pentekontaetia) between the end of the Persian Wars and the beginning of the Peloponnesian War (479–431 BC). A revised edition of Hill's book was published by R. Meiggs and A. Andrewes in 1951.[2] We decided to work on the original version of Hill not only because it is out of copyright, but also because it still represents a fundamental work for establishing a new digital comprehensive guide to the Pentekontaetia and the Peloponnesian War.[3] Another point for working on this edition is that we are interested in collections of heterogeneous sources, and not only in isolated authors.

Within the topics addressed by Hill in his edition, we chose to work with the sources on the Athenian suppression of the revolt of Samos (441–439 BC) since they are a good test case for

1 The repository of the project is on GitHub at http://digitalhill.github.io. The printed edition of the book is freely available at https://archive.org/details/sourcesforgreekh00hilluoft.

2 Meiggs/Andrewes (1951).

3 On the fragmentary state of most of the sources concerning the Pentekontaetia, see Martin/Berti (forthcoming).

showing the tools that we have been using and the methodology that we have been devising for establishing a possible model for producing a digital version of the whole collection.[4]

2. The Sources on the Revolt of Samos (441-439 BC)

The sources on the Pentekontaetia collected by G. F. Hill are arranged by topic in eight chapters starting with the origin and organization of the Athenian confederacy and ending with the Western Greeks.[5] The sources on the revolt of Samos are printed in chapter 3 – which is about the external history of Athens, her allies, and colonies – and include both literary and epigraphic texts.[6]

The project is focussed on three main goals: 1) to produce XML files of the sources on the revolt of Samos following the EpiDoc TEI XML subset;[7] 2) to produce linguistic annotations of the literary sources on the revolt of Samos according to the Ancient Greek and Latin Dependency Treebank 2.0 guidelines;[8] 3) to produce translation alignments of the literary sources on the revolt of Samos using the Alpheios alignment editor.[9]

In order to produce these annotations, the first part of the work is devoted to listing the sources on the revolt of Samos collected by Hill and to checking which were already available in an XML format in the Perseus Digital Library.[10] The sources are constituted by Greek and Latin literary texts and inscriptions, and they have been arranged into a spreadsheet.[11] The spreadsheet includes different pieces of information: 1) editions used by Hill (when this is referred to by the editor);[12] 2) links to the XML files in the Perseus Digital Library or in other available repositories; 3) links to treebank and text alignment files that have been created as part of the

---

4 For a study of the fragmentary sources on the revolt of Samos with bibliography, see Berti (2013), 269–288. For a synoptical representation of the primary sources on the revolt of Samos, see http://demo.fragmentarytexts.org/en/revolt-of-samos.html. For further information about the revolt and its chronology see Pritchard (2012), 39, Phillips (2010), 2 and Fornara/Lewis (1979), 7–10. For the dating see also the text-alignment of Thuc I. 115, 2 and Schol. in Arist. Vesp. 283 (http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=15868), and the treebank files of Plut. Per. 28 (http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=10&doc=11142) and Schol. in Arist. Vesp. 283 (http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=2&doc=10318 and http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=8&doc=10318).

5 The revised edition by Meiggs and Andrewes has a different internal structure, because sources are printed in alphabetical order, but arranged by topic in rich and detailed indices at the end of the book.

6 Hill (1897), 137–146.

7 http://sourceforge.net/p/epidoc/wiki/Home/

8 https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md

9 http://alpheios.net/

10 The sources not available in a digital format have been digitized and manually annotated.

11 We chose a Google Drive spreadsheet for no particular reason. The same results could have been achieved with Open Source software like Davros for the storage and EtherCalc for the spreadsheet. This spreadsheet is accessible at https://docs.google.com/spreadsheets/d/1dDuAS9vXrrvMczAIja8oUhmPhU-nHQ7lPsTT3wDsVdg/edit

12 Hill, for example, in his collection doesn't print the text of Thucydides, Xenophon, and the Aristotelian *Athenaion Politeia* for reason of space and "because they can best be supplied from the shelves of those who are likely to consult this work" (Hill (1897), vi).

project (see below);[13] 4) portions of the source texts left out by Hill;[14] 5) links to the EpiDoc files that were manually produced as part of the project; 6) additional notes and a legend explaining the meaning of the coloured cells.

3. Linguistic Annotations of the Sources on the Revolt of Samos

One of the main goals of the project was the production of morphosyntactic annotations of the sources on the revolt of Samos. In order to produce these annotations, we followed the Ancient Greek and Latin Dependency Treebank 2.0 guidelines[15] and the Arethusa interface openly available through Perseids, which is a collaborative platform for editing and annotating ancient source documents.[16]

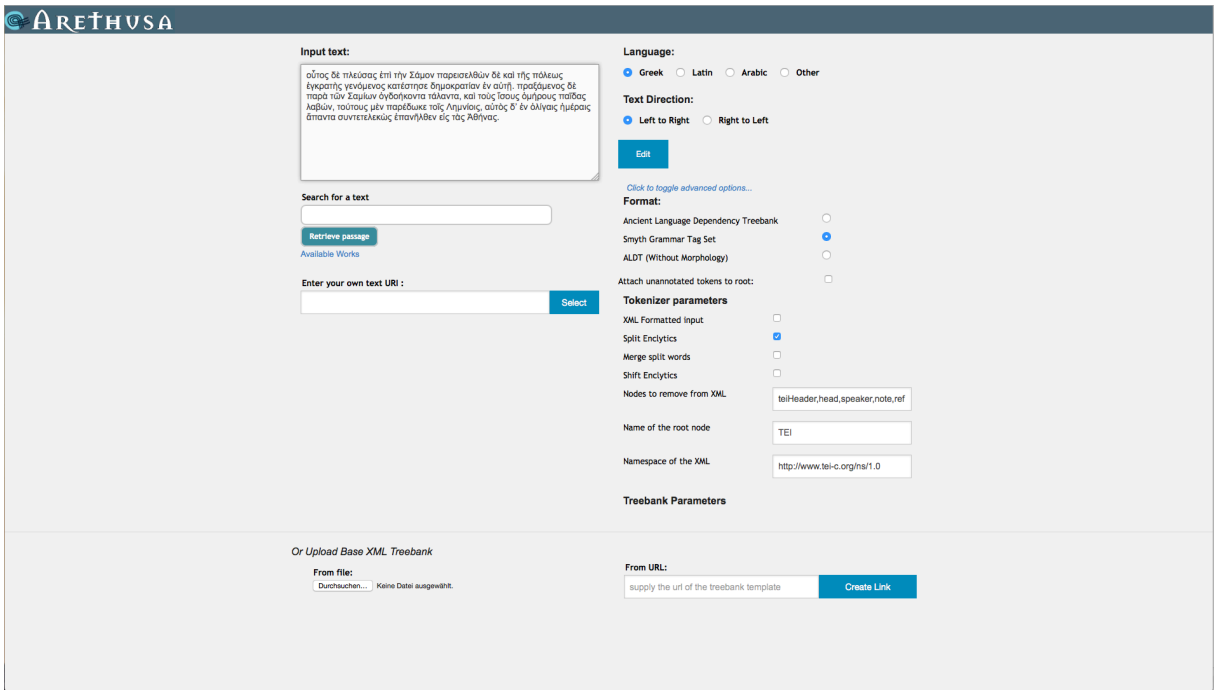


Fig. 1: Screenshot from the Arethusa treebank file creation mask (Diod. XII.27.2.1).

Fig. 1 shows an example of a treebank file of Diod. XII.27.2.1 using the Arethusa interface. The language has been automatically set up to Greek and the 'Smyth Greek Grammar Tag Set' provides morphological, syntactic, and semantic annotations. After setting up these options, the 'Edit' button allows to create the treebank file shown in fig. 2.

13 The column of the spreadsheet containing links to the translation alignment files is split into two separate columns: one contains the actual link to the file, the other one contains information about the type of alignment (for example, if it is a partial or a full alignment and, in the case of a full alignment, which translation has been used).

14 When collecting sources, Hill prints only the passage of text which is relevant to the event he is dealing with, and sometimes he leaves out parts of the text. This left out text was added in a specific column of the spreadsheet.

15 See note 9 for the guidelines. The guidelines are based on Herbert Smyth's Greek grammar (http://www.perseus.tufts.edu/hopper/text?doc=Smyth+grammar+1&fromdoc=Perseus%3Atext%3A1999.04.0007).

16 http://perseids.org

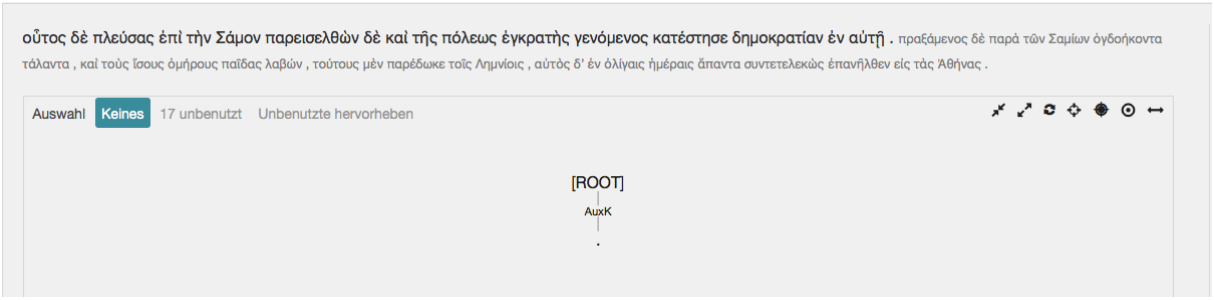


Fig. 2: Detailed screenshot from Arethusa treebank file edit mask (Diod. XII.27.2.1).

When the tool has finished processing, the edit mask automatically opens and the passage is ready to be treebanked. Using the mouse and the drag function, it is possible to toggle a word and align it to other words depending on the ROOT-node. On the top right side of the interface there are several buttons which provide services like saving, downloading the XML file, several other options, or switching the language (they are not shown in the screenshots). Below those buttons there is a menu-bar that provides several tabs necessary for the annotation of the words (fig. 3).

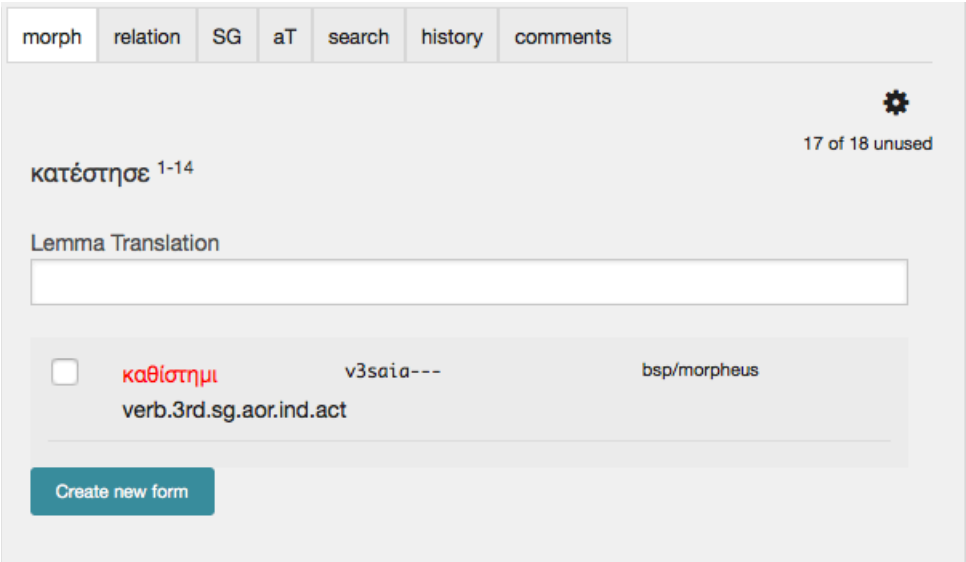


Fig. 3: Detail of the Arethusa menu-bar with a toggled word.

When a word is toggled, it is possible to annotate its morphological layer in the 'morph'-tab, the syntactic layer in the 'relation'-tab, and the semantic layer in the 'SG'-tab. The 'aT'-tab enables users to add elliptical nodes that help to annotate according to the guidelines, for example in sentences where the main verb is missing. In the interface, words are coloured depending on their morphological function.[17] Once the annotation is done, this feature allows to visualize the morphological layer very clearly (see fig. 5).

17 For example, verbs are displayed in red, nouns in green, pronouns in violet, adverbs in orange, etc.

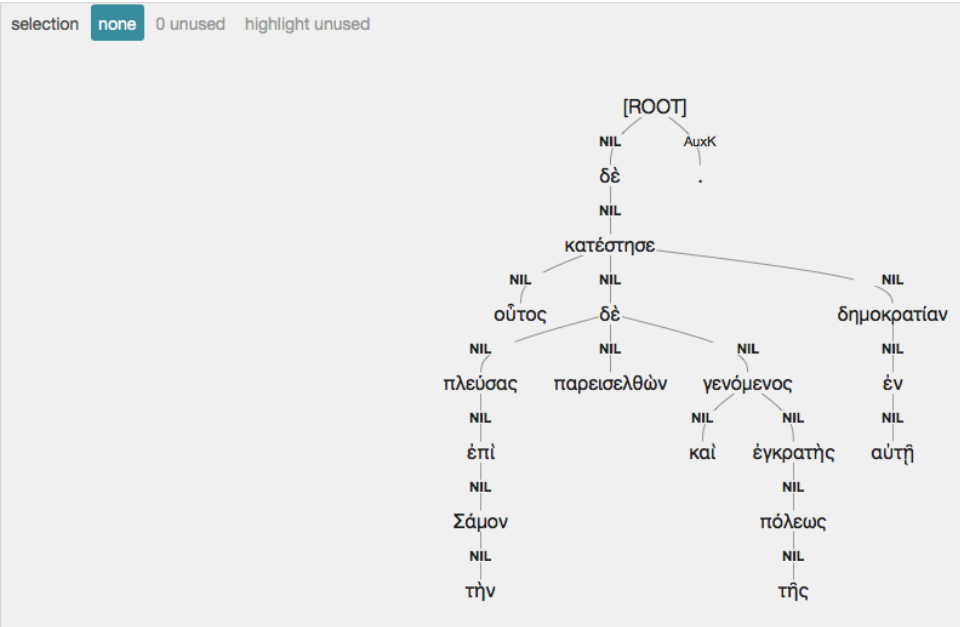


Fig. 4: A sentence before the annotation of the morphological layer (see fig. 5 for the annotation).

In order to add the morphological annotation, a word has to be selected by clicking on it. Then either one of the proposed words has to be chosen or a new word has to be added with the 'create new form' function. A new form needs additional information depending on its part of speech. Nouns need different types of information than verbs or numerals. The interface provides drop-down-menus for all the required pieces of information. When all the information is gathered, the new form is added by using the 'Save'-button. The selection of a word is undone using the 'esc'-key or by clicking on the word it depends on.[18] It is necessary to undo the selection of a word in order to continue the annotation, otherwise the former selected word would become annotated to the next selected word, and the sentence tree would become messed up. Fortunately, it is pretty easy to correct this mistake should it occur by either selecting again the correct node or using the 'undo'-button. It is possible to add to the morphological layer a lemma translation in a letterbox (see fig. 3). The guidelines deal with the way this translation should be done.

18 In the example, the word Σάμον would not be selected anymore when the user clicks on ἐπὶ.

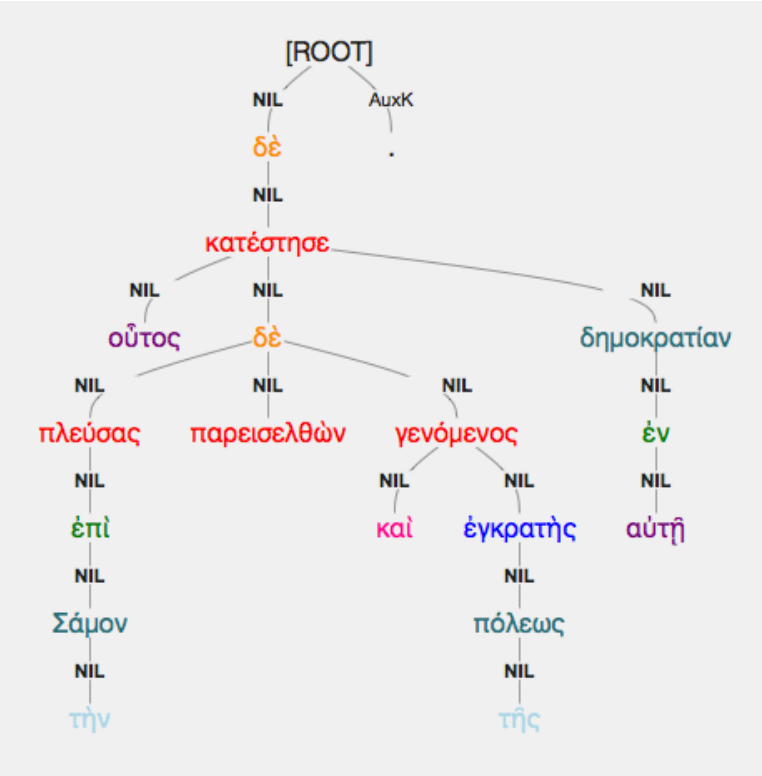


Fig. 5: The same sentence of fig. 4 after annotation of the morphological layer.

Once the morphological layer is annotated, the syntactic layer may be added. To add this layer, a word has to be selected and the tab 'relation' has to be chosen. A drop-down menu presents various choices for this layer, and the word may be annotated according to the guidelines. This layer may be annotated without another layer previously annotated.[19]

19 The semantic layer, on the other hand, may only be annotated once the morphological layer has been finished since it depends on the former.

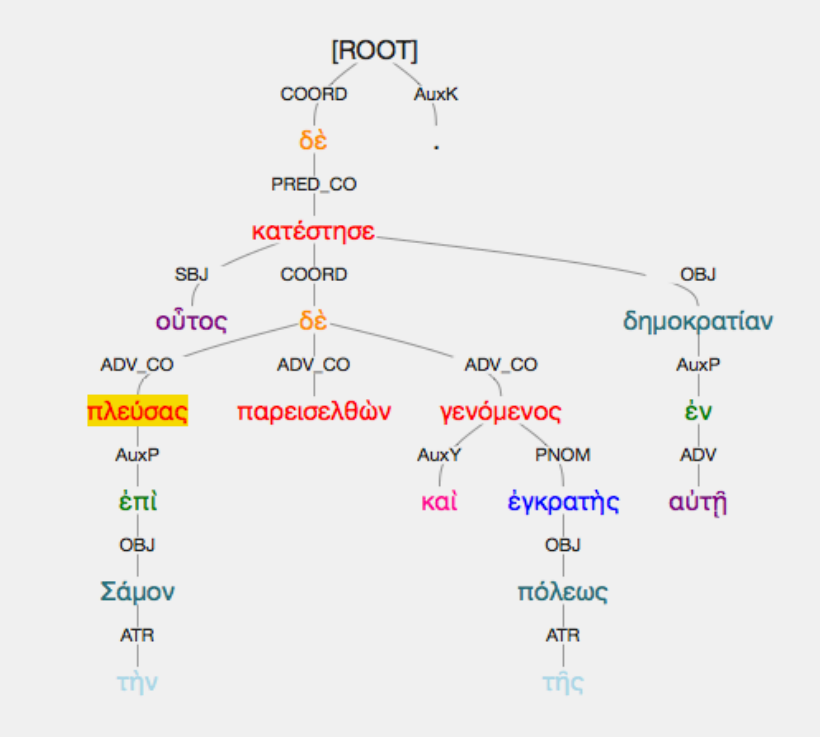


Fig. 6: The same sentence of fig. 5 after annotation of the semantic layer.

The next step is the annotation of the semantic layer. To annotate this layer, a word has to be toggled again and the tab 'SG' has to be chosen. Depending on the choice, the morphological layer allows different options. The favoured option may be chosen using the drop-down menu. Examples for those drop-down menus are given in fig. 7–8. The numbers next to the words refer to the sentence of the passage and the number of the word in it. For example, 1–14 stands for sentence 1 and word 14, 1–15 stands for sentence 1 and word 15, and so on.

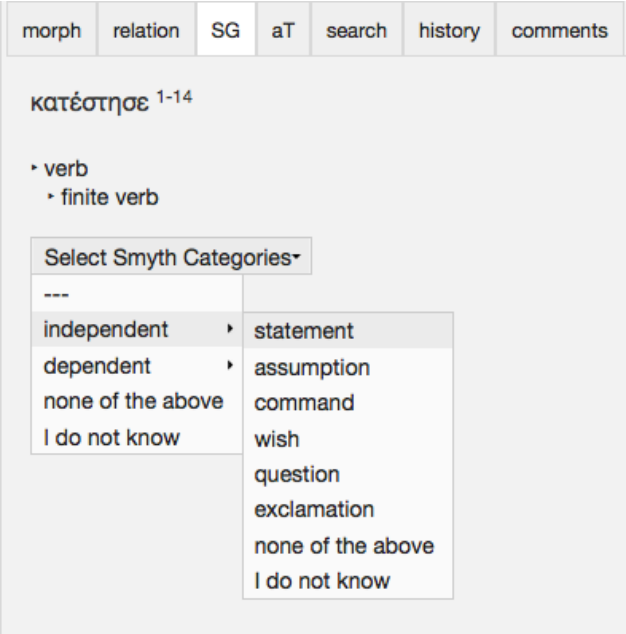


Fig. 7: Drop-down menu for the word κατέστησε treated as a verb in the morphological layer.

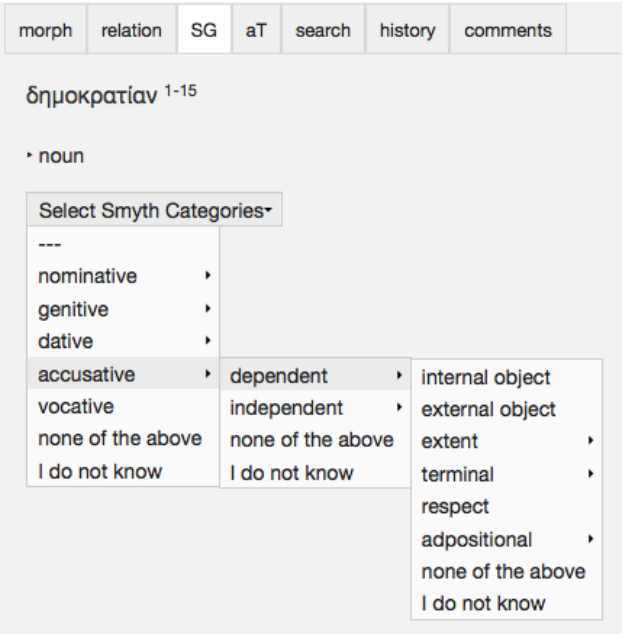


Fig. 8: Drop-down menu for the word δημοκρατίαν treated as a noun in the morphological layer.

Morphosyntactic and semantic annotations are important and useful for different linguistic analyses and interpretations. We will show the example of the sentence displayed in the next

figures, taken from Diod. XII.27.2.1: οὗτος (sc. Pericles) δὲ πλεύσας ἐπὶ τὴν Σάμον[20] παρεισελθὼν δὲ καὶ τῆς πόλεως ἐγκρατὴς γενόμενος κατέστησε δημοκρατίαν ἐν αὐτῇ.
According to the syntactic layer, the second particle δέ and the conjunction καὶ coordinate the second part of the sentence with the first one. The main verb and predicate of the sentence is κατέστησε, which is labelled as PRED_CO. The dependent subject is οὗτος and is labelled as SBJ. The adverbial phrases πλεύσας, παρεισελθών and γενόμενος depend on the predicate and are coordinated by the conjunction καὶ, which is thus labelled as COORD, the particle δέ is labelled as AuxY, and have their own dependencies in ἐπὶ τὴν Σάμον and τῆς πόλεως ἐγκρατής respectively. The object of the main sentence is δημοκρατίαν, which is complemented by the adverbial phrase ἐν αὐτῇ and labelled as OBJ. The subject οὗτος refers to a previous sentence and stands for Pericles. All the particles may be treated as temporal particles and present a sequence; they are therefore labelled as ADV or ADV_CO. Appositions are always labelled as AuxP, and articles as ATR. The words following the appositions are treated as adverbial phrases of places and thus labelled as ADV. Only the full stop is automatically labelled as AuxK as this is the common label for final punctuation. Since ἐγκρατής is dependent on the copulative verb γενόμενος, it is labelled as PNOM with πόλεως as its argument, which was thus labelled as OBJ.
The number of choices for dependent nominatives is quite limited – actually there is only one possibility – and thus, it is only annotated as a dependent nominative. On the other hand, the number of choices for verbs is quite numerous. The only finite verb in the sentence of Diodorus is the predicate, thus, it has to be independent. Given that the sentence places a statement, the predicate is treated as such and its object is annotated as an external object. Both δέ are annotated as particles. As already mentioned, the participles are annotated as temporal sequences and the words following an apposition are annotated as a terminal accusative and a dative of place. According to the guidelines, predicate nominals are annotated as dependent nominatives and ἐγκρατής is accordingly annotated.
There is no semantic annotation for conjunctions, appositions, and articles. With all these pieces of information a translation of the sentence might be as follows: And after sailing to Samos, after reaching and after mastering the city, he (*sc*. Pericles) established democracy in it.

In order to get consistent work when treebanking, sometimes it has been necessary to add technical nodes, which are called ‘elliptical nodes’, that would act as predicate forms (PRED), as the sentences do not contain a finite verb that would serve this function.[21] Yet the other parts of the sentence are dependent on this predicate form – except for the coordinating conjunctions or particles, in the most cases δέ. It is possible to download the XML-file that is the foundation of each treebank file. If the file has been created by retrieving the text, it contains a CTS-URN, should the source contain such an URN.[22]

20 The text provided by Hill shows two asterisks here to indicate a possible lacuna after Σάμον. Unfortunately there is no way to deal with lacunas in Arethusa, yet. For that reason, the text has been treated as it is, without that lacuna, to present one way, how issues like that may be treated.

21 See http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=9&doc=11140 for an example. Those nodes are recognisable due to the fact that they are displayed smaller than the other words of the sentence.

22 Since the treebank files for this project were produced manually and not by retrieving text automatically, they do not contain a CTS-URN. It would not have been possible to retrieve the text for all the sources anyway, since Hill does not always quote the edition that he used. In the future, it may be considered to use texts containing a CTS-URN, for example, by using the capiTains API, available at http://cts.perseids.org/.

3.1 Translation Alignment of the Sources on the Revolt of Samos

The interface used for creating translation and textual alignments is called Alpheios and it is part of the Perseids annotation environment.[23] It allows to align two texts in two different languages or in the same language as well. Fig. 9 shows the mask for aligning the Greek text of the sentence of Diod. XII.27.2.1 (see previous paragraph) with its English translation.

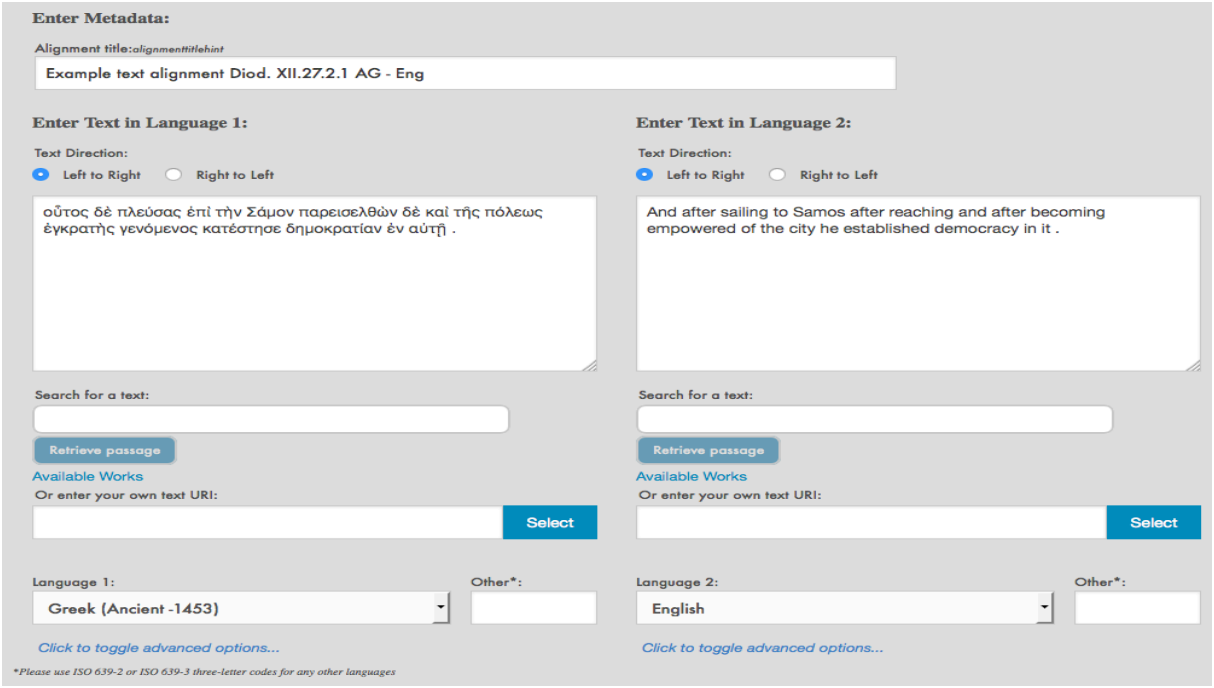


Fig. 9: Creation mask for text-alignments in Perseids (Diod. XII.27.2.1).

After pasting, retrieving or putting the relevant URI of the two texts that have to be aligned in the two boxes of the interface, it is possible to select the respective languages and start the alignment by pressing the 'Align'-button on the right side (see fig. 10).

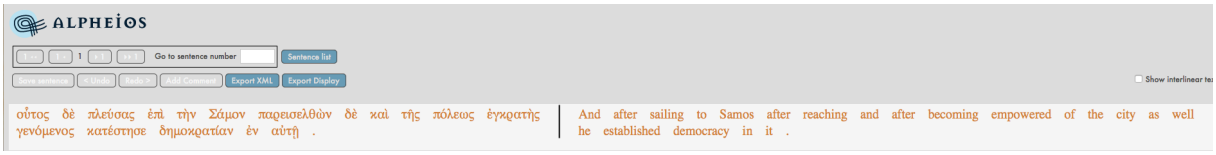


Fig. 10: The alignment interface for Greek and English (Diod. XII.27.2.1).

Not aligned words are displayed in orange, aligned words in black. Selected words have a purple box around them. Words are aligned by clicking on them and then clicking on the corresponding word in the other text. It is possible to align more than one word to one single word and vice versa. One word is aligned to several other words when it is aligned to a word that is already aligned to those words. In this example οὗτος is aligned to 'he', δέ is aligned to 'and', πλεύσας to 'after sailing', ἐπί to 'to', Σάμον to 'Samos', παρεισελθών to 'after reaching', the καὶ to 'and', τῆς to 'the', πόλεως to 'city', ἐγκρατής and γενόμενος to 'after mastering', κατέστησε to 'established', δημοκρατίαν to 'democracy', ἐν to 'in' and αὐτῇ to 'it'. The article before Σάμον (which is τήν) and the second δὲ could not be aligned. One word, πόλεως, is

23 http://perseids.org/. For a description of Alpheios see also http://alpheios.net/.

aligned to two words, since its case demands that apposition in English. For translation issues, ἐγκρατής and γενόμενος were aligned to the same two words and not to one single word each at a time.

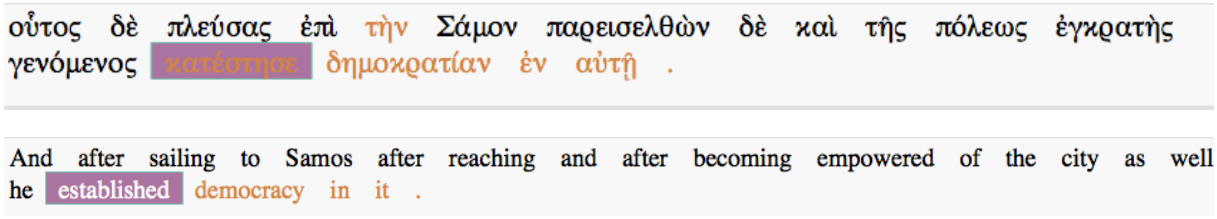


Fig. 11: One of the phases of the text-alignment (Diod. XII.27.2.1).

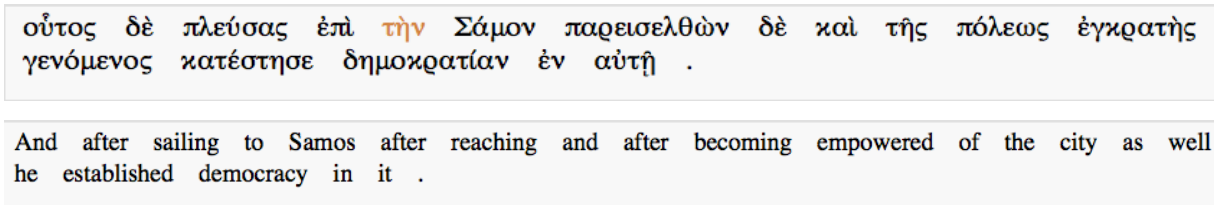


Fig. 12: Final text-alignment (Diod. XII.27.2.1).

The interface also allows to show the alignment as 'interlinear text' and the result of the alignment can be exported both in HTML (the actual display of the interface) or in an XML file (fig. 13).

Given that there are no guidelines for aligning texts, some additional notes have to be given here in order to keep the work consistent and to explain how to work with texts in different languages.[24] Even if it is possible to work with text passages containing more than 100 characters, the interface presents several sentences as one big block. In the XML file of the translation alignment, each sentence is treated separately. As part of the Digital Hill project, several Greek-English and Greek-German alignments have been created. Moreover, given that we have different ancient sources dealing with the same event concerning the revolt of Samos, we have also been producing Greek-Greek and Greek-Latin alignments.

A word-by-word alignment between ancient Greek texts and their modern translations is not possible in most cases. One reason is that word endings in ancient Greek contain information that in many modern languages is translated with personal pronomina. For example δοκοῦσιν is translated into English with 'they seem', and so a word-by-word alignment is not possible. The same happens with tense forms for verbs and with articles accompanying personal names in ancient Greek (e.g., ὁ Περικλῆς, which is only 'Pericles' in the English translation).

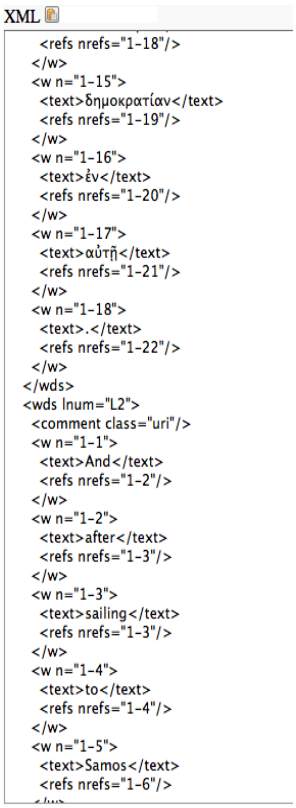


Fig. 13: Extract from the XML-view provided by Perseids.

24 One of the purposes of producing translation alignments of the sources on the revolt of Samos has been not only to try to analyse textual evidence on this historical event with digital tools, but also to provide a rich set of test cases for building in the future translation alignment guidelines.

Furthermore, it is possible that two words are contracted together in one language and not in the other (e.g., the German 'in dem' that may become 'im'). These are just very few examples, but it is important to keep them in mind when trying to create word-by-word-alignments.
The XML files resulting from the alignment and the HTML visualization in Perseids do not display punctuation except for full stops.[25] This depends on the fact that at the beginning only texts up to 100 characters could be processed by Perseids.[26] Considering the limitations resulting from not visualizing punctuation, the visualization of the alignments in the GitHub webpage of the project provides texts with punctuation.[27]
There are also grammatical differences that have to be taken into account when working with translation alignments. For example different languages may use different cases for expressing the same conditions and there are many particles in ancient Greek that cannot be translated into modern languages.
To make the problems more explicit, here is a concrete example from a passage of Arist., *Rhet.* 1411a1:[28]

> τῶν δὲ μεταφορῶν τεττάρων οὐσῶν εὐδοκιμοῦσι μάλιστα αἱ κατ᾽ ἀναλογίαν, ὥσπερ Περικλῆς ἔφη τὴν νεότητα τὴν ἀπολομένην ἐν τῷ πολέμῳ οὕτως ἠφανίσθαι ἐκ τῆς πόλεως ὥσπερ εἴ τις τὸ ἔαρ ἐκ τοῦ ἐνιαυτοῦ ἐξέλοι.

Here are three different translations of this passage:

> *- Of the four kinds of metaphor the most popular are those based on proportion. Thus, Pericles said that the youth that had perished during the war had disappeared from the State as if the year had lost its springtime.*[29]
>
> *- Of the metaphors, which are four, those about proportions seem most popular, as for example, when Pericles said, that the youth, who had been killed during the war, had been stolen from the city in this way as if someone had taken away the spring from the year.*[30]
>
> *- Und von den Metaphern, es sind vier, erscheinen die über Proportionen besonders gut, zum Beispiel sagte Perikles, dass die Jugend der im Krieg Gefallenen auf diese Weise so aus der Stadt geraubt wurde, als ob irgendwer den Frühling aus dem Jahr entfernte.*[31]

Freese's translation is pretty free, as it is possible to see in the first part of the sentence which is treated as a single isolated one. An accurate sentence alignment is not possible because we have one Greek sentence opposed to two English ones. Yet it is possible to align those sentences as the algorithm does not divide text blocks according to full stops, unlike the algorithm of Arethusa for treebanking. Both English sentences are now seen as one block. Furthermore, some words have been omitted, as for example τις, or added, as for example 'kinds' in the first part of the sentence.[32] The genitive μεταφορῶν at the beginning of the sentence is a genitive of the divided whole and for that reason contains the condition which is expressed with the preposition 'of'.[33] αἱ is treated as the subject of the sentence and the translation could be 'of the four

---

25 The semicolon may be used too, as it corresponds to a question mark in Ancient Greek.

26 This limitation does not exist anymore, at least in the instance of the alignment tool provided by the Perseids web page http://sosol.perseids.org. There is still a limitation in the instance provided by http://alpheios.net/.

27 http://digitalhill.github.io/

28 This passage is mentioned among other passages in Hill, No. 267.

29 The translation is by Freese: Freese (1926).

30 Translated by Marcel Mernitz.

31 Translated by Marcel Mernitz.

32 See http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=15904 for the alignment.

33 http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=15901 shows the treebank file of this sentence.

metaphors those that ...'. In this translation μεταφορῶν has to be aligned to two words ('kinds' and 'metaphors'). The annotation tree in Arethusa would have a different shape as well.
In the second translation, the genitive μεταφορῶν is treated as a genitive of connection, stating a remark,[34] dependent to εὐδοκιμοῦσι.[35] The participle οὐσῶν is translated as an attribute to μεταφορῶν in this translation with τεττάρων as its predicate nominal, but it could be translated as 'the metaphors, being four, about proportions seem very good' as well to almost provide a word-by-word alignment. Here κατά is aligned to 'about'. Furthermore, it is not possible to translate εὐδοκιμοῦσι as a single word, so a word-by-word alignment becomes impossible. If such an alignment were be pursued, ὥσπερ would just have to be translated as 'as'. Although in this case, the coordinating particle δέ has not been translated, as it is not needed for the right speech flow in English, it might as well be translated due to a word-by-word alignment. The Accusativus cum Infinitivo, short A.c.I., that follows ἔφη, is introduced by 'that' in the translation, yet the Greek original has no need for it. The modal adverbial οὕτως can be translated with more than one word as 'in this way', by which the Greek word would have to be aligned to three words, but also simply as 'so'.[36]
In the German translation, a word-by-word alignment is pursued, but it is evident that this is also impossible. The A.c.I. is introduced by a conjunction in the German translation as well, which is not needed in the Greek version. The predicate cannot be displayed as one word, either, and also, an adjectival translation would make no difference, as a verb would still be needed (e.g., 'sind wohlscheinend'). In addition, it does not seem desirable to switch the word type, if a word-by-word alignment is pursued. Although it is possible to express the modal verb with one word in German, it appears more prominent if the expression contains three words. The problem regarding the melted article has already been mentioned above.[37]
It is not always necessary to align the entire text block. This is especially the case with ancient texts that deal with the same topic, such as for alignments of Greek-Greek or Greek-Latin texts. In these cases, partial alignments can help to highlight the similarities. We are going to show an example aligning two extracts from the *Timotheus* of Cornelius Nepos and the *De Permutatione* of Isocrates:

> - *in quo oppido oppugnando superiore bello Athenienses mille et ducenta talenta consumpserant, id ille sine ulla publica impensa populo restituit.*[38]
> - μετὰ δὲ ταύτας τὰς πράξεις ἐπὶ Σάμον στρατεύσας ἣν Περικλῆς ἀπὸ διακοσίων νεῶν καὶ χιλίων ταλάντων κατεπολέμησε.[39]

Both passages deal with the costs of the Samian war.[40] As the aim of this alignment is to highlight similarities, the words that should be aligned are *mille* with χιλίων, *talenta* with ταλάντων, and *consumpserant* with κατεπολέμησε.[41] An interesting fact is that Nepos speaks of 1200 talents, which were spent by the Athenians for the siege, while Isocrates states that

---

34 Since our reference grammar is Smyth, the work is based on his grammar. Cfr. Smyth (1956): SG 1381 (genitive of connection).

35 http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=12115 shows the sentence tree of this variation.

36 The text alignment looks like this: http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=15903.

37 Here, the text alignment is as follows: http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=15902.

38 Corn. Nep., *Timoth*. I, 2 = Hill, No. 259.

39 Isocrat., *De Perm*. 111 = Hill, No. 253.

40 See also ML-55 = IG I3 363 = CIA I 177.

41 Here is the text-alignment: http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=12406.

they spent an amount of 1000 talents for conquering the Samians. In this case, the matter is not a translation, so a word-by-word alignment would not be too useful. Furthermore, Nepos reports that Timotheus was able to conquer Samos later without any expenses for the Athenean people. Isocrates on the other hand reports that, in addition to the expenses of 1000 talents, the Athenians maintained 200 ships for the siege. Diodorus only reports the payment of a fee of 200 talents, which occurred during the siege:

> - κολάσας δὲ τοὺς αἰτίους ἐπράξατο τοὺς Σαμίους τὰς εἰς τὴν πολιορκίαν γεγενημένας δαπάνας, τιμησάμενος αὐτὰς ταλάντων διακοσίων.[42]

In this passage it is not mentioned how the persons responsible for the riot (τοὺς αἰτίους) were punished. The payment of 200 talents is inflicted to all the Samians (τοὺς Σαμίους). In addition to this payment, Diodorus reports a penalty that the Samians had to pay during the first Athenian invasion combined with the provision of hostages. They had to provide 80 hostages and as many talents.[43] Combined with the amount they had to pay, the total is 280 talents. The difference between this amount and the amount in Nepos' report is 920 talents and the difference from Isocrates' report is 200 ships and 720 talents, that would have been spent for the remaining war. Diodorus only estimates these as 280 talents for the first invasion, in which Samos seemed not to have resisted, and the second invasion and the accompanying siege.[44] Plutarch tells of reports according to which the hostages had to pay a talent each, but he rejects those reports as propaganda. Yet Plutarch reports that a part of the penalty had to be paid immediately and the rest - which is not clarified – had to be paid by a stated time (ἐν χρόνῳ ῥητῷ). In addition, the Samians would have to provide hostages again.[45] The *Corpus Inscriptionum Atticarum* (CIA) I 177[46] states an amount up to 1404 talents paid by three Ἑλληνοταμίαι[47], while another

---

42 Diod. XII. 28, 3 = Hill, No. 238.

43 Diodorus varies here from the report of Thucydides. According to Thucydides, the Samians had to provide 100 hostages, 50 children and men at a time. For the text alignment of the two passages, see http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=15880. Diodorus also conceals the construction of a garrison, but he reports the payment of 80 talents.

44 Cfr. Legon (1972), 149. See also http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=6&doc=10318 for the treebank file of Schol. in Arist. Vesp. 283 in which a certain Carystion warned the Athenians about the Samian warcraft and earned the Athenian civil right in this way. According to this account, the Samians were resisting though.

45 See http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=11142 for the treebank file of Plut. Per. 28. A scholion, that reports the blackmail of money during the Samian campaign, is intercessional for the provision of hostages. See therefore http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=11094 for the treebank file of Schol. in Arist. Pax 697.

46 See https://github.com/DigitalHill/EpiDoc-files/blob/master/cia_i_177_epidoc.xml for the EpiDoc-file.

47 See CIA I 177 = IG I³ 363 (FR. A). There are three different amounts that are mentioned in the inscriptions: 128 talents (line 5), 368 talents (line 12) and at last 908 talents (line 17). In line 19, the total amount is up to 1400 talents, but this line is utterly mutilated and barely readable, so it depends on interpretation. The three amounts sum up to 1404 talents. Cfr. also Gabrielsen (2008), p. 46-73 and Fornara/Lewis (1979), 9–11. In Isocrates, Nepos and Diodorus, the siege represents the entire war, but they leave out the first invasion, so they ignore the first amount. During this invasion, democracy had been established and had probably been secured for another two or three months. The expenditures for that would be displayed by the first amount. Pericles landed with 40 triremes. At maintenance costs of one talent a month, they would add up to 40 talents, for three months they would add up to 120 talents. The remaining eight could have been used to secure the political change, cfr. also Fornara/Lewis (1979), 11f. According to Gabrielsen the first amount represents the expenditures for the first invasion and the following two the expenditures for the siege of Samos, so he starts his calculations for the monthly expenditures at 1276 talents and adds up to 2, 3 talents per ship a month. Cfr. Gabrielsen (1994), 115. Pritchard ignored the first invasion as well and determined the expenditures at 1276 talents: cfr. Pritchard (2012), 39. Ἑλλενοταμίαι were the Greek treasurers who organized the Athenian expenses. Given that they were in charge for one year, Fornara and Lewis saw their assumption about the duration of the war confirmed, as three different treasures are mentioned. Cfr. Pritchard (2012), 41 and Fornara/Lewis (1979), 12.

inscription mentions Ἑλλενοταμίαι who received 57 talents and 1000 drachmas of Samos.[48] Thucydides mentions only that the Samians had to pay an agreed sum within a certain time, but he does not mention the amount of this levy and whether it also contained the toll of ships.[49] Apparently Samos was not punished in another fashion as other revolting members of the League. The island was only required to provide hostages and to host an Athenian garrison. This is probably due to the fact that the Samians hardly resisted during the first invasion and in fact they could keep their exceptional position, that is the fleet, walls, and freedom from tributes.[50]

A comparison of these war expenses shows additionally that literary sources (independently from their authors and temporal distance, and topics) derive from the inscriptions.[51] This short passage clearly shows that working with text alignments is very useful to produce and answer historical questions.

We also aligned the passages of Harpocration concerning Aspasia (s. v. Ἀσπασία) and the entry about the *demopoietos* from Suidas (s. v. Δημοποίητος):

> - Harp: δοκεῖ δὲ καὶ ἐξ αὐτῆς ἐσχηκέναι ὁ Περικλῆς τὸν ὁμώνυμον αὐτῷ Περικλέα τὸν νόθον ὡς ἐμφαίνει καὶ Εὔπολις ἐν τοῖς Δήμοις.
> - Suid: ὅμως γε μὴν ἀντιβολοῦντος καὶ δεκάσαντος τοὺς ἐντεῦθεν ζῶντας ὀψὲ καὶ μόλις τὸν νόθον οἱ παῖδα τὸν ἐξ Ἀσπασίας τῆς Μιλησίας ἐποίησε δημοποίητον.

These passages are not translations of each other, thus a word-by-word alignment would not be too useful. The words that should be aligned are τὸν νόθον and τὸν νόθον with παῖδα, as well as ἐξ αὐτης ἐσχηκέναι and ἐξ Ἀσπασίας.[52] Aligning the verb ἐσχηκέναι is considered heavy alignment, since it has no equivalent in the second sentence, yet I think it needs to be aligned, since it carries the meaning that is implemented in the genitive in the passage of Suidas. This example clearly shows that it is not possible to provide a word-by-word alignment. It is also not always necessary that the cases in Greek-Greek alignments are constantly the same. An alignment of Photius and Aelian clearly points this out.

> - Photius: οἱ δὲ ὅτι Ἀθηναῖοι μὲν τοὺς ληφθέντας ἐν πολέμω Σαμίους ἔστιζον γλαυκὶ Σάμιοι δὲ τοὺς Ἀθηναίους τῇ σαμαίνῃ ὅ ἐστι πλοῖον δίκροτον ὑπὸ Πολυκράτους πρῶτον παρασκευασθὲν τοῦ Σαμίων τυράννου ὡς Λυσίμαχος ἐν β Νοστῶν·
> - Aelian: τούς γε μὴν ἁλισκομένους αἰχμαλώτους Σαμίων στίζειν κατὰ τοῦ προσώπου καὶ εἶναι τὸ στίγμα γλαῦκα καὶ τοῦτο Ἀττικὸν ψήφισμα.

The aligned words are τοὺς and τοὺς, ληφθέντας ἐν πολέμῳ and ἁλοσκομένους αἰχναλώτους, Σαμίους and Σαμίων as well as γλαυκί and γλαῦκα.[53]

Is it always possible to align a part of speech to the exact same part of speech? The answer is no, this is not always possible. Sometimes participles and adjectives have to be substantivized in translations. The following sentence should serve as an example. The Greek passage is once again taken from the speech of Isocrates mentioned above.

---

48 See CIA I 188 = IG I² 304. For the text included in an EpiDoc file see: https://github.com/DigitalHill/EpiDoc-files/blob/master/cia_i_188_epidoc.xml.

49 See Thuc. I. 117, 3.

50 Cfr. Legon (1972), 150 and 153f. for establishing a garrison. See also http://www.perseids.org/tools/arethusa/app/#/perseids?chunk=10&doc=10318 for the treebank file of Schol. in Arist. Vesp. 283.

51 This is probably due to the preference of the authors for round numbers. The inscriptions on the other hand provide solid information. Cfr. Burrer/Müller (2008), 10.

52 The text-alignment looks like this: http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=12404.

53 See http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=12550 for the text alignment.

- μετὰ δὲ ταύτας τὰς πράξεις ἐπὶ Σάμον στρατεύσας ἣν Περικλῆς ἀπὸ διακοσίων νεῶν καὶ χιλίων ταλάντων κατεπολέμησε.
- Und nach diesen Taten zog er gegen Samos, das Perikles mit zweihundert Schiffen und tausend Talenten unterwarf.[54]

These are the words that have been aligned: μετὰ - nach, δὲ - und, ταύτας τὰς - diesen, πράξεις - Taten, ἐπὶ - gegen, Σάμον - Samos, στρατεύσας - zog er, ἣν - das, Περικλῆς - Perikles, ἀπὸ - mit, διακοσίων - zweihundert, νεῶν - Schiffe, καὶ - und, χιλίων - tausend, ταλάντων - Talente, κατεπολέμησε - unterwarf. The coordinating δέ could be aligned to a conjunction, as well as the prepositions μετά and ἐπί could be aligned to prepositions and the nouns πράξεις and Σάμον, as well as the nouns of the second part of the sentence could be aligned to nouns. This is also the case for the relative pronoun ἣν, the noun that serves as subject Περικλῆς, and the numerals. The verb could be aligned to one word as well, as the subject is expressed by a word on its own. The pronoun ταύτας and the article τάς have been aligned to the pronoun 'diesen', since it already contains the demonstrative function of the pronoun. Also the participle στρατεύσας has been aligned to two words, namely a verb and a pronoun, which is already contained in the Greek word. A translation as close to the original as that one provides a text alignment with the same parts of speech. An example for a less close translation is the following one by Norlin:[55]

- After these exploits he led an expedition against Samos which Pericles reduced with a fleet of two hundred ships and the expenditure of a thousand talents.

In this case, the aligned words are as follows:[56] μετὰ - after, δὲ - /, ταύτας τὰς - these, πράξεις - exploits, ἐπὶ - against, Σάμον - Samos, στρατεύσας - he led an expedition, ἣν - which, Περικλῆς - Pericles, ἀπὸ - with, διακοσίων - two hundred, νεῶν - ships, καὶ - and, χιλίων - a thousand, ταλάντων - talents, κατεπολέμησε - reduced. This time, it was not been possible to align all of the words, so the Greek δέ has no equivalent and on the English side 'a fleet of' and 'the expenditure of' can not find any partners as well. In the translation, the numerals have been split into two words each. Aligned to the participle is the phrase „he led an expedition". Still, prepositions could be aligned to prepositions, nouns to nouns and conjunctions to conjunctions. As in the German example, ταύτας τὰς is to be aligned to a single word that contains both the demonstrative function and the function of the article. A word-by-word alignment could not be achieved in any of these examples.

3.2 Visualizing the sources on the Revolt of Samos

After producing morphosyntactic analyses and translation alignments of the sources on the revolt of Samos, the last part of the project has been devoted to the creation of a HTML-page for the visualization of the alignments of the sources (alignments between ancient languages and bilingual alignments), combining these results with treebank data.

The aim of the HTML-page is to provide a rather slim HTML-body with processing and design taking place in outsourced files. We have done that for two reasons. First, the page should be loaded very quickly and without causing too much traffic, which is achieved by sourcing out the processing files. Second, sourcing out those files enables easy recycling of the functions and designs in other files, and it also provides an easy way to apply changes to all files using

54 Translated by Marcel Mernitz.

55 The text has been taken from http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0144%3Aspeech%3D15%3Asection%3D111, last visited on 21.05.2015 at 13:41.

56 This is the link for the text-alignment: http://sosol.perseids.org/alpheios/app/align-editsentence-perseids.xhtml?s=1&numSentences=1&doc=12875.

these functions and designs. All design information is contained in one stylesheet that is encoded within the header section as well as the used jQuery scripts that grant the interaction. Each script has been programmed to start operating once the page has been completely loaded. This has been achieved by writing the script code as a function of this command between the curly brackets:

```
$(document).ready(function() {});
```
[57]

After a short introduction, the HTML document contains several <h2>-elements, one for each chapter including the preface and the table of contents. After each of those elements, there are -elements that contain the subchapters of the chapter as <h3>-elements.[58] These subchapters are divided into bilingual and ancient alignments inside one -element that has two <h4>-elements, one for the bilingual alignments and one for the ancient ones. After a short <p>-element that contains the introduction and interactive buttons, the content of the subchapter can be found inside a -element. This content is presented in various tables, one for each source mentioned by Hill without the inscriptions that are introduced by a <p>-element right before each table[59]. There are also some tables to display the results for the Greek-German alignments. The reason for this nesting is that the user should be allowed to hide content that is not of interest to her or him. When the page is loaded, all chapters, subchapters and tables are hidden. The -elements contain 'id'-attributes as well as the <h2>-elements and all the tables, whereas the <h3>-, <h4>- and the <p>-elements which appear before each table contain 'class'-attributes. Those attributes are used to hide or show their content using jQuery scripts. In most scripts, whenever an id or a class is used, they are stored in a variable which is referenced to further steps of the scripts. Both the name of the variable and the name of the id or class are arbitrary, yet they need to be consistent for the script to work.
To illustrate that the buttons are interactive, the cursor changes its appearance when the user hovers the mouse over it. This is achieved by a jQuery file by the command:

```
$('#chap3 .rosmore .rosamore').css('cursor', 'pointer');
```
[60]
before the lines that have been written for the remaining processing.

Most of the words within the tables have been marked up by enclosing them with -tags. These tags contain the attributes 'class' and 'title'. According to their class, the words would get a new color when the user clicks on one of the three buttons.
It does not matter, if the user first chooses a passage and then changes the appearance of the aligned words by clicking on one of the pencil-buttons or vice versa. These words are colored according to their color in Arethusa. Changing the colors can be achieved by adding or removing a class to the marked up words. The user may mark up all the words or only nouns or verbs. Since it is possible to click the buttons before selecting a passage, a note is written in the console to check if the buttons are operational. That note will not be noticed by most users.

When clicking on a passage from the list, the passage is presented twice, once in Ancient Greek or Latin and once in English or German. Above the text, an abbreviation may be found to

57 The programming code is written to be as easy to read as possible by humans.

58 By the time of the writing of this article, the only content available is in chapter 3.

59 Instead of tables -blocks could have been used as well and would have operated in a similar way.

60 There is a similar line for the pen-buttons in another script. The first bracket contains one id and 2 classes, using '#' for the id and '.' for the classes.

specify the language in the column's headline.[61] Instead of the passage's name, the list provides the text 'undo selection'. Once you click on this button, the passage is hidden again and the name of the passage is displayed again in the list. How does this work? Each of the tables is cached in a variable according to its own unique identification string, similar to the -elements earlier. From then on, the tables are only referenced by this variable. This table is then immediately hidden using the hide()-function. The selector of each passage is marked up in the HTML file as a link and thus written inside an <a>-tag, that has its own class. This class bears no further information, yet it allows the jQuery file to work with each tag in a different way. Depending on the link and thus on the referring class, another passage is shown. If the class is being hidden, it is slowly faded in, if it is displayed it slowly fades out by the command 'fadeToggle('slow')'. The name of the passage in the list is cached in the variable $link. Then an if-else-loop checks the content of the link and changes the text either to 'undo selection', if the passage was hidden, or back to the name of the passage, if the passage was shown. To prevent reloading the page and thus jumping to its top and causing traffic, the standard function of the link was disabled with this command:

```
return false;
```

It is possible to select several passages.

Clicking on the button to mark up the words, every passage is marked up. Sometimes several words of the same part of speech are standing next to each other. Because of that, it is not always clear on the first sight which word of the translation is aligned to which word in the original passage. For that reason, another jQuery file has been written and implemented into the head section that adds the class, that is called 'highlight', to the aligned words. This class adds a new background color to the words when the user hovers the cursor over it. It has been necessary to add a title to the span that surrounds each aligned word for this jQuery file to work. Hovering above a span will first cache the spans 'title'-attribute. Afterwards if-loops check this value and add the class to the respective -tags. It does not matter which side of the table is hovered over, because the programme is functional on both sides. In this way, all other spans with the same title would be highlighted. Some words in the translation have been titled the same, so all of them would be highlighted when hovered over. The problem with this code is, that words in other passages would be highlighted as well, which would only cause confusion. Thus, it should not be aimed at. For this cause, another variable had to be added. The programme runs through the document object model and looks for the ancestors of the span, that are called 'table' according to this code:

```
var $elementTable = $(this).parents('table');
```

The next step has been to check if this code works, and so we have written an if-loop that should write the 'id' of the table-tag and the 'title' of the span-tag into the console. Then, in the console two lines containing the values of the 'title' of the span and the unique identification string of the table are shown. Since this id is unique, no span from another table will be highlighted, once the user hovers above a span with the same title. The following lines of source code are responsible for the programme to do what it is supposed to do. Depending from the 'title'-attribute, all spans with the same 'title' are searched within the table and then they receive the class 'highlight' as long as the cursor hovers over the word. As mentioned above, the semantic layer could not be applied to articles, and thus, sometimes nouns have been aligned

61 In an earlier version of this page, the passage and its translation had been loaded into the HTML document into two -blocks. Unfortunately, only one passage can be selected at a time, and once another passage has been chosen, the former selection is undone.

to several words. After this has been done, the remaining layout of the page has been adapted according to the other pages of the fragmentary texts.[62]
For all the sources mentioned by Hill, EpiDoc files have been created. After the XML-declaration, a processing instruction that introduces a schema and the embracing <TEI>-tag which contains the namespace as its attribute, all files contain a TEI header that provides information about the title of the project and its contributor in the title statement, about the license, the publisher and the filename in the publication statement, about Hill's work and the edition he used in the two <bibl>-tags of the source description and finally some information about the file itself in the encoding description. If possible, CTS-links have been added as attributes to <author>- and <title>-tags of the second <bibl>-tag. The header follows the TEI guidelines.[63] After this header, the text of the source has been written in a simple XML file. Whenever Hill leaves out text, this text has been written in the file, but marked as a comment with a specific tag. If the XML file already existed, it has been copied without any changes to its tagset. The optional <profileDesc>-, <xenoData>- and <recisionDesc>-elements have been left out.

During the creation of the files for the inscriptions, we dealt with issues concerning Greek lowercase letters that are used in modern editions instead of uppercase ones that appear in the original epigraphic sources. Moreover, editions used aspiration and stress marks that don't occur in the original inscriptions, except for the rough breathing which is displayed either as H or later as Ⱶ.[64] The inscriptions mentioned here belong to the type of inscription that is called στοιχηδόν. A special feature of these inscriptions is that letters are written in vertical and horizontal lines next to or above each other. There are no spaces between single words, and the stonecutters did not care for union of words or syllables, thus it may occur that words are continued in the following line. In this case, the separation of words is not always a simple task, especially if the inscription is mutilated.[65]
In modern editions, inscriptions are conventionally transcribed in lowercase letters with accents, which is a problem for our EpiDoc files. An example for that may be the word οἷς, that is written in the inscription as HOIΣ, in Hill as ʽοῖς, and 'usually' as οἷς. This problem was solved in the EpiDoc files using a <choice>-tag. We used lowercase letters for inscriptions following the method of Hill. Such a tag looks as follows:

```
<choice>
        <orig>&#x0371;οις</orig>[66]
        <reg source="#hill">ʽοῖς</reg>
        <reg resp="#berti">οἷς</reg>[67]
</choice>.[68]
```

Yet these tags have only been used for articles and for words containing aspiration marks or accents, for which no ASCII code is available, as for example, an omicron or an epsilon with a circumflex accent. Aspiration marks were applied to words that start with a rho, as well, as

62 The code of all the scripts can be seen here: https://github.com/DigitalHill/digitalhill.github.io/tree/master/javascripts.

63 These guidelines are available at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

64 This is already a kind of interpretation.

65 The word στοιχηδόν is a scientific term. It is not known how ancient Greeks called those inscriptions, or if they had a particular word for them at all: cfr. Klaffenbach (1957), 48.

66 The character string ͱ generates the symbol „ Ⱶ".

67 The name given in exclamation marks shows the person who worked as editor.

68 The example shows the tag that was used finally.

they demand a rough breathing, which were omitted by Hill always. Yet, these changes have only been applied on words, where those characteristics occur.

Most of the inscriptions are quota-lists. They contain the fees and tributes of the members of the Delian League in amounts of drachmas and obols. Due to those lists, it is obvious that there are difficulties to display those combined amounts. One of the problems is that the symbol that acts as the unit for 'drachma' also contains the worth of one drachma amidst the amount. An amount of 36 drachmas and four obols is displayed in the inscription as ΔΔΔΓ𐅄IIII. At first, it has been considered if it is useful to treat the amounts separately, thus at first, the amount of drachmas would be gathered in an own <num>-tag, and then, the amount of obols would be gathered in an own <num>-tag as well. The result of this version for the given example would be as follows:

<num value="36">ΔΔΔΓ<g type="drachma">𐅄</g></num>
<num value="4">IIII</num>.

Another concern has been to display the amount as a value. As result for the example the following would be expected:

<measure type="currency" unit="drachma" value="36.67">.

Yet, none of the versions is very satisfying. Thus a combination of both versions has been applied, as the value should not be treated separately and the symbol that indicates the unit should be marked. The solution for the example above looks as follows:

<measure type="currency" unit="drachma" value="36.67">ΔΔΔΓ
<g type="drachma">𐅄</g>IIII</measure>.

In two instances, the numbers used by Hill could not be confirmed. These instances are CIA I 177, line 19 and CIA I 240 (IG I^3 272, IG I^3 279), column two, line 27. The line in the first instance (CIA I 177) is missing and, when it was double checked in the IG (= IG I^3 363) and in the edition of Meiggs and Lewis (= ML 55), it was noticed that the editors used different symbols that indicate the unit of talents. Those symbols were included as a comment in the EpiDoc file. No ASCII code could be confirmed for the symbol in the second instance.[69] The symbol resembles the Greek acrophonic Hermionian fifty:𐅅.[70]
Treebank files were created according to the passages quoted by Hill, even though he sometimes leaves out text passages or even entire sentences. Also, no treebank files have been created for inscriptions and quota-lists. The fragment of Duris has insofar an exceptional position, as it provides the text of Harpocration, which has been treated separately, and thus, no treebank file for Duris has been created. Furthermore, no EpiDoc files have been created for the passages of Thucydides, since files for these were already existing.[71]

---

69 The symbol was changed to 𐅆 in the EpiDoc file, which represents 50. I preferred that symbol to H, which represents 100, due to the context. Whenever this symbol occurs in combination with H it always follows H and stands before Δ. My interpretation is also supported by the similarity of the symbol of the acrophonic Hermionian fifty mentioned below. The amount of the tax in this instance would thus be 283 drachmas and four obols, a number that is confirmed by Larfeld (1898), 29.

70 The difference between the two symbols is that the symbol found in Hill and the CIA as well has a closing line on its top. The unicode number for the acrophonic Hermionian fifty is 10168.

71 For the complete list of EpiDoc files see https://github.com/DigitalHill/EpiDoc-files.

4. Conclusion

Producing morphosyntactic annotations and translation alignments of literary sources is a good exercise for achieving different results, such as detecting recurring syntactic features and textual reuses in ancient sources, exploring and highlighting the vocabulary concerning a specific historical event (in this case the revolt of Samos), and providing users with different translations of the same terms and expressions in ancient sources and in modern editions. In this case, the alignment of inscriptions is less useful – unless one wants to show similar inscription patterns – and, for that reason, no alignments of inscriptions were produced.[72]

As mentioned before, in many cases print and editorial reasons obliged Hill to shorten the text of the sources he quotes. We adopted a different criterion, given that we worked in a digital environment and we decided to reproduce the complete text of the sources with links to the whole works.

The Digital Hill is an ongoing project and the aim is to extend the work to other chapters of the book and to add more digital resources addressing computational and textual issues. The final goal is to provide users with a sort of companion to the book with external digital resources and visualization tools for many different possible linguistic, historical, and computational outputs.

72 The EAGLE project provides an interesting approach on aligning inscriptions. See http://www.eagle-network.eu/ for further information about the project.

5. Bibliography

5.1 Sources

G. F. Hill, Sources for Greek History between the Persian and Peloponnesian Wars, Oxford 1897.

Russell Meiggs u. David Lewis (Hgg.), A Selection of Greek Historical Inscriptions to the end of the fifth century BC, Oxford 1988.

5.2 Literature

Berti (2013): M. Berti, "Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres", Ancient Society 43, 269–288.

Burrer/Müller (2008): Friedrich Burrer u. Holger Müller, "Einleitung", in: Friedrich Burrer u. Holger Müller (Hgg.), Kriegskosten und Kriegsfinanzierung in der Antike, Darmstadt, S. 9–18.

Freese (1926): H. Freese, "Aristotle", http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0060%3Abook%3D3%3Achapter%3D10%3Asection%3D7 (Stand 18.05.2016).

Fornara/Lewis (1979): Charles W. Fornara u. D. M. Lewis, "On the Chronology of the Samian War", The Journal of Hellenic Studies Vol. 99, 7–19.

Gabrielsen (2008): Vincent Gabrielsen, "Die Kosten der athenischen Flotte in klassischer Zeit", in: Friedrich Burrer u. Holger Müller (Hgg.), Kriegskosten und Kriegsfinanzierung in der Antike, Darmstadt, 46–73.

Gabrielsen (1994): Vincent Gabrielsen, Financing the Athenian fleet: public taxation and social relations, Baltimore (u. a.).

Hill (1897): G. F. Hill, Sources for Greek History between the Persian and Peloponnesian Wars, Oxford.

Klaffenbach (1957): Günther Klaffenbach, Griechische Epigraphik, Göttingen.

Larfeld (1898): Wilhelm Larfeld, Handbuch der griechischen Epigraphik. Zweiter Band. Die attischen Inschriften. Erste Hälfte, Leipzig.

Legon (1972): Ronald P. Legon, "Samos in the Delian League", Historia. Zeitschrift für alte Geschichte. Bd. 21 Heft 2, 145–158.

Martin/Berti (forthcoming): T. R. Martin u. M. Berti, "Open Greek and Latin Data for the Challenges of the Fragmentary State of the Primary Sources for the Pentekontaetia", Mouseion (forthcoming).

Meiggs/Andrewes (1951): R. Meiggs u. A. Andrewes (Hgg.), Sources for Greek History Between the Persian and Peloponnesian Wars. Collected and Arranged by G.F. Hill, Oxford.

Phillips (2010): David J. Phillips, "Thucydides 1.99: Tribute and Revolts in the Athenian Empire", ASCS 31. Proceedings: classics.uwa.edu.au/ascs31, 1–14.

Pritchard (2012): David M. Pritchard, "Costing Festivals and War: Spending Priorities of the Athenian Democracy", Historia. Zeitschrift für alte Geschichte. Bd. 61 Heft 1, 18 – 65.

Smyth (1956): Herbert Weir Smyth, "Greek Grammar", Cambridge, http://www.perseus.tufts.edu/hopper/text?doc=Smyth+grammar+1&fromdoc=Perseus%3Atext%3A1999.04.0007 (Stand 27.11.2016).

Autorenkontakt[73]

Marcel Mernitz M.A.

Universität Leipzig
Institut für Informatik
Digital Humanities
Augustusplatz 10
04109 Leipzig

Email: mernitz@informatik.uni-leipzig.de

73 Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.

DIGITAL CLASSICS ONLINE

An Automated Approach to Syntax-based Analysis of Classical Latin

Anjalie Field

Abstract: The goal of this study is to present an automated method for analyzing the style of Latin authors. Many of the common automated methods in stylistic analysis are based on lexical measures, which do not work well with Latin because of the language's high degree of inflection and free word order. In contrast, this study focuses on analysis at a syntax level by examining two constructions, the ablative absolute and the *cum* clause. These constructions are often interchangeable, which suggests an author's choice of construction is typically more stylistic than functional. We first identified these constructions in hand-annotated texts. Next we developed a method for identifying the constructions in unannotated texts, using probabilistic morphological tagging. Our methods identified constructions with enough accuracy to distinguish among different genres and different authors. In particular, we were able to determine which book of Caesar's *Commentarii de Bello Gallico* was not written by Caesar. Furthermore, the usage of ablative absolutes and cum clauses observed in this study is consistent with the usage scholars have observed when analyzing these texts by hand. The proposed methods for an automatic syntax-based analysis are shown to be valuable for the study of classical literature.

1. Introduction

Over the past 50 years, computational methods have become an essential tool for analyzing the style and authorship of texts. Common problems in authorship analysis include plagiarism detection, author profiling, detection of stylistic inconsistencies, and authorship verification.[1] Style and authorship studies are particularly applicable to classical texts, which often involve authorship controversies. Furthermore, since texts from the Greek and Roman era have barely survived, it is in a classicist's best interest to glean as much information as possible from each text. Additionally, classical texts are works that scholars actually care about, as they contribute to scholarship rather than transient interest. Many authorship studies focus on modern documents like newspaper articles, even though most scholars care more about the use of literary devices in Vergil than in the Wall Street Journal. Finally, the volume of electronically available classical texts, especially in Latin, far surpasses the hand-analysis abilities of a small community of scholars.[2]

One of the central ideas behind stylistic and authorship analysis is that certain features of writing are unique to an author, so that even across different genres, texts by the same author will have certain similarities.[3] Most analyses seek to distinguish an author's style by choosing a set of features and using classifiers or distance functions to compare various texts. The traditional

1 Stamatatos (2009).

2 Bamman / Crane (2006).

3 Diederich et al. (2003).

feature sets tend to be at a word or character level, such as the frequency of function words, n-grams of words, characters, or parts of speech, i.e. how often the phrase „if you give," the letters „ify," or the combination „conjunction pronoun verb" occur.[4] These traditional feature sets fail to work well with Latin, primarily because Latin is a highly inflected language with a very free word order. In English, function words, such as prepositions, are used to convey the relation among words, but in Latin the relation among words is often expressed by the form of each word rather than by a function word, so measuring the frequency of function words is far less informative.

Furthermore, the prevalence of inflection in Latin makes metrics like counting the most common words difficult. In order to count word frequencies, it is necessary to lemmatize each word in the text, i.e. associate the words „walking" and „walked" as forms of the same word, „to walk." Because of the many overlapping forms of words, it can be difficult to determine the stem of a word without completely parsing the text. Attempts to parse Latin have been successful on a limited scale: Covington and Koch have both proposed methods for parsing Latin that focus on a small subset of the language, Passarotti reports a high rate of accuracy for dependency parsing medieval Latin, and Koster presents a rule-based method capable of parsing simple sentences.[5] However, there does not yet exist a parser capable of handling the complicated syntax of classical Latin with high accuracy. Because parsing classical Latin remains a non-trivial task, precise lemmatization also remains difficult. Finally, the high degree of inflection in Latin allows for very free word order. Some loose conventions exist, such as placing the subject of a sentence near the beginning and the verb at the end, but in general Latin words can occur almost anywhere in a sentence. Not only does this make parsing more difficult, it also makes metrics like n-gram frequencies less meaningful.

Because of the difficultly of applying lexically based methods to stylistic analysis, one of the goals of this study was to analyze authors' style at a syntax level. Latin has the ability to express the same idea in many ways by using various types of grammar constructions. Purpose can be expressed using a purpose clause, a relative clause of purpose, a gerund, a gerundive, or a supine.[6] Lexical measures in Latin often fail to represent syntax and grammar, like the difference between a gerund and a supine, which can be strong indicators of style. Additionally, a syntax-based approach has more potential for cross-language analysis than lexically based methods. For example, A.D. Leeman suggests that the Roman historian Sallust was greatly influenced by the Greek historian Thucydides.[7] Jonas Grethlein questions this comparison and instead suggests that Sallust's writing contains elements of Herodotus.[8] Since many grammar constructions exist in both Latin and Ancient Greek, and some commonalities are easy to identify across languages, a syntax-based approach offers a way to quantify these cross-language comparisons. In general, a syntax-level analysis more closely represents how a classicist might approach reading a text, by paying attention to the use of grammar constructions.

Our paper specifically focuses on two such grammar constructions: the ablative absolute and the *cum* clause. The first part of the study involved developing methods for identifying ablative absolutes and *cum* clauses in texts. Since parsing in Latin is still a difficult problem, we

4 Mosteller / Wallace (1964), Stamatatos (2009).

5 Covington (1990), Koch (1994), Passarotti / Dell'Orletta (2010), Koster (2005).

6 Moreland / Fleischer (1990).

7 Leeman (1963).

8 Grethlein (2006).

aimed to show that it is possible to analyze grammar without full scale parsing. Instead, the two constructions were identified by combining part-of-speech tags with a rule-based approach. Furthermore, unlike previous syntax-based analyses, our methods were designed to work on texts without any annotations. The second part of this study involved finding ablative absolutes and *cum* clauses in a variety of texts and looking for patterns in usage across authors and genres.

2. The Ablative Absolute and the *Cum* Clause

The ablative absolute usually consists of a participle, often a perfect passive participle, and a noun in the ablative case. The construction may not have a participle, consisting simply of a noun or an adjective in the ablative, and it may contain additional words, such as objects, adjectives or other qualifiers. It is also grammatically independent from the rest of the sentence, with a different subject and not directly referring to any words in the rest of the sentence. Hence it is „absolute." An example is: *his responsis ad Caesarem relatis, iterum ad eum Caesar legatos cum his mandatis mittit,* which translates: "When these answers were reported to Caesar, he sends ambassadors to him a second time with this message" (Caesar, Commentarii de Bello Gallico, 2.5, Translator W. S. Bohn).

The ablative absolute is typically used to provide background or contextual information. It can express time, condition, opposition, cause, or attendant circumstance, and is often best translated with „when", „since," or „although."[9] The construction is especially common in military and historical accounts, because it allows the author to convey information concisely.[10]

The *cum* clause is also often used adverbially to provide background information. It consists of the conjunction *cum* with a verb in either the indicative or the subjunctive. With the indicative, it is almost always temporal, translated as „when". With the subjunctive, it can also be causal or concessive, translated as „since" or „although."[11] Thus, the *cum* clause and the ablative absolute are both adverbial clauses, used to express contextual information, and in many cases, are somewhat interchangeable.[12] There are other types of phrases used to express contextual information, and there are situations in which ablative absolutes and *cum* clauses are not interchangeable. However, authors generally use these constructions in similar ways.
The similarities between these two constructions suggest the following hypotheses:

1. An author's decision to use a *cum* clause or an ablative absolute is often more stylistic than functional, so the relative frequency of cum clauses and ablative absolutes in an author's work is indicative of the author's style.

2. The relative frequencies of *cum* clauses and ablative absolutes are similar for a given author, with some consistency even across genres, but vary significantly across different authors.

3. Authors writing in the same genre use more similar distributions of ablative absolutes and *cum* clauses than authors writing in different genres.

---

9 Bennett (1918).

10 Leeman (1963); Von Albrecht (1979).

11 Bennett (1918).

12 Moreland / Fleischer (1990).

3. Methodology

3.1 Description of Data Sets

Parts of this study relied on the use of the Latin Dependency Treebank (LDT), a corpus of syntactically tagged Latin sentences.[13] Table 1 describes the texts included in the annotated data set. The total number of tokens in this data set is 53,143 (48,521 excluding punctuation). Each word in the corpus has been hand-annotated with morphological and part-of-speech information. Each sentence has been further parsed according to a dependency grammar.

Perseus ID	Author	Title	Word Count	Time Period
1999.02.0002	Caesar	Commentarii de Bello Gallico	1,383	1st century B.C.
1999.02.0010	Cicero	In Catilinam	5,582	1st century B.C.
1999.02.0060	Jerome	Vulgata	8,382	5th century A.D.
1999.02.0055	Vergil	Aeneid	2,311	1st century B.C.
1999.02.0029	Ovid	Metamorphoses	4,285	1st century B.C.
2007.01.0001	Petronius	Satyricon	11,247	1st century A.D.
1999.02.0066	Propertius	Elegies	4,395	1st century B.C.
2008.01.0002	Sallust	Bellum Catilinae	10,936	1st century B.C

Table 1: Description of annotated data set

Perseus ID	Author	Title	Word Count	Time Period
1999.02.0002	Caesar	Commentarii de Bello Gallico	51,305	1st century B.C.
1999.02.0010	Cicero	In Catilinam	12,605	1st century B.C.
1999.02.0120	Cicero	De Oratore	61,570	1st century B.C.
1999.02.0077	Tactius	Annales	88,412	1st century A.D.
2007.01.0014-6	Seneca	De Clementia; De Ira; De Brevitate Vitae	37,032	1st century A.D.
2008.01.0002	Sallust	Bellum Catilinae	10,936	1st century B.C

Table 2: Description of unannotated data set

13 Bamman et al. (2007), http://nlp.perseus.tufts.edu/syntax/treebank/.

For this study, we also compiled a data set of unannotated texts, which are summarized in Table 2. The data set consists of classical prose, and we specifically chose it to represent a range of authors and genres. Notably, the data set includes two different works by Cicero and works by several historians. All texts were obtained from the Perseus Project.[14]

3.2 Identification of Ablative Absolutes and *Cum* Clauses in Hand-Annotated Data

We developed rules for identifying ablative absolutes and *cum* clauses in the annotated texts that target how the texts were annotated. *Cum* clauses are relatively straightforward constructions, consisting of simply the word *cum* functioning as a conjunction with a finite verb in either the indicative or the subjunctive. However, the word *cum* can be used as either a conjunction or a preposition in Latin. In the hand-annotated data, a *cum* signifying a *cum* clause can be easily distinguished from cum the preposition, because words are tagged with their part-of-speech. Thus, according to our rules, any instance of the word *cum*, where *cum* was tagged as a conjunction, was counted as a *cum* clause. The number of *cum* clauses in each text was counted using a python script.

In contrast, the ablative absolute is a more ambiguous construction that can take various forms. While the basic form involves a noun and a participle in the ablative case, it is also possible to omit the participle and have simply a noun, usually with an adjective, in the ablative case. Furthermore, the participle can govern other objects or qualifiers, such as adjectives or prepositional phrases. Since this study aimed to count the number of ablative absolutes in a text, we focused on identifying the nouns and participles that signify an ablative absolute, without considering other words that might be part of the construction. The annotation guidelines describe how ablative absolutes were annotated: "the noun should be annotated as the subject of the participle, with the participle (as the head of the ablative absolute phrase) depending on the main verb as an adverbial."[15] However, the actual annotations are not so clear-cut. Cases where ablative absolutes have multiple participles and nouns are annotated differently. For example, participles do not always depend on the predicate in the sentence; they can instead depend on other words, like conjunctions.

Because the defined description of the annotation of ablative absolutes is too simplistic, we tested various restrictions to determine a set of rules that most accurately finds ablative absolutes. The testing started with a very broad definition, namely flagging any clause that contains an ablative participle, and then we added in more constraints to eliminate false positives. These trials focused on finding the most accurate system for identifying ablative absolutes with only a few rules, rather than trying to cover all possible combinations of conjunctions, subordinate clauses, and commas.
The final criteria we used to identify ablative absolutes are:

1. The phrase must contain a participle in the ablative case with an adverbial relation
2. The phrase must contain a noun with a subject relation
3. The noun must meet one of the following:

14 http://www.perseus.tufts.edu.

15 Bamman (2007).

- Depend on the participle
- Depend on a conjunction that depends on the participle (indicates two nouns, 1 participle)
- Depend on conjunction that participle also depends on (indicates 2 participles, 1 noun)

This classification excludes some constructions, notably, ablative absolutes containing no participle. However, this restriction significantly reduces the number of false positives, and it simply focuses the study on a more specific construction: an ablative absolute containing a participle.

Our search counts the number of ablative absolutes in a text by the number of participle-noun pairs, thus an ablative absolute with 1 noun and 2 particles would count as 1 ablative absolute. We implemented these rules by using a python script to search through each text.

3.3 Identification of Ablative Absolutes and *Cum* Clauses in Unannotated Data

The purpose of this study was to automate the analysis of large corpora, rather than just small hand-annotated corpora. This goal necessitated methods for identifying constructions in unannotated texts.

In order to better handle ambiguous word forms, TreeTagger was used to assign part-of-speech tags and case tags to each word in the text. This program uses decisions trees to conduct probabilistic tagging.[16] The hand-annotated data in the LDT was used as training data, since using these data resulted in higher accuracy than the provided training files. In processing a text, we first tagged the entire text for part-of-speech and for case. Then, we divided the text into clauses according to all punctuation markers, including periods, commas, semicolons, parentheses, brackets, and quotation marks. Finally, the text was searched for *cum* clauses and ablative absolutes.

Like the search for *cum* clauses in the hand-annotated texts, the search for *cum* clauses in the unannotated texts used the part-of-speech tags assigned by TreeTagger to distinguish between *cum* the preposition and *cum* the conjunction. Any clause containing the word *cum* that was tagged as a conjunction by TreeTagger was considered to be a *cum* clause.
Ablative absolutes were identified as:

1. The clause contains a word tagged as a participle and tagged as the ablative case
2. The clause contains a noun that could match in gender, number, and case with the participle

The phrasing "could match" refers to the ambiguity of Latin word forms. More specifically, for a clause containing a participle, rule 2 above is satisfied if any other word in the clause can be interpreted as noun matching in gender, number, and case with the participle, even if the word can be interpreted in a different way. Thus if a clause contains a masculine participle and a noun that could be masculine or feminine, rule 2 would be satisfied. Lemmatization of words, or the identification of their possible forms was performed using the Morpheus Engine developed by the Perseus Project. When queried, Morpheus provides all possible forms of the given word. Texts were pre-processed by querying Morpheus for all words in the text and sto-

16 Schmid (1994).

ring lemmatization information in a local database, which was then used to lemmatize words in a text while searching for ablative absolutes.

We automated tagging with TreeTagger and implementation of the rules using python scripts. Each punctuation-separated clause was counted at most once, even if it contained multiple ablative absolutes.

4. Results

4.1 Identification of Syntactic Constructions in Hand-Annotated Texts

Figures 1 and 2 show the frequencies of ablative absolutes and *cum* clauses for each author in the hand-annotated corpus. For texts where fewer than 20 clauses were identified as containing the given syntactic construction, each clause was hand-checked to determine if it contained an ablative absolute or *cum* clause. For texts where more than 20 clauses were identified (Ovid, Petronius, Propertius, Sallust for ablative absolutes; Cicero, Jerome, Petronius for *cum* clauses), 1 in 5 constructions was hand-checked.

For ablative absolutes, 1 false positive was identified in Cicero, and for *cum* clauses, 1 false positive was identified in Petronius, which appears to be the result of an incorrect tag. The lack of false positives suggests that these constructions were found with high precision. Hand-checking the found constructions does not ensure that the search method had a high recall rate, but the search for *cum* clauses was very straightforward with little room for error. Additionally, the use of broad search criteria when identifying ablative absolutes, which relied primarily on the definition of the construction, suggests that this method was very inclusive with few false negatives.

While *cum* clauses are unmistakable, it can be ambiguous whether or not a construction is strictly an ablative absolute. Our definition of an ablative absolute is very broad. For example, one construction identified in Ovid was: "sic aquilam penna fugiunt trepidante columbae" (Ovid, *Metamorphoses*, 1.506), which translates "thus doves flee from an eagle with a trembling wing." The phrase „trembling wing" was considered an ablative absolute. However, this phrase could also be taken as a simple ablative of description, depicting what the doves look like, or even as an ablative of means, explaining how the doves flee. The phrase „ablative absolute" describes a particular usage of the ablative case, but usages of the ablative are not always easy to classify. Thus, this study more generally focused on ablative participles that are used adverbially in a sentence.

Figure 1 shows that Caesar uses ablative absolutes most frequently, while Jerome uses no ablative absolutes and Cicero uses very few. In contrast, Figure 2 shows that Cicero uses *cum* clauses very frequently. The lowest usage rates of *cum* clauses are found in Sallust and in Vergil.

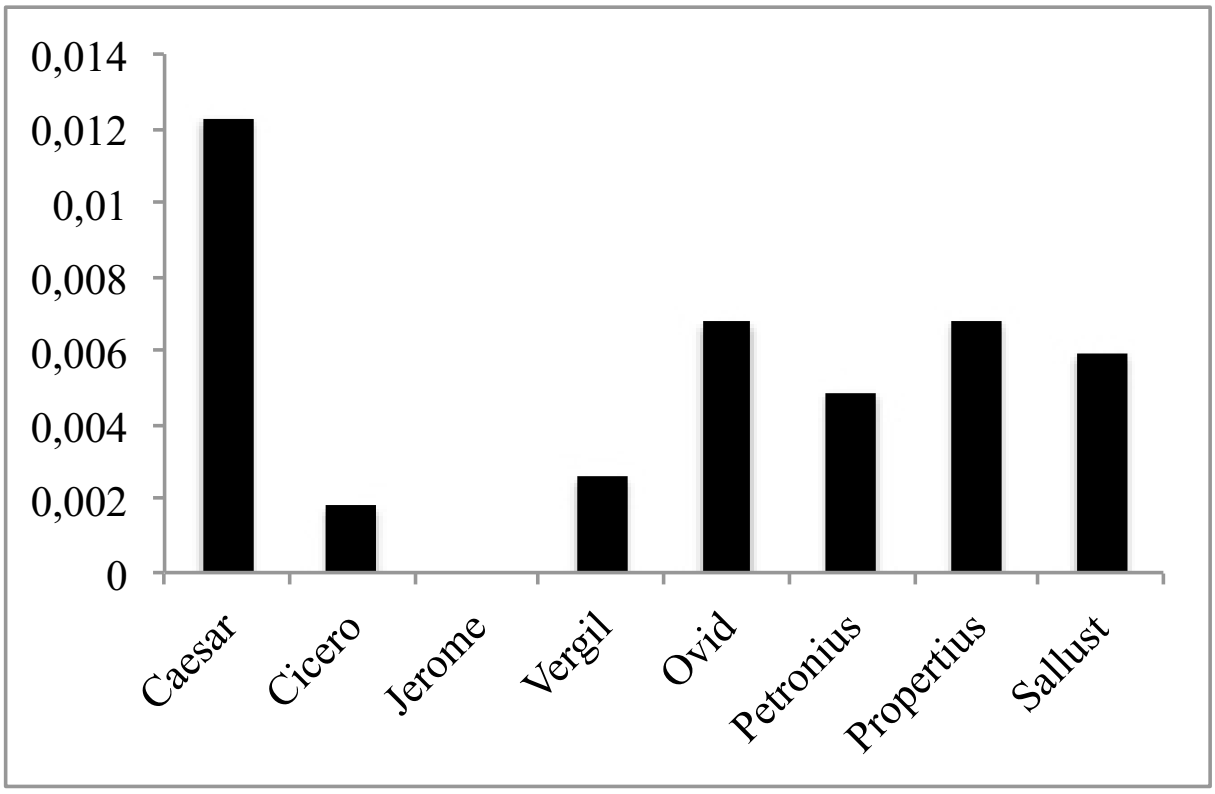


Figure 1: Rates of ablative absolutes in hand-annotated texts, expressed as the number of ablative absolutes found divided by the number of words in each text

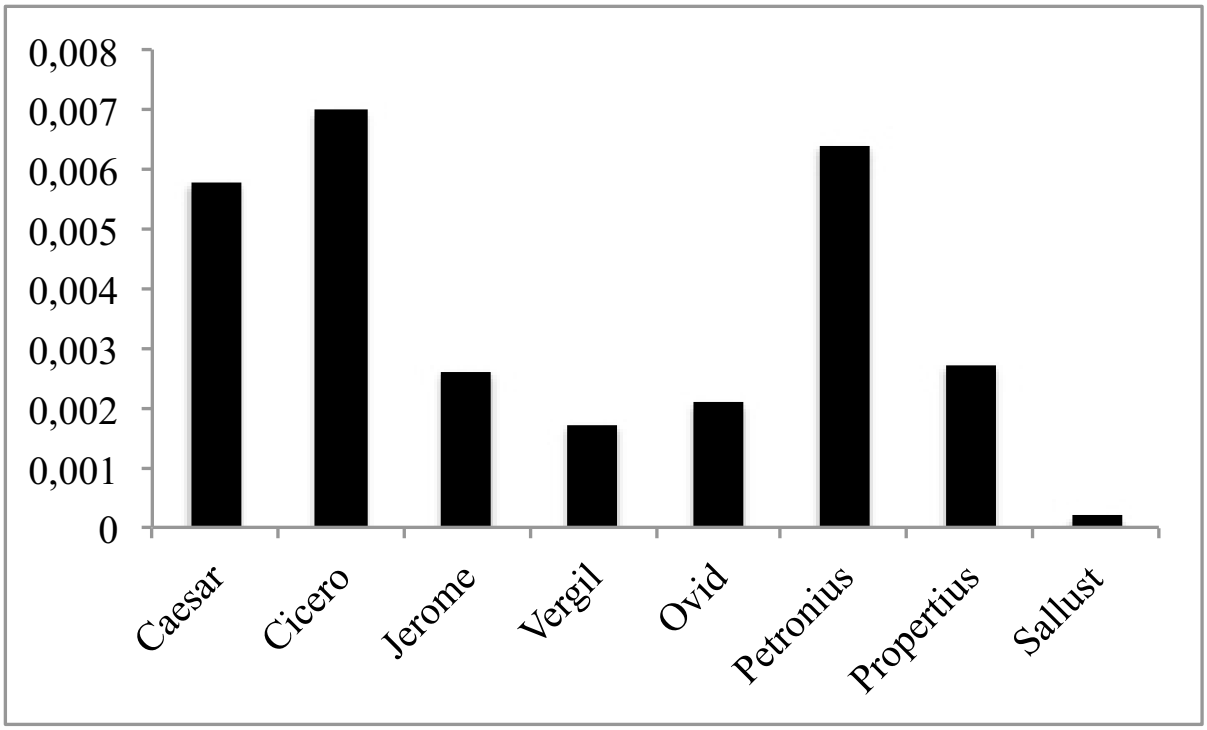


Figure 2: Rates of *cum* clauses in hand-annotated texts, expressed as the number of *cum* clauses found in each text divided by the number of words, excluding punctuation, in the text

4.2 Identification of Syntactic Constructions in Unannotated Texts

To identify the target structures in unannotated texts, we used TreeTagger to tag the texts followed by the application of rules to identify constructions. In order to measure the accuracy of this method, it was tested on the hand-annotated data, but ignoring annotations. The sentences in the hand-annotated text were randomly divided into 10 equal sections. For each of the 10 sections, 1 section was used as test data and the remaining 9 sections were used as training data. The accuracy rates are calculated as: total number of correct tags / total number of tags across all 10 tests. Figure 3 shows the accuracy rates for using TreeTagger on the hand-annotated data. In using TreeTagger for identifying ablative absolutes and *cum* clauses, only part-of-speech (95.5% accuracy) and case tags (84.0%) were considered.

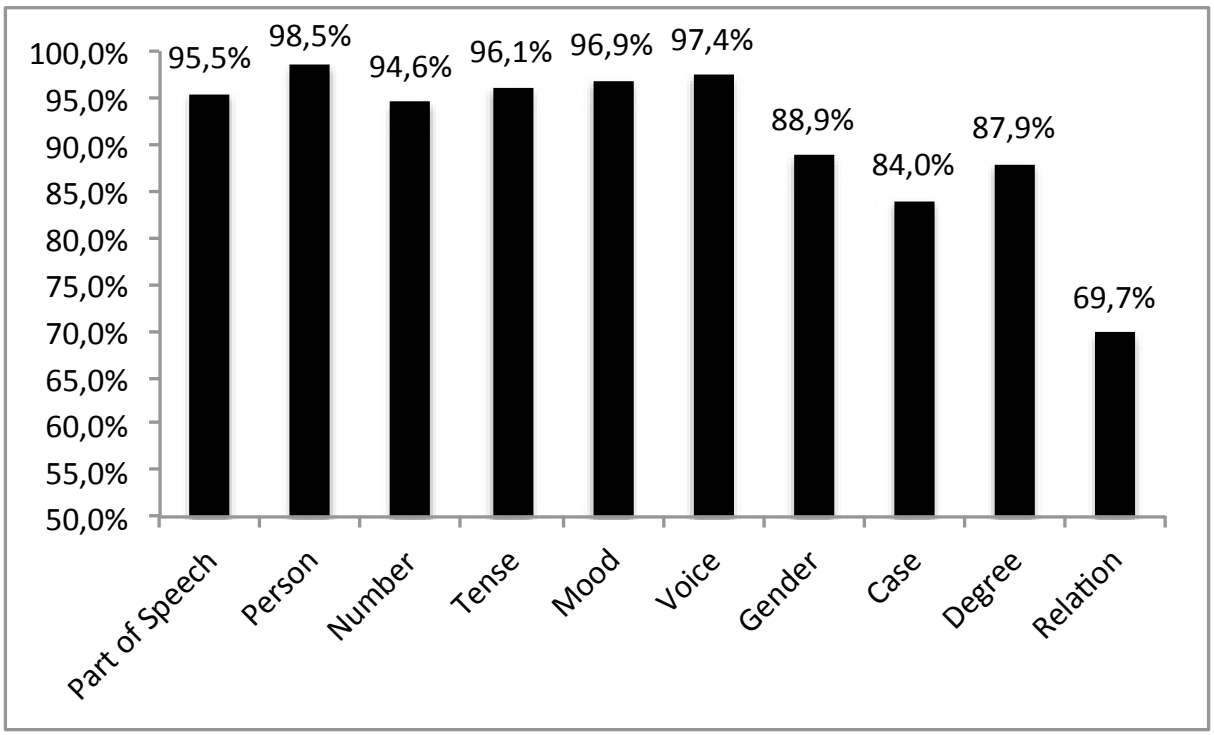


Figure 3: Accuracy of TreeTagger in tagging morphological information in the hand-annotated texts

Tables 3 and 4 show the accuracy rates for identifying syntactic constructions in the hand-annotated data, ignoring annotations. Precision was measured as (number of clauses found with and without using annotations) / (number of clauses found without using annotations). Recall was measured as (number of clauses found with and without using annotations) / (number of clauses found using annotations). Thus, precision measures how many of the clauses identified actually contained the correct constructions, while recall measures how many of the clauses containing the correct constructions were identified. F-Score, essentially a weighted average of precision and recall, is defined as: 2 * (precision * recall) / (precision + recall).

The lowest accuracy rates occur for identifying ablative absolutes in Ovid and Vergil. However, the search was able to identify *cum* clauses in Caesar with 100% accuracy. The size of the data set for Vergil is very small, with only 6 ablative absolutes and 4 *cum* clauses. Similarly, the analyzed section of Sallust only contains 2 *cum* clauses. A larger data set would likely result in more reliable estimates of precision and accuracy. Because the data set was small, the variations in the accuracy could be exaggerated. Jerome is omitted from Table 3, since this text contains no ablative absolutes.

Text	Precision	Recall	F-Score
Caesar (Commentarii de Bello Gallico)	80.00%	47.06%	59.26%
Cicero (In Catilinam)	57.14%	80.00%	66.67%
Ovid (Metamorphoses)	47.37%	62.07%	53.73%
Vergil (Aeneid)	30.00%	50.00%	37.50%
Petronius (Satyricon)	56.36%	56.36%	56.36%
Propertius (Elegies)	44.44%	66.67%	53.33%
Sallust (Bellum Catilinae)	55.56%	53.85%	54.69%

Table 3: Accuracy of identifying ablative absolutes in unannotated texts

Text	Precision	Recall	F-Score
Caesar (Commentarii de Bello Gallico)	100.00%	100.00%	100.00%
Cicero (In Catilinam)	85.71%	61.54%	71.64%
Jerome (Vulgata)	95.24%	90.91%	93.02%
Ovid (Metamorphoses)	66.67%	44.44%	53.33%
Vergil (Aeneid)	100.00%	75.00%	85.71%
Petronius (Satyricon)	95.83%	63.89%	76.67%
Propertius (Elegies)	100.00%	83.33%	90.91%
Sallust (Bellum Catilinae)	5.56%	50.00%	10.00%

Table 4: Accuracy of identifying *cum* clauses in unannotated texts

Figure 4 shows the results of counting *cum* clause and ablative absolute frequencies in a variety of texts. Frequencies were calculated as: (the number of constructions identified / the number of words in the text segment). Each point represents the frequencies in a text segment. For Caesar, Tacitus, and Cicero, each point represents a book of the specified work. For Seneca, each point represents a complete essay or a fragment of an essay (*De Brevitate Vitae*, *De Ira*, and *De Clementia*). For Sallust, the *Bellum Catilinae* was divided into 3 segments of equal length.

These frequencies are normalized for text length, but the length of each text fragment varied by author. Chapters of Cicero's *De Oratore* contain as many as 26,865 words, while fragments of

Tacitus's *Annales* contain as few as 526 (Book 5) words. In general, the texts group by author. The Cicero texts, both from *In Catilinam* and *De Oratore* are clustered in the same group, as are all of the chapters of Tacitus's *Annales*. The sections of Sallust are more varied in their frequencies of ablative absolutes, but all have very few cum clauses. Similarly, the books of Caesar's *Commentarii de Bello Gallico* vary in their frequencies of ablative absolutes, but are more consistent in their frequencies of *cum clauses*, with the exception of Book 8 (Figure 4b). Book 8 is an outlier in this data set, falling 1.83 IQRs beyond the 3rd quartile. Figure 4c, a more detailed view of the books of Tacitus's *Annales*, suggests these books are more similar than the books of Caesar's *Commentarii de Bello Gallico*.

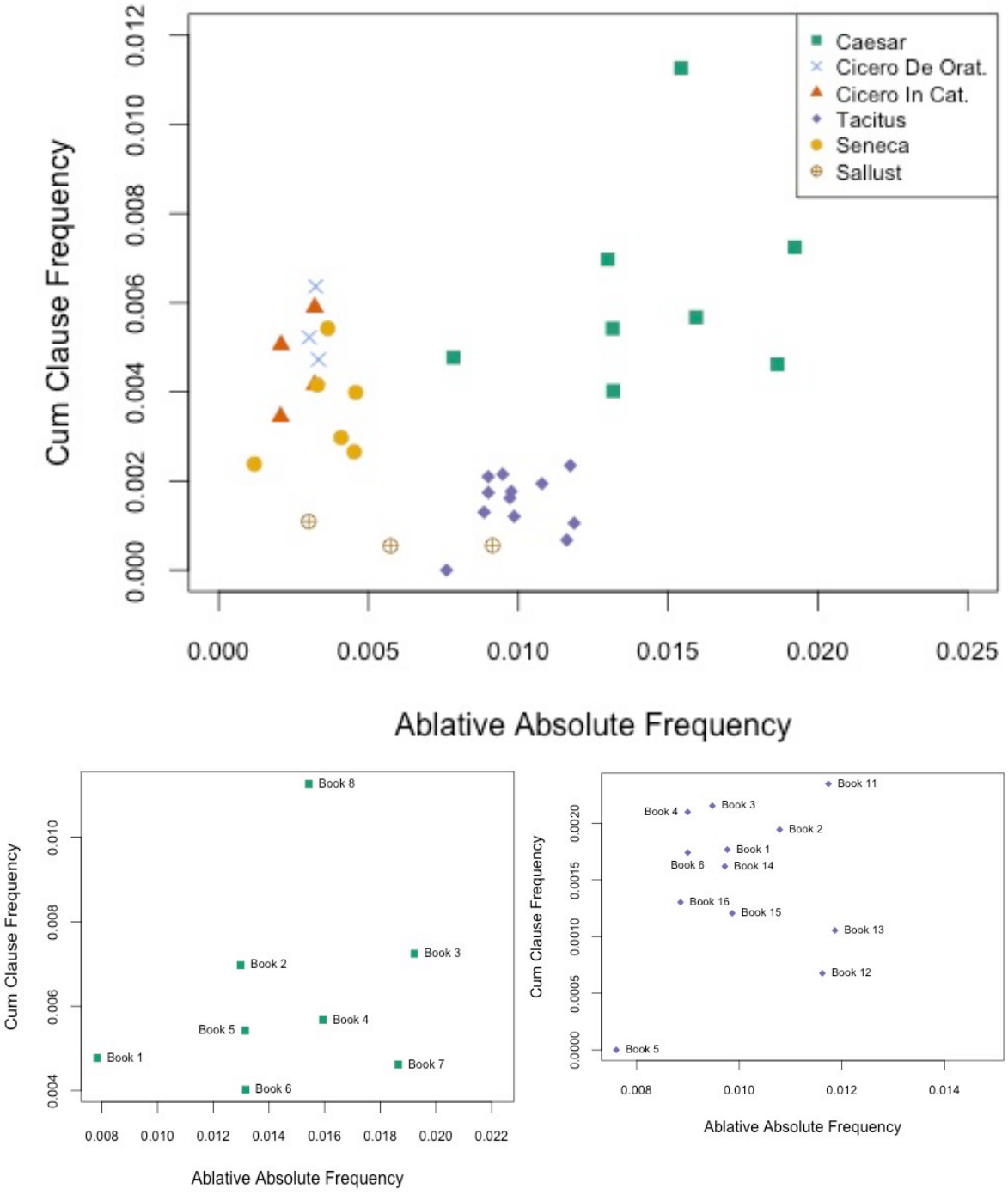


Figure 4: A comparison of *cum* clause frequencies and ablative absolute frequencies for a variety of authors; top (a), frequency rates in all authors; bottom left (b), frequency rates in Caesar; bottom right (c), frequency rates in Tacitus

Figure 5 offers a different perspective on the relationship between ablative absolutes and *cum* clauses in Caesar and Tacitus. We show what percentage of the clauses identified as either ablative absolutes or *cum* clauses were identified as ablative absolutes (i.e. number ablative absolutes identified / total number of clauses identified * 100). Figure 5a (Caesar) shows an increase in the percentage of ablative absolutes across Books 1 to 7, with a sharp decrease in Book 8. Figure 5b (Tacitus) shows no clear progression in the usage of ablative absolutes vs. *cum* clauses.

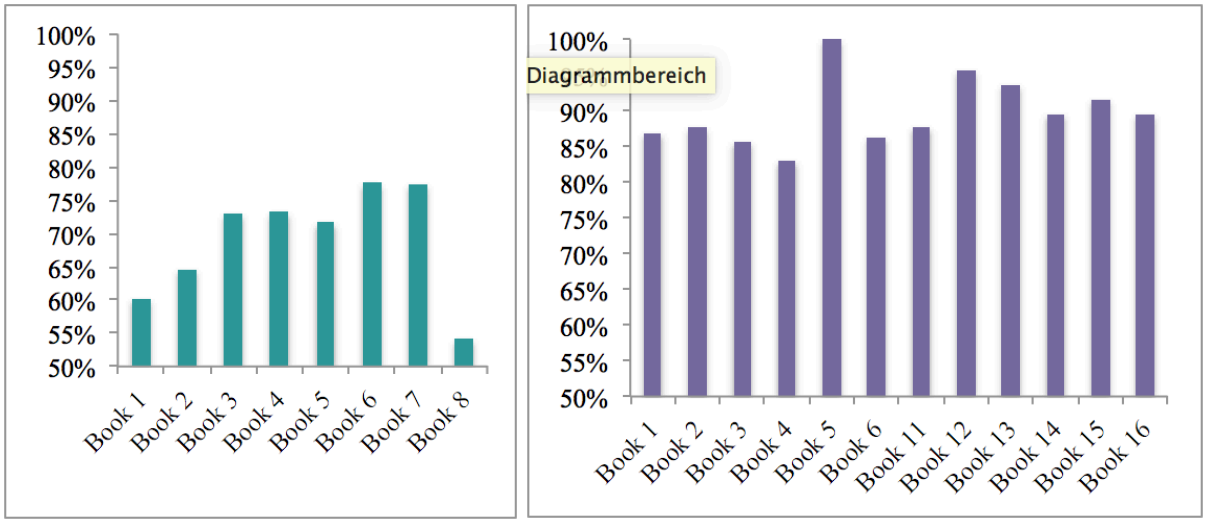


Figure 5: Percentage of ablative absolutes out of identified clauses; left, in Caesar (a); right, in Tacitus (b)

5. Discussion

In the following sections we discuss our main results.

5.1 Differentiation of Book 8 of *Commentarii de Bello Gallico*

Figures 1, 2, and 4 all demonstrate how our method for identifying syntactic constructions can distinguish between texts. In particular, Figures 4b and 4c suggest that our method identifies stylistic differences, rather than just genre or content differences. One of the classic problems in stylistic analysis, and especially in authorship attribution, is the tendency to extract features representative of the text's content, rather than of the author's style.[17] However, examining books within the same works, namely Caesar's *Commentarii de Bello Gallico* and Tacitus's *Annales*, minimizes any content differences.

Figures 4b and 5a show a clear difference between Book 8 of Caesar's *Commentarii de Bello Gallico* and the rest of the work, specifically because Book 8 contains a much higher frequency of *cum* clauses than any other book. This distinction is somewhat expected, as Book 8 was not written by Caesar. Instead, it is attributed to one of his officers, Aulus Hirtius. By counting syntactic constructions, it is possible to determine that Book 8 of the Commentarii de Bello Gallico is quite different from the other books, and even without further information, this difference raises the question of authorship. These results suggest that Aulus Hirtius has

17 Gamon (2004).

a unique style distinguishable from Caesar and possibly distinguishable from other historians of his time. Other scholars have also observed a difference between Hirtius's section of the work and Caesar's sections. Kathryn Welch notes, "Book 8 reveals proportionally more about the legates," and calls it "arguably the most boring book in the Caesarian Corpus."[18] Closer examination of Book 8 and comparison with works of disputed authorship, like the *Alexandrine, African*, and *Spanish Wars*, could help determine whether or not these disputed works were written by Aulus Hirtius, as some scholars believe.[19] In analyzing the style and language of *Bellum Alexandrinum*, Gaertner and Hausburg observe a generally heterogeneous style and also mention the usage of subordinate conjunctions, including the use of cum + subjunctive instead of *post(ea)quam* in certain sections.[20]

Although Book 8 offers the clearest example of how our syntax-based analysis can be applied to open questions of authorship and style, the trends in other books also reflect previously observed stylistic differences. Much of the scholarship on the style of Caesar's *Commentarii* focuses on the literary nature of the work. While *commentarii* in general were though to follow the same style as *annales*, consisting of a plain recording of events so that other historians could use them as a basis for more ornate works, Caesar's works show more attention to language and style than is thought to be typical of the genre.[21] In particular, some research has shown that the *Commentarii de Bello Gallico* becomes more literary over course of the work, straying further and further from the expected style. First, most scholars agree that Book 1 is considerably different from the rest of the work. J. J. Schlicher describes it as: "a book of argument as much as it is a book of war and conquest," and Kathryn Welch notes, "the legates have little or no role in its action".[22] Figures 4b and 5a show that Book 1 contains ablative absolutes the least frequently, and even strays close to the styles of Seneca and Cicero. The ablative absolute is a very succinct construction, useful primarily for narrating events concisely. The highly rhetorical nature of Book 1 can explain why the usage of ablative absolutes is so low, especially as compared with other books.

Schlicher further analyzes the development of Caesar's style, claiming that it becomes more periodic over the course of *Commentarii de Bello Gallico* as Caesar uses participles in place of subordinate clauses. Gotoff also observes how Caesar uses the ablative absolute to facilitate a periodic style, though argues that Caesar's style is reasonably consistent across the work.[23] In contrast, Eden and later Kraus agree with Schlicher, commenting on the heterogeneous style of the *Commentarii de Bello Gallico* and its inability to fit into a traditional genre.[24] Our results show the same progression of style observed by Schlicher and Eden. Specifically, Schlicher counts occurrences of subordinate clauses (temporal or circumstantial), ablative absolutes, and participial phrases, and reports the percentage of each of these 3 constructions out of the total counted clauses, comparable to Figure 5a. His counts show the percentage of ablative absolutes increases from Books 1 to 2 and Books 2 to 3, drops slightly from Books 3 to 4, drops more significantly from Books 4 to 5, and then increases for Books 6 and 7. As his analysis focuses on Caesar, he does not report counts for Book 8. Although we focus only on the ablative abso-

18 Welch (1998).

19 Daly (1951).

20 Gaertner / Hausburg (2013).

21 Eden (1962).

22 Welch (1998); Schlicher (1936).

23 Gotoff (1984).

24 Eden (1962); Kraus (2005).

lute and one type of subordinate clause, the *cum clause*, we see almost the exact same trend in Figure 5a. Furthermore, we see a large drop in the percentage of ablative absolutes from Book 7 to Book 8, showing how Hirtius's book does not fit the progression of Caesar's style. Overall, our automated method was able to identify the same increase in the usage of ablative absolutes over subordinate clauses as Schlicher's hand analysis.

5.2 Consistency of Tacitus's Syntax in *Annales*

While Figures 4 and 5 show a range in Caesar's usage of ablative absolutes and *cum* clauses, Tacitus's style remains fairly consistent across *Annales*. As with Caesar, scholars have debated the development of Tacitus's style over time and within *Annales*. Most agree that his style from his earlier works, like *Histories*, to his later works, like *Annales*, becomes more compressed, stronger, and more „Tacitean", but some authors claim that the final books of *Annales* regress to a more „Ciceroian" style.[25] F.R.D. Goodyear questions this claim, suggesting that Books 13–16 might have some unusual vocabulary, but that Tacitus's overall style and syntax remain relatively consistent. Goodyear examines some specific markers, focusing on lexical measures like frequencies of certain adjectives and prepositions, but he suggests that a closer examination of the ablative absolute might offer further insight into the consistency of Tacitus's style.[26] Figure 4c reveals little difference between the final books of Tacitus's *Annales* (Books 13–16) and the rest of the work. While Book 5 has an unusually low number of *cum* clauses, and Books 11–13 have higher numbers of ablative absolutes, Books 13-16 form no group distinguishable from the rest of the work. The unusual syntax in Book 5 likely occurs because this book has only survived as a fragment and contains less than 600 words. Thus, this fragment is too small to accurately demonstrate Tacitus's style. Overall, these data support Goodyear's claim, that there is evidence of continuous style between the final books of the *Annales* and the rest of the work.

5.3 Syntax Usage Varies Across Genres

Although, the use of syntactic constructions varies enough within *Commentarii de Bello Gallico* to distinguish features of different books, the variation greatly increases across authors and genres. The hand-annotated data, which can be considered highly accurate, shows a high frequency of ablative absolutes in the works of Caesar and much lower frequencies in the works of Vergil and Cicero. The ablative absolute is commonly associated with the style of military reports. Adams finds ablative absolutes in Plautus's parodies of such reports and notes their frequency in texts that summarize military events.[27] Caesar's *Commentarii de Bello Gallico* falls into this category, as it relates the events of the Gallic Wars, essentially a military history. Although Caesar's style varies within the work, the frequent use of the ablative absolute demonstrates Caesar's overall adherence to the normal style of military descriptions, rather than a more rhetorical or poetic style, as in Cicero's *In Catilinam* or Vergil's *Aeneid*, where ablative absolutes are scarce. While the ablative absolute's ability to convey information concisely

25 Löfstedt (1948).

26 Goodyear (1968).

27 Adams (2005).

makes it useful for military reports and historical accounts, such utilitarian language does not belong in poetry or stylized prose.
Analysis of the unannotated data presented in Figure 4 confirms the same trend; specifically Caesar and Tacitus use ablative absolutes more frequently than Cicero and Seneca. Authors like Seneca and Cicero prefer to take more time to express ideas, especially in orations, since the speaker wants to give the listener time to process information. A.D. Leeman has observed the different usage of ablative absolutes in Caesar and Cicero, and he estimates that Caesar uses about 10 times as many ablative absolutes as Cicero.[28] While the ratio in these data is closer to 5:1, the difference in usage is still clear. By counting the frequencies of ablative absolutes in a range of texts, we are able to systematically observe the usage patterns identified by other scholars and to generalize them across authors.

5.4 Variation of Syntactic Constructions Among Historians

Furthermore, our method highlights deviations from these trends. Although the frequency of ablative absolutes generally distinguishes between the plainer style of military and historical accounts and the more ornate style of philosophy and orations, Sallust, a historian, defies the pattern by using far fewer ablative absolutes than either Tacitus or Caesar. The deviation that our method detects coincides with theories about Sallust's motivations and style. Sallust wrote *Bellum Catilinae* earlier in his life than when most historians begin writing, and some have speculated that he had ulterior motives in writing the work, more than just recording history for future generations. Scholars have observed some peculiarities in his style, including his tendency to use the historical infinitive where most authors would use the imperfect tense.[29] More generally, his style is also thought to be especially poetic and paratactic.[30] Parataxis is a writing form that uses short parallel sentences, rather than nested clauses and subordination. Our results reveal Sallust's tendency to use few *cum* clauses and few ablative absolutes, which could reflect a more general avoidance of subordinate clauses as a result of a paratactic style.

When comparing Sallust with Tacitus and Caesar, who both use ablative absolutes frequently, the difference between Caesar and Sallust is not wholly unexpected. Previous scholars have observed the high frequency of ablative absolutes in Caesar as compared to Sallust.[31] However, the difference between Sallust and Tacitus is less expected, since Tacitus's style is often thought to be Sallustian.[32] The lack of similarity demonstrates that ablative absolutes and *cum* clauses are a very small subset of an author's style. Although Sallust and Tacitus differ in the use these particular constructions, they may have similarities in other aspects of their styles, such as vocabulary choice or other syntax.
We can further compare Caesar and Tacitus. There is conflicting literature on how much these authors differ in their use of ablative absolutes. Leeman observes more ablative absolutes in Caesar than in Tacitus.[33] In contrast, J.N. Adams claims that the descriptions of battles in

28 Leeman (1963).

29 Von Albrecht (1979).

30 Leeman (1963).

31 Von Albrecht (1979); Leeman (1963).

32 Goodyear (1968).

33 Leeman (1963).

Tacitus use the language of military reports, as indicated by the frequency of ablative absolutes, and even compares Tacitus's militaristic style in battle scenes to Caesar's works.[34] From Figure 4, the high frequency of ablative absolutes we find in Tacitus lends support to Adams's claim. Nevertheless, a clear distinction occurs between Tacitus and Caesar, since Caesar uses *cum* clauses much more frequently than Tacitus.

5.6 Stylistic Implications of Varied Accuracy Rates

Although our method was able to identify syntactic constructions with enough accuracy to detect patterns in usage, accuracy remains a limiting factor in analyzing the unannotated data. However, although the accuracy of identifying constructions was very low for some authors, these accuracy rates are also a reflection on the author's style. For example, the imperfect recall of *cum* clauses in Cicero in Table 4 largely occurs because of 1 particular sentence: *cum arma, cum securis, cum fascis, cum tubas, cum signa militaria, cum aquilam illam argenteam…scirem esse praemissam,* which translates: "When I knew that arms, that the axes, the fasces, and trumpets, and military standards and that silver eagle...had been sent on?" (Cicero, In Catilinam 2.6, Translator C. D. Yonge).

This sentence contains a series of 6 *cum*-noun pairings. Unsurprisingly, TreeTagger tags all 6 of these *cums* as prepositions, which seems natural, since they are all in self-contained clauses followed by nouns. However, on closer inspection, this sentence actually consists of a series of parallel clauses, in which Cicero repeats the conjunction *cum* with each noun in the sentence and omits a verb. Because applying TreeTagger to this sentence results in all 6 *cums* tagged as prepositions instead of conjunctions, this one construction greatly contributes to the error rate of identifying *cum* clauses in Cicero. The repeated conjunction translates awkwardly into English, but this type of construction is not uncommon in Latin. Cicero in particular uses such repetition frequently, and similar constructions occur throughout *In Catilinam*. The first section alone contains two examples, where Cicero repeats the word *nihil* and then the word *quid* (Cicero, In Catilinam 1.1). In this way, the accuracy of identifying syntactic constructions reflects Cicero's style just as much as the actual construction identified.

The relationship between style and accuracy of identifying syntactic constructions becomes more apparent by looking at a different stylistic element: non-projectivity. The non-projectivity rate refers to how often constituents of a phrase are broken up by other constituents. For example, Vergil writes, *Troiae qui primus ab oris* (Vergil, Aeneid, 1.1), breaking up the phrase "Trojan shore" by separating the words *Troiae* and *oris*. High rates of non-projectivity can make text analyses, such as parsing, more difficult.[35] Bamman and Crane calculate the non-projectivity rates for some of the text segments in this data set, displayed in Table 5. For reference, the non-projectivity rate in Swedish is approximately 0.94% and in Czech is approximately 1.81%.[36]

Both Jerome and Caesar write in fairly straightforward prose. They have low non-projectivity rates, and syntactic analysis was very accurate in both authors. In contrast, Cicero's *In Catilinam* has a high rate of non-projectivity, reflecting the deeply stylized nature of Roman oratory. Similarly, the poet Vergil also has a high rate of non-projectivity. It seems logical that poetry inherently involves manipulation of word order, because poets must arrange their verse to fit

34 Adams (1973).

35 Nivre / Nilsson (2005).

36 Bamman / Crane (2006).

the meter. High rates of non-projectivity have been observed in Ancient Greek poetry as well, which suggests that discontinuous constituents could be an expected feature of poetry in a free word order language.[37] The accuracy rates for identifying syntactic constructions in both Cicero and Vergil are lower than the accuracy rates for Jerome and Caesar, suggesting that non-projectivity could influence the accuracy of these methods. Specifically, a low accuracy implies high non-projectivity.

Author	Non-Projectivity Rate
Jerome	1.8%
Caesar	2.9%
Cicero	5.8%
Vergil	12.2%

Table 5: Non-projectivity rates in segments of hand-annotated texts

Although the similarities between constructions identified in the annotated data and the unannotated suggest our method does reflect true syntax usage in unannotated texts, accurate identification of syntactic constructions is not strictly necessary for distinguishing between the styles of various authors. Even if the texts written by Cicero seem to have a low rate of ablative absolutes simply because identifying ablative absolutes in Cicero is difficult, the fact the finding ablative absolutes in Cicero is difficult reflects unique elements about Cicero's style. How well syntactic analysis (specifically parsing) works on different texts has been used as a feature in authorship attribution studies.[38]

6. Related Work

The main advantage of our method for syntactic analysis is its applicability to unannotated texts without automated parsing. The concept of a syntax-based method for stylistic analysis is not new. Baayen et al. developed a method for authorship attribution that focused on syntactic rewrite rules and resulted in higher accuracy than word-based methods. However, their method was only tested on annotated texts.[39] Similarly, Bamman, Passarotti and Crane analyzed Latin syntax change over time, specifically in the shift from an *Accusativus cum Infinitivo* (ACI) construction to *quia/quod* clauses. The study was able to identify a shift in usage by examining two sets of hand-annotated data: the Latin Dependency TreeBank (LDT), consisting of classical Latin, and Index Thomisticus (IT-TB) consisting of the works of Thomas Aquinas, written about 13 centuries later.[40] That study was very similar to this one in that it focused on counting specific grammar constructions in hand-annotated data and comparing their frequencies across different time periods. However, because our method was not limited to annotated data, we were able to examine a broader range of texts, comparing constructions across different authors and genres, not just different time periods.

Other studies have examined unannotated texts but require automated parsing. Stamatatos et al. propose a computer-based method for authorship attribution that uses low-level markers, like sentence boundaries, and syntactic-level markers, like noun phrase counts. Although, this method was able to distinguish between authors of Modern Greek, which is also a highly in-

37 Mambrini / Passarotti (2013).

38 Stamatatos et. al. (2001).

39 Baaeyn et al. (1996).

40 Bamman et al. (2008).

flected language with variable word order, it requires constituent parsing.[41] Not only is Latin parsing difficult, studies involving Latin parsing typically focus on dependency parsers, in which words are linked to their immediate head, over constituent parsers, in which words are grouped in phrasal categories. Bamman and Crane used a set of 30K hand-annotated words to train a Latin dependency parser and Lee, Naradowsky, and Smith used an expanded version of this data set (53K) to train a combined approach to morphological and syntactic tagging.[42] Neither method achieved an accuracy rate greater than 65%.

However, Bamman and Crane were still able to use the tags generated by their parser to extract valuable information about selectional preferences. Breaking down the accuracy of Bamman and Crane's parser, their precision rates ranged from 34% to 68%, and their recall rates ranged from 27% to 71% for tagging relationships between words. These accuracy rates are comparable with the precision and recall rates of our method for syntactic construction identification (Tables 3 and 4).[43]

7. Conclusions

Comparison between constructions in the hand-annotated data and in the unannotated data suggests that the methods proposed in this study are accurate enough to facilitate a syntax-based analysis of classical Latin. Furthermore, the distribution of ablative absolutes and *cum* clauses identified in various authors is generally consistent with past analyses of classical Latin. This consistency confirms some the observations of scholars who examined these texts, including the frequent use of ablative absolutes in history and military accounts and the infrequent use of ablative absolutes in more ornate prose.

A more in depth analysis of specific authors, Caesar in particular, also reflects observations about these authors and contributes evidence to open debates about their style, such as how Caesar's style changes throughout the *Commentarii de Bello Gallico*. Similarly, the analysis of Caesar's *Commentarii de Bello Gallico* demonstrates that our methods can help resolve questions of authorship attribution. Our methods were able to distinguish between Book 8 of *Commentarii de Bello Gallico*, which was not written by Caesar, and the rest of the work. Overall, the consistency with manual research affirms the usefulness of these methods, indicating that they are accurate enough to contribute to the study of classical literature.

More generally, this study demonstrates that an automated syntax-based analysis of Latin is both useful and possible. Analysis of specific constructions can distinguish between the style of different authors. Furthermore, unlike traditional lexically based measures, such as word-frequencies or n-gram frequencies, this sort of analysis can target constructions that classicists are interested in studying. Automatic identification of syntax can be applied to existing literature to help answer questions that classicists have been asking for centuries.

41 Stamatatos et al. (2001).

42 Bamman / Crane (2008), Lee et al. (2011).

43 Bamman / Crane (2008).

8. References

Adams (1973): J.N. Adams, "The Vocabulary of the Speeches in Tacitus' Historical Works", Bulletin of the Institute of Classical Studies 20, 124–144.

Adams (2005): J.N. Adams, "The Bellum Africum" in: Adams, J. N., Lapidge, M., and Reinhardt, T. (Hrsg.), Aspects of the Language of Latin Prose, Oxford/New York, 73–96.

Albrecht (1979): M. V. Albrecht, Masters of Roman Prose: from Cato to Apuleius, Francis Cairns.

Baaeyn et al. (1996): H. Baayen / H. Van Halteren / F. Tweedie, "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution", Literary and Linguistic Computing 11, 121-132.

Bamman / Crane (2006): D. Bamman / G. Crane, "The Design and Use of a Latin Dependency Treebank", in: Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT, 2006), Prague, 67–78.

Bamman / Crane (2008): D. Bamman / G. Crane, "Building a Dynamic Lexicon from a Digital Library", in: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (ACM, 2008), New York, 67–78.

Bamman et al. (2007): D. Bamman / M. Passarotti / G. Crane / S. Raynaud, "Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3)", Technical report. Medford: Tufts Digital Library.

Bamman et al. (2008): D. Bamman / M. Passarotti / G. Crane, "A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin", The Prague Bulletin of Mathematical Linguistics 90, 109–122.

Bennett (1918): C.E. Bennett, A New Latin Grammar, Allyn and Bacon.
G. Celano / G. Crane / G. Almas et al, "The Ancient Greek and Latin Dependency Treebanks", https://perseusdl.github.io/treebank_data/, (Accessed 2015).

Covington (1990): M. Covington, "A Dependency Parser for Variable-Word-Order Languages", Technical Report AI-1990-01, Artificial Intelligence Programs, The University of Georgia Athens.

Daly (1951): L. Daly, "Aulus Hirtius and the Corpus Caesarianum", The Classical Weekly 44, 113–117.

Diederich et al. (2003): J. Diederich / J. Kindermann / E. Leopold / G. Paass, "Authorship Attribution with Support Vector Machines", Applied Intelligence 19, 109–123.

Eden (1962): P. T. Eden, "Caesar's Style: Inheritance versus Intelligence, Glotta 40, no. ½, 74–117.

Gaertner / Hausburg (2013): J. Gaertner / B. Hausburg, Caesar and the Bellum Alexandrinum Göttingen.

Gamon (2004): M. Gamon, "Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features", in: Proceedings of the 20th Annual Conference on Computational Linguistics (ACL, 2004), Morristown, NJ, 611–617.

Goodyear 1968: F.R.D. Goodyear, "Development of Language and Style in the Annals of Tacitus", The Journal of Roman Studies 58, 22–31.

Gotoff (1984): H.C. Gotoff, „Towards a practical criticism of Caesar's prose style", Illinois Classical Studies 9.1, 1–18.

Grethlein (2006): J. Grethlein, "The Unthucydidean Voice of Sallust", Transactions of the American Philological Association 136, 299–327.

Koch (1994): U. Koch, "The Enhancement of a Dependency Parser for Latin", Research Report AI-1993-03, Artificial Intelligence Programs, The University of Georgia Athens.

Koster (2005): C. Koster, "Constructing a parser for Latin", Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Berlin – Heidelberg, 48–59.

Kraus (2005): C. Kraus, "Hair, Hegemony, and Historiography: Caesar's Style and its Earliest Critics", Proceedings-British Academy, Vol. 129. Oxford University Press Inc.

Lee et al. (2011): J. Lee / J. Naradowsky / D. Smith, "A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing", in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (ACL, 2011), 885–894.

Leeman (1963): A.D. Leeman, Oratonis Ratio: The Stylistic Theories and Practice of the Roman Orators Historians and Philosophers, A.M. Hakkert.

Löfstedt (1948): E. Löfstedt, "On the Style of Tacitus", The Journal of Roman Studies 38, 1–8.

Mambrini / Passarotti (2013): F. Mambrini / M. Passarotti, "Non-projectivity in the Ancient Greek Dependency Treebank", in: Proceedings of the Second International Conference on Dependency Linguistics (DepLing, 2013), Prague, 177–186.

Mosteller / Wallace (1964): F. Mosteller / D. L. Wallace, Inference and Disputed Authorship: The Federalist, Addison-Wesley.

Moreland / Fleischer (1990): F.L. Moreland / R.M. Fleischer, Latin: An Intensive Course, University of California Press.

Nivre / Nilsson (2005): J. Nivre / J. Nilsson, "Pseudo-projective Dependency Parsing", in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL, 2005), Ann Arbor, Michigan, 99–106.

Passarotti / Dell'Orletta (2010): M. Passarotti / F. Dell'Orletta, "Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin", in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC, 2010), Valletta, Malta, 1964–1971.

"The Perseus Digital Library", http://www.perseus.tufts.edu/hopper/ (Accessed 2015).

Schlicher (1936): J.J. Schlicher, „The development of Caesar's narrative style", Classical Philology 31.3, 212–224.

Schmid (1994): H. Schmid, "Probabilistic Part-of-speech Tagging Using Decision Trees", in: Proceedings of the International Conference on New Methods in Language Processing, 44–49.

H. Schmid, "TreeTagger - A Part-of-speech Tagger for Many Languages", http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Accessed 2015).

Stamatatos et al. (2001): E. Stamatatos / N. Fakotakis / G. Kokkinakis, "Computer-based Authorship Attribution without Lexical Measures", Computers and the Humanities 35, 193–214.

Stamatatos (2009): E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology 60, 538–556.

Welch (1998): K. Welch, „Caesar and his officers in the Gallic War commentaries", Julius Caesar as Artful Reporter, 85–110.

Autorenkontakt[44]

Anjalie Field
Princeton University
Department of Computer Science

Email: aefield@alumni.princeton.edu

44 Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.